Proyecto final IA para ingenieros (Don’t Get Kicked!)

José Alejandro Urrego Pabón

Mateo toro molina

Inteligencia artificial para ingenieros

Raul Ramos Pollan

Docente

19 de noviembre de 2023

Facultad de ingeniería

Universidad de Antioquia

Medellín, Antioquia

Tabla de contenido

1. Introducción.

Los concesionarios que compran autos en gran cantidad durante las subastas son propensos a comprar autos en mal estado o que necesitan gran cantidad de reparaciones debido precisamente a que esos autos se camuflan con los que están en buen estado, los concesionarios y la comunidad automovilista llama a estas desafortunadas compras "Kicks" o patadas.

Los autos "pateados" o comprados con alguna avería son aquellos que cuentan con tacómetros manipulados, problemas en la carrocería que son imperceptibles, pero ponen en riesgo la vida del conductor, problemas con el traspaso de la propiedad del vehículo y muchos problemas más, estos percances hacen que los concesionarios pierdan mucho dinero en tratar de reparar las averías para revender el vehículo, porque inclusive muchas veces la propia avería es más costosa que el precio del auto. Nosotros como estudiantes de ingeniería mecánica hemos detectado que con un pronóstico más preciso usando técnicas de machine learning se le podría dar herramientas a los concesionarios sobre posibles autos averiados y de esta forma poder ahorrar mucho dinero en compras "patada".

El dataset implementado lo tomamos de una competencia realizada en Kaggle donde reportan 121756 autos recopilados entre los años 2009 y 2010, dicho dataset cuenta con 33 columnas de variables que permiten la correcta identificación y clasificación del auto. El 60% del dataset fue destinado para el proceso de entrenamiento y el otro 40% para el proceso de testeo, además, cerca del 40% de las variables son categóricas y el otro 60% son variables numéricas, además, se tienen muchos valores faltantes o nulos en por lo menos 5 columnas del dataset.

El dataset cuenta con 4 archivos donde se explican cada una de las variables que conforman dichos archivos (Carvana_Data_Dictionary.txt), un archivo destinado al entrenamiento del modelo (Training.csv), otro archivo destinado al testeo de este (test.csv) y un archivo final con las entradas para el testeo (Example_entry.csv) **[1].**

1.1 Descripción del problema

En este problema tenemos un dataset donde se encuentran muchos registros recopilados entre los años 2009 y 2010 de donde se pretende realizar una serie de modelos predictivos de mahcine learning que nos permita conocer con base a las características de un vehículo cuando es una mala compra o no.

1.2 Modelos implementados

En este trabajo se realizarán 4 modelos de machine learning los cuales estarán divididos en supervisados y no supervisados, siendo los supervisados la regresión logística y el árbol de decisión y los no supervisados PCA + árbol de decisión y PCA + regresión logística.

1.3 Métricas de desempeño

En este problema como solo se define si la futura compra en mala o buena (posible patada o no), se tiene un problema binario donde 1 correspondería a si y 0 correspondería a no, por lo que se pretenden calcular dos métricas de desempeño.

1.3.1 Precisión

Esta métrica de desempeño me indica que porcentaje de valores que calificaron como positivos son realmente positivos, es decir, cuales de los autos que son predichos como posibles patadas en realidad lo son. **[2, 3]**

$$Presicion = \frac{Verdadero\ positivo}{Verdadero\ positivo\ +\ Falso\ positivo} = \frac{TP}{TP + FN}$$

1.3.2 Accuracy

Esta métrica es muy similar a la anterior solo que en este caso se cuantifica la totalidad de los valores que han sido fielmente clasificados. **[2, 3]**

$$Accuracy = \frac{Verdadero\ positivo\ +\ Verdadero\ negativo}{Verdadero\ positivo\ y\ negativo\ +\ Falso\ positivo\ y\ negativo} = \frac{TP\ + TN}{TP + TN + FP + TN}$$

2. Exploración descriptiva del dataset

Se realiza un análisis exploratorio de los datos tanto numéricos como categóricos, para filtrar la cantidad de datos faltantes en cada una de las variables, obteniendo un resultado del 6,19% de datos faltantes en todo el dataset. Luego se procede a analizar las variables categóricas y numéricas individualmente.

