



Ebay Reviews

Adrià Orozco 1490952

<https://github.com/AdriaOrozco/CasKaggle>

INDICE

EBAY REVIEWS



Introducción



Análisis de Datos



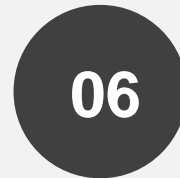
Pre procesamiento



Métodos utilizados



Resultados



Conclusiones

04

Introducción



Ebay Reviews

Reseñas de usuarios sobre productos

Base de datos con 4 atributos



Análisis de datos

01

Categoría (String)

02

Título de la reseña (String)

03

Descripción de la reseña (String)

04

Valoración (Integer)

Pre Procesado (Encoding, Normalization)

Resultados (dataset.head)

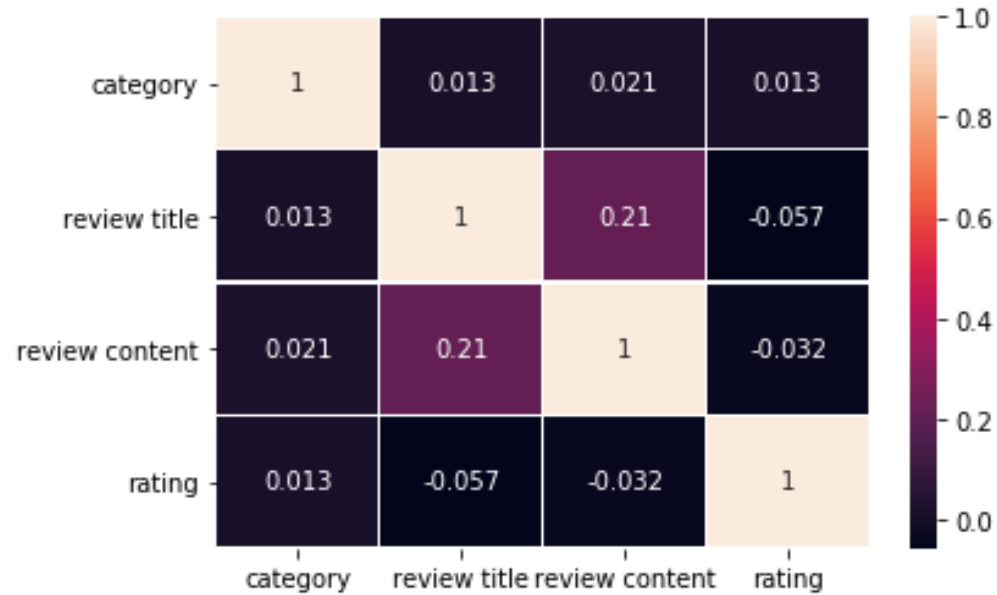
	category	review title	review content	rating
0	Headsets	Wireless gaming headset	This gaming headset ticks all the boxes # look...	5
1	Headsets	Good for those with a big head, low budget	Easy setup, rated for 6 hours battery but mine...	3
2	Headsets	MezumiWireless Gaming Headset	I originally bought this wireless headset for ...	5
3	Headsets	HW- S2 great headset.	This is my 2nd Mezumi headset, It kills the fi...	5
4	Headsets	BEST HEADPHONES I'VE PURCHASED IN MY ENTIRE LIFE	This is probably the best headset I've purchas...	5

