Aplicación de técnicas de clustering en datos de robos de autos

Trabajo Práctico Final

Cluster AI- 2021

Cifuentes, Juan Martín

Kalousek, Santiago

Resumen

Se buscará explicar relaciones existentes entre variables mediante aprendizaje no supervisado y algoritmos de clustering en robos de autos en Provincia de Buenos Aires durante 2020

Palabras claves

Clustering, robos, autos, EDA, PSA

1 INTRODUCCION

El robo de autos es una problemática que afecta con frecuencia a usuarios de los mismos y lo que se buscó en este trabajo es encontrar alguna posible relación entre los robos efectuados durante 2020 y las características de los vehículos robados, lugar y época de los hechos ocurridos, para investigar mediante técnicas de aprendizaje no supervisado,, usando algoritmos de clustering, si existe alguna relación que pueda explicar la mayor o menor ocurrencia de los mismos y determinados patrones de ocurrencia.

Para poner en contexto, según datos de la DNRPA[i], en 2019 se denunciaron 36833 robos, el número más alto en 12 años y en 2020, el año investigado, que si bien hubo una considerable baja (24199 robos), esta se puede explicar por la menor circulación producto de la pandemia y sus restricciones a la circulación, ya que los primeros dos meses fue dónde más robos ocurrieron, mismos meses dónde aún las restricciones no existían.

Según la AFAC en 2019* había una flota circulante de 14.301.842, por lo que esos 36833 representan un 0,25% del total de autos en el periodo de un año.

Por la característica del dataset con el que nos dispusimos a trabajar, consideramos que algunas de las herramientas provistas por la cátedra, como los modelos de regresión, no son adecuados para analizar este tipo de datos.

2 ANALISIS EXPLORATORIO DE DATOS (EDA)

2.1 Selección y limpieza de datos

Nuestro trabajo comenzó con la exploración de datos, para ello primero importamos todos los datasets provistos por Datos Argentina correspondientes a robo de autos en 2020[ii], los cuales correspondían a los 12 meses del año 2019, y los concatenamos para que pertenezcan a una misma base de datos. Esta concatenación nos reporto un total de 24546 samples con 25 features, que corresponden a todos los hechos denunciados y los casos de recupero de automoviles en 2020 con sus características. Luego se procedió a eliminar aquellas features que consideramos que no eran relevantes para el estudio, eliminando 18 de las mismas y posteriormente agregando la de precio nominal.

registro_seccional_descripcion	automotor_anio_modelo	automotor_marca_descripcion	Precio_nominal
Lugar donde se denuncion el hecho	Fecha de fabricación del auto	Marca del auto	Precio a valores actuales

titular_genero	titular_anio_nacimiento	mes
Género	Año de nacimiento dueño	Mes de la denuncia

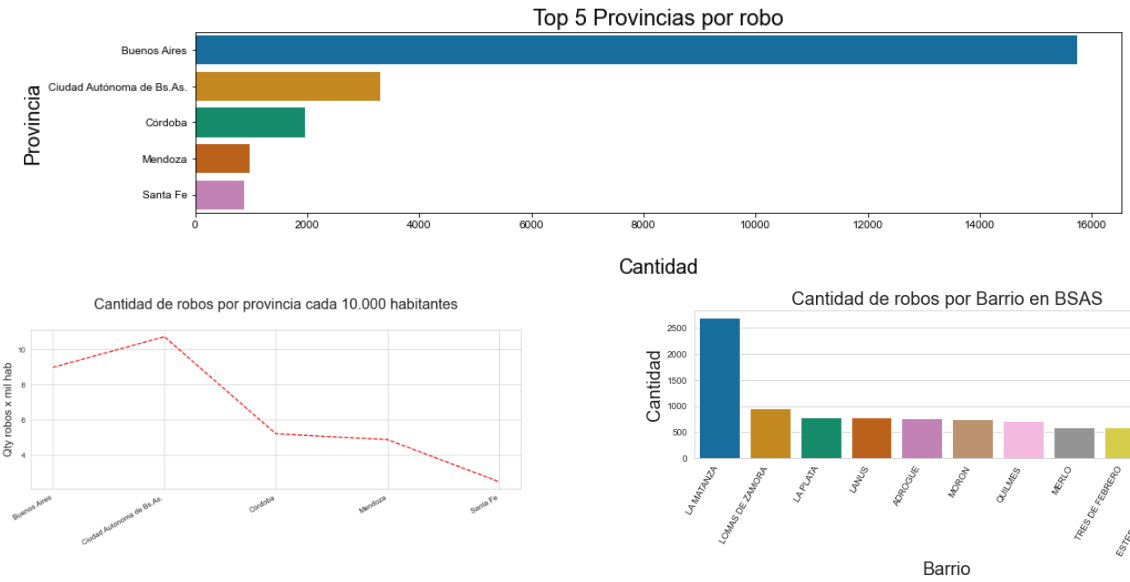
El siguiente paso fue eliminar todos registros con los valores nulos (1100 samples) y los casos de recupero de automóviles incluidos en el dataset ya que solo nos interesaban los casos de robo.

