Universidad
Rey Juan Carlos

ESCUELA TÉCNICA SUPERIOR
DE INGENIERÍA INFORMÁTICA

Asignatura: APRENDIZAJE AUTOMÁTICO I
Grado en Ciencia e Ingeniería de Datos

Diapositivas de la asignatura

(Fecha del material: Diciembre 2023)

Curso académico 2023-2024

Material docente en abierto de la Universidad Rey Juan Carlos

Autores: Carmen Lancho, Isaac Martín de Diego

Índice de las diapositivas

Introducción al Aprendizaje Automático

Aprendizaje Automático 1 - Grado en Ciencia e Ingeniería de Datos

Curso académico 2023-2024

Asignatura

- Aprendizaje Automático 1
- Grado en Ciencia e Ingeniería de Datos

¿Qué es Machine Learning?

We are drowning in information and starving for knowledge. — John Naisbitt

This deluge of data calls for automated methods of data analysis, which is what machine learning provides. In particular, we define ***machine learning*** *as a set of methods that can automatically detect patterns in data, and then use the uncovered patterns to predict future data, or to perform other kinds of decision making under uncertainty (such as planning how to collect more data!)*

Murphy, K. P. (2012). *Machine learning: a probabilistic perspective*. MIT press.

Un paso atrás... Inteligencia Artificial

Sistemas que piensan como humanos	Sistemas que piensan racionalmente
«El nuevo y excitante esfuerzo de hacer que los computadores piensen… máquinas con mentes, en el más amplio sentido literal». (Haugeland, 1985)	«El estudio de las facultades mentales mediante el uso de modelos computacionales». (Charniak y McDermott, 1985)
«[La automatización de] actividades que vinculamos con procesos de pensamiento humano, actividades como la toma de decisiones, resolución de problemas, aprendizaje…» (Bellman, 1978)	«El estudio de los cálculos que hacen posible percibir, razonar y actuar». (Winston, 1992)
Sistemas que actúan como humanos	**Sistemas que actúan racionalmente**
«El arte de desarrollar máquinas con capacidad para realizar funciones que cuando son realizadas por personas requieren de inteligencia». (Kurzweil, 1990)	«La Inteligencia Computacional es el estudio del diseño de agentes inteligentes». (Poole *et al.*, 1998)
«El estudio de cómo lograr que los computadores realicen tareas que, por el momento, los humanos hacen mejor». (Rich y Knight, 1991)	«IA… está relacionada con conductas inteligentes en artefactos». (Nilsson, 1998)

Figura 1.1 Algunas definiciones de inteligencia artificial, organizadas en cuatro categorías.

Figure 1: Russell, S. J., & Norvig, P. (2010). Artificial intelligence a modern approach. London.

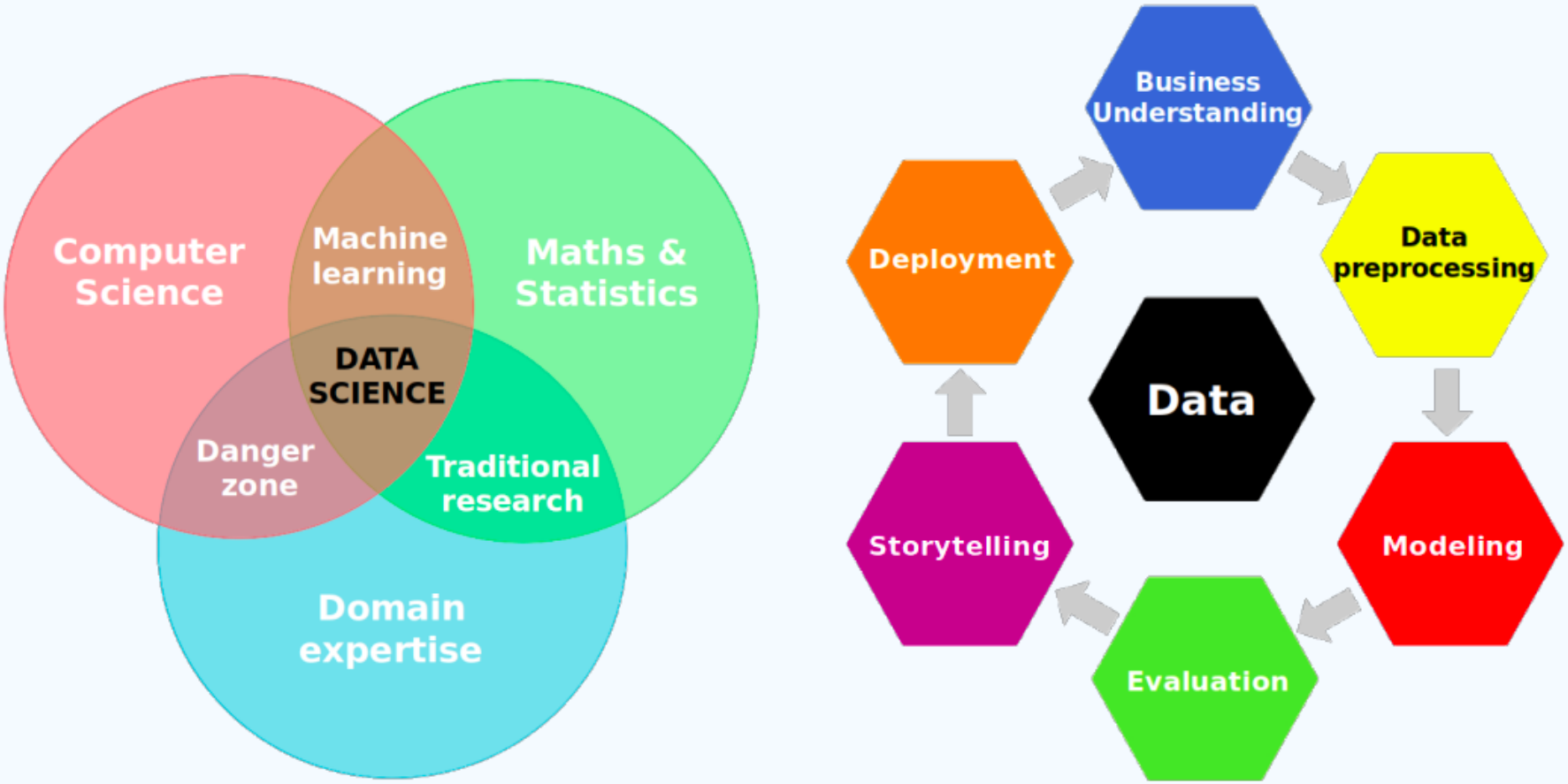
Inteligencia Artificial

La **inteligencia artificial** es un campo de las ciencias de la computación que se enfoca en el desarrollo de sistemas y programas informáticos capaces de realizar tareas que, cuando se ejecutan por parte de seres humanos, requieren inteligencia y aprendizaje. Estos sistemas de IA pueden aprender de datos, adaptarse a nuevas situaciones, tomar decisiones, resolver problemas y realizar tareas específicas sin intervención humana directa. La IA busca imitar y replicar procesos cognitivos y de toma de decisiones humanas, permitiendo a las máquinas realizar actividades que normalmente requerirían la inteligencia humana.

Multitud de disciplinas interconectadas:

- Robótica
- **Aprendizaje Automático**
- Procesamiento del Lenguaje Natural (NLP)
- Visión por computador
- Lógica difusa
- Redes neuronales artificiales –> Deep Learning
- Reconocimiento de patrones
- Computación evolutiva

¿Cómo encajamos todo esto en GCID?



(a) Foundations (b) Applications

Figure 1.1: Data Science

Cross Industry Standard Process for Data Mining (CRISP-DM)

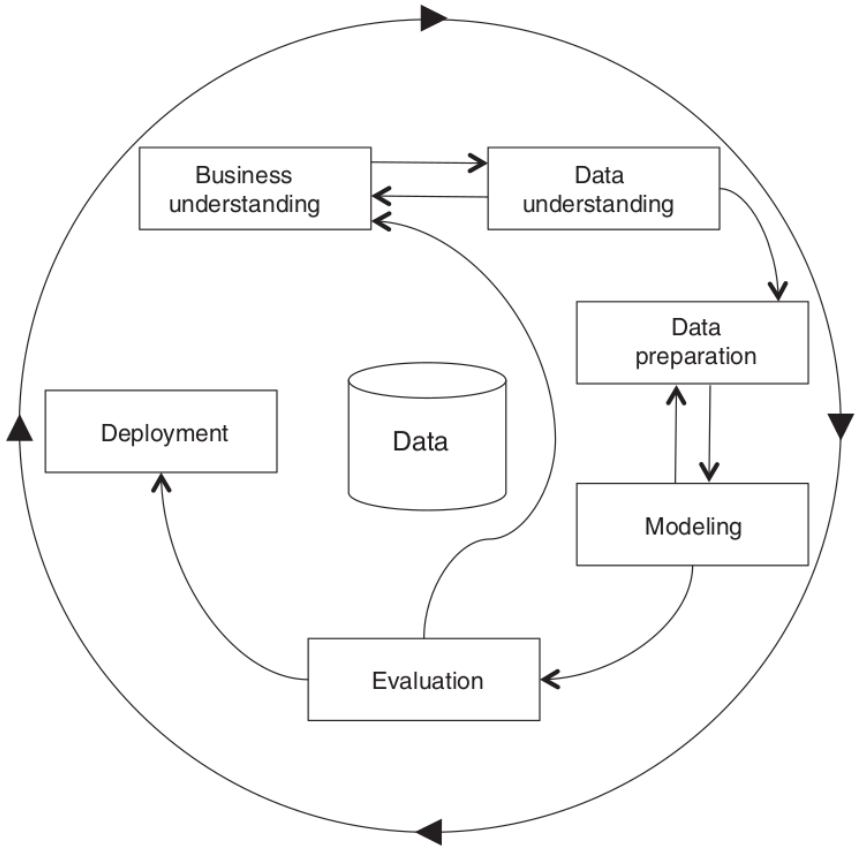


Figure 4 The CRISP-DM life cycle (based on figure 2 in Chapman, Clinton, Kerber, et al. 1999).

Figure 2: Kelleher, J. D., & Tierney, B. (2018). Data science. MIT Press.

CRISP-DM

- **Business Understanding**: Comprender plenamente el problema empresarial que se aborda y diseñar una solución de análisis de datos para el mismo
- **Data Understanding**: Comprender las diferentes fuentes de datos disponibles en la entidad y los diferentes tipos de datos que contienen dichas fuentes
- **Data Preparation**: Poner las distintas fuentes de datos disponibles en un formato adecuado a partir del cual puedan inducirse modelos de aprendizaje automático
- **Modeling**: Crear distintos modelos de aprendizaje automático y selecccionar el mejor para su implantación
- **Evaluation**: Estudiar y validar el rendimiento del modelo para confirmar que es capaz de hacer predicciones precisas antes de ser desplegado
- **Deployment**: Integrar con éxito el modelo de aprendizaje automático en el proceso de la empresa/organización

Tipos de Machine Learning

- **Aprendizaje supervisado**: Datos $\mathbf{x}_i$ con etiquetas, La etiqueta y_i refleja el valor de la variable objetivo (respuesta). El objetivo es "aprender" una función f de los inputs $\mathbf{x}_i$ para predecir los outputs y_i dado un conjunto $\mathcal{D}$ de datos formado por n pares de observaciones:

$$\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$$

  - **Clasificación**: variable objetivo es cualitativa $y_i \in \{1, 2, \ldots, n\}$. Ej: predecir si un paciente tiene una enfermedad o no
  - **Regresión**: variable objetivo es cuantitativa $y_i \in \mathbb{R}$. Ej: predecir la resistencia de un material en base a sus características

Tipos de Machine Learning

- **Aprendizaje no supervisado**: Datos no etiquetados. Se parte de un conjunto de datos $\mathcal{D} = \{\mathbf{x}_i\}_{i=1}^n$ y el objetivo es descubrir patrones de interés entre los datos, describir los datos encontrando grupos de observaciones similares entre sí (Clustering). Ej: Encontrar grupos de clientes similares para una campaña de marketing.

Tipos de Machine Learning

Algunos libros añaden la siguiente categoría:

- **Aprendizaje por refuerzo**. Para aprender cómo actuar a través de recompensas o castigos (un bebé aprendiendo a caminar)

El mundo con datos

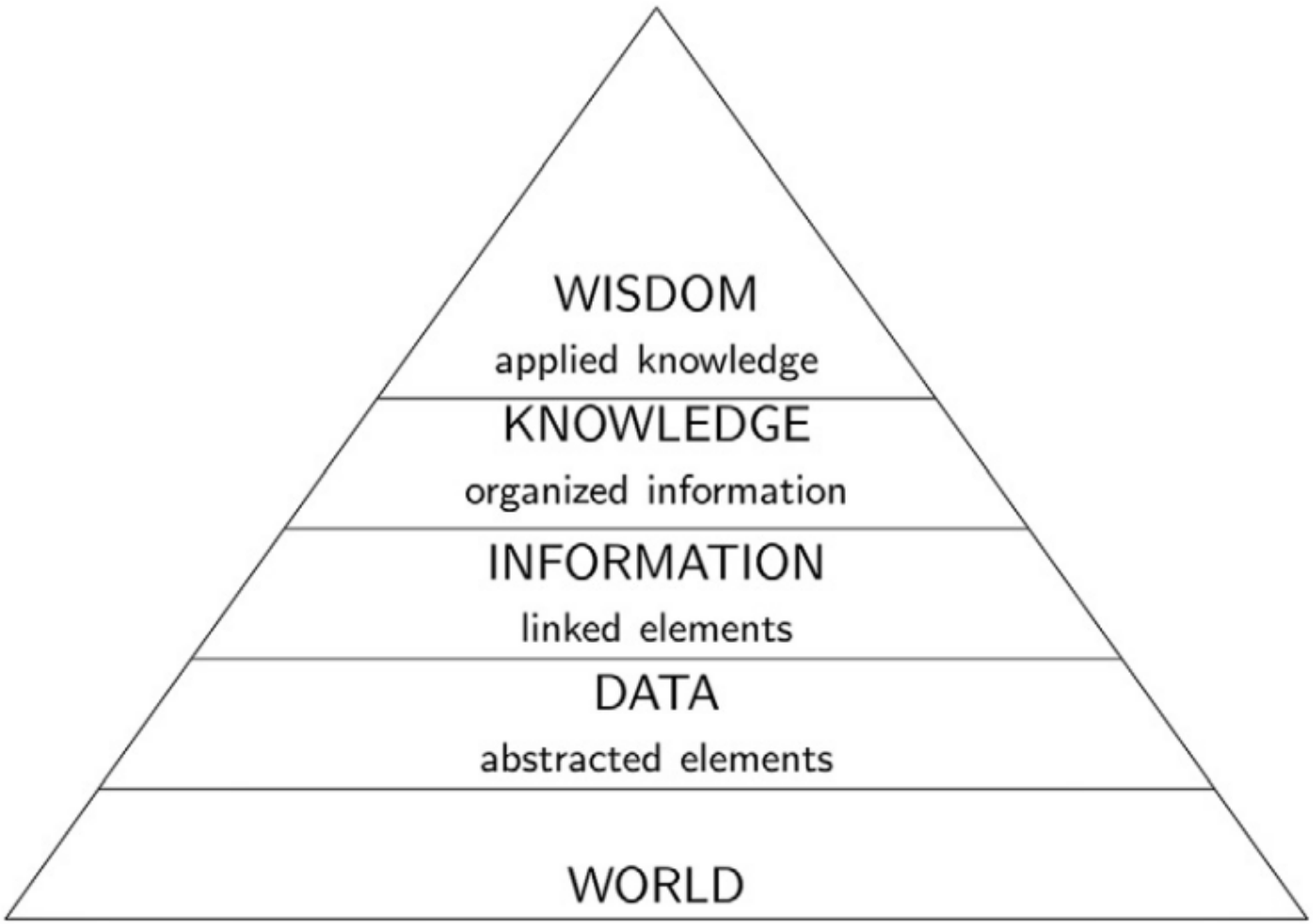


Figure 2 The DIKW pyramid (adapted from Kitchin 2014a).

Figure 3: Kelleher, J. D., & Tierney, B. (2018). Data science. MIT Press.

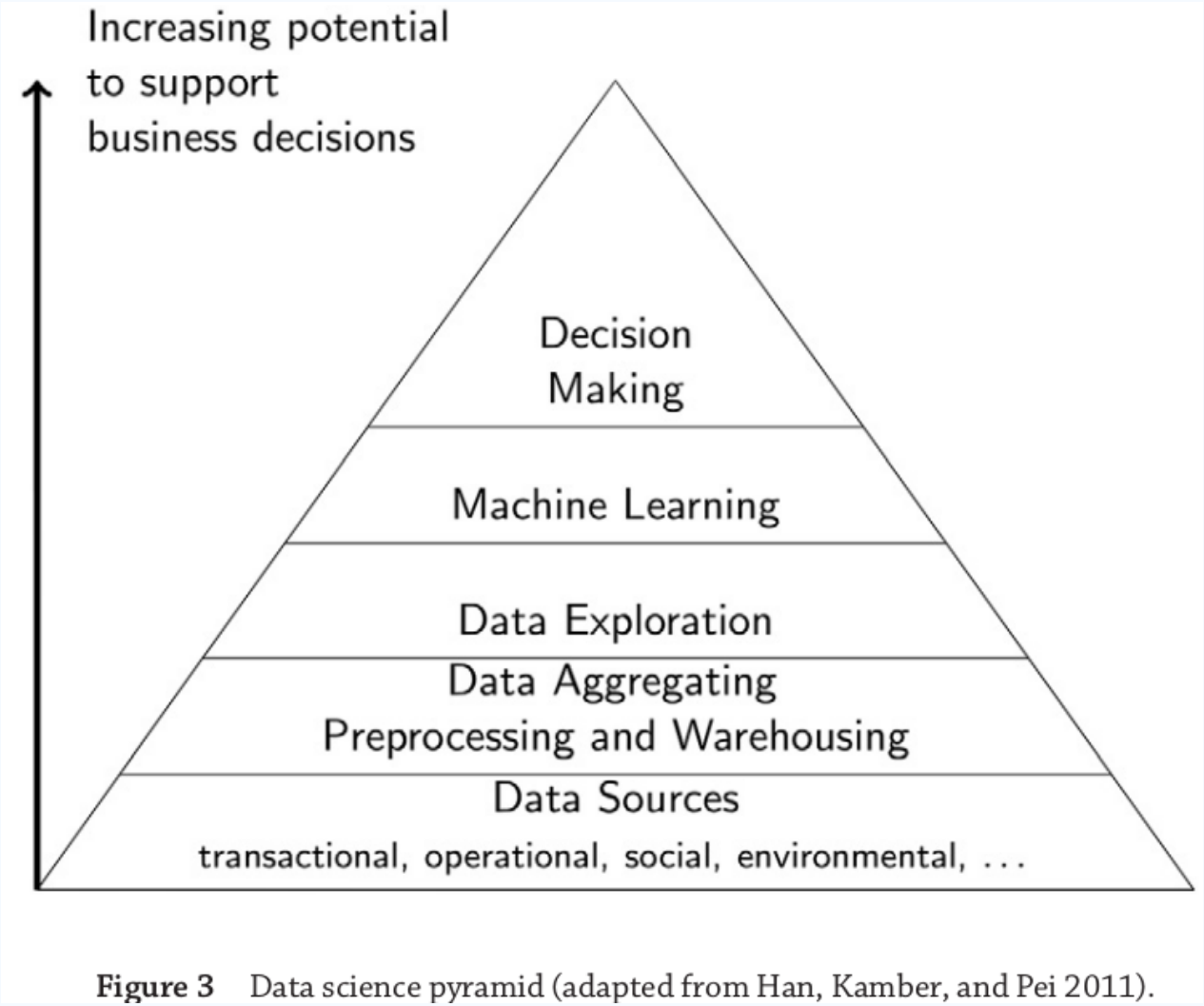


Figure 3 Data science pyramid (adapted from Han, Kamber, and Pei 2011).

Figure 4: Kelleher, J. D., & Tierney, B. (2018). Data science. MIT Press.

Referencias

Murphy, K. P. (2012). *Machine learning: a probabilistic perspective*. MIT press.

Russell, S. J., & Norvig, P. (2010). *Artificial intelligence a modern approach*. London.

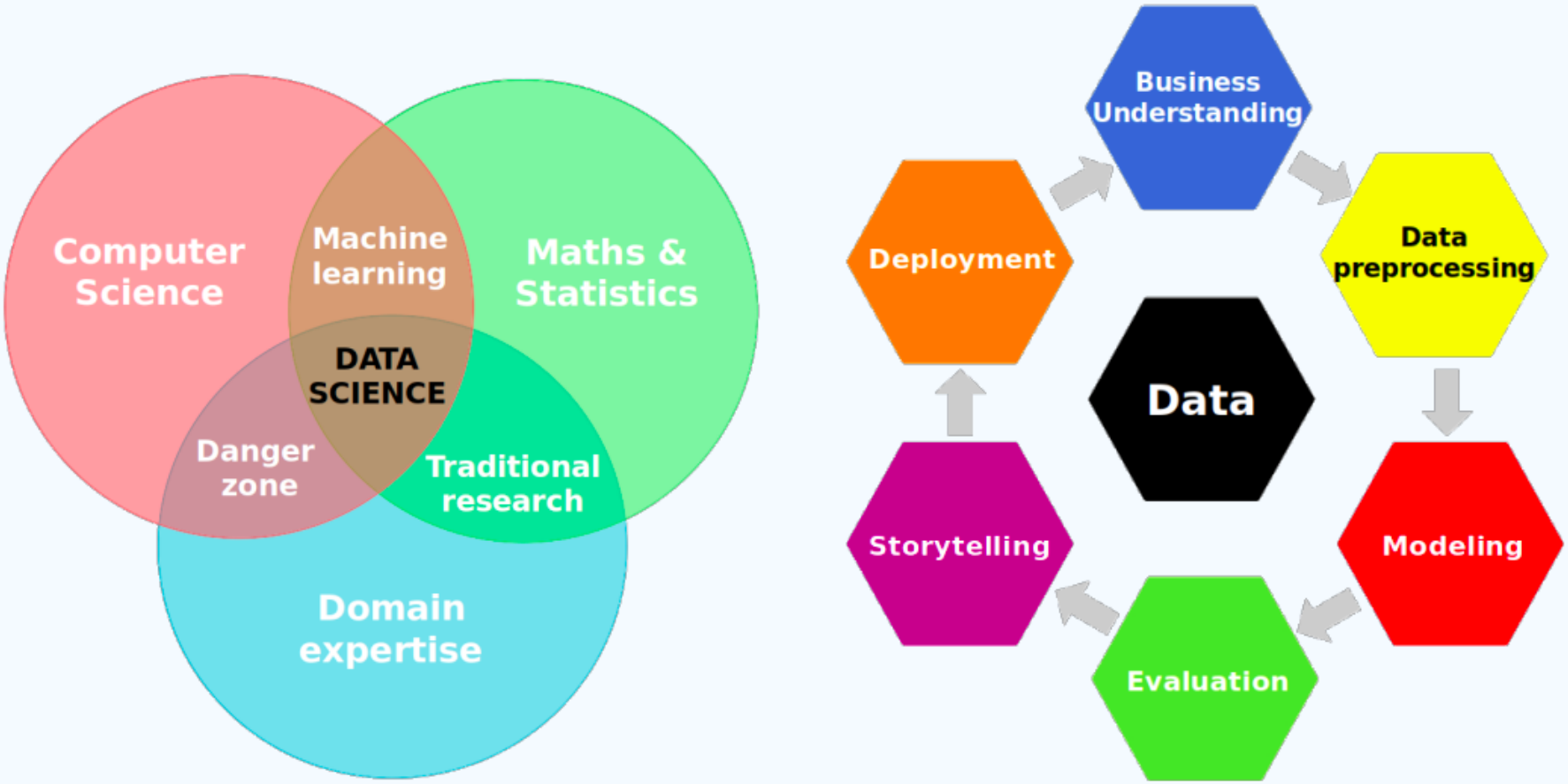
Kelleher, J. D., & Tierney, B. (2018). *Data science*. MIT Press.

Datos

Aprendizaje Automático 1 - Grado en Ciencia e Ingeniería de Datos

Curso académico 2023-2024

Datos en la Ciencia de Datos



(a) Foundations (b) Applications

Figure 1.1: Data Science

Tipos de datos

- Según la **estructuras**: Datos estructurados vs no estructurados
- Según el **comportamiento en el tiempo**: Datos estáticos vs datos dinámicos

Datos estructurados vs no estructurados

- Datos **estructurados**: poseen longitud, tipo, formato y tamaño definidos. Se organizan en formatos de bases de datos, por ejemplo, tablas.

	A	B	C	D	E	F	G	H	I
1	rowid	species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex	year
2	1	Adelie	Torgersen	39.1	18.7	181	3750	male	2007
3	2	Adelie	Torgersen	39.5	17.4	186	3800	female	2007
4	3	Adelie	Torgersen	40.3	18	195	3250	female	2007
5	4	Adelie	Torgersen	NA	NA	NA	NA	NA	2007
6	5	Adelie	Torgersen	36.7	19.3	193	3450	female	2007
7	6	Adelie	Torgersen	39.3	20.6	190	3650	male	2007
8	7	Adelie	Torgersen	38.9	17.8	181	3625	female	2007
9	8	Adelie	Torgersen	39.2	19.6	195	4675	male	2007
10	9	Adelie	Torgersen	34.1	18.1	193	3475	NA	2007
11	10	Adelie	Torgersen	42	20.2	190	4250	NA	2007
12	11	Adelie	Torgersen	37.8	17.1	186	3300	NA	2007

- Datos **no estructurados**: Carecen de formato específico. Documentos de texto, vídeo, datos de redes sociales, correos electrónicos, etc. Se almacenan en su formato original y requieren un procesamiento para ser analizados.

Datos estáticos vs dinámicos

- **Estáticos**: no varían a lo largo del tiempo. Ejemplo: censo, datos de natalidad.
- **Dinámicos**: evolucionan con el tiempo. Ejemplo: base de datos de una tienda con productos y precios

Obtención de datos

Recopilación de información en un dominio específico.

Obtención datos --> Procesamiento de datos

Algunas técnicas de obtención de datos:

- Encuestas y entrevistas
- Toma de muestras
- Web scraping
- Sensores y dispositivos IoT
- etc

Fuentes de datos (¡para practicar!)

- UCI Machine Learning repository
- OpenML
- Kaggle
- KEEL dataset repository (Artículo de referencia)
- Penn Machine Learning Benchmarks (Artículo de referencia)
- Eurostat
- Datos abiertos del Gobierno de España

Datos en R

- R incluye en sus librerías distintos conjuntos de datos
- Librería "datasets" contiene bastantes. Para ver la lista completa, basta ejecutar `library(help = "datasets")`

Datos en R

```r
# install.packages('palmerpenguins')
library(palmerpenguins)

str(penguins)
```

```
tibble [344 x 8] (S3: tbl_df/tbl/data.frame)
 $ species          : Factor w/ 3 levels "Adelie","Chinstrap",..: 1 1 1 1 1 1 1 1 1
 $ island           : Factor w/ 3 levels "Biscoe","Dream",..: 3 3 3 3 3 3 3 3 3 3 .
 $ bill_length_mm   : num [1:344] 39.1 39.5 40.3 NA 36.7 39.3 38.9 39.2 34.1 42 ...
 $ bill_depth_mm    : num [1:344] 18.7 17.4 18 NA 19.3 20.6 17.8 19.6 18.1 20.2 ...
 $ flipper_length_mm: int [1:344] 181 186 195 NA 193 190 181 195 193 190 ...
 $ body_mass_g      : int [1:344] 3750 3800 3250 NA 3450 3650 3625 4675 3475 4250 .
 $ sex              : Factor w/ 2 levels "female","male": 2 1 1 NA 1 2 1 2 NA NA ..
 $ year             : int [1:344] 2007 2007 2007 2007 2007 2007 2007 2007 2007 2007
```

¿Con qué tipo de datos vamos a trabajar?

- Datos tabulares: $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$
- Observaciones (filas): items, instancias, puntos, elementos, objetos, etc. $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ij})$
- Variables (columnas): atributos, características (del inglés *features*) $\mathbf{f}_j = \mathbf{x}_j = (x_{j1}, \ldots, x_{jn})$

Primer paso: **Entender los datos**

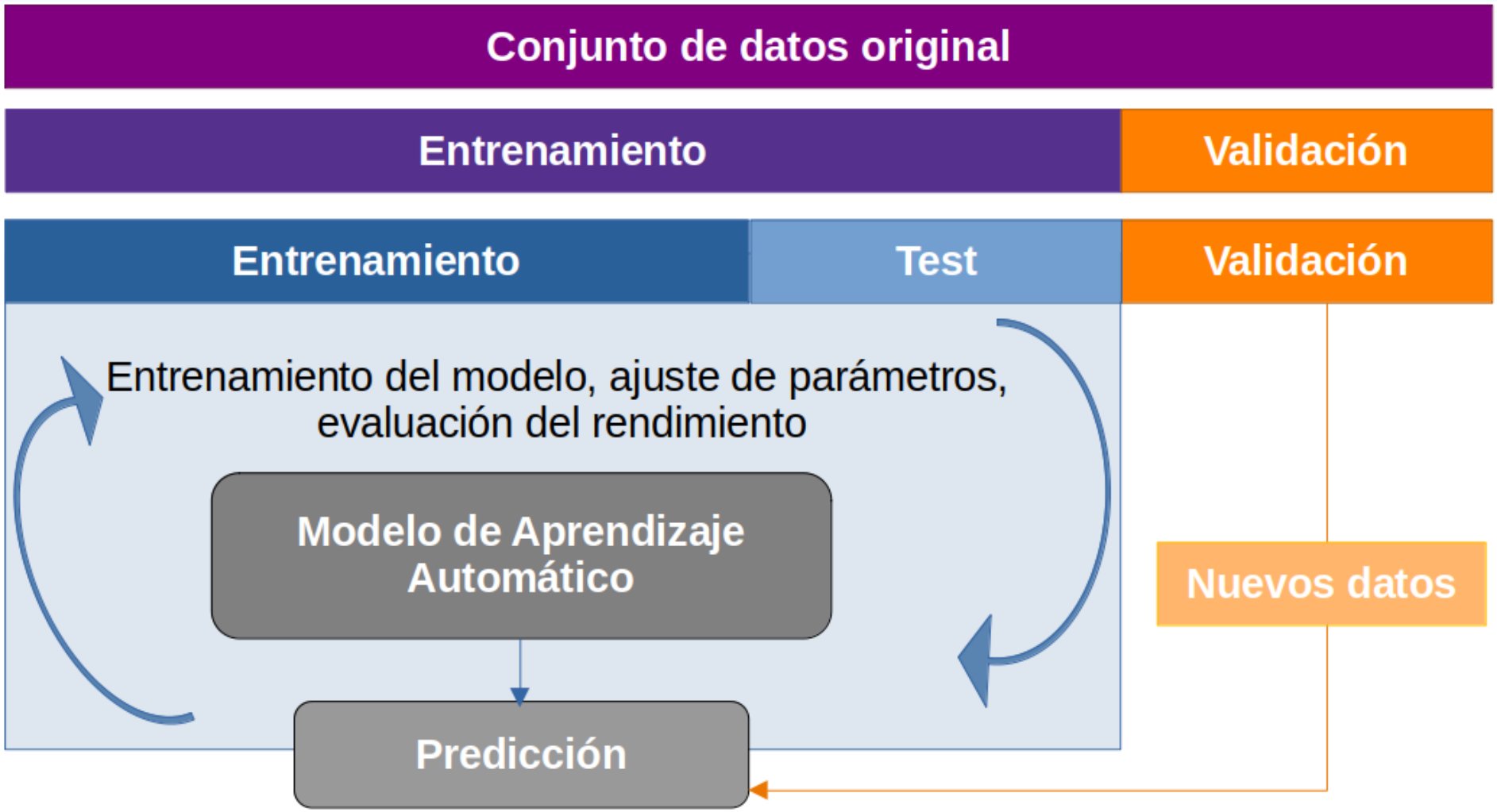
- ¿Cuál es la dimensión de los datos? ¿Cuál es el número de filas (instancias) y de columnas (variables)?
- ¿Qué significan las variables?
- ¿Hay datos erróneos?
- ¿Hay datos faltantes?

