

Tratamiento Estadístico de Datos

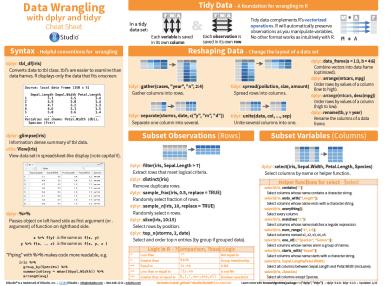


Preprocesamiento

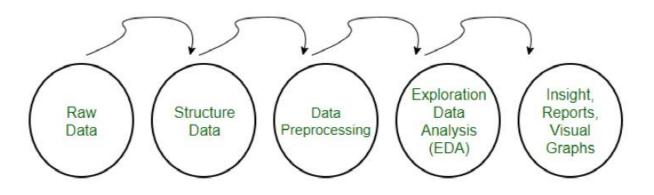
→ Métodos de Remuestreo → Evaluación de las Métricas

Preprocesamiento

Analizaremos las diferentes técnicas que tenemos a nuestra disposición para realizar un buen preprocesamiento de datos y, así, mejorar el rendimiento de los modelos que generaremos posteriormente. Se debe tener en cuenta si en el modelado supervisado realizado el problema es de regresión o clasificación.



Data Wrangling



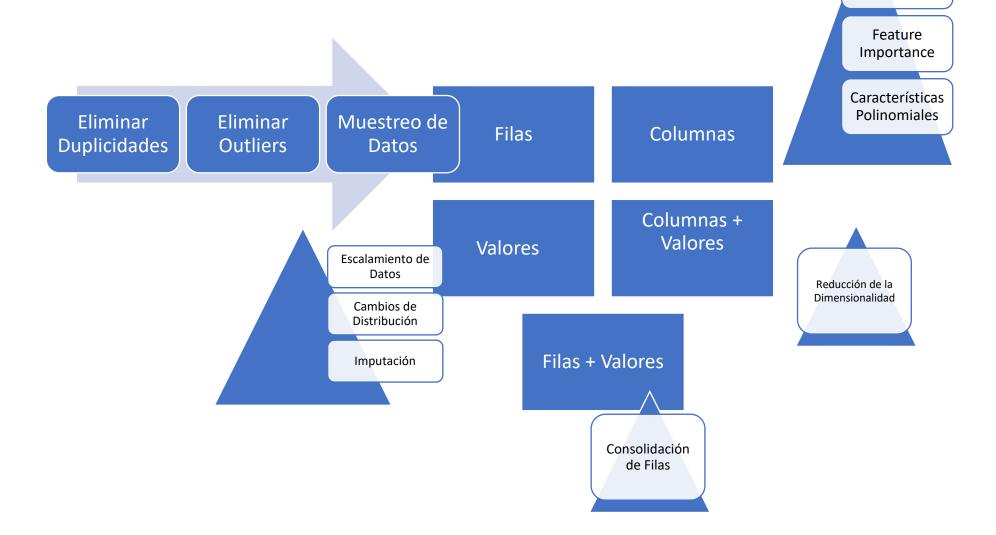


Preprocesamiento



Feature

Selection





APRENDIZAJE SUPERVISADO



Permite a las organizaciones comprender y prevenir los resultados no deseados o impulsar los resultados deseados para lo que sea que estén tratando de predecir

Los modelos deben ser reconstruidos periódicamente con el fin de mantener sus predicciones sin que se conviertan en obsoletas

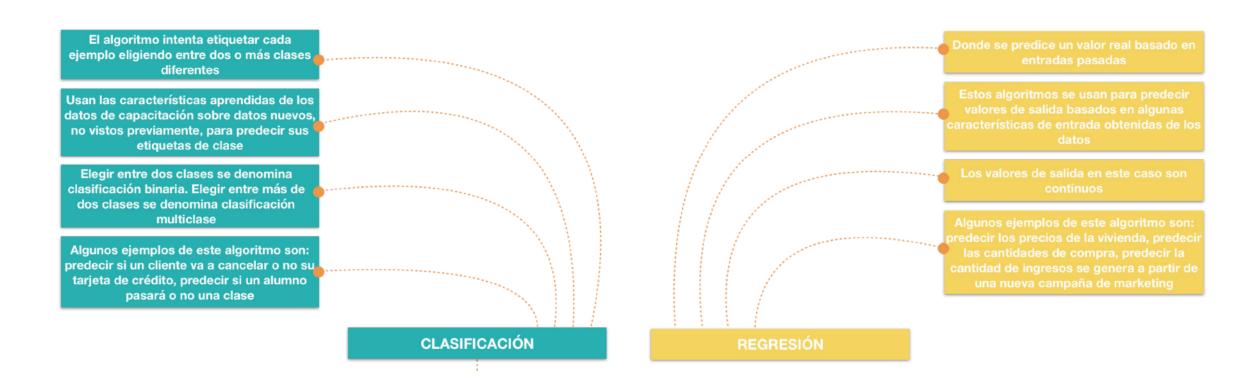
Proporciona una ruta directa para convertir datos en información real y procesable Este aprendizaje es uno de los motores más potentes que permite que los sistemas de inteligencia artificial tomen decisiones empresariales de forma más rápida y precisa que los humanos

APRENDIZAJE SUPERVISADO - IMPORTANCIA



PROCESO DE MACHINE LEARNING SINTETIZADO





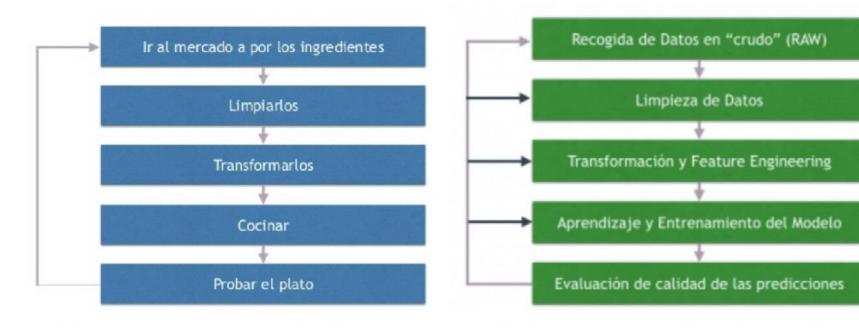






80%

20%



(a) Analogía con la vida diaria.

(b) Proceso completo de machine learning.