Luego procedimos a explorar las marcas de los robos de automóviles registrados. Notamos que muchas veces la misma estaba escrita de formas diferente o algunos casos no se podían interpretar ya que no correspondían a marcas reconocibles. Unificamos las marcas para que estén escritas de la misma manera y aquellos registros de marcas que no se podían identificar los eliminamos del dataset. Se realizo un trabajo similar para las seccionales donde se denunciaron los robos y los modelos de autos.

2.2 Visualización de datos

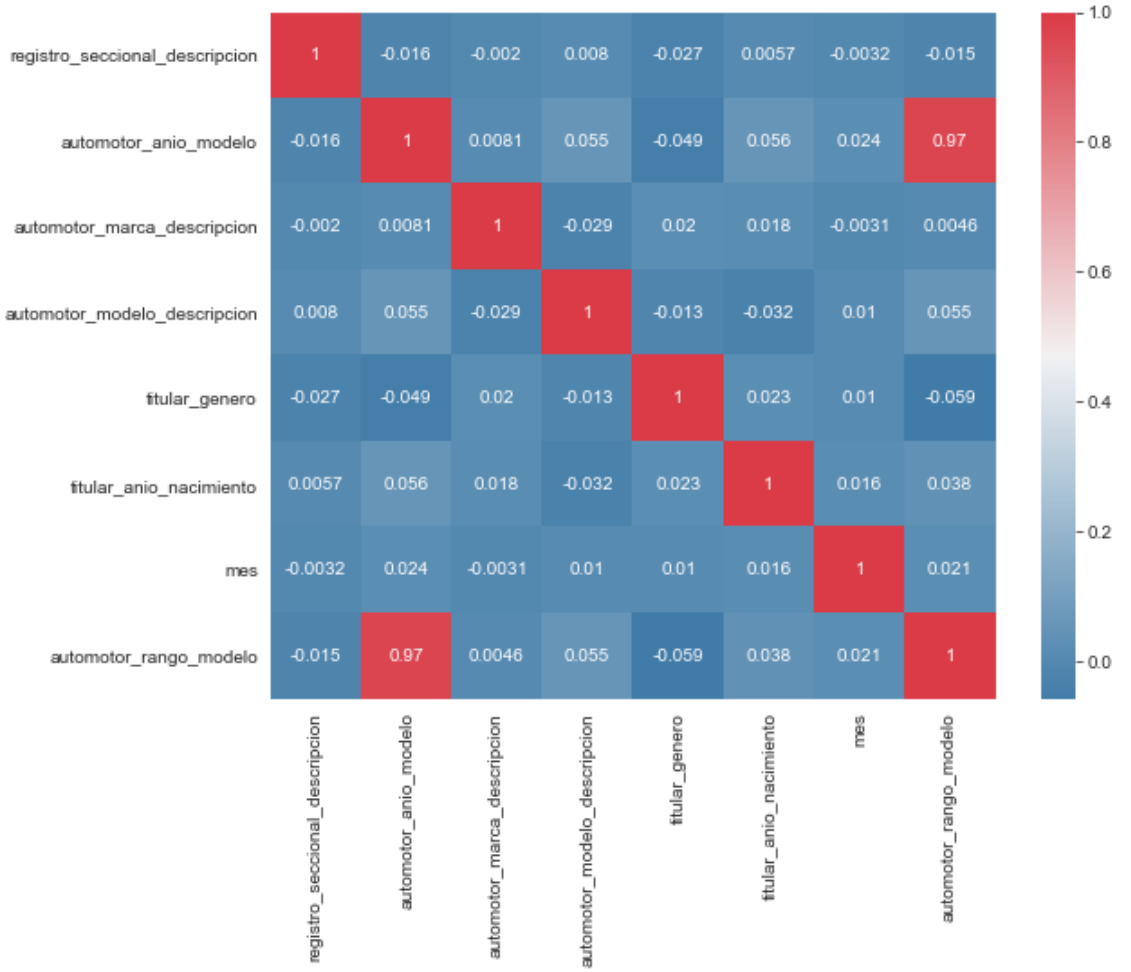
Comenzamos por averiguar en qué zona del país ocurría la mayor cantidad de denuncias de robos de autos, observando que claramente ocurre la mayor cantidad en la Provincia de Buenos Aires. Por esta razón decidimos centrar nuestro trabajo en esta zona, para poder acercarnos a obtener conclusiones más precisas.