¡Practiquemos un poco más con estos datos en R!

Particiones de los datos

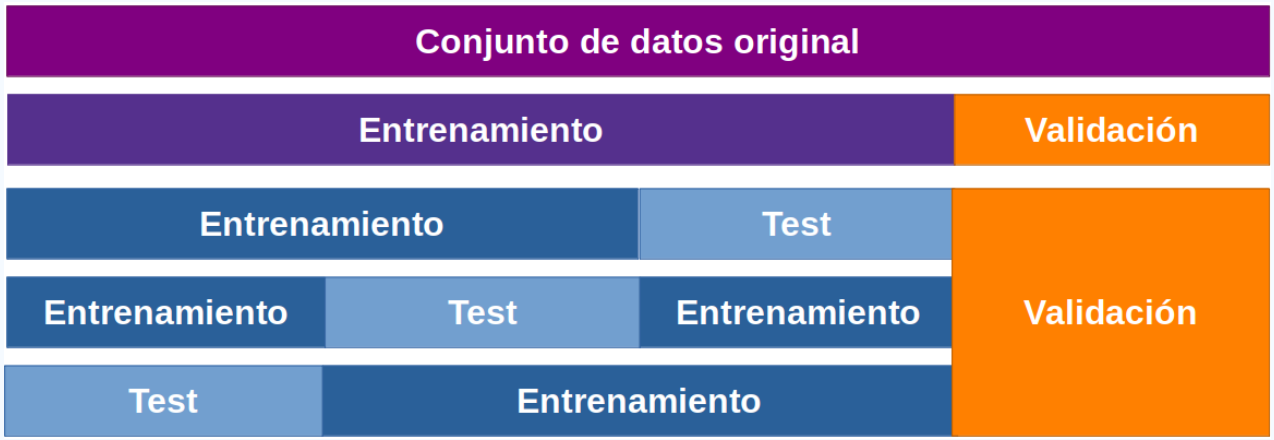
- **Entrenamiento (Training)**: Muestra para entrenar el modelo, el modelo aprenderá el comportamiento de los datos con esta muestra
- **Test**: Para probar el modelo entrenado y comparar el rendimiento en entrenamiento y test. En base a los resultados, se puede cambiar de modelo o realizar ajustes sobre él (reentrenar el modelo)
- **Validación (Validation)**: Para reflejar el comportamiento del modelo en un entorno real con nuevos datos. ¡No se usa para reentrenar!

Particiones de los datos



Particiones de los datos

- Construcción de las particiones: Train 60% - Test 20% - Validación 20% (aproximadamente)
- Los % anteriores dependerán del volumen de los datos y los objetivos del problema
- k-fold cross validation. Se obtienen k valores del error –> media y desviación



Particiones de los datos: Muestreo

- ¿Por qué funcionan bien las particiones?
- Muestreo aleatorio
- Muestreo estratificado –> guiado por la variable objetivo

Bases de datos

- **Relacionales**. Siguen el modelo entidad-relación, también llamado modelo relacional, en donde cada una de las tablas (o entidades) presenta algún tipo de enlace con otras (relaciones).
  - SQL: Structured Query Language
- **No relacionales** (no SQL). Representar datos de forma más flexible

Infraestructuras para datos

- Bases de datos relacionales y no relaciones
- Almacenamiento de datos en la nube
- Almacenamiento en memoria
- Almacenamiento distribuido
  - Federated learning

Calidad de los datos

- Clave en cualquier proyecto que involucre datos –> influye directamente en la confiabilidad y el valor de los resultados
- Precisión
- Integridad
- Consistencia
- Relevancia
- Actualización
- Limpieza
- Documentación

Ética, privacidad y seguridad en los datos

- La ética, privacidad y seguridad en los datos son aspectos entrelazados y fundamentales para garantizar que la recopilación, el análisis y el uso de datos se realicen de manera responsable y en beneficio de la sociedad
- ¿Algún ejemplo de falta de ética?
- ¿Algún ejemplo de falta de privacidad?

https://www.unesco.org/en/artificial-intelligence/recommendation-ethics/cases

Referencias

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction* (Vol. 2, pp. 1-758). New York: springer.

Saltz, J., Skirpan, M., Fiesler, C., Gorelick, M., Yeh, T., Heckman, R., ... & Beard, N. (2019). Integrating ethics within machine learning courses. *ACM Transactions on Computing Education (TOCE)*, *19*(4), 1-26.

Análisis exploratorio de datos

Aprendizaje Automático 1 - Grado en Ciencia e Ingeniería de Datos

Curso académico 2023-2024

Exploratory Data Analysis (EDA)

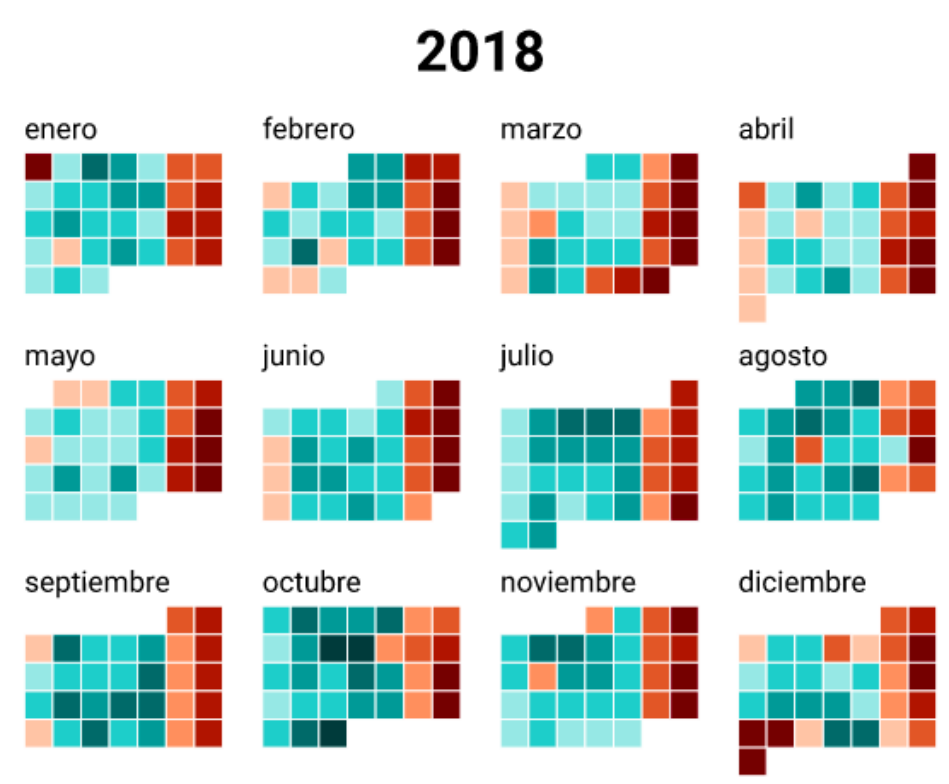
"Exploratory data analysis can never be the whole story, but nothing else can serve as the foundation stone as the first step"

"Exploratory Data Analysis is detective work"

John Tukey

Importancia del EDA

Estudio natalidad: https://www.eldiario.es/nidos/no-ninos-nacen-toca-dar-luz-semana-21-probable-hacerlo-lunes-viernes_1_6400307.html



Introducción

- Primer análisis sobre cualquier conjunto de datos
- ¡Entender los datos!
- El análisis exploratorio de datos es un conjunto de técnicas que permiten resumir las características más importantes de un conjunto de datos, normalmente con especial énfasis en el uso de métodos de visualización gráfica
- No hay un guión estricto para realizar un EDA (¡somos detectives!)
- Fundamental adquirir conocimiento de los datos antes de usar un modelo de Aprendizaje Automático

Preguntas

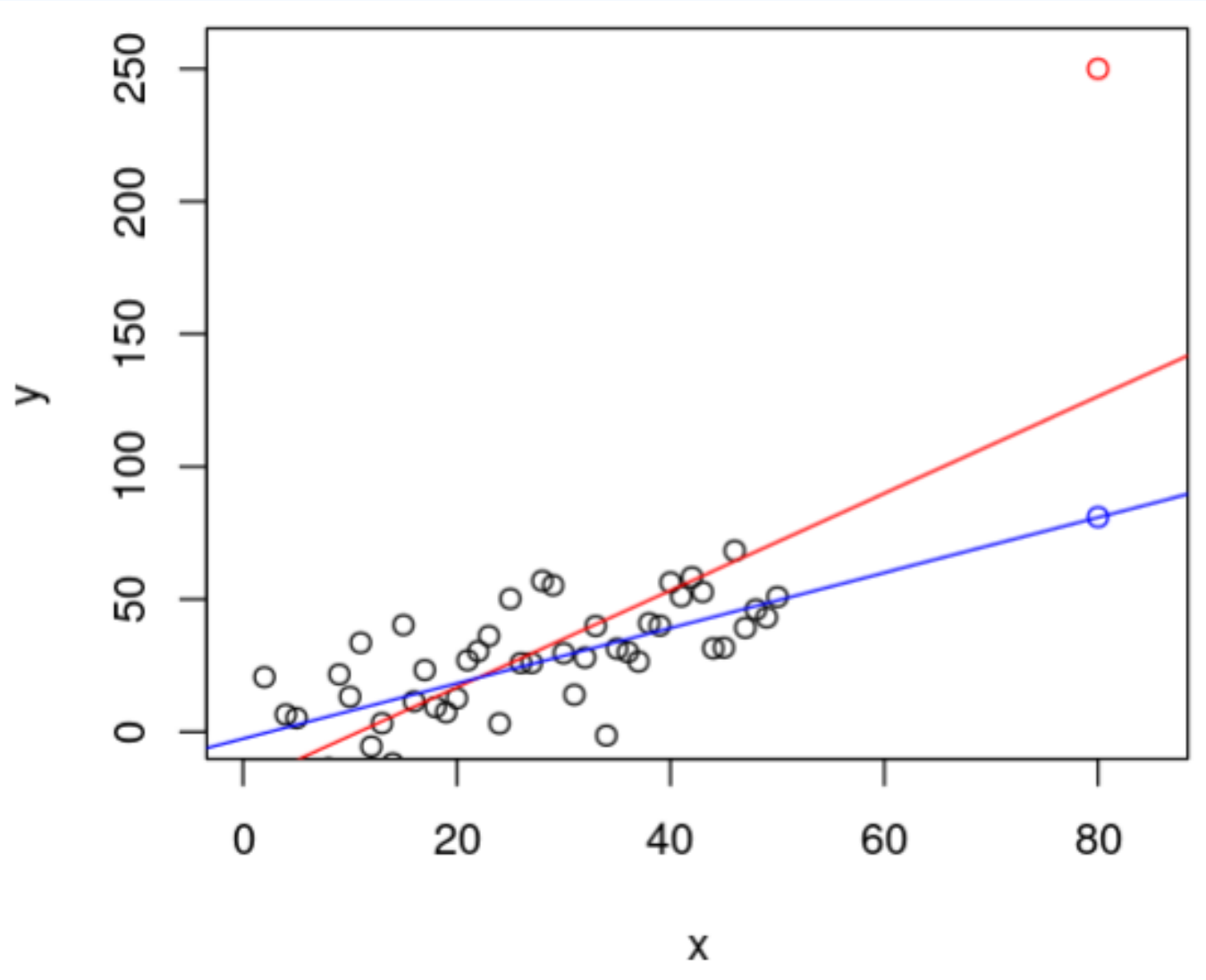
- Objetivo EDA: comprensión profunda de los datos
- ¿Cómo hacerlo? Planteando preguntas:
  - ¿Tamaño de los datos?
  - ¿Tipos de variables?
  - ¿Hay variable objetivo? ¿Cómo es?
  - ¿Hay errores?
  - ¿Hay variables irrelevantes?
  - ¿Están las variables relacionadas?
  - ¿Hay atípicos?
  - ¿Tengo suficiente capacidad de cómputo para procesar los datos?

Datos atípicos (outliers)

- Datos que no están en consonancia con el resto, que destaca por ser distinto del resto
- Causas:
  - Errores de medición (humano o del sistema). Ej: Peso de un paciente: 800 kg o un medidor manipulado
  - Contaminación: la muestra contiene datos de una población distinta a la de interés
  - Desviaciones naturales
- Solo se modifica el dato si es un error. Se busca el valor real y, si no es posible, se pone como missing

Datos atípicos

¿Diferencia entre el valor atípico rojo y el azul?



Datos faltantes (missing)

- Dato vacío, dato perdido, NA
- Causas:
  - Error en la medición, la transcripción
  - No se puede lograr el dato
- Acciones:
  - Trabajar únicamente con los datos sin valores faltantes (representan un % bajo del total de los datos)
  - Imputación de missing (media o mediana de la variable, el valor de los puntos más similares, predicción de un modelo de ML)
  - Agrupar los missings en una nueva categoría ¡fácilmente distinguible! Ej: 9999

Tipos de variables

- **Cualitativa** (también llamada categórica): refleja una cualidad de la realidad, su valor no se representa con un número. Pueden ser:
  - Dicotómicas (Ej: Sí o no) o politómicas (Ej: grupo sanguíneo)
  - Nominales (Ej: color de ojos) u ordinales (Ej: nota de un examen Suspenso-Aprobado-Notable-Sobresaliente)
- **Cuantitativa**: su valor se indica con un número, se corresponde con características que representan cantidades. Ej: distancia en km, nivel de colesterol, temperatura, etc. Pueden ser:
  - Discretas: Toma un número finito o infinito numerable de valores. Ej: números naturales o un recuento
  - Continuas: Puede tomar infinitos valores. Ej: altura, peso

Resumen numérico de variables cualitativas

- **Moda**: valor más frecuente de la distribución
- **Tabla de frecuencias** o **tabla de contingencia**: Muestra, para cada valor que tome una variable categórica, o para cada combinación de valores de dos o más variables categóricas, el número de casos que aparecen con dicho valor o combinación de valores

```
table(penguins$species)
```

```
   Adelie Chinstrap    Gentoo
      152        68       124
```

```
prop.table(table(penguins$species))
```

```
   Adelie Chinstrap    Gentoo
0.4418605 0.1976744 0.3604651
```

Medidas de centralidad

- **Media**. Dada la variable $\mathbf{x} = (x_1, \ldots, x_n)$ medida en n observaciones, su media es

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

  Muy afectada por valores atípicos/extremos

- **Mediana**: valor que ocupa la posición central de los datos, i.e., deja el 50% de los puntos a su izquierda (por debajo de él) y el otro 50% a la derecha (por encima de él). Sea $\mathbf{x}$ una variable con n observaciones, ordenados de menor a mayor, entonces:
  - Si n es impar, la mediana es justamente el valor que ocupa justamente la posición central $\lfloor n/2 \rfloor + 1$, $Med(\mathbf{x}) = x_{(\lfloor n/2 \rfloor + 1)}$
  - Si n es par, la mediana será la media de los dos valores centrales, esto es, $Med(\mathbf{x}) = \frac{x_{(n/2)} + x_{(n/2+1)}}{2}$

Resumen numérico de variables cuantitativas

Medidas de posición

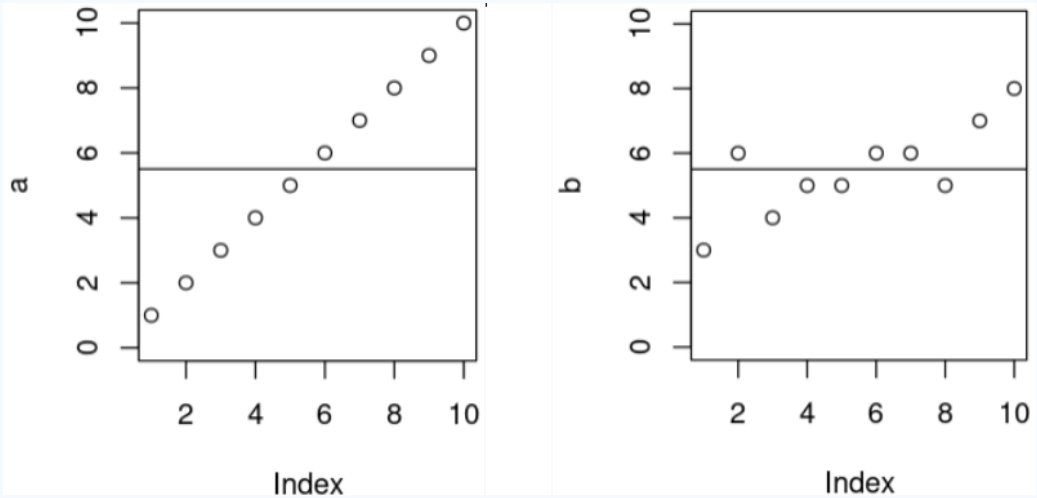
- Valores **mínimo** y **máximo** de la variable: x_{min}, x_{max}
- Primer y tercer **cuartil**: Los valores que dejan por debajo un $p\%$ de los datos, siendo $p = 25\%$ en el caso del primer cuartil (Q_1) y $p = 75\%$ en el caso del tercer cuartil (Q_3). El segundo cuartil es la mediana.
- **Deciles**: Mismo concepto que los cuartiles pero de 10 en 10

Resumen numérico de variables cuantitativas

Medidas de dispersión: ¿Cómo varían los datos en torno a los valores centrales?

```
a <- c(1, 2, 3, 4, 5, 6, 7, 8 , 9, 10)
b <- c(3, 6, 4, 5, 5, 6, 6, 5, 7, 8)
cat("Media a: ", mean(a), ', Desviación típica a: ', round(sd(a),2), '\n',
             "Media b: ", mean(b), ', Desviación típica b: ', round(sd(b),2))
```

```
Media a:  5.5 , Desviación típica a:  3.03
 Media b:  5.5 , Desviación típica b:  1.43
```



Medidas de dispersión

- **Rango** o **recorrido**: $Rango = x_{max} - x_{min}$
- **Varianza**: Mide la dispersión de los valores de la variable respecto a la media
  - Varianza **muestral**: $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$
  - Varianza **poblacional**: $\sigma^2 = \frac{1}{N}\sum_{i=1}^{n}(x_i - \mu)^2$, siendo N el tamaño de la población y μ su media
- **Desviación típica**: raíz cuadrada de la varianza muestral o poblacional.
  - Desviación típica **muestral**: $s = \sqrt{s^2}$
  - Desviación típica **poblacional**: $\sigma = \sqrt{\sigma^2}$

  Interpretación más sencilla al medir la dispersión en las mismas unidades que la variable

Medidas de dispersión

- **Rango intercuartílico**: diferencia entre el tercer y el primer cuartil $IQR = Q_3 - Q_1$
- **Coeficiente de variación**: representa la desviación típica en unidades de la media $CV = s/\bar{x}$. Se suele expresar en porcentaje. Por ejemplo, CV $= 60\%$ indica que el valor de la desviación típica es 0.6 veces la magnitud de la media.

Diagrama de barras

```
library(ggplot2)
ggplot(penguins, aes(species)) + geom_bar(fill = "orange") +
  theme_bw() +
  labs(title = "Pingüinos por especie",
       x = "Especie",
       y = "Frecuencia absoluta")
```

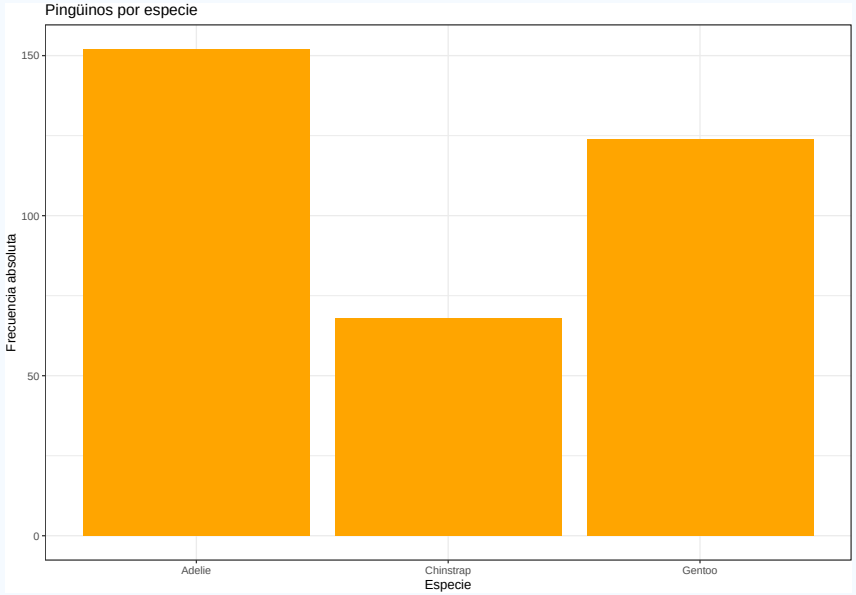
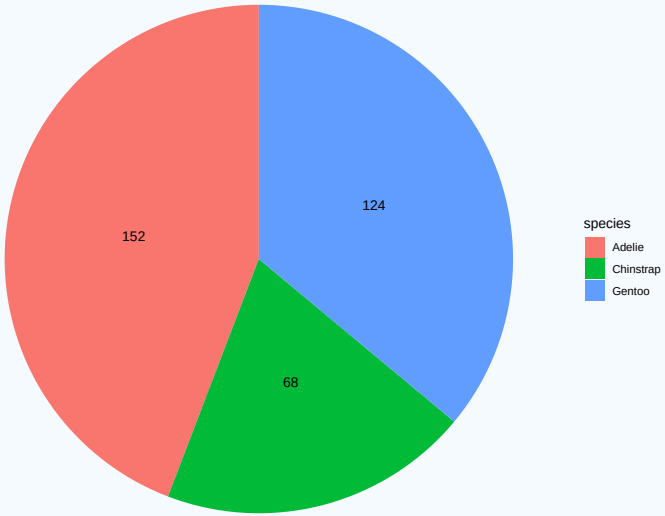


Gráfico de tarta

```
penguins %>% count(species) %>% ggplot(aes(x="", y=n, fill=species)) +
  geom_bar(stat="identity", width=1) +
  coord_polar("y", start=0) +
  geom_text(aes(label = n),
            position = position_stack(vjust = 0.5))+
  theme_void()
```



Problema: ojo humano tiene problemas para percibir correctamente diferencias en sectores angulares,

Gráficos de gofre (*waffle*)

```
library(waffle)
library(extrafont)
iron(
  waffle(c(classic = 49,hybrid=33, electric = 18), rows = 5,  glyph_size = 6,
         colors = c("#A91C68","#53D5D9", "#AFD953"), title = "City A"),
  waffle(c(classic = 69, hybrid = 26, electric = 5), rows = 5, glyph_size = 6,
         colors = c("#A91C68","#53D5D9", "#AFD953"), title = "City B")
)
```

Diagrama de cajas y bigotes (boxplot)

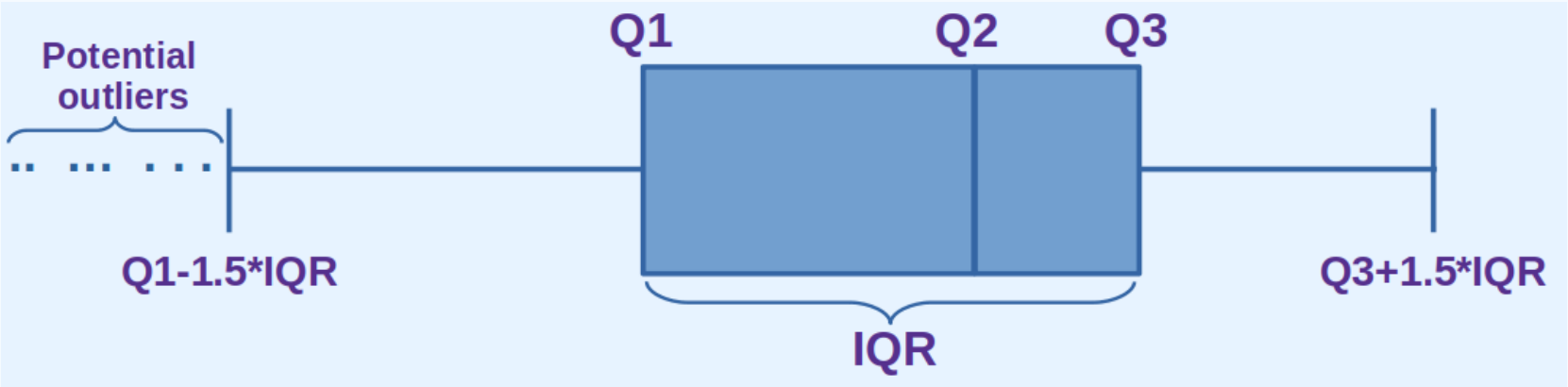
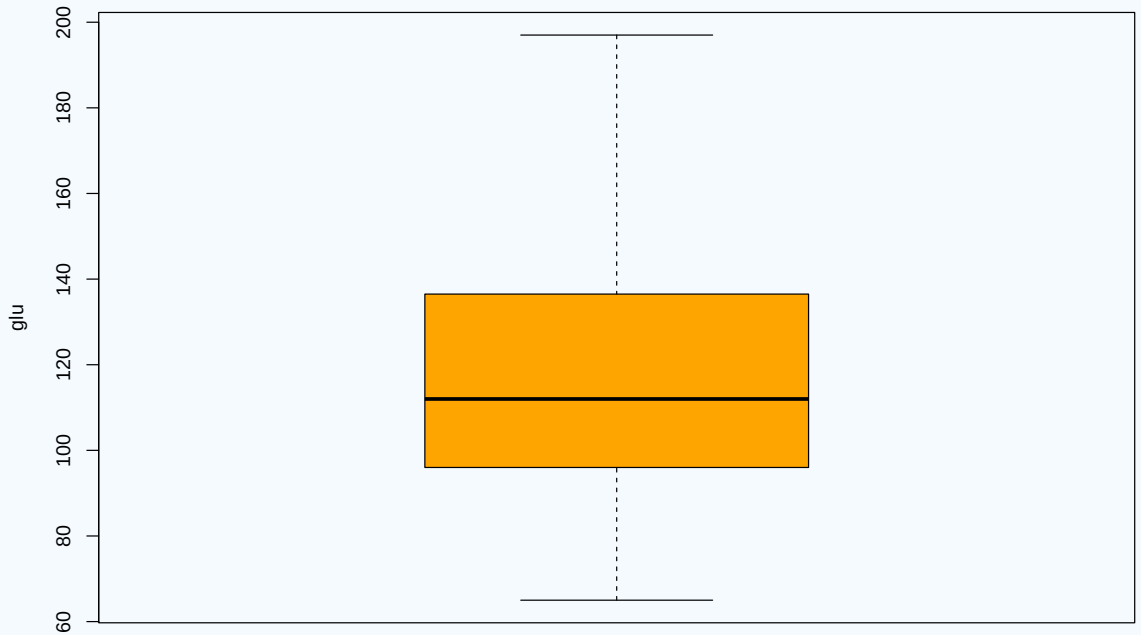


Diagrama de cajas y bigotes (boxplot)

```
library(car)
library(Hmisc)
library(MASS)
Boxplot(~glu, data = Pima.te, col= 'orange')
```



Histograma

```
# Histograma
ggplot(Pima.te, aes(x = glu)) +
geom_histogram(fill="white", colour="black") +
ggtitle('Histograma glucose')
```

