Diplomatura en Ciencia de Datos, Aprendizaje Automático y sus Aplicaciones 2019

ESTIMACIÓN DE PESO Y DIMENSIONES DE LOS ENVÍOS DE MERCADO LIBRE

Sergio Martín Buzzi

Mentor: Diego Piloni

TP 1: Análisis y Visualización de Datos

Variables

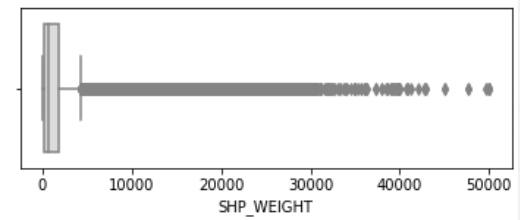
- ITEM_ID: id unívoco de cada item publicado. (Ofuscado)
- **SHP_WEIGHT**: peso del paquete informado por el correo.
- **SHP_LENGTH**: largo del paquete informado por el correo.
- **SHP_WIDTH**: ancho del paquete informado por el correo.
- **SHP_HEIGHT**: altura del paquete informado por el correo.
- ATTRIBUTES: atributos como marca y modelo, entre otros, en formato json-lines.
- CATALOG_PRODUCT_ID: id del catálogo (ofuscado).
- CONDITION: condición de venta (nuevo o usado).
- DOMAIN_ID: id de la categoría a la que pertenece la publicación.
- PRICE: precio en reales.
- SELLER_ID: id del vendedor (ofuscado).
- STATUS: estado de la publicación (activa, cerrada, pausada, etc.)
- TITLE: título de la publicación.

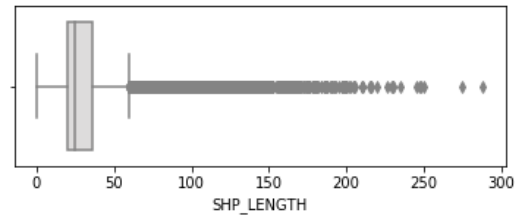
Dataset

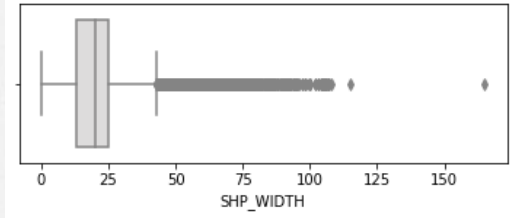
ITEM_ID	SHP_WEIGHT	SHP_LENGTH	SHP_WIDTH	SHP_HEIGHT	ATTRIBUTES	CATALOG_PRODUCT_ID	CONDITION	DOMAIN_ID	PRICE	SELLER_ID	TITLE
R49NJK99F7	300	22	19	8	[{'id': 'BLOUSE_MATERIAL', 'name': 'Material d...	H53U1H7Q5G	new	MLB-BLOUSES	26.99	T77B008Y9D	Blusa Cirre Manga Longa
ID6G5YZ2B7	49	0	0	0	[{'id': 'BRAND', 'name': 'Marca', 'value_id': ...	PE74WC2QH9	new	MLB-INK_CARTRIDGES	69.9	Otro	Cartucho Lexmark Original 100xl Amarelo 10,6ml
H3JUCUHEOF	220	25	15	11	[{'id': 'ACCESSORIES_INCLUDED', 'name': 'Acess...	H53U1H7Q5G	new	Otro	42.8	Otro	Case Hd 2,5 Usb 3.0 Sata Hdd Ssd Gaveta Note F...
H99ZF9RONF	680	29	19	10	[{'id': 'BRAND', 'name': 'Marca', 'value_id': ...	H53U1H7Q5G	new	Otro	139.99	Otro	Tênis Lilica Ripilica Menina Preto - Original ...
SAJVUQHROW	218	40	20	5	[{'id': 'BRAND', 'name': 'Marca', 'value_id': ...	H53U1H7Q5G	new	MLB-LEARNING_TOYS	39.9	G2MKR2YZSM	Calendário De Rotina Semanal Personalizado - G...
...	...	...	...	...	...	...	...	...	...	...	...

