TESIS DOCTORAL

El fraude fiscal en España. Propuestas para su detección y análisis mediante técnicas de Big Data Analytics. Aplicación al IRPF

Autor:

César Pérez López

Directoras:

María Jesús Delgado Rodríguez
Sonia de Lucas Santos

Programa de Doctorado en Ciencias Sociales y Jurídicas

Escuela Internacional de Doctorado

2019

ÍNDICE

INTRODUCCIÓN

La detección del fraude fiscal y su cuantificación es uno de los objetivos más importantes a llevar a cabo por las Administraciones Tributarias de los distintos países.

Las metodologías para la detección del fraude y de la evasión fiscal establecidas tanto por el Banco Mundial como por la Comisión Europea se basan fundamentalmente en métodos cuantitativos.

Estas metodologías se dividen en dos grandes grupos: el enfoque de "abajo a arriba" (*botton-up*) y el enfoque de "arriba abajo" (*top-down*).

El enfoque de "abajo a arriba" (*botton-up*) parte de datos de declarantes y establece perfiles que permiten asignar probabilidades de fraude a individuos o empresas. Se trata, por tanto, de un enfoque microeconómico. Este enfoque es muy eficiente en la medición del fraude fiscal, pero exige datos habitualmente restringidos por la Administración. Exige además el uso de técnicas cuantitativas modernas avanzadas como son las Técnicas de Big Data, Minería de Datos y Machine Learning. Este es el enfoque que se desarrollará en esta tesis doctoral.

El enfoque de "arriba abajo" (*top-down*) parte de variables agregadas a nivel de país y estima cuánto debería recaudarse si todo el mundo cumpliese la regulación, es decir, estima la brecha fiscal o *Fiscal Gap*. Se trata, por tanto, de un enfoque macroeconómico. Las técnicas más utilizadas de este grupo son el método monetario, el método del consumo de electricidad, el método de la Contabilidad Nacional y el método MIMIC.

Existe mucha literatura relativa al enfoque *top-down*. Se han hecho muchas estimaciones macro de la economía sumergida utilizando diferentes metodologías. No ocurre lo mismo con el enfoque *botton-up*. Las razones fundamentale pueden ser la dificultad en acceder a los microdatos relativos a los impuestos y la alta capacidad de procesamiento necesaria para aplicar modelos avanzados a grandes conjuntos de datos. Debemos movernos entonces en el mundo de los grandes datos.

Se conoce como Big Data el tratamiento y análisis de grandes cantidades de datos, cuyo tamaño hace imposible manejarlos con las herramientas de bases de datos y analíticas convencionales. La gran cantidad de información que se almacena en las bases de datos relativas a impuestos y en otras bases de datos de la Hacienda Pública, como por ejemplo las relativas a las transacciones en el juego online (que inciden en el imposición), conlleva la necesidad de movernos en el mundo de los grandes datos. Es necesario el uso de herramientas de Big Data para su análisis.

Los análisis de datos de hoy en día requieren el uso de técnicas estadísticas para aprender de los datos, de patrones de relieve y anomalías, de predicciones y de profesionales que sepan utilizarlas. El empleo de tecnologías Big Data no solo permite aumentar la capacidad de procesamiento de los datos, sino que también facilita su análisis. Por esta razón, el Data Mining, el Machine Learning y el Big Data (*Técnicas de*

Analytics) caminan juntos para la explotación óptima de la información. Sobre una buena estructura de Big Data, que aporta el procesamiento a gran escala de la información, se implementan las técnicas y métodos avanzados de análisis de la información contenidos en la Minería de Datos y en el Machine Learning. Estos métodos suelen englobarse bajo el nombre de Analytics (*Big Data Analytics*) o Data Science.

En los sucesivos capítulos de esta tesis doctoral se dearrollarán estos métodos de Big Data Analytics con la finalidad de analizar, ordenar y cuantificar los factores que inciden en el fraude fiscal. También se aplicarán estos métodos para calcular la probabilidad que tiene cualquier individuo actual o futuro de ser defraudador. De esta forma se segmentarán los contribuyentes por nivel de propensión al fraude. Esta información sería de gran utilidad para la autoridad tributaria.

Las técnicas se desrrollarán con datos del Impuesto sobre la Renta de las Personas Físicas (IRPF), pero la metodología sería perfectamente replicable para los datos de el resto de los impuestos del sistema fiscal español.

Capítulo 1

BIG DATA ANALYTICS, MINERÍA DE DATOS Y MACHINE LEARNING

1.1 TÉCNICAS MODERNAS DE TRATAMIENTO DE DATOS

La disponibilidad de grandes volúmenes de datos y el uso generalizado de herramientas informáticas ha transformado la investigación y el análisis de datos, orientándolo hacia determinadas técnicas especializadas englobadas bajo el nombre de Minería de datos (Data Mining). El Data Mining puede definirse como un proceso de descubrimiento de nuevas y significativas relaciones, patrones y tendencias al examinar y tratar grandes cantidades de datos organizados según las técnicas de Big Data. Estamos así ante el proceso de extracción del conocimiento a través de los datos o KDD (*Knowledge Discovery in Databases*).

Por su parte, las herramientas de Minería de Datos son muy variadas y permiten la modelización, la segmentación o perfilado a través de patrones, descubrir relaciones, regularidades, tendencias, reglas de asociación, etc.

Conviene destacar que la analítica avanzada no necesariamente tiene que estar relacionada con el Big Data. El Business Intelligence puede ser igualmente eficaz para mejorar la forma en que las empresas puedan obtener información valiosa de sus bases de datos, ya sean grandes o pequeñas. Estos datos, por lo general, incluyen cifras e información de áreas como ventas, demografía de marketing, registros de CRM y otras funciones básicas del negocio. Estos son los datos estructurados que las organizaciones saben utilizar bien.

El empleo de tecnologías Big Data permite aumentar la capacidad de procesamiento. Las técnicas de Data Mining, como parte fundamental de las técnicas de Big Data Analytics, contribuyen a implementar los análisis que permitan extraer el conocimiento subyacente en los datos.

La figura 1-1 muestra el proceso de interrelación entre Big Data y las técnicas de Analytics (esencialmente Business Intelligence y Data Mining).

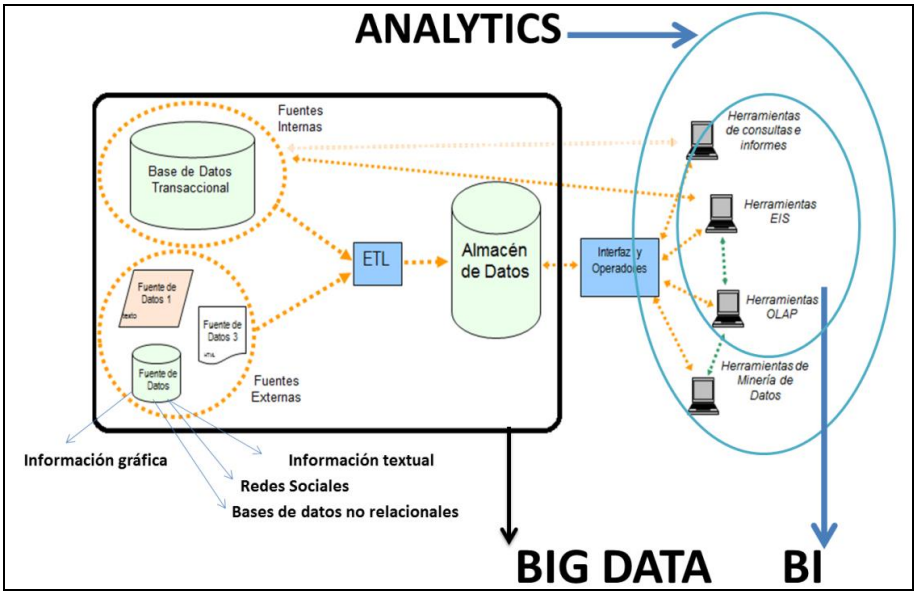


Figura 1-1

La parte inferior izquierda de la figura anterior muestra las fuentes externas de las que se extrae la información para nuestro trabajo. Estas fuentes tienen distinta naturaleza combinando información gráfica, con información textual, con información de redes sociales, con información de bases de datos no relacionales y con información de otras fuentes. Precisamente, la realización de estas tareas constituye la parte esencial del Big Data. Esta información se agrupa mediante técnicas ETL y se almacena en el Data Warehouse (Almacén de Datos). De esta forma, la información obtenida de fuentes internas y externas es almacenada adecuadamente y puesta a disposición para el análisis mediante la interfaz adecuada de las herramientas de Big Data. En la parte derecha de la figura se muestra cómo la información obtenida implementando técnicas de Big Data es tratada mediante técnicas de Analytics como Data Mining (Minería de Datos) y Business Intelligence (BI) para obtener el conocimiento almacenado en los datos. Dentro de las técnicas de Business Intellegence destacan las herramientas de consultas e informes, las herramientas EIS (Executive Information System) y las herramientas de procesamiento de transacciones (OLAP), que no son de interés en esta tesis doctoral.

1.2 ANALYTICS Y LAS TÉCNICAS DE MINERÍA DE DATOS. METODOLOGÍAS

El Data Mining (Minería de Datos) puede definirse como un proceso de descubrimiento de nuevas y significativas relaciones, patrones y tendencias al examinar grandes cantidades de datos. Las técnicas de Minería de Datos persiguen el descubrimiento automático del conocimiento contenido en la información almacenada de modo ordenado en grandes bases de datos. Estas técnicas tienen como objetivo descubrir patrones, perfiles y tendencias a través del análisis de los datos utilizando técnicas estadísticas avanzadas de análisis multivariante de datos.

La meta es permitir al investigador-analista encontrar una solución útil al problema planteado a través de una mejor comprensión de los datos existentes.

Las técnicas de Data Mining existen hace años, pero en su desarrollo actual han convergido los siguientes factores:

- Cantidad de datos disponibles
- La potencia de los ordenadores
- Las estructuras de Big Data
- Fuerte presión de la competencia (multidisciplinar)
- Software especializado de Data Mining (Enterprise Miner (SAS), Modeler (SPSS), etc.)
- El desarrollo de metodologías adecuadas (SEMMA, CRISP-DM, etc.)

Las herramientas de Data Mining siguen el ciclo del análisis de datos que se muestra en la figura 1-2. Cuando tenemos un problema, la primera tarea es comprenderlo bien y pensar en el conjunto de datos que mejor podría ayudarnos.

Comprender los datos está directamente asociado con comprender el problema.

A continuación, será necesario preparar adecuadamente los datos para utilizar modelos u otras herramientas de análisis sobre los mismos con la finalidad de extraer el conocimiento.

Estos modelos han de ser evaluados adecuadamente siguiendo las reglas técnicas de la Estadística y las Matemáticas para ser implementados finalmente.

El ciclo del análisis de datos se muestra en la Figura 1-2.

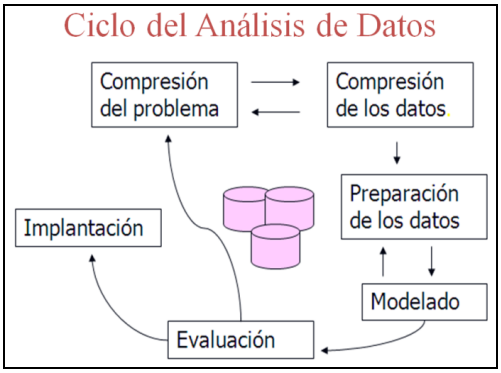


Figura 1-2

Directamente asociadas a las fases del ciclo del análisis de datos se encuentran las diferentes tareas que constituyen el proceso de la minería de datos, tal y como se muestra en la figura 1-3. Después de haber comprendido el problema y los datos es necesaria la preparación de los mismos. Esta tarea se lleva a cabo mediante las fases de selección, limpieza y codificación que obtienen los datos objetivo de las fuentes origen y los procesan y transforman para hacerlos susceptibles de análisis. Posteriormente se aplican las técnicas de Data Mining, que habitualmente conllevan modelización. Una vez evaluados e interpretados los modelos, se extrae el conocimiento contenido en los datos.

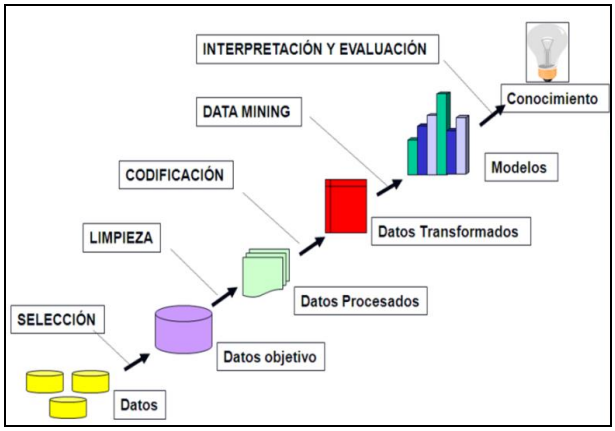


Figura 1-3

1.2.1 Metodología SEMMA

Dentro de las metodologías de Minería de Datos destacan la metodología SEMMA de SAS Institute y la metodología CRISP-DM de IBM.

La metodología SEMMA de SAS Institute divide el proceso de la Minería de Datos en las cinco fases relativas a las siglas SEMMA, que son: SAMPLE, EXPLORE, MODIFY, MODEL y ASSES (muestreo o selección, exploración, modificación, modelización y evaluación o valoración). La figura 1-4 ilustra estas fases.

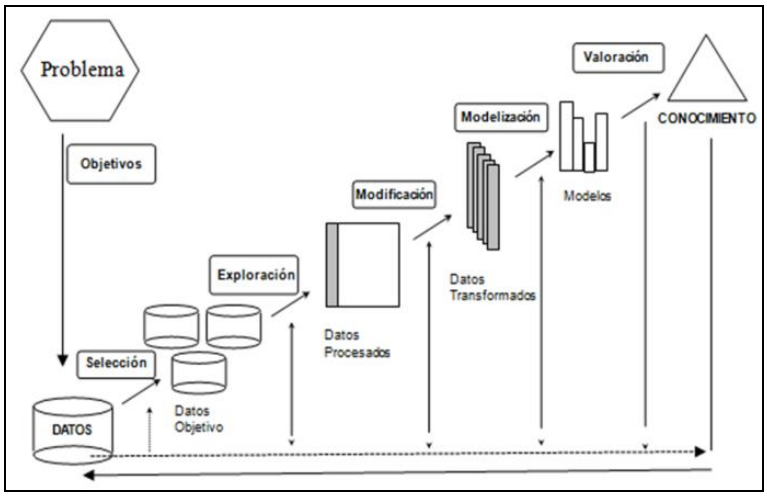


Figura 1-4

Observamos que la primera tarea en cualquier proceso de Minería de Datos es la *Selección* de la información necesaria para el análisis. Aquí tiene mucha importancia el proceso previo de tratamiento de grandes datos (Big Data). El origen de los datos es habitualmente una base de datos de alto tamaño de la cual se extrae la información y que debe de tener habilitadas las capacidades de procesamiento paralelo, computación distribuida y otras características de computación avanzada. La tecnología

Big Data, a través de Hadoop y otras aplicaciones desde el lado del software (Spark, Apache, etc.) y a través de servidores avanzados desde el lado del hardware, permite habilitar estas características. Hoy en día esto ya no es un problema en la computación. Por otra parte, cuando no es posible trabajar con todo el origen los datos debido a su gran tamaño o a su dificultad de tratamiento, se utilizan las Técnicas de Muestreo Estadístico necesarias para trabajar con una muestra lo suficientemente representativa de la población y con un tamaño adecuado para permitir el procesamiento.

Una vez que hemos seleccionado adecuadamente los datos objeto de análisis del origen de datos inicial, será necesario explorarlos convenientemente mediante las técnicas de *Exploración* de datos. Para ello contamos con la metodología estadística del Análisis Exploratorio de Datos, que nos ofrece un camino ordenado para estudiar la naturaleza de las variables del análisis, su distribución, la presencia de valores atípicos, la incidencia de los valores missing, las correlaciones entre las variables, las tendencias y muchas otras propiedades que nos capacitarán para aplicar técnicas estadísticas avanzadas a los datos. Jamás debemos implementar una técnica estadística o econométrica sin realizar previamente un análisis exploratorio de las variables implicadas. La exploración nos obligará a un procesamiento de los datos para transformar variables, eliminar valores atípicos e imputar valores mising.

Una vez realizado el análisis exploratorio de los datos, suele ser necesario llevar a cabo algunas transformaciones de las variables. En la actualidad, las variables de cualquier proceso suelen estar correlacionadas (variables del mismo negocio tienden a la correlación), lo que nos lleva a tener que aplicar transformaciones como las Técnicas de Reducción de la Dimensión (Análisis en Componentes Principales, Análisis Factorial, etc.). Mediante estas técnicas se elimina la correlación entre las variables para que los análisis realizados con las mismas estén libres del ruido introducido

por la correlación. Cualquier modelo predictivo no admite variables independientes correladas (originan multicolinealidad). Lo mismo le ocurre a las técnicas estadísticas descriptivas, como es el caso del análisis cluster, que no permite como variables de segmentación aquellas que estén muy correladas. Por lo tanto la fase de *Modificación* de los datos a través de transformaciones matemáticas de los mismos es fundamental en la Minería de Datos.

Una vez transformados los datos ya podemos aplicar Técnicas de Análisis de Datos para analizar la información. Estas técnicas podrán ser predictivas o descriptivas, según las necesidades del análisis. La mayoría de ellas se basan en modelos matemáticos formales que nos llevarán a los resultados requeridos. Por lo tanto, la fase de *Modelización* es fundamental en la Minería de datos. En esta fase también se pueden utilizar técnicas descriptivas habitualmente enfocadas a la segmentación, como es el caso del análisis cluster, el análisis de correspondencias y el escalamiento multidimensional.

Una vez establecido el modelo en la fase de modelización es necesario evaluarlo en la fase de *Evaluación* o Valoración. El modelo tiene que superar la diagnosis estadística correspondiente y hay que valorar su capacidad predictiva y su grado de eficiencia a través de herramientas como las curvas ROC, las matrices de confusión y otros instrumentos de valoración de modelos.

Superada la fase de valoración del modelo se lleva a cabo su estimación, cálculo de predicciones, clasificaciones y todas las tareas que nos permiten extraer el conocimiento a través de los datos.

En la tabla 1-1 se muestra un resumen de las tareas que comprende cada fase de la metodología SEMMA.

METODOLOGIA SEMMA

Sample →	**Orígenes de datos y Muestreo**
Explore →	**Análisis exploratorio de datos**
	Outliers, missing data, imputación…
Modify →	**Transformación de datos**
	Reducción de la dimensión…
Model →	**Modelado**
	Técnicas predictivas
Asses →	**Evaluación**
	Comparación de modelos, ROC,…

Tabla 1-1

La metodología SEMMA la utilizan especialmente las herramientas de Minería de Datos de SAS Institute y en concreto SAS ENTERPRISE MINER, que es la herramienta de Minería de Datos por excelencia de SAS Institute. No obstante, la mayoría de las herramientas modernas de minería de datos utilizan actualmente la metodología SEMMA. Esta es la metodología que se utiliza en esta tesis doctoral.

1.2.2 Metodologías CRISP-DM

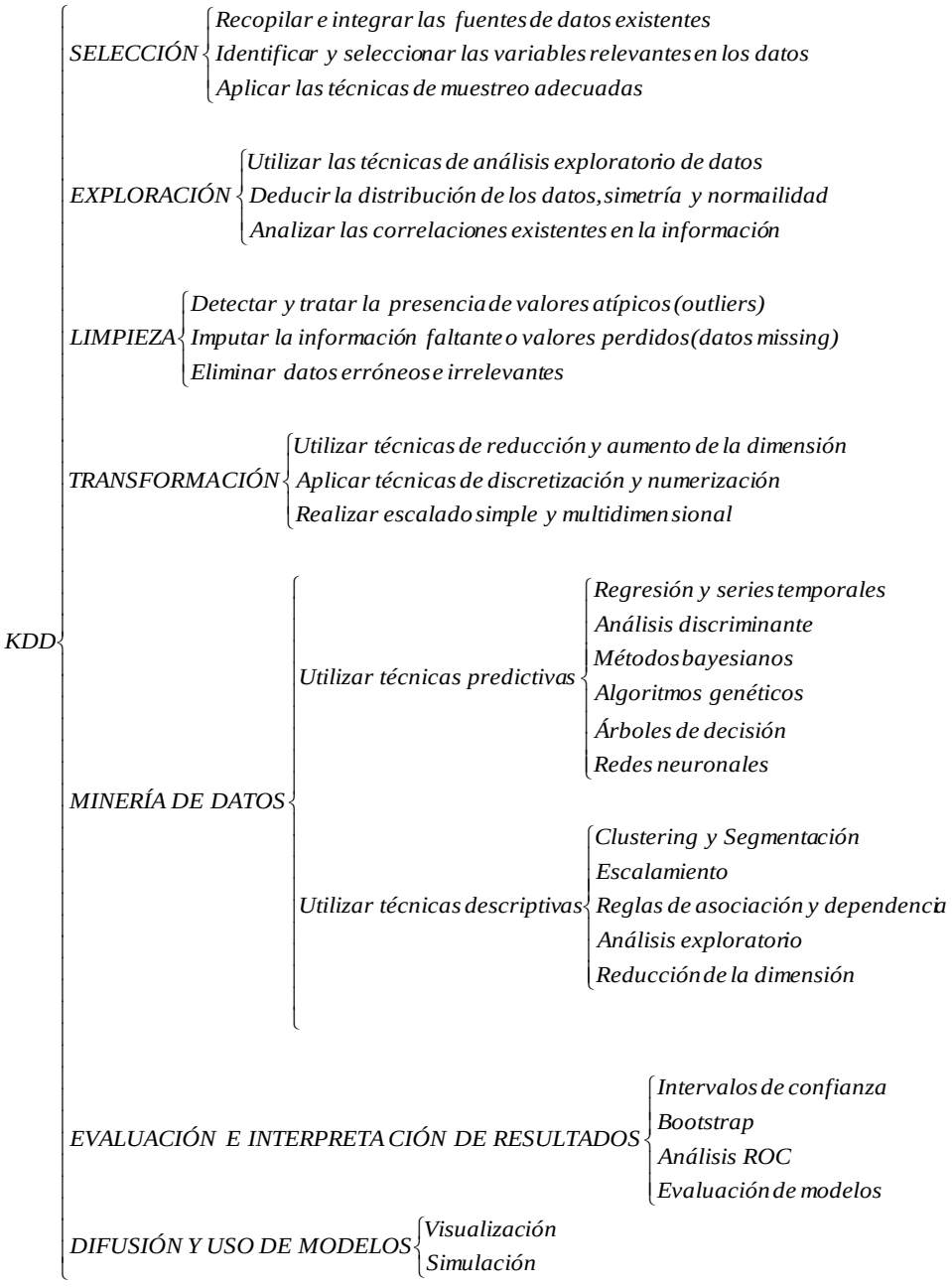
Al igual que SAS, IBM provee una metodología completa para ordenar las tareas de Minería de Datos denominada CRISP-DM (*CRoss-Industry Standard Process for Data Mining*). Inicialmente ideada para su aplicación en procesos industriales, el fundamento de CRISP_DM es similar al de la metodología SEMMA de SAS Institute. CRISP-DM considera el proceso de extracción del conocimiento a partir de los datos englobado en 6 fases, de modo que cada fase de la metodología comprende las tareas que se indican de modo resumido en la tabla 1-2.

Tabla 1-2

La metodología CRISP-DM la utilizan habitualmente las herramientas de Minería de Datos de IBM y en concreto IBM SPSS MODELER, que es la herramienta de Minería de Datos por excelencia de IBM.

1.3 EL PROCESO DE EXTRACCIÓN DEL CONOCIMIENTO

La minería de datos y todas las técnicas de Big Data Analytics son solamente una evolución del clásico *proceso de extracción de conocimiento a partir de datos*, conocido en la literatura estadística con las siglas KDD (*Knowledge Discovery in Databases*). Este proceso consta de varias fases como la preparación de datos (selección, limpieza, y transformación), su exploración y auditoría, minería de datos propiamente dicha (desarrollo de modelos y análisis de datos), evaluación, difusión y utilización de modelos (*output*). Además, el proceso de extracción del conocimiento incorpora muy diferentes técnicas (árboles de decisión, regresión lineal, redes neuronales artificiales, técnicas bayesianas, máquinas de soporte vectorial, etc) de campos diversos (aprendizaje automático e inteligencia artificial)., estadística, bases de datos, etc.) y aborda una tipología variada de problemas (clasificación, categorización, estimación/regresión, agrupamiento, etc.). El esquema siguiente muestra las etapas del KDD.


KDD
SELECCIÓN
Recopilar e integrar las fuentes de datos existentes
Identificar y seleccionar las variables relevantes en los datos
Aplicar las técnicas de muestreo adecuadas
EXPLORACIÓN
Utilizar las técnicas de análisis exploratorio de datos
Deducir la distribución de los datos, simetría y normalidad
Analizar las correlaciones existentes en la información
LIMPIEZA
Detectar y tratar la presencia de valores atípicos (outliers)
Imputar la información faltante o valores perdidos (datos missing)
Eliminar datos erróneos e irrelevantes
TRANSFORMACIÓN
Utilizar técnicas de reducción y aumento de la dimensión
Aplicar técnicas de discretización y numerización
Realizar escalado simple y multidimensional
MINERÍA DE DATOS
Utilizar técnicas predictivas
Regresión y series temporales
Análisis discriminante
Métodos bayesianos
Algoritmos genéticos
Árboles de decisión
Redes neuronales
Utilizar técnicas descriptivas
Clustering y Segmentación
Escalamiento
Reglas de asociación y dependencia
Análisis exploratorio
Reducción de la dimensión
EVALUACIÓN E INTERPRETACIÓN DE RESULTADOS
Intervalos de confianza
Bootstrap
Análisis ROC
Evaluación de modelos
DIFUSIÓN Y USO DE MODELOS
Visualización
Simulación

Se observa que las etapas del KDD son similares a las etapas de la minería de datos, con la salvedad de que denomina técnicas de minería de datos a las técnicas de modelado de las metodologías SEMMA y CRISP-DM. Si en la clasificación de las técnicas de extracción del conocimiento KDD cambiamos la fase de minería de datos por la fase de modelado, tendrías la clasificación actual de las técnicas de minería de datos.

1.4 TÉCNICAS DE MODELADO EN MINERÍA DE DATOS

La clasificación inicial de las técnicas de modelado en minería de datos distingue entre técnicas predictivas, en las que las variables pueden clasificarse inicialmente en dependientes e independientes (similares a las técnicas del análisis de la dependencia o métodos explicativos del análisis multivariante), técnicas descriptivas, en las que todas las variables tienen inicialmente el mismo estatus (similares a las técnicas del análisis de la interdependencia o métodos descriptivos del análisis multivariante) y técnicas auxiliares.

Las *técnicas predictivas* especifican el modelo para los datos en base a un conocimiento teórico previo. El modelo supuesto para los datos debe contrastarse después del proceso de minería de datos antes de aceptarlo como válido. Formalmente, la aplicación de todo modelo debe superar las fases de *identificación objetiva* (a partir de los datos se aplican reglas que permitan identificar el mejor modelo posible que ajuste los datos), *estimación* (proceso de cálculo de los parámetros del modelo elegido para los datos en la fase de identificación), *diagnosis* (proceso de contraste de la validez del modelo estimado) y *predicción* (proceso de utilización del modelo identificado, estimado y validado para predecir valores futuros de las variables dependientes). En algunos casos, el modelo se obtiene como mezcla del conocimiento obtenido antes y después del Data Mining y también debe contrastarse antes de aceptarse como válido.

Por ejemplo, las *redes neuronales* permiten descubrir modelos complejos y afinarlos a medida que progresa la exploración de los datos. Gracias a su capacidad de aprendizaje, permiten descubrir relaciones complejas entre variables sin ninguna intervención externa. Podemos incluir entre estas técnicas todos los tipos de regresión, series temporales, análisis de la varianza y covarianza, análisis discriminante, modelos logísticos, árboles de decisión, redes neuronales, algoritmos genéticos y técnicas bayesianas. Los árboles de decisión, las redes neuronales, los modelos logísticos y el análisis discriminante son a su vez *técnicas de clasificación* que pueden extraer perfiles de comportamiento o clases, siendo el objetivo construir un modelo que permita clasificar cualquier nuevo dato en las categorías de la variable dependiente.

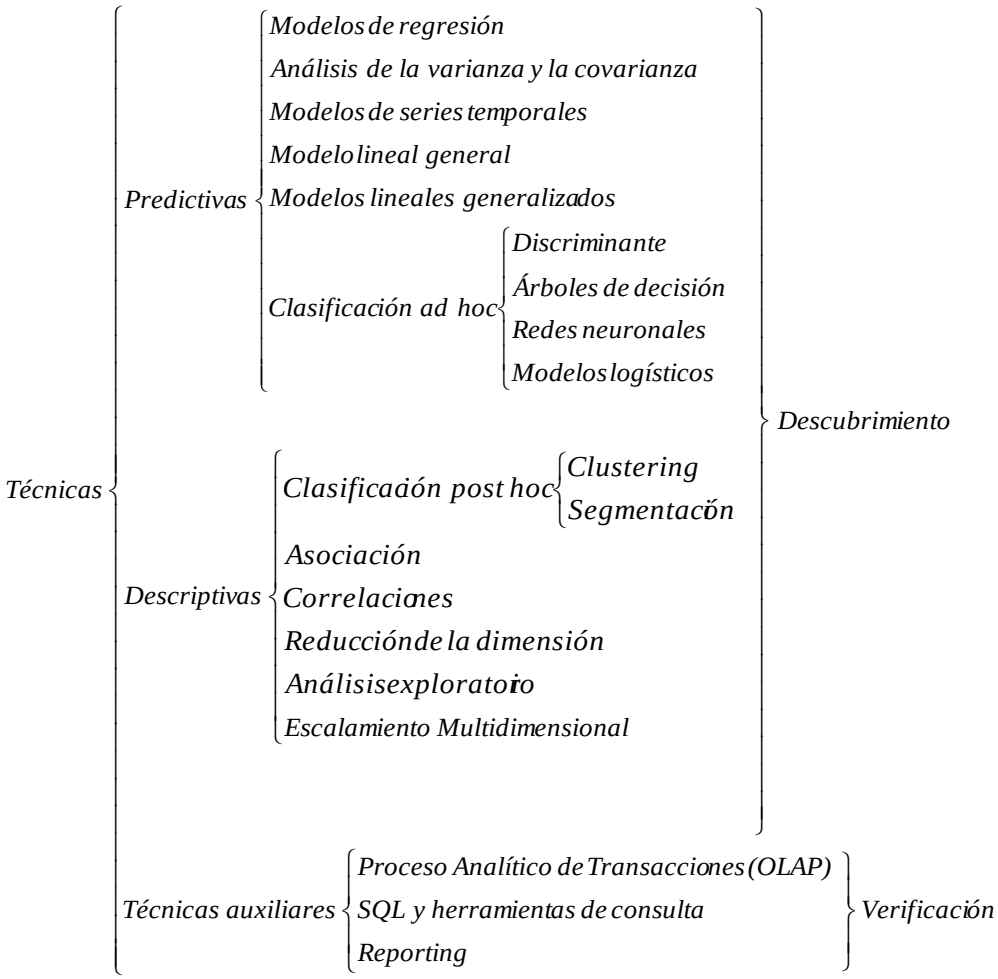
La mayoría de las técnicas predictivas se utilizarán en esta tesis doctoral.

En las *técnicas descriptivas* no se asigna ningún papel predeterminado a las variables. No se supone la existencia de variables dependientes ni independientes y tampoco se supone la existencia de un modelo previo para los datos. Los modelos se crean automáticamente partiendo del reconocimiento de patrones. En este grupo se incluyen las técnicas de clustering y segmentación (que también son *técnicas de clasificación*), las técnicas de asociación y dependencia, las técnicas de análisis exploratorio de datos y las técnicas de reducción de la dimensión (factorial, componentes principales, correspondencias, etc.) y de escalamiento multidimensional.

Tanto las técnicas predictivas como las técnicas descriptivas están enfocadas al *descubrimiento del conocimiento* embebido en los datos.

Las *técnicas auxiliares* son herramientas de apoyo más superficiales y limitadas. Se trata de métodos basados en técnicas estadísticas descriptivas, consultas e informes y enfocados en general hacia la *verificación*.

A continuación se muestra una *clasificación de las técnicas de Modelado* en al minería de datos.



Se observa que las *técnicas de clasificación* pueden pertenecer tanto al grupo de técnicas predictivas (discriminante, árboles de decisión, redes neuronales y modelos logísticos) como a las descriptivas (clustering y segmentación). Las técnicas de clasificación predictivas suelen denominarse *técnicas de clasificación ad hoc* ya que clasifican individuos u observaciones dentro de grupos previamente definidos que son las

categorías de la variable dependiente del modelo. Las técnicas de clasificción descriptivas se denominan *técnicas de clasificación post hoc* porque realizan clasificación sin especificación previa de los grupos. Los grupos se obtienen a posteriori después de aplicar la técnica.

Capítulo 2

FACTORES QUE AFECTAN AL FRAUDE FISCAL A TRAVÉS DE TÉCNICAS DE MINERÍA DE DATOS. APLICACIÓN AL IRPF MEDIANTE ÁRBOLES DE DECISIÓN

2.1 INTRODUCCIÓN

Este capítulo tiene como finalidad estudiar los factores que afectan al fraude fiscal (causas de fraude) en el Impuesto sobre la Renta de las Personas Físicas y cuantificar, ordenar y probabilizar su incidencia. El hecho de disponer de grandes conjuntos de datos con información relativa a impuestos permite ampliar las posibilidades de análisis cuantitativo y utilizar las nuevas prestaciones que aportan el Big Data y la Minería de Datos. En este trabajo se trata de mostrar el uso de los modelos de Árboles de Decisión aplicados a las muestras de IRPF del IEF con la finalidad de estudiar los factores que afectan al fraude fiscal en este impuesto. A partir de una muestra anual de IRPF se construirán modelos de árboles de decisión que cuantificarán la incidencia de las distintas causas que delimitan el fraude fiscal en IRPF basándose exclusivamente en la información que se declara a la Agencia Tributaria en el modelo correspondiente a este impuesto. Asimismo se elaborarán modelos predictivos que permiten

cuantificar la probabilidad que tiene cualquier contribuyente futuro de ser defraudador por cada factor de fraude una vez que presente su declaración de IRPF. Estos modelos permitirán segmentar a los declarantes del impuesto por nivel de propensión al fraude y causas del mismo. Esta metodología es generalizable para cuantificar la propensión al fraude en cualquier otro impuesto según los factores que lo determinan y para cuantificar y ordenar la incidencia de dichos factores en el fraude.

2.2 MARCO METODOLÓGICO: LOS ÁRBOLES DE DECISIÓN

Las técnicas estadísticas de la dependencia o técnicas predictivas se caracterizan por el hecho de que alguna (o algunas) de las variables en estudio destaca como dependiente principal. Este concepto está en contraposición con las técnicas estadísticas de la interdependencia o técnicas descriptivas en los que ninguna variable destaca como dependiente.

En el caso de los métodos de la dependencia es necesario utilizar técnicas multivariantes analíticas o inferenciales considerando la variable dependiente como explicada por las demás variables independientes explicativas, y tratando de relacionar todas las variables por medio de una posible ecuación o modelo que las ligue (por ejemplo, el modelo de regresión que para varias variables dependientes se generaliza a los modelos de ecuaciones simultáneas). Una vez configurado el modelo matemático se podrá llegar a predecir el valor de la variable (o variables) dependiente conocido el perfil de las demás, es decir, conocidos los valores de las variaablers independientes.

El análisis de la regresión múltiple es una técnica estadística utilizada para analizar la relación entre una variable dependiente (o endógena) métrica (cuantitativa) y varias variables independientes (o exógenas) también métricas. El objetivo esencial del análisis de la regresión

múltiple es utilizar las variables independientes, cuyos valores son conocidos, para predecir la única variable criterio (dependiente) seleccionada por el investigador. La expresión funcional del análisis de la regresión múltiple es la siguiente:

$$y = F(x_1, x_2, \cdots, x_n)$$

donde inicialmente, tanto la variable dependiente *y* como las independientes x_i son métricas. Asimismo la regresión múltiple admite la posibilidad de trabajar con variables independientes no métricas si se emplean variables ficticias (modelos de regresión con variables ficticias) para su transformación en métricas

Si la variable dependiente fuera cualitativa podrá usarse como clasificadora, estudiando su relación con el resto de variables clasificativas (independientes). Si la variable dependiente cualitativa observada constatara la asignación de cada individuo a grupos previamente definidos (dos, o más de dos), puede ser utilizada para clasificar nuevos casos en que se desconozca el grupo a que probablemente pertenecen, que resuelve el problema de asignación en función de un perfil cuantitativo de variables clasificativas (las independientes del modelo). En este caso de variable dependiente cualitativa, si las variables independientes fuesen cuantitativas estaríamos ante el caso del modelo discriminante o el modelo logístico. Pero si las variables independientes fuesen también cualitativas estaríamos ante un modelo de árboles de decisión.

Como en toda técnica de la dependencia subyace un modelo, suelen asociarse estas técnicas a los modelos predictivos o modelos econométricos.

En los modelos predictivos (métodos de la dependencia) subyace una relación general de dependencia entre las variables independientes $x_1, x_2, \cdots, x_n$ y las dependientes $y_1, y_2, \cdots, y_n$ del tipo genérico:

$$G(y_1, y_2, \cdots, y_n) = F(x_1, x_2, \cdots, x_n).$$

La naturaleza de las variables caracterizará cada modelo.

Un árbol de decisión analiza la relación entre una variable dependiente (o endógena) no métrica (cualitativa) y varias variables independientes (o exógenas) no métricas (cualitativas). Su expresión es $y = F(x_1, x_2, \cdots, x_n)$.

Al igual que en el caso de la regresión logística y el análisis discriminante, la finalidad es predecir la categoría de la variable dependiente cualitativa en la que se clasifican los individuos según los valores de sus variables independientes cualitativas.

Habitualmente, en los árboles de decisión también se utilizan variables métricas agrupando previamente sus valores en intervalos (a lo sumo cinco en la práctica).

Los árboles de decisión son modelos predictivos, que trata de resolver los problemas de discriminación en una población segmentando de forma progresiva la muestra para obtener finalmente una clasificación fehaciente en grupos homogéneos, según la variable de interés denominada variable de segmentación.

En los árboles de decisión la segmentación de la población se realiza según los valores de la variable de interés que juega el papel de variable dependiente del modelo predictivo subyacente en el árbol (variable cualitativa).

La asignación de un elemento poblacional a un segmento se realiza de acuerdo a los valores de determinadas variables medidas sobre él que constituyen las variables independientes del modelo (habitualmente también variables cualitativas, aunque también suelen utilizarse variables cuantitativas con sus valores agrupados en un número pequeño de intervalos).

Se trata, por tanto, de seleccionar las variables explicativas que son más discriminantes para la variable dependiente y de construir una regla de decisión que permita asignar un nuevo individuo a un valor o clase de la variable dependiente.

El método consiste en buscar la variable independiente x_j que mejor explique a la variable dependiente y. Esta variable define una primera división de la muestra en dos subconjuntos, llamados segmentos. Después se reitera el procedimiento en el interior de cada uno de estos dos segmentos buscando la segunda mejor variable y así sucesivamente. No obstante, en la práctica esta tarea se realiza ordenadamente a través del dendograma del árbol.

En nuestro caso utilizaremos un modelo de árbol de decisión que explica el fraude global (variable dependiente) en función de las causas o factores de fraude más comunes en el Impuesto sobre la Renta de las Personas Físicas (variables independientes). En primer lugar obtendremos la causa de fraude que mejor explica el frade global y reiteraremos el procedimiento para ordenar las distintas causas de fraude de acuerdo a su incidencia sobre el fraude global.

Posteriormente segmentaremos los individuos por su propensión al fraude calculando la probabilidad que tiene cada individuo de ser defraudador.

Los tres tipos de árboles más utilizados hoy en día son: los árboles CHAID, los árboles CART y los árboles QUEST.

El método CHAID (*Chi-square Automatic Interaction Detector*) es la conclusión de una serie de métodos basados en el detector Automático de Interacciones (AID) de Morgan y Sonquist. La variable dependiente suele ser cualitativa (nominal u ordinal) o cuantitativa. Para variables cualitativas, el análisis lleva a cabo una serie de análisis $\chi 2$ entre las variables dependiente y predictora. En el caso de variables dependientes cuantitativas, se categorizan reduciéndolas a pocos intervalos recurriendo a métodos de análisis de varianza, en los que los intervalos (divisiones) se determinan óptimamente para las variables independientes, de forma que maximicen la capacidad para explicar la varianza de la variable dependiente. Este método ahorra bastante tiempo de computación, pero no garantiza que sea capaz de encontrar realmente la mejor división posible en cada modo.

Para garantizar el hallazgo de la división más significativa se utiliza el *método CHAID exhaustivo*, que trata a todas las variables por igual, independientemente del tipo de variable y del número de categorías. Por otro lado, este método permite trabajar con variables dependientes categóricas y métricas. Las variables categóricas utilizan el estadístico $\chi 2$ y dan lugar a un *árbol de clasificación*. Las variables métricas utilizan el estadístico F y dan lugar a lo que se conoce como *árboles de regresión*. También permite utilizar predictores de tipo métrico, mediante su conversión previa en variables categóricas. Los métodos CHAID producen divisiones de la validación cruzada en más de dos grupos, lo cual siempre es un valor añadido.

El método CART (*Classification And Regression Trees*) o C&RT es una alternativa al CHAID exhaustivo para *árboles de clasificación* (variables dependientes categóricas). Este método nació para intentar superar algunas de las deficiencias y debilidades que por entonces mostraba la formulación

original del CHAID, que estaba limitado inicialmente a variables dependientes nominales y variables independientes categóricas hasta la aparición de su versión exhaustiva. Estaba claro que se necesitaba utilizar predictores de cualquier nivel de medida. Además, CART tiene una estructura estadística más fuerte que CHAID, lo que le llevó a ser utilizado en campos de la investigación como la medicina y el marketing. CART se utiliza para árboles de clasificación con variable dependiente cualitativa y para árboles de regresión con variable dependiente cuantitativa.

El método comienza dividiendo la muestra en subconjuntos y evaluando cada predictor cuantitativo para encontrar el mejor punto de corte o cada predictor categórico y para encontrar las mejores agrupaciones de categorías. A continuación se comparan también los predictores, seleccionándose el predictor y la división que produce la mayor bondad de ajuste. Para predictores cuantitativos suele utilizarse la minimización del error cuadrático o de la desviación media absoluta respecto de la mediana. Para predictores cualitativos suele utilizarse el coeficiente Gini para evaluar la probabilidad de una mala clasificación (valor cero para clasificación perfecta y valor uno para una mala clasificación). No debemos de olvidar que los métodos CHAID producen divisiones de la validación cruzada en más de dos grupos, mientras que el método CART sólo produce divisiones binarias.

Los árboles QUEST (*Quick, Unbiased, Efficient, Statistical Tree*) consisten en un algoritmo de clasificación arborescente creado específicamente para solventar dos de los principales problemas que presentan métodos como CART y CHAID exhaustivo, a la hora de dividir un grupo de sujetos en función de una variable independiente. Este tipo de árboles mitigan la complejidad computacional (enfoque de cálculo más sencillo) y los sesgos en la selección de variables. Se trata de evitar que se seleccionen aquellas variables que cuentan con un mayor número de categorías. QUEST intenta seleccionar el mejor predictor y su mejor punto

de corte como tareas separadas, calculando en cada nodo la asociación entre cada predictor y la variable dependiente mediante el estadístico F del ANOVA o la F de Levenne para predictores continuos y ordinales o mediante una $\chi 2$ de Pearson para predictores nominales. Se consiguen divisiones binarias de la variable dependiente mediante la creación de dos superclases en el predictor, aplicando un algoritmo conglomerativo. Por último, para eliminar el sesgo en la selección de variables, se elige el predictor que tiene la mayor asociación con la variable dependiente.

En cuanto a la valoración de los métodos de construcción de árboles, podría establecerse un orden de jerarquía (nunca absoluto) que sitúe el método QUEST como superior a CART y este último método superior a CHAID. No olvidemos que QUEST admite métodos de validación mediante poda y permite utilizar combinaciones lineales de variables. Pero debe quedar claro que esta evaluación sólo es válida en líneas generales.

En nuestro modelo de árbol, tanto la variable dependiente como las independientes son categóricas, ya que modelizamos el fraude global (variable dependiente) en función de las causas o factores de fraude más comunes en el Impuesto sobre la Renta de las Personas Físicas (variables independientes). Además, todas las variables son binarias, ya que todas ellas se miden en término de fraude (categoría 1) o no fraude (categoría 0). Por lo tanto, podremos utilizar todas las tipologías de árboles y compararlas entre sí.

2.3 FASE DE SELECCIÓN DE LA INFORMACION: LOS DATOS

La aplicación de cualquier técnica de minería de datos comienza por la fase de selección, intentando identificar la fuente óptima de datos para la aplicación de la técnica. Lo ideal sería utilizar toda la base de datos de IRPF de la Agencia Tributaria. Pero sus más de 25 millones de registros y cerca de 800

variables hacen imposible la utilización de técnicas estadísticas complejas sobre esta gran base de datos. Por lo tanto utilizaremos una muestra seleccionada según criterios matemáticos formales de muestreo.

En la aplicación que aquí se presenta, se utiliza como fuente de datos la muestra del Impuesto sobre la Renta de las Personas Físicas (IRPF) que proporciona el Instituto de Estudios Fiscales (IEF). Los trabajos de investigación sobre fiscalidad llevados a cabo en España antes de la publicación de estas muestras por el IEF han tenido que realizarse bien con datos fiscales anticuados (caso del Panel de IRPF 1982-1998), bien con datos agregados (BADESPE y Memorias de la Agencia Tributaria), o bien con datos de origen no fiscal que habitualmente minoran los ingresos reales (Encuestas de Presupuestos Familiares, Panel de Hogares de la Unión Europea). Las muestras de IRPF del IEF vienen a cubrir esta importante laguna, proporcionando una base de datos de gran amplitud (más de dos millones de observaciones por año) y detalle (cerca de 300 variables personales, familiares y fiscales).

El origen fiscal de la muestra proporciona, por tanto, unos datos de gran precisión, y en los que además no aparecen los problemas de infrarrepresentación y falta de respuesta habituales de las encuestas. Por consiguiente, la riqueza de estos datos permite realizar múltiples análisis que están vedados a otras muestras de origen no fiscal.

Hay que tener presente que disponer de una estructura de hardware y software que implemente procesamiento de grandes datos (*Big Data*) es esencial en nuestro caso. De esta forma, las técnicas de minería de datos que vamos a utilizar se encuadran dentro de las técnicas de *Big Data Analytics*.

2.3.1 La muestra de IRPF: ámbitos, unidades de muestreo y tipo de muestreo

Como ya se ha comentado anteriormente, se seleccionan más de 2 millones de registros de la población total de declarantes mediante muestreo.

En cuanto al *ámbito poblacional*, la población objetivo para el muestreo son las declaraciones presentadas del Impuesto sobre la Renta de las Personas Físicas (IRPF) correspondientes al ejercicio correspondiente. El *ámbito geográfico* lo constituye el Territorio de Régimen Fiscal Común (no incluye País Vasco y Navarra). El *ámbito temporal* es el ejercicio correspondiente, teniendo presente que las muestras se elaboran y publican todos los años a partir de 2002.

Las *unidades de muestreo* son las declaraciones de IRPF del año correspondiente y la *población marco*, que la constituyen el conjunto de unidades de entre las cuales se selecciona efectivamente la muestra, es la lista de declaraciones del Modelo 100 de IRPF en el año.

En cuanto al *tipo de muestreo*, se ha utilizado muestreo estratificado aleatorio. En cuanto a la formación de los estratos se han considerado en primer lugar las provincias españolas del Territorio Fiscal Común (49, ya que Ceuta y Melilla se han considerado de forma conjunta). En un segundo nivel de estratificación se han considerado 12 tramos de renta (rentas negativas y cero, intervalos de rentas de 6.000 € hasta 60.000 € y un último intervalo de rentas superiores a 60.000 €) y en un tercer nivel de estratificación se han considerado las declaraciones individuales y las conjuntas. Por lo tanto, el número de estratos de último nivel es 49x2x12=1.176, no existiendo estratos vacíos.

La variable utilizada para definir los tramos de renta ha sido la variable *Renta = Saldo Neto de Rendimiento e Imputaciones de Renta + Base Imponible del Ahorro*.

El muestreo estratificado puede aportar información más precisa de algunas subpoblaciones que varían bastante en tamaño y propiedades entre sí, pero que son homogéneas dentro de sí. Por ejemplo, será posible realizar estimaciones por provincias y comunidades autónomas con la misma precisión que cualquier estimación a nivel nacional. Asimismo, pueden realizarse estimaciones para áreas pequeñas, como municipios, con bastante fiabilidad, dado que la muestra está muy repartida por toda la población debido a la propia definición de la estratificación. No obstante, el error de cualquier estimación de área pequeña es calculable con las fórmulas del error que se verán posteriormente.

El uso adecuado del muestreo estratificado genera ganancia en precisión respecto del muestreo aleatorio simple siempre y cuando los estratos sean muy homogéneos dentro de sí respecto de la variable de estratificación. En nuestro caso, el hecho de definir estratos por zonas geográficas y tramos de renta, produce homogeneidad dentro de los estratos y además dispersa muy bien la muestra por la geografía nacional (cruzar 12 tramos de renta con todas las provincias españolas produce un reparto óptimo de la muestra).

2.3.2 La muestra de IRPF: afijación de la muestra

El reparto de la muestra en los estratos se ha realizado mediante *afijación de mínima varianza*. La afijación de mínima varianza o afijación de Neyman consiste en determinar los valores de n_h (número de unidades que se extraen del estarto h-ésimo para la muestra) de forma que para un tamaño de muestra fijo igual a *n* la varianza de los estimadores sea mínima. Matemáticamente resulta que:

$$n_h = n \cdot \frac{N_h S_h}{\sum_{h=1}^{L} N_h S_h} = n \cdot \frac{\frac{N_h}{N} S_h}{\sum_{h=1}^{L} \frac{N_h}{N} S_h} = n \cdot \frac{W_h S_h}{\sum_{h=1}^{L} W_h S_h}$$

S_h^2 = cuasivarianza poblacional en el estrato *h*, N_h es el tamaño poblacional del estrato y *N* es el tamaño de la población = número de declaraciones. $W_h = N_h/N$ son los coeficientes de ponderación de suma unitaria. L es el número de estratos.

Teóricamente, la afijación de mínima varianza es el tipo de afijación más preciso en muestreo estratificado. La utilidad de la afijación de mínima varianza es mayor si hay grandes diferencias en la variabilidad entre los estratos, que es el caso que nos ocupa (las rentas son muy variables entre los estratos). En otro caso, la mayor sencillez y autoponderación de la afijación proporcional (extracción de un número de elementos para la muestra proporcional al tamaño del estrato) harían preferible el empleo de ésta, pero no es nuestro caso.

En general, el muestreo estratificado con afijación de mínima varianza es más preciso que el muestreo estratificado con afijación proporcional y que el aleatorio simple, siendo además el estratificado con afijación proporcional más preciso que el aleatorio simple. En nuestro caso se comprobó que, para el mismo error de muestreo, se necesitaba el doble de tamaño muestral si se utilizaba afijación proporcional que en el caso de utilizar afijación de mínima varianza.

El *tamaño de la muestra* está calculado para un error relativo de muestreo $e_{r\alpha}$ menor del 1,1% con un nivel de confianza λ_α del 3 por mil ($\overline{X}$ = renta media poblacional).

$$e_{r\alpha}{}^2 = \lambda_\alpha^2 \frac{\frac{1}{n}\left(\sum_{h=1}^{L} N_h S_h\right)^2 - \sum_{h=1}^{L} N_h S_h^2}{\overline{X}^2} \Rightarrow n \approx 2000000 de\ declaraciones$$

2.3.3 La muestra de IRPF: estimadores y errores

En cuanto a los *estimadores*, el *estimador de cualquier total poblacional* X en muestreo estratificado aleatorio es la suma de los estimadores del total en cada uno de los L estratos. Se tiene:

$$\hat{X}_{st} = \sum_{h=1}^{L} \hat{X}_h = \sum_{h=1}^{L} N_h \overline{x}_h = \sum_{h=1}^{L} \frac{N_h}{n_h} x_h = \sum_{h=1}^{L} fe_h x_h$$

$$\begin{cases} \overline{x}_h = media\ muestral\ en\ el\ estrato h \\ x_h = total\ muestral\ en\ el\ estrato h \\ N_h = tamaño\ poblacional\ del\ estrato h \\ n_h = tamaño\ muestral\ del\ estrato h \\ fe_h = \ factor\ de\ elevación\ del\ estrato h \end{cases}$$

Por lo tanto, para estimar cualquier total poblacional se suman los productos de los factores de elevación fe_h por los totales muestrales en cada estrato x_h.

El *estimador de cualquier media* en muestreo estratificado aleatorio es la media ponderada de los estimadores de la media en cada estrato, siendo los coeficientes de ponderación $W_h = N_h/N$ de suma unitaria (N_h es el tamaño poblacional del estrato y N es el tamaño de la población = número de declaraciones).

$$\hat{\bar{X}}_{st} = \bar{x}_{st} = \sum_{h=1}^{L} W_h \bar{x}_h = \sum_{h=1}^{L} \underbrace{\frac{N_h}{N}}_{W_h} \frac{1}{n_h} x_h = \frac{1}{N} \sum_{h=1}^{L} \frac{N_h}{n_h} x_h = \frac{1}{N} \sum_{h=1}^{L} fe_h x_h$$

Por lo tanto, para estimar cualquier media poblacional se suman los productos de los factores de elevación por los totales muestrales en cada estrato y se divide por el tamaño poblacional.

Las varianzas de los estimadores para medir los errores absolutos y sus estimaciones son ($f_h = n_h / N_h$):

$$V\left(\hat{X}_{st}\right) = \sum_{h=1}^{L} N_h^2 (1 - f_h) \frac{S_h^2}{n_h}, \quad V(\bar{x}_{st}) = V\left(\sum_{h=1}^{L} W_h \bar{x}_h\right) = \sum_{h=1}^{L} W_h^2 (1 - f_h) \frac{S_h^2}{n_h}$$

$$\hat{V}\left(\hat{X}_{st}\right) = \sum_{h=1}^{L} N_h^2 (1 - f_h) \frac{\hat{S}_h^2}{n_h}, \quad \hat{V}\left(\bar{X}_{st}\right) = \sum_{h=1}^{L} W_h^2 (1 - f_h) \frac{\hat{S}_h^2}{n_h}$$

S_h^2 = cuasivarianza poblacional en el estrato *h*, $\hat{S}_h^2$ = cuasivarianza muestral en el estrato *h*

Los errores relativos y sus estimaciones se calculan mediante las expresiones de los correspondientes coeficientes de variación:

$$\hat{C}_v\left(\hat{X}_{st}\right) = \frac{\sqrt{\hat{V}(\hat{X}_{st})}}{\hat{X}_{st}} \qquad \hat{C}_v\left(\bar{x}_{st}\right) = \frac{\sqrt{\hat{V}(\bar{x}_{st})}}{\bar{x}_{st}}$$

Estas fórmulas para el cláculo de variaanzas, coeficientes de variación y sus estimaciones se simplifican bastante para afijación de mínima varianza, sobre todo a la hora de realizar los cálculos prácticos.

2.3.4 La muestra de IRPF: estimadores de errores para áreas pequeñas

Este apartado es muy importante cuando se utiliza la muestra para hacer estimaciones en áreas pequeñas, por ejemplo municipios.

Cuando se trabaja a nivel provincial o autonómico, las estimaciones se obtienen con el error general de muestreo (1,1% con una confianza adicional del 3 por mil como ya se indicó antes), ya que las zonas citadas son zonas de estratificación.

Existe una propiedad del muestreo estratificado que asegura que las estimaciones en cualquier estrato tienen el mismo error que las estimaciones en toda la población, por lo tanto, cuando se trabaja en regiones no inferiores a los niveles de estratificación no es necesario calcular el error de muestreo.

Pero si se realizan estimaciones en áreas geográficas inferiores a la provincia, por ejemplo en municipios, es estrictamente necesario calcular el error de muestreo de cualquier estimación que se realice. El coeficiente de variación estimado debe de ser inferior al 15% si se aplica un criterio estricto, o al 20% si se utiliza un criterio más amplio.

Matemáticamente se demuestra que para afijación de mínima varianza, los errores de muestreo se calculan como se indica en los párrafos siguientes.

Una vez calculados los n_h para afijación de mínima varianza con la fórmula ya especificada anteriormente (en la práctica se trata de los tamaños muestrales de los estratos que son un dato conocido si tenemos laa muestra), vamos a ver cuánto vale la varianza del estimador de la media para este tipo de afijación en el caso de la estimación del total. Tenemos:

$$V\left(\hat{X}_{st}\right)=\sum_{h=1}^{L}N_h^2\cdot\frac{S_h^2}{n_h}\left(1-\frac{n_h}{N_h}\right)=\sum_{h=1}^{L}N_h^2\cdot\frac{S_h^2}{n_h}-\sum_{h=1}^{L}N_h\cdot\frac{S_h^2}{N}=\sum_{h=1}^{L}N_h^2\cdot\frac{S_h^2}{n.\dfrac{N_hS_h}{\sum_{h=1}^{L}N_hS_h}}-$$

$$-\sum_{h=1}^{L}N_h\cdot\frac{S_h^2}{N}=\sum_{h=1}^{L}\frac{N_hS_h}{n.\dfrac{1}{\sum_{h=1}^{L}N_hS_h}}-\frac{1}{N}\sum_{h=1}^{L}N_hS_h^2=\frac{1}{n}\left(\sum_{h=1}^{L}N_hS_h\right)^2-\frac{1}{N}\sum_{h=1}^{L}N_hS_h^2$$

Si se quiere la afijación y la expresión de la varianza para el estimador del total de clase basta sustituir en la fórmula anterior S_h^2 por $P_hQ_hN_h/(N_h-1)$.

En el caso del estimador de cualquier media tenemos:

$$V\left(\bar{x}_{st}\right)=\sum_{h=1}^{L}W_h^2\cdot\frac{S_h^2}{n_h}\left(1-\frac{n_h}{N_h}\right)=\sum_{h=1}^{L}W_h^2\cdot\frac{S_h^2}{n_h}-\sum_{h=1}^{L}W_h\cdot\frac{S_h^2}{N}=\sum_{h=1}^{L}W_h^2\cdot\frac{S_h^2}{n.\dfrac{W_hS_h}{\sum_{h=1}^{L}W_hS_h}}-$$

$$-\sum_{h=1}^{L}W_h\cdot\frac{S_h^2}{N}=\sum_{h=1}^{L}\frac{W_hS_h}{n.\dfrac{1}{\sum_{h=1}^{L}W_hS_h}}-\frac{1}{N}\sum_{h=1}^{L}W_hS_h^2=\frac{1}{n}\left(\sum_{h=1}^{L}W_hS_h\right)^2-\frac{1}{N}\sum_{h=1}^{L}W_hS_h^2$$

Por lo tanto, podemos estimar los errores de muestreo para totales y medias en cualquier área pequeña inferior a la provincia mediante las siguientes fórmulas:

$$\hat{V}\left(\hat{X}_{st}\right)=\frac{1}{n}\left(\sum_{h=1}^{L}N_h\hat{S}_h\right)^2-\frac{1}{N}\sum_{h=1}^{L}N_h\hat{S}_h^2$$

$$\hat{V}\left(\bar{x}_{st}\right)=\frac{1}{n}\left(\sum_{h=1}^{L}W_h\hat{S}_h\right)^2-\frac{1}{N}\sum_{h=1}^{L}W_h\hat{S}_h^2$$

Si se quiere la afijación y la expresión de la varianza mínima para el estimador de la proporción basta sustituir en la fórmula anterior $\hat{S}_h^2$ por $\hat{P}_h \ \hat{Q}_h$ $N_h/(N_h-1)$.

Estas estimaciones de varianzas son muy fáciles de calculaar en la práctica, ya que $\hat{S}_h^2$ es la cuasivariaanza de los estratos, que se calcula fácilmente cuando tenemos la muestra. Por otra parte $W_h = N_h/N$ siendo N_h es el tamaño poblacional del estrato y N es el tamaño de la población = número de declaraciones. N_h es fácilmente calculable a partir de los factores de elevación f_h que vienen dados en la muestra. Tenemos que $N_h=n_hfh$.

Pero la magnitud que se utiliza más asiduamente para medir el error estimado es el coeficiente de variación estimado, ya que, al ser una medida relativa del error, puede interpretarse en términos de tanto por ciento, resultando su magnitud perfectamente inteligible para el usuario. Hay que procurar que el coeficiente de variación estimado no supere mucho el 20% en ningún caso. Los coeficientes de variación estimados se calculan de la siguiente forma:

$$\hat{C}_v\left(\hat{X}_{st}\right)=\frac{\sqrt{\hat{V}(\hat{X}_{st})}}{\hat{X}_{st}} \qquad \hat{C}_v\left(\bar{x}_{st}\right)=\frac{\sqrt{\hat{V}\left(\bar{x}_{st}\right)}}{\bar{x}_{st}}$$

Como ejemplo práctico, la Diputación Provincial de Barcelona realizó un análisis completo de la renta en sus municipios mayores de 5 mil habitantes utilizando la muestra de IRPF del IEF y en todos los casos las estimaciones fueron correctas a nivel municipal, ya que todos los coeficientes de variación estimados (errores relativos de muestreo) fueron inferiores al 20%.

Otro ejemplo representativo es la Comunidad de Madrid, que realizó el análisis de la renta a nivel de distrito municipal utilizando la muestra de IRPF del IEF, resultando las estimaciones significativas al 20% o menos.

2.3.5 La muestra de IRPF: las variables de la muestra

La muestra e IRPF aporta cerca de 300 variables, tanto partidas económicas como otras variables numéricas y no numéricas, todas ellas contenidas como casillas en el modelo 100 de declaración del IRPF

La tabla 2-1 muestra las primeras partidas económicas de la muestra de IRPF.

par1	Numérico	8	2	Rdto. del trabajo Dinerarios
par2	Numérico	8	2	Retribuciones en especie (valoración)
par3	Numérico	8	2	Retribuciones en especie (ingresos a cuenta)
par4	Numérico	8	2	Retribuciones en especie (ingresos a cuenta repercutidos)
par5	Numérico	8	2	Rdto. del Trabajo En especie.
par6	Numérico	8	2	Contribuciones Planes Pensiones.
par7	Numérico	8	2	Aportaciones recibidas al patrimonio protegido de las personas con discapacidad del que es titular el contribuyente
par8	Numérico	8	2	Reducciones Art. 18 apartados 2 y 3, y dispos. trans. 11ª y 12ª Ley del Impuesto
par9	Numérico	8	2	Total ingresos integros computables [(01)+(05)+(06)+(07)-(08)]
par10	Numérico	8	2	Cotizac. Seguridad Social, Mutualidad Funcionarios, detracciones derechos pasivos y Coleg.Huérfanos.
par11	Numérico	8	2	Cuotas satisfechas a sindicatos
par12	Numérico	8	2	Cuotas satisfechas a colegios profesionales (si la colegiación es obligatoria y con un máximo de 500 euros anuales)
par13	Numérico	8	2	Gastos de defensa jurídica derivados directamente de litigios con el empleador (máximo: 300 euros anuales)
par14	Numérico	8	2	Gastos deducibles.
par15	Numérico	8	2	Rendimiento neto.Trabajo
par16	Numérico	8	2	Reducción de rendimientos acogidos al régimen especial "33.ª Copa del América" (disposición adicional séptima de la Ley
par17	Numérico	8	2	Reducción por obtención rdto. trabajo.Cuantía aplicable con carácter general.
par18	Numérico	8	2	Reducción por obtención rdto. trabajo.Incremento para trabajadores activos mayores de 65 años que continuen o prolonguen
par19	Numérico	8	2	Reducción por obtención rdto. trabajo.Incremento para contrib. desempleados que acepten un puesto que exija traslado de
par20	Numérico	8	2	Reducción por obtención rdto. trabajo.Reducción adicional para trabajadores activos que sean presonas con discapacidad.
par21	Numérico	8	2	Rendimiento neto reducido.Trabajo
par22	Numérico	8	2	Rend. Cap. Mobiliario. Intereses de cuentas, depósitos y activos financieros
par23	Numérico	8	2	Rend. Cap. Mobiliario. Intereses de activos financieros con bonificación

Tabla 2-1

Asimismo, la base de datos contiene otro tipo de variables numéricas y no numéricas relativas a sexo, edad, provincia, estado civil… de los declarantes. Estas variables permitirían estudiar el fraude con análisis de género, análisis geográfico, etc.

La tabla 2-2 muestra este tipo de variables.

cdpost	Numérico	8	2	Código postal
estcv	Cadena	3	0	Estado civil de declarante
sexo	Cadena	3	0	Sexo del declarante
dec	Cadena	3	0	Tipo de declaración
prov	Cadena	6	0	Provincia
ejnacd	Cadena	12	0	Ejercicio de nacimiento del declarante
ejnacc	Cadena	12	0	Ejercicio de nacimiento del cónyuge
minusd	Cadena	9	0	Grado de minusvalía del declarante
minusc	Cadena	9	0	Grado de minusvalía del cónyuge
nmdesc	Numérico	8	2	Número total de descendientes
nmdesc0	Numérico	8	2	Número de descendientes <3 años
nmdesc3	Numérico	8	2	Número de descendientes >= 3 y < 16 años
nmdesc16	Numérico	8	2	Número de descendientes >= 16 y < 18 años
nmdesc18	Numérico	8	2	Número de descendientes >= 18 y < 25 años
nmdescr	Numérico	8	2	Número de descendientes >=25 años
nmdescd	Numérico	8	2	Número de descendientes con edad desconocida
nmdesm0	Numérico	8	2	Número de descendientes sin minusvalía

Tabla 2-2

2.4 FASE DE EXPLORACIÓN DE LA INFORMACION

En nuestro caso, las variables que intervienen en los modelos de árboles de decisión son todas dicotómicas con valores cero y uno. El valor cero corresponde a individuos que no defraudan por la citada causa de fraude y el valor uno corresponde a individuos que defraudan.

Dado que no hay valores missing en la muestra y que las variables son dicotómicas, no será necesario analizar ninguna técnica de análisis exploratorio de datos para las variables del modelo.

La tabla 2-3 muestra los primeros registros de la base de datos relativos a las variables del modelo.

id	f_capinm	f_tmg	f_nhijos	f_gastos	f_planp	f_aaee	marca
1919997,00	,00	1,00	,00	,00	1,00	1,00	1,00
1920051,00	,00	,00	1,00	1,00	,00	,00	1,00
1921327,00	,00	,00	,00	,00	,00	,00	,00
1921802,00	,00	,00	,00	,00	,00	,00	,00
1922582,00	,00	,00	,00	,00	,00	,00	,00
1923448,00	,00	,00	,00	,00	,00	,00	,00
1924037,00	,00	,00	,00	,00	,00	1,00	1,00
1924774,00	,00	,00	,00	,00	,00	1,00	1,00
1924907,00	,00	1,00	,00	,00	,00	,00	1,00
93081,00	,00	1,00	,00	,00	1,00	1,00	1,00
93378,00	1,00	1,00	,00	,00	1,00	1,00	1,00
114059,00	1,00	1,00	,00	,00	,00	1,00	1,00
131426,00	,00	,00	,00	,00	1,00	1,00	1,00
172766,00	,00	,00	,00	,00	,00	,00	,00
214367,00	,00	,00	1,00	,00	,00	,00	1,00
288025,00	,00	,00	,00	,00	,00	,00	,00
363350,00	1,00	1,00	,00	,00	,00	1,00	1,00
531793,00	1,00	,00	,00	,00	,00	1,00	1,00
563282,00	,00	,00	,00	,00	1,00	1,00	1,00
657412,00	,00	,00	,00	,00	,00	,00	,00
688185,00	,00	,00	,00	,00	,00	1,00	1,00
850963,00	,00	,00	,00	,00	,00	,00	,00
882297,00	,00	,00	,00	,00	,00	,00	,00
930120,00	,00	,00	,00	,00	,00	,00	,00
1185853,00	,00	,00	,00	,00	,00	,00	,00

Tabla 2-3

Los factores de fraude más comunes que aparecen definidos en la base de datos utilizada son los siguientes:

- La variable *f_tmg* indica fraude relativo al tipo marginal,
- la variable *f_capinm* indica fraude relativo a los rendimientos de capital inmobiliario,
- la variable *f_aaee* indica fraude relativo a la declaración de actividades económicas,
- la variable *f_gastos* indica fraude relativo a las deducciones de gastos,
- la variable *f_planp* indica fraude relativo a las declaraciones de planes de pensiones
- la variable *f_nhijos* indica fraude relativo a la declaración del número de hijos y ascendientes.

Estas son las causas de fraude más habituales, pero podrían considerarse de igual modo todas las causas de fraude que la Agencia Tributaria registra en la base de datos completa de IRPF una vez inspeccionada.

Por motivos de confidencialidad legalmente exigidos y escrupulosamente respetados en esta investigación, los datos muestrales de individuos defraudadores y no defraudadores, tanto para el fraude global como para las distintas causas de fraude, siguen la pauta real sin ser exactamente coincidentes con los datos concretos por motivos de confidencialidad. Además, se utiliza una base de datos totalmente anonimizada. En la práctica, serían defraudadores los individuos de la muestra que la inspección ha determinado fehacientemente como tales defraudadores, tanto globalmente, como por las distintas causas.

Dado que el fraude debe de ser fehaciente y que simultáneamente están abiertos a inspección por ley los últimos 4 años de declaraciones, la base de datos a utilizar deberá ser al menos de 5 ejercicios anteriores al actual (ultimo año del que se tiene datos). Si además, se consideran todos los problemas de fraude sujetos a recursos en los tribunales, cerrar una base de datos inspeccionada completa cuesta mucho tiempo.

No obstante, esta investigación es independiente del ejercicio de datos que se considere, ya que se busca una metodología para cuantificar la incidencia sobre el fraude global de las distintas causas de fraude. Como además la metodología se basa en un modelo predictivo, se obtiene una función de predicción de fraude que es válida para varios ejercicios futuros consecutivos. No es necesario estimar el modelo que predice el fraude todos los años.

2.5 FASE DE TRANSFORMACIÓN DE LA INFORMACION

En nuestro modelo de árbol, tanto la variable dependiente como las independientes son categóricas, ya que modelizamos el fraude global (variable dependiente) en función de los factores de fraude más comunes en el Impuesto sobre la Renta de las Personas Físicas (variables independientes). Además, todas las variables son binarias, ya que todas ellas se miden en término de fraude (categoría 1) o no fraude (categoría 0). Por lo tanto, podremos utilizar todas las tipologías de árboles y compararlas entre sí sin necesidad de aplicar ninguna transformación de las variables.

Por lo tanto, en este caso no será necesario utilizar reducción de la dimensión ni ninguna otra transformación de los datos iniciales similar.

2.6 FASES DE MODELIZACIÓN Y EVALUACIÓN: ESTIMACION, DIAGNOSIS Y EXTRACCIÓN DEL CONOCIMIENTO EN LOS ÁRBOLES DE DECISIÓN

En la aplicación que aquí se presenta, se utiliza la muestra de IRPF que proporciona el Instituto de Estudios Fiscales, para construir un modelo de árbol de decisión tomando como variable dependiente una variable dicotómica que toma el valor 1 si el individuo defrauda y el valor cero si el individuo no defrauda (variable *marca*).

Las variables independientes son variables dicotómicas que toman el valor 1 para individuos que defraudan por una determinada causa de fraude y el valor cero si no defraudan por esa causa de fraude.

De esta forma modelizaremos el fraude global en función de las distintas causas de fraude y ordenaremos estas causas de acuerdo a su incidencia en el fraude global (variables que comienzan por f_).

Como ya hemos visto anteriormente, la variable *f_tmg* indica fraude relativo al tipo marginal, la variable *f_capinm* indica fraude relativo a los rendimientos de capital inmobiliario, la variable *f_aaee* indica fraude relativo a la declaración de actividades económicas, la variable *f_gastos* indica fraude relativo a las deducciones de gastos, la variable *f_planp* indica fraude relativo a las declaraciones de planes de pensiones y la variable *f_nhijos* indica fraude relativo a la declaración del número de hijos. Es importante volver a remarcar que quí se han considerado las causas de fraude más comunes, pero podrían considerarse todas las causas de fraude que la Agencia Tributaria registra en la base de datos completa de IRPF una vez inspeccionada y que habitualmente no proporciona a los usuarios.

La tabla 2-4 muestra en la base de datos las características de variables de fraude más comúnmente utilizadas.

id	Numérico	8	2	Identificador del perceptor
f_tmg	Numérico	8	2	Fraude que afecta al tipo marginal
f_capinm	Numérico	8	2	Fraude que afecta a los rendimientos de capital inmobiliario
f_nhijos	Numérico	8	2	Fraude que afecta a las declaraciones de número e hijos y ascendieentes y descendientes
f_aaee	Numérico	8	2	Fraude que afecta a la declaración de actividaes económicas
f_planp	Numérico	8	2	Fraude que afecta a la desgrbación por planes de pensiones
f_gastos	Numérico	8	2	Fraude que afecta a la declarción de gastos
marca	Numérico	8	2	fraude global

Tabla 2-4

2.6.1 Modelo CHAID Exhaustivo

Como ya se ha indicado anteriormente, dado que nuestro modelo tiene todas sus variables (dependiente e independientes) categóricas binarias, será posible utilizar cualquiera de las tipologías de árboles citadas en el apartado de metodología.

Comenzaremos considerando un árbol CHAID exhaustivo. En la figura 2-1 se muestra el árbol de decisión estimado para este método.

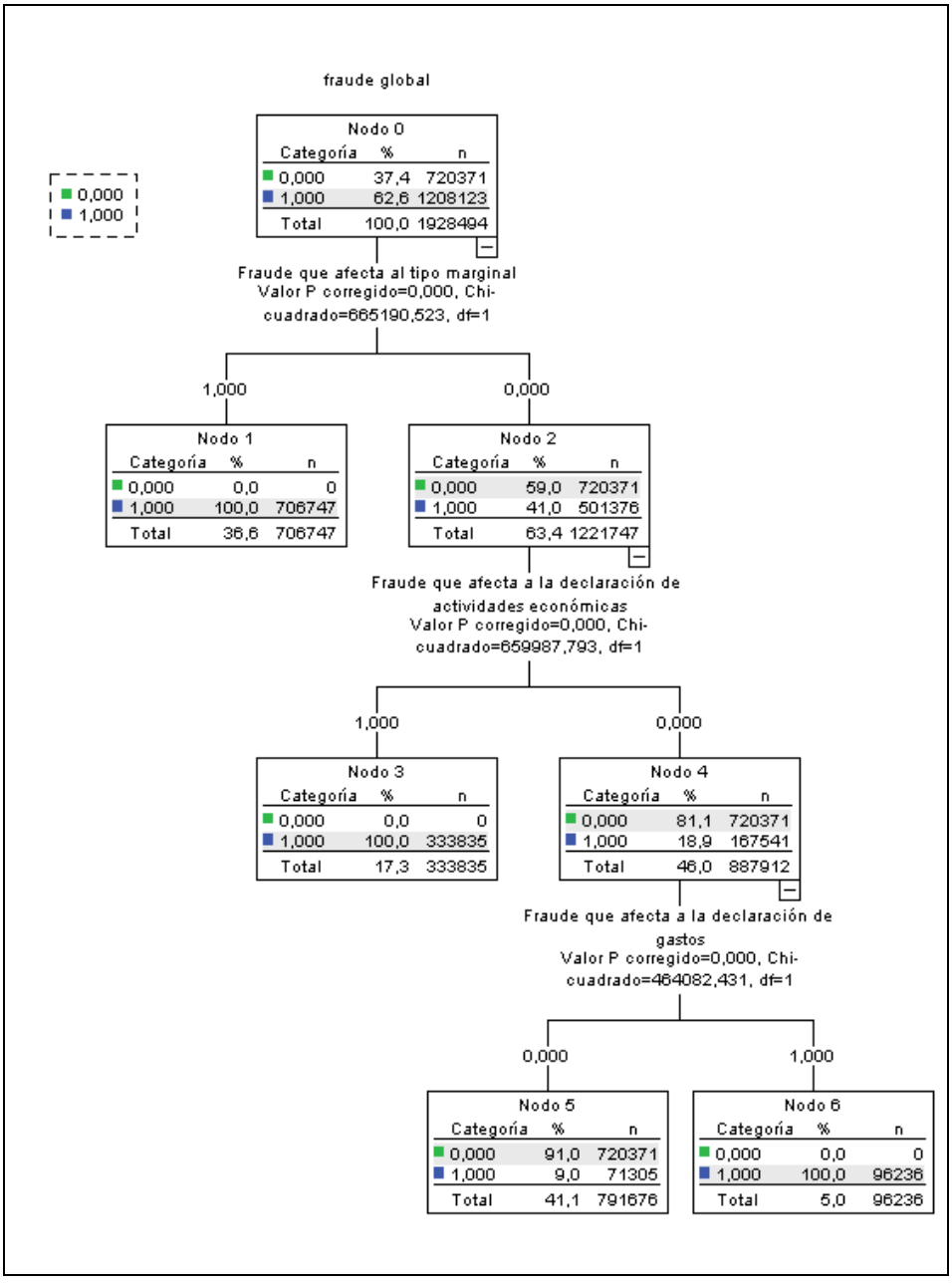

fraude global
0,000
1,000
Nodo 0
Categoría % n
0,000 37,4 720371
1,000 62,6 1208123
Total 100,0 1928494
Fraude que afecta al tipo marginal
Valor P corregido=0,000, Chi-cuadrado=665190,523, df=1
1,000
0,000
Nodo 1
Categoría % n
0,000 0,0 0
1,000 100,0 706747
Total 36,6 706747
Nodo 2
Categoría % n
0,000 59,0 720371
1,000 41,0 501376
Total 63,4 1221747
Fraude que afecta a la declaración de actividades económicas
Valor P corregido=0,000, Chi-cuadrado=659987,793, df=1
1,000
0,000
Nodo 3
Categoría % n
0,000 0,0 0
1,000 100,0 333835
Total 17,3 333835
Nodo 4
Categoría % n
0,000 81,1 720371
1,000 18,9 167541
Total 46,0 887912
Fraude que afecta a la declaración de gastos
Valor P corregido=0,000, Chi-cuadrado=464082,431, df=1
0,000
1,000
Nodo 5
Categoría % n
0,000 91,0 720371
1,000 9,0 71305
Total 41,1 791676
Nodo 6
Categoría % n
0,000 0,0 0
1,000 100,0 96236
Total 5,0 96236

Figura 2-1

Se observa que la causa principal de fraude es la manipulación del tipo marginal aplicable. Suele ser habitual la presencia de actividades cuyas rentas eluden la tributación, bien por no ser declaradas o bien por no estar registradas constituyendo economía sumergida. De esta forma, el tipo marginal correspondiente a la declaración resulta inferior al real, manipulándose así el resultado de la liquidación. Las cuantías defraudadas por esta causa suelen ser de elevada magnitud.

En cuanto a la diagnosis, la significatividad del fraude relativo al tipo marginal es muy alta en el modelo de árbol CHAID ya que el p-valor del estadístico Chi-cuadrado es muy bajo (casi nulo).

La segunda causa en importancia de incidencia en el fraude resulta ser la declaración incorrecta de actividades económicas, bien sea por ocultación de sus rentas, o bien sea por el no registro fraudulento de las mismas que las llevan a formar parte también de la economía sumergida. El difícil control de las rentas de autónomos lleva habitualmente a ingresos no declarados de sus actividades. A pesar del control exhaustivo de la inspección sobre la declaración de las actividades económicas, pidiendo libros de contabilidad y documentación asociada, el fraude no se mitiga lo suficiente. El p-valor del estadístico Chi-cuadrado para esta causa de fraude es muy bajo, con lo cual es muy significativa en el modelo de árbol CHAID.

La tercera causa de fraude resulta ser la incorrecta declaración de gastos desgravables, bien sea por la inobservancia de las normas legales o por la realización de artificios engañosos para eludirlas. En el caso de las subvenciones y ayudas, la mayor parte se registran como gasto desde el lado del pagador, existiendo otras que se registran minorando ingresos impositivos o de cotizaciones sociales. La incorrecta aplicación de la normativa puede llevar en esta caso a la deducción de gastos incorrecta o al fraude por manipulación del tipo marginal aplicable. En el caso de los gastos financieros, se observa que en la normativa del IRPF, los gastos financieros relativos a

viviendas no alquiladas o a inversiones financieras no resultan deducibles, siendo esta norma habitualmente vulnerada. En general, la incorrecta declaración de gastos derivados de actividades económicas suele ser otra fuente de fraude. Esta causa de fraude también presenta p-valor nulo, lo cual indica su alta significatividad en el modelo de árbol CHAID.

Observamos que el método CHAID exhaustivo considera variables no significativas para determinar el fraude global las siguientes causas: el fraude en la desgravación por planes de pensiones, el fraude en la declaración del número de hijos, ascendientes y descendientes y el fraude derivado de la declaración de los rendimientos del capital inmobiliario. Por esta razón consideraremos métodos más potentes para construir árboles, como es el caso del método CRT.

Los resultados de la diagnosis del modelo CHAID (tabla 2-5) presentan una estimación de la función de riesgo muy baja con un error despreciable. Asimismo, se obtiene una matriz de confusión con porcentajes correctos de pronóstico muy alto.

Riesgo

Estimación	Error estándar
,037	,000

Método de crecimiento: EXHAUSTIVE CHAID
Variable dependiente: fraude global

Clasificación

Observado	Pronosticado		Porcentaje correcto
	,00	1,00	
,00	720371	0	100,0%
1,00	71305	1136818	94,1%
Porcentaje global	41,1%	58,9%	96,3%

Método de crecimiento: EXHAUSTIVE CHAID
Variable dependiente: fraude global

Tabla 2-5

En cuanto a la curva ROC del modelo (Figura 2-2), observamos que encierra un área de 0,970 (tabla 2-6), que resulta ser muy alto por estar muy cercano a uno, lo que indica que el modelo CHAID exhaustivo es adecuado.

Área bajo la curva

Área
,970

Tabla 2-6

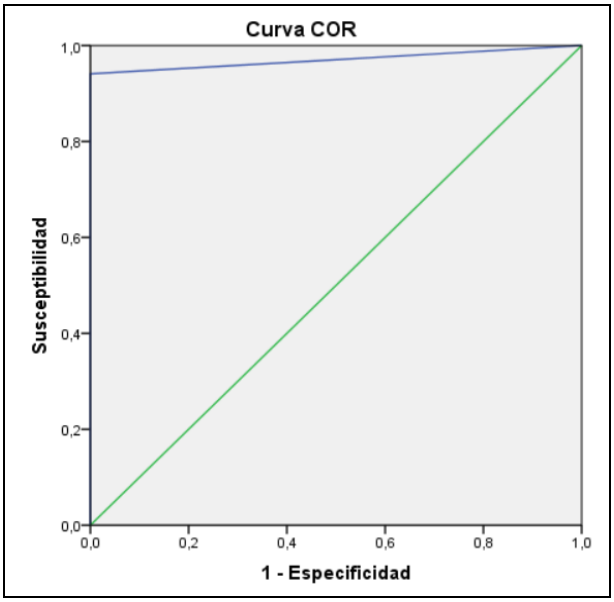


Figura 2-2

En cuanto a la propensión al fraude, el modelo de árbol CHAID exhaustivo también permite calcular las probabilidades de fraude (*PredictedProbability_2*) y no fraude (*PredictedProbability_1*) para cada individuo. También predice el segmento de fraude o no fraude (*Predictedvalue*) en el que se clasifica cada individuo (*Predictedvalue* = 1 indica que el individuo defrauda y *Predictedvlue* = 0 indica que el individuo no defrauda). La tabla 2-7 muestra los primeros registros.

PredictedValue	PredictedProbability_1	PredictedProbability_2
1,0000	,0000	1,0000
1,0000	,0000	1,0000
,0000	,9099	,0901
,0000	,9099	,0901
,0000	,9099	,0901
,0000	,9099	,0901
1,0000	,0000	1,0000
1,0000	,0000	1,0000

Tabla 2-7

Habitualmente el árbol de decisión no es el mejor instrumento para predecir probabilidades de fraude de los individuos, ya que los valores de esas probabilidades suelen ser muy extremos. Por la misma razón la curva ROC es muy lineal. No obstante el segmento de clasificación del individuo como fraudulento o no, suele ser muy acertado.

Para analizar los perfiles de fraude de los individuos calculamos la función de densidad de la probabilidad de fraude, pero al ser las probabilidades tan extremas no tiene mucho sentido considerar esta función de densidad.

Esto ocurre siempre con los árboles de decisión. Son una herramienta magnífica para discriminar entre las causas de fraude, pero no es la mejor herramienta para calcular probabilidades de fraude.

2.6.2 Modelo CRT (Classification Regresion Tree)

Consideraremos ahora un árbol CRT. En la figura 2-3 se muestra el árbol de decisión estimado para este método.