Preparación atributos

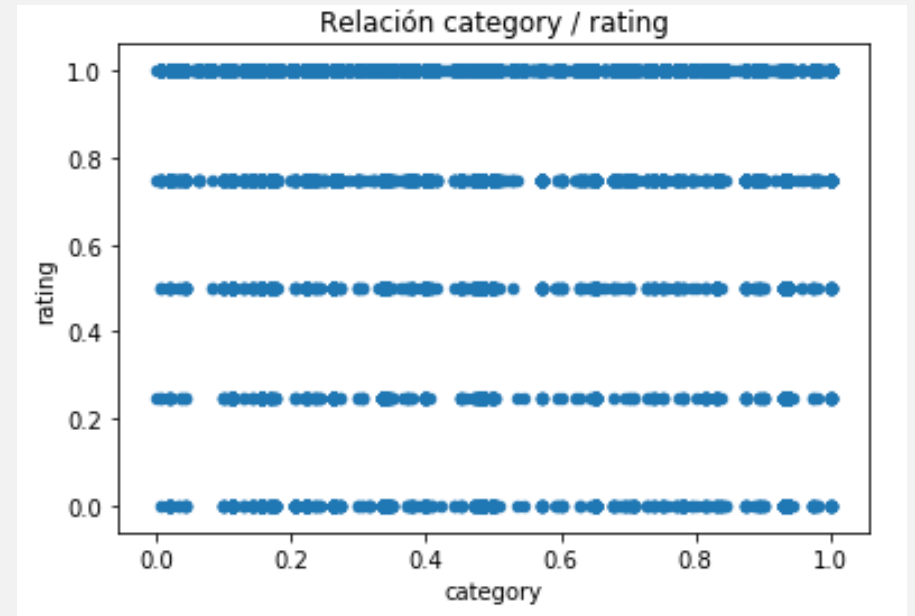
- Valores nulos ----> No existen
- Outliers ----> No se detectan
- Variables categóricas ----> Codificadas con labelEncoder
- Normalización ----> Realizado con MinMaxScaler
- OverSampling ----> Igualando el numero de muestras de cada rating

