FUNDAMENTOS EN INGENIERÍA DE DATOS

Grupo III
Amalio Cabeza
Rafael Delgado
Álvaro Navarro
Lea Ross

1
Introducción

2
Algoritmo No supervisado

3
Visualización

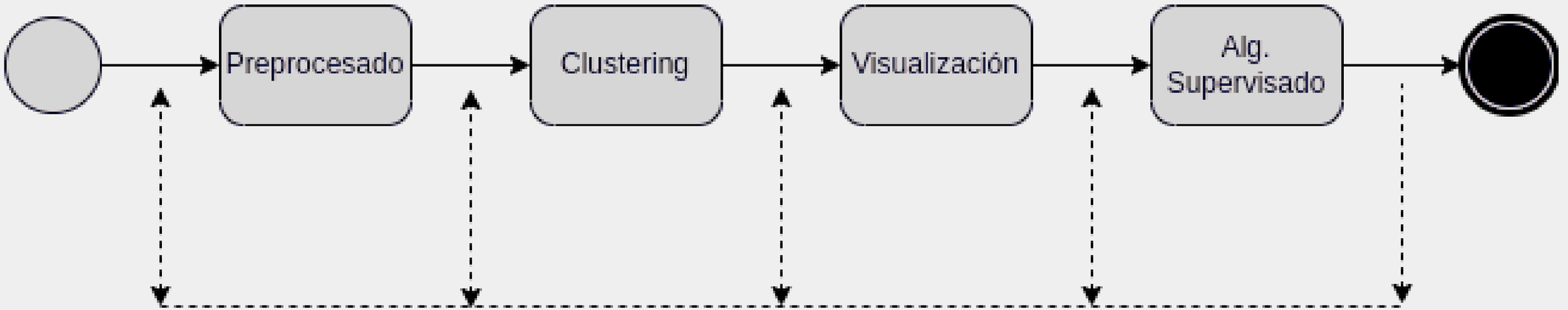
4
Algoritmos Supervisado

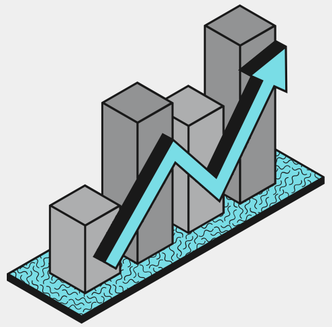
5
Detección de anomalías

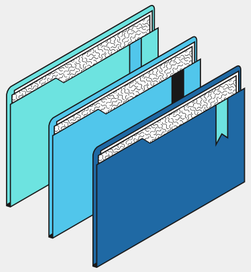
6
Conclusiones

ÍNDICE DE CONTENIDO

Introducción







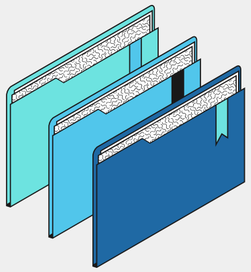
Los datos

kaggle

kaggle.com

droom India's Most Trusted Motorplace

droom.in



Los datos

1

Eliminación de texto innecesario

Sobre las columnas kms_driven, mileage y power

2

Descartar información redundante

Sobre las columnas model_name

3

Creación de nuevas columnas

Se ha creado una nueva columna para almacenar la cilindrada, la marca del vehículo y si es eléctrico

4

Trasnformación de las columnas de texto a numérico

Sobre las columnas kms_driven, mileage, power y owner

5

Tratamiento preeliminar de NA

Sobre el dataset.

Aprendizaje No Supervisado

Aprendizaje no supervisado

¿Qué técnica hemos empleado?

¿Qué parámetros hemos optimizado?

¿Con qué objetivo?

Aprendizaje no supervisado

¿Qué técnica hemos empleado?