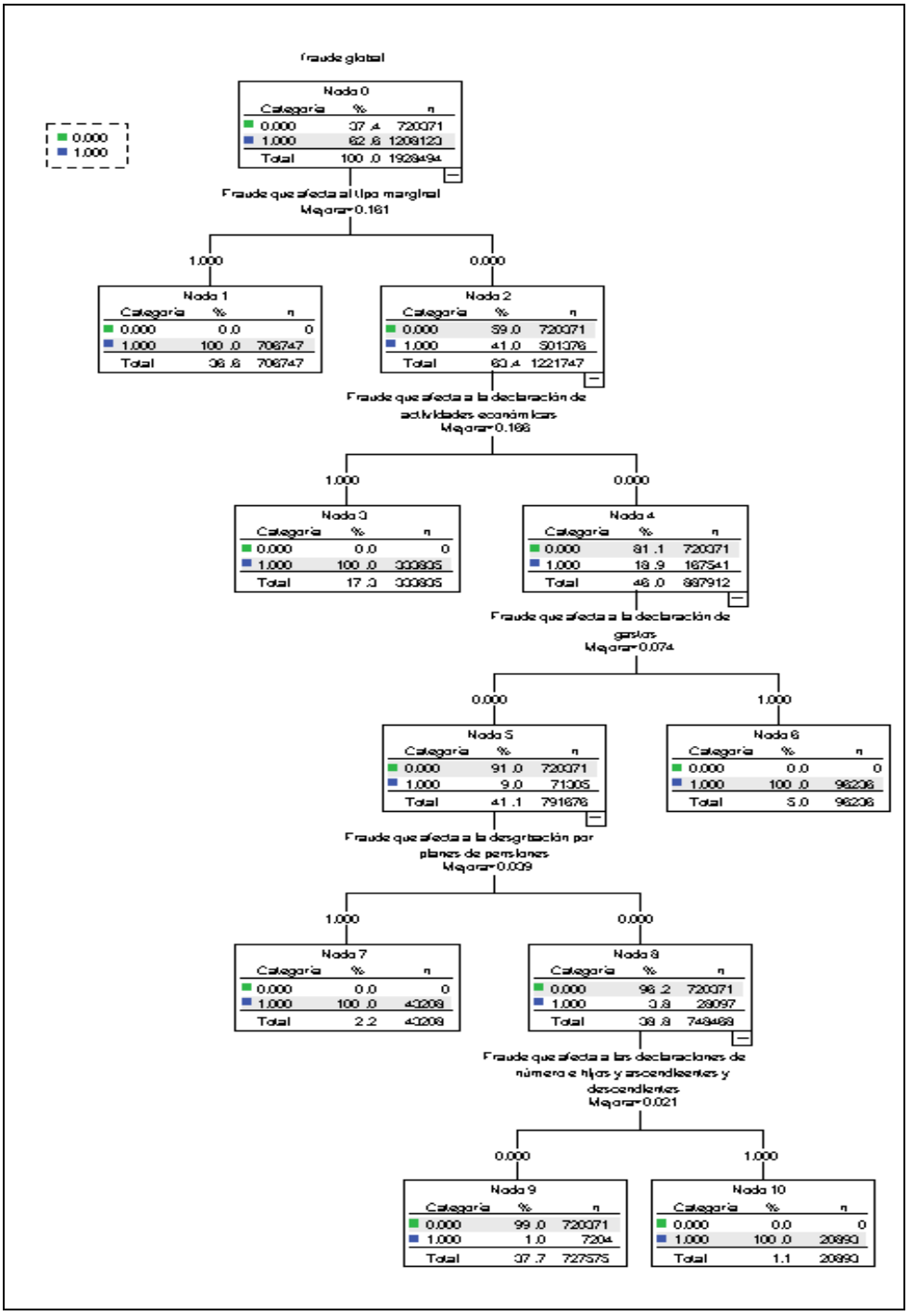

Fraude global
0.000
1.000
Nodo 0
Categoría % n
0.000 37.4 720071
1.000 62.6 1208123
Total 100.0 1928494
Fraude que afecta al tipo marginal
Mejora=0.161
1.000
0.000
Nodo 1
Categoría % n
0.000 0.0 0
1.000 100.0 706747
Total 36.6 706747
Nodo 2
Categoría % n
0.000 59.0 720071
1.000 41.0 501376
Total 63.4 1221747
Fraude que afecta a la declaración de actividades económicas
Mejora=0.166
1.000
0.000
Nodo 3
Categoría % n
0.000 0.0 0
1.000 100.0 333935
Total 17.3 333935
Nodo 4
Categoría % n
0.000 81.1 720071
1.000 18.9 167541
Total 46.0 887912
Fraude que afecta a la declaración de ganias
Mejora=0.074
0.000
1.000
Nodo 5
Categoría % n
0.000 91.0 720071
1.000 9.0 71305
Total 41.1 791676
Nodo 6
Categoría % n
0.000 0.0 0
1.000 100.0 96236
Total 5.0 96236
Fraude que afecta a la desgrtación por planes de pensiones
Mejora=0.009
1.000
0.000
Nodo 7
Categoría % n
0.000 0.0 0
1.000 100.0 43208
Total 2.2 43208
Nodo 8
Categoría % n
0.000 96.2 720071
1.000 3.8 28097
Total 38.8 748468
Fraude que afecta a las declaraciones de número e hijas y ascendientes y descendientes
Mejora=0.021
0.000
1.000
Nodo 9
Categoría % n
0.000 99.0 720071
1.000 1.0 7204
Total 37.7 727575
Nodo 10
Categoría % n
0.000 0.0 0
1.000 100.0 20890
Total 1.1 20890

Figura 2-3

Se observa que la cuantificación de la incidencia de las distintas causas de fraude sobre el fraude global sigue el mismo orden que en el caso del árbol CHAID (tipo marginal, declaración de actividades económicas y declaración de gastos, por este orden). No obstante, al tratarse de un árbol de decisión más potente, observamos que incluye como significativas nuevas causas de fraude que no recogía el árbol CHAID. Además todas las causas de fraude son altamente significativas (p-valores muy pequeños).

La cuarta causa de fraude significativa (p-valor pequeño) por incidencia en el fraude global resulta ser el fraude que afecta a la desgravación de planes de pensiones. Esta rúbrica del IRPF fue durante un tiempo el refugio de las rentas altas, ya que desgrava de la base y además por cantidades importantes hasta que se acotó el máximo deducible. Por lo tanto era objeto de especial tratamiento por los declarantes de IRPF con peligro de deducciones fraudulentas ilegales que acentuó la vigilancia de la inspección.

La quinta causa de fraude significativa (p-valor bajo) resulta ser el fraude que afecta a las declaraciones del número de hijos y ascendientes y descendientes, que habitualmente eran simultáneamente desgravadas por los dos padres (separados, divorciados o en otras situaciones) en el caso de los hijos o por diferentes hermanos en el caso de los ascendientes.

Mediante este método, la única causa de fraude no significativa resulta ser el fraude relativo a la declaración de los rendimientos de capital inmobiliario.

La diagnosis del ajuste del árbol por el método CRT presenta la salida que muestra la tabla 2-8.

Riesgo

Estimación	Error estándar
,004	,000

Método de crecimiento: CRT
Variable dependiente: fraude global

Clasificación

Observado	Pronosticado		
	,00	1,00	Porcentaje correcto
,00	720371	0	100,0%
1,00	7204	1200919	99,4%
Porcentaje global	37,7%	62,3%	99,6%

Método de crecimiento: CRT
Variable dependiente: fraude global

Tabla 2-8

Se observa que el riesgo estimado tiene un valor muy pequeño (0,04), inferior al caso del árbol CHAID exhaustivo estimado anteriormente (0,037). También se obtiene una matriz de confusión con porcentajes correctos de pronóstico muy alto. Asimismo, se observa que estos porcentajes son más altos que en el caso del árbol CHAID exhaustivo. Esto nos lleva a concluir que el árbol CRT es superior al árbol CHAID exhaustivo para clasificar las causas de fraude según su incidencia en el fraude global.

En cuanto a la curva ROC del modelo de árbol CRT (figura 2-4), observamos que encierra un área de 0,997 (tabla 2-9), que resulta ser muy alto por estar muy cercano a uno, lo que indica que el modelo CRT es más preciso que el CHAID exhaustivo por ser el área bajo la curva ROC mayor (en el caso del árbol CHAID el área 0,970).

Área bajo la curva

Variables resultado de contraste: Predicted Probability for marca=1

Área	Error típ.[a]	Sig. asintótica[b]	Intervalo de confianza asintótico al 95%	
			Límite inferior	Límite superior
,997	,000	,000	,997	,997

Tabla 2-9

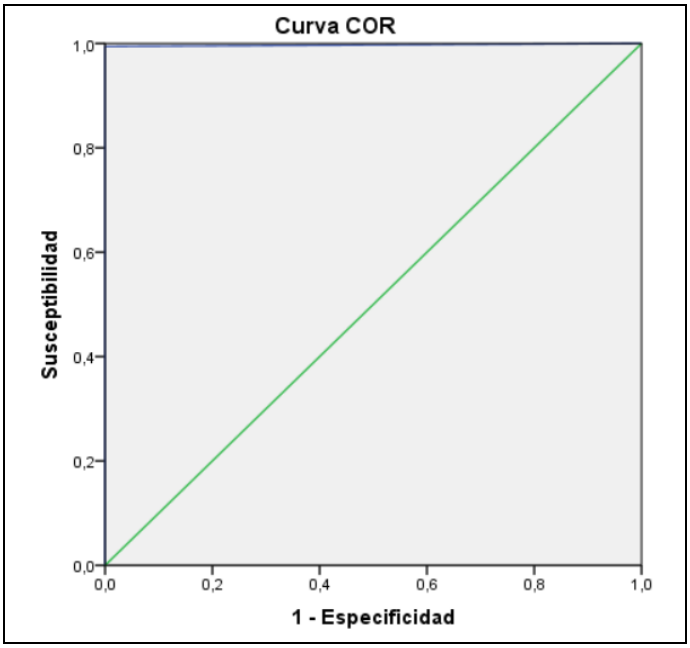


Figura 2-4

En cuanto a la propensión al fraude, el modelo de árbol CRT también permite calcular las probabilidades de fraude (*PredictedProbability_2_1*) y no fraude (*PredictedProbability_1_1*) para cada individuo. También predice el segmento de fraude o no fraude (*Predictedvalue_1*) en el que se clasifica cada individuo (*Predictedvalue_1* = 1 indica que el individuo defrauda y *Predictedvlue_1* = 0 indica que el individuo no defrauda). La tabla 2-10 muestra los primeros registros.

PredictedValue_1	PredictedProbability_1_1	PredictedProbability_2_1
1,0000	,0000	1,0000
1,0000	,0000	1,0000
,0000	,9901	,0099
,0000	,9901	,0099
,0000	,9901	,0099
,0000	,9901	,0099
1,0000	,0000	1,0000
1,0000	,0000	1,0000
1,0000	,0000	1,0000

Tabla 2-10

Los valores de las probabilidades de fraude de los individuos en el árbol CRT vuelven a ser ser muy extremos, incluso más que en el árbol CHAID. Por la misma razón la curva ROC es todavía más lineal, confundiéndose prácticamente con el triángulo restángulo superior de la figura de la curva ROC, que sería el caso ideal de curva (curva ROC óptima). No obstante el segmento de clasificación del individuo como fraudulento o no sigue siendo muy acertado (más que en el árbol CHAID exhaustivo poque ahora la diagnosis es mejor).

Para analizar los perfiles de fraude de los individuos no tiene mucho sentido calcular la función de densidad de la probabilidad de fraude, por la misma razón que en caso del árbol CHAID.

Mediante el método CRT, la única causa de fraude no significativa resulta ser el fraude relativo a la declaración de los rendimientos de capital inmobiliario. Ante este resultado, intentaremos mejorar este modelo de árbol aplicando adicionalmente el método QUEST, que metodológicamente podría superar al método CRT.

2.6.3 Modelo QUEST

El árbol estimado obtenido por el método QUEST se muestra en la figura 2-5.

Efectivamente se observa que la clasificación de la incidencia de las distintas causas de fraude significativas sobre el fraude global coincide en los métodos CRT y QUEST.

Además, el método QUEST no introduce nuevas variables significativas en el modelo de árbol, resultando finalmente que la causa de fraude relativa a la declaración de rendimientos de capital inmobiliario no es significativamente incidente en el fraude global.

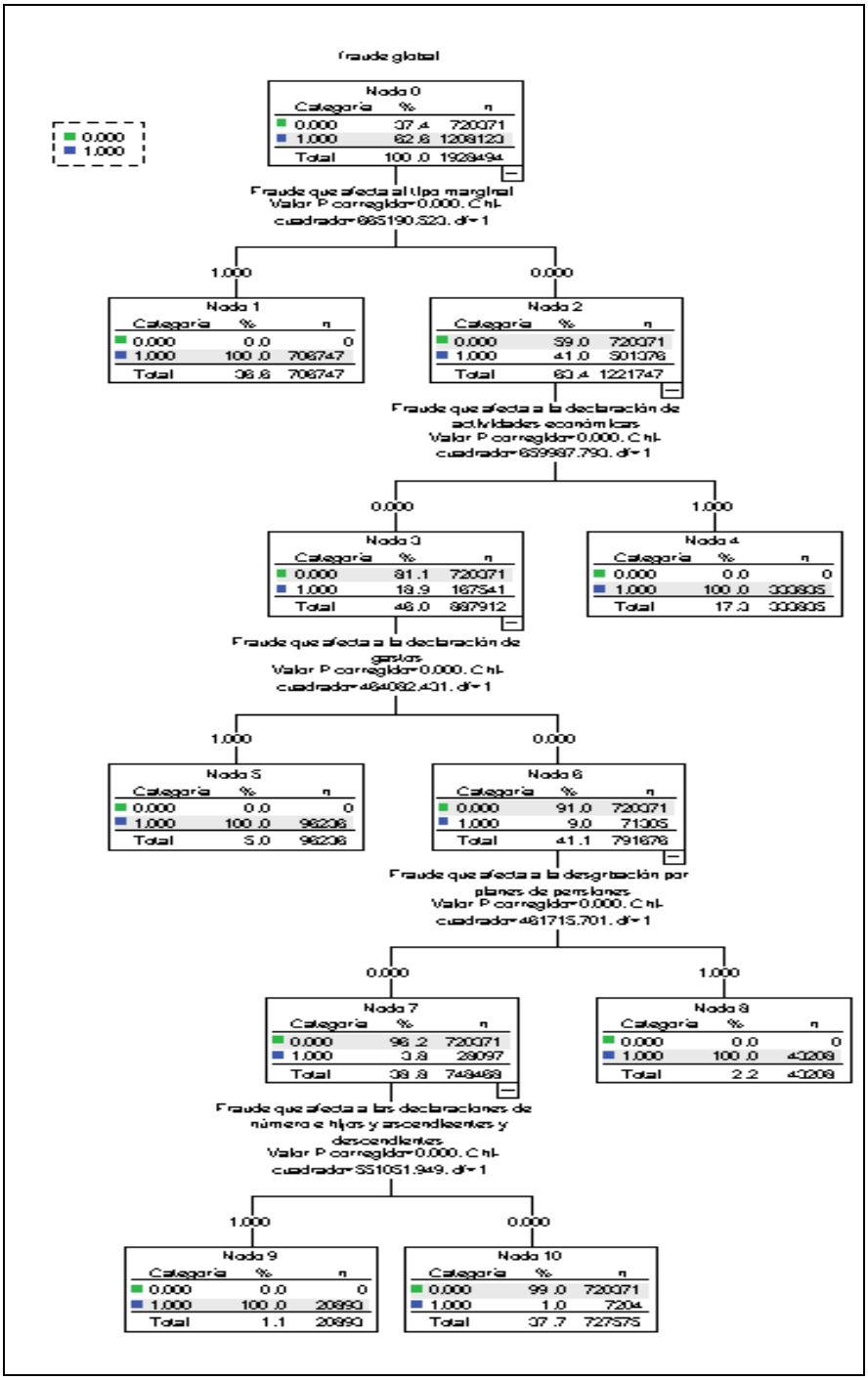

fraude global
0,000
1,000
Nodo 0
Categoría % n
0,000 37,4 720071
1,000 62,6 1208123
Total 100,0 1928194
Fraude que afecta al tipo marginal
Valor P corregido=0,000. Chi-cuadrado=865190,523. df=1
1,000
0,000
Nodo 1
Categoría % n
0,000 0,0 0
1,000 100,0 706747
Total 36,6 706747
Nodo 2
Categoría % n
0,000 59,0 720071
1,000 41,0 501376
Total 63,4 1221747
Fraude que afecta a la declaración de actividades económicas
Valor P corregido=0,000. Chi-cuadrado=659987,790. df=1
0,000
1,000
Nodo 3
Categoría % n
0,000 81,1 720071
1,000 18,9 167541
Total 46,0 887912
Nodo 4
Categoría % n
0,000 0,0 0
1,000 100,0 333835
Total 17,3 333835
Fraude que afecta a la declaración de ganancias
Valor P corregido=0,000. Chi-cuadrado=464082,431. df=1
1,000
0,000
Nodo 5
Categoría % n
0,000 0,0 0
1,000 100,0 96236
Total 5,0 96236
Nodo 6
Categoría % n
0,000 91,0 720071
1,000 9,0 71305
Total 41,1 791676
Fraude que afecta a la desgravación por planes de pensiones
Valor P corregido=0,000. Chi-cuadrado=461715,701. df=1
0,000
1,000
Nodo 7
Categoría % n
0,000 96,2 720071
1,000 3,8 28097
Total 38,8 748468
Nodo 8
Categoría % n
0,000 0,0 0
1,000 100,0 43208
Total 2,2 43208
Fraude que afecta a las declaraciones de número e hijos y ascendientes y descendientes
Valor P corregido=0,000. Chi-cuadrado=551051,949. df=1
1,000
0,000
Nodo 9
Categoría % n
0,000 0,0 0
1,000 100,0 20890
Total 1,1 20890
Nodo 10
Categoría % n
0,000 99,0 720071
1,000 1,0 7204
Total 37,7 727575

Figura 2-5

Los resultados de la diagnosis para el método QUEST se presentan en la tabla 2-11.

Riesgo

Estimación	Error estándar
,004	,000

Método de crecimiento: QUEST
Variable dependiente: fraude global

Clasificación

Observado	Pronosticado		Porcentaje correcto
	,00	1,00	
,00	720371	0	100,0%
1,00	7204	1200919	99,4%
Porcentaje global	37,7%	62,3%	99,6%

Método de crecimiento: QUEST
Variable dependiente: fraude global

Tabla 2-11

Se observa un riesgo estimado de magnitud 0,04 exactamente igual al riesgo estimado en el método CRT. Asimismo, se observa una matriz de confusión con probabilidades de porcentaje correcto pronosticado exactamente igual al caso del árbol. Este resultado nos lleva a pensar que la clasificación de la importancia de las causas de fraude sobre el fraude global no sufrirá variaciones y coincidirá por ambos métodos.

En cuanto a la curva ROC del modelo de árbol QUEST (figura 2-6), observamos que encierra un área de 0,997 (tabla 2-12), exactamente igual al caso del árbol CRT. Las curvas ROC también son coincidentes, lo que vuelve a indicar que los modelos CRT y QUEST son muy similares e igualmente precisos, pero ambos más precisos que el modelo CHAID exhaustivo.

Área bajo la curva

Variables resultado de contraste: Predicted Probability for marca=1

Área	Error típ.[a]	Sig. asintótica[b]	Intervalo de confianza asintótico al 95%	
			Límite inferior	Límite superior
,997	,000	,000	,997	,997

Tabla 2-12

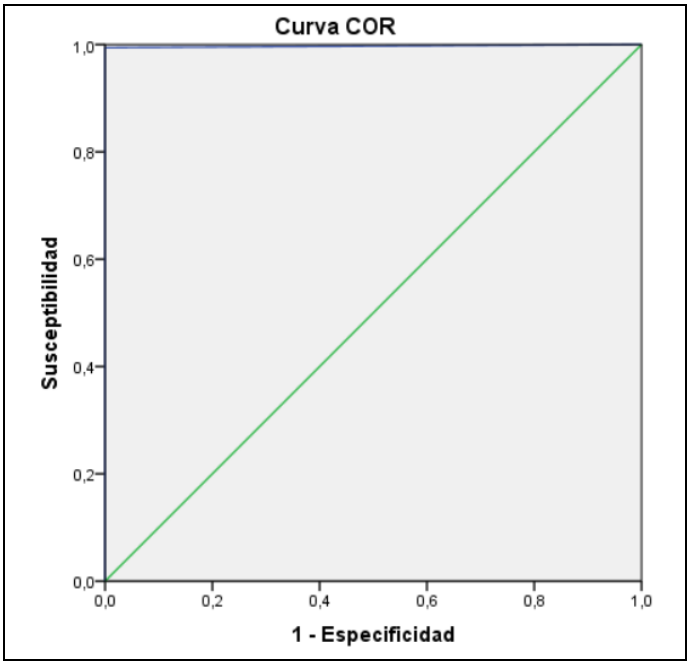


Figura 2-6

En cuanto a la propensión al fraude, el modelo de árbol QUEST también permite calcular las probabilidades de fraude (*PredictedProbability_2_2*) y no fraude (*PredictedProbability_1_2*) para cada individuo. También predice el segmento de fraude o no fraude (*Predictedvalue_2*) en el que se clasifica cada individuo (*Predictedvalue_2* = 1 indica que el individuo defrauda y *Predictedvlue_2* = 0 indica que el individuo no defrauda). La tabla 2-13 muestra los primeros registros.

PredictedValue_2	PredictedProbability_1_2	PredictedProbability_2_2
1,0000	,0000	1,0000
1,0000	,0000	1,0000
,0000	,9901	,0099
,0000	,9901	,0099
,0000	,9901	,0099
,0000	,9901	,0099
1,0000	,0000	1,0000

Tabla 2-13

Los valores de las probabilidades de fraude de los individuos en el árbol QUEST resultan ser los mismos que que en el árbol CRT, lo que corrobora la semejanza de estas técnicas..

Para analizar los perfiles de fraude de los individuos no tiene mucho sentido calcular la función de densidad de la probabilidad de fraude, por la misma razón que en caso de los árboles CRT y CHAID.

2.7 SEGMENTACIÓN DE LAS CAUSAS DE FRAUDE A TRAVÉS DE LOS ÁRBOLES DE DECISIÓN

Para segmentar las causas de fraude utilizaremos las técnicas estadísticas del Escalamiento Multidimensional y Analisis Cluster por variables.

2.7.1 Escalamiento Multidimensional

El Escalamiento multidimensional es una técnica descriptiva de minería de datos que permite segmenntar variables de un conjunto de datos agrupándolas por similitud en un mapa perceptual.

Si aplicamos escalamiento multidimensional para todas las variables de fraude, obtenemos el mapa perceptual de la figura 2-7.

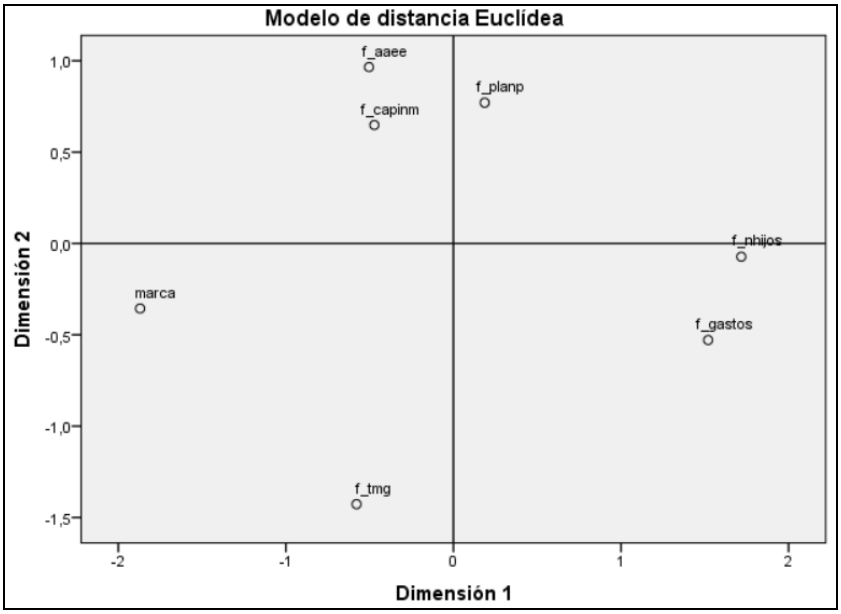


Figura 2-7

La segmentación del mapa perceptual nos indica que el fraude en actividades económicas, en planes de pensiones y en rendimientos capital inmobiliario tienen un comportamiento similar. Lo mismo ocurre con el fraude por declaración incorrecta de gastos y número de hijos y ascendientes. El fraude en la alteración del tipo marginal se comporta más aisladamente.

También es importante observar que el fraude por tipo marginal es el más cercano en el mapa al fraude global (marca), lo que indica que será el más incidente en el fraude global, tal y como ya habíamos visto anteriormente.

El fruade por actividades económicas ocupa el segundo lugar en cercanía al fraude global en el mapa perceptual. Ello indica que el fraude en actividades económicas ocupa el segundo lugar en influencia sobre el fraude global.

El fraude por incorrecta declaración de gastos ocupa el tercer lugar en cercanía al fraude global, luego será la tercera causa de fraude en incidencia sobre el fraude lugar.

El cuarto lugar en incidencia sobre el fraude global lo ocupa el fraude en planes de pensiones y el quinto lugar lo ocupa el fraude por incorrecta declaración del número de hijos y ascendientes.

Para evaluar este tipo de escalamiento se utiliza el gráfico de disparidades de la figura 2-8.

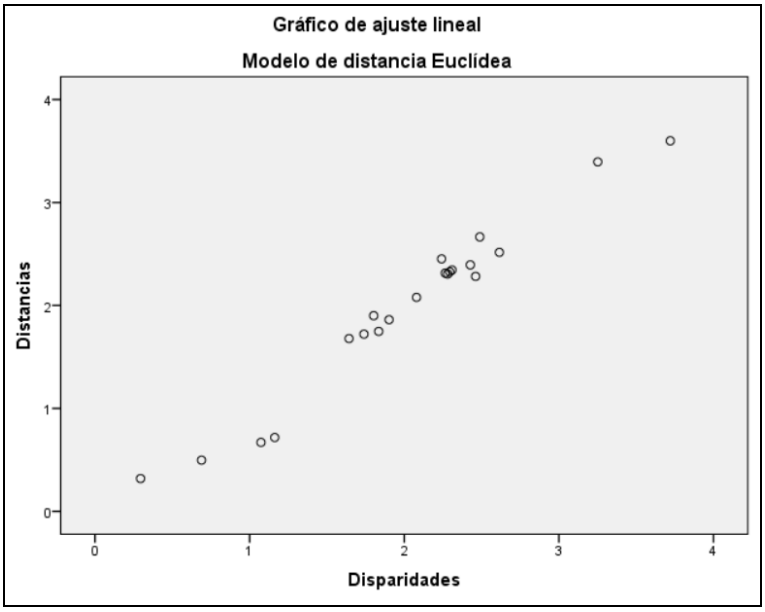


Figura 2-8

El escalamiento multidimensional utilizado es correcto porque el gráfico de disparidades presenta una nube de puntos que se ajusta bien a la diagonal del primer cuadrante.

Además, el estadístico S-Stress toma un valor bajo cercano a cero y el estadístico RSQ toma un valor alto cercano a la unidad

```
Stress  =   ,07669      RSQ =  ,96808
```

Concluimos que la segmentación realizada de las causa de fraude es correcta.

También podemos aplicar escalamiento multidimensional para ver como se relacionan las distintas causas de fraude con la probabilidad de fraude. Al realizar el citado escalamiento multidimnensional para las probabilidades de fraude obtenidas por el método CHAID exhaustivo, obtenemos el mapa perceptual de la Figura 2-9.

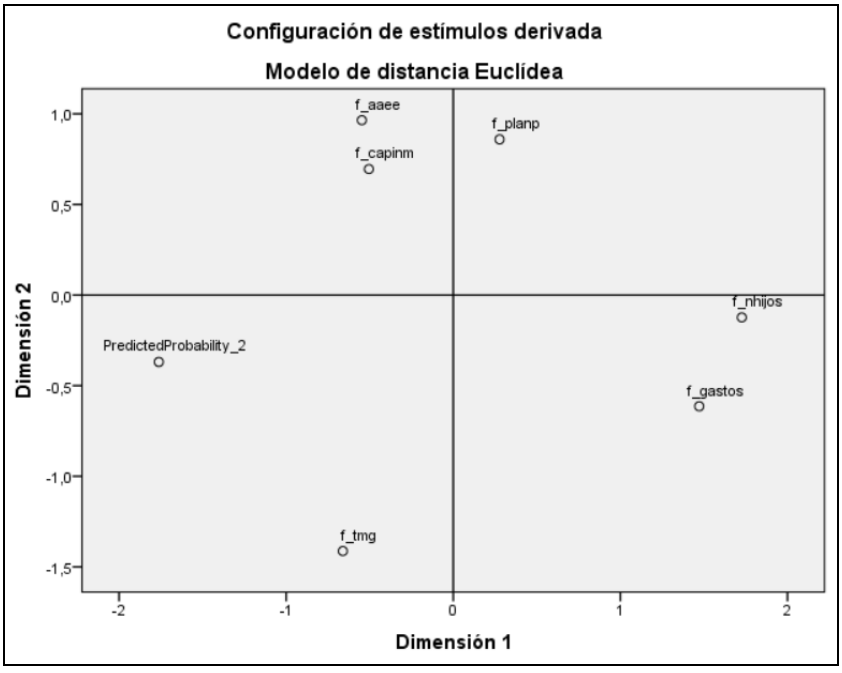


Figura 2-9

Se observa que la segmentación de la incidencia de las distintas causas de fraude sobre el fraude global coincide con la segmentación de las causa de fraude con la probabilidad de fraude por el método CHAID exhaustivo. El mapa es prácticamente igual al sustituir el fraude global por la probabilidad de fraude.

El escalamiento multidimensional vuelve a ser muy bueno, ya que el valor del estadístico S-Stress es bajo y el valor del esatdístico RSQ es muy alto.

```
Stress  =   ,08550      RSQ =  ,96040
```

No obstante, estos valores indican que la diagnosis del escalamiento anterior es ligeramente mejor que este. Es más efectivo valorar la incidencia de las distintas causas de fraude sobre el fraude global que sobre la probabilidad de fraude.

En este caso la nube de puntos del gráfico de disparidades (Figura 2-10) también se ajusta bien a una recta, indicando que el escalamiento es efectivo.

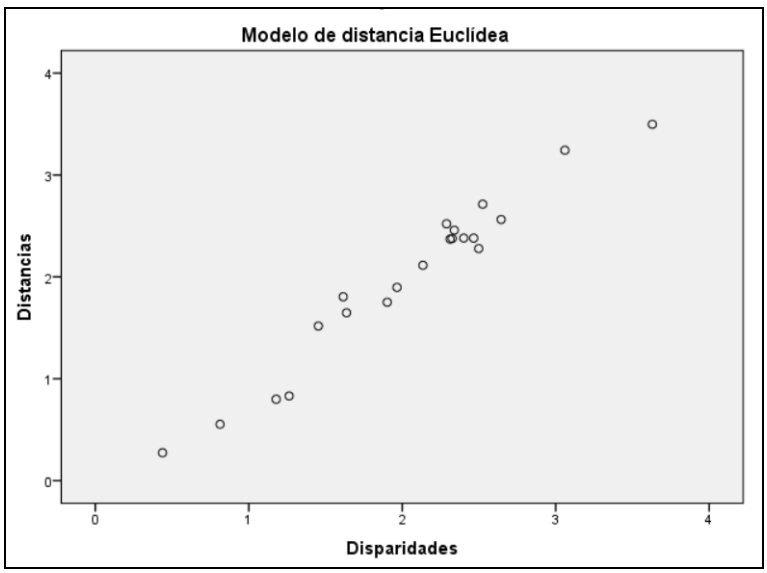


Figura 2-10

Realizamos ahora la segmentación de las causas de fraude según su relación con la probabilidad de fraude utilizando el método CRT. Se obtiene el mapa perceptual de la figura 2-11.

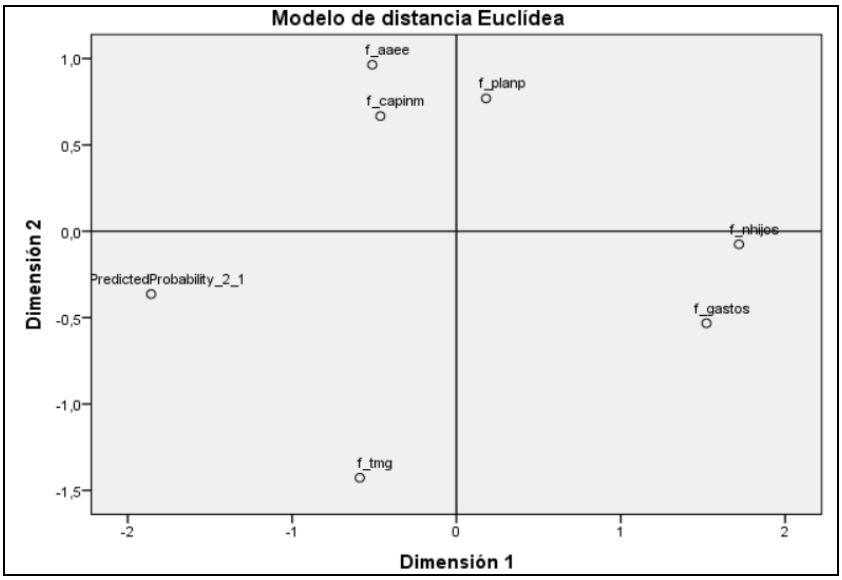


Figura 2-11

El mapa perceptual para el método CRT sigue siendo prácticamente el mismo que para el caso del método CHAID exhaustivo. Los estadísticos de diagnosis también presentan valores similares.

```
Stress  =   ,07828      RSQ =  ,96726
```

La nube de puntos de la gráfica de disparidades (Figura 2-12) también se ajusta bien a la diagonal del primer cuadrante.

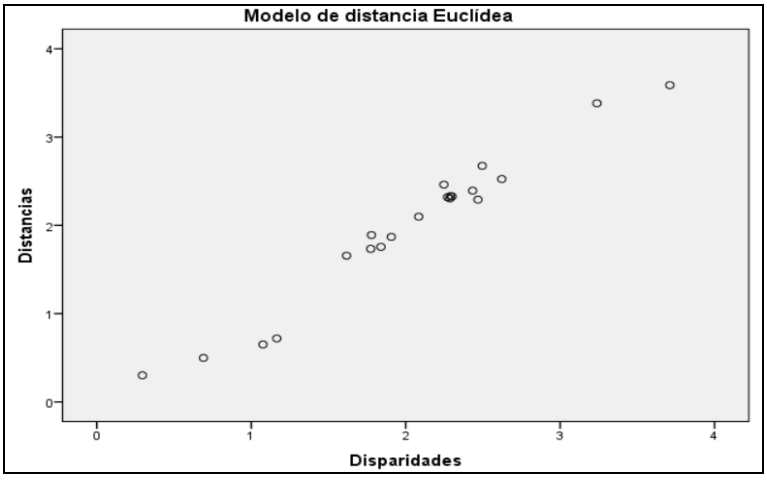


Figura 2-12

Para el método QUEST no es necesario realizar el ejercicio ya que las probabilidades de fraude coincidían con las del método CRT. Por lo tanto las segmentaciones serán exactamente iguales.

Siendo precisos, la segmentación de las causas de fraude mediante los métodos QUEST y CRT es más precisa que utilizando el método CHAID exhaustivo (el valor del estadístico S-Stress es más pequeño y el de RSQ más elevado). Este resultado lo habíamos obtenido ya anteriormente.

2.7.2 Análisis Cluster

El Análisis clúster es otra técnica descriptiva de minería de datos que permite segmentar variables de un conjunto de datos agrupándolas por similitud en un dendograma. Si aplicamos análisis clústerjerárquico por el método de Ward para todas las variables que inciden en el fraude global, obtenemos el dendograma de la figura 2-13:

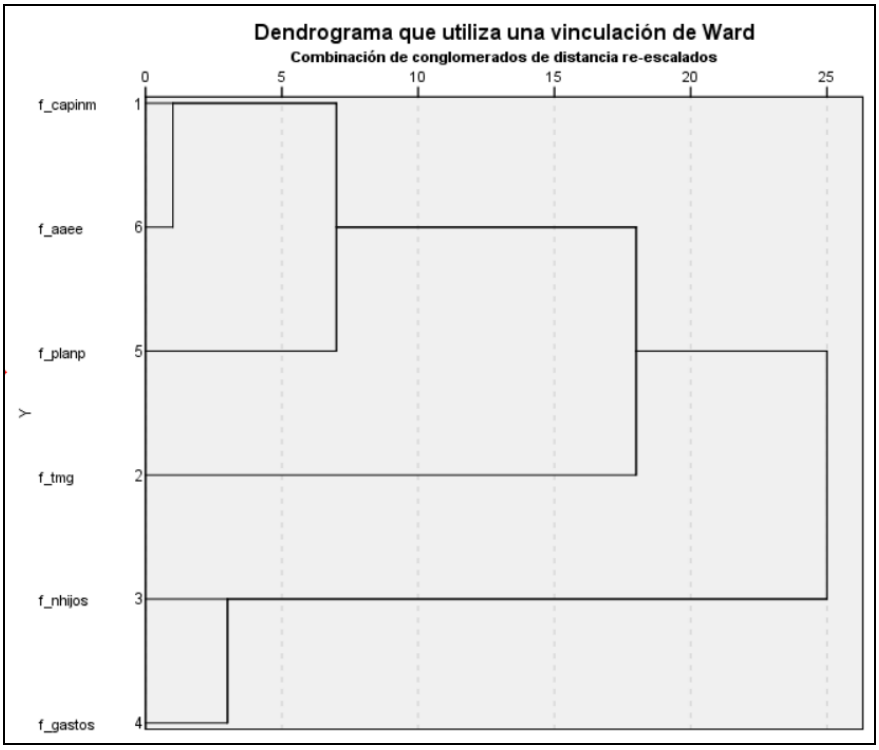


Figura 2-13

Se observa que la segmentación de las causas de frude es idéntica a la obtenida mediante escalamiento multidimensional. Se comportan de modo similar, por pertenecer al mismo cluster de nivel dos, el fraude en actividades económicas, el fraude en planes de pensiones y el fraude en rendimientos de capital inmobiliario. También se comportan de modo similar el fraude por incorrecta deducción de gastos y por declaración del número de hijos y ascendientes. El fraude por tipo marginal está aislado en el cluster porque su comportamiento es independiente del resto de las causas de fraude, resultado que también conocíamos.

Capítulo 3

FACTORES QUE AFECTAN AL FRAUDE FISCAL A TRAVÉS DE TÉCNICAS DE MINERÍA DE DATOS. APLICACIÓN AL IRPF MEDIANTE ANÁLISIS DISCRIMINANTE

3.1 INTRODUCCIÓN

En este capítulo se elaborarán modelos predictivos que permiten cuantificar la probabilidad que tiene cualquier contribuyente actual o futuro de ser defraudador por cada factor o causa de fraude una vez que presente su declaración de IRPF, basándose en los datos de la muestra de IRPF.

A partir de la muestra de IRPF y mediante la aplicación de funciones discriminantes de Fisher se buscarán modelos de análisis discriminante que permitan asignar una probabilidad de defraudar por cualquier causa de fraude a cualquier declarante de IRPF actual o futuro basándose exclusivamente en la información que declara a la Agencia tributaria en el modelo correspondiente.

De esta forma, los modelos discriminantes permitirán segmentar a los declarantes del impuesto por nivel de propensión al fraude y causas del mismo. También se estimarán perfiles de fraude de los contribuyentes y se segmentarán por afinidad las causa de fraude.

Esta metodología es generalizable para cuantificar la propensión al fraude en cualquier otro impuesto según los factores que lo determinan y para cuantificar y ordenar la incidencia de dichos factores en el fraude.

3.2 MARCO METODOLÓGICO: LOS MODELOS DE ANÁLISIS DISCRIMINANTE

El análisis discriminante es una técnica que tiene como finalidad construir un modelo predictivo para pronosticar el grupo al que pertenece una observación a partir de determinadas características observadas que delimitan su perfil. Se trata de una técnica estadística que permite asignar o clasificar nuevos individuos u observaciones dentro de grupos o segmentos previamente definidos (valores de la variable categórica dependiente), razón por la cual es una técnica de clasificación y segmentación ad hoc. El análisis discriminante se conoce en ocasiones como análisis de la clasificación, ya que su objetivo fundamental es producir una regla o un esquema de clasificación que permita a un investigador predecir la población a la que es más probable que tenga que pertenecer una nueva observación o individuo.

Las dos grandes finalidades perseguidas en el uso del análisis discriminante son la clasificación de los individuos en los grupos de la variable categórica dependiente y la predicción de pertenencia a los citados grupos.

Las características usadas para realizar esta clasificación de individuos en grupos reciben el nombre de *variables discriminantes*. La predicción de pertenencia a los grupos se lleva a cabo determinando una o más ecuaciones

matemáticas, denominadas *funciones discriminantes*, que permitan la clasificación de nuevos casos a partir de la información que poseemos sobre ellos. Estas ecuaciones combinan una serie de características o variables de tal modo que su aplicación a un caso nos permite identificar el grupo al que más se parece. En este sentido podremos hablar del carácter predictivo del análisis discriminante.

El modelo predictivo que pronostica el grupo de pertenencia de una observación en virtud de su perfil define la relación entre una variable dependiente (o endógena) no métrica (categórica) y varias variables independientes (o exógenas) métricas. Por tanto, la expresión funcional del análisis discriminante puede escribirse como:

$$y = F(x_1, x_2, \cdots, x_n)$$

con la variable dependiente no métrica y las variables independientes métricas. Las categorías de la variable dependiente definen los posibles grupos de pertenencia de las observaciones o individuos y las variables independientes definen el perfil conocido de cada observación. El objetivo esencial del análisis discriminante es utilizar los valores conocidos de las variables independientes medidas sobre un individuo u observación (perfil) para predecir con qué categoría de la variable dependiente se corresponden para clasificar al individuo en la categoría adecuada.

En el análisis discriminante, una vez comprobado el cumplimiento de los supuestos subyacentes al modelo matemático, se persigue estimar una serie de funciones lineales a partir de las variables independientes que permitan interpretar las diferencias entre los grupos y clasificar a los individuos en alguna de las subpoblaciones definidas por la variable dependiente. Estas funciones lineales se denominan funciones discriminantes y son combinaciones lineales de las variables discriminantes.

Cada una de las funciones discriminantes D_i se obtiene como función lineal de las *k* variables explicativas *X*, es decir:

$$D_i = u_{i1}X_1 + u_{i2}X_2 + \cdots + u_{ik}X_k \quad i=1,2$$

Halladas las funciones discriminantes, y fijado el número de ellas que se retiene, es necesario interpretar el significado de las mismas.

El análisis discriminante, decíamos anteriormente, puede ser utilizado con dos finalidades básicas: interpretar las diferencias existentes entre varios grupos o pronosticar la clasificación de los sujetos. Para el investigador interesado en obtener una regla de decisión que permita clasificar nuevos casos, el número de dimensiones consideradas en el espacio discriminante y su significado posiblemente no atraigan su atención. Puede ser más interesante la utilización de las funciones discriminantes para pronosticar el grupo al que quedará adscrito un nuevo caso no contemplado al extraer las funciones. *Un primer criterio sería clasificar al individuo en el grupo para el que su función discriminante, aplicada en los valores de las variables independientes del individuo concreto (puntuación discriminante), tiene un valor mayor (no olvidemos que hay tantas funciones discriminantes como grupos en la variable dependiente, en nuestro caso 2).* Este procedimiento de clasificación resulta muy sensible a la violación del supuesto de igualdad de matrices de varianzas-covarianzas. Cuando no se verifica dicho supuesto, los casos tienden a ser clasificados en el grupo en el que se registra la mayor dispersión.

Otro de los procedimientos seguidos para asignar un caso a uno de los grupos es utilizar las *probabilidades de pertenencia al grupo*. Un caso se clasifica en el grupo al que su pertenencia resulta más probable. El cálculo de probabilidad de pertenencia a un grupo asume que todos los grupos tienen un tamaño similar. No se tiene en cuenta que a priori es posible anticipar una mayor probabilidad de pertenencia a un determinado grupo cuando en la

población el porcentaje de sujetos que pertenece a cada grupo es muy diferente. En tal situación, conviene incorporar al cálculo las *probabilidades a priori*, con lo que se consigue mejorar la predicción final y reducir lo errores de clasificación. De acuerdo con este planteamiento, la regla de Bayes sería útil para calcular la probabilidad posterior de pertenencia del caso a un grupo (*probabilidad a posteriori*), conocida la probabilidad a priori para el mismo. Un caso será clasificado en el grupo en el que su pertenencia cuenta con una mayor probabilidad a posteriori. Podría ocurrir que dos casos que son clasificados en el mismo grupo tengan probabilidades bastante diferentes, o que las probabilidades de que un sujeto pertenezca a dos grupos distintos no sean muy diferentes entre sí, en cuyo caso, aun asignándolo a la clase en la que cuenta con mayor probabilidad, su clasificación no sería tan clara. Por ese motivo, resulta interesante conocer para cada individuo no sólo la *máxima probabilidad*, sino también las probabilidades de pertenecer a otros grupos.

La probabilidad de pertenecía de un individuo a un grupo *i* de la variable dependiente se evalúa mediante:

$$P_i = \frac{e^{F_i}}{\sum_i e^{F_i}}$$

F_i son las puntuaciones de las funciones discriminantes en el grupo *i*.

Si se utilizan propiedades a priori π_i diferentes de pertenecía a los grupos, la probabilidad anterior tiene la siguiente expresión:

$$P_i = \frac{\pi_i e^{F_i}}{\sum_i \pi_i e^{F_i}}$$

En la aplicación que aquí se presenta, se utiliza la muestra de IRPF, tomándose como variables independientes del modelo discriminante las

partidas económicas declaradas por el individuo en el modelo 100 de IRPF (más de 200 variables) y como variable dependiente una variable dicotómica que toma el valor 1 si el individuo defrauda por una determinada causa de fraude y toma el valor 0 si el individuo no defrauda. Con el modelo discriminante se buscará predecir la probabilidad que tiene cualquier individuo de defraudar o no, para cada causa de fraude, según los valores declarados en las variables del modelo 100. Buscamos por tanto, perfiles de fraude que puedan ayudar en el futuro a la labor inspectora.

Hbíamos comentado que las dos grandes finalidades perseguidas en el uso del análisis discriminante son la clasificación de los individuos en los grupos de la variable categórica dependiente (el individuo defrauda o no defrauda) y la predicción de pertenencia a los citados grupos (probabilidad que tiene cada indivuiduo de defraudar y no defraudar).

3.3 FASE DE SELECCIÓN DE LA INFORMACION: LOS DATOS

Al igual que en el caso de los árboles de decisión, se utiliza como fuente de datos la muestra del Impuesto sobre la Renta de las Personas Físicas (IRPF) que proporciona el Instituto de Estudios Fiscales (IEF). El origen fiscal de la muestra aporta, por tanto, unos datos de gran precisión, y en los que además no aparecen los problemas de infrarrepresentación y falta de respuesta habituales de las encuestas. Por consiguiente, la riqueza de estos datos permite realizar múltiples análisis que están vedados a otras muestras de origen no fiscal. Hay que seguir teniendo presente que disponer de una estructura de hardware y software que implemente procesamiento de grandes datos (*Big Data*) es esencial en nuestro caso. De esta forma, las técnicas de minería de datos que vamos a utilizar, en este caso el análisis discriminante, se encuadran dentro de las técnicas de *Big Data Analytics.*

Recordamos que se seleccionan más de 2 millones de registros de la población total de declarantes mediante muestreo estratificado por provincias, tramos de renta y fuente de renta utilizando afijación proporcional. Este método de selección expande adecuadamente la muestra por toda la población resultando muy significativa. Esta muestra alimemtará nuestros modelos discriminantes.

Además, la muestra de IRPF aporta cerca de 300 variables, tanto partidas económicas como otras variables numéricas y no numéricas, todas ellas contenidas como casillas en el modelo 100 de declaración del IRPF.

La tabla 3-1 muestra las primeras partidas económicas de la muestra de IRPF. En el caso de los modelos discriminantes, estas partidas económicas serán candidatas inicialmente a ser las variables independientes métricas de los modelos

par1	Numérico	8	2	Rdto. del trabajo Dinerarios
par2	Numérico	8	2	Retribuciones en especie (valoración)
par3	Numérico	8	2	Retribuciones en especie (ingresos a cuenta)
par4	Numérico	8	2	Retribuciones en especie (ingresos a cuenta repercutidos)
par5	Numérico	8	2	Rdto. del Trabajo En especie.
par6	Numérico	8	2	Contribuciones Planes Pensiones.
par7	Numérico	8	2	Aportaciones recibidas al patrimonio protegido de las personas con discapacidad del que es titular el contribuyente
par8	Numérico	8	2	Reducciones Art. 18 apartados 2 y 3, y dispos. trans. 11ª y 12ª Ley del Impuesto
par9	Numérico	8	2	Total ingresos integros computables [(01)+(05)+(06)+(07)-(08)]
par10	Numérico	8	2	Cotizac. Seguridad Social, Mutualidad Funcionarios, detracciones derechos pasivos y Coleg.Huérfanos.
par11	Numérico	8	2	Cuotas satisfechas a sindicatos
par12	Numérico	8	2	Cuotas satisfechas a colegios profesionales (si la colegiación es obligatoria y con un máximo de 500 euros anuales)
par13	Numérico	8	2	Gastos de defensa jurídica derivados directamente de litigios con el empleador (máximo: 300 euros anuales)
par14	Numérico	8	2	Gastos deducibles.
par15	Numérico	8	2	Rendimiento neto.Trabajo
par16	Numérico	8	2	Reducción de rendimientos acogidos al régimen especial "33.ª Copa del América" (disposición adicional séptima de la Ley
par17	Numérico	8	2	Reducción por obtención rdto. trabajo.Cuantía aplicable con carácter general.
par18	Numérico	8	2	Reducción por obtención rdto. trabajo.Incremento para trabajadores activos mayores de 65 años que continuen o prolonguen
par19	Numérico	8	2	Reducción por obtención rdto. trabajo.Incremento para contrib. desempleados que acepten un puesto que exija traslado de
par20	Numérico	8	2	Reducción por obtención rdto. trabajo.Reducción adicional para trabajadores activos que sean presonas con discapacidad.
par21	Numérico	8	2	Rendimiento neto reducido.Trabajo
par22	Numérico	8	2	Rend. Cap. Mobiliario. Intereses de cuentas, depósitos y activos financieros
par23	Numérico	8	2	Rend. Cap. Mobiliario. Intereses de activos financieros con bonificación

Tabla 3-1

Asimismo, la base de datos contiene otro tipo de variables numéricas y no numéricas relativas a sexo, edad, provincia, estado civil… de los declarantes (tabla 3-2). Estas variables permitirían estudiar el fraude con análisis de género, análisis geográfico, etc.

cdpost	Numérico	8	2	Código postal
estcv	Cadena	3	0	Estado civil de declarante
sexo	Cadena	3	0	Sexo del declarante
dec	Cadena	3	0	Tipo de declaración
prov	Cadena	6	0	Provincia
ejnacd	Cadena	12	0	Ejercicio de nacimiento del declarante
ejnacc	Cadena	12	0	Ejercicio de nacimiento del cónyuge
minusd	Cadena	9	0	Grado de minusvalía del declarante
minusc	Cadena	9	0	Grado de minusvalía del cónyuge
nmdesc	Numérico	8	2	Número total de descendientes
nmdesc0	Numérico	8	2	Número de descendientes <3 años
nmdesc3	Numérico	8	2	Número de descendientes >= 3 y < 16 años
nmdesc16	Numérico	8	2	Número de descendientes >= 16 y < 18 años
nmdesc18	Numérico	8	2	Número de descendientes >= 18 y < 25 años
nmdescr	Numérico	8	2	Número de descendientes >=25 años
nmdescd	Numérico	8	2	Número de descendientes con edad desconocida
nmdesm0	Numérico	8	2	Número de descendientes sin minusvalía

Tabla 3-2

Como variables dependientes de los modelos discriminantes se utilizarán las variables de fraude más comúnmente utilizadas (tabla 3-3).

id	Numérico	8	2	Identificador del perceptor
f_tmg	Numérico	8	2	Fraude que afecta al tipo marginal
f_capinm	Numérico	8	2	Fraude que afecta a los rendimientos de capital inmobiliario
f_nhijos	Numérico	8	2	Fraude que afecta a las declaraciones de número e hijos y ascendieentes y descendientes
f_aaee	Numérico	8	2	Fraude que afecta a la declaración de actividaes económicas
f_planp	Numérico	8	2	Fraude que afecta a la desgrbación por planes de pensiones
f_gastos	Numérico	8	2	Fraude que afecta a la declarción de gastos
marca	Numérico	8	2	fraude global

Tabla 3-3

´ Los factores de fraude más comunes que aparecen definidos en la base de datos utilizada son los siguientes:

- La variable *f_tmg* indica fraude relativo al tipo marginal,

- la variable *f_capinm* indica fraude relativo a los rendimientos de capital inmobiliario,
- la variable *f_aaee* indica fraude relativo a la declaración de actividades económicas,
- la variable *f_gastos* indica fraude relativo a las deducciones de gastos,
- la variable *f_planp* indica fraude relativo a las declaraciones de planes de pensiones
- la variable *f_nhijos* indica fraude relativo a la declaración del número de hijos y ascendientes.

Al igual que en el caso de los modelos de árboles de decisión, estas son las causas de fraude más habituales, pero podrían considerarse de igual modo todas las causas de fraude que la Agencia Tributaria registra en la base de datos completa de IRPF una vez inspeccionada.

3.4 FASE DE EXPLORACIÓN DE LA INFORMACIÓN

En nuestro caso, las variables dependientes que intervienen en los modelos discriminantes son todas dicotómicas con valores cero y uno. El valor cero corresponde a individuos que no defraudan por la citada causa de fraude y el valor uno corresponde a individuos que defraudan. Sin embargo, las variables independientes son variables métricas que constituyen partidaas de laa muestras de IRPF. Ante esta circunstancia nos encontramos ante modelos que deberán cumplir determinadas condiciones antes de ser utuilizados, todas relacionadas con las hipótesis que habitualmente se le exigen a todos los modelos econométricos. No olvidemos que el modelo discriminante es una caso particular del modelo de regresión múltiple cuando la variable dependiente es categórica y las independientes son métricas (cuantitativas).

El modelo subyacente en el análisis discriminante requiere de una comprobación de determinados supuestos. Para comenzar, la aplicación del

modelo de análisis discriminante requiere que contemos con un conjunto de variables independientes discriminantes (características conocidas de los individuos) y una variable dependiente nominal que define dos o más grupos (cada modalidad de la variable nominal se corresponde con un grupo diferente). Además, los datos deben corresponder a individuos o casos clasificados en dos o más grupos mutuamente excluyentes. Es decir, cada caso corresponde a un grupo y sólo a uno. Por otra parte, las variables discriminantes han de estar medidas en una escala de intervalo o de razón, lo cual permitiría el cálculo de medias y varianzas y la utilización de éstas en ecuaciones matemáticas. Teóricamente, no existen límites para el número de variables discriminantes, salvo la restricción de que no debe ser nunca superior al número de casos en el grupo más pequeño, pero sí es conveniente contar al menos con 20 sujetos por cada variable discriminante si queremos que las interpretaciones y conclusiones obtenidas sean correctas. Todas estas condiciones se cumplen con creces en la aplicación que aquí se trata. En el modelo 100 tenemos más de 400 variables utilizándose en este trabajo prácticamente 300 de ellas. La muestra de IRPF tiene más de dos millones de declarantes, con lo que el tamaño muestral es suficientemente alto. Las dos categorías de la variable dependiente son mutuamente excluyentes, ya que se trata de defraudadores y no defraudadores.

En cuanto a la presencia de datos desaparecidos (*missing*), hay que tener presente que cuando corresponden a la variable de clasificación, los individuos afectados podrían ser excluidos del análisis a la hora de determinar las funciones discriminantes. Si los datos desaparecidos están en variables independientes, hay que asegurarse de que los individuos en los que se registra la ausencia de datos no posean características diferenciales respecto al resto de los individuos, modificando las características de la muestra con la que trabajamos. Si se diera esta circunstancia, sería necesario recurrir a alguno de los procedimientos para tratar los casos desaparecidos (imputación

por la media, por regresión, por métodos especiales etc.). En nuestro caso, los datos missing fueron imputados por la Agencia Tributaria por lo que no tendremos este problema. No obstante, los datos missing deben distribuirse aleatoriamente por toda la muestra, situación ideal ante este tipo de problema. Hay contrastes formales, como el contraste de Little, el contaste de las pruebas pareadas y el contraste de la matriz de correlaciones dicotomizadas para constatar este hecho. Por otra parte, la variable dependiente tampoco tiene datos missing.

Por otro lado, la aplicación del análisis discriminante se apoya en una serie de supuestos básicos como la normalidad multivariante, homogeneidad de matrices de varianza-covarianza (homoscedasticidad), linealidad y ausencia de multicolinealidad.

El supuesto de *normalidad* exige que cada grupo represente una muestra aleatoria extraída de una población con distribución normal multivariable sobre las variables discriminantes. La normalidad univariante no implica la multivariante, pero como esta última es difícil de comprobar, se contrasta la normalidad univariante mediante pruebas clásicas como la prueba de bondad de ajuste basada en *Chi-cuadrado,* la prueba de *Kolmogorov-Smirnov,* el test de *Shapiro-Wilk* o las pruebas de significación basadas en la asimetría y la curtosis. En nuestro caso, sabemos que las variables de renta no son normales y que suelen seguir una distribución paretiana truncada. Este problema se solventa utilizando como variables discriminantes los factores resultantes de aplicar un análisis de componentes principales con rotación ortogonal varimax sobre las variables independientes iniciales. Dada la cantidad de variables, la cantidad de factores y el tamaño de la muestra, puede presuponerse la convergencia a la normalidad de los factores por aplicación del teorema central del límite. Además, el uso de factores refuerza la confidencialidad de las variables con más incidencia en el fraude fiscal.

En cuanto a los casos aislados (*outliers*), es necesario detectar su existencia en cada una de las variables consideradas por separado. Para la detección de casos aislados multivariantes podría recurrirse al cálculo de la distancia de Mahalanobis de cada individuo respecto al centro del grupo o a un método gráfico. En nuestro caso, el uso de factores derivados de la reducción de la dimensión y que engloban cada uno de ellos varias variables iniciales, minimiza el efecto de los valores atípicos.

El supuesto de homogeneidad de matrices de varianza-covarianza (*homoscedasticidad*) obliga a que las matrices de varianzas-covarianzas para las poblaciones de las que fueron extraídos los grupos sean iguales, hipótesis que suele probarse mediante la prueba de *M* de Box, que no es más que una generalización del test de Barlett para la comprobación de la homogeneidad de varianzas univariadas y que se basa en los determinantes de las matrices de varianzas-covarianzas para cada grupo. Por otro lado, el supuesto de *linealidad* implica que existen relaciones lineales entre las variables dentro de cada grupo y suele comprobarse a partir de los diagramas de dispersión de las variables o mediante el cálculo de coeficientes de correlación lineal de Pearson. La matriz de correlaciones de las variables también se utiliza para detectar la *multicolinealidad* (variables con correlación muy alta pueden ser redundantes), que puede ser muy nociva en la inversión de matrices requeridas en los algoritmos discriminantes.

En nuestro caso, estos problemas también desaparecen al utilizar factores en vez de variables iniciales como variables independientes (variables discriminantes) del modelo discriminante. No obstante, se presentarán contrastes formales para estas hipótesis. El problema de la multicolinealidad queda perfectamente resuelto con la utilización de los factores.

En el análisis discriminante, una vez comprobado el cumplimiento de los supuestos subyacentes al modelo matemático, se persigue estimar una serie de funciones lineales a partir de las variables independientes que permitan interpretar las diferencias entre los grupos y clasificar a los individuos en alguna de las subpoblaciones definidas por la variable dependiente. Estas funciones lineales se denominan funciones discriminantes y son combinaciones lineales de las variables discriminantes.

Es importante tener presente también que esta investigación es independiente del ejercicio de datos que se considere, ya que se busca una metodología para cuantificar la incidencia sobre el fraude global y los distintas causas de fraude de las partidas económicas del IRPF. Como además la metodología se basa en un modelo predictivo, se obtiene una función de predicción de fraude que es válida para varios ejercicios futuros consecutivos. No es necesario estimar el modelo que predice el fraude todos los años.

3.5 FASE DE TRANSFORMACIÓN DE LA INFORMACIÓN

Hemos visto en el apartado anterior que las técnicas de Análisis Discriminante, lo mismo que todas las técnicas de Minería de Datos, requieren una exploración de los datos antes de ser aplicadas. Los datos provienen de la explotación que realiza la Agencia Tributaria y no presentan información faltante, lo que evita problemas de imputación de datos mising. No obstante, dada la naturaleza de las variables, existen valores atípicos que no son tratables. Por ejemplo, un contribuyente con muchos ingresos presenta un atípico superior en esa variable que no se puede discutir porque los datos son así. Para minorar el efecto de estos valores atípicos en las técnicas de Minería de Datos será necesario transformar las variables aplicando reducción de la dimensión previa.

Por otra parte, como las variables de la muestra están muy correladas, dada su naturaleza, será necesario transformarlas para eliminar el ruido que provoca la correlación en las técnicas de Análisis Disriminante. Para ello reduciremos las variables iniciales de la muestra altamente correladas a un grupo mucho menor de variables incorreladas mediante el análisis de componentes principales.

En cuanto a la normalidad, ya sabemos que las variables de renta no son normales y que suelen seguir una distribución paretiana truncada. No olvidemos que la distribución de Pareto nació para modelizar variables de renta. Aunque la mayoría de las técnicas de Minería de Datos no exigen la normalidad de las variables, es muy conveniente su presencia. Este problema también se solventa utilizando como variables de las técnicas los factores resultantes de aplicar un análisis de componentes principales con rotación ortogonal varimax sobre las variables iniciales. Dada la cantidad de variables, la cantidad de factores y el tamaño de la muestra, puede presuponerse matemáticamente la convergencia a la normalidad de los factores. Además, el uso de factores refuerza la confidencialidad de las variables con más incidencia en el fraude fiscal.

Asimismo, el uso de factores en lugar de variables iniciales evitara problemas de multicolinealidad en los modelos predictivos inmersos en las técnicas de Minería de Datos e introduce confidencialidad en los datos al transformas las variables iniciales. Además, las puntuaciones correspondientes a las componentes principales tienen valores muy uniformes, lo que siempre es beneficioso para cualquier técnica. En el caso de los modelos predictivos se evitan problemas de heteroscedasticidad, autocorrelación y no normalidad residual.

3.5.1 Componentes Principales

El análisis en componentes principales es una técnica de análisis estadístico multivariante que se clasifica entre los métodos de interdependencia. Se trata de un método multivariante de simplificación o reducción de la dimensión y que se aplica cuando se dispone de un conjunto elevado de variables con datos cuantitativos correlacionadas entre sí persiguiendo obtener un menor número de variables, combinación lineal de las primitivas e incorrelacionadas, que se denominan componentes principales o factores, que resuman lo mejor posible a las variables iniciales con la mínima pérdida de información y cuya posterior interpretación permitirá un análisis más simple del problema estudiado. Esta reducción de muchas variables a pocas componentes puede simplificar la aplicación sobre estas últimas de otras técnicas multivariantes (regresión, clusters, discriminante, etc.).

El elevado número de variables iniciales x_1, x_2,…,x_p se resumen en unas pocas variables C_1, C_2,…,C_k (*componentes principales*) *perfectamente calculables* ($k<<p$) combinanción lineal de las iniciales y que sintetizan la mayor parte de la información contenida en sus datos. Inicialmente se tienen tantas componentes como variables:

$$\begin{aligned} C_1 &= a_{11}x_1 + a_{12}x_2 + \cdots + a_{1p}x_p) \\ &\vdots \\ C_p &= a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{pp}x_p) \end{aligned}$$

Pero sólo se retienen las k componentes principales que explican un porcentaje alto de la variabilidad de las variables iniciales (C_1, C_2,…, C_k).

La primera componente principal, al igual que las restantes, se expresa como combinación lineal de las variables originales como sigue:

$$C_{1i} = u_{11}X_{1i} + u_{12}X_{2i} + \cdots + u_{1p}X_{pi} \qquad i=1,..,n$$

Para el conjunto de las *n* observaciones muestrales y para todas las componentes tenemos:

$$\begin{bmatrix} C_{11} \\ C_{12} \\ \vdots \\ C_{1n} \end{bmatrix} = \begin{bmatrix} X_{11} & X_{21} \cdots & X_{p1} \\ X_{12} & X_{22} \cdots & X_{p2} \\ & \vdots & \\ X_{1n} & X_{2n} \cdots & X_{pn} \end{bmatrix} \begin{bmatrix} u_{11} \\ u_{12} \\ \vdots \\ u_{1p} \end{bmatrix}$$

En notación abreviada tendremos: $C_1 = X\,u_1$ y:

$$V(C_1) = \frac{\sum_{i=1}^{n} C_{1i}^2}{n} = \frac{1}{n} C_1{}' C_1 = \frac{1}{n} u_1' X' X u_1 = u_1' \left[\frac{1}{n} X' X \right] u_1 = u_1' V u_1$$

La primera componente C_1 se obtiene de forma que su varianza sea máxima sujeta a la restricción de que la suma de los pesos u_{1j} al cuadrado sea igual a la unidad, es decir, la variable de los pesos o ponderaciones (u_{11}, u_{12},...,u_{1p})' se toma normalizada. Se trata entonces de hallar C_1 maximizando $V(C_1) = u_1'Vu_1$, sujeta a la restricción:

$$\sum_{j=1}^{p} u_{1i}^2 = u_1' u_1 = 1$$

Se demuestra que, para maximizar $V(C_1)$ se toma el mayor valor propio λ de la matriz V. Sea λ_1 el citado mayor valor propio de V y tomando u_1 como su vector propio asociado normalizado ($u_1'u_1$=1), ya tenemos definido el vector de ponderaciones que se aplica a las variables iniciales para obtener la primera componente principal, componente que vendrá definida como:

$$C_1 = u_1 X = u_{11} X_1 + u_{12} X_2 + \cdots + u_{1p} X_p$$

Para maximizar $V(C_2)$ hemos de tomar el segundo mayor valor propio λ de la matriz V (el mayor ya lo había tomado al obtener la primera componente principal) .

Tomando λ_2 como el segundo mayor valor propio de V y tomando u_2 como su vector propio asociado normalizado ($u_2'u_2=1$), ya tenemos definido el vector de ponderaciones que se aplica a las variables iniciales para obtener la segunda componente principal, componente que vendrá definida como:

$$C_2 = u_2 X = u_{21}X_1 + u_{22}X_2 + \cdots + u_{2p}X_p$$

De forma similar, la componente principal h-ésima se define como $C_h = Xu_h$ donde u_h es el vector propio de V asociado a su h-ésimo mayor valor propio. Suele denominarse también a u_h eje factorial h-ésimo.

Se demuestra que la proporción de la variabilidad total recogida por la componente principal h-ésima (porcentaje de inercia explicada por la componente principal h-ésima) vendrá dada por:

$$\frac{\lambda_h}{\sum_{h=1}^{p}\lambda_h} = \frac{\lambda_h}{traza(V)}$$

Si las variables están tipificadas, traza(V) = p, con lo que la proporción de la componente h-esima en la variabilidad total será $\lambda h/p$. También se define el porcentaje de inercia explicada por las k primeras componentes principales (o ejes factoriales) como:

$$\frac{\sum_{h=1}^{k}\lambda_h}{\sum_{h=1}^{p}\lambda_h} = \frac{\sum_{h=1}^{k}\lambda_h}{traza(V)}$$

Cuando las variables originales están muy correlacionadas entre sí, la mayor parte de su variabilidad se puede explicar con muy pocas componentes. Si las variables originales estuvieran completamente incorrelacionadas entre sí, entonces el análisis de componentes principales carecería por completo de interés, ya que en ese caso las componentes principales coincidirían con las variables originales.

Como *criterio general para precisar el número de componentes a retener*, se seleccionan aquellas componentes cuya raíz característica λ_j excede de la media de las raíces características. Recordemos que la raíz característica asociada a una componente es precisamente su varianza. Analíticamente este criterio implica retener todas aquellas componentes en que se verifique que:

$$\lambda_h > \overline{\lambda} = \frac{\sum_{j=1}^{p}\lambda_h}{p}$$

Si se utilizan variables tipificadas, entonces, como ya se ha visto, se verifica que $\sum_{j=1}^{p}\lambda_h = p$,

Con lo que para variables tipificadas se retiene aquellas componentes tales que $\lambda_h > 1$. La representación gráfica de este criterio se conoce como *gráfico de sedimentación*.

La dificultad en la interpretación de los componentes estriba en la necesidad de que tengan sentido y midan algo útil en el contexto del fenómeno estudiado. Por tanto, es indispensable considerar el *peso que cada variable original tiene dentro del componente elegido*, así como las correlaciones existentes entre variables y factores. Un componente es una función lineal de todas las variables, pero puede estar muy bien correlacionado con algunas de ellas, y menos con otras. Ya hemos visto que el coeficiente de correlación entre una componente y una variable se calcula multiplicando el peso de la variable en esa componente por la raíz cuadrada de su valor propio:

$$r_{jh} = u_{hj}\sqrt{\lambda_h}$$

Se demuestra también que estos coeficientes *r* representan la parte de varianza de cada variable que explica cada factor. De este modo, cada variable puede ser representada como una función lineal de los *k* componentes retenidos, donde los pesos o cargas de cada componente o factor (*cargas factoriales*) en la variable coinciden con los coeficientes de correlación.

El cálculo matricial permite obtener de forma inmediata la tabla de coeficientes de correlación variables-componentes (*pxk*), que se denomina matriz de cargas factoriales. Las ecuaciones de las variables en función de las componentes (factores), traspuestas las inicialmente planteadas, son de mayor utilidad en la interpretación de los componentes, y se expresan como sigue:

$$\begin{array}{ll} C_1 = r_{11}X_1 + \cdots + r_{1p}X_p & \\ C_2 = r_{21}X_1 + \cdots + r_{2p}X_p & \Rightarrow \\ \vdots & \\ C_k = r_{k1}X_1 + \cdots + r_{kp}X_p & \end{array} \quad \begin{array}{l} X_1 = r_{11}C_1 + \cdots + r_{k1}C_k \\ X_2 = r_{12}C_1 + \cdots + r_{k2}C_k \\ \vdots \\ X_p = r_{1p}C_1 + \cdots + r_{kp}C_k \end{array}$$

Es frecuente no encontrar interpretaciones verosímiles a los factores (componentes) obtenidos. Sería deseable, para una más fácil interpretación, que cada componente estuviera relacionada muy bien con pocas variables (coeficientes de correlación *r* próximos a 1 ó -1) y mal con las demás (*r* próximos a 0). Esta optimización se obtiene por una adecuada rotación de los ejes que definen los componentes principales.

Rotar un conjunto de componentes no cambia la proporción de inercia total explicada, como tampoco cambia las comunalidades de cada variable, que no son sino la proporción de varianza explicada por todos ellos. Las rotaciones más utilizadas son la rotación VARIMAX y la QUARTIMAX (ortogonales) y PROMAX (oblicua).

Sin embargo, los coeficientes, que dependen directamente de la posición de los componentes respecto a las variables originales (cargas factoriales y valores propios), se ven alterados por la rotación.

3.5.2 Cálculo de las Componentes Principales e interpretación

En nuestro caso, la reducción de la dimesión es procedente porque el determinante de la matriz de correlaciones de las variables iniciales es prácticamente nulo. Además, las comunalidades de las variables están muy cercanas a la unidad (tabla 3-4).

El proceso de reducción nos lleva a 65 componentes principales Fi (factores) que explican más del 80% de la variabilidad inicial de los datos, resultando así una buena reducción. En concreto explican el 80,123% de la variabilidad, tal y como indica la tabla 3-5.

Matriz de correlaciones[a,b]

a. Determinante = ,000

Comunalidades

	Inicial	Extracción
Rdto. del trabajo Dinerarios	1,000	,966
Retribuciones en especie (valoración)	1,000	,992
Retribuciones en especie (ingresos a cuenta)	1,000	,986
Retribuciones en especie (ingresos a cuenta repercutidos)	1,000	,987
Rdto. del Trabajo En especie.	1,000	,992

Tabla 3-4

Varianza total explicada

Componente	Autovalores iniciales			Sumas de extracción de cargas al cuadrado			Sumas de rotación de cargas al cuadrado		
	Total	% de varianza	% acumulado	Total	% de varianza	% acumulado	Total	% de varianza	% acumulado
1	25,072	13,196	13,196	25,072	13,196	13,196	22,371	11,774	11,774
2	10,982	5,780	18,976	10,982	5,780	18,976	9,928	5,225	17,000
3	6,148	3,236	22,212	6,148	3,236	22,212	5,984	3,149	20,149
4	5,181	2,727	24,939	5,181	2,727	24,939	4,948	2,604	22,753
5	5,065	2,666	27,605	5,065	2,666	27,605	4,845	2,550	25,303
6	4,598	2,420	30,025	4,598	2,420	30,025	4,485	2,360	27,663
7	4,092	2,154	32,178	4,092	2,154	32,178	4,108	2,162	29,826
8	3,348	1,762	33,940	3,348	1,762	33,940	4,002	2,106	31,932
9	3,142	1,654	35,594	3,142	1,654	35,594	3,809	2,005	33,937
10	2,811	1,479	37,073	2,811	1,479	37,073	3,547	1,867	35,803
11	2,777	1,462	38,535	2,777	1,462	38,535	2,771	1,458	37,261
12	2,545	1,340	39,875	2,545	1,340	39,875	2,505	1,319	38,580
13	2,474	1,302	41,177	2,474	1,302	41,177	2,378	1,251	39,831
14	2,196	1,156	42,332	2,196	1,156	42,332	2,132	1,122	40,954
15	2,153	1,133	43,465	2,153	1,133	43,465	2,094	1,102	42,056
16	2,127	1,120	44,585	2,127	1,120	44,585	2,028	1,068	43,123
17	2,094	1,102	45,687	2,094	1,102	45,687	2,014	1,060	44,184
18	2,040	1,073	46,760	2,040	1,073	46,760	2,008	1,057	45,241
19	2,027	1,067	47,827	2,027	1,067	47,827	2,007	1,056	46,297
20	2,000	1,053	48,880	2,000	1,053	48,880	2,000	1,053	47,350
21	1,981	1,043	49,922	1,981	1,043	49,922	1,999	1,052	48,402

22	1,974	1,039	50,961	1,974	1,039	50,961	1,992	1,049	49,450
23	1,963	1,033	51,995	1,963	1,033	51,995	1,981	1,043	50,493
24	1,946	1,024	53,019	1,946	1,024	53,019	1,979	1,042	51,535
25	1,922	1,011	54,030	1,922	1,011	54,030	1,973	1,039	52,573
26	1,900	1,000	55,030	1,900	1,000	55,030	1,973	1,038	53,612
27	1,842	,970	56,000	1,842	,970	56,000	1,958	1,031	54,642
28	1,815	,955	56,955	1,815	,955	56,955	1,938	1,020	55,662
29	1,800	,947	57,903	1,800	,947	57,903	1,932	1,017	56,679
30	1,702	,896	58,799	1,702	,896	58,799	1,923	1,012	57,691
31	1,699	,894	59,693	1,699	,894	59,693	1,912	1,006	58,697
32	1,668	,878	60,571	1,668	,878	60,571	1,891	,995	59,693
33	1,633	,859	61,430	1,633	,859	61,430	1,829	,962	60,655
34	1,592	,838	62,268	1,592	,838	62,268	1,790	,942	61,597
35	1,490	,784	63,052	1,490	,784	63,052	1,759	,926	62,523
36	1,424	,749	63,802	1,424	,749	63,802	1,610	,848	63,370
37	1,407	,741	64,542	1,407	,741	64,542	1,426	,750	64,121
38	1,249	,657	65,200	1,249	,657	65,200	1,408	,741	64,862
39	1,220	,642	65,842	1,220	,642	65,842	1,232	,648	65,510
40	1,189	,626	66,468	1,189	,626	66,468	1,213	,639	66,149
41	1,169	,615	67,083	1,169	,615	67,083	1,191	,627	66,776
42	1,125	,592	67,675	1,125	,592	67,675	1,179	,620	67,396
43	1,099	,578	68,253	1,099	,578	68,253	1,172	,617	68,013
44	1,087	,572	68,825	1,087	,572	68,825	1,171	,616	68,629
45	1,083	,570	69,395	1,083	,570	69,395	1,147	,604	69,233
46	1,076	,566	69,962	1,076	,566	69,962	1,145	,603	69,835
47	1,066	,561	70,523	1,066	,561	70,523	1,110	,584	70,420
48	1,048	,552	71,074	1,048	,552	71,074	1,091	,574	70,994
49	1,038	,547	71,621	1,038	,547	71,621	1,067	,561	71,555
50	1,033	,544	72,165	1,033	,544	72,165	1,033	,544	72,099
51	1,024	,539	72,704	1,024	,539	72,704	1,032	,543	72,642
52	1,022	,538	73,242	1,022	,538	73,242	1,028	,541	73,184
53	1,021	,537	73,779	1,021	,537	73,779	1,028	,541	73,725
54	1,014	,534	74,312	1,014	,534	74,312	1,026	,540	74,265
55	1,011	,532	74,845	1,011	,532	74,845	1,026	,540	74,805
56	1,008	,531	75,375	1,008	,531	75,375	1,021	,537	75,342
57	1,007	,530	75,905	1,007	,530	75,905	1,019	,536	75,878
58	1,005	,529	76,434	1,005	,529	76,434	1,018	,536	76,414
59	1,004	,528	76,963	1,004	,528	76,963	1,013	,533	76,947
60	1,003	,528	77,490	1,003	,528	77,490	1,010	,532	77,479
61	1,001	,527	78,017	1,001	,527	78,017	1,010	,531	78,010
62	1,001	,527	78,544	1,001	,527	78,544	1,007	,530	78,540
63	1,000	,526	79,070	1,000	,526	79,070	1,007	,530	79,070
64	1,000	,526	79,597	1,000	,526	79,597	1,001	,527	79,597
65	1,000	,526	80,123						
66	,998	,525	80,648						
67	,997	,525	81,173						

Método de extracción: análisis de componentes principales.

Tabla 3-5

Las puntuaciones de estos factores serán utilizadas como nuestras 65 variables para las técnicas de Minería de Datos. Evidentemente también conocemos las expresiones de las combinaciones lineales que nos definen cada factor Fi en función de las variables iniciales Xi a partir de los valores de la matriz de cargas factoriales rotadas según una rotación Varimax. También conocemos la naturaleza de las 65 variables latentes resultantes de la reducción que se nombran a continuación. Para cada factor se presentan las variables que lo componen y las cargas factoriales correspondientes que permitirán escribir cada factor en función de las variables iniciales más correladas con él y de modo disjunto.

Factor 1: Base general y cuotas

par694	Cuota autonómica o complementaria correspondiente a la base liquidable general	0.97607
par693	Cuota estatal correspondiente a la base liquidable general	0.97603
par690	Gravamen autonómico correspondiente a la base liquidable general.	0.97486
par689	Gravamen estatal correspondiente a la base liquidable general.	0.97454
par620	Base liquidable general sometida a gravamen.	0.96654
par618	Base liquidable general.	0.96640
par455	Base imponible general	0.96440
par452	Saldo neto de los rendimientos a integrar en la base imponible general y de las imputaciones de renta.	0.96356
par21	Rendimiento neto reducido. Trabajo	0.94615
par742	Retenciones y pagos a cuenta por rendimientos del trabajo	0.94528
par15	Rendimiento neto.Trabajo	0.94513
par9	Total ingresos integros computables	0.94408

	[(01)+(05)+(06)+(07)-(08)]	
par754	Suma de pagos a cuenta	0.90529
par720	Cuota líquida estatal	0.88247
par730	Cuota líquida estatal incrementada	0.88238
par698	Cuota íntegra estatal	0.88188
par732	Cuota líquida incrementada total	0.87750
par771	Parte de las cuotas integras del ejercicio 2009 que corresponde a la Comunidad Autonoma [50% de (698) + (699)]	0.87617
par779	Importe del IRPF que corresponde a la Comunidad Autónoma de residencia del contribuyente	0.87320
par741	Cuota resultante de la autoliquidación	0.87315
par1	Rdto. del trabajo Dinerarios	0.86906
par721	Cuota líquida autonómica	0.86898
par731	Cuota líquida autonómica incrementada	0.86891
par699	Cuota íntegra autonómica o complementaria	0.86764

FAC1=0,97607*par694+0,97603*par693+0,97486*par690+0.97454*par689+0,96654*par620+0,96640*par618+0.96440*par455+0.96356*par452+0,94615*par21+0.94528*par742+0,94513*par15+0,94408*par9+0,90529*par754+0,88247*par720+0,88238*par730+0,88188*par698+0,87750*par732+0,87617*par771+0,87320*par779+0.87315*par741+0,86906*par1+0,86898*par721+0,86891*par731+0,86764*par699

Factor 2: Base del ahorro

par460	Saldo positivo de los rendimientos del capital mobiliario a integrar en la base imponible del ahorro	0.98945
par31	Rendimiento neto [(29)-(30)]	0.98917
par29	Ingresos íntegros Cap. Mobiliario a integrar en la base imponible del ahorro.	0.98915

par35	Rdto. Neto Reducido Cap. Mobiliario a integrar en la base imponible del ahorro [(31)-(32)].	0.98898
par743	Retenciones y pagos a cuenta por rendimientos del capital mobiliario	0.98615
par24	Rend. Cap. Mobiliario. Dividendos y Rendtos. Partic. Fondos Prop.	0.98393
par697	Cuota autonómica o complementaria correspondiente a la base liquidable del ahorro	0.72498
par695	Base liquidable del ahorro sometida a gravamen	0.72496
par630	Base liquidable del ahorro.	0.72494
par465	Base imponible del ahorro (457 – 458 + 460 – 461).	0.72491
par696	Cuota estatal correspondiente a la base liquidable del ahorro	0.72491

FAC2=0,98945*par460+0,98917*par31+0,98915*par29+0,98898*par35+0,98615*par743+0,98393*par24+0,72498*par697+0,72496*par695+0,72494*par630+0,72491*par465+0,72491*par696

Factor 3: Saldo neto y resultado

par457	Saldo neto positivo de ganancias y pérdidas patrimoniales imputables a 2009 integrar en B.I.del ahorro	0.95778
par755	Cuota diferencial (741 – 754)	0.93434
par760	Resultado de la declaración (755 – 756 + 757 – 758 + 759)	0.93433

FAC3 = 0,95778*par457+0,93434*par755+0,93433*par760

Factor 4: Deducciones generales

par727	Intereses de demora de deducciones generales de 1997 a	0.99796

	2008 a las que se ha perdido el derecho. Parte autonómica.	
par725	Intereses demora de deducciones generales de 1997 a 2008 a las que se ha perdido el derecho. Parte estatal.	0.99755
par724	Importe de las deducciones generales de 1997 a 2008 a las que se ha perdido el derecho. Parte estatal.	0.99750
par726	Importe de las deducciones generales de 1997 a 2008 a las que se ha perdido el derecho. Parte autonómica.	0.99750
par775	Incrementos de la cuota líquida autonómica por perdida del derecho a determinadas deducciones en ejercicios anteriores	0.96733

FAC4 = 0,99796*par727+0,99755*par725+0,99750*par724+0,99750*par726+0,96733*par775

Factor 5: Capital Inmobiliario

par70	Ingresos íntegros Cap.Inmobiliario	0.96582
par79	Rendimiento neto reducido del capital inmobiliario: la cantidad mayor de (075 - 076 - 077) y 078	0.95278
par85	Suma de rendimientos netos reducidos del capital inmobiliario.	0.95278
par75	Rendimiento neto (070 - 071 - 072 - 074).	0.94860
par744	Retenciones y pagos a cuenta por arrendamientos de inmuebles urbanos	0.65431
par72	Gastos deducibles. Importe de 2009 que se aplica en esta declaración.	0.44420

FAC5=0,96582*par70+0,95278*par79+0,95278*par85+0,94860*par75+0,65431*par744+0,44420*par72

Factor 6: Deducciones vivienda habitual

par700	Deducc. por adquisición o rehabilitación de la vivienda habitual, parte estatal	0.91907
par701	Deducc. por adquisición o rehabilitación de vivienda habitual, parte autonómica	0.91704
par772	Parte de la deducción por inversión en vivienda habitual que corresponde a la Comunidad Autonoma [50% de (700) + (701)]	0.89811
par777	50% de la compensación fiscal por deducción en adquisición de vivienda habitual [50% de (738)]	0.86965
par738	Compensación fiscal por deducción en adquisición de vivienda habitual, para viviendas adquiridas antes del 20-01-2006	0.86960

FAC6 =
0,91907*par700+0,91704*par701+0,89811*par772+0,86965*par777+
0,86960*par738

Factor 7: Retribuciones en especie

par2	Retribuciones en especie (valoración)	0.89391
par5	Rdto. del Trabajo En especie.	0.89305
par4	Retribuciones en especie (ingresos a cuenta repercutidos)	0.88687
par3	Retribuciones en especie (ingresos a cuenta)	0.88382
par8	Reducciones Art. 18 apartados 2 y 3, y dispos. trans. 11ª y 12ª Ley del Impuesto	0.67341

FAC7 =
0,89391*par2+0,89305*par5+0,88687*par4+0,88382*par3+0,67341*par8

Factor 8: Capital mobiliario

par47	Rendimiento neto [(45)-(46)]	0.97217
par50	Rdto. Neto Reducido Cap.Mobiliario a integrar en la base imponible general.	0.96866
par44	Otros rendimientos del capital mobiliario a integrar en la base imponible general.	0.94312
par45	Ingresos íntegros Cap. Mobiliario a integrar en la base imponible general.	0.92420

FAC8 =0,97217*par47+0,96866*par50+0,94312*par44+0,92420*par45

Factor 9: Deducciones autonómicas

par705	Deducc. por cantidades o bienes donados a determinadas entidades parte autonómica	0.92436
par704	Deducc. por cantidades o bienes donados a determinadas entidades parte estatal	0.92436
par774	Suma deducciones autonómicas [(717)]	0.92402
par717	Suma de deducciones autonómicas	0.92402

FAC9
=0,92436*par705+0,92436*par704+0,92402*par774+0,92402*par717

Factor 10: Mínimos personales y familiares

par692	Gravamen autonómico correspondiente a la base liquidable general. Importe mínimo personal y familiar.	0.89706
par691	Gravamen estatal correspondiente a la base liquidable general. Importe mínimo personal y familiar.	0.89556
par680	Importe del mínimo personal y familiar que forma parte de la base liquidable general.	0.89540

par735	Deducción por obtención de rendimientos del trabajo o de actividades económicas.	0.62668

FAC10=0,89706*par692+0,89556*par691+0,89540*par680+0,62668*par735

Factor 11: Deducciones Islas Canarias

par708	Deducc. por dotaciones a la Reserva para Inversiones en Canarias, parte estatal	0.99185
par709	Deducc. por dotaciones a la Reserva para Inversiones en Canarias, parte autonómica	0.99184
par773	Parte de las decmás deducciones generales que corresponde a la Comunidad Autónoma [50% de la suma (702) a (716)]	0.87484

FAC11 =0,99185*par708+0,99184*par709+0,87484*par773

Factor 12: Compensación rendimientos capital mobiliario y seguros de vida

par739	Compensación fiscal por percepción de rdtos. del capital mob. con período de generación superior a dos años	0.96570
par778	50% de la compensacion fiscal por percepción de determinados rendimientos del capital mobiliario [50% de (739)]	0.96567
par27	Rend. Cap. Mobiliario. Rendtos. Contratos Seguros Vida o Inv.	0.78461

FAC12 =0,96570*par739+0,96567*par778+0,78461*par27

Factor 13: **Actividades económicas**

par140	Rdto. Neto reducido total act. econ. Est. Directa	0.93401
par745	Retenciones y pagos a cuenta por actividades económicas	0.84244
par750	Pagos fraccionados ingresados (actividades económicas).	0.46731

FAC13 =0,93401*par140+0,84244*par745+0,46731*par750

Factor 14: Reducción tributación conjunta y circunstancias personales y familiares.