Pre Procesado

Matriz de correlación



Ejemplo nube de puntos respecto dos atributos

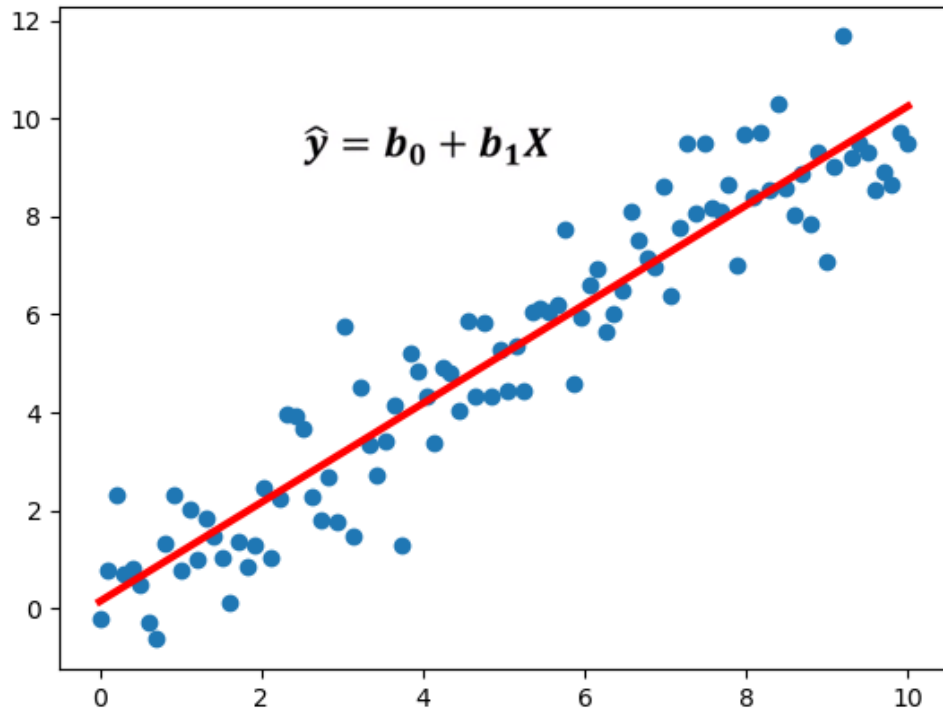


Preparación atributos

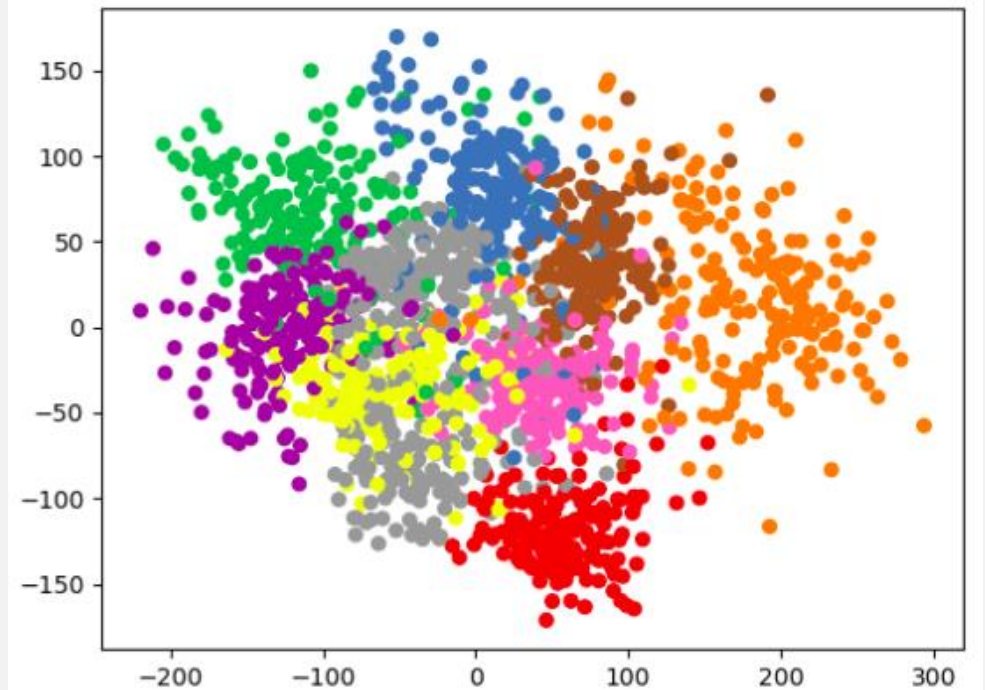
- Se eliminan los atributos review title y review content

Métodos Utilizados

Regresión Lineal

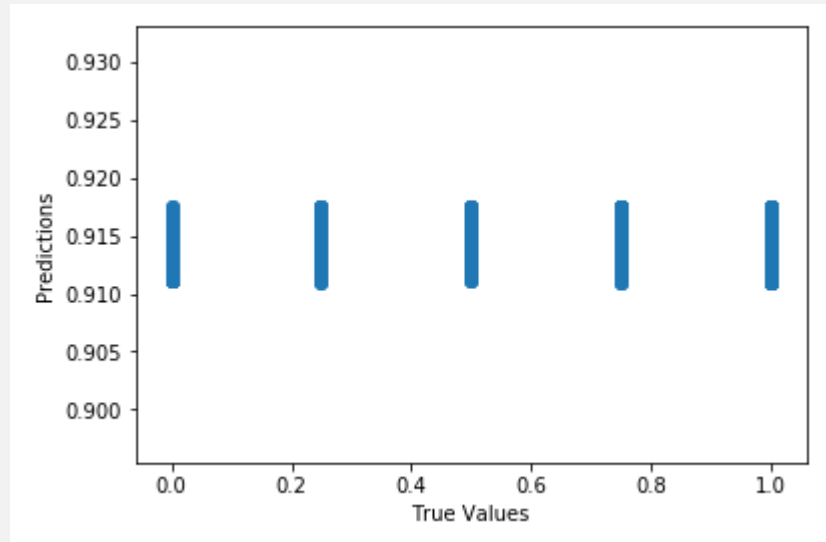


KNN



Resultados Regresión (Lineal)

Predictions-True Values



MSE

0.045

SCORE

0.00026

Se transforman los ratings en tres categorías (mala, normal buena), para utilizar algoritmos de clasificación