También realizamos gráficos cargando otro dataset para poder medir cantidad de robos denunciados por cantidad de habitantes y un top 10 de zonas dentro de la misma dónde se produjo la mayor cantidad de denuncias dentro de la Provincia de Buenos Aires.



Se agrego una nueva feature con el precio de los autos del dataset utilizado. Con este nuevo dato se realizó un boxplot entre el top 10 de marcas con mayores robos vs su respectivo precio

Cerramos esta etapa realizando una correlación de Pearson para poder observar qué relación existe entre las distintas features que decidimos utilizar para la base de datos.



3 Materiales y Métodos

3.1 Clustering

A la hora de plantear algoritmos de clustering se propuso trabajar con el dataset obtenido luego del EDA dado la baja cantidad de features finalmente obtenidos. A este data set lo sometidos a dos algoritmos de clustering distintos como son el K-Means y Clustering Jerárquico, se estudiaron los resultados utilizando distintos números de cluster, a su vez se aplicaron valores escalados y sin escalar, finalmente se aplico un modelo de reducción de la dimensionalidad para poder representar los resultados en dos dimensiones.

Algoritmo K-Means[iiiiv]: Cada cluster estará identificado por un centroide.

● Una muestra será asignada al cluster cuyo centroide este más cerca. Para medir la similaridad entre muestras se utiliza la distancia euclidiana cuadrática

$$d(x_i, x_{i'}) = \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 = ||x_i - x_{i'}||^2$$

Cada muestra Xi será asignada al cluster "k" que presente la distancia cuadrática más cercana con su centroide "m".

$$C(i) = \underset{1 \le k \le K}{\operatorname{argmin}} ||x_i - m_k||^2$$

● El algoritmo de k-Means es iterativo: durante el proceso de aprendizaje los centroides se recalculan y en consecuencia la pertenencia de las muestras en los clusters.

Algoritmo de Clustering Jerárquico[v]

● El algoritmo construye representaciones jerárquicas en donde los clusters de cada nivel de

jerarquía son creados agrupando los clusters del nivel inmediatamente inferior, donde en el nivel más bajo posible, cada cluster contiene una sola muestra.

● El clustering jerárquico aglomerativo implica agrupar desde el nivel inferior de a dos clusters para formar uno en el nivel inmediato superior. El par de clusters seleccionados para

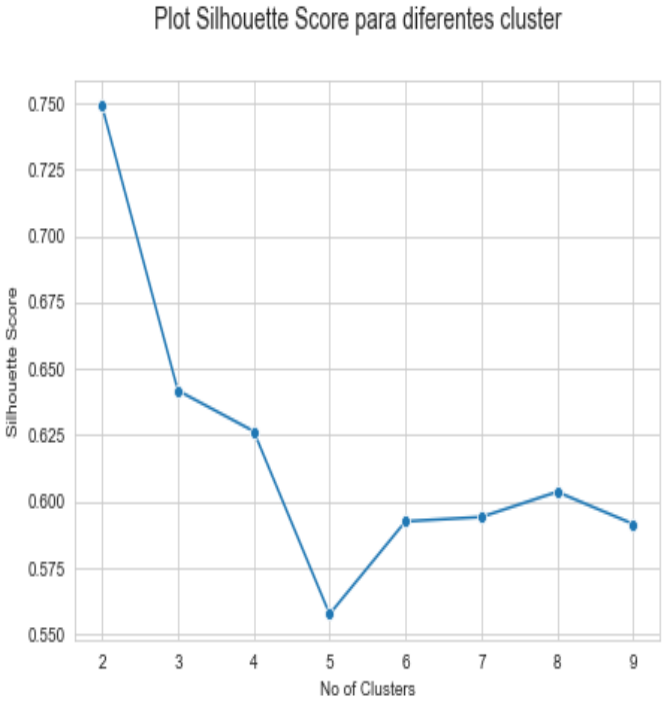
Evaluación y Métricas

Al haber utilizado un método de aprendizaje no supervisado no contábamos con las etiquetas de nuestros datos, por lo que, para la medición de los resultados utilizamos únicamente el Silhouttte Score. Este índice mide cuan similar es una muestra respecto a su propio clúster (cohesión) comparado con el resto de los clúster (separación) y varía entre -1 y 1, mientras mas cercano a 1 significa que los clusters asignados están claramente separados y definidos

No se pudo medir los resultados a través del rand índex ya que no contábamos con las etiquetas reales de los registros. Esta métrica se utiliza para validar, luego del aprendizaje, la calidad de los clusters obtenidos.