2.1 Variables numéricas

Se filtran los datos para obtener la cantidad de datos faltantes entre las variables numéricas, obteniendo un total de 1332 datos faltantes. Luego se procede a calcular las características de cada variable para lograr una comprensión mayor de los datos que se tienen. A continuación, se muestra la relación que tiene cada variable con respecto a las demás (figura 1).

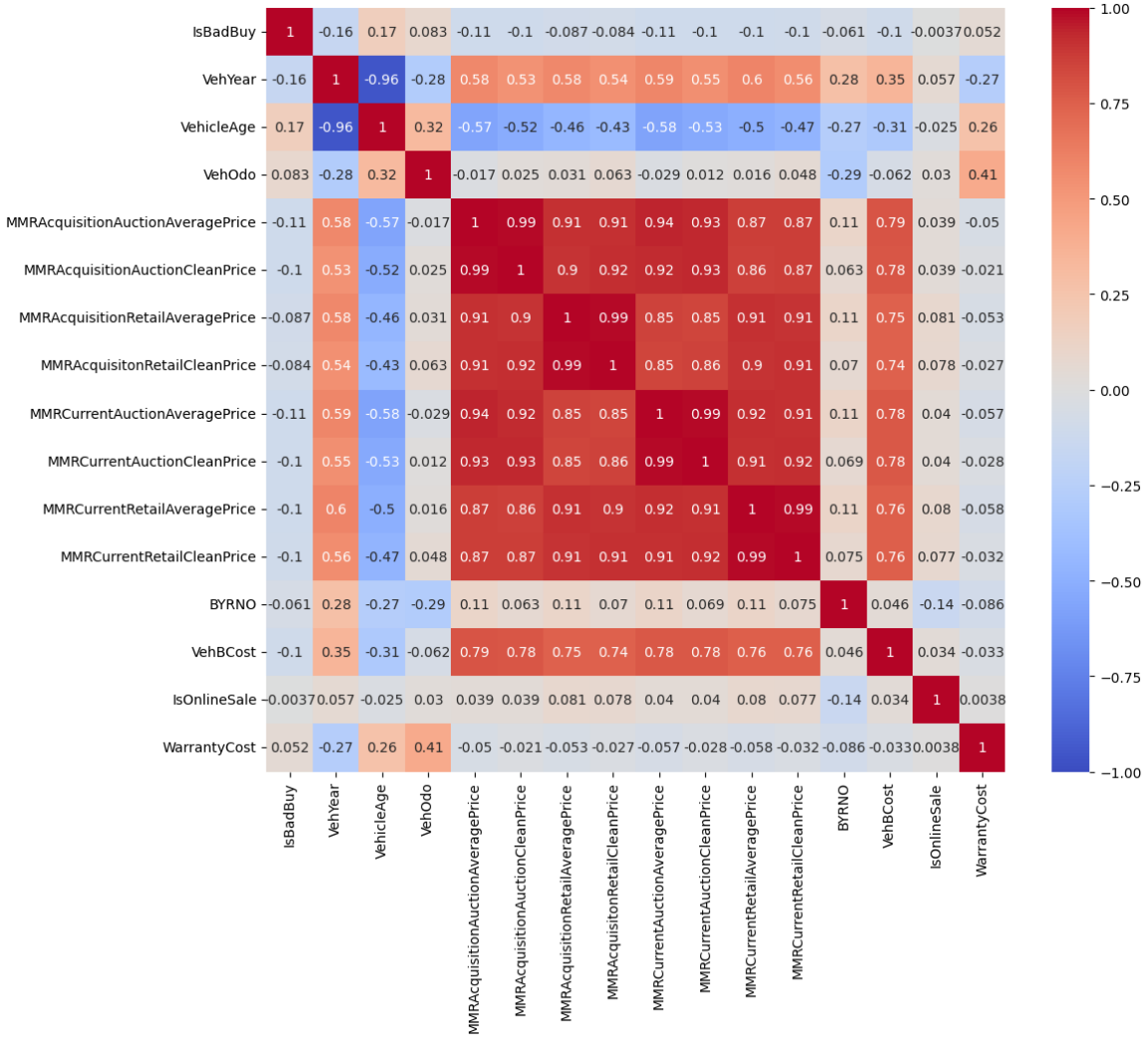

Figura 1. Diagrama de correlación de las variables numéricas.

2.2 Variables categóricas

Se filtran los datos para obtener la cantidad de datos faltantes entre las variables categóricas, obteniendo un total de 144662 datos faltantes. En base a lo anterior, se procede a elegir 6 variables categóricas que se consideraron las más importantes debido a que son las que, en nuestra opinión, afectan en mayor medida si se realizó una buena o mala compra. Las 6 variables elegidas son: auction, make, color, nationality, size y VNST. Las variables elegidas se contrastan con respecto a la variable IsBadBuy para saber cuántos datos se tienen, obteniendo los siguientes resultados (figura 2).

NOTA: Las variables VNST (Lugar de fabricación del vehículo por estado en EE. UU.) fue posteriormente eliminada por motivos prácticos, pero se quiso dejar para ilustrar la tendencia de autos que son malas compras en EE. UU.

