Hacienda Pública Española/Review of Public Economics 246-(3/2023):
© 2023, Instituto de Estudios Fiscales
https://dx.doi.org/10.7866/HPE-RPE.23.3.5

Modelización de los factores que afectan al fraude fiscal con técnicas de minería de datos: aplicación al Impuesto de la Renta en España

CÉSAR PÉREZ LÓPEZ*
Instituto de Estudios Fiscales y *Universidad Rey Juan Carlos*

M.ª JESÚS DELGADO RODRÍGUEZ**
Universidad Rey Juan Carlos

SONIA DE LUCAS SANTOS***
Universidad Autónoma de Madrid

Recibido: Mayo, 2022
Aceptado: Marzo, 2023

Resumen

Este trabajo presenta una propuesta para modelizar y predecir el comportamiento de los contribuyentes del Impuesto de la Renta de las Personas Físicas (IRPF) con técnicas de minería de datos. Se combinan los árboles de decisión y el análisis discriminante para cuantificar la propensión al fraude de cada contribuyente usando los componentes del impuesto con mayor incidencia en el fraude. El modelo alcanza una eficiencia media en las predicciones superior al 89%, permitiendo segmentar a los declarantes por nivel de propensión al fraude. La propuesta puede ser usada en el proceso de auditoría y control que realiza la Agencia Tributaria.

Palabras clave: fraude fiscal, minería de datos, IRPF, predicción.

Clasificación JEL: H26, C55, C38.

1. Introducción

Uno de los grandes retos al que se enfrentan los países en la actualidad es la reducción del fraude fiscal, de ahí el interés por contribuir a su detección y control con nuevas estrategias (Prichard *et al*., 2019). La expansión de las tecnologías y las posibilidades de comunicación

* ORCID ID: 0000-0001-8524-2751.

** ORCID ID: 0000-0003-3830-2701.

*** ORCID ID: 0000-0002-1490-3820.

global, aunque han limitado algunas de los canales de fraude (Alognon *et al*, 2021) y han favorecido los acuerdos internacionales de intercambio de información (O'Reilly *et al*., 2021), han hecho, a su vez, posible nuevas formas de evasión y fraude que generan un gran coste a la sociedad e incrementan las desigualdades económicas (Alm, 2012, 2021). Estas actividades fraudulentas provocan la reducción de la capacidad de generar ingresos públicos, perjudicando el objetivo de estabilidad presupuestaria e incrementando los efectos distorsionadores del sistema fiscal. Por ello, son de gran interés los trabajos tanto a nivel internacional (Feldman y Slemrod, 2007, Feige y Cebula, 2012, Buehn y Schneider, 2016, Alstadsæter *et al*., 2019, European Commission, 2021) como para la economía española (Domínguez-Barrero *et al*, 2015a, 2015b, Almunia y López-Rodríguez, 2018, Torregrosa-Hetland, 2020), que ponen de manifiesto la dimensión del fraude fiscal y reclaman estrategias correctoras.

Las auditorías fiscales que realiza la Agencia Tributaria son una fuente directa sobre el cumplimiento fiscal y ofrece una valiosa información sobre el comportamiento de los contribuyentes, pero son costosas y limitadas, dado el elevado número de contribuyentes a analizar (Chica *et al*., 2021). Para avanzar en la lucha contra el fraude fiscal, las Agencias Tributarias están realizando una gran apuesta por el uso de técnicas de procesamiento basadas en la minería de datos. El desarrollo de modelos predictivos que faciliten la detección del fraude, a partir de información de ejercicios anteriores, puede apoyar al sistema tradicional de selección, contribuyendo a reforzar su efectividad. Las posibilidades de análisis que ofrece la minería de datos permiten abordar, desde nuevas perspectivas, el estudio del fraude fiscal, aportando procedimientos sistematizados y efectivos para gestionar la información que disponen las Agencias Tributarias y así reforzar el cumplimiento fiscal (Brondolo *et al*., 2022).

Nuestro trabajo se encuadra en esta línea de investigación y contribuye a la actual literatura en varios aspectos. En primer lugar, este trabajo amplía la literatura sobre el fraude fiscal en el IRPF al ofrecer una propuesta metodológica que, combinando las técnicas de árboles de decisión y el análisis discriminante, consigue aumentar la eficiencia de las predicciones de fraude en este impuesto y ofrecer nueva evidencia dentro de la aplicación práctica. Además, en esta propuesta es posible la identificación de los factores que están relacionados con el comportamiento de fraude del contribuyente (a partir de la información que ofrece el impuesto) y modelizar de manera precisa cada uno de ellos, lo que permitirá cuantificar su incidencia. No son frecuentes los trabajos que pueden realizar este tipo de análisis, a pesar de que su estudio aporta una mayor capacidad para predecir y caracterizar al defraudador, facilitando el diseño de las reformas fiscales. Una de las principales limitaciones para llevar a cabo estos trabajos es la falta de información sobre los contribuyentes que defraudan que pueda ser utilizada en este tipo de análisis. El desarrollo de datos fiscales de gran amplitud y detalle como es el caso de la muestra del IRPF del IEF (https://www.ief.es/badespe) ha permitido avanzar en la realización de trabajos sobre este impuesto, pero al no disponer de datos sobre defraudadores hace necesario complementar esta base. Por ello, para poder llevar a cabo los objetivos de esta investigación, en este artículo se ha accedido a declaraciones de IRPF anonimizadas que ha sido facilitada por la Agencia tributaria. Procesar la gran cantidad de datos de la muestra, así como el ajuste del modelo lineal propuesto han requerido el uso del *software IBM SPSS Modeler* que implementa la tecnología de Minería de Datos para po-

sibilitar el procesamiento a gran escala. Un aspecto de interés de esta propuesta es el hecho de que el esquema que se presenta puede ser útil también para otros impuestos.

El apartado siguiente ofrece la revisión de la literatura. En la sección tercera se presenta la propuesta metodológica para modelizar y predecir el comportamiento fiscal de los contribuyentes y, a continuación, en la sección cuarta se describe la base de datos utilizada para su aplicación. El apartado siguiente presenta los resultados obtenidos con la propuesta realizada y la cuantifica del nivel de fraude de forma agregada, además se realiza un análisis complementario sobre la evolución del índice de fraude fiscal en España. En el último apartado se presentan las principales conclusiones.

2. Revisión de la literatura

El uso de las técnicas de minería de datos se ha extendido a numerosas áreas de investigación con el objetivo de encontrar patrones y tendencias que existen en grandes conjuntos de datos (Wendler y Gröttrups, 2021). Su análisis hace posible obtener estructuras comprensibles para su uso posterior, bien para prever el comportamiento futuro de la variable analizada (técnicas de aprendizaje supervisado) o para lograr describirla ayudando a su comprensión (técnicas de aprendizaje no supervisado).

Entre los principales retos a los que se enfrentan los trabajos que usan estas herramientas está el identificar la técnica apropiada para usar en cada uno de los contextos y para los objetivos en los que se emplean. Cada técnica constituye un enfoque conceptual para extraer la información de los datos, y, en general, es implementada por varios algoritmos. Su elección en el análisis del fraude fiscal viene determinada, en gran medida, por la posibilidad de disponer de datos de los que aprender (datos de defraudadores) o, en caso contrario, de recurrir a supuestos implícitos a partir de los cuales es posible establecer los casos de fraude como situaciones anómalas o menos frecuentes que las situaciones consideradas como normales. En el primer caso, sería posible aplicar los métodos de aprendizaje supervisado. Esta metodología ofrece la posibilidad de evaluar la técnica empleada en función de la mayor o menor precisión de la predicción que se obtiene y ofrece una mayor interpretabilidad a los resultados alcanzados, siendo por tanto la opción más atractiva cuando se dispone de los datos que hacen posible su aplicación (Hilal *et al*., 2022). En segundo caso, estaríamos aplicando técnicas de aprendizaje no supervisado, que son utilizadas para encontrar patrones en los datos que indiquen características potenciales para los contribuyentes defraudadores. En este último caso los resultados son menos precisos, ya que pueden incluir como potenciales defraudadores contribuyentes con irregulares o anómalos resultados. No obstante, ofrece posibilidades muy interesantes cuando no es posible o tiene un elevado coste acceder a la información (Tian *et al.*, 2016, Wei *et al*., 2019). En estos trabajos es muy importante el algoritmo que se seleccione para tratar los valores atípicos o anómalos (De Roux *et al*., 2018, Mehta *et al*., 2020, González *et al*., 2021).

Una alternativa a estas técnicas es el uso de herramientas de aprendizaje semi-supervisado (Kleanthous y Chatzis, 2020), con las que es posible entrenar modelos en los que no se dispo-

ne de un número elevado de datos etiquetados o datos de contribuyentes que hayan cometido fraude fiscal. En la literatura también podemos encontrar con trabajos que utilizan modelos híbridos que combinan distintas técnicas para mejorar la precisión obtenidos validando el grupo de individuos seleccionados como potenciales defraudades (Savić *et al*., 2022, González y Caballero, 2023).