K-Means Clustering

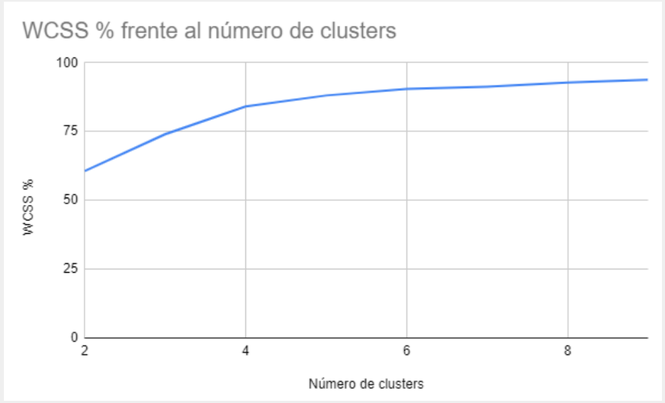
Aprendizaje no supervisado

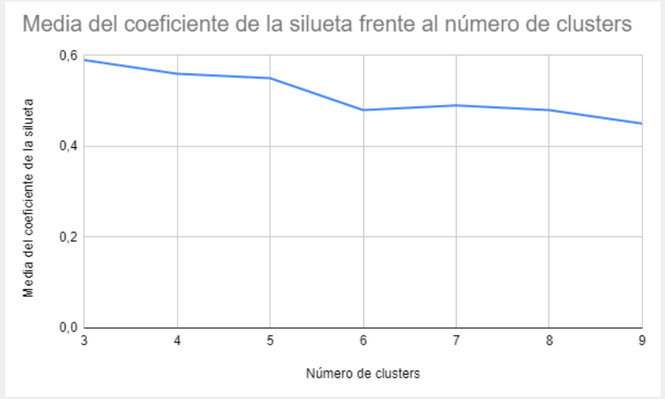
¿Qué parámetros hemos optimizado?

El número de centroides óptimo

2 Métricas empleadas

Within cluster sum of squares.
Average silhouette width.





Aprendizaje no supervisado

¿Con qué objetivo?

Mejorar el tratamiento de los datos.

Remplazo de valores nulos

Calculando la media de los valores
pertenecientes al mismo cluster

Visualización

PowerBI

- Funciona bien con el formato csv
- Tenemos experiencia porque lo hemos utilizado en el seminario
- Acceso gratuito

Procedimiento

1

Conectar los datos en GitHub con PowerBI

2

Cambiar los números a números sin decimales para simplificar la visualización

3

Creación de grupos de datos para simplificar la visualización y aumentar la pertinencia (precio, kilometers driven & BHP)