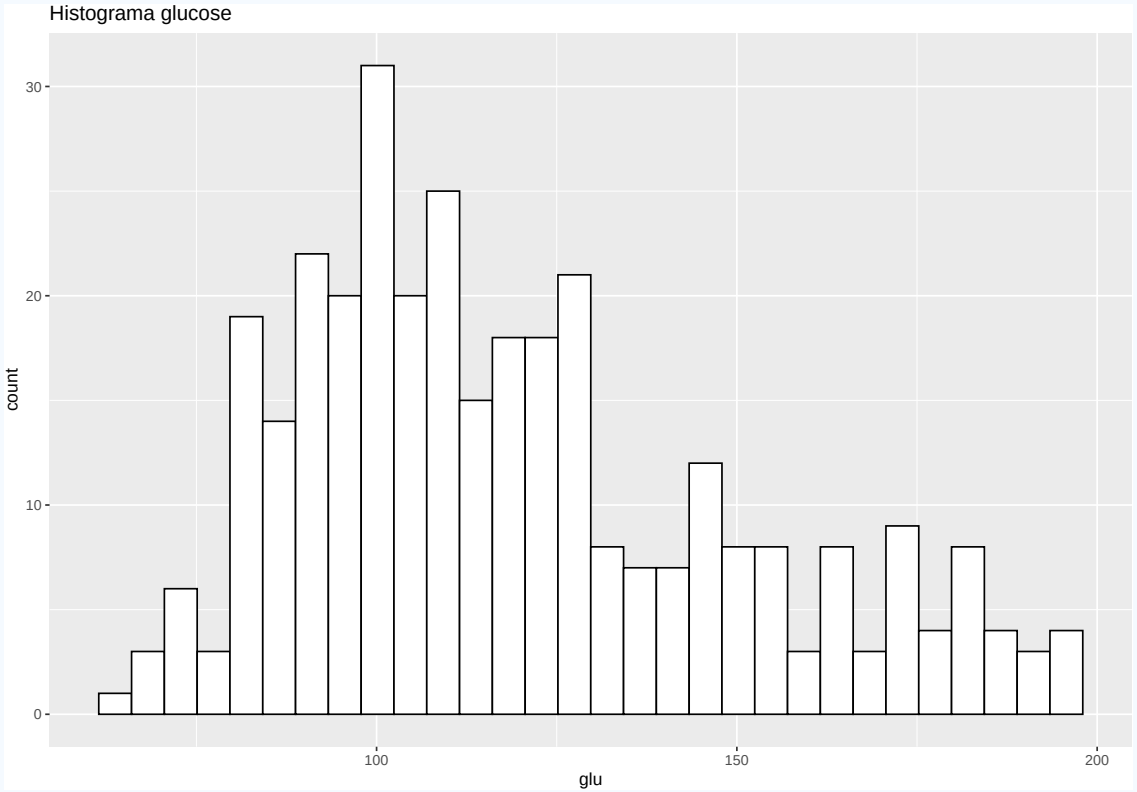
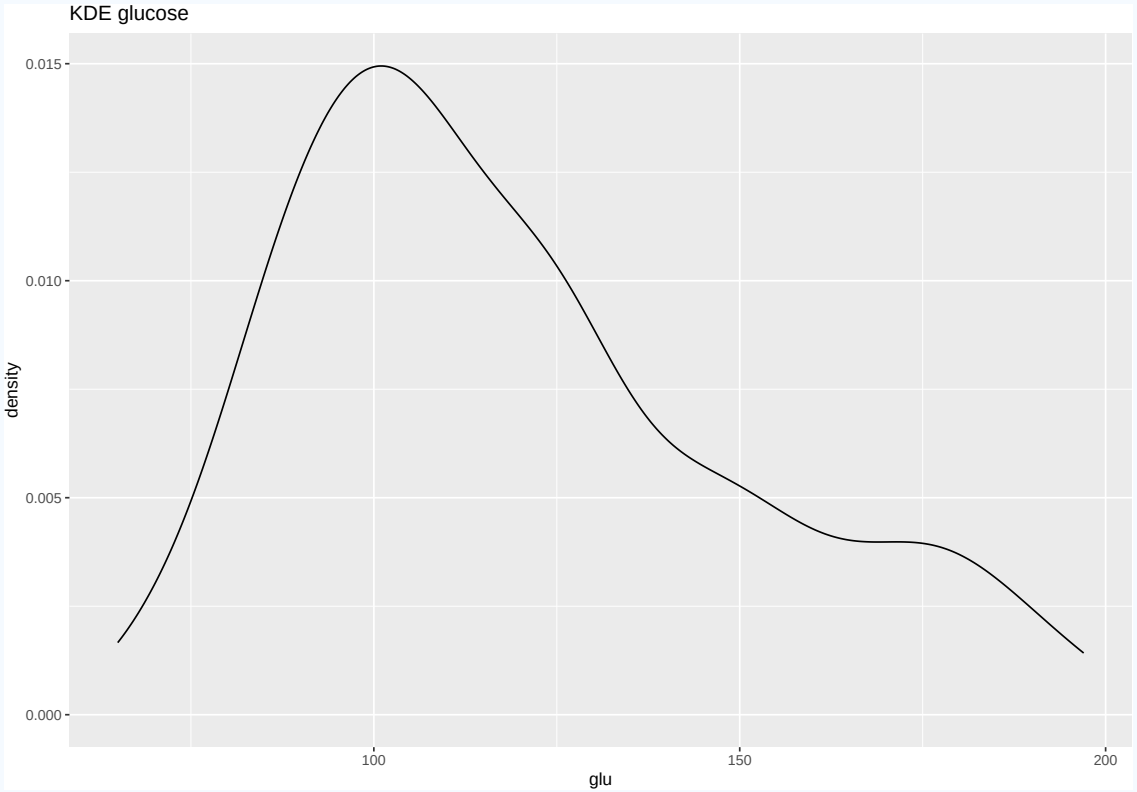


Gráfico de densidad

```
ggplot(Pima.te, aes(x = glu)) +
geom_density() +
ggtitle('KDE glucose')
```



Histograma + gráfico de densidad

```
ggplot(Pima.te, aes(x = glu)) +
 geom_histogram(aes(y = after_stat(density)),
                position = "identity",
                color = "darkblue", fill = "lightblue", bins=20) +
  geom_density(lwd = 1, colour = 1)
```

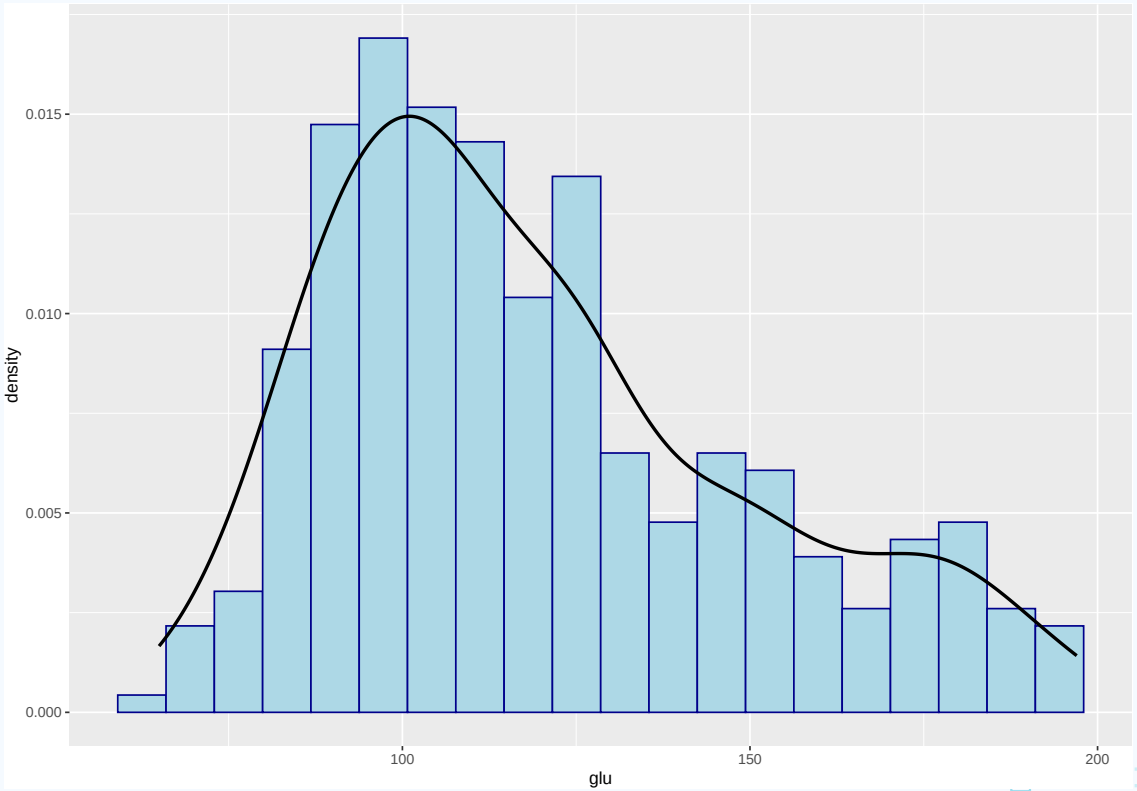


Diagrama de barras

```
library(gridExtra)

# Grouped
barplot_grouped <- ggplot(diamonds, aes(color, fill=cut)) + geom_bar(position="dodg
# Stacked
barplot_stacked <- ggplot(diamonds, aes(color, fill=cut)) + geom_bar(position="stac
grid.arrange(barplot_grouped, barplot_stacked, ncol=2)
```

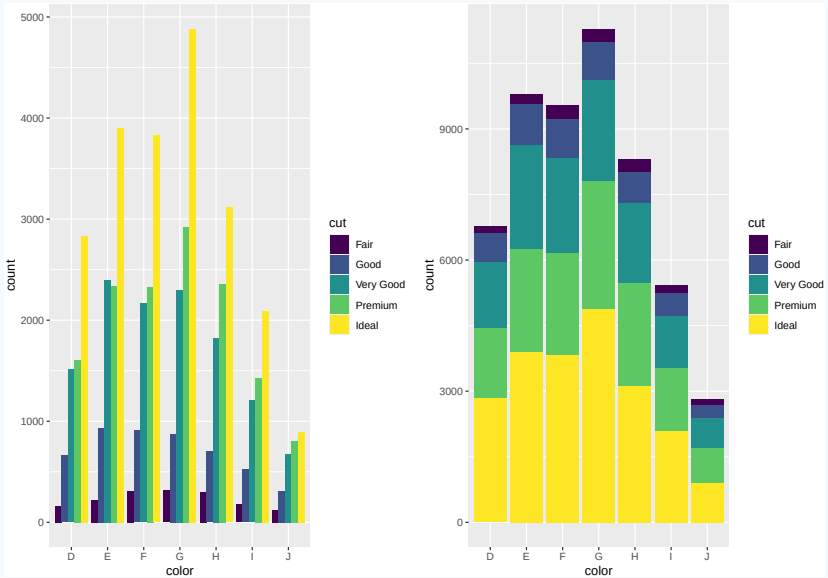
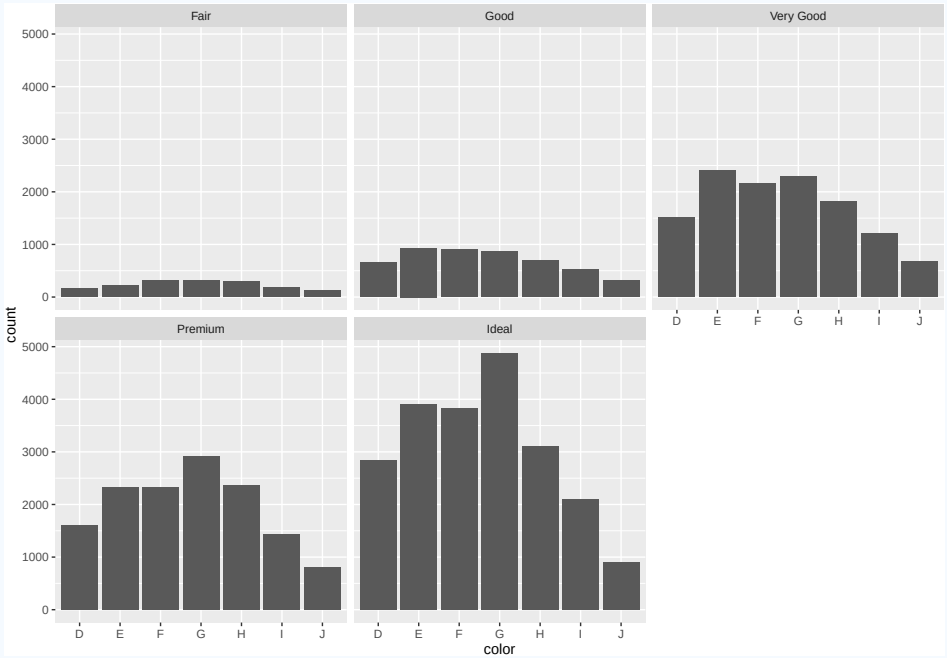
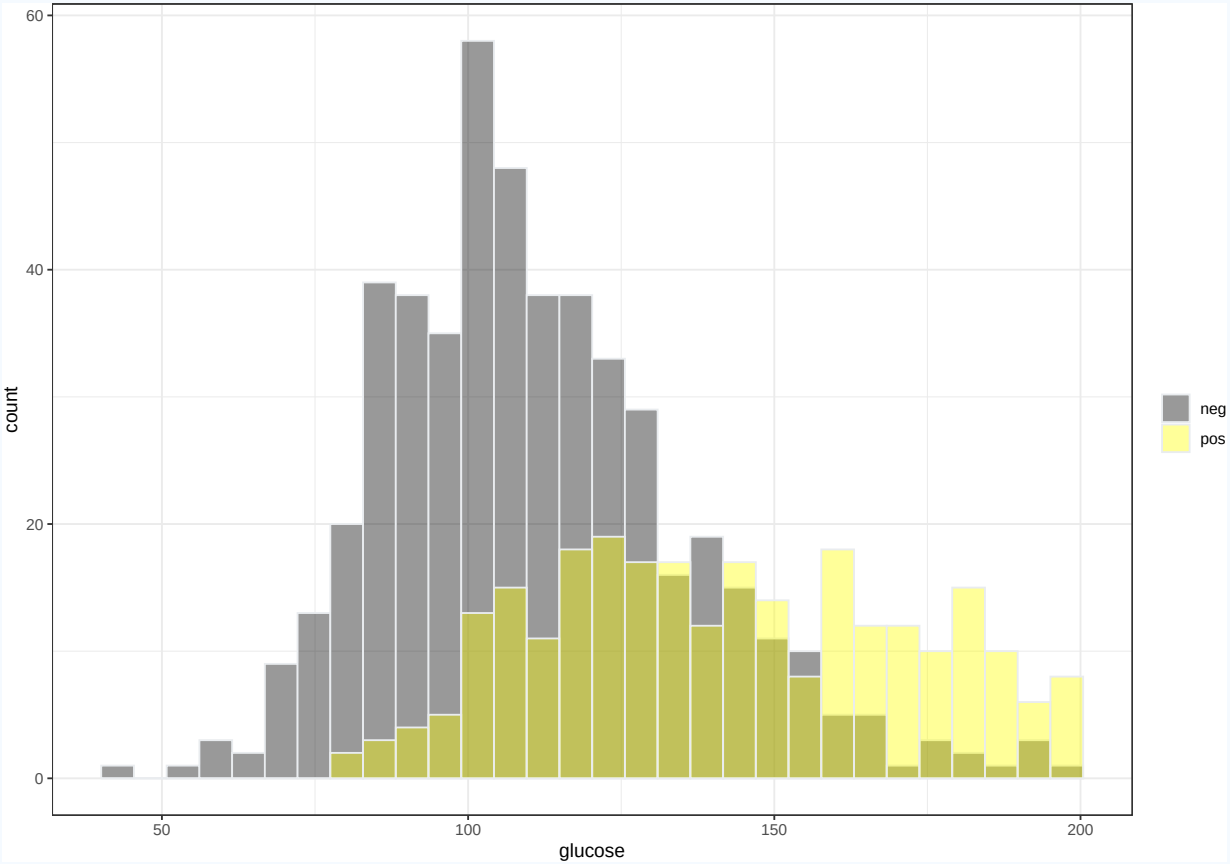


Diagrama de barras

```
# Por paneles
ggplot(diamonds, aes(color)) + geom_bar() +
facet_wrap(~ cut)
```

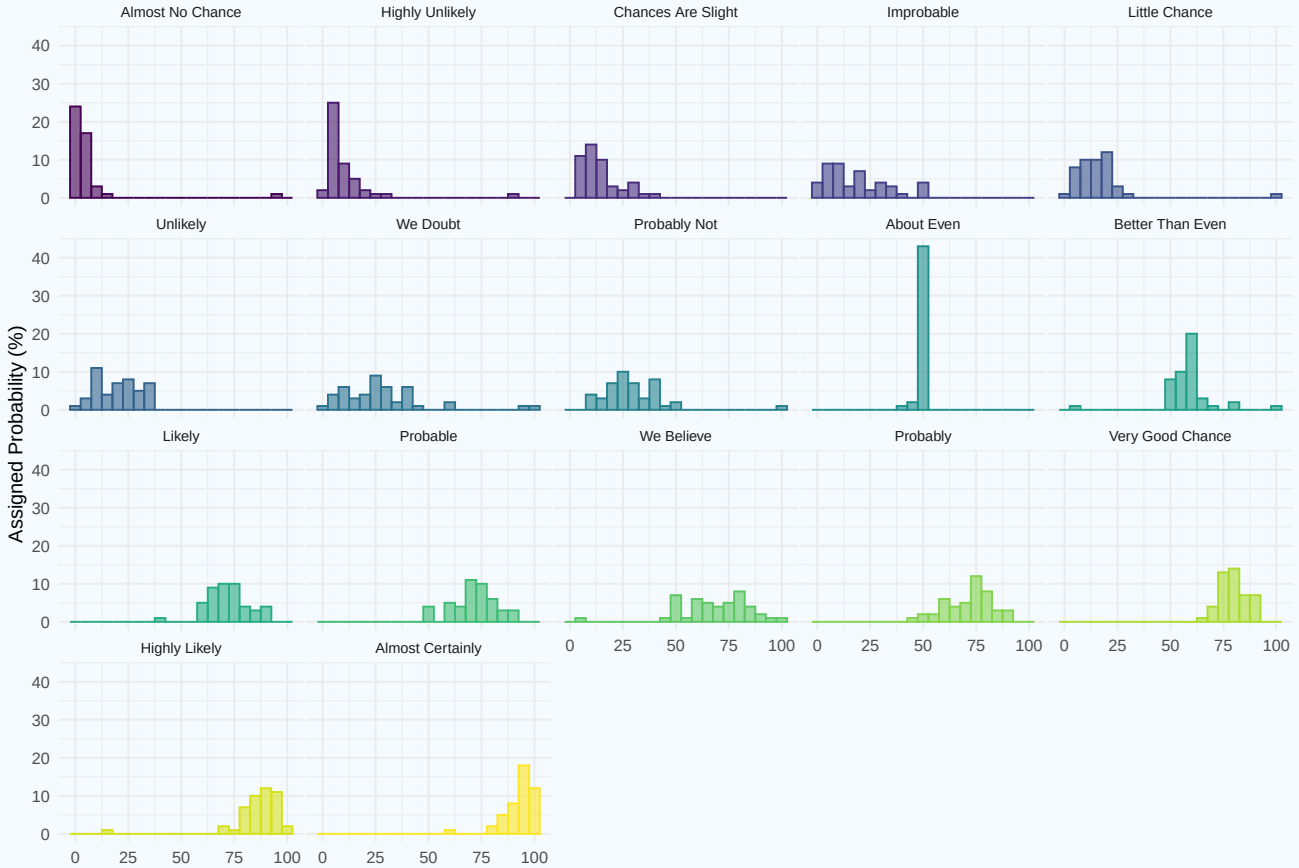


Histogramas conjuntos

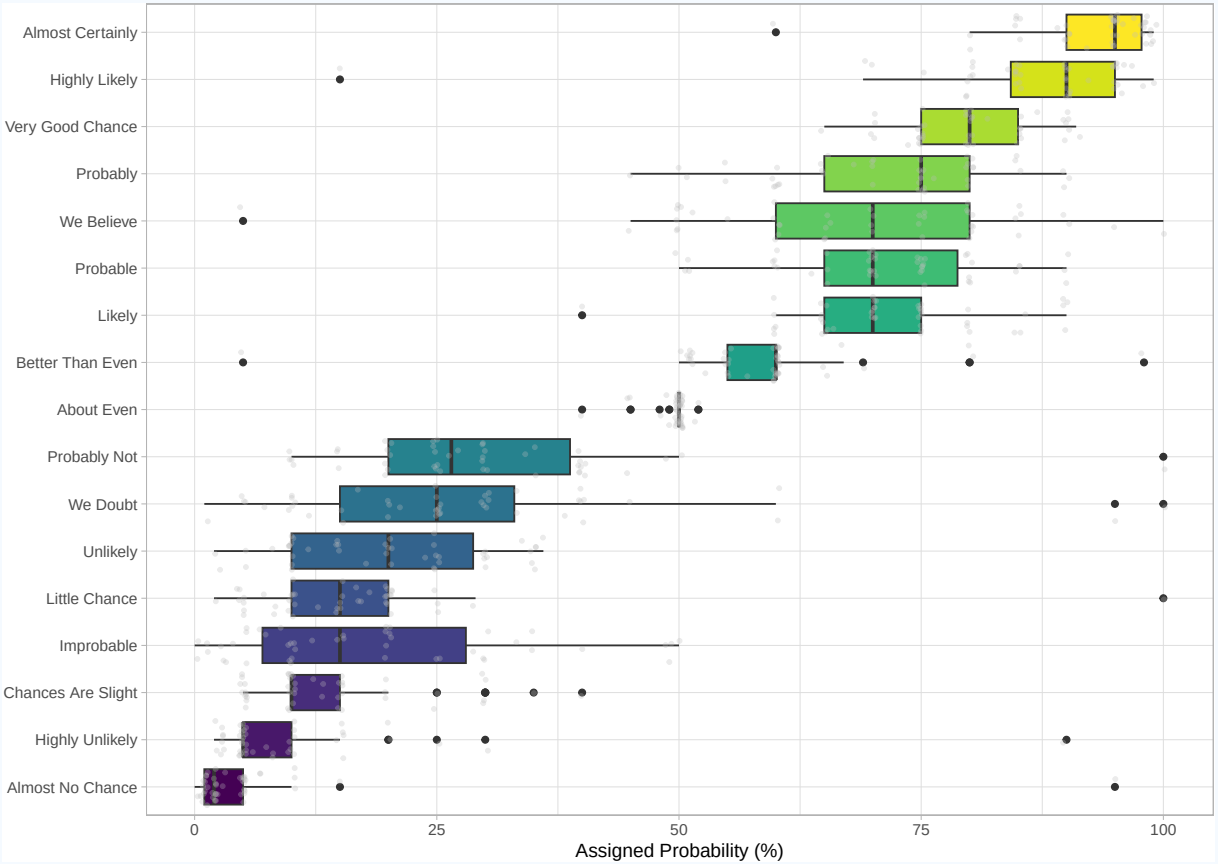


Histogramas conjuntos

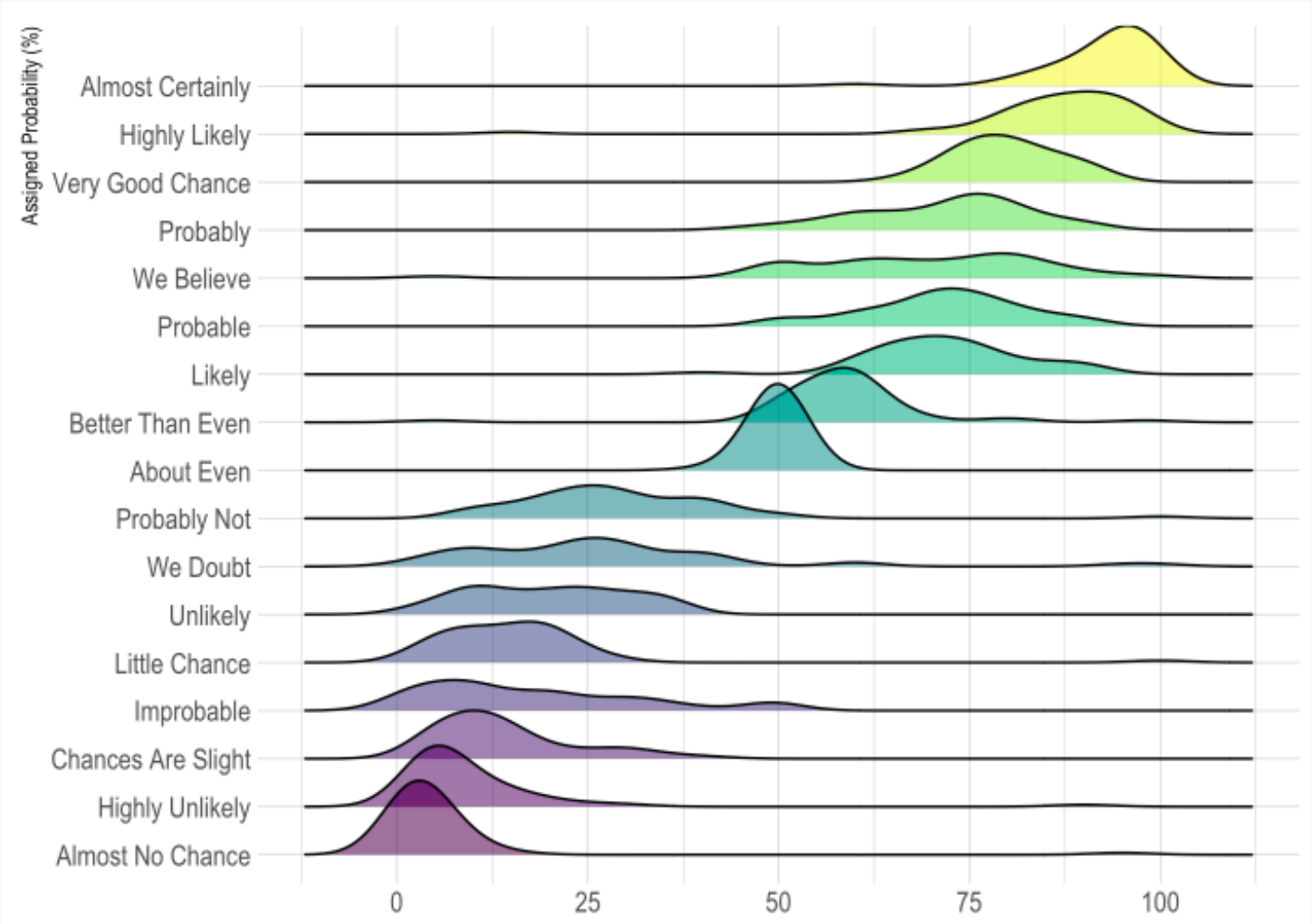
Origen del gráfico: From Data to Viz



Origen del gráfico: From Data to Viz



Origen del gráfico: From Data to Viz



Corrplot

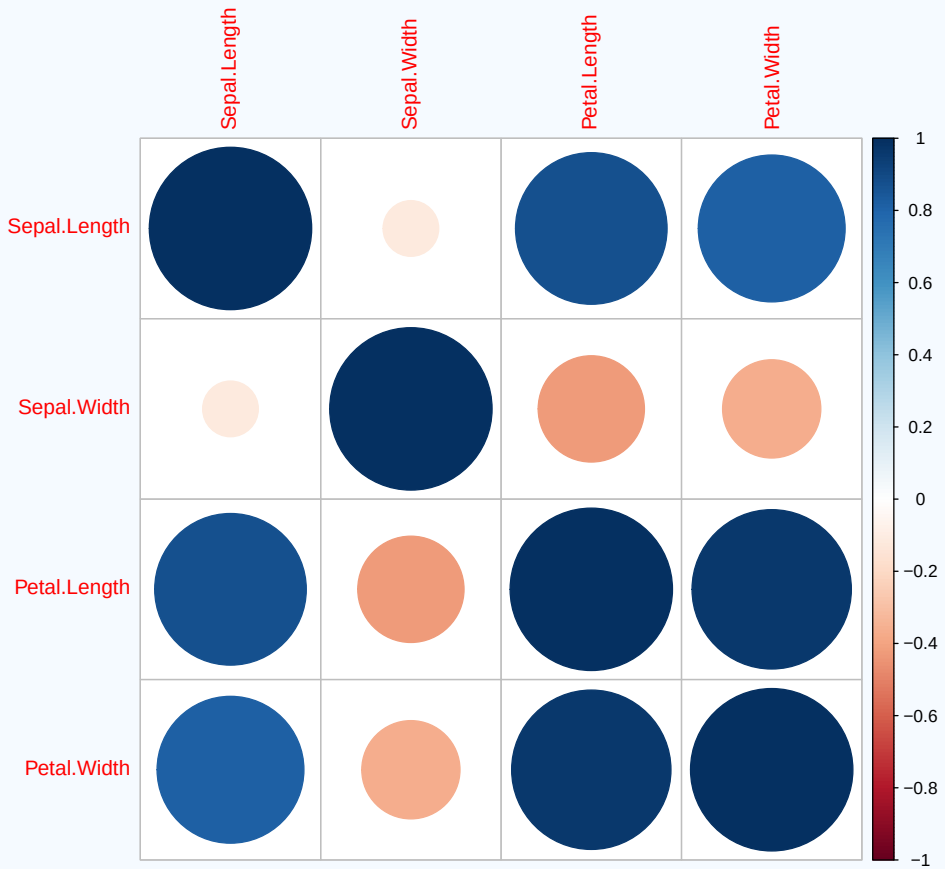

Sepal.Length
Sepal.Width
Petal.Length
Petal.Width
Sepal.Length
Sepal.Width
Petal.Length
Petal.Width
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1