Los trabajos que utilizan aprendizaje supervisado (Tabla 1) han usado, entre los algoritmos de clasificación más frecuentes: árboles de decisión (Mittal *et al*., 2018), redes neuronales (Rosid, 2022), reglas de asociación (Matos *et al*, 2015) y redes bayesianas (Castellón y Velásquez, 2013), a partir de las auditorías realizadas por las Agencias Tributarias. Los avances en la lucha contra el fraude fiscal a nivel internacional han ampliado las posibilidades de análisis a numerosos países. Podemos encontrar trabajos con datos de Marruecos (Ameur y Tkiouat, 2012), Italia (Basta *et al*., 2009), Brasil (da Silva *et al*., 2016) y la India (Mittal *et al*., 2018), aunque una buena parte se centran en datos de empresas que han defraudado en IVA (Wu *et al.*, 2012, Zumaya *et al*., 2021) o en el I. de Sociedades (Xu *et al*., 2022). La información sobre individuos que han defraudado es menos accesible, aunque también se ha conseguido que las Agencias Tributarias ofrezcan información para realizar investigaciones (Murorunkwere *et al*., 2022). Hasta ahora, en España el único trabajo que ha realizado este tipo de análisis para el IRPF es el de Pérez *et al.*, 2019, que utiliza redes neuronales para llevar a cabo su investigación.

Tabla 1
TRABAJOS SOBRE FRAUDE FISCAL CON TÉCNICAS DE APRENDIZAJE SUPERVISADO

Autores	Principales herramientas	Ámbito de estudio	Muestra
Ameur y Tkiouat, 2012	Árboles de decisión	I. Sociedades	Empresas-Marruecos
Basta *et al*., 2009	Reglas Básicas de Decisión	IVA	Empresas-Italia
Castellón y Velásquez, 2013	Árboles de decisión, redes neuronales y bayesianas	IVA	Empresas-Chile
Da Silva *et al*., 2016	Redes Bayesianas	I. Renta	Individuos-Brasil
Matos *et al*., 2015	Reglas de asociación	IVA	Empresas- Brasil
Mittal *et al*., 2018	Árboles de decisión	IVA	Empresas-Delhi (India)
Murorunkwere *et al*., 2022	Redes Neuronales	I. Renta	Individuos-Ruanda
Pérez *et al*., 2019	Redes Neuronales	IRPF	Individuos-España
Rosid, 2022	Redes Neuronales	I. Sociedades	Empresas-Indonesia
Xu *et al*., 2022	Árboles de decisión y redes neuronales	I. Sociedades	Empresas-China
Wu *et al*., 2012	Reglas de asociación	IVA	Empresas -Taiwan
Zumaya *et al*., 2021	Redes Neuronales y árboles de decisión	IVA	Individuos y empresas-México

Fuente: elaboración propia.

Nota: I.: Impuesto.

3. Marco metodológico: los árboles de decisión y el análisis discriminante

En este trabajo se han seleccionado las técnicas de árboles de decisión y el análisis discriminante[1] para ofrecer una propuesta metodológica que permite, en primer lugar, identificar los factores o componentes que, en la propia estructura del impuesto, afectan al fraude fiscal. A partir de la información obtenida, es posible, a continuación, caracterizar a los contribuyentes declarantes en el IRPF, al detectar patrones de comportamiento de fraude en base a la información tributaria disponible. De este modo, se obtiene una modelización del fraude que permite apoyar la fiscalización de este tipo de casos por parte de la Agencia Tributaria.

Los árboles de decisión son modelos predictivos, que tratan de resolver los problemas de discriminación en una población. Para ello, se segmentan de forma progresiva la muestra con el fin de obtener una clasificación fehaciente en grupos homogéneos, según la variable de interés denominada *variable de segmentación*. 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.

En nuestro caso utilizaremos un modelo de árbol de decisión que explica el fraude global (variable dependiente) en función de los factores determinantes (variables independientes). En primer lugar, obtendremos el factor que mejor explica el fraude global y reiteraremos el procedimiento para ordenar los distintos determinantes de acuerdo con su incidencia sobre el fraude global. Para llevar a cabo este análisis se aplicarán los tres tipos de árboles más utilizados hoy en día son: los árboles CHAID y CHAID exhaustivo, 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 (1963). La variable dependiente suele ser cualitativa (nominal u ordinal) o cuantitativa con valores agrupados en pocos intervalos. Para variables cualitativas, el proceso lleva a cabo una serie de análisis $\chi 2$ entre las variables dependiente y predictoras. 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 ya 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 o el *marketing*. CART se utiliza para árboles de clasificación con variable dependiente cualitativa y para árboles de regresión con variable dependiente cuantitativa.

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. Para un mayor detalle de las técnicas puede consultarse Pérez y Santín (2007) y Pérez (2009, 2010, 2011a y 2011b).

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 IRPF (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.

Una vez seleccionados los factores relacionados con el fraude, en este trabajo se elaborarán modelos predictivos 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, basándose en los datos de la muestra de IRPF utilizando, para ello, el 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, 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.

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:

$$y = F(x_1, x_2, \cdots, x_n) \tag{1}$$

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 para predecir con qué categoría de la variable dependiente se corresponden para clasificar al individuo en la categoría adecuada o perfil.

En este trabajo vamos a utilizar una muestra de IRPF elaborada para este análisis que toma como variables independientes del modelo discriminante las partidas económicas declaradas por el individuo en el modelo 100 de IRPF (prácticamente 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 factor 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.

Como las variables independientes o explicativas de nuestro trabajo están muy correladas, las reduciremos a un grupo mucho menor de variables incorreladas mediante el análisis de componentes principales. De esta forma evitaremos problemas de multicolinealidad, influencia de valores atípicos, normalidad de la variables y confidencialidad en los datos.

El paso siguiente en nuestro análisis consiste en la aplicación del análisis discriminante con las dos grandes finalidades de clasificar a 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 discriminan-*

tes, 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.

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.

Siguiendo la metodología de Fisher, cada una de las funciones discriminantes D_i se obtiene como función lineal de las k variables explicativas X, es decir:

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

Habrá tantas funciones discriminantes como categorías tenga la variable dependiente del modelo. En nuestro caso habrá dos funciones discriminantes, una función F1 correspondiente a la categoría fraude de la variable dependiente y otra función F2 correspondiente a la categoría no fraude de dicha variable. Para clasificar un individuo en una categoría existen dos criterios fundamentales. 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. Un segundo criterio sería utilizar las probabilidades de pertenencia a cada categoría. Un individuo se clasifica en la categoría a la que su pertenencia resulta más probable.

La probabilidad de pertenencia de un individuo a una categoría o grupo *i* de la variable dependiente se evalúa mediante:

$$P_i = \frac{e^{F_i}}{\sum_i e^{F_i}} \tag{3}$$

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

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

$$P_i = \frac{\pi_i e^{F_i}}{\sum_i \pi_i e^{F_i}} \tag{4}$$

4. Datos: la Muestra del IRPF del IEF

El punto de partida de una buena parte de los trabajos realizados sobre el IRPF es la muestra de IRPF del IEF (https://www.ief.es/badespe), que proporciona una base de datos de

gran amplitud (casi dos millones de observaciones por año) y detalle (más de 200 variables personales, familiares y fiscales).[2] Para este trabajo, se ha completado la información que ofrece el IEF con una muestra facilitada de contribuyentes de IRPF con objetivos de investigación por la Agencia Tributaria que permite disponer de datos de gran precisión sobre contribuyentes defraudadores, 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 o con un menor detalle.

En cuanto al *ámbito poblacional*, la población objetivo son las declaraciones presentadas del 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.

Las *unidades de muestreo* son las declaraciones de IRPF del año analizado y la *población marco* la constituye el conjunto de unidades obtenidas a partir de las declaraciones del Modelo 100 de la Agencia Tributaria, al que se puede acceder en: https://sede.agenciatributaria.gob.es/Sede/procedimientoini/G229.shtml. En nuestro análisis es posible considerar totas las opciones de fraude que aparecen en las declaraciones que la Agencia Tributaria registra, una vez inspeccionadas, en sus bases de datos del IRPF.

La muestra de IRPF disponible para este trabajo permite la construcción de un modelo de árbol de decisión 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 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 será posible modelizar el fraude global en función de los distintos factores de fraude y ordenar estos factores de acuerdo con su incidencia en el fraude global.

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. 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 de cualquier ejercicio. Si, además, se consideran todos los problemas de fraude sujetos a recursos en los tribunales, cerrar una base de datos inspeccionada cuesta mucho tiempo. En este trabajo la base de datos que se utiliza es la de 2009, que cumple los requisitos para su aplicación.[3]

5. Resultados y análisis

5.1. Selección de variables relacionadas con el comportamiento del fraude fiscal

El primer objetivo de nuestro trabajo es estimar cuáles han sido las variables relacionadas con el comportamiento de fraude en el IRPF aplicando, para ello, las tipologías de árboles descritas. La finalidad del árbol es estimar cuáles han sido las variables independientes relacionadas con el comportamiento de fraude en el IRPF y ordenar estas variables según su incidencia en el fraude global.