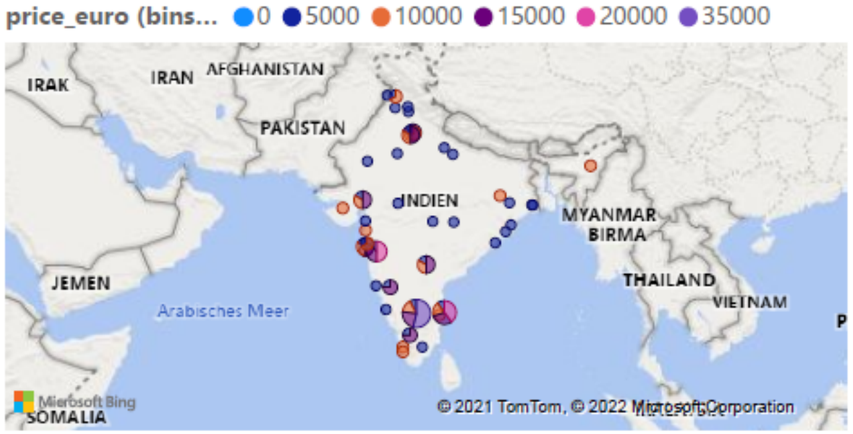
4

Creación de una nueva columna con el precio en euros

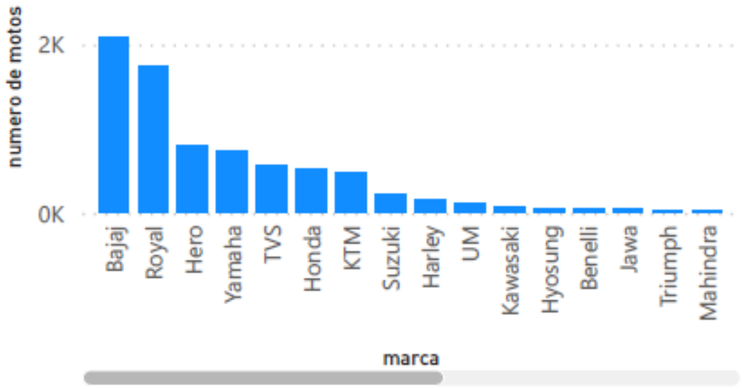
5

Visualización de datos con modelos diferentes

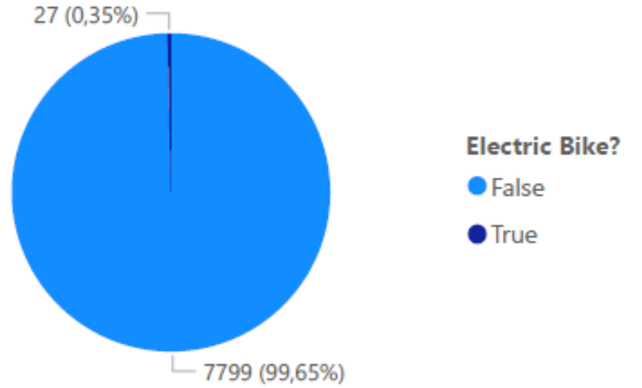
Average price per location



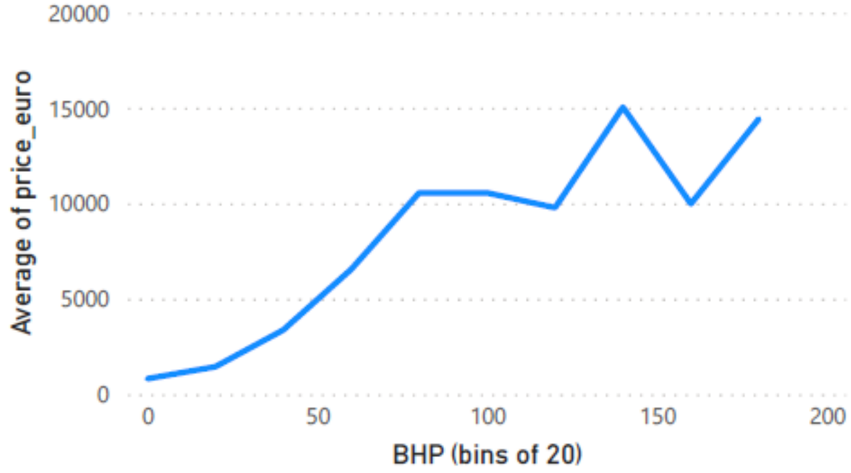
Number of bikes per brand



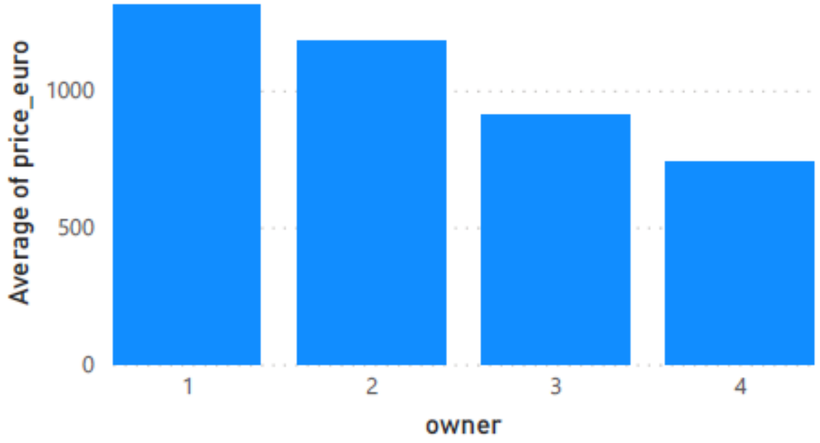
Number of electric bikes



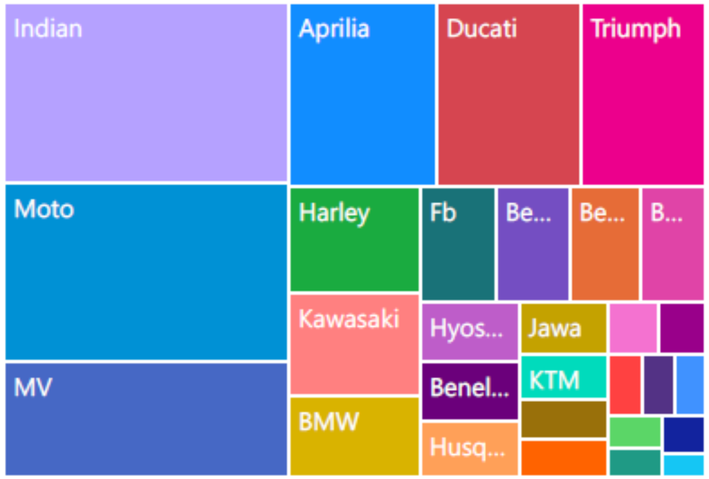
Average price group per BHP group