Par470	Reducción de la Base Imponible por tributación conjunta	0.95046
Par610	Reducc. B I General por tributación conjunta.	0.90411
Par676	Adecuación del impuesto a las circunstancias personales y familiares.Mínimo por descendientes.	0.50451

FAC14 =0,95046*par470+0,90411*par610+0,50451*par676

Factor 15: Deducción por incentivos y estímulos a la inversión empresarial

par706	Deducc. por incentivos y estímulos a la inversión empresarial, parte estatal	0.99526
par707	Deducc. por incentivos y estímulos a la inversión empresarial, parte autonómica	0.99526

FAC15 =0,99526*par706+0,99526*par707

Factor 16: Deducciones por rentas obtenidas en Ceuta y Melilla y doble imposición.

par713	Deducc. por rentas obtenidas en Ceuta y Melilla parte autonómica	0.98779
par712	Deducc. por rentas obtenidas en Ceuta y Melilla parte estatal	0.98779
par733	Deducción por doble imposición de dividendos pendientes de aplicar de 2005 y 2006. Importe que se aplica.	0.25565

FAC16 =0,98779*par713+0,98779*par712+0,25565*par733

Factor 17: Deducciones autonómicas a las que se ha perdido derecho

par728	Importe de las deducciones autonómicas de 1998 a 2008 a las que se ha perdido el derecho	0.98974
par729	Intereses demora de deducciones autonómicas de 1998 a 2008 a las que se ha perdido el derecho	0.98919

FAC17 =0,98974*par728+0,98919*par729

Factor 18: Deducción por rendimientos en Canarias.

par710	Deducc. por rendimientos derivados de la venta bienes corporales producidos en Canarias, parte estatal	0.99954
par711	Deducc. por rendimientos derivados de la venta bienes corporales producidos en Canarias, parte autonómica	0.99954

FAC18 =0,99954*par710+0,99954*par711

Factor 19: Gastos deducibles y cotizaciones

par10	Cotizac. Seguridad Social, Mutualidad Funcionarios, detracciones derechos pasivos y Coleg.Huérfanos.	0.99384
par14	Gastos deducibles.	0.99335

FAC19 =0,99384*par10+0,99335*par14

Factor 20: Deducción cuentas ahorro empresa

par714	Deducc. por cantidades depositadas en cuentas ahorro-empresa parte estatal	0.99994
par715	Deducc. por cantidades depositadas en cuentas ahorro-empresa parte autonómica	0.99994

FAC20 =0,99994*par714+0,99994*par715

Factor 21: Reducción Base imponible por alimentos

par585	Reducción de la Base Imponible por pensiones compensatorias al cónyuge y anualidades por alimentos	0.97815
par622	Reducc. Base imponible del ahorro por pensiones compensatorias y anualidades por alimentos.	0.97400

FAC21 =0,97815*par585+0,97400*par622

Factor 22: Reducción Base imponible por discapacidad

par560	Reducción de la Base Imponible por aportaciones a los patrimonios protegidos de las personas con discapacidad	0.99698
par614	Reducc. B I General por aportaciones a patrimonios protegidos de personas con discapacidad.	0.99696

FAC22 =0,99698*par560+0,99696*par614

Factor 23: Deducción por niños

par758	Deducción por nacimiento o adopción: importe de la deducción	0.97929
par759	Deducción por nacimiento o adopción: cantidades percibidas en concepto de abono anticipado	0.97813

FAC23 =0,97929*par758+0,97813*par759

Factor 24: Deducciones por doble imposición

par776	50% de las deducciones por doble imposición [50% de (733) + (734) + (736) + (737)]	0.97980
par734	Deducción por doble imposición internacional, por las rentas obtenidas y gravadas en el extranjero	0.97850

FAC24 =0,97980*par776+0,97850*par734

Factor 25: Rentas inmobiliarias

par69	Imputación rentas inmobiliarias.	0.96367
par80	Suma rentas inmob. Imputadas.	0.96367

FAC25 =0,96367*par69+0,96367*par80

Factor 26: Reducción BI aportaciones previsión social

par613	Reducc. B I General por aportaciones y contribuciones a sistemas de previsión social constituidos a favor de personas con discapacidad	0.98878

par530	Reducción de la Base Imponible por aportaciones y contribuciones a sistemas de previsión social de personas discapacidad	0.98876

FAC26 =0,98878*par613+0,98876*par530

Factor 27: Reducción mutualidades deportistas

par600	Reducción de la Base Imponible por aportaciones a Mutualidades de Previsión Social de deportistas profesionales	0.98682
par617	Reducc. B I General por aportaciones a la mutualidad de previsión social de deportistas profesionales.	0.98641

FAC27 =0,98682*par600+0,98641*par617

Factor 28: Deducciones maternidad

par756	Deducción por maternidad: importe de la deducción	0.95540
par757	Deducción por maternidad: cantidades percibidas en concepto de abono anticipado	0.95382

FAC28 =0,95540*par756+0,95382*par757

Factor 29: Reducción por aportaciones de previsión social

par611	Reducc. B I General por aportaciones y contribuciones a sistemas de previsión social (régimen general).	0.81858
par500	Reducción de la Base Imponible por aportaciones y contribuciones a sistemas de previsión social, régimen general	0.75479
par6	Contribuciones Planes Pensiones.	0.72417

FAC29 =0,81858*par611+0,75479*par500+0,72417*par6

Factor 30: Deducciones inversión cultural

par702	Deducc. por inversiones o gastos en bienes de interés cultural parte estatal	0.98021
par703	Deducc. por inversiones o gastos en bienes de interés cultural parte autonómica	0.98020

FAC30 =0,98021*par702+0,98020*par703

Factor 31: Reducciones por previsiones sociales de uno mismo y del cónyuge.

Par 505	Reducción de la base imponible por aportaciones y contribuciones a sistemas de previsión social del cónyuge	0.9611
Par 612	Reduccion B General por aportaciones a sistemas de previsión social de los que es partícipe, mutualista o titular con el	0.96791

FAC31 =0,9611*par505+0,96791*par612

Factor 32: Adecuaciones del impuesto por situaciones familiares y personales.

Par 675	Adecuación del impuesto a las circunstancias personales y familiares. Mínimo del contribuyente	0.96923
Par 679	Adecuación del impuesto a las circunstancias personales y familiares. Mínimo personal y familiar	0.94729

FAC32 =0,96923*par675+0,94729*par679

Factor 33: Atribuciones de rentas.

Par 223	Atribuciones de rentas: Rendimientos de actividades económicas	0.92263
Par 746	Retenciones e ingresos a cuenta atribuidos por régimen especial de atribución de rentas	0.85226

FAC33 =0,92263*par223+0,85226*par746

Factor 34: Gastos e ingresos deducibles.

Par 46	Gastos deducibles	0.90455
Par 40	Rend. Cap. Mob. Rendtos procedentes del arrendamiento de bienes muebles, negocios o minas de subarrendamientos	0.90410

FAC34 =0,90455*par46+0,90410*par40

Factor 35: Rendimientos debidos a bienes de derechos de imagen.

Par 265	Imputación de rentas por la cesión de derechos de imagen	0.92471
Par 737	Deducc. Doble imposición, régimen imputación de rentas derivadas de la cesión de derechos de imagen	0.84417

FAC35 =0,92471*par265+0,84417*par737

Factor 36: Rendimientos y reducciones de impuestos por capital inmobiliario

par459	Saldo negativo de los rendimientos del capital	0.89508

	mobiliario a integrar en la base imponible del ahorro	
par32	Reducciones Disp.Transitoria 4ª de la Ley del Impuesto.	0.74814
par26	Rend. Cap. Mobiliario. Rendtos. Transmisión o Amortización otros activos	-0.48019

FAC36 =0,89508*par459+0,74814*par32-0,48019*par26

Factor 37: Retenciones, ingresos e imputación en régimen de transparencia fiscal y empresas

par747	Retenciones e ingresos a cta. por imputaciones de agrupaciones de interés económico y uniones temporales de empresas	0.84456
par245	Imputación de entidades en régimen de transparencia fiscal	0.84379

FAC37 =0,84456*par747+0,84379*par245

Factor 38: Adecuación, reducción e importes por discapacidad

par678	Adecuación del impuesto a las circunstancias personales y familiares.Mínimo por discapacidad.	0.78204
par20	Reducción por obtención rdto. trabajo.Reducción adicional para trabajadores activos que sean presonas con discapacidad.	0.64119
par621	Reducc. Base imponible del ahorro por tributación conjunta.	0.37638
par681	Importe del mínimo personal y familiar que forma parte de la base liquidable del ahorro.	0.33936

FAC38
=0,78204*par678+0,64119*par20+0,37638*par621+0,33936*par681

Factor 39: Retenciones, saldos netos negativos y rendimiento por mobiliario y patrimonio

par752	Retenciones a cuenta efectivamente practicadas art. 11 Directiva 2003/48/CE	0.69448
par458	Saldos netos negativos de ganancias y pérdidas patrimoniales de 2005-2008 a integrar en la parte especial de la renta de	0.51650
par22	Rend. Cap. Mobiliario. Intereses de cuentas, depósitos y activos financieros	0.39438

FAC39=0,69448*par752+0,51650*par458+0,39438*par22

Factor 40: Saldo neto negativo y pérdidas patrimoniales imputables a 2009 a integrar en B.I

par456	Saldo neto negativo ganancias y pérdidas patrim. imput. a 2009 a integrar en B.I. Gral.: imp. pendte. compensar 4 ejerci	0.88305
par454	Saldo neto negativo de ganancias y pérdidas patrimoniales imputables 2009 a integrar en B.I. general	0.64919

FAC40=0,88305*par456+0,64919*par454

Factor 41: Capital mobiliario, letras del tesoro, ganancias patrimoniales y gastos deducibles

par25	Rend. Cap. Mobiliario. Rendtos. Transmisión o	0.63533

	Amortización Letras Tesoro	
par749	Por ganancias patrimoniales, incluidos premios	0.56988
par30	Gastos Deducibles	0.35874

FAC41=0,63533*par25+0,56988*par749+0,35874*par30

Factor 42: Anualidades de alimentos (pensiones y a favor de los hijos)

par615	Reducc. B I General por pensiones compensatorias y anualidades por alimentos.	0.73441
par688	Anualidades por alimentos en favor de los hijos satisfechas por resolución judicial.	0.73373

FAC42=0,73441*par615+0,73373*par688

Factor 43: Arrendamiento de inmuebles y gastos fiscales deducibles

par76	Reducción por arrendamiento de inmuebles destinados a vivienda (artículo 23.2 de la Ley del Impuesto).	0.74476
par74	Otros gastos fiscalmente deducibles.	-0.65241

FAC43=0,74476*par76-0,65241*par74

Factor 44: Gastos deducibles por importes pendientes de ejercicios anteriores

par71	Gastos deducibles. Importe pendiente de deducir del ejercicio 2008 que se aplica en esta declaración	0.74668
par73	Gastos deducibles. Importe de 2009 pendiente de	0.61235

	deducir en los 4 años siguientes.	

FAC44=0,74668*par71+0,61235*par73

Factor 45: Gastos en defensa jurídica, colegios profesionales y sindicatos

par13	Gastos de defensa jurídica derivados directamente de litigios con el empleador (máximo: 300 euros anuales)	0.70542
par12	Cuotas satisfechas a colegios profesionales (si la colegiación es obligatoria y con un máximo de 500 euros anuales)	0.62152
par11	Cuotas satisfechas a sindicatos	0.33483

FAC45=0,70542*par13+0,62152*par12+0,33483*par11

Factor 46: Ingresos por actividad económica.

par170	Rdto. Neto reducido total act. econ. Est. Objetiva	0.69259
par17	Reducción por obtención rdto. trabajo. Cuantía aplicable con carácter general.	-0.58191

FAC46=0,69259*par170-0.58191*par17

Factor 47: Atribuciones de rentas.

par221	Atribución de rentas: Rendimientos capital mobiliario a integrar en la base imponible del ahorro.	0.59682
par222	Atribucion de rentas: Rendimientos capital inmobiliario	0.57640

FAC47=0,59682*par221+0,57640*par222

Factor 48: Ingresos y reducciones procedentes actividad no económica.

par42	Rend.procedentes de la propiedad intelectual cuando el contribuyente no sea el autor	0.64962
par48	Reducciones Art. 26.2 de la Ley del Impuesto.	0.58683
par450	Saldo neto positivo de ganancias y pérdidas patrimoniales imputables a 2009 a integrar en B.I. general	-0.47739

FAC48=0,64962*par42+0,58683*par48-0,47739*par450

Factor 49: Saldos negativos patrimoniales.

par451	Saldos netos negativos de ganancias y pérdidas patrimoniales de 2005-2008 a integrar en la parte general de la renta del	0.73125
par453	Resto de los saldos netos negativos de ganancias y pérdidas patrimoniales de 2005-2008 a integrar en la parte general de	0.71533

FAC49=0,73125*par451+0,71533*par453

Factor 50: Ganancias agrarias y adecuación de impuesto

par197	Rdto. Neto Módulos Agrarios	0.68557
par677	Adecuación del impuesto a las circunstancias personales y familiares. Mínimo por ascendientes.	0.53345

FAC50=0,68557*par197+0,53345*par677

Factor 51: Intereses demora e importe deducciones de 1996 y ejercicios anteriores

par723	Intereses demora de deducciones de 1996 y ejercicios anteriores a las que se ha perdido el derecho	0.72069
par722	Importe de las deducciones de 1996 y ejercicios anteriores a las que se ha perdido el derecho	0.71546

FAC51=0,72069*par723+0,71546*par722

Factor 52: Rendimiento Capital Mobiliario con bonif. y retenciones deducibles de rendimientos bonif

par23	Rend. Cap. Mobiliario. Intereses de activos financieros con bonificación	0.70874
par740	Retenciones deducibles correspondientes a rendimientos bonificados	0.70132

FAC52=0,70874*par23+0,70132*par740

Factor 53: Rendimientos capital mobiliario 2007-2008 y rendimientos rentas imps. Capital

par461	Saldo neto negativo de rendimientos del capital mobiliario de 2007 y 2008 a integrar en la base imponible del ahorro	0.69808
par28	Rend. rentas que tengan por causa la imp. cap. y otros rend. del cap. mob. a integrar en la base imp. ahorro	0.60377

FAC53=0,69808*par461+0,60377*par28

Factor 54: Reducción por obtención rendimiento trabajo. Incremento para trabajadores activos mayores de 65 años que continúen o prolonguen

par18	Reducción por obtención rdto. trabajo. Incremento para trabajadores activos mayores de 65 años que continúen o prolonguen	0.70123

FAC54=0,70123*par18

Factor 55: Deducciones por alquiler de vivienda y rendimientos trabajo

par716	Deducc. por alquiler de la vivienda habitual	0.69945
par19	Reducción por obtención rdto. trabajo.Incremento para contrib. desempleados que acepten un puesto que exija traslado	0.67526

FAC55=0,69945*par716+0,67526*par19

Factor 56: Deducciones e imputaciones de rentas aplicando régimen de transparencia fiscal internacional

par736	Deducc. doble imposición internaci. habiendo aplicado el régimen de transp. fiscal internacional	0.70983
par255	Imputaciones de rentas positivas en el régimen de transparencia fiscal internacional	0.70685

FAC56=0,70983*par736+0,70685*par255

Factor 57: Reducciones por rendimientos generados en más de 2 años o en caso de parentesco

par77	Reducción por rendimientos generados en más de 2	0.67232

	años u obtenidos de forma notoriamente irregular (art. 23.3 de la Ley d	
par78	Rendimiento mínimo computable en caso de parentesco (Art. 24 de la ley del impuesto)	-0.49059

FAC57=0,67232*par77-0,49059*par78

Factor 58: Rendimientos procedentes de la propiedad industrial que no se encuentre afectada a una actividad económica.

par43	Rend.procedentes de la propiedad industrial que no se encuentre afecta a una actividad económica	0.96166

FAC58=0,96166*par43

Factor 59: Reducción de rendimientos acogidos al régimen especial 33.ª Copa del América.

par16	Reducción de rendimientos acogidos al régimen especial "33.ª Copa del América" (disposición adicional séptima de la Ley	0.88703

FAC59=0,88703*par16

Factor 60: Impuestos aplicados a las aportaciones realizadas a los partidos políticos por parte de sus afiliados.

par623	Reducc. Base imponible del ahorro. Cuotas de afiliación y demás aportaciones a los partidos políticos realizadas por afiliados	0.85066
par616	Reducc. B I General Cuotas de afiliación y demás aportaciones a los partidos políticos realizadas por	0.52034

	afiliados	

FAC60=0,85066*par623+0,52034*par616

Factor 61: Imputación de rentas derivadas y compensación de bases liquidables

par275	Imputación de rentas derivadas participación Instituciones Inversión Colectiva en paraisos fiscales	0,29589
par619	Compensación de bases liquidables generales negativas de 2005 a 2008,	0,89093

FAC61=0,29589*par275+0,89093*par619

Factor 62: Rentas exentas de IRPF

par686	Rentas exentas del IRPF, excepto para determinar el tipo de gravamen, De la base liquidable del ahorro,	0,69664
par687	Rentas exentas del IRPF, excepto para determinar el tipo de gravamen, De la base liquidable general,	0,68799

FAC62=0,69664*par686+0,68799*par687

Factor 63: Rendimientos de la prestación técnica y del capital inmobiliario

par41	Rend,procedentes de la prestación de asistecia técnica, salvo en el ámbito de una actividad económica	0,77114
par220	Atribución de rentas: Rendimientos capital mobiliario a integrar en la base imponible general,	0,58224

FAC63=0,77114*par41+0,58224*par220

Factor 64: Cuotas del impuesto sobre la renta de no residentes

par751	Cuotas del Impuesto sobre la Renta de no Residentes	0,99298

FAC64=0,99298*par751

Factor 65: Aportaciones recibidas al patrimonio protegido de las personas con discapacidad del que es titular el contribuyente

par7	Aportaciones recibidas al patrimonio protegido de las personas con discapacidad del que es titular el contribuyente	0,9894

FAC65=0,9894*par7

3.5.3 Puntuaciones de las componentes

Las puntuaciones de las componentes son las variables reducidas incorreladas que sustituirán a las variables iniciales que constituyen las partidas esconómicas del IRPF y que están muy correladas y por ello no pueden ser variables independientes de ningún modelo por el problema de la multicolinealidad.

A partir de ahora, los modelos tendrán como variables independientes las puntuaciones de las componentes C1, C2, …. para evitar la multicolinealidad.

Parte de las puntuaciones de las componentes obtenidas se presentan en la tabla 3-6.

FAC1_1	FAC2_1	FAC3_1	FAC4_1	FAC5_1	FAC6_1	FAC7_1	FAC8_1	FAC9_1	FAC10_1
-,28733	,02082	,06897	-,00307	,41413	-,88081	,23402	,07471	-,00050	2,92372
-,32042	,00424	,06667	-,00395	-,21645	-,90257	,21766	-,00792	-,08000	2,29451
-,24166	-,01114	,00228	-,00183	-,09887	-,32273	,05859	,00086	-,00143	-,90528
-,24364	-,01063	-,00346	-,00177	-,09721	-,29709	,05883	-,00249	,00278	-,97691
-,25142	-,00848	,01474	-,00218	-,11037	-,43155	,08886	,00531	-,01596	-,37856
-,24370	-,01137	,01199	-,00201	-,10834	-,37403	,06858	,00273	-,01158	-,62209
-,24881	-,00826	-,01222	-,00215	-,16421	-,17482	,12582	-,01412	,01986	,86438
-,18107	-,03787	,01079	-,00196	-,17678	-,49507	,05854	-,01819	-,03597	,70086
-,29364	,10862	-,08900	-,00267	,18870	-,34214	,14316	-,02691	-,03847	-,15804
-,07748	-,02901	-,03972	-,00294	-,16371	-,67197	,04466	-,02207	-,04326	1,56624
,19409	,02077	-,11588	-,00275	-,13130	-,54946	-,14866	-,01053	,02726	,43033
,42319	-,08268	-,08273	-,00254	-,19831	1,42078	-,19330	-,04591	-,07591	,91602
-,18636	-,02137	,01242	-,00530	-,16369	,80438	,16216	-,00311	-,03867	,63321
-,43981	,07769	,12848	-,00413	,08397	-1,11599	,49600	,10589	-,09213	-1,52525
-,27315	,00531	,00732	,00015	-,08598	-,28397	,06995	,09276	,02176	-,87602
-,26406	-,00790	,02342	-,00244	-,14630	-,53585	,12427	-,00392	-,03470	,56151
,21007	-,06666	-,04748	-,00357	-,22155	2,59063	-,05183	-,03047	-,07008	-,14517
,17008	,05202	,06008	,00186	,17625	1,47654	,10388	,02869	,05567	,22449

Tabla 3-6

Por cuestiones de comodidad y facilidad de escritura e interpretación, en lo que sigue sustituiremos los nombres de las componentes FAC1_1, FAC2_1,… por C1, C2,…

3.6 FASES DE MODELIZACIÓN Y EVALUACIÓN: ESTIMACIÓN Y DIAGNOSIS DE LOS MODELOS DE ANÁLISIS DISCRIMINANTE

A continuación estimaremos los modelos discriminantes para el fraude global y para las distintas causas de fraude que indicen en el fraude global utilizando como variables independientes las puntuaciones de las componentes principales que resumen las partidas económicas del impuesto y como variable depediente el fraude global o cualquier causa de fraude..

3.6.1 Modelo discriminante para el fraude global

El modelo para el fraude global tendrá la expresión:

$$Marca = F(C_1, C_2, \cdots, C_n) + e$$

Estimaremos las funciones discriminantes de Fisher:

$$D_i = u_{i1}C_1 + u_{i2}C_2 + \cdots + u_{ik}C_k \qquad i=1,2$$

correspondientes a los grupos de fraude y no fraude (categorías de la variable dependiente del modelo *Marca*).

A continuación se muestran los coeficientes estimados de la dos funciones discriminantes de Fisher (tabla 3-7).

Coeficientes de función de clasificación

	fraude global	
	,00	1,00
REGR factor score 1 for analysis 1	-,378	,226
REGR factor score 2 for analysis 1	-,005	,003
REGR factor score 3 for analysis 1	,010	-,006
REGR factor score 4 for analysis 1	-,002	,001
REGR factor score 5 for analysis 1	-,168	,100
REGR factor score 6 for analysis 1	-,501	,299
REGR factor score 7 for analysis 1	,116	-,069
REGR factor score 8 for analysis 1	,002	-,001
REGR factor score 9 for analysis 1	-,019	,011
REGR factor score 10 for analysis 1	-,719	,429

Tabla 3-7

Resulta entonces que una vez estimado el modelo discriminante tomando como variables independientes los 64 factores (ninguno resulta expulsado del modelo por los criterios de selección de variabes discriminantes), se obtienen los coeficientes de las dos funciones discriminantes de Fisher.

$$\mathbf{D0 = - 0{,}378C_1 - 0{,}05C_2 +\text{--------} - 0{,}776}$$

$$\mathbf{D1 = \ 0{,}226C_1 + 0{,}03C_2 +\text{--------} - 0{,}942}$$

En cuanto a la diagnosis del modelo, observamos los resultados del contraste de la Lambda de Wilks cuyo p-valor (Sig.) pequeño valida la significatividad del modelo discriminante en su conjunto (tabla 3-8).

También observamos los resultados de la prueba M de Box, cuyo p-valor pequeño muestra la ausencia de heteroscedasticidad en el modelo (tabla 3-9). La ausencia de multicolinealidad viene determinada por el uso de los factores que son incorrelados y la hipótesis de normalidad le asegura el uso de factores rotados y el teorema central del límite, como ya se ha indicado.

Lambda de Wilks

Prueba de funciones	Lambda de Wilks	Chi-cuadrado	gl	Sig.
1	,651	828541,892	64	,000

Tabla 3-8

M de Box		144316437,7
F	Aprox.	69380,371
	df1	2080
	df2	7,111E+12
	Sig.	,000

Prueba la hipótesis nula de las matrices de covarianzas de población iguales.

Tabla 3-9

A continuación se muestra la matriz de confusión que muestra el porcentaje de individuos muestrales bien clasificados por el modelo discriminante (tabla 3-10). Si el porcentaje de casos correctamente clasificados es alto, cabe esperar que las funciones discriminantes también proporcionen buenos resultados a la hora de predecir el grupo al que se adscribirá cualquier nuevo sujeto perteneciente a la misma población de donde fue extraída la muestra. Este porcentaje puede ser tomado como una medida no sólo de la bondad de la clasificación, sino también de las diferencias existentes entre los grupos; si la clasificación es buena se deberá a que las variables discriminantes permiten diferenciar entre los grupos. En nuestro caso se observa que se clasifican bien el 79,1% de los individuos muestrales.

Resultados de clasificación[a]

			Pertenencia a grupos pronosticada		
		fraude global	,00	1,00	Total
Original	Recuento	,00	653735	66636	720371
		1,00	335498	872625	1208123
	%	,00	90,7	9,3	100,0
		1,00	27,8	72,2	100,0

a. 79,1% de casos agrupados originales clasificados correctamente.

Tabla 3-10

A continuación se muestra el área bajo la curva ROC correspondiente al modelo discriminante (tabla 3-11). El área óptima es la unidad. Se observa que en nuestro modelo el área es 0,886, valor bastante alto que indica que le modelo presenta un ajuste de calidad.

También se presenta la gráfica de la curva ROC (Figura 3-1) que incluye una porción de área por encima de la diagonal baste considerable. La curva ROC óptima es la que ocupa todo el triángulo rectángulo superior de la figura (el área de este triángulo es la unidad).

Finalmente se muestra una tabla con el grupo de pertenencia de los individuos declarantes de IRP (1=fraude, 0=no fraude), los valores de las dos funciones discriminantes de Fisher para cada individuo (Dis1_1=función discriminante para el grupo de fraude, Dis1_2=función discriminante para el grupo de no fraude) y las probabilidades de fraude global de los individuos, que cuantifican su propensión al fraude (tabla 3-12).

Área bajo la curva

Variable(s) de resultado de prueba: Probabilidades de pertenencia a grupo 1

Área	Error estándar[a]	Significación asintótica[b]	95% de intervalo de confianza asintótico	
			Límite inferior	Límite superior
,886	,000	,000	,886	,887

Tabla 3-11

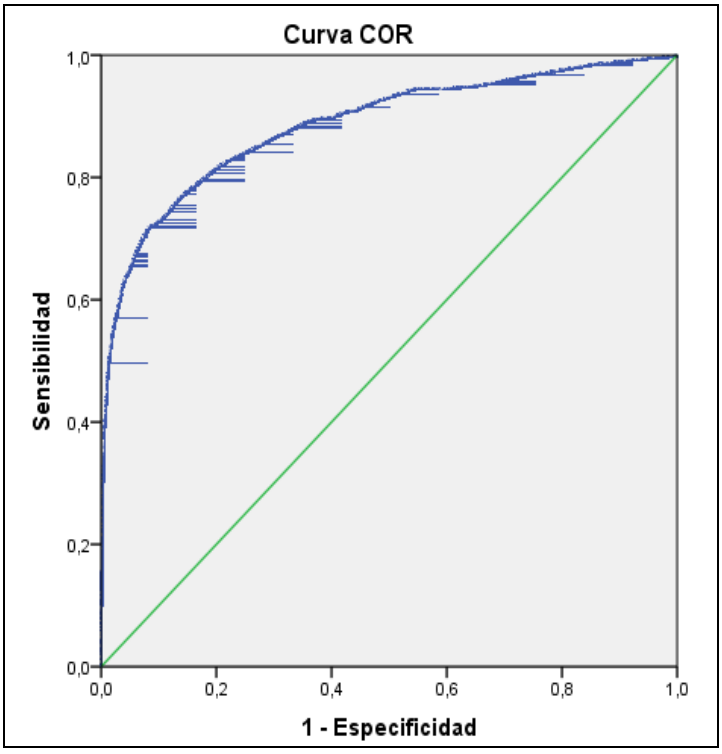


Figura 3-1

Dis_1	Dis1_1	Dis1_2	Probabilidad_Fraude_Global
1,00	,86908	,16712	,83288
1,00	2,10327	,03002	,96998
,00	-1,43787	,86847	,13153
,00	-1,31299	,84533	,15467
,00	-1,22853	,82785	,17215
,00	-1,42975	,86706	,13294
1,00	,13323	,37947	,62053
1,00	-,11969	,47283	,52717
1,00	,14150	,37652	,62348
1,00	1,03191	,13554	,86446
1,00	,48407	,26442	,73558
1,00	1,47543	,07416	,92584
1,00	,69154	,20795	,79205
1,00	,95179	,15040	,84960
,00	-,90405	,74632	,25368
,00	-,64634	,66570	,33430

Tabla 3-12

Para calcular la propensión al fraude de un individuo futuro (individuo que no está en la muestra) tenemos que tener presente que las funciones discriminantes se calculan a partir de las componentes principales y que los valores de las componentes se calculan a partir de sus combinaciones lineales en función de las variables iniciales, tal y como ya hemos visto en el apartado anterior.

Por lo tanto, una vez calculadas las componentes Ci en función de las variables independientes iniciales del modelos, se calculan ya las funciones discriminantes en función de las componentes ya conocidas numéricaamente

$$\mathbf{D0 = -0{,}776 - 0{,}378C_1 +\text{--------}}$$

$$\mathbf{D1 = -0{,}942 + 0{,}226C_1 +\text{--------}}$$

Finalmente se puede calcular la probabilidad de que el individuo no defraude y defraude mediante las expresiones:

$$P_0 = \frac{e^{D_0}}{e^{D_0} + e^{D_1}}$$

$$P_1 = \frac{e^{D_1}}{e^{D_0} + e^{D_1}}$$

P1 resulta ser la propensión al fraude del individuo futuro externo a la muestra.

Las funciones discriminantes no es necesario calcularlas todos los años. El poder predictivo de las mismas puede ser válido para un período en el que no haya reformas impositivas significativas.

3.6.2 Modelo discriminante para el fraude relativo al tipo marginal

El modelo para el fraude relativo al tipo marginal tendrá la expresión:

$$f\_tmg = F(C_1, C_2, \cdots, C_n) + e$$

Estimaremos las funciones discriminantes de Fisher:

$$D_i = u_{i1}C_1 + u_{i2}C_2 + \cdots + u_{ik}C_k \quad i=1,2$$

correspondientes a los grupos de fraude y no fraude (categorías de la variable dependiente del modelo *f_tmg*).

A continuación se muestran los coeficientes estimados de las dos funciones discriminantes de Fisher (tabla 3-13).

Coeficientes de función de clasificación

	Fraude que afecta al tipo marginal	
	,00	1,00
REGR factor score 1 for analysis 1	-,115	,199
REGR factor score 2 for analysis 1	-,015	,026
REGR factor score 3 for analysis 1	-,013	,022
REGR factor score 4 for analysis 1	,000	-,001
REGR factor score 5 for analysis 1	-,157	,272
REGR factor score 6 for analysis 1	-,038	,066
REGR factor score 7 for analysis 1	,033	-,058
REGR factor score 8 for analysis 1	-,003	,005

Tabla 3-13

Se obtienen los coeficientes de las dos funciones discriminantes de Fisher.

$$\mathbf{D0 = -0{,}115C1 - 0{,}15C_2 + \text{--------} - 0{,}776}$$

$$\mathbf{D1 = 0{,}199C1 + 0{,}026C_2 + \text{-------} - 0{,}942}$$

En cuanto a la diagnosis del modelo, observamos los resultados del contraste de la Lambda de Wilks cuyo p-valor (Sig.) pequeño valida la significatividad del modelo discriminante en su conjunto (tabla 3-14). También observamos los resultados de la prueba M de Box, cuyo p-valor pequeño muestra la ausencia de heteroscedasticidad en el modelo (tabla 3-15). La ausencia de multicolinealidad viene determinada por el uso de los factores que son incorrelados y la hipótesis de normalidad la asegura el uso de factores rotados y el teorema central del límite, como ya se ha indicado.

Lambda de Wilks

Prueba de funciones	Lambda de Wilks	Chi-cuadrado	gl	Sig.
1	,777	487629,705	64	,000

Tabla 3-14

Resultados de pruebas

M de Box		59467686,78
F	Aprox.	28589,180
	df1	2080
	df2	6,791E+12
	Sig.	,000

Prueba la hipótesis nula de las matrices de covarianzas de población iguales.

Tabla 3-15

Si analizamos la matriz de confusión (tabla 3-16) que se presenta a continuación, se observa que se clasifican bien el 77,5% de los individuos muestrales, valor bastante elevado que indica un buen ajuste del modelo. En el caso del fraude global este porcentaje de clasificación correcta era ligeramente superior (79,1%).

Resultados de clasificación[a]

		Fraude que afecta al tipo marginal	Pertenencia a grupos pronosticada		Total
			,00	1,00	
Original	Recuento	,00	1046113	175634	1221747
		1,00	258136	448611	706747
	%	,00	85,6	14,4	100,0
		1,00	36,5	63,5	100,0

a. 77,5% de casos agrupados originales clasificados correctamente.

Tabla 3-16

A continuación se muestra el área bajo la curva ROC correspondiente al modelo discriminante. Se observa que en nuestro modelo el área es 0,832 (tabla 3-17), valor bastante alto que indica que el modelo presenta un ajuste de calidad. En el caso del fraude global esta área era mayor (0,886). También se presenta la gráfica de la curva ROC (figura 3-2).

Área bajo la curva

Variable(s) de resultado de prueba: Probabilidades de pertenencia a grupo 1

Área	Error estándar[a]	Significación asintótica[b]	95% de intervalo de confianza asintótico	
			Límite inferior	Límite superior
,832	,000	,000	,831	,832

Tabla 3-17

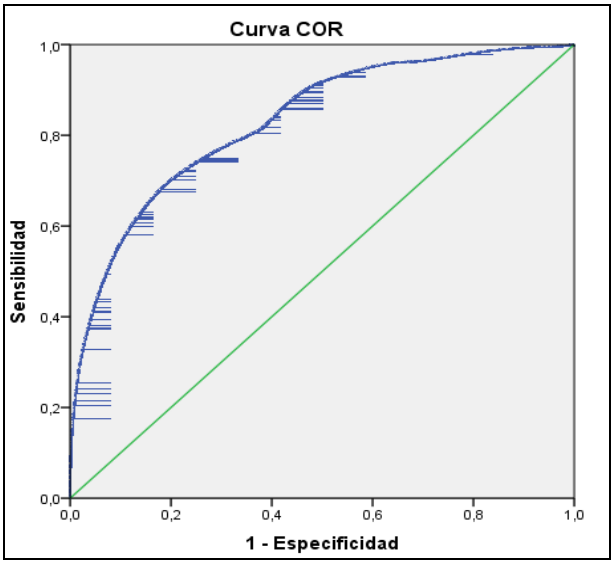


Figura 3-2

A continuación se muestra una tabla con el grupo de pertenencia de los individuos declarantes de IRP (1=fraude, 0=no fraude en la variable Disk_2), los valores de las dos funciones discriminantes de Fisher para cada individuo (Dis1_3=función discriminante para el grupo de fraude, Dis1_4=función discriminante para el grupo de no f raude) y las probabilidades de fraude por tipo marginal de los individuos, que cuantifican su propensión al fraude (tabla 3-18).

Dis_2	Dis1_3	Dis1_4	Probabilidad_Fraude_Tmg
1,00	1,73730	,14573	,85427
1,00	,58243	,38157	,61843
,00	-,94629	,77186	,22814
,00	-,85639	,75376	,24624
,00	-,86204	,75493	,24507
,00	-,95817	,77418	,22582
,00	-,15189	,58286	,41714
,00	-,19548	,59461	,40539
1,00	1,13868	,24935	,75065
1,00	,30226	,45735	,54265
1,00	,40508	,42911	,57089
1,00	,21997	,48016	,51984
,00	-,09108	,56632	,43368
,00	-,06236	,55845	,44155
,00	-,82094	,74636	,25364
,00	-,55466	,68630	,31370

Tabla 3-18

Para calcular la propensión al fraude por tipo marginal de un individuo futuro externo a la muestra se utilizara la expresión:

$$P_1 = \frac{e^{D_1}}{e^{D_0} + e^{D_1}}$$

3.6.3 Modelo discriminante para el fraude relativo a las actividades económicas

El modelo para el fraude que afecta a la declaración de actividades económicas tendrá la expresión:

$$f\_aaee = F(C_1, C_2, \cdots, C_n) + e$$

Estimaremos las funciones discriminantes de Fisher:

$$D_i = u_{i1}C_1 + u_{i2}C_2 + \cdots + u_{ik}C_k \quad i=1,2$$

correspondientes a los grupos de fraude y no fraude (categorías de la variable dependiente del modelo *f_aaee*).

A continuación se muestran los coeficientes estimados de las dos funciones discriminantes de Fisher (tabla 3-19).

Coeficientes de función de clasificación

	Fraude que afecta a la declaración de actividades económicas	
	,00	1,00
REGR factor score 1 for analysis 1	-,404	,762
REGR factor score 2 for analysis 1	,026	-,049
REGR factor score 3 for analysis 1	,061	-,115
REGR factor score 4 for analysis 1	,001	-,003
REGR factor score 5 for analysis 1	,037	-,069
REGR factor score 6 for analysis 1	-,620	1,170
REGR factor score 7 for analysis 1	,146	-,276

Tabla 3-19

Se obtienen los coeficientes de las dos funciones discriminantes de Fisher.

$$\mathbf{D0 = -0{,}404C1 + 0{,}26C_2 + \text{-------} - 0{,}003C_{64} - 1{,}006}$$

$$\mathbf{D1 = -0{,}762C1 - 0{,}049C_2 + \text{-------} + 0{,}005C_{64} - 1{,}808}$$

En cuanto a la diagnosis del modelo, observamos los resultados del contraste de la Lambda de Wilks cuyo p-valor (Sig.) pequeño valida la significatividad del modelo discriminante en su conjunto (tabla 3-20).

También observamos los resultados de la prueba M de Box, cuyo p-valor pequeño muestra la ausencia de heteroscedasticidad en el modelo (tabla 3-21). La ausencia de multicolinealidad viene determinada por el uso de los factores que son incorrelados y la hipótesis de normalidad la asegura el uso de factores rotados y el teorema central del límite, como ya se ha indicado.

Lambda de Wilks

Prueba de funciones	Lambda de Wilks	Chi-cuadrado	gl	Sig.
1	,458	1504390,149	64	,000

Tabla 3-20

Resultados de pruebas

M de Box		49123772,16
F	Aprox.	23616,299
	df1	2080
	df2	5,906E+12
	Sig.	,000

Prueba la hipótesis nula de las matrices de covarianzas de población iguales.

Tabla 3-21

Si analizamos la matriz de confusión que se presenta en la tabla 3-22, se observa que se clasifican bien el 89,7% de los individuos muestrales, valor bastante elevado que indica un buen ajuste del modelo. En el caso del fraude global y del fraude por tipo marginal este porcentaje de clasificación correcta era inferior.

Resultados de clasificación[a]

		Fraude que afecta a la declaración de actividades económicas	Pertenencia a grupos pronosticada		Total
			,00	1,00	
Original	Recuento	,00	1155248	105345	1260593
		1,00	92732	575169	667901
	%	,00	91,6	8,4	100,0
		1,00	13,9	86,1	100,0

a. 89,7% de casos agrupados originales clasificados correctamente.

Tabla 3-22

A continuación se muestra el área bajo la curva ROC correspondiente al modelo discriminante. Se observa que en nuestro modelo el área es 0,947 (tabla 3-23), valor bastante alto que indica que el modelo presenta un ajuste de calidad. En el caso del fraude global y el fraude por tipo marginal esta área era menor. También se presenta la gráfica de la curva ROC (Figura 3-3).

Área bajo la curva

Variable(s) de resultado de prueba: Probabilidades de pertenencia a grupo 1

Área	Error estándar[a]	Significación asintótica[b]	95% de intervalo de confianza asintótico	
			Límite inferior	Límite superior
,947	,000	,000	,946	,947

Tabla 3-23

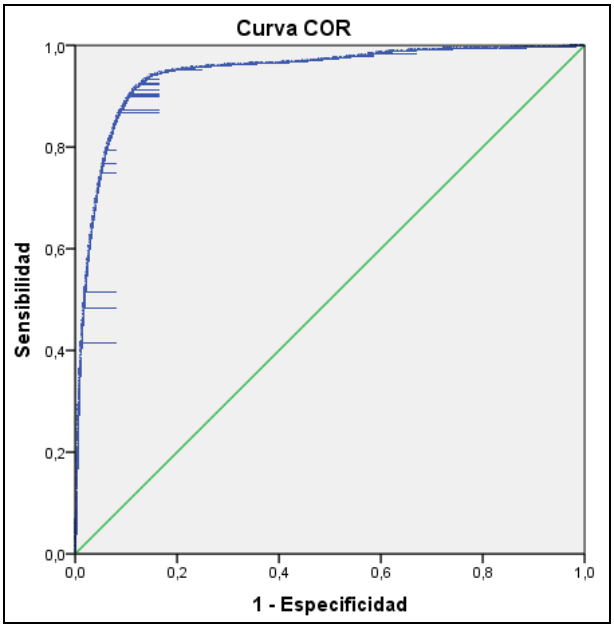


Figura 3-3

A continuación se muestra la tabla 3-24 con el grupo de pertenencia de los individuos declarantes de IRP (1=fraude, 0=no fraude en la variable Dis_3), los valores de las dos funciones discriminantes de Fisher para cada individuo (Dis1_5=función discriminante para el grupo de fraude,

Dis1_6=función discriminante para el grupo de no fraude) y las probabilidades de fraude por actividades económicas de los individuos, que cuantifican su propensión al fraude.

Dis_3	Dis1_5	Dis1_6	Probabilidad_Fraude_aaee
1,00	,52388	,40256	,59744
,00	,08071	,64970	,35030
,00	-1,28680	,97684	,02316
,00	-1,33086	,97901	,02099
,00	-,86824	,94190	,05810
,00	-1,12296	,96668	,03332
1,00	,38381	,48131	,51869
,00	,08621	,64683	,35317
,00	-,77019	,92836	,07164
1,00	,76325	,28056	,71944
,00	,19387	,58884	,41116
1,00	2,38234	,00956	,99044
1,00	1,25870	,11169	,88831
1,00	1,91006	,02760	,97240
,00	-1,32609	,97879	,02121
,00	-,36022	,83550	,16450

Tabla 3-24

Para calcular la propensión al fraude por actividades económicas de un individuo futuro externo a la muestra se utilizara la expresión:

$$P_1 = \frac{e^{D_1}}{e^{D_0} + e^{D_1}}$$

3.6.4 Modelo discriminante para el fraude relativo a la declaración de gastos

El modelo para el fraude que afecta a la declaración de gastos tendrá la expresión:

$$f\_gastos = F(C_1, C_2, \cdots, C_n) + e$$

Estimaremos las funciones discriminantes de Fisher:

$$D_i = u_{i1}C_1 + u_{i2}C_2 + \cdots + u_{ik}C_k \quad i=1,2$$

correspondientes a los grupos de fraude y no fraude (categorías de la variable dependiente del modelo *f_gastos*).

En la tabla 3-25 se muestran los coeficientes estimados de las dos funciones discriminantes de Fisher.

Coeficientes de función de clasificación

	Fraude que afecta a la declaración de gastos	
	,00	1,00
REGR factor score 1 for analysis 1	-,019	,130
REGR factor score 2 for analysis 1	,000	,002
REGR factor score 3 for analysis 1	-,011	,073
REGR factor score 4 for analysis 1	-,002	,011
REGR factor score 5 for analysis 1	-,007	,044
REGR factor score 6 for analysis 1	-,020	,137

Tabla 3-25

Se obtienen los coeficientes de las dos funciones discriminantes de Fisher.

$$\mathbf{D0 = -0{,}019C1 + \text{--------} - 0{,}003C_{64} - 0{,}712}$$

$$\mathbf{D1 = -0{,}130C1 + \text{--------} + 0{,}005C_{64} - 1{,}557}$$

En cuanto a la diagnosis del modelo, observamos los resultados del contraste de la Lambda de Wilks cuyo p-valor (Sig.) pequeño valida la significatividad del modelo discriminante en su conjunto (tabla 3-26). También observamos los resultados de la prueba M de Box, cuyo p-valor pequeño muestra la ausencia de heteroscedasticidad en el modelo (tabla 3-27). La ausencia de multicolinealidad viene determinada por el uso de los factores

que son incorrelados y la hipótesis de normalidad la asegura el uso de factores rotados y el teorema central del límite, como ya se ha indicado.

Lambda de Wilks

Prueba de funciones	Lambda de Wilks	Chi-cuadrado	gl	Sig.
1	,797	437948,523	64	,000

Tabla 3-26

Log determinante

Fraude que afecta a la declaración de gastos	Rango	Determinante de logaritmo
,00	64	-34,241
1,00	64	14,560
Dentro de grupos combinados	64	-,227

Los logaritmos naturales y los rangos de determinantes impresos son los de las matrices de covarianzas de grupo.

Resultados de pruebas

M de Box		53489354,03
F	Aprox.	25713,765
	df1	2080
	df2	5,907E+11
	Sig.	,000

Prueba la hipótesis nula de las matrices de covarianzas de población iguales.

Tabla 3-27

Si analizamos la matriz de confusión de la tabla 3-28, se observa que se clasifican bien el 89,5% de los individuos muestrales, valor bastante elevado que indica un buen ajuste del modelo. En el caso del fraude por actividades económicas este porcentaje de clasificación correcta era levemente superior (89,7%).

Resultados de clasificación[a]

		Fraude que afecta a la declaración de gastos	Pertenencia a grupos pronosticada ,00	Pertenencia a grupos pronosticada 1,00	Total
Original	Recuento	,00	1590981	89451	1680432
		1,00	112372	135690	248062
	%	,00	94,7	5,3	100,0
		1,00	45,3	54,7	100,0

a. 89,5% de casos agrupados originales clasificados correctamente.

Tabla 3-28

A continuación se muestra el área bajo la curva ROC correspondiente al modelo discriminante (tabla 3-29). Se observa que en nuestro modelo el área es 0,856, valor bastante alto que indica que el modelo presenta un ajuste de calidad. En el caso del fraude global y el fraude por tipo marginal esta área era menor. También se presenta la gráfica de la curva ROC (figura 3-4).

Área bajo la curva

Variable(s) de resultado de prueba: Probabilidades de pertenencia a grupo

Área	Error estándar[a]	Significación asintótica[b]	95% de intervalo de confianza asintótico	
			Límite inferior	Límite superior
,856	,000	,000	,855	,857

Tabla 3-29

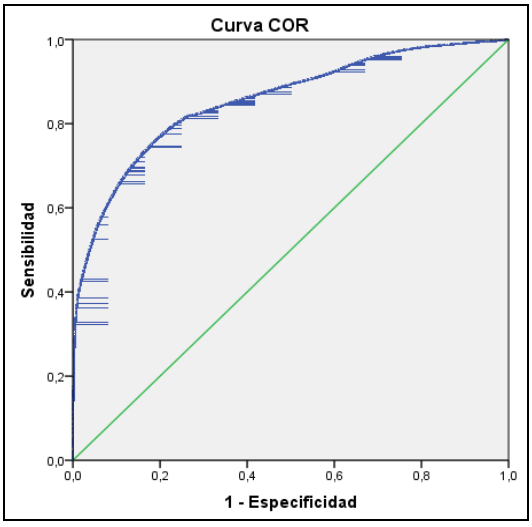


Figura 3-4

A continuación se muestra una tabla con el grupo de pertenencia de los individuos declarantes de IRP (1=fraude, 0=no fraude en la variable Dis_4), los valores de las dos funciones discriminantes de Fisher para cada individuo (Dis1_7=función discriminante para el grupo de fraude, Dis1_8=función discriminante para el grupo de no fraude) y las probabilidades de fraude por declaración de gastos de los individuos, que cuantifican su propensión al fraude (tabla 3-30).

Dis_4	Dis1_7	Dis1_8	Probabiliad_Fraude_Gastos
,00	-1,56812	,96120	,03880
1,00	3,95301	,00596	,99404
,00	-,63408	,85828	,14172
,00	-,33970	,79528	,20472
,00	-,80047	,88615	,11385
,00	-,87056	,89639	,10361
,00	,08104	,67316	,32684
,00	-,00367	,70062	,29938
,00	,42152	,55206	,44794
1,00	,58391	,49102	,50898
,00	,18447	,63796	,36204
,00	,31442	,59159	,40841
,00	,26032	,61114	,38886
1,00	1,60426	,17153	,82847
,00	,22777	,62275	,37725

Tabla 3-30

Para calcular la propensión al fraude por actividades económicas de un individuo futuro externo a la muestra se utilizara la expresión:

$$P_1 = \frac{e^{D_1}}{e^{D_0} + e^{D_1}}$$

3.6.5 Modelo discriminante para el fraude relativo a los planes de pensiones

El modelo para el fraude que afecta a los planes de pensiones tendrá la expresión:

$$f\_planp = F(C_1, C_2, \cdots, C_n) + e$$

Estimaremos las funciones discriminantes de Fisher:

$$D_i = u_{i1}C_1 + u_{i2}C_2 + \cdots + u_{ik}C_k \qquad i=1,2$$

correspondientes a los grupos de fraude y no fraude (categorías de la variable dependiente del modelo *f_planp*).

En la tabla 3-31 se muestran los coeficientes estimados de las dos funciones discriminantes de Fisher.

Coeficientes de función de clasificación

	Fraude que afecta a la desgrbación por planes de pensiones	
	,00	1,00
REGR factor score 1 for analysis 1	-,203	,529
REGR factor score 2 for analysis 1	,011	-,029
REGR factor score 3 for analysis 1	,020	-,051
REGR factor score 4 for analysis 1	-,002	,004
REGR factor score 5 for analysis 1	,005	-,014
REGR factor score 6 for analysis 1	-,243	,634
REGR factor score 7 for analysis 1	,061	-,158

Tabla 3-31

Se obtienen las dos funciones discriminantes de Fisher.

$$\mathbf{D0 = -0{,}203C1 + 0{,}011C_2 + \text{--------} + 0{,}001C_{64} - 0{,}830}$$

$$\mathbf{D1 = 0{,}529C1 - 0{,}029C_2 + \text{--------} - 0{,}003C_{64} - 1{,}625}$$

En cuanto a la diagnosis del modelo, observamos los resultados del contraste de la Lambda de Wilks cuyo p-valor (Sig.) pequeño valida la significatividad del modelo discriminante en su conjunto (tabla 3-32). También observamos los resultados de la prueba M de Box, cuyo p-valor pequeño muestra la ausencia de heteroscedasticidad en el modelo (tabla 3-33). La ausencia de multicolinealidad viene determinada por el uso de los factores que son incorrelados y la hipótesis de normalidad la asegura el uso de factores rotados y el teorema central del límite, como ya se ha indicado.

Log determinante

Fraude que afecta a la desgrbación por planes de pensiones	Rango	Determinante de logaritmo
,00	64	-33,072
1,00	64	9,087
Dentro de grupos combinados	64	-,539

Los logaritmos naturales y los rangos de determinantes impresos son los de las matrices de covarianzas de grupo.

Resultados de pruebas

M de Box		40196734,63
F	Aprox.	19324,494
	df1	2080
	df2	3,379E+12
	Sig.	,000

Prueba la hipótesis nula de las matrices de covarianzas de población iguales.

Tabla 3-32

Lambda de Wilks

Prueba de funciones	Lambda de Wilks	Chi-cuadrado	gl	Sig.
1	,583	1040037,721	64	,000

Tabla 3-33

Si analizamos la matriz de confusión que se presenta en la figura 3-34, se observa que se clasifican bien el 85,7% de los individuos muestrales, valor elevado que indica un buen ajuste del modelo. En el caso del fraude por actividades económicas, por gastos y por tipo marginal, este porcentaje de clasificación correcta era levemente superior.

Resultados de clasificación[a]

		Fraude que afecta a la desgrbación por planes de pensiones	Pertenencia a grupos pronosticada		Total
			,00	1,00	
Original	Recuento	,00	1259546	134262	1393808
		1,00	141127	393559	534686
	%	,00	90,4	9,6	100,0
		1,00	26,4	73,6	100,0

a. 85,7% de casos agrupados originales clasificados correctamente.

Tabla 3-34

En la figura 3-35 se muestra el área bajo la curva ROC correspondiente al modelo discriminante. Se observa que en nuestro modelo el área es 0,900, valor muy alto que indica que el modelo presenta un ajuste de calidad. También se presenta la gráfica de la curva ROC (figura 3-5).

Área bajo la curva

Variable(s) de resultado de prueba: Probabilidades de pertenencia a grupo 1

Área	Error estándar[a]	Significación asintótica[b]	95% de intervalo de confianza asintótico	
			Límite inferior	Límite superior
,900	,000	,000	,899	,900

Tabla 3-35

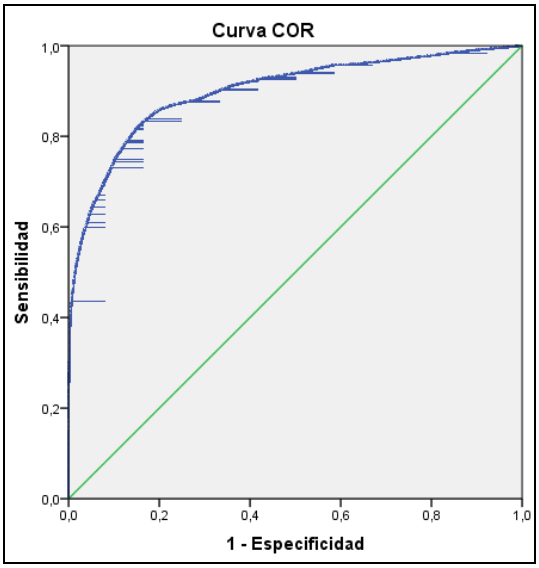


Figura 3-5

A continuación se muestra una tabla con el grupo de pertenencia de los individuos declarantes de IRP (1=fraude, 0=no fraude en la variable Dis_5), los valores de las dos funciones discriminantes de Fisher para cada individuo (Dis1_9=función discriminante para el grupo de fraude, Dis1_10=función discriminante para el grupo de no fraude) y las probabilidades de fraude por declaración de planes de pensiones de los individuos, que cuantifican su propensión al fraude (tabla 3-36).

Dis_5	Dis1_9	Dis1_10	Probabilidad_Fraude_Planp
,00	-,07328	,71768	,28232
,00	,06075	,66370	,33630
,00	-,98941	,93483	,06517
,00	-,94446	,92946	,07054
,00	-,81240	,91125	,08875
,00	-,96454	,93191	,06809
,00	,07840	,65622	,34378
,00	-,05105	,70910	,29090
,00	-,46021	,84075	,15925
,00	,31511	,54969	,45031
,00	,04201	,67156	,32844
1,00	,93014	,27644	,72356
1,00	,68645	,37709	,62291
1,00	,73008	,35794	,64206
,00	-,94644	,92970	,07030
,00	-,54881	,86190	,13810

Tabla 3-36

3.6.6 Modelo discriminante para el fraude que afecta a las declaraciones del número de hijos y ascendientes y descendientes

El modelo para el fraude que afecta a las declaración del número de hijos, ascendientes y descendientes tendrá la expresión:

$$f\_nhijos = F(C_1, C_2, \cdots, C_n) + e$$

Estimaremos las funciones discriminantes de Fisher:

$$D_i = u_{i1}C_1 + u_{i2}C_2 + \cdots + u_{ik}C_k \qquad i=1,2$$

correspondientes a los grupos de fraude y no fraude (categorías de la variable dependiente del modelo *f_nhijos*).

A continuación se muestran los coeficientes estimados de las dos funciones discriminantes de Fisher (tabla 3-37).

Coeficientes de función de clasificación

	Fraude que afecta a las declaraciones de número e hijos y ascendieentes y descendientes	
	,00	1,00
REGR factor score 1 for analysis 1	-,014	,299
REGR factor score 2 for analysis 1	,000	-,008
REGR factor score 3 for analysis 1	,001	-,020
REGR factor score 4 for analysis 1	-,001	,015
REGR factor score 5 for analysis 1	,002	-,053
REGR factor score 6 for analysis 1	-,017	,357
REGR factor score 7 for analysis 1	,003	-,069
REGR factor score 8 for analysis 1	,001	-,016

Tabla 3-37

Se obtienen las dos funciones discriminantes de Fisher.

$$\mathbf{D0 = -0{,}0014C1 + \text{--------} - 0{,}697}$$

$$\mathbf{D1 = 0{,}299C1 + \text{--------} - 2{,}367}$$

En cuanto a la diagnosis del modelo, observamos los resultados del contraste de la Lambda de Wilks cuyo p-valor (Sig.) pequeño valida la significatividad del modelo discriminante en su conjunto (tabla 3-38). También observamos los resultados de la prueba M de Box, cuyo p-valor pequeño muestra la ausencia de heteroscedasticidad en el modelo (tabla 3-39). La ausencia de multicolinealidad viene determinada por el uso de los factores que son incorrelados y la hipótesis de normalidad la asegura el uso de factores rotados y el teorema central del límite, como ya se ha indicado.

Log determinante

Fraude que afecta a las declaraciones de número e hijos y ascendieentes y descendientes	Rango	Determinante de logaritmo
,00	64	-7,992
1,00	64	-11,237
Dentro de grupos combinados	64	-,146

Los logaritmos naturales y los rangos de determinantes impresos son los de las matrices de covarianzas de grupo.

Resultados de pruebas

M de Box		15411781,40
F	Aprox.	7407,665
	df1	2080
	df2	6,798E+10
	Sig.	,000

Prueba la hipótesis nula de las matrices de covarianzas de población iguales.

Tabla 3-38

Lambda de Wilks

Prueba de funciones	Lambda de Wilks	Chi-cuadrado	gl	Sig.
1	,864	281648,295	64	,000

Tabla 3-39

Si analizamos la matriz de confusión de la tabla 3-40, se observa que se clasifican bien el 86,7% de los individuos muestrales, valor elevado que indica un buen ajuste del modelo. En el caso del fraude por actividades económicas, por gastos y por tipo marginal, este porcentaje de clasificación correcta era levemente superior. En el caso de los planes de pensiones ese porcentaje es prácticamente el mismo.

Resultados de clasificación[a]

		Fraude que afecta a las declaraciones de número e hijos y ascendieentes y descendientes	Pertenencia a grupos pronosticada		
			,00	1,00	Total
Original	Recuento	,00	1603994	237999	1841993
		1,00	19144	67357	86501
	%	,00	87,1	12,9	100,0
		1,00	22,1	77,9	100,0

a. 86,7% de casos agrupados originales clasificados correctamente.

Tabla 3-40

A continuación se muestra el área bajo la curva ROC correspondiente al modelo discriminante (tabla 3-41). Se observa que en nuestro modelo el área es 0,895, valor alto que indica que el modelo presenta un ajuste de calidad. También se presenta la gráfica de la curva ROC (figura 3-6).

Área bajo la curva

Variable(s) de resultado de prueba: Probabilidades de pertenencia a grupo 1

Área	Error estándar[a]	Significación asintótica[b]	95% de intervalo de confianza asintótico	
			Límite inferior	Límite superior
,895	,000	,000	,894	,896

Tabla 3-41

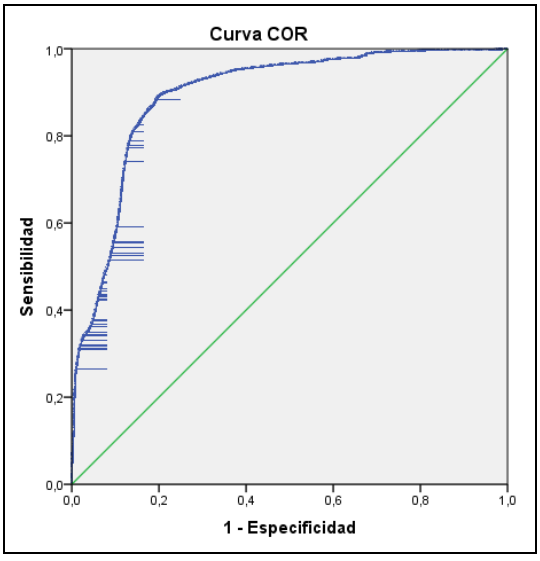


Figura 3-6

A continuación se muestra una tabla con el grupo de pertenencia de los individuos declarantes de IRP (1=fraude, 0=no fraude en la variable Dis_6), los valores de las dos funciones discriminantes de Fisher para cada individuo (Dis1_11=función discriminante para el grupo de fraude, Dis1_12=función discriminante para el grupo de no fraude) y las probabilidades de fraude por declaración del número de hijos, ascendientes y descendientes de los individuos, que cuantifican su propensión al fraude (tabla 3-42).

Dis_6	Dis1_11	Dis1_12	Probabilidad_Fraude_Nhijos
,00	,32472	,74049	,25951
1,00	3,05509	,01503	,98497
,00	-1,10163	,97771	,02229
,00	-1,10938	,97803	,02197
,00	-,74864	,95709	,04291
,00	-,96735	,97136	,02864
,00	,13679	,80353	,19647
,00	-,18169	,88275	,11725
,00	-,65170	,94879	,05121
1,00	2,02714	,09858	,90142
,00	,75416	,55620	,44380
,00	,84520	,51283	,48717
,00	-,16795	,87999	,12001

Tabla 3-42

3.7 EXTRACCIÓN DEL CONOCIMIENTO Y ANÁLISIS DE LOS PERFILES DE FRAUDE A TRAVÉS DEL ANÁLISIS DISCRIMINANTE

En este capítulo, se han elaborado modelos predictivos de Análisis Discriminante que permiten cuantificar la probabilidad que tiene cualquier contribuyente actual o futuro de ser defraudador por cada factor de fraude una vez que presente su declaración de IRPF. Estos modelos permitirán segmentar a los declarantes del impuesto por nivel de propensión al fraude y causas del mismo.

El ajuste matemático de todos los modelos de Análisis Discriminante también ha sido de calidad, con significatividad alta de los modelos, matrices de confusión con alto porcentaje de aciertos (según las tablas 3-43 a 3-48, 79,1% para el fraude global, 77,5% para el fraude por tipo marginal, 89,7% para el fraude por actividades económicas, 89,5% para el fraude por desgravación de gastos, 85,7% para el fraude por planes de pensiones y 86,7% para el fraude por declaración de hijos, ascendientes y descendientes) y áreas bajo la curva ROC muy cercanas a la unidad.

Resultados de clasificación[a]

			Pertenencia a grupos pronosticada		
		fraude global	,00	1,00	Total
Original	Recuento	,00	653735	66636	720371
		1,00	335498	872625	1208123
	%	,00	90,7	9,3	100,0
		1,00	27,8	72,2	100,0

a. 79,1% de casos agrupados originales clasificados correctamente.

Tabla 3-43

Resultados de clasificación[a]

			Pertenencia a grupos pronosticada		
		Fraude que afecta al tipo marginal	,00	1,00	Total
Original	Recuento	,00	1046113	175634	1221747
		1,00	258136	448611	706747
	%	,00	85,6	14,4	100,0
		1,00	36,5	63,5	100,0

a. 77,5% de casos agrupados originales clasificados correctamente.

Tabla 3-44

Resultados de clasificación[a]

			Pertenencia a grupos pronosticada		
		Fraude que afecta a la declaración de actividades económicas	,00	1,00	Total
Original	Recuento	,00	1155248	105345	1260593
		1,00	92732	575169	667901
	%	,00	91,6	8,4	100,0
		1,00	13,9	86,1	100,0

a. 89,7% de casos agrupados originales clasificados correctamente.

Tabla 3-45

Resultados de clasificación[a]

			Pertenencia a grupos pronosticada		
		Fraude que afecta a la declaración de gastos	,00	1,00	Total
Original	Recuento	,00	1590981	89451	1680432
		1,00	112372	135690	248062
	%	,00	94,7	5,3	100,0
		1,00	45,3	54,7	100,0

a. 89,5% de casos agrupados originales clasificados correctamente.

Tabla 3-46

Resultados de clasificación[a]

			Pertenencia a grupos pronosticada		
		Fraude que afecta a la desgrbación por planes de pensiones	,00	1,00	Total
Original	Recuento	,00	1259546	134262	1393808
		1,00	141127	393559	534686
	%	,00	90,4	9,6	100,0
		1,00	26,4	73,6	100,0

a. 85,7% de casos agrupados originales clasificados correctamente.

Tabla 3-47

Resultados de clasificación[a]

			Pertenencia a grupos pronosticada		
		Fraude que afecta a las declaraciones de número e hijos y ascendieentes y descendientes	,00	1,00	Total
Original	Recuento	,00	1603994	237999	1841993
		1,00	19144	67357	86501
	%	,00	87,1	12,9	100,0
		1,00	22,1	77,9	100,0

a. 86,7% de casos agrupados originales clasificados correctamente.

Tabla 3-48

Los métodos predictivos para el análisis del fraude permiten calcular perfiles de fraude a partir de la función de densidad de la probabilidad de fraude o propensión al fraude (global o por causas de fraude, para todos los individuos de la muestra). En las figuras 3-7 a 3-12 se comparan las

densidades de probabilidad para las propensiones al fraude, calculadas mediante el algoritmo del Kernel, para las distintas causas de fraude.

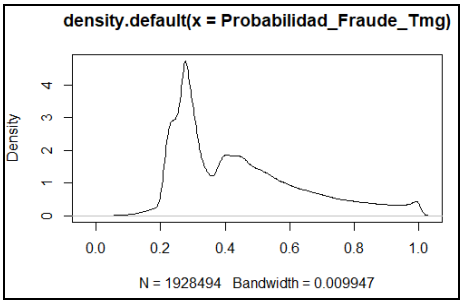


Figura 3-7

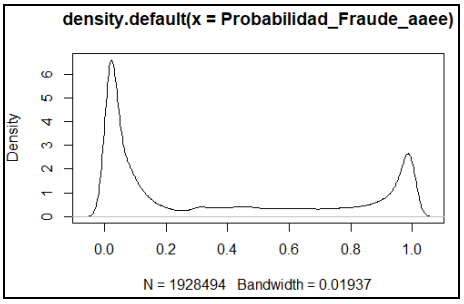


Figura 3-8

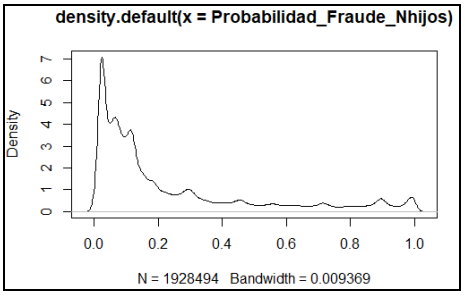


Figura 3-9

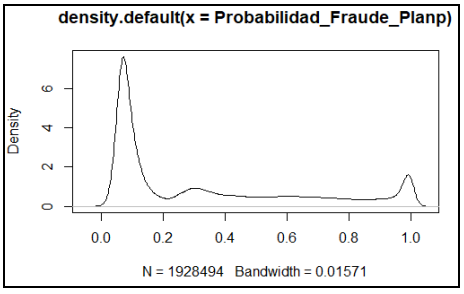


Figura 3-10

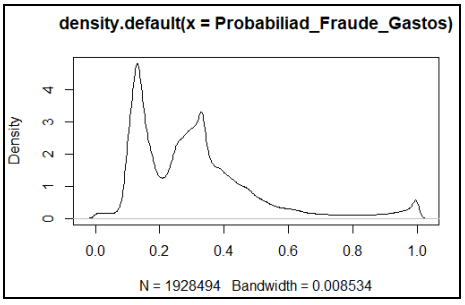


Figura 3-11

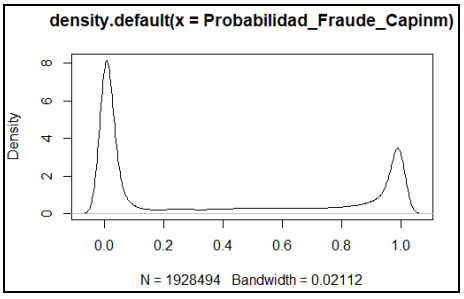


Figura 3-12

Se observa que la densidad de probabilidad para las diferentes causas de fraude es muy similar. Particularmente las propensiones al fraude en IRPF por actividades económicas, planes de pensiones y rendimientos de capital inmobiliario se comportan de forma muy similar. Lo mismo ocurre con las propensiones al fraude en IRPF por tipo marginal, número de hijos y

desgravación de gastos. Vemos así que los patrones de fraude no difieren mucho para las diferentes causas de fraude. Además, esta segmentación por patrones de fraude de las distintas cuasas de fraude coincide con la ya obtenida mediante árboles utilizando escalamiento multidimensional y que se presenta en la figura 3-13.

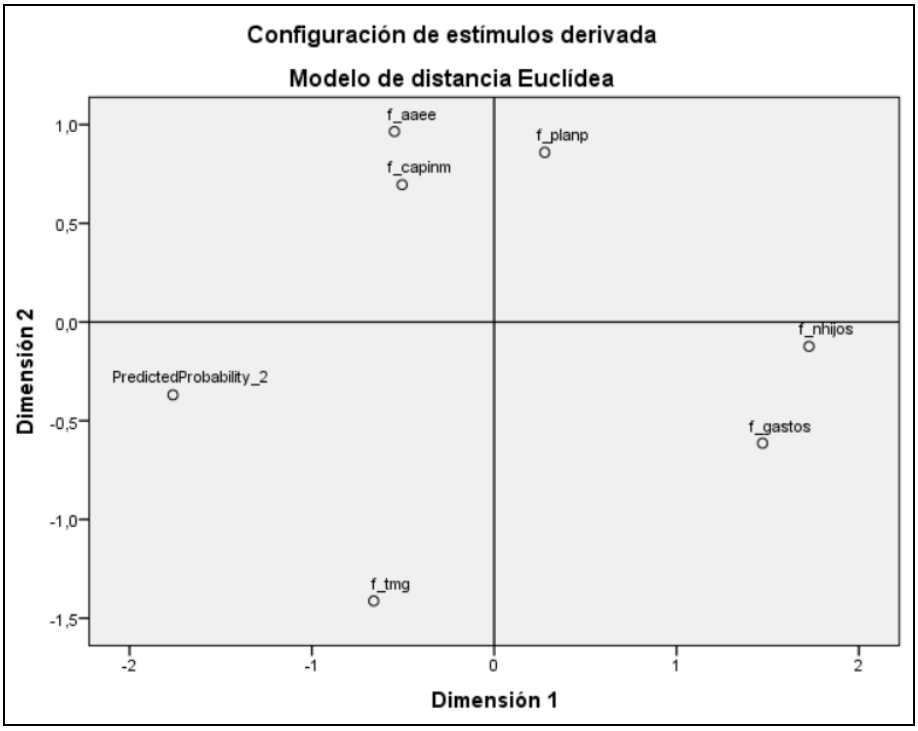


Figura 3-13

El Perfil del fraude global, a partir de su función de densidad se `presenta en la Figura 3-14.

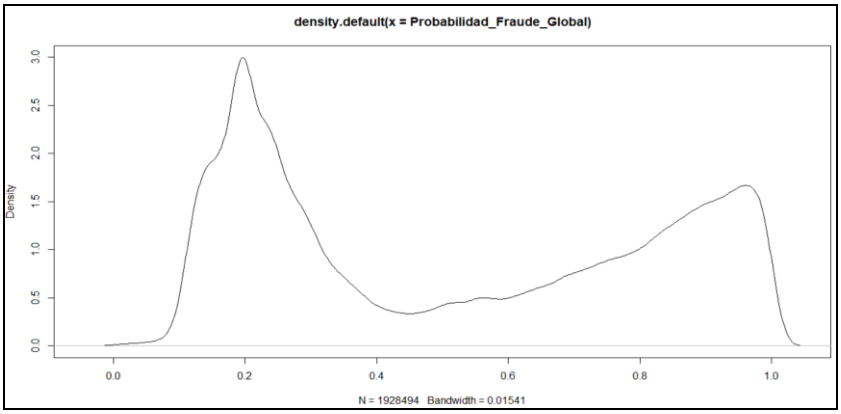


Figura 3-14

Observamos que para probabilidades de fraude bajas se encuentra la mayor densidad de contribuyentes, que presenta su máximo alrededor de la probabilidda de fraude 0,2. Se observa también que probabilidades de fraude altas entre 0,8 y 1 repunta la densidad de contribuyentes.

Por lo tanto observamos que la mayoría de contibuyentes no defraudan, pero los que defraudan tiene una propensión al fraude muy alta.

Observamos también que para los diferentes tipos de fruade, los perfiles de fraude dados por sus funciones de densidad, nos llevan a la misma conclusión que para el fraude global. Sencillamente vale con observar sus gráficas.

Un valor añadido de los modelos predictivos para el análisis del fraude, incluido el análisis discriminante, es la capacidad de predecir la probabilidad de fraude de cualquier nuevo contribuyente por IRPF que presente su modelo de declaración. Basta sustituir los valores de las variables del modelo para ese contribuyente en las dos funciones discriminantes (D_i) y calcular la probabilidad de fraude mediante la expresión:

$$P_1 = \frac{e^{D_1}}{e^{D_0} + e^{D_1}}$$

Finalmente, es necesario tener presente que el aspecto cuantitativo de este trabajo se desarrolla en el campo de los grandes datos (Big Data). Los modelos se ajustan para dos millones de registros y todos los gráficos tienen dos millones de puntos. Con software normal no sería posible realizar estas tareas. Se ha utilizado software de IBM y SAS que incorpora la posibilidad de trabajar en los campos del Big Data y la Minería de Datos.

3.8 SEGMENTACIÓN DE LAS CAUSAS DE FRAUDE A TRAVÉS DEL ANÁLISIS DISCRIMINANTE

Para segmentar las causas de fraude utilizaremos las técnicas estadísticas del Escalamiento Multidimensional y Analisis Cluster por variables.

3.8.1 Escalamiento Multidimensional

Recordamos que el Escalamiento multidimensional es una técnica descriptiva de minería de datos que permite segmentar variables de un conjunto de datos agrupándolas por similitud en un mapa perceptual.

Si aplicamos escalamiento multidimensional para ver como se relacionan las diversas causas de fraude con la probabilidad de fraaude global, obtenemos el mapa perceptual de la figura 3-15.

La segmentación del mapa perceptual nos indica que el fraude en actividades económicas, en planes de pensiones y en rendimientos capital inmobiliario tienen una incidencia similar en la probabilidda de fraude global. Lo mismo ocurre con el fraude por declaración incorrecta de gastos y número de hijos y ascendientes. El fraude en la alteración del tipo marginal se comporta aisladamente de las demás causas.

Para evaluar este tipo de escalamiento se utiliza el gráfico de disparidades de la figura 3-16.

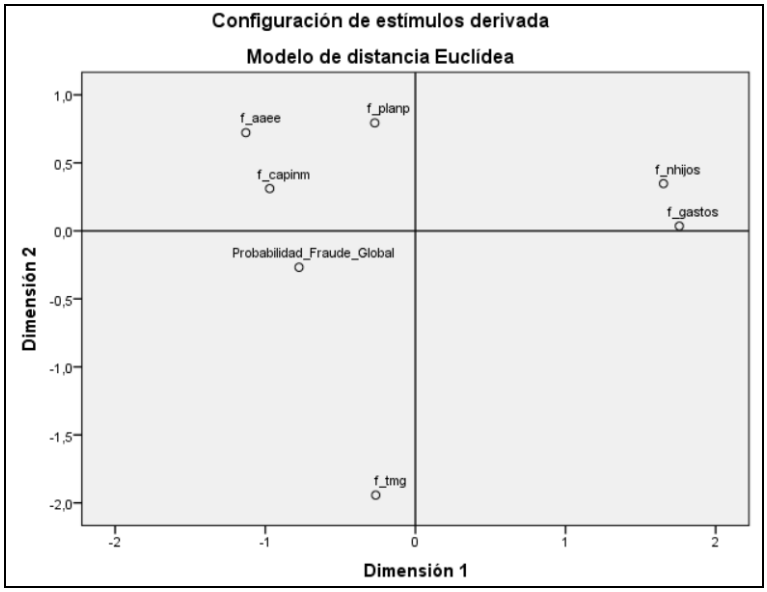


Figura 3-15

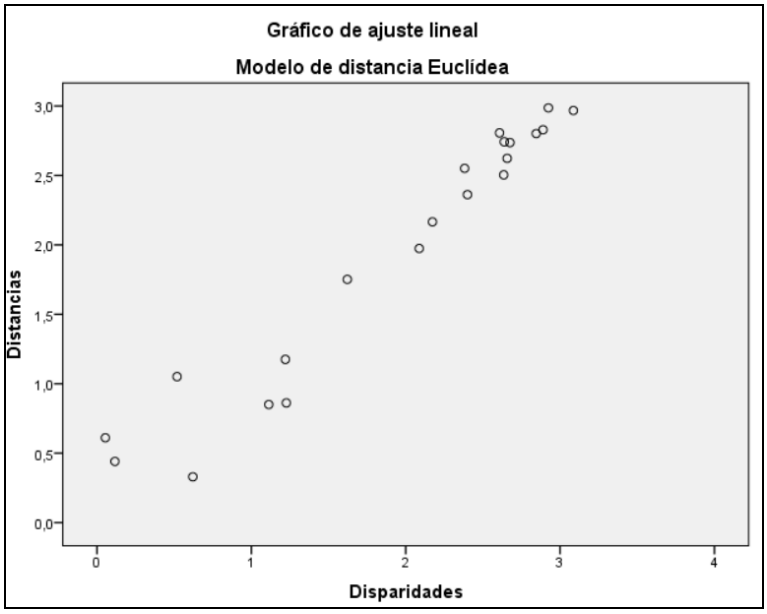


Figura 3-16

El escalamiento multidimemnsional utilizado es correcto porque el gráfico de disparidades presenta una nube de puntos que se ajusta bien a la diagonal del primer cuadrante.

Además, el estadístico S-Stress toma un valor bajo cercano a cero y el estadístico RSQ toma un valor alto cercano a la unidad

```
Stress  =   ,10807      RSQ =  ,94256
```

Concluimos que la segmentación realizada de las causa de fraude es correcta.

Si aplicamos escalamiento multidimensional para ver como se relacionan las probabilidades de las diversas causas de fraude con la probabilidad de fraude global, obtenemos el mapa perceptual de la figura 3-17.

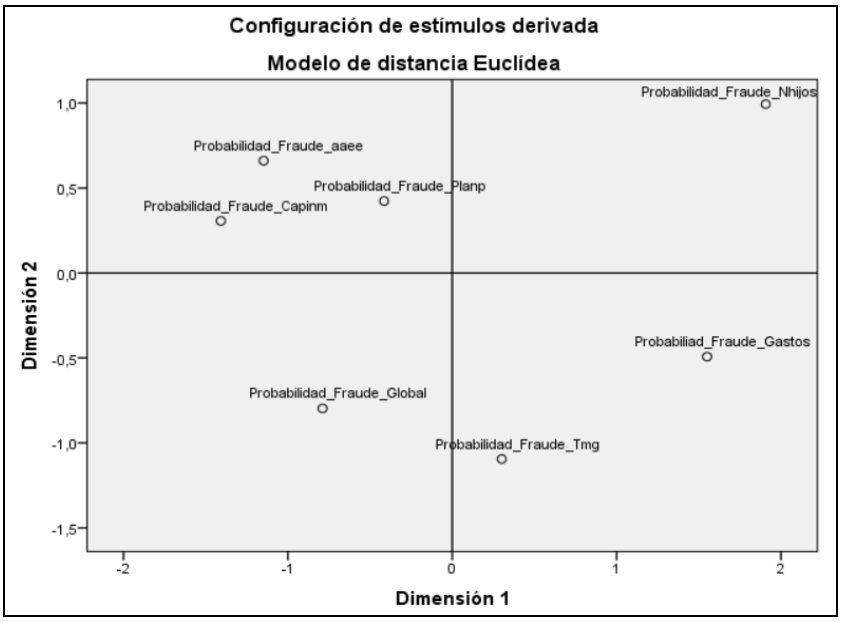


Figura 3-17

La segmentación del mapa perceptual nos indica que la probabilidad de fraude en actividades económicas, en planes de pensiones y en rendimientos de capital inmobiliario forman un segmento que indica que el fraude para esas tres causas se comporta de forma similar. Lo mismo ocurre con la probabilidad de fraude global y la del tipo marginal, lo cual puede indicar una fuerte incidencia del fraude en tipo marginal sobre el fraude global tal y como ya hemos visto en el caso de los árboles de decisión. También podría considerarse un segmento, pero no tan claro, la probabilidad de fraude por declaración incorrecta de gastos y número de hijos y ascendientes. Esta segmentación no dista mucho de la que venimos obteniendo para otras técnicas.

Como siempre, para evaluar este tipo de escalamiento se utiliza el gráfico de disparidades (figura 3-18):

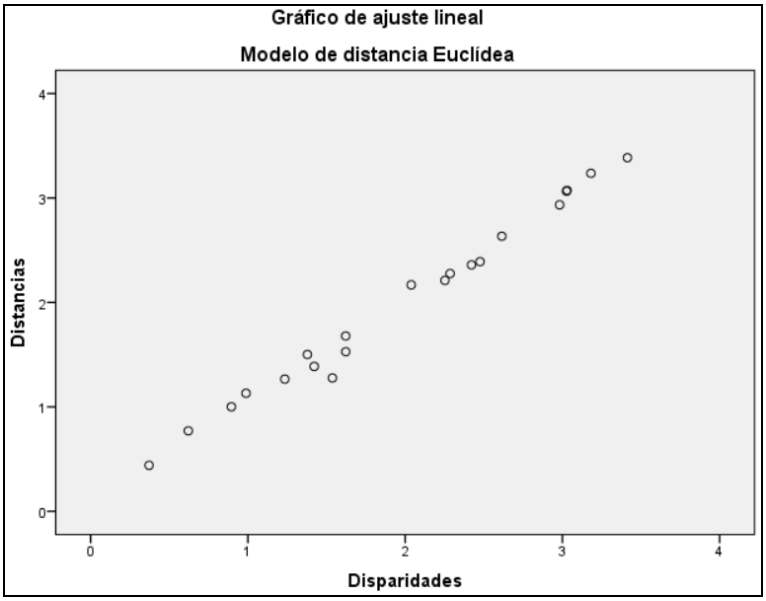


Figura 3-18

El escalamiento multidimemnsional utilizado es correcto porque el gráfico de disparidades presenta una nube de puntos que se ajusta bien a la diagonal del primer cuadrante.

Además, el estadístico S-Stress toma un valor bajo cercano a cero y el estadístico RSQ toma un valor alto cercano a la unidad

```
Stress  =   ,04468      RSQ =  ,98846
```

Concluimos que la segmentación realizada de las causa de fraude es correcta.

3.8.2 Análisis Cluster

El Análisis clúster es otra técnica descriptiva de minería de datos que permite segmentar variables de un conjunto de datos agrupándolas por similitud en un dendograma. Si aplicamos análisis clúster jerárquico por el método de Ward para segmentar las distintas tipologías de frude y el fraude global, obtenemos el dendograma de la figura 3-19.

Se observa que la segmentación de las causas de fraude es idéntica a la obtenida mediante escalamiento multidimensional. Se comportan de modo similar, por pertenecer al mismo cluster de nivel dos, el fraude en actividades económicas, el fraude en planes de pensiones y el fraude en rendimientos de capital inmobiliario. También se comportan de modo similar el fraude por incorrecta deducción de gastos y por declaración del número de hijos y ascendientes. El fraude por tipo marginal está aislado en el cluster porque su comportamiento es independiente del resto de las causas de fraude, resultado que también conocíamos y que coincide con lo ya contrastado en el capítulo anterior para el caso de los árboles de decisión.

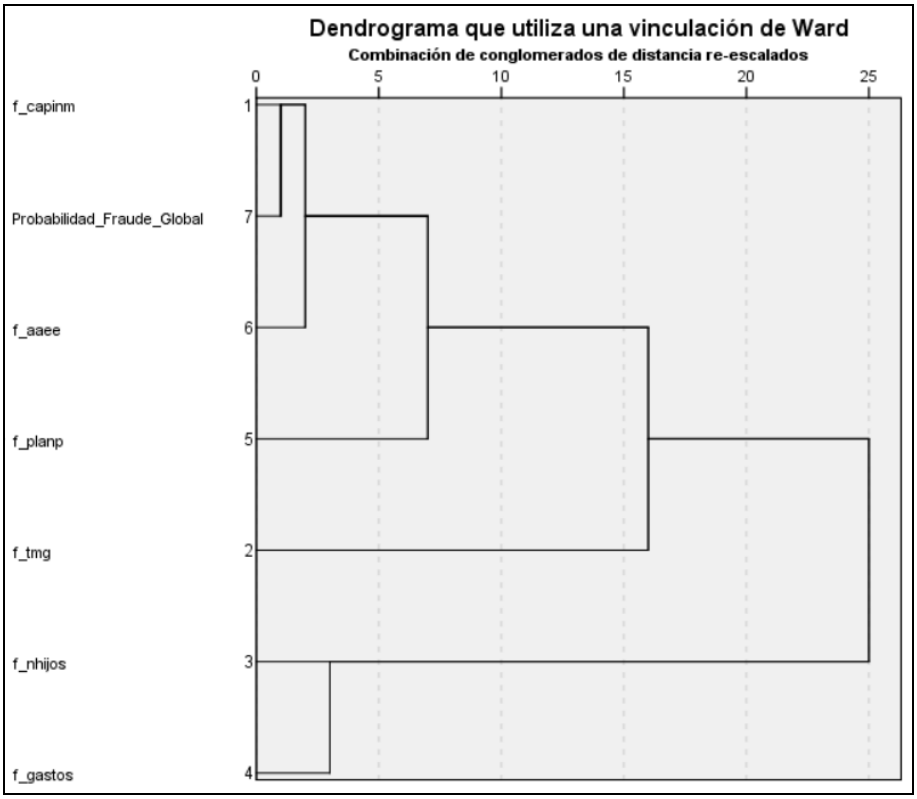


Figura 3-19

Si aplicamos análisis clúster jerárquico por el método de Ward para segmentar las probabilidades por las distintas tipologías de fraude y por fraude global, obtenemos el dendograma de la figura 3-20.

Se observa que la segmentación de las probabilidades de fraude para las distintas causas del mismo es idéntica a la obtenida mediante escalamiento multidimensional. Se comportan de modo similar, por pertenecer al mismo cluster de nivel dos, la probabilidad de fraude en actividades económicas, en planes de pensiones y en rendimientos de capital inmobiliario. También se comportan de modo similar la probablidad de fraude por incorrecta deducción de gastos y por declaración del número de hijos y ascendientes, comportamiento que se observa mucho mejor en el análisis cluster que en el

escalamiento multidimensional. La probabilidad de fraude por tipo marginal está muy asociada con la probabilidad de fraude global, tal y como ya hemos visto en el caso de los árboles de decisión.

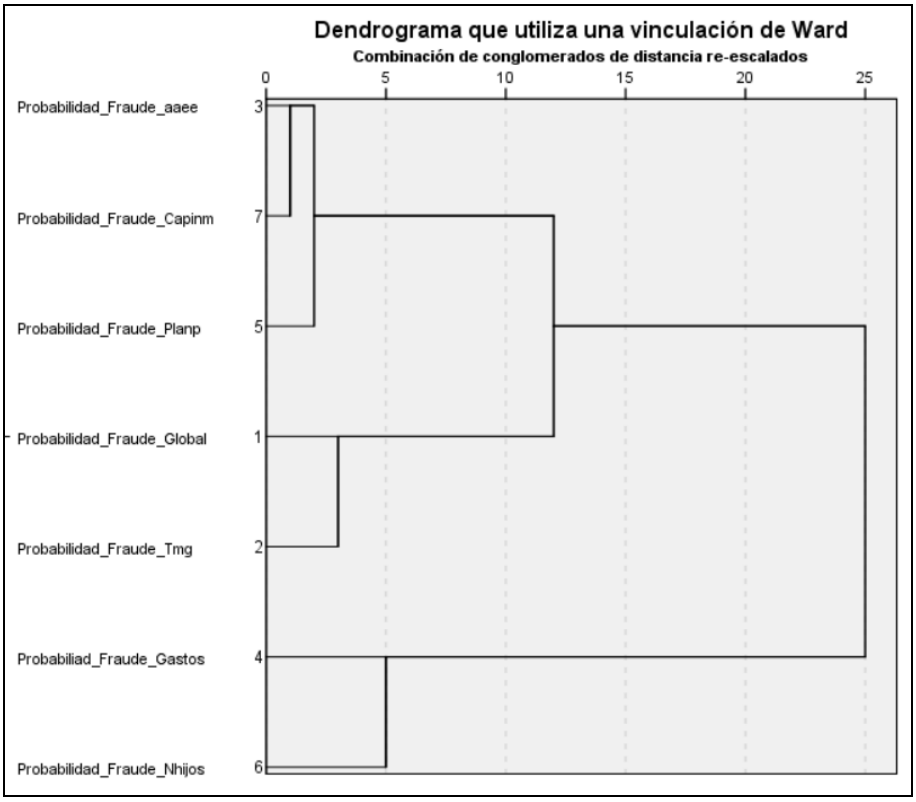


Figura 3-20

Capítulo 4

DETECCIÓN DEL FRAUDE FISCAL A TRAVÉS DE REDES NEURONALES

4.1 INTRODUCCIÓN

Este capítulo tiene como finalidad estudiar el fraude fiscal en el Impuesto sobre la Renta de las Personas Físicas mediante el uso de herramientas predictivas avanzadas de Machine Learning y Minería de Datos y en concreto mediante modelos de redes neuronales. Al igual que en todos los capítulos, se utilizará una metodología generalizable para cuantificar la propensión al fraude en cualquier otro impuesto según las causas que lo determinan.