Necesidad de Preprocesamiento

• Los datos en bruto, sin analizar de antemano, muy difícilmente nos va a dar precisiones robustas a la hora de realizar modelos de machine learning ya que, entre otros aspectos, muchos de ellos requieren los datos en una forma específica, por lo que debemos realizar transformación del conjunto de datos. Por otro lado, otros algoritmos pueden funcionar mejor si los datos se preparan de una manera específica, por ejemplo, los algoritmos tipo árbol con atributos características tipo nominal. Por lo tanto, es de vital importancia realizar un preprocesamiento de nuestros datos como parte fundamental de un proyecto de Machine Learning.





Consideraciones a tomar

La aplicación de algoritmos de Machine Learning para la predicción es una *ciencia empírica*, a pesar de las guías generales que podemos tener, la única forma de saber que algoritmos usar para nuestro conjunto de datos es a través de nuestra experiencia y conocimiento conjunto de nuestros datos. Debemos probar una variedad de transformaciones de datos con un conjunto de algoritmos diferentes que nos ayudará a descubrir qué algoritmos pueden ser los **mejores candidatos** para un conjunto determinado de datos. A grande rasgos:

- Los métodos basados en instancias (filas) son más efectivos si los atributos de entrada tienen la misma escala.
- Los métodos de regresión pueden funcionar mejor si los atributos de entrada están estandarizados.



¿Por qué transformar?



En algunos casos al realizar el modelo predictivo de Machine Learning y al momento de cuantificar categorías el modelo le da *mayor relevancia a una que otra*, cuando tendría que haber ponderado a ambas por igual. Por ejemplo 3 variables analizadas (edad, salario que aspira el candidato, contratado). Donde Edad es la edad del postulante a una oferta de trabajo, el Salario al que aspira es lo que pide por realizar ese trabajo y donde contratado indica si el postulante fue contratado o no (Si-No). En ese caso la **edad** es de 2 dígitos y una categoría de **salario** posiblemente de 4, 5 o más dígitos. En otros casos la diferencia puede ser mayor, con lo que el modelo podría otorgar más importancia al salario que la edad por esa razón necesitamos realizar el **escalamiento/transformación** a nuestro conjunto de datos.

Cabe resaltar que si todas las características están dentro de un rango similar entre sí, entonces no hay necesidad real de transformar.



Transformación en R



Librería `caret` de R.

Técnicas principales para transformar los datos:

- * Autónomo: Las transformaciones se pueden modelar a partir de los datos de entrenamiento y se pueden aplicar a multiples conjunto de datos, se usará lo siguiente:
 - PreProcess(): Entrena los datos
 - predict() : Aplica la transformación
- * Entrenamiento: Las transformaciones se preparan y aplican automáticamente durante a evaluación del modelo, se usará lo siguiente:
 - preProcess() y dentro de preProces(...,.., train(),...,..).

Debemos considerar que lo que se va a realizar es para variables o datos numéricos, R omite los datos no numéricos sin error. Existe una probabilidad de que los algoritmos para métodos basados en instancias tales como regresión, métodos basados en instancias (como k-NN y LVQ), Support Vector Machine (SVM) y redes neuronales; y menos útiles para métodos basados en reglas y en árboles, aunque no olvidemos es **empírico**.

Lin para revisar información sobre caret: https://cran.r-project.org/web/packages/caret/caret.pdf,

Link para revisar información sobre Pre-Processing: http://topepo.github.io/caret/pre-processing.html



Métodos de Transformación



Veremos los más conocidos métodos de transformación admitidos dentro de la función 'preProcess()' de la librería 'caret' que se aplica a la data de entrenamiento.

Posibles valores "Box-Cox", "YeoJohnson", "expoTrans", "center", "scale", "range", "knnImpute", "bag-Impute", "medianImpute", "pca", "ica", "spatialSign", "corr", "zv", "nzv", y "conditionalX".

- BoxCox: aplica una transformada de Box-Cox, los valores deben ser distintos de cero y positivos.
- YeoJohnson: aplica una transformada de Yeo-Johnson, como un BoxCox, pero los valores pueden ser negativos o cero.
- expoTrans: aplica una transformada de potencia como BoxCox y YeoJohnson.
- zv: elimina los atributos con una varianza cero (todos con el mismo valor).
- nzv: elimina los atributos con una varianza cercana a cero (cerca del mismo valor).
- center: resta la media de los valores.
- scale: dividir los valores mediante la desviación estándar.
- range: normaliza los valores.
- pca: transformación de datos mediante componentes principales.
- ica: transformación de datos mediante componentes independientes.
- spatialSign: proyecta datos en un círculo unitario.



Escalamiento de Datos



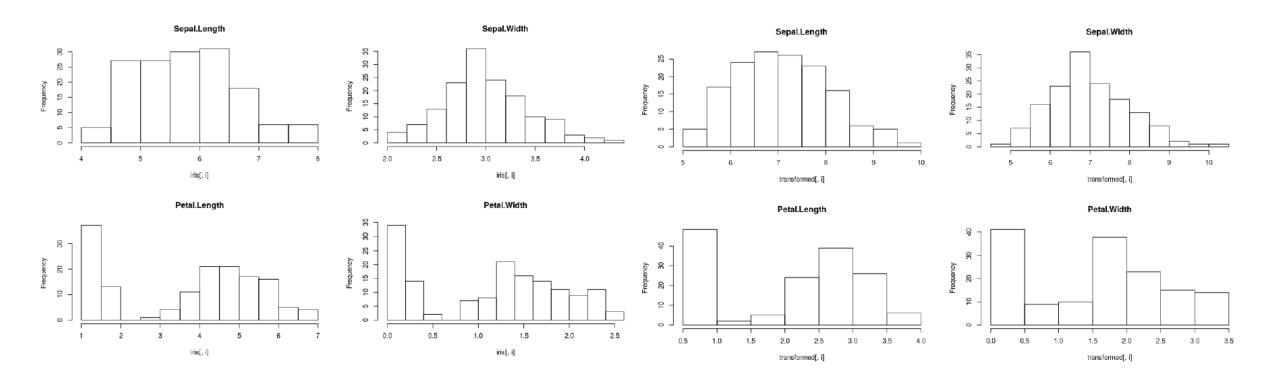
• La transformación `scale()` calcula la desviación estándar de un atributo y divide cada valor por esa desviación estándar. Este escalamiento es útil principalmente para los datos que provienen de una distribución gaussiana lo hace de manera consistente. Los algoritmos de optimización: *Gradient Descending*. Algoritmos ponderan entradas: *Regression* y *Neural Network (NN)*. Usan medidas de distancia: *K-Nearest Neighbor (K-NN)*.

```
library(lattice)
library(ggplot2)
library(caret)
data("iris")
summary(iris[,1:4])
preProcessParams <- preProcess(iris[,1:4], method = c("scale"))
print(preProcessParams)
transformed <- predict(preProcessParams, iris[,1:4])
summary(transformed)</pre>
```