Diagrama de dispersión

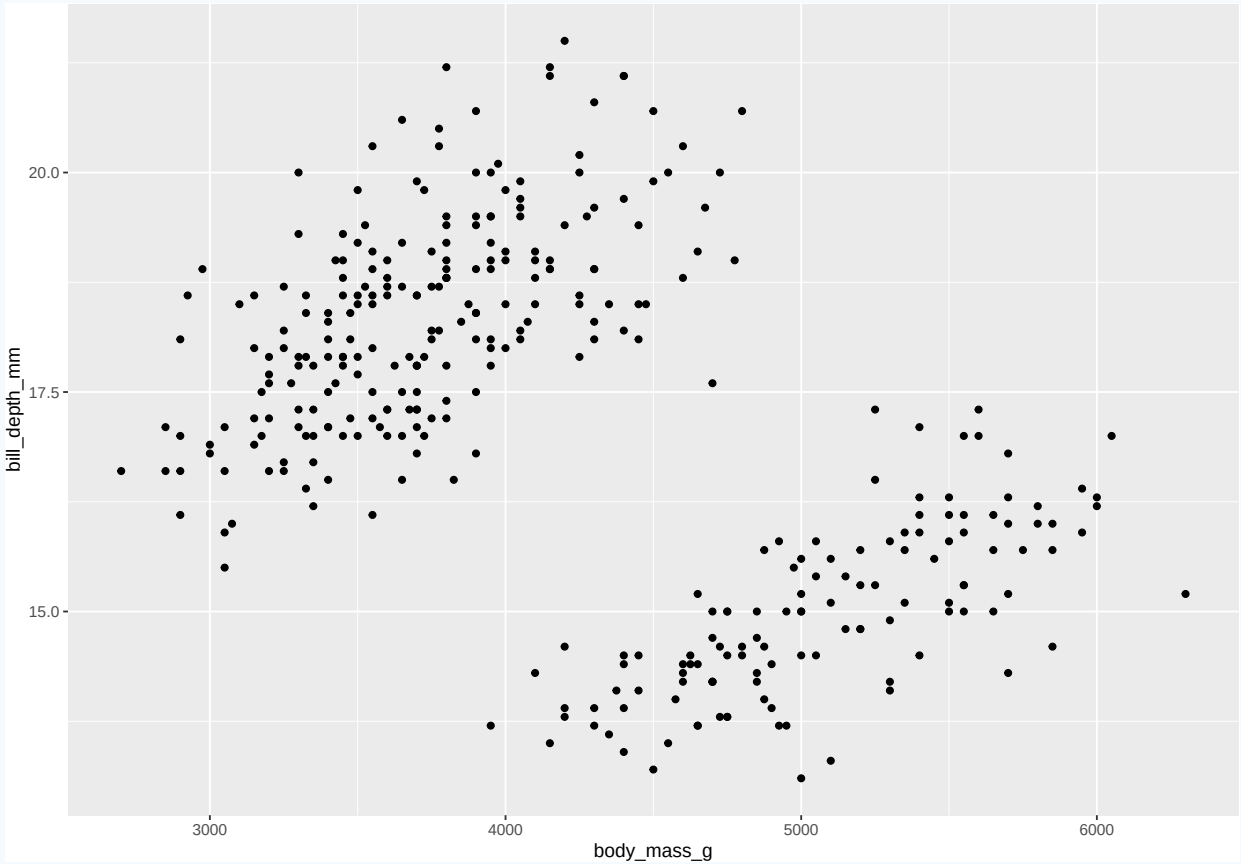


Diagrama de dispersión

Diagrama de dispersión

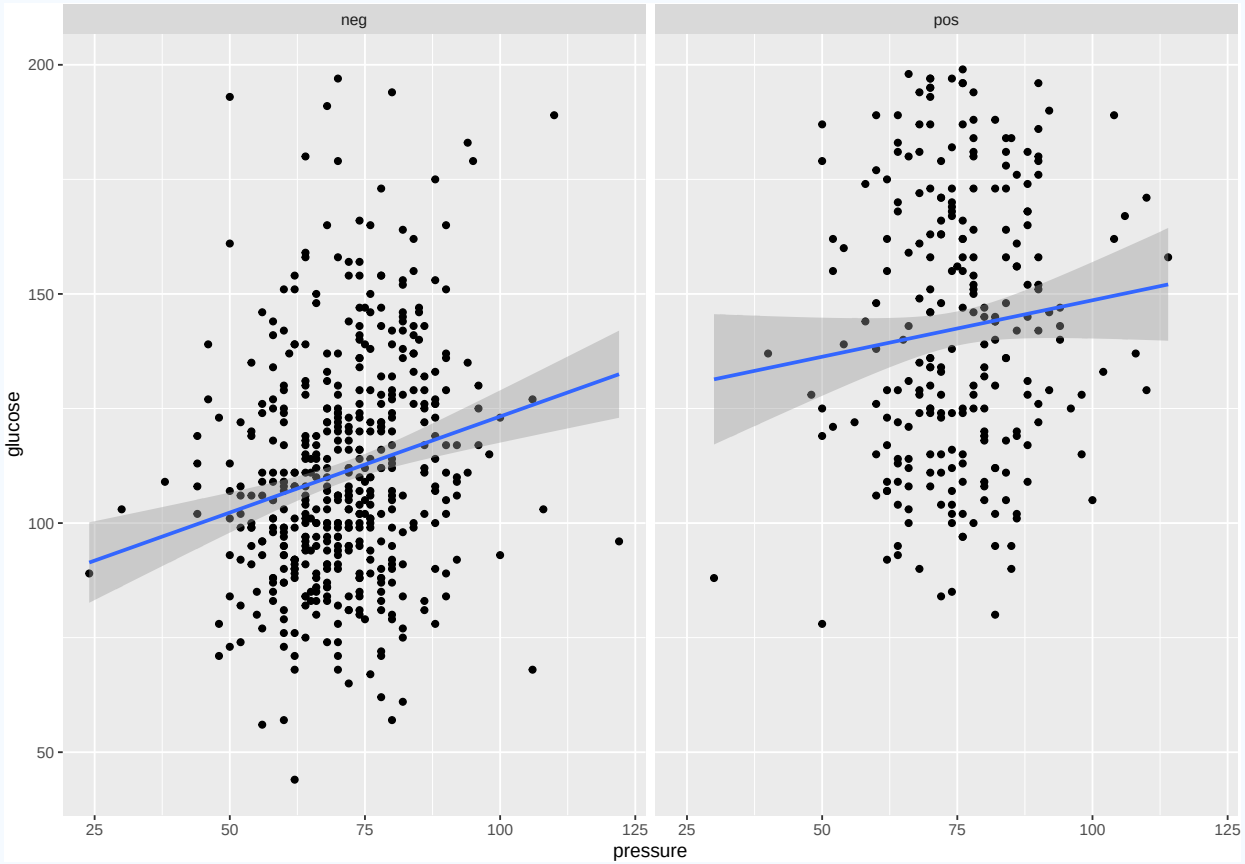


Diagrama de puntos (dotplot): categórica vs continua

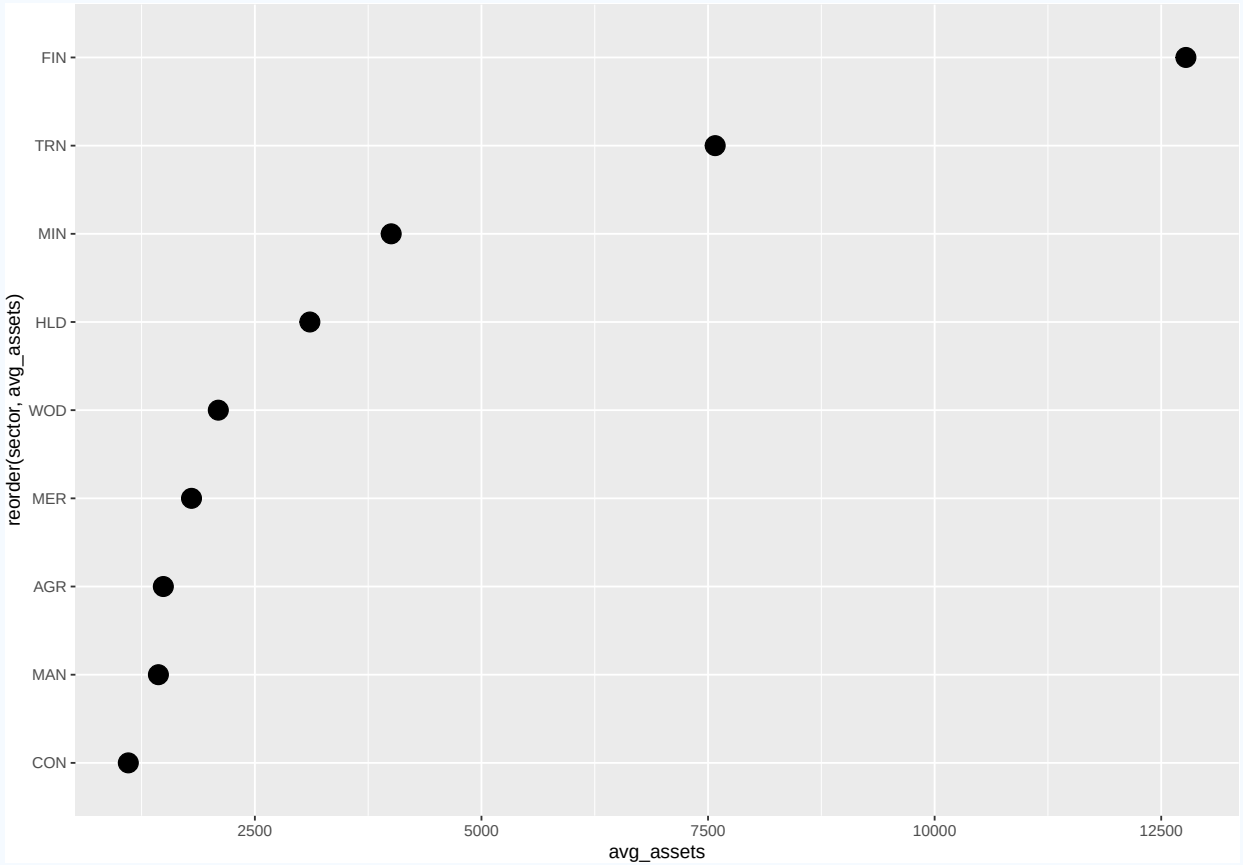
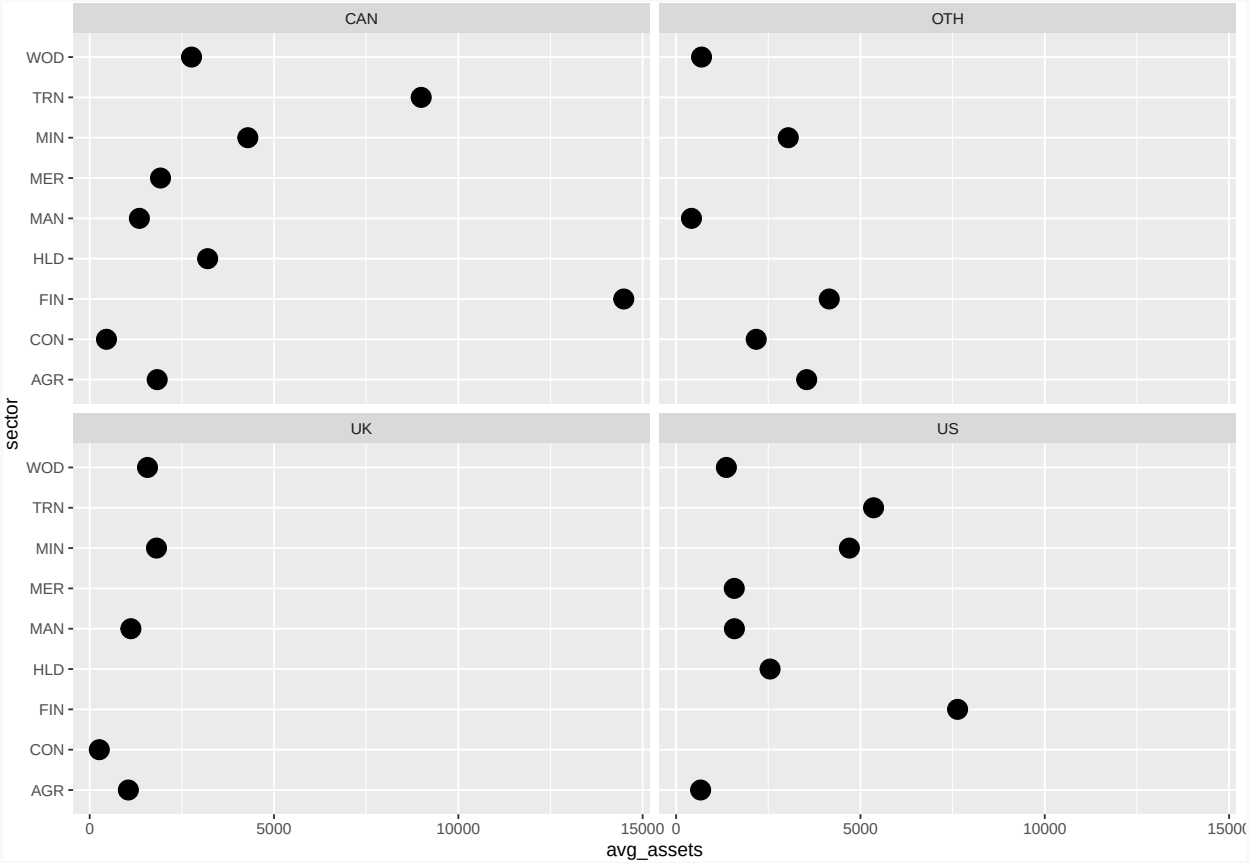
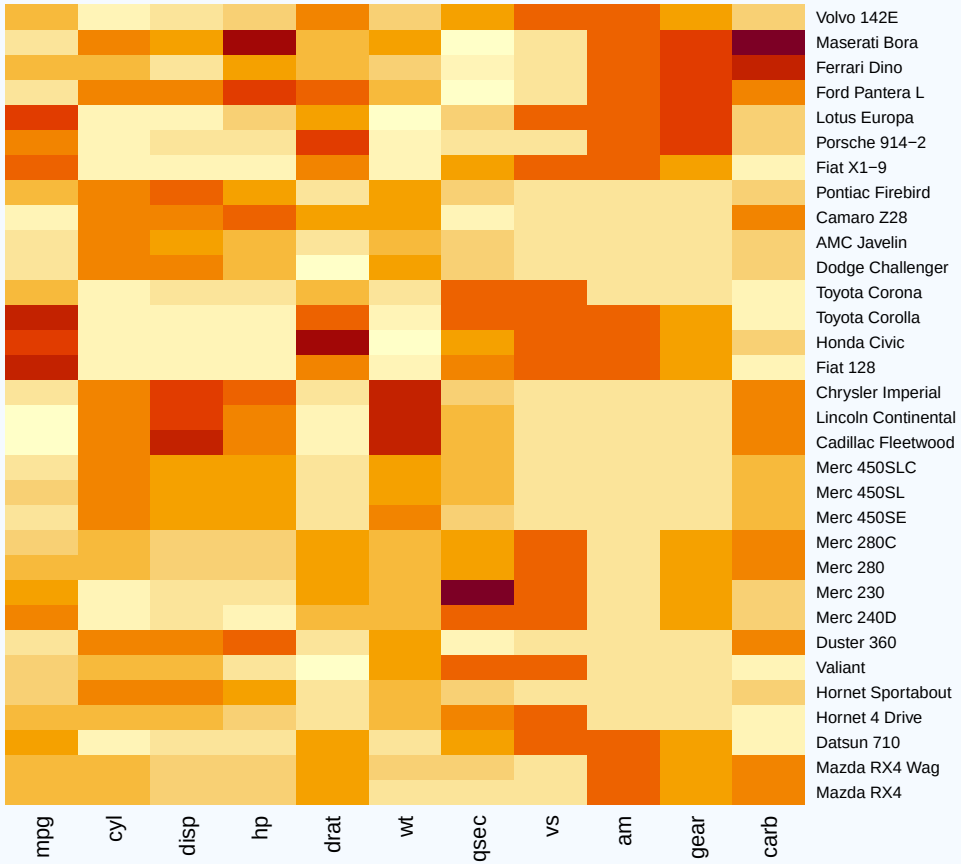


Diagrama de dispersión por categorías

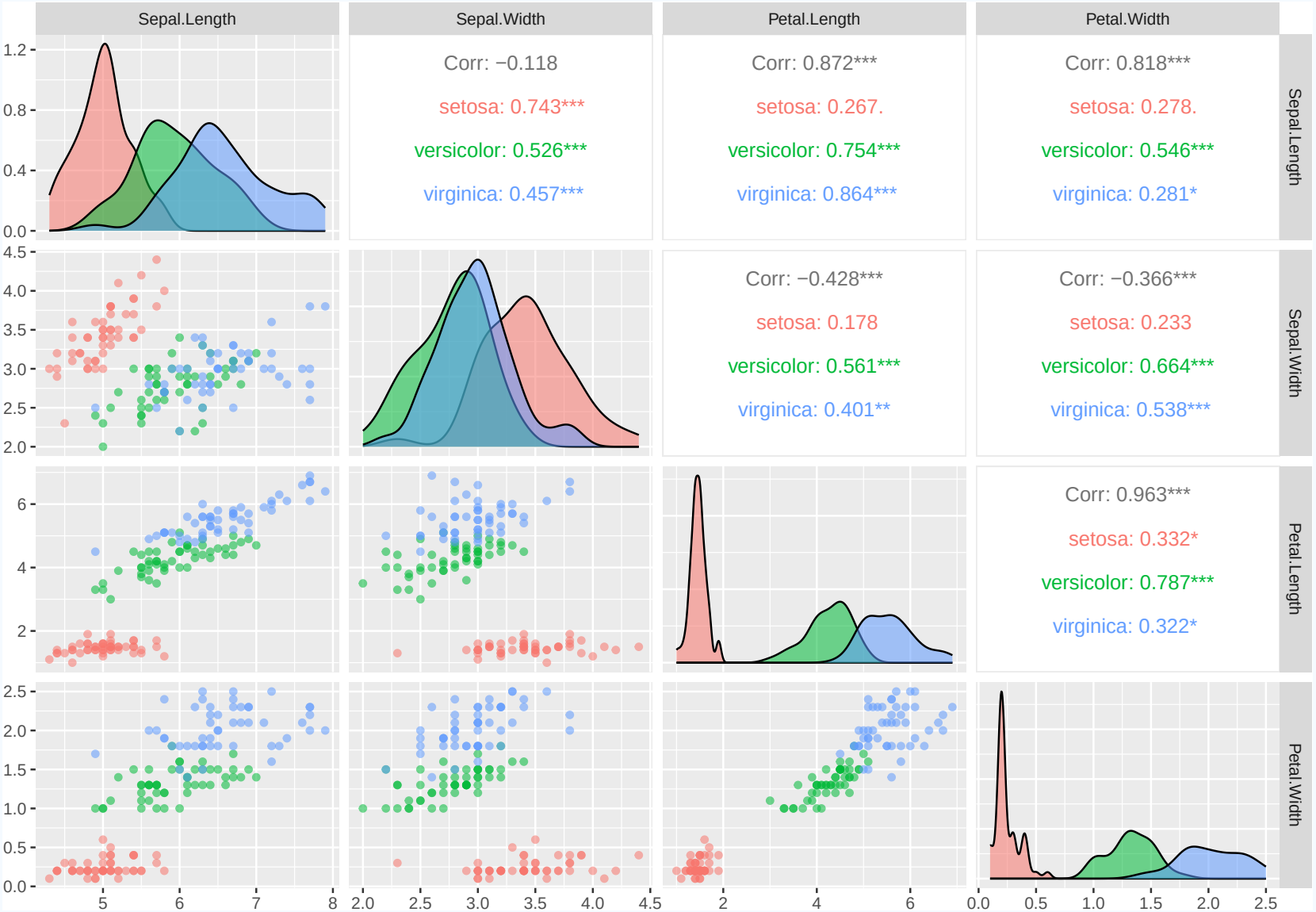


Heatmap (mapa de calor)



ggpairs

https://r-charts.com/es/correlacion/ggpairs/



Transformación de variables

- Distribución adecuada de la variable (Ej: distribución Normal)
- Relación con otras variables y visualización
- Igualar dispersión entre variables –> variables en escalas comparables
- Variables cuantitativas –> variables categóricas

Transformación de variables

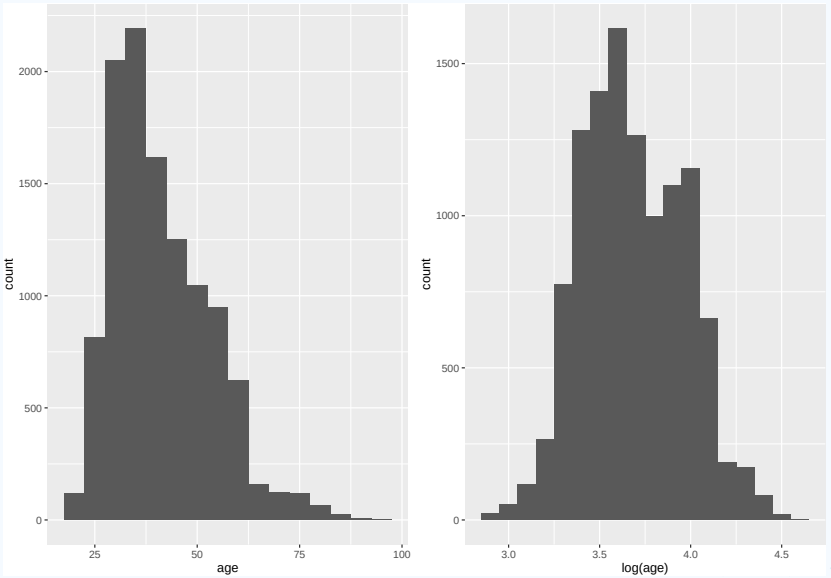
- Distribución adecuada de la variable (Ej: distribución Normal)

```
bank = read.csv('https://raw.githubusercontent.com/rafiag/DTI2020/main/data/bank.cs

    hist_var <- ggplot(data = bank) +
      geom_histogram(mapping = aes(x = age), binwidth = 5)

    hist_log <- ggplot(data = bank) +
      geom_histogram(mapping = aes(x = log(age)), binwidth = .1)

    grid.arrange(hist_var, hist_log, ncol=2)
```



Transformación de variables

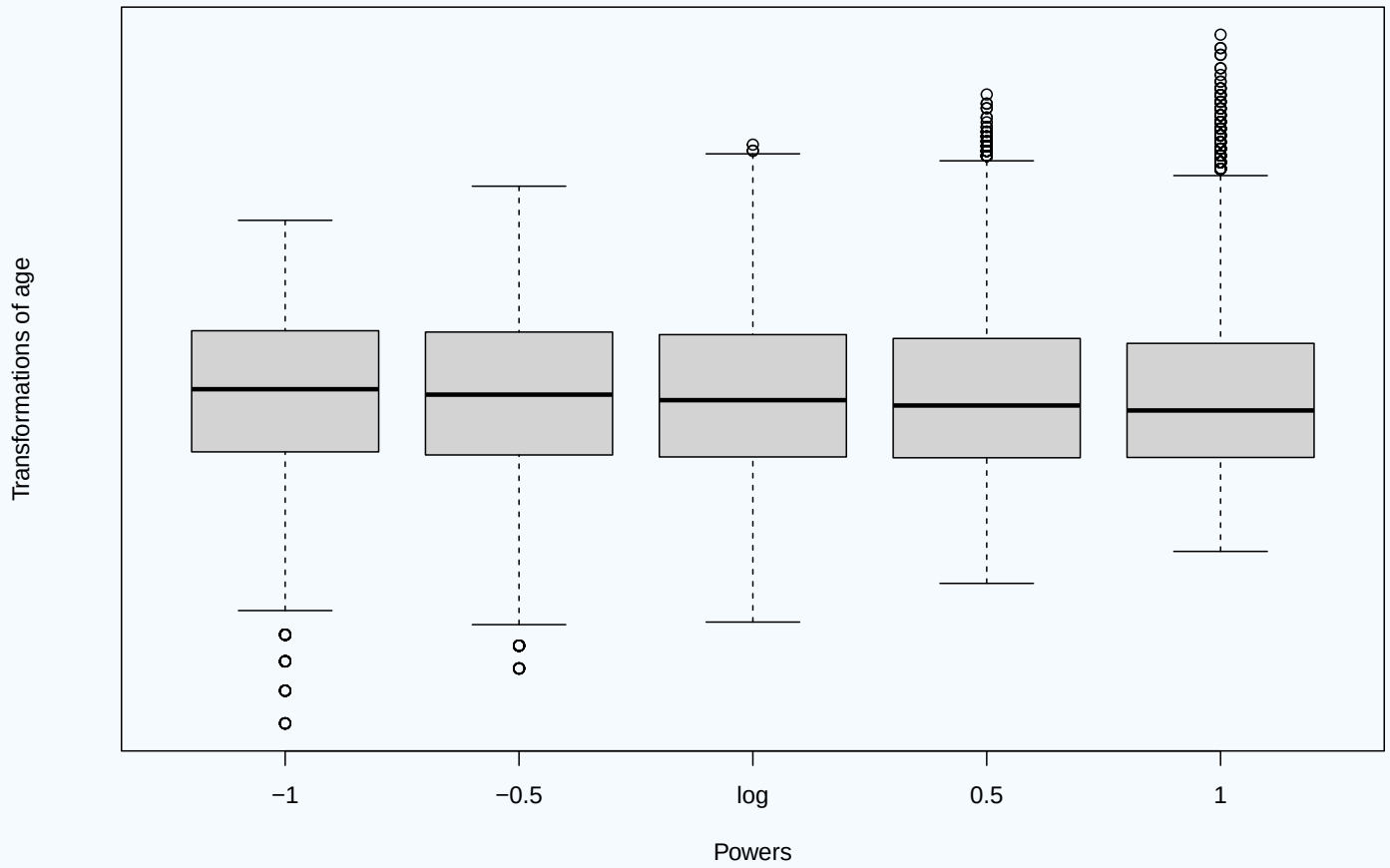
Obtener una variable cuya distribución de valores sea:

- Más simétrica y con menor dispersión que la original
- Más semejante a una distribución normal (e.g. para algunos modelos lineales)
- Restringida en un intervalo de valores (e.g. [0,1])

Transformaciones de escala-potencia o transformaciones Box-Cox:

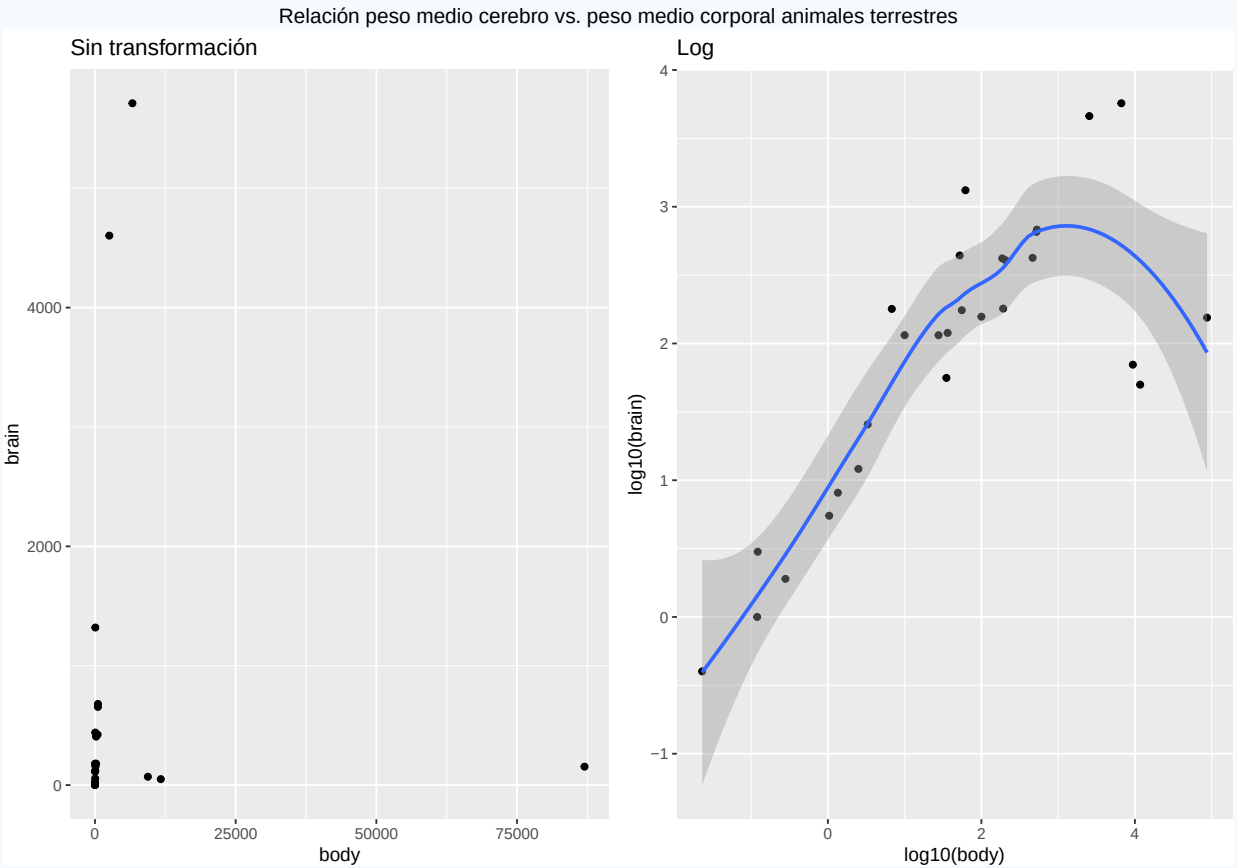
$$x(\lambda) = \begin{cases} \frac{x^{\lambda}-1}{\lambda}, & \text{cuando } \lambda \neq 0, \\ log_e(x), & \text{cuando } \lambda = 0 \end{cases}$$

Transformación de variables



Transformación de variables

- Relación con otras variables y visualización



Transformación de variables

Igualar dispersión entre variables –> variables en escalas comparables

- **Reescalado o cambio de escala**: Sumar o restar una constante a un vector, y luego multiplicar o dividir por una constante. Por ejemplo, para transformar la unidad de medida de una variable (grados Farenheit $\rightarrow$ grados Celsius).
- **Normalización**: Dividir por la norma de un vector, por ejemplo para hacer su distancia euclídea igual a 1.
- **Estandarización**: Consiste en restar a un vector una medida de localización o nivel (e.g. media, mediana) y dividir por una medida de escala (dispersión). Sea X una variable aleatoria con media $\bar{x}$ y desviación típica s:
  $Estandarizacin \rightarrow Y = \frac{X - \bar{x}}{s}$
  Distribución con media 0 y desviación típica 1
  $Escaladomin - max \rightarrow Y = \frac{X - min_x}{max_x - min_x}$

Transformación de variables

Variables cuantitativas –> variables categóricas:

- Agrupación de datos numéricos. Ej: Edad -> “menos de 18 años”, “18-30 años”, “31-45 años”, “46-60 años”, “mayor de 60”
- Creación de variables binarias. Ej: clientes satisfechos / insatisfechos

Más y más

- Series temporales
- Mapas
- Gráficos específicos de clustering
- Pirámides de población
- QQplot
- etc

Más y más

https://elartedeldato.com/

https://rkabacoff.github.io/datavis/ Modern Data Visualization with R

https://r-graph-gallery.com/ggplot2-package.html

https://r-graph-gallery.com/

https://www.data-to-viz.com/

Bibliografía de utilidad para el lector

- "Fundamentos de ciencia de datos con R" coordinado por Gema Fernández-Avilés y José-María Montero: https://cdr-book.github.io/
- Weiss, N. A., & Weiss, C. A. (2017). *Introductory statistics*. London: Pearson.
- "Estadística Aplicada a las Ciencias y la Ingeniería" escrito por Emilio L. Cano. https://emilopezcano.github.io/estadistica-ciencias-ingenieria/index.html
- R for Data Science: https://r4ds.hadley.nz/eda
  - Primera versión en castellano: https://es.r4ds.hadley.nz/

Técnicas de reducción de la dimensionalidad

Aprendizaje Automático 1 - Grado en Ciencia e Ingeniería de Datos

Curso académico 2023-2024

Reducción de la dimensión

- **Objetivo**: identificar un conjunto más pequeño de variables que capturen la mayor parte de la información esencial contenida en el total de las variables originales
- **Ventajas**:
  - Reducir la cantidad de información utilizada, especialmente útil cuando se trabaja con grandes conjuntos de datos
  - Eliminación de problemas de correlación entre variables –> elimina la redundancia en los datos y previene posibles distorsiones en los resultados del análisis
  - Posibilidad de visualizar los datos de manera sencilla (a veces en 2D) –> facilita la interpretación y la comunicación de resultados
- **Desventaja**: falta de explicabilidad cuando las nuevas variables son una combinación de las originales (e.g, PCA)

Técnicas

- **Análisis de componentes principales (PCA)**
- Escalado multidimensional (MDS)
- Análisis de correspondencias
- **Selección de variables (Feature Selection)**
- Autoencoders
- t-SNE (t-Distributed Stochastic Neighbor Embedding)

Selección de variables

- Seleccionar el subconjunto de variables que proporcionen la misma información que el total de variables (el rendimiento del modelo debe ser igual o superior)
- Tipos:
  - Filter
  - Wrapper
  - Embedded

Selección de variables: Filter

- Previos al entrenamiento de un modelo de ML
- Estudian la relación entre la variable objetivo y el resto de variables usando alguna medida de relevancia
- En función de dicha medida, proporcionan un ranking de variables
- Poco costosos computacionalmente
- Medidas de relevancia:
  - Correlación de Pearson
  - Tests estadísticos: T-test, Chi-cuadrado

Selección de variables: Wrapper

- Involucran modelos de ML
- Usan modelos de ML para evaluar la calidad de distintos subconjuntos de variables en función de su capacidad predictiva
- Tienen en cuenta la interacción entre las variables
- Costoso computacionalmente
- Ejemplos:
  - Forward selection
  - Backward elimination
  - Stepwise selection

Selección de variables: Embedded

- Involucran modelos de ML
- Realizan la selección y evaluación de las variables como parte del entrenamiento del modelo de ML
- Ejemplos: árboles de decisión, random forest, regresión lasso

Análisis de componentes principales

- Se debe a Pearson (1901) y a Hotteling (1933)
- **Objetivo**: representar la información recogida en el conjunto de datos original mediante un número menor de variables que son combinaciones lineales de las originales y que están incorreladas entre sí
- Herramienta exploratoria
- Información - Variabilidad (Varianza)

Dada una matrix $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_p)^T$ formada por n elementos (filas) y p variables (columnas), las componentes principales son p nuevas variables $\mathbf{z_j}, j = 1, \ldots, p$ construidas como una combinación lineal de las originales:

$$\mathbf{Z} = \mathbf{XA} = \mathbf{a}_1\mathbf{x}_1 + \cdots + \mathbf{a}_p\mathbf{x}_p$$

donde $\mathbf{A}^t\mathbf{A} = \mathbf{I}$, es decir, las nuevas variables $\mathbf{Z}$ están incorreladas entre sí. Es decir, obtener las componentes principales $\mathbf{Z}$ es realmente hacer una transformación ortogonal de las variables originales $\mathbf{X}$ para lograr nuevas variables incorreladas entre sí.

