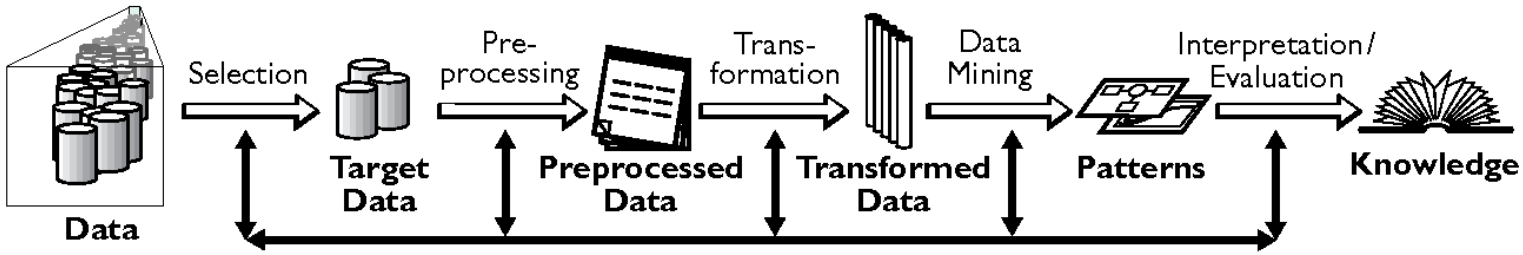
Análisis trayectos de taxis de NY

Daniel Arellano Martínez
Bruno González Llaga
Carlos González Arenas
Diego Monsalves Vázquez
Víctor Manuel Vázquez García

Fundamentos de Ingeniería de Datos

Índice

Introducción



kaggle

Conjunto de datos

Conjunto de entrenamiento

Conjunto de test

Sample submission

id. taxi

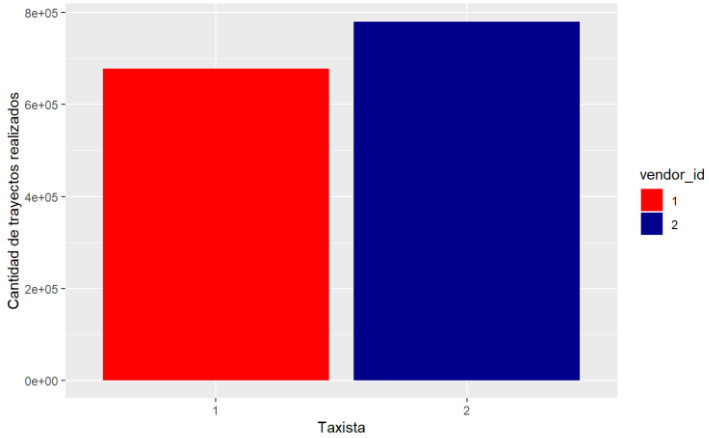
vendedor

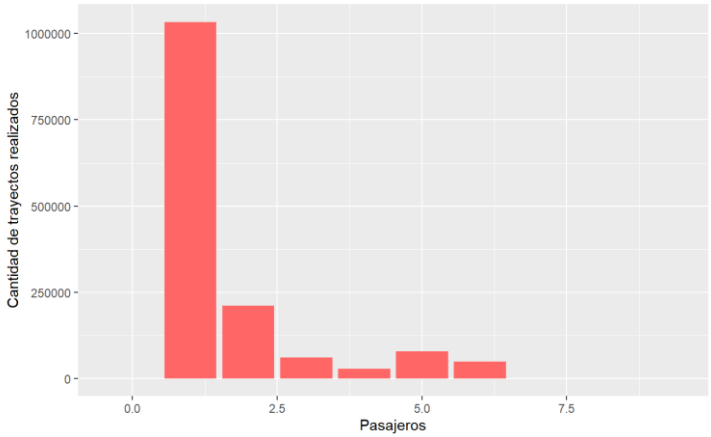
pasajeros

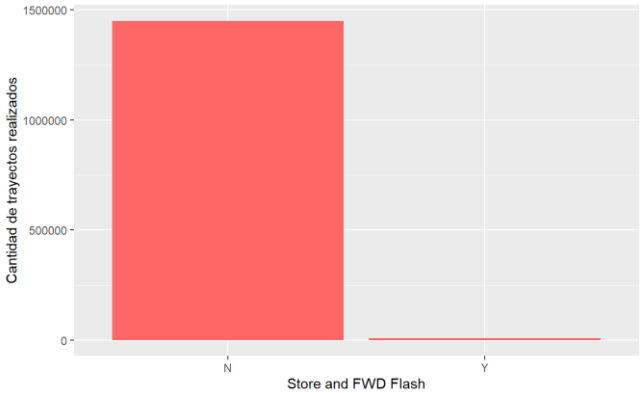
posición baj./sub.