Escalamiento de Datos (HISTOGRAMAS)

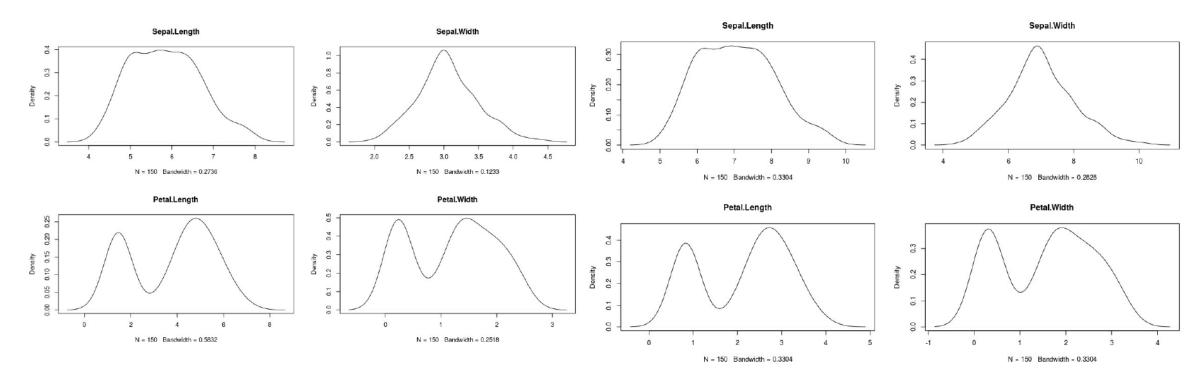






Escalamiento de Datos (DENSIDAD)







Centrado de Datos



• La transformación `center` calcula la media de un atributo y, posteriormente, se lo resta de cada valor.

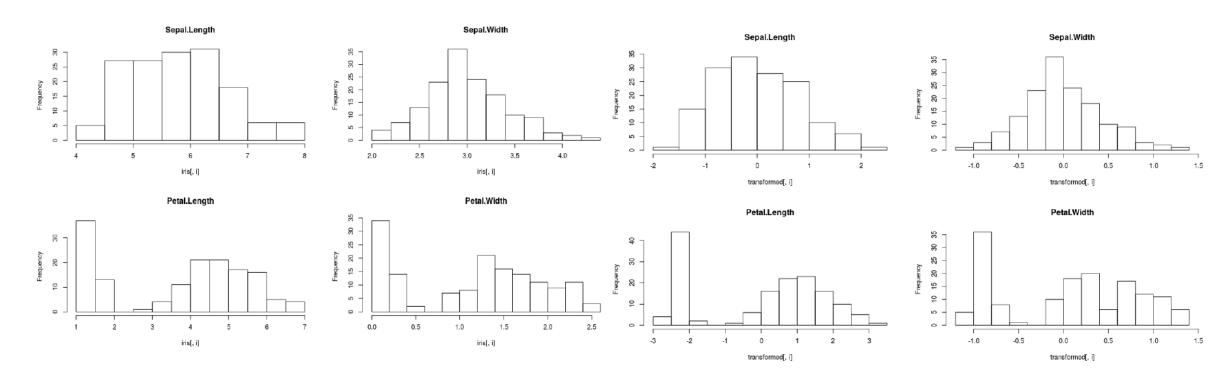
```
summary(iris[,1:4])

preProcessParams <- preProcess(iris[,1:4], method=c("center"))
print(preProcessParams)
transformed <- predict(preProcessParams, iris[,1:4])
summary(transformed)
View(transformed)</pre>
```



Centrado de Datos (HISTOGRAMAS)

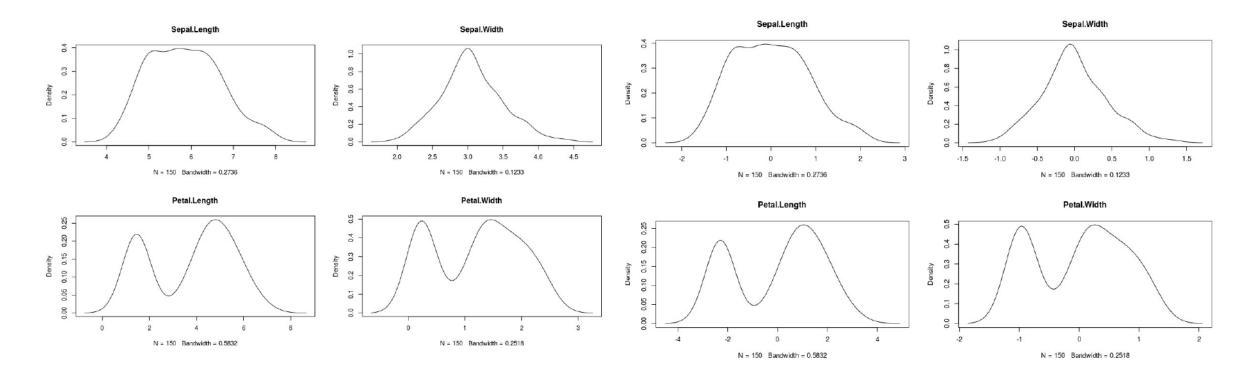






Centrado de Datos (DENSIDAD)