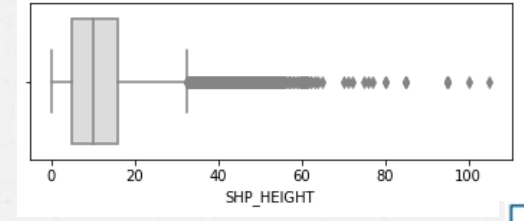
UNC FAMAF Córdoba Technology Cluster mercado libre

Etiquetas





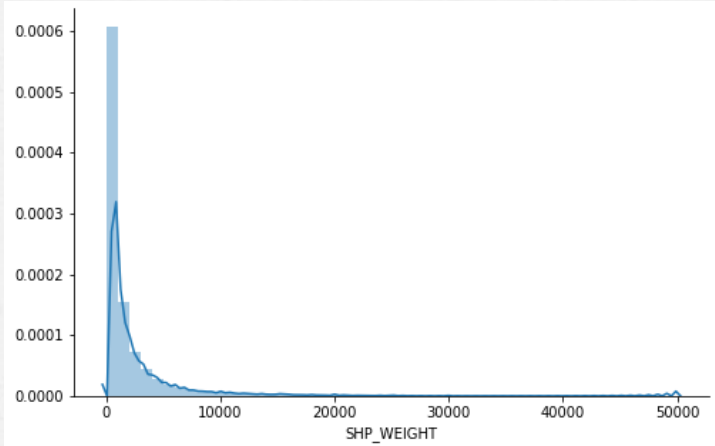


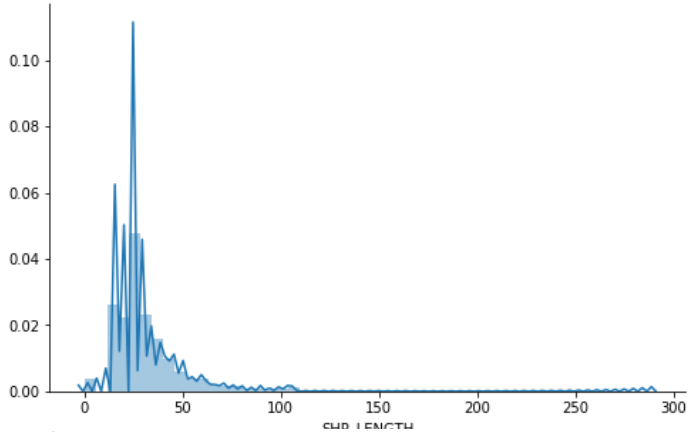


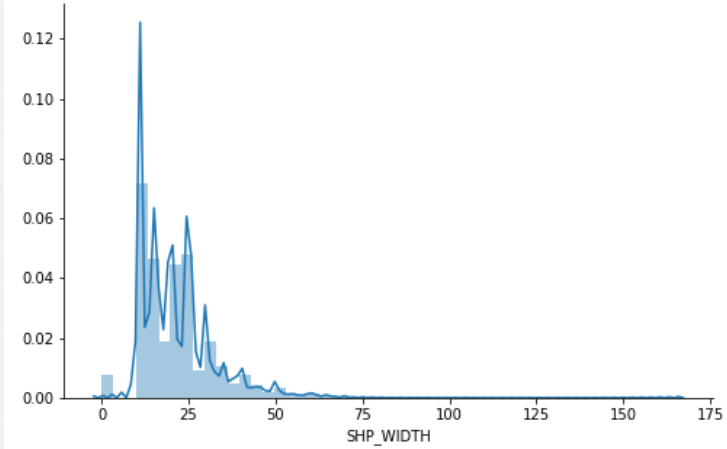
	SHP_WEIGHT	SHP_LENGTH	SHP_WIDTH	SHP_HEIGHT
count	347751	347751	347751	347751
mean	1854.78	31.451148	21.29	11.57
std	3257.88	18.33	11.21	8.31
min	1	0	0	0
25%	260	20	13	5
50%	675	25	20	10
75%	1900	36	25	16
max	50000	288.2	165	105

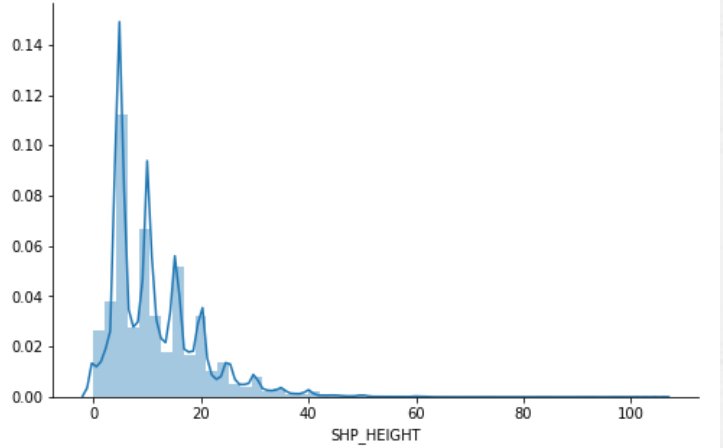
UNC FAMAF Córdoba Technology Cluster mercado libre