En cuanto a los resultados de los tres modelos, que se recogen en las Figuras 2-4 del Anexo, se observa que la cuantificación de la incidencia de los distintos factores sobre el fraude global sigue el mismo orden en los tres árboles y con niveles de significación altos. Las causas o factores de fraude encontradas por los tres modelos permiten obtener una información muy importante sobre las declaraciones en las que se comprueba que la principal causa de fraude procede de la incorrecta declaración de las fuentes de renta y el resto de los componentes de la base liquidable que afecta al tipo marginal. De todas las fuentes de renta, la declaración de actividades económicas es la que tiene una mayor incidencia en el fraude fiscal, siendo a su vez las declaraciones de gastos deducibles que afecta al cálculo de los rendimientos de las actividades económicas, la siguiente variable relaciona con el comportamiento del fraude más importante. Sin embargo, en el caso del árbol CHAID exhaustivo al ser un árbol de decisión menos potente no incluye como significativas otras causas de fraude, detectada por los métodos CRT y QUEST, como el fraude en la desgravación por planes de pensiones y el fraude en la declaración del número de hijos, ascendientes y descendientes.

Tabla 2
DIAGNOSIS DE LOS MODELOS DE ÁRBOL UTILIZADOS PARA DETECTAR EL FRAUDE FISCAL

CHAID EXHAUSTIVO		**CRT**		**QUEST**	
Riesgo		**Riesgo**		**Riesgo**	
Estimación	**Error estándar**	**Estimación**	**Error estándar**	**Estimación**	**Error estándar**
0,037	0,000	0,004	0,000	0,004	0,000
Matriz de confusión:		*Matriz de confusión:*		*Matriz de confusión:*	
96,3 % global clasificación		99,6% global de clasificación		99,6% global de clasificación	
Fraude por tipo impositivo					
36,6%(p-valor:0,00)		36,6%(p-valor:0,00)		36,6%(p-valor:0,00)	
Fraude por declaración de actividades económicas					
17,3%(p-valor:0,00)		17,3%(p-valor:0,00)		17,3%(p-valor:0,00)	
Fraude por declaración de gastos					
5 %(p-valor:0,00)		5 %(p-valor:0,00)		5 %(p-valor:0,00)	
Fraude por desgravación de planes de pensiones					
—		2,2 %(p-valor:0,00)		2,2 %(p-valor:0,00)	
Fraude por declaración de hijos y ascendientes y descendientes					
—		1,1 %(p-valor:0,00)		1,1 %(p-valor:0,00)	

Fuente: elaboración propia.

Los resultados en cuanto a la diagnosis de los tres modelos de árbol utilizados (EXHAUSTIVE CHAID, CRT y QUEST), se recogen en la Tabla 2, donde se confirma lo comentado anteriormente. Se observa que el ajuste matemático de todos los modelos de árbol ha sido de calidad, con una significatividad alta de los factores principales que inciden sobre el fraude fiscal, riesgos pequeños y matrices de confusión con alto porcentaje de aciertos para todos los métodos. No obstante, los modelos CRT y QUEST son superiores al CHAID exhaustivo con porcentajes de riesgo más pequeños y porcentajes de aciertos en la matriz de confusión superiores.

A continuación, se analizan los factores detectados, presentándose por orden de importancia e incidencia sobre el fraude fiscal.

El factor principal de fraude es la manipulación del tipo marginal aplicable, ya que es el factor o causa más incidente en el fraude global, y la primera variable en poder discriminatorio del árbol con un 36,6% de resultado. 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.

Al reiterar el procedimiento para ordenar el resto de los factores, comprobamos que de las fuentes de rentas la que está más relacionada con el comportamiento del fraude (con una incidencia del 17,3%), 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 o al indebido tratamiento de algunas de las partidas declaradas. 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 importante papel que tienen los Rendimientos de Actividades Económicas en el análisis del fraude se corrobora al analizar el tercer componente o factor con mayor incidencia en el fraude: la incorrecta declaración de gastos desgravables (con un 5% de incidencia en el fraude), 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 este caso a la deducción de gastos incorrecta o a 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.

También aparece entre los factores de fraude significativos con incidencia en el fraude global, con un 2,2%, el fraude que afecta a la desgravación de planes de pensiones. Esta rú-

brica 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. Pero es interesante este resultado, puesto que se comprueba lo adecuado de las medidas para su control.

Por último, el análisis ha señalado como otro de los factores relacionados con el comportamiento del fraude en el ejercicio analizado (con un 1,1% de incidencia), al 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.

El resto de las variables analizadas no han resultado que tengan una incidencia significativa en el fraude global. Este resultado difiere de los obtenidos en otros trabajos, como el de Domínguez-Barrero *et al*., 2017, en el que, aplicando la metodología de Felman y Slemrod (2007), encuentran aumentos en los niveles de fraude en fuentes de rentas como las de capital inmobiliario. No obstante, tal y como los autores señalan, estos resultados presentan limitaciones, ya que no siempre es posible interpretar como fraude fiscal los casos en los que se detectan declaraciones con ingresos inferiores a los esperados en función de la información presentada. En este sentido, hay que destacar el valor del análisis realizado en este trabajo, al disponer de una muestra que cuenta con la información sobre los contribuyentes que han sido señalados como defraudadores y se cuenta con la información sobre el fraude cometido.

5.2. Modelización discriminante para detectar el posible fraude fiscal

Una vez determinados los principales factores relacionados con el fraude fiscal a partir de las técnicas aplicadas a los datos disponibles, se han elaborado modelos predictivos de análisis discriminante que permiten cuantificar la probabilidad que tiene cualquier contribuyente actual o futuro de ser defraudador globalmente y por cada factor de fraude, una vez que presente su declaración de IRPF, basándose en los datos previos de la muestra de IRPF utilizados. La caracterización de los contribuyentes según el factor de fraude nos permite llevar a cabo la segmentación de los declarantes del impuesto por nivel de propensión al fraude y factor.

Para la selección de las variables independientes se ha aplicado el Análisis de Componentes Principales (ACP) por las razones comentadas anteriormente. Además, la reducción por componentes principales es procedente porque el determinante de la matriz de correlaciones de las variables iniciales es prácticamente nulo y las comunalidades de las variables están muy cercanas a la unidad. El resultado de la reducción ofrece 64 componentes principales Ci (factores) que explican cerca del 80% de la variabilidad inicial de los datos, resultando así una buena reducción. En concreto explican el 79,597% de la variabilidad. La matriz factorial resultante después de una rotación VARIMAX permite obtener las combinaciones lineales de las componentes en función de las variables iniciales, que serán las siguientes:

$$
\begin{aligned}
C_1 &= 0{,}0723X_1 + 0{,}492X_2 + 0{,}502X_3 + 0{,}497X_4 + \cdots \\
C_2 &= -0{,}329X_1 - 0{,}307X_2 - 0{,}310X_3 - 0{,}307X_4 + \cdots \\
C_3 &= 0{,}132X_1 - 0{,}442X_2 - 0{,}441X_3 - 0{,}443X_4 + \cdots \\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots
\end{aligned}
\tag{5}
$$

El siguiente paso en nuestro análisis será la estimación de los modelos discriminantes para el fraude global y para los distintos factores de fraude que indicen en el fraude global utilizando como variables independientes las puntuaciones de las componentes principales.

Resulta entonces que una vez estimados los modelos discriminantes Fisher para el fraude global y para los cinco factores de fraude detectados en el apartado anterior, tomando como variables independientes los 64 factores (ninguno resulta expulsado del modelo por los criterios de selección de variables discriminantes), se obtienen los coeficientes de las funciones discriminantes cuyos resultados se presentan en la Tabla 3. Se puede observar según los estadísticos de la tabla que el ajuste matemático de todos los modelos de análisis discriminante ha sido de calidad, con significatividad alta de los modelos, matrices de confusión con alto porcentaje de aciertos (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 (*Receiver Operating Characteristic* o Característica Operativa del Receptor) muy cercanas a la unidad, como se puede comprobar en la Figura 5 del Anexo. En cuanto a la diagnosis de los modelos, observamos los resultados del contraste de la Lambda de Wilks cuyo p-valor (Sig.) pequeño valida la significatividad de cada uno de los modelos discriminantes en su conjunto. Además, los resultados de la prueba M de Box, cuyo p-valor es pequeño, muestra la ausencia de heteroscedasticidad en los modelos.