En este proceso se quiere maximizar la información, es decir, la variabilidad recogida, por tanto se buscan transformaciones que maximicen la varianza.

Cálculo de la primera componente

- La primera componente será $\mathbf{z}_1 = \mathbf{a}_1^t \mathbf{X}$, $\mathbf{a}_1$ es un vector de constantes
- Para asegurar la ortogonalidad de la transformación $\mathbf{a}_1^t \mathbf{a}_1 = 1$
- El cálculo de $\mathbf{z}_1$ depende de $\mathbf{a}_1$. Así, buscamos $\mathbf{a}_1$ tal que $\mathbf{z}_1$ tenga máxima varianza y se cumpla que $\mathbf{a}_1^t \mathbf{a}_1 = 1$
- Máxima varianza –> maximizar $Var(\mathbf{z}_1) = \mathbf{a}_1^t \Sigma \mathbf{a}_1$, siendo Σ la matriz de varianzas covarianzas

Cálculo de la primera componente

- Multiplicadores de Lagrante para maximizar una función ($max\ Var(\mathbf{z}_1) = \mathbf{a}_1^t \Sigma \mathbf{a}_1$) sujeta a restricciones ($\mathbf{a}_1^t \mathbf{a}_1 = 1$):

$$L(\mathbf{a}_1) = \mathbf{a}_1^t \Sigma \mathbf{a}_1 - \lambda(\mathbf{a}_1^t \mathbf{a}_1 - 1)$$

Derivamos e igualamos a 0:

$$\frac{\partial L}{\partial \mathbf{a}_1} = 2\Sigma \mathbf{a}_1 - 2\lambda \mathbf{a}_1 = 0 \leftrightarrow \Sigma \mathbf{a}_1 = \lambda \mathbf{a}_1$$

Es decir, $\mathbf{a}_1$ es un autovector de la matriz Σ y λ su autovalor

Cálculo de la primera componente

- ¿Qué autovalor? Multiplicamos por $\mathbf{a}_1^t$: $\underbrace{\mathbf{a}_1^t \Sigma \mathbf{a}_1}_{Var(\mathbf{z}_1)} = \lambda \underbrace{\mathbf{a}_1^t \mathbf{a}_1}_{} 1 = \lambda$

  Es decir, λ es la varianza de la primera componente principal. Como se busca que sea máxima, será el mayor autovalor de Σ.

- La matriz de varianzas covarianzas Σ, de tamaño $(p \times p)$ es definida positiva: tiene p autovalores distintos $\lambda_1, \ldots, \lambda_p$

Segunda componente principal

Cálculo de la segunda componente principal $\mathbf{z}_2$: similar pero añadiendo la condición de estar incorrelada con $\mathbf{z}_1$ esto es,

$$Cov(\mathbf{z}_2, \mathbf{z}_1) = 0 \Leftrightarrow Cov(\mathbf{z}_2, \mathbf{z}_1) = Cov(\mathbf{a}_2^t\mathbf{X}, \mathbf{a}_1^t\mathbf{X}) = \mathbf{a}_2^t\Sigma\mathbf{a}_1 = 0$$

Como $\Sigma\mathbf{a}_1 = \lambda\mathbf{a}_1$: $\mathbf{a}_2^t\Sigma\mathbf{a}_1 = \lambda\mathbf{a}_2^t\mathbf{a}_1 = 0$ –> vectores ortogonales
Así, ahora, buscamos $max\ Var(\mathbf{z}_2) = \mathbf{a}_2^t\Sigma\mathbf{a}_2$ sujeta a $\mathbf{a}_2^t\mathbf{a}_2 = 1$ y $\mathbf{a}_2^t\mathbf{a}_1 = 0$
Y volvemos a llegar a $\Sigma\mathbf{a}_2 = \lambda\mathbf{a}_2$, escogiendo en este caso λ como el segundo mayor autovalor de la matriz de varianzas covarianzas

Finalmente obtenemos $\mathbf{z} = \mathbf{AX}$, siendo $\mathbf{z}_1, \ldots, \mathbf{z}_p$ variables incorreladas entre sí y con varianza

$$Var(\mathbf{z}_1) = \lambda_1, \ldots, Var(\mathbf{z}_p) = \lambda_p$$

que son los autovalores de la matriz de covarianzas.

Estos autovalores indican la variabilidad que recoge cada componente principal. Para hablar en términos de porcentaje de variabilidad recogido por cada componente:

$$\frac{\lambda_i}{\sum_{i=1}^{p} \lambda_i}$$

Objetivo: seleccionar el menor número $m < p$ de componentes principales que recojan un % alto de la variabilidad total

- El cálculo se basa en la matriz de varianzas covarianzas (no la de correlaciones) –> PCA depende de la escala
- PCA debe aplicarse a datos en los que las variables tengan escalas similares (comparables)

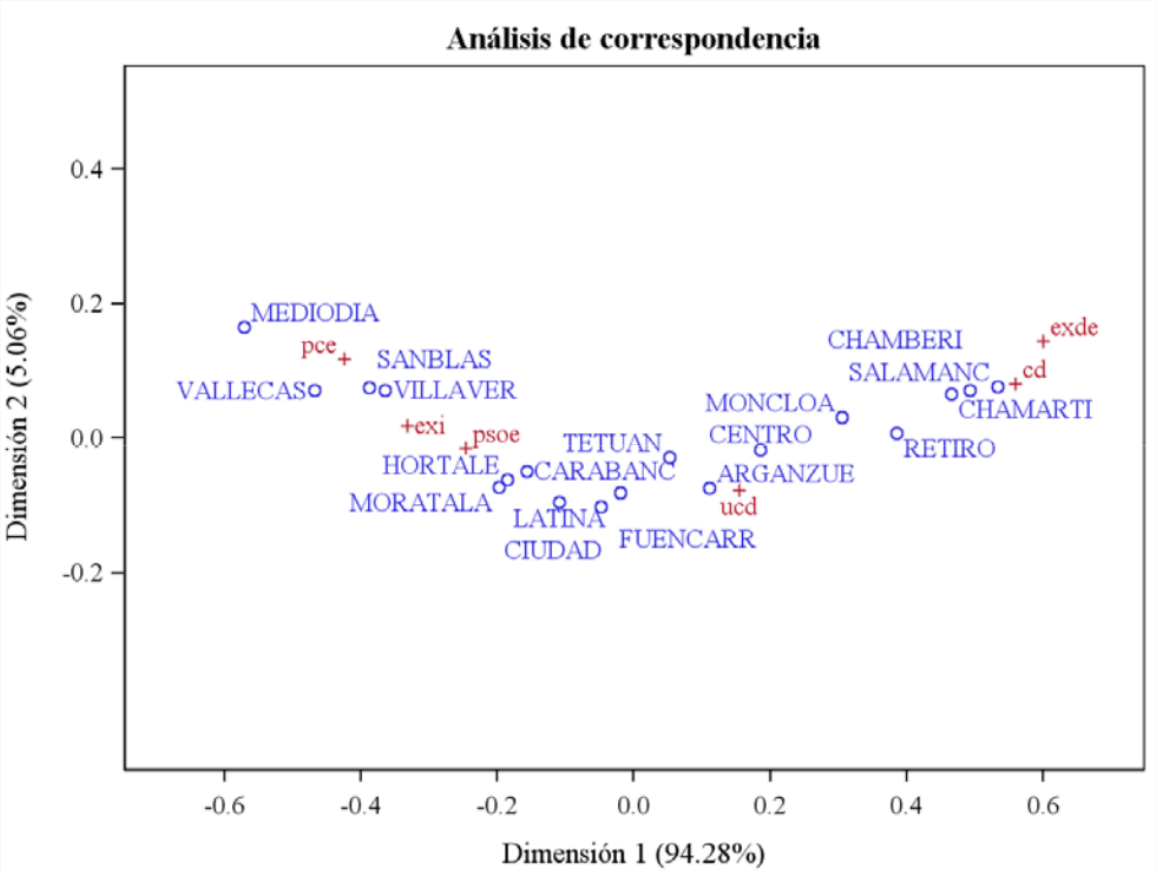
¿Qué ocurre si las variables originales están incorreladas?

Escalado multidimensional (MDS)

- Generalización del concepto de PCA
- Los datos son una matriz de distancias o de similaridades –> no se tienen observaciones y variables, se tienen distancias (o similaridades) entre ellos (Ej: comparación de productos para marketing)
- Traduce información de distancias entre n elementos en una representación de n puntos en el espacio más pequeño (2 o 3 dimensiones es lo ideal).
- La visualización en un espacio más pequeño permite entender la estructura, ver si hay grupos, qué elementos se asemejan más, etc.
- ¿Cómo funciona?
  - Ofrece unas coordenadas iniciales
  - Va modificando dichas coordenadas buscando que las distancias de las observaciones en las nuevas coordenadas sean lo más parecidas posibles o sus distancias en los datos originales.

Análisis de correspondencias

- Caso particular de MDS usando la distancia Chi-cuadrado
- Input: tabla de contingencia con las frecuencias absolutas observadas de 2 variables cualitativas. Ej: Color de ojos y color de pelo, Distritos de Madrid y partidos políticos



Referencias

John, George H, Ron Kohavi, y Karl Pfleger. 1994. «Irrelevant features and the subset selection problem». En *Machine learning proceedings 1994*, 121-29. Elsevier.

Guyon, I., Gunn, S., Nikravesh, M., & Zadeh, L. A. (Eds.). (2008). *Feature extraction: foundations and applications* (Vol. 207). Springer.

Peña, D. (2013). *Análisis de datos multivariantes*. Cambridge: McGraw-Hill España.

Cuadras, C. M. (2007). *Nuevos métodos de análisis multivariante* (Vol. 20). Barcelona: CMC editions.

Aprendizaje no supervisado

Aprendizaje Automático 1 - Grado en Ciencia e Ingeniería de Datos

Curso académico 2023-2024

Introducción

- No se dispone de etiqueta, no hay clases a aprender
- **Objetivo**: particionar el conjunto de datos en grupos de observaciones donde cada observación se parezca lo más posible a las observaciones de su mismo grupo y lo menos posible a las observaciones de los otros grupos
- Grupos se llaman conglomerados o clústeres
- Ejemplos: segmentación del mercado, visualización, detección de anomalías, imputación de valores faltantes, etc.

Clasificación

- **Agrupamiento jerárquico.** Se trata de estructurar los elementos de un conjunto de forma jerárquica y en función de su similitud. Una clasificación jerárquica conlleva que los datos se ordenan en niveles y los niveles superiores contienen a los inferiores. Cuando una observación forma parte de un cluster, permanece en él.
- **Agrupamiento no jerárquico (particiones de los datos)**. Se dividen los datos en un número de grupos de tal modo que cada elemento pertenezca a uno y sólo uno de los grupos, todo elemento quede clasificado y cada grupo sea internamente homogéneo. En los algoritmos de partición de datos todos los clusters se encuentran de forma simultánea.

Medida de similitud

- En el análisis de conglomerados las agrupaciones se hacen en función de la semejanza entre los individuos

Medida de desemejanza o similitud

La medida de desemejanza entre dos variables x e y es una función $\delta(x, y)$ que cumple

- $\delta(x, y) = 0 \Leftrightarrow x = y$
- $\delta(x, y) \geq 0$ (no negatividad)
- $\delta(x, y) = \delta(y, x)$ (simetría)

Si además verifica la desigualdad triangular se trataría de una distancia:

$$\delta(x, y) \leq \delta(x, z) + \delta(y, z)$$

La idea de la distancia y de la medida de desemejanza es que cuanto mayor sean, menos parecido habrá entre los dos puntos que se estudian

Parámetros de un modelo de ML

- Un **parámetro** es un valor que el algoritmo del modelo de ML ajusta durante el proceso de entrenamiento para hacer que el modelo se adapte mejor a los datos de entrenamiento y, en última instancia, haga predicciones más precisas en datos no vistos.
- Los parámetros son esenciales para definir la estructura y el comportamiento del modelo
- Tipos de parámetros:
  - **Parámetros** del modelo: componentes internos del modelo, se obtienen de los datos. Ej: los coeficientes asociados a las variables en una regresión lineal o logística

  son valores que no se obtienen a partir de los datos, sino que los propone el científico de datos
  - **Hiperparámetros** del modelo: Los establece el científico de datos antes del entrenamiento y controlan aspectos más generales del modelo. Ej: el valor k en el modelo de los k vecinos (me fijo en $k = 2$ vecinos para determinar la clase de un punto).
- Ajuste de parámetros e hiperparámetros: clave para el desarrollo de modelos existosos

Partición de los datos: k-medias

- El algoritmo de las k-medias es el algoritmo de aprendizaje automático no supervisado más utilizado para agrupar un conjunto de observaciones en un conjunto de grupos o clústeres
- K representa el número de grupos pre-especificados por el científico de datos
- **Idea**: definir clústeres de tal modo que se minimice la variabilidad total dentro de los clústeres (*within-cluster variation*):

$$W(S, c) = \sum_{k=1}^{K} W(S_k) = \sum_{k=1}^{K} \sum_{\mathbf{x}_i \in S_k} (\mathbf{x}_i - \mathbf{c}_k)^2$$

siendo K el número total de clusters, S_k el cluster k, x_i cada una de las observaciones del conjunto de datos y $\mathbf{c}_k$ el centroide del cluster S_k (la media de los elementos de dicho cluster). Nótese que los clústeres resultantes $S = \{S_1, \ldots, S_K\}$ son disjuntos, es decir, $S_1 \cap \cdots \cap S_K = \emptyset$

Algoritmo de las k-medias

1. **Inicialización**. Se elige el número K de clusters y se escogen al azar K centroides del conjunto de datos
2. **Actualización de los clústeres**. Dados los K centroides $\mathbf{c}_k$, $k = 1, \ldots, K$, cada punto $\mathbf{x}_i$ se asigna al centroide del que menos dista, es decir, al que minimice $(\mathbf{x}i - \mathbf{c}_k)^2$. Los elementos atribuidos a cada centroide $\mathbf{c}_k$ forman el cluster S_k.
3. **Actualización de los centroides**. En cada cluster S_k, $k = 1, \ldots, K$ se calcula la media, que se denota por $\mathbf{c}_k^{'}$ y se convierte en el nuevo centroide.
4. **Test de los centroides**. Se comparan los nuevos centroides $\mathbf{c}_k^{'}$ con los de la iteración previa ($\mathbf{c}_k$). Si $\mathbf{c}_k^{'} = \mathbf{c}_k \;\; \forall k = 1, \ldots, K$, el algoritmo para y se devuelven los clústeres S_k y sus centroides $\mathbf{c}_k^{'}$, $\forall k = 1, \ldots, K$. En caso contrario, se hace $\mathbf{c}_k = \mathbf{c}_k^{'}$ y se vuelve al paso 2.

Los clústeres están representados por su centroide, entendiendo éste como un punto de referencia

Versiones de las k-medias

- **Algoritmo de Forgy/Lloyd**. Algoritmo en modo batch. Se caracteriza porque cuando se realiza una transformación se aplica a todos los elementos a la vez. Por ejemplo, cuando se actualizan los centroides se actualizan todos al mismo tiempo.
- **Algoritmo de MacQueen.** Es un algoritmo iterativo, también llamado online o incremental. Los centroides se recalculan cada vez que un elemento cambia de cluster y después de pasar por todos los elementos.

- **Algoritmo de Hartigan & Wong**. Busca la partición que minimice la suma de los cuadrados de los errores. Esto significa que se puede asignar un elemento a otro grupo, aunque ya pertenezca al cluster cuyo centroide es el más cercano, siempre que con esto se minimice $W(S, c)$.

Ventajas e inconvenientes

Ventajas

- Simple y rápido
- Escalable: aplicable con facilidad a grandes conjuntos de datos
- Computacionalmente mejor que los algoritmos jerárquicos. Eficiente en tiempo y memoria
- Tiene su propia función objetivo, la cual se pretende minimizar, que permite hacerse una idea de cómo de buena es la solución
- Sólo depende de un parámetro k

Ventajas e inconvenientes

Desventajas

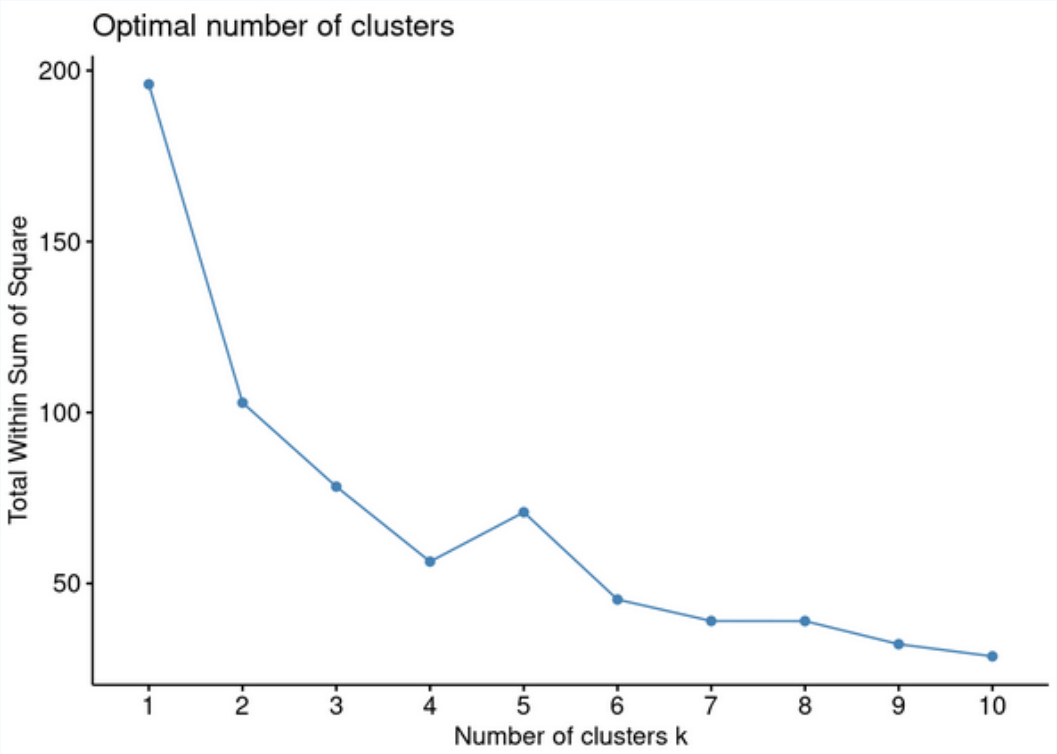
- Hay que seleccionar el valor de k y los centroides iniciales
- Puede converger a mínimos locales –> puede no ser la partición óptima
- Depende de la inicialización
  - Solución: replicar el algoritmo con distintas inicializaciones
- Tiende a crear grupos del mismo tamaño y con forma globular. Resultados pobres si los grupos son no convexos
- Usa la media –> se ve afectado por atípicos.
  - Solución: usar medoides en lugar de centroides. Los medoides son obligatoriamente puntos de la muestra

Métodos de elección de k

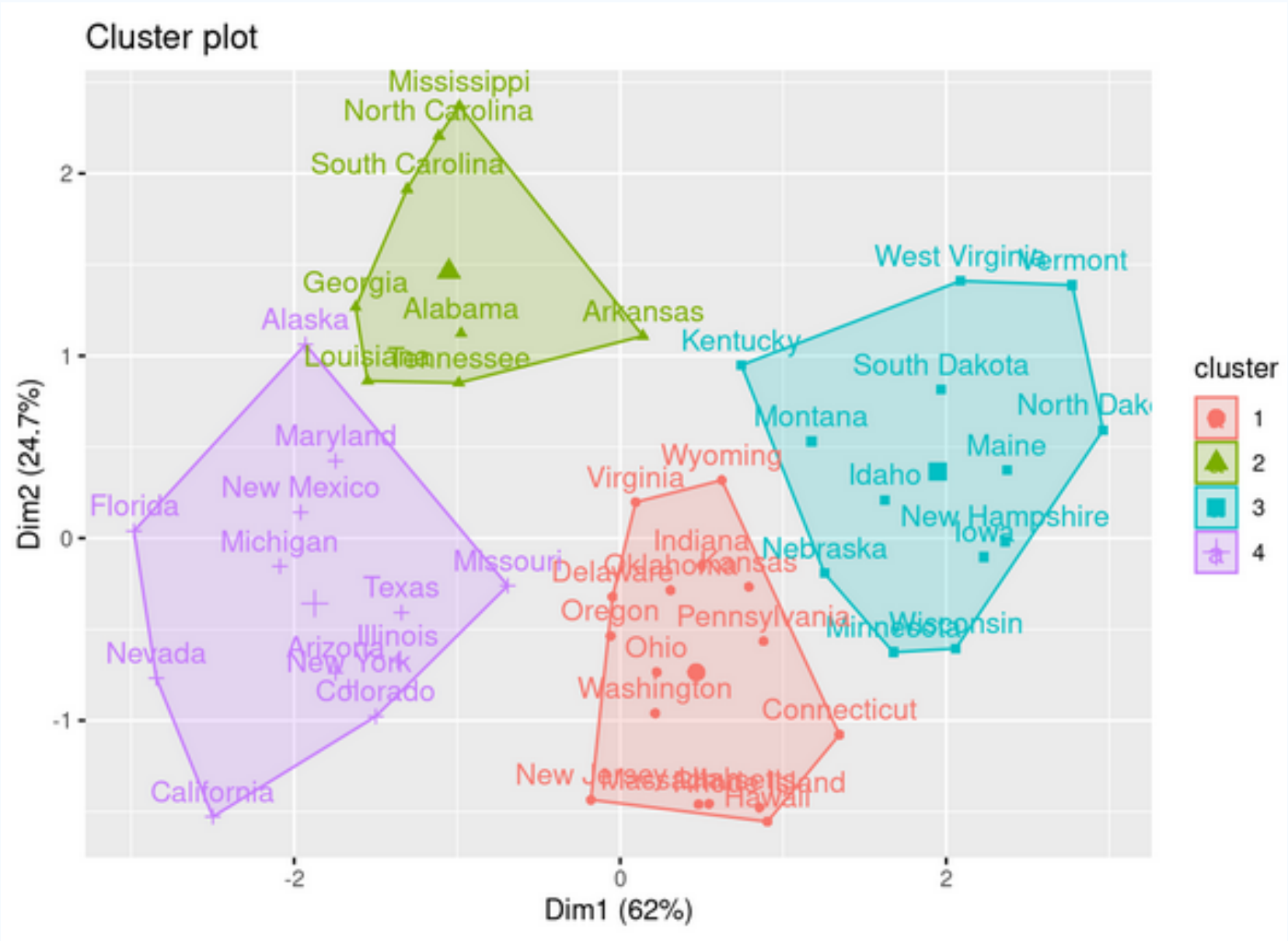
- Existen distintas aproximaciones para elegir el valor de k
  - Paquete NbClust de R contiene 30 criterios distintos
- Algunos de los más conocidos:
  - Método del codo
  - Método de la silueta (Silhouette)
  - Gap

Elección de k: método del codo

- **Objetivo de las k-medias**: construir grupos que minimicen la variación total dentro de los clústeres
- Calcular dicha variación para distintos valores de k y graficarla
- Un "codo" en el gráfico avisará del número apropiado de grupos, es decir, cuando el descenso de la variación de un k al siguiente no sea llamativa



Elección de k: método del codo



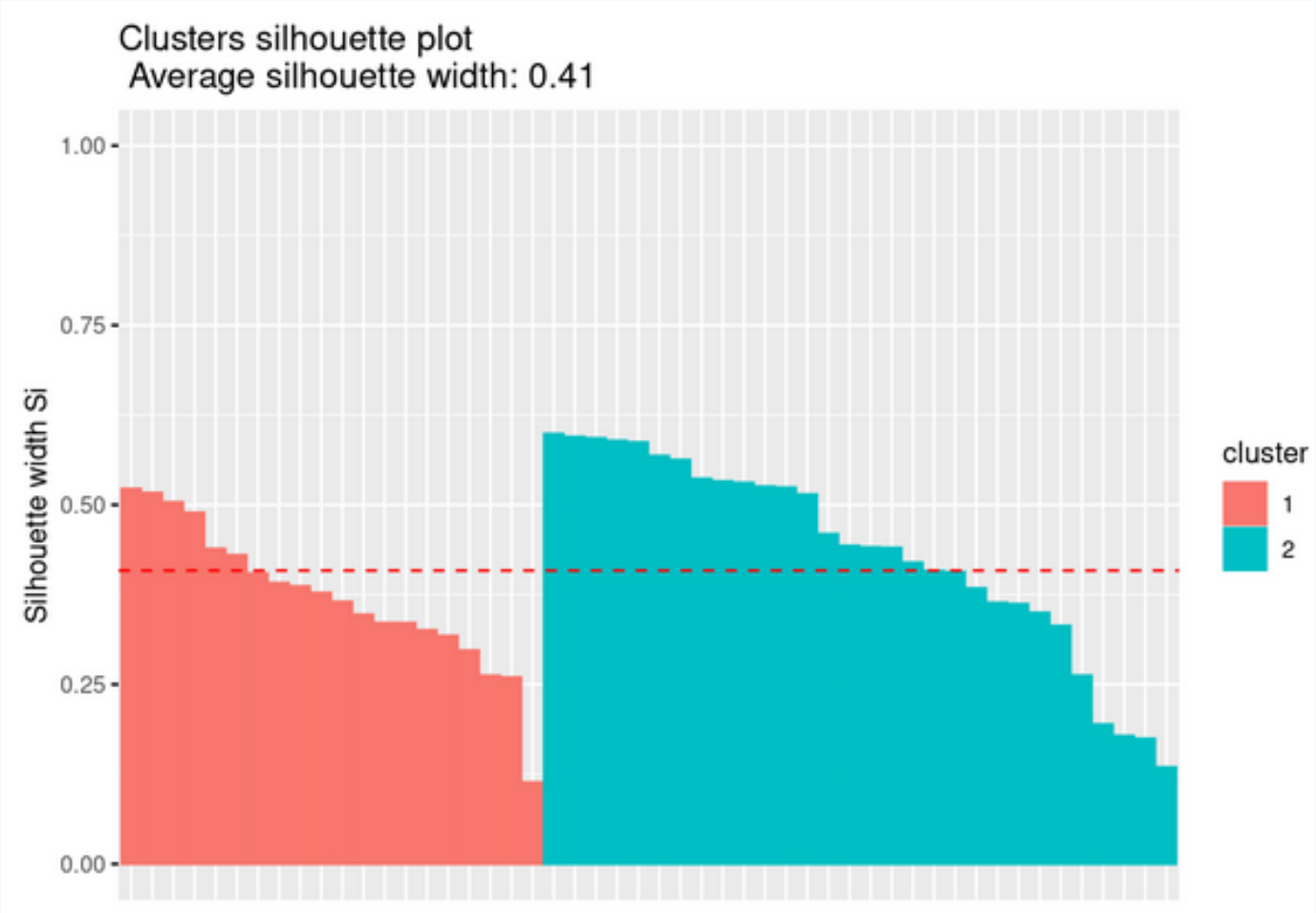
Elección de k: método de la silueta

- Técnica no supervisada para valorar la coherencia o calidad de la partición
- La silueta determina hasta qué punto cada elemento se encuentra dentro de su agrupación.
- Para cada observación $\mathbf{x}_i$, la silueta se calcula como

$$s(\mathbf{x}_i) = \frac{b(\mathbf{x}_i) - a(\mathbf{x}_i)}{max(a(\mathbf{x}_i), b(\mathbf{x}_i))}$$

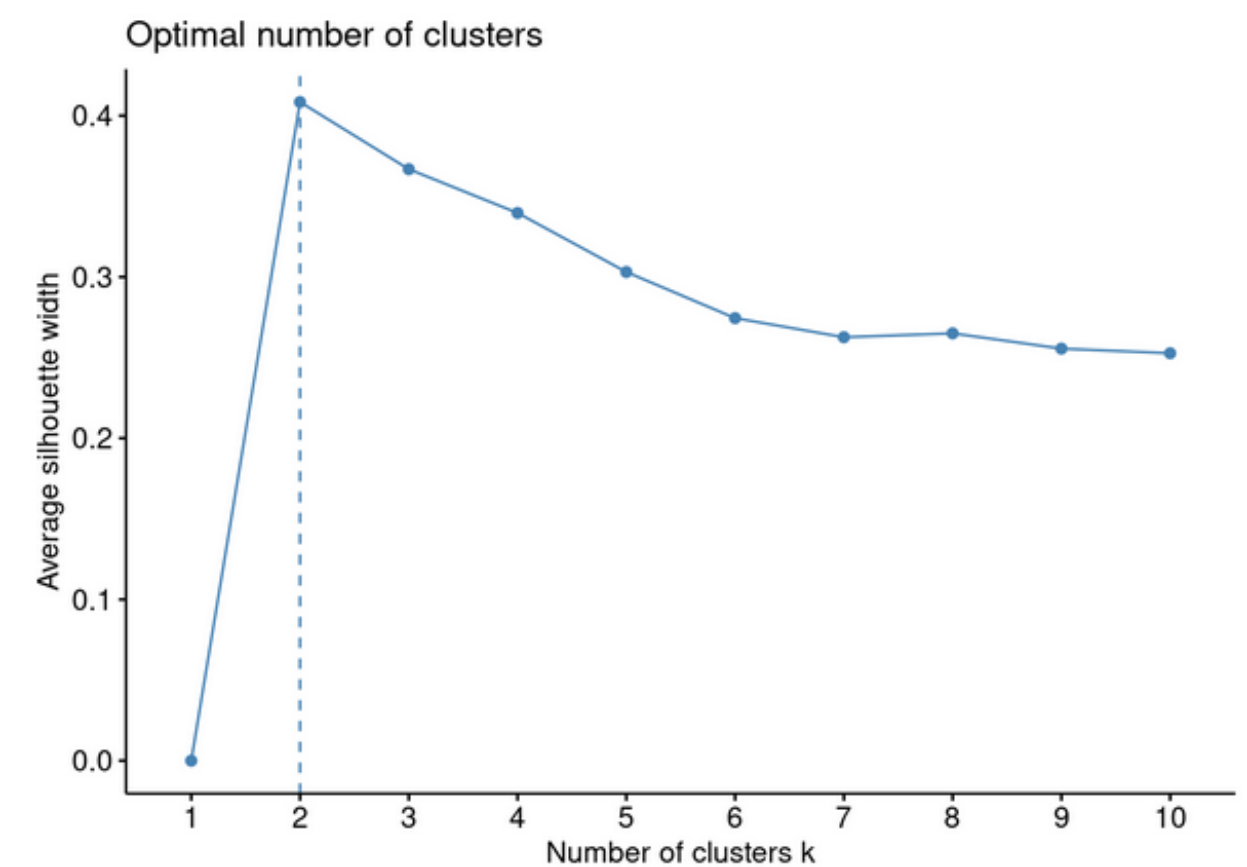
  donde $a(\mathbf{x}_i)$ es la media de las distancias de la observación $\mathbf{x}_i$ a los puntos de su propio cluster y $b(\mathbf{x}_i)$ es la media de las distancias de $\mathbf{x}_i$ a los puntos de su cluster más cercano (excluyendo el suyo)
- Interpretación:
  - Valores altos: el punto encaja bien en su cluster
  - Valores bajos o negativos: el punto no encaja bien en su cluster

Elección de k: método de la silueta



Elección de k: método de la silueta

- La silueta media es la media de los valores silueta de todos los puntos.
- Se calcula para varios valores de k. El k adecuado es aquel que maximiza la silueta media.