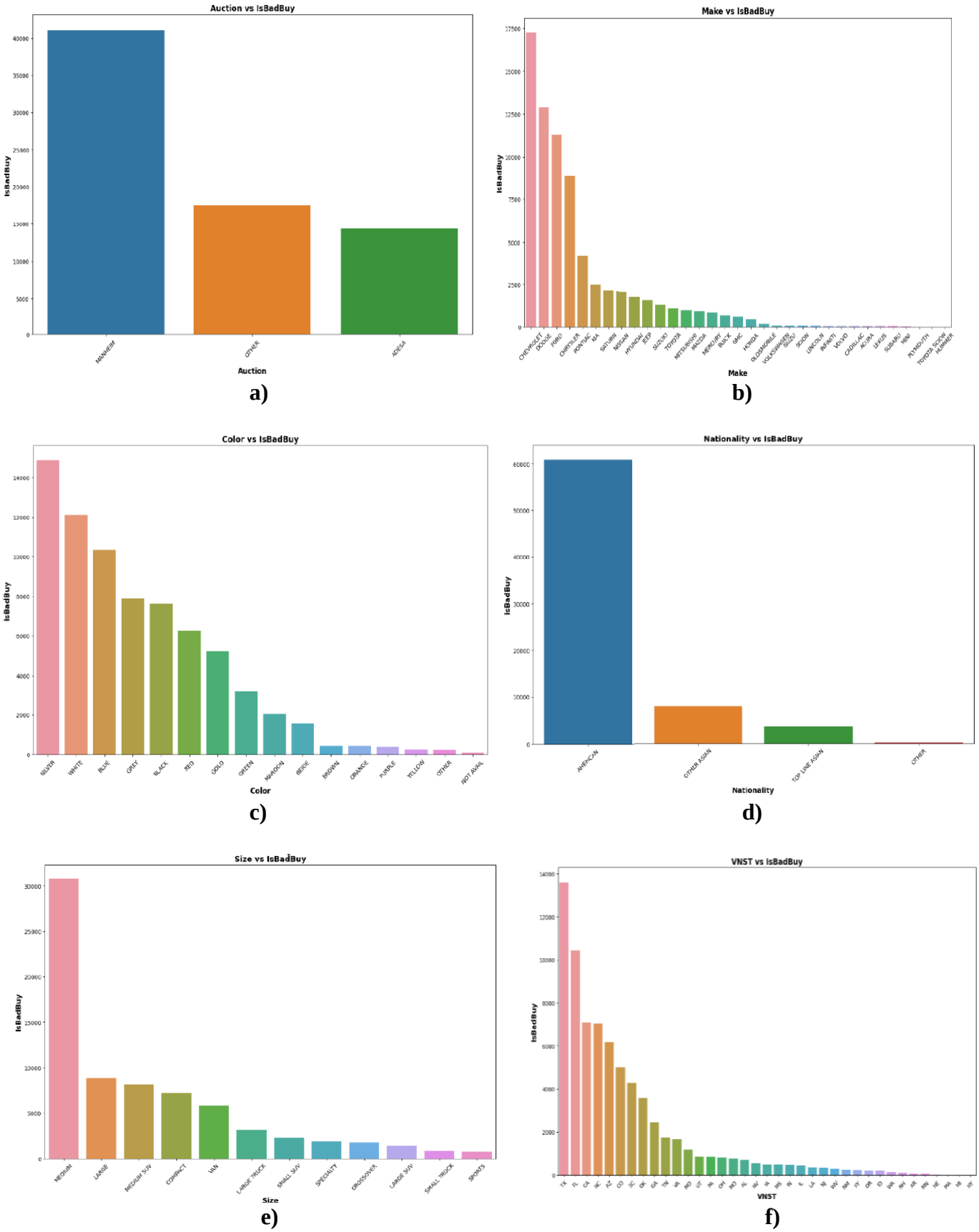

Figura 2. Variables categoricas: a) Auction vs IsBadBuy, b) Make vs IsBadBuy, c) Color vs IsBadBuy, d) Nationality vs IsBadBuy, e) Size vs IsBadBuy, f) VNST vs IsBadBuy.

3. Iteraciones de desarrollo

3.1 Preprocesado de datos

Las siguientes variables fueron eliminas ya que no otorgaban información relevante o les faltaban muchos datos que harían muy difícil el análisis posterior.

1. **Trim:** Nivel de equipamiento del vehículo.
2. **Wheel Type y Wheel type ID:** tipo de ruedas con su ID.
3. **PRIMEUNIT:** Identifica si el vehículo en cuestión tiene una demanda elevada.
4. **AUCGUART:** Nivel de garantía otorgada por la subasta para el vehículo.

De este primer análisis se concluye que las variables que involucran algún reporte por parte de la subasta sobre el estado pasado (1), presente (2,3) y futuro (4) del vehículo subastado tiene muchos datos faltantes ya sea por negligencia o con el objetivo de sacar algún provecho por parte de la subasta ocultando estos datos.

Además, se eliminan las siguientes columnas ya que en nuestra opinión no aportan mucha información al análisis.

- ✓ **RefID:** Ya que esta columna solo nos da información de la numeración de las filas dentro del mismo dataframe.

- ✓ **PurchDate:** La fecha en la que se compró el vehículo nos parece poco relevante porque en cualquier temporada del año se puede o no obtener una compra patada.

- ✓ **VehicleAge:** Se elimina la edad del vehículo ya que su fecha de fabricación nos parece más relevante (VehYear) debido a que el posible averió del vehículo pudo haber venido de fábrica y esto se determina a partir del lote o fecha de fabricación.

- ✓ Se eliminan igualmente las columnas modelo (**Model**) y submodelo (**SubModel**) ya que, con el solo fabricante, es decir, "make" se puede determinar si el vehículo tiene cierta predisposición a ser una mala compra.

- ✓ **Color:** Se elimina esta variable ya que la apariencia superficial del vehículo no nos dice nada, porque un vehículo puede tener cierto color que lo predispone a ser "exclusivo" como el negro, pero podría representar una posible compra patada.

- ✓ **BYRNO:** Esta columna correspondiente al NIT o representación numérica del comprador no nos da más información, salvo conocer el comprador.

