Tarea 5. Aprendizaje no supervisado

Hernández, Gerardo

22 de julio de 2024

1. Introducción

Cuando se vende un auto de segunda mano, no se sabe on certeza el precio exacto con el que se debe vender, muchas de las veces por la necesidad de venderlo lo más rápido posible se venden a precios muy por debajo de lo habitual, y ésta como algunas otras situaciones impactan financieramente, tanto a los vendedores como a los compradores del vehículo.

Entonces, lo que se busca con este proyecto es poder encontrar algún modelo que nos pueda dar el valor adecuado de un auto, de acuerdo a las características del mismo. Y a su vez, reconocer que características son las que hay que tomar en consideración.

1.1. Fuente de información

La información que se utiliza para este proyecto fue tomada de la siguiente liga, en la página de Kaggle:

https://www.kaggle.com/datasets/zeeshanlatif/used-car-price- prediction-dataset?select=train.csv

Trabajaremos más adelante la limpieza y creación de más información a partir de columnas ya establecidas.

También se está tomando como referencia el siguiente proyecto para la estructura de este documento para cumplir con la tarea 5 de la materia de Aprendizaje Automático:

https://github.com/suhasmaddali/Car-Prices-Prediction/blob/main/README.md

1.2. Librerías

Las librerías y funciones que se tendrán que coargar al procesador de python serán las siguientes:

```
import matplotlib.pyplot as plt
import numpy as np
import pandas as pd
import plotly.graph_objects as go
import seaborn as sns
import scipy.stats as stats
import statsmodels.api as sm
import warnings
```

Figura 1: Librerías

```
from matplotlib import pyplot as plt
from mlxtend.feature_selection import ExhaustiveFeatureSelector as EFS
from mlxtend.feature_selection import SequentialFeatureSelector as SFS
from scipy.cluster.hierarchy import fcluster
from scipy.cluster.hierarchy import linkage, dendrogram
from scipy.stats import multivariate_normal
from scipy.stats import normaltest
from scipy.stats import shapiro
from scipy.spatial.distance import pdist
from sklearn.decomposition import PCA
from sklearn.ensemble import RandomForestClassifier
from sklearn.feature_selection import f_regression
from sklearn.feature_selection import r_regression
from sklearn.feature_selection import VarianceThreshold
from sklearn.feature_selection import mutual_info_regression
from sklearn.feature_selection import SelectKBest, chi2,f_regression,f_classif
from sklearn.preprocessing import LabelEncoder
from sklearn.preprocessing import MinMaxScaler
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_absolute_error,mean_squared_log_error
from sklearn.metrics import r2_score, mean_squared_error, accuracy_score
from sklearn.metrics import silhouette_score
from sklearn.metrics import mean_absolute_percentage_error as mape
from sklearn.model_selection import train_test_split
from sklearn import linear_model
from sklearn.svm import SVR
from sklearn.neighbors import KNeighborsRegressor
from sklearn.tree import DecisionTreeRegressor
```

Figura 2: Funciones

2. Descripción de los datos

Los datos que se utilizaron, es una lista de registros de autos vendidos con 12 columnas, incluida una para del precio al que se vendió el auto.

2.1. Estructura de los datos

Los datos constan de 54,273 refistros (filas) y 12 variables (columnas).

A continuación, se describen las variables:

Nombre	ID_Variable	Descripción	Ejemplo
ID	id	Es solo un identificador de los registros.	11
Marca	brand	Es el nombre de la marca del auto.	Chevrolet
Modelo	model	Es el nombre del modelo del auto.	Tahoe Premier
Año del modelo	model_year	Es el año en que se lanzó el auto a la venta.	2017

Nombre	ID_Variable	Descripción	Ejemplo
Kilometraje	milage	Es el kilometraje que tenía el auto en el momento de la venta.	92728
Tipo de combustible	fuel_type	Es el tipo de combustible que maneja el auto.	Gasoline
Motos	engine	Es el motor que tiene el auto.	355.0HP 5.3L 8 Cylinder Engine Gasoline Fuel
Transmisión	transmis-sion	Es el tipo de transmisión que tiene el auto	A/T
Color exterior	ext_col	Es el color de la parte externa del auto.	Silver
Color interior	int_col	Es el color de los interiores del auto.	Silver
Accidente	accident	Es el registro de si el auto sufrio algun accidente o no.	None reported
Titulo limpio	clean_title	Desconozco la variable, pero más adelante se eliminará ya que es el mismo dato para todos los registros.	Yes
Precio	price	Es el precio al que se vendió el auto.	49900

Como podemos observar, en el ejemplo de los datos, la mayoría de las variables son de tipo categórico, por lo que hay que tenerlo en consideración para hacer los análisis, ya sea que debamos transformarlos en índices, crear clasificaciones diferentes u obtener datos númericos de esas variables.