Estandarización de Datos



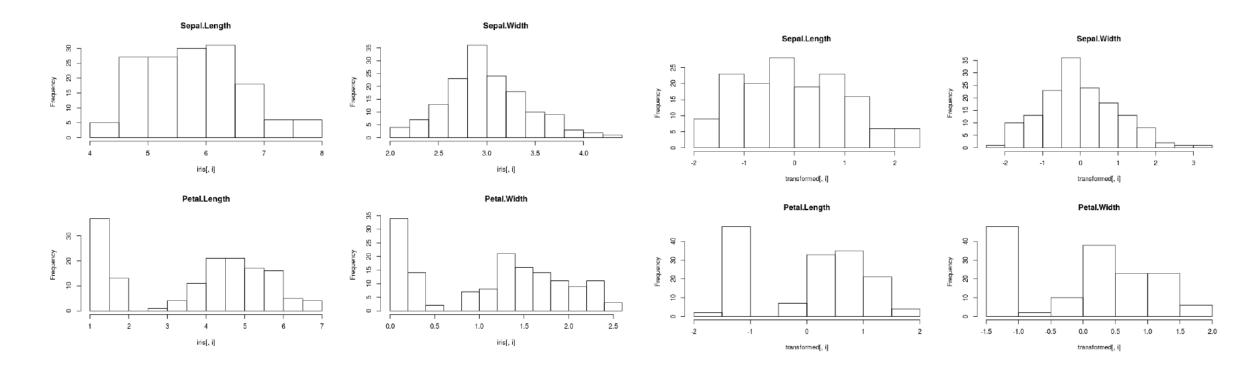
• La transformación `scale` y `center` realizan la estandarización de nuestros datos. En este caso, los atributos obtendrán un valor medio de 0 y una desviación estándar de 1. Para poder la realizar la estandarización de nuestros datos tenemos que especificar el parámetro `method` de la función `preProcess()` ambas funciones para la transformación `train()`. Es bueno para algoritmos Linear Regression (LiR) , Logistic Regression (LoR) y Linear Discriminant Analysis (LDA), entre otros.

```
summary(iris[,1:4])
summary(iris[,1:4])
preProcessParams <- preProcess(iris[,1:4], method=c("center", "scale"))
print(preProcessParams)
transformed <- predict(preProcessParams, iris[,1:4])
summary(transformed)</pre>
```



Estandarización de Datos (HISTOGRAMAS)

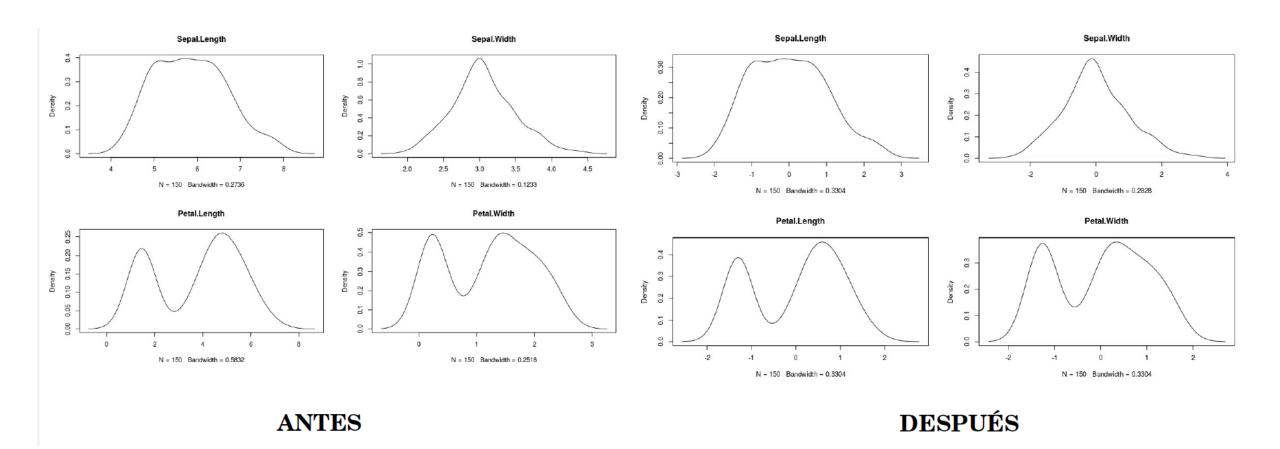






Estandarización de Datos (DENSIDAD)







Normalización de Datos



• Los valores de los datos se pueden escalar en el rango de [0,1] en otras palabras con un valor mínimo a 0 y el *valor máximo a 1.* A este proceso se le denomina normalización. En general la mayoría de algoritmos se beneficia más de una estandarización que de una normalización. Como ejemplo tenemos los análisis de agrupamiento, la normalización puede ser especialmente crucial al momento de comparar similitudes entre las características en base a ciertas medidas de distancia. Otro ejemplo es en el análisis de componentes principales, en el que generalmente preferimos la estandarización a la escala mínima, ya que estamos interesados en maximizar la varianza (depende de la pregunta y si la PCA calcula los componentes mediante la matriz de correlación en lugar de la matriz de covarianza). La escala Min-Max es útil para la aplicación popular del procesamiento de imágenes, donde la intensidad de los pixels deben normalizarse para ajustarse dentro de un cierto rango (es decir, de 0 a 255 para el rango de color RGB). Además los algoritmos de redes neuronales requieren datos en una escala de 0 a1. Útil en algoritmos NN y K-NN.



Normalización de Datos



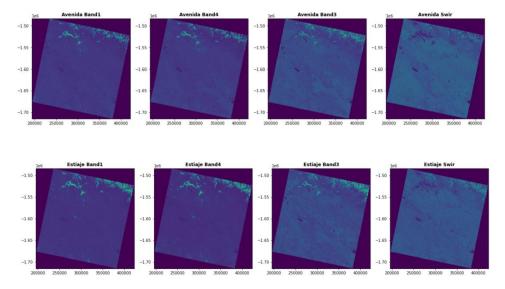


Figure 3: Bandas Obtenidas del Landsat8

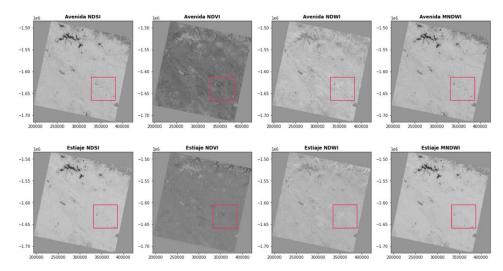


Figure 4: Índices Obtenidos del Procesamiento

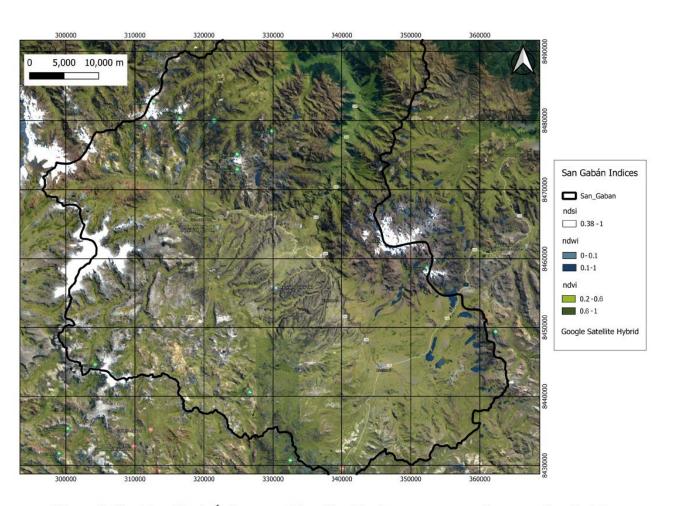


Figure 5: Combinación de Índices para Identificación de aguas, vegetación y nevado - Estiaje