Ampliamos aquí las posibilidades de análisis cuantitativo para estudiar el fraude al utilizar las nuevas prestaciones que aportan el Big Data, la Minería de Datos y las técnicas de Machine Learning desde la óptica de las Redes Neuronales. Dentro del grupo de técnicas de aprendizaje supervisado en el Machine Learning y dentro del grupo de técnicas predictivas de la Minería de Datos se encuentran las redes neuronales. En este trabajo se trata de mostrar el uso de los modelos de Redes Neuronales aplicados a las muestras de IRPF del IEF con la finalidad de estudiar las

partidas económicas más incidentes en el fraude fiscal en este impuesto y cuantificar la propensión al fraude de los declarantes. Precisamente, el valor añadido de este capítulo es profundizar en el estudio de las variables del modelo 100 de IRPF que más inciden en el fraude una vez quer conocemos los individuos que han defreaudado o no. En el capítulo anterior se calculó la probabilidad de fraude de los contribuyentes y en este capítulo se mejorará el cálculo de esa probabilidad de fraude haciendo incidir en ella las partidas económicas del IRPF.

A partir de una muestra de gran tamaño de IRPF se construirán modelos de Redes Neuronales Perceptrón Multicapa (MLP) y Función de Base Radial (RBF) que cuantificarán la incidencia de las distintas partidas económicas del impuesto en el fraude fiscal en IRPF. Asimismo se utilizarán estos modelos predictivos para mejorar la cuantificación de la probabilidad que tiene cualquier contribuyente de ser defraudador.

De esta forma se segmentarán los declarantes del impuesto por nivel de propensión al fraude y causas del mismo.

Dado el tamaño del conjunto de datos, el elevado número de variables y la complejidad de los algoritmos implícitos en las técnicas que se utilizan, es necesario trabajar con herramientas de Big Data que implementen paralelismo en la computación, distribución, tolerancia a fallos y otras propiedades que permitan la computación con grandes datos. A estos efectos, utilizaremos software de IBM y SAS.

.

4.2 MARCO METODOLÓGICO: LAS REDES NEURONALES

Podemos definir una *red neuronal* como un conjunto de elementos de procesamiento de la información altamente interconectados, que son capaces de aprender con la información que se les alimenta. La principal

característica de esta nueva tecnología de redes neuronales es que puede aplicarse a gran número de problemas que pueden ir desde problemas complejos reales a modelos teóricos sofisticados como por ejemplo reconocimiento de imágenes, reconocimiento de voz, análisis y filtrado de señales, clasificación, discriminación, análisis financiero, predicción dinámica, etc.

Las Redes Neuronales tratan de emular el sistema nervioso, de forma que son capaces de reproducir algunas de las principales tareas que desarrolla el cerebro humano, al reflejar las características fundamentales de comportamiento del mismo. Lo que realmente intentan modelizar las redes neuronales es una de las estructuras fisiológicas de soporte del cerebro, la neurona y los grupos estructurados e interconectados de varias de ellas, conocidos como redes de neuronas. De este modo, construyen sistemas que presentan un cierto grado de inteligencia. No obstante, debemos insistir en el hecho de que los Sistemas Neuronales Artificiales, como cualquier otra herramienta construida por el hombre, tienen limitaciones y sólo poseen un parecido superficial con sus contrapartidas biológicas. Las redes neuronales, en relación con el procesamiento de información, heredan tres características básicas de las redes de neuronas biológicas: paralelismo masivo, respuesta no lineal de las neuronas frente a las entradas recibidas y procesamiento de información a través de múltiples capas de neuronas.

Una de las principales propiedades de estos modelos es su capacidad de aprender y generalizar a partir de ejemplos reales. Es decir, la red aprende a reconocer la relación (que no deja de ser equivalente a estimar una dependencia funcional) que existe entre el conjunto de entradas proporcionadas como ejemplos y sus correspondientes salidas, de modo que, finalizado el aprendizaje, cuando a la red se le presenta una nueva entrada (aunque esté incompleta o posea algún error), en base a la relación funcional establecida en el mismo, es capaz de generalizarla ofreciendo una salida. En

consecuencia, podemos definir una red neuronal artificial como un sistema inteligente capaz, no sólo de aprender, sino también de generalizar.

Una red neuronal está formada por unidades de procesamiento que reciben el nombre de neuronas o nodos. Estos nodos están organizados en grupos que se llaman "capas". Generalmente existen tres tipos de capas: una capa de entrada, una o varias capas ocultas y una capa de salida. Las conexiones se establecen entre los nodos de cada capa adyacente. La capa de entrada, mediante la cual se presentan los datos a la red, está formada por nodos de entrada que reciben la información directamente del exterior. La capa de salida representa la respuesta de la red a una entrada dada, siendo esta información transferida al exterior. Las capas ocultas o intermedias se encargan de procesar la información y se interponen entre las capas de entrada y salida y son las únicas que no tienen conexión con el exterior.

La estructura de red más habitual es la denominada red alimentada hacia delante o *feedforward*, ya que las conexiones entre neuronas se establecen en un único sentido, por el siguiente orden: capa de entrada, capa(s) oculta(s) y capa de salida. Por ejemplo, en la figura 4-1 se muestra una red con dos capas ocultas alimentada hacia adelante.

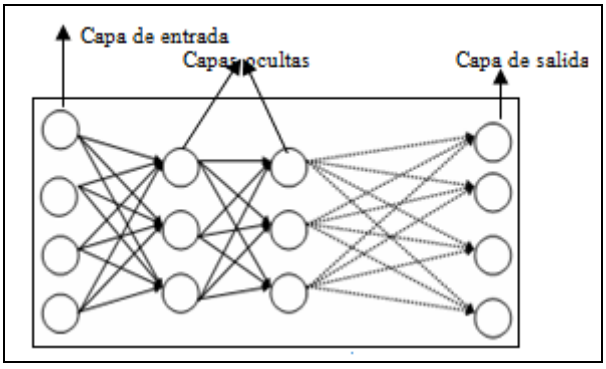


Figura 4-1

No obstante, existen también redes retroalimentadas o *feedback*, que pueden tener conexiones hacia atrás, es decir, de nodos de una capa a elementos de proceso de capas anteriores, así como redes recurrentes, que pueden poseer conexiones, tanto entre neuronas de una misma capa, como de un nodo a sí mismo. La figura 4-2 ilustra un modelo de red en que coexisten los distintos tipos de conexiones que hemos comentado, es decir, hacia delante, hacia atrás y recurrentes, mostrando una interconexión total.

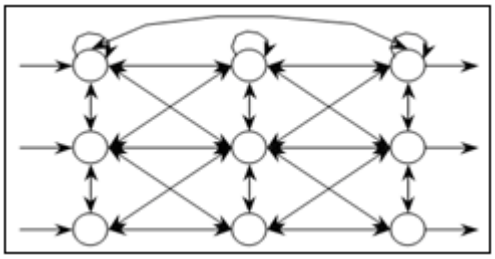


Figura 4-2

La red neuronal totalmente interconectada es aquella en la que los nodos de cada capa están conectados con todos los nodos de la capa siguiente. La capa de entrada tiene por única misión la de distribuir la información que se le presenta a la red neuronal para el procesamiento en la capa siguiente. Los nodos de las capas ocultas y de la capa de salida procesan las señales aplicando factores de procesamiento, llamados *pesos sinápticos*. Cada capa tiene un nodo adicional llamado sesgo (*bias*), que añaden un término adicional a la salida de todos los nodos de la capa. Todas las entradas en un nodo son ponderadas, combinadas y procesadas a través de una función, llamada *función de transferencia o función de activación* que controla el flujo de salida de ese nodo para conectar con todos los nodos de la capa siguiente. Esta función de transferencia sirve para normalizar la salida.

Una red neuronal artificial no es más que la conexión de varias neuronas. Así, las neuronas artificiales, denominadas también unidades,

nodos o elementos de proceso, constituyen la unidad básica de una red neuronal (análoga a la neurona biológica). Dichas neuronas artificiales operan a modo de microprocesadores simples, cuya función consiste en dar respuesta a un determinado patrón de entrada. Cada elemento de proceso, al igual que ocurre en una neurona biológica, recibe entradas procedentes de otros nodos vecinos, o del exterior, en el caso de la capa de entrada, y su función consiste en transformar, mediante sencillos cálculos internos, dichas entradas en un sólo valor de salida que envía al resto de nodos (constituyendo la entrada de éstos) o bien, al exterior, si la neurona en cuestión pertenece a la capa de salida. Las conexiones entre elementos de proceso llevan asociadas un peso o fuerza de conexión *W* que determina cuantitativamente el efecto que producen unos elementos sobre otros. Es decir, en los pesos se almacena la información de la red, al igual que sucede en las redes de neuronas biológicas.

El que una entrada tenga un efecto excitatorio o inhibitorio, depende de que el signo del peso correspondiente sea, respectivamente, positivo o negativo. La efectividad de las entradas está determinada por la fuerza de la conexión, representada por el valor absoluto de los pesos. Así, cada uno de los elementos W_{ij} de la matriz de pesos *W*, conocida como patrón de conexiones, representa la intensidad y sentido de la relación del elemento de proceso *j*, con respecto al elemento de proceso *i*.

El proceso de transformación de las entradas en salidas, en una red neuronal artificial alimentada hacia delante, con *r* entradas, una única capa oculta, compuesta de *q* elementos de proceso, y una unidad de salida puede resumirse en la siguiente formulación de la *función de salida de la red*:

$$\hat{f}(x,W) = F(\beta_0 + \sum_{j=1}^{q} \beta_j G(x'\gamma_j))$$

donde, $\hat{f}(x,W)$ es la salida de la red, el vector $x = (1, x_1, x_2, ..., x_r)'$ representa las entradas de la red (el 1 se corresponde con el sesgo de un

modelo tradicional), $\gamma_j = (\gamma_{j0}, \gamma_{j1}, ..., \gamma_{ji}, ..., \gamma_{jr})' \in \Re^{r+1}$ son los pesos de las neuronas de la capa de entrada a las de la intermedia u oculta, β_j, $j = 0, ...,$ q, representa la fuerza de conexión de las unidades ocultas a las de salida (j=0 indexa la unidad sesgo), q es el número de unidades intermedias, es decir, el número de nodos de la capa oculta, F: $\Re \rightarrow \Re$ es la función de activación de la unidad de salida y G: $\Re \rightarrow \Re$ se corresponde con la función de activación de las neuronas intermedias. W es un vector que incluye todos los pesos de la red (pesos sinápticos), es decir, γ_j y β_j. La figura 4-3 representa la función $\hat{f}(x, W)$.

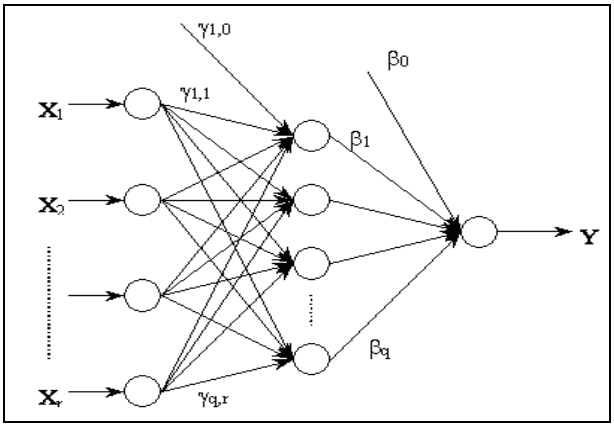


Figura 4-3

Históricamente, se emplearon como funciones de activación *funciones de umbral*, cuyo efecto es que las unidades se activan bruscamente, esto es, o no se activan, o se activan de golpe. La respuesta sólo puede ser blanco o negro, por ello, estas funciones son adecuadas para tareas de clasificación y reconocimiento. Con el tiempo, se introdujeron funciones de activación que permiten que las neuronas se activen gradualmente a medida que el nivel de actividad de sus entradas aumenta, en lugar de que su estado pueda ser, únicamente, activación-desactivación. En concreto, suele utilizarse la función sigmoidal o logística $G(a)$ =

1/(1+exp(-a)), que produce una respuesta sigmoidal alisada. También suele utilizarse la función tangente hiperbólica

Si en la expresión de $\hat{f}(x,W)$ consideramos que $a = x'\gamma_j$, nos encontramos con que $G(x'\gamma_j)$ se corresponde con el *modelo logit binario*.

En general, las funciones *F* y *G* pueden adoptar cualquier forma en la expresión de $\hat{f}(x,W)$. Ahora bien, es práctica habitual considerar, bien que la función de activación de las neuronas de salida y de las intermedias es idéntica, *F(a) = G(a)*, y que se corresponde con la función sigmoidal o arco tangente, o bien, que *F(a) = a*, es decir, que es la función identidad y que *G(a)* se corresponde con la función sigmoidal o arco tangente, lo que es equivalente a considerar que sólo existe función de activación (la sigmoidal o arcotangente) en las unidades ocultas. Esta última hipótesis es la que suponemos, a partir de ahora, en nuestro planteamiento, porque además de simplificar enormemente la notación, es la que con mayor frecuencia se adopta en la construcción de redes neuronales artificiales. También debemos apuntar que emplearemos la función arcotangente, habitualmente utilizada, ya que sus propiedades permiten emplear algoritmos de aprendizaje como el de retropropagación de errores, que utilizamos en nuestro trabajo.

Suponiendo, como hemos indicado, que sólo existe función de activación en las neuronas intermedias y que ésta se corresponde con la sigmoidal, tenemos:

$$\hat{f}(x,W) = \beta_0 + \sum_{j=1}^{q} \beta_j G(x'\gamma_j)$$

Otra posibilidad, de gran utilidad en aplicaciones econométricas, es considerar que en la red que representamos, una red neuronal artificial alimentada hacia delante, con *r* entradas, una única capa oculta, compuesta

de q elementos de proceso, y una unidad de salida, también existen conexiones directas entre la capa de entrada y la de salida. En este caso, la salida de la red se obtiene mediante la siguiente expresión:

$$\hat{f}(x,W) = x'\alpha + \beta_0 + \sum_{j=1}^{q} \beta_j G(x'\gamma_j)$$

donde α es un vector de dimensión $r \times 1$ que representa los pesos sinápticos de las conexiones directas entre las capas de entrada y salida. Como es lógico, ahora *W*, que recoge la totalidad de pesos de la red o *pesos sinápticos*, se compone de α, γ_j y β_j.

Después de diseñar una red neuronal artificial, lo que pretendemos conseguir con la misma es que, para ciertas entradas, o patrones ejemplo que suministramos a la red, ésta sea capaz de generar una salida deseada. Para ello, además de que la topología de la red (entendida como la estructura de la red) sea adecuada, se requiere que la misma aprenda a proporcionar soluciones correctas, es decir, es necesario someter a la red a un proceso de aprendizaje o entrenamiento. El aprendizaje puede entenderse como un procedimiento de prueba y error que permite la estimación estadística de los parámetros del modelo de red neuronal empleado.

Las redes neuronales con aprendizaje supervisado, que suelen venir asociadas al perceptrón multicapa (Multilayer Perceptron MLP) y la función de base radial RBF, presentan un patrón de salida o variable dependiente que les permite contrastar y corregir los datos. Las redes neuronales con patrón de salida suelen ser una técnica utilizada para la clasificación como para la predicción con ello se pueden segmentar mercados, posicionar productos, realizar previsiones de demanda, evaluaciones de expedientes de crédito o de análisis del valor de acciones en bolsa y un sinfín de aplicaciones más. El modelo de red neuronal perceptrón multicapa se fundamenta en el aprendizaje por retropropagación del error (Back-

Propagation) y utiliza habitualmente el algoritmo por retropropagación, el algoritmo del gradiente descendente (conjugate gradient descent) y el algoritmo de Levenberg-Marquardt.

Así como en el percetrón multicapa todas las capas tienen la misma estructura (lineal), en la función de base radial (FBR), la capa intermedia tiene precisamente una estructura radial. La función de base radial es una función supervisada con patrón de salida que realiza clasificaciones (y o previsiones) a partir de elipses e hiperelipses que parten el espacio de entrada de datos. La función de base radial presenta ciertas ventajas con respecto a las redes neuronales multicapas entre las que destacan que se puede modelizar usando nada más que una capa intermedia en vez de varias y que su algoritmo es más rápido y no se queda nunca en una solución local.

En nuestro caso utilizaremos redes neuronales artificiales alimentadas hacia adelante, con *r* entradas (variables o partidas más importantes del Impuesto sobre la Renta de las Personas Físicas con posible incidencia sobre el fraude fiscal), una única capa oculta, compuesta de *q* elementos de proceso, y una unidad de salida (variable fraude que indica si el individuo ha defraudado o no). A través de esta red se estimarán las propensiones al fraude de los individuos y se analizará la incidencia de las partidas de IRPF sobre el fraude fiscal.

4.3 FASE DE SELECCIÓN DE LA INFORMACIÓN: LOS DATOS

Siguiendo la pauta de los capítulos anteriores, en la aplicación que aquí se presenta, se utiliza como fuente de datos la muestra del Impuesto sobre la Renta de las Personas Físicas (IRPF) que proporciona el Instituto de Estudios Fiscales (IEF). Las muestras de IRPF del IEF son bases de datos de registros administrativos de gran amplitud (casi dos millones de

observaciones por año) y detalle (cerca de 300 variables personales, familiares y fiscales).

Habitualmente se entiende por registro administrativo la información referida a personas físicas, empresas, hogares y otras unidades individuales, cuyo diseño, recogida y mantenimiento corresponden a la administración pública. Los registros administrativos, las encuestas y los censos, constituyen las tres fuentes de datos estadísticos clásicas.

Dado su cuantioso volumen y la gran diversidad de variables que contienen, los registros administrativos constituyen hoy en día un componente esencial en la infraestructura estadística y son muy apropiados para la investigación social y económica.

El uso de registros administrativos presenta grandes ventajas, entre las que destacan su gran volumen y cobertura, que permite tipos de análisis imposibles de realizar con otras fuentes como las encuestas, pues se incurriría en costos elevados. Por otro lado, los registros administrativos permiten altos niveles de desagregación de las variables enriqueciendo los análisis. Su amplia cobertura también es esencial, ya que posibilitan investigaciones a nivel regional, municipal y otras áreas pequeñas. Dado que el registro administrativo suele tener como dato la fecha, permite utilizar estructuras de datos de series temporales, de sección cruzada, de pool de datos y de panel.

Los registros administrativos tienen también como ventaja importante ahorrar los elevados costes del trabajo de campo de las encuestas y sobre todo, descargar al informante de la pesadez de rellenar largos cuestionarios para aportar la información. En cuanto a este asunto, el Consejo Superior de Estadística dicta criterios para la incorporación de

Operaciones Estadísticas al Plan Estadístico Nacional entre los que prioriza que la información se obtenga de registros administrativos.

Entre las desventajas más típicas del uso de registros administrativos se encuentra la dificultad de acceso a los mismos debido, entre otras causas, a la reticencia de determinadas administraciones a hacer públicos estos datos, problema que se acentúa cuando los datos son de carácter tributario. Es muy necesario mantener la confidencialidad de los datos, pero esta razón es utilizada habitualmente en exceso por las administraciones para evitar la disponibilidad de datos. Los registros administrativos pueden y deben anonimizarse para cumplir las condiciones de confidencialidad que marcan la Ley General Tributaria, la Ley de la Estadística Pública y la Ley de Protección de Datos. Cumplidas estas condiciones, y teniendo en cuenta otras cuestiones como por ejemplo no cargar excesivamente de trabajo a las unidades de la Administración responsables de la difusión, los registros administrativos debieran de ser de acceso público.

La puesta en práctica del Plan de Transparencia en el Ministerio de Hacienda involucró especialmente a la Agencia Estatal de Administración Tributaria (AEAT), al Instituto de estudios Fiscales (IEF), a la Intervención General de la Administración del estado IGAE y a otras unidades ministeriales que disponen de datos tributarios.

En lo referente a la Agencia Tributaria y al Instituto de Estudios Fiscales se estableció un convenio de colaboración para la difusión de información con fines estadísticos que dio como resultado la puesta a disposición de los investigadores de muestras anuales de microdatos de declarantes del impuesto sobre la Renta de las Personas Fisicas (IRF) y de paneles de datos con información longitudinal del mismo impuesto que distribuye periódicamente el IEF y que son accesibles públicamente en la

página web del IEF a través de un protocolo de petición de datos. El IEF difunde muestras anuales de IRPF desde el año 2002.

Este es el origen de los datos que alimenta este trabajo. El origen fiscal de la muestra proporciona, por tanto, unos datos de gran precisión, y en los que además no aparecen los problemas de infrarrepresentación y falta de respuesta habituales de las encuestas. Por consiguiente, la riqueza de estos datos permite realizar múltiples análisis que están vedados a otras muestras de origen no fiscal.

La información de muestras y paneles se estratifica por criterios geográficos, por tramos de renta y por fuentes de renta, lo que origina una expansión de la información por todo el territorio nacional que aporta una gran representatividad de la misma y posibilita análisis de la información e investigaciones a nivel regional, incluidas áreas pequeñas. El diseño de muestreo con criterios matemáticos precisos, tanto de las muestras como del panel de IRPF permite calcular errores de las estimaciones realizadas que nos permiten valorar la validez y la precisión de las mismas.

4.4 FASE DE EXPLORACIÓN DE LA INFORMACIÓN

En nuestro modelo de redes neuronales la variable dependiente es el fraude global (variable que toma valores cero y uno según que el individuo no defraude o defraude respectivamente).

En la muestra de IRPF se ofrecen las partidas económicas del impuesto sobre la renta de las personas físicas que son candidatas inicialmente a ser las variables independientes del modelo de red neuronal. La tabla 4-1 muestra las primeras partidas económicas de la muestra de IRPF del IEF.

par1	Numérico	8	2	Rdto. del trabajo Dinerarios
par2	Numérico	8	2	Retribuciones en especie (valoración)
par3	Numérico	8	2	Retribuciones en especie (ingresos a cuenta)
par4	Numérico	8	2	Retribuciones en especie (ingresos a cuenta repercutidos)
par5	Numérico	8	2	Rdto. del Trabajo En especie.
par6	Numérico	8	2	Contribuciones Planes Pensiones.
par7	Numérico	8	2	Aportaciones recibidas al patrimonio protegido de las personas con discapacidad del que es titular el contribuyente
par8	Numérico	8	2	Reducciones Art. 18 apartados 2 y 3, y dispos. trans. 11ª y 12ª Ley del Impuesto
par9	Numérico	8	2	Total ingresos integros computables [(01)+(05)+(06)+(07)-(08)]
par10	Numérico	8	2	Cotizac. Seguridad Social, Mutualidad Funcionarios, detracciones derechos pasivos y Coleg.Huérfanos.
par11	Numérico	8	2	Cuotas satisfechas a sindicatos
par12	Numérico	8	2	Cuotas satisfechas a colegios profesionales (si la colegiación es obligatoria y con un máximo de 500 euros anuales)
par13	Numérico	8	2	Gastos de defensa jurídica derivados directamente de litigios con el empleador (máximo: 300 euros anuales)
par14	Numérico	8	2	Gastos deducibles.
par15	Numérico	8	2	Rendimiento neto.Trabajo
par16	Numérico	8	2	Reducción de rendimientos acogidos al régimen especial "33.ª Copa del América" (disposición adicional séptima de la Ley
par17	Numérico	8	2	Reducción por obtención rdto. trabajo.Cuantía aplicable con carácter general.
par18	Numérico	8	2	Reducción por obtención rdto. trabajo.Incremento para trabajadores activos mayores de 65 años que continuen o prolonguen
par19	Numérico	8	2	Reducción por obtención rdto. trabajo.Incremento para contrib. desempleados que acepten un puesto que exija traslado de
par20	Numérico	8	2	Reducción por obtención rdto. trabajo.Reducción adicional para trabajadores activos que sean presonas con discapacidad.
par21	Numérico	8	2	Rendimiento neto reducido.Trabajo
par22	Numérico	8	2	Rend. Cap. Mobiliario. Intereses de cuentas, depósitos y activos financieros
par23	Numérico	8	2	Rend. Cap. Mobiliario. Intereses de activos financieros con bonificación

Tabla 4-1

Pero con el objeto de simplificar el trabajo, se utilizarán como variables independientes del modelo de red neuronal las partidas económicas más importantes del impuesto, que habitualmente son las más incidentes en el fraude fiscal. Estas partidas incluyen prácticamente a todas las demás mediante sumas de las mismas. Las 4-2 a 4-8 presentan las citadas partidas agrupadas por los diferentes conceptos del impuesto.

Concepto	Casilla
Rendimientos	
Ingresos íntegros del trabajo (dinerarios)	par1
Rendimiento neto del trabajo	par15
Ingresos íntegros del capital mobiliario	par29+par45
Rendimiento neto del capital mobiliario	par31+par47
Rendimientos netos reducidos del capital mobiliario	par35+par50
Ingresos íntegros del capital inmobiliario	par70
Rendimientos netos del capital inmobiliario	par75
Rendimiento neto reducido del capital inmobiliario	par79=par85
Rendimientos neto reducido total de actividades económicas en régimen de estimación directa	par140
Rendimientos netos de actividades económicas en estimación objetiva (excepto agrícolas, ganaderas y forest.)	par170
Rendimientos netos de actividades agrícolas, ganaderas y forestales en estimación objetiva	par197
Saldo neto positivo de ganancias y pérdidas patrimoniales	par450+par457

Tabla 4-2

Concepto	Casilla
Mínimos y bases	
Base imponible general	par455
Base imponible del ahorro	par465
Mínimo personal y familiar, aplicado parte general	par680
Mínimo personal y familiar, aplicado parte del ahorro	par681
Base liquidable general sometida a gravamen	par620
Base liquidable del ahorro	par630
Cuotas	
Cuota íntegra estatal	par698
Cuota íntegra autonómica	par699
Cuota líquida estatal	par720
Cuota líquida autonómica	par721
Cuota resultante de la autoliquidación	par741
Cuota diferencial	par755
Resultado de la declaración	par760

Tabla 4-3

Concepto	Casillas
Rendimiento neto reducido del trabajo	par21
Rendimiento neto reducido del capital mobiliario	par35+par50
Rendimiento neto reducido del capital inmobiliario	par85
Rendimiento neto reducido total ED	par140
Rdto. neto reducido total est. objetiva excepto agr., gan. y for.	par170
Rdto. neto reducido total act. agr., gan. y for. est. objetiva	par197
Saldo neto positivo de ganancias y pérdidas patrimoniales imputables a 2009 a integrar en B. I. general	par450
Saldo neto positivo de ganancias y pérdidas patrimoniales imputables a 2009 integrar en B. I. del ahorro	par457

Tabla 4-4

Concepto	Casillas
Renta del periodo	par455+par456+par20+par19+par18+par17
Base imponible	par455+par456
Base liquidable	par618+par630
Cuota íntegra	par698+par699
Cuota líquida	par720+par721
Cuota resultante de la autoliquidación	par741
Cuota real (*)	par741−par756−par758
Resultado de la declaración	par760

Tabla 4-5

También se considerarán las reducciones de la base imponible (tabla 4-6)

par470	Reducción de la Base Imponible por tributación conjunta
par500	Reducción de la Base Imponible por aportaciones y contribuciones a sistemas de previsión social, régimen general
par505	Reducción de la Base Imponible por aportaciones y contribuciones a sistemas de previsión social del cónyuge
par530	Reducción de la Base Imponible por aportaciones y contribuciones a sistemas de previsión social de personas discapacitadas
par560	Reducción de la Base Imponible por aportaciones a los patrimonios protegidos de las personas con discapacidad
par585	Reducción de la Base Imponible por pensiones compensatorias al cónyuge y anualidades por alimentos
par600	Reducción de la Base Imponible por aportaciones a Mutualidades de Previsión Social de deportistas profesionales
par610	Reducc. B. I. General por tributación conjunta
par611	Reducc. B. I. General por aportaciones y contribuciones a sistemas de previsión social (régimen general)
par612	Reducc. B. I. General por aportaciones a sistemas de previsión social de los que es partícipe, mutualista o titular el cónyuge
par613	Reducc. B. I. General por aportaciones y contribuciones a sistemas de previsión social constituidos a favor de personas con discapacidad
par614	Reducc. B. I. General por aportaciones a patrimonios protegidos de personas con discapacidad
par615	Reducc. B. I. General por pensiones compensatorias y anualidades por alimentos
par616	Reducc. B. I. General Cuotas de afiliación y demás aportaciones a los partidos políticos realizadas por afiliados, adheridos y simpatizantes
par617	Reducc. B. I. General por aportaciones a la mutualidad de previsión social de deportistas profesionales
par618	Base liquidable general
par619	Compensación de bases liquidables generales negativas de 2005 a 2008
par620	Base liquidable general sometida a gravamen
par621	Reducc. Base imponible del ahorro por tributación conjunta

Tabla 4-6

Consideraremos como total de reducciones de la base imponible la siguiente variable:

```
REDUCCIONESBASEIMPONIBLE=par470+par500+par505+par530+par560+par585+par600.
```

Otro grupo de variables importante del impuesto lo constituyen las deducciones por vivienda, donativos, autonómicas, incentivos y estímulos a la inversión y otras deducciones (tablas 4-7 y 4-8).

par700	Deducc. por adquisición o rehabilitación de la vivienda habitual, parte estatal
par701	Deducc. por adquisición o rehabilitación de vivienda habitual, parte autonómica
par702	Deducc. por inversiones o gastos en bienes de interés cultural parte estatal
par703	Deducc. por inversiones o gastos en bienes de interés cultural parte autonómica
par704	Deducc. por cantidades o bienes donados a determinadas entidades parte estatal
par705	Deducc. por cantidades o bienes donados a determinadas entidades parte autonómica
par706	Deducc. por incentivos y estímulos a la inversión empresarial, parte estatal
par707	Deducc. por incentivos y estímulos a la inversión empresarial, parte autonómica
par708	Deducc. por dotaciones a la Reserva para Inversiones en Canarias, parte estatal
par709	Deducc. por dotaciones a la Reserva para Inversiones en Canarias, parte autonómica
par710	Deducc. por rendimientos derivados de la venta bienes corporales producidos en Canarias, parte estatal
par711	Deducc. por rendimientos derivados de la venta bienes corporales producidos en Canarias, parte autonómica
par712	Deducc. por rentas obtenidas en Ceuta y Melilla parte estatal
par713	Deducc. por rentas obtenidas en Ceuta y Melilla parte autonómica

Tabla 4-7

par714	Deducc. por cantidades depositadas en cuentas ahorro-empresa parte estatal
par715	Deducc. por cantidades depositadas en cuentas ahorro-empresa parte autonómica
par716	Deducc. por alquiler de la vivienda habitual
par717	Suma de deducciones autonómicas
par720	Cuota líquida estatal
par721	Cuota líquida autonómica
par722	Importe de las deducciones de 1996 y ejercicios anteriores a las que se ha perdido el derecho
par723	Intereses demora de deducciones de 1996 y ejercicios anteriores a las que se ha perdido el derecho
par724	Importe de las deducciones generales de 1997 a 2008 a las que se ha perdido el derecho. Parte estatal
par725	Intereses demora de deducciones generales de 1997 a 2008 a las que se ha perdido el derecho. Parte estatal
par726	Importe de las deducciones generales de 1997 a 2008 a las que se ha perdido el derecho. Parte autonómica
par727	Intereses de demora de deducciones generales de 1997 a 2008 a las que se ha perdido el derecho. Parte autonómica
par728	Importe de las deducciones autonómicas de 1998 a 2008 a las que se ha perdido el derecho
par729	Intereses demora de deducciones autonómicas de 1998 a 2008 a las que se ha perdido el derecho
par730	Cuota líquida estatal incrementada
par731	Cuota líquida autonómica incrementada
par732	Cuota líquida incrementada total
par733	Deducción por doble imposición de dividendos pendientes de aplicar de 2005 y 2006. Importe que se aplica
par734	Deducción por doble imposición internacional, por las rentas obtenidas y gravadas en el extranjero
par735	Deducción por obtención de rendimientos del trabajo o de actividades económicas

Tabla 4-8

Las deducciones finalmente aplicables se han agrupado en variables como sigue:

```
DEDUCCIONESVIVIENDA=par700+par701+par716.

DEDUCCIONESDONATIVOS=par704+par705.

OTRASDEDUCCIONES=par702+par703+par712+par713+par714+par715.

DEDUCCIONESAUTONOMICAS=par717.

DEDUCCIONESINCENTIVOSINVERSION=par706+par707.
```

También se tienen en cuenta el total de gastos deducibles por rendimientos del trabajo y por rendimientos del capital mobiliario (par14+par30+par46) recogidos en la variable:

```
GASTOSDEDUCIBLESTOTALES=par14+par30+par46.
```

Esta investigación es independiente del ejercicio de datos que se considere, ya que se busca una metodología para cuantificar la incidencia sobre el fraude global de las distintas partidas del impuesto. Como además la metodología se basa en un modelo predictivo, se obtiene una función de predicción de fraude que permite cuantificar la propensión al fraude de los individuos declarantes de IRPF.

4.5 FASE DE TRANSFORMACIÓN DE LOS DATOS

En nuestro modelo tenemos las 44 variables independientes cuantitativas más importantes del IRPF, que iniciamente están correladas entre sí y que provocarían un problema de multicolinealidad en cualquier modelo a estimar. Por lo tanto será necesario reducir estas variables a sus componentes principales, que están incorreladas. Al ajustar el modelo en las componentes, se elimina el problema de la multicolinealidad, se reduce el efecto de los valores atípicos y se induce normalidad en las variables. Por lo tanto, el modelo ajustado en componentes tendrá propiedades óptimas.

El análisis en componentes principales es una técnica de análisis estadístico multivariante que se clasifica entre los métodos de interdependencia. Se trata de un método multivariante de simplificación o reducción de la dimensión y que se aplica cuando se dispone de un conjunto elevado de variables con datos cuantitativos correlacionadas entre sí persiguiendo obtener un menor número de variables, combinación lineal de las primitivas e incorreladas, que se denominan componentes principales o factores, que resuman lo mejor posible a las variables iniciales con la mínima pérdida de información y cuya posterior interpretación permitirá un análisis más simple del problema estudiado. Esta reducción de muchas variables a pocas componentes puede simplificar la aplicación sobre estas últimas de otras técnicas multivariantes (regresión, clusters, logística, etc.).

En nuestro caso, la reducción es procedente porque el determinante de la matriz de correlaciones de las variables iniciales es prácticamente nulo. Además las comunalidades de las variables son altas y muchas de ellas cercanas a la unidad (tabla 4-9).

Matriz de correlaciones[a,b]

a. Determinante = ,000

b. Esta matriz no es cierta positiva.

Comunalidades

	Inicial	Extracción
Rdto. del trabajo Dinerarios	1,000	,787
Rdto. del Trabajo En especie.	1,000	,426
Contribuciones Planes Pensiones.	1,000	,509
Rendimiento neto. Trabajo	1,000	,987
INGRESOSICMOBILIARIO	1,000	,997
RENDIMIENTONCMOBILIARIO	1,000	,997
RENDIMIENTONRCMOBILIARIO	1,000	,997

Tabla 4-9

En nuestro caso hemos obtenido 11 componentes principales Ci (factores) que explican cerca del 85% de la variabilidad inicial de los datos, resultando así una buena reducción. En concreto explican el 84,882% de la variabilidad, tal y como indica la tabla 4-10.

Varianza total explicada

Componente	Autovalores iniciales			Sumas de extracción de cargas al cuadrado			Sumas de rotación de cargas al cuadrado		
	Total	% de varianza	% acumulado	Total	% de varianza	% acumulado	Total	% de varianza	% acumulado
1	15,753	35,801	35,801	15,753	35,801	35,801	13,356	30,354	30,354
2	5,946	13,513	49,314	5,946	13,513	49,314	5,854	13,304	43,659
3	3,843	8,734	58,048	3,843	8,734	58,048	5,231	11,889	55,547
4	3,248	7,382	65,430	3,248	7,382	65,430	3,877	8,811	64,358
5	1,679	3,816	69,246	1,679	3,816	69,246	1,771	4,026	68,384
6	1,519	3,451	72,697	1,519	3,451	72,697	1,655	3,761	72,145
7	1,282	2,913	75,610	1,282	2,913	75,610	1,351	3,072	75,216
8	1,056	2,401	78,011	1,056	2,401	78,011	1,208	2,746	77,962
9	1,012	2,300	80,311	1,012	2,300	80,311	1,018	2,314	80,276
10	1,010	2,295	82,607	1,010	2,295	82,607	1,017	2,311	82,587
11	1,001	2,276	84,882	1,001	2,276	84,882	1,010	2,295	84,882
12	,994	2,259	87,142						
13	,988	2,244	89,386						
14	,955	2,170	91,556						
15	,925	2,102	93,658						
16	,850	1,932	95,590						
17	,785	1,784	97,374						
18	,560	1,274	98,648						
19	,234	,532	99,179						
20	,124	,282	99,461						
21	,100	,227	99,688						
22	,043	,099	99,786						
23	,033	,076	99,862						
24	,027	,062	99,925						
25	,013	,029	99,953						
26	,011	,024	99,977						
27	,007	,015	99,992						
28	,001	,003	99,995						
29	,001	,001	99,996						
30	,000	,001	99,998						
31	,000	,001	99,999						
32	,000	,001	99,999						
33	,000	,000	99,999						
34	6,092E-5	,000	100,000						
35	5,341E-5	,000	100,000						
36	5,171E-5	,000	100,000						
37	3,857E-5	8,765E-5	100,000						
38	2,955E-5	6,716E-5	100,000						
39	1,556E-5	3,536E-5	100,000						
40	1,545E-6	3,511E-6	100,000						
41	9,026E-15	2,051E-14	100,000						
42	3,795E-17	8,625E-17	100,000						
43	-2,033E-14	-4,620E-14	100,000						
44	-4,889E-14	-1,111E-13	100,000						

Método de extracción: análisis de componentes principales.

Tabla 4-10

Después de una rotación VARIMAX se obtiene la matriz factorial que se presenta a continuación y que permite en primer lugar asociar cada una de las componentes con las variables iniciales de las que es combinación lineal. Esta asociación debe hacerse de modo disjunto, es decir, las variables que intervienen en una componente no pueden estar

incluidas en ninguna otra. Los términos de la matriz factorial (tabla 4-11) representan la correlación de las variables con las componentes. Por lo tanto, asociaremos cada componente principal con aquellas variables iniciales con las que tenga mayor correlación.

Matriz de componente rotado[a]

	Componente										
	1	2	3	4	5	6	7	8	9	10	11
Base liquidable general sometida a gravamen.	,971	-,048	-,080	,056	,004	,114	,160	,039	,002	,018	-,006
Base imponible general	,969	-,048	-,079	,056	,005	,120	,162	,058	,002	,017	-,006
RENTAPERIODO	,969	-,048	-,079	,055	,005	,122	,160	,058	,001	,017	-,006
BASEIMPONIBLE	,969	-,048	-,079	,056	,005	,120	,162	,058	,002	,017	-,006
Rendimiento neto reducido.Trabajo	,959	-,098	-,083	-,032	,008	,095	-,192	,068	,003	-,021	-,005
Rendimiento neto.Trabajo	,958	-,098	-,083	-,033	,008	,097	-,194	,068	,002	-,021	-,005
Cuota líquida estatal	,880	,329	,326	,038	,039	,028	,064	,012	-,018	,011	,001
Cuota íntegra estatal	,879	,328	,326	,038	,048	,036	,069	,013	-,017	,012	-,009
CUOTALIQUIDA	,875	,335	,334	,037	,036	,025	,063	,011	-,018	,011	,000
CUOTAINTEGRA	,874	,335	,333	,037	,049	,033	,068	,012	-,016	,012	-,008
Cuota resultante de la autoliquidación	,869	,340	,338	,038	,029	,022	,063	,010	-,017	,011	,001
CUOTAREAL	,869	,340	,338	,038	,029	,021	,063	,010	-,017	,011	,001
Cuota líquida autonómica	,866	,346	,347	,036	,032	,020	,061	,009	-,017	,010	,000
Cuota íntegra autonómica o complementaria	,864	,347	,346	,036	,050	,027	,065	,010	-,016	,011	-,008
Rdto. del trabajo Dinerarios	,815	-,069	-,059	-,032	,007	,244	-,178	,125	-,073	-,009	-,040
BASELIQUIDABLE	,637	,526	,550	,033	,075	,047	,030	,022	,003	-,017	-,003
Rdto. del Trabajo En especie.	,555	-,080	-,072	-,009	-,005	-,252	-,104	-,133	,099	-,032	,056
SALDOPATRIMONIALES	,095	,971	,042	-,007	,122	-,014	-,106	,012	,007	-,005	,003
Saldo neto positivo de ganancias y pérdidas patrimoniales imputables a 2009 integrar en B.I.del ahorro	,094	,971	,042	-,007	,123	-,014	-,104	,012	,008	-,045	,003
Cuota diferencial (741 – 754)	,168	,944	,023	,055	,055	-,012	,194	-,027	-,003	,067	,004
Resultado de la declaración (755 – 756 + 757 – 758 + 759)	,168	,944	,023	,055	,055	-,012	,194	-,027	-,003	,067	,004
RENDIMIENTONCMOBILIARIO	,126	,004	,990	,014	,005	-,010	,007	-,006	-,004	,008	-,003
INGRESOSICMOBILIARIO	,126	,004	,990	,014	,005	-,009	,007	-,006	-,005	,008	-,003
RENDIMIENTONRCMOBILIARIO	,126	,004	,990	,014	,004	-,010	,007	-,006	-,004	,008	-,003
Base imponible del ahorro (457 – 458 + 460 – 461).	,151	,667	,716	,004	,088	-,017	-,066	,006	,003	-,031	,001
Base liquidable del ahorro.	,152	,667	,716	,004	,088	-,015	-,066	,001	,003	-,031	,001
RENDIMIENTONRCINMOBILIARIO	,041	,022	,013	,990	,003	,014	,000	-,008	,006	,000	,003
Suma de rendimientos netos reducidos del capital inmobiliario.	,041	,022	,013	,990	,003	,014	,000	-,008	,006	,000	,003
Rendimiento neto (070 - 071 - 072 - 074).	,043	,022	,013	,984	,004	,019	,001	-,003	,005	,000	,000
Ingresos íntegros Cap.Inmobiliario	,051	,024	,016	,950	,006	,028	,006	,001	,003	,001	,000
DEDUCCIONESDONATIVOS	,080	,147	,054	,021	,922	,020	,017	,007	-,012	-,002	-,005
DEDUCCIONESAUTONOMICAS	,030	,179	,020	-,006	,921	,018	-,019	-,004	-,005	-,005	-,001
Importe del mínimo personal y familiar que forma parte de la base liquidable general.	,203	-,005	,005	,097	,042	,735	,060	,143	,053	-,002	,016

DEDUCCIONESVIVIENDA	,146	-,010	-,010	-,042	,003	,699	-,029	,082	-,060	-,027	-,081
Importe del mínimo personal y familiar que forma parte de la base liquidable del ahorro.	-,042	,029	,024	-,016	,001	-,459	-,069	,064	-,140	-,034	-,119
Rdto. Neto reducido total act. econ. Est. Directa	,189	,091	-,017	-,004	-,007	,064	,938	-,011	-,056	-,061	-,050
REDUCCIONESBASEIMPONIBLE	,072	-,006	,009	,012	,012	-,071	,104	,828	,052	,012	,044
Contribuciones Planes Pensiones.	,106	-,010	-,017	-,027	-,009	,211	-,120	,656	-,069	-,016	-,046
Rdto. Neto reducido total act. econ. Est. Objetiva	-,008	,011	,015	,022	,001	,146	,023	-,036	,725	,057	,081
DEDUCCIONESINCENTIVOSINVERSION	,010	-,014	,015	,007	,023	,016	,197	-,004	-,399	,111	,155
GASTOSDEDUCIBLESTOTALES	,016	,026	,028	,008	-,009	,356	-,129	-,030	-,358	-,027	-,099
Saldo neto positivo de ganancias y pérdidas patrimoniales imputables a 2009 a integrar en B.I. general	,024	,044	,000	-,001	-,007	,001	-,044	-,004	-,023	,988	-,017
Rdto. Neto Módulos Agrarios	-,009	,007	,011	,012	,001	,043	,051	,022	,160	,005	,799
OTRASDEDUCCIONES	,007	-,008	,021	,012	,011	,031	,120	,030	,353	,032	-,545

Método de extracción: análisis de componentes principales.
Método de rotación: Varimax con normalización Kaiser.
a. La rotación ha convergido en 7 iteraciones.

Tabla 4-11

Analizando la matriz factorial se observa que la primera componente C1 recoge las 17 primeras variables, que incluyen los *rendimientos, bases y cuotas*.La segunda componente C2 incluye cuatro variables relativas a *saldos patrimoniales, cuota diferencial y resultado* de la declaración. La tercera componente C3 incluye 5 variables relativas a *capital mobiliario y base del ahorro*. La cuarta componente C4 contiene 4 variables relativas a *capital inmobiliario*. La componente C5 incluye 2 variables relativas a *deducciones autonómicas y por donativos*. La componente C6 contienes tres variables relativas a *deducciones por vivienda y mínimos personal y familiar*. La componente C7 contiene una única variable relativa a *actividades económicas*. La componente C8 incluye dos variables relativas *deducciones de la base imponible y planes de pensiones*. La componente C9 contiene 3 variables relativas a *gastos deducibles totales y deducción por incentivos a la inversión*. La componente C10 incluye una única variable relativa al saldo neto positivo de las *ganancias y pérdidas patrimoniales*. Por último, la componente C11 contiene dos variables relativas al rendimiento neto de los *módulos agrarios y otras deducciones*. Por lo tanto, la matriz factorial permite expresar cada una de las 11 componentes principales como combinación lineal de las variables iniciales que la componen y de modo disjunto.

Las puntuaciones de las 11 componentes principales obtenidas serán las variables de entrada del modelo de red neuronal (variables independientes). La variable de salida es la variable dicotómica *marca* cuyo valor 1 indica fraude y cuya valor cero indica no fraude.

La tabla 4-12 muestra los primeros registros de las puntuaciones de las componentes principales, que en este capítulo serán las variables independientes de nuestros modelos.

FAC1_1	FAC2_1	FAC3_1	FAC4_1	FAC5_1	FAC6_1	FAC7_1	FAC8_1	FAC9_1	FAC10_1	FAC11_1
-,26837	-,04803	-,01433	,42697	,00397	,87164	,03425	,23296	,16112	-,00401	,10967
-,35675	,03143	,09309	-,15568	,00948	1,33734	,25668	-,16666	5,00323	,38099	,64635
-,24990	-,00388	-,01806	-,14133	-,03114	-,70729	-,01263	-,33916	,01372	,00883	,00165
-,25331	-,00219	-,01659	-,14099	-,03100	-,71295	-,00641	-,33852	,02022	,00980	,00358
-,25235	-,00486	-,01704	-,13698	-,02785	-,43532	-,01349	-,33076	,02111	,00572	,01186
-,24699	-,00506	-,01938	-,13906	-,02810	-,61662	-,00950	-,32303	,03121	,00739	,00463
-,23683	-,04163	-,00363	-,13725	,04990	,33922	,05439	-,29456	,36883	,01930	-,51051
-,17914	-,02702	-,01948	-,13360	-,00433	,06145	,15061	-,26420	,71341	,06810	-1,04154
-,26584	-,06521	,14791	,37306	,01169	-,61340	,07596	-,07150	,40736	-,00772	-,99262
-,06275	-,07081	-,04254	-,12444	,01186	,52974	-,02872	,12613	,06969	-,00484	,06552
,18738	-,09107	,01993	-,09020	,04191	,02254	-,09708	,11134	-,04334	-,00207	,01577
,43477	-,10147	-,09435	-,18194	-,05898	1,10557	-,21905	-,37695	-,12121	-,05091	-,07057
-,12064	-,01210	-,04338	-,18283	-,04201	,97091	-,13398	-,21042	-,09584	-,05372	-,11155
-,23799	-,01254	-,02475	-,13564	-,02442	-,62210	,02014	,00456	,05349	,01551	,03037
-,29164	,02317	,00115	-,13734	-,03286	-,72776	-,33605	-,33396	,03214	,57665	,01196
-,26242	-,00369	-,01318	-,12809	-,01937	,04755	,00947	-,30534	,08284	,00481	,04419
,24298	-,03714	-,06498	-,21046	-,08946	1,30135	-,16955	-,43817	-,19876	-,06893	-,17523
,14605	-,04697	-,05680	-,21821	-,08817	1,48120	-,26337	-,44264	-,25009	-,09271	-,20637
-,06867	-,03738	-,04536	-,13638	-,02776	,04552	-,05724	-,16863	,00641	-,00210	,02096

Tabla 4-12

4.6 FASES DE MODELIZACIÓN Y EVALUACIÓN: ESTIMACIÓN Y DIAGNOSIS DE LOS MODELOS DE REDES NEURONALES

Para la creación y aplicación de una red neuronal a un problema concreto, hemos de distinguir los siguientes pasos:

Conceptualización del modelo para el estudio del problema concreto. En este Modelo debemos señalar las entradas, las salidas y la información de que se dispone.

Adecuación de la información de que se dispone a la estructura de la red a crear. Hay que especificar la parte de la información que va a ser utilizada para el entrenamiento o aprendizaje de la red y la parte de la información que va a ser utilizada como validación de la red.

Fase de aprendizaje. Se le van presentando a la red sucesivos patrones de datos de entrenamiento y validación y la red va proporcionando una salida. Este proceso se repite un cierto número de etapas comparando las salidas con las salidas esperadas. Los diversos algoritmos de aprendizaje intentan minimizar el error que hay entre la salida proporcionada por la red y la salida esperada. Así se consigue un patrón de datos de entrenamiento lo suficientemente eficiente.

Fase de validación. Se presenta a la red entrenada el patrón de datos de validación, y se ve el error cometido por la red en este conjunto. Este error es una medida de la bondad de la red.

Fase de generalización. Si hemos conseguido una red adecuada, se procede a utilizar la red como modelo predictor, aportándole una nueva entrada que la red la procesará para dar una salida.

Suelen considerarse tres tipos básicos de aprendizaje que dan lugar a diferentes tipos de redes neuronales. Cuando el entrenador proporciona a la red la salida deseada, se dice que el *aprendizaje* es *supervisado*. En caso contrario, nos encontramos ante un *aprendizaje no supervisado*. Por último, un tipo intermedio de *aprendizaje* es el *reforzado o híbrido*, en el cuál el entrenador sólo proporciona a la red una indicación de si la respuesta a una entrada dada es buena o mala.

Las redes neuronales con *aprendizaje no supervisado* son aquéllas que entrenan sin necesidad de un *supervisor* o *entrenador* externo que

proporcione a la red la salida deseada, pues son capaces de organizar sus parámetros internamente adaptándose al entorno del mejor modo posible. La red, una vez se le presentan las entradas, es capaz de determinar, por sí sola, las características, correlaciones, regularidades o categorías de las mismas, proporcionando una salida codificada. Por ello, podemos afirmar que estas redes poseen propiedades de autoorganización (redes autoorganizativas). Los sistemas neuronales con aprendizaje no supervisado se caracterizan por poseer arquitecturas simples, puesto que las leyes de aprendizaje ya complican bastante su funcionamiento. En segundo término, la mayor parte de ellas son redes alimentadas hacia adelante, o *feed-forward* con una sola capa intermedia u oculta. Los modelos más característicos que entrenan mediante aprendizaje no supervisado son las *redes de Kohonen* (1977,1984) y *Grossberg* (1976). Las redes no supervisadas suelen utilizarse para la clasificación. Concretamente las redes de *Cohonen* suelen utilizarse cuando uno de los objetivos del análisis sea la visualización sencilla e intuitiva de los conglomerados, cuando se desconoce su forma u cuando existan casos atípicos o errores en los datos. Se utilizan asiduamente en Deep Learning. Las redes neuronales con aprendizaje no supervisado son herramientas fundamentales en el desarrollo de la técnica moderna denominada *Unsupervised Machine Learning*.

En cuanto al aprendizaje supervisado, las redes neuronales utilizadas, que como ya sabemos suelen venir asociadas al *perceptrón multicapa* (*Multilayer Perceptron* MLP) y la *función de base radial* RBF, presentan un patrón de salida o variable dependiente que les permite contrastar y corregir los datos. Las redes neuronales con patrón de salida suelen ser una técnica utilizada para la clasificación como para la predicción y suelen ser ajustadas a través de algoritmos matemáticos como el *algoritmo de Levenberg-Marquardt* o el *algoritmo del gradiente conjugado*. Así como en el perceptrón multicapa todas las capas tienen la misma estructura (lineal), en la función de base radial (FBR), la capa intermedia tiene

precisamente una estructura radial. La función de base radial es una función supervisada con patrón de salida que realiza clasificaciones (y o previsiones) a partir de elipses e hiperelipses que parten el espacio de entrada de datos. La función de base radial presenta ciertas ventajas con respecto a las redes neuronales multicapas entre las que destacan que se puede modelizar usando nada más que una capa intermedia en vez de varias y que su algoritmo es más rápido y no se queda nunca en una solución local. Las redes neuronales con aprendizaje supervisado son herramientas fundamentales en el desarrollo de la técnica moderna denominada *Supervised Machine Learning*.

Debemos destacar que, en ocasiones, aunque sea posible aplicar el aprendizaje supervisado, los métodos de aprendizaje no supervisado pueden resultar de gran utilidad, e incluso ofrecer mejores resultados. Por ejemplo, el *Algoritmo de Retropropagación de Errores (Back-Propagation)* en redes multicapa es muy lento, como consecuencia de que el valor que adopta cada peso depende de los valores que toma en las demás capas. Para evitar este problema podría emplearse, bien un método de aprendizaje no supervisado, o bien un sistema híbrido, que permita a algunas capas autoorganizarse antes de que sus salidas pasen a la red supervisada. Por otra parte, debemos destacar que puede ser aconsejable efectuar algún tipo de entrenamiento no supervisado a redes previamente entrenadas mediante mecanismos de aprendizaje supervisado. La finalidad de este modo de proceder es permitir que la red se adapte paulatinamente a los posibles cambios del entorno.

El *aprendizaje reforzado o híbrido* es intermedio entre el supervisado y el no supervisado. En este tipo de aprendizaje, al igual que en el supervisado, existe un *profesor* o *supervisor externo*. Sin embargo, se diferencian en que el "entrenador" no proporciona a la red las salidas deseadas, pues su comportamiento se evalúa de manera global, esto es, sólo

es posible decidir e indicar a la red si su respuesta es buena o mala y en que grado se comporta bien. El fundamento del aprendizaje reforzado reside en que se deben reforzar aquellas acciones que generan una mejora en el comportamiento y respuesta de la red neuronal.

Análogamente al aprendizaje supervisado, la red neuronal responde generando un conjunto de salidas, correspondientes a los patrones de entrada que se le presentan. Ahora bien, como no se proporcionan salidas deseadas al sistema, es imposible computar la fracción de error que comete cada una de las unidades de salida. Tan sólo se dispone de un indicador del éxito o fracaso de la red, similar a una función de utilidad, que la evalúa de forma global. Esto exige el empleo de algoritmos de aprendizaje mucho más complejos que en el supervisado, así como mayores exigencias en cuanto a tamaño de la muestra.

Formalmente, el proceso del aprendizaje consiste en resolver un problema de mínimos cuadrados no lineales. Para ello, hay que emplear métodos numéricos de optimización como el de *retropropagación de errores* (*Back-propagation*), que se fundamenta en el algoritmo de aproximación estocástica de Robbins y Monro (1951) aplicado a mínimos cuadrados no lineales. Actualmente es el algoritmo más utilizado.

Una vez finalizado el aprendizaje se debe proceder a testear la red. La fase de test consiste en introducir nuevos patrones de entrada y comprobar la eficacia del sistema generado. Si no resulta aceptable se repite la fase de entrenamiento utilizando nuevos patrones-ejemplo, e incluso puede ser necesario modificar la estructura de la red.

El modelo de red neuronal que vamos a utilizar aquí es un modelo de aprendizaje supervisado basado en el perceptrón multicapa, que se fundamenta en el *aprendizaje por retropropagación del error* (*Back-*

Propagation) y utiliza habitualmente el *algoritmo por retropropagación,* el *algoritmo del gradiente descendente* (*conjugate gradient descent*) o el *algoritmo de Levenberg-Marquardt*.

El proceso de aprendizaje o entrenamiento de la red consiste en ir presentando a la red el conjunto de patrones un determinado número de etapas prefijadas de antemano, de forma a minimizar el error de aprendizaje, entendiendo éste como la diferencia cuadrática entre la salida esperada y la salida que aporta la red. En la primera etapa, la red tienen unos pesos de interconexión elegidos de forma aleatoria, a la red se le presenta un vector de entrada en la primera etapa, constituido por el primer patrón, éste se va propagando a través de todas las capas hasta proporcionar una salida, la señal de salida se compara con la salida deseada en todos los nodos de la capa de salida. Este proceso se realiza para todos los patrones del conjunto de aprendizaje, y la suma de los errores cuadráticos de todos los patrones será el error cometido por la red en esa primera etapa.

El objetivo es ir cambiando o actualizando para la segunda etapa los pesos de interconexión de forma a disminuir el error total. La idea del *algoritmo back-propagation* consiste en actualizar los pesos de interconexión de forma que la señal de error se transmita hacia atrás partiendo de la capa de salida; sin embargo estas unidades intermedias sólo reciben una fracción de error proporcional a la contribución relativa que haya aportado a la salida. Este proceso se repite capa por capa hasta que todos los nodos hayan recibido una señal de error que describa su contribución al mismo. Una vez hemos actualizado los pesos, se repite el proceso de presentar de nuevo los patrones de aprendizaje y el cálculo de error, este proceso acaba bien porque el error total es menor que uno prefijado, bien porque hemos concluido con el número de etapas prefijado.

4.6.1 Estimación y diagnosis del modelo de red Perceptrón Multicapa

En primer lugar estimamos un modelo de red neuronal de tipo Perceptrón Multicapa cuyas variables de entrada son las 11 componentes principales obtenidas en el apartado anterior y cuya variable de salida es la variable marca.

Utilizaremos un 70% de los datos para la fase de entrenamiento y un 30% para la fase de pruebas (tabla 4-13). En total tenemos prácticamente 2.000.000 de filas en la base de datos tal y como indica la tabla siguiente (exactamente 1.928.494) de las cuales 1.350.974 se utilizan para entrenar la red y el resto para prueba. No hay datos faltantes en la base de datos.

Perceptrón multicapa

Resumen de procesamiento de casos

		N	Porcentaje
Ejemplo	Entrenamiento	1350974	70,1%
	Pruebas	577520	29,9%
Válido		1928494	100,0%
Excluido		0	
Total		1928494	

Tabla 4-13

En la tabla 4-14 se observa que se han utilizado como variables de entrada de la red las componentes principales, lo cual supone varias ventajas. En primer lugar se elimina el efecto de los valores atípicos ya que las puntuaciones de las componentes tienen un rango muy inferior al de las variables iniciales. Al ser las componentes combinaciones lineales de las variables iniciales, se aminora el efecto de los valores atípicos. En segundo lugar se aumenta la confidencialidad de los datos ya que es muy difícil identificar individuos a partir de los valores de las componentes. En tercer lugar se eliminan problemas de multicolinealidad en el modelo, ya que las componentes están incorreladas. En cuarto lugar se induce normalidad en

las variables del modelo debido al comportamiento asintóticamente normal de las componentes principales. Por último, los modelos predictivos que tienen variables independientes derivadas de una reducción de la dimensión presentan siempre un buen ajuste con una diagnosis muy favorable. También observamos en la tabla que se ha utilizado como función de activación en las capas ocultas la función tangente hiperbólica. La función de activación en la capa de resultado es la función Softmax. Asimismo se observa que hemos utilizado una única capa oculta en la red.

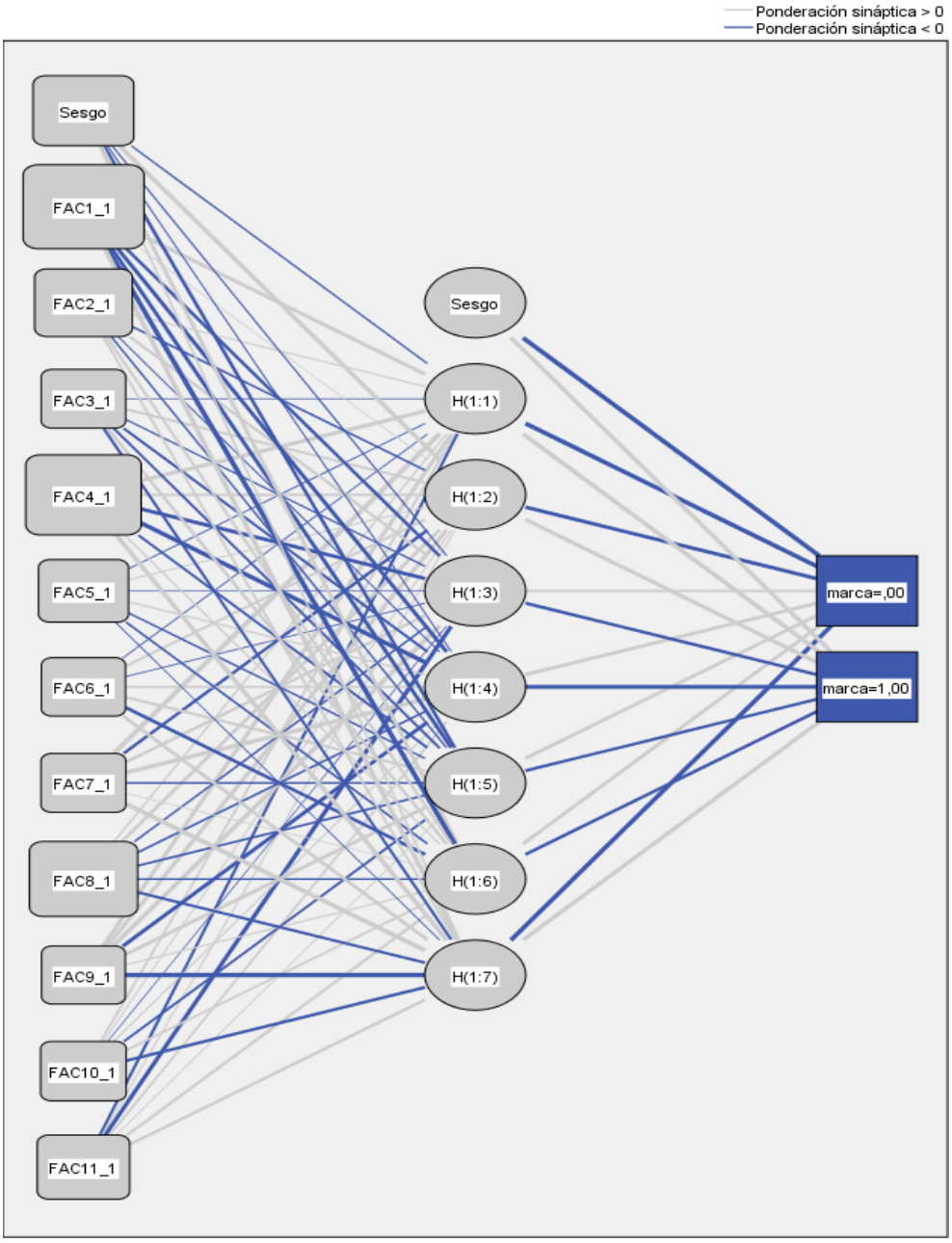
Información de red

Capa de entrada	Covariables	1	RENDIMIENTOS, BASES Y CUOTAS
		2	SALDOS PATRIMONIALES, CUOTA DIFERENCIAL Y RESULTADO
		3	CAPITAL MOBILIARIO Y BASE DEL AHORRO
		4	CAPITAL INMOBILIARIO
		5	DEDUCCIONES AUTONÓMICAS Y DONATIVOS
		6	DEDUCCIONES VIVIENDA Y MÍNIMO PERSONAL Y FAMILIAR
		7	ACTIVIDADES ECONÓMCAS
		8	REDUCCIONES BASE IMPONIBLE Y PLANES DE PENSIONES
		9	GASTOS DEDUCIBLES TOTALES Y DEDUCCION INCENTIVOS INVERSION
		10	GANANCIAS Y PÉRDIDAS PATRIMONIALES
		11	MÓDULOS AGRARIOS Y OTRAS DEDUCCIONES
	Número de unidades[a]		11
	Método de cambio de escala para las covariables		Estandarizados
Capas ocultas	Número de capas ocultas		1
	Número de unidades en la capa oculta 1[a]		7
	Función de activación		Tangente hiperbólica
Capa de salida	Variables dependientes	1	fraude global
	Número de unidades		2
	Función de activación		Softmax
	Función de error		Entropía cruzada

a. Excluyendo la unidad de sesgo

Tabla 4-14

La figura 4-4 muestra la estructura de la red neuronal con sus once nodos correspondientes a las variables de entrada o variables independientes (componentes principales), su única capa oculta cuyos nodos están etiquetados con las etiquetas de los pesos sinápticos y un nodo de salida etiquetado con las dos categorías de la variable dependiente del modelo de red.



Función de activación de capa oculta: Tangente hiperbólica

Función de activación de capa de resultado: Softmax

Figura 4-4

El tamaño de los nodos de entrada indica la magnitud del efecto de las correspondientes variables independientes sobre la variable dependiente. Rectángulos más grandes indican más efecto de la correspondiente variable independiente sobre la respuesta. Por ejemplo, la primera componente, la octava y la cuarta son las que mayor cuantía de efecto tienen sobre el fraude.

A continuación se presenta la tabla 4-15 que indica que el porcentaje de pronósticos incorrectos en la fase de entrenamiento es solamente del 15,8 por ciento. El mismo porcentaje se observa en la fase de pruebas.

Resumen del modelo

Entrenamiento	Error de entropía cruzada	447871,088
	Porcentaje de pronósticos incorrectos	15,8%
	Regla de parada utilizada	Se ha superado el número máximo de épocas (100)
	Tiempo de preparación	0:02:15,04
Pruebas	Error de entropía cruzada	191581,542
	Porcentaje de pronósticos incorrectos	15,8%

Variable dependiente: fraude global

Tabla 4-15

La tabla 4-16 muestra las estimaciones de los pesos sinápticos de la red. Ya sabemos que el proceso de transformación de las entradas en salidas, en una red neuronal artificial alimentada hacia delante, con *r* entradas, una única capa oculta, compuesta de *q* elementos de proceso, y una unidad de salida puede resumirse en la siguiente formulación de la función de salida de la red:

$$\hat{f}(x,W) = F(\beta_0 + \sum_{j=1}^{q} \beta_j G(x'\gamma_j))$$

$x = (1, x_1, x_2, ..., x_r)'$	Entradas de la red
$\gamma_j = (\gamma_{j0}, \gamma_{j1}, ..., \gamma_{ji}, ..., \gamma_{jr})'$	Pesos de las neuronas de la capa de entrada a la oculta
β_j	Fuerza de conexión de las unidades ocultas a las de salida
W	Matriz de pesos sinápticos (γ_j β_j) o patrón de conexiones
$\hat{f}(x,W)$	Salida de la red
$F: \Re \to \Re$	Función de activación de la unidad de salida
$G: \Re \to \Re$	Función de activación de las neuronas intermedias

Las primeras siete columnas de la tabla siguiente estiman los pesos sinápticos de las neuronas de la capa de entrada a la oculta (γ_j) y las dos últimas columnas estiman los pesos sinápticos de las neuronas de la capa oculta a la capa de salida (βj).

El vector *x* recoge las puntuaciones de las componentes principales, que son las variables de entrada de la red.

La matriz *W* recoge todos los pesos sinápticos.

Como función *G* de activación de las neuronas intermedias se ha utilizado una función tangente hiperbólica.

Como función *F* de activación de la unidad de salida se ha utilizado una función softmax.

Estimaciones de parámetro

Predictor		Pronosticado								
		Capa oculta 1							Capa de salida	
		H(1:1)	H(1:2)	H(1:3)	H(1:4)	H(1:5)	H(1:6)	H(1:7)	[marca=,00]	[marca=1,00]
Capa de entrada	(Sesgo)	-,274	1,639	-,167	-,122	-,834	,954	,306		
	FAC1_1	1,173	,029	-,794	-1,110	-,988	-2,318	1,828		
	FAC2_1	,187	-,319	,649	-,104	-,278	,508	,430		
	FAC3_1	-,035	,488	-,210	-,490	-,437	,713	-,514		
	FAC4_1	1,496	,536	-1,207	-1,885	-,466	,200	1,700		
	FAC5_1	-,101	,157	-,052	,284	-,298	-,330	-,074		
	FAC6_1	-,098	,467	-,083	,416	,671	-,892	,898		
	FAC7_1	4,657	-,763	,750	1,847	-,314	,072	2,417		
	FAC8_1	,289	,889	-,300	-,342	-,429	-,245	-,697		
	FAC9_1	1,272	1,280	1,850	-1,144	1,084	,195	-4,070		
	FAC10_1	,401	,148	-,022	,181	-,437	,436	-,833		
	FAC11_1	-,541	,164	-1,281	,343	,087	,682	,767		
Capa oculta 1	(Sesgo)								-2,256	1,514
	H(1:1)								-2,131	2,104
	H(1:2)								-1,212	1,280
	H(1:3)								,633	-,874
	H(1:4)								1,102	-1,594
	H(1:5)								,911	-,773
	H(1:6)								,995	-,860
	H(1:7)								-1,894	1,579

Tabla 4-16

En cuanto a la diagnosis del modelo de red, vemos en primer lugar la matriz de confusión que presenta altos porcentajes de acierto en los valores pronosticados (tabla 4-17).

Clasificación

Ejemplo	Observado	Pronosticado		
		,00	1,00	Porcentaje correcto
Entrenamiento	,00	441525	63016	87,5%
	1,00	150694	695739	82,2%
	Porcentaje global	43,8%	56,2%	84,2%
Pruebas	,00	188867	26963	87,5%
	1,00	64279	297411	82,2%
	Porcentaje global	43,8%	56,2%	84,2%

Variable dependiente: fraude global

Tabla 4-17

Otro elemento de diagnosis es la curva ROC de la red. En la figura 4-4a se observa la curva ROC para el fraude y para el no fraude, presentando ambas un área muy elevada entre las curvas y la diagonal. El

área bajo la curva ROC se estima en 0,918 (tabla 4-18), valor muy cercano a la unidad, lo que indica que la capacidad predictiva de la red es muy alta

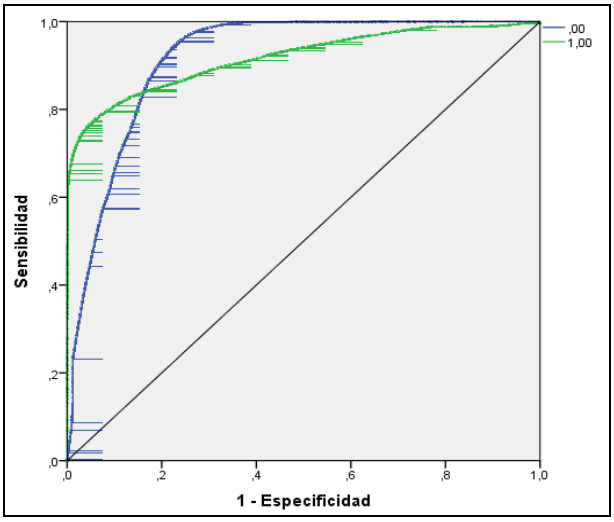


Figura 4-4a

Área bajo la curva

		Área
Tabla		
fraude global	,00	,918
	1,00	,918

Tabla 4-18

La curva de ganancias (figura 4-5) es otro elemento de diagnosis para comparar modelos alternativos. Para porcentajes entre el 40% y el 70% se obtiene la zona de mayor anchura entre las dos curvas del gráfico. Un modelo predice mejor que otro cuando la anchura entre las dos curvas es mayor para los mismos porcentajes. Es decir, a mayor ganancia para el mismo porcentaje, mejor predice el modelo.

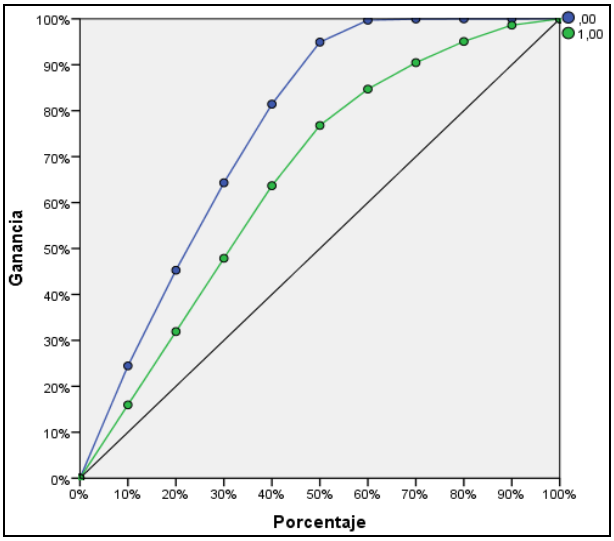


Figura 4-5

El gráfico de elevación de la figura 4-6 es un gráfico alternativo al gráfico de ganancias para comparar la capacidad predictiva de dos modelos. A mayor elevación para el mismo porcentaje, mejor predice el modelo.

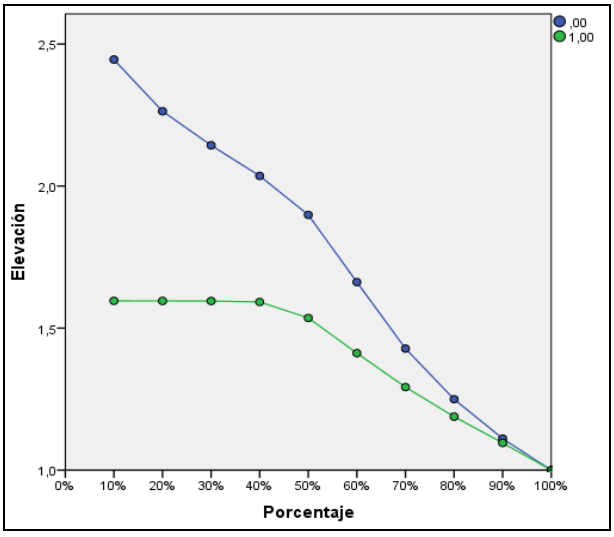


Figura 4-6

En cuanto a la importancia de los predictores (variables independientes) sobre la propensión al fraude (tabla 4-19) vemos que la variable más incidente en el fraude es la componente relativa a rendimientos, bases y cuotas. Le sigue la componente relativa a saldos patrimoniales, cuota diferencial y resultado de la declaración. También tienen mucha influencia sobre el fraude las variables relativas a capital mobiliario e inmobiliario, así como las deducciones autonómicas, las deducciones por donativos y las deducciones por vivienda y mínimos personal y familiar. Por último también afectan significativamente al fraude las actividades económicas, las deducciones de base imponible y planes de pensiones y los gastos deducibles totales. Estos resultados son equivalntes a los obtenidos en capítulos anteriores.

Importancia de las variables independientes

	Importancia	Importancia normalizada
RENDIMIENTOS, BASES Y CUOTAS	,252	100,0%
SALDOS PATRIMONIALES, CUOTA DIFERENCIAL Y RESULTADO	,106	42,2%
CAPITAL MOBILIARIO Y BASE DEL AHORRO	,024	9,4%
CAPITAL INMOBILIARIO	,217	86,2%
DEDUCCIONES AUTONÓMICAS Y DONATIVOS	,059	23,2%
DEDUCCIONES VIVIENDA Y MÍNIMO PERSONAL Y FAMILIAR	,022	8,6%
ACTIVIDADES ECONÓMCAS	,026	10,4%
REDUCCIONES BASE IMPONIBLE Y PLANES DE PENSIONES	,171	68,0%
GASTOS DEDUCIBLES TOTALES Y DEDUCCION INCENTIVOS INVERSION	,019	7,4%
GANANCIAS Y PÉRDIDAS PATRIMONIALES	,031	12,5%
MÓDULOS AGRARIOS Y OTRAS DEDUCCIONES	,073	29,0%

Tabla 4-19

La figura 4-7 muestra la importancia de cada componente sobre la probabilidad de defraudar de los individuos declarantes del impuesto. Esta figura es la representación gráfica de los datos de la tabla anterior.

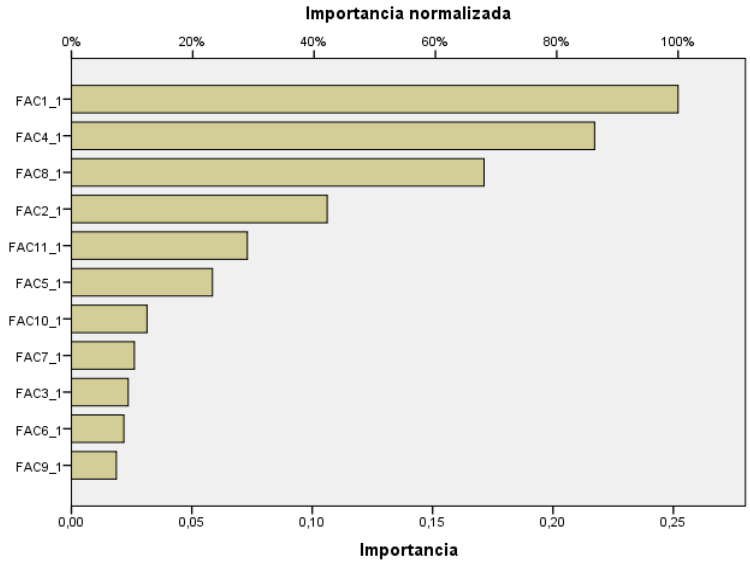
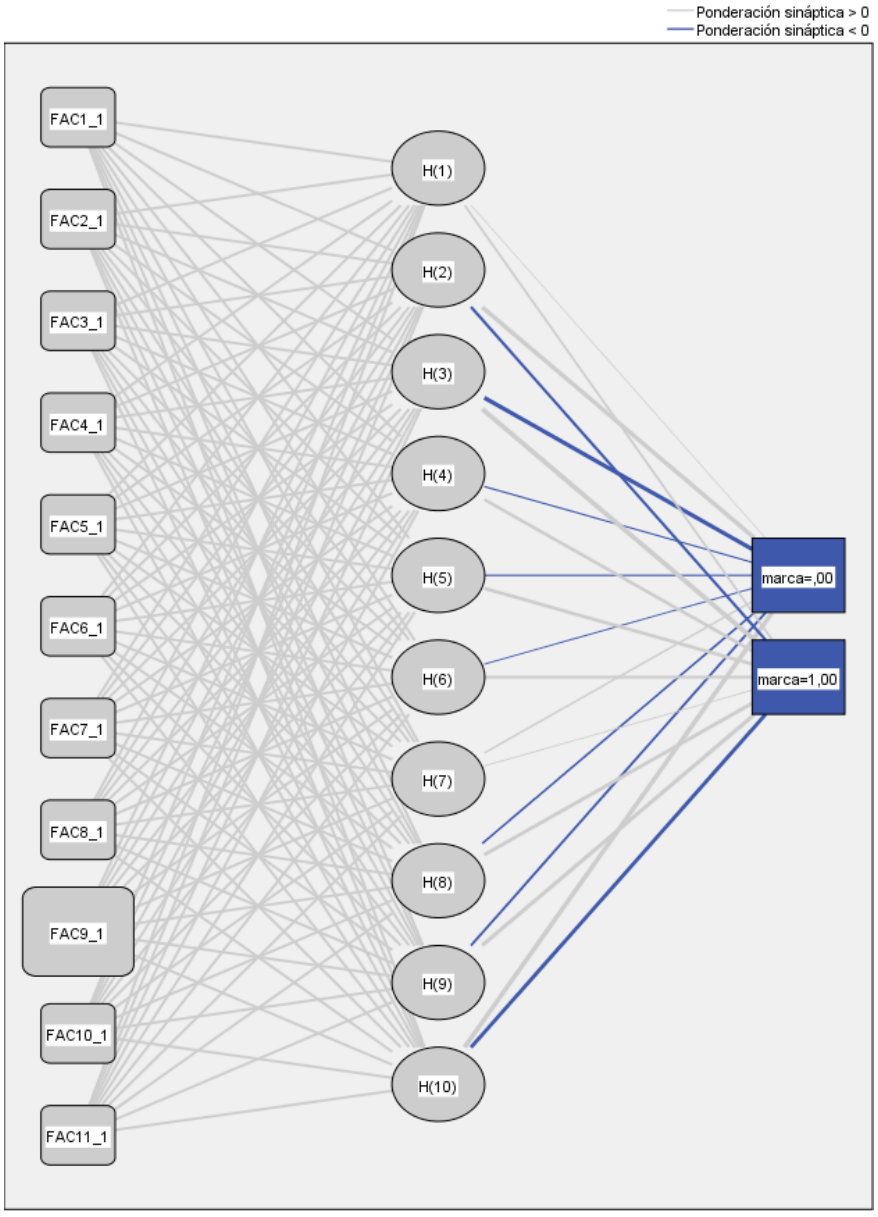


Figura 4-7

4.6.2 Estimación y diagnosis del modelo de red Función de Base Radial

Si alternativamente consideramos la red neuronal Función de Base Radial utilizando también un 70% de los casos para entrenamiento y un 30% para pruebas y una capa oculta, tenemos la estructura que presenta el gráfico de la figura 4-8 para la red estimada. En cuanto a la importancia de los predictores (variables independientes) sobre la propensión al fraude vemos que no hay una distinción clara, ya que el área de los rectángulos relativos a los factores es prácticamente constante salvo en un caso aislado. Por tanto, la red no discrimina bien la importancia de los predictores.



Función de activación de capa oculta: Softmax

Función de activación de capa de resultado: Identidad

Figura 4-8

Observamos que el número de nodos para la capa oculta de la red ahora es 10, es decir, ha aumentado respecto del Perceptrón Multicapa. Por lo tanto, el número de pesos sinápticos a estimar será superior, tal y como se observa en la tabla de Estimaciones de parámetros. Este hecho lleva a que el porcentaje de pronósticos incorrectos sea superior al caso del Perceptrón Multicapa. En la tabla de Resumen del modelo (tabla 4-20) se ve que este porcentaje es ahora el 23,1 por ciento, frente al 15,8% que teníamos en el Perceptrón Multicapa. La tabla 4-21 muestra la estimación de los pesos sinápticos de la red

Resumen del modelo

Entrenamiento	Error de suma de cuadrados	203372,555
	Porcentaje de pronósticos incorrectos	23,1%
	Tiempo de preparación	0:00:37,33
Pruebas	Error de suma de cuadrados	87264,741[a]
	Porcentaje de pronósticos incorrectos	23,2%

Variable dependiente: fraude global

a. El número de unidades ocultas se determina por el criterio de los datos de prueba: El "mejor" número de unidades ocultas es la que produce el menor error en los datos de prueba.

Tabla 4-20

Estimaciones de parámetro

		Pronosticado											
		Capa oculta[a]										Capa de salida	
Predictor		H(1)	H(2)	H(3)	H(4)	H(5)	H(6)	H(7)	H(8)	H(9)	H(10)	[marca=,00]	[marca=1,00]
Capa de entrada	FAC1_1	5,991	-,079	3,018	,375	,314	,142	-,034	,224	,307	-,234		
	FAC2_1	6,662	,059	,216	,098	,136	,069	-,021	-,054	-,049	-,004		
	FAC3_1	2,041	,127	1,185	,073	-,001	,047	-,010	-,031	-,054	-,008		
	FAC4_1	1,899	,273	,150	,435	-,053	7,083	-,086	-,150	-,123	-,134		
	FAC5_1	1,333	,041	,432	-,010	,009	,001	-,009	,044	-,030	-,018		
	FAC6_1	-,416	-,274	-,169	,304	,563	,095	-,275	1,316	,885	-,506		
	FAC7_1	2,625	,231	2,284	-,080	2,463	-,082	-,034	-,153	-,449	-,032		
	FAC8_1	,851	,231	1,274	-,020	3,681E-5	-,104	-,074	-,132	2,551	-,211		
	FAC9_1	-2,546	1,513	-,512	,533	-,083	-,148	-,025	-,135	-,261	,003		
	FAC10_1	4,824	,331	-,032	,064	-,055	-,048	-,003	-,080	-,093	,018		
	FAC11_1	-2,266	,728	-,279	,250	-,089	-,044	-,001	-,108	-,153	,021		
Ancho de unidad oculta		32,226	4,173	4,597	1,079	,847	1,941	,360	,380	,708	,311		
Capa oculta	H(1)											,120	,880
	H(2)											2,485	-1,485
	H(3)											-2,319	3,319
	H(4)											-,662	1,662
	H(5)											-,665	1,665
	H(6)											-,508	1,508
	H(7)											,755	,245
	H(8)											-,807	1,807
	H(9)											-,939	1,939
	H(10)											2,830	-1,830

a. Muestra el vector del centro para cada unida oculta.

Tabla 4-21

Por otra parte, la matriz de confusión, presentada en la tabla 4-22 muestra porcentajes de aciertos en la clasificación inferiores a los mostrados en la matriz de confusión del Perceptrón Multicapa, pero también son altos. Esto indica que, aunque le modelo de Función de Base Radial prediga con menos precisión que le Perceptrón Multicapa, sigue siendo efectivo.

Clasificación

		Pronosticado		
Ejemplo	Observado	,00	1,00	Porcentaje correcto
Entrenamiento	,00	345414	158426	68,6%
	1,00	153492	692364	81,9%
	Porcentaje global	37,0%	63,0%	76,9%
Pruebas	,00	148627	67904	68,6%
	1,00	66147	296120	81,7%
	Porcentaje global	37,1%	62,9%	76,8%

Variable dependiente: fraude global

Tabla 4-22

En cuanto al área bajo la curva ROC (Figura 4-9), observamos un valor de 0,858 (tabla 4-23) que, aunque es menor que en el caso dela red Perceptrón Multicapa, sería bastante aceptable.

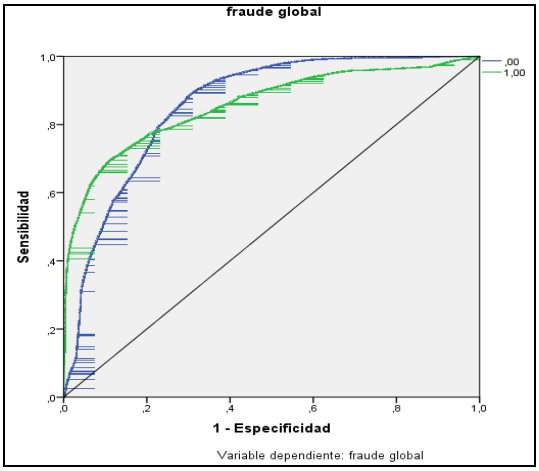


Figura 4-9

		Área
fraude global	,00	,858
	1,00	,858

Tabla 4-23

Las figuras 4-10 y 4-11 muestran el gráfico de ganancias y el gráfico de elevación de la red neuronal de función de base radial RBF.