2.2. Análisis exploratorio de los datos

Una vez cargados los datos al procesador de python, eliminamos la columna "Título limpio" `ya` que contiene el mismo valor para todos los registros por lo que no sera significativa para los análisis posteriores y el "ID" `ya` que no representa nada más que el identificador de cada registro.

Al hacer una lectura de valores unicos de cada variable, obtenemos los siguientes resultados:

```
[ ]  autos.nunique()

id                54273
brand                53
model              1827
model_year           34
milage             3212
fuel_type             7
engine             1061
transmission         46
ext_col             260
int_col             124
accident              2
price              1481
horsepower          342
litre-capacity       61
dtype: int64
```

Donde podemos observar que no hay ninguna variable que tenga valores únicos tan grandes como los registros, por lo que podemos continuar con las variables.

En la Figura 3, se muestra el top 10 de marcas de auto que contiene la base de datos y en la Figura 4 el top 10 de modelos de auto. Esto nos hace dar cuenta que la base de datos está orientada a los autos de gama alta en el mercado, como lo es la BMW, Audi, Lexus, etc.

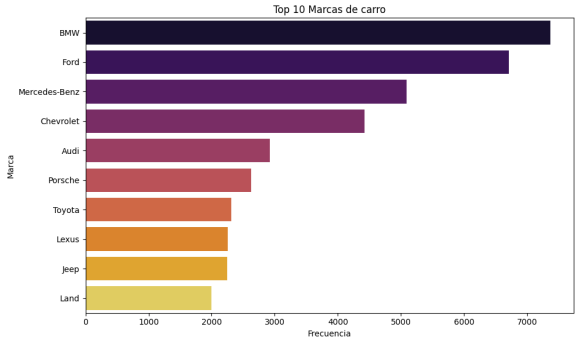

Figura 3: Estas son las 10 marcas de autos que más contiene la base de datos.

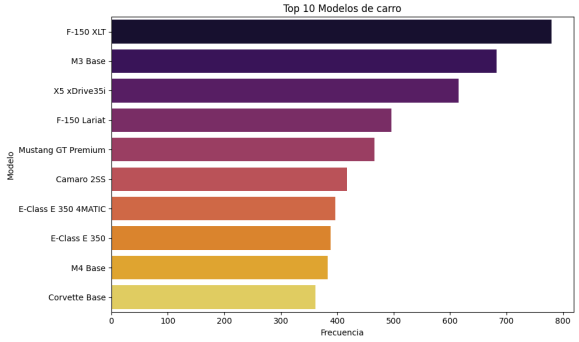

Figura 4: Estos son los 10 modelos que más contiene la base de datos.

2.3. Preprocesamiento

En el procesamiento, se crearon 2 nuevas variables para la información, las cuales fueron extraidas de la variable “Motor”. Estas variables nuevas son “Caballos de fuerza ” y “Capacidad de litros”, los caballos de fuerza miden la potencia de un auto, entre mayor sea este número, mayor es la potencia del auto, y la capacidad de litros es una medida que nos indica cuánta cantidad de aire y combustible puede consumir el motor en cada ciclo de funcionamiento.

Por ejemplo, tenemos un registro con el motor 355.0HP 5.3L 8 Cylinder Engine Gasoline Fuel, de este dato se extrae el número 355, que está antes de los caracteres ’HP’, y el número 5.3, que está antes del caracter ’L’, dandonos así los datos de los caballos de fuerza y la capacidad de litros, respectivamente.

Y para los registros que en su motor no tengan estos datos, se utilizó el promedio general de cada variable de los datos obtenidos. Enriqueciendo nuestra base de datos con 2 variables numéricas más en la base de datos.

2.4. Estadística descriptiva básica

Ya que hemos agregado las variables explicadas previamente, la base de datos cuenta ahora con 5 variables numéricas que son, "Año del modelo", "Kilometraje","Precio", "Caballos de fuerza" y "Capacidad de litros".