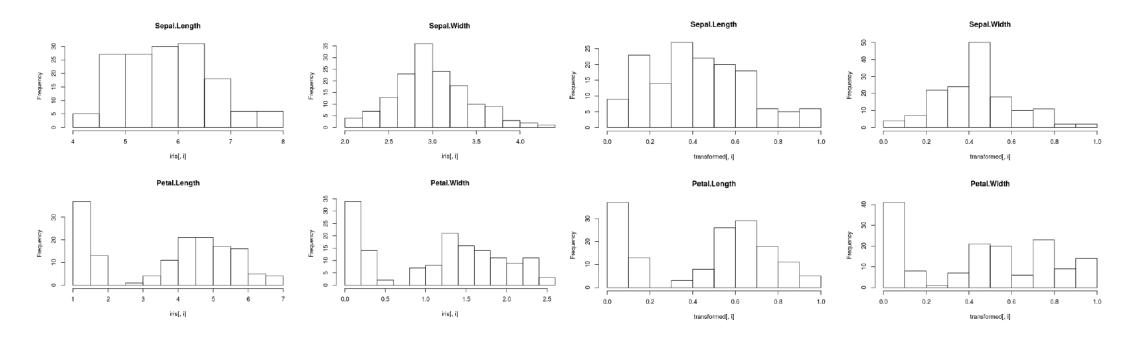
Normalización Código en R

```
summary(iris[,1:4])
preProcessParams <- preProcess(iris[,1:4], method=c("range"))
print(preProcessParams)
transformed <- predict(preProcessParams, iris[,1:4])
summary(transformed)</pre>
```



Normalización de Datos (HISTOGRAMAS)

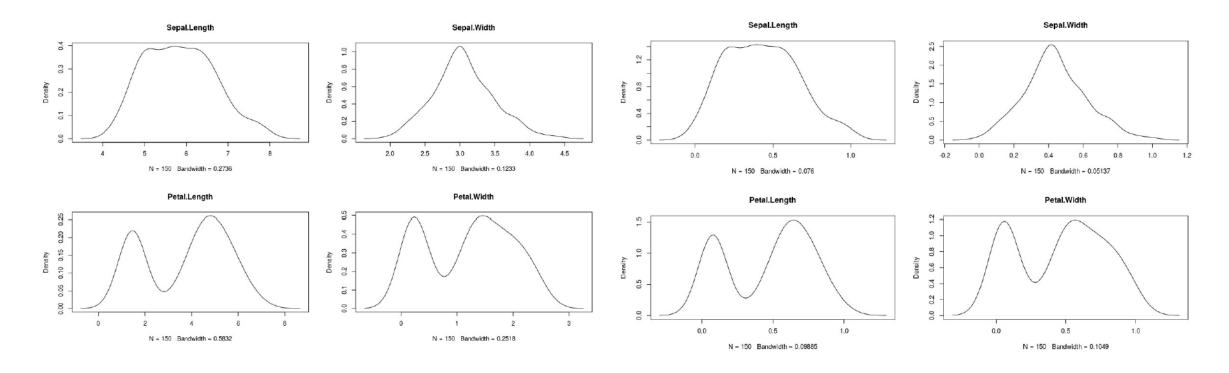






Normalización de Datos (DENSIDAD)







Transformación Box-Cox

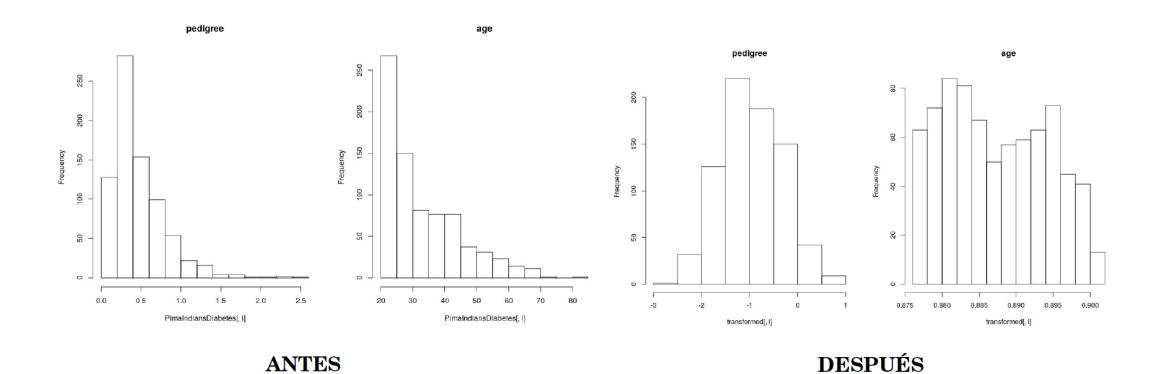


• La transformación Box-Cox son una familia de transformaciones potenciales usadas en estadística para corregir sesgos en la distribución de errores, para corregir varianzas desiguales (para diferentes valores de la variable predictora) y principalmente para corregir la no linealidad en la relación (mejora la correlación lineal entre variables). En muchos casos los atributos tienen sesgo o inclinación, que es una distribución Gaussiana pero desplazada. Para estos casos se usa la transformación Box-Cox. Se aplicará por recomendación en el modelado solo a atributos que tienen sesgo. Box-Cox asume todos los atributos positivos.