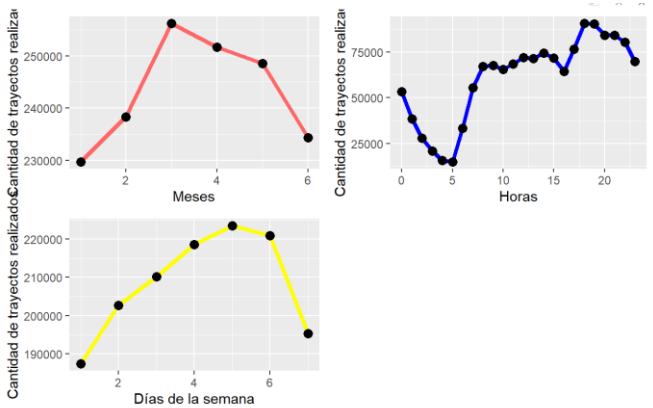
duración

Visualización

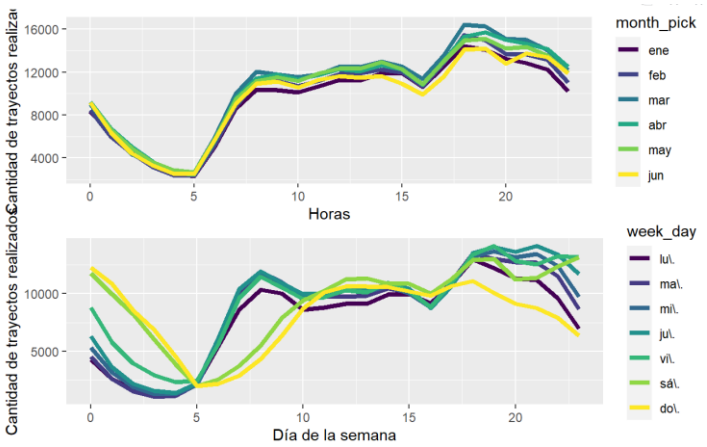


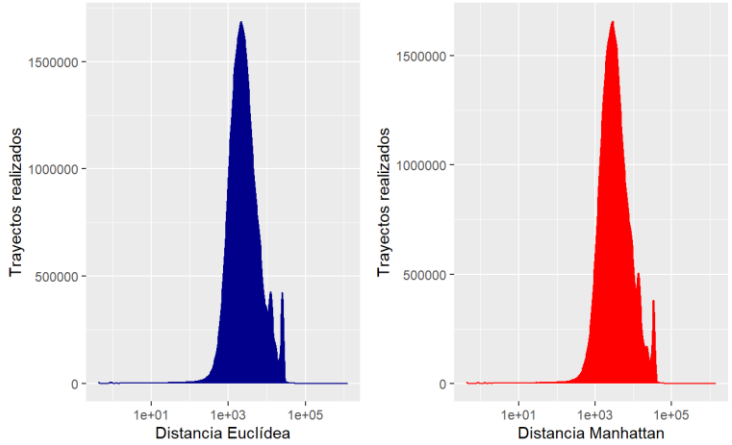


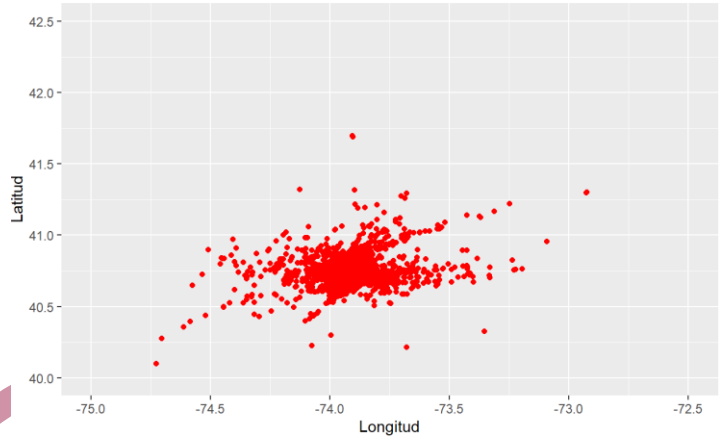


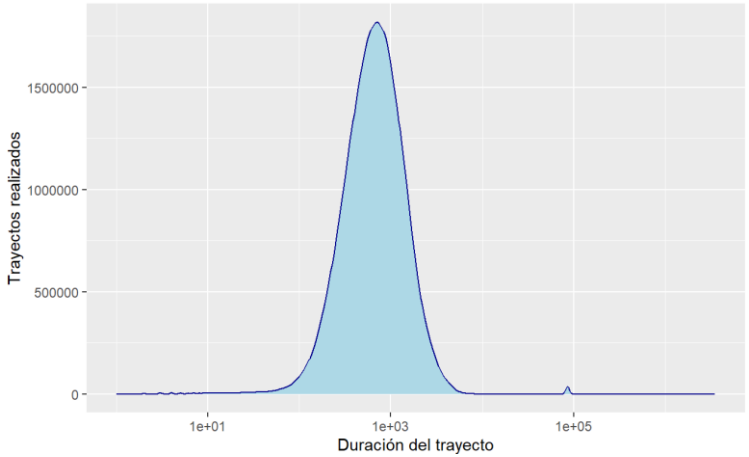


Visualización

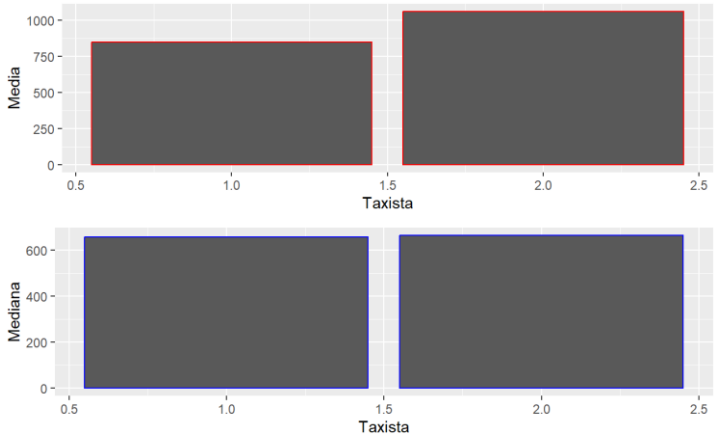


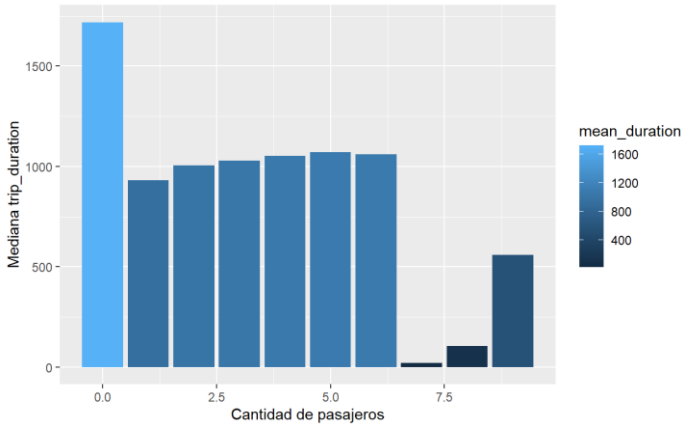


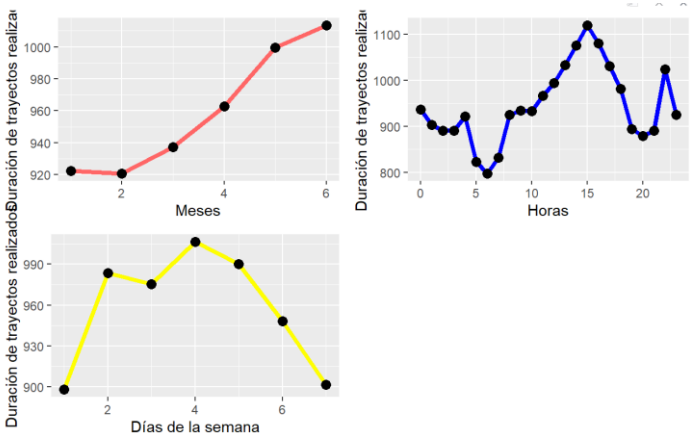


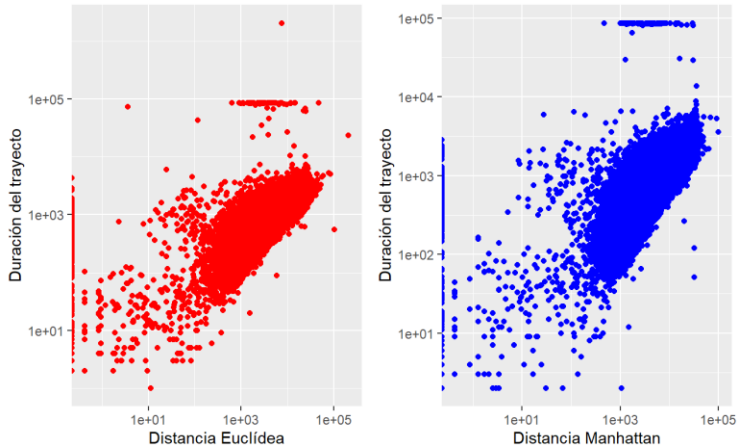


Visualización







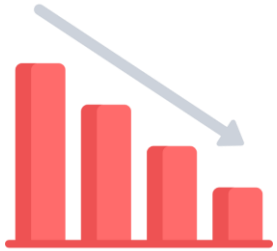


Preprocesamiento

recopilación de datos

transformación

limpieza

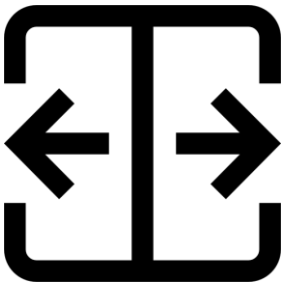


reducción

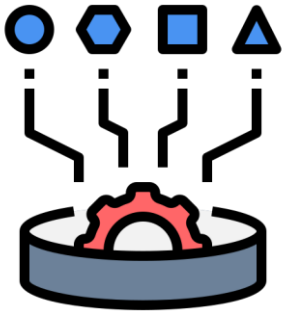
distancia

normalización
acotar

Aprendizaje supervisado



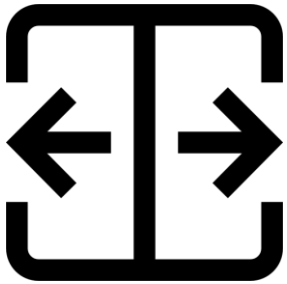