A continuación se obtiene el siguiente resumen de los estadísticos básicos de las variables numéricas:

	count	mean	std	min	25%	50%	75%	max	Sesgo	Moda
model_year	54273.0	2015.091979	5.588909	1974.00	2012.0	2016.0	2019.0	2024.0	-0.940515	2018.000000
milage	54273.0	72746.175667	50469.490448	100.00	32268.0	66107.0	102000.0	405000.0	0.856366	60000.000000
price	54273.0	39218.443333	72826.335535	2000.00	15500.0	28000.0	45000.0	2954083.0	23.628974	15000.000000
horsepower	54273.0	331.698323	103.936245	76.00	261.0	329.0	395.0	1020.0	0.659195	331.698323
litre-capacity	54273.0	3.717009	1.328685	0.65	3.0	3.5	4.6	8.4	0.531840	3.000000

Figura 5: Tabla de estadísticos básicos

De estas variables, podemos observar que las variables "Kilometraje" y "Precio"tienen una desviación estandar amplia, y al ver los cuartiles vemos que puede haber valores atípicos en la variable "Precio".

2.5. Correlación

La correlación nos puede proporcionar información inicial de si alguna de las variables está relacionada con alguna otra, pero más importante con la variable a predecir.

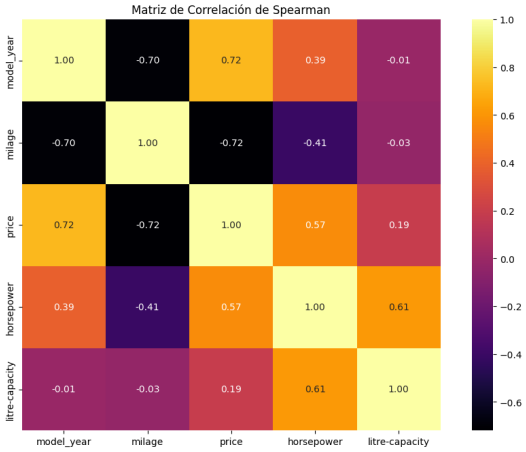

Figura 6: Correlación de Spearman sobre las variables numéricas.

En al Figura 6, podemos observar que la correlación entre las variables “Año de modelo”vs “Kilometraje” y “Precio”vs “Kilometraje" es alta pero negativa, es decir que cuando una variable sube, la otra baja. Pero el “Año de modelo”vs “Precio”tienen una correlación alta y positiva.

2.6. Codificación de los datos

Para crear mejores análisis, vammos a transformar las variables categóricas en variables numéricas.

Con la paquetería de python “category encoders”será muy sencillo, ya que en esta paquetería ya se encuentran funciones que codifican las variables categóricas, solamente hay que seleccionar los mejores métodos para hacerlo. Por el tipo de datos que tenemos, vamos a utilizar métodos:

- El método *Target Encoding*
- El método *Base N Encoding*

El primero, se utiliza comunmente para las variables que tienen un gran número de categorías, como lo son las variables “Marca”, “Modelo” y “Motor”, lo que realiza este método de codificación es asignar a cada categoría el valor del promedio de la variable objetivo, que en este caso es el “Precio”. Y el segundo método, asigna un código basado en el número de digitos que quieras que lo explique, por ejemplo, si utilizamos 2 dígitos, las categorías serán remplazadas por ceros y unos, pero en este caso como tenemos un gran número de categorías para algunas variables, utilizaremos un numero base mayor, 7 números, ya que si utilizamos bases más pequeñas, se incrementarían las columnas que alojan el código.

Ahora que ya tenemos “limpia”la información y las variable categóricas codificadas, es hora de navegar en algunos análisis.

3. Metodología

Previo a realizar los análisis que se presentarán, aplicaremos algunos métodos de filtro para seleccionar las características de las mejores variables y reducir la dimensión de la base de datos.

3.1. Selección de características

3.1.1. Valor F

Con este primer filtro, vamos a revisar que variables son las que muestran una relación lineal alta con las demás variables, una vez realizado el análisis se grafican los resultados.

En la Figura 7, podemos observar que las variables “Motor” y “Modelo”son las variables con una alta relación lineal.

3.1.2. R de correlación

Revisaremos la correlación que tienen las variables, como lo visto anteriormente, y gráficarlo para definir que variables están más correlacionadas.