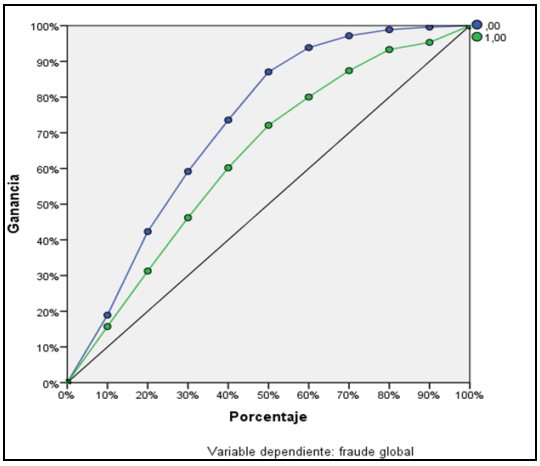


Figura 4-10

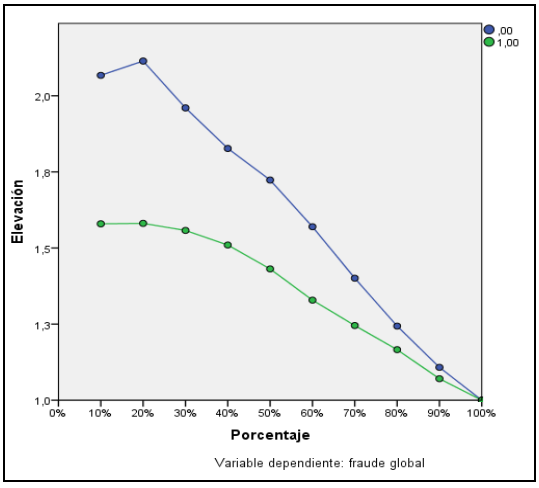


Figura 4-11

Las figuras 4-12 y 4-13 comparan los gráficos de ganancias de la red Perceptron Multicapa con la red Función de Base Radial. Como para los mismos porcentajes es mejor gráfico el que tiene más alta la ganacia,

observamos que es más eficiente la red Perceptrón Multicapa. Este criterio suele ser equivalente a considerar mejor gráfico de ganancia aquel que presenta mayor área entre las dos curvas de la figura.

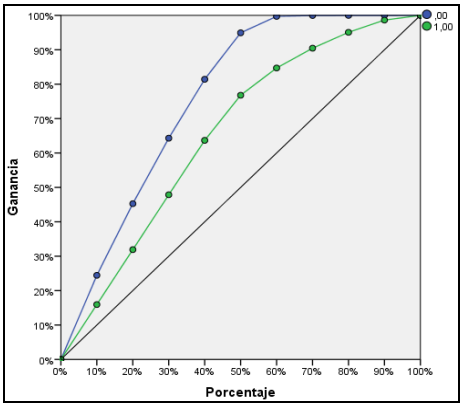


Figura 4-12

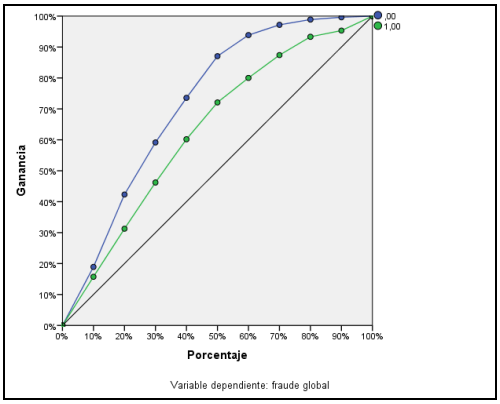


Figura 4-13

Las dos figuras 4-14 y 4-15 comparan los gráficos de elevación de la red Perceptron Multicapa con la red Función de Base Radial. Sabemos que a mayor elevación para el mismo porcentaje mejor predice el modelo. Se observa que predice mejor la red Perceptrón Multicapa. Este criterio suele ser equivalente a considerar mejor gráfico de elevación aquel que más separa las dos curvas de la figura.

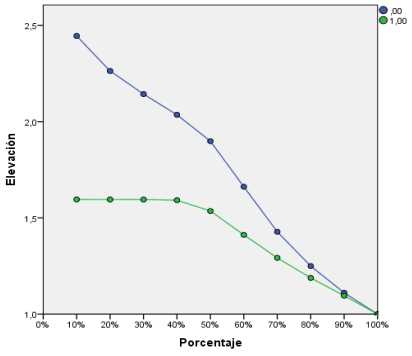


Figura 4-14

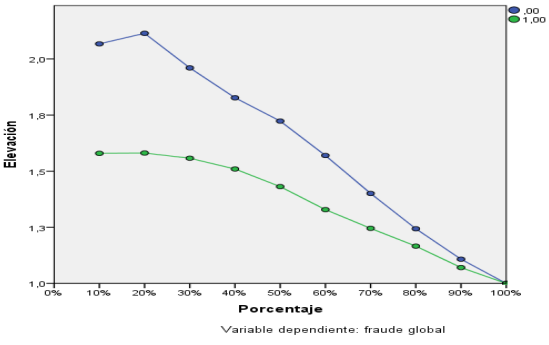


Figura 4-15

4.6.3 Cálculo de las probabilidades de fraude. Propensión al fraude

La tabla 4-24 muestra el cálculo de las probabilidades de fraude (propensiones al fraude) de cada declarante para el Perceptrón Multicapa y la Función de Base Radial. La columna *MLP-PredictiveValue* muestra si el individuo se clasificó en fraudulento (valor 1) o no fraudulento (valor cero) según el Perceptrón Multicapa, la columna *MLP_PseudoProbability_2* muestra la probabilidad de fraude para cada individuo declarante del impuesto según el Perceptrón Multicapa y la columna *MLP_PseudoProbability_1* muestra la probabilidad de no fraude. Alternativamente, la columna *RBF-PredictiveValue* muestra si el individuo se clasificó en fraudulento (valor 1) o no fraudulento (valor cero) según la Red de Base Radial, la columna *RBF_PseudoProbability_2* muestra la probabilidad de fraude para cada individuo declarante del impuesto según la Red de Base Radial y la columna *RBF_PseudoProbability_1* muestra la probabilidad de no fraude.

Tenemos así cuantificadas las propensiones al fraude de los individuos declarantes del Impuesto sobre la Renta de las Personas Físicas.

MLP_PredictedValue	MLP_PseudoProbability_1	MLP_PseudoProbability_2	RBF_PredictedValue	RBF_PseudoProbability_1	RBF_PseudoProbability_2
1,00	,000	1,000	1,00	,141	,859
1,00	,009	,991	,00	,572	,428
,00	,832	,168	,00	,784	,216
,00	,831	,169	,00	,780	,220
,00	,795	,205	,00	,811	,189
,00	,815	,185	,00	,829	,171
,00	,531	,469	1,00	,191	,809
1,00	,011	,989	1,00	,239	,761
1,00	,001	,999	1,00	,242	,758
1,00	,050	,950	1,00	,130	,870
1,00	,001	,999	1,00	,232	,768

Tabla 4-24

4.7 ANÁLISIS DE LOS PERFILES DE FRAUDE Y EXTRACCIÓN DEL CONOCIMIENTO

Una de las ventajas de los modelos predictivos para la detección del fraude radica en la posibilidad de poder calcular probabilidades de fraude individuales para los contribuyentes. La red neuronal ofrece como salida la clasificación de cada declarante como fraudulento o no fraudulento y adicionalmente muestra las propensiones al fraude de cada declarante. Es decir, no sólo clasifica los individuos como propensos o no al fraude, sino que también computa la probabilidad de fraude de cada declarante. La figura 4-16 muestra la densidad de probabilidad de la propensión al fraude mediante el Perceptrón Multicapa, que era la red más efectiva, con mejores gráficos de ganancia y elevación y por lo tanto con mayor capacidad predictiva. Se observa que para probabilidades de fraude bajas hay más densidad de contribuyentes, aunque para probabilidades de fraude alrededor de 0,8 la densidad de declarantes vuelve a repuntar. Esto mismo ya ocurrió cuando consideramos los modelos de análisis discriminante.

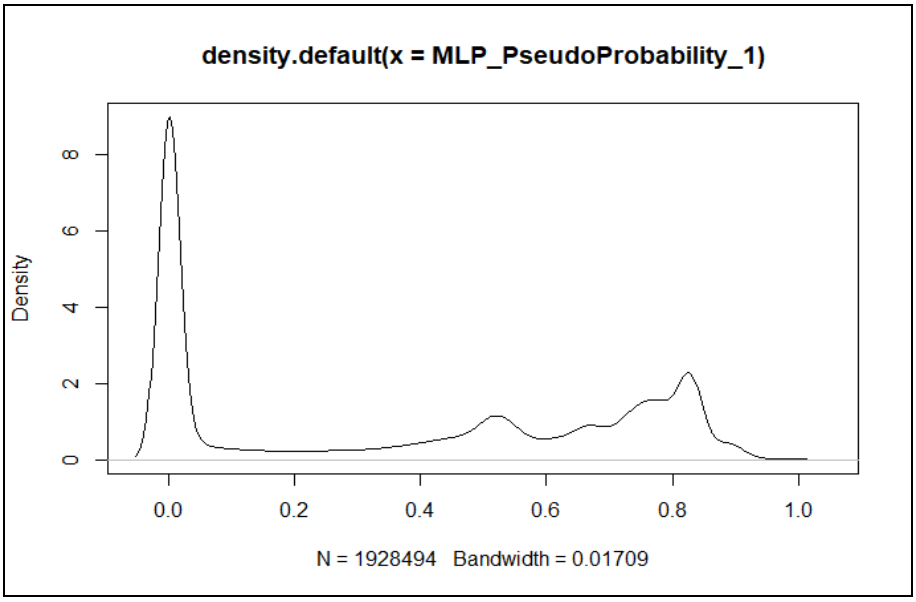


Figura 4-16

Por otra parte, basándonos en la valoración importancia de las variables independientes sobre el fraude global que hemos obtenido al ajustar la red, podremos deducir qué variables son las más influyentes sobre el fraude global.

El análisis de las causas de fraude derivado de la estimación de la Red Perceptrón Multicapa sitúa como partidas más influyentes en el fraude fiscal en IRPF a los rendimientos, bases y cuotas. Los rendimientos ocultos no declarados suelen ser la causa principal de fraude. Estos rendimientos ocultos llevan a la manipulación de las bases imponible y liquidable y por tanto a la minoración del tipo aplicable. Por lo tanto la cuota resultante es inferior a la que debería de ser. Suele ser habitual la presencia de actividades cuyas rentas eluden la tributación, bien por no ser declaradas o bien por no estar registradas constituyendo economía sumergida. De esta forma, el tipo marginal correspondiente a la declaración resulta inferior al real, manipulándose así el resultado de la liquidación. Las cuantías defraudadas por esta causa suelen ser de elevada magnitud.

Los saldos netos positivos de ganancias y pérdidas patrimoniales constituyen una partida también muy influyente en el fraude fiscal en IRPF. Son ganancias y pérdidas patrimoniales las variaciones en el valor del patrimonio del contribuyente que se pongan de manifiesto con ocasión de cualquier alteración en el patrimonio, salvo que sean rendimientos. El cálculo del importe de las ganancias y pérdidas patrimoniales suele ser causa de fraude, bien sea por la mala aplicación de la norma general, por la incorrecta consideración de transmisiones a título oneroso o lucrativo, por la mala aplicación de las normas específicas de valoración, por la incorrecta computación de las ganancias excluidas de gravamen en supuestos de reinversión o por la mala consideración de las ganancias patrimoniales no justificadas. Este apartado está relacionado con el cálculo de la cuota diferencial, ya que en ella interviene la deducción por doble imposición

internacional que se aplica, entre otros supuestos, cuando entre las rentas del contribuyente figuren rendimientos o ganancias patrimoniales obtenidos y grabados en el extranjero. La cuota diferencial se obtiene minorando la cuota líquida total del impuesto por las deducciones y retenciones que marca la ley. La cuota líquida suele ser también objeto de fraude ya que se suelen cometer incorrecciones al minorar la cuota íntegra por las inversiones en empresas de nueva o reciente creación, por las deducciones en actividades económicas, por las deducciones por donativos y otras aportaciones. Evidentemente, cualquier componente que influya en las cuotas, influirá en el resultado de la declaración.

La siguiente partida a considerar por su influencia en el fraude fiscal son los rendimientos de capital mobiliario. Esta partida incluye los rendimientos obtenidos por la participación en los fondos propios de cualquier tipo de entidad (dividendos, rendimientos de activos, primas de emisión de acciones y otros rendimientos), los dividendos obtenidos por la cesión a terceros de capitales propios (contraprestación por cuentas en entidades financieras, rendimientos procedentes de cualquier instrumento de giro o de la cesión temporal de activos financieros u otros rendimientos), rendimientos procedentes de operaciones de capitalización y contratos de seguro de vida o invalidez y otros rendimientos de capital mobiliario (propiedad intelectual de no autores, asistencia técnica, arrendamientos de bienes muebles y negocios, cesión de derechos de imagen y otros rendimientos). Estos rendimientos suelen ser susceptibles de ocultación por parte de los declarantes minorando así los rendimientos íntegros de capital mobiliario. Además, para determinar el rendimiento neto se aplican deducciones de gastos y reducciones (gastos de administración y depósito de valores negociables, gastos necesarios para la obtención de rendimientos provenientes de la prestación de asistencia técnica, del arrendamiento de bienes muebles o del subarrendamiento). Del mismo modo que los rendimientos son susceptibles de ocultación, los gastos son susceptibles de

aumento fraudulento. Los rendimientos de capital mobiliario citados anteriormente (salvo otros rendimientos de capital mobiliario) forman parte de la renta del ahorro y por lo tanto influyen en la base imponible y liquidable del ahorro. Por lo tanto, estas bases del ahorro son también susceptibles de fraude fiscal.

A los rendimientos de capital mobiliario le siguen en importancia sobre el fraude fiscal los rendimientos de capital inmobiliario (rendimientos provenientes de la titularidad de bienes rústicos y urbanos o de derechos reales sobre ellos). La computación de estos rendimientos también suele ser fraudulenta. Lo mismo ocurre con la computación de los gastos deducibles y reducciones para obtener los rendimientos netos del capital inmobiliario. Suelen inflarse los gastos necesarios para la obtención de los rendimientos.

Hasta aquí hemos visto la incidencia en el fraude fiscal de las partidas que conforman los rendimientos del impuesto. Ahora nos ocuparemos de las deducciones y reducciones más influyentes en el fraude fiscal en IRPF. Observamos que las deducciones autonómicas y por donativos suelen incidir en el fraude (estas últimas ya las habíamos citado al hablar de ganancias y pérdidas patrimoniales). Lo mismo ocurre con las deducciones por vivienda habitual y mínimo personal y familiar. Las deducciones por vivienda habitual que todavía perduran dependen del año de construcción del inmueble y este hecho debe de ser objeto de especial vigilancia. El mínimo personal y familiar constituye la parte de la base liquidable que, por destinarse a satisfacer las necesidades básicas personales y familiares del contribuyente, no se somete a tributación en el IRPF. Los mínimos por descendientes, ascendientes y discapacidad también deben de ser objeto de especial vigilancia. Suelen ser habitual el fraude que afecta a las declaraciones del número de hijos y ascendientes y descendientes, que habitualmente eran simultáneamente desgravadas por los dos padres (separados, divorciados o en otras situaciones) en el caso de los hijos o por

diferentes hermanos en el caso de los ascendientes. Otra partida a vigilar son los gastos deducibles totales y los límites de determinadas deducciones con especial referencia a las deducciones por incentivos a la inversión.

En general, la incorrecta declaración de deducciones derivadas de actividades económicas suele ser otra fuente de fraude. También es necesario vigilar el cálculo de los rendimientos íntegros de actividades económicas, la correcta aplicación de las reglas generales del cálculo del rendimiento neto, los elementos patrimoniales afectos a la actividad económica, las normas para la determinación del rendimiento neto en estimación directa y objetiva y las reducciones.

Asimismo las reducciones de la base imponible también suelen ser susceptibles de fraude y más en concreto las reducciones por aportaciones a sistemas de previsión social entre las que se encuentran incluidas especialmente las aportaciones realizadas a planes de pensiones. Esta rúbrica del IRPF fue durante un tiempo el refugio de las rentas altas, ya que desgrava de la base y además por cantidades importantes hasta que se acotó el máximo deducible. Por lo tanto, era objeto de especial tratamiento por los declarantes de IRPF con peligro de deducciones fraudulentas ilegales que acentuó la vigilancia de la inspección. Estas reducciones de base imponible deben de ser objeto de especial vigilancia porque inciden en el tipo a aplicar. Una minoración del tipo es muy incidente en el resultado de la declaración.

De las ganancias y pérdidas patrimoniales ya habíamos hablado al referirnos a los saldos netos positivos de las mismas. Habrá que vigilar que no existe ganancia o pérdida patrimonial en reducciones de capital, con ocasión de transmisiones lucrativas por causa de muerte del contribuyente, en la extinción del régimen económico matrimonial de separación de bienes y con ocasión de las aportaciones a los patrimonios protegidos constituidos a favor de personas con discapacidad. También habrá que tener presente que

no existe alteración de la composición del patrimonio en los supuestos de división de la cosa común, en la disolución de la sociedad de gananciales en la extinción del régimen económico matrimonial de participación y en la disolución de comunidades de bienes o en los casos de separación de comuneros. Todas estas consideraciones hay que tenerlas muy presentes al realizar el cálculo de las ganancias y pérdidas patrimoniales, ya que estas cifras suelen ser fuente común de fraude fiscal en IRPF.

Por último, tenemos como partidas menos incidentes el fraude fiscal los rendimientos netos en módulos agrarios y otros rendimientos. No obstante, estas partidas deben de ser también susceptibles de vigilancia.

Resumiendo un poco el análisis de la importancia de las partidas del impuesto (variables independientes del modelo) sobre el fraude fiscal, vemos que las partidas relativas a rendimientos son las más incidentes en el fraude y en concreto las ganancias y pérdidas patrimoniales, los rendimientos del capital mobiliario e inmobiliario y los rendimientos de actividades económicas.

El otro grupo de partidas más incidentes en el fraude son los gastos deducibles y las reducciones con especial incidencia de las deducciones autonómicas y donativos, las deducciones por vivienda, las deducciones por incentivos a la inversión y las gastos deducibles totales. Esta última partida implica la vigilancia de cualquier tipo de gasto deducible.

En cuanto a las reducciones, serán objeto de vigilancia cualquier tipo de reducción de la base imponible, con especial interés en las reducciones por aportaciones a sistemas de previsión social entre las que se encuentran incluidas especialmente las aportaciones realizadas a planes de pensiones.

No hay que olvidarse tampoco de la vigilancia de los mínimos personales y familiares, por descendientes y ascendientes y por discapacidad.

4.8 SEGMENTACIÓN DE LAS CAUSAS DE FRAUDE

Para segmentar las causas de fraude utilizaremos la técnics estadístics del Escalamiento Multidimensional. Recordamos que el Escalamiento multidimensional es una técnica descriptiva de minería de datos que permite segmentar variables de un conjunto de datos agrupándolas por similitud en un mapa perceptual.

Si aplicamos escalamiento multidimensional para ver como se relacionan las diversas causas de fraude con la probabilidad de fraude global según el modelo de redes neuronales, obtenemos el mapa perceptual de la figura 4-17.

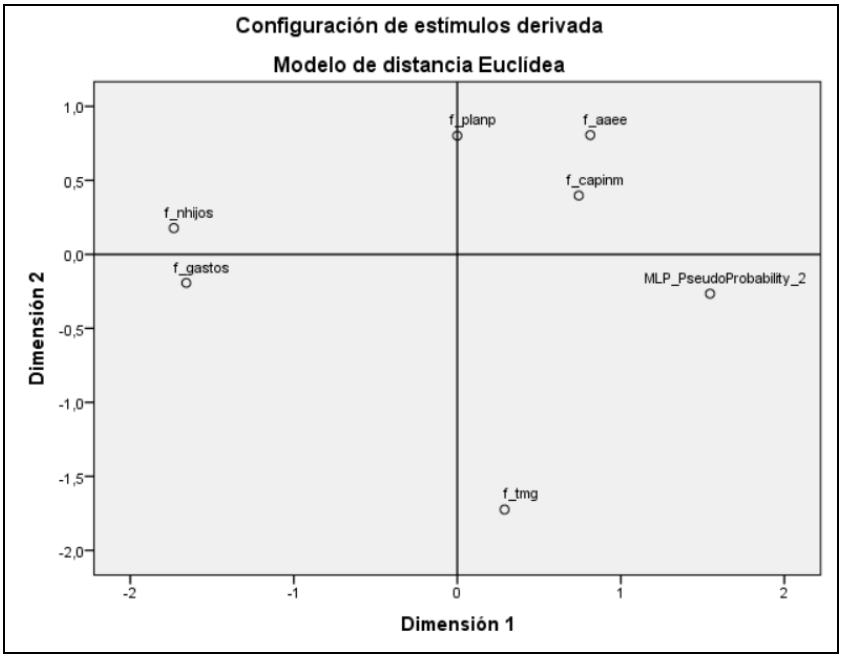


Figura 4-17

La segmentación del mapa perceptual nos indica que el fraude en actividades económicas, en planes de pensiones y en rendimientos capital inmobiliario tienen una incidencia similar en la probabilidda de fraude global. Lo mismo ocurre con el fraude por declaración incorrecta de gastos y número de hijos y ascendientes. El fraude en la alteración del tipo marginal se comporta aisladamente de las demás causas. Este resultado es el mismo que hemos obtenido en los árboles de decisión y en los modelos de análisis discriminante.

Para evaluar este tipo de escalamiento se utiliza el gráfico de disparidades de la figura 4-18

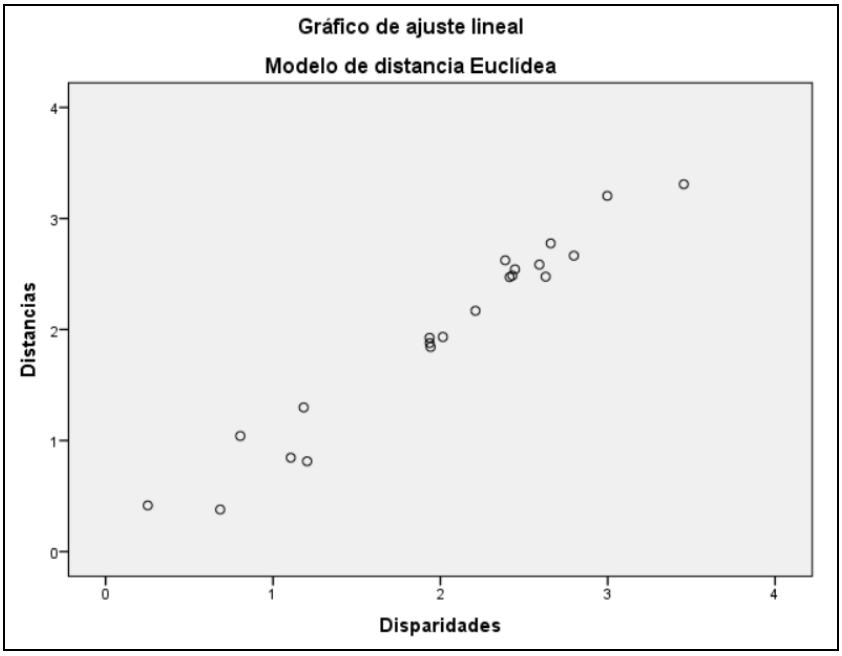


Figura 4-18

El escalamiento multidimensional utilizado es correcto porque el gráfico de disparidades presenta una nube de puntos que se ajusta bien a la diagonal del primer cuadrante.

Además, el estadístico S-Stress toma un valor bajo cercano a cero y el estadístico RSQ toma un valor alto cercano a la unidad

Stress = ,07959 RSQ = ,96036

Concluimos que la segmentación realizada de las causa de fraude es correcta.

4.9 ANÁLISIS DE LOS MODELOS DE REDES NEURONALES PARA LAS CAUSAS DE FRAUDE

Analizaremos ahora los modelos de redes neuronales que tienen como variables dependientes las distintas causas de fraude. Hasta ahora solamente se había utilizado como variaable dependiente el fraude global. Realmente, la tarea se puede utilizar una única red neuronal que tiene como entradas las componentes principales de las partidas económicas del IRPF y como salidas las variables relativas a las distintas causas de fraude.

4.9.1 Estimación de una red neuronal Perceptrón Multicapa Múltiple

Estimamos ahora un modelo de red neuronal de tipo Perceptrón Multicapa múltiple cuyas variables de entrada son las 11 componentes principales obtenidas en la reducción de la dimensión y cuyas variables de salida son las variables dicotómicas relativas a las distintas causas de fraude.

A continuación se analizan los resultados de la estimación de esta la red neuronal múltiple.

Se han utilizado un 70% de los datos para la fase de entrenamiento y un 30% para la fase de pruebas (tabla 4-25). En total tenemos prácticamente 2.000.000 de filas en la base de datos tal y como indica la tabla siguiente

(exactamente 1.928.494) de las cuales 1.350.974 se utilizan para entrenar la red y el resto para prueba. No hay datos faltantes en la base de datos.

Resumen del procesamiento de los casos

		N	Porcentaje
Muestra	Entrenamiento	1350974	70,1%
	Prueba	577520	29,9%
Válidos		1928494	100,0%
Excluidos		0	
Total		1928494	

Tabal 4-25

En la tabla 4-26 se observa que se han utilizado como variables de entrada de la red las componentes principales, lo cual supone las ventajas ya conocidas. Se aminora el efecto de los valores atípicos, se eliminan problemas de multicolinealidad en el modelo, se induce normalidad en las variables del modelo, se uniformiza la escala de medidad de las variables independientes para no tener problemas de no normalidad residual, heteroscedasticidad y autocorrelación y finalmente el modelo suele tener una buena diagnosis de ajuste.

También observamos en la tabla 4-26 que se ha utilizado una capa de entrada con las 11 nodos relativos a las variables independientes, una única capa oculta con 11 nodos y una capa de salida con 4 nodos relativos a las 4 causas variables que delimitan las caussaa de fraude.

Como función de activación en las capas ocultas se ha utilizado la función tangente hiperbólica. La función de activación en la capa de resultado (salida) es la función Softmax. Asimismo se observa que hemos utilizado una única capa oculta en la red.

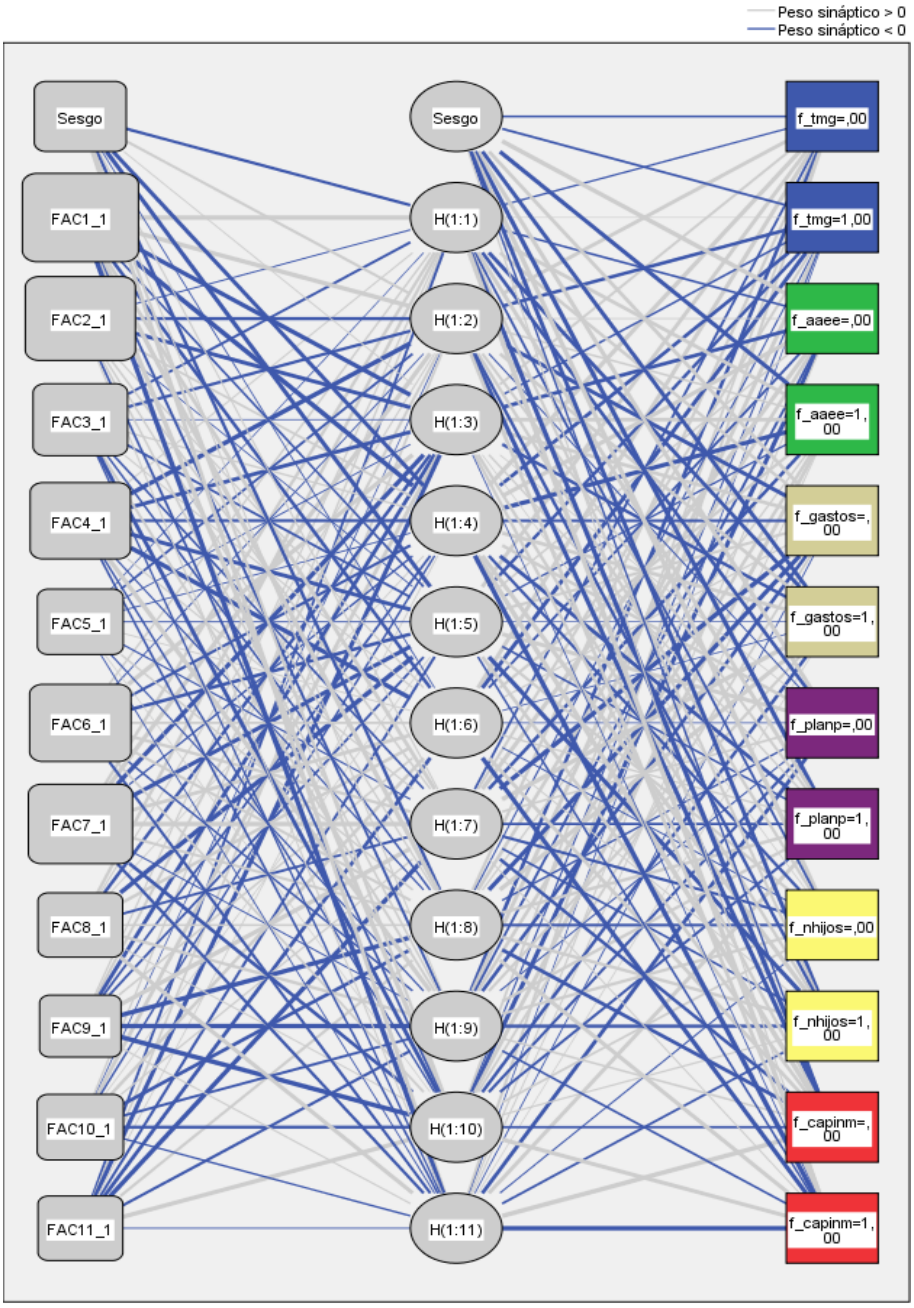
Información sobre la red

Capa de entrada	Covariables	1	RENDIMIENTOS, BASES Y CUOTAS
		2	SALDOS PATRIMONIALES, CUOTA DIFERENCIAL Y RESULTADO
		3	CAPITAL MOBILIARIO Y BASE DEL AHORRO
		4	CAPITAL INMOBILIARIO
		5	DEDUCCIONES AUTONÓMICAS Y DONATIVOS
		6	DEDUCCIONES VIVIENDA Y MÍNIMO PERSONAL Y FAMILIAR
		7	ACTIVIDADES ECONÓMCAS
		8	REDUCCIONES BASE IMPONIBLE Y PLANES DE PENSIONES
		9	GASTOS DEDUCIBLES TOTALES Y DEDUCCION INCENTIVOS INVERSION
		10	GANANCIAS Y PÉRDIDAS PATRIMONIALES
		11	MÓDULOS AGRARIOS Y OTRAS DEDUCCIONES
	Número de unidades[a]		11
	Método de cambio de escala para las covariables		Tipificados
Capas ocultas	Número de capas ocultas		1
	Número de unidades de la capa oculta 1[a]		11
	Función de activación		Tangente hiperbólica
Capa de salida	Variables dependientes	1	Fraude que afecta al tipo marginal
		2	Fraude que afecta a la declaración de actividades económicas
		3	Fraude que afecta a la declaración de gastos
		4	Fraude que afecta a la desgrbación por planes de pensiones
		5	Fraude que afecta a las declaraciones de número e hijos y ascendieentes y descendientes
		6	Fraude que afecta a los rendimientos de capital inmobiliario
	Número de unidades		12
	Función de activación		Softmax
	Función de error		Entropía cruzada

a. Sin incluir la unidad de sesgo

Tabla 4-26

La figura 4-19 muestra la estructura de la red neuronal con sus once nodos correspondientes a la variables de entrada o variables independientes (componentes principales), su única capa oculta cuyos nodos están etiquetados con las etiquetas de los pesos sinápticos y un nodo de salida etiquetado con las dos categorías de las variables dependientes del modelo de red.



Función de activación de capa oculta: Tangente hiperbólica

Función de activación de capa de salida: Softmax

Figuar 4-19

4.9.2 Diagnosis de la red neuronal Perceptrón Multicapa Múltiple

En cuanto a la diagnosis del modelo de red, vemos en primer lugar las matrices de confusión para variable de salida de la red (tablas 4-27 a 4-33). Observamos que estas matrices presentan altos porcentajes de acierto en los valores pronosticados. Los porcentajes de acierto más aaltos se observan para el fraude relativo a actividades económicas, rendimientos de capital inmobiliario, declaración de gastos y tipo marginal. El porcentaje global conjunto de aciertos es del 90,7 por ciento, lo cual indica un buen ajuste de la red neuronal múltiple Perceptrón Multicapa.

Fraude que afecta al tipo marginal

		Pronosticado		
Muestra	Observado	,00	1,00	Porcentaje correcto
Entrenamiento	,00	768606	87250	89,8%
	1,00	251696	243422	49,2%
	Porcentaje global	75,5%	24,5%	74,9%
Prueba	,00	328718	37173	89,8%
	1,00	107327	104302	49,3%
	Porcentaje global	75,5%	24,5%	75,0%

Tabla 4-27

Fraude que afecta a la declaración de actividades económicas

		Pronosticado		
Muestra	Observado	,00	1,00	Porcentaje correcto
Entrenamiento	,00	850179	32718	96,3%
	1,00	50898	417179	89,1%
	Porcentaje global	66,7%	33,3%	93,8%
Prueba	,00	363921	13775	96,4%
	1,00	21706	178118	89,1%
	Porcentaje global	66,8%	33,2%	93,9%

Tabla 4-28

Fraude que afecta a la declaración de gastos

		Pronosticado		
Muestra	Observado	,00	1,00	Porcentaje correcto
Entrenamiento	,00	1165012	12356	99,0%
	1,00	64380	109226	62,9%
	Porcentaje global	91,0%	9,0%	94,3%
Prueba	,00	497632	5432	98,9%
	1,00	27743	46713	62,7%
	Porcentaje global	91,0%	9,0%	94,3%

Tabla 4-29

Fraude que afecta a la desgrbación por planes de pensiones

		Pronosticado		
Muestra	Observado	,00	1,00	Porcentaje correcto
Entrenamiento	,00	914894	61343	93,7%
	1,00	125402	249335	66,5%
	Porcentaje global	77,0%	23,0%	86,2%
Prueba	,00	391514	26057	93,8%
	1,00	53641	106308	66,5%
	Porcentaje global	77,1%	22,9%	86,2%

Tabla 4-30

Fraude que afecta a las declaraciones de número e hijos y ascendieentes y descendientes

		Pronosticado		
Muestra	Observado	,00	1,00	Porcentaje correcto
Entrenamiento	,00	1289777	645	100,0%
	1,00	59937	615	1,0%
	Porcentaje global	99,9%	0,1%	95,5%
Prueba	,00	551309	262	100,0%
	1,00	25679	270	1,0%
	Porcentaje global	99,9%	0,1%	95,5%

Tabla 4-31

Fraude que afecta a los rendimientos de capital inmobiliario

		Pronosticado		
Muestra	Observado	,00	1,00	Porcentaje correcto
Entrenamiento	,00	866762	5170	99,4%
	1,00	3156	475886	99,3%
	Porcentaje global	64,4%	35,6%	99,4%
Prueba	,00	370948	2256	99,4%
	1,00	1328	202988	99,4%
	Porcentaje global	64,5%	35,5%	99,4%

Tabla 4-32

Porcentaje global correcto

Muestra	Porcentaje global correcto
Entrenamiento	90,7%
Prueba	90,7%

Tabla 4-33

Un elemento esencial de diagnosis lo constituyen las curvas ROC de la red. La tabla siguiente muestra el área bajos las curvas ROC relativas a las causas de fraude (tabla 4-34). Se constata que las áreas más elevadas son las relativas a rendimientos de capital inmobiliario, actividades económicas, declaración de gastos, planes de pensiones y tipo marginal.

Área debajo de la curva

		Área
Fraude que afecta al tipo marginal	,00	,796
	1,00	,796
Fraude que afecta a la declaración de actividades económicas	,00	,980
	1,00	,980
Fraude que afecta a la declaración de gastos	,00	,910
	1,00	,910
Fraude que afecta a la desgrbación por planes de pensiones	,00	,902
	1,00	,902
Fraude que afecta a las declaraciones de número e hijos y ascendieentes y descendientes	,00	,761
	1,00	,761
Fraude que afecta a los rendimientos de capital inmobiliario	,00	1,000
	1,00	1,000

Tabla 4-34

En las gráficas siguientes (figuras 4-20 a 4-25) se observan las curvas ROC para las distintas causas de fraude, presentando ambas un área muy elevada entre las curvas y la diagonal.

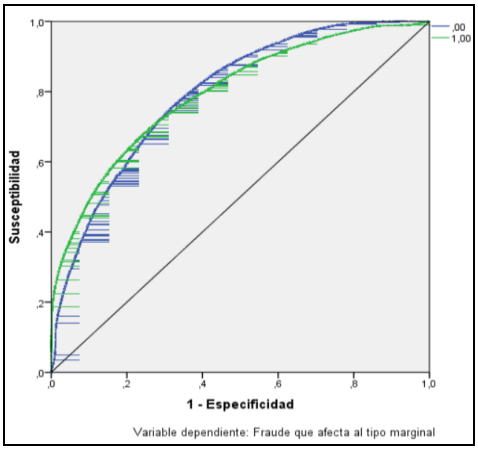


Figura 4-20

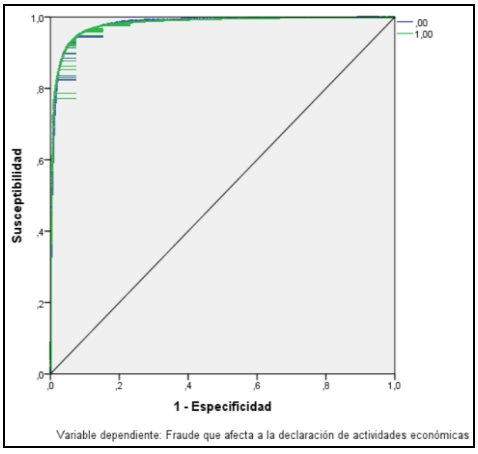


Figura 4-21

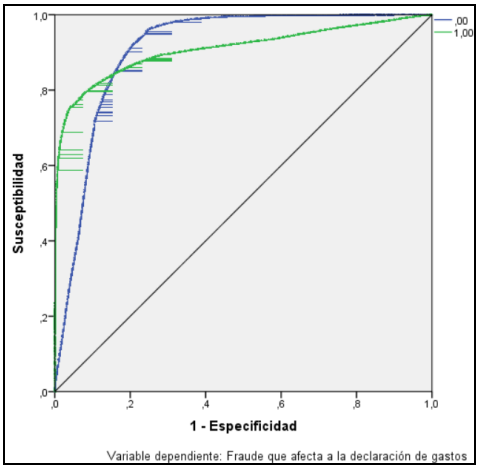


Figura 4-22

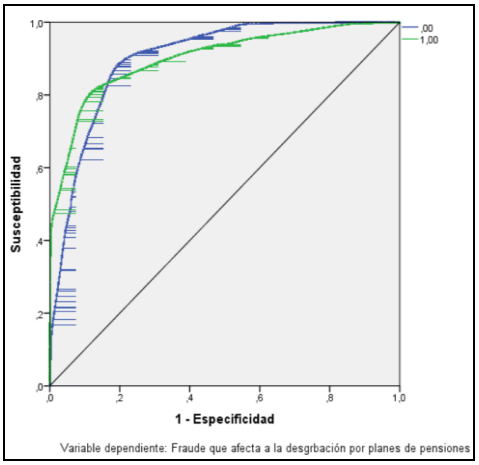


Figura 4-23

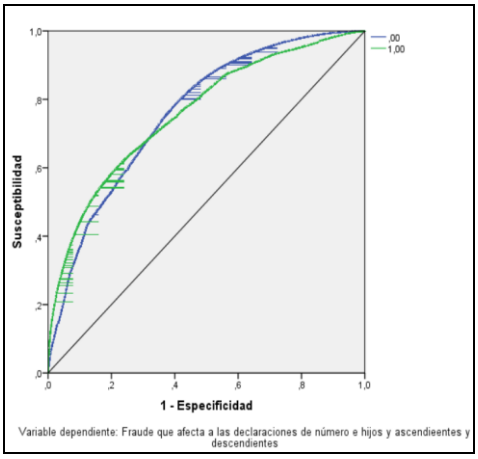


Figura 4-24

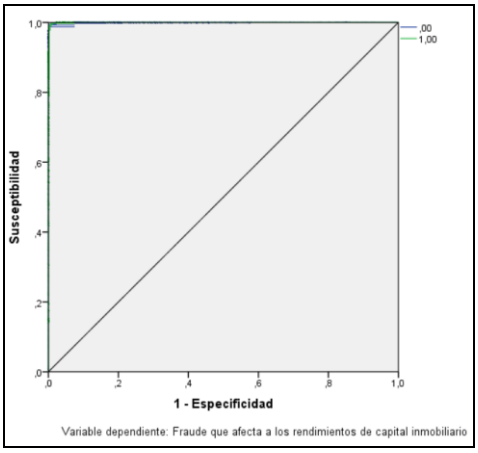


Figura 4-25

La curva de ganancias es otro elemento de diagnosis para comparar modelos alternativos (figuras 4-26 a 4-31). Para porcentajes entre el 40% y el 70% se obtiene la zona de mayor anchura entre las dos curvas del gráfico. Un modelo predice mejor que otro cuando la anchura entre las dos curvas es mayor para los mismos porcentajes. Es decir, a mayor ganancia para el mismo porcentaje, mejor predice el modelo. los gráficos siguientes muestran las curvas de ganacia de la red para cada causa de fraude.

Se observa que las ganacias se ordenan para las distintas causa de fraude de la misma forma que se habían ordenado las áreas bajo la curva ROC y el porcentraje de aciertos de las maatrices de confusión. Las mayores ganacias son las relativas a rendimientos de capital inmobiliario, actividades económicas, declaración de gastos, planes de pensiones y tipo marginal.

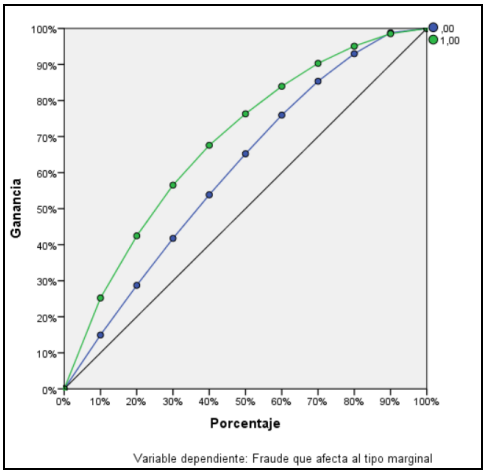


Figura 4-26

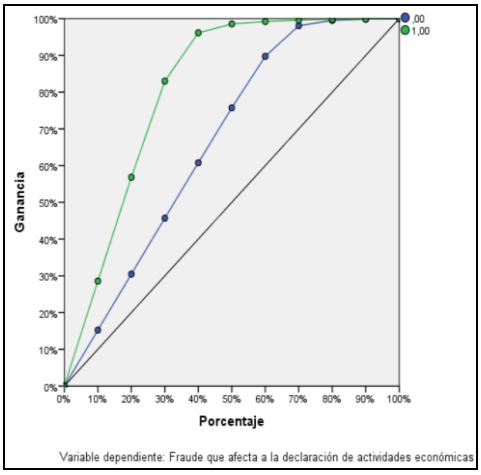


Figura 4-27

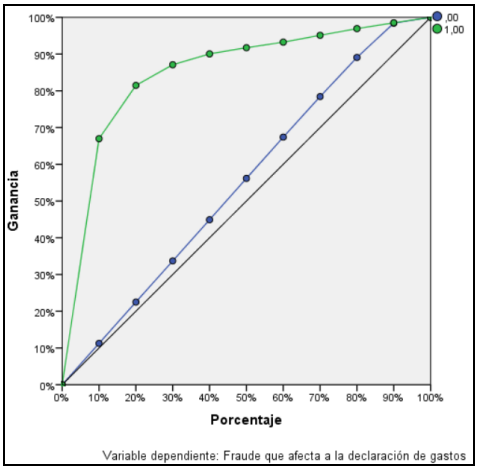


Figura 4-28

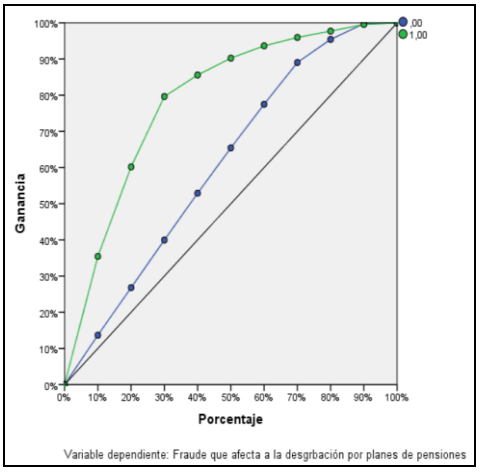


Figura 4-29

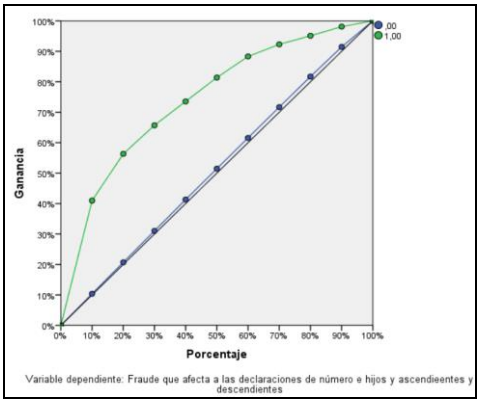


Figura 4-30

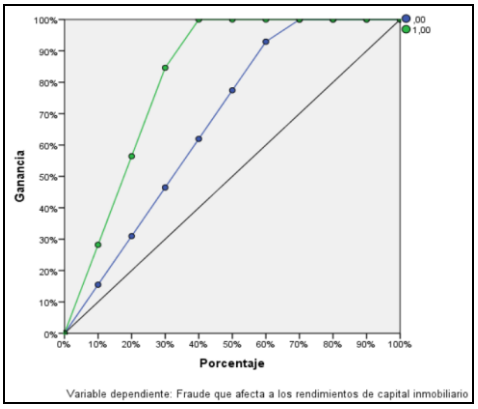


Figura 4-31

El gráfico de elevación es un gráfico alternativo al gráfico de ganancias para comparar la capacidad predictiva de dos modelos. A mayor elevación para el mismo porcentaje, mejor predice el modelo. En las figuras 4-32 a 4-37 se muestran los gráficos de elevación de la red para las distintas causas de fraude. Se observa que las elevaciones se ordenan para las distintas causas de fraude de la misma forma que se habían ordenado las áreas bajo la curvaa ROC, las ganacias y el porcentraje de aciertos de las matrices de confusión. Las mayores elevaciones son las relativas a rendimientos de capital inmobiliario, actividades económicas, declaración de gastos, planes de pensiones y tipo marginal.

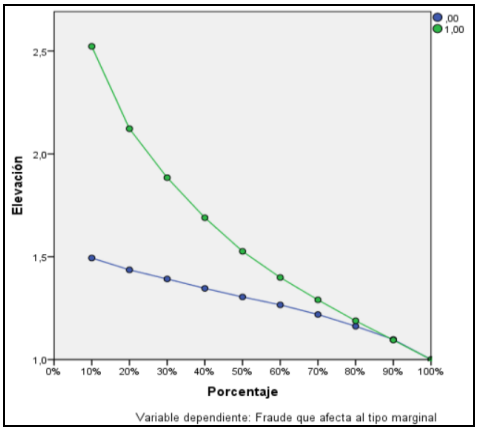


Figura 4-32

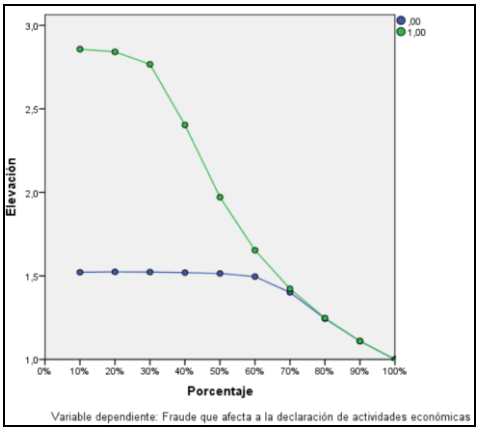


Figura 4-33

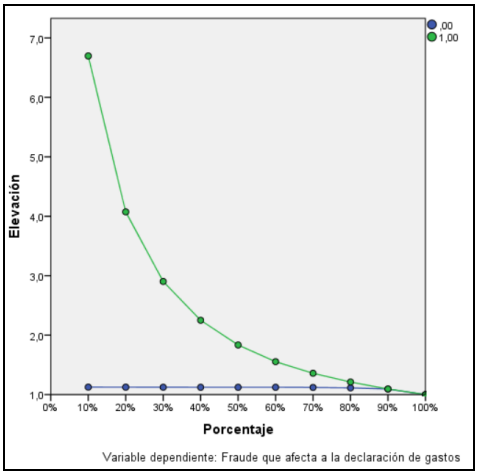


Figura 4-34

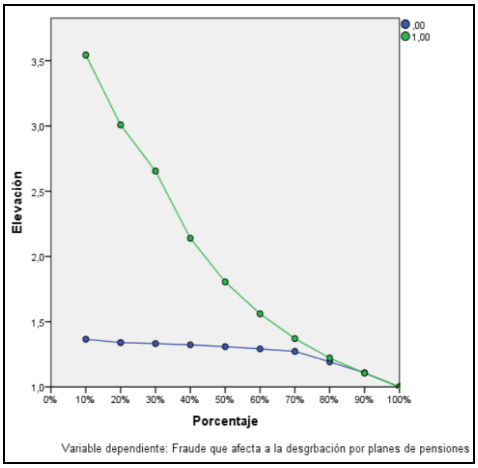


Figura 4-35

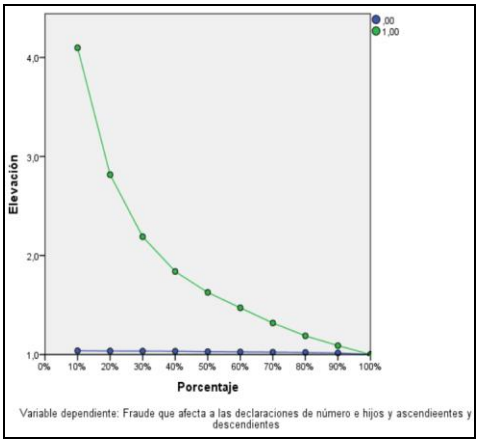


Figura 4-36

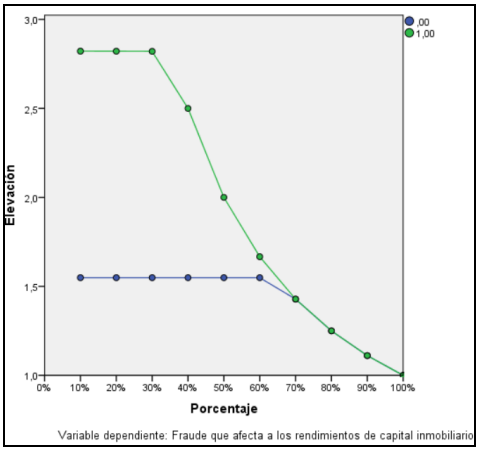


Figura 4-37

Las áreas más elevadas son las relativas a rendimientos de capital inmobiliario, actividades económicas, declaración de gastos, planes de pensiones y tipo marginal.

En cuanto a la importancia de los predictores (variables independientes) sobre la propensión al fraude vemos en la tabla 4-35 que la variable más incidente en el fraude es la componente relativa a rendimientos, bases y cuotas.

Le sigue la componente relativa a saldos patrimoniales, cuota diferencial y resultado de la declaración. Es muy destacado también el fraude por actividades económicas.

También tienen mucha influencia sobre el fraude las variables relativas a capital mobiliario e inmobiliario, así como las deducciones autonómicas, las deducciones por donativos y las deducciones por vivienda y mínimos personal y familiar.

Por último también afectan significativamente al fraude las actividades económicas, las deducciones de base imponible y planes de pensiones y los gastos deducibles totales.

Estos resultados son equivalentes a los obtenidos en el caso de la red neuronal simple para el fraude global y están muy en línea con la diagnosis gráfica de la red que se acaba de realizar.

Importancia de las variables independientes

	Importancia	Importancia normalizada
RENDIMIENTOS, BASES Y CUOTAS	,159	100,0%
SALDOS PATRIMONIALES, CUOTA DIFERENCIAL Y RESULTADO	,139	87,8%
CAPITAL MOBILIARIO Y BASE DEL AHORRO	,088	55,4%
CAPITAL INMOBILIARIO	,107	67,4%
DEDUCCIONES AUTONÓMICAS Y DONATIVOS	,059	37,3%
DEDUCCIONES VIVIENDA Y MÍNIMO PERSONAL Y FAMILIAR	,110	69,5%
ACTIVIDADES ECONÓMCAS	,120	75,4%
REDUCCIONES BASE IMPONIBLE Y PLANES DE PENSIONES	,056	35,2%
GASTOS DEDUCIBLES TOTALES Y DEDUCCION INCENTIVOS INVERSION	,045	28,1%
GANANCIAS Y PÉRDIDAS PATRIMONIALES	,060	38,0%
MÓDULOS AGRARIOS Y OTRAS DEDUCCIONES	,056	35,5%

Tabla 4-35

La figura 4-38 muestra la importancia de las distintas partidas del impuesto ya comentadas de forma gráfica.

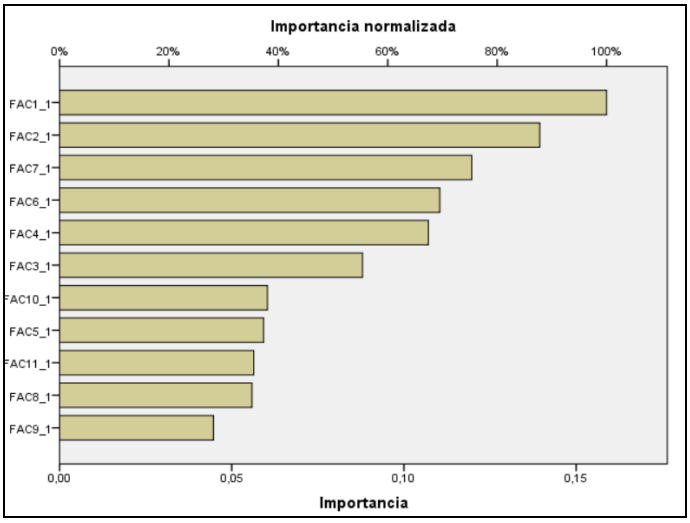


Figura 4-38

Por último, la diagnosis de la red presenta un análisis de errores del modelo (tabla 4-36).

Resumen del modelo

Entrenamiento	Error de entropía cruzada		1841948,319
	Pronósticos incorrectos de porcentaje promedio		9,3%
	Porcentaje de pronósticos incorrectos para dependientes categóricas	Fraude que afecta al tipo marginal	25,1%
		Fraude que afecta a la declaración de actividades económicas	6,2%
		Fraude que afecta a la declaración de gastos	5,7%
		Fraude que afecta a la desgrbación por planes de pensiones	13,8%
		Fraude que afecta a las declaraciones de número e hijos y ascendieentes y descendientes	4,5%
		Fraude que afecta a los rendimientos de capital inmobiliario	0,6%
	Regla de parada utilizada		Número máximo de épocas (100) superado
	Tiempo de entrenamiento		0:04:10,44
Prueba	Error de entropía cruzada		787235,772
	Pronósticos incorrectos de porcentaje promedio		9,3%
	Porcentaje de pronósticos incorrectos para dependientes categóricas	Fraude que afecta al tipo marginal	25,0%
		Fraude que afecta a la declaración de actividades económicas	6,1%
		Fraude que afecta a la declaración de gastos	5,7%
		Fraude que afecta a la desgrbación por planes de pensiones	13,8%
		Fraude que afecta a las declaraciones de número e hijos y ascendieentes y descendientes	4,5%
		Fraude que afecta a los rendimientos de capital inmobiliario	0,6%

Tabla 4-36

4.9.3 Cálculo de las probabilidades de fraude. Propensión al fraude

La tabla 4-37 muestra el grupo en el que se clasifican los constribuyentes (fraude o no fraude) para las seis causas de fraude consideradas (primeros registros de entre los 2 millones totales). La columna *MLP-PredictiveValue_i* muestra si el individuo se clasificó en fraudulento (valor 1) o no fraudulento (valor cero) según el Perceptrón Multicapa para la causa de fraude iésima.

MLP_PredictedValue_1	MLP_PredictedValue_2	MLP_PredictedValue_3	MLP_PredictedValue_4	MLP_PredictedValue_5	MLP_PredictedValue_6
1,00	,00	,00	,00	,00	,00
,00	,00	1,00	,00	,00	,00
,00	,00	,00	,00	,00	,00
,00	,00	,00	,00	,00	,00
,00	,00	,00	,00	,00	,00
,00	,00	,00	,00	,00	,00
,00	,00	,00	,00	,00	,00
,00	,00	,00	,00	,00	,00
1,00	,00	1,00	,00	,00	,00
,00	1,00	,00	1,00	,00	,00
1,00	1,00	,00	1,00	,00	1,00

Tabla 4-37

En las tablas 4-38 y 4-39 se presentan los primeros registros del cálculo de las probabilidades de fraude (propensiones al fraude) de cada declarante para el Perceptrón Multicapa. La *columna MLP_PseudoProbability_i_2* muestra la probabilidad de fraude para cada individuo declarante del impuesto según el Perceptrón Multicapa según la causa de fraude iésima y la columna *MLP_PseudoProbability_i_1* muestra la probabilidad de no fraude para los mismos individuos.

Tenemos así cuantificadas las propensiones al fraude de los individuos declarantes del Impuesto sobre la Renta de las Personas Físicas para las distintas causas de fraude.

MLP_PseudoProbability_1_1	MLP_PseudoProbability_1_2	MLP_PseudoProbability_2_1	MLP_PseudoProbability_2_2	MLP_PseudoProbability_3_1	MLP_PseudoProbability_3_2
,006	,994	,844	,156	,930	,070
,589	,411	,997	,003	,084	,916
,903	,097	,992	,008	,967	,033
,903	,097	,993	,007	,960	,040
,891	,109	,979	,021	,972	,028
,894	,106	,990	,010	,962	,038
,836	,164	,987	,013	,589	,411
,901	,099	,976	,024	,653	,347
,027	,973	,999	,001	,354	,646
,535	,465	,136	,864	,969	,031
,252	,748	,035	,965	,985	,015
,665	,335	,012	,988	,966	,034
,673	,327	,055	,945	,984	,016
,716	,284	,989	,011	,895	,105
,956	,044	,965	,035	1,000	,000
,834	,166	,966	,034	,934	,066

Tabla 4-38

MLP_PseudoProbability_4_1	MLP_PseudoProbability_4_2	MLP_PseudoProbability_5_1	MLP_PseudoProbability_5_2	MLP_PseudoProbability_6_1	MLP_PseudoProbability_6_2
,647	,353	,918	,082	,999	,001
,638	,362	,979	,021	,960	,040
,980	,020	,992	,008	1,000	,000
,981	,019	,991	,009	1,000	,000
,973	,027	,987	,013	1,000	,000
,977	,023	,990	,010	1,000	,000
,941	,059	,946	,054	1,000	,000
,898	,102	,964	,036	,999	,001
,925	,075	,920	,080	1,000	,000
,360	,640	,957	,043	,815	,185
,154	,846	,979	,021	,010	,990
,652	,348	,937	,063	,000	1,000
,586	,414	,962	,038	,889	,111
,894	,106	,967	,033	1,000	,000
,987	,013	,983	,017	1,000	,000
,951	,049	,964	,036	1,000	,000
,778	,222	,961	,039	,000	1,000
,743	,257	,972	,028	,000	1,000

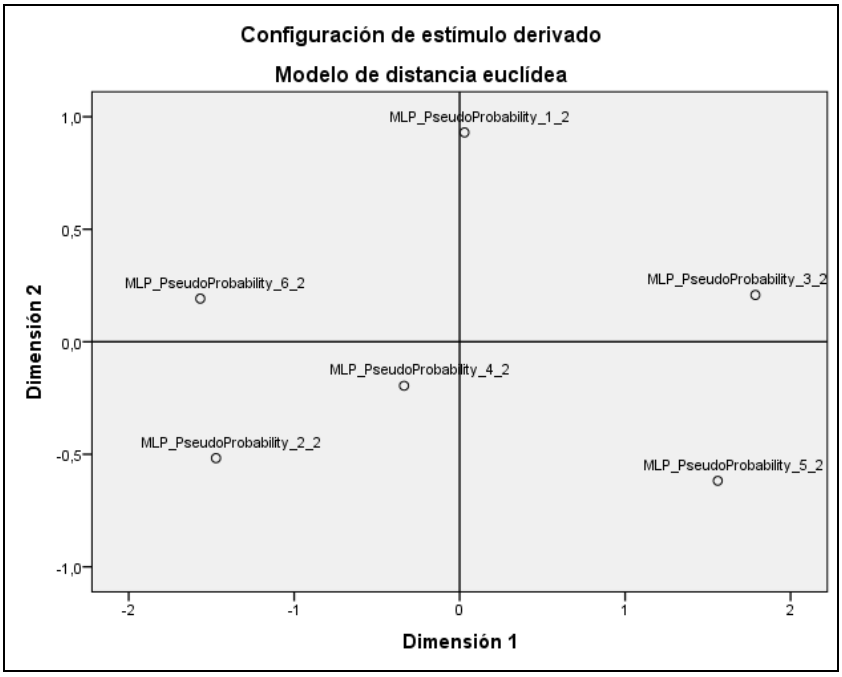
Tabla 4-39

4.10 SEGMENTACIÓN DE LAS CAUSAS DE FRAUDE

Para segmentar las causas de fraude utilizaremos la técnica estadística del Escalamiento Multidimensional.

Recordamos que el Escalamiento multidimensional es una técnica descriptiva de minería de datos que permite segmentar variables de un conjunto de datos agrupándolas por similitud en un mapa perceptual.

Si aplicamos escalamiento multidimensional para ver como se relacionan las probabilidades de las diversas cuasas de fraude según el modelo múltiple de redes neuronales, obtenemos el mapa perceptual de la figura 4-39.



Figuar 4-39

La segmentación del mapa perceptual nos indica que el fraude en actividades económicas (*MLP_PseudoProbability_2_2*), en planes de pensiones (*MLP_PseudoProbability_4_2*) y en rendimientos capital inmobiliario ((*MLP_PseudoProbability_6_2*) tienen una incidencia similar en la probabilidad de fraude global. Lo mismo ocurre con el fraude por declaración incorrecta de gastos (*MLP_PseudoProbability_3_2*) y número de hijos y ascendientes (*MLP_PseudoProbability_5_2*). El fraude en la alteración del tipo marginal se comporta aisladamente de las demás causas (*MLP_PseudoProbability_1_2*). Estre resultado es el mismo que hemos obtenido en los árboles de decisión y en los modelos de análisis discriminnate

Para evaluar este tipo de escalamiento se utiliza el gráfico de disparidades de la figura 4-40.

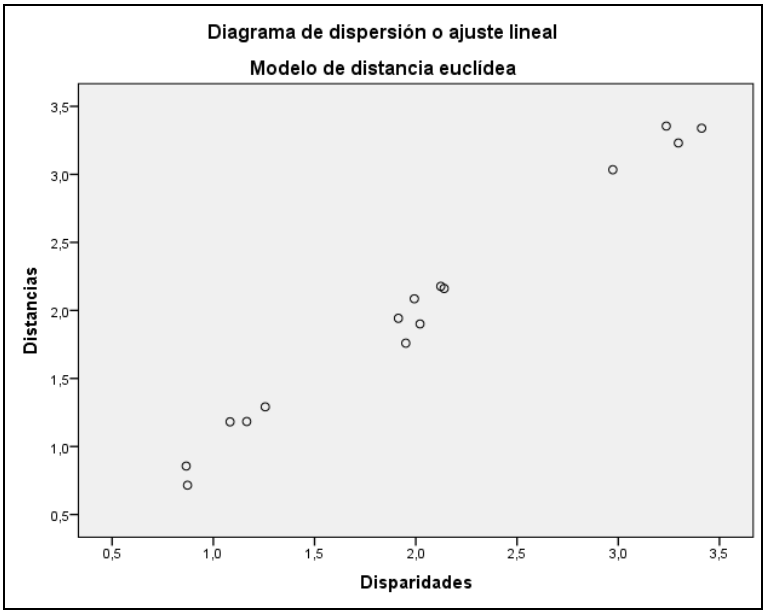


Figura 4-40

El escalamiento multidimemnsional utilizado es correcto porque el gráfico de disparidades presenta una nube de puntos que se ajusta bien a la diagonal del primer cuadrante.

Además, el estadístico S-Stress toma un valor bajo cercano a cero y el estadístico RSQ toma un valor alto cercano a la unidad

```
Stress  =   ,04204      RSQ =  ,98870
```

Concluimos que la segmentación realizada de las causas de fraude es correcta.

4.11 ANÁLISIS DE LOS PERFILES DE FRAUDE Y EXTRACCIÓN DEL CONOCIMIENTO

Los métodos predictivos para el análisis del fraude permiten calcular perfiles de fraude a partir de la función de densidad de la probabilidad de fraude o propensión al fraude (global o por causas de fraude, para todos los individuos

de la muestra). A continuación compararemos las densidades de probabilidad para las propensiones al fraude, calculadas mediante el algoritmo del Kernel, para las distintas causas de fraude (figuras 4-41 a 4-46).

Una de las ventajas de los modelos predictivos para la detección del fraude radica en la posibilidad de poder calcular probabilidades de fraude individuales para los contribuyentes. La red neuronal ofrece como salida la clasificación de cada declarante como fraudulento o no fraudulento y adicionalmente muestra las propensiones al fraude de cada declarante. Es decir, no sólo clasifica los individuos como propensos o no al fraude, sino que también computa la probabilidad de fraude de cada declarante. La figura siguiente muestra la densidad de probabilidad de la propensión al fraude mediante el Perceptrón Multicapa, que era la red más efectiva, con mejores gráficos de ganancia y elevación y por lo tanto con mayor capacidad predictiva. Se observa que para probabilidades de fraude bajas hay más densidad de contribuyentes, aunque la para probabilidades de fraude altas la densidad de declarantes vuelve a repuntar. Esto mismo ya ocurrió cuando consideramos los modelos de análisis discriminante.

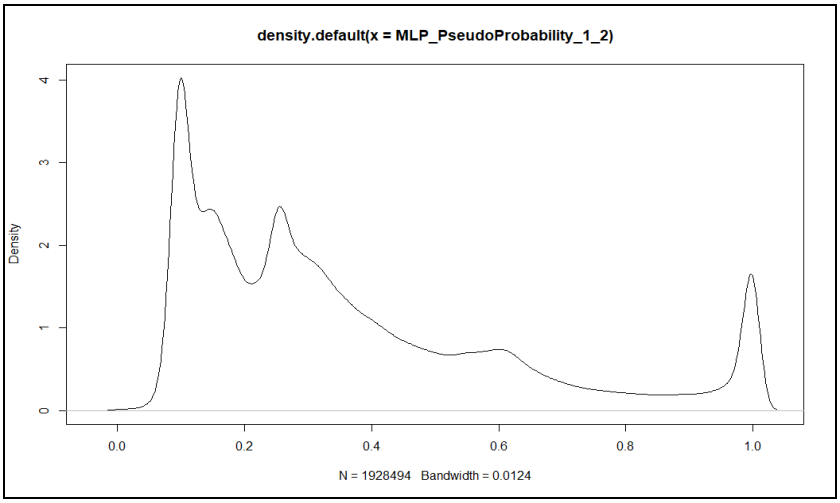


Figura 4-41

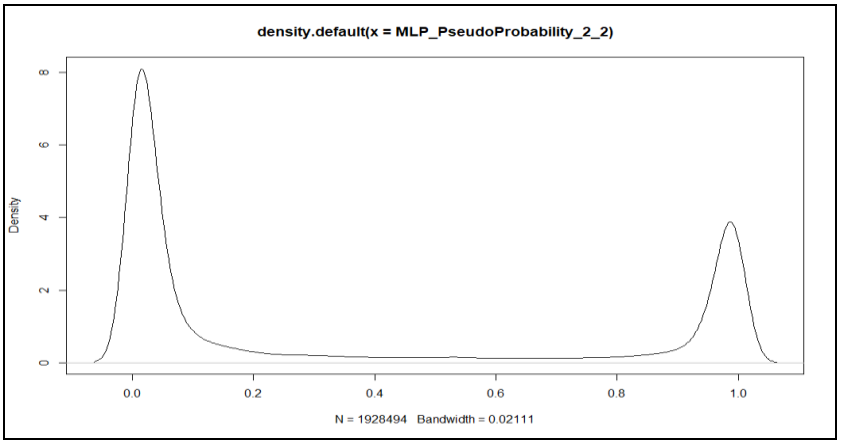


Figura 4-42

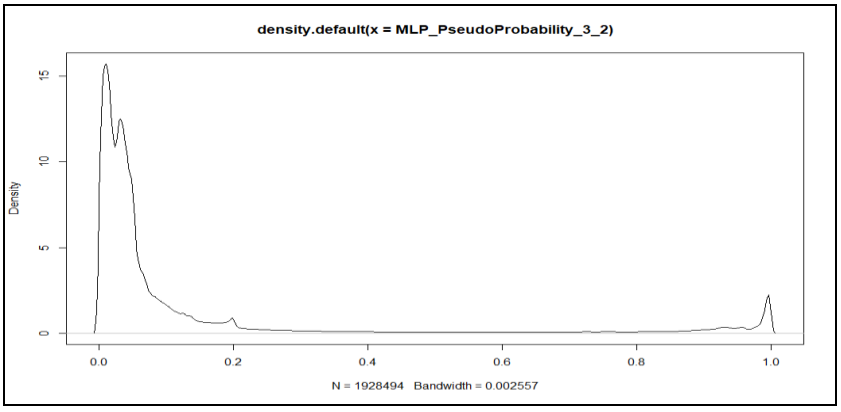


Figura 4-43

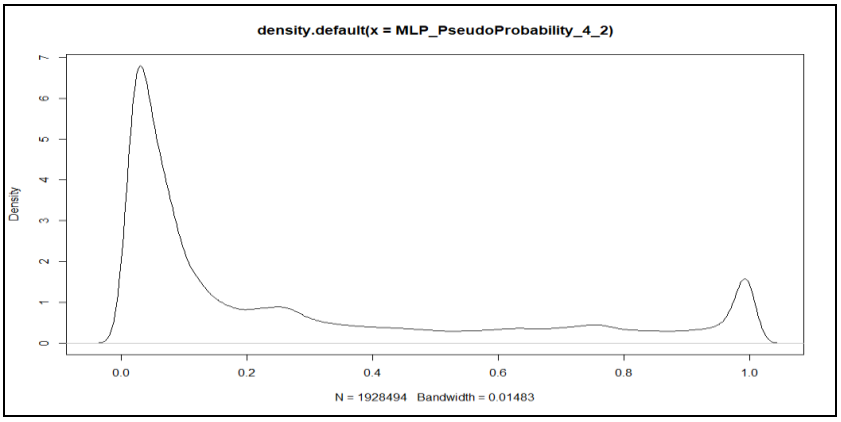


Figura 4-44

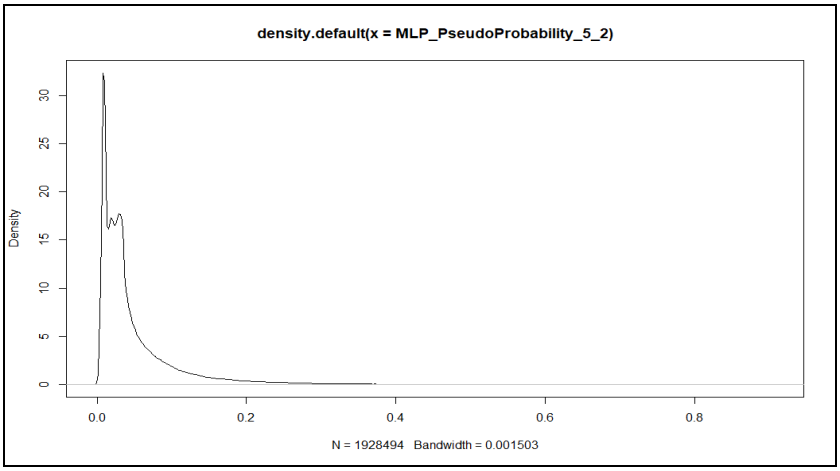


Figura 4-45

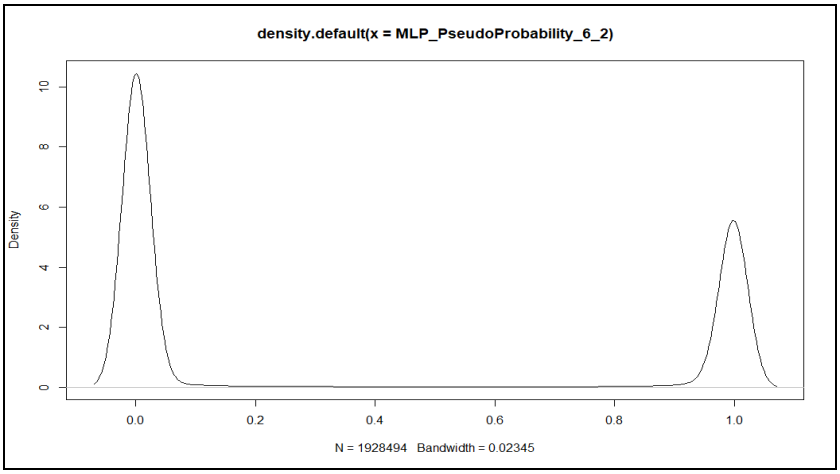


Figura 4-46

Se observa que la densidad de probabilidad para las diferentes causas de fraude es muy parecida. Particularmente las propensiones al fraude en IRPF por actividades económicas, planes de pensiones y rendimientos de capital inmobiliario se comportan de forma muy similar. Lo mismo ocurre con las propensiones al fraude en IRPF por número de hijos y desgravación de gastos. La propensión al fraude por tipo marginal sigue un patrón más aislado. Vemos así que los patrones de fraude no difieren mucho para las diferentes causas de fraude. Además, esta segmentación por patrones de

fraude de las distintas cuasas de fraude coincide con la ya obtenida mediante árboles de decisión, mediante modelos de análisis discriminante y utilizando escalamiento multidimensional.

Observamos que para probabilidades de fraude bajas se encuentra la mayor densidad de contribuyentes. Se observa también que para probabilidades de fraude altas repunta la densidad de contribuyentes. Por lo tanto vemos que la mayoría de contibuyentes no defraudan, pero los que defraudan tiene una propensión al fraude muy alta. Observamos también que para los diferentes tipos de fraude, los perfiles de fraude dados por sus funciones de densidad, nos llevan a la misma conclusión que para el fraude global.

Por otra parte, basándonos en la valoración importancia de las variables independientes sobre el fraude global que hemos obtenido al ajustar la red múltiple, podremos deducir qué variables son las más influyentes sobre el fraude global y su relación con las causas de fraude que acabamos de analizar.

Las reducciones de base imponible y planes de pensiones también resultan partidas influyentes en el fraude, que están directamente relacionadas con las causas de fraude por tipo marginal y por planes de pensiones.

En cuanto a la importancia de los predictores (variables independientes) sobre la propensión al fraude vemos que la variable más incidente en el fraude es la componente relativa a rendimientos, bases y cuotas. Le sigue la componente relativa a saldos patrimoniales, cuota diferencial y resultado de la declaración. Estas partidas están muy relacionadas con la alteración del tipo marginal correspondiente a las declaraciones, que era la primera causa de fraude. Los rendimientos ocultos no declarados llevan a la manipulación de las bases imponible y liquidable y por tanto a la minoración del tipo aplicable. Por lo tanto la cuota resultante

es inferior a la que debería de ser. También suele ser habitual la presencia de actividades cuyas rentas eluden la tributación, bien por no ser declaradas o bien por no estar registradas constituyendo economía sumergida. De esta forma, el tipo marginal correspondiente a la declaración resulta inferior al real, manipulándose así el resultado de la liquidación.

Las ganancias y pérdidas patrimoniales son variaciones en el valor del patrimonio del contribuyente. Su cálculo suele ser causa de fraude, bien sea por la mala aplicación de la norma general, por la mala aplicación de las normas específicas de valoración o por otras causas ya analizadas previamente que pueden incidir en el tipo marginal a aplicar. Además, este apartado está relacionado con el cálculo de la cuota diferencial, que se obtiene minorando la cuota líquida total del impuesto por las deducciones y retenciones que marca la ley. La cuota líquida suele ser también objeto de fraude ya que se suelen cometer incorrecciones al minorar la cuota íntegra por las inversiones en empresas de nueva o reciente creación, por las deducciones en actividades económicas, por las deducciones por donativos y otras aportaciones. Evidentemente, cualquier componente que influya en las cuotas, influirá en el resultado de la declaración.

Otros predictores influyentes en el fraude son los rendimientos de capital mobiliario e inmobiliario (rendimientos provenientes de la titularidad de bienes rústicos y urbanos o de derechos reales sobre ellos). La computación de estos rendimientos también suele ser fraudulenta. Lo mismo ocurre con la computación de los gastos deducibles y reducciones para obtener los rendimientos netos del capital inmobiliario. Suelen inflarse los gastos necesarios para la obtención de los rendimientos. Por lo tanto esta partida es influyente en las causas de fraude relativas a la alteración del tipo marginal y la incorrecta computación de gastos a deducir, que eran causas de fraude que ocupaban lugares importantes en la incidencia sobre el fraude.

Son partidas a vigilar también los mínimos por descendientes, ascendientes y discapacidad. Suelen ser habitual el fraude que afecta a las declaraciones del número de hijos y ascendientes y descendientes, que habitualmente eran simultáneamente desgravadas por los dos padres (separados, divorciados o en otras situaciones) en el caso de los hijos o por diferentes hermanos en el caso de los ascendientes. Hemos visto ya que la declaración errónea del número de hijos y ascendientes es también una de las causas de fraude destacadas.

Otra de las partidas influyentes en el fraude es la incorrecta declaración de deducciones derivadas de actividades económicas, lo mismo que los rendimientos íntegros de actividades económicas y los elementos patrimoniales afectos a la actividad económica. Estas partidas tienen que ver con la causa de fraude relativa a actividades económicas.

Otra partida a vigilar son los gastos deducibles totales y los límites de determinadas deducciones con especial referencia a las deducciones por incentivos a la inversión. El fraude en estas partidas está directamente relacionada con la cause de fraude relativa a la incorrecta deducción de gastos.

Vemos así la existencia de una correlación muy alta entre las partidas de la declaración más incidentes en el fraude y las causas generales de fraude.

Estos resultados son muy similares a los ya obtenidos en otras modelizaciones del fraude estudiadas anteriormente.

.

Capítulo 5

INVESTIGACIÓN DEL FRAUDE FISCAL CON TÉCNICAS DE MACHINE LEARNING. MODELOS LINEALES GENERALIZADOS Y REDES NEURONALES MÚLTIPLES

5.1 INTRODUCCIÓN

En este capítulo la finalidad esencial es estudiar el fraude fiscal en el Impuesto sobre la Renta de las Personas Físicas mediante el uso de técnicas de Machine Learning. Como en todas las técnicas presentadas anteriormente, se utilizará una metodología generalizable para cuantificar la propensión al fraude en cualquier otro impuesto según las causas que lo determinan, siempre y cuando se disponga de la base de datos correspondiente.

Se utilizarán las nuevas prestaciones que aportan las técnicas de Machine Learning aplicadas a impuestos para ampliar las posibilidades de análisis cuantitativo. Estas técnicas, junto con la Minería de datos, cuando se utilizan sobre plataformas Big Data son el corazón del Big Data Analytics. En este capítulo se trata de mostrar el uso de las técnicas de

Machine Learning relativas a Modelos Lineales Generalizados y Redes Neuronales Múltiples aplicadas a las muestras de IRPF del IEF con la finalidad de estudiar las variables más incidentes que afectan al fraude fiscal en este impuesto. Así mismo, se utilizarán modelos predictivos de aprendizaje automático para cuantificar la probabilidad que tiene cualquier contribuyente de ser defraudador.

Por otra parte, en este capítulo se utilizará una reducción de la dimensión mucho más amplia que engloba a todas las partidas del IRPF sin simplificación alguna. En el capítulo anterior se utilizaban las variables más importantes del IRPF y su reducción a 11 componentes principales. En este capítulo se utilizarán todas las partidas económicas de la base de datos de IRPF y su reducción a 65 componentes principales que ya fue estudiada en el capítulo 2.

5.2 MARCO METODOLÓGICO: LAS TÉCNICAS DE MACHINE LEARNING

El proceso de Machine Learning puede definirse como una tarea de descubrimiento de nuevas y significativas relaciones, patrones y tendencias al examinar grandes cantidades de datos. Las técnicas de Machine Learning persiguen el descubrimiento automático del conocimiento contenido en la información almacenada de modo ordenado en grandes bases de datos. Estas técnicas tienen como objetivo descubrir patrones, perfiles y tendencias a través del análisis de los datos utilizando técnicas estadísticas avanzadas de análisis multivariante de datos.

La meta es permitir al investigador-analista encontrar una solución útil al problema planteado a través de una mejor comprensión de los datos existentes. Las herramientas de Machine Learning siguen el ciclo del análisis del análisis de datos, que indica que cuando tenemos un problema, la primera tarea es comprenderlo bien y pensar en el conjunto de datos que

mejor podría ayudarnos. Comprender los datos está directamente asociado con comprender el problema. A continuación, será necesario preparar adecuadamente los datos para utilizar modelos u otras herramientas de análisis sobre los mismos con la finalidad de extraer el conocimiento. Estos modelos han de ser evaluados adecuadamente siguiendo las reglas técnicas de la Estadística y las Matemáticas para ser implementados finalmente. Por lo tanto, las fases del trabajo en el Machine Learning son muy similares a la Minería de Datos.

Según *Mathworks* en su publicación "*Statistics and Machine Learning Toolbox"*, las técnicas de Machine Learning se dividen en dos grandes grupos: Técnicas de aprendizaje supervisado y técnicas de aprendizaje no supervisado tal y como indica el esquema de la figura 5-1.

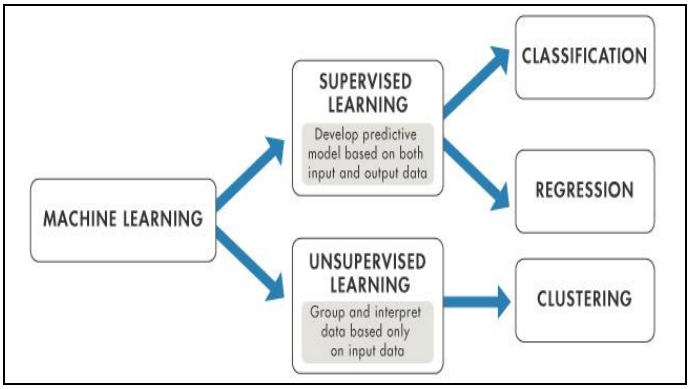


Figura 5-1

Dentro de las técnicas de aprendizaje supervisado se incluyen todas aquellas que desarrollan modelos predictivos basados en datos de entrada y datos de salida. Los datos de entrada vienen representados por las variables independientes del modelo y los datos de salida se representan por las variables dependientes. Incluiremos aquí todas las técnicas que conllevan modelos y cuya finalidad esencial es la predicción. Por lo tanto, este grupo de técnicas se corresponden con las técnicas predictivas o técnicas de la

dependencia de la Minería de Datos, ampliándolas con nuevas técnicas automáticas para la predicción.

Estas técnicas de aprendizaje supervisado se pueden subdividir a su vez en dos grupos: técnicas de regresión y técnicas de clasificación. Podemos incluir en las técnicas de regresión todos aquellos modelos que tienen variables dependientes cuantitativas (modelo lineal de regresión múltiple, modelo lineal generalizado GLM, Support Vector Machine Regression SVR, Gaussian Procces Regression GPR, árboles de decisión, redes neuronales y otras técnicas). Podemos incluir dentro de las técnicas de clasificación de análisis supervisado todos aquellos modelos que tienen variables dependientes categóricas (modelos de variable dependiente limitada como los modelos logit y probit, modelos lineales generalizados, modelos de Análisis Discriminante, Árboles de Decisión, Support Vector Machine Classification, Naive Bayes, Vecino más cercano kNN y redes neuronales para la clasificación. Estas técnicas supervisadas de Machinne Learning para la clasificación suelen determinarse técnicas de clasificación ad hoc, ya que clasifican a los elementos de la población en grupos previamente conocidos que se corresponden con las categorías de la variable dependiente.

Dentro de las técnicas de aprendizaje no supervisado se incluyen todas aquellas basados en datos sólo de entrada. Estas técnicas, a través de los datos de entrada, encuentra patrones de comportamiento en los datos, realizan perfilado, segmentación y otras tareas de obtención del conocimiento contenido en los datos. Por lo tanto, este grupo de técnicas se corresponden con las técnicas descriptivas o técnicas de la interdependencia de la Minería de Datos. Podemos incluir aquí las técnicas de reducción de la dimensión, el escalamiento multidimensional, las técnicas de Análisis Cluster jerárquico y no jerárquico, los modelos de Markov y otras técnicas de segmentación. Estas técnicas no supervisadas de Machinne Learning para

la clasificación suelen denominarse técnicas de clasificación post hoc, ya que clasifican a los elementos de la población en grupos que no se conocen hasta que finaliza la ejecución de la técnica.

El esquema de la figura 5-2 presenta la mayoría de las técnicas de Machine Learning, tanto de aprendizaje supervisado, como no superviasado.

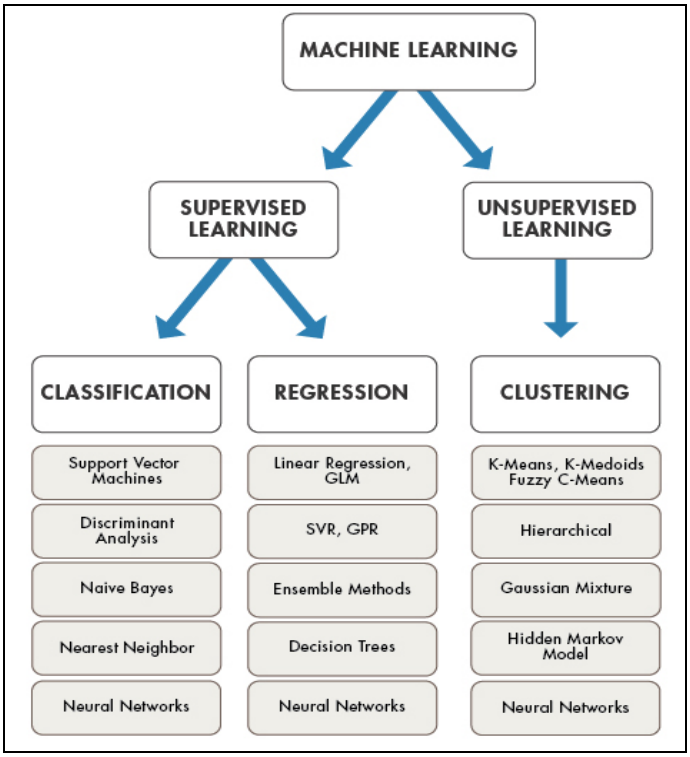


Figura 5-2

En este capítulo utilizaremos técnicas de clasificación de aprendizaje supervisado para cuantificar la propensión individual al fraude en el impuesto sobre la renta de las personas físicas, así como para cuantificar la incidencia de las distintas partidas del impuesto en el fraude fiscal en IRPF. Se cuantificará la probabilidad que tiene cualquier contribuyente de ser defraudador. De esta forma se segmentan los declarantes del impuesto por

nivel de propensión al fraude y causas del mismo. En concreto se utilizarán los Modelos Lineales Generalizados GLM y las Redes Neuronales Múltiples. En capítulos anteriores ya habíamos utilizado los Árboles de Decisión y el Análisis Discriminante, que pueden considerarse técnicas predictivas de Minería de Datos y a la vez técnicas de análisis supervisado de Machine Learning.