Resultados

De la tabla 2 se puede observar que el mejor Silhouttte Score obtenido corresponde al modelo realizado de K-Means con los datos sin escalar utilizando 2 clusters, donde el valor obtenido es de 0.7489.

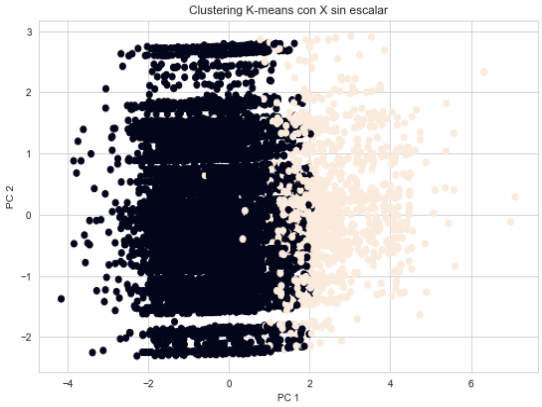


.

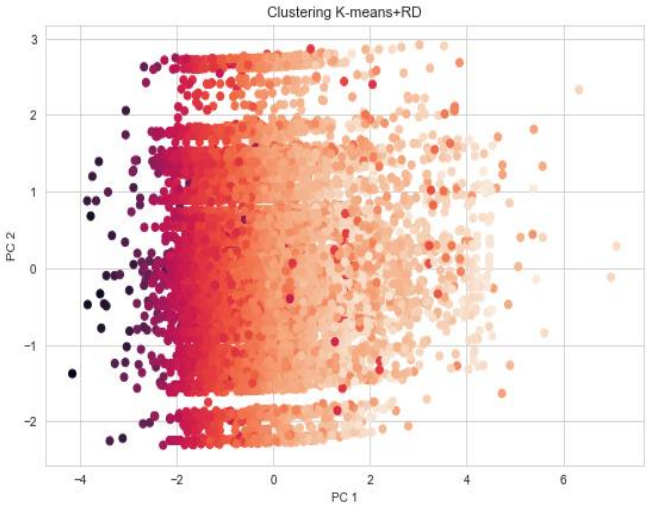
A su vez se puede observar que los resultados también son mejores cuando se utiliza el método de clustering jerárquico para datos sin escalar.

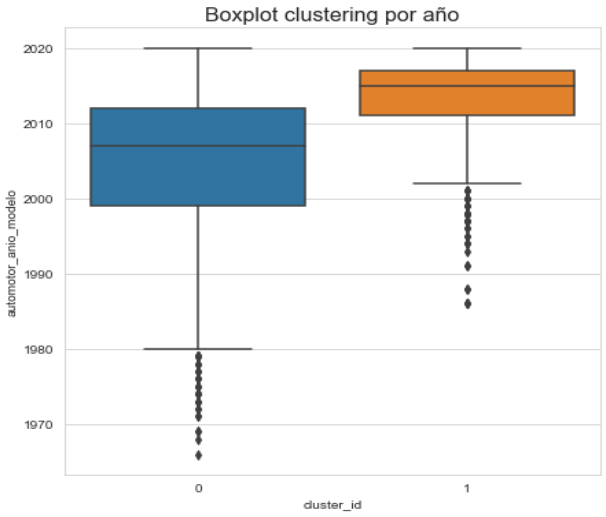
Características del modelo	N° Clusters	Silhouttte Score
K-Means sin escalar los datos	2	0.7489
	3	0.6381
	4	0.6173
	5	0.5601
K-Means con datos escalados	2	0.2309
	3	0.2903
	4	0.3114
	5	0.3125
Clustering Jerárquico sin escalar	5	0.5647
	6	0.5837
	7	0.6036
	8	0.6195
Clustering Jerárquico con datos escalados	5	0.2766
	6	0.2635
	7	0.2818
	8	0.2871
PCA + K-Means con 2 PC	2	0.3322

Luego de este análisis se tomó el método de aprendizaje no supervisado de K-Means de dos clusters con datos sin escalar para poder realizar las conclusiones. Para esto se identificaron los labels y centroides correspondientes y se obtuvieron la cantidad de muestras por cluster. Luego se realizó un análisis de componentes principales tomando únicamente dos dimensiones para poder visualizar los clusters obtenido.