Se puede observar en la gráfica de la Figura 8, que de nuevo dominan las variables “Motor” y “Modelo”, por lo que se encamina a que estas 2 variables si o si aparezcan en nuestros modelos, y la variable “Caballos de fuerza”tienen una correlación negativa.

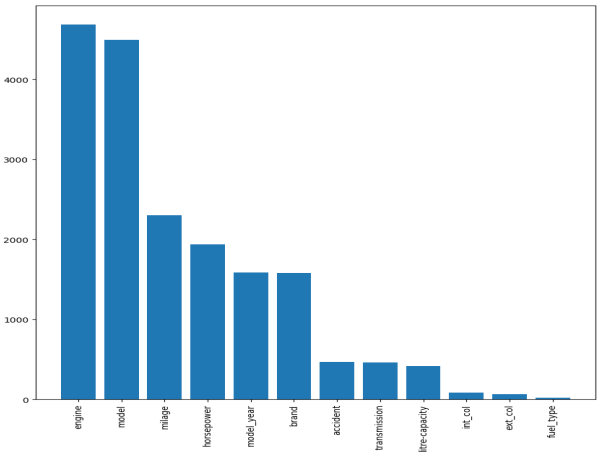

Figura 7: Resultado de los valores F de las variables.

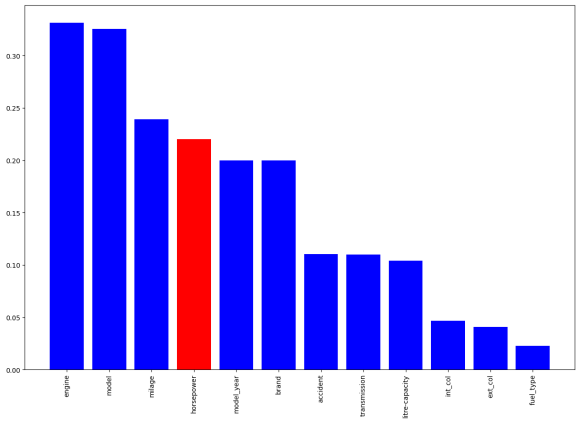

Figura 8: Resultados de correlación.

3.1.3. Umbral de varianza

Para este método, necesitaremos estandarizar las variables y despues obtener los indices de varianza para poder concluir a cerca de que variables son las que aportan mejor variabilidad a los datos.

Para la consideración de este filtro, si las varianzas están por debajo de 0.2 son variables que no aportan información al modelo, dado este criterio se tendrían que descartar todas, como se observa en la Figura 9, es por eso que no solo podemos considerar un solo filtro para las variables.

3.1.4. Información Mutua

El último método de filtro que se aplica es la Información Mutua, que nos dara un panorama de que variables pueden ser independientes.

Los resultados de este filtro nos proporciona si hay dependencia entre las variables, en este caso todas son mayor a 0, por lo cual no son independientes.

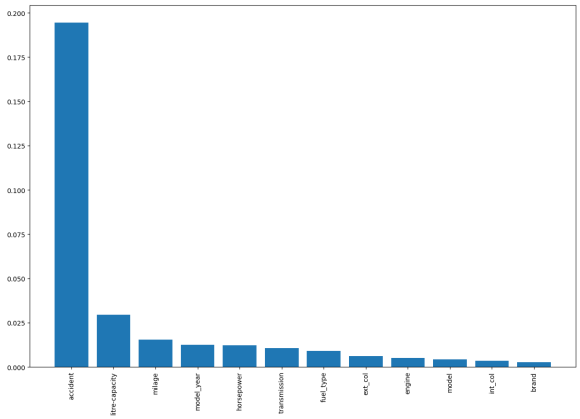

Figura 9: Umbral de varianza de los datos.

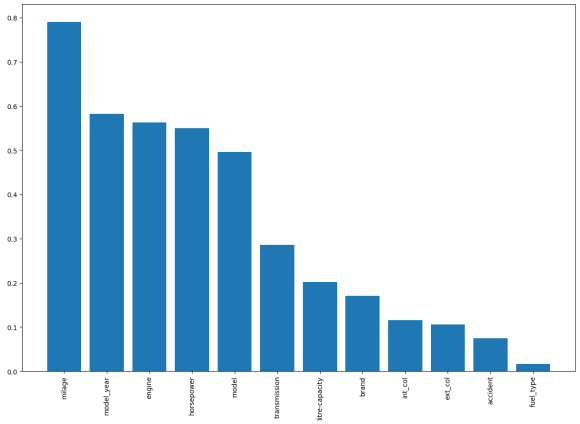

Figura 10: Información mutua

Y dando como resultado, la variable “Kilometraje" con mayor dependencia, siguiendo con poca diferencia entre ellas, las variables “Año del modelo”, “Motor”, “Caballos de fuerza” y “Modelo”.