Etiquetas



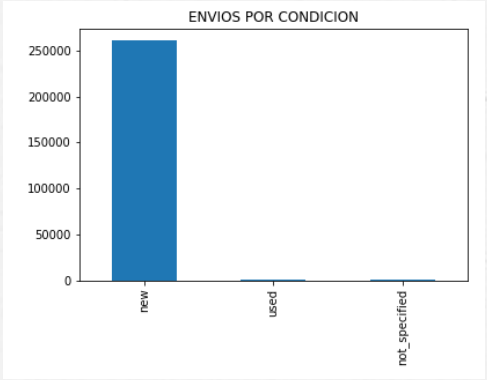


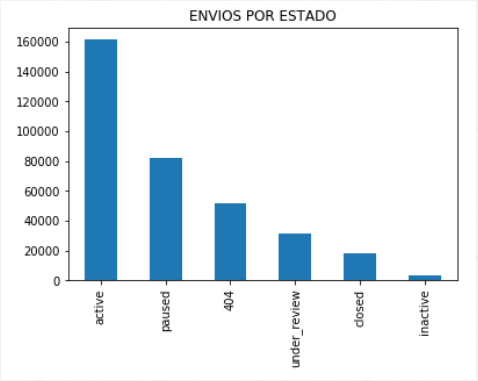




UNC FAMAF Córdoba Technology Cluster mercado libre

Features

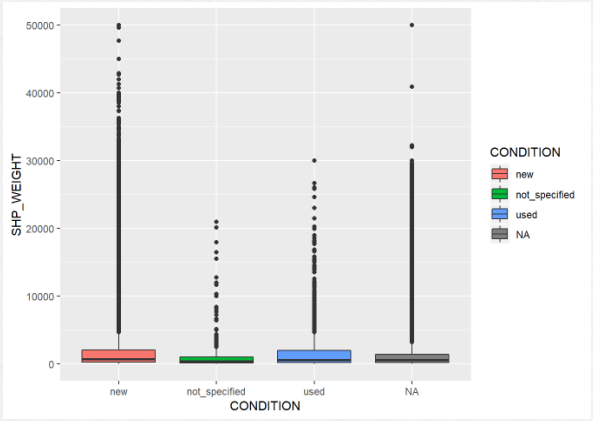


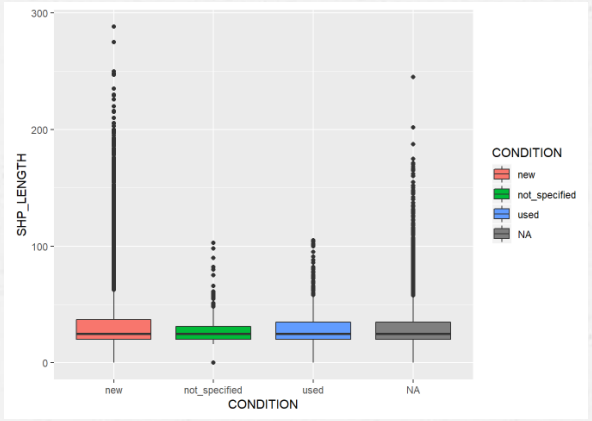


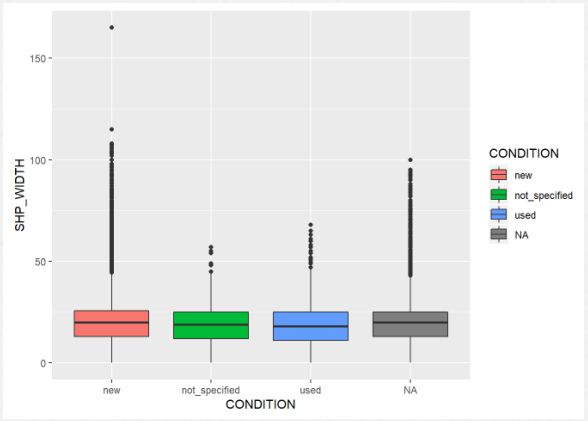
ATTRIBUTES	CATALOG_PRODUCT_ID	CONDITION	DOMAIN_ID	PRICE	SELLER_ID	TITLE
[{'id': 'BLOUSE_MATERIAL', 'name': 'Material d...	H53U1H7Q5G	new	MLB-BLOUSES	26.99	T77B008Y9D	Blusa Cirre Manga Longa
[{'id': 'BRAND', 'name': 'Marca', 'value_id': ...	PE74WC2QH9	new	MLB-INK_CARTRIDGES	69.9	Otro	Cartucho Lexmark Original 100xl Amarelo 10,6ml
[{'id': 'ACCESSORIES_INCLUDED', 'name': 'Acess...	H53U1H7Q5G	new	Otro	42.8	Otro	Case Hd 2,5 Usb 3.0 Sata Hdd Ssd Gaveta Note F...
[{'id': 'BRAND', 'name': 'Marca', 'value_id': ...	H53U1H7Q5G	new	Otro	139.99	Otro	Tênis Lilica Ripilica Menina Preto - Original ...
[{'id': 'BRAND', 'name': 'Marca', 'value_id': ...	H53U1H7Q5G	new	MLB-LEARNING_TOYS	39.9	G2MKR2YZSM	Calendário De Rotina Semanal Personalizado - G...
...	...	...	...	...	...	...