división del conjunto



importancia de las variables

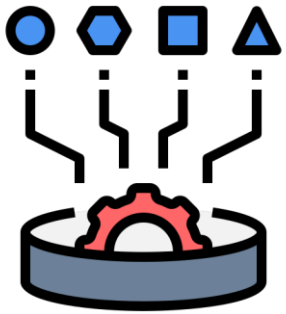
entrenamiento

Aprendizaje supervisado

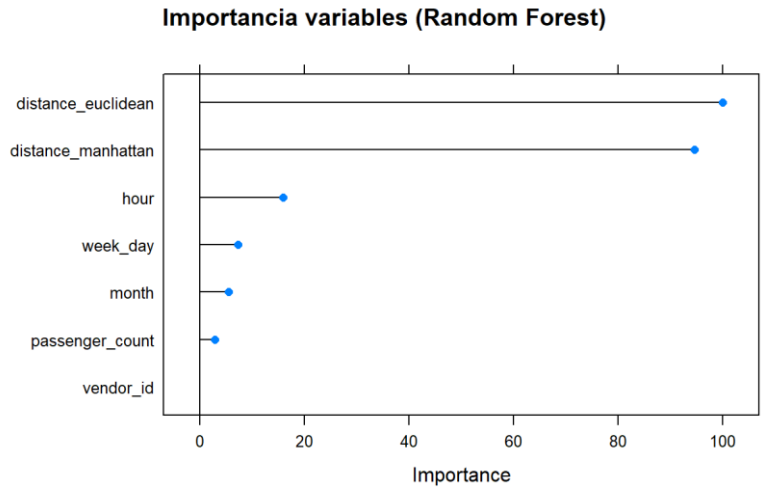


división del conjunto

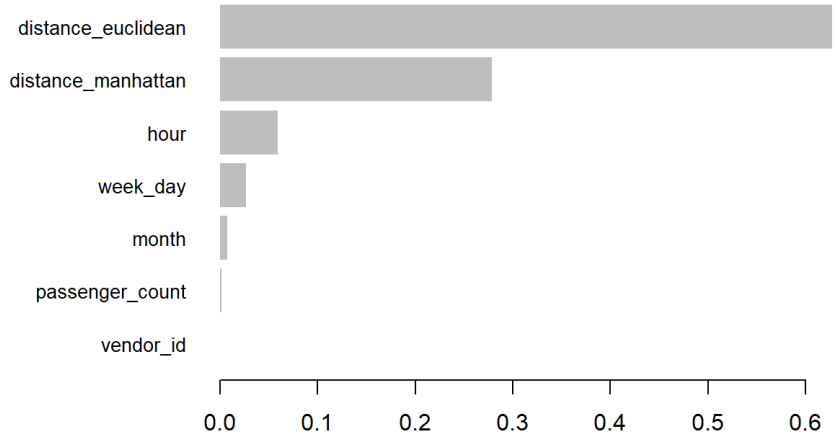
Aprendizaje supervisado



importancia de las variables



xgb



Aprendizaje supervisado

entrenamiento

- Regresión lineal
- XGBoost
- Random Forest

Aprendizaje supervisado

entrenamiento

Regresión Lineal

- **Distancia Euclídea**

RMSE	Rsquared	MAE
0.0001128188	0.5936826823	0.0000828220

- **Distancia Manhattan**

RMSE	Rsquared	MAE
0.0001127156	0.5929620046	0.0000825294

Aprendizaje supervisado

Regresión Lineal

entrenamiento

```
Call:
summary.resamples(object = resam)

Models: EUC, MAN
Number of resamples: 25

MAE
             Min.      1st Qu.       Median         Mean      3rd Qu.         Max.
NA's
EUC 7.470956e-05 7.934337e-05 8.173332e-05 8.183042e-05 8.430475e-05 8.893851e-05
0
MAN 7.751571e-05 8.091833e-05 8.244996e-05 8.268354e-05 8.392152e-05 9.077866e-05
0

RMSE
             Min.      1st Qu.       Median         Mean      3rd Qu.         Max.
NA's
EUC 9.908712e-05 0.0001131512 0.0001202717 0.0001203689 0.0001312543 0.0001383157
0
MAN 1.090085e-04 0.0001145401 0.0001225887 0.0001221160 0.0001269971 0.0001435306
0

Rsquared
         Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
EUC 0.5414675 0.5949269 0.6294260 0.6225607 0.6446965 0.7064989    0
MAN 0.5158429 0.5884741 0.6139354 0.6133371 0.6589657 0.6885963    0
```