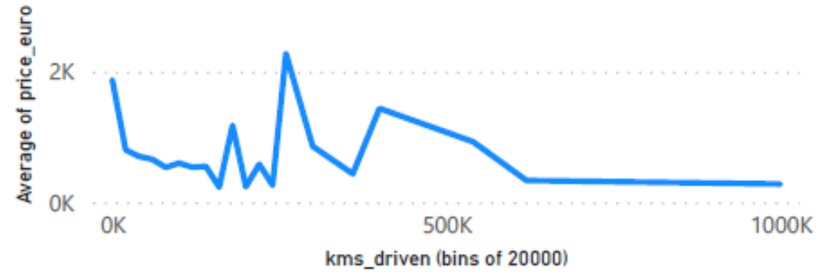
Average price per number of owners



Average price per brand



Average price per Kilometer driven



Aprendizaje Supervisado

Objetivo

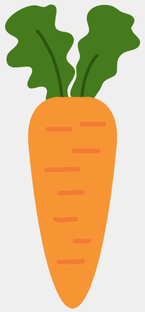
Training/testing

0.75 / 0.25

Evaluación

Herramientas

+



+

Selección de variables

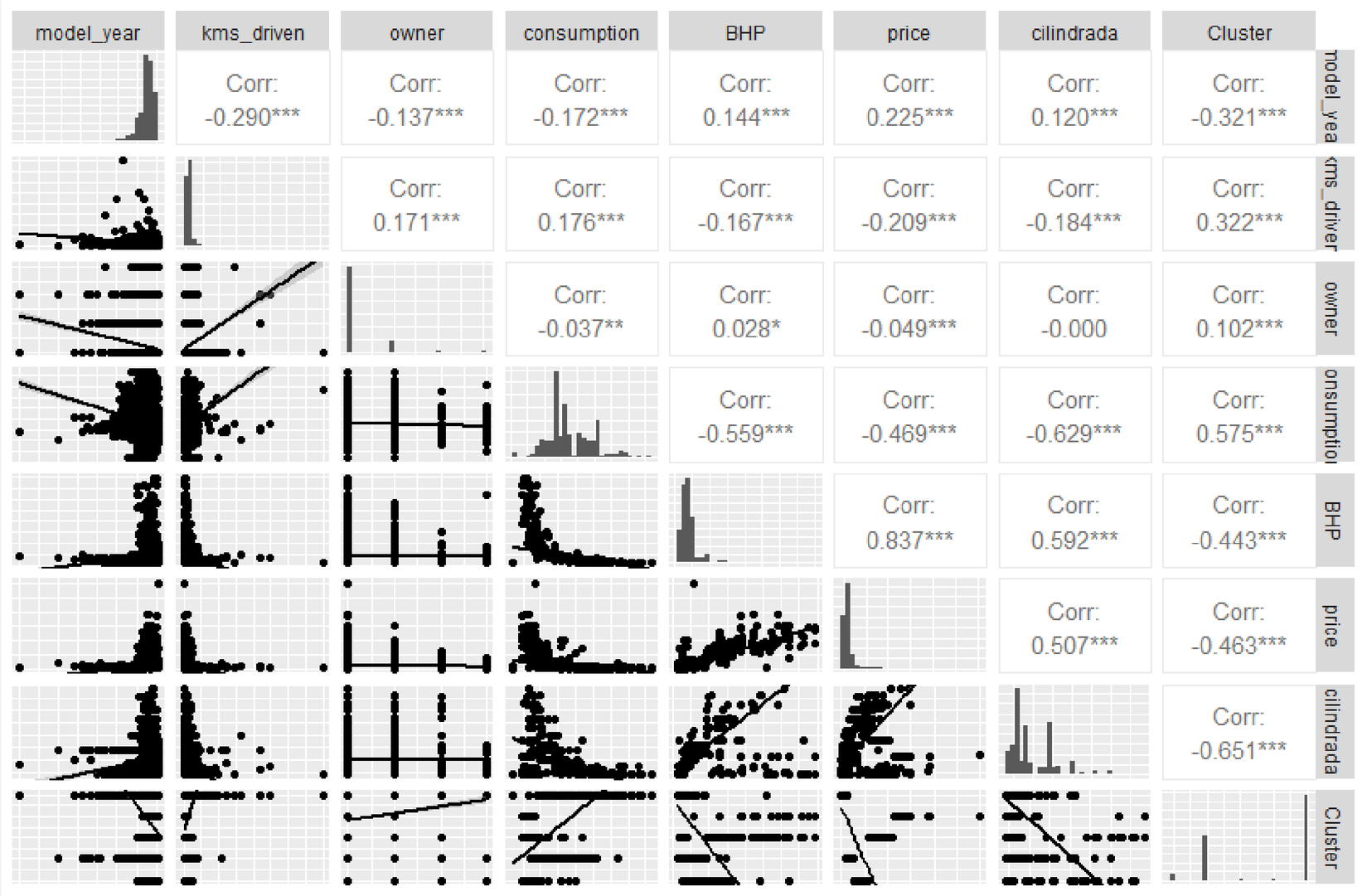
Numéricas

Correlacionadas
con el precio

kms_driven owner cilindrada

model_year BHP consumption

GGally

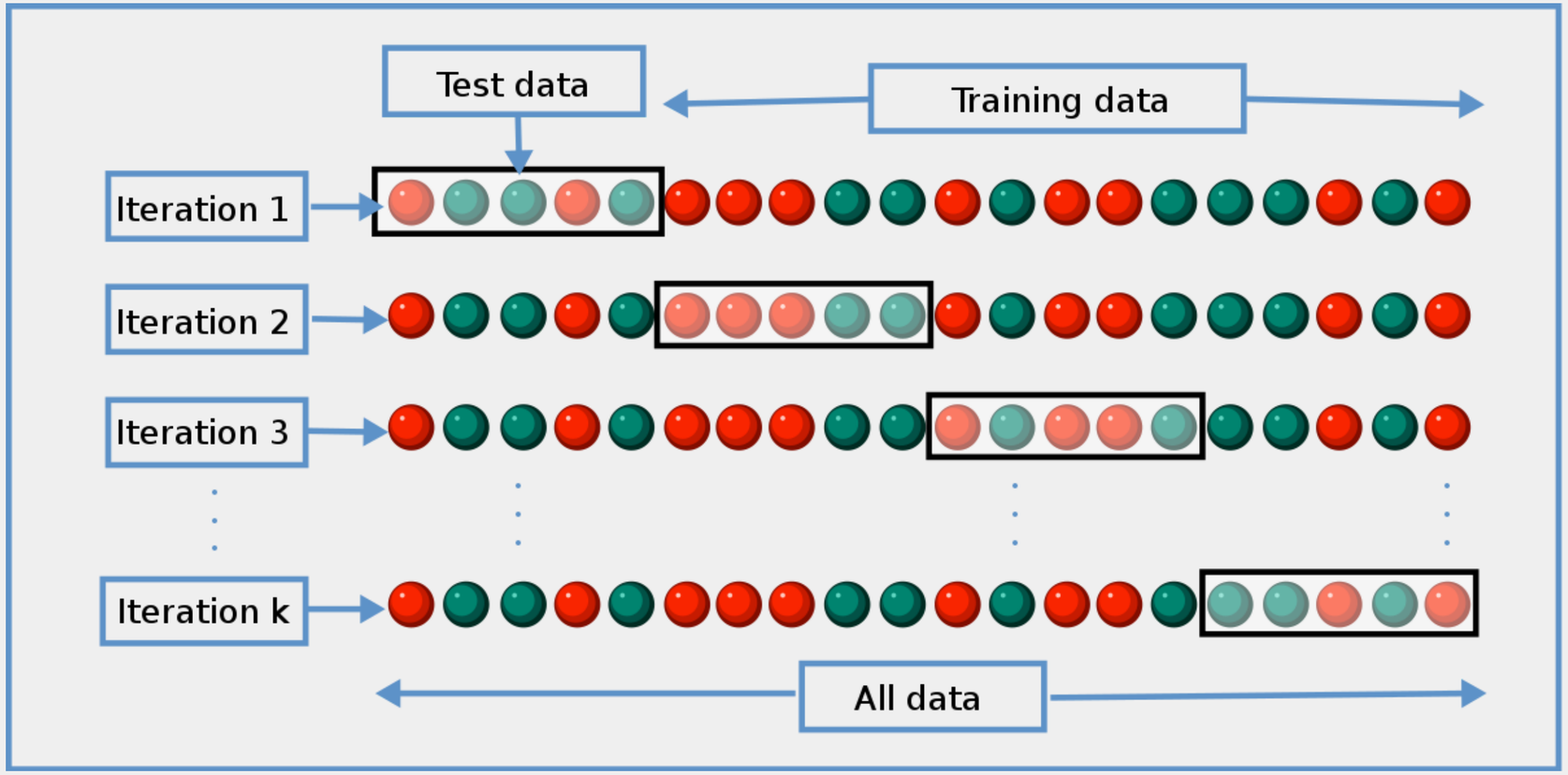