Transformación Box-Cox de Datos (HISTOGRAMAS)

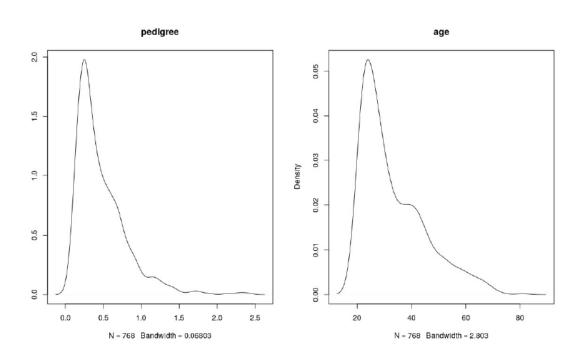


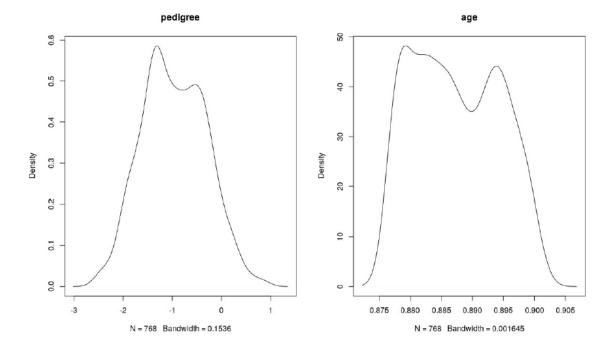




Transformación Box-Cox de Datos (DENSIDAD)









Transformación Yeo-Jhonson

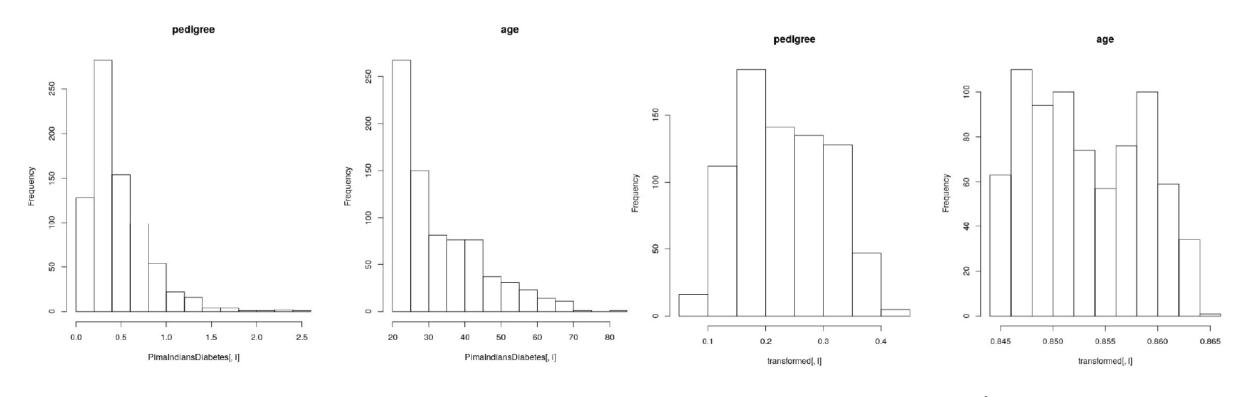


• La transformación Yeo-Jhonson es otra transformación de potencia como Box-Cox, pero soporta valores en bruto (sin procesar) que son iguales a cero o negativos.



Transformación Yeo-Jhonson de Datos (HISTOGRAMAS)

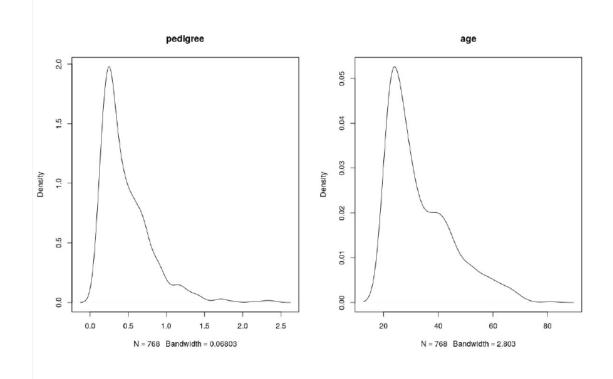


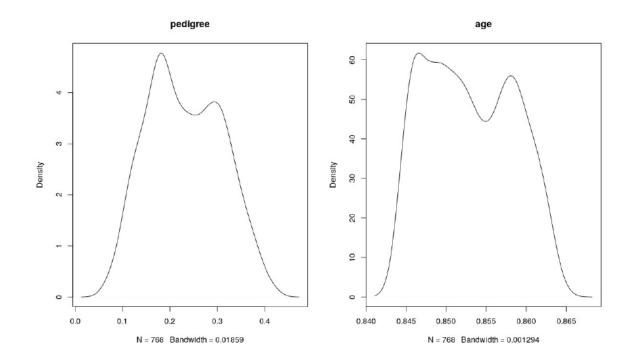




Transformación Yeo-Jhonson de Datos (DENSIDAD)







ANTES

DESPUÉS



PCA (Principal Component Analysis)



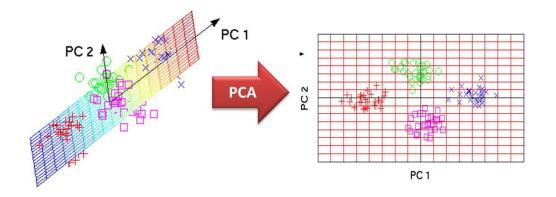
- Busca explotar la estructura existente de manera no supervisada para simplificar los datos y reducirlos o comprimirlos. Son útiles para visualizar datos o para simplificar un conjunto de variables que luego puede usar el algoritmo supervisado. El PCA es un procedimiento estadístico que utiliza una transformación ortogonal que convierte el conjunto de variables correlacionada en un conjunto de variables no correlacionadas. El PCA identifica pautas en nuestros datos basados en las correlaciones entre características. Esta correlación implica redundancia en nuestros datos, en otras palabras, que hay parte de los datos que se pueden explicar por relaciones con otras partes de los mismos. Estos datos correlacionados no son necesarios para el aprendizaje correcto del modelo, y por tanto pueden ser eliminados. Se puede eliminar borrando directamente ciertas columnas (características) o combinando un número de ellos y obteniendo nuevos datos que contengan la mayoría de la información.
- Eliminar problema de dimensionalidad y overfitting.





• El PCA transforma los datos para devolver sólo los componentes principales, una técnica de estadística multivariada y álgebra lineal. La transformación mantiene esos componente por encima del umbral de varianza (valor predeterminado = 0.95) o se puede especificar el número de componentes (`pcaComp`). El resultados son atributos no correlacionados, útiles para algoritmos como *Linear Regression* y *Generalized Linear Regression*, *entre otros*.