Aprendizaje supervisado

entrenamiento

Regresión Lineal

- **Distancia Euclídea CV**

RMSE	Rsquared	MAE
0.0001128188	0.5936826823	0.0000828220

- **Distancia Manhattan CV**

RMSE	Rsquared	MAE
0.0001127156	0.5929620046	0.0000825294

Aprendizaje supervisado

Regresión Lineal

entrenamiento

```
Call:
summary.resamples(object = resam)

Models: EUC, MAN
Number of resamples: 25

MAE
           Min.      1st Qu.       Median         Mean      3rd Qu.         Max.
NA's
EUC 7.270360e-05 7.802745e-05 8.044021e-05 8.115232e-05 8.560012e-05 9.128758e-05
0
MAN 6.970988e-05 7.891683e-05 8.269996e-05 8.220356e-05 8.556611e-05 9.092459e-05
0

RMSE
           Min.      1st Qu.       Median         Mean      3rd Qu.         Max.
NA's
EUC 9.443334e-05 0.0001081502 0.0001171875 0.0001171842 0.0001276460 0.000138404
0
MAN 9.444028e-05 0.0001051338 0.0001202042 0.0001189592 0.0001302339 0.000152033
0

Rsquared
         Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
EUC 0.5354619 0.5739745 0.6143498 0.6206134 0.6779628 0.6950297    0
MAN 0.3758607 0.5472309 0.6237546 0.6090115 0.6766204 0.7409798    0
```

Aprendizaje supervisado

XGBoost

entrenamiento

- Implementación eficiente del algoritmo de boosting gradient descent.
- Boosting: se entrenan varios modelos.
- Utiliza árboles de decisión como modelos base y los entrena utilizando una función de pérdida y el algoritmo de gradiente descendente.
- XGBoost tiene varios hiperparámetros que puedes ajustar para mejorar el rendimiento del modelo.
- Rápido y eficiente.

Aprendizaje supervisado

XGBoost

entrenamiento

```
[13:43:19] WARNING: amalgamation/../src/objective/regression_obj.cu:203:
reg:linear is now deprecated in favor of reg:squarederror.
[1]     train-rmse:303.883722
[2]     train-rmse:229.012598
[3]     train-rmse:180.576932
[4]     train-rmse:150.687317
[5]     train-rmse:130.740838
[6]     train-rmse:119.400281
[7]     train-rmse:112.800194
[8]     train-rmse:108.174040
[9]     train-rmse:105.209131
[10]    train-rmse:103.157635
      RMSE   Rsquared        MAE
76.0376857  0.5258062 60.9132809
```

Aprendizaje supervisado

entrenamiento

Random Forest

RMSE	Rsquared	MAE
1.010730e-04	6.767166e-01	7.108522e-05

Aprendizaje supervisado

entrenamiento

Conclusión

Regresión lineal

RMSE	Rsquared	MAE
0.0001128188	0.5936826823	0.0000828220

XGBoost

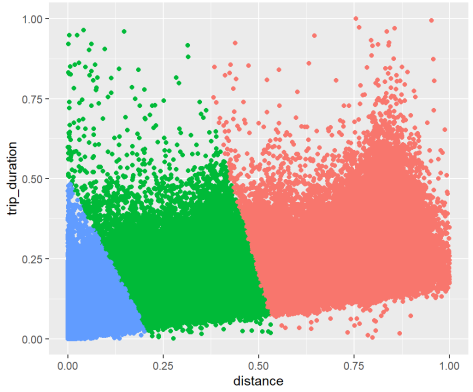
RMSE	Rsquared	MAE
76.0376857	0.5258062	60.9132809

Random Forest

RMSE	Rsquared	MAE
1.010730e-04	6.767166e-01	7.108522e-05

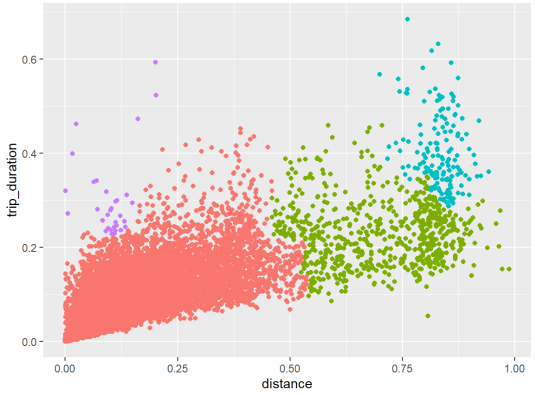
Aprendizaje no supervisado

Clustering con k-means



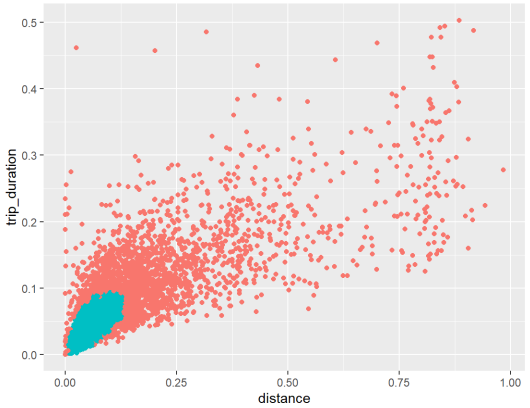
- 100% muestras
- k -> WSS

Clustering con árbol jerárquico



- 20k muestras
- matriz distancias
- dendograma
- k -> WSS
- Corte dend.