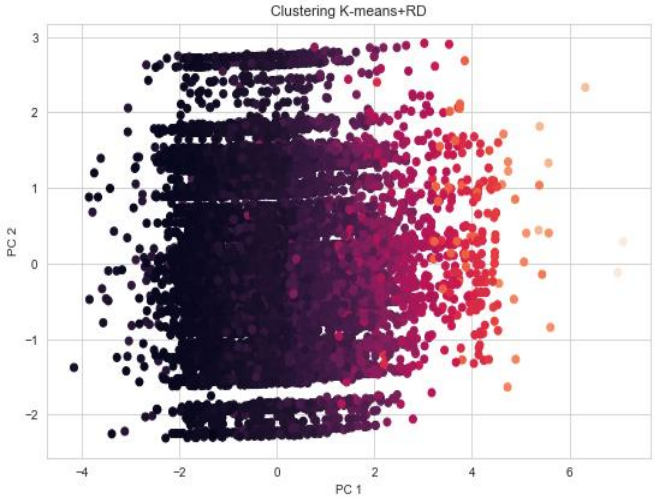


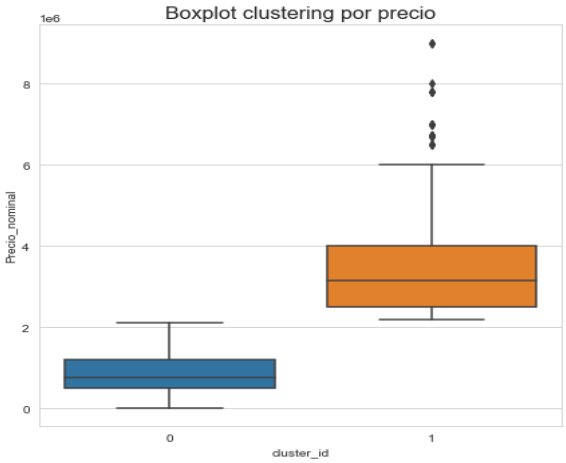
Visualización influencia del año del auto en el método de clusterización





Visualización influencia del precio del auto en el método de clusterización

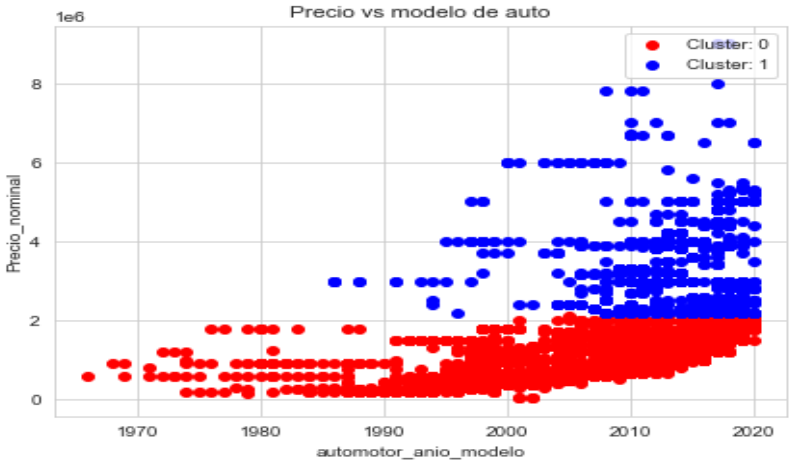




Conclusiones

Se observo que se obtiene el mejor modelo cuando se aplica K-Means para datos no escalados y se utilizan dos clusters, donde estos clusters tienen la siguientes características:

- Cluster 1: Los registros dentro de este clúster se caracterizan por ser más antiguos, dado que el 75% de los registros son previos al 2010 y más económicos, ya que el 100% de los datos tienen un precio menor a 2.000.000 de pesos.
- Cluster 2: Los autos dentro de este clúster se caracterizan por ser más nuevos, dado que el 75% de los registros son posteriores al 2010 y son más caros, ya que el 100% de los datos tienen un precio mayor a 2.000.000 de pesos.



Bibliografía:

[i] https://datos.gob.ar/dataset/justicia-robos-recuperos-autos (DNRPA, s.f.)

[ii]http://datos.jus.gob.ar/dataset/robos-y-recuperos-de-autos/archivo/1adf676e-d6cc-4787-a700-4289b7c07ef0 (DNRPA, s.f.)

[iii] Jain, Anil K. (2010) “Data clustering: 50 years beyond K-means” in *Pattern Recognition Letters,* 31(8), 651–666. Michigan State University

[iv] T. Hastie , R. Tibshirani y J. Friedman (2009) The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2da ed). Springer

[v] Xu, Rui y Wunsch, Donald C. (2005) “Survey of clustering algorithms” IEEE Transactions on Neural Networks, 16(3), 645–678. Institute of Electrical and Electronics Engineers