3.1.5. Estandarización de resultados

Al hacer una estandarización de los resultados de los filtros, podemos otener una métrica general, al obtener la media de cada resultado, que aporte todos los resultados de cada filtro para poder seleccionar las variables con mejores rendimientos.
Hecho lo anterior, la siguiente gráfica muestra los resultados, y se puede observar que hay un rendimiento muy parejo entre las variables.

Las variables “Motor” y “Modelo”deben ser variables que formen parte del modelo, ya que tienen los resultados más altos, y de las demas variables se seleccionan las variables “Caballos de fuerza”, “Año del modelo”, “Kilometraje”,“Accidentes” y “Marca”, ya que en las variables restantes, al pasar del resultado de la variable “Marca” a “Transmisión”se nota una mayor diferencia.

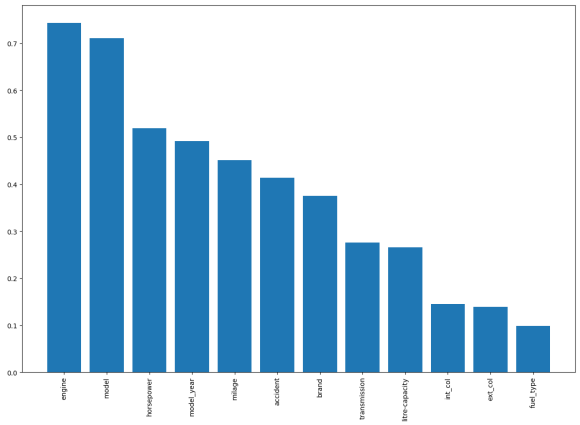

Figura 11: Indices de filtro estandarizados.

Por lo tanto, las variables a las que se les dará más importancia en los modelos serán:

- "Motor"
- "Modelo"
- "Caballos de fuerza"
- "Año del modelo"
- "Kilometraje"
- "Accidentes"
- "Marca"

3.2. Aprendizaje No Supervisado

3.2.1. Prueba de Componentes Principales (PCA)

Con todas las variables de la base de datos

Para iniciar el análisis de PCA, utilizaremos los datos estandarizados para posteriormente realizar la gráfica de la varianza explicada de cada componente.

Al ver la varianza acumulada en la Figura 12, se puede notar que con 3 componentes se puede tomar hasta un `85%` de la varianza de los datos, lo cual es suficiente para el análisis.

Por lo que, se utilizarán 3 componentes principales.

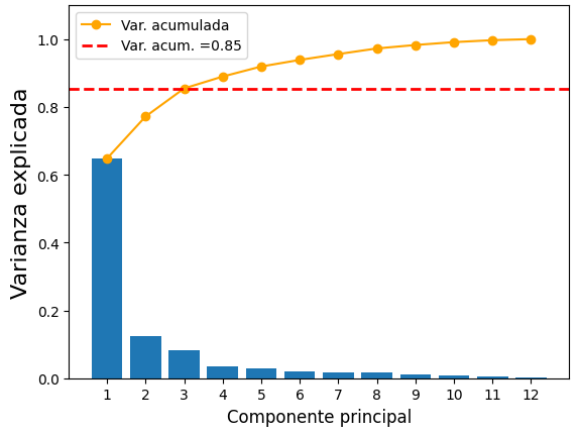

Figura 12: Resultados de varianza explicada por componente.

Al obtener más información de las cargas para cada componente, se realiza la siguiente tabla:

Concepto	PCA1	PCA2	PCA3
Eigenvalores	0.198	0.038	0.025
Prop. Varianza	0.648	0.124	0.083
Prop. Var. Acum.	0.648	0.772	0.855
Marca	-0.016	0.097	-0.049
Modelo	-0.038	0.192	-0.142
Año del modelo	-0.065	0.173	-0.521
Kilometraje	0.094	-0.210	0.579
Tipo de combustible	-0.006	0.017	-0.042
Motor	-0.044	0.226	-0.187
Transmisión	-0.025	0.075	-0.275
Color exterior	-0.009	0.018	-0.104
Color interior	-0.005	0.016	-0.055
Accidente	-0.989	-0.102	0.103
Caballos de fuerza	-0.056	0.464	-0.097
Capacidad de litros	-0.030	0.772	0.473

Con el gráfico de la Figura 13, se puede visualizar de mejor manera las cargas importantes para cada componente.