Clustering basado en densidad



- 5k muestras
- matriz distancias
- dbscan
- eps -> k
- min_pts

Aprendizaje no supervisado

Clustering con k-means

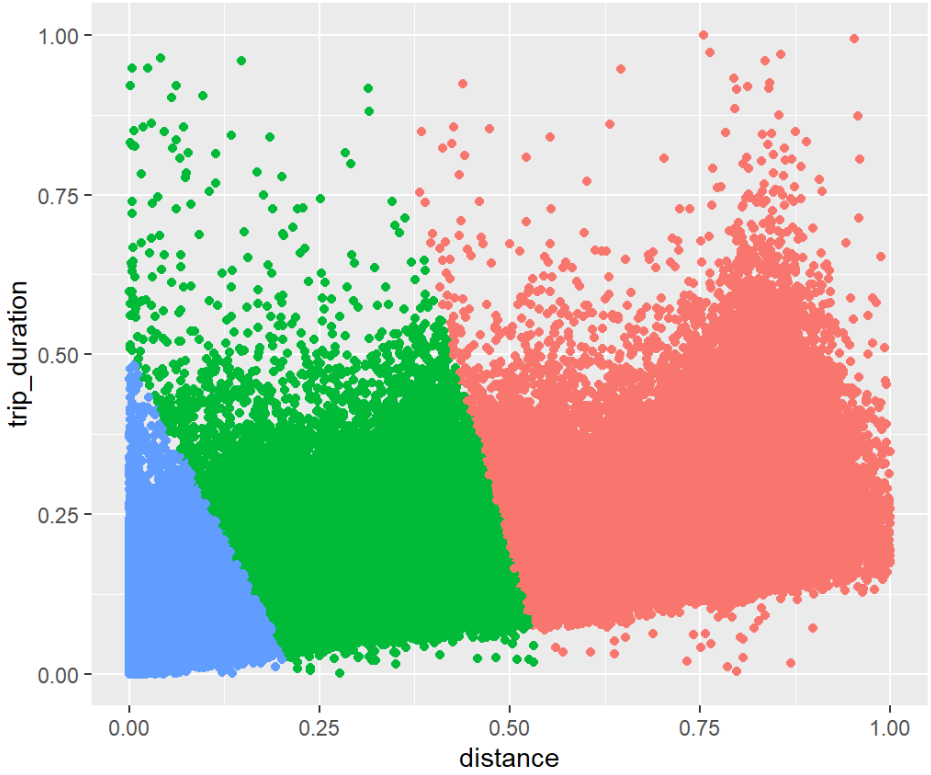
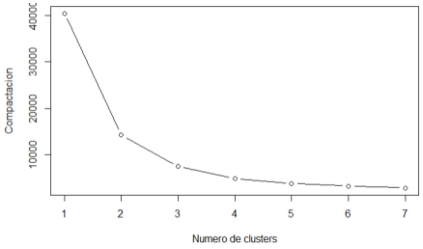