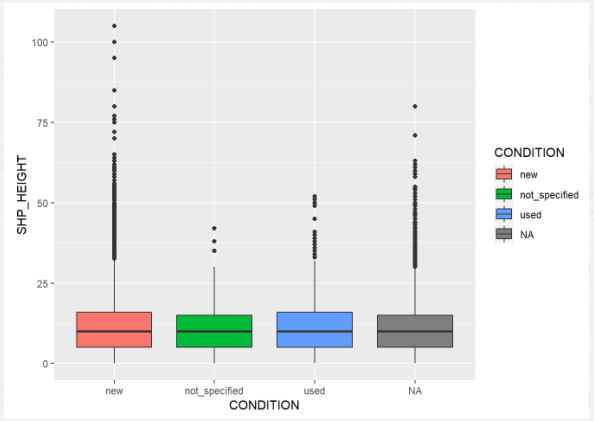
UNC FAMAF Córdoba Technology Cluster mercado libre

Features









UNC FAMAF Córdoba Technology Cluster mercado libre

TP 2: Análisis y Curación de Datos

Resumen de tareas

- Eliminación de registros con STATUS='404'.
- Eliminación de registros con faltantes en las variables 'SHP'.
- Agrupación por ITEM_ID y reemplazo por mediana de peso y medidas.
- Imputación de faltantes de PRICE por KNN. Leakage!
- Parseo de ATRIBUTES >>> BRAND y MODEL.
- OneHotEncoding sobre: CONDITION, DOMAIN_ID, SELLER_ID, CATALOG_PRODUCT_ID, BRAND y MODEL. Ingeniería de features.
- Tratamiento de outliers.

One Hot Encoding

FEATURES	VALORES ÚNICOS
CATALOG_PRODUCT_ID	9939
CONDITION	3
DOMAIN_ID	2572
SELLER_ID	33015
BRAND	39536
MODEL	96902

Total de Registros **124927**

One Hot Encoding

FEATURES	VALORES ÚNICOS
CATALOG_PRODUCT_ID	9939
CONDITION	3
DOMAIN_ID	2572
SELLER_ID	33015
BRAND	39536
MODEL	96902
Total de Registros	**124927**