- En el primer componente, la mayor carga se la lleva la variable “Accidentes”.
- En el segundo, la variable “Capacidad de litros”.
- Y por último, la variable “Kilometraje”, “Año del modelo” y “Capacidad de litros”se llevan la mayor carga en el componente 3.

Esto es de sorprender, ya que la variable “Marca” o “Modelo”no aparecen con una carga importante, ya que son de las mejores variables que obtuvimos en la selección de características.

Por lo que este análisis, puede no ser bueno, ya que incorpora las cargas más importantes en las variables que no fueron seleccionadas como mejores, es por eso que es adecuado hacer un análisis para seleccionar las mejores variables para los modelos.

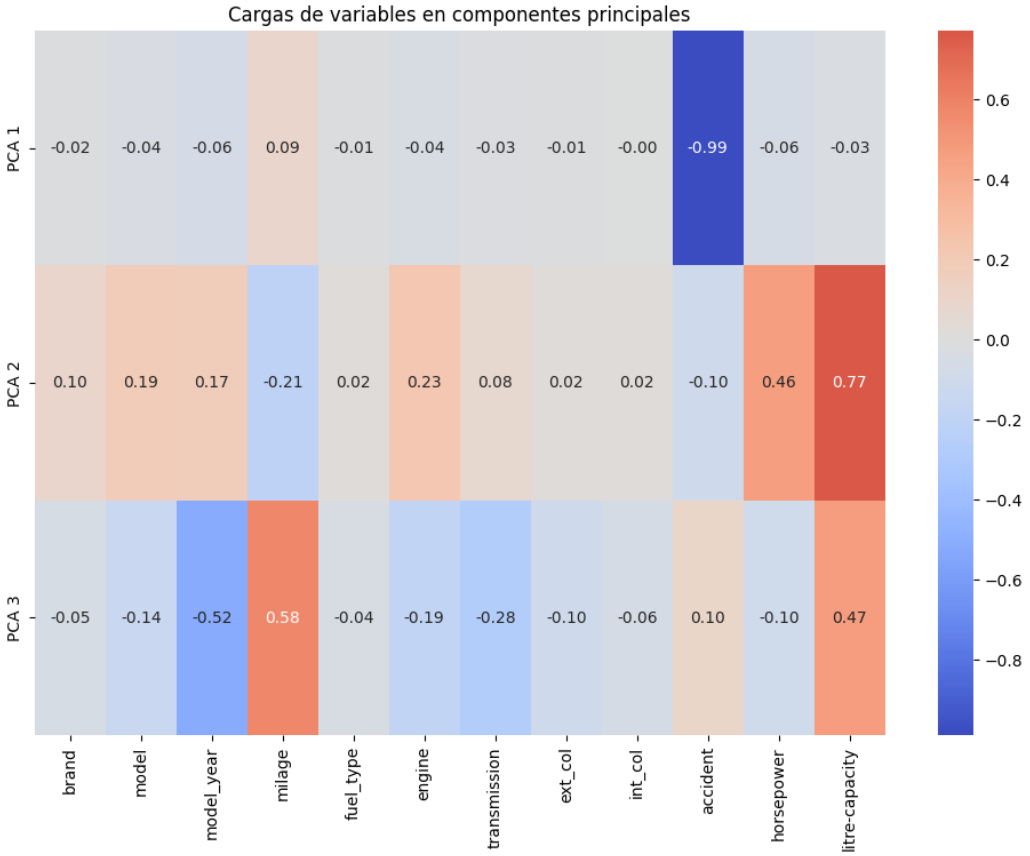

Figura 13: Cargas de los factores de PCA con todas las variables.

Con las mejores variables de la base de datos

Se realiza el mismo análisis solamente que en esta ocasión se toman solo las mejores variables que fueron seleccionadas en la sección de *Selección de características*.