ggpairs()

Modelos usados

1. Regresión lineal múltiple
2. KNN
3. Random Forest

Train Control



Resultados

métrica: R2

RLM < KNN < RF

Resultados

métrica: R2

RLM < **KNN** < **RF**

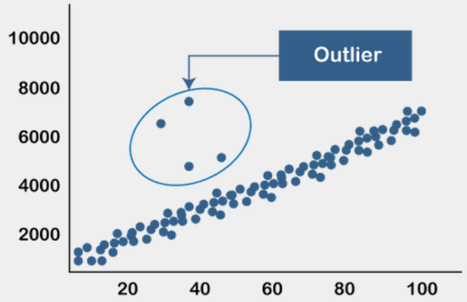
BigML

ANOMALY DETECTION

Detección de anomalías

Anomalía (outlier)

un valor atípico que es numéricamente distante del resto de los datos



Detección de anomalías

una forma de detectar instancias inusuales en su conjunto de datos

Aplicación

- Detectar comportamiento malicioso poco común
- Alerta a los técnicos de servicio
- Filtrado de anomalías para un aprendizaje supervisado "más limpio"
- Evaluación de la competencia del modelo

la interfaz **BigML** permite ver y interactuar fácilmente con las anomalías detectadas en el conjunto de datos

paso 1

muestra las 10 anomalías principales en el conjunto de datos

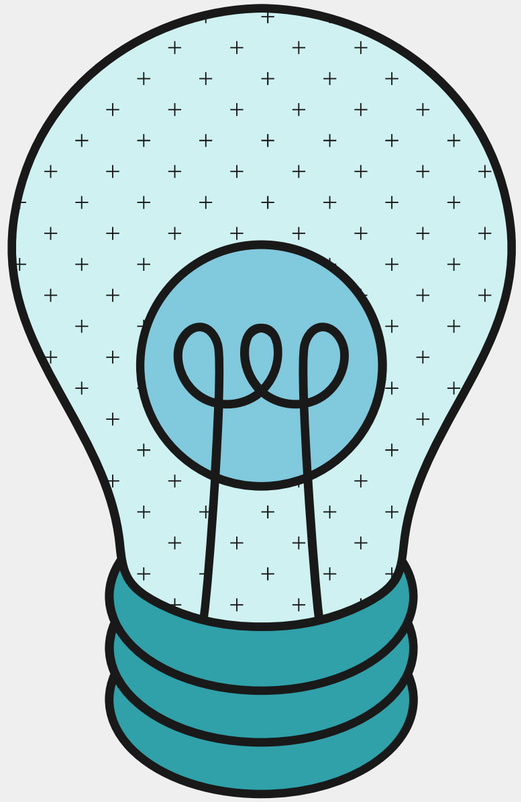
paso 2

obtener las puntuaciones de anomalías de todos los puntos en el conjunto de datos