Asimismo, se utilizarán técnicas de clasificación de aprendizaje no supervisado para analizar la relación entre las distintas causas de fraude y para segmentar las partidas del impuesto según su incidencia en el fraude fiscal. Se utilizarán técnicas de clustering y de escalamiento multidimensional para realizar segmentaciones.

5.3 LOS DATOS: SELECCIÓN, EXPLORACIÓN Y TRANSFORMACIÓN DE LA INFORMACIÓN

Al igual que en el caso de todas las técnicas anteriores, se utiliza como fuente de datos la muestra del Impuesto sobre la Renta de las Personas Físicas (IRPF) que proporciona el Instituto de Estudios Fiscales (IEF). El origen fiscal de la muestra aporta, por tanto, unos datos de gran precisión, y en los que además no aparecen los problemas de infrarrepresentación y falta de respuesta habituales de las encuestas. Por consiguiente, la riqueza de estos datos permite realizar múltiples análisis incluidos los relativos a las técnicas de Machine Learning. Hay que seguir teniendo presente que disponer de una estructura de hardware y software que implemente procesamiento de grandes datos (*Big Data*) es esencial en nuestro caso, ya que las técnicas de Machine Learning que vamos a utilizar en este capítulo, se encuadran dentro de las técnicas de *Big Data Analytics*.

Recordamos que se seleccionan más de 2 millones de registros de la población total de declarantes mediante muestreo estratificado por provincias, tramos de renta y fuente de renta utilizando afijación

proporcional. Este método de selección expande adecuadamente la muestra por toda la población resultando muy significativa. Esta muestra alimentará nuestros modelos de Machine Learning ya que aporta cerca de 300 variables, tanto partidas económicas como otras variables numéricas y no numéricas, todas ellas contenidas como casillas en el modelo 100 de declaración del IRPF. La tabla 5-1 muestra las primeras partidas económicas de la muestra de IRPF. En el caso de los modelos de Machine Learning, estas partidas económicas serán candidatas inicialmente a ser las variables independientes métricas de los modelos.

Como variables dependientes de los modelos de Machine Learning se utilizarán las variables de fraude más comunes.

id	Numérico	8	2	Identificador del perceptor
f_tmg	Numérico	8	2	Fraude que afecta al tipo marginal
f_capinm	Numérico	8	2	Fraude que afecta a los rendimientos de capital inmobiliario
f_nhijos	Numérico	8	2	Fraude que afecta a las declaraciones de número e hijos y ascendieentes y descendientes
f_aaee	Numérico	8	2	Fraude que afecta a la declaración de actividaes económicas
f_planp	Numérico	8	2	Fraude que afecta a la desgrbación por planes de pensiones
f_gastos	Numérico	8	2	Fraude que afecta a la declarción de gastos
marca	Numérico	8	2	fraude global

Tabla 5-1

´ Como indica la figura anterior, los factores de fraude más comunes que aparecen definidos en la base de datos utilizada son los siguientes:

- La variable *marca* indica fraude global.
- La variable *f_tmg* indica fraude relativo al tipo marginal,
- La variable *f_capinm* indica fraude relativo a los rendimientos de capital inmobiliario.
- La variable *f_aaee* indica fraude relativo a la declaración de actividades económicas.
- La variable *f_gastos* indica fraude relativo a las deducciones de gastos.

- La variable *f_planp* indica fraude relativo a las declaraciones de planes de pensiones.
- La variable *f_nhijos* indica fraude relativo a la declaración del número de hijos y ascendientes.

En cuanto a la transformación de la información, en nuestro caso tenemos en los modelos cerca de 300 variables independientes cuantitativas muy correladas entre sí que provocarían un problema de multicolinealidad en el modelo a estimar. Por lo tanto será necesario reducir estas variables a sus componentes principales, que están incorreladas. Al ajustar el modelo en las componentes se elimina el problema de la multicolinealidad.

En cuanto a la presencia de datos desaparecidos (missing), hay que tener presente que la Agencia Tributaria proporciona la base de datos ya imputada.

Por otro lado, los modelos de Machine Learning se optimizan para regresores normales y poco correlados entre sí para garantizar la ausencia de multicolinealidad. En nuestro caso, sabemos que las variables de renta no son normales y que suelen seguir una distribución paretiana truncada. Este problema se solventa utilizando como variables regresoras los factores resultantes de aplicar un análisis de componentes principales con rotación ortogonal varimax sobre las variables independientes iniciales. Dada la cantidad de variables, la cantidad de factores y el tamaño de la muestra, puede presuponerse la convergencia a la normalidad de los factores por aplicación del teorema central del límite. Además, el uso de factores refuerza la confidencialidad de las variables con más incidencia en el fraude fiscal, ya que son combinaciones lineales de las variables iniciales.

En cuanto a los datos atípicos (outliers), es necesario detectar su existencia en cada una de las variables consideradas por separado. Para la

detección de casos atípicos multivariantes podría recurrirse al cálculo de la distancia de Mahalanobis de cada individuo respecto al centro del grupo o a un método gráfico. En nuestro caso, el uso de factores que engloban varias variables iniciales, minimiza el efecto de los valores atípicos.

La matriz de correlaciones de las variables suele utilizarse para detectar la multicolinealidad (variables con correlación muy alta pueden ser redundantes), que puede ser muy nociva en la estimación del modelo logístico. En nuestro caso, estos problemas también desaparecen al utilizar factores en vez de variables iniciales como variables independientes de los modelos de Machine learning. El problema de la multicolinealidad queda perfectamente resuelto con la utilización de los factores.

Por lo tanto transformaremos las variables iniciales a utilizar como independientes en los modelos (partidas económicas de la base de datos de IRPF) mediante reducción de la dimensión aplicando componentes principales.

Las puntuaciones de las componentes son las variables reducidas incorreladas que sustituirán a las variables iniciales (tabla 5-2) que constituyen las partidas esconómicas del IRPF y que están muy correladas y por ello no pueden ser variables independientes de ningún modelo por el problema de la multicolinealidad.

A partir de ahora, los modelos tendrán como variables independientes las puntuaciones de las componentes C1, C2, …. para evitar la multicolinealidad. Parte de las puntuaciones de las componentes obtenidas se presentan en la tabla 5-2.

FAC1_1	FAC2_1	FAC3_1	FAC4_1	FAC5_1	FAC6_1	FAC7_1	FAC8_1	FAC9_1	FAC10_1
-,28733	,02082	,06897	-,00307	,41413	-,88081	,23402	,07471	-,00050	2,92372
-,32042	,00424	,06667	-,00395	-,21645	-,90257	,21766	-,00792	-,08000	2,29451
-,24166	-,01114	,00228	-,00183	-,09887	-,32273	,05859	,00086	-,00143	-,90528
-,24364	-,01063	-,00346	-,00177	-,09721	-,29709	,05883	-,00249	,00278	-,97691
-,25142	-,00848	,01474	-,00218	-,11037	-,43155	,08886	,00531	-,01596	-,37856
-,24370	-,01137	,01199	-,00201	-,10834	-,37403	,06858	,00273	-,01158	-,62209
-,24881	-,00826	-,01222	-,00215	-,16421	-,17482	,12582	-,01412	,01986	,86438
-,18107	-,03787	,01079	-,00196	-,17678	-,49507	,05854	-,01819	-,03597	,70086
-,29364	,10862	-,08900	-,00267	,18870	-,34214	,14316	-,02691	-,03847	-,15804
-,07748	-,02901	-,03972	-,00294	-,16371	-,67197	,04466	-,02207	-,04326	1,56624
,19409	,02077	-,11588	-,00275	-,13130	-,54946	-,14866	-,01053	,02726	,43033
,42319	-,08268	-,08273	-,00254	-,19831	1,42078	-,19330	-,04591	-,07591	,91602
-,18636	-,02137	,01242	-,00530	-,16369	,80438	,16216	-,00311	-,03867	,63321
-,43981	,07769	,12848	-,00413	,08397	-1,11599	,49600	,10589	-,09213	-1,52525
-,27315	,00531	,00732	,00015	-,08598	-,28397	,06995	,09276	,02176	-,87602
-,26406	-,00790	,02342	-,00244	-,14630	-,53585	,12427	-,00392	-,03470	,56151
,21007	-,06666	-,04748	-,00357	-,22155	2,59063	-,05183	-,03047	-,07008	-,14517
,17008	,05202	,06008	,00186	,17625	1,47654	,10388	,02869	,05567	,22440

Tabla 5-2

Por cuestiones de comodidad y facilidad de escritura e interpretación, en lo que sigue sustituiremos los nombres de las componentes FAC1_1, FAC2_1,… por C1, C2,…

5.4 MODELOS DE APRENDIZAJE SUPERVISADO: MODELO LINEAL GENERALIZADO

El modelo lineal generalizado amplía el modelo lineal de regresión múltiple $E[y_i] = x_i'\beta$, de manera que la variable dependiente y está relacionada linealmente con las variables independientes mediante una determinada *función de enlace g*, de modo que: $E[y_i] = g^{-1}(x_i'\beta)$. Si μ_i = E[y_i] entonces $\eta_i = g(u_i) = x_i'\beta$. El otro elemento característico de los modelos lineales generalizados es la *familia de probabilidades* de la variable dependiente. Cuando la función de enlace es la identidad y la familia de probabilidades es la normal, estamos ante el modelo lineal de regresión múltiple como caso particular de los modelos lineales generalizados.

Pero en general, la posibilidad de especificar una distribución específica para la variable dependiente que no sea la normal y la posibilidad de especificar

una función de enlace que no sea la identidad, es la principal mejora que aporta el modelo lineal generalizado respecto al modelo lineal general.

Ya sabemos que los ***modelos de elección discreta*** predicen directamente la probabilidad de un suceso que tienen dos o más posibilidades de ocurrencia. Cuando el suceso tiene solo dos posibilidades de ocurrencia mutuamente excluyentes estamos ante una distribución binomial. Como los valores de una probabilidad están entre cero y uno, las predicciones realizadas con los modelos de elección discreta deben estar acotadas para que caigan en el rango entre cero y uno. El modelo general que cumple esta condición tiene la forma funcional:

$$P = F(X, \beta) + u$$

Se observa que si F es la función de distribución de una variable aleatoria, entonces P varía entre cero y uno.

En el caso particular en que la función F es la función logística estaremos ante el ***modelo Logit o Regresión Logística***, cuya forma funcional será la siguiente:

$$P = F(X, \beta) + u = \frac{e^{X\beta}}{1 + e^{X\beta}} + u$$

Se observa que:

$$P = F(X, \beta) = \frac{e^{X\beta}}{1 + e^{X\beta}} \Rightarrow \log\left(\frac{P}{1 - P}\right) = X\beta$$

Por lo tanto, el modelo logit también se puede expresar en la forma:

$$\log\left(\frac{P}{1-P}\right) = X\beta + u$$

Se trata de un modelo lineal generalizado donde la función de enlace resulta ser $\log\left(\frac{P}{1-P}\right)$ que se denomina ***función de enlace logit*** y la distribución de probabilidad pertenece a la ***familia binomial***.

En el caso particular en que la función *F* es la función de distribución de una normal unitaria estaremos ante el ***modelo Probit***, cuya forma funcional será la siguiente:

$$P = F(X, \beta) + u = (2\pi)^{-\frac{1}{2}} \int_{-\infty}^{X\beta} e^{-\frac{t^2}{2}} dt + u$$

Tenemos:

$$P = (2\pi)^{-\frac{1}{2}} \int_{-\infty}^{X\beta} e^{-\frac{t^2}{2}} dt = \Phi(X\beta) \Rightarrow \Phi^{-1}(P) = X\beta$$

Por lo tanto, el modelo logit también se puede expresar en la forma:

$$\Phi^{-1}(P) = X\beta + u$$

siendo Φ la función de distribución de una normal (0,1).

Por lo tanto, el modelo probit es un modelo lineal generalizado cuya función de enlace resulta ser $\Phi^{-1}(P)$ que se denomina *función de enlace probit* y la distribución de probabilidad también pertenece a la familia binomial.

Los modelos lineales generalizados son recientes, están muy elaborados matemáticamente y son muy eficientes para computar grandes cantidades de datos. Por esta razón se utilizarán aquí para estimar las regresiones Logit y Probit, como alternativa a las estimaciones clásicas de los modelos de variable dependiente limitada.

En la aplicación que aquí se presenta, se utiliza la muestra de IRPF que proporciona el Instituto de Estudios Fiscales, para considerar modelos logit y probit tomado como variable dependiente una variable dicotómica que toma el valor 1 si el individuo defrauda y el valor cero si el individuo no defrauda (variable *marca*). Las variables independientes son los factores resultantes de la reducción de la dimensión previamente realizada.

Se utiliza una base de datos totalmente anonimizada. En la práctica, serían defraudadores los individuos de la muestra que la inspección ha determinado fehacientemente como tales defraudadores.

Esta investigación es independiente del ejercicio de datos que se considere, ya que se busca una metodología para predecir la probabilidad de fraude de cada contribuyente en IRPF. Como además la metodología se basa en un modelo predictivo, se obtiene una función de predicción de fraude que es válida para varios ejercicios futuros consecutivos.

No es necesario estimar el modelo que predice el fraude todos los años. Bastará con hacerlo cuando haya cambios legislativos apreciables en el impuesto o cambios significativos en la coyuntura económica.

5.4.1 Estimación del modelo logit y resultados

Realizada la estimación del modelo logístico con las componentes como regresores, se obtienen resultados óptimos. En primer lugar observamos la salida de la tabla 5-3 que indica que la estimación del modelo

logit se ha realizado utilizando modelos lineales generalizados mediante una distribución de probabilidad Binomial y una función de enlace Logit. Observamos también que se han procesado 1.928.494 declaraciones de IRPF con los porcentajes de fraude global que se indican en la información de la variable categórica.

Modelos lineales generalizados

Información de modelo

Variable dependiente	fraude global[a]
Distribución de probabilidad	Binomial
Función de enlace	Logit

a. El procedimiento modela 0 como la respuesta, tratando a 1 como la categoría de referencia.

Resumen de procesamiento de casos

	N	Porcentaje
Incluido	1928494	100,0%
Excluido	0	0,0%
Total	1928494	100,0%

Tabla 5-3

A continuación se observan medidas de la bondad de ajuste a través de los estadísticos de la cantidad de información del modelo (tabla 5-4). Para un modelo aislado estas medidas no tienen mucho interés. Su utilidad radica en la comparación de modelos alternativos.

Un modelo es mejor que otro, en las mismas condiciones de estimación, cuando sus valores de los estadísticos de información son menores.

Bondad de ajuste[a]

	Valor	gl	Valor/gl
Desvianza	841325,237	1910144	,440
Desvianza escalada	841325,237	1910144	
Chi-cuadrado de Pearson	2,996E+215	1910144	1,568E+209
Chi-cuadrado de Pearson escalado	2,996E+215	1910144	
Logaritmo de verosimilitud[b]	-420664,515		
Criterio de información Akaike (AIC)	841461,029		
AIC corregido para muestras finitas (AICC)	841461,034		
Criterio de información bayesiana (BIC)	842284,198		
AIC consistente (CAIC)	842350,198		

Variable dependiente: fraude global

Tabla 5-4

La tabla siguiente muestra el contraste Omnibus para la significatividad global del modelo estimado. Dado que el p-valor del contraste es mucho menor que 0,05, se acepta la significatividad conjunta del modelo estimado al 95 % de confianza.

Contraste Omnibus[a]

Chi-cuadrado de razón de verosimilitud	gl	Sig.
1707351,063	65	,000

Variable dependiente: fraude global

Tabla 5-5

Para contrastar la significatividad individual de los parámetros estimados observamos los p-valores del contraste de la Chi-cuadrado de Wald (columna Sig. de la tabla 5-6). Vemos que los parámetros estimados tienen una significatividad muy alta salvo raras excepciones ya que los p-valores son prácticamente nulos. Esto siempre ocurre cuando las variables independientes del modelo son fruto de una reducción de la dimensión.

Los errores estándar de las estimaciones son muy bajos y los intervalos de confianza al 95% para los parámetros estimados son muy estrechos, lo que indica una alta fiabilidad de las estimaciones.

La dificultad de interpretar los parámetros estimados del modelo logit en términos de efectos sobre la variable dependiente, nos lleva a utilizar los odds-ratio. La columna Exp(B) presenta los odds-ratio estimados, que se interpretan como la razón de ventajas de la probabilidad de que exista fraude respecto de que no exista.

$$Odds\text{-}ratio = P(fraude=1)/P(fraude=0).$$

Odds-ratio mayores que la unidad indican mejor significatividad de la estimación de fraude respecto de no fraude. Aquellos parámetros estimados con Odds-ratio superior a la unidad son más significativos y su efecto es mayor sobre la variable dependiente. De esta manera podemos cuantificar el efecto de cada variable independiente sobre el fraude. A mayor odds-ratio mayor efecto y odds-ratio superiores a la unidad indican efecto importante sobre el fraude. También se obtienen intervalos de confianza al 95% para los odds-ratio (muy estrechos).

Estimaciones de parámetro

Parámetro	B	Error estándar	95% de intervalo de confianza de Wald		Contraste de hipótesis			Exp(B)	95% de intervalo de confianza de Wald para Exp(B)	
			Inferior	Superior	Chi-cuadrado de Wald	gl	Sig.		Inferior	Superior
(Intercepción)	-17,852	,0577	-17,965	-17,739	95719,749	1	,000	1,766E-8	1,577E-8	1,978E-8
FAC1	-52,542	,3602	-53,248	-51,836	21278,209	1	,000	1,517E-23	7,490E-24	3,074E-23
FAC2	2,149	,0351	2,081	2,218	3746,482	1	,000	8,579	8,009	9,190
FAC3	2,607	,0329	2,543	2,672	6300,172	1	,000	13,564	12,718	14,466
FAC4	-6,882	,0947	-7,068	-6,697	5280,867	1	,000	,001	,001	,001
FAC5	-13,712	,0838	-13,876	-13,548	26775,386	1	,000	1,109E-6	9,412E-7	1,307E-6
FAC6	-1,886	,0202	-1,925	-1,846	8735,325	1	,000	,152	,146	,158
FAC7	-51,873	,5655	-52,981	-50,765	8414,242	1	,000	2,964E-23	9,783E-24	8,978E-23
FAC8	,752	,0197	,714	,791	1464,910	1	,000	2,122	2,042	2,206
FAC9	,764	,0505	,665	,863	228,842	1	,000	2,148	1,945	2,371
FAC10	-1,649	,0231	-1,694	-1,604	5104,566	1	,000	,192	,184	,201
FAC11	-57,488	,7830	-59,022	-55,953	5390,248	1	,000	1,080E-25	2,328E-26	5,012E-25
FAC12	2,829	,0115	2,806	2,851	60730,504	1	,000	16,924	16,547	17,309
FAC13	-10,109	,0600	-10,227	-9,991	28368,312	1	,000	4,071E-5	3,620E-5	4,580E-5
FAC14	-1,845	,0073	-1,860	-1,831	63733,596	1	,000	,158	,156	,160
FAC15	-57,980	1,9809	-61,863	-54,098	856,689	1	,000	6,598E-26	1,359E-27	3,203E-24
FAC16	-,894	,0243	-,942	-,847	1356,163	1	,000	,409	,390	,429
FAC17	-1,582	,0526	-1,685	-1,478	902,992	1	,000	,206	,186	,228
FAC18	937,699	3,4145	931,006	944,391	75418,691	1	,000	.[a]	.[a]	.[a]
FAC19	-10,285	,0465	-10,376	-10,194	48885,291	1	,000	3,413E-5	3,116E-5	3,739E-5

FAC20	,087	,0080	,071	,103	116,795	1	,000	1,091	1,074	1,108
FAC21	-,243	,0051	-,253	-,233	2249,132	1	,000	,784	,776	,792
FAC22	-,206	,0080	-,221	-,190	655,012	1	,000	,814	,801	,827
FAC23	,102	,0027	,097	,107	1469,676	1	,000	1,107	1,101	1,113
FAC24	1,844	,0266	1,792	1,896	4796,688	1	,000	6,323	6,002	6,662
FAC25	-28,097	,0925	-28,278	-27,915	92250,411	1	,000	6,278E-13	5,237E-13	7,525E-13
FAC26	-,379	,0098	-,399	-,360	1506,986	1	,000	,684	,671	,698
FAC27	-1,101	,2227	-1,537	-,665	24,445	1	,000	,333	,215	,515
FAC28	,105	,0030	,099	,111	1228,690	1	,000	1,111	1,104	1,118
FAC29	-12,356	,0647	-12,483	-12,230	36519,420	1	,000	4,302E-6	3,790E-6	4,883E-6
FAC30	-137,066	1,8148	-140,623	-133,509	5704,297	1	,000	2,970E-60	8,473E-62	1,041E-58
FAC31	-3,065	,0398	-3,143	-2,987	5939,736	1	,000	,047	,043	,050
FAC32	-2,361	,0155	-2,391	-2,330	23157,495	1	,000	,094	,092	,097
FAC33	-8,809	,0503	-8,908	-8,711	30713,217	1	,000	,000	,000	,000
FAC34	5,815	,0310	5,754	5,875	35109,416	1	,000	335,171	315,393	356,190
FAC35	-8,129	,3411	-8,797	-7,460	567,894	1	,000	,000	,000	,001
FAC36	,142	,0248	,093	,190	32,722	1	,000	1,152	1,098	1,210
FAC37	-2,112	,0304	-2,172	-2,052	4811,492	1	,000	,121	,114	,128
FAC38	-,155	,0039	-,163	-,147	1568,394	1	,000	,856	,850	,863
FAC39	-8,715	,1916	-9,091	-8,340	2068,091	1	,000	,000	,000	,000
FAC40	-5,040	,0529	-5,144	-4,937	9065,759	1	,000	,006	,006	,007
FAC41	,795	,0190	,757	,832	1756,017	1	,000	2,214	2,133	2,297
FAC42	-1,013	,0268	-1,066	-,961	1434,012	1	,000	,363	,345	,383
FAC43	-,203	,0332	-,268	-,138	37,376	1	,000	,816	,765	,871
FAC44	-,737	,0193	-,775	-,699	1459,001	1	,000	,479	,461	,497
FAC45	-,227	,0113	-,249	-,205	403,904	1	,000	,797	,780	,815
FAC46	-2,422	,0091	-2,440	-2,405	70525,498	1	,000	,089	,087	,090
FAC47	-13,991	,0725	-14,133	-13,849	37215,000	1	,000	8,388E-7	7,277E-7	9,670E-7
FAC48	-2,899	,0336	-2,965	-2,833	7440,226	1	,000	,055	,052	,059
FAC49	-1,358	,0146	-1,386	-1,329	8593,632	1	,000	,257	,250	,265
FAC50	,663	,0084	,647	,680	6197,623	1	,000	1,941	1,909	1,973
FAC51	-3,632	,8323	-5,263	-2,000	19,041	1	,000	,026	,005	,135
FAC52	-1,054	,0263	-1,106	-1,003	1612,343	1	,000	,348	,331	,367
FAC53	,009	,0251	-,040	,058	,132	1	,716	1,009	,961	1,060
FAC54	,321	,0104	,301	,341	949,615	1	,000	1,379	1,351	1,407
FAC55	,555	,0043	,546	,563	16856,874	1	,000	1,741	1,727	1,756
FAC56	-2,764	,1822	-3,121	-2,407	230,258	1	,000	,063	,044	,090
FAC57	,008	,0313	-,053	,069	,065	1	,799	1,008	,948	1,072
FAC58	,229	,0116	,206	,252	390,527	1	,000	1,257	1,229	1,286
FAC59	4,442	,0513	4,341	4,542	7492,908	1	,000	84,906	76,783	93,889
FAC60	,941	,0243	,894	,989	1496,748	1	,000	2,563	2,444	2,689
FAC61	,000	,0150	-,030	,029	,001	1	,974	1,000	,970	1,029
FAC62	,214	,0111	,192	,236	370,242	1	,000	1,238	1,212	1,266
FAC63	2,673	,0236	2,627	2,720	12787,009	1	,000	14,487	13,831	15,174
FAC64	2,652	,1267	2,403	2,900	437,843	1	,000	14,177	11,059	18,174
FAC65	,229	,0045	,220	,238	2595,548	1	,000	1,257	1,246	1,268
(Escala)	1[b]									

Tabla 5-6

El modelo Logit estimado, que predice probabilidades de fraude es el siguiente:

$$P(\mathit{fraude}) = \frac{1}{1+e^{-(-17{,}852-52{,}542FAC1+2{,}149FAC2+2{,}607FAC3+\cdots+2{,}652CFAC64+0{,}229FAC65)}}$$

Hay que tener en cuenta que esta función de predicción de probabilidad de fraude es válida para los individuos de la muestra y para los que no están en la muestra. Además, también será válida para cualquier individuo futuro que realice su declaración de la renta.

En cuanto a los principales efectos de cada partida del impuesto sobre el fraude (efectos relativos computados hasta 5 milésimas), podemos realizar la ordenación de la tabla 5-7.

Parámetro	**Efecto**
Mínimos personales y familiares y deducciones por rendimientos del trabajo y actividades económicas	0,172
Deducciones vivienda habitual	0,122
Base general y cuotas	0,091
Reducciones de la base imponible por aportaciones y contribuciones a sistemas de previsión social y planes de pensiones	0,082
Rentas inmobiliarias, cuota diferencial y resultado declaración	0,075
Actividades económicas -rendimientos, retenciones y pagos a cuenta	0,062
Reducciones de la base imponible por tributación conjunta y circunstancias personales y familiares	0,054
Gastos en defensa jurídica, colegios profesionales y sindicatos	0,048
Capital inmobiliario	0,040
Gastos deducibles y cotizaciones	0,032
Ganancias agrarias y mínimo por ascendientes	0,031
Retribuciones en especie	0,028
Rendimientos capital mobiliario, letras del tesoro, ganancias patrimoniales y gastos deducibles	0,017
Atribuciones de rentas (rendimientos de capital inmobiliario)	0,013
Deducciones por doble imposición	0,010
Reducción por obtención de rendimientos del trabajo (mayores de 65 años)	0,009
Arrendamiento de inmuebles y gastos fiscales deducibles	0,007
Rendimientos asistencia técnica y capital mobiliario	0,006
Rentas exentas de IRPF	0,005

Tabla 5-7

Al analizar la tabla anterior de efectos sobre el fraude de las diferentes partidas del Impuesto sobre la Renta, observamos el papel preponderante de las deducciones por rendimientos del trabajo y actividades económicas, actividades económicas (retenciones y pagos a cuenta), base general y cuotas, reducciones de la base imponible por aportaciones y contribuciones a sistemas de previsión social y planes de pensiones, Rentas y capital inmobiliario, cuota diferencial y resultado declaración, gastos deducibes y cotizaciones y otros efectos.

Se observa que estos efectos sobre el fraude de las diferentes partidas del Impuesto sobre la Renta están muy en línea con las causas generales de fraude, en concreto con el fraude en el tipo marginal, el fraude en actividades económicas, el fraude en la declaración de gastos, el fraude en los planes de pensiones y el fraude relativo al capital inmobiliario.

En esta ocasión se incorporan como partidas incidentes el fraude las retribuciones en especie, la deducciones por doble imposición, los rendimientos del trabajo de mayores de 65 años, los arrendamientos de inmuebles y gastos fiscales deducibles y las rentas exentas de IRPF.

En cuanto a la diagnosis del modelo, la matriz de correlaciones de los estimadores de los parámetros presenta valores muy bajos, lo que certifica la significatividad de la estimación.

A continuación se muestra la tabla 5-8 que presenta los estadísticos Leverage y Distancia de Cook para evaluar las observaciones influyentes en el modelo.

Se observa que todos estos valores son muy pequeños, lo que indica que no hay observaciones influyentes en el modelo.

También se observan en la tabla las probabilidades de fraude estimadas para cada contribuyente y la categoría de la variable dependiente en la que se clasificarían (fraude=1 o no fraude=0)

Leverage	Residual	CooksDistance	Probabilidadfraude	Categoríapredicha
,000	-,002	,000	,99844	1
,000	-,091	,000	,90907	1
,000	,101	,000	,10062	0
,000	,136	,000	,13614	0
,000	,130	,000	,13021	0
,000	,049	,000	,04911	0
,000	-,350	,000	,64968	1
,000	-,330	,000	,67036	1
,000	,000	,000	1,00000	1
,000	-,005	,000	,99469	1
,000	,000	,000	1,00000	1
,000	,000	,000	,99995	1
,000	-,077	,000	,92321	1
,002	,964	,001	,96411	1
,000	-,750	,000	,24997	0
,000	,151	,000	,15062	0
,000	,000	,000	1,00000	1
,000	-,035	,000	,96526	1
,000	-,038	,000	,96205	1
,000	,053	,000	,05319	0
,000	-,749	,000	,25133	0

Tabla 5-8

En la tabla 5-8 también se muestran los residuos del modelo, cuyo análisis se presenta a continuación.

Los estadísticos descriptivos de los residuos muestran que la esperanza residual es nula prácticamente, lo mismo que la mediana (tabla 5-9). Como media y mediana son muy similares, se deduce la simetría de los residuos.

Por otra parte, el teorema de normalidad débil asegura que si los coeficientes de asimetría y curtosis de los residuos varían en el intervalo (-2,2), dichos residuos pueden considerarse normales.

En nuestro caso la asimetría cumple la condición y la curtosis se sale ligeramente del intervalo por la parte superior, indicando que los residuos son un poco más apuntados que la campana de Gauss. Este no es un problema importante para la normalidad.

Quizá este hecho indique la presencia de algún valor atípico residual no demasiado significativo.

Asimismo, la media recortada al 95% difiere levemente de la media, lo que indica la presencia de atípicos no muy significativos.

Lo mismo indican los Estimadores-M, que son estadísticos robustos de centralización para los residuos. Los Estimadores-M calculan la media ponderada residual utilizando diferentes sistemas de ponderación que ponderan más los valores centrales de la distribución residual y menos los valores extremos de las dos colas. De esta forma minoran el efecto de los valores extremos sobre la media sin prescindir de ningún valor residual. En nuestro caso hay ligeras diferencias entre los estimadores-M, pero todos son pequeños y bastante parecidos en valor, lo que indica que el efecto de los atípicos residuales es pequeño.

Descriptivos

			Estadístico	Error estándar
Residuo bruto	Media		-,00007	,000194
	95% de intervalo de confianza para la media	Límite inferior	-,00045	
		Límite superior	,00031	
	Media recortada al 5%		,01524	
	Mediana		,00000	
	Varianza		,072	
	Desviación estándar		,268767	
	Mínimo		-1,000	
	Máximo		1,000	
	Rango		2,000	
	Rango intercuartil		,120	
	Asimetría		-1,172	,002
	Curtosis		3,382	,004

Estimadores-M

	Estimador-M de Huber[a]	Biponderado de Tukey[b]	Estimador-M de Hampel[c]	Onda de Andrews[d]
Residuo bruto	,02898	,03331	,04244	,03321

Tabla 5-9

Si observamos el gráfico Q-Q de normalidad residual (figura 5-3), vemos que su nube de puntos se ajusta bastante bien a la diagonal, lo que indica presencia de normalidad en los residuos.

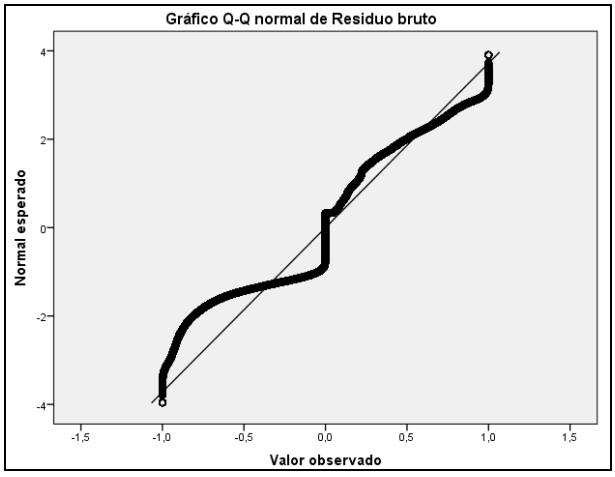


Figura 5-3

También podemos evaluar el ajuste del modelo Logit a través de la curva ROC. El área bajo la curva ROC (tabla 5-10) y por encima de la diagonal del primer cuadrante evalúa el modelo. El valor óptimo de esta área es el uno, caso en que la curva ROC (Figura 5-4) y la diagonal del primer cuadrante forman un triángulo rectángulo. En nuestro caso este área vale 0,959 valor muy cercano a la unidad, lo que indica que el ajuste del modelo Logit es óptimo.

Área bajo la curva

Variable(s) de re:
Área
,959

Tabla 5-10

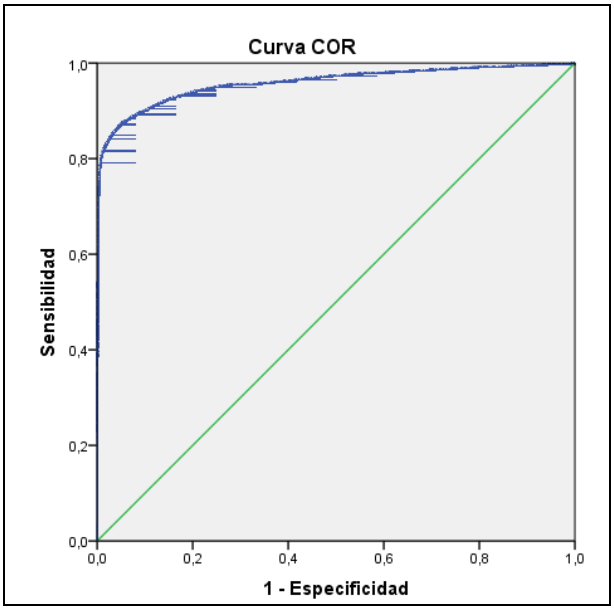


Figura 5-4

La tabla 5-11 siguiente muestra el cálculo de las probabilidades de fraude de cada individuos (propensiones al fraude) según un modelo Logit.

PredictedVal...	Leverage	Residual	CooksDistan...	ProbFraudeLogit
1	,000	,000	,000	1,00000
1	,000	-,004	,000	,99603
0	,000	,080	,000	,08005
0	,000	,097	,000	,09686
0	,000	,212	,000	,21176
0	,000	,043	,000	,04274
1	,000	-,254	,000	,74635
1	,001	-,424	,000	,57618
1	,000	,000	,000	1,00000
1	,000	,000	,000	,99953

Tabla 5-11

Si segmentamos mediante escalamiento multidimensional las causas de fraude junto con la probabilidad de fraude, obtenemos el mapa perceptual que se muestra en la Figura 5-5).

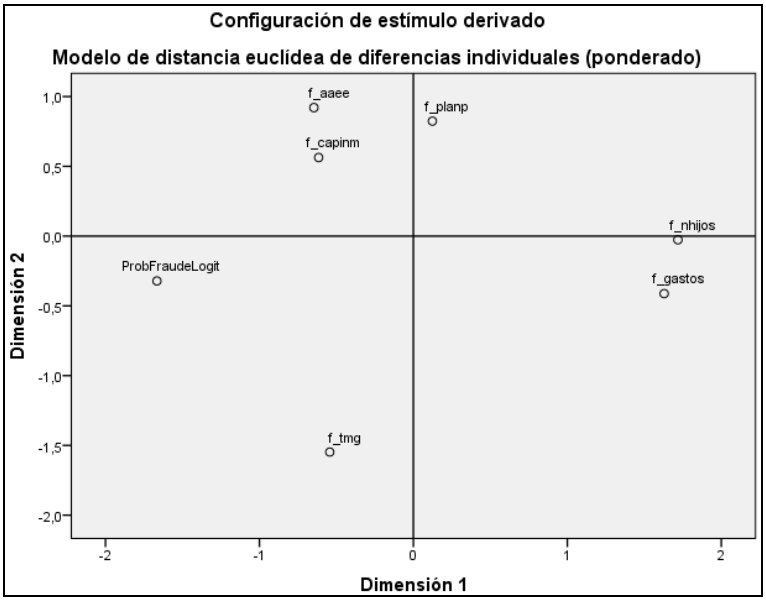


Figura 5-5

Se observa que el fraude en actividades económicas, en capital inmobiliario y en planes de pensiones influyen de modo similar en la probabilidad de fraude. De la misma forma, el fraude en deducción de

gastos y declaración del número de hijos y ascendientes tienen parecida incidencia sobre la probabilidad de fraude, aunque el valor de esa incidencia es menor que en el caso del grupo anterior (más lejanía en el mapa perceptual). El fraude en tipo marginal incide de modo aislado sobre la probabilidad de fraude.

La calidad de la segmentación por escalamiento multidimensional es alta dado el valor bajo de estadístico S-stress y el valor alto del estadístico RSQ.

```
Stress  =   ,08499      RSQ =  ,9594
```

Si realizamos la segmentación anterior mediante análisis clúster por variables jerárquico utilizando el método de Ward, obtenemos el dendograma de la figura 5-6.

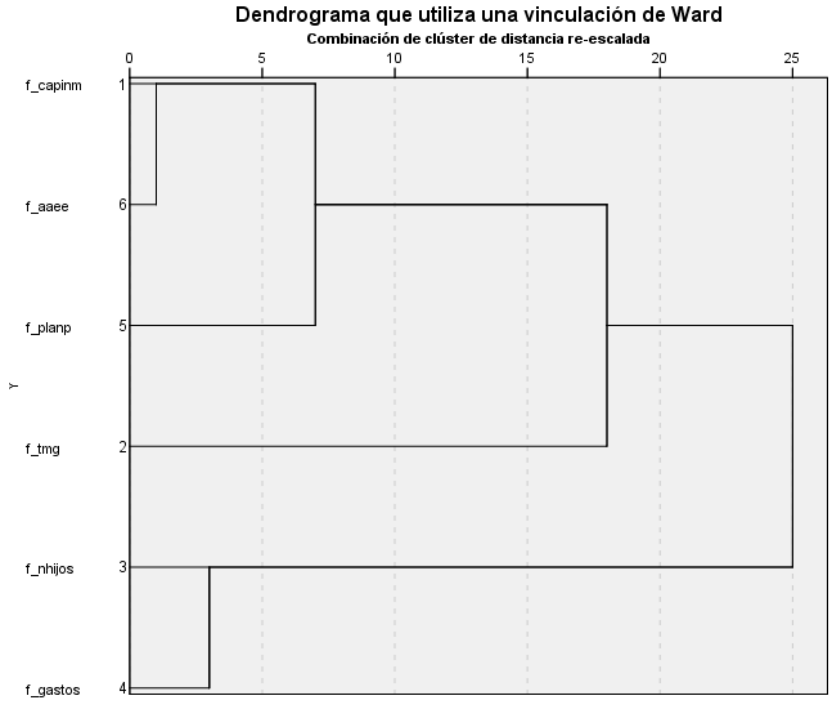


Figura 5-6

Vemos que un cluster de nivel dos está formado por el fraude en capital inmobiliario, el fraude en actividades económicas y el fraude en planes de pensiones. Un segundo cluster está formado únicamente por el fraude en tipo marginal. Un tercer cluster lo forman el fraude por incorrecta deducción de gastos y por inadecuada deducción por número de hijos y ascendientes. Se observa que el resultado de la segmentación es el mismo que en el caso de las redes neuronales.

5.4.2 Análisis de la sensibilidad. Estimación y diagnosis del modelo probit a través de los modelos lineales generalizados

Alternativamente podemos estimar nuestro modelo de fraude a través de un modelo probit desde la óptica de los modelos lineales generalizados. Ahora observamos en la tabla 5-12 que la función de enlace del modelo lineal generalizado es la función Probit y la distribución de probabilidad es la Binomal .

Modelos lineales generalizados

Información de modelo

Variable dependiente	fraude global[a]
Distribución de probabilidad	Binomial
Función de enlace	Probit

a. El procedimiento modela 0 como la respuesta, tratando a 1 como la categoría de referencia.

Resumen de procesamiento de casos

	N	Porcentaje
Incluido	1928494	100,0%
Excluido	0	0,0%
Total	1928494	100,0%

Información de variable categórica

			N	Porcentaje
Variable dependiente	fraude global	0	720371	37,4%
		1	1208123	62,6%
		Total	1928494	100,0%

Tabla 5-12

En la tabla 5-13 observamos que los estadísticos de la cantidad de información son ligeramente mayores que en el caso de la regresión logística, lo que indica que el modelo Probit es ligeramente inferior al modelo logit para la predicción del fraude con nuestra base de datos.

Bondad de ajuste[a]

	Valor	gl	Valor/gl
Desvianza	1047877,111	1910144	,549
Desvianza escalada	1047877,111	1910144	
Chi-cuadrado de Pearson	5,780E+272	1910144	3,026E+266
Chi-cuadrado de Pearson escalado	5,780E+272	1910144	
Logaritmo de verosimilitud[b]	-523940,452		
Criterio de información Akaike (AIC)	1048012,903		
AIC corregido para muestras finitas (AICC)	1048012,908		
Criterio de información bayesiana (BIC)	1048836,072		
AIC consistente (CAIC)	1048902,072		

Variable dependiente: fraude global

Tabla 5-13

La prueba Omnibus del contraste de significatividad conjunta tiene un p-valor prácticamente nulo (tabla 5-14), lo que indica que la significatividad conjunta de los parámetros estimados es muy alta.

Contraste Omnibus[a]

Chi-cuadrado de razón de verosimilitud	gl	Sig.
1500799,189	65	,000

Variable dependiente: fraude global

Tabla 5-14

Las estimaciones de los parámetros del modelo probit (tabla 5-15) son muy significativas, ya que los p-valores (columna Sig.) son muy pequeños salvo raras excepciones. Por otra parte, los odds-ratio (columna Exp(B)) que valoran el efecto de las variables independientes sobre el fraude, presentan muchos valores superiores a la unidad correspondientes a variables con efecto alto sobre el fraude.

Estimaciones de parámetro

Parámetro	B	Error estándar	95% de intervalo de confianza de Wald		Contraste de hipótesis			Exp(B)	95% de intervalo de confianza de Wald para Exp(B)	
			Inferior	Superior	Chi-cuadrado de Wald	gl	Sig.		Inferior	Superior
(Interceptación)	-4,509	,0126	-4,533	-4,484	127249,255	1	,000	,011	,011	,011
FAC1	-20,377	,0968	-20,567	-20,187	44349,921	1	,000	1,414E-9	1,170E-9	1,709E-9
FAC2	1,194	,0101	1,174	1,214	13867,093	1	,000	3,300	3,235	3,366
FAC3	1,140	,0104	1,120	1,161	12077,534	1	,000	3,127	3,064	3,192
FAC4	-,206	,0173	-,240	-,173	142,467	1	,000	,813	,786	,842
FAC5	-3,862	,0208	-3,903	-3,821	34393,600	1	,000	,021	,020	,022
FAC6	,416	,0055	,405	,427	5728,847	1	,000	1,516	1,499	1,532
FAC7	-28,176	,1523	-28,474	-27,877	34213,316	1	,000	5,801E-13	4,303E-13	7,819E-13
FAC8	,026	,0128	,001	,051	4,157	1	,041	1,026	1,001	1,052
FAC9	,169	,0118	,146	,193	205,010	1	,000	1,185	1,157	1,212
FAC10	-,074	,0065	-,087	-,062	132,526	1	,000	,928	,917	,940
FAC11	-2,533	,0813	-2,692	-2,374	971,490	1	,000	,079	,068	,093
FAC12	-,070	,0020	-,074	-,066	1188,709	1	,000	,932	,929	,936
FAC13	-,480	,0144	-,508	-,452	1116,061	1	,000	,619	,602	,637
FAC14	-,314	,0017	-,318	-,311	34238,462	1	,000	,730	,728	,733
FAC15	,008	,0241	-,039	,056	,119	1	,730	1,008	,962	1,057
FAC16	-,173	,0073	-,187	-,159	563,729	1	,000	,841	,829	,853
FAC17	-,043	,0027	-,048	-,037	242,103	1	,000	,958	,953	,964
FAC18	,025	,0045	,016	,034	31,044	1	,000	1,026	1,016	1,035
FAC19	-5,103	,0227	-5,148	-5,059	50399,724	1	,000	,006	,006	,006
FAC20	,007	,0045	-,002	,015	2,223	1	,136	1,007	,998	1,016
FAC21	-,045	,0134	-,071	-,019	11,312	1	,001	,956	,931	,981
FAC22	-,002	,0032	-,008	,004	,474	1	,491	,998	,992	1,004
FAC23	,024	,0013	,021	,026	345,099	1	,000	1,024	1,022	1,027
FAC24	,935	,0079	,920	,951	14182,297	1	,000	2,548	2,509	2,588
FAC25	-5,892	,0158	-5,923	-5,861	138185,071	1	,000	,003	,003	,003
FAC26	-,039	,0043	-,047	-,031	82,888	1	,000	,962	,954	,970
FAC27	,376	,0126	,351	,400	892,357	1	,000	1,456	1,420	1,492
FAC28	,051	,0014	,048	,053	1244,135	1	,000	1,052	1,049	1,055
FAC29	-1,362	,0176	-1,396	-1,327	6009,097	1	,000	,256	,248	,265
FAC30	-,386	,0566	-,497	-,275	46,483	1	,000	,680	,608	,760
FAC31	-,243	,0095	-,262	-,224	656,513	1	,000	,784	,770	,799
FAC32	-1,094	,0062	-1,106	-1,082	31323,251	1	,000	,335	,331	,339
FAC33	-,400	,0105	-,421	-,379	1451,493	1	,000	,670	,657	,684
FAC34	-,045	,0035	-,052	-,039	172,823	1	,000	,956	,949	,962
FAC35	-1,739	,0946	-1,924	-1,553	337,860	1	,000	,176	,146	,212
FAC36	-,054	,0067	-,067	-,041	65,775	1	,000	,947	,935	,960
FAC37	-,004	,0044	-,013	,005	,802	1	,370	,996	,987	1,005
FAC38	-,016	,0017	-,020	-,013	94,924	1	,000	,984	,981	,987
FAC39	,099	,0052	,089	,109	357,296	1	,000	1,104	1,093	1,115
FAC40	-,737	,0095	-,756	-,718	5978,458	1	,000	,479	,470	,488
FAC41	-,029	,0036	-,036	-,022	63,351	1	,000	,972	,965	,979
FAC42	,130	,0077	,115	,145	285,174	1	,000	1,139	1,122	1,156
FAC43	-,462	,0071	-,476	-,448	4204,728	1	,000	,630	,621	,639
FAC44	-,077	,0027	-,082	-,072	844,646	1	,000	,926	,921	,931
FAC45	,211	,0038	,204	,218	3146,309	1	,000	1,235	1,226	1,244
FAC46	-,538	,0022	-,542	-,534	59567,524	1	,000	,584	,581	,586
FAC47	-,280	,0090	-,298	-,263	975,852	1	,000	,756	,743	,769
FAC48	-,017	,0062	-,029	-,005	7,330	1	,007	,983	,971	,995
FAC49	-,009	,0053	-,020	,001	2,989	1	,084	,991	,980	1,001
FAC50	-,080	,0025	-,085	-,075	991,827	1	,000	,923	,919	,928
FAC51	,049	,0016	,046	,052	926,094	1	,000	1,050	1,047	1,053
FAC52	,019	,0034	,013	,026	32,320	1	,000	1,020	1,013	1,026
FAC53	,064	,0022	,060	,069	823,606	1	,000	1,066	1,062	1,071
FAC54	,055	,0032	,049	,062	296,594	1	,000	1,057	1,050	1,064
FAC55	,072	,0012	,069	,074	3284,166	1	,000	1,074	1,072	1,077
FAC56	,125	,0047	,116	,134	698,743	1	,000	1,133	1,123	1,144
FAC57	,057	,0047	,048	,067	150,945	1	,000	1,059	1,049	1,069
FAC58	,045	,0037	,037	,052	142,454	1	,000	1,046	1,038	1,053
FAC59	,108	,0037	,101	,115	838,254	1	,000	1,114	1,106	1,122
FAC60	-,022	,0017	-,025	-,018	153,211	1	,000	,979	,975	,982
FAC61	-,527	,0071	-,541	-,513	5500,720	1	,000	,591	,582	,599
FAC62	-,045	,0042	-,053	-,036	113,619	1	,000	,956	,949	,964
FAC63	-,032	,0025	-,037	-,027	159,642	1	,000	,969	,964	,973
FAC64	,016	,0016	,013	,020	108,795	1	,000	1,017	1,013	1,020
FAC65	-,114	,0128	-,139	-,089	79,562	1	,000	,892	,870	,915
(Escala)	1[a]									

Tabla 5-15

En cuanto al efecto de cada partida del impuesto sobre el fraude, se obtiene la misma clasificación que en caso del modelo logit.

La matriz de correlaciones de los estimadores de los parámetros en el modelo probit tiene muchos valores muy bajos, lo que indica que el ajuste del modelo probit es bueno.

El modelo probit estimado, que predice probabilidades de fraude es el siguiente:

$$P(\mathit{fraude}) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{-4{,}5-20{,}37FAC1+1{,}185C2+\cdots+0{,}016FAC64-0{,}114FAC65} e^{-\frac{t^2}{2}} dt$$

Las probabilidades de fraude predichas por el modelo Probit se encuentran en la columna *Probabilidadfraude1* de la tabla 5-16. En la columna *PredictedValue_1* se muestran las categorías de fraude asociadas a cada contribuyente. Estas categorías de fraude coinciden prácticamente con las estimadas mediante el modelo Logit. Ambos modelos son bastante equivalentes. Por otra parte, las columnas *Leverage_1* y *CooksDistance_1* presentan valores muy pequeños, lo que indica que no existen valores influyentes en la regresión.

PredictedValue_1	Leverage_1	Residual_1	CooksDistance_1	ProbFraudeProbit
1	,000	-,097	,000	,90302
1	,000	-,108	,000	,89153
0	,000	,121	,000	,12066
0	,000	,154	,000	,15355
0	,000	,215	,000	,21494
0	,000	,061	,000	,06072
1	,000	-,316	,000	,68377
1	,000	-,310	,000	,69018
1	,000	,000	,000	1,00000

Tabla 5-16

Si realizamos un escalamiento multidimensional para segmentar las cuasas de fraude según su incidencia en el fraude global, obtenemos el mapa

perceptual de la figura 5-7. Se observa un comportamiento totalmente similar al caso del modelo probit

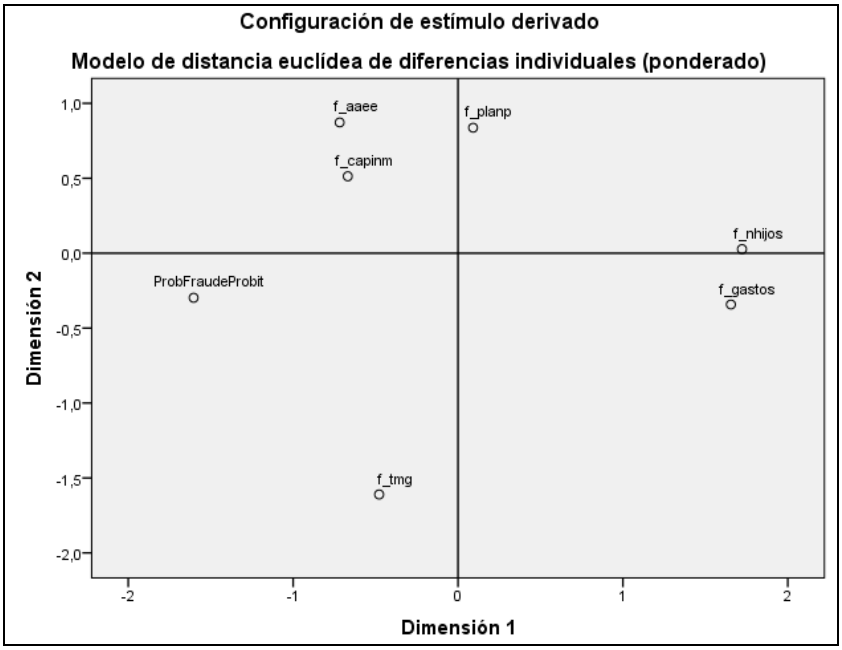


Figura 5-7

La calidad de la segmentación es alta dado el valor bajo de estadístico S-stress y el valor alto del estadístico RSQ.

```
Stress  =   ,08638      RSQ =  ,95630
```

Adicionalmente se observa que el área bajo la curva ROC (Figura 5-8) para el modelo Probit es 0,950 (tabla 5-17), valor muy cercano a 1 que indica un ajuste muy bueno del modelo Probit. No obstante, esta área es ligeramente inferior a la misma área para el modelo Logit (0,959). Por lo tanto, como ya indicaron los estadísticos de la cantidad de información, el modelo Logit es ligeramente superior al modelo Probit para estimar las probabilidades de fraude de los contribuyentes por IRPF.

Área bajo la curva

Variable(s) de re:

Área
,950

Tabla 5-17

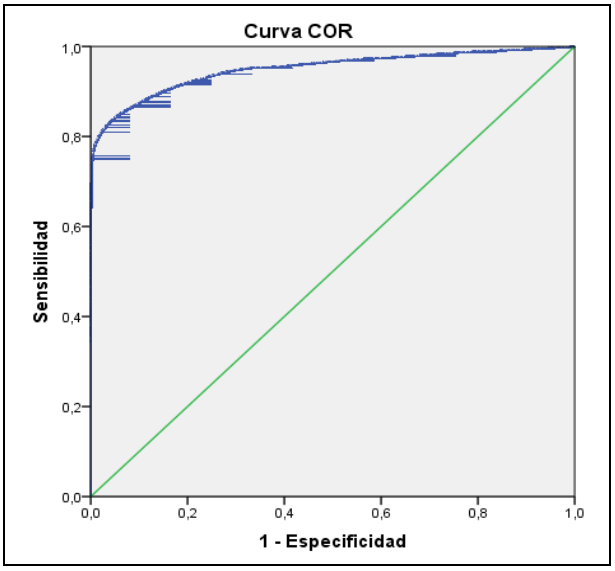


Figura 5-8

Para finalizar la diagnosis se realiza un análisis de los residuos del modelo Probit.

Si observamos la tabla 5-18, vemos que el valor esperado de los residuos es prácticamente nulo. Además, como media y mediana tiene valores muy cercanos, la distribución de los residuos es simétrica, lo que posibilita la normalidad. Por otra parte, los coeficientes de asimetría están dentro del intervalo (-2,2) y el de curtosis está muy cercano a 2, lo que implica normalidad de los residuos. Asimismo, los estimadores robustos tienen valores algo diferentes, lo que puede indicar la presencia de algún valor atípico residual no demasiado incidente.

Descriptivos

			Estadístico	Error típ.
Residuo bruto	Media		-,00043	,000207
	Intervalo de confianza para la media al 95%	Límite inferior	-,00083	
		Límite superior	-,00002	
	Media recortada al 5%		,01333	
	Mediana		,00000	
	Varianza		,083	
	Desv. típ.		,287288	
	Mínimo		-1,000	
	Máximo		1,000	
	Rango		2,000	
	Amplitud intercuartil		,150	
	Asimetría		-1,028	,002
	Curtosis		2,224	,004

Estimadores-M

	Estimador-M de Huber[a]	Biponderado de Tukey[b]	Estimador-M de Hampel[c]	Onda de Andrews[d]
Residuo bruto	,03664	,04985	,05479	,04981

Tabla 5-18

La figura 5-9 muestra el gráfico Q-Q de normalidad para los residuos del modelo probit. Como la nube de puntos se ajusta bien a la diagonal, los residuos son normales.

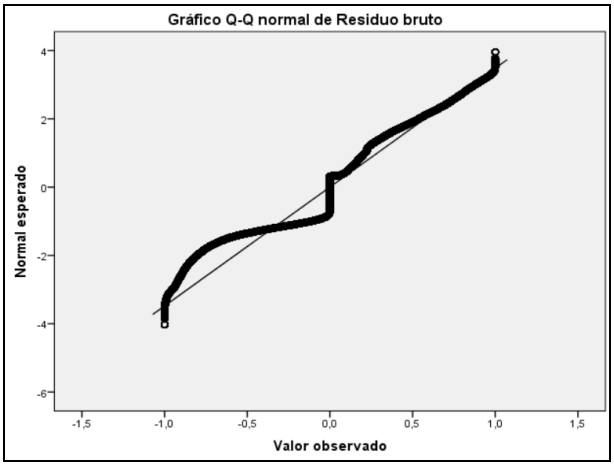


Figura 5-9

5.4.3 Análisis de la propensión al fraude a través de redes neuronales

Las Redes Neuronales constituyen el corazón de las técnicas de Machine Learning. En este apartado del trabajo se construye un modelo de red neuronal de aprendizaje supervisado tipo Perceptrón Multicapa tomando como variable dependiente (única variable de salida de la red) una variable dicotómica que toma el valor 1 si el individuo defrauda y el valor cero si el individuo no defrauda (variable marca). Las variables independientes (variables de entrada de la red) son las variables que constituyen las partidas más importantes del Impuesto sobre la Renta de las Personas Físicas.

Con el modelo de red neuronal se buscará predecir la probabilidad que tiene cualquier individuo de defraudar o no, según los valores declarados en las variables del modelo 100.

También se trata de investigar qué partidas del IRPF tienen mayor incidencia en el fraude fiscal. Con el objeto de simplificar el trabajo, se utilizarán como variables independientes del modelo de red neuronal las 65 componentes principales previamente calculadas en el proceso de reducción de la dimensión.

Se utilizan un 70% de los datos para la fase de entrenamiento y un 30% para la fase de pruebas. En total tenemos prácticamente 2.000.000 de filas en la base de datos tal y como indica la tabla siguiente (exactamente 1.928.494) de las cuales 1.350.974 se utilizan para entrenar la red y el resto para prueba. No hay datos faltantes en la base de datos.

La red estimada se presenta en la Figura 5-10.

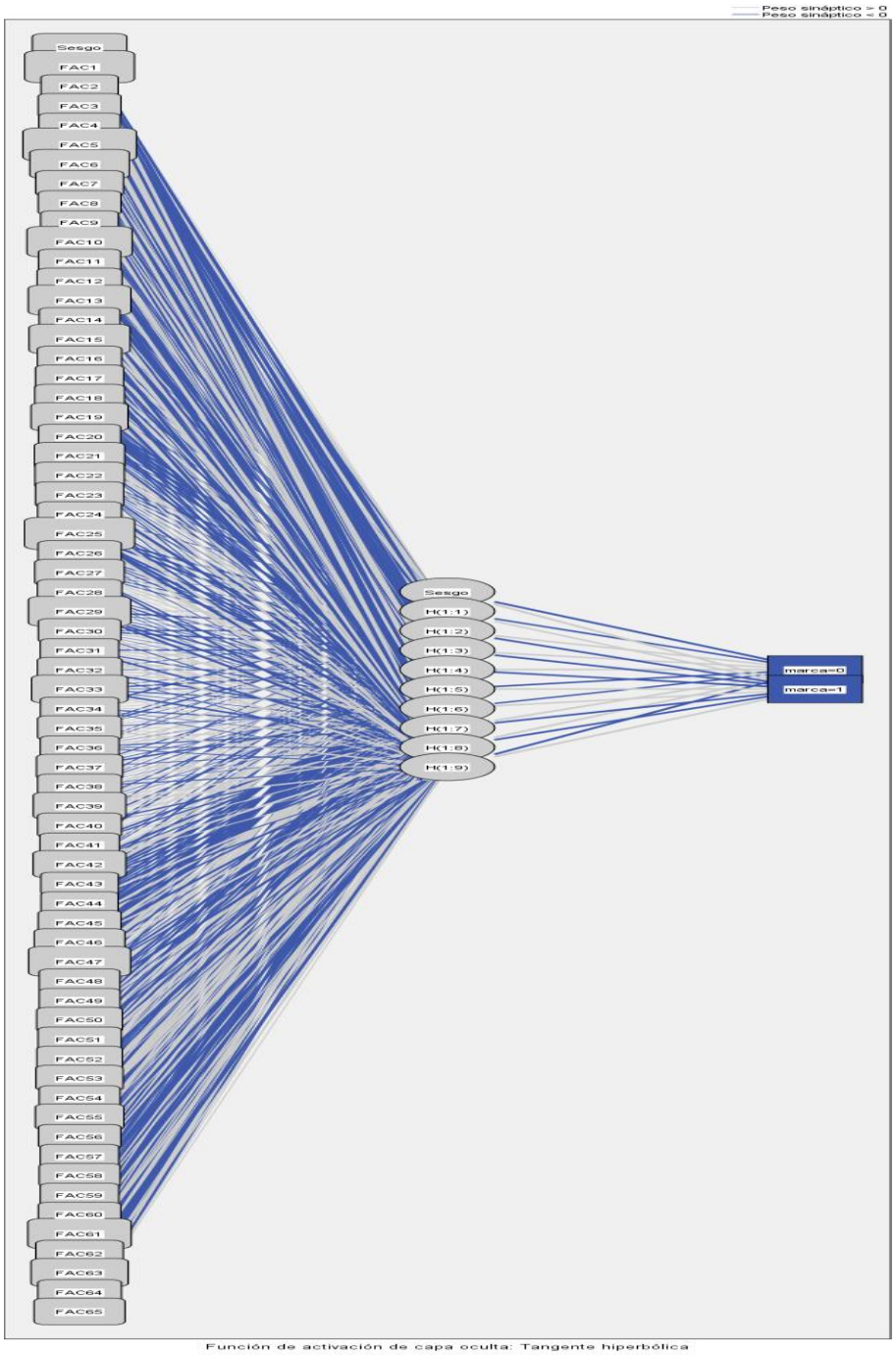

Peso sináptico > 0
Peso sináptico < 0
Sesgo
FAC1
FAC2
FAC3
FAC4
FAC5
FAC6
FAC7
FAC8
FAC9
FAC10
FAC11
FAC12
FAC13
FAC14
FAC15
FAC16
FAC17
FAC18
FAC19
FAC20
FAC21
FAC22
FAC23
FAC24
FAC25
FAC26
FAC27
FAC28
FAC29
FAC30
FAC31
FAC32
FAC33
FAC34
FAC35
FAC36
FAC37
FAC38
FAC39
FAC40
FAC41
FAC42
FAC43
FAC44
FAC45
FAC46
FAC47
FAC48
FAC49
FAC50
FAC51
FAC52
FAC53
FAC54
FAC55
FAC56
FAC57
FAC58
FAC59
FAC60
FAC61
FAC62
FAC63
FAC64
FAC65
Sesgo
H(1:1)
H(1:2)
H(1:3)
H(1:4)
H(1:5)
H(1:6)
H(1:7)
H(1:8)
H(1:9)
marca=0
marca=1
Función de activación de capa oculta: Tangente hiperbólica
Función de activación de capa de salida: Softmax

Figura 5-10

La estructura de la red neuronal incluye en la capa de entrada 65 nodos correspondientes a las variables de entrada o variables independientes (componentes principales), una única capa oculta cuyos nodos están etiquetados con las etiquetas de los pesos sinápticos y un nodo de salida etiquetado con las dos categorías de la variable dependiente del modelo de red (fraude y no fraude).

En cuanto a la diagnosis del modelo de red, vemos en primer lugar la matriz de confusión (tabla 5-19) que presenta altos porcentajes de acierto en los valores pronosticados.

Clasificación

		Pronosticado		
Ejemplo	Observado	0	1	Porcentaje correcto
Entrenamiento	0	475973	28568	94,3%
	1	75685	770748	91,1%
	Porcentaje global	40,8%	59,2%	92,3%
Pruebas	0	203454	12376	94,3%
	1	32478	329212	91,0%
	Porcentaje global	40,9%	59,1%	92,2%

Variable dependiente: fraude global

Tabla 5-19

Asimismo, el porcentaje de pronósticos incorrectos del modelo de red se valora en un 7% aproximadamente (tabla 5-20).

Resumen del modelo

Entrenamiento	Error de entropía cruzada	266541,696
	Porcentaje de pronósticos incorrectos	7,7%
	Regla de parada utilizada	Se ha superado el número máximo de épocas (100)
	Tiempo de preparación	0:05:13,58
Pruebas	Error de entropía cruzada	115158,987
	Porcentaje de pronósticos incorrectos	7,8%

Variable dependiente: fraude global

Tabla 5-20

Otro elemento de diagnosis es la curva ROC de la red. En la figura 5-11 se observa la curva ROC para el fraude y para el no fraude, presentando ambas un área muy elevada entre las curvas y la diagonal. El área bajo la curva ROC se estima en 0,972 (tabla 5-21), valor muy cercano a la unidad, lo que indica que la capacidad predictiva de la red es muy alta. También se presenta el gráfico de ganacias (Figura 5-12) y el gráfico de elevación de la red (Figura 5-13). Los altos valores de ganacias y elevaciones indican un buen ajuste de la red neuronal.

Un valor añadido esencial del modelo de red neuronal es la evaluación de la importancia de las variables independientes sobre la dependiente. En nuestro caso, la tabla de importancias de la red (tabla 5-22) cuantifica el grado de incidencia de las partidas de IRPF en el fraude fiscal.

Área bajo la curva

		Área
fraude global	0	,972
	1	,972

Tabla 5-21

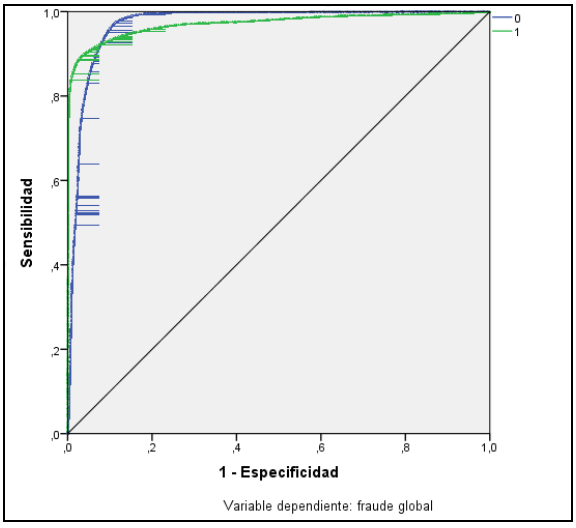


Figura 5-11

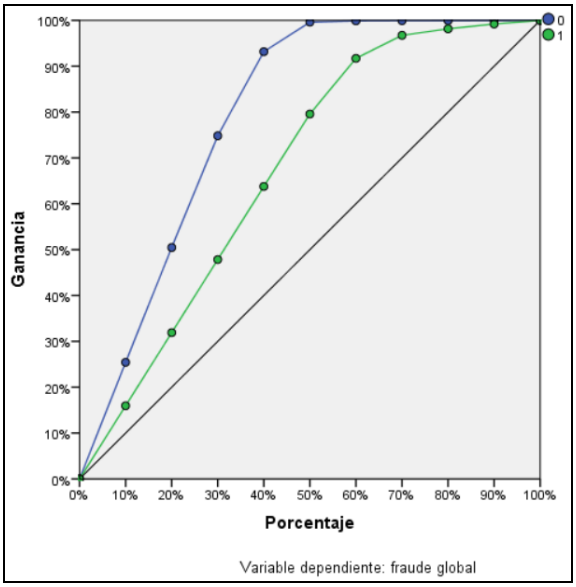


Figura 5-12

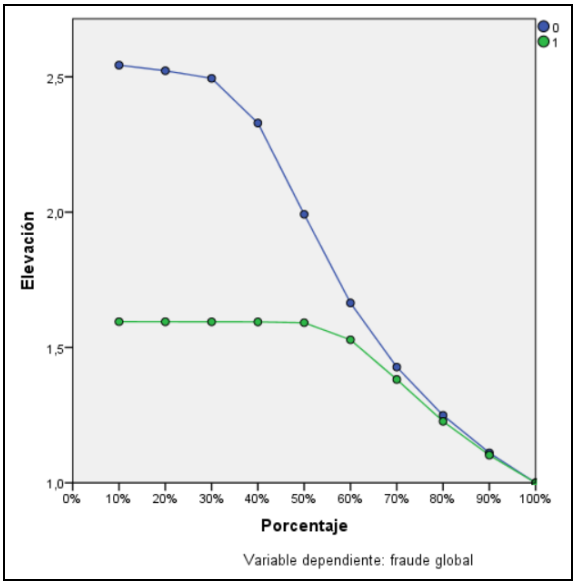


Figura 5-13

	Importancia	Importancia normalizada
Capital Inmobiliario	0,048	100,00%
Base General y Cuotas	0,044	90,40%
Rentas Inmobiliarias	0,044	91,20%
Mínimos personal y familiar	0,038	78,00%
Imputación rentas paraísos fiscales y compensación bases liquidables	0,035	73,00%
Actividades económicas	0,034	71,20%
Reducción aportaciones previsión social	0,034	70,70%
Atribuciones de renta: Capital mobiliario e inmobiliario	0,034	69,50%
Deducción por incentivos y estímulos a la inversión	0,032	66,60%
Deducciones vivienda habitual	0,03	63,00%
Rendimientos asistencia técnica y capital mobiliario	0,028	57,00%
Gastos deducibles y cotizaciones	0,027	56,50%
Atribuciones de rentas: Actividades económicas	0,026	53,40%
Retenciones, saldos negativos y rendimientos capital mobiliario	0,022	44,70%
Anualidades de alimentos	0,022	44,90%
Reducción Base Imponible por alimentos	0,019	38,80%
Ingresos por activiad económica	0,019	39,20%
Deducción venta bienes Canarias	0,018	37,20%
Reducción mutualidades deportivas	0,018	36,80%
Reducción Base Imponible por discapacidad	0,017	36,10%
Deducciones autonómicas pérdida de derecho	0,016	33,20%
Deducción autonómica por maternidad	0,016	33,20%
Deducciones alquiler vivienda y rendimientos del tabajo	0,016	34,10%
Retribuciones en especie	0,015	31,90%
Rentas exentas de IRPF	0,015	31,00%
Saldo neto rendimientos capital mobiliario	0,014	30,00%
Retenciones e ingresos UTEs e imputación transparencia fiscal	0,013	27,70%
Ganancias agrarias y adecuación del impuesto	0,013	26,20%
Compensacion rendimientos C.M. y Seguros vida	0,012	24,50%
Deducciones a las que se ha perdido derecho	0,012	25,90%
Deducciones Ceuta y Melilla y doble imposición	0,011	22,50%
Deducciones por doble imposición	0,011	21,80%
Deducciones maternidad	0,011	23,00%
Adecuaciones situación familiar y personla	0,011	23,20%
Rendimiento bienes derecho imagen	0,011	21,90%
Reducción BI aportaciones previsión social	0,01	20,20%
Adecuación, reducción e importes por discapacidad	0,01	21,50%
Saldo neto negativo y pérdidas patrimoniales	0,01	21,50%
Intereses de demora y deducciones ejercicios anteriores	0,01	20,40%
Cuotas impuesto rentas no residentes	0,01	21,10%
Saldo neto y Resultado	0,009	18,70%
Capital Mobiliario	0,009	19,00%
Deducciones inversión cultural	0,009	18,20%
Gastos defensa jurídica, colegios profesionales y sindicatos	0,009	19,50%
Reducciones cuotas de afiliación partidos políticos	0,009	18,00%
Deducción cuentas ahorro empresa	0,008	17,00%
Deducciones doble imposición y Transparencia fiscal	0,008	17,40%
Deducciones generales	0,007	15,30%
Reducción Base Imponible	0,007	13,80%
Saldos negativos patrimoniales	0,007	13,70%
Reducción rendimiento trabajo activos > 65 años	0,007	13,60%
Rendimientos propiedad industrial no actividad económica	0,007	13,70%
Ingresos y reducciones actividad no económica	0,006	13,20%
Gastos deducibles, letras Tesoro y gastos deducibles	0,005	9,90%
Reducciones previsiones sociales declarnte y cónyuge	0,004	7,60%

Rendimientos y reducciones por capital mobiliario	0,004	8,10%
Reducciones rendimientos parentesco en más de 2 años	0,004	7,80%
Deducciones autonómicas	0,003	6,00%
Gastos e ingresos deducibles	0,003	6,40%
Arrendamineto inmuebles y gastos fiscales deducibles	0,003	5,30%
Base del Ahorro	0,002	4,00%
Gastos deducibles importes ejercicios anteriores	0,002	3,90%

Tabla 5-22

Al observar esta tabla de importancias vemos que contiene partidas muy similares a la tabla de efectos de los modelos logit y probit, sobre todo para las partidas de mayores efectos. El gráfico de importancias de la Figura 5-14 ilustra un poco más la tabla anterior.

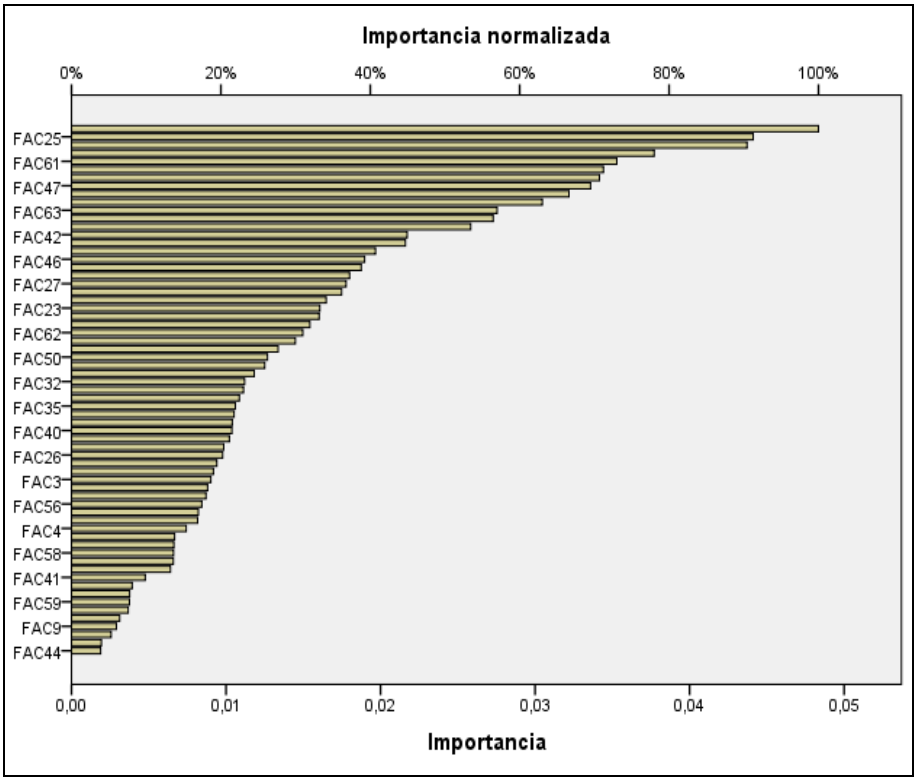


Figura 5-14

Una de las ventajas de los modelos predictivos para la detección del fraude radica en la posibilidad de poder calcular probabilidades de fraude individuales para los contribuyentes. La red neuronal ofrece como salida (tabla 5-23) la clasificación de cada declarante como fraudulento o no fraudulento (variable *MLP_PredictiveValue_A*) y adicionalmente muestra las propensiones al fraude de cada declarante (variable *MLP_PseudoProbability_2_A*). Es decir, no sólo clasifica los individuos como propensos o no al fraude, sino que también computa la probabilidad de fraude de cada declarante.

MLP_PredictedValue_A	MLP_PseudoProbability_1_A	MLP_PseudoProbability_2_A
1	,003	,997
1	,006	,994
0	,941	,059
0	,942	,058
1	,381	,619
0	,954	,046
1	,084	,916
1	,347	,653
1	,000	1,000
1	,002	,998
1	,000	1,000
1	,000	1,000
1	,001	,999
0	,699	,301
0	,600	,400
0	,897	,103
1	,000	1,000
1	,023	,977
1	,001	,999
0	,956	,044
0	,773	,227

Tabla 5-23

La figura 5-15 muestra la densidad de probabilidad de la propensión al fraude mediante el Perceptrón Multicapa. Se observa que la probabilidad de fraude es más densa para sus valores pequeños, pero también tiene valores altos alrededor de la probabilidad 0,8. Por lo tanto, la mayoría de los contribuyentes no defraudan, pero sí repunta un grupo de contribuyentes con probabilidad de fraude alta (alrededor de 0,8).

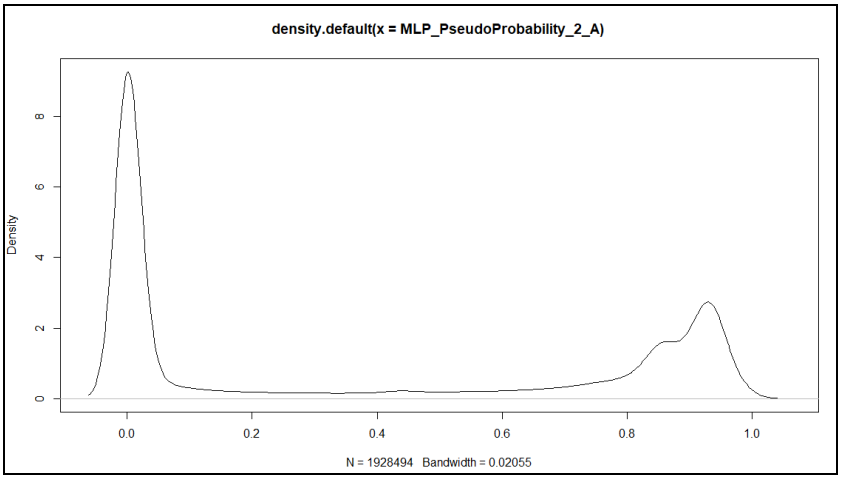


Figura 5-15

5.5 ANÁLISIS DE LA PROPENSIÓN AL FRAUDE A TRAVÉS DE REDES NEURONALES MÚLTIPLES CON REDUCCIÓN DE LA DIMENSIÓN

Ya se ha comentado que las Redes Neuronales constituyen el corazón de las técnicas de Machine Learning. En este apartado del trabajo se construye un modelo de red neuronal múltiple tipo Perceptrón Multicapa tomando inicialmente como variables dependientes las diversas causas de fraude (capa de salida de la red con seis nodos, uno para cada causa de fraude) todas ellas variables dicotómicas que toma el valor 1 si el individuo defrauda y el valor cero si el individuo no defrauda. Las variables independientes (variables de entrada de la red) son las variables que constituyen las partidas del Impuesto sobre la Renta de las Personas Físicas reducidas a 65 componentes principales (capa de entrada de la red con 65 nodos). A diferencia del capítulo anterior, aquí ampliaremos la red utilizando todas las partidas económicas reducidas a sus 65 componentes principales especificadas en el capítulo 3. En el capítulo 4 se había utilizado una reducción a 11 componentes principales. Una de las características esenciales de las herramientas de Machine Learning es que permiten el trabajo con Redes Neuronales en toda su extensión. Si observamos la figura

5-2 de la clasificación de las Técnicas de Machine Learning presentada al principio de este capítulo, vemos que las Redes Neuronales son una técnica aplicable para análisis supervisado de clasificación, para análisis supervisado de Regresión y para análisis no supervisado. Esta es la razón por la que en este capítulo se presenta el valor añadido de trabajar las redes neuronales en toda su extensión, tanto la red simple, como la red múltiple.

Con el modelo de red neuronal se buscará predecir la probabilidad que tiene cualquier individuo de defraudar o no, según los valores declarados en las variables del modelo 100. También se trata de investigar qué partidas del IRPF tienen mayor incidencia en el fraude fiscal.

Se utilizan un 70% de los datos para la fase de entrenamiento y un 30% para la fase de pruebas. En total tenemos prácticamente 2.000.000 de filas en la base de datos tal y como indica la tabla siguiente (exactamente 1.928.494) de las cuales 1.350.974 se utilizan para entrenar la red y el resto para prueba (tabla 5-24). No hay datos faltantes en la base de datos.

Perceptrón multicapa

Resumen de procesamiento de casos

		N	Porcentaje
Ejemplo	Entrenamiento	1350974	70,1%
	Pruebas	577520	29,9%
Válido		1928494	100,0%
Excluido		0	
Total		1928494	

Tabla 5-24

La estructura de la red neuronal (figura 5-16) incluye en la capa de entrada 65 nodos correspondientes a la variables de entrada o variables independientes (componentes principales), una única capa oculta cuyos nodos están etiquetados con las etiquetas de los pesos sinápticos y con 6 nodos en la capa de salida, uno para cada causa de fraude, etiquetados con las dos categorías de la variable dependiente del modelo de red (fraude y no fraude).

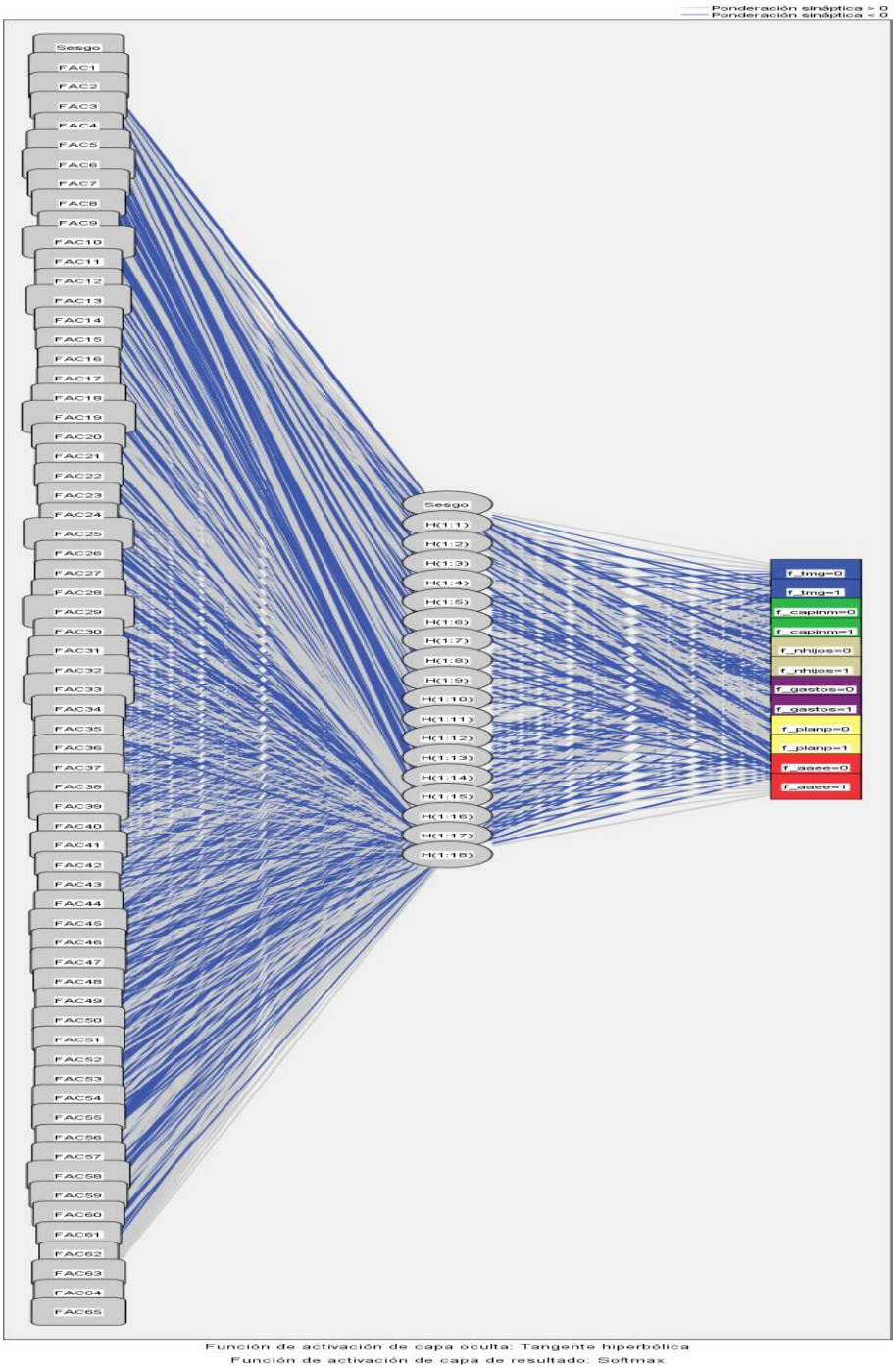

Ponderación sináptica > 0
Ponderación sináptica < 0
Sesgo
FAC1
FAC2
FAC3
FAC4
FAC5
FAC6
FAC7
FAC8
FAC9
FAC10
FAC11
FAC12
FAC13
FAC14
FAC15
FAC16
FAC17
FAC18
FAC19
FAC20
FAC21
FAC22
FAC23
FAC24
FAC25
FAC26
FAC27
FAC28
FAC29
FAC30
FAC31
FAC32
FAC33
FAC34
FAC35
FAC36
FAC37
FAC38
FAC39
FAC40
FAC41
FAC42
FAC43
FAC44
FAC45
FAC46
FAC47
FAC48
FAC49
FAC50
FAC51
FAC52
FAC53
FAC54
FAC55
FAC56
FAC57
FAC58
FAC59
FAC60
FAC61
FAC62
FAC63
FAC64
FAC65
Sesgo
H(1:1)
H(1:2)
H(1:3)
H(1:4)
H(1:5)
H(1:6)
H(1:7)
H(1:8)
H(1:9)
H(1:10)
H(1:11)
H(1:12)
H(1:13)
H(1:14)
H(1:15)
H(1:16)
H(1:17)
H(1:18)
f_Imig=0
f_Imig=1
f_capinm=0
f_capinm=1
f_nhijos=0
f_nhijos=1
f_gastos=0
f_gastos=1
f_planp=0
f_planp=1
f_aaee=0
f_aaee=1
Función de activación de capa oculta: Tangente hiperbólica
Función de activación de capa de resultado: Softmax

Figura 5-16

En cuanto a la diagnosis del modelo de red múltiple, vemos en primer lugar un resumen técnico del modelo (tabla 5-25) que presentan porcentajes de pronóstico incorrectos muy bajos para las variables dependientes categóricas.

Resumen del modelo			
Entrenamiento	Error de entropía cruzada		869461,321
	Promedio de porcentaje de pronósticos incorrectos		3,5%
	Porcentaje de pronósticos incorrectos para dependientes categóricas	Fraude que afecta al tipo marginal	5,1%
		Fraude que afecta a los rendimientos de capital inmobiliario	1,3%
		Fraude que afecta a las declaraciones de número e hijos y ascendieentes y descendientes	2,7%
		Fraude que afecta a la declaración de gastos	5,2%
		Fraude que afecta a la desgrbación por planes de pensiones	5,0%
		Fraude que afecta a la declaración de actividades económicas	1,6%
	Regla de parada utilizada		Se ha superado el tiempo máximo de preparación (15 minutos)
	Tiempo de preparación		0:15:13,55
Pruebas	Error de entropía cruzada		374075,523
	Promedio de porcentaje de pronósticos incorrectos		3,5%
	Porcentaje de pronósticos incorrectos para dependientes categóricas	Fraude que afecta al tipo marginal	5,1%
		Fraude que afecta a los rendimientos de capital inmobiliario	1,3%
		Fraude que afecta a las declaraciones de número e hijos y ascendieentes y descendientes	2,6%
		Fraude que afecta a la declaración de gastos	5,3%
		Fraude que afecta a la desgrbación por planes de pensiones	5,0%
		Fraude que afecta a la declaración de actividades económicas	1,7%

Tabla 5-25

En cuanto a la diagnosis del modelo de red múltiple, vemos en primer lugar las matrices de confusión (tablas 5-26 a 5-31), que presentan altos porcentajes de acierto en los valores pronosticados.