Dimensionality Reduction Principal Component Analysis

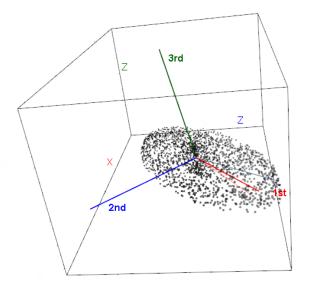






```
data("iris")
summary(iris)
preProcessParams <- preProcess(iris, method=c("center", "scale", "pca"))
print(preProcessParams)
transformed <- predict(preProcessParams, iris)
summary(transformed)</pre>
```

PCA applied to an ellipsoidically shaped point cloud

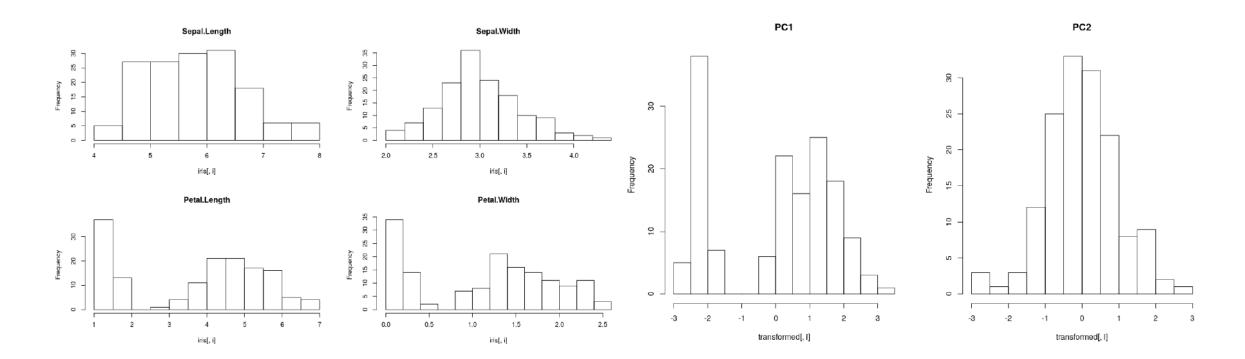


more information: www.joyofdata.de/blog/illustration-of-principal-component-analysis-pca





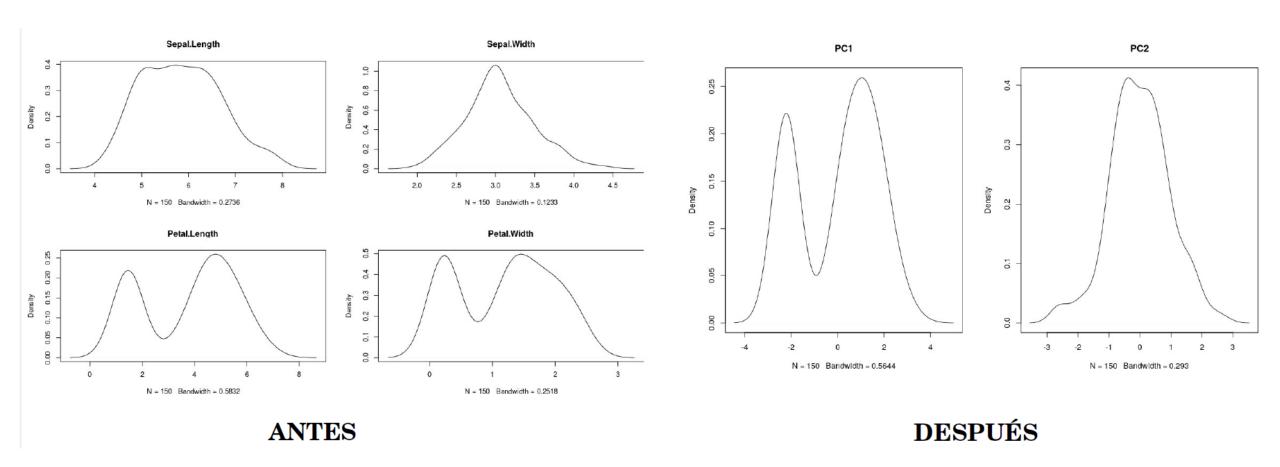






PCA (DENSIDAD)







ICA (Independent Component Analysis)



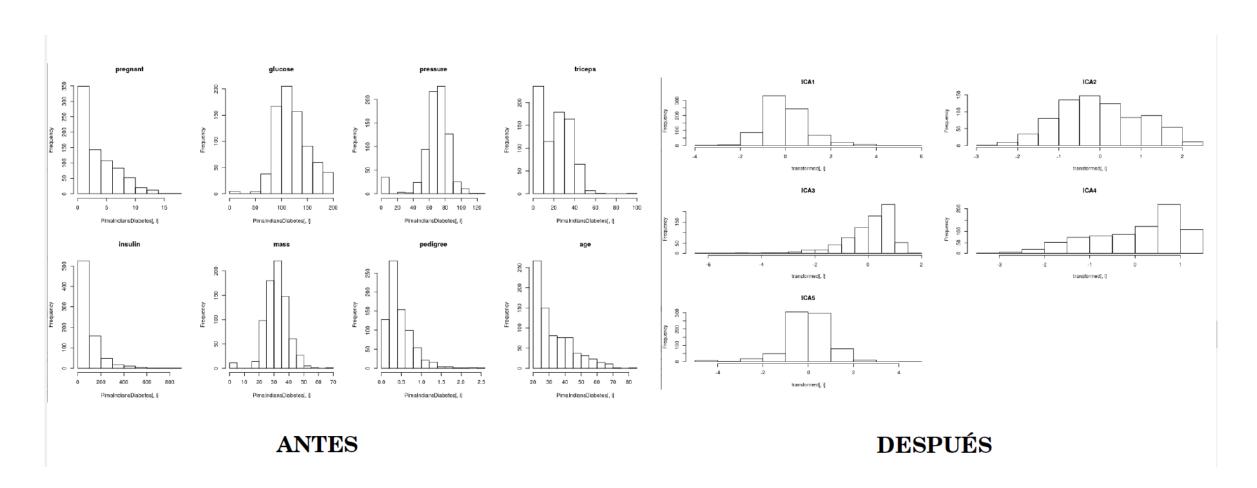
• El ICA es un método computacional que sirve para separar una señal multivariante en subcomponentes aditivos suponiendo que la señal de origen tienen una independencia estadística y es no-Gaussiana. En este contexto, ICA es una generalización de PCA, en ambos casos se practica una transformación lineal de los datos originales, aunque la diferencia básica es que el ICA no requiere que las variables tengan una distribución gaussiana. ICA conserva aquellos componentes que son independientes. Debe especificar el número de componentes independientes deseados con el argumento `n.comp`. Esta transformación puede ser útil para algoritmos como *Naive Bayes*.

```
97 data(PimaIndiansDiabetes)
98 summary(PimaIndiansDiabetes[,1:8])
99 preProcessParams <- preProcess(PimaIndiansDiabetes[,1:8],
100 method=c("center", "scale", "ica"), n.comp=5)
101 transformed <-predict(preProcessParams, PimaIndiansDiabetes[,1:8])
102 summary(transformed)</pre>
```