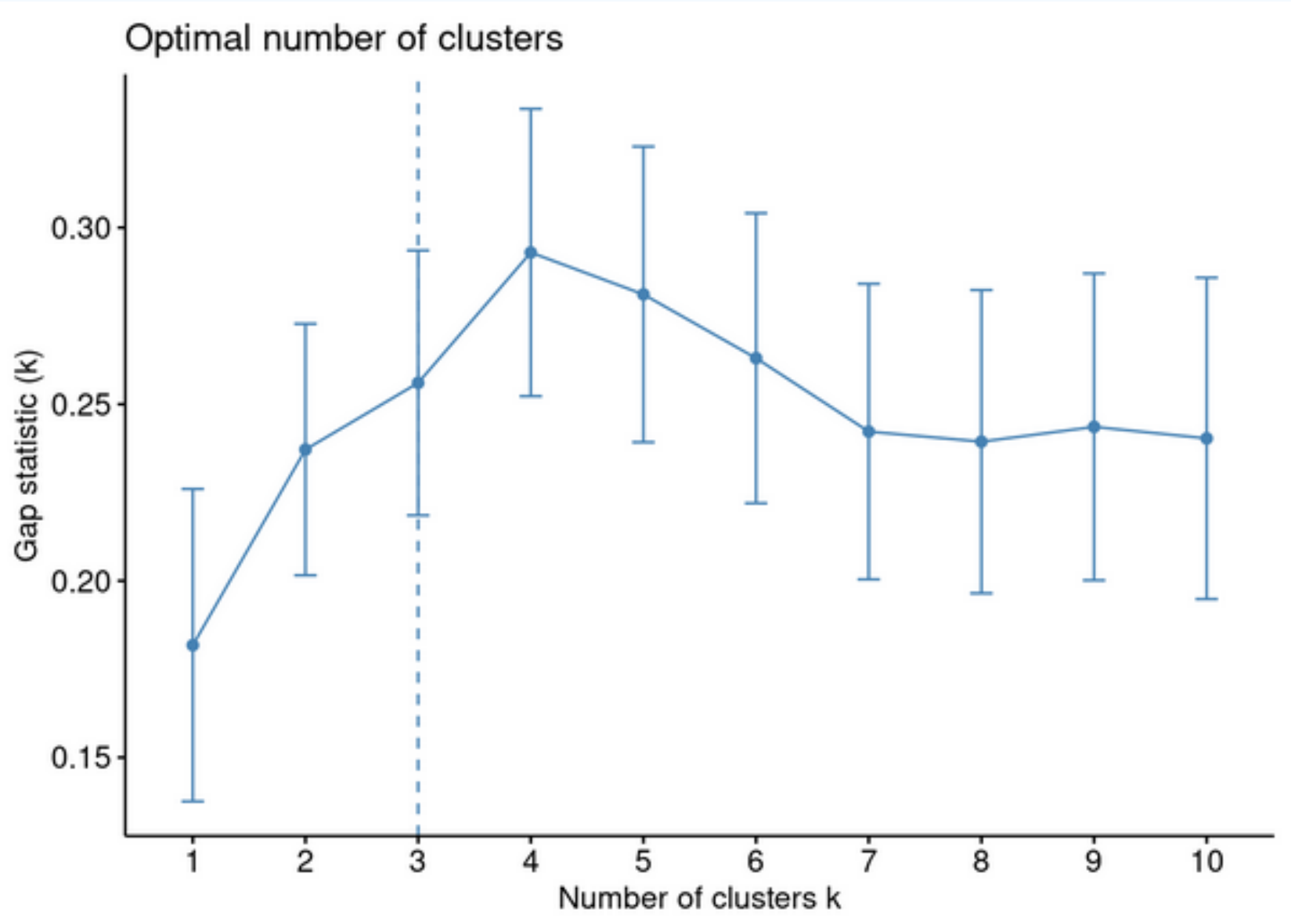
Elección de k: Gap

- Compara los resultados del clustering con datos aleatorios sin patrones de clustering (datos uniformes)
- Se realiza el clustering para distintos valores de k y se calcula la varianza total intracluster W_k
- Se simulan B conjuntos de datos aleatorios y se les aplica el mismo algoritmo de clustering con los distintos k y se calcula su varianza intracluster W_{kb}, $b = 1, \ldots, B$
- Cálculo del estadístico Gap para cada k. Se comparan ambas varianzas intracluster. Cuanto mayor sea la diferencia, más sólida es la partición para dicho k.

$$Gap(k) = \frac{1}{B} \sum_{b=1}^{B} log(W_{kb}) - log(W_k)$$

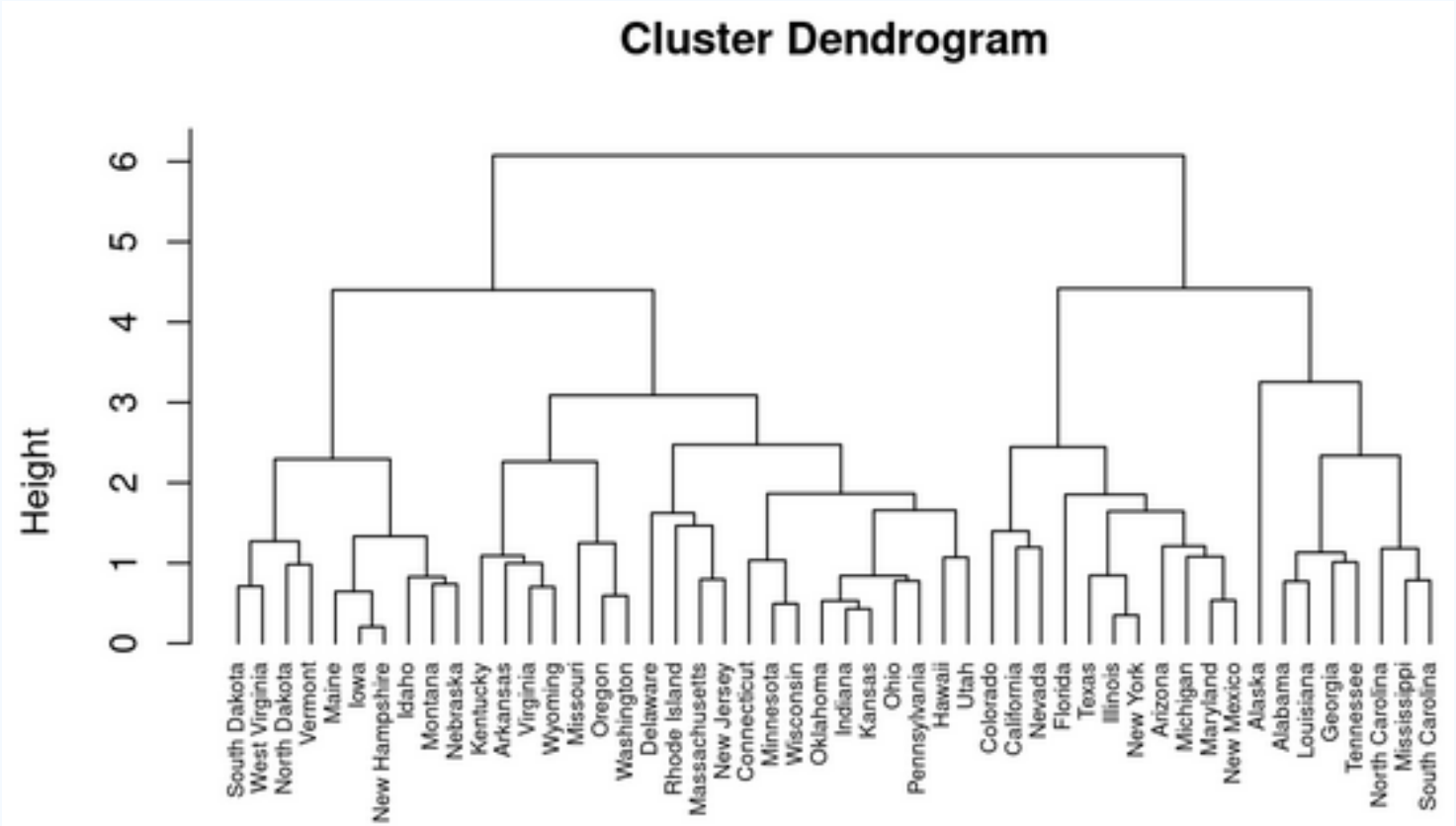
- Se escoge el k que maximiza dicha diferencia

Elección de k: Gap



Cluster jerárquico

- Generan una clasificación iterativa de clústeres anidados mediante la unión o la separación de clústeres creados en etapas anteriores



Cluster jerárquico

- Tipos:
  - **Aglomerativos**. Cada observación comienza siendo un clúster, y en cada iteración se unen los dos clústeres más similares, hasta alcanzar una situación final en la que todos las observaciones pertenecen a un único clúster. Esta versión se conoce como **AGNES** ("Agglomerative Nesting").
  - **Divisivos**. Todas las observaciones comienzan en un único clúster y se va dividiendo hasta que cada observación forma un único clúster individual. Esta versión se conoce como **DIANA** ("Divise Analysis").

Cluster jerárquico: criterio de conexión

- **Criterio de conexión** o "***linkage***" especifica cómo se determina el parecido (o la disimilitud) entre dos clústeres.
- Algunos de los criterios más comunes son:
  - Método de Ward
  - Agrupamiento de enlace completo
  - Agrupamiento de enlace promedio
  - Agrupamineto de enlace mínimo o simple
  - Agrupamiento de enlace de centroides
- El criterio de agrupación o conexión es un parámetro fundamental en el resultado final del clustering jerárquico

Cluster jerárquico: criterio de conexión

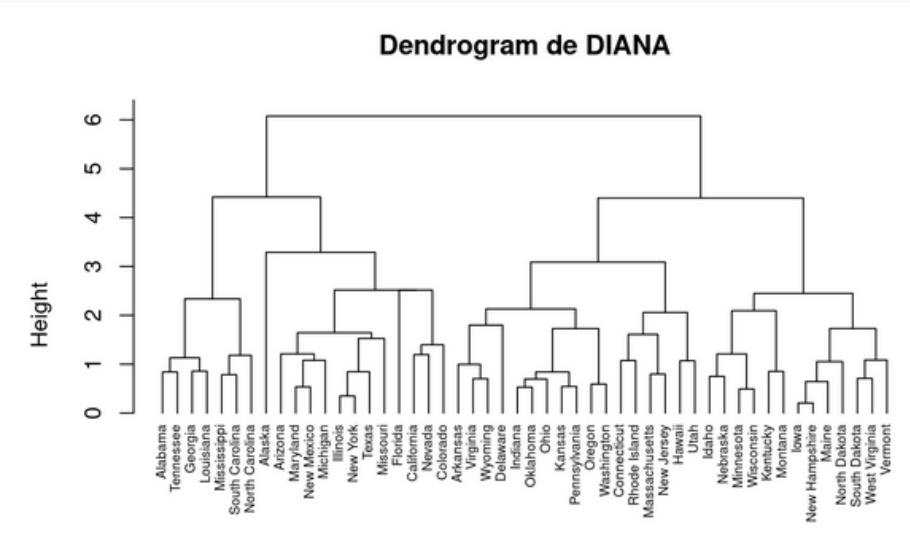
- **Agrupamiento de enlace mínimo o simple.** Minimiza las disimilitudes entre las observaciones más cercanas de dos clústeres. Es decir, calcula todas las disimilitudes por pares entre los elementos del conglomerado A y los elementos del conglomerado B. La disimilitud entre ambos conglomerados será la disimilitud de sus dos puntos más cercanos. Finalmente se unirán aquellos conglomerados con menor disimilitud.
- **Agrupamiento de enlace completo.** Similar al anterior pero con la disimilitud máxima. Minimiza la disimilitud máxima entre las observaciones de dos clústeres.
- **Agrupamiento de enlace promedio.** Minimiza el promedio de las disimilitudes entre las observaciones de dos clústeres. Calcula todas las disimilitudes por pares entre los elementos del conglomerado A y los elementos del conglomerado B, y considera la media de estas disimilitudes como la distancia entre los dos conglomerados.

Cluster jerárquico: criterio de conexión

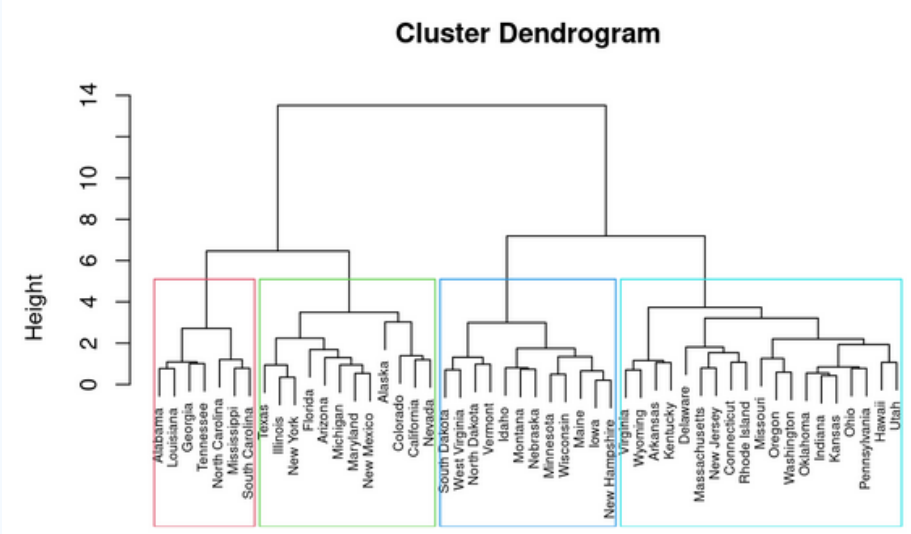
- **Método de Ward.** Minimiza la suma de las diferencias cuadradas dentro de los clústeres, es decir, la varianza intracluster. En cada paso, agrupa los clústeres que provocan el menor incremento de la varianza intracluster $W(S)$.
- **Agrupamiento de enlace de centroides.** La disimilitud entre los conglomerados A y B es la disimilitud entre sus centroides.

Cluster jerárquico: dendograma

Divisivo



Aglomerativo (Ward)

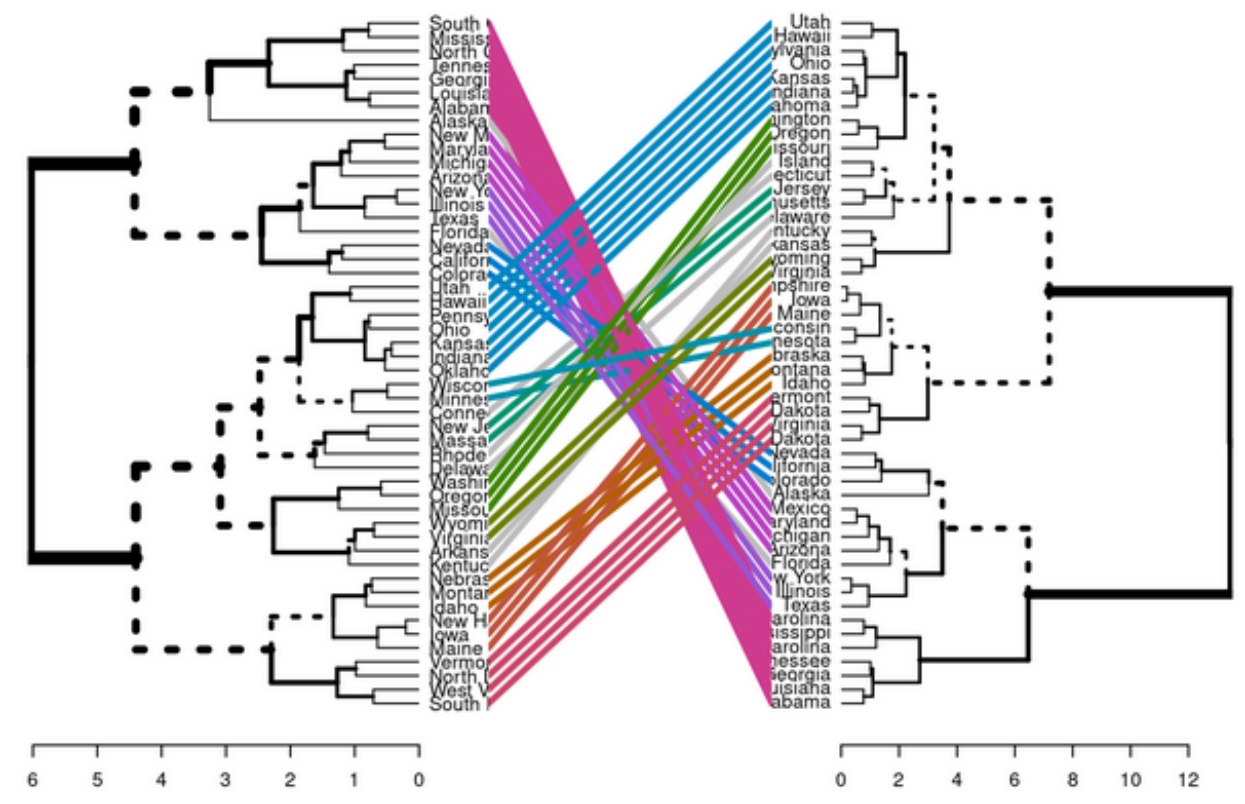


Cluster jerárquico: dendograma

- Cada hoja corresponde a una observación
- La altura de la fusión indica la disimilitud entre las observaciones. Cuanto mayor es la altura de la fusión, menos parecidas son las observaciones
- **¡Ojo!** Cuando empleamos un Dendrograma, las conclusiones sobre la proximidad de dos observaciones sólo pueden extraerse a partir de la altura a la que se fusionan las ramas que contienen primero esas dos observaciones. No podemos utilizar la proximidad de dos observaciones a lo largo del eje horizontal como criterio de su similitud.
- La **altura del corte del dendograma controla el número de clusters** (similar al k en las k-medias)

Cluster jerárquico: comparación de dendogramas

Comparación de enlace Ward y completo:



- Líneas discontinuas: elementos no presentes en el otro dendograma
- Función de entrelazamiento: calidad de alineación entre los dos dendogramas. 1 (entrelazamiento total) y 0 (sin entrelazamiento)

Cluster jerárquico: elección del número de grupos

- Al igual que en el cluster no jerárquico, se pueden aplicar métodos para elegir el número de conglomerados
- Los métodos vistos anteriormente (método del codo, silueta, estadístico Gap) son perfectamente aplicables

Cluster jerárquico: ventajas

- **Jerarquía de clusters**. Permite análisis a distintos niveles
- **Interpretación visual** de cómo se agrupan los datos y se relacionan
- **No requiere especificar previamente el número de grupos**
- **Identificación de subgrupos**. Permite detectar clases dentro de clases
- **Detección de outliers**
- **No sensible a la inicialización**
- **Análisis exploratorio de datos**. Visión general de cómo se agrupan naturalmente los datos sin necesidad de conocimiento previo.

Cluster jerárquico: desventajas

- **Requiere definir un criterio de corte** para convertir la jerarquía en clusters
- **No hay una única respuesta correcta.** Deben considerarse los diversos resultados con información interesante
- **No es óptimo para todo tipo de datos**. Funcionan mejor cuando los datos tienen una estructura jerárquica natural
- **No es adecuado para datos de alta dimensión**
- **Resultados no siempre reproducibles**
- **Sin capacidad de predicción**. No son útiles para predecir a qué clúster pertenece una nueva observación

Mapas auto-organizados de Kohonen

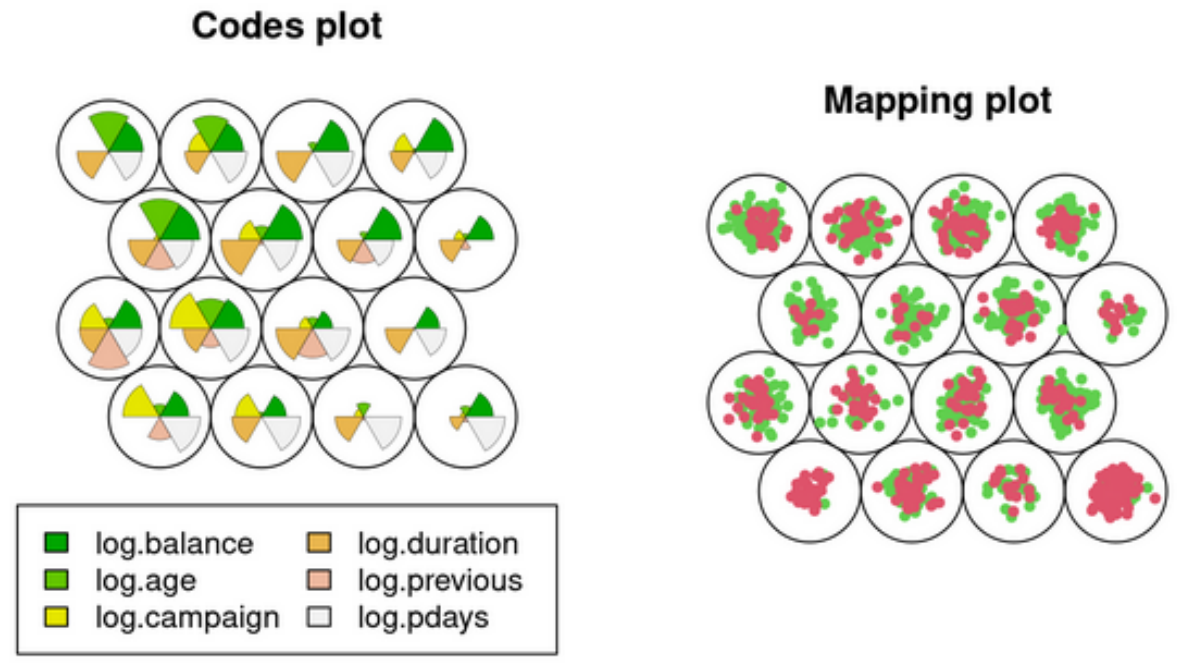
- SOM: Self-Organizing Maps
- Herramienta de **reducción de la dimensión**
- **Algoritmo de clustering**: organizar datos en grupos tal que los elementos dentro de un mismo grupo sean similares entre sí en función de ciertas características.
- **Red neuronal** bidimensional:
  - Capa de entrada con tantas neuronas como variables
  - Capa para representar en 2 dimensiones el total de observaciones

Mapas auto-organizados de Kohonen: algoritmo

1. **Inicialización**: Creación de la red de neuronas bidimensional, conocida como "mapa auto-organizativo (SOM)". Cada neurona representa una ubicación en el espacio SOM.
2. **Asignación de Pesos**: Cada neurona en el SOM tiene asociado un vector de pesos que es del mismo tamaño que los datos originales
3. **Entrenamiento**: Se presentan los datos al SOM, y cada dato se asigna a la neurona cuyos pesos son más similares a los atributos del dato. Las neuronas ganadoras (aquellas a las que se asigna un dato) y sus vecinas en el mapa SOM se ajustan para que se parezcan más al dato presentado. Este proceso de aprendizaje se repite varias veces.
4. **Agrupación en Clusters**: Después del entrenamiento, las neuronas en el mapa SOM que están cerca una de la otra representan clusters de datos. Los datos que se asignaron a estas neuronas durante el entrenamiento se consideran miembros de un mismo cluster.

Mapas auto-organizados de Kohonen

Mapa con estructura 4x4



- **Codes plot**: En cada neurona se muestra el peso de cada variable. Se aprecia la estrucutra de similitud: las variables dominan por zonas
- **Mapping plot:** Densidad de puntos por neurona. En este caso se colorean los puntos en función del target (estudio supervisado)

Mapas auto-organizados de Kohonen: ventajas

- **Topología Preservada**: Los clusters en el SOM reflejan la estructura de vecindad en los datos originales -> facilita la interpretación de los resultados
- **Escalabilidad**: Pueden manejar grandes conjuntos de datos y dimensiones elevadas
- **Visualización**
- **Exploración interactiva:** Los usuarios pueden navegar por el mapa para inspeccionar las regiones y sus contenidos
- **Reducción de ruido:** Pueden ayudar a reducir el ruido y la redundancia en los datos, lo que mejora la calidad del análisis

Mapas auto-organizados de Kohonen: desventajas

- **Sensibilidad a la inicialización** de los pesos de las neuronas
- **Determinación del tamaño del mapa**: Si el mapa es demasiado pequeño, puede no capturar la estructura de los datos correctamente, mientras que si es demasiado grande, puede sobreajustarse a los datos y perder la capacidad de generalización
- **Interpretación de los resultados**: Se dificulta en mapas de gran tamaño
- **Requiere ajuste de hiperparámetros** y el rendimiento depende de ellos
- **Puede converger a mínimos locales**
- **Requiere grandes conjuntos de datos**: Pueden no funcionar bien en conjuntos de datos pequeños o altamente desequilibrados, ya que su eficacia se basa en la capacidad de aprender patrones significativos a partir de una cantidad suficiente de datos

Referencias

Morissette, L., & Chartier, S. (2013). The k-means clustering technique: General considerations and implementation in Mathematica. *Tutorials in Quantitative Methods for Psychology*, *9*(1), 15-24.

Chiang, M. M. T., & Mirkin, B. (2010). Intelligent choice of the number of clusters in k-means clustering: an experimental study with different cluster spreads. *Journal of classification*, *27*, 3-40.

Gan, G., Ma, C., & Wu, J. (2020). *Data clustering: theory, algorithms, and applications*. Society for Industrial and Applied Mathematics.

Mirkin, B. (2005). *Clustering for data mining: a data recovery approach*. Chapman and Hall/CRC.

Xu, D., & Tian, Y. (2015). A comprehensive survey of clustering algorithms. *Annals of Data Science*, *2*, 165-193.

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction* (Vol. 2, pp. 1-758). New York: springer.

Medidas de rendimiento

Aprendizaje Automático 1 - Grado en Ciencia e Ingeniería de Datos

Curso académico 2023-2024

Medidas de rendimiento

All models are wrong but some are useful

Now it would be very remarkable if any system existing in the real world could be exactly represented by any simple model. However, cunningly chosen parsimonious models often do provide remarkably useful approximations. For example, the law $PV = RT$ relating pressure P, volume V and temperature T of an "ideal" gas via a constant R is not exactly true for any real gas, but it frequently provides a useful approximation and furthermore its structure is informative since it springs from a physical view of the behavior of gas molecules.

For such a model there is no need to ask the question "Is the model true?". If "truth" is to be the "whole truth" the answer must be "No". The only question of interest is "Is the model illuminating and useful?".

(Box, GEP, 1979, Robustness in the strategy of scientific model building, *Robustness in Statistics*, Academic Press, pp.201-236)

Medidas de rendimiento

$$DATOS = MODELO + ERROR$$

- **Datos**. La realidad que se quiere comprender, predecir o mejorar
- **Modelo**. Representación **simplificada** de la realidad que se propone para describirla e interpretarla más fácilmente
- **Error**. Diferencia entre la representación simplificada de la realidad (modelo) y los datos (describen la realidad de forma precisa)

Una vez construido un modelo –> evaluación del rendimiento –> ¿error aceptable?

Complejidad del modelo

- Más información en modelo (más variables) –> error suele reducirse
- Más variables –> más complejo es el modelo
- ¿Esto es bueno?
  - **Principio de parsimonia**. Priorizar modelos sencillos. Navaja de Occam
  - **Pérdida de generalidad**. Si añado demasiados parámetros de entrada a un modelo puedo representar exactamente la información de los datos que tengo, pero no funcionará bien con nuevos datos –> **sobreajuste** ("*overfitting*").