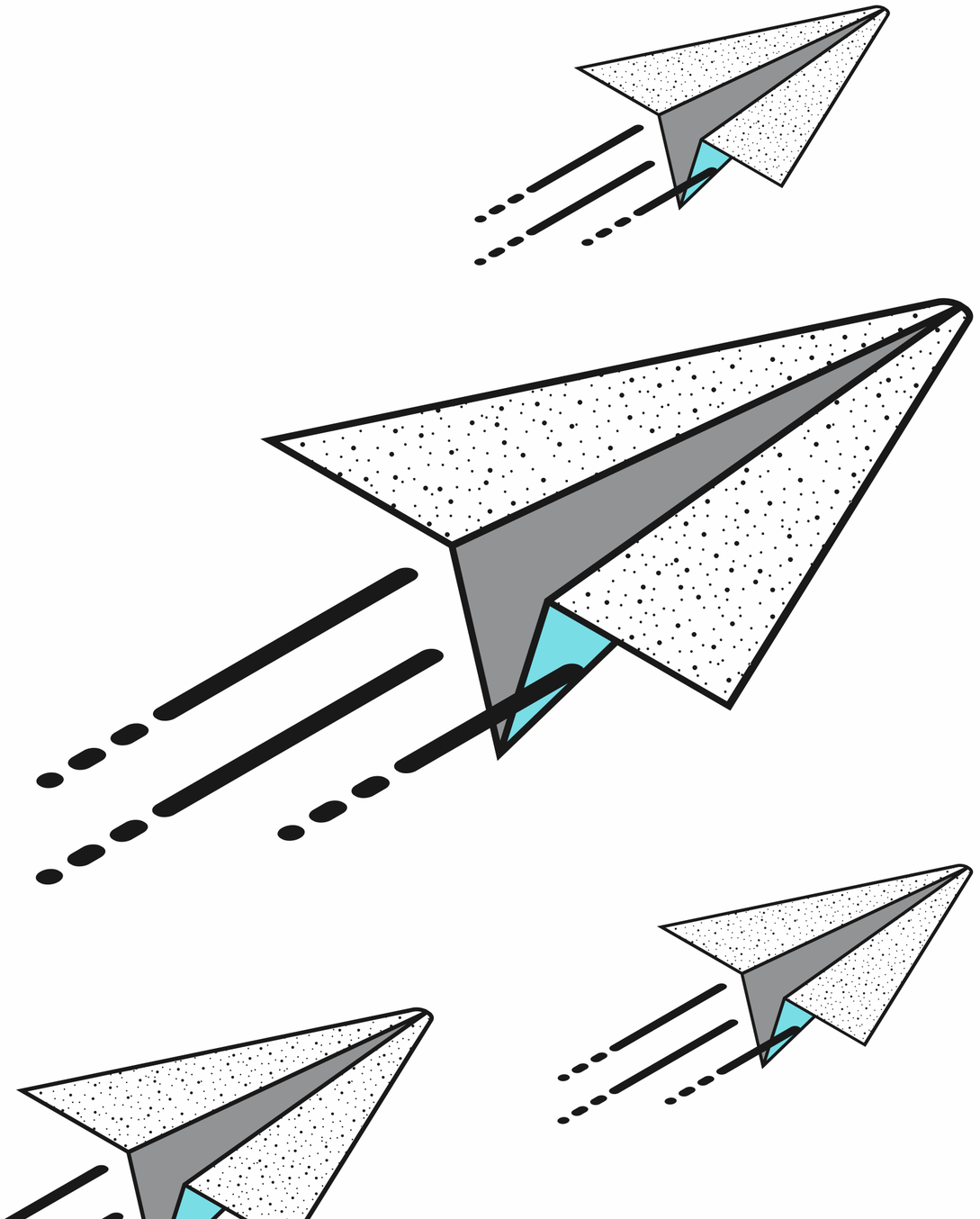
Conclusiones

- La importancia de preprocesar los datos.
- Los distintos usos de los algoritmos.

- Aplicar técnicas de detección de anomalías
- Comparaciones con otros lenguajes.
- Aprovechar las variables de texto.



¿Alguna pregunta?



https://github.com/AmalioF96/FID