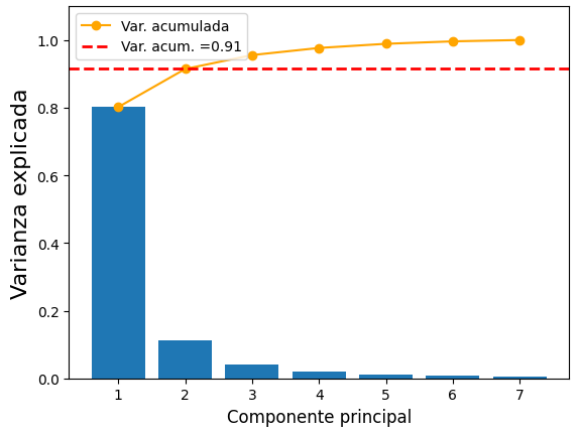

Figura 14: Cargas de los factores de PCA con las mejores variables.

Al ver la varianza acumulada en el gráfico, se puede notar que con 2 componentes se puede tomar hasta un 91 % de la varianza de los datos, lo cual es suficiente para el análisis.

Por lo que, se utilizarán 2 componentes principales.

Al obtener más información de las cargas para cada componente, se realiza la siguiente tabla:

Concepto	PCA1	PCA2
Eigenvalores	0.198	0.028
Prop. Varianza	0.802	0.113
Prop. Var. Acum.	0.802	0.915
Marca	-0.016	-0.118
Modelo	-0.037	-0.260
Año del modelo	-0.064	-0.506
Kilometraje	0.094	0.581
Motor	-0.043	-0.314
Accidente	-0.990	0.138
Caballos de fuerza	-0.054	-0.455

Con el gráfico de la Figura 14, se puede visualizar de mejor manera las cargas importantes para cada componente.

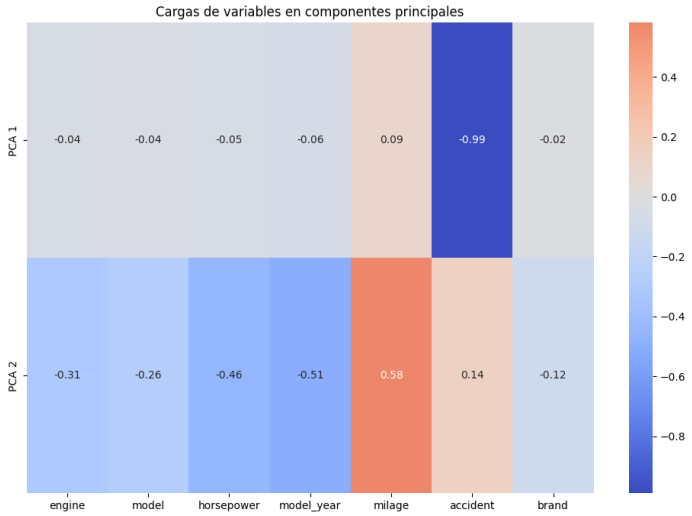

Figura 15: Cargas de los factores de PCA con las mejores variables.

En este caso la variable "Accidentes"sigue dominando el primer componente, lo cual es un poco extraño, ya que esta es una variable categórica solo con 2 opciones, es por eso que es mejor seleccionar 2 componentes, ya que en el segundo se puede ver que las demás variables tienen una carga bien distribuida en cad una de ellas, y dándole menos peso a "Accidentes".

Al hacer la transformación de los datos originales a los principales componentes, y al graficarlos se puede observar claramente 2 grupos de datos.Esto me hace suponer que el precio de los autos va a cambiar o va a ser relevante el cambio, si el auto tuvo algún accidente o no.

Interesantes resultados.

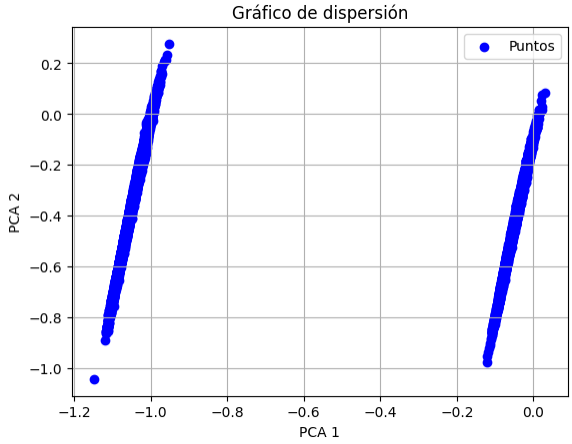

Figura 16: Dispersión de PCA 1 y PCA2