CATALOG_PRODUCT_ID:
Una categoría representa casi el 90%

UNC FAMAF Córdoba Technology Cluster mercado libre

One Hot Encoding

FEATURES	VALORES ÚNICOS
CATALOG_PRODUCT_ID	9939
CONDITION	3
DOMAIN_ID	2572
SELLER_ID	33015
BRAND	39536
MODEL	96902

CATALOG_PRODUCT_ID: Una categoría representa casi el 90%

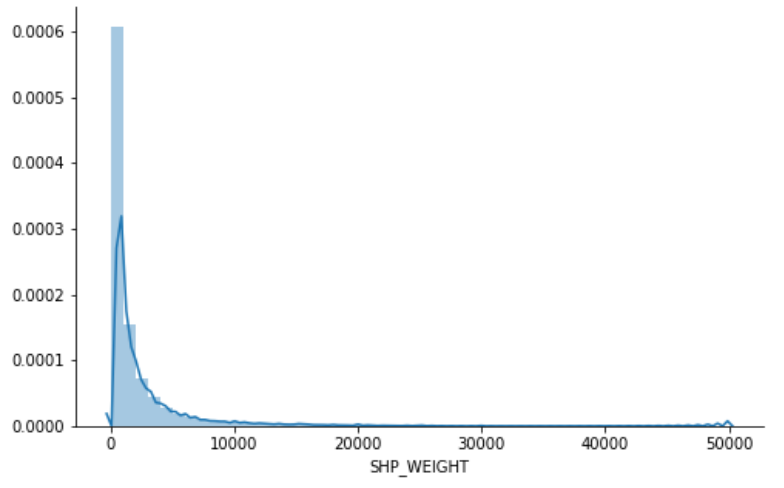
Se usan las categorías mas frecuentes

Total de Registros **124927**

UNC FAMAF Córdoba Technology Cluster mercado libre

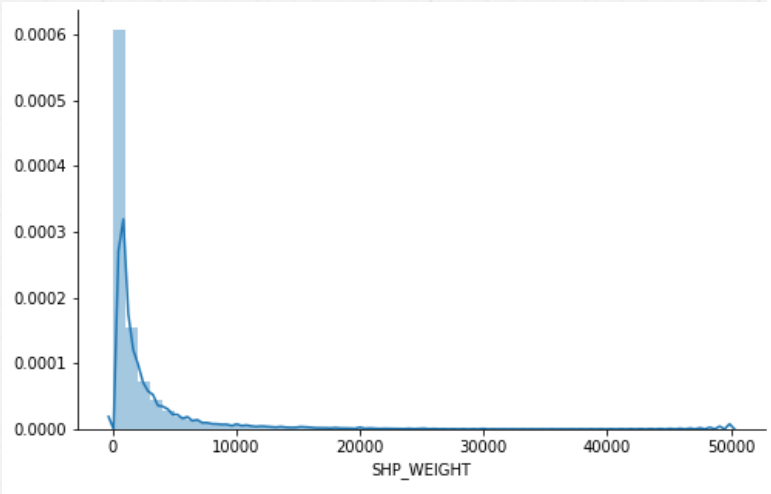
Outliers

Datos originales: SHP_WEIGHT

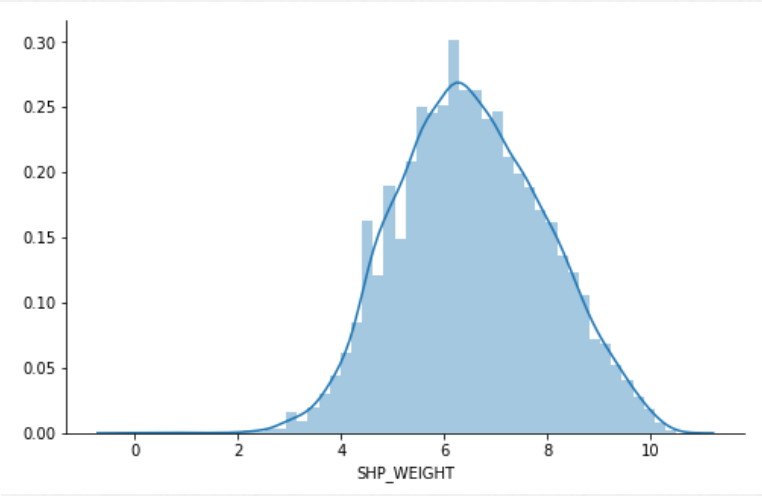


Outliers?

Datos originales: SHP_WEIGHT



Transformación: log(SHP_WEIGHT)



UNC FAMAF Córdoba Technology Cluster mercado libre

TP 3: Introducción al Aprendizaje Supervisado

Modelos de regresión

- Ridge
- Lasso
- Elastic Net
- KNN

TP 3: Introducción al Aprendizaje Supervisado

Modelos de regresión

- Ridge
- Lasso
- Elastic Net
- KNN

Modelo	MAE_train	MAE_test
Ridge	0.72	0.94
LASSO	1.1	1.11
Elastic Net	1.09	1.11
KNN	0.89	1.03

Media de log(SHP_WEIGHT) = 6.53

UNC FAMAF Córdoba Technology Cluster mercado libre

TP 4: Aprendizaje Supervisado

Modelos de regresión

- Random forest
- Gradient Boosting
- Red neuronal (1 capa)
- MLP

TP 4: Aprendizaje Supervisado

Modelos de regresión

- Random forest
- Gradient Boosting
- Red neuronal (1 capa)
- MLP

} Multioutput & univariante

Resultados

	Modelos Multivariantes					
	Random forest		Perceptron (1 capa)		MLP	
	MAE_train	MAE_test	MAE_train	MAE_test	MAE_train	MAE_test
SHP_WEIGHT	876.56	1518.36	1959.22	1837.00	1024.11	1572.17
SHP_LENGTH	7.99	10.61	45.39	45.64	12.34	12.43
SHP_WIDTH	5.85	7.85	31.67	31.94	12.27	11.81
SHP_HEIGHT	4.58	6.06	17.10	16.27	11.43	11.67