- ✓ **VNZIP1:** El código postal donde se compró el auto tampoco tiene mucha relevancia según nuestro juicio.

- ✓ **VNST:** El estado donde fue comprado tampoco nos es de utilidad.

3.1.1 Imputación de los datos

Luego de realizar la eliminación de las variables que a nuestro parecer no eran relevantes se procedió con la imputación de los datos.

1. para los datos numéricos la teoría recomienda la imputación de los datos faltantes a partir de la media.

2. Para los datos categóricos la teoría recomienda la imputación de los datos faltantes a partir de la moda.

NOTA: Este análisis realizo tanto para el dataset de entrenamiento como para el dataset de testeo.

3.1.2 Hot encoding (Codificación en caliente)

En este análisis se procede a realizar la conversión de las variables categóricas a números binarios (Codificación en caliente), donde 0 indica inexistente y 1 indica que, si existe, de esta forma obtenemos finalmente un dataset con un total de 55 columnas.

NOTA: Este análisis realizo tanto para el dataset de entrenamiento como para el dataset de testeo.

3.1.3 Revisiones finales

- ✓ Después de haber concatenado las columnas separadas para la eliminación, limpieza, imputación, entre otros, se evidencio que había más columnas en el dataset de entrenamiento que de testeo, es decir, muchas mas columnas que la variable "IsBadBuy" (figura 3), después de buscar, se evidencio que existían tres tipos de nombrado para "MANUAL", por tal motivo se renombraron todas de esta forma.

```
Elementos diferentes en los dataframes: {'Make_HUMMER', 'Transmission_Manual', 'Make_TOYOTA SCION', 'Make_PLYMOUTH', 'IsBadBuy'}
```

Figura 3. Elementos diferentes entre el entrenamiento y testeo.

- ✓ Existían también 3 marcas de autos de más en el dataset de entrenamiento por lo que se procedió con la eliminación y se comprobó nuevamente, evidenciándose que esta vez solo había una variable de diferente, es decir, "IsBadBuy".

3.2 Modelos supervisados

Antes de presentar este modelo se realizó el mismo preprocesado para obtener los dataframes para el análisis de los modelos.

3.2.1 Árbol de decisión

Este modelo predictivo nos permite fabricar diagramas de decisión dado un conjunto de datos, que nos permite representar y categorizar las decisiones de forma sucesiva en la búsqueda de la solución de un problema. Tiene una estructura de árbol jerárquica, que consta de un nodo raíz, ramas, nodos internos y nodos hoja, tal y como se ve en la siguiente imagen:

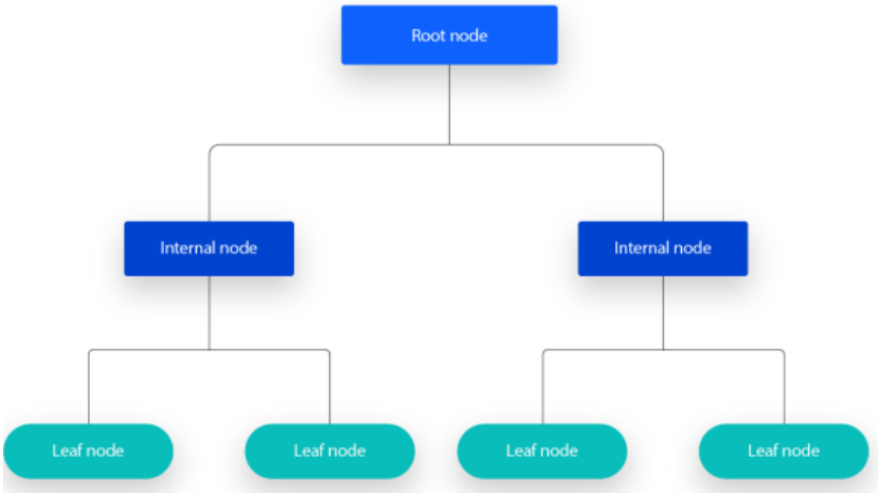

Figura 4. Diagrama árbol de decisión [4].