Tabla 3
RESULTADOS DE LAS FUNCIONES DISCRIMINANTES DE FISHER (2) PARA EL FRAUDE FISCAL

Modelo discriminante	Funciones discriminantes de Fisher $D_i = u_{i1}C_1 + u_{i2}C_2 + \ldots + u_{ik}C_k \quad i = 1,2$	Lamda-Wilks	M-Box	Area ROC	% clas.
	D0 = -0,378C_1 - 0,05C_2 +…- 0,005C_{64} − 0,776	0,651	69380,4	0,886	79,1
	D1 = 0,226C_1 + 0,03C_2 +…+ 0,008C_{64} − 0,942	(0,00)	(0,00)	(0,00)	(0,00)
$f\_tmg = F(C_1, C_2, \ldots, C_n) + e$	D0 = -0,115C_1 - 0,15C_2 +…- 0,005C_{64} − 0,776	0,77	28589,8	0,832	77,5
	D1 = 0,199C_1 + 0,026C_2 +…+ 0,008C_{64} − 0,942	(0,00)	(0,00)	(0,00)	
$f\_aaee = F(C_1, C_2, \ldots, C_n) + e$	D0 = -0,404C_1 + 0,26C_2 +…- 0,003C_{64} − 1,006	0,458	23616,3	0,947	89,7
	D1 = -0,762C_1 - 0,049C_2 +…+ 0,005C_{64} − 1,808	(0,00)	(0,00)	(0,00)	
$f\_gastos = F(C_1, C_2, \ldots, C_n) + e$	D0 = -0,019C_1 +…- 0,003C_{64} − 0,712	0,797	25713,8	0,856	89,5
	D1 = -0,130C_1 +…+ 0,005C_{64} − 1.557	(0,00)	(0,00)	(0,00)	
$f\_planp = F(C_1, C_2, \ldots, C_n) + e$	D0 = -0,203C_1 + 0,011C_2 +…+ 0,001C_{64} − 0,830	0,583	19324,5	0,9	85,7
	D1 = 0,529C_1 − 0,029C_2 +…− 0,003C_{64} − 1,625	(0,00)	(0,00)	(0,00)	
$f\_nhijos = F(C_1, C_2, \ldots, C_n) + e$	D0 = -0,0014C_1 +…− 0,002C_{64} − 0,697	0,864	7407,7	0,895	86,7
	D1 = 0,299C_1 +…− 0,041C_{64} − 2,367	(0,00)	(0,00)	(0,00)	

Fuente: elaboración propia.

Nota: Entre paréntesis aparece el p-valor (significatividad).

Los métodos predictivos para el análisis del fraude permiten calcular la probabilidad de fraude o propensiones al fraude (global o por causas de fraude, para todos los individuos de la muestra) de acuerdo con las ecuaciones (3)-(4). Los resultados de las propensiones al fraude se recogen en la Figura 6 del Anexo, donde se muestra una comparación entre las densidades de probabilidad para las propensiones al fraude, calculadas mediante el algoritmo del Kernel, para las distintas causas de fraude resultantes de nuestra investigación.

En general, la densidad de probabilidad para los diferentes factores de fraude es muy similar. En concreto, se observa la tendencia bimodal de las propensiones al fraude, donde la densidad es alta para valores pequeños de probabilidad de fraude (hay muchos declarantes con probabilidad de fraude baja asociados al primer cuartil de la distribución) y desciende según va aumentando la propensión al fraude (entre el primer y el tercer cuartil) tendiendo a ser constante hasta propensiones de valor 0,8. Finalmente, la densidad es alta para valores altos de probabilidad de fraude (vuelve a haber muchos declarantes con probabilidad de fraude alta por encima del tercer cuartil).

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 los diferentes factores de fraude.

Los resultados obtenidos la diagnosis de los modelos empleados (Tablas 2 y 3) son muy satisfactorios, con una ratio de eficiencia media que alcanza un 89,3%, que mejora los obtenidos en otros trabajos para España y centrados también en el IRPF, como son los realizados por Pérez *et al*., 2019 y González *et al*., 2021. En primer lugar, la evidencia obtenida confirma las mayores posibilidades de análisis de la propuesta realizada basada en el uso de técnicas de aprendizaje supervisado frente a los trabajos que usan aprendizaje no supervisado, como es el caso de González *et al*., 2021, que requiere asumir supuestos para la detección de los casos de fraude a partir de los datos considerados como anómalos por el modelo. Además, también mejora otros que usan el aprendizaje supervisado, pero con herramientas alternativos, como es el análisis de redes neuronales de Pérez *et al*., 2019, en el que se obtiene un 84,2% de eficiencia, inferior a la obtenida en nuestro trabajo.

5.3. Cuantificación agregada del nivel de fraude fiscal y evolución del índice de fraude en IRPF en España

Con el objetivo de realizar un análisis complementario empleando la metodología propuesta, se ha considerado llevar a cabo también una cuantificación agregada del nivel de fraude.[4] Este análisis permite reforzar el objetivo esencial de este trabajo que es establecer una metodología para el análisis del fraude fiscal en IRPF, aplicable incluso a otras figuras impositivas y que permite realizar una cuantificación agregada del nivel de fraude. Dado que en el trabajo se calculan las probabilidades de fraude asignadas a cada contribuyente, es posible utilizar el valor esperado medio de estas probabilidades para computar el fraude

agregado en IRPF y comprobar la robustez de nuestros resultados. La estimación realizada nos da como resultado una cifra del 28,2%.

Por otra parte, también es muy interesante calcular la evolución en el tiempo del fraude en IRPF. Inicialmente podría pensarse en aplicar la metodología aquí expuesta para calcular año a año las probabilidades de fraude basándose en un conjunto de datos como es el Panel de IRPF elaborado por el Instituto de Estudios Fiscales con datos de la Agencia Tributaria. Pero este método tiene dos desventajas claras. La primera es la disponibilidad de datos de fraude para los sucesivos años y la segunda es que, como la mayoría de los contribuyentes del panel son los mismos en los diferentes años, la variable fraude no recogería la variabilidad real del fraude en la población de todos los declarantes. No olvidemos que el hecho de que un individuo haya sido detectado como fraudulento en un año dado, condiciona su comportamiento futuro respecto al fraude.

No obstante, puede calcularse la evolución del fraude en IRPF por otra vía alternativa. Por una parte, disponemos de los ingresos declarados por los contribuyentes en los sucesivos años a través de los microdatos del Panel de Renta o de las muestras anuales de IRPF. Al tratarse de registros administrativos a nivel micro, esta información es fehaciente y presenta poco error. La metodología utilizada para calcular los índices de fraude en IRPF sigue la propuesta en Lagares (1987) y adaptada posteriormente en Pulido (2014). De esta forma disponemos de la recaudación efectiva o ingresos declarados. Por otra parte, disponemos de la recaudación potencial en IRPF a través de las cifras de la Contabilidad Nacional[5], es decir, disponemos de la recaudación que percibiría el Estado sí todos los contribuyentes cumplieran de forma estricta la norma.

La comparación entre la recaudación potencial (lo que se debiera recaudar) y la recaudación efectiva (lo que realmente se recauda), que se recogen en las columnas 2 y 3 de la tabla 5 del anexo, puede utilizarse para cuantificar el grado de cumplimiento y el índice de fraude en IRPF. De esta forma podríamos valorar adecuadamente la evolución del fraude en IRPF en el tiempo.

Podemos definir el grado de cumplimiento como:

$$Grado\ de\ cumplimiento = \frac{Ingresos\ declarados}{Recaudación\ potencial} \tag{6}$$

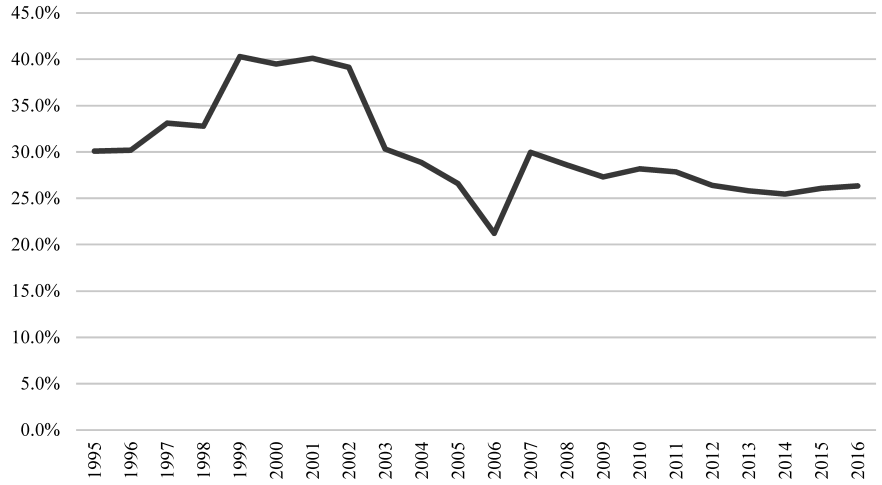
A continuación, se define el índice de fraude como:

$$Índice\ de\ Fraude = 1 - Grado\ de\ Cumplimiento \tag{7}$$

En cuanto a la evolución del fraude (Figura 1), si se evalúa la evolución del Índice de fraude en IRPF en el periodo analizado a partir de la información disponible tanto en el panel de declarantes, como en la Contabilidad Nacional, se observa que en períodos de crecimiento económico el nivel de fraude decrece, mientras que en períodos de crisis el nivel de fraude aumenta volviendo posteriormente a su cauce normal a medida que la economía se recupera. Por otra parte, para el año 2009 el fraude estimado en IRPF por la vía de las probabilidades

de fraude de los contribuyentes se valoró en un 28,2%, mientras que a través de la evolución de los índices de fraude se valoró en un 27,3%. Los resultados no son muy diferentes.