Complejidad del modelo

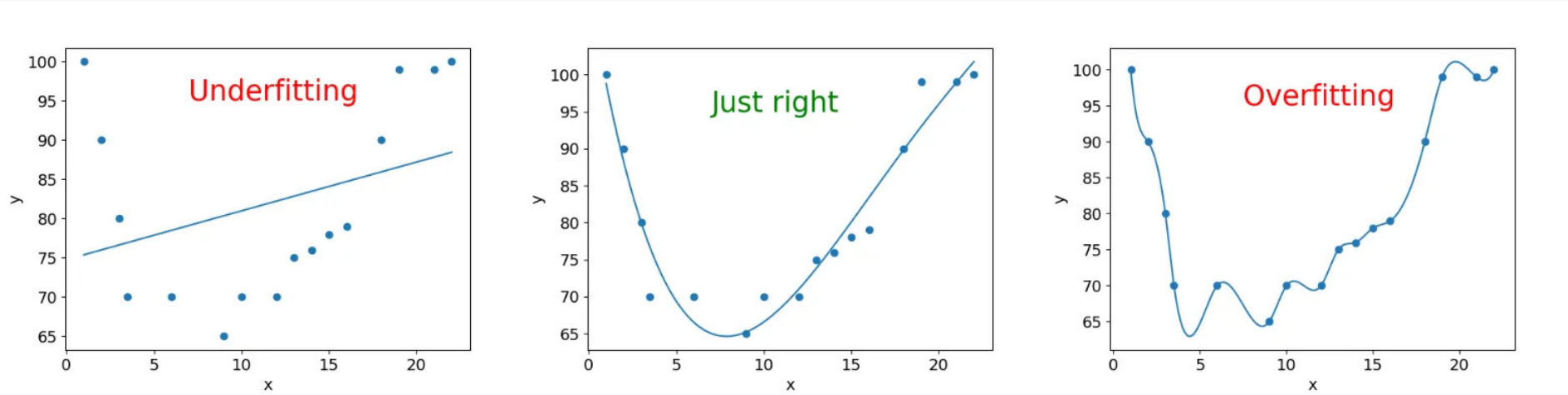


Figure 1: https://medium.com/@kiprono_65591/regularization-a-technique-used-to-prevent-over-fitting-886d5b361700

Balance sesgo-varianza

- **Sesgo**. Diferencia entre la predicción y lo real. Se refiere a la simplificación excesiva de un modelo, asumiendo que los datos de entrenamiento siguen una cierta estructura o patrón predefinido.
  - Sesgo alto –> subajusta los datos, no captura la complejidad de los datos ni representa la relación entre las variables
- **Varianza**. Sensibilidad de un modelo a las fluctuaciones en los datos de entrenamiento
  - Varianza alta –> demasiado ajuste en training –> mal rendimiento en nuevos datos

Balance sesgo-varianza

Equilibrio sesgo-varianza -> Modelos eficaces y con capacidad de generalización

- Modelo con **sesgo alto y varianza baja**: más simple y tiende a subajustar los datos
- Modelo con **sesgo bajo y varianza alta** se ajusta muy bien a los datos de entrenamiento pero generaliza mal

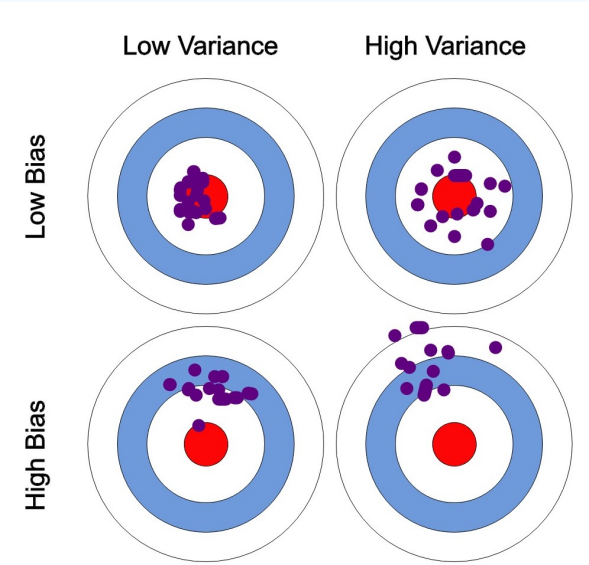


Figure 2: https://nvsyashwanth.github.io/machinelearningmaster/bias-variance/

Balance sesgo-varianza

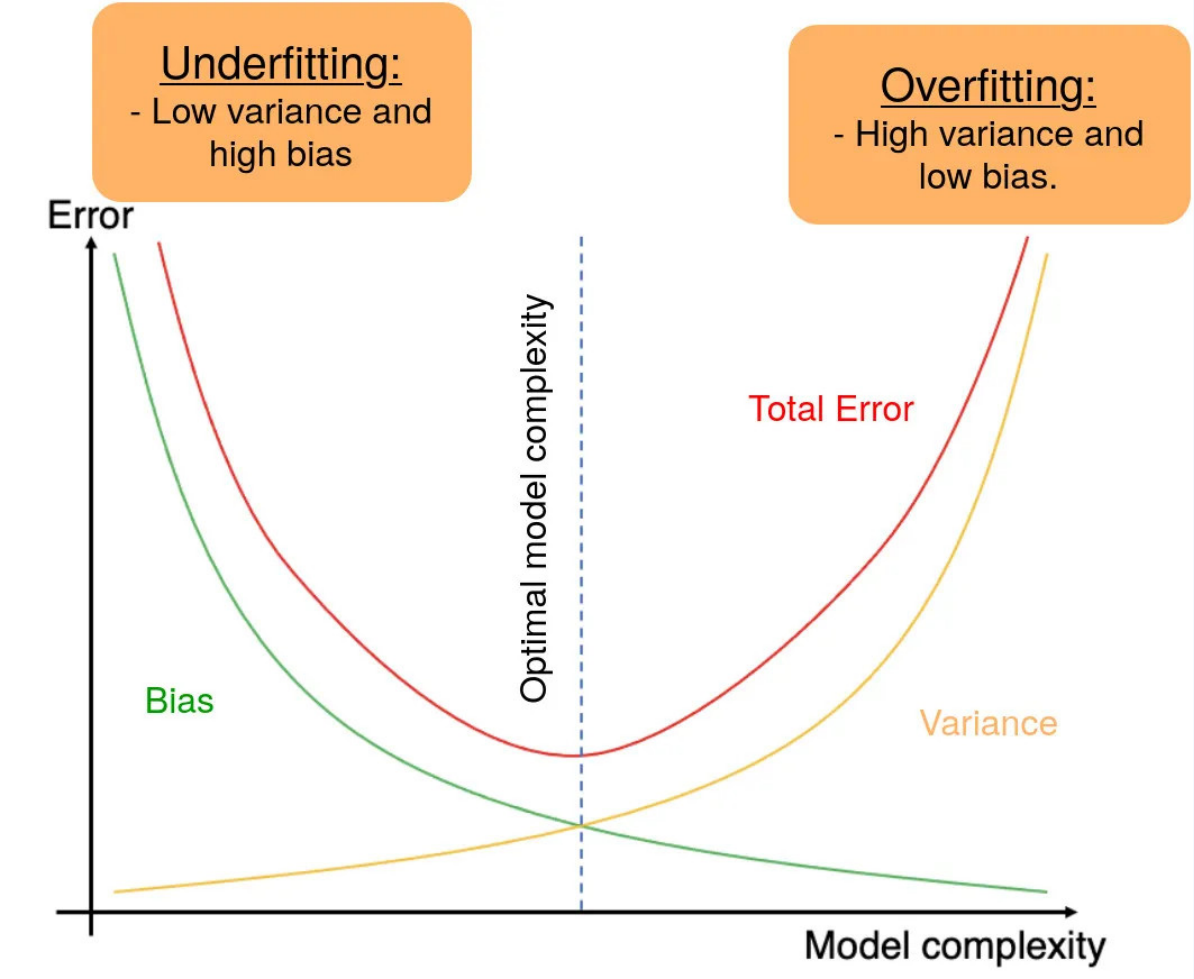


Figure 3: https://medium.com/@kiprono_65591/regularization-a-technique-used-to-prevent-over-fitting-886d5b361700

¿Cómo evitar el sobreajuste?

- Mayor tamaño muestral
- Validación cruzada
- Selección de variables (evitar variables redundantes)
- Modelos simples
- Partición de los datos

Medidas de evaluación

- Estudio del error del modelo
- Evaluación si todas las variables son útiles
- Comparación de modelos
- Comparación del error en distintos conjuntos de datos (train-test-validation) –> capacidad de generalización
- Distintas técnicas para regresión y clasificación
- **Error**: diferencia (en base a alguna medida) entre el valor observado y el predicho

$$error_i \propto targetobservado_i - targetpredicho_i$$

Medidas para regresión

- **Modelo de regresión**: técnica estadística que analiza la relación entre una variable dependiente (o de respuesta) continua y una o más variables independientes (o predictoras)
- **Error en regresión**: diferencia entre los valores predichos y los observados

Medidas para regresión

- **Error Cuadrático Medio (Mean Squared Error, MSE)**: Promedio de las diferencias al cuadrado entre las predicciones del modelo y los valores reales. Sensible a los errores grandes debido al término de cuadrado. Cuanto menor sea el MSE, mejor será el ajuste del modelo:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2$$

  donde y_i es el valor de la variable objetivo en la observación $\mathbf{x}_i$ con $i = 1, \ldots, n$, y $f(x_i)$ es la predicción del modelo de ML
- **Error Absoluto Medio (Mean Absolute Error, MAE)**: Similar al MSE, pero con el valor absoluto. Menos sensible a los errores extremos.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - f(x_i)|$$

Medidas para regresión

- **Raíz del Error Cuadrático Medio (Root Mean Squared Error, RMSE)**: Raíz cuadrada del MSE. Ofrece una medida del error en la misma unidad que la variable objetivo, lo que facilita su interpretación.
- **R-cuadrado (R-squared, R^2)**: Proporciona una medida de la proporción de la variabilidad en la variable dependiente que es explicada por el modelo. Un R^2 más alto indica un mejor ajuste del modelo a los datos, con un valor máximo de 1.

$$R^2 = \frac{\sum_{i=1}^{n}(f(x_i) - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

donde $\bar{y}$ es el valor medio de la variable objetivo

- **Error Porcentual Absoluto Medio (Mean Absolute Percentage Error, MAPE)**: Calcula el porcentaje promedio de error absoluto en relación con los valores reales.

$$MAPE = \frac{100}{n} \sum_{i=1}^{n} |\frac{y_i - f(x_i)}{y_i}|$$

Medidas para clasificación

- **Clasificación binaria**: dividir las observaciones dadas en dos clases mutuamente excluyentes $\{-1,+1\}$
- Medidas se obtienen a partir de la **matriz de confusión**:

		Valor observado	
		-1	1
Valor predicho	-1	TN	FN
	1	FP	TP

  - **TP**: "True positive", 1 clasificados como 1
  - **TN**: "True negative" -1 clasificados como -1
  - **FP**: "False positive" -1 erróneamente clasificados como 1
  - **FN**: "False negative" 1 erróneamente clasificados como -1
- Nótese que: $n = TP + FP + TN + FN$
- Importancia relativa de FP y FN. Ejemplo: control de accesos

Medidas para clasificación

- **Exactitud (Accuracy)**: Medida más común. Representa la proporción de observaciones correctamente predichas, es decir:

$$Accuracy = \frac{TP + TN}{n}$$

- **Error**: recíproco de la exactitud:

$$Error = \frac{FP + FN}{n}$$

- **Sensibilidad (recall)**: también conocida como Recuperación o Tasa de Verdaderos Positivos (TPR). Puede verse como la probabilidad de que un 1 observado sea clasificado efectivamente como 1:

$$Recall = \frac{TP}{TP + FN}$$

Medidas para clasificación

- **Especificidad (specificity)**: también conocida Tasa de Verdaderos Negativos puede verse como la probabilidad de que un -1 observado sea clasificado efectivamente como -1:

$$Specificity = \frac{TN}{TN + FP}$$

- **Precisión**: también conocida Valor Predictivo Positivo puede verse como la probabilidad de que acierto cuando se predice un valor 1:

$$Precision = \frac{TP}{TP + FP}$$

- **Valor Predictivo Positivo (NPV, "Negative Predictive Value")**: tasa de acierto cuando se predice un valor -1:

$$NPV = \frac{TN}{TN + FN}$$

- **F1-score**: media armónica de Precisión y Recuperación:

$$F_1 - score = 2\frac{Precision * Recall}{Precision + Recall}$$

- **F-score generalizado**: media armónica ponderada de Precisión y Recall:

$$F_\beta = (1 + \beta^2)\frac{Precision * Recall}{\beta^2 Precision + Recall}$$

Cuando $\beta = 1$, se tiene la medida anterior: $F_1 - score$. F_2 da el doble de peso a la Recall que a la Precisión. Por contra, $F_{0.5}$ da el doble de peso a la Precisión que a la Recall.

Rendimiento de las particiones

- Partición Train-Test-Validation –> 3 medidas de rendimiento
- Validación cruzada en t trozos
  - t medidas de rendimiento (referentes a Train-Test) –> media y desviación típica
  - La medida de rendimiento de Validation

Ajuste hiperparámetros

- Modelo knn
- Elección del número de vecinos k: validación cruzada (t trozos) probando distintos valores de k, por ejemplo:

Número de vecinos k	Media (Desv. Std.)
1	1,54 (0,33)
2	1,75 (0,41)
3	1,68 (0,28)
4	1,68 (0,35)
5	1,70 (0,34)
6	1,89 (0,38)
7	1,91 (0,36)
8	2,12 (0,45)
9	2,25 (0,46)
10	2,34 (0,34)
11	2,33 (0,38)

Comparación de modelos

- Modelos devuelven como salida, para cada observación, la **probabilidad de pertenencia** a las diferentes clases de la variable respuesta
- Ejemplo: knn devuelve % de tus vecinos que pertenecen a cada clase
- Esta probabilidad se **binariza** para determinar a qué clase pertenece cada observación
- ¿Con qué **umbral**? –> En general 0.5 (**¡Ojo! otro parámetro!!**)
- ¿Datos desbalanceados?

Comparación de modelos: Curvas ROC

- **Curva ROC ("*Receiver Operating Characteristic curve*", o curva característica de operación)**: método gráfico para ilustrar la capacidad predictiva de un modelo de ML binario

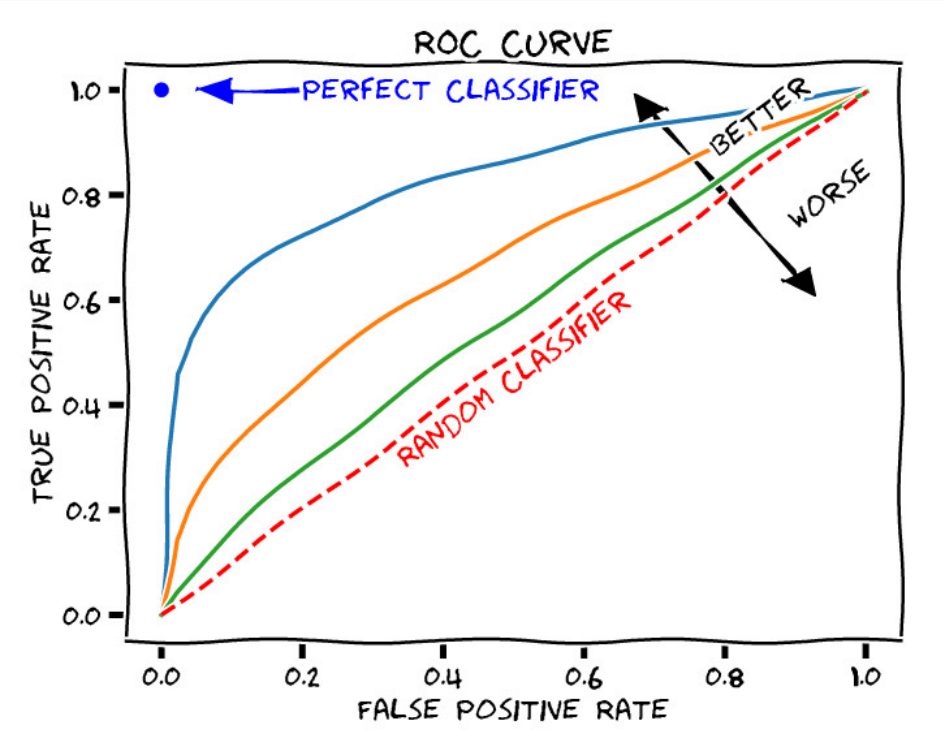


Figure 4:
https://sefiks.com/2020/12/10/a-gentle-introduction-to-roc-curve-and-auc/

Comparación de modelos: Curvas ROC

Cálculo de los valores *False Positive Rate* y *True Positive Rate* de la curva ROC*:*

- Se fija un umbral para binarizar
- Se obtiene la matriz de confusión y, con ella, el valor de FPR y TPR
- Se grafica el punto
- Se repite el proceso con otro valor del umbral (en orden creciente, ej: 0.1, 0.2,..., 0.9, 1)
- Finalmente se unen todos los puntos

Comparación de modelos: Curvas ROC

- **¿Mejor solución?** (*FPR=0, TPR=1*) –> no hay errores en la clasificación
- Curva ROC sirve para elegir el mejor umbral (el más cercano al punto ideal (0,1))
- También sirve para elegir en qué modo de funcionamiento ajustar nuestro modelo
  - Modelo con muy alta recall (*TPR*), sacrificando la *FPR* (baja especificidad)
  - Modelo con baja recall y alta especificidad
  - Deseable: recall y especificidad altos

Comparación de modelos: AUC

- **Área bajo la curva (AUC, "Area Under Curve"):** medida resumen de la curva ROC
  - $AUC \approx 1$ –> mejor modelo
  - $AUC \approx 0.5$ –> predicción cercana al azar

Referencias

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112, p. 18). New York: springer.

Murphy, K. P. (2012). *Machine learning: a probabilistic perspective*. MIT press.

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction* (Vol. 2, pp. 1-758). New York: springer.

Aprendizaje supervisado

Aprendizaje Automático 1 - Grado en Ciencia e Ingeniería de Datos

Curso académico 2023-2024

Introducción

- Se conoce la variable objetivo que se desea predecir o clasificar
- **Variable objetivo**: contiene la información que se quiere explicar o entender en base al resto de las variables del conjunto de datos
- **Objetivo**: desarrollar modelos que capturen patrones y relaciones entre las variables con el fin de realizar predicciones precisas
  - **Clasificación**. Variable objetivo es categórica (problemas binarios o multiclase). Asignar observaciones a las diferentes categorías o clases
    - Ej: Predecir si un mail es span o no, si un paciente tiene una enfermedad
  - **Regresión**. Variable objetivo es continua. Predicciones numéricas.
    - Ej: Precio de una casa, demanda de productos

Algunos modelos

- Modelo lineal: Análisis Discriminante Lineal
- k-vecinos más próximos
- Árboles de decisión
- Métodos ensamblados (Random Forest, Bagging, Boosting)
- Naïve Bayes
- Modelos de mezcla de Gaussianas

Análisis Discriminante Lineal (ADL)

- Algoritmo de **clasificación**
  - Target: variable categórica
  - Variables explicativas continuas: distribución Normal
- **Objetivo**: Construir la combinación lineal de las variables explicativas que mejor discrimina las clases. En base a dicho hiperplano separador se clasificarán las nuevas observaciones

Análisis Discriminante Lineal (ADL)

- Teorema de Bayes para clasificación:

$$P(C = 1|\mathbf{X} = \mathbf{x}) = \frac{f_1(\mathbf{x})P(1)}{\sum_{c=1}^{C_n} f_c(\mathbf{x})P(c)}$$

siendo $P(C = 1|\mathbf{X} = \mathbf{x})$ la probabilidad a posteriori, es decir, la probabilidad de la clase 1 dada la observación $\mathbf{x}$. $P(c)$ es la probabilidad a priori de la clase c. Nótese que $\sum_{c=1}^{C_n} P(c) = 1$. Finalmente $f_c(\mathbf{x})$ es la función de densidad condicional de las $\mathbf{X}$ en la clase c

- En el caso del ADL, se asume que la densidad de cada clase es una Normal multivariante

$$f_c(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\Sigma_k|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\mu_k)^t \Sigma_k^{-1}(\mathbf{x}-\mu_k)}$$

- También se asume que las clases comparten la misma matriz de varianzas covarianzas $\Sigma_c = \Sigma,\ \forall c$

En base a estas asunciones, comparemos las probabilidades de ambas clases:

$$log\left(\frac{P(C=1|\mathbf{X}=\mathbf{x})}{P(C=0|\mathbf{X}=\mathbf{x})}\right) = log\left(\frac{f_1(\mathbf{x})}{f_0(\mathbf{x})}\right) + log\left(\frac{P(1)}{P(0)}\right) =$$

$$= log\left(\frac{P(1)}{P(0)}\right) - \frac{1}{2}(\mu_1+\mu_0)^t\Sigma^{-1}(\mu_1-\mu_0) + \mathbf{x}^t\Sigma^{-1}(\mu_1-\mu_0)$$

Obtenemos una ecuación que es lineal en $\mathbf{x}$. Es decir, la frontera de decisión entre las clases 0 y 1 es una ecuación lineal, un hiperplano en dimensión p

En la práctica, se desconocen los parámetros de la distribución Normal. Se estiman con los datos de entrenamiento:

- $\widehat{P(c)} = n_c/n$, siendo n_c el tamaño de la clase c y n el total
- $\widehat{\mu}_c$ media muestral de los elementos de la clase c
- $\widehat{\Sigma}$ varianza muestral de los elementos de la clase c

Análisis Discriminante Lineal (ADL)

Clasificación: ADL clasificará una obervación como de la clase 1 si

$$\mathbf{x}^t \widehat{\Sigma}^{-1}(\widehat{\mu}_1 - \widehat{\mu}_0) > \frac{1}{2}(\widehat{\mu}_1 + \widehat{\mu}_0)^t \widehat{\Sigma}^{-1}(\widehat{\mu}_1 - \widehat{\mu}_0) - log(n_1/n_0)$$

Si llamamos $\mathbf{w} = \widehat{\Sigma}^{-1}(\widehat{\mu}_1 - \widehat{\mu}_0)$, podemos reescribir lo anterior como

$$\mathbf{x}^t \mathbf{w} > \frac{1}{2}(\widehat{\mu}_1 + \widehat{\mu}_0)^t \mathbf{w} - log(n_1/n_0)$$

La frontera de decisión entre ambas clases es

$$\mathbf{x}^t \mathbf{w} = \frac{1}{2}(\widehat{\mu}_1 + \widehat{\mu}_0)^t \mathbf{w} - log(n_1/n_0)$$

que es una combinación lineal de las variables explicativas

Interpretación: Proyectar el punto a clasificar y las medias de las clases en una recta y asignar la observación a la clase con media más cercana

Análisis Discriminante Lineal (ADL)

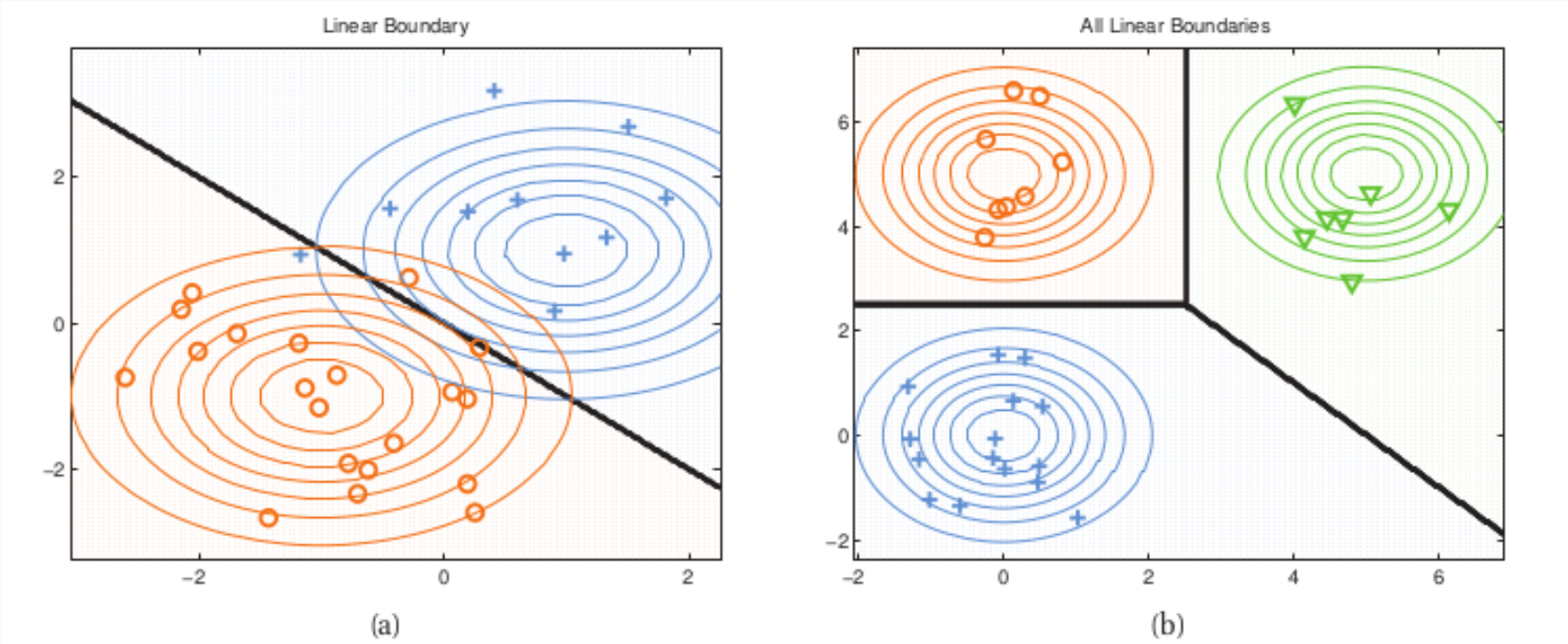


Figure 4.5 Linear decision boundaries in 2D for the 2 and 3 class case. Figure generated by `discrimAnalysisDboundariesDemo`.

Figure 1: Murphy, K. P. (2012). Machine learning: a probabilistic perspective. MIT press.

Análisis Discriminante Lineal: ventajas y desventajas

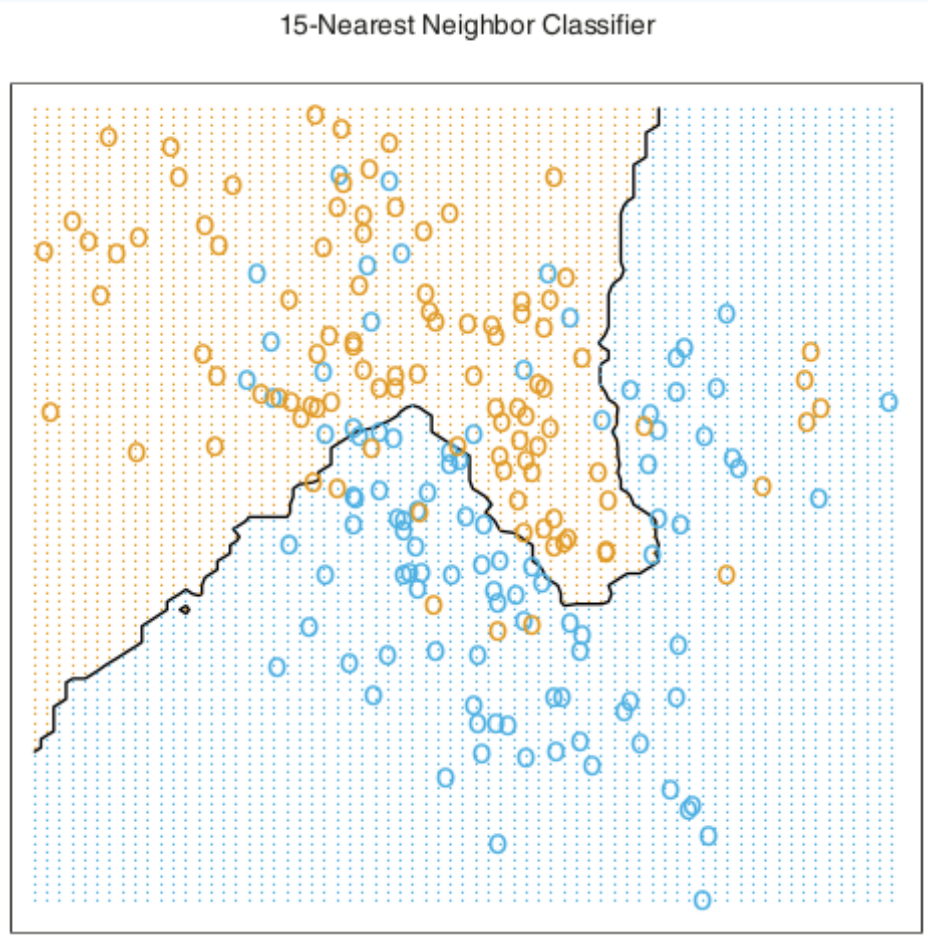
- **Ventajas**
  - Simple y rápido
  - Eficiente cuando se cumple las hipótesis
  - Clasificación de observaciones en grupos determinados –> facilita la interpretación
  - Combinación de información para la frontera de decisión
- **Desventajas**
  - Asume normalidad e igualdad de varianzas
  - Sensible a outliers
  - Es un clasificador lineal
  - Requiere cierto tamaño de muestra

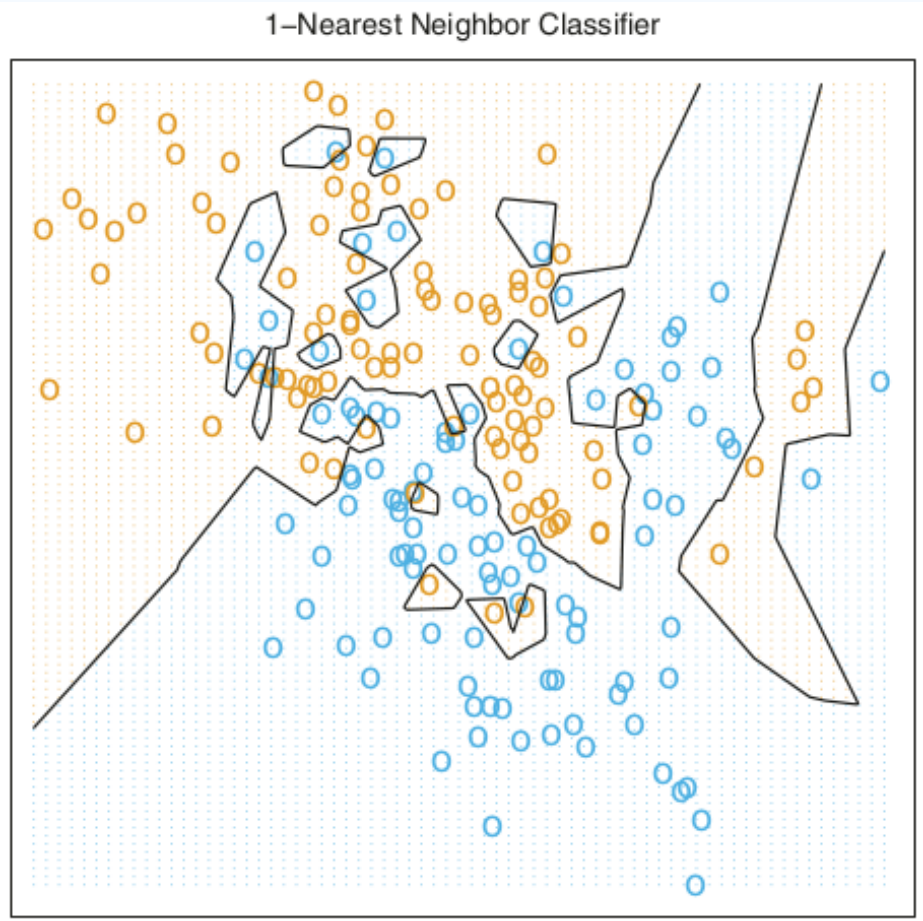
k-vecinos más próximos (k-nn)

- k-nn: k nearest neighbors
- Se basa en la noción de similitud o distancia entre individuos, en la idea de que observaciones similares se encuentran próximas
- Desafíos:
  - Métrica de similaridad utilizada para evaluar el parecido entre observaciones
  - Noción de cercanía: ¿Qué k elegir? –> Podemos usar técnicas como validación cruzada y grid search
- Funcionamiento:
  - Clasificación: devuelve la clase predominante entre los k vecinos (la moda)
  - Regresión: devuelve la media de la variable respuesta en los k vecinos

- **Ventajas**
  - Sencillo y de fácil implementación
  - No asume una distribución específica sobre los datos
  - Adaptabilidad a datos cambiantes
  - Interpretabilidad
  - Robusto frente al ruido
  - Sirve para regresión y clasificación (tanto binaria como multiclase)
- **Desventajas**
  - Sensible a la elección de k y de la métrica
  - Coste computacional en alta dimensionalidad
  - Datos desbalanceados: sesgo hacia la clase mayoritaria

k-vecinos más próximos (k-nn)





Origen de las imágenes: Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction* (Vol. 2, pp. 1-758). New York: springer.