Como puede ver en el diagrama anterior, un árbol de decisión comienza con un nodo raíz, que no tiene ramas entrantes. Las ramas salientes del nodo raíz alimentan los nodos internos, también conocidos como nodos de decisión. En función de las características disponibles, ambos tipos de nodos realizan evaluaciones para formar subconjuntos homogéneos, que se indican mediante nodos hoja o nodos terminales. Los nodos hoja representan todos los resultados posibles dentro del conjunto de datos.

3.2.1.1 Resultados

Para este modelo se obtuvo un accuracy de aproximadamente el 63%, un valor que puede parecer bajo, pero supone una gran pérdida para los estafadores ya que se tienen demasiados autos que pueden llegar a costar mucho dinero.

Por otro lado, se obtuvo una precisión de aproximadamente el 17%, un resultado que era de esperarse ya que en el dataset utilizado se tienen muy pocos datos que tienen la variable IsBadBuy en comparación al total.

3.2.1.2 Curva de aprendizaje

A continuación, se presenta la curva de aprendizaje para el modelo de árbol de decisión:

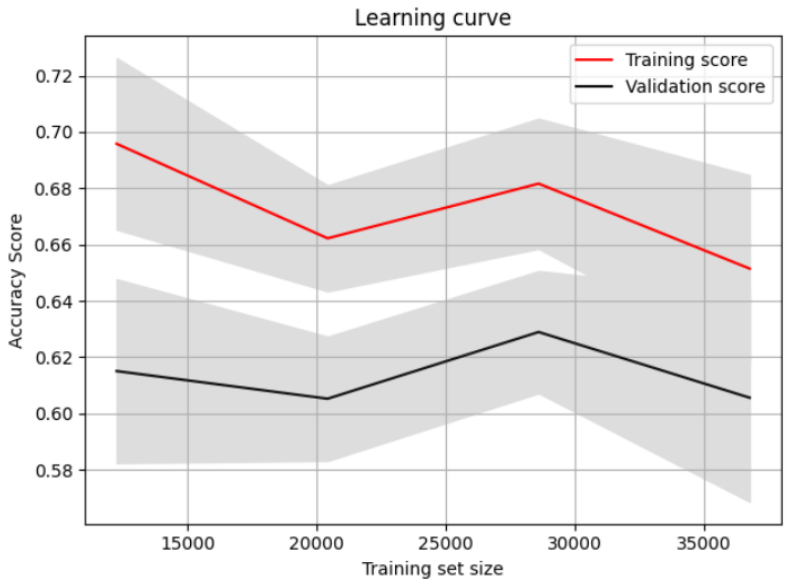

Figura 5. Curva de aprendizaje árbol de decisión

3.2.2 Regresión logística

Este modelo predictivo nos permite caracterizar algún evento con base a sus variables categóricas, en este caso no permitirá predecir el tipo de compra con base a las características físicas del auto. Algunos de los beneficios que nos ofrece este modelo son:

- Simplicidad debido a que el modelo de regresión logística es matemáticamente menos complejo que otros métodos de ML.

- Velocidad ya que los modelos de regresión logística pueden procesar grandes volúmenes de datos a alta velocidad porque requieren menos capacidad computacional, como memoria y potencia de procesamiento.

- Flexibilidad debido a que se puede usar la regresión logística para encontrar respuestas a preguntas que tienen dos o más resultados finitos.

3.2.2.1 Resultados

Para este modelo se obtuvo un accuracy de aproximadamente el 60%, un valor que puede parecer bajo, pero supone una gran pérdida para los estafadores ya que se tienen demasiados

autos que pueden llegar a costar mucho dinero. Si se hace una comparación entre este modelo y el anterior, el modelo anterior tiene mayores beneficios para el usuario.