Figura 1
EVOLUCIÓN DEL ÍNDICE DE FRAUDE EN IRPF EN ESPAÑA (1995-2016)



Fuente: elaboración propia a partir del Panel de declarantes de IRPF (Agencia Tributaria: Estadística de los declarantes del IRPF) y Contabilidad Nacional (Ministerio de Hacienda y Función Pública).

6. Conclusiones

Un aspecto importante a la hora de controlar el fraude fiscal es disponer de herramientas que faciliten la caracterización y detección de los defraudadores. Con este objetivo, en este trabajo se ofrece una propuesta de análisis en la que, en primer lugar, se identifican las variables del impuesto más relacionadas con el comportamiento de fraude y se cuantifica su incidencia mediante el empleo de árboles de decisión. La diagnosis de estos modelos ha resultado muy correcta, superando a los resultados de otros trabajos realizados sobre este impuesto en España (Perez *et al*., 2019, González *et al*., 2021). La propuesta de análisis que se ofrece en esta investigación constituye una contribución importante a la hora de establecer planes de prevención del fraude y reformas fiscales orientadas a aquellas tipologías o factores de fraude más habitualmente utilizadas por los contribuyentes.

El análisis de los resultados obtenidos muestra que el principal factor que incide en el fraude en IRPF es la manipulación del tipo marginal aplicable, que viene derivada principalmente de la ocultación de ingresos por diferentes fuentes de renta en la base liquidable del impuesto y no por variables relacionadas con la consignación de las deducciones en la cuota. De esta forma, el tipo marginal correspondiente a la declaración resulta inferior al real, alterando así el resultado de la liquidación. Las cuantías defraudadas por esta causa suelen ser de elevada magnitud. Al reiterar el procedimiento, comprobamos que, de las distintas fuentes

de renta del contribuyente, la que tiene mayor incidencia en el fraude en el IRPF resulta ser la declaración incorrecta de actividades económicas. El difícil control de las rentas de autónomos lleva habitualmente a ingresos no declarados de sus actividades. Nuestros resultados también han puesto de manifiesto que es la incorrecta declaración de gastos desgravables en el cálculo de los rendimientos de actividades económicas, bien sea por la inobservancia de las normas legales o por la realización de artificios engañosos para eludirlas, el factor de fraude más destacado en este análisis. Para la Agencia Tributaria el control de las rentas de los autónomos es uno de los principales aspectos en los que debe centrar su atención.

A partir de la información obtenida es posible caracterizar a los contribuyentes según el factor de fraude. Para ello, se han elaborado modelos predictivos de Análisis Discriminante que permiten cuantificar la probabilidad o propensión que tiene cualquier contribuyente actual o futuro al fraude global y a cada factor de fraude, una vez que presente su declaración de IRPF. Estos modelos han hecho posible la segmentación de los declarantes del impuesto por nivel de propensión al fraude, lo que supone una propuesta de gran interés para las Agencias Tributarias, ya que les permitiría priorizar sus inspecciones fiscales a partir de los modelos obtenidos, siendo este uno de los objetivos que destacan en las directrices generales del Plan Anual de Control Tributario y Aduanero (*BOE*, 2022). De hecho, dentro de las medidas de prevención y control del fraude fiscal, la AEAT pretende mejorar la valoración del riesgo recaudatorio a través de herramientas informáticas de minería de datos. Una vez realizada esta definición de riesgos, se diseñarán los perfiles que combinen y ponderen los riesgos definidos, creando grupos para seleccionar contribuyentes que sean objeto de especial seguimiento.

Las orientaciones para la actuación de la Agencia Tributaria en los próximos años están en línea con los objetivos propuestos en este trabajo. Entre las acciones dirigidas a mejorar el cumplimiento tributario se encuentra el uso de técnicas basadas en la comprensión del comportamiento del contribuyente, como la que se plantea en este trabajo. Además, pretende desarrollar actuaciones de comprobación e investigación sobre los obligados tributarios en lo que concurran perfiles de riesgo, lo que requiere la definición previa de los criterios básicos y de las áreas de riesgo fiscal que se consideren de atención prioritaria para el ejercicio. Para llevar a cabo esta tarea el uso de las herramientas propuestas en este trabajo puede ser de gran interés, ya que para la definición de estos perfiles de riesgo se crearán grupos para seleccionar contribuyentes que sean objeto de especial seguimiento.

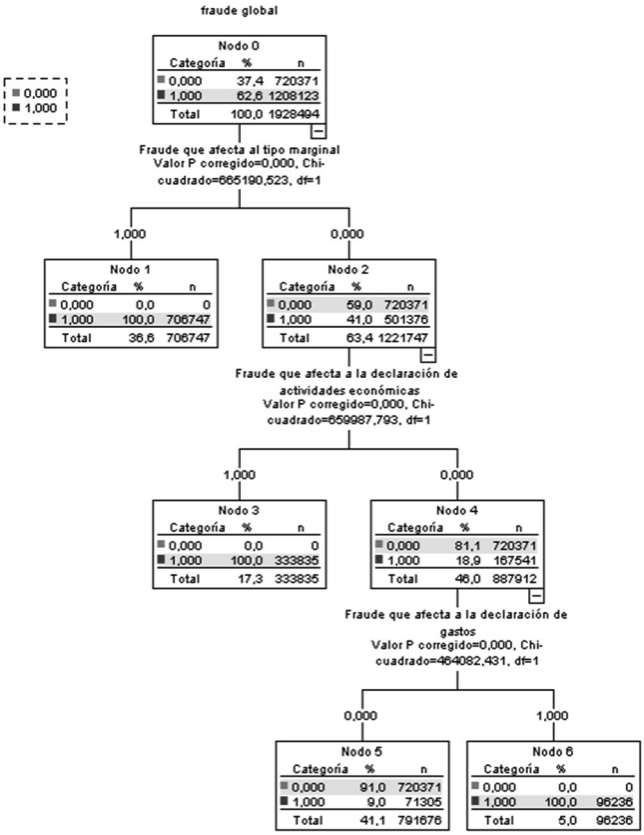
Otra de las medidas planteadas por la Agencia Tributaria se centra en el empleo de la técnica de trabajo, basada en los principios de la OCDE conocidos como “Behavioural insights” consistente en complementar el plan extensivo anual de visitas a determinadas actividades empresariales con la emisión de un número de cartas de aviso dirigidas a aquellos contribuyentes, de los sectores elegidos, que incurran en determinados parámetros de riesgo de incumplimiento cuando se observe que dichos parámetros se vienen manteniendo de forma continuada a lo largo de los últimos ejercicios. En este proceso también pueden ser de interés las herramientas propuestas en este trabajo con las que es posible determinar la probabilidad de fraude fiscal de los contribuyentes. En el análisis realizado, uno de los patrones de mayor interés son los contribuyentes que realizan actividades económicas, por lo que la evidencia obtenida puede contribuir a reforzar este tipo de acciones de la Agencia Tributaria.

Finalmente, un aspecto de interés de la propuesta realizada es la posibilidad de ser aplicada a cualquier muestra de declarantes para los que se disponga de información sobre los contribuyentes señalados por la Administración como defraudadores. Para las Agencias Tributarias realizar este tipo de análisis puede ampliar su capacidad de control del fraude fiscal. Además, la metodología expuesta en este trabajo es aplicable al análisis del fraude en cualquier otro impuesto y para otros años disponibles de IRPF, cuando sea posible disponer de una muestra con las mismas características que la que hemos dispuesto en este trabajo. En este sentido, en el futuro también sería muy interesante explorar las posibilidades que ofrece utilizar la metodología presentada con otras muestras, como los paneles de declarantes. Del mismo modo, explorar con más detalle la evolución del incumplimiento de contribuyentes ya investigados y analizar la evolución de la composición y magnitud relativa de sus fuentes de renta. Estas extensiones de esta investigación pueden ser muy útiles para el avance del estudio del fraude fiscal en España.

Es necesario tener presente que el aspecto cuantitativo de este trabajo se desarrolla en el campo de los grandes datos (Big Data). Los análisis realizados se han ajustado para dos millones de registros. Con software normal no sería posible realizar estas tareas. Se ha utilizado software de IBM que incorpora la posibilidad de trabajar en los campos del Big Data y la Minería de Datos.