	Medias
SHP_WEIGHT	1854.78
SHP_LENGTH	31.45
SHP_WIDTH	21.29
SHP_HEIGHT	11.57

	Modelos univariantes					
	Random forest		Perceptron (1 capa)		Gradient boosting	
	MAE_train	MAE_test	MAE_train	MAE_test	MAE_train	MAE_test
SHP_WEIGHT	876.00	1513.42	1901.35	1778.96	1515.23	1453.33
SHP_LENGTH	7.01	10.66	11.90	11.57	10.47	10.33
SHP_WIDTH	5.01	7.77	8.16	7.98	7.67	7.66
SHP_HEIGHT	4.01	6.30	5.95	5.95	5.80	6.17

UNC FAMAF Córdoba Technology Cluster mercado libre

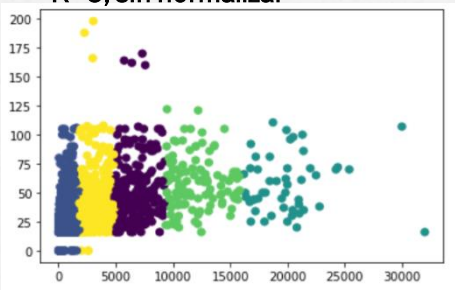
TP 5: Aprendizaje no Supervisado

Resumen

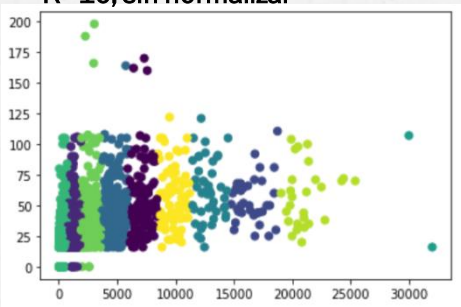
- PCA sobre el conjunto de features
- K- means sobre SHP_WEIGHT y SHP_LENGHT, con y sin normalización
- Mixtura de Gaussianas

K-means (SHP_WEIGHT y SHP_LENGHT)

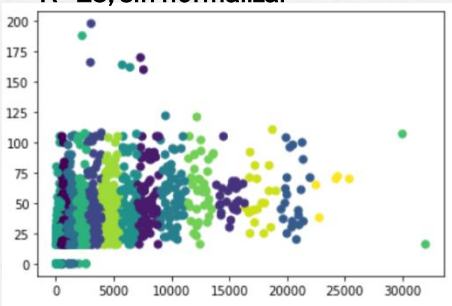
K=5, sin normalizar



K=10, sin normalizar

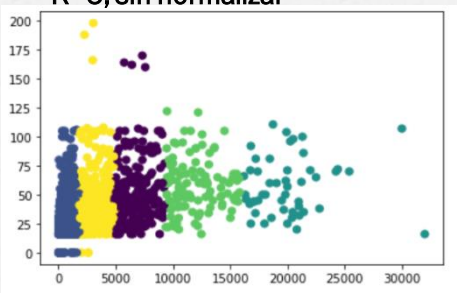


K=15, sin normalizar

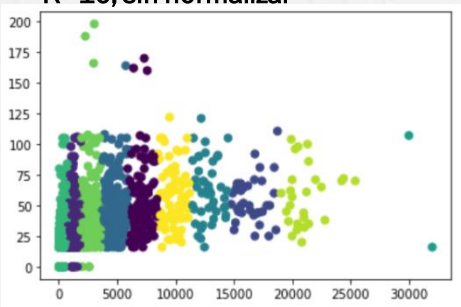


K-means (SHP_WEIGHT y SHP_LENGHT)