ICA (HISTOGRAMAS)

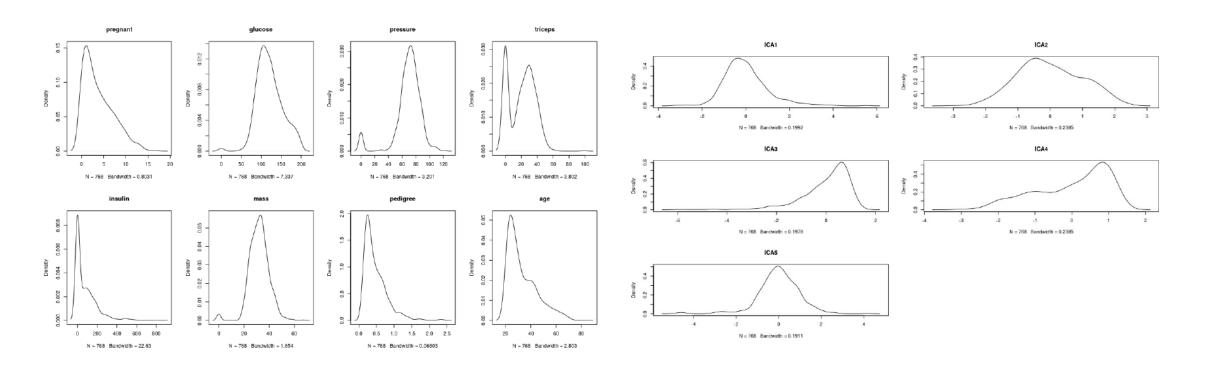






ICA (DENSIDAD)









Consejos

- Las transformaciones enseñadas en general tienen un gran impacto en la precisión de modelos finales por eso *debemos usarlas*.
- En los problemas de "modelado" debemos probara diferentes transformaciones de los datos con diferentes algoritmos de machine learning por eso *usemos en variedad*.
- Es una recomendación muy importante resumir y visualizar los datos antes y después de aplicar una transformación lo podemos hacer numérica o gráficamente (summary(), hist(), density()).

BIBLIOGRAFÍA:



Preprocesado de datos:

https://www.diegocalvo.es/preprocesado-de-datos-en-r/

Tratamiento de NA:

https://www.guru99.com/r-replace-missing-values.html

PCA:

https://www.aprendemachinelearning.com/comprende-principal-component-analysis/ https://towardsdatascience.com/principal-component-analysis-for-dimensionality-reduction-115a3d157bad https://medium.com/towards-artificial-intelligence/machine-learning-dimensionality-reduction-via-principal-component-analysis-1bdc77462831

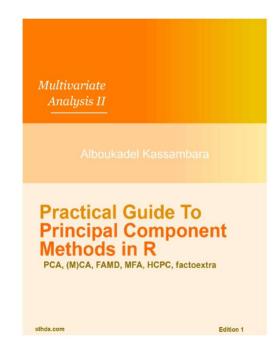
Curso de R con ejemplos:

http://www.estadistica-dma.ulpgc.es/cursoR4ULPGC/

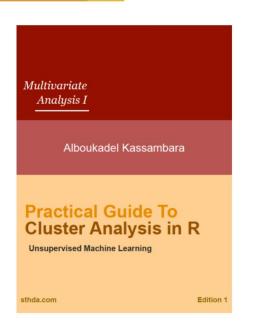
Cluster:







Use F

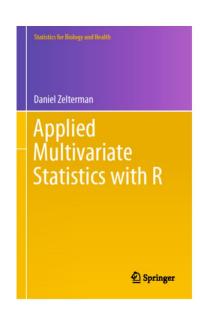


Yosef Cohen and Jeremiah Y. Cohen

with R

Statistics and Data

An Applied Approach Through Examples



An Introduction to Applied Multivariate Analysis with R

Springer Texts in Statistics

Gareth James Daniela Witten Trevor Hastie Robert Tibshirani

An Introduction to Statistical Learning

with Applications in R



BIBLIOGRAFÍA ESTADÍSTICA APLICADA A GEOLOGÍA EXTRA:

Analytics AoZ

Practical Methods for Data Analysis (US EPA QA/G-9, 2000)

Helsel, D. R., & Hirsch, R. M. (2002). *Statistical methods in water resources* (Vol. 323). Reston, VA: US Geological Survey.

Salvador Figu eras, M y Gargallo, P. (2003): "Análisis Exploratorio de Datos, 5campus.com, Estadística http://www.5campus.com/leccion/aed>

Ramalle-Gómara, E., & De Llano, J. A. (2003). Utilización de métodos robustos en la estadística inferencial. *Atención Primaria*, 32(3), 177-182.

Verzani, J. (2005). Using R for introductory statistics. CRC press.

Cohen, Y., & Cohen, J. Y. (2008). Statistics and Data with R: An applied approach through examples. John Wiley & Sons.

Arnaldo Mangeaud (2014). Estadística aplicada a las Ciencias Geológicas. Universidad nacional de Córdova.

Helsel, D.R., Hirsch, R.M., Ryberg, K.R., Archfield, S.A., and Gilroy, E.J., 2020, *Statistical methods in water resources: U.S. Geological Survey Techniques and Methods*, book 4, chapter A3, 458 p.



"LO QUE ESCUCHO LO OLVIDO. LO QUE VEO LO RECUERDO. PERO LO QUE HAGO, LO ENTIENDO." AUTOR: ANÓNIMO

VAMOS A RESOLVER UN EJERCICIO