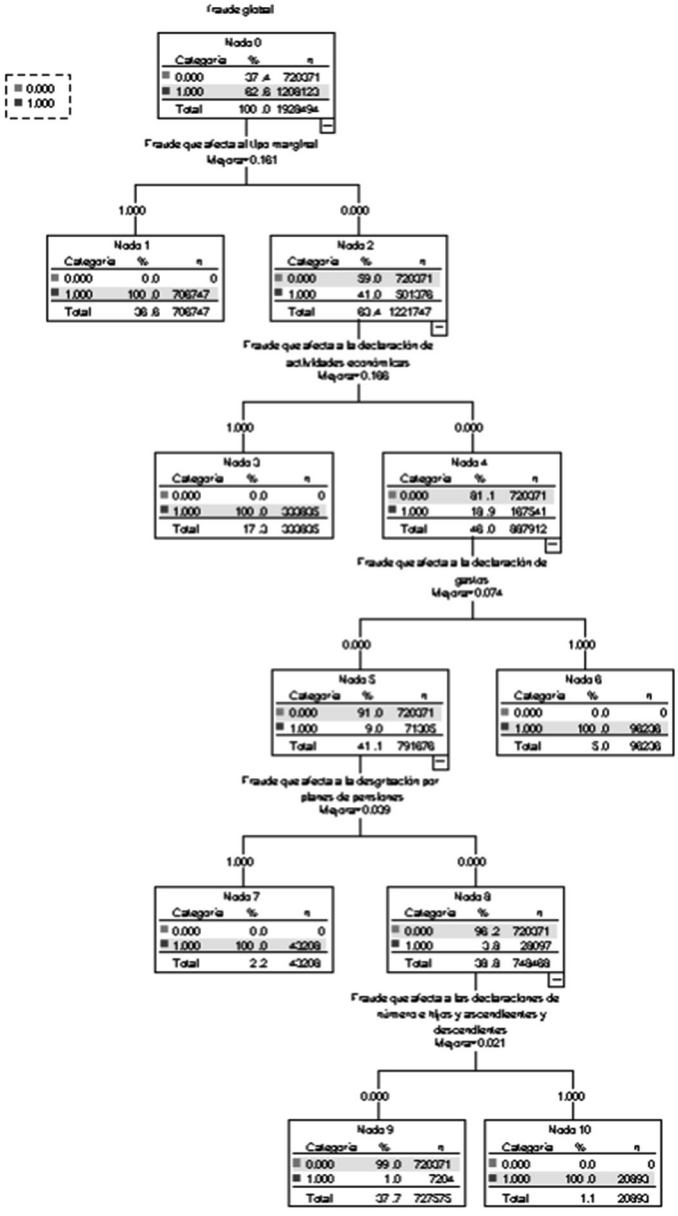
Anexo

Figura 2
RESULTADOS DE LA ESTIMACIÓN DEL ÁRBOL DE CLASIFICACIÓN CHAID EXHAUSTIVO PARA EL FRAUDE



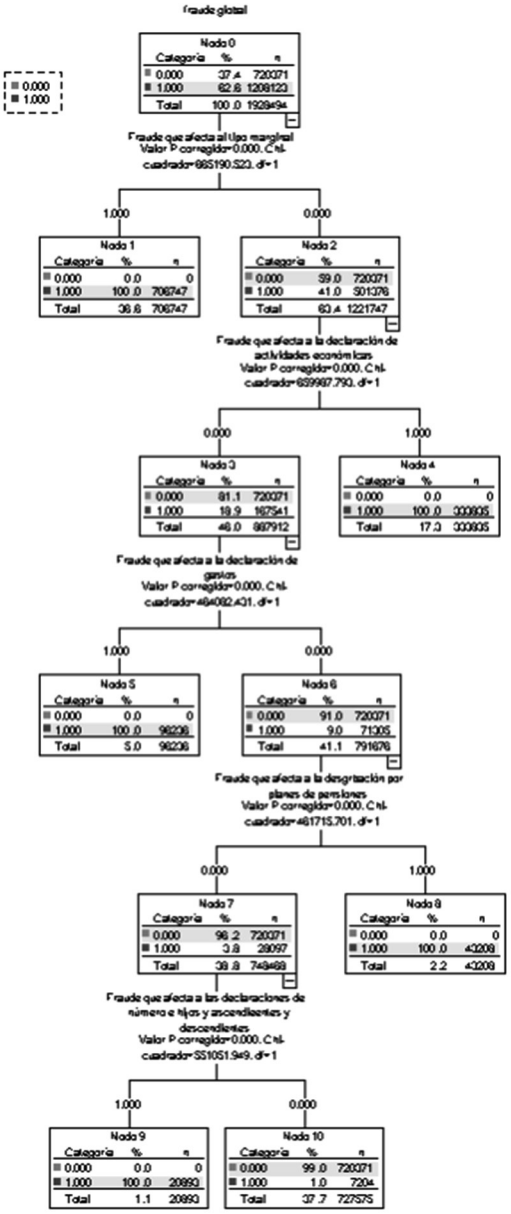
Fuente: elaboración propia.

Figura 3
RESULTADOS DE LA ESTIMACIÓN DEL ÁRBOL DE CLASIFICACIÓN CRT PARA EL FRAUDE



Fuente: elaboración propia.

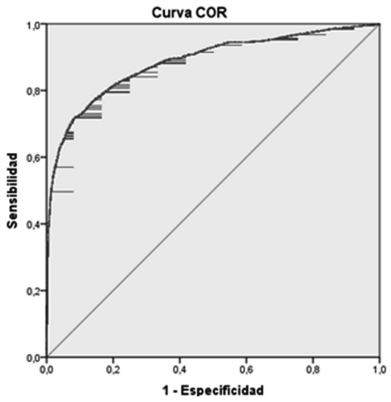
Figura 4
RESULTADOS DE LA ESTIMACIÓN DEL ÁRBOL DE CLASIFICACIÓN QUEST PARA EL FRAUDE



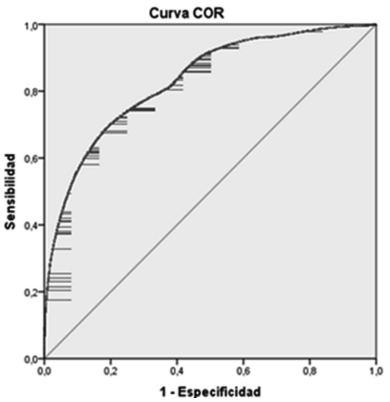
Fuente: elaboración propia.