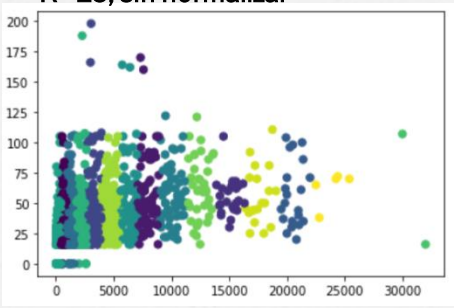
K=5, sin normalizar



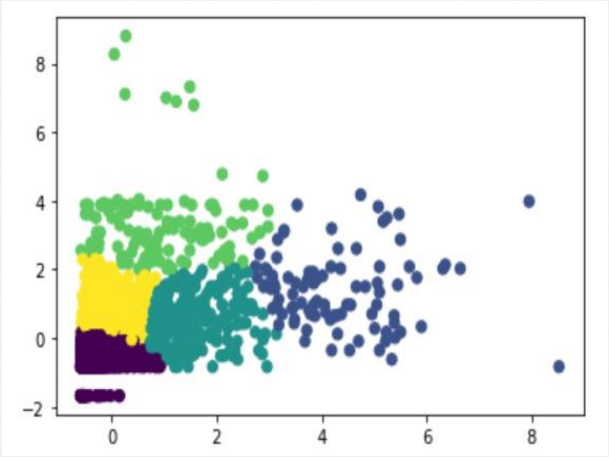
K=10, sin normalizar



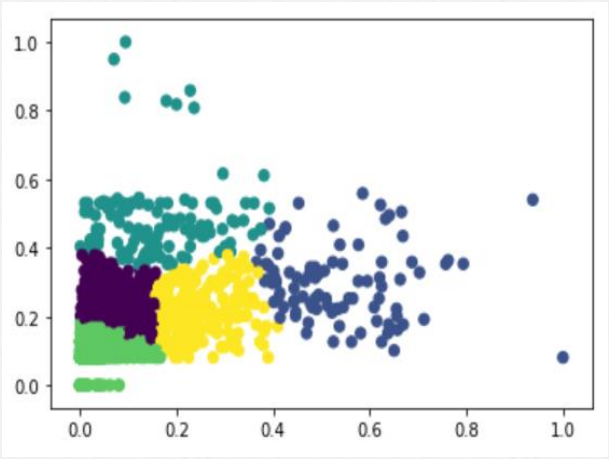
K=15, sin normalizar



K=5, StandardScaler

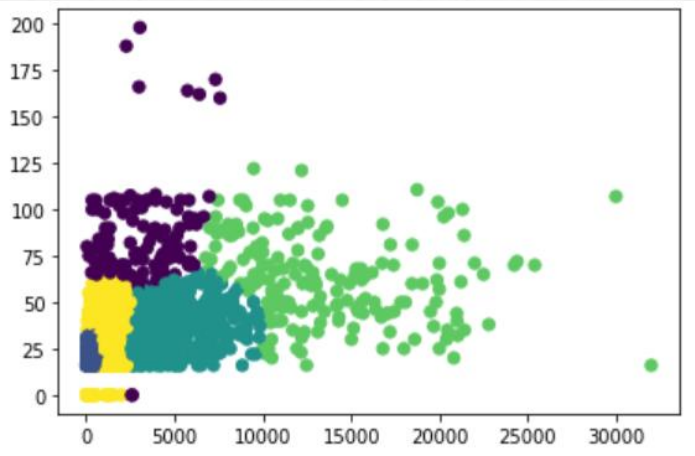


K=5, MinMaxScaler

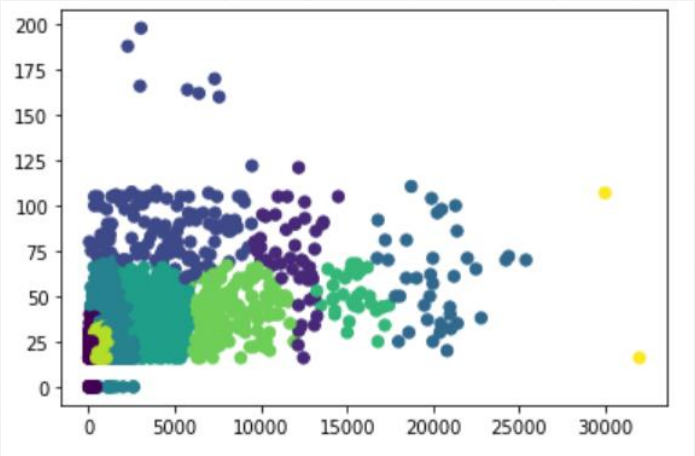


Mixtura de Gaussianas (SHP_WEIGHT y SHP_LENGHT)