Fraude que afecta al tipo marginal

		Pronosticado		
Ejemplo	Observado	0	1	Porcentaje correcto
Entrenamiento	0	840710	14446	98,3%
	1	54216	440324	89,0%
	Porcentaje global	66,3%	33,7%	94,9%
Pruebas	0	360382	6209	98,3%
	1	23310	188897	89,0%
	Porcentaje global	66,3%	33,7%	94,9%

Tabla 5-26

Fraude que afecta a los rendimientos de capital inmobiliario

		Pronosticado		
Ejemplo	Observado	0	1	Porcentaje correcto
Entrenamiento	0	862409	8606	99,0%
	1	8425	470256	98,2%
	Porcentaje global	64,5%	35,5%	98,7%
Pruebas	0	370515	3606	99,0%
	1	3644	201033	98,2%
	Porcentaje global	64,6%	35,4%	98,7%

Tabla 5-27

Fraude que afecta a las declaraciones de número e hijos y ascendieentes y descendientes

		Pronosticado		
Ejemplo	Observado	0	1	Porcentaje correcto
Entrenamiento	0	1280094	8962	99,3%
	1	27002	33638	55,5%
	Porcentaje global	96,8%	3,2%	97,3%
Pruebas	0	549134	3803	99,3%
	1	11289	14572	56,3%
	Porcentaje global	96,8%	3,2%	97,4%

Tabla 5-28

Fraude que afecta a la declaración de gastos

Ejemplo	Observado	Pronosticado		
		0	1	Porcentaje correcto
Entrenamiento	0	1167950	8436	99,3%
	1	61833	111477	64,3%
	Porcentaje global	91,1%	8,9%	94,8%
Pruebas	0	500319	3727	99,3%
	1	26781	47971	64,2%
	Porcentaje global	91,1%	8,9%	94,7%

Tabla 5-29

Fraude que afecta a la desgrbación por planes de pensiones

Ejemplo	Observado	Pronosticado		
		0	1	Porcentaje correcto
Entrenamiento	0	965496	9993	99,0%
	1	57450	316757	84,6%
	Porcentaje global	75,8%	24,2%	95,0%
Pruebas	0	414154	4165	99,0%
	1	24925	135554	84,5%
	Porcentaje global	75,9%	24,1%	95,0%

Tabla 5-30

Fraude que afecta a la declaración de actividades económicas

Ejemplo	Observado	Pronosticado		
		0	1	Porcentaje correcto
Entrenamiento	0	873528	7870	99,1%
	1	14365	453933	96,9%
	Porcentaje global	65,8%	34,2%	98,4%
Pruebas	0	375814	3381	99,1%
	1	6378	193225	96,8%
	Porcentaje global	66,0%	34,0%	98,3%

Tabla 5-31

Las matrices anteriores presentan los porcentajes de acierto de la clasificación para las distintas causas de fraude. En la tabla 5-32 se presenta el porcentaje global de aciertos en la clasificación de la red.

Porcentaje global correcto

Ejemplo	Porcentaje global correcto
Entrenamiento	96,5%
Pruebas	96,5%

Tabla 5-32

Todos estos valores son mucho más altos que en el caso de la red del capítulo anterior construida con 11 componentes principales. Hemos consigido aumentar la eficiencia de la red al considerar 65 componentes principales.

Otro elemento de diagnosis son las curvas ROC de la red. En la tabla siguiente se muestran las áreas bajo las curvas ROC (tabla 5-33), que presentan valores muy cercanos a la unidad, lo que indica que la capacidad predictiva de la red es muy alta.

Área bajo la curva

		Área
Fraude que afecta al tipo marginal	0	,980
	1	,980
Fraude que afecta a los rendimientos de capital inmobiliario	0	,999
	1	,999
Fraude que afecta a las declaraciones de número e hijos y ascendieentes y descendientes	0	,967
	1	,967
Fraude que afecta a la declaración de gastos	0	,924
	1	,924
Fraude que afecta a la desgrbación por planes de pensiones	0	,965
	1	,965
Fraude que afecta a la declaración de actividades económicas	0	,998
	1	,998

Tabla 5-33

A continuación se presentan las 6 curvas ROC de la red múltiple (Figuras 5-17 a 5-22). Como hemos visto, todas ellas encierran un área elevada muy cercana a la unidad, lo que indica un ajuste excelente de la red.

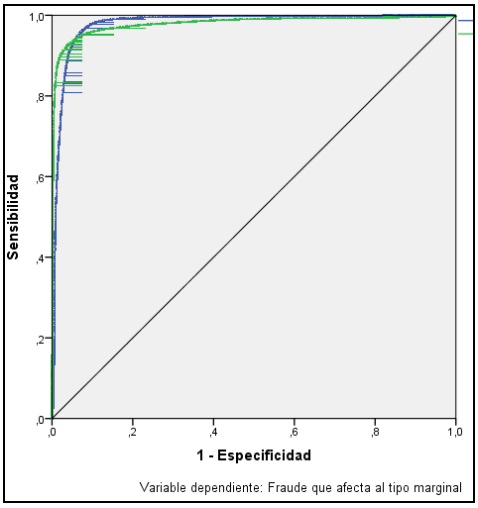


Figura 5-17

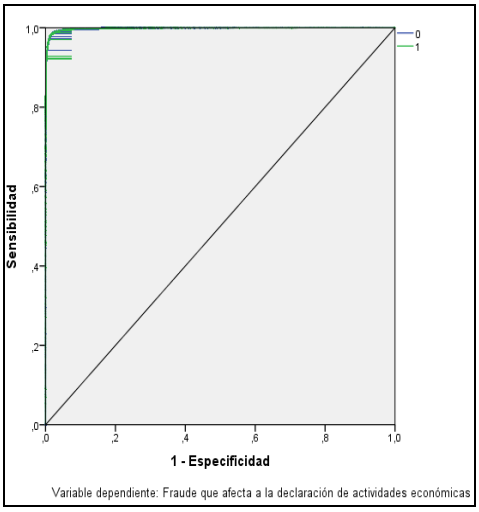


Figura 5-18

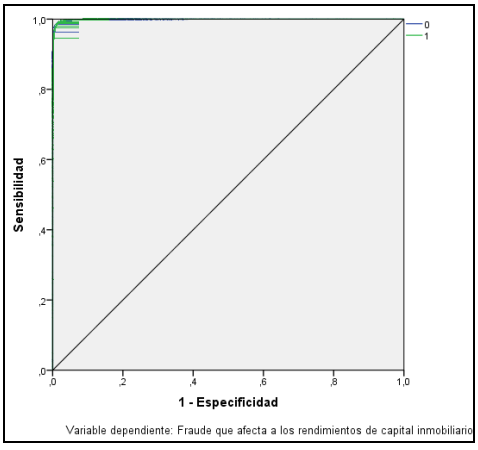


Figura 5-19

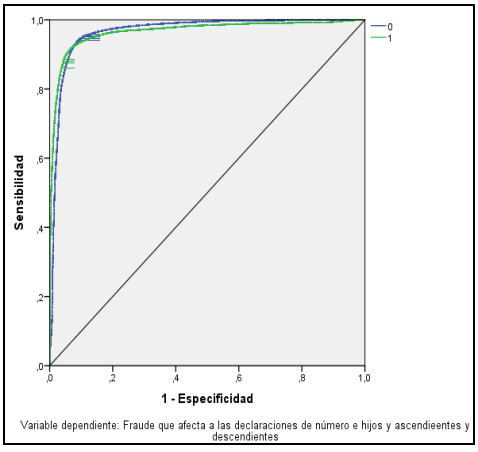


Figura 5-20

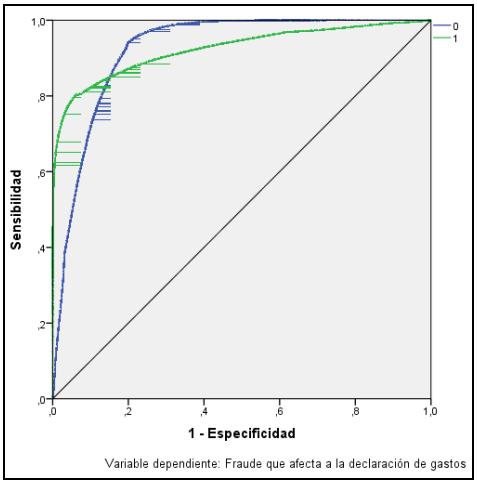


Figura 5-21

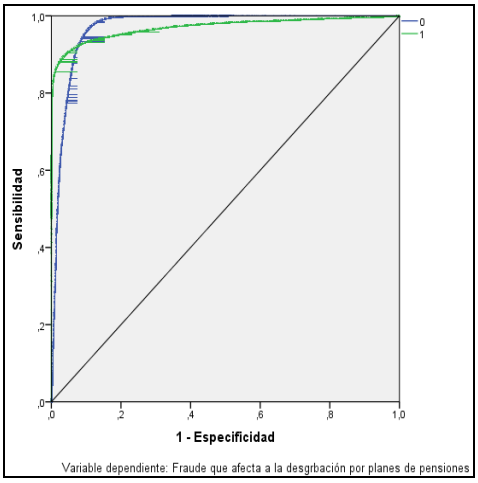


Figura 5-22

Las figuras siguientes presentan los 6 gráficos de ganancias de la red múltiple (Figura 5-23 a 5-28). Todas ellas presentan ganancias elevadas encierran un área significativa entre las dos curvas, lo que favorece el ajuste de la red.

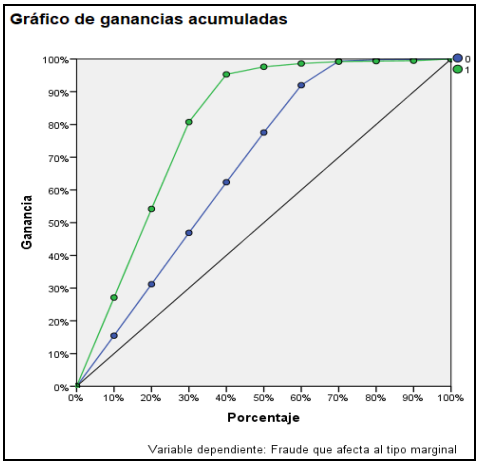


Figura 5-23

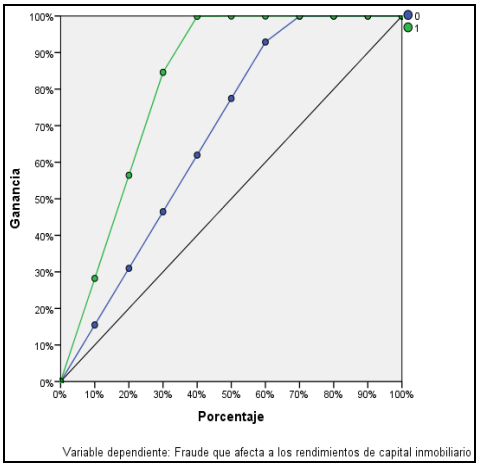


Figura 5-24

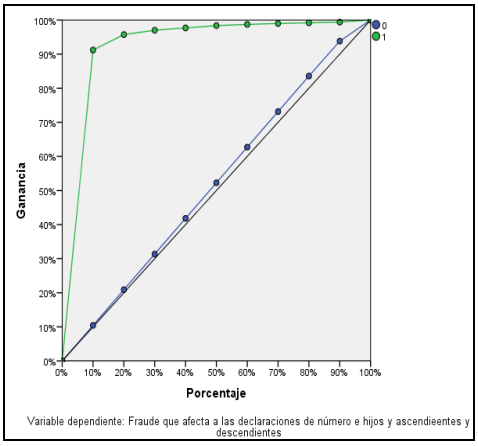


Figura 5-25

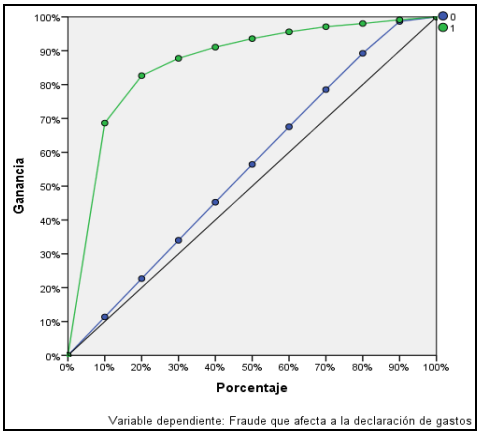


Figura 5-26

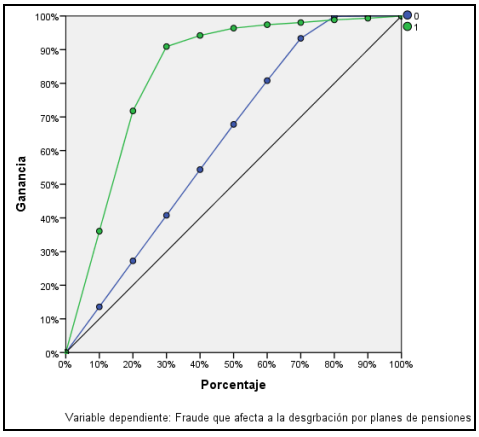


Figura 5-27

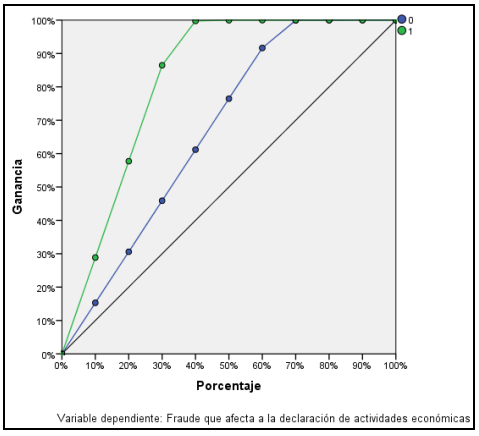


Figura 5-28

A continuación se presentan los 6 gráficos de elevación de la red neuronal múltiple (Figuras 5-29 a 5-34). Todos ellos alcanzan cuotas de elevación altas para indicar un buen ajuste de la red.

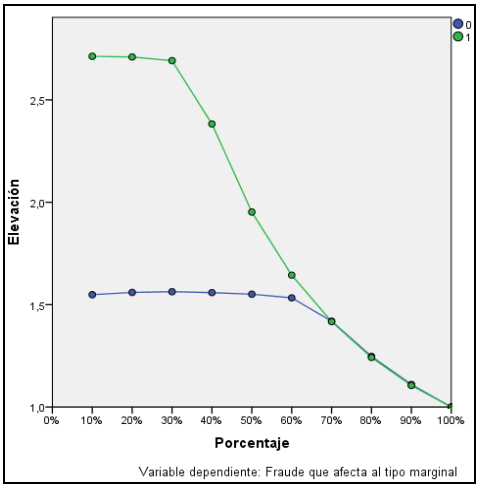


Figura 5-29

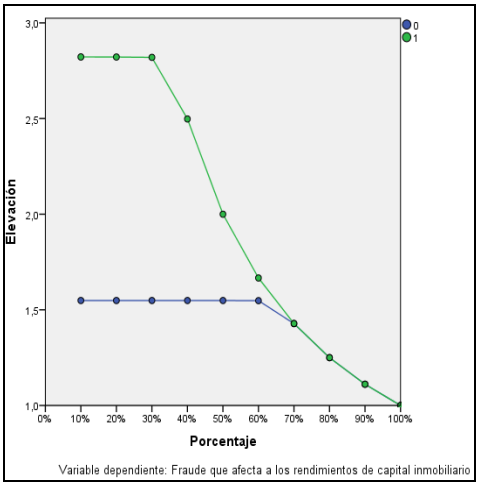


Figura 5-30

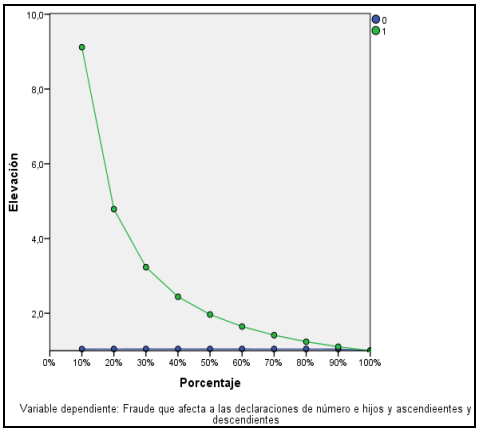


Figura 5-31

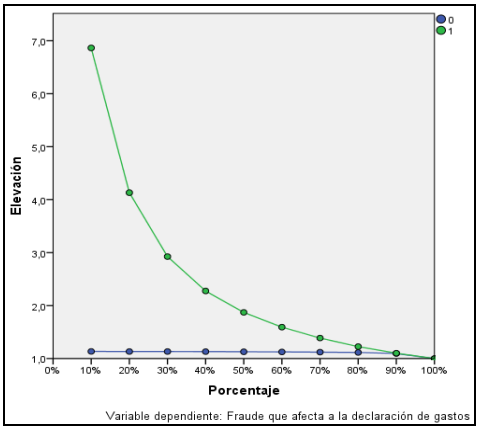


Figura 5-32

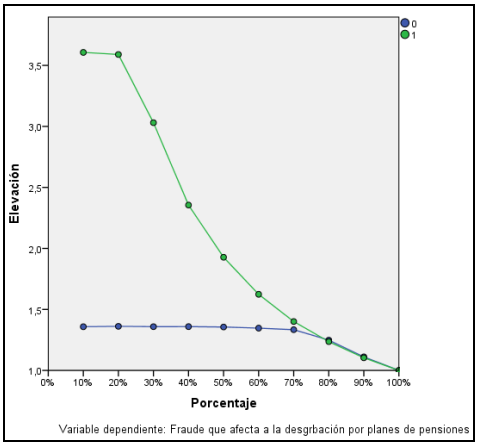


Figura 5-33

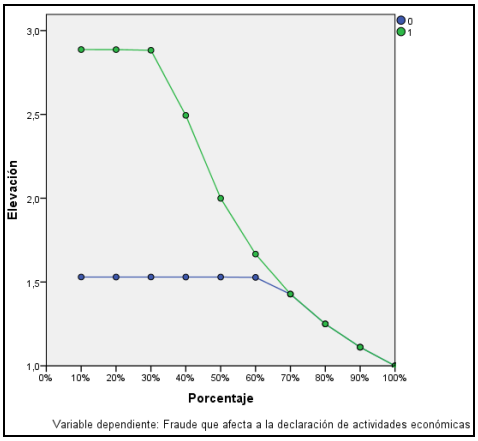


Figura 5-34

Un valor añadido esencial del modelo de red neuronal es la evaluación de la importancia de las variables independientes sobre la dependiente. En nuestro caso, la tabla de importancias de la red (tabla 5-34) cuantifica el grado de incidencia de las partidas de IRPF en el fraude fiscal.

	IMPORTANCIA	IMPORTANCIA NORMALIZADA
Deducciones vivienda habitual	0,027	98,00%
Mínimos personal y familiar	0,027	97,80%
Gastos deducibles y cotizaciones	0,027	100,00%
Reducción aportaciones previsión social	0,026	95,70%
Atribuciones de rentas: Actividades económicas	0,026	94,10%
Rentas Inmobiliarias	0,025	92,20%
Actividades económicas	0,022	82,00%
Capital Inmobiliario	0,021	78,80%
Adecuaciones situación familiar y personal	0,021	76,30%
Rendimientos propiedad industrial no actividad económica	0,021	76,90%
Retribuciones en especie	0,02	75,10%
Adecuación, reducción e importes por discapacidad	0,02	72,20%
Base General y Cuotas	0,019	69,50%
Base del Ahorro	0,019	69,20%
Reducciones previsiones sociales declarante y cónyuge	0,019	70,40%
Retenciones, saldos negativos y rendimientos capital mobiliario	0,019	69,70%
Gastos defensa jurídica, colegios profesionales y sindicatos	0,018	66,80%
Reducción rendimientos Copa América	0,018	66,30%
Saldo neto y Resultado	0,017	63,70%
Gastos deducibles, letras Tesoro y gastos deducibles	0,017	63,90%
Atribuciones de renta: Capital mobiliario e inmobiliario	0,017	61,10%

Capital Mobiliario	0,016	58,10%
Deducción venta bienes Canarias	0,016	58,70%
Deducciones maternidad	0,016	57,50%
Gastos e ingresos deducibles	0,016	57,30%
Ingresos por actiniad económica	0,016	60,40%
Aportaciones patrimonio protegido personas con discapacidad	0,016	58,80%
Deducción cuentas ahorro empresa	0,015	56,80%
Ganancias agrarias y adecuación del impuesto	0,015	53,60%
Reducción rendimiento trabajo activos > 65 años	0,015	54,40%
Deducciones alquiler vivienda y rendimientos del tabajo	0,015	56,10%
Reducciones rendimientos parentesco en más de 2 años	0,015	56,60%
Rendimientos asistencia técnica y capital mobiliario	0,015	55,90%
Retenciones e ingresos UTEs e imputación transparencia fiscal	0,014	50,40%
Ingresos y reducciones actividad no económica	0,014	50,80%
Intereses de demora y deducciones ejercicios anteriores	0,014	52,90%
Reducciones cuotas de afiliación partidos políticos	0,014	52,40%
Deducciones generales	0,013	46,20%
Reducción Base Imponible	0,013	46,00%
Reducción Base Imponible por discapacidad	0,013	46,20%
Deducciones inversión cultural	0,013	49,00%
Rendimiento bienes derecho imagen	0,013	47,20%
Rendimientos y reducciones por capital mobiliario	0,013	49,30%
Gastos deducibles importes ejercicios anteriores	0,013	47,20%
Saldo neto rendimientos capital mobiliario	0,013	46,50%
Cuotas impuesto rentas no residentes	0,013	49,10%
Deducciones Islas Canarias	0,012	43,40%
Compensación rendimientos C.M. y Seguros vida	0,012	42,40%
Reducción BI aportaciones previsión social	0,012	42,60%
Reducción mutualidades deportivas	0,012	45,10%
Anualidades de alimentos	0,012	43,30%
Saldos negativos patrimoniales	0,012	42,70%
Deducciones a las que se ha perdido derecho	0,012	43,30%
Deducción por incentivos y estímulos a la inversión	0,011	40,40%
Reducción Base Imponible por alimentos	0,011	40,30%
Deducciones doble imposición y Transparencia fiscal	0,011	41,70%
Deducciones autonómicas pérdida de derecho	0,01	35,30%
Imputación rentas paraísos fiscales y compensación bases liquidables	0,01	38,40%
Deducciones Ceuta y Melilla y doble imposición	0,009	32,30%
Deducción autonómica por maternidad	0,009	34,30%
Saldo neto negativo y pérdidas patrimoniales	0,009	32,80%
Arrendamineto inmuebles y gastos fiscales deducibles	0,009	32,30%
Deducciones autonómicas	0,008	30,60%
Deducciones por doble imposición	0,007	25,00%
Rentas exentas de IRPF	0,007	26,90%

Tabla 5-34

Las importancias de la tabla anterior pueden representarse en el gráfico de importancias de la Figura 5-35.

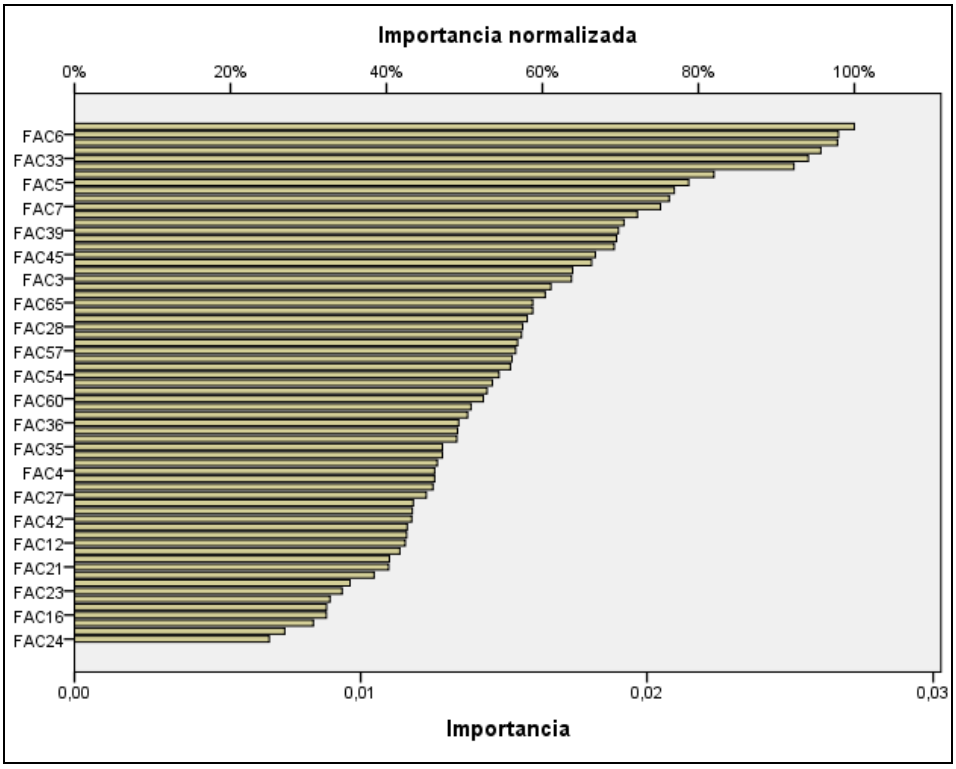


Figura 5-35

Al observar esta tabla de importancias vemos que contiene partidas muy similares a la tabla de efectos de los modelos logit, probit y redes simples, sobre todo para las partidas de mayores efectos.

5.5.1 Cálculo de las probabilidades de fraude de los contribuyentes

Una de las ventajas de los modelos predictivos para la detección del fraude radica en la posibilidad de poder calcular probabilidades de fraude individuales para los contribuyentes. La red neuronal múltiple ofrece como salida la clasificación de los contribuyentes en fraudulentos o no según las 6 causas de fraude consideradas (tabla 5-35).

MLP_PredictedValue_1	MLP_PredictedValue_2	MLP_PredictedValue_3	MLP_PredictedValue_4	MLP_PredictedValue_5	MLP_PredictedValue_6
1	0	0	0	1	1
0	0	1	1	0	0
0	0	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0	1
0	0	0	0	0	1
1	0	0	0	0	0
0	0	0	0	1	1
1	1	0	0	1	1
1	1	0	0	0	1

Tabla 5-35

Además, también se obtienen como salida las probabilidades de fraude de cada contribuyente para las 6 causas de fraude distintas (tabla 5-36).

MLP_Pseudo Probability_1 _1	MLP_Pseudo Probability_1 _2	MLP_Pseudo Probability_2 _1	MLP_Pseudo Probability_2 _2	MLP_Pseudo Probability_3 _1	MLP_Pseudo Probability_3 _2	MLP_Pseudo Probability_4 _1	MLP_Pseudo Probability_4 _2	MLP_Pseudo Probability_5 _1	MLP_Pseudo Probability_5 _2	MLP_Pseudo Probability_6 _1	MLP_Pseudo Probability_6 _2
,999	,001	,034	,966	1,000	,000	,990	,010	,051	,949	,017	,983
,999	,001	,990	,010	,253	,747	,000	1,000	,930	,070	1,000	,000
1,000	,000	,969	,031	1,000	,000	,970	,030	,975	,025	1,000	,000
1,000	,000	,972	,028	1,000	,000	,951	,049	,981	,019	1,000	,000
1,000	,000	,889	,111	,998	,002	,986	,014	,841	,159	,999	,001
1,000	,000	,969	,031	1,000	,000	,982	,018	,976	,024	1,000	,000
1,000	,000	,908	,092	,997	,003	,983	,017	,718	,282	,242	,758
,999	,001	,962	,038	1,000	,000	,983	,017	,897	,103	,071	,929
1,000	,000	,000	1,000	,993	,007	,823	,177	,985	,015	1,000	,000
,793	,207	,855	,145	,973	,027	,960	,040	,083	,917	,002	,998
,007	,993	,000	1,000	1,000	,000	,955	,045	,126	,874	,000	1,000
,000	1,000	,147	,853	,858	,142	,894	,106	,813	,187	,000	1,000
,984	,016	,950	,050	1,000	,000	,962	,038	,001	,999	,024	,976
,999	,001	,905	,095	,999	,001	,848	,152	,733	,267	1,000	,000
1,000	,000	,997	,003	,091	,909	,993	,007	,945	,055	,999	,001
1,000	,000	,980	,020	,998	,002	,987	,013	,997	,003	1,000	,000

Tabla 5-36

La red neuronal también ofrece como salida la probabilidad de fraude de cada declarante (*MLP_PseudoProbability_2_B*) y la de no fraude (*MLP_PseudoProbability_2_A*) para la red global, así como el grupo de clasificación global (tabla 5-37).

MLP_PredictedValue_B	MLP_PseudoProbability_1_B	MLP_PseudoProbability_2_B
1	,104	,896
0	,990	,010
0	,982	,018
0	,983	,017
0	,852	,148
0	,976	,024
0	,939	,061
0	,983	,017
1	,000	1,000
0	,783	,217
1	,000	1,000
1	,151	,849

Tabla 5-37

5.5.2 Análisis de los perfiles de fraude y segmentación de las causas de fraude

La figura 5-36 muestra la densidad de probabilidad de la propensión al fraude global de la red múltiple mediante el Perceptrón Multicapa. Se observa que la probabilidad de fraude es más densa para sus valores pequeños, pero también tiene valores altos alrededor de probabilidaes elevadas.

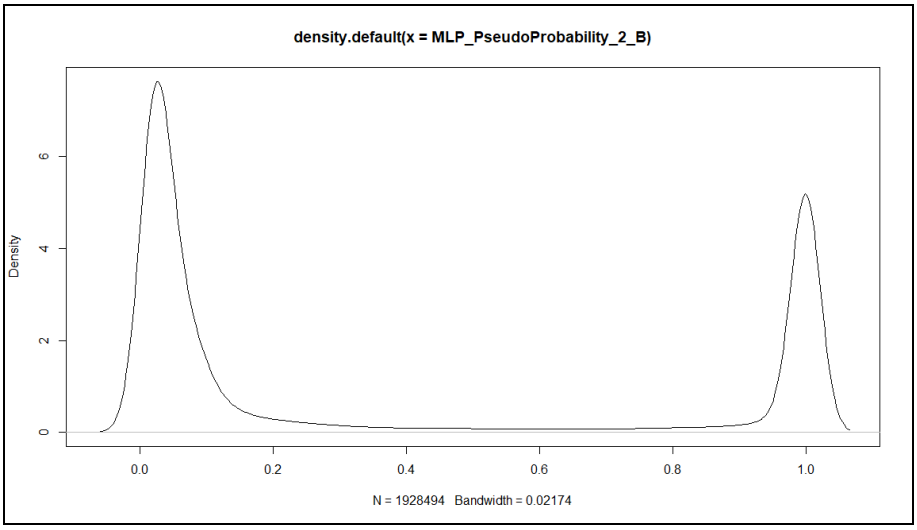


Figura 5-36

En cuanto a la segmentación de las causas de fraude según su incidencia en el fraude global a través de la red neuronal múltiple, observamos los mismos segmentos que en los modelos anteriores en la salida del escalamiento multidimesional (Figura 5-37)

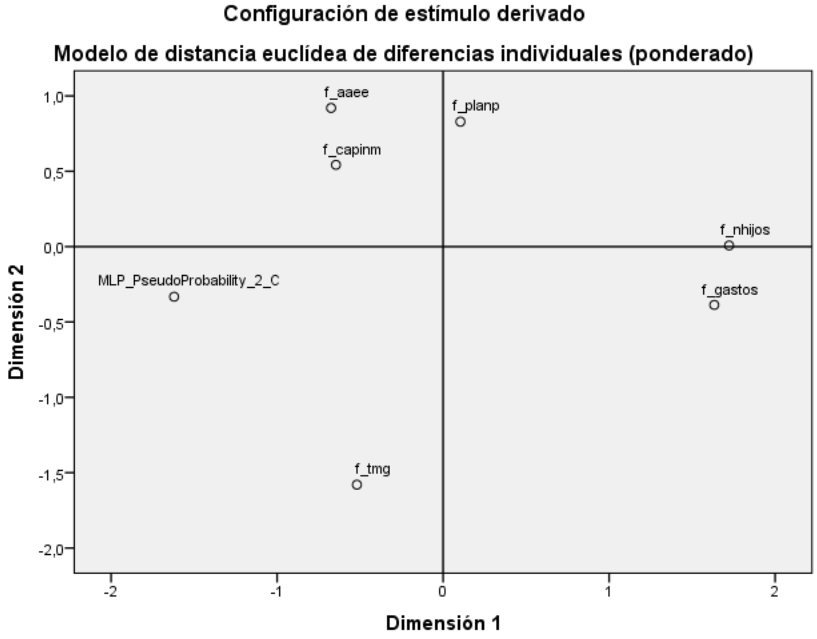


Figura 5-37

El valor pequeño del estadístico Stress y valor alto de RSQ validan con creces el escalamiento como una técnica de segmentación adecuada en este caso.

```
Stress  =   ,08142      RSQ =  ,96114
```

En las figuras 5-38 a 5-43 se muestran las densidades de probabilidad de las diferentes causas de fraude derivadas del ajuste de la red neuronal múltiple.

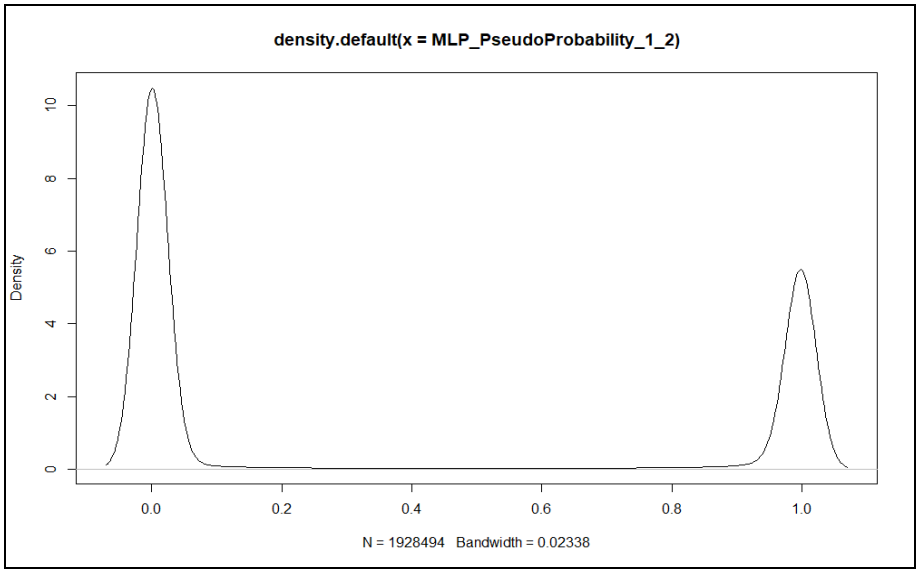


Figura 5-38

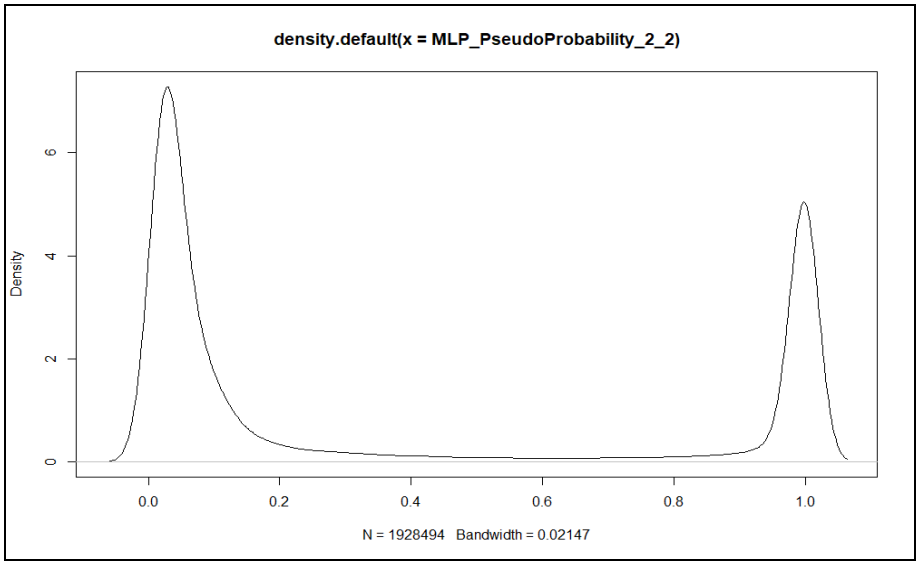


Figura 5-39

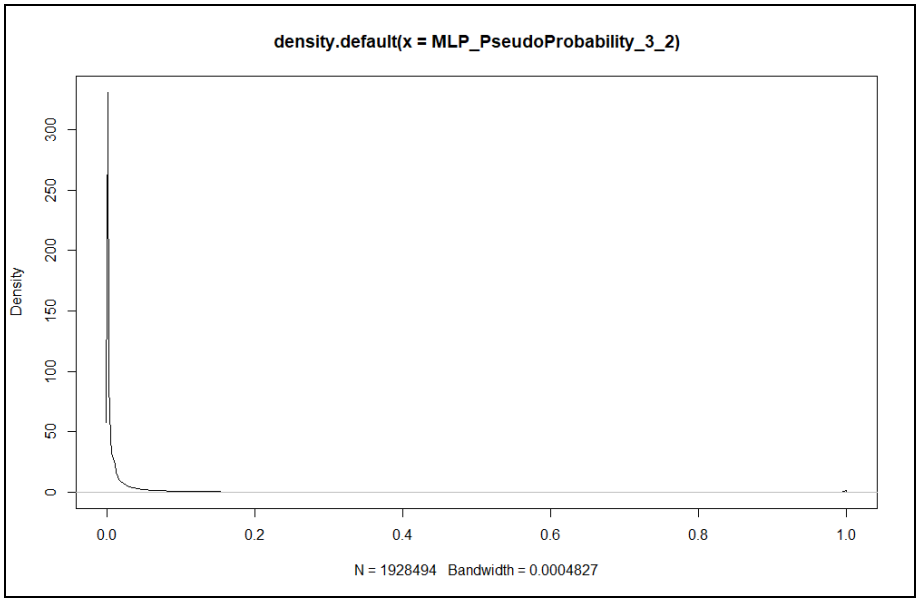


Figura 5-40

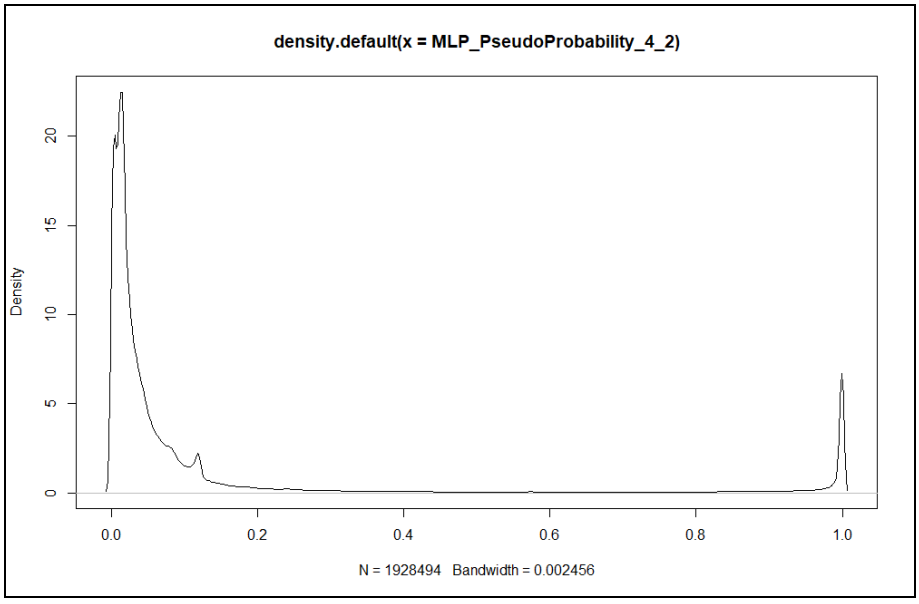


Figura 5-41

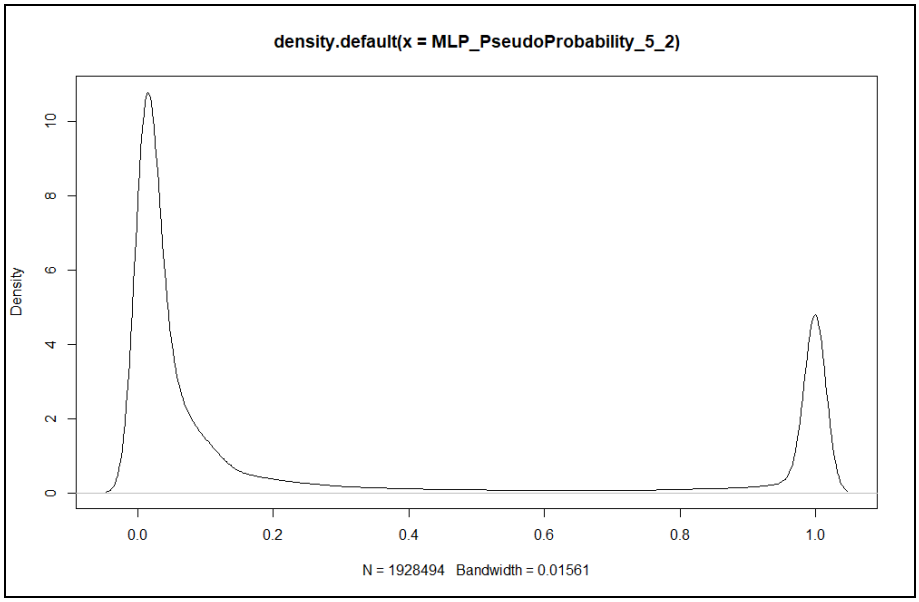


Figura 5-42

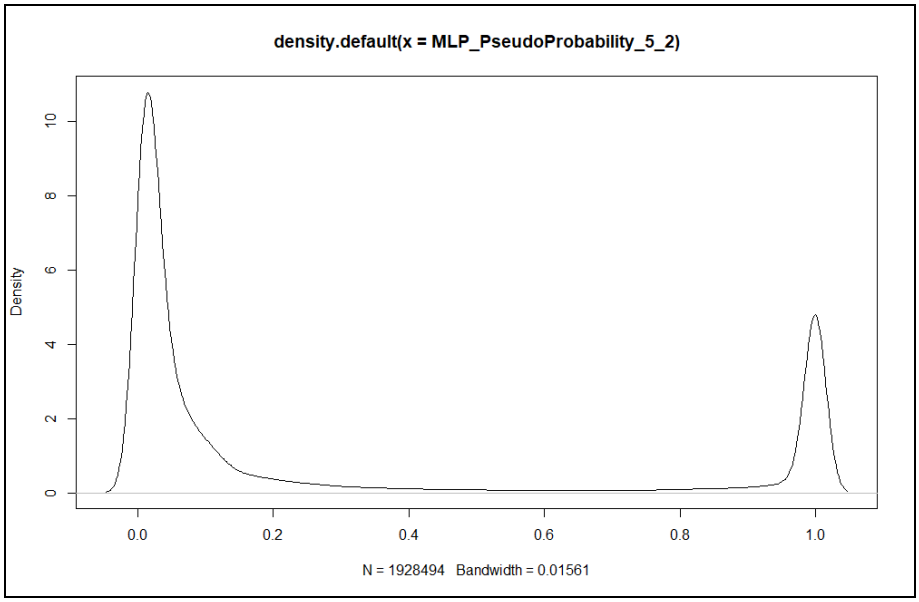


Figura 5-43

La figura 5-44 presenta el escalamiento multidimensional para la segmentación de las diferentes causas de fraude. El gráfico de disparidades de la figura 5-45 y los valores de los estadísticos Stress y RSQ validan el escalamiento.

Con la misma finalidad se muestra el dendograma de la clasificación cluster jerárquica por le método de Ward de las probabilidades de fraude para las distintas causas de fraude (figura 5-46).

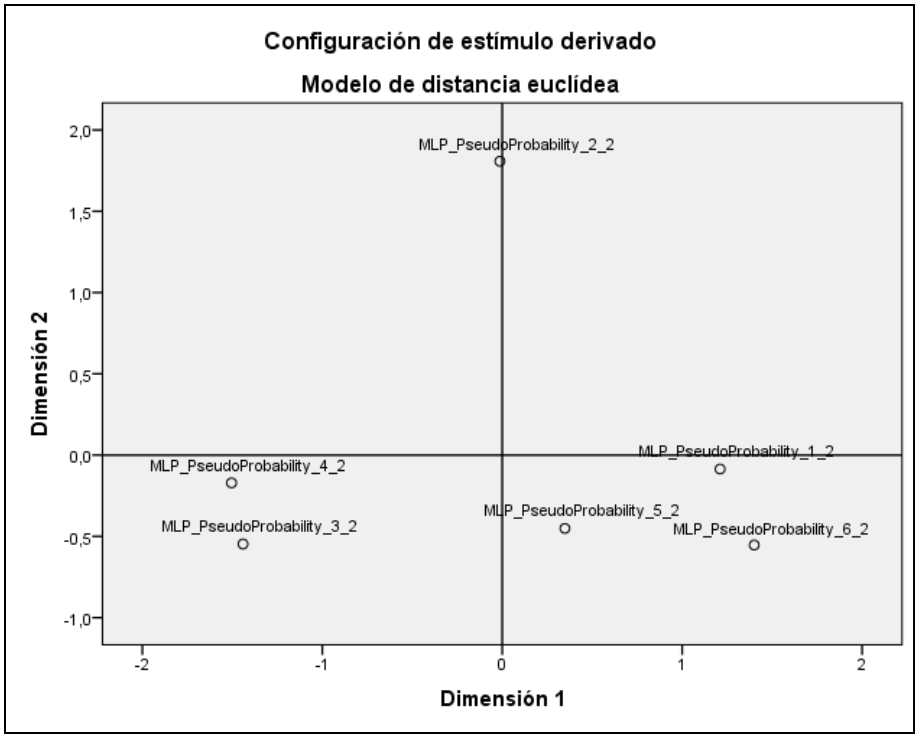


Figura 5-44

Stress = ,03550 RSQ = ,99200

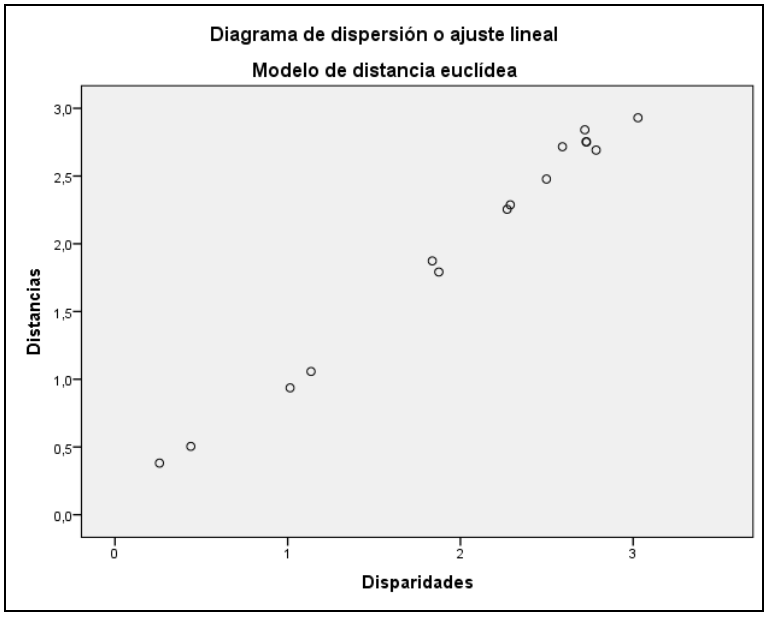


Figura 5-45

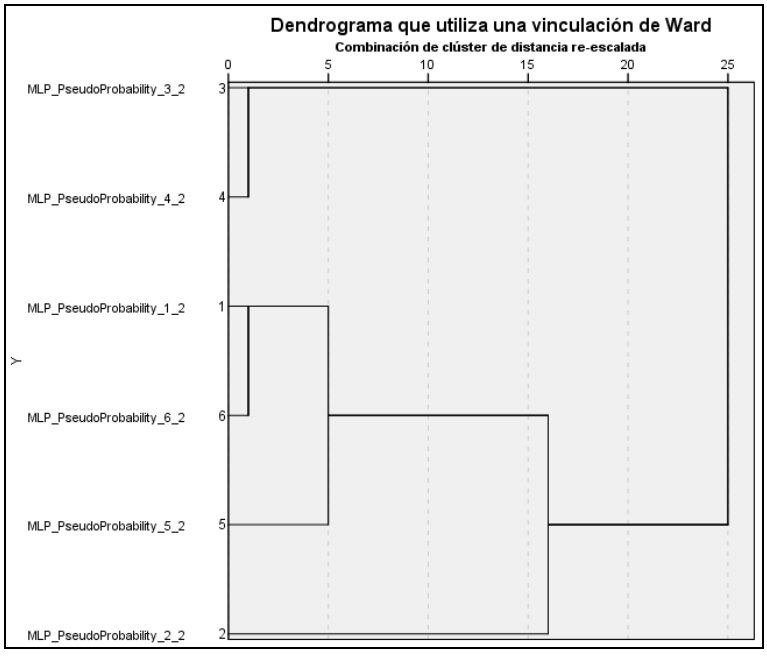


Figura 5-46

5.6 ANÁLISIS DE LA PROPENSIÓN AL FRAUDE A TRAVÉS DE REDES NEURONALES MÚLTIPLES SIN REDUCCIÓN DE LA DIMENSIÓN

También podemos considerar una red neuronal de gran dimensión cuya capa de entrada corresponda a todas las partidas económicas del IRPF sin reducir la dimensión (más de 200 nodos) y cuya capa de salida está formada por un único nodo correspondiente a la variable dicotómica fraude global.

Al ajustar esta red, se obtiene una matriz de confusión con altos porcentajes de acierto (tabla 5-38) y área alta bajo la curva ROC en la tabla 5-39 (0,972), lo que indica que la estimación de la red neuronal es correcta.

La figura 5-47 muestra la curva ROC y las figuras 5-48 y 5-49 muestran el gráfico de ganancias y el gráfico de elevación. Todos estos gráficos indican un buena capacidad predictiva de la red neuroal sin utilizar reducción de la dimensión.

Clasificación

		Pronosticado		
Ejemplo	Observado	0	1	Porcentaje correcto
Entrenamiento	0	483142	21267	95,8%
	1	86568	759056	89,8%
	Porcentaje global	42,2%	57,8%	92,0%
Pruebas	0	206775	9187	95,7%
	1	36942	325557	89,8%
	Porcentaje global	42,1%	57,9%	92,0%

Variable dependiente: fraude global

Tabla 5-38

Área bajo la curva

		Área
fraude global	0	,972
	1	,972

Tabla 5-39

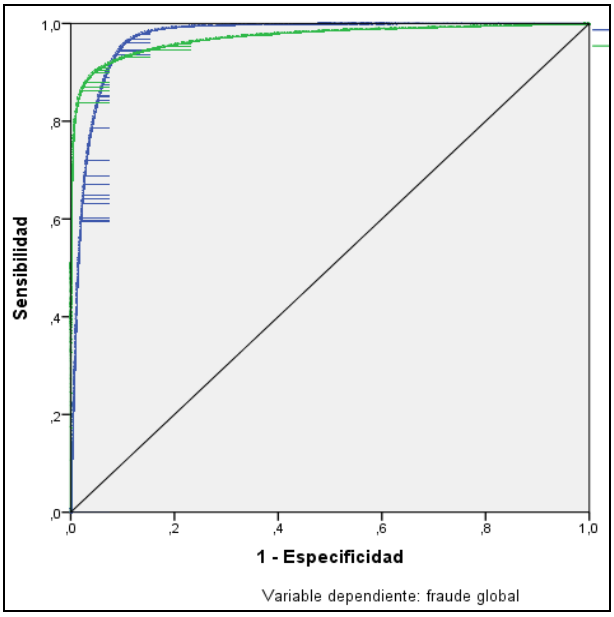


Figura 5-47

A continuación se presenta el gráfico de ganancias y el gráfico de elevación. Ambos indican un buena capacidad predictiva de la red neuroal sin utilizar reducción de la dimensión.

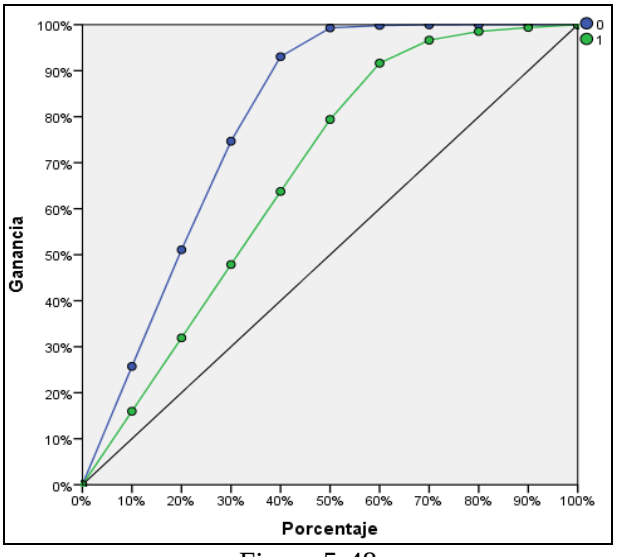


Figura 5-48

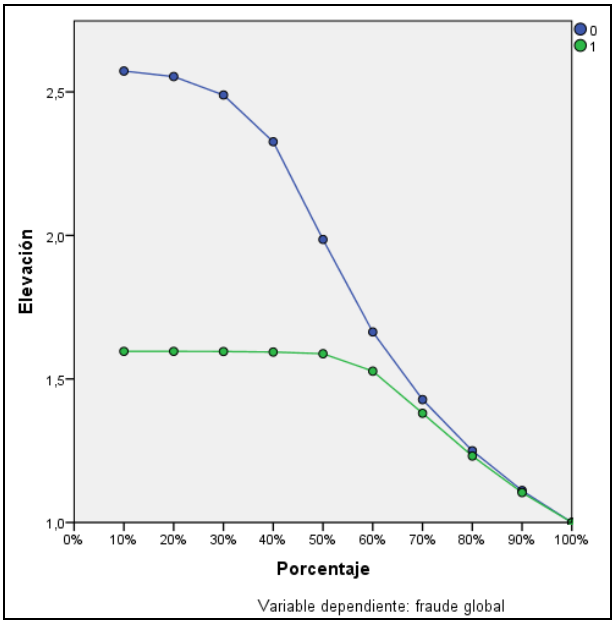


Figura 5-49

En la tabla 5-40 se muestra la importancia de cada partida económica del IRPF sobre el fraude global. Observamos que las partidas con más incidencia están muy relacionadas con las ya estudiadas para todas las técnicas anteriores.

	Importancia	Importancia normalizada
Cuota diferencial	,013	100,0%
Base imponible del ahorro	,012	87,9%
Suma rentas inmob. Imputadas.	,011	85,5%
Reducción de la Base Imponible por aportaciones y contribuciones a sistemas de previsión social	,011	83,9%
Cuota líquida incrementada total	,011	83,1%
Retenciones y pagos a cuenta por actividades económicas	,011	82,8%
Rdto. Neto Reducido Cap.Mobiliario a integrar en la base imponible general.	,011	79,7%
Reducción por arrendamiento de inmuebles destinados a vivienda (artículo 23.2 de la Ley del Impuesto).	,010	74,9%
Atribucion de rentas: Rendimientos capital inmobiliario	,010	74,5%
Saldo positivo de los rendimientos del capital mobiliario a integrar en la base imponible del ahorro	,009	70,1%
Atribucion de rentas: Rendimientos capital mobiliario a integrar en la base imponible del ahorro.	,009	67,8%
Reducc. Base imponible del ahorro por pensiones compensatorias y anualidades por alimentos.	,009	65,9%
Deducc. Por rendimientos derivados de la venta bienes corporales	,009	65,9%

producidos en Canarias, parte estatal		
Rend. Cap. Mobiliario. Intereses de cuentas, depósitos y activos financieros	,009	64,8%
Gravamen autonómico correspondiente a la base liquidable general.	,009	63,8%
Cotizac. Seguridad Social, Mutualidad Funcionarios, detracciones derechos pasivos y Coleg.Huérfanos.	,008	60,5%
Rendimiento neto	,008	59,8%
Rendimiento neto reducido.Trabajo	,008	59,1%
Ingresos íntegros Cap. Mobiliario a integrar en la base imponible general.	,008	59,0%
Imputacion rentas inmobiliarias.	,008	58,8%
Gastos Deducibles	,008	58,2%
Retenciones y pagos a cuenta por arrendamientos de inmuebles urbanos	,008	57,9%
Pagos fraccionados ingresados (actividades económicas).	,008	57,6%
Base liquidable general.	,008	57,5%
Reducc. B I General por aportaciones y contribuciones a sistemas de previsión social (régimen general).	,007	54,8%
Ingresos íntegros Cap.Inmobiliario	,007	54,3%
Rendimiento neto reducido del capital inmobiliario:	,007	53,8%
Saldos netos negativos de ganancias y pérdidas patrimoniales de 2005-2008 a integrar en la parte general de la renta del	,007	53,7%
Otros gastos fi scalmente deducibles.	,007	53,6%
Rend. Cap. Mobiliario. Rendtos. Contratos Seguros Vida o Inv.	,007	52,9%
Compensación de bases liquidables generales negativas de 2005 a 2008.	,007	52,5%
Base imponible general	,007	51,4%
Base liquidable del ahorro.	,007	51,3%
50% de las deducciones por doble adquisición	,007	48,4%
Rend. Cap. Mobiliario. Dividendos y Rendtos. Partic. Fondos Prop.	,006	47,9%
Rdto. Del Trabajo En especie.	,006	46,7%
Suma de rendimientos netos reducidos del capital inmobiliario.	,006	46,5%
Atribucion de rentas: Rendimientos actividades economicas	,006	46,5%
Cuota estatal correspondiente a la base liquidable general	,006	46,2%
Rendimiento neto.Trabajo	,006	46,0%
Rendimiento neto	,006	45,6%
Reducc. B I General por aportaciones a la mutualidad de previsión social de deportistas profesionales.	,006	45,3%
Importe de las deducciones autonómicas de 1998 a 2008 a las que se ha perdido el derecho	,006	44,0%
Rend. Cap. Mobiliario. Rendtos. Transmisión o Amortización Letras Tesoro	,006	43,9%
Intereses de demora de deducciones generales de 1997 a 2008 a las que se ha perdido el derecho.Parte autonómica.	,006	43,7%
Rdto. Neto reducido total act. Econ. Est. Objetiva	,006	43,7%
Otros rendimientos del capital mobiliario a integrar en la base imponible general.	,006	43,6%
Gastos deducibles.	,006	43,4%
Cuota líquida estatal incrementada	,006	43,4%
Deducc. Por incentivos y estímulos a la inversión empresarial, parte autonómica	,006	43,2%
Aportaciones recibidas al patrimonio protegido de las personas con discapacidad del que es titular el contribuyente	,006	42,8%
Cuotas satisfechas a sindicatos	,006	42,7%
Suma deducciones dquisición	,006	42,6%
Gastos deducibles. Importe pendiente de deducir del ejercicio 2008	,006	41,9%

que se aplica en esta declaración		
Reducc. Base imponible del ahorro. Cuotas de afiliación y demás aportaciones a los partidos políticos realizadas por afiliados	,006	41,7%
Base liquidable general sometida a gravamen.	,006	41,6%
Saldo neto de los rendimientos a integrar en la base imponible general y de las imputaciones de renta.	,006	40,9%
Cuota líquida estatal	,005	40,6%
Intereses demora de deducciones autonómicas de 1998 a 2008 a las que se ha perdido el derecho	,005	40,6%
Rend. Cap. Mobiliario. Rendtos. Transmisión o Amortización otros activos	,005	40,5%
Reducc. B I General por pensiones compensatorias y anualidades por alimentos.	,005	40,4%
Saldo neto negativo de ganancias y pérdidas patrimoniales imputables 2009 a integrar en B.I. general	,005	40,2%
Rend. Cap. Mob. Rendtos. Procedentes del arrendamiento de bienes muebles, negocios o minas o de subarrendamientos.	,005	39,8%
Base liquidable del ahorro sometida a gravamen	,005	39,8%
Reducciones Disp.Transitoria 4ª de la Ley del Impuesto.	,005	39,8%
Rdto. Neto Reducido Cap. Mobiliario a integrar en la base imponible del ahorro	,005	39,8%
Saldo neto negativo ganancias y pérdidas patrim. Imput. A 2009 a integrar en B.I. Gral.: imp. Pendte. Compensar 4 ejercicios	,005	39,7%
Cuotas del Impuesto sobre la Renta de no Residentes	,005	39,5%
Retenciones e ingresos a cuenta atribuidos por dquisi especial de atribución de rentas	,005	39,4%
Cuota autonómica o complementaria correspondiente a la base liquidable del ahorro	,005	39,1%
Rdto. Neto Módulos Agrarios	,005	38,5%
Intereses demora de deducciones generales de 1997 a 2008 a las que se ha perdido el derecho. Parte estatal.	,005	38,2%
Rentas exentas del IRPF, excepto para determinar el tipo de gravamen. De la base liquidable general.	,005	38,1%
Reducción de la Base Imponible por pensiones compensatorias al cónyuge y anualidades por alimentos	,005	37,7%
Reducc. B I General por aportaciones y contribuciones a sistemas de previsión social constituidos a favor de personas co	,005	37,4%
Reducción de rendimientos acogidos al régimen especial “33.ª Copa del América”	,005	37,3%
Importe de las deducciones de 1996 y ejercicios anteriores a las que se ha perdido el derecho	,005	37,2%
Reducc. B I General por aportaciones a patrimonios protegidos de personas con discapacidad.	,005	37,1%
Deducc. Por dotaciones a la Reserva para Inversiones en Canarias, parte autonómica	,005	36,6%
Importe del IRPF que corresponde a la Comunidad Autonoma de residencia del contribuyente	,005	36,4%
Cuota líquida autonómica	,005	36,2%
Parte de la adquisición en vivienda habitual que corresponde a la Comunidad Autonoma	,005	35,9%
Cuota autonómica o complementaria correspondiente a la base liquidable general	,005	35,9%
Imputación de rentas derivadas participación Instituciones Inversión Colectiva	,005	35,8%
Deducc. Por rentas obtenidas en Ceuta y Melilla parte autonómica	,005	35,8%
Adecuación del impuesto a las circunstancias personales y familiares.Mínimo del contribuyente.	,005	35,8%

Rend.procedentes de la propiedad industrial que no se encuentre afecta a una actividad económica	,005	35,7%
Deducción por doble imposición internacional,por las rentas obtenidas y gravadas en el extranjero	,005	35,7%
Deducc. Por rentas obtenidas en Ceuta y Melilla parte estatal	,005	35,7%
50% por adquisició de determinados rendimientos del capital mobiliario	,005	35,7%
Retribuciones en especie (valoración)	,005	35,7%
Retenciones e ingresos a cta. Por imputaciones de agrupaciones de interés económico y uniones temporales de empresas	,005	35,6%
Deducc. Doble imposición, régimen imputación de rentas derivadas de la cesión de derechos de imagen	,005	35,6%
Gravamen estatal correspondiente a la base liquidable general.	,005	35,5%
Reducciones Art. 18 apartados 2 y 3, y dispos. Trans. 11ª y 12ª Ley del Impuesto	,005	35,4%
Rend. Cap. Mobiliario. Intereses de activos financieros con bonificación	,005	35,3%
Deducc. Por inversiones o gastos en bienes de interés cultural parte autonómica	,005	35,3%
Adecuación del impuesto a las circunstancias personales y familiares.Mínimo por descendientes.	,005	35,2%
Suma de deducciones autonómicas	,005	35,1%
Deducc. Por dotaciones a la Reserva para Inversiones en Canarias, parte estatal	,005	35,1%
Gastos deducibles. Importe de 2009 pendiente de deducir en los 4 años siguientes.	,005	35,0%
Gastos Deducibles	,005	34,9%
Importe de las deducciones generales de 1997 a 2008 a las que se ha perdido el derecho. Parte estatal.	,005	34,8%
Cuota líquida autonómica incrementada	,005	34,8%
Suma de pagos a cuenta	,005	34,8%
Retenciones a cuenta efectivamente practicadas art. 11 Directiva 2003/48/CE	,005	34,8%
Saldo negativo de los rendimientos del capital mobiliario a integrar en la base imponible del ahorro	,005	34,8%
Reducción de la Base Imponible por aportaciones y contribuciones a sistemas de previsión social del conyuge	,005	34,7%
Resto de los saldos netos negativos de ganancias y pérdidas patrimoniales de 2005-2008 a integrar en la parte general de	,005	34,7%
Retribuciones en especie (ingresos a cuenta)	,005	34,7%
Deducc. Por rendimientos derivados de la venta bienes corporales producidos en Canarias, parte autonómica	,005	34,6%
Deducción por nacimiento o adopción: importe de la deduccion	,005	34,6%
Saldos netos negativos de ganancias y pérdidas patrimoniales de 2005-2008 a integrar en la parte especial de la renta de	,005	34,6%
Parte de las dqui deducciones generales que corresponde a la Comunidad Autonoma	,005	34,6%
Cuota íntegra estatal	,005	34,6%
Deducción por doble imposición de dividendos pendientes de aplicar de 2005 y 2006. Importe que se aplica.	,005	34,6%
Parte de las cuotas integras del ejercicio 2009 que corresponde a la Comunidad Autonoma	,005	34,5%
Retribuciones en especie (ingresos a cuenta repercutidos)	,005	34,5%
Deducc. Doble imposición internaci. Habiendo aplicado el régimen de transp. Fiscal internacional	,005	34,5%
Retenciones y pagos a cuenta por rendimientos del trabajo	,005	34,4%
Deducc. Por incentivos y estímulos a la inversión empresarial, parte	,005	34,4%

estatal		
Deducc. Por cantidades o bienes donados a determinadas entidades parte autonómica	,005	34,4%
Imputaciones de rentas positivas en el régimen de transparencia fiscal internacional	,005	34,3%
Adecuación del impuesto a las circunstancias personales y familiares.Mínimo personal y familiar.	,005	34,3%
Retenciones deducibles correspondientes a rendimientos bonificados	,005	34,3%
Cuota resultante de la autoliquidación	,005	34,3%
Rend.procedentes de la propiedad intelectual cuando el contribuyente no sea el autor	,005	34,3%
Ingresos a cuenta del artículo 92.8 de la Ley del Impuesto	,005	34,2%
Saldo neto positivo de ganancias y pérdidas patrimoniales imputables a 2009 integrar en B.I.del ahorro	,005	34,2%
Saldo neto positivo de ganancias y pérdidas patrimoniales imputables a 2009 a integrar en B.I. general	,005	34,2%
Incrementos de la cuota liquida dquisició por perdida del derecho a determinadas deducciones en ejercicios anteriores	,005	34,1%
Reducción de la Base Imponible por aportaciones y contribuciones a sistemas de previsión social de personas discapacitad	,005	34,1%
Deducc. Por cantidades depositadas en cuentas ahorro-empresa parte autonómica	,005	34,1%
Rentas exentas del IRPF, excepto para determinar el tipo de gravamen. De la base liquidable del ahorro.	,005	34,1%
Saldo neto negativo de rendimientos del capital mobiliario de 2007 y 2008 a integrar en la base imponible del ahorro	,005	34,1%
Importe de las deducciones generales de 1997 a 2008 a las que se ha perdido el derecho. Parte autonómica.	,005	33,9%
Rendimiento mínimo computable en caso de parentesco (Art. 24 de la ley del impuesto)	,005	33,9%
Deducc. Por cantidades o bienes donados a determinadas entidades parte estatal	,005	33,9%
Deducc. Por inversiones o gastos en bienes de interés cultural parte estatal	,005	33,8%
Cuota íntegra autonómica o complementaria	,005	33,8%
Imputación de rentas por la cesión de derechos de imagen	,005	33,8%
Anualidades por alimentos en favor de los hijos satisfechas por resolución judicial.	,005	33,8%
Reducción de la Base Imponible por aportaciones a Mutualidades de Previsión Social de deportistas profesionales	,005	33,8%
Rend.procedentes de la prestación de asistecia técnica, salvo en el ámbito de una actividad económica	,005	33,7%
Rdto. Del trabajo Dinerarios	,005	33,6%
Deducc. Por cantidades depositadas en cuentas ahorro-empresa parte estatal	,005	33,6%
Por ganancias patrimoniales, incluidos premios	,005	33,5%
Rdto. Neto reducido total act. Econ. Est. Directa	,004	33,4%
Reducciones Art. 26.2 de la Ley del Impuesto.	,004	33,2%
Cuota estatal correspondiente a la base liquidable del ahorro	,004	33,2%
Gastos deducibles. Importe de 2009 que se aplica en esta declaración.	,004	33,0%
Compensación fiscal por percepción de rdtos. Del capital mob. Con período de generación superior a dos años	,004	32,4%
Deduccion por maternidad: cantidades percibidas en concepto de abono anticipado	,004	32,3%
Reducción de la Base Imponible por aportaciones a los patrimonios	,004	31,9%

protegidos de las personas con discapacidad		
Deduccion por maternidad: importe de la deduccion	,004	31,9%
Deduccion por nacimiento o adopción: cantidades percibidas en concepto de abono anticipado	,004	31,7%
Deducc. Por alquiler de la vivienda habitual	,004	31,5%
Gravamen autonómico correspondiente a la base liquidable general. Importe mínimo personal y familiar.	,004	31,1%
Reducc. B I General por aportaciones a sistemas de previsión social de los que es partícipe, mutualista o titular el cón	,004	30,1%
Atribucion de rentas: Rendimientos capital mobiliario a integrar en la base imponible general.	,004	29,7%
Retenciones y pagos a cuenta por rendimientos del capital mobiliario	,004	29,3%
Reducción por rendimientos generados en más de 2 años u obtenidos de forma notoriamente irregular (art. 23.3 de la Ley d	,004	28,6%
Gravamen estatal correspondiente a la base liquidable general. Importe mínimo personal y familiar.	,004	28,4%
Reducción por obtención rdto. Trabajo.Incremento para contrib. Desempleados que acepten un puesto que exija traslado de	,004	28,0%
Importe del mínimo personal y familiar que forma parte de la base liquidable general.	,004	27,9%
Gastos de defensa jurídica derivados directamente de litigios con el empleador (máximo: 300 euros anuales)	,004	27,9%
Reducción por obtención rdto. Trabajo.Incremento para trabajadores activos mayores de 65 años que dquisici o prolonguen	,004	27,6%
Reducción de la Base Imponible por tributación conjunta	,004	27,5%
Contribuciones Planes Pensiones.	,003	25,1%
Ingresos íntegros Cap. Mobiliario a integrar en la base imponible del ahorro.	,003	24,5%
Rendimiento neto	,003	23,9%
50% de la dquisición fiscal por dquisici en dquisición de vivienda habitual	,003	22,9%
Adecuación del impuesto a las circunstancias personales y familiares.Mínimo por ascendientes.	,003	22,2%
Importe del mínimo personal y familiar que forma parte de la base liquidable del ahorro.	,003	21,6%
Compensación fiscal por deducción en adquisición de vivienda habitual, para viviendas adquiridas antes del 20-01-2006	,003	21,3%
Imputación de entidades en régimen de transparencia fiscal	,003	20,7%
Reducción por obtención rdto. Trabajo.Reducción adicional para trabajadores activos que sean presonas con discapacidad.	,003	20,6%
Deducc. Por adquisición o rehabilitación de la vivienda habitual, parte estatal	,003	19,9%
Reducc. B I General Cuotas de afiliación y demás aportaciones a los partidos políticos realizadas por afiliados, adherid	,003	19,2%
Adecuación del impuesto a las circunstancias personales y familiares.Mínimo por discapacidad.	,002	18,1%
Deducc. Por adquisición o rehabilitación de vivienda habitual, parte autonómica	,002	17,8%
Cuotas satisfechas a colegios profesionales (si la colegiación es obligatoria y con un máximo de 500 euros anuales)	,002	14,8%
Total ingresos integros computables	,002	14,5%
Rend. Rentas que tengan por causa la imp. Cap. Y otros rend. Del cap. Mob. A integrar en la base imp. Ahorro	,002	14,5%
Reducc. Base imponible del ahorro por tributación conjunta.	,001	11,1%
Reducción por obtención rdto. Trabajo.Cuantía aplicable con	,001	8,0%

carácter general.		
Reducc. B I General por tributación conjunta.	,001	7,4%
Resultado de la declaración	,000	3,7%
Deducción por obtención de rendimentos del trabajo o de actividades económicas.	,000	3,3%

Tabla 5-40

El valor añadido fundamenal de esta red es precisamente que desciende hasta el nivel partida original del IRPF sin transformar, para calcular su importancia sobre el fraude global. Esto es prácticamente imposible de conseguir con cualquier otra técnica.

Lo mismo que para las restantes técnicas, se calcula la probabilidad de fraude global de cada contribuyente (*MLP_PseudoProbability_2_C*), la de no fraude ((*MLP_PseudoProbability_1_C*) y el grupo de fraude en el que se clasifica (*MLP_PredictedVaue_C*) como muestra la tabla 5-41.

MLP_PredictedValue_C	MLP_PseudoProbability_1_C	MLP_PseudoProbability_2_C
1	,058	,942
1	,000	1,000
0	,934	,066
0	,941	,059
0	,796	,204
0	,935	,065
1	,140	,860
1	,464	,536
1	,000	1,000

Tabla 5-41

5.7 EXTRACCIÓN DEL CONOCIMIENTO Y CONCLUSIONES

La detección del fraude fiscal y su cuantificación es uno de los objetivos más importantes a llevar a cabo por las Administraciones Tributarias de los distintos países.

Las metodologías para la detección del fraude y de la evasión fiscal establecidas tanto por el Banco Mundial como por la Comisión Europea se basan fundamentalmente en métodos cuantitativos y se dividen en dos

grandes grupos: el enfoque de "abajo a arriba" (botton-up) que parte de datos de declarantes y establece perfiles que permiten asignar probabilidades de fraude a individuos o empresas) y el enfoque de "arriba abajo" (top-down) que parte de variables agregadas a nivel de país y estima cuanto debería recaudarse si todo el mundo cumpliese la regulación, es decir, estima la brecha fiscal o Fiscal Gap.

En este capítulo se han utilizado técnicas de "arriba abajo" (top-down) para estudiar el fraude fiscal en el Impuesto sobre la Renta de las Personas Físicas mediante el uso de herramientas predictivas avanzadas de Machine Learning y en concreto mediante modelos lineales generalizados y modelos de redes neuronales.

Los modelos de redes neuronales habitualmente perfeccionan y superan en capacidad predictiva al resto de los modelos predictivos lineales y no lineales.

Cuantitativamente suelen consistir en combinaciones óptimas de linealidades y no linealidades que permite predecir mejor y realizar estimaciones mucho más precisas que otros tipos de modelos.

A cambios de estas bondades hay que pagar un precio en términos de disponibilidad de software (no es muy común el software que habilita estas técnicas), de capacidad de computación (los algoritmos de redes son complicados y necesitan de recursos de hardware adecuados para su convergencia) y de control metodológico (las técnicas de machine Learning no son triviales en cuanto a metodología).

Un valor añadido muy importante de este trabajo es la utilización de bases de datos provenientes de fuentes administrativas oficiales.

Ello conlleva que no hay problemáticas de datos missing ni de otras imperfecciones en los datos.

La información que se utiliza son los datos oficiales de la Agencia Tributaria Española que están perfectamente depurados y explorados, ya que son la base del análisis tributario y de la inspección de contribuyentes.

En concreto, la muestra de IRPF que se utiliza en este trabajo es el instrumento fundamental para el análisis fiscal en la Administración española, en la Academia y en el sector privado.

Por otro lado, el uso de estos conjuntos de datos tan ricos y extensos también tiene un precio en términos de computación.

Utilizar modelos tan complejos sobre una cantidad de datos tan elevada nos obliga a introducirnos en el mundo del Big Data.

Se ha utilizado software y hardware de IBM y SAS que ha permitido llegar a la convergencia de algoritmos de modelos lineales generalizados y redes neuronales aplicados a millones de datos y a cientos de variables.

Por otra parte, cualquier representación gráfica que involucra millones de puntos no puede realizarse sin una infraestructura de grandes datos.

Otra característica destacable de este trabajo consiste en que la metodología utilizada es generalizable para cuantificar la propensión al fraude en cualquier otro impuesto según las causas que lo determinan.

El hecho de disponer de grandes conjuntos de datos con información relativa a cada impuesto permite utilizar una metodología genérica para ampliar las posibilidades de análisis cuantitativo y utilizar las nuevas prestaciones que aportan el Big Data, la Minería de Datos y las técnicas de

Machine Learning. Los Modelos Lineales Generalizados y de Redes Neuronales aplicados a las muestras de IRPF en este trabajo con la finalidad de estudiar las variables más incidentes que afectan al fraude fiscal en este impuesto y cuantificar la propensión al fraude de los declarantes, son aplicables a cualquier otro impuesto disponiendo de la base de datos de declarantes del mismo. La Agencia Tributaria Española dispone de estas bases de datos.

El elevado número de variables en el modelo ha llevado al uso de técnicas de reducción de la dimensión para disminuir el número de variables independientes de los modelos de redes, con la finalidad de facilitar su convergencia y simplificar la interpretación de los resultados.

Asimismo, la reducción de la dimensión elimina problemas de multicolinealidad en los modelos, ya que las variables reducidas son incorreladas matemáticamente.

Por otra parte, como las variables reducidas son combinaciones lineales de las variables iniciales, se aminora mucho el problema de los datos atípicos, que como sabemos son enemigos acérrimos para la convergencia de cualquier tipo de modelo econométrico.

Además, las variables reducidas son asintóticamente normales, lo que facilita el trabajo con los modelos, aunque no sea una hipótesis estrictamente necesaria. Por otra parte, las variables reducidas tienen una escala similar, lo que lleva a que los datos sean uniformes en cuanto a medida. Este hecho elimina posibles problemas de no normalidades residuales y de falta de aleatoriedad residual.

Por otro lado, las variables reducidas introducen confidencialidad en la información. Los datos originales pueden presentar problemas de identificación de contribuyentes perdiéndose la confidencialidad

estrictamente necesaria en el tratamiento de estos datos. Pero las variables reducidas, al ser combinación lineal de las variables iniciales imposibilitan la identificación de contribuyentes e introducen un alto grado de confidencialidad.

Los Modelos Lineales Generalizados considerados para este problema predictivo han sido el Modelo Logit y el Modelo Probit. El modelo de Red Neuronal considerado ha sido el Perceptrón Multicapa. Después de estimar ambos tipos de modelos, una diagnosis bastante amplia nos llevó a concluir que el tipo de modelo más adecuado a utilizar es la Red Perceptrón Multicapa. Su capacidad predictiva es superior a la del resto de modelos y su poder discriminatorio en cuanto a las causas de fraude es mucho más alto.

El análisis de las causas de fraude derivado de la estimación de la Red Perceptrón Multicapa sitúa como influyente en el fraude fiscal en IRPF a los rendimientos de capital inmobiliario (rendimientos provenientes de la titularidad de bienes rústicos y urbanos o de derechos reales sobre ellos). La computación de estos rendimientos suele ser fraudulenta. Lo mismo ocurre con la computación de los gastos deducibles y reducciones para obtener los rendimientos netos del capital inmobiliario. Suelen inflarse los gastos necesarios para la obtención de los rendimientos.

También es una partida muy influyente los rendimientos, bases y cuotas. Los rendimientos ocultos no declarados suelen ser la causa principal de fraude. Estos rendimientos ocultos llevan a la manipulación de las bases imponible y liquidable y por tanto a la minoración del tipo aplicable. Por lo tanto la cuota resultante es inferior a la que debería de ser. Suele ser habitual la presencia de actividades cuyas rentas eluden la tributación, bien por no ser declaradas o bien por no estar registradas constituyendo economía sumergida. De esta forma, el tipo marginal correspondiente a la declaración resulta inferior al real, manipulándose así el resultado de la liquidación. Las cuantías defraudadas por esta causa suelen ser de elevada magnitud.

El mínimo personal y familiar constituye la parte de la base liquidable que, por destinarse a satisfacer las necesidades básicas personales y familiares del contribuyente, no se somete a tributación en el IRPF. Los mínimos por descendientes, ascendientes y discapacidad también deben de ser objeto de especial vigilancia. Suelen ser habitual el fraude que afecta a las declaraciones del número de hijos y ascendientes y descendientes, que habitualmente eran simultáneamente desgravadas por los dos padres (separados, divorciados o en otras situaciones) en el caso de los hijos o por diferentes hermanos en el caso de los ascendientes.

La imputación de rentas en paraísos fiscales y la compensación de bases liquidables generales negativas también son partidas susceptibles de fraude.

Mención especial merecen las actividades económicas. En general, la incorrecta declaración de deducciones derivadas de actividades económicas suele ser otra fuente de fraude. También es necesario vigilar el cálculo de los rendimientos íntegros de actividades económicas, la correcta aplicación de las reglas generales del cálculo del rendimiento neto, los elementos patrimoniales afectos a la actividad económica, las normas para la determinación del rendimiento neto en estimación directa y objetiva y las reducciones.

Las reducciones de la base imponible también suelen ser susceptibles de fraude y más en concreto las reducciones por aportaciones a sistemas de previsión social entre las que se encuentran incluidas especialmente las aportaciones realizadas a planes de pensiones. Esta rúbrica del IRPF fue durante un tiempo el refugio de las rentas altas, ya que desgrava de la base y además por cantidades importantes hasta que se acotó el máximo deducible.

Por lo tanto, era objeto de especial tratamiento por los declarantes de IRPF con peligro de deducciones fraudulentas ilegales que acentuó la vigilancia de la inspección. Estas reducciones de base imponible deben de ser objeto de especial vigilancia porque inciden en el tipo a aplicar. Una minoración del tipo es muy incidente en el resultado de la declaración.

Las atribuciones de renta por rendimientos de capital mobiliario, inmobiliario y actividades económicas también constituyen partidas a vigilar.

También es necesario vigilar las deducciones y reducciones más influyentes en el fraude fiscal en IRPF. Observamos que las deducciones autonómicas, las deducciones por incentivos y estímulos a la inversión y las deducciones por vivienda habitual suelen ser problemáticas. Las deducciones por vivienda habitual que todavía perduran dependen del año de construcción del inmueble y este hecho debe de ser objeto de especial vigilancia. El mínimo personal y familiar constituye la parte de la base liquidable que, por destinarse a satisfacer las necesidades básicas personales y familiares del contribuyente, no se somete a tributación en el IRPF. Los mínimos por descendientes, ascendientes y discapacidad también deben de ser objeto de especial vigilancia. Suele ser habitual el fraude que afecta a las declaraciones del número de hijos y ascendientes y descendientes, que habitualmente eran simultáneamente desgravadas por los dos padres (separados, divorciados o en otras situaciones) en el caso de los hijos o por diferentes hermanos en el caso de los ascendientes. Otra partida a vigilar son los gastos deducibles totales y los límites de determinadas deducciones con especial referencia a las deducciones por incentivos a la inversión. También es necesaria la vigilancia de las cotizaciones.

Otras partidas susceptibles de fraude suelen ser: retribuciones en especie, rentas exentas de IRPF, deducciones por doble imposición, ganancias agrarias, arrendamiento de inmuebles con gastos fiscales

deducibles, la imputación de rentas en paraísos fiscales y la compensación bases liquidables.

Resumiendo un poco el análisis de la importancia de las partidas del impuesto (variables independientes del modelo) sobre el fraude fiscal, vemos que las partidas relativas a rendimientos son las más incidentes en el fraude y en concreto los rendimientos del capital mobiliario e inmobiliario y los rendimientos de actividades económicas.

El otro grupo de partidas más incidentes en el fraude son los gastos deducibles y las reducciones con especial incidencia de las deducciones autonómicas, las deducciones por vivienda, las deducciones por incentivos a la inversión y los gastos deducibles totales. Esta última partida implica la vigilancia de cualquier tipo de gasto deducible.

En cuanto a las reducciones, serán objeto de vigilancia cualquier tipo de reducción de la base imponible, con especial interés en las reducciones por aportaciones a sistemas de previsión social entre las que se encuentran incluidas especialmente las aportaciones realizadas a planes de pensiones.

No hay que olvidarse tampoco de la vigilancia de los mínimos personales y familiares, por descendientes y ascendientes y por discapacidad.

Por otra parte, todas las técnicas consideradas, ofrecen como salida la clasificación de cada declarante como fraudulento o no fraudulento y adicionalmente muestra las propensiones al fraude de cada declarante basadas en las ecuaciones estimadas de los correspondientes modelos predictivos. Es decir, no sólo clasifican los individuos como propensos o no al fraude, sino que también computan la probabilidad de fraude de cada declarante (de la muestra, de la población y nuevos declarantes). Este hecho es especialmente importante para la labor inspectora. A la hora de

establecer planes de inspección pueden tomarse decisiones de incluir en el plan a todos aquellos declarantes que superen una determinada cota de probabilidad de fraude o al menos una muestra de los mismos en caso de que los recursos para la inspección no sean suficientes.

También conviene destacar que las ecuaciones de los modelos predictivos para asignar probabilidades de fraude a los individuos son válidas para varios ejercicios consecutivos, no siendo necesario realizar estimaciones cada año, salvo en el caso de cambios legislativos profundos en el impuesto.

La representación de la densidad de probabilidad de la propensión al fraude mediante el Perceptrón Multicapa (la mejor de las técnicas, según habíamos visto anteriormente) muestra que la probabilidad de fraude es más densa para valores pequeños del mismo, lo cual es lógico, ya que siempre hay muchos más declarantes no defraudadores que defraudadores. Pero a partir de probabilidades de fraude superiores a 0,5 vemos que la densidad es creciente hasta valores cercanos a la probabilidad de fraude 0,8 o 0,9. Este hecho indica que existe una bolsa de fraude no despreciable con valores altos de probabilidad de fraude. No deja de ser curioso que la densidad de fraude sea más alta, tanto para valores muy pequeños de fraude, como valores altos de fraude. Podríamos hablar entonces de una polarización de la propensión al fraude.

Capítulo 6

INVESTIGACIÓN DEL FRAUDE FISCAL CON TÉCNICAS DE MACHINE LEARNING. REDES NEURONALES BAYESIANAS Y MÉTODO kNN

6.1 INTRODUCCIÓN

Este capítulo tiene como finalidad estudiar los factores que afectan al fraude fiscal (causas de fraude) en el Impuesto sobre la Renta de las Personas Físicas y cuantificar, ordenar y probabilizar su incidencia a través de técnicas especializadas, como son las Redes Neuronales Bayesianas y el método kNN (*k-nearest neighbours*) o vecino más cercano. Estas técnicas utilizan las prestaciones que aporta el Big Data y permiten la computación con millones de datos, como es el caso que nos ocupa. Como en el resto de las técnicas, a partir de una muestra anual de IRPF se construirán modelos de Redes Bayesianas que cuantificarán la incidencia de las distintas causas que delimitan el fraude fiscal en IRPF basándose exclusivamente en la información que se declara a la Agencia Tributaria en el modelo correspondiente a este impuesto. Asimismo, se elaborarán modelos predictivos tipo kNN que permiten cuantificar la probabilidad que tiene cualquier contribuyente de ser defraudador por cada factor de fraude una vez que presente su declaración de IRPF. Con en el resto de las técnicas,

estos modelos permitirán segmentar a los declarantes del impuesto por nivel de propensión al fraude y causas del mismo. Esta metodología es generalizable para cuantificar la propensión al fraude en cualquier otro impuesto según los factores que lo determinan y para cuantificar y ordenar la incidencia de dichos factores en el fraude.