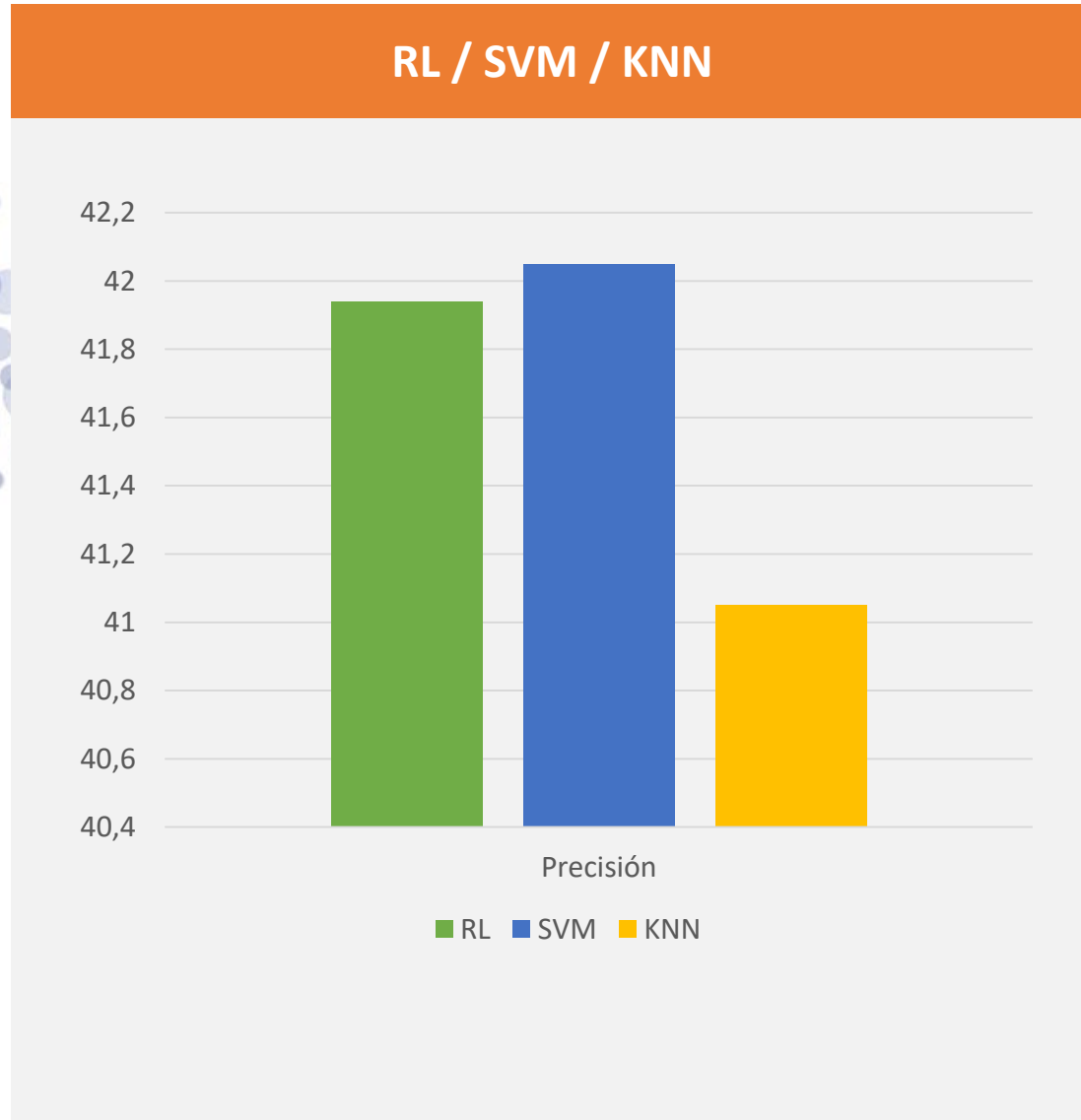
Resultados

RL	Precision	Recall	F1-score	Support	Train set accuracy	Test set accuracy
0	0.34	0.43	0.38	783	46.30%	41.94%
1	0.03	0.33	0.05	58		
2	0.75	0.42	0.54	1853		

SVC	Precision	Recall	F1-score	Support	Train set accuracy	Test set accuracy
0	0.34	0.43	0.38	779	46.15%	42.05%
1	0.03	0.34	0.05	56	Kernel: RBF	
2	0.76	0.42	0.54	1859	C=1.0	

KNN	Precision	Recall	F1-score	Support	Train set accuracy	Test set accuracy
0	0.71	0.40	0.51	1758	43.70%	41.05%
1	0.03	0.24	0.05	82	N_neighbors=3	
2	0.38	0.46	0.41	854		

Comparativa RL, SVM, KNN





Conclusiones

01

El mejor clasificador depende de la naturaleza del problema

02

Muy difícil clasificar según categorías generales

03

El reducido numero de variables afecta a la complejidad del problema

04

No se puede determinar la reseña de un producto por su categoría