Por otro lado, se obtuvo una precisión de aproximadamente el 17%, un resultado que era de esperarse ya que en el dataset utilizado se tienen muy pocos datos que tienen la variable IsBadBuy en comparación al total.

Cabe resaltar que ambos modelos (Árbol de decisión y regresión logística), dieron resultados muy similares, tanto en accuracy como en precisión. Lo anterior se debe a que la cantidad de autos que eran mala compra son relativamente poco en comparación con el total.

3.2.2.2 Curva de aprendizaje

A continuación, se presenta la curva de aprendizaje para el modelo de regresión logística:

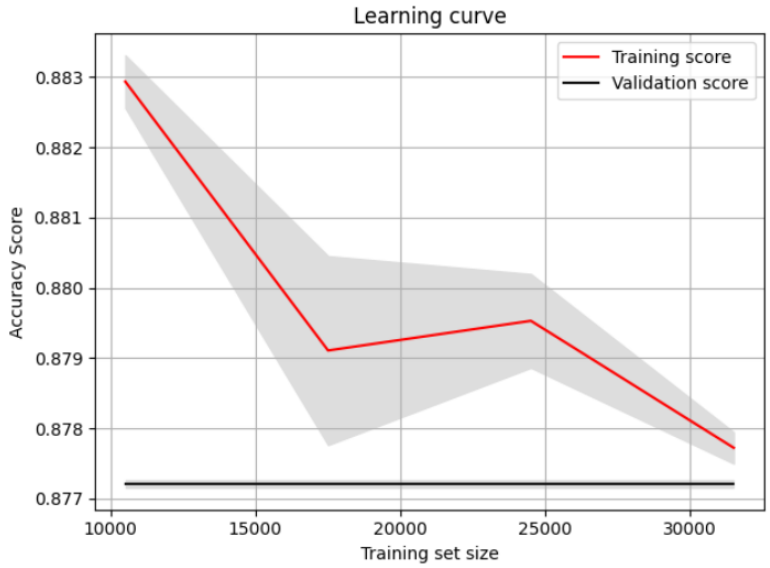

Figura 5. Curva de aprendizaje regresión logística.

.

3.3 Modelos no supervisados

Antes de presentar este modelo se realizó el mismo preprocesado para obtener los dataframes para el análisis de los modelos.

3.3.1 PCA + Árbol de decisión

El PCS (Principal Component Analyss) o análisis de componentes principales es un algoritmo no supervisado permite reducir la dimensionalidad de los datos, es decir, intenta mantener todas las variables posibles pero se podrán prescindir de las menos importantes, de esta forma se podrá saber cuáles de las variables son más o menos valiosas, a diferencia de una eliminación tradicional, nuestras nuevas variables son combinaciones de todas las variables originales, es decir, aunque algunas se eliminen se seguirá manteniendo la información útil de las variables iniciales.

Después de aplicar el modelo se graficó el número de componentes más influyentes e importantes entre sí. **[5]**

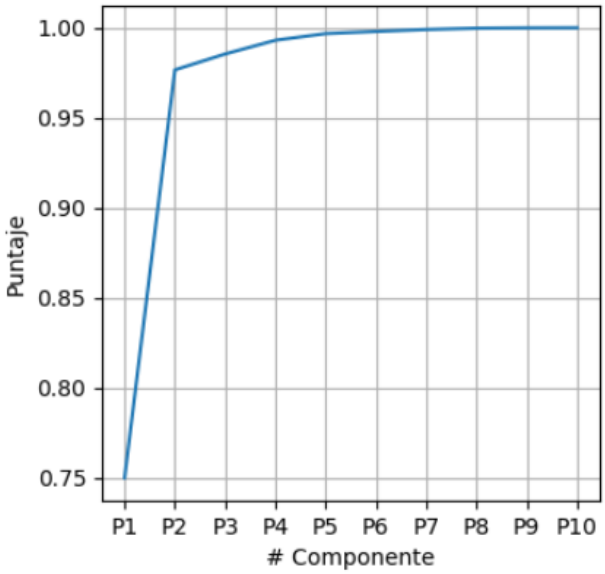

Figura 6. Grafico de variabilidad explicada acumulada.