Figura 5
CURVAS ROC DE LOS MODELOS DISCRIMINANTES PARA EL FRAUDE

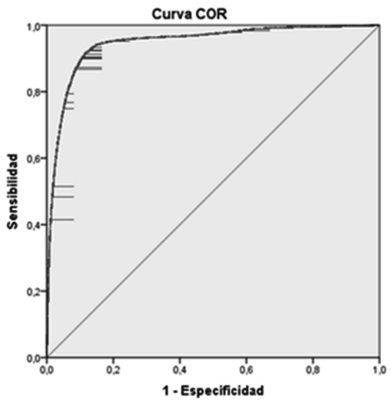
I. *f_Global*



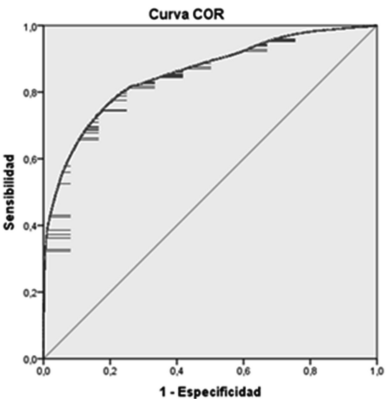
II. *f_tmg*



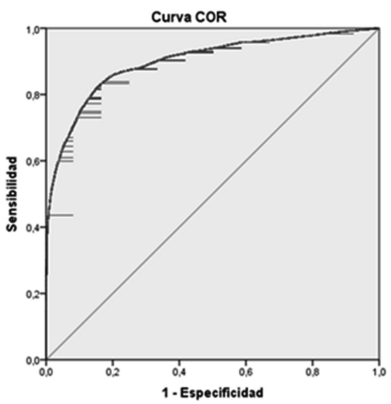
III. *f_aaee*



IV. *f_gastos*



V. *f_nhijos*



VI. *f_planp*

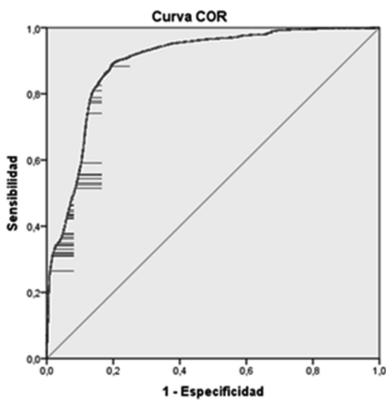
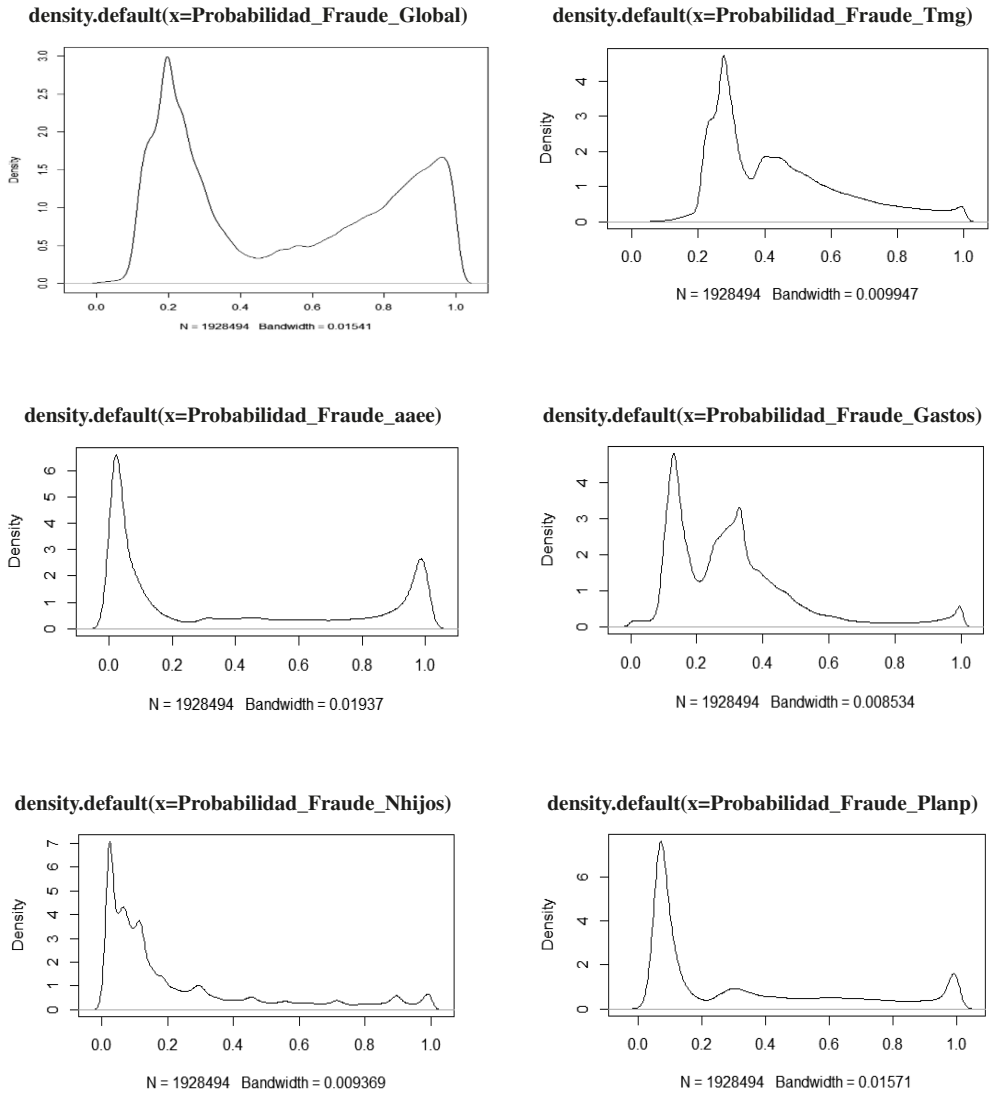


Figura 6
FUNCIONES DE DENSIDAD DE PROBABILIDAD DE PROPENSIONES AL FRAUDE FISCAL



Fuente: elaboración propia.

Tabla 4
DATOS SOBRE LA RECAUDACIÓN DE IRPF EN ESPAÑA (1995-2016)

Años	Recaudación Potencial (millones euros)	Ingresos declarados (millones de euros)	Cumplimiento Fiscal Ecuación (6)	Índice de Fraude en IRPF Ecuación (7)
1995	262,605	183,580	69,9%	30,1%
1996	279,339	195,025	69,8%	30,2%
1997	298,794	199,823	66,9%	33,1%
1998	318,591	214,187	67,2%	32,8%
1999	339,055	202,424	59,7%	40,3%
2000	368,059	222,653	60,5%	39,5%
2001	402,412	241,053	59,9%	40,1%
2002	422,137	256,854	60,8%	39,2%
2003	450,916	314,114	69,7%	30,3%
2004	474,399	337,423	71,1%	28,9%
2005	504,575	370,313	73,4%	26,6%
2006	535,145	421,576	78,8%	21,2%
2007	566,129	396,453	70,0%	30,0%
2008	577,692	412,399	71,4%	28,6%
2009	553,862	402,651	72,7%	27,3%
2010	542,565	389,610	71,8%	28,2%
2011	543,611	392,244	72,2%	27,8%
2012	505,145	371,795	73,6%	26,4%
2013	498,573	369,913	74,2%	25,8%
2014	503,117	375,010	74,5%	25,5%
2015	530,444	392,187	73,9%	26,1%
2016	546,644	402,617	73,7%	26,3%

Fuente: elaboración propia a partir del Panel de declarantes de IRPF (Agencia Tributaria: Estadística de los declarantes del IRPF) y Contabilidad Nacional (Ministerio de Hacienda y Función Pública.

Notas

1. Para la implementación de estas técnicas se ha utilizado el *software IBM SPSS Modeler*.
2. En Onrubia *et al*. (2011 y 2012) y en Pérez *et al*., (2012) se describe la muestra y el panel de declarantes que se elabora en el Instituto de Estudios Fiscales.
3. Para un mayor detalle sobre la muestra de declarantes y el esquema de IRPF en ese año, puede consultarse Pérez *et al*., 2012.
4. Agradecemos la sugerencia realizada por uno de los evaluadores de este trabajo para incorporar este análisis.
5. Para más detalle sobre esta información se puede acceder en IGAE: Contabilidad Nacional (hacienda.gob.es).

Referencias

Alm, J. (2021), "Tax evasion, technology and inequality", *Economics of Governance*, 22: 321-343. https://link.springer.com/article/10.1007/s10101-021-00247-w.

Alm, J. (2012), "Measuring, explaining, and controlling tax evasion: lessons from theory, experiments, and field studies", *International Tax and Public Finance*, 19: 54-77. https://doi.org/10.1007/s10797-011-9171-2.

Alognon, A. D., Koumpias, A. M. y Martínez-Vazquez, J. (2021), "The impact of plastic money use on VAT compliance: evidence from EU countries", *Hacienda Pública Española/Review of Public Economics*, 239(4): 5-26. https://dx.doi.org/10.7866/HPE-RPE.21.4.1.

Almunia, M. y Lopez-Rodríguez, D. (2018), "Under the radar: the effects of monitoring firms on tax compliance", *American Economic Journal: Economic Policy*, 20(1): 1-38. https://doi.org/10.1257/pol.20160229.

Alstadsæter, A., Johannesen, N. y Zucman, G. (2019), "Tax evasion and inequality", *American Economic Review*, 109(6): 2073-2103. https://doi.org/10.1257/aer.20172043.

Ameur, F. y Tkiouat, M. (2012), "Taxpayers fraudulent behaviour modeling the use of datamining in fiscal fraud detecting Moroccan case", *Applied Mathematics*, 3: 1207-13. http://dx.doi.org/10.4236/am.2012.310176.

Basta, S., Fassetti, F., Guarascio, M., Manco, G., Giannotti, F., Pedreschi, D., Spinsanti, L., Papi, G. y Pisani, S. (2009), "High Quality True-Positive Prediction for Fiscal Fraud Detection" *IEEE International Conference on Data Mining Workshops*, 7-12. https://doi.org/10.1109/ICDMW.2009.59.

BOE (2022), "Plan General de Control Tributario y Aduanero", Ministerio de Hacienda y Función Pública, *BOE*-A-2022-1453.pdf. https://www.boe.es/eli/es/res/2022/01/26/(1).

Brondolo, J., Chooi, A., Schloss, T. y Siouclis, A. (2022), "Compliance risk management: developing compliance improvement plans", *Technical Notes and Manuals*, N.º 2022/001, IMF, Washington.

Buehn, A. y Schneider, F. (2016), "Size and development of Tax Evasion in 38 OECD countries: What do we (not) know?", *Journal of Economics and Political Economy*, 3(1): 1-11. www.kspjournals.org.

Castellón, P. y Velásquez, J. D. (2013), "Characterization and detection of taxpayers with false invoices using data mining techniques", *Expert systems with Applications*, 40: 1427-1436. https://doi.org/10.1016/j.eswa.2012.08.051.

Chica, M., Hernández, J. M., Manriquede Lara Penate, C. y Chiong, R. (2021), "An evolutionary game model for understanding fraud in consumption taxes", *IEEE Computational Intelligence Magazine*, 16(2): 62-76. https://doi.org/10.1109/MCI.2021.3061878.

Da Silva, L. S., Rigitano, H., Carvalho, R. y Souza, C. (2016), "Bayesian networks on income tax audit selection-a case study of Brazilian tax administration", en Carvalho, R. N. y Laskey, K. B. (eds.), *CEUR Workshop Proceedings*, vol. 1663. www.ceur-ws-org.

De Roux, D., Pérez, B., Moreno, A., Villamil, M. P. y Figueroa, C. (2018), "Tax Fraud Detection for Under-Reporting Declarations Using an Unsupervised Machine Learning Approach", *Proceedings del 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. https://doi.org/10.1145/3219819.3219878.

Domínguez-Barrero, F., López-Laborda, J. y Rodrigo-Sauco, F. (2015a), "El hueco que deja el diablo: Una estimación del fraude en el IRPF con Microdatos Tributarios", *Revista de Economía Aplicada*, 68(XXIII): 181-102.

Domínguez-Barrero, F., López-Laborda, J. y Rodrigo-Sauco, F. (2015b), "Fraude en el IRPF por fuentes de renta, 2005-2008: del impuesto sintético al impuesto dual", *EEE2015-14*, Madrid.