6.2 MARCO METODOLÓGICO: LAS REDES NEURONALES BAYESIANAS

Una red bayesiana es un modelo gráfico probabilístico que ofrece como salida un grafo acíclico dirigido que representa un conjunto de variables (nodos) y sus dependencias condicionales probabilísticas (codificado en sus arcos). Los nodos pueden representar cualquier variable. Existen algoritmos eficientes que realizan inferencia y aprendizaje en las redes neuronales bayesianas con la finalidad de predecir en qué categoría de la variable dependiente cualitativa del modelo se clasifica un individuo para el que conocemos los valores de los predictores del modelo, habitualmente también cualitativos.

Podríamos considerar una red bayesiana como un grafo dirigido, en el cual los diferentes nodos se conectan mediante relaciones de implicación que poseen una determinada direccionalidad, y con un algoritmo de modificación de los pesos de conexión basado en el Teorema de Bayes.

Hablaremos de una "configuración" refiriéndonos a redes bayesianas como a un conjunto determinado de sus pesos de conexión. Esto es,

$$\Phi_i \rightarrow \left(\omega_{A,i}, \omega_{B,i}\right)$$

Con una configuración determinada, podríamos introducir los valores iniciales y la red nos proporcionaría una salida. Una idea

fundamental que hay que tener en cuenta en una red neuronal bayesiana es que la red sería todo el conjunto de posibles configuraciones, cada una con su respectiva probabilidad. Esto es, una red bayesiana, a diferencia de una red neuronal, no es una única configuración de pesos de conexiones, sino el conjunto de todas las posibles configuraciones, asociada cada una con una cierta probabilidad. En un modelo continuo, tendríamos, por ejemplo:

$$\Phi = p_1\Phi_1 + p_2\Phi_2 + \ldots + p_{10}\Phi_{10}$$

Aprender significa modificar las probabilidades de cada máquina. Esto es, no se modifican en sí mismos los valores de los pesos de conexión entre los nodos. Las diferentes posibilidades siguen existiendo. Lo que se modifican son las diferentes probabilidades de cada máquina. ¿Cómo se modifican estas probabilidades a lo largo del entrenamiento de la red, según se van presentando los diferentes ensayos? La forma de modificar las diferentes probabilidades viene determinada por el Teorema de Bayes:

$$p_i^{n+1} = \frac{p_i^n \Phi_i}{\sum_j p_j^n \Phi_j}$$

Estas diferentes conceptualizaciones entre las redes bayesianas y las redes neuronales (existencia de diferentes máquinas y aprendizaje como variación de las probabilidades de cada máquina en el caso de redes bayesianas y existencia de una máquina concreta y aprendizaje como variación de los parámetros de esa máquina en el caso de redes neuronales) plantean un problema a la hora de buscar un framework básico que englobe a ambas. De hecho, el funcionamiento de una red bayesiana se puede asociar a una red neuronal determinada. Pero el aprendizaje ya no se puede asociar, porque se perdería información: diferentes redes bayesianas se podrían asociar con la misma red neuronal, y sin embargo aprenderían de forma

diferente. Cómo resolver este punto clave podría llevarnos a construir un modelo globalizador entre ambos tipos de redes.

Podríamos clasificar las redes bayesianas como técnicas estadísticas de la dependencia o técnicas predictivas que se caracterizan por el hecho de que alguna (o algunas) de las variables en estudio destaca como dependiente principal.

En los modelos predictivos (métodos de la dependencia) subyace una relación general de dependencia entre las variables independientes $x_1, x_2, \cdots, x_n$ y las dependientes $y_1, y_2, \cdots, y_n$ del tipo genérico:

$$G(y_1, y_2, \cdots, y_n) = F(x_1, x_2, \cdots, x_n)$$

La naturaleza de las variables caracterizará cada modelo.

Igual que en el caso de una árbol de decisión, también podría considerarse que en una red neuronal de tipo bayesiano se analiza la relación entre una variable dependiente (o endógena) no métrica (cualitativa) y varias variables independientes (o exógenas) no métricas (cualitativas).

Al igual que en el caso de la regresión logística, el análisis discriminante, los árboles de decisión y la redes neuronales, la finalidad de una red neuronal bayesiana es predecir la categoría de la variable dependiente cualitativa en la que se clasifican los individuos según los valores de sus variables independientes cualitativas.

Las redes bayesianas son modelos predictivos, que trata de resolver los problemas de discriminación en una población segmentando de forma progresiva la muestra para obtener finalmente una clasificación fehaciente en

grupos homogéneos, según la variable de interés denominada variable de segmentación (variable dependiente del modelo).

En la redes bayesianas la segmentación de la población se realiza según los valores de la variable de interés que juega el papel de variable dependiente del modelo predictivo subyacente en la red (variable cualitativa). La asignación de un elemento poblacional a un segmento se realiza de acuerdo a los valores de determinadas variables medidas sobre él que constituyen las variables independientes del modelo (habitualmente también variables cualitativas, aunque también suelen utilizarse variables cuantitativas con sus valores agrupados en un número pequeño de intervalos).

En nuestro caso utilizaremos una red bayesiana que explica el fraude global (variable dependiente) en función de las causas o factores de fraude más comunes en el Impuesto sobre la Renta de las Personas Físicas (variables independientes). Se ordenarán las distintas causas de fraude de acuerdo a su incidencia sobre el fraude global. Posteriormente segmentaremos los individuos por su propensión al fraude calculando la probabilidad que tiene cada individuo de ser defraudador.

6.3 DATOS: FASES DE SELECCIÓN, EXPLORACIÓN Y TRANSFORMACIÓN DE LA INFORMACIÓN

Al igual que en el caso de los árboles de decisión, se utiliza como fuente de datos la muestra del Impuesto sobre la Renta de las Personas Físicas (IRPF) que proporciona el Instituto de Estudios Fiscales (IEF). Hay que seguir teniendo presente que disponer de una estructura de hardware y software que implemente procesamiento de grandes datos (*Big Data*) es esencial en nuestro caso. De esta forma, las técnicas de Machine Learning que vamos a utilizar, en este caso las Redes Neuronales Bayesianas, se encuadran dentro de las técnicas de *Big Data Analytics*.

Recordamos que se seleccionan más de 2 millones de registros de la población total de declarantes mediante muestreo estratificado por provincias, tramos de renta y fuente de renta utilizando afijación proporcional. Este método de selección expande adecuadamente la muestra por toda la población resultando muy significativa. Esta muestra alimentará nuestros modelos de redes neuronales bayesianas. La muestra de IRPF aporta cerca de 300 variables, tanto partidas económicas como otras variables numéricas y no numéricas, todas ellas contenidas como casillas en el modelo 100 de declaración del IRPF.

Asimismo, la base de datos contiene otro tipo de variables numéricas y no numéricas relativas a sexo, edad, provincia, estado civil… de los declarantes. Estas variables permitirían estudiar el fraude con análisis de género, análisis geográfico, etc.

Como variables independientes de los modelos de redes bayesianas se utilizarán las variables de fraude más comúnmente utilizadas y como variable dependiente se utilizará la variable de fraude global (*marca*). No se realizan transformaciones en los datos. La tabla 6-1 muestra estas variables.

id	Numérico	8	2	Identificador del perceptor
f_tmg	Numérico	8	2	Fraude que afecta al tipo marginal
f_capinm	Numérico	8	2	Fraude que afecta a los rendimientos de capital inmobiliario
f_nhijos	Numérico	8	2	Fraude que afecta a las declaraciones de número e hijos y ascendieentes y descendientes
f_aaee	Numérico	8	2	Fraude que afecta a la declaración de actividaes económicas
f_planp	Numérico	8	2	Fraude que afecta a la desgrbación por planes de pensiones
f_gastos	Numérico	8	2	Fraude que afecta a la declarción de gastos
marca	Numérico	8	2	fraude global

Tabla 6-1

´ Los factores de fraude más comunes que aparecen definidos en la base de datos utilizada son los siguientes:

- La variable *f_tmg* indica fraude relativo al tipo marginal.

- La variable *f_capinm* indica fraude relativo a los rendimientos de capital inmobiliario,
- La variable *f_aaee* indica fraude relativo a la declaración de actividades económicas.
- La variable *f_gastos* indica fraude relativo a las deducciones de gastos.
- La variable *f_planp* indica fraude relativo a las declaraciones de planes de pensiones.
- La variable *f_nhijos* indica fraude relativo a la declaración del número de hijos y ascendientes.

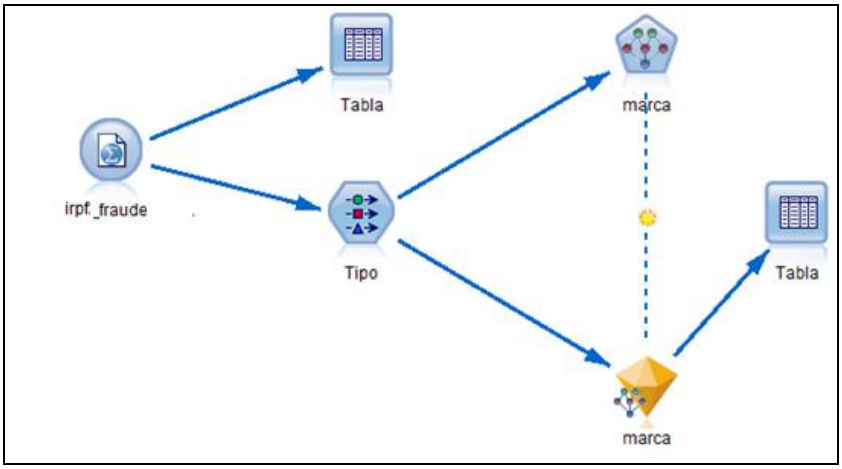
Al igual que en el caso de los otros modelos ya estudiados, estas son las causas de fraude más habituales, pero podrían considerarse de igual modo todas las causas de fraude que la Agencia Tributaria registra en la base de datos completa de IRPF una vez inspeccionada.

6.4 FASE DE MODELADO: ESTIMACIÓN DEL MODELO DE RED BAYESIANA

Determinadas herramientas de Machine Learning y Minería de Datos, utilizan marcos de trabajo amigables en los que se pueden secuenciar mediante nodos las distintas fases del trabajo hasta llegar a estimar la red. En la figura 6-1 se observa un grafo relativo a un proyecto de IBM SPSS Modeler que contiene las fase de trabajo a tener en cuenta para estimar una red neuronal bayesiana.

El primer nodo del proyecto está etiquetado con el nombre del conjunto que aporta los datos (fase de selección) que ha sido seleccionado para trabajar con la red neuronal bayesiana. El nodo Tabla se utiliza para explorar los datos del conjunto de datos importado en el nodo anterior. El nodo Tipo se utiliza para definir los tipos de variables que van a intervenir en el modelo de red bayesiana y para realizar las transformaciones previas necesarias. Superadas las fases de

selección, exploración y transformación de los datos, nos introducimos ya en la fase de modelado y utilizaremos el nodo Red Neuronal Bayesiana, etiquetado con el nombre de la variable dependiente del modelo (*marca*). Una vez que se ejecuta el nodo de Red, se obtienen los resultados de la estimación que se representan en el grafo a través de un diamante también etiquetado con el nombre de la variable dependiente de la red. Estos resultados se analizarán a continuación. Finalmente con otro nodo Tabla exploraremos los resultados con las predicciones de fraude y con la clasificación de los contribuyentes en fraudulentos o no.



Figuar 6-1

Al estimar la Red Neuronal Bayesiana obtemos el grafo dirigido de la figura 6-2.

En el grafo de la red se observa que el fraude que afecta al tipo marginal, a las actividades económicas, a la declaración de gastos, a los rendimientos del capital inmobiliario, al número de hijos y a los rendimientos por planes de pensiones son sucesores inmediatos del fraude global (hay una flecha directa desde el fraude global hacia cada tipo de fraude), lo que indica que todos influyen sobre el fraude global.

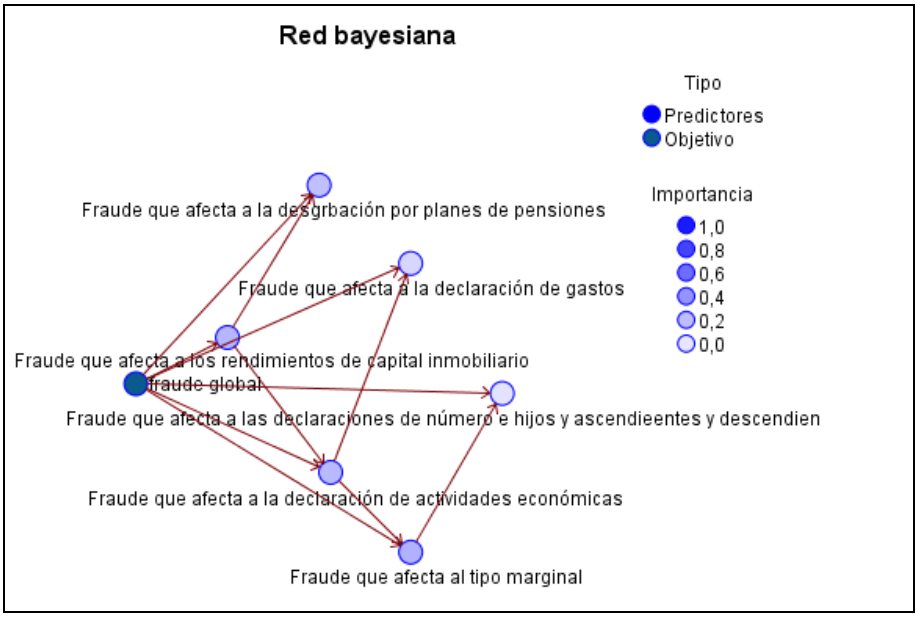


Figura 6-2

Por otra parte, el fraude que afecta al tipo marginal tiene como antecesor inmediato en el grafo dirigido al fraude que afecta a la declaración de actividades económicas, luego el fraude por tipo marginal está condicionado por el fraude en actividades económicas en su influencia sobre el fraude global.

Asimismo, el fraude por actividades económicas tiene como antecesor inmediato al fraude que afecta a la declaración de los rendimientos de capital inmobiliario. Por lo tanto, el fraude en actividades económicas está condicionado por el fraude relativo a los rendimientos de capital inmobiliario en su influencia sobre el fraude global.

También, el fraude por número de hijos y ascendientes tiene como antecesor inmediato al fraude que afecta a la declaración del tipo marginal. Por lo tanto, el fraude por número de hijos y ascendientes está condicionado por el fraude relativo al tipo marginal en su influencia sobre el fraude global.

De la misma forma podrían analizarse el resto de dependencias en el grafo de la red.

La red neuronal bayesiana también permite medir la importancia de los predictores sobre la probabilidad de fraude global. La figura 6-3 indica que el fraude que afecta al tipo marginal es el que más influencia tiene sobre el fraude global, seguido del fraude por rendimeinto del capital inmobiliario, por actividades económicas, por planes de pensiones, por declaración de gastos y por el número de hijos y ascendientes. Los resultados difieren muy poco del resultado obtenido con los árboles de decisión y otras técnicas ya aplicadas.

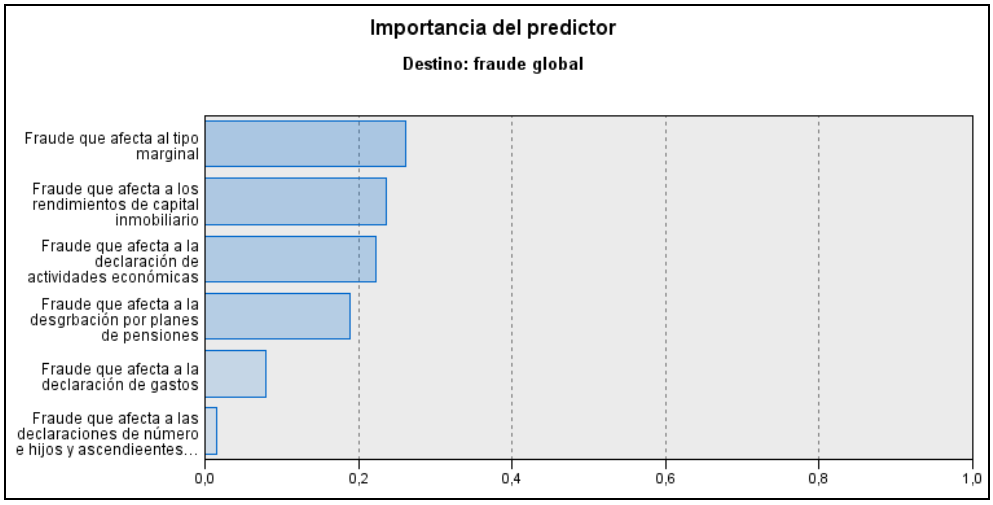


Figura 6-3

Asimismo, tal y como indica la tabla 6-2, se pueden calcular las probabilidades de fraude de cada contribuyente $BP-1 y el grupo de clasificación del mismo (fraude o no fraude) dado por la variable $B_marca.

$B-marca	...	$BP-0	$BP-1
1.000	...	0.000	1.000
1.000	...	0.000	1.000
0.000	...	0.920	0.080
0.000	...	0.920	0.080
0.000	...	0.920	0.080
0.000	...	0.920	0.080
1.000	...	0.000	1.000
1.000	...	0.000	1.000
1.000	...	0.000	1.000
1.000	...	0.000	1.000
1.000	...	0.000	1.000
1.000	...	0.000	1.000
1.000	...	0.000	1.000
0.000	...	0.920	0.080
1.000	...	0.000	1.000
0.000	...	0.920	0.080
1.000	...	0.000	1.000
1.000	...	0.000	1.000
1.000	...	0.000	1.000
0.000	...	0.920	0.080

Tabla 6-2

6.5 ANALISIS DEL PERFIL DE FRAUDE PARA LA RED BAYESIANA

Para analizar el perfil de los defraudadores representamos la función de densidad de la probabilidad de fraude, calculada a partir del método del kernel, que se muestra en la figura 6-4:

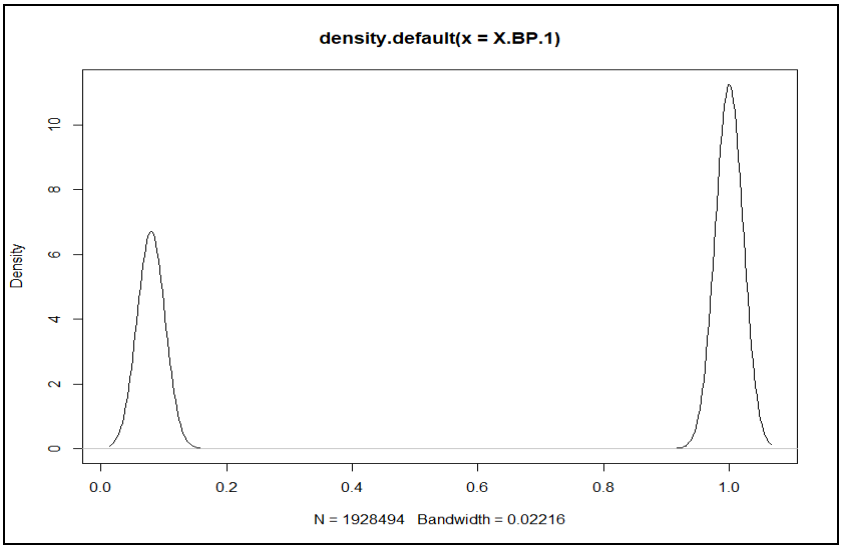


Figura 6-4

El perfil de los defraudadores resulta similar al obtenido para otros modelos previos. Para probabilidades bajas de fraude hay una densidad alta de contribuyentes. La densidad se mantiene baja hasta aproximarse a probabilidades de fraude altas, superiores a 0,9, donde vuelve a aparecer una densidad de defraudores mayor.

Comparando este perfil de fraude con el de otros modelos previos, vemos que la red neuronal bayesiana detecta mejor los contribuyentes fraudulentos con probabilidades de fraude altas.

6.6 MARCO METODOLÓGICO: METODO KNN

El algoritmo k-Nearest Neighbor (vecino más cercano) es un método que simplemente busca en las observaciones más cercanas a la que se está tratando de predecir y clasifica el punto de interés basado en la mayoría de datos que le rodean. Se trata de una técnica de Machine Learning de aprendizaje supervisado.

El algoritmo kNN sirve para estimar la función de densidad F(x/Cj)} de las variables predictoras del modelo para cada clase de predicción Cj. Se trata de un método de clasificación no paramétrico, que estima el valor de la función de densidad de probabilidad o directamente la probabilidad a posteriori de que un elemento pertenezca a la clase Cj a partir de la información proporcionada por el conjunto de prototipos. El elemento se clasifica en la clase a la que tenga mayor probabilidad de pertenecer. En el proceso de aprendizaje no se hace ninguna suposición acerca de la distribución de las variables predictoras.

En nuestro caso utlizaremos el algoritmo kNN para predecir si un individuo defrauda o no y con qué probabilidad lo hace, basándonos en las partidas económicas del Impuesto sobre la Renta de las Personas Físicas.

Las fases de selección y exploración son similares al caso de las redes neuronales bayesianas. La fase de transformación consiste en reducir a 65 componentes principales las más de 200 variables económicas del IRPF. Finalmente el modelo tiene como variable dependiente el fraude global y como variables independientes las componentes principales resultantes de la reducción de la dimensión de las partidas económicas del IRPF.

6.7 ESTIMACIÓN Y DIAGNOSIS DEL METODO KNN

Realizada la estimación del método kNN se obtienen las probabilidades de fraude de cada uno de los contribuyentes, recogidas en la columna P_marca de la tabla 6-3.

P_marca	R_marca	@_N1	@_N2	@_N3	@_N4	@_N5	@_N6	@_N7	@_N8
,94	,06	1292926	398	1786711	1382273	443233	1230328	5518	605578
1,00	,00	1270689	2690	1480299	9564	232503	449897	1060454	700548
,00	,00	931662	1771460	1173465	1197553	90782	1051978	250256	1554137
,00	,00	1657649	1232116	1117198	1887745	1102321	1065107	1470697	48187
,06	-,06	1920948	1322489	1431425	1502136	1719574	449362	569762	87687
,00	,00	1124946	465731	1333300	1463534	1047460	1167199	331489	68786
,44	,56	11683	7116	11682	3956	3670	11353	11283	11662
,94	,06	11743	14427	3913	1662	211646	14365	3916	11672
1,00	,00	429406	882548	257851	1407897	1618448	1298970	651190	1533289
1,00	,00	1467193	1467186	691958	1541994	1046814	1471417	572723	1123805
1,00	,00	120516	531001	1345331	572811	1786894	1107039	76077	790411

Tabla 6-3

En la tabla 6-4 se presenta resultados sobre la diagnosis del modelo.

Estadísticos de ajuste	Etiqueta de estadísticos	Entrenamiento
NW	número de pesos estimados	65
NOBS	suma de frecuencias	1915390
SUMW	suma de frec temporales de peso...	1915390
DFT	grados de libertad totales	1915390
DFM	grados de libertad del modelo	65
DFE	grados de libertad para el error	1915325
ASE	error cuadrático medio	0.0332
RASE	error cuadrático medio de la raíz	0.182209
DIV	divisor para ASE	1915390
SSE	suma de errores cuadráticos	63591.25
MSE	error cuadrático de la media	0.033201
RMSE	error cuadrático medio de la raíz	0.182212
AVERR	función de error de promedio	0.0332
ERR	función de error	63591.25
MAX	error absoluto máximo	0.999341
FPE	error de predicción final	0.033202
RFPE	error de predicción final de la raíz	0.182215
AIC	criterio de información de Akaike	-6522157
SBC	criterio bayesiano de Schwarz	-6521347

Tabla 6-4

Se observan magnitudes de error pequeñas, valores negativos de los estadísticos de la cantidad de información (AIC y SBC), lo que indica una capacidad predictiva del modelo alta.

6.8 ANALISIS DEL PERFIL DE FRAUDE PARA EL MÉTODO KNN

Si segmentamos las causas de fraude para ver como afectan al fraude global estimado por el método kNN, observamos un resultado similar a los modelos ya estudiados, tal y como indica el mapa perceptual del escalamiento multidimesional (figura 6-5).

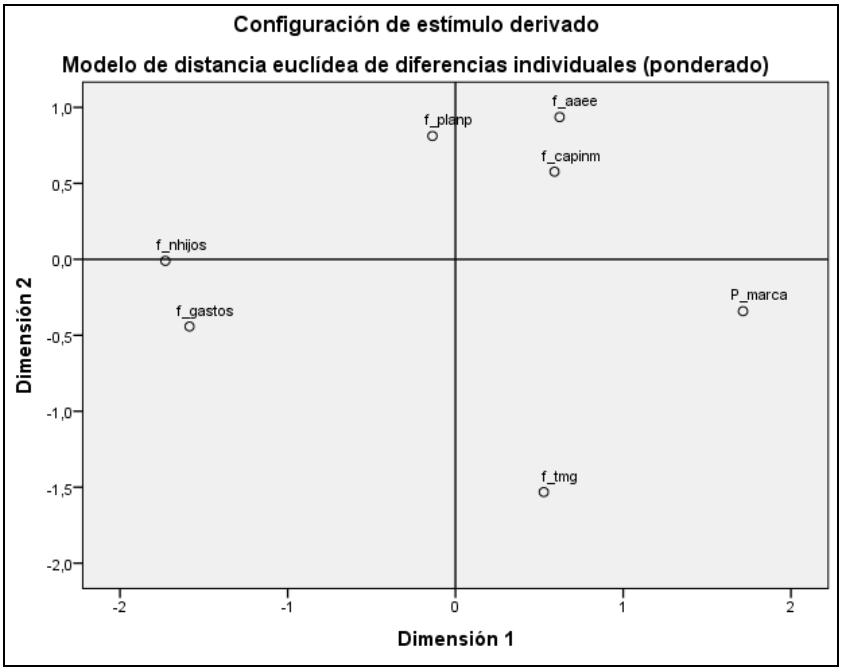


Figura 6-5

Para analizar el perfil de los defraudadores representamos la función de densidad de la probabilidad de fraude, calculada a partir del método kNN, que se muestra en la figura 6-6.

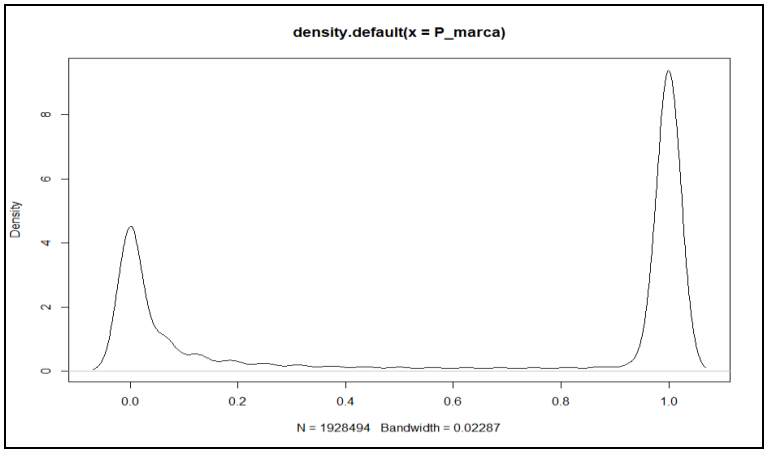


Figura 6-6

El perfil es muy similar al caso de la red Bayesiana. Para probabilidades bajas de fraude bajas hay una densidad alta de contribuyentes. La densidad se mantiene baja hasta aproximarse a probabilidades de fraude altas, superiores a 0,9, donde vuelve a aparecer una densidad de defraudores mayor.

6.9 EVALUACIÓN DE MODELOS

Las metodologías de Data Mining y Machine Learning contemplan una última fase de evaluación de modelos que consiste en comparar simultáneamente varios de los modelos ya estudiados.

En nuestro caso utilizaremos el marco de trabajo de la metodología SEMMA implementada sobre la herramienta de Data Mining y Machine Learning SAS Enterprise Miner. Mediante el proyecto que se presenta en la figura 6-7, se comparan para predecir las probabiliddaes de fraude, el modelo Logit (etiquetado como Regresion), el modelo Probit (etiquetado cono Regresion 2), la Red Neuronal Perceptrón Multicapa (etiquetado Red neuronal), la red neuronal automática óptima (etiquetado Autoneural) y el método kNN (etiquetado como MBR o Razonamiento Basado en la Memoria)

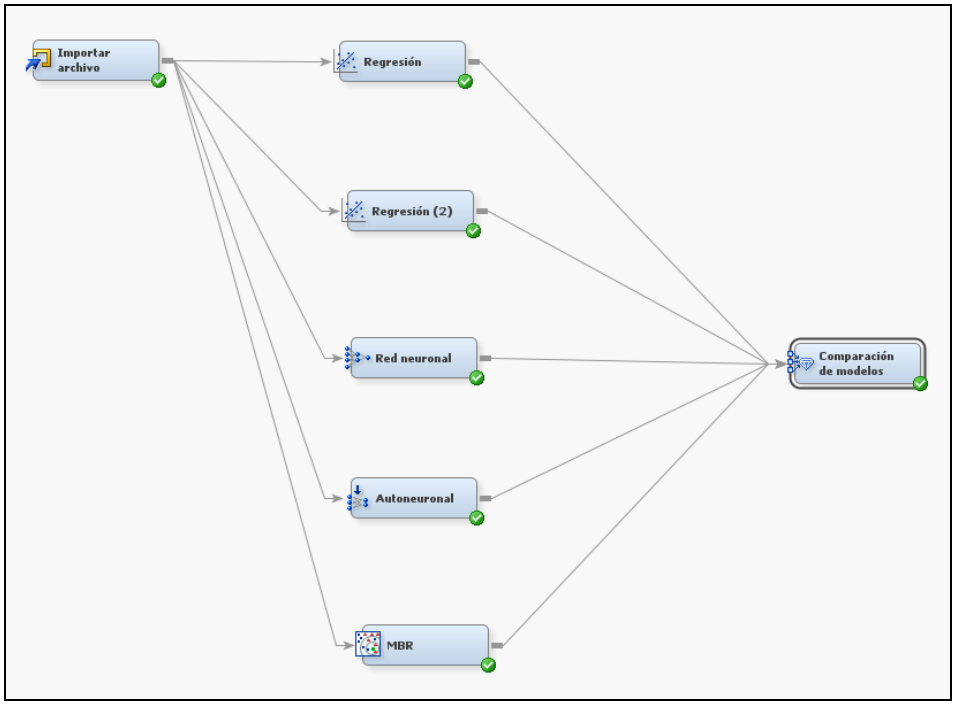


Figura 6-7

Como resultado de la comparación de modelos se obtiene la tabla 6-5 que presenta estadísticos de comparación.

Estadísticos de ajuste	Etiqueta de estadísticos	MBR	Neural	AutoNeural	Reg	Reg2
AIC	Entrenar: criterio de información de Akaike	-6522157	-4926394	-4640065	-3629444	-3629444
ASE	Entrenar: error cuadrático medio	0.0332	0.077713	0.090158	0.152273	0.152273
AVERR	Entrenar: función de error de promedio	0.0332	0.077713	0.090158	0.152273	0.152273
CRITERION	Criterio de selección: Entrenar: error cuadrático medio	0.0332	0.077713	0.090158	0.152273	0.152273
DFE	Entrenar: grados de libertad para el error	1915325	1928292	1928359	1928428	1928428
DFM	Entrenar: grados de libertad del modelo	65	202	135	66	66
DFT	Entrenar: grados de libertad totales	1915390	1928494	1928494	1928494	1928494
DIV	Entrenar: divisor para ASE	1915390	1928494	1928494	1928494	1928494
ERR	Entrenar: función de error	63591.25	149868.2	173868.4	293658.2	293658.2
FPE	Entrenar: error de predicción final	0.033202	0.077729	0.09017	0.152284	0.152284
MAX	Entrenar: error absoluto máximo	0.999341	1.533835	1.822287	25.96735	25.96735
MISC	Entrenar: tasa de clasificación errónea	.	.	.	.	.
MSE	Entrenar: error cuadrático de la media	0.033201	0.077721	0.090164	0.152279	0.152279
NOBS	Entrenar: suma de frecuencias	1915390	1928494	1928494	1928494	1928494
NW	Entrenar: número de pesos de los estimadores	65	202	135	66	66
RASE	Entrenar: suma del promedio de la raíz de cuadrados	0.182209	0.27877	0.300263	0.390222	0.390222
RFPE	Entrenar: error de predicción final de la raíz	0.182215	0.278799	0.300284	0.390236	0.390236
RMSE	Entrenar: error cuadrático medio de la raíz	0.182212	0.278784	0.300273	0.390229	0.390229
SBC	Entrenar: criterio bayesiano de Schwarz	-6521347	-4923874	-4638381	-3628621	-3628621
SSE	Entrenar: suma de errores cuadráticos	63591.25	149868.2	173868.4	293658.2	293658.2
SUMW	Entrenar: suma de frec temporales de pesos de caso	1915390	1928494	1928494	1928494	1928494
WRONG	Entrenar: número de clasificaciones erróneas	.	.	.	.	.

Tabla 6-5

Observamos que el método que tiene menores errores y menor estadístico de la cantidad de informacion de Akaike (AIC) es el método kNN.

Además la diagnosis gráfica de comparación de curvas características de operación, indica que la curva dominante es la relativa al método kNN, tal y como se muestra en la figura 6-8.

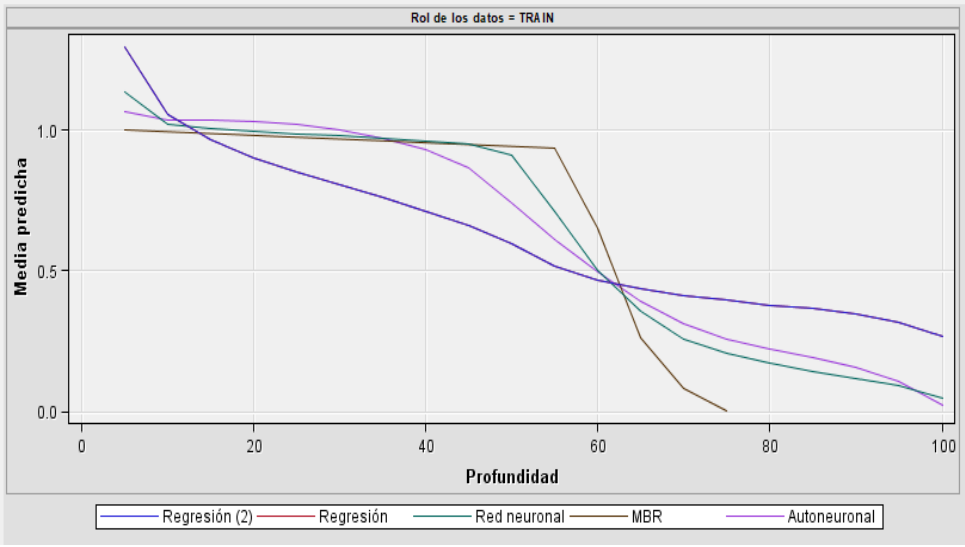


Figura 6-8

BIBLIOGRAFÍA

Agresti, A. Categorical Data Analysis, 2nd Ed. Hoboken, NJ: John Wiley & Sons, Inc., 2002.

Ahumada H., Alvaredo, F., Canavese, A. "The monetary method to measure the shadow economy: The forgotten problem of the initial conditions", 2008. Economics Letters 101 (2), 97–99.

Alañón-Pardo, A., Gómez de Antonio, M. "Estimating the size of the shadow economy in Spain: a structural model with latent variables", 2005. Applied Economics, 37 (9), 1011-1025.

Allwein, E., R. Schapire, Y. Singer. "Reducing multiclass to binary: A unifying approach for margin classifiers." Journal of Machine Learning Research. Vol. 1, 2000, pp. 113–141.

Alpaydin, E. "Combined 5 x 2 CV F Test for Comparing Supervised Classification Learning Algorithms." Neural Computation, Vol. 11, No. 8, 1999, pp. 1885–1992.

Arrazola, M., Hevia, J., Mauleón, I., Sánchez, R. "Estimación del volumen de economía sumergida en España", 2011. Cuadernos de Información Económica, 220, 81-87.

Ayala, L.. "La Desigualdad en España: Fuentes, Tendencias y Comparaciones Internacionales". Estudios sobre la Economía Española - 2016/24

Ayala, L., Onrubia J., Rodado, M.C. "El tratamiento de las fuentes de renta en el IRPF y su influencia en la desigualdad y la redistribución". Papeles de trabajo del Instituto de Estudios Fiscales. Serie economía, ISSN 1578-0252, Nº 25, 2006, págs. 7-57.

Ayala, L., Onrubia J., Ruiz-Huerta, J. "Modelos de microsimulación: aplicaciones a partir del Panel de Declarantes por IRPF del Instituto de Estudios Fiscales". 2004. Cuadernos económicos de I.C.E. N.º 68

Bates, D. M., D. G. Watts. Nonlinear Regression Analysis Its Applications. Hoboken, NJ: John Wiley & Sons, Inc., 1988.

Battiti, R., "First second order methods for learning: Between steepest descent Newton's method," Neural Computation, Vol. 4, No. 2, 1992, pp. 141–166.

Belsley, D. A., E. Kuh, R. E. Welsch. Regression Diagnostics. Hoboken, NJ: John Wiley & Sons, Inc., 1980.

Berry, M. W., et al. "Algorithms Applications for Approximate Nonnegative Matrix Factorization." Computational Statistics Data Analysis. Vol. 52, No. 1, 2007, pp. 155–173.

Bezdek, J.C., Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum Press, New York, 1981.

Blackard, J. A. D. J. Dean. "Comparative accuracies of artificial neural networks discriminant analysis in predicting forest cover types from cartographic

variables". Computers Electronics in Agriculture Vol. 24, Issue 3, 1999, pp. 131–151.

Bottou, L., Chih-Jen Lin. "Support Vector Machine Solvers." Large Scale Kernel Machines (L. Bottou, O. Chapelle, D. DeCoste, J. Weston, eds.). Cambridge, MA: MIT Press, 2007.

Bouckaert, R. E. Frank. "Evaluating the Replicability of Significance Tests for Comparing Learning Algorithms." In Advances in Knowledge Discovery Data Mining, 8th Pacific-Asia Conference, 2004, pp. 3–12.

Bouckaert. R. "Choosing Between Two Learning Algorithms Based on Calibrated Tests." International Conference on Machine Learning, pp. 51–58, 2003.

Box, G. E. P., W. G. Hunter, J. S. Hunter. Statistics for Experimenters. Hoboken, NJ: Wiley-Interscience, 1978.

Breiman, L. "Random Forests." Machine Learning. Vol. 4, 2001, pp. 5–32.

Breiman, L. Bagging Predictors. Machine Learning 26, 1996, pp. 123–140. [9] Breiman, L. Random Forests. Machine Learning 45, 2001, pp. 5–32.

Breiman, L., J. H. Friedman, R. A. Olshen, C. J. Stone. Classification Regression Trees. Boca Raton, FL: Chapman & Hall, 1984.

Brindusa, A., Vázquez, P. "Economía sumergida: Comparativa internacional y métodos de estimación" 2010. Círculo de Empresarios (ed.), Implicaciones de la economía sumergida en España, Círculo de Empresarios, 17-44,

Bulmer, M. G. Principles of Statistics. Mineola, NY: Dover Publications, Inc., 1979.

Cantarero, D., Blázquez, C. "Una aproximación a la magnitud de la economía sumergida en Cantabria (2009-2012)", 2013. Departamento de Economía, Universidad de Cantabria,

Caudill, M., C. Butler, Understanding Neural Networks: Computer Explorations, Vols. 1 2, Cambridge, MA: The MIT Press, 1992.

Caudill, M., Neural Networks Primer, San Francisco, CA: Miller Freeman Publications, 1989.

Charalambous, C.,"Conjugate gradient algorithm for efficient training of artificial neural networks," IEEE Proceedings, Vol. 139, No. 3, 1992, pp. 301–310.

Chengyu, G., K. Danai, "Fault diagnosis of the IFAC Benchmark Problem with a model-based recurrent neural network," Proceedings of the 1999 IEEE International Conference on Control Applications, Vol. 2, 1999, pp. 1755–1760.

Chiu, S., "Fuzzy Model Identification Based on Cluster Estimation," Journal of Intelligent & Fuzzy Systems, Vol. 2, No. 3, Sept. 1994.

Christianini, N., J. Shawe-Taylor. An Introduction to Support Vector Machines Other Kernel-Based Learning Methods. Cambridge, UK: Cambridge University Press, 2000.

Collett, D. Modeling Binary Data. New York: Chapman & Hall, 2002.

Comisión del Fraude Fiscal. "Evaluación del fraude en el Impuesto sobre la Renta de las Personas Físicas. Ejercicio 1979-1986", 1988. Instituto de Estudios Fiscales, Madrid.

Conover, W. J. Practical Nonparametric Statistics. Hoboken, NJ: John Wiley & Sons, Inc., 1980.

Consejo Económico y Social de la Región De Murcia. "La economía sumergida en la Región de Murcia", Estudio 20. 2006. Dell'Anno, R., Gómez-Antonio, M., Alañon-Pardo, A. "The shadow economy in three Mediterranean countries: France, Spain and Greece. A MIMIC approach" 2013, Empirical Economics, 33(1), 197.

Consejo Económico y Social, CES. "Informe sobre la distribución de la renta en España: desigualdad, cambios estructurales y ciclo", 2013. Informe 3/2013, Madrid

Cook, R. D., S. Weisberg. Residuals Influence in Regression. New York: Chapman & Hall/CRC Press, 1983.

Davidian, M., D. M. Giltinan. Nonlinear Models for Repeated Measurements Data. New York: Chapman & Hall, 1995.

De Jesús, O., J.M. Horn, M.T. Hagan, "Analysis of Recurrent Network Training Suggestions for Improvements," Proceedings of the International Joint Conference on Neural Networks, Washington, DC, July 15–19, 2001, pp. 2632–2637.

De Jesús, O., M.T. Hagan, "Backpropagation Algorithms for a Broad Class of Dynamic Networks," IEEE Transactions on Neural Networks, Vol. 18, No. 1, January 2007, pp. 14 -27.

De Jesús, O., M.T. Hagan, "Backpropagation Through Time for a General Class of Recurrent Network," Proceedings of the International Joint Conference on Neural Networks, Washington, DC, July 15–19, 2001, pp. 2638–2642.

De Jesús, O., M.T. Hagan, "Forward Perturbation Algorithm for a General Class of Recurrent Network," Proceedings of the International Joint Conference on Neural Networks, Washington, DC, July 15–19, 2001, pp. 2626–2631.

Demidenko, E. Mixed Models: Theory Applications. Hoboken, NJ: John Wiley & Sons, Inc., 2004.

Dietterich, T. "Approximate statistical tests for comparing supervised classification learning algorithms." Neural Computation, Vol. 10, No. 7, 1998, pp. 1895–1923.

Dietterich, T., G. Bakiri. "Solving Multiclass Learning Problems Via Error-Correcting Output Codes." Journal of Artificial Intelligence Research. Vol. 2, 1995, pp. 263–286.

Dobson, A. J. An Introduction to Generalized Linear Models. New York: Chapman & Hall, 1990.

Domínguez, F., López, J. y Rodrigo, F. "El hueco que deja el diablo: una estimación del fraude en el IRPF con microdatos tributarios", 2013. Documento de Trabajo, 728, FUNCAS.

Draper, N. R., H. Smith. Applied Regression Analysis. Hoboken, NJ: Wiley-Interscience, 1998.

Dubois, D. H. Prade, Fuzzy Sets Systems: Theory Applications, Academic Press, New York, 1980.

DuMouchel, W. H., F. L. O'Brien. "Integrating a Robust Option into a Multiple Regression Computing Environment." Computer Science Statistics: Proceedings of the 21st Symposium on the Interface. Alexandria, VA: American Statistical Association, 1989.

Dunn, O.J., V.A. Clark. Applied Statistics: Analysis of Variance Regression. New York: Wiley, 1974.

Efron, B., R. J. Tibshirani. An Introduction to the Bootstrap. New York: Chapman & Hall, 1993.

Esteller, A. "Economía sumergida, fraude fiscal y amplitud de bases", 2014. IEB Report 3/2014, 3-6

Esteller, A.. "Incumplimiento fiscal en el IRPF (1993-2000). Un análisis de sus factores determinantes", 2005. Universitat de Barcelona e Instituto de Economía de Barcelona

Evans, M., N. Hastings, B. Peacock. Statistical Distributions. 2nd ed., Hoboken, NJ: John Wiley & Sons, Inc., 1993, pp. 50–52, 73–74, 102–105, 147, 148.

Fan, R.-E., P.-H. Chen, C.-J. Lin. "Working set selection using second order information for training support vector machines." Journal of Machine Learning Research, Vol 6, 2005, pp. 1889–1918.

Feige, E.. "How big is the irregular Economy", 1979. Challenge (November-December), 5-13.

Feld, L., Schneider, F. "Survey on the shadow economy and undeclared earnings in OECD countries", 2010. German Economic Review, 11 (2), 109-149.

Feng, J., C.K. Tse, F.C.M. Lau, "A neural-network-based channel-equalization strategy for chaos-based communication systems," IEEE Transactions on Circuits Systems I: Fundamental Theory Applications, Vol. 50, No. 7, 2003, pp. 954–957.

Ferraro, F.J., Campayo, C., Rubio, C., Millán, C. "La economía sumergida en Andalucía", 2002. CES, Andalucía

Foresee, F.D., M.T. Hagan, "Gauss-Newton approximation to Bayesian regularization," Proceedings of the 1997 International Joint Conference on Neural Networks, 1997, pp. 1930–1935.

Fortín, B., Lacroix, G., Pinard, D. "Evaluation of the Underground Economy in Quebec: A Microeconomic Approach", 2010. IZA DP, 5384

Freund, Y. R. E. Schapire. "A Decision-Theoretic Generalization of On-Line Learning an Application to Boosting". J. of Computer System Sciences, Vol. 55, 1997, pp. 119–139.

Frey B., Weck-Hannemann H. "The hidden economy as an "unobservable" variable", 1984. European Economic Review 26, 33–53.

Friedman, J. H. "Greedy function approximation: a gradient boosting machine." The Annals of Statistics. Vol. 29, No. 5, 2001, pp. 1189-1232.

Friedman, J., T. Hastie, R. Tibshirani. Additive logistic regression: A statistical view of boosting. Annals of Statistics, Vol. 28, No. 2, 2000, pp. 337–407.

Fundación de Estudios Financieros, FEF. "La economía sumergida en España", 2013. Documentos de Trabajo 4, Madrid

Gallego, C. , Labeaga, J.M. "Estimación del fraude fiscal procedente del IVA en España a través de la comparación entre las recaudaciones real y potencial (2006- 2011)", 2014.

García-Herrera C., "Las medidas de lucha contra el fraude fiscal" (1ª parte) Fraude fiscal: Dimensión nacional". Encuentro de Derecho Financiero y Tributario (2ª ed.). Documentos - Instituto de Estudios Fiscales, ISSN 1578-0244, Nº. 16, 2013, 298 págs.

García-Herrera C., "Las medidas de lucha contra el fraude fiscal" (2ª parte) Fraude fiscal: Dimensión internacional". Encuentro de Derecho Financiero y Tributario (2ª ed.). Documentos - Instituto de Estudios Fiscales, ISSN 1578-0244, Nº. 17, 2013, 92 págs.

Genz, A., F. Bretz. "Numerical Computation of Multivariate t Probabilities with Application to Power Calculation of Multiple Contrasts." Journal of Statistical Computation Simulation. Vol. 63, 1999, pp. 361–378.

GESTHA. "Informe de Economía Sumergida 2000-2009", 2010.

GESTHA. "Reducir el fraude fiscal y la economía sumergida. Una medida vital e imprescindible para superar la crisis", 2011

GESTHA-URV. "La economía sumergida pasa factura. El avance del fraude en España durante la crisis", 2014. Gestha- Fundación URV, Madrid

Giachi, S. "Dimensiones sociales del fraude fiscal: confianza y moral fiscal en la España contemporánea", 2014. Revista Española de Investigación Social, 145: 73-98.

Gianluca, P., D. Przybylski, B. Rost, P. Baldi, "Improving the prediction of protein secondary structure in three eight classes using recurrent neural networks profiles," Proteins: Structure, Function, Genetics, Vol. 47, No. 2, 2002, pp. 228–235.

Gibbons, J. D. "Nonparametric Statistical Inference". New York: Marcel Dekker, 1985.

Gill, P.E., W. Murray, M.H. Wright, "Practical Optimization", New York: Academic Press, 1981.

Gómez de Antonio, M., Alañón, A. "Evaluación y análisis espacial del grado de incumplimiento fiscal para las provincias españolas (1980-2000)", 2004. Hacienda Pública Española, 171, 9-32.

Gómez de Enterría, P., Melis, F., Romero, D. "Evaluación del cumplimiento en el IVA: Revisión de las estimaciones años 1990 a 1994", 1998. Papeles de Trabajo, 18, Instituto de Estudios Fiscales.

Gómez, A.P. "El fraude fiscal en España en el impuesto sobre sociedades. Medidas para combatirlo" 2019. Universidad Politécnica de Valencia. Master en dirección financiera y fiscal.

González M., González, Mª. C. "Análisis de la economía sumergida en las Comunidades Autónomas. Una aproximación a través del enfoque de la demanda de efectivo", 2013. XXII Congreso Nacional de ACDE, septiembre, Málaga,

Gutmann, P. M. "The subterranean Economy", 1977. Financial Analysts Journal, 33 (6), 26-34.

Hagan, M.T., H.B. Demuth, "Neural Networks for Control," Proceedings of the 1999 American Control Conference, San Diego, CA, 1999, pp. 1642–1656.

Hagan, M.T., H.B. Demuth, M.H. "Neural Network Design", Boston, MA: PWS Publishing, 1996.

Hagan, M.T., M. Menhaj, "Training feed-forward networks with the Marquardt algorithm," IEEE Transactions on Neural Networks, Vol. 5, No. 6, 1999, pp. 989–993, 1994.

Hagan, M.T., O. De Jesus, R. Schultz, "Training Recurrent Networks for Filtering Control," Chapter 12 in Recurrent Neural Networks: Design Applications, L. Medsker L.C. Jain, Eds., CRC Press, pp. 311–340.

Hahn, Gerald J., S. S. Shapiro. "Statistical Models in Engineering". Hoboken, NJ: John Wiley & Sons, Inc., 1994, p. 95.

Hald, A. "Statistical Theory with Engineering Applications". Hoboken, NJ: John Wiley & Sons, Inc., 1960.

Harman, H. H. "Modern Factor Analysis". 3rd Ed. Chicago: University of Chicago Press, 1976.

Hastie, T., R. Tibshirani, J. Friedman. "The Elements of Statistical Learning", second edition. New York: Springer, 2008.

Hastie, T., R. Tibshirani, J. H. Friedman. "The Elements of Statistical Learning". New York: Springer, 2001.

Hill, P. D. "Kernel estimation of a distribution function." Communications in Statistics – Theory Methods. Vol. 14, Issue 3, 1985, pp. 605-620.

Himmelblau, D.M., "Applied Nonlinear Programming", New York: McGraw-Hill, 1972.

Ho, C. H. C. J. Lin. "Large-Scale Linear Support Vector Regression." Journal of Machine Learning Research, Vol. 13, 2012, pp. 3323–3348.

Ho, T. K. "The random subspace method for constructing decision forests". IEEE Transactions on Pattern Analysis Machine Intelligence, Vol. 20, No. 8, 1998, pp. 832–844.

Hochberg, Y., A. C. Tamhane. "Multiple Comparison Procedures". Hoboken, NJ: John Wiley & Sons, 1987.

Hoerl, A. E., R. W. Kennard. "Ridge Regression: Applications to Nonorthogonal Problems." Technometrics. Vol. 12, No. 1, 1970, pp. 69–82.

Hoerl, A. E., R. W. Kennard. "Ridge Regression: Biased Estimation for Nonorthogonal Problems." Technometrics. Vol. 12, No. 1, 1970, pp. 55–67.

Hogg, R. V., J. Ledolter. "Engineering Statistics". New York: MacMillan, 1987.

Holland, P. W., R. E. Welsch. "Robust Regression Using Iteratively Reweighted Least-Squares." Communications in Statistics: Theory Methods, A6, 1977, pp. 813–827.

Hollander, M., D. A. Wolfe. "Nonparametric Statistical Methods". Hoboken, NJ: John Wiley & Sons, Inc., 1999.

Horn, J.M., O. De Jesús M.T. Hagan, "Spurious Valleys in the Error Surface of Recurrent Networks - Analysis Avoidance," IEEE Transactions on Neural Networks, Vol. 20, No. 4, pp. 686-700, April 2009.

Hsieh, C. J., K. W. Chang, C. J. Lin, S. S. Keerthi, S. Sundararajan. "A Dual Coordinate Descent Method for Large-Scale Linear SVM." Proceedings of the 25th International Conference on Machine Learning, ICML '08, 2001, pp. 408–415.

Hu, Q., X. Che, L. Zhang, D. Yu. "Feature Evaluation Selection Based on Neighborhood Soft Margin." Neurocomputing. Vol. 73, 2010, pp. 2114–2124.

Huang, P. S., H. Avron, T. N. Sainath, V. Sindhwani, B. Ramabhadran. "Kernel methods match Deep Neural Networks on TIMIT." 2014 IEEE International Conference on Acoustics, Speech Signal Processing. 2014, pp. 205–209.

Huber, P. J. "Robust Statistics". Hoboken, NJ: John Wiley & Sons, Inc., 1981.

Hunt, K.J., D. Sbarbaro, R. Zbikowski, P.J. Gawthrop, Neural Networks for Control System — A Survey," Automatica, Vol. 28, 1992, pp. 1083–1112.

Jackson, J. E. "A User's Guide to Principal Components". Hoboken, NJ: John Wiley Sons, 1991.

Jain, A., R. Dubes. "Algorithms for Clustering Data". Upper Saddle River, NJ: Prentice-Hall, 1988.

Jang, J.-S. R. C.-T. Sun, "Neuro-fuzzy modeling control," Proceedings of the IEEE, March 1995.

Jang, J.-S. R. C.-T. Sun, "Neuro-Fuzzy Soft Computing: A Computational Approach to Learning Machine Intelligence", Prentice Hall, 1997.

Jang, J.-S. R., "ANFIS: Adaptive-Network-based Fuzzy Inference Systems," IEEE Transactions on Systems, Man, Cybernetics, Vol. 23, No. 3, pp. 665-685, May 1993.

Jang, J.-S. R., "Fuzzy Modeling Using Generalized Neural Networks Kalman Filter Algorithm," Proc. of the Ninth National Conf. on Artificial Intelligence (AAAI-91), pp. 762-767, July 1991.

Jarque, C. M., A. K. Bera. "A test for normality of observations regression residuals." International Statistical Review. Vol. 55, No. 2, 1987, pp. 163–172.

Jayadeva S.A.Rahman, "A neural network with O(N) neurons for ranking N numbers in O(1/N) time," IEEE Transactions on Circuits Systems I: Regular Papers, Vol. 51, No. 10, 2004, pp. 2044–2051.

Johnson, N. L., N. Balakrishnan, S. Kotz. "Continuous Multivariate Distributions". Vol. 1. Hoboken, NJ: Wiley-Interscience, 2000.

Johnson, N. L., S. Kotz, A. W. Kemp. "Univariate Discrete Distributions". Hoboken,

Johnson, N. L., S. Kotz, N. Balakrishnan. "Continuous Univariate Distributions". Vol. 1, Hoboken, NJ: Wiley-Interscience, 1993.

Johnson, N. L., S. Kotz, N. Balakrishnan. "Continuous Univariate Distributions". Vol. 2, Hoboken, NJ: Wiley-Interscience, 1994.

Johnson, N. L., S. Kotz, N. Balakrishnan. "Discrete Multivariate Distributions". Hoboken, NJ: Wiley-Interscience, 1997.

Johnson, N., S. Kotz. "Distributions in Statistics: Continuous Univariate Distributions-2". Hoboken, NJ: John Wiley & Sons, Inc., 1970, pp. 130–148, 189–200, 201–219.

Jolliffe, I. T. "Principal Component Analysis". 2nd ed., New York: Springer-Verlag, 2002.

Jolliffe, I.T., "Principal Component Analysis", New York: Springer-Verlag, 1986.

Jones, M.C. "Simple boundary correction for kernel density estimation." Statistics Computing. Vol. 3, Issue 3, 1993, pp. 135-146.

Jöreskog, K. G. "Some Contributions to Maximum Likelihood Factor Analysis."Psychometrika. Vol. 32, 1967, pp. 443–482.

Kamwa, I., R. Grondin, V.K. Sood, C. Gagnon, Van Thich Nguyen, J. Mereb, "Recurrent neural networks for phasor detection adaptive identification in power

system control protection," IEEE Transactions on Instrumentation Measurement, Vol. 45, No. 2, 1996, pp. 657–664.

Kaufman L., P. J. Rousseeuw. Finding Groups in Data: An Introduction to Cluster Analysis. Hoboken, NJ: John Wiley & Sons, Inc., 1990.

Kaufmann, A. M.M. Gupta, Introduction to Fuzzy Arithmetic, V.N. Reinhold, 1985. [10] Lee, C.-C., "Fuzzy logic in control systems: fuzzy logic controller-parts 1 2," IEEE Transactions on Systems, Man, Cybernetics, Vol. 20, No. 2, pp 404-435, 1990.

Kaufmann, D., Kaliberda, A. "Integrating the Unofficial Economy into de Dynamics of Post Socialist Economies: a Framework of Analyses and Evidence", 1996. Policy Research Working Paper, 1691,

Kecman V., T. -M. Huang, M. Vogt. "Iterative Single Data Algorithm for Training Kernel Machines from Huge Data Sets: Theory Performance." In Support Vector Machines: Theory Applications. Edited by Lipo Wang, 255–274. Berlin: Springer-Verlag, 2005.

Kendall, David G. "A Survey of the Statistical Theory of Shape." Statistical Science. Vol. 4, No. 2, 1989, pp. 87–99.

Kohavi, R. "Scaling Up the Accuracy of Naive-Bayes Classifiers: a Decision-Tree Hybrid." Proceedings of the Second International Conference on Knowledge Discovery Data Mining, 1996.

Kohonen, T., "Self-Organization Associative Memory", 2nd Edition, Berlin: Springer-Verlag, 1987.

Kohonen, T., "Self-Organizing Maps", Second Edition, Berlin: Springer-Verlag, 1997.

Kotz, S., S. Nadarajah. "Extreme Value Distributions: Theory Applications". London: Imperial College Press, 2000.

Krzanowski, W. J. "Principles of Multivariate Analysis: A User's Perspective". New York: Oxford University Press, 1988.

Labeaga, J.M. "Estimación del volumen de economía sumergida a través del método monetario", 2014, mimeo.

Lafuente, A. "Una medición de la economía oculta en España.", 1980. Boletín de Estudios Económicos, 111, 581-593, Universidad de Deusto.

Lancaster, H.O. "Significance Tests in Discrete Distributions." JASA, Vol. 56, Number 294, 1961, pp. 223–234.

Lawley, D. N., A. E. Maxwell. "Factor Analysis as a Statistical Method". 2nd ed. New York: American Elsevier Publishing, 1971.

Li, J., A.N. Michel, W. Porod, "Analysis synthesis of a class of neural networks: linear systems operating on a closed hypercube," IEEE Transactions on Circuits Systems, Vol. 36, No. 11, 1989, pp. 1405–1422.

Lilliefors, H. W. "On the Kolmogorov-Smirnov test for normality with mean variance unknown." Journal of the American Statistical Association. Vol. 62, 1967, pp. 399–402.

Lilliefors, H. W. "On the Kolmogorov-Smirnov test for the exponential distribution with mean unknown." Journal of the American Statistical Association. Vol. 64, 1969, pp. 387–389.

Lindstrom, M. J., D. M. Bates. "Nonlinear mixed-effects models for repeated measures data." Biometrics. Vol. 46, 1990, pp. 673–687.

Lippman, R.P., "An introduction to computing with neural nets," IEEE ASSP Magazine, 1987, pp. 4–22.

Little, Roderick J. A., Donald B. Rubin. "Statistical Analysis with Missing Data". 2nd ed., Hoboken, NJ: John Wiley & Sons, Inc., 2002.

Loh, W.Y. "Regression Trees with Unbiased Variable Selection Interaction Detection." Statistica Sinica, Vol. 12, 2002, pp. 361–386.

Loh, W.Y. Y.S. Shih. "Split Selection Methods for Classification Trees." Statistica Sinica, Vol. 7, 1997, pp. 815–840.

López J, Vallés J., Zárate, A. "Cumplimiento fiscal en el IRPF a nivel regional: medición y estimación de sus factores explicativos". Estudios sobre la economía española. 2018-15.

MacKay, D.J.C., "Bayesian interpolation," Neural Computation, Vol. 4, No. 3, 1992, pp. 415–447.

Mamdani, E.H. S. Assilian, "An experiment in linguistic synthesis with a fuzzy logic controller," International Journal of Man-Machine Studies, Vol. 7, No. 1, pp. 1-13, 1975.

Mamdani, E.H., "Advances in the linguistic synthesis of fuzzy controllers,"International Journal of Man-Machine Studies, Vol. 8, pp. 669-678, 1976.

Mardia, K. V., J. T. Kent, J. M. Bibby. "Multivariate Analysis". Burlington, MA: Academic Press, 1980.

Marquardt, D. W., R.D. Snee. "Ridge Regression in Practice." The American Statistician. Vol. 29, No. 1, 1975, pp. 3–20.

Marquardt, D., "An Algorithm for Least-Squares Estimation of Nonlinear Parameters," SIAM Journal on Applied Mathematics, Vol. 11, No. 2, June 1963, pp. 431–441.

Marquardt, D.W. "Generalized Inverses, Ridge Regression, Biased Linear Estimation, Nonlinear Estimation." Technometrics. Vol. 12, No. 3, 1970, pp. 591–612.

Marsaglia, G., W. Tsang, J. Wang. "Evaluating Kolmogorov's Distribution." Journal of Statistical Software. Vol. 8, Issue 18, 2003.

Marsaglia, G., W. W. Tsang. "A Simple Method for Generating Gamma Variables."ACM Transactions on Mathematical Software. Vol. 26, 2000, pp. 363–372.

Martínez-López, D. "How different are the Spanish self-employed workers by underreporting their incomes?" 2012. XIX Encuentro de Economía Pública, Santiago de Compostela.

Martínez-López, D. "Las agencias tributarias autonómicas: una visión panorámica". Administración & cidadanía: revista da Escola Galega de Administración Pública, ISSN-e 1887-0279, Vol. 1, Nº. 2, 2006, págs. 109-129

Martínez-Vázquez, J., Torgler, B. "The Evolution of Tax Morale in Modern Spain", 2009. Journal of Economic Issues, 43, 1-28.

Martínez-Vázquez, J., Torgler, B. "The Evolution of Tax Morale in Modern Spain". 2005 Working Paper, 33. Center for Research in Economics, Management and Arts.

Massey, F. J. "The Kolmogorov-Smirnov Test for Goodness of Fit." Journal of the American Statistical Association. Vol. 46, No. 253, 1951, pp. 68–78.

Mauleón, I. "Cuantificación reciente de la economía sumergida y el fraude fiscal en España", 2014. IEB report 2014, 3, 7-10.

Mauleón, I., Escobedo, M. I. "Demanda de dinero y economía sumergida", 1991. Hacienda Pública Española, 119, 105-122.

Mauleón, I., Sardá, J. "La Economía Sumergida en Navarra". 2014. Informe al Parlamento de Navarra.

Mauleón, I., Sardá. J. "Estimación cuantitativa de la economía sumergida en España", 1997. Ekonomiaz, 39, 125-135.

McCullagh, P., J. A. Nelder. "Generalized Linear Models". New York: Chapman & Hall, 1990.

McLachlan, G., D. Peel. "Finite Mixture Models". Hoboken, NJ: John Wiley & Sons, Inc., 2000.

Medsker, L.R., L.C. Jain, "Recurrent neural networks: design applications", Boca Raton, FL: CRC Press, 2000.

Meinshausen, N. "Quantile Regression Forests." Journal of Machine Learning Research, Vol. 7, 2006, pp. 983–999.

Meng, Xiao-Li, Donald B. Rubin. "Maximum Likelihood Estimation via the ECM Algorithm." Biometrika. Vol. 80, No. 2, 1993, pp. 267–278.

Moller, M.F., "A scaled conjugate gradient algorithm for fast supervised learning," Neural Networks, Vol. 6, 1993, pp. 525–533.

Moltó, M. A. "La economía irregular. Una primera aproximación al caso español.", 1980. Revista Española de Economía, julio-septiembre: 33-52.

Montgomery, D. C. "Design Analysis of Experiments". Hoboken, NJ: John Wiley & Sons, Inc., 2001.

Mood, A. M., F. A. Graybill, D. C. Boes. "Introduction to the Theory of Statistics". 3rd ed., New York: McGraw-Hill, 1974. pp. 540–541.

Mosteller, F., J. Tukey. "Data Analysis Regression". Upper Saddle River, NJ: Addison-Wesley, 1977.

Murphy, R. "Closing the European Tax Gap. A report for the Group of the Progressive Alliance of Socialists & Democrats in the European Parliament", 2012.

Murray, R., D. Neumerkel, D. Sbarbaro, "Neural Networks for Modeling Control of a Non-linear Dynamic System," Proceedings of the 1992 IEEE International Symposium on Intelligent Control, 1992, pp. 404–409.

Narendra, K.S., S. Mukhopadhyay, "Adaptive Control Using Neural Networks Approximate Models," IEEE Transactions on Neural Networks, Vol. 8, 1997, pp. 475–485.

Narendra, Kumpati S. Kannan Parthasarathy, "Learning Automata Approach to Hierarchical Multiobjective Analysis," IEEE Transactions on Systems, Man Cybernetics, Vol. 20, No. 1, January/February 1991, pp. 263–272.

Nocedal, J. S. J. Wright. "Numerical Optimization", 2nd ed., New York: Springer, 2006.

Onrubia J, Picos F, Pérez C. "Panel de declarantes de IRPF, 1999-2007: diseño, metodología y guía de utilización". ISBN: 978-84-8008-334-8. Instituto de Estudios Fiscales.

Onrubia J., Rodado M.C., Sarralde S., Pérez C. "Progresividad y redistribución a través del IRPF español: un análisis del bienestar social para el periodo 1982-1998". "Hacienda Pública Española/Revista de Economía Pública" ISSN 0210-1173, Volumen 183 año 2007 páginas 81-124. Papeles de trabajo del Instituto de Estudios Fiscales. Serie economía, ISSN 1578-0252, Nº 23, 2006, pags. 7-52

Patel, J. K., C. H. Kapadia, D. B. Owen. "Handbook of Statistical Distributions". New York: Marcel Dekker, 1976.

Pérez C. "Análisis multivariante de datos: aplicaciones con IBM SPSS, SAS y STATGRAPHICS". 2013. ISBN: 978-84-1545-273-7. Garceta Editorial.

Pérez C. "Data Mining with Matlab. Neural Networks Applications". 2019. ISBN: 978-1099211638. Amazon Independently published

Pérez C. "Modelos econométricos con SPSS". 2009. ISBN: 978-84-9372-086-5. Garceta Editorial.

Pérez C. "Técnicas estadísticas multivariantes con SPSS". 2009. ISBN: 978-84-9281-200-4 Editorial RAMA.

Pérez C., Villanueva, J., Molinero I. "La muestra de IRPF de 2015: descripción general y principales magnitudes" 2018. Instituto de Estudios Fiscales

Pérez C., Villanueva, J., Molinero I. "Panel de declarantes de IRPF 1999-2014: metodología, estructura y variables" 2018. Instituto de Estudios Fiscales.

Pérez C., Delgado, M.J., De Lucas, S. "Tax fraud detection through neural networks: and application using a sample of personal income taxpayers", 2019. *Future Internet*, 11, 86-99.

Pérez, C, "Data Mining: Soluciones con Enterprise Miner". 2005. ISBN: 84-7897-695-7. RAMA Editorial.

Pérez, C. "Applications with neural networks: fit regression models, clustering, pattern recognition time series models. Examples with Matlab". 2019. ISBN: 978-1093526677. Amazon Independently published

Pérez, C. "Base de Datos Económicos del Sector Público Español (BADESPE)". Indice: revista de estadística y sociedad, ISSN 1696-9359, Nº. 41 (Julio), 2010 (Ejemplar dedicado a: Sector público), pags. 9-13.

Pérez, C. "BIG DATA ANALYTICS con herramientas de SAS, IBM, ORACLE y Microsoft". 2015. ISBN: 978-84-1622-835-5. Garceta Editorial

Pérez, C. "BIG DATA ANALYTICS: Cluster analysis pattern recognition. Examples with Matlab" 2019. ISBN-13: 978-1092678889. Amazon Independently published

Pérez, C. "BIG DATA ANALYTICS: Neural networks applications. Examples with Matlab". 2019. ISBN: 978-1092795029. Amazon Independently published

Pérez, C. "BIG DATA DEEP LEARNING. Examples with Matlab. ISBN: 978-1092973991. 2019. Amazon Independently published

Pérez, C. "BIG DATA with MATLAB". 2019. ISBN-13: 978-1092649759. Amazon Independently published

Pérez, C. "Data Mining and Big Data Analytics with Neural Networks using Matlab". ISBN: 978-1099696282. 2019. Amazon Independently published

Pérez, C. "Data Mining with IBM SPSS through examples". 2013. ISBN:9781490541945. Createspace Independent Publishing Platform

Pérez, C. "Data Mining with Matlab. Classification Predictive Techniques". 2019. ISBN: 978-1098736682. Amazon Independently published

Pérez, C. "Data Mining with Matlab. Descriptive classification techniques". 2019. ISBN: 978-1097686827. Amazon Independently published

Pérez, C. "Data Mining with Matlab. Descriptive Classification Techniques with Neural Networks". 2019. ISBN: 978-1099506505. Amazon Independently published

Pérez, C. "Data Mining with Matlab. Pattern Recognition". 2019. ISBN: 978-1098917593. Amazon Independently published

Pérez, C. "Data Mining with Matlab. Predictive Models Regression". 2019. ISBN: 978-1098567132. Amazon Independently published

Pérez, C. "Data Mining with Matlab. Predictive Techniques with Neural Networks". 2019. ISBN: 978-1099167669. Amazon Independently published

Pérez, C. "Data Mining, Big Data Analytics Deep Learning with Matlab". 2019. ISBN: 978-1070189048. Amazon Independently published

Pérez, C. "Data Mining, Big Data Analytics Machine Learning with Neural Networks using Matlab". 2019. ISBN-13: 978-1099848148. Amazon Independently published

Pérez, C. "Data Mining. Soluciones con Enterprise Miner". 2006. ISBN: 9789701511909 Editorial Alfa Omega.

Pérez, C. "Diseño de experimentos: técnicas y herramientas". 2013. ISBN: 978-84-1545-239-3. Garceta Editorial.

Pérez, C. "Econometría avanzada: técnicas y herramientas". 2008. ISBN: 978-84-8322-479-3. Pearson, Prentice Hall.

Pérez, C. "Econometría avanzada: técnicas y herramientas". 2011. ISBN: 978-84-9281-298-1. Garceta Editorial.

Pérez, C. "Econometría básica: aplicaciones con EVIEWS, STATA, SAS y SPSS". 2012. ISBN: 978-84-1545-202-7. Garceta Editorial.

Pérez, C. "Econometría básica: técnicas y herramientas". 2007. ISBN: 978-84-8322-384-0. Pearson, Prentice Hall.

Pérez, C. "Econometría de las series temporales". 2006. ISBN: 84-8322-290-6 Pearson, Prentice Hall.

Pérez, C. "El sistema estadístico SAS". 2001. ISBN: 84-205-3168-5. Pearson, Prentice Hall

Pérez, C. "El sistema estadístico SAS". 2010. ISBN: 978-84-9281-238-7. Garceta Editorial

Pérez, C. "Fiscal panels data. Application to income TAX panel data (IRPF) of spanish institute for fiscal studies: Methodology, estimators errors" . Revista BEIO, Boletín de Estadística e Investigación Operativa, ISSN 1889-3805, Vol. 27, Nº. 3, 2011, págs. 204-220

Pérez, C. "IBM SPSS estadística aplicada: conceptos y ejercicios resueltos". 2013. ISBN: 978-84-1545-271-3. Garceta Editorial.

Pérez, C. "Investigación del fraude fiscal mediante análisis discriminante: aplicaciones con las muestras y paneles de IRPF del IEF". XX Encuentro de Economía Pública 2013, ISBN 978-84-695-6945-0, pág. 6

Pérez, C. "Investigación operativa: Técnicas y herramientas". 2013. ISBN: 978-84-1545-240-9 Garceta Editorial.

Pérez, C. "Machine learning techniques: supervised learning classification. Examples with Matlab", 2019. ISBN: 978-1096996545. Amazon Independently published

Pérez, C. "Machine learning techniques: supervised learning classification. Examples with Matlab". 2019. ISBN: 978-1096996545. Amazon Independently published

Pérez, C. "Machine learning techniques: Unsupervised learning. Examples with Matlab". 2019. ISBN:978-1096928331. Amazon Independently published

Pérez, C. "Machine learning with Matlab. Supervised learning regression". 2019. ISBN: 978-1096963707. Amazon Independently published

Pérez, C. "Machine learning with neural networks. Examples with Matlab. 2019. ISBN: 978-1092551939. Amazon Independently published

Pérez, C. "Métodos estadísticos avanzados con SPSS". 2005. ISBN: 84-9732-387-4. Editorial Thomson.

Pérez, C. "Métodos estadísticos con Statgraphics para Windows: técnicas básicas". 1998. ISBN: 84-7897-305-2. Editorial RA-MA.

Pérez, C. "Mineria de Datos. Redes Neuronales y Arboles de Decisión. Ejemplos con SAS Enterprise Miner". 2013. ISBN: 9781493768400. Createspace Independent Publishing Platform

Pérez, C. "Modelos con herramientas de Mineria de Datos. Ejercicios resueltos con MODELER y SAS Miner". 2013. Createspace Independent Publishing Platform.

Pérez, C. "Muestreo estadístico a través de ejemplos. Aplicaciones con Excel, SPSS, SAS y STATA. 2017. ISBN: 978-84-1622-885-0. Garceta Editorial.

Pérez, C. "Muestreo estadístico: conceptos y problemas resueltos". 2005. ISBN: 84-205-4411-6. Pearson, Prentice Hall.

Pérez, C. "Neural networks theory and examples with Matlab". 2019. ISBN-13: 978-1093618778. Amazon Independently published

Pérez, C. "Panel de datos del Impuesto sobre la Renta de las Personas Físicas del Instituto de Estudios Fiscales". Indice: revista de estadística y sociedad, ISSN-e 1696-9359, Nº. 45 (Marzo), 2011 (Ejemplar dedicado a: Estadísticas tributarias), págs. 31-35

Pérez, C. "Problemas resueltos de econometría". 2006. ISBN: 84-9732-376-9. Thomson

Perez, C. "R. Lenguaje de programación y análisis estadístico de datos". 2015. ISBN: 978-84-1622-812-6. Garceta Editorial.

Pérez, C. "Segmentation with Matlab. Cluster Analisys Nearest Neighbors (kNN)". 2019. ISBN: 978-1091196360. Amazon Independently published

Pérez, C. "Series temporales: técnicas y herramientas". 2011. ISBN: 978-84-9281-288-2. Garceta Editorial

Pérez, C. "Statistical Research on Public Policy". Revista BEIO, Boletín de Estadística e Investigación Operativa, ISSN 1889-3805, Vol. 25, Nº. 3, 2009, pags. 250-255.

Pérez, C. "Técnicas avanzadas de predicción". 2016. ISBN: 978-84-1622-857-7. Garceta Editorial.

Pérez, C. "Técnicas de análisis de datos con SPSS 15". 2009. ISBN: 978-84-8322-601-8. Pearson, Prentice Hall.

Pérez, C. "Técnicas de Minería de Datos e Inteligencia de Negocios con IBM SPSS Modeler". 2014. ISBN: 9788415452904. Garceta Editorial

Pérez, C. "Técnicas de muestreo estadístico: Teoría, práctica y aplicaciones informáticas". 1999. ISBN: 84-7897-345-1. Editorial RAMA.

Pérez, C. "Técnicas de muestreo estadístico". 2010. ISBN: 978-84-9281-210-3. Garceta Editorial.

Pérez, C. "Técnicas de segmentación: conceptos, herramientas y aplicaciones". 2011. ISBN: 978-84-9281-219-6. Garceta Editorial.

Pérez, C. "Técnicas estadísticas con SPSS 12: aplicaciones al análisis de datos". 2005. ISBN: 84-205-4410-8. Pearson, Prentice Hall.

Pérez, C. "Técnicas estadísticas con variables categóricas": IBM SPSS. 2014. ISBN: 987-84-1545-293-5". Garceta Editorial

Pérez, C. "Técnicas estadísticas predictivas con IBM SPSS : Modelos". 2014. ISBN: 978-84-1545-287-4. Garceta Editorial

Pérez, C., González, C. "Combining the Predictive Ability of Factorial Analysis Transfer Functions for VAT Revenue Forecasting". Journal of Mathematics Statistical Science vol3, edición de agosto de 2017). Págs: 248 – 269

Pérez, C., Martín A. "Time Series Analysis Forecasting" Págs: 105-112 Springer. ISSN: 1431-1968

Pérez, C., Moral I. "Técnicas de evaluación de impacto". 2015. ISBN: 978-84-1622-836-2. Garceta Editorial.

Pérez, C., Santín D. "Minería de Datos. Técnicas y Herramientas". 2007. ISBN: 9788497324922. Editorial Thomson International.

Pickhardt, M., Sardá, J. "Size and causes of the underground economy in Spain: A correction of the record and new evidence from the MCDR approach", 2015. European Journal and Law and Economics, 39 (2), 403-429.

Pickhardt, M., Sardá, J. "The size of the underground economy in Germany: a correction of the record and new evidence from the modified-cash-deposit-ratio approach", 2011. European Journal of Law and Economics (32), 143-163.

Pinheiro, J. C., D. M. Bates. "Approximations to the log-likelihood function in the nonlinear mixed-effects model." Journal of Computational Graphical Statistics. Vol. 4, 1995, pp. 12–35.

Prado J. "Una estimación de la economía informal en España según un enfoque monetario, 1964-2001", 2004. El Trimestre Económico, 71 (82), 417-452.

Pulido, E. J. "EL FRAUDE FISCAL EN ESPAÑA. UNA ESTIMACIÓN CON DATOS DE CONTABILIDAD NACIONAL" Tesis doctoral. Universidad de Salamanca.

Rice, J. A. Mathematical Statistics Data Analysis. Pacific Grove, CA: Duxbury Press, 1994.

Riedmiller, M., H. Braun, "A direct adaptive method for faster backpropagation learning: The RPROP algorithm," Proceedings of the IEEE International Conference on Neural Networks, 1993.

Roman, J., A. Jameel, "Backpropagation recurrent neural networks in financial analysis of multiple stock market returns," Proceedings of the Twenty-Ninth Hawaii International Conference on System Sciences, Vol. 2, 1996, pp. 454–460.

Rosenblatt, F., "Principles of Neurodynamics", Washington, D.C.: Spartan Press, 1961.

Ruesga, S. "Economía sumergida y fraude fiscal", 1994. Praxis Fiscal. Jurisprudencia y Comentarios, septiembre, 668-678.

Ruesga, S. "La economía sumergida en tiempos de la Gran Depresión 2.0", 2013. Revista de Responsabilidad Social de la Empresa, 14, 49-76.

Ruesga, S., Carbajo D., Trujillo, M. "La economía sumergida y el ciclo económico", 2013. Atlantic Review of Economics, 2(1), 37.

Rumelhart, D.E., G.E. Hinton, R.J. Williams, "Learning representations by back-propagating errors," Nature, Vol. 323, 1986, pp. 533–536.

Rumelhart, D.E., J.L. McClelland, the PDP Research Group, Eds., "Parallel Distributed Processing", Vols. 1 2, Cambridge, MA: The M.I.T. Press, 1986.

Sachs, L. "Applied Statistics: A Handbook of Techniques". New York: Springer-Verlag, 1984, p. 253.

Salinas, F..J, Delgado J. "Impuestos y crecimiento económico: una panorámica" RAE: Revista Asturiana de Economía, ISSN 1134-8291, Nº. 42, 2008 (Ejemplar dedicado a: Crisis económica y finanzas públicas), págs. 9-30.

Salinas, F. J. "Economía Política de la cooperación de las políticas fiscales nacionales en la Unión Europea". Logros, iniciativas y retos institucionales y económicos : la Unión Europea del siglo XXI / coord. por Isabel Vega Mocoroa, 2005, ISBN 84-8406-648-7, págs. 369-389.

Sánchez-Maldonado J., Ávila, A. J., Avilés C. A. "Economía irregular y evasión fiscal. Análisis económico y aplicaciones regionales a la economía española", 1997. Ed. Analistas económicos de Andalucía. Colección Documentos y Estudios, 2.

Scales, L.E., "Introduction to Non-Linear Optimization", New York: Springer-Verlag, 1985.

Schapire, R. E. et al. "Boosting the margin: A new explanation for the effectiveness of voting methods". Annals of Statistics, Vol. 26, No. 5, 1998, pp. 1651–1686.

Schapire, R., Y. Singer. "Improved boosting algorithms using confidence-rated predictions". Machine Learning, Vol. 37, No. 3, 1999, pp. 297–336.

Schneider F., Enste D. "Shadow economies: size, causes, and consequences". 2000. Journal Economic Literature, 38, 77–115.

Schneider, F. "Size and development of the shadow economy of 31 European countries from 2003 to 2010", 2010. Working Paper, Universidad de Linz, Austria

Schneider, F. "Size and Development of the Shadow Economy of 31 European and 5 other OECD Countries from 2003 to 2012: Some New Facts", 2012. University of Linz, Austria

Schneider, F. "The Shadow Economy: An Essay", 2014

Schneider, F., Buehn, A. "Shadow Economies in Highly Developed OECD Countries: What Are the Driving Forces?", 2012. IZA DP, 6891.

Schneider, F., Raczkowski, K., Mroz, B. "Shadow economy and tax evasion in EU", 2015. Journal of Money Laundering Control, 18 (1), 34-51,

Scott, D. W. "Multivariate Density Estimation: Theory, Practice, Visualization". John Wiley & Sons, 2015.

Searle, S. R., F. M. Speed, G. A. Milliken. "Population marginal means in the linear model: an alternative to least-squares means." American Statistician. 1980, pp. 216–221.

Seber, G. A. F. "Multivariate Observations". Hoboken, NJ: John Wiley & Sons, Inc., 1984.

Seber, G. A. F. A. J. Lee. "Linear Regression Analysis". 2nd ed. Hoboken, NJ: Wiley-Interscience, 2003.

Seber, G. A. F., C. J. Wild. "Nonlinear Regression". Hoboken, NJ: Wiley-Interscience, 2003.

Seiffert, C., T. Khoshgoftaar, J. Hulse, A. Napolitano. "RUSBoost: Improving classification performance when training data is skewed". 19th International Conference on Pattern Recognition, 2008, pp. 1–4.

Serrano-Sanz, J.M.., Bandrés, E., Gadea, M.D., Sanaú, J. "Desigualdades territoriales en la economía sumergida", 1998. Instituto Aragonés de Desarrollo

Sexton, Joe, A. R. Swensen. "ECM Algorithms that Converge at the Rate of EM."Biometrika. Vol. 87, No. 3, 2000, pp. 651–662.

Silverman, B.W. "Density Estimation for Statistics Data Analysis". Chapman & Hall/CRC, 1986.

Snedecor, G. W., W. G. Cochran. "Statistical Methods". Ames, IA: Iowa State Press, 1989.

Soloway, D., P.J. Haley, "Neural Generalized Predictive Control", Proceedings of the 1996 IEEE International Symposium on Intelligent Control, 1996, pp. 277–281.

Spath, H. "Cluster Dissection Analysis: Theory", FORTRAN Programs, Examples. Translated by J. Goldschmidt. New York: Halsted Press, 1985.

Stephens, M. A. "Use of the Kolmogorov-Smirnov, Cramer-Von Mises Related Statistics Without Extensive Tables." Journal of the Royal Statistical Society. Series B, Vol. 32, No. 1, 1970, pp. 115–122.

Street, J. O., R. J. Carroll, D. Ruppert. "A Note on Computing Robust Regression Estimates via Iteratively Reweighted Least Squares." The American Statistician. Vol. 42, 1988, pp. 152–154.

Student. "On the Probable Error of the Mean." Biometrika. Vol. 6, No. 1, 1908, pp. 1–25.

Sugeno, M., "Industrial applications of fuzzy control", Elsevier Science Pub. Co., 1985.

Tanzi, V. "The Underground Economy in the United States and Abroad", 1982. Lexington: Lexington Books.

Tanzi, V. "The underground economy in the United States: Estimates and Implications", 1980. Banco Nazionale del Lavoro, Quarterly Review, 135: 428-453.

Truyols, M.A. "El Impuesto sobre Sociedades en términos de la Contabilidad Nacional", 1994. Hacienda Pública Española, 130, 127-150.

Vellemen, P. F., D. C. Hoaglin. "Application, Basics, Computing of Exploratory Data Analysis". Pacific Grove, CA: Duxbury Press, 1981.

Vogl, T.P., J.K. Mangis, A.K. Rigler, W.T. Zink, D.L. Alkon, "Accelerating the convergence of the backpropagation method," Biological Cybernetics, Vol. 59, 1988, pp. 256–264.

Wang, L.-X., "Adaptive fuzzy systems control: design stability analysis", Prentice Hall, 1994.

Warmuth, M., J. Liao, G. Ratsch. "Totally corrective boosting algorithms that maximize the margin". Proc. 23rd Int'l. Conf. on Machine Learning, ACM, New York, 2006, pp. 1001–1008.

Wasserman, P.D., "Advanced Methods in Neural Computing", New York: Van Nostrand Reinhold, 1993.

Weigend, A. S., N. A. Gershenfeld, eds., "Time Series Prediction: Forecasting the Future Understanding the Past", Reading, MA: Addison-Wesley, 1994.

Widrow, B. D. Stearns, "Adaptive Signal Processing", Prentice Hall, 1985.

Yager, R. D. Filev, "Generation of Fuzzy Rules by Mountain Clustering," Journal of Intelligent & Fuzzy Systems, Vol. 2, No. 3, pp. 209-219, 1994.

Zadeh, L.A., "Fuzzy Logic," Computer, Vol. 1, No. 4, pp. 83-93, 1988.

Zadeh, L.A., "Fuzzy sets," Information Control, Vol. 8, pp. 338-353, 1965. [22] Zadeh, L.A., "Outline of a new approach to the analysis of complex systems decision processes, "IEEE Transactions on Systems, Man, Cybernetics, Vol. 3, No. 1, pp. 28-44, Jan. 1973.

Zadeh, L.A., "Knowledge representation in fuzzy logic," IEEE Transactions on Knowledge Data Engineering, Vol. 1, pp. 89-100, 1989.

Zadrozny, B., J. Langford, N. Abe. "Cost-Sensitive Learning by Cost-Proportionate Example Weighting." Third IEEE International Conference on Data Mining, 435–442. 2003.

Zahn, C. T. "Graph-theoretical methods for detecting describing Gestalt clusters." IEEE Transactions on Computers. Vol. C-20, Issue 1, 1971, pp. 68–86.

Zhou, Z.-H. X.-Y. Liu. "On Multi-Class Cost-Sensitive Learning". Computational Intelligence. Vol. 26, Issue 3, 2010, pp. 232–257 CiteSeerX.