tabla minable



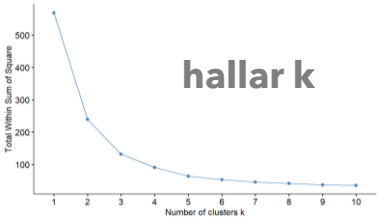
determinar k en función de WSS

obtención resultados

Aprendizaje no supervisado

Clustering con árboles jerárquicos

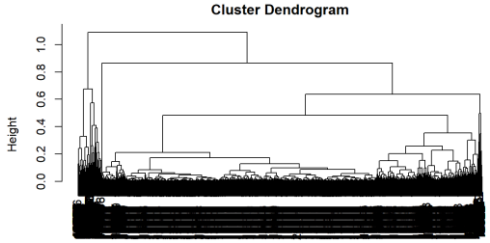
tabla minable



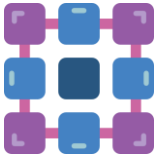
hallar k

corte dendograma

reducción a 20k muestras



unir dataset con clúster



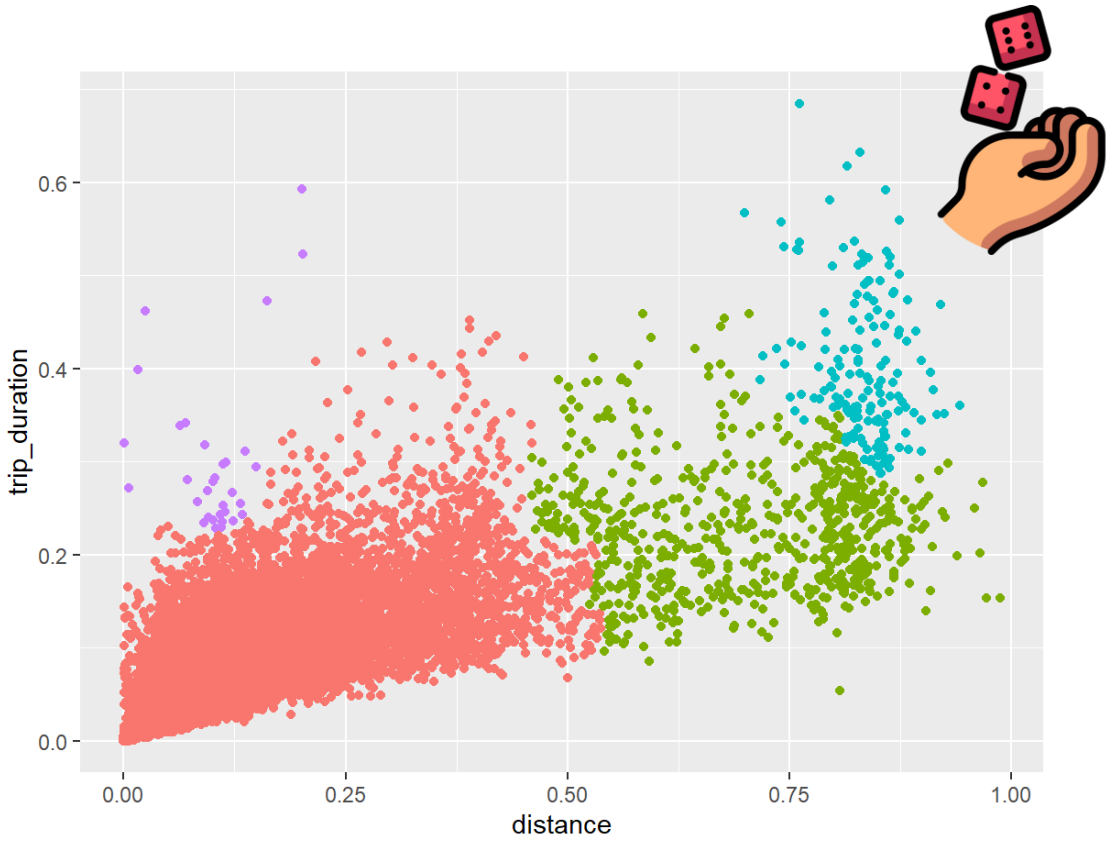
matriz de distancias

función hclust

obtención resultados

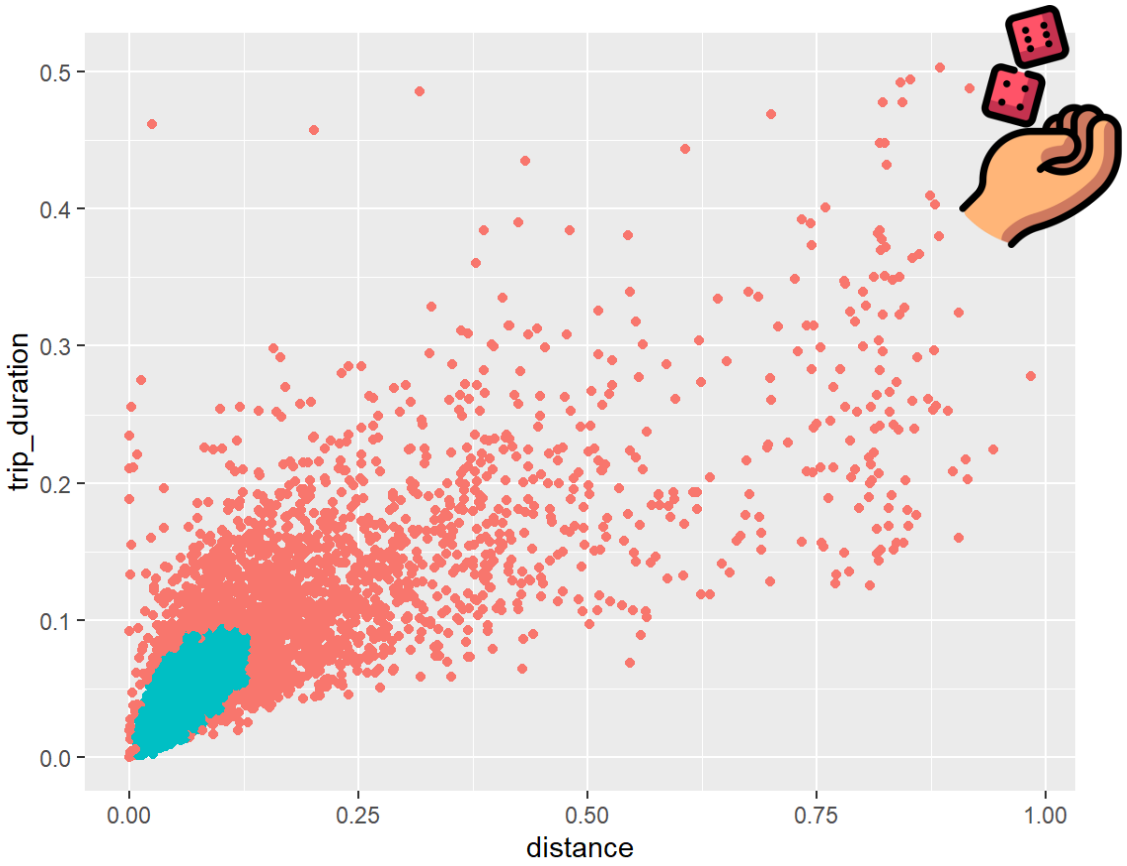
Aprendizaje no supervisado

Clustering con árboles jerárquicos