Domínguez-Barrero, F., López-Laborda, J. y Rodrigo-Sauco, F. (2017), "Tax evasión in Spanish Personal Income Tax by income sources, 2005-2008: from the synthetic to the dual tax", *European Journal of Law and Economics*, 44: 47-65. https://doi.org/10.1007/s10657-016-9553-0 (FALTA).

European Commission (2021), Directorate-General for Taxation and Customs Union.

Poniatowski, G., Bonch-Osmolovskiy, M., Śmietanka A., *VAT gap in the EU: report 2021*, Publications Office. https://data.europa.eu/doi/10.2778/447556 (FALTA).

Feige, E. y Cebula, R. J. (2012), "America's underground economy: measuring the size, growth and determinants of income tax evasion in the U.S", *Crime, Law and Social Change*, 57(3): 265-285. https://doi.org/10.1007/s10611-011-9346-x.

Feldman, N. E. y Slemrod, J. (2007): "Estimating tax noncompliance with evidence from unaudited tax returns", *Economic Journal*, 117: 327-352. https://doi.org/10.1111/j.1468-0297.2007.02020.x.

González, C., Delgado, M. J. y De Lucas, S. (2021) "Segmentation of potential fraud taxpayers and characterization in personal income tax using data mining techniques", *Hacienda Pública Española/ Review of Public Economics*, 239(4): 127-157. https://dx.doi.org/10.7866/HPE-RPE.21.4.4.

González, I., y Caballero, A. (2023), "A Comparison Between Bayesian Dialysis and Machine Learning to Detect Tax Fraud and Its Causes: The Case of Vat", *Corporate Tax and Customs Duties in Spain. SN COMPUT. SCI*, 4: 80. https://doi.org/10.1007/s42979-022-01483-5.

Hilal, W., Gadsden, A. y Yawney, J. (2022), "Financial fraud: a Review of anomaly detection techniques and recent advances", *Expert Systems with Applications*, 193(1): 116429. https://doi.org/10.1016/j.eswa.2021.116429.

Kleanthous, C. y Chatzis, S. (2020), "Gated mixture variational autoencoders for value added tax audit case selection", *Knoledge-Based Systems*, 188(5): 105048. https://doi.org/10.1016/j.knosys.2019.105048.

Lagares Calvo, M. J. (1987): "Metodología utilizada en la estimación del fraude fiscal", *Papeles de Economía Española*, 30-31/1987: 85-89. https://www.funcas.es/wp-content/uploads/Migracion/Articulos/FUNCAS_PEE/030art09.pdf.

Matos, T., de Macedo, J. y Monteiro, J. M. (2015), "An empirical method for discovering tax fraudsters: A real case study of Brazilian fiscal evasion", *Proceedings of the 19th International Database Engineering &Applications Symposium*, 41-48. https://doi.org/10.1145/2790755.2790759.

Mehta, P., Mathews, J., Bisht, D., Suryamukhi, K., Kumar, S. y Babu, Ch. (2020), "Detecting tax evaders using TrustRank and spectral clustering", *International Conference on Business Information Systems*. https://link.springer.com/chapter/10.1007/978-3-030-53337-3_13.

Mittal, S., Reich, O. y Mahajan, A. (2018), "Who is bogus? Using one-sided labels to identify fraudulent firms from tax returns", *Proceedings I ACM SIGCAS Conference on Computing and Sustainable Societies*, New York. http://dx.doi.org/10.1145/3209811.3209824.

Murorunkwere, B. F., Tuyishimire, O., Haughton, D. y Nzabanita, J. (2022), "Fraud Detection Using Neural Networks: A Case Study of Income Tax", *Future Internet*, 14: 168. https://doi.org/10.3390/fi 14060168.

Onrubia, J., Picos, F. y Pérez, C. (2011), "Panel de declarantes de IRPF 1999-2007: diseño, metodología y guía de utilización", Instituto de Estudios Fiscales, Madrid.

Onrubia, J., Picos, F., Pérez, C. y Gallego, M. C. (2012), "Panel de declarantes del IRPF 1999-2008: metodología, estructura y variables", *Documento* 12/2012, Instituto de Estudios Fiscales, Madrid.

O'Reilly, P., Parra, K. y Stemmer, M. (2021), "Exchange of information and bank deposits in International Finances Centres", *Hacienda Pública Española/Review of Public Economics*, 239(4): 27-69. https://dx.doi.org/10.7866/HPE-RPE.21.4.2.

Pérez, C., Delgado, M. J, y De Lucas, S. (2019), "Tax fraud detection through neural networks: an application using a sample of personal income taxpayers", *Future Internet*, 11: 86. http://dx.doi.org/10. 3390/fi11040086.

Pérez, C. y Santín, D. (2007), *Minería de datos. Técnicas y herramientas*, Thomson, Madrid.

Pérez, C. (2009), *Técnicas de análisis multivariante con SPSS*, Garceta Editorial, Madrid.

Pérez, C. (2010), *Técnicas de muestreo estadístico*, Garceta Editorial, Madrid.

Pérez, C. (2011a), *El sistema Estadístico SAS*, Garceta Editorial, Madrid.

Pérez, C. (2011b), *Técnicas de segmentación. Conceptos, herramienta y aplicaciones*, Garceta Editorial, Madrid.

Pérez, C., Burgos, J., Huete, S. y Gallego, C. (2012), "La muestra de declarantes de IRPF 2009", D*ocumento de trabajo*, 11, Instituto de Estudios Fiscales, Madrid.

Pulido Alba, E. J. (2014), *El fraude fiscal en España. Una estimación con datos de Contabilidad Nacional*. http://hdl.handle.net/10366/125760.

Prichard, W., Custers, A., Dom, R., Daveport, S. y Roscitt, M. (2019), "Innovations in tax compliance. Conceptual framework", *Policy Research working paper*, N.º WPS 9032, Washington, D. C., World Bank Group. https://doi.org/10.1596/1813-9450-9032.

Rosid, A. (2022), "Predicting firm's taxpaying behaviour using artificial neural networks: the case of Indonesia", *Working Paper Series*, 22-08, Directorate General of Taxation (DGT). http://dx.doi.org/ 10.2139/ssrn.4185966.

Savić, M., Atanasijević, J., Jakovetić, D. y Krejić, N. (2022), "Tax evasion risk management using a Hybrid Unsupervised Outlier Detection method", *Expert Systems with Applications*, 193: 116409, https://doi.org/10.1016/j.eswa.2021.116409.

Wei, R., Dong, B., Quinghua, Z., Ruan, J. y He, H. (2019), "Unsupervised conditional adversarial networks for tax evasion detection", *IEEE International Conference on Big Data (Big Data)*. https://doi.org/10.1109/BigData47090.2019.9005656.

Wendler, T. y Gröttrup, S. (2021), *Data Mining with SPSS Modeler: Theory, Exercises and Solutions*, Springer Nature, Suiza. https://doi.org/10.1007/978-3-030-54338-9.

Wu, R-S., Ou, C. S., Lin, H-Y., Chang, S. I. y Yen, D. C. (2012), "Using data mining technique to enhance tax evasión detection performance", *Expert Systems with Applications*, 39: 8769-8777. https://doi.org/10.1007/978-3-030-63833-7_12.

Tian, F., Lan, T., Chao, K., Godwin, N., Zheng, Q., Shah, N. y Zhang, F. (2016), "Mining suspicious tax evasion groups in big data", *IEEE Transactions on Knowledge and Data Engineering*, 28(10): 2651-2664. https://doi.org/10.1109/ICDE.2017.19.

Torregrosa-Hetland, S. (2020), "Inequality in tax evasion: the case of the Spanish income tax", *Applied Economic Analysis*, 28(63): 89-109. https://doi.org/10.1108/AEA-01-2020-0007.

Xu, X., Xiong, F. y Zhe, A. (2022), "Using machine learning to predict corporate fraud: evidence based on the GONE Framework", *Journal of Business Ethics*. https://doi.org/10.1007/s10551-022-05120-2.

Zumaya, M., Guerrero, R., Islas, E., Pineda, O., Gershenson, C., Iñiguez, G. I. y Pineda, C. (2021), "Identifying Tax Evasion in Mexico with Tools from Network Science and Machine Learning", en: Granados, O. M. y Nicolás-Carlock, J. R. (eds.), *Corruption Networks. Understanding Complex Systems*, Springer, Cham. https://doi.org/10.1007/978-3-030-81484-7_6.

Abstract

This paper presents a proposal to model and predict the behavior of Personal Income Tax (IRPF) taxpayers with data mining techniques. Decision trees and discriminant analysis are combined to quantify each taxpayer's propensity to fraud using the tax components with the highest incidence of fraud. The model achieves an average efficiency in predictions of more than 89%, allowing respondents to be segmented by level of propensity to fraud. The proposal can be used in the audit and control process carried out by the Tax Agency.

Keywords: Tax fraud, Data mining, Personal income tax, Prediction.

JEL Classification: H26, C55, C38.