Árbol de decisión

- Algoritmos de ML basados en la información
- Determinan qué variables explicativas proporcionan la mayor **ganancia de información** para medir la variable objetivo
- Proceso de particionamiento en el conjunto de variables explicativas
  - **Nodo raíz**: Nodo origen, todas las observaciones forman parte
  - **Nodos internos**: Nodos que se crean al definir reglas sobre una variable explicativa
  - **Nodos hojas**: Nodos terminales del árbol (donde se realiza la clasificación)

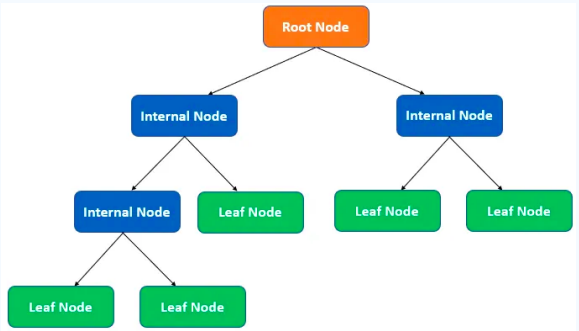


Figure 2: https://iprathore71.medium.com/complete-guide-to-decision-tree-cee0238128d

Árbol de decisión

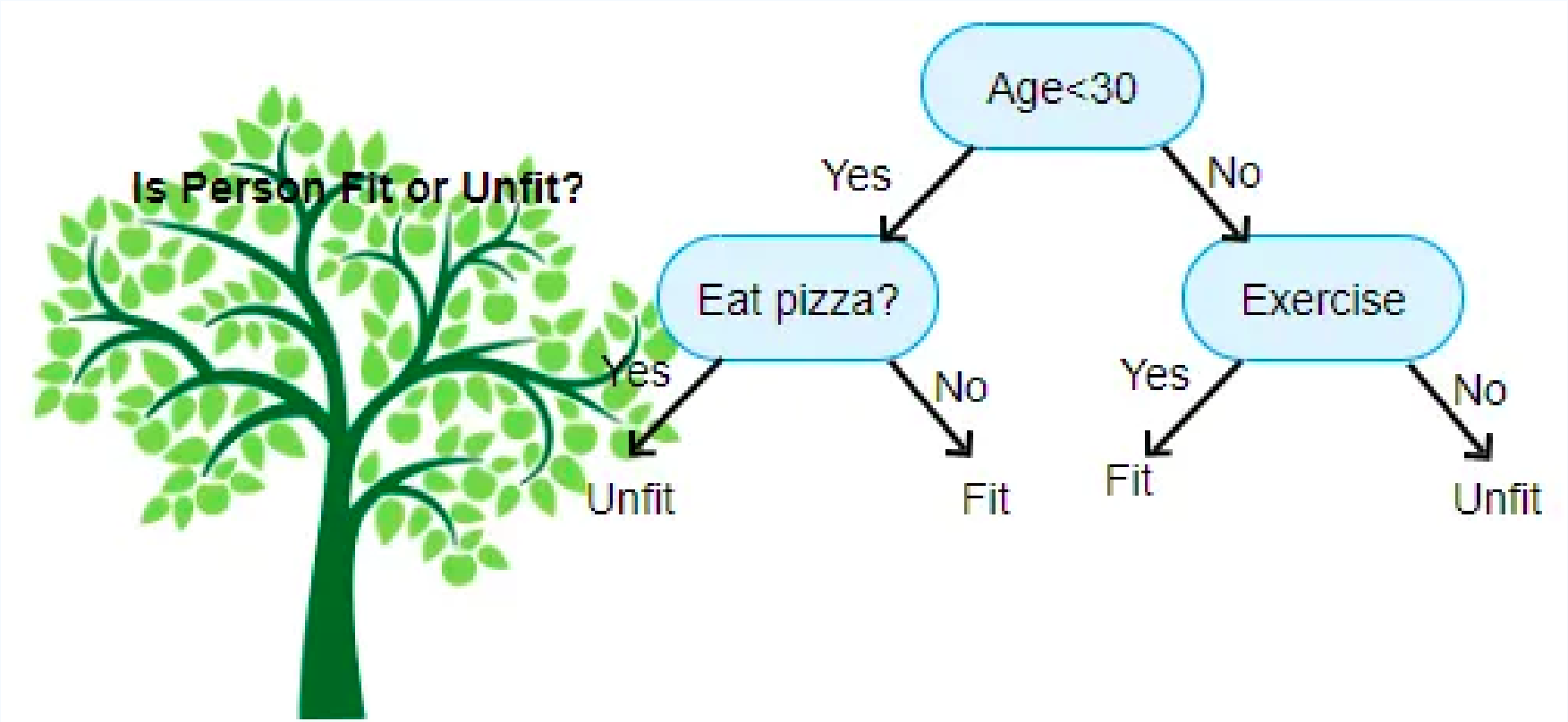


Figure 3:
https://iprathore71.medium.com/complete-guide-to-decision-tree-cee0238128d

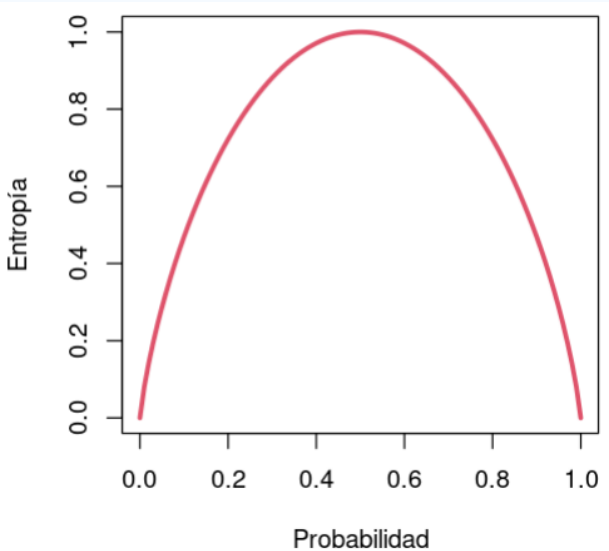
Árbol de decisión

- Las variables que más discriminen las clases están en la parte superior del árbol
- ¿Cómo evaluar esta discriminación? –> Métricas
  - Entropía
  - Índice de Gini

Árbol de decisión: Entropía

- La **entropía** es una medida teórica de la "*incertidumbre*" contenida en un conjunto de datos
- La **entropía** de un conjunto de n valores distintos, igualmente probables, es el menor número de preguntas de *sí/no* necesarias para determinar un valor desconocido extraído de las posibilidades:

$$Entropa = -\sum_{c=1}^{C} p_c \log_2(p_c)$$



Árbol de decisión: Entropía

Procedimiento del árbol usando la entropía

1. Calcular la **entropía** del conjunto original de datos
2. Para cada variable explicativa, se crean los conjuntos resultantes **dividiendo las observaciones** en el conjunto de datos utilizando un umbral para dicha variable. Se calcula la entropía de cada nodo individual de división y la media ponderada de todos los nodos hijos disponibles en una división
3. Calcular la **Ganancia de Información** restando la entropía restante (paso 2) del valor de entropía original (paso 1)

Para cada nodo en el árbol, se elige para la división aquella variable que maximiza la ganancia de información

Árbol de decisión: Índice de Gini

- Mide la pureza de los nodos
- El índice de Gini es el error esperado. p_c es la probabilidad de que una entrada cualquiera de la hoja pertenezca a la clase c y $(1 - p_c)$ es la probabilidad de que sea erróneamente clasificada:

$$Gini = \sum_{c=1}^{C} p_c(1 - p_c) = 1 - \sum_{c=1}^{C} p_c^2$$

  siendo C es el total de clases
- Comportamiento similar a la entropía
- Buscamos que la impureza de los nodos vaya disminuyendo –> índice de Gini bajo

Árbol de regresión: reducción de la varianza

- Reducción de la varianza: equivalente de la entropía o el índice de Gini para árboles de regresión
- Se utiliza la varianza para encontrar la mejor división
- Se calcula la varianza de cada división como la media ponderada de la varianza de cada nodo resultado de dicha división

Árbol de decisión: criterio de parada

- Error 0 de clasificación –> ¿Qué implicaría esto?
- Profundidad específica
- Número mínimo de observaciones en el nodo
- Cantidad de mejora

Árbol de decisión: ventajas

- Explicabilidad
- Compatibilidad con datos mixtos
- No requiere escalado de variables
- Manejo de datos faltantes
- Robustez ante valores atípicos
- Captura de no linealidades
- Eficiencia computacional

Árbol de decisión: desventajas

- Sensibilidad a cambios en los datos. Pequeños cambios en los datos pueden implicar un árbol diferente
- Naturaleza jerárquica –> propagación del error
- Propensión al sobreajuste
- Dificultad para modelar relacionales lineales
- No son óptimos para datos de alta dimensión
- Difícil interpretación cuando los árboles son muy profundos

Métodos de ensamblado

- **La unión hace la fuerza**: se basa en la idea de que la unión de múltiples modelos puede mejorar significativamente el rendimiento de predicción en comparación con un solo modelo
- Interpetación del Machine Learning de la **sabiduría colectiva**
- Clave: **diversidad** entre los distintos modelos (entrenando con distintas muestras del conjunto de datos o con distintos conjuntos de variables)
- Los más famosos:
  - Bagging
  - Boosting
  - Random Forest

Bagging

- **Bagging (Bootstrap Aggregating)**: técnica diseñada para mejorar la precisión y estabilidad de los modelos predictivos. Combina una serie de modelos "débiles"
- **Algoritmo**:
  1. **Bootstrap**: Se obtienen m muestras con reemplazamiento de tamaño n (tamaño real del conjunto de datos). Estas serán las muestras de entrenamiento
  2. **Modelo base**: Se entrenan m modelos (partiendo de un modelo base como un DT) usando las m muestras de entrenamiento
  3. **Predicciones**: Se obtienen un total de m predicciones del conjunto de test, 1 por cada modelo
  4. **Combinación**: Se combinan las predicciones de los m modelos para dar una predicción final
     - Clasificación: regla del voto mayoritario
     - Regresión: media

- **Ventajas**
  - Reducción de la varianza. Más robusto y menos propenso al sobreajuste
  - Mayor precisión
  - Estabilidad
  - Mayor capacidad de generalización
- **Desventajas**
  - Mayor complejidad computacional
  - Menor interpretabilidad
  - No garantiza la mejora
  - Menos efectivo con modelos base inestables

Random Forest

- Bosque formado por múltiples Árboles de Decisión
- Los distintos árboles se entrenan con un subconjunto aleatorio de observaciones (Bagging) y también un subconjunto aleatorio de variables –> Diversidad
- Cada árbol individual no es potente, pero la combinación de todos, que han aprendido cosas distintas, sí resulta potente
- Ofrece información sobre la importancia de las variables
- ¿Ventajas y desventajas?

Boosting

- Se centra en mejorar iterativamente un modelo "débil"
- **Algoritmo**
  1. Se entrena un modelo base en el conjunto de datos original
  2. Se evalúa su rendimiento. Se da más peso a las observaciones clasificadas erróneamente
  3. Se entrena otro modelo débil usando los pesos del paso previo
  4. Se combina la predicción de los modelos base ponderando sus predicciones en función de su rendimiento. Los mejores modelos tendrán más peso en la predicción final.
- Cada modelo nuevo se centra en corregir las deficiencias del previo
- La combinación final es un ensamblado fuerte
- Algoritmos Boosting populares: AdaBoost, Gradient Boosting, XGBoost

- **Ventajas**
  - Mejora del rendimiento
  - Reducción del sesgo
  - Gestión de datos desequilibrados
  - Capacidad para detectar patrones complejos
- **Desventajas**
  - Coste computacional
  - Tiempo de entrenamiento (secuencial)
  - Menor interpretabilidad
  - Ajuste de hiperparámetros

Bagging vs Boosting

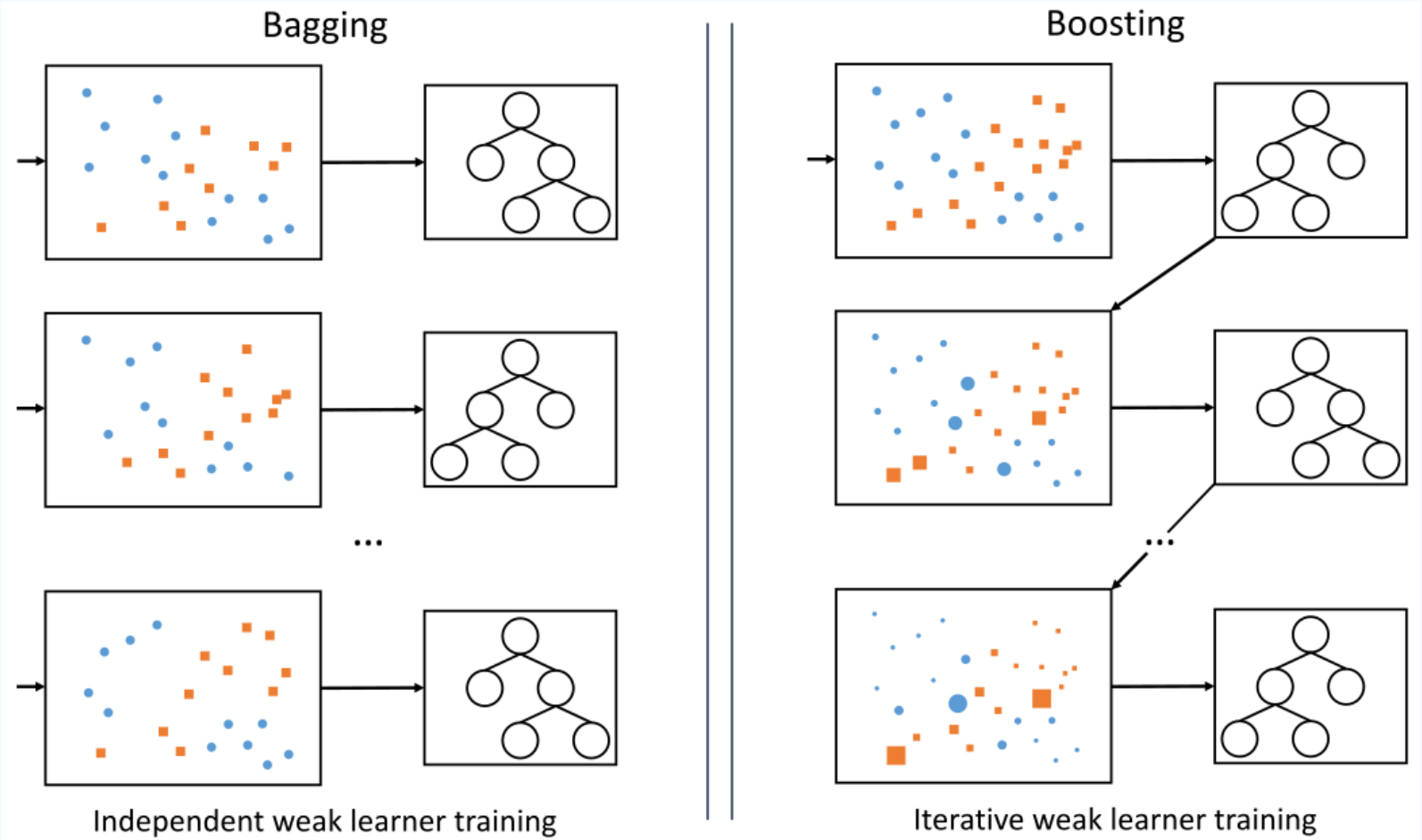


Figure 4: Imagen de González, S., García, S., Del Ser, J., Rokach, L., & Herrera, F. (2020). A practical tutorial on bagging and boosting based ensembles for machine learning: Algorithms, software tools, performance study, practical perspectives and opportunities. Information Fusion, 64, 205-237.

- **Clasificador ingenuo de Bayes** (en inglés "*Naïve Bayes*") es un clasificador sencillo que se basa en el conocido **teorema de Bayes**
- Teorema de Bayes: relaciona la probabilidad condicional de dos eventos A y B

$$P(A \cap B) = P(A, B) = P(A)P(B|A) = P(B)P(A|B) \Rightarrow$$

$$\Rightarrow P(B|A) = \frac{P(B)P(A|B)}{P(A)}$$

Aplicación en un problema de clasificación:

Supongamos $X_1, X_2, \ldots, X_n$ variables explicativas independientes dado el valor de la variable objetivo Y_k. Es decir,

$$P(X_1|X_2, \ldots, X_n, Y_k) = P(X_1|Y_k); \quad P(X_2|X_3, \ldots, X_n, Y_k) = P(X_2|Y_k); \quad .$$

Aplicando el teorema de Bayes recursivamente

$$P(X_1, X_2, \ldots, X_n, Y_k) = P(Y_k)P(X_n|Y_k)P(X_{n-1}|Y_k)\cdots P(X_1|Y_k)$$

Por tanto

$$P(Y_k|X_1, \ldots, X_n) = \frac{P(Y_k)\prod_{j=1}^{n} P(X_j|Y_k)}{P(X_1, X_2, \ldots, X_n)}$$

Aplicación en un problema de clasificación:

$$P(Y_k|X_1,\ldots,X_n) = \frac{P(Y_k)\prod_{j=1}^{n} P(X_j|Y_k)}{P(X_1,X_2,\ldots,X_n)}$$

- Denominador constante
- La probabilidad de clase $P(Y_k)$: probabilidad a priori
- Verosimilitud:

$$\prod_{j=1}^{n} P(X_j|Y_k)$$

  Mide cómo de verosímiles son las observaciones de las variables explicativas dado el valor de la respuesta Y_k
  - Variable cuantitativa: se asume distribución Normal
  - Variable cualitativa: distribución multinomial
- Se predice la clase con mayor probabilidad de clase condicionada $P(Y_k|X_1,\ldots,X_n)$

Mezcla de Gaussianas

- **GMM: Gaussian Mixture Models**
- Se basan en la idea de que un conjunto de datos está compuesto por varias distribuciones gaussianas (normales)
- GMM permiten descomponer un conjunto de datos en múltiples componentes gaussianas, cada una de las cuales describe una parte de la distribución de los datos
- Cada componente gaussiana representa una subdistribución de datos y se caracteriza por su media (promedio) y desviación estándar (dispersión)
- Estimación de los parámetros: algoritmos de optimización para maximizar la probabilidad conjunta de los datos observados
- Cada observación se asigna a la gaussiana a la que pertenece con mayor probabilidad
- Sensibles al número de componentes y la inicialización

Referencias

González, S., García, S., Del Ser, J., Rokach, L., & Herrera, F. (2020). A practical tutorial on bagging and boosting based ensembles for machine learning: Algorithms, software tools, performance study, practical perspectives and opportunities. *Information Fusion*, *64*, 205-237.

Murphy, K. P. (2012). *Machine learning: a probabilistic perspective*. MIT press.

Kelleher, John D, y Brendan Tierney. 2018. *Data science*. MIT Press.

Zhou, Z. H. (2012). *Ensemble methods: foundations and algorithms*. CRC press.

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction* (Vol. 2, pp. 1-758). New York: springer.

Reglas de asociación

Aprendizaje Automático 1 - Grado en Ciencia e Ingeniería de Datos

Curso académico 2023-2024

Reglas de asociación

Las reglas de asociación son un enfoque de ML diseñado para identificar patrones de co-ocurrencia entre elementos o características en conjuntos de datos. Estas reglas permiten revelar conexiones interesantes y relaciones significativas que a menudo pasan desapercibidas a simple vista.

Reglas de asociación

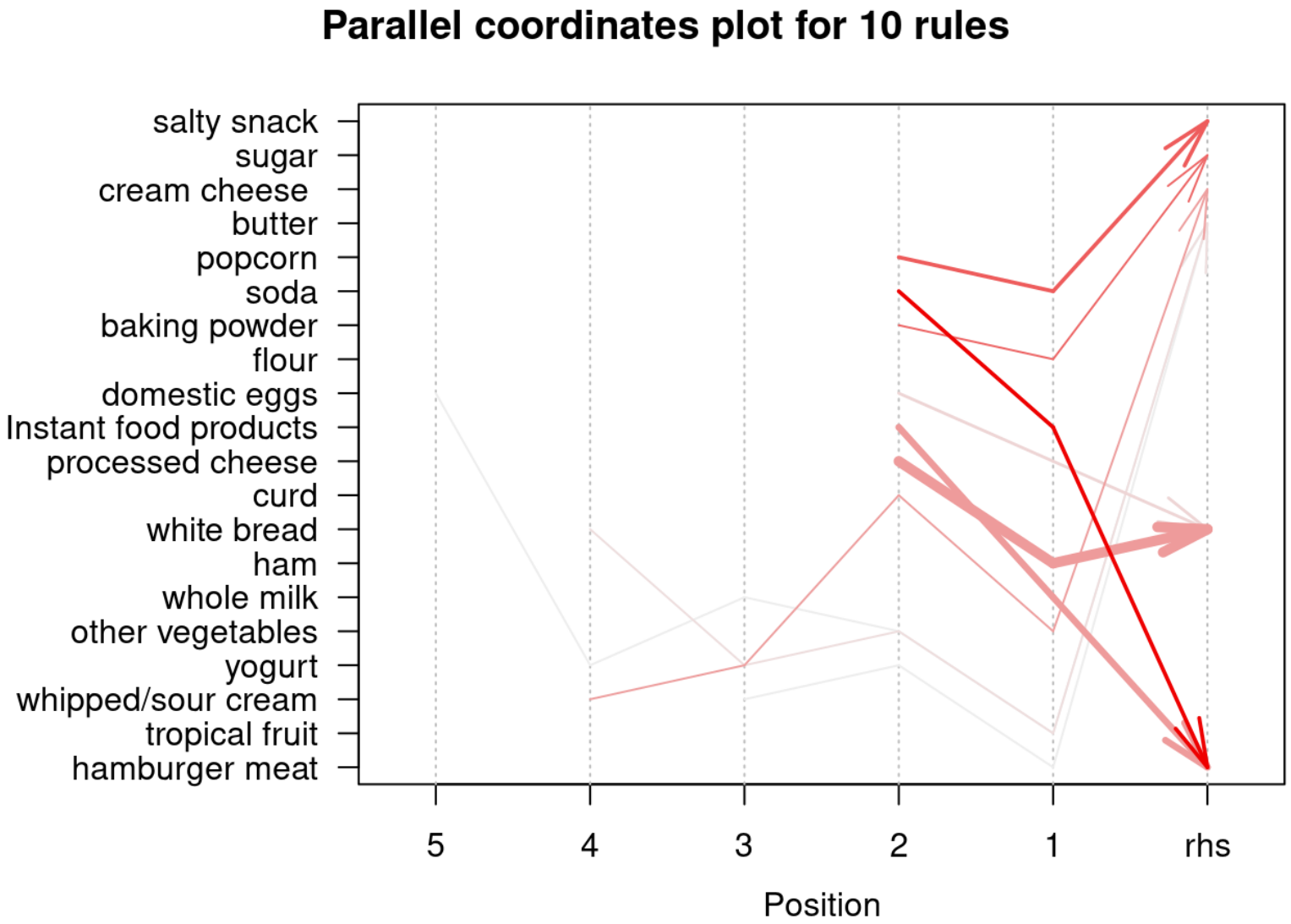
Supongamos que disponemos de un conjunto de datos de compras de clientes en una tienda en línea.

Transacción	Productos comprados
1	Pan, Leche, Huevos
2	Leche, Queso, Yogur
3	Pan, Leche, Mantequilla
4	Pan, Huevos
5	Leche, Huevos, Queso

Cada fila representa una transacción y los productos comprados en esa transacción. Las reglas de asociación se utilizan para descubrir patrones de co-ocurrencia entre productos. En el ejemplo: *Si un cliente compra Pan y Leche, entonces también compra Huevos. Si un cliente compra Leche y Queso, entonces también compra Yogur*.

Reglas de asociación

- **"Si...Entonces..."**: relación entre los elementos o características.
- **Fuerza en la asociación**: el soporte mide la frecuencia con la que aparece una asociación en el conjunto de datos (ocurrencia). La confianza mide la probabilidad de que se cumpla una regla dada.
- **Algoritmos**: Apriori, FP-Growth, entre otros.


Parallel coordinates plot for 10 rules
salty snack
sugar
cream cheese
butter
popcorn
soda
baking powder
flour
domestic eggs
Instant food products
processed cheese
curd
white bread
ham
whole milk
other vegetables
yogurt
whipped/sour cream
tropical fruit
hamburger meat
5
4
3
2
1
rhs
Position

Referencias

Piatetsky-Shapiro, G. (1991). Discovery, analysis, and presentation of strong rules. *Knowledge Discovery in Data-bases*, 229-248.

Pang-Ning, T., Steinbach, M., & Kumar, V. (2005). Introduction to data mining Addison-Wesley. *Princeton, USA: Independence Way*, *4*, 76-79.

Nuevas Tendencias

Aprendizaje Automático 1 - Grado en Ciencia e Ingeniería de Datos

Curso académico 2023-2024

Ética en el Aprendizaje Automático

- La toma de decisiones automatizada basada en datos plantea desafíos éticos y sociales que deben ser abordados de manera responsable

Ética

- Disciplina filosófica que estudia el bien y el mal y sus relaciones con la moral y el comportamiento humano.
- Conjunto de costumbres y normas que dirigen o valoran el comportamiento humano en una comunidad.
- La ética en el Aprendizaje Automático se centra en garantizar que los sistemas de IA tomen decisiones justas, imparciales y éticas.

Aprendizaje Automático Explicable

El aprendizaje máquina explicable (**XML**, "Explainable Machine Learning" en inglés) se refiere a la capacidad de los modelos de ML para proporcionar explicaciones claras y comprensibles de sus decisiones.

Métodos basados en reglas

- **Reglas de decisión**: Estos modelos generan reglas lógicas que explican el razonamiento detrás de las predicciones del modelo.
- **Árboles de decisión**: Los árboles muestran la secuencia de decisiones tomadas por el modelo en cada nodo, lo que facilita la interpretación.

Métodos basados en características

- **Importancia de características**: Calcula la importancia de cada característica en el modelo, lo que permite identificar las variables más influyentes en las predicciones.
- **Análisis de efectos parciales**: Evalúa el impacto de una sola característica en las predicciones, manteniendo las demás constantes.

Métodos basados en ejemplos

- **Prototipos**: Encuentra ejemplos representativos o prototipos de datos que explican cómo el modelo toma decisiones.
- **Casos de prueba**: Genera instancias que muestran cómo el modelo reacciona a diferentes escenarios.

Métodos de aproximación

- **Modelos locales interpretables**: Crea modelos más simples (lineales, regresiones, etc.) en regiones locales del espacio de características para comprender decisiones en áreas específicas.
- **Regresiones lineales localmente ponderadas (LWLR)**: Asigna pesos a las instancias cercanas para ajustar una regresión lineal local.

Aprendizaje Automático Explicable

Métodos basados en atención

- **Atención y atención saliente**: Modelos basados en atención destacan características o regiones de interés que influyen en las predicciones.
- **Redes neuronales con atención**: Las redes neuronales con mecanismos de atención permiten entender qué partes de la entrada son relevantes para la salida.

Métodos basados en métricas

- **Métricas de proximidad**: Evalúan la similitud entre entradas y cómo se relacionan con las predicciones.
- **SHAP (SHapley Additive exPlanations)**: Utiliza conceptos de teoría de juegos para asignar valores de importancia a las características.

Métodos basados en resumen

- **Reglas de decisión generadas por el modelo**: El modelo crea reglas que resumen su comportamiento.
- **Análisis de componentes**: Reduce la dimensión de los datos para visualizar y resumir características significativas.

Métodos de muestreo y generación de datos

- **Perturbación de datos**: Se modifican las características de entrada para entender cómo afectan a las predicciones.
- **Muestreo de datos contrapuestos**: Se generan ejemplos que muestran cómo las predicciones cambiarían si los datos fueran diferentes.

Métodos de comparación

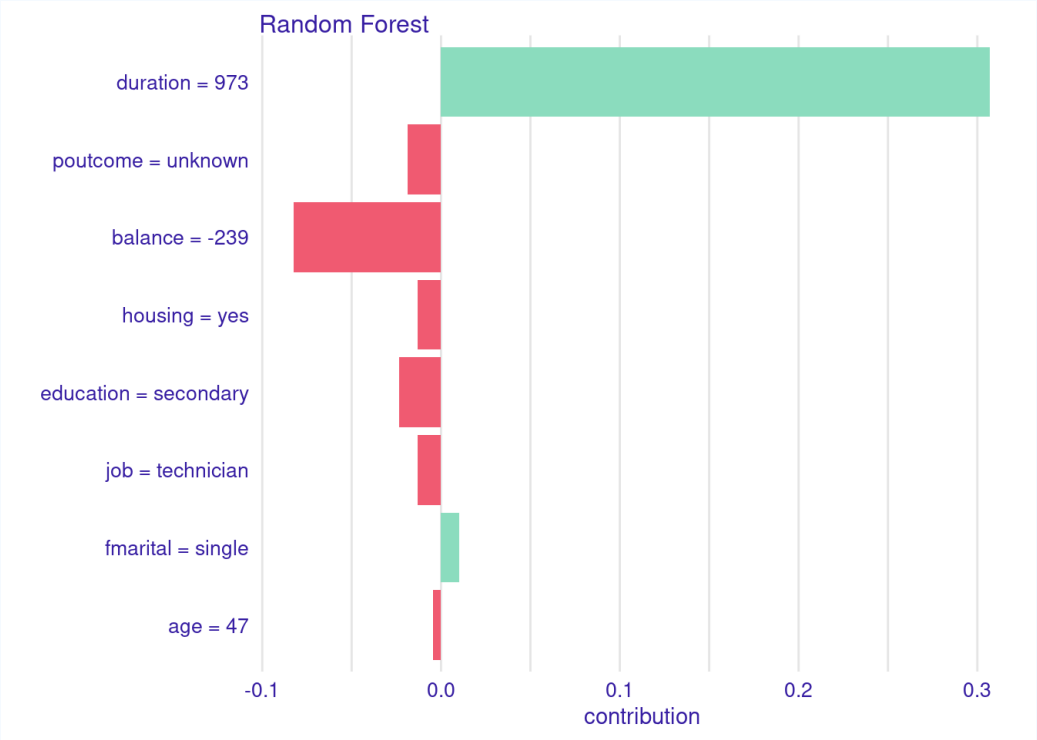
- **Modelos interpretables frente a modelos de caja negra**: Compara modelos interpretables con modelos complejos en términos de rendimiento y capacidad de explicación.

Herramientas de visualización

- **Gráficos interactivos**: Visualizaciones que muestran cómo las características afectan a las predicciones.
- **Heatmaps y perfiles de importancia**: Muestran la importancia de las características en un formato visual.

SHAP

SHAP se basa en el principio de que cada característica o atributo de entrada de un modelo contribuye de alguna manera a la predicción final. SHAP cuantifica esta contribución para cada característica, permitiendo una interpretación más profunda de cómo el modelo llega a sus conclusiones.



Contrafácticos

- Los contrafácticos son preguntas o declaraciones que plantean "¿Qué habría ocurrido si...?" con el propósito de analizar cómo un modelo habría respondido si las condiciones o las entradas hubieran sido diferentes.
- Esta técnica se utiliza para obtener información sobre cómo un modelo realiza sus predicciones y para explicar su razonamiento.

Deriva conceptual

La deriva conceptual es un fenómeno crítico que se refiere a la evolución de los datos y la cambiante naturaleza de las relaciones entre las variables a lo largo del tiempo.

Cambios

- en la **distribución de los datos**
- en la **importancia relativa de las características**
- en las **relaciones entre variables**
- **nuevos patrones** que antes no estaban presentes.

Deriva conceptual

Para abordar la deriva conceptual, es necesario implementar técnicas de adaptación de modelos que permitan a los algoritmos de ML ajustarse a los cambios en los datos.

- **reentrenamiento periódico** de los modelos
- **monitorización constante de la calidad** del modelo
- **identificación de los momentos** en que se produce una deriva conceptual significativa.

Referencias

Molnar, C. (2020). *Interpretable machine learning*. Lulu. com.

Lu, J., Liu, A., Dong, F., Gu, F., Gama, J., & Zhang, G. (2018). Learning under concept drift: A review. *IEEE transactions on knowledge and data engineering*, *31*(12), 2346-2363.