Del grafico anterior se evidencia que con aproximadamente 3 componentes se llega a mas del 95% del puntaje, se dejo hasta 10 para evidenciar como la variabilidad explicada se estabilizada en 1.

Luego se usa el modelo previo para convertir los datos de X_train y X_test para ponerlos a prueba en el modelo predictivo “Árbol de decisión”.

3.3.1.1 Resultados

Para este modelo se obtuvo un accuracy de aproximadamente el 65%, un valor que es mayor en comparación al obtenido con los modelos no supervisados.

Por otro lado, se obtuvo una precisión de aproximadamente el 19%, un resultado que, aunque es mayor que el valor obtenido con los modelos no supervisados sigue siendo muy bajo debido a la poca cantidad de datos que tienen la variable que nos muestra si se hizo una mala compra en comparación al resto de los datos.

3.3.1.2 Curva de aprendizaje

A continuación, se presenta la curva de aprendizaje para el modelo de PCA + Árbol de decisión:

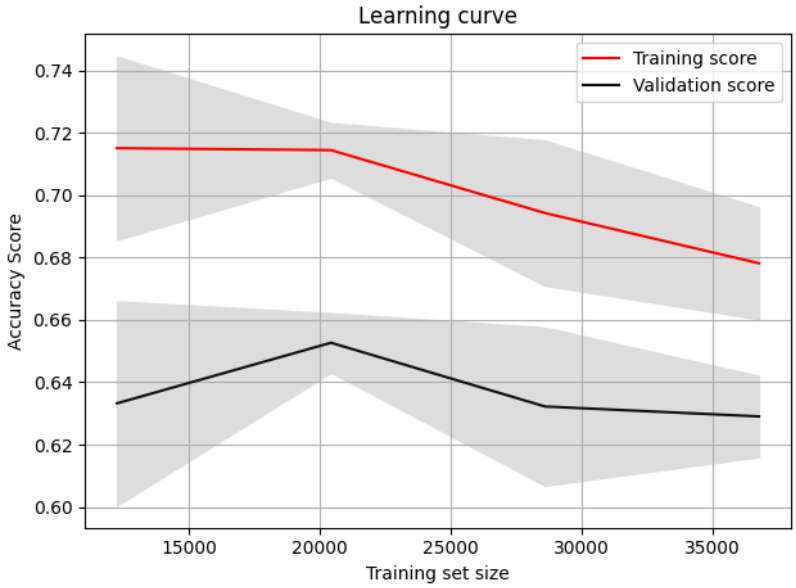

Figura 7. Curva de aprendizaje PCA + Árbol de decisión.

3.3.2 PCA + Regresión logística

El PCS (Principal Component Analyss) o análisis de componentes principales es un algoritmo no supervisado permite reducir la dimensionalidad de los datos, es decir, intenta mantener todas las variables posibles pero se podrán prescindir de las menos importantes, de esta forma se podrá saber cuáles de las variables son más o menos valiosas, a diferencia de una eliminación tradicional, nuestras nuevas variables son combinaciones de todas las variables originales, es decir, aunque algunas se eliminen se seguirá manteniendo la información útil de las variables iniciales.

Después de aplicar el modelo se graficó el número de componentes más influyentes e importantes entre sí. **[5]**

3.3.2.1 Resultados

Para este modelo se obtuvo un accuracy de aproximadamente el 63%, un valor que es mayor en comparación al obtenido con los modelos no supervisados.

Por otro lado, se obtuvo una precisión de aproximadamente el 18%, siendo el menos valor de los obtenidos con los modelos no supervisados.

3.3.2.2 Curva de aprendizaje

A continuación, se presenta la curva de aprendizaje para el modelo de PCA + Regresión logística, donde se evidencia que por parte de la validación los datos son constantes en el tiempo.