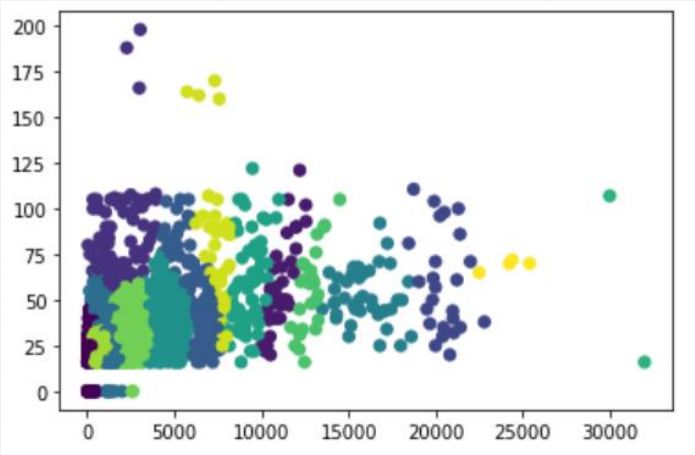
K=5



K=10

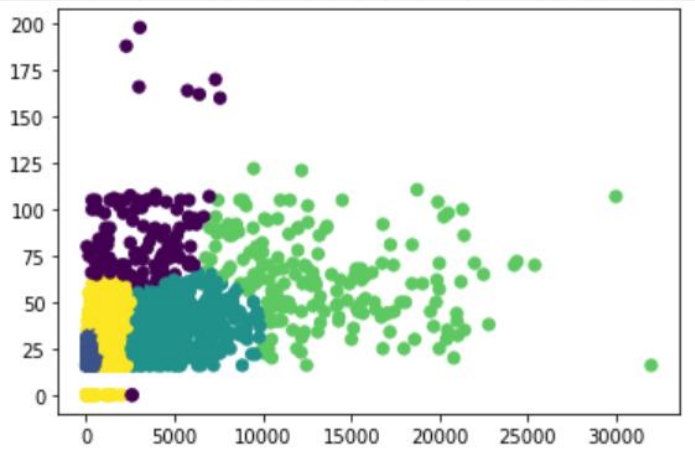


K=15

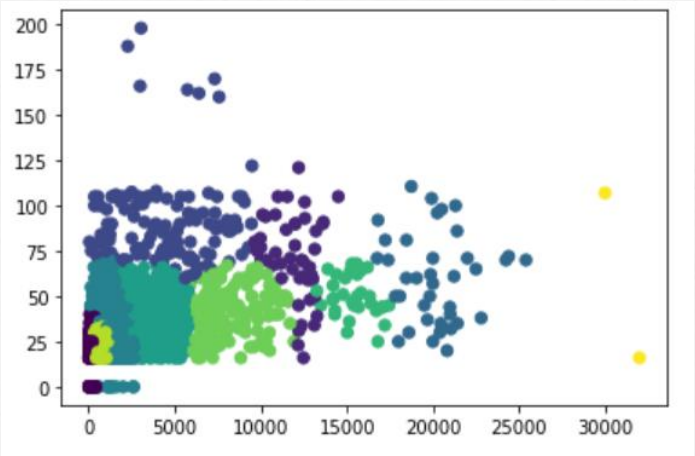


Mixtura de Gaussianas (SHP_WEIGHT y SHP_LENGHT)

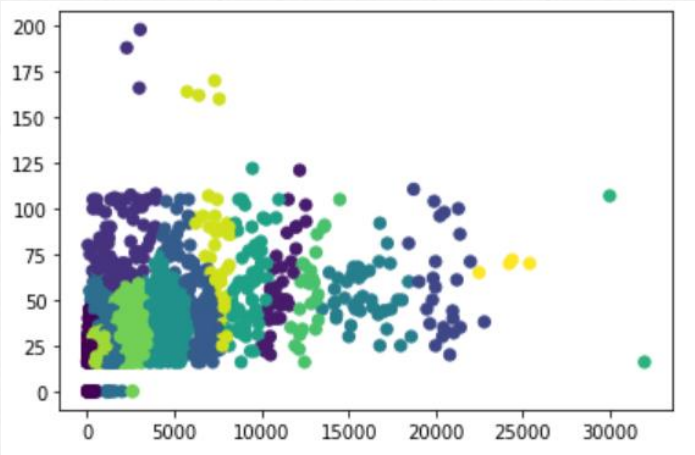
K=5



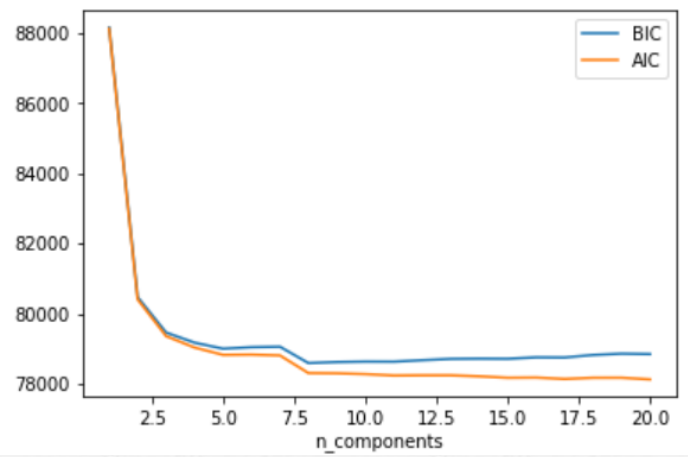
K=10



K=15



Criterios de información



Para seguir trabajando…

- Correr con todo el dataset.
- Feature importance.
- Incorporación del título (NLP).
- Comparar con modelos de clasificación.
- Estimación directa del costo de los envíos.

GRACIAS!