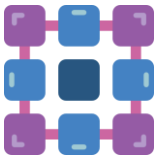


Aprendizaje no supervisado

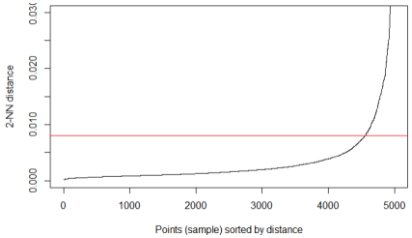
Clustering basado en densidad



= anterior con 5k muestras



matriz de distancias



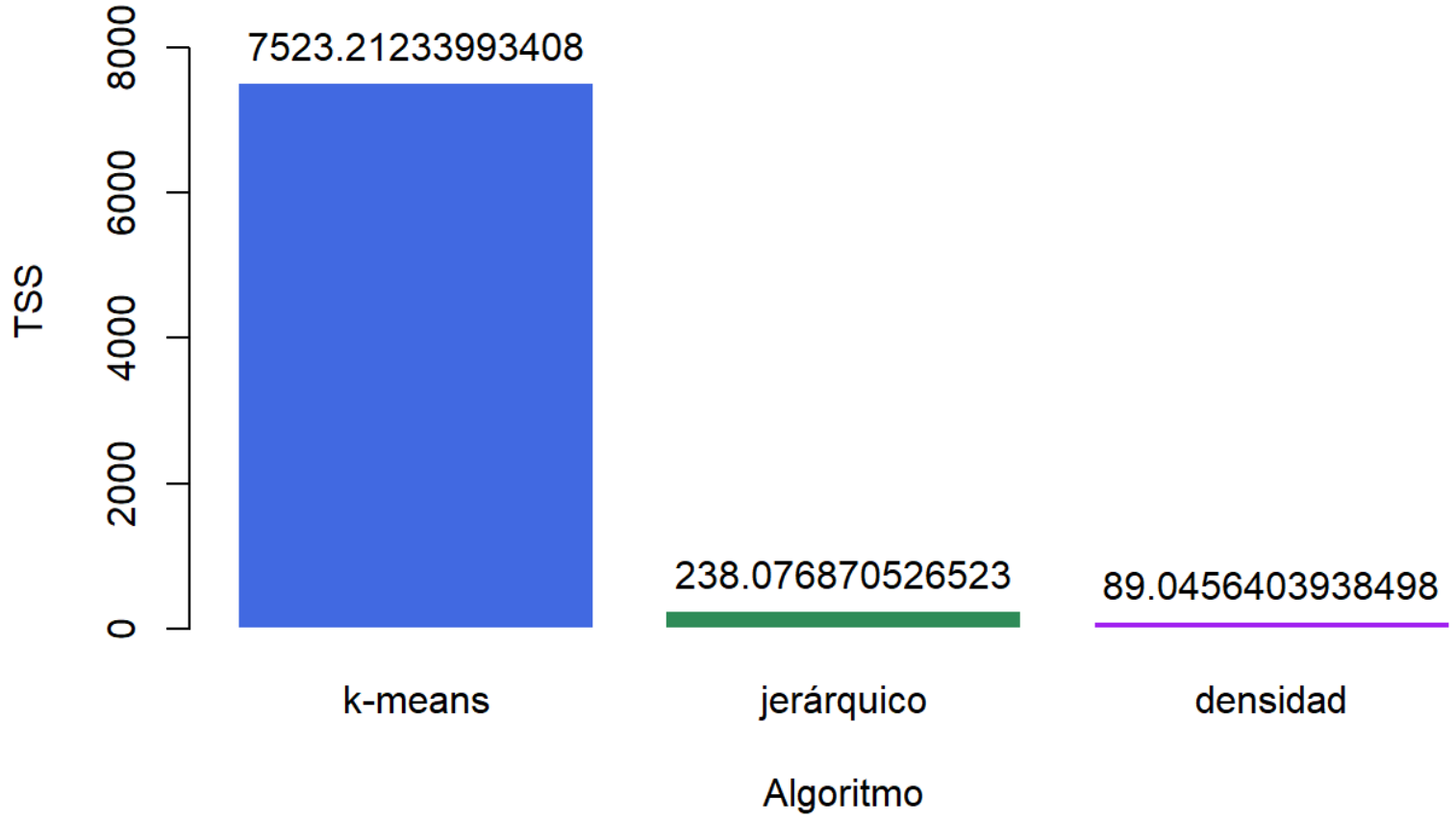
obtener EPS

aplicar dbscan

= anterior

Aprendizaje no supervisado

Comparativa algoritmos



Conclusiones

Bibliografía

ICONOS

- https://icon-icons.com/
- https://www.vectorlogo.zone/
- https://upload.wikimedia.org/wikipedia/commons/5/58/Scrum_process.svg
- https://www.flaticon.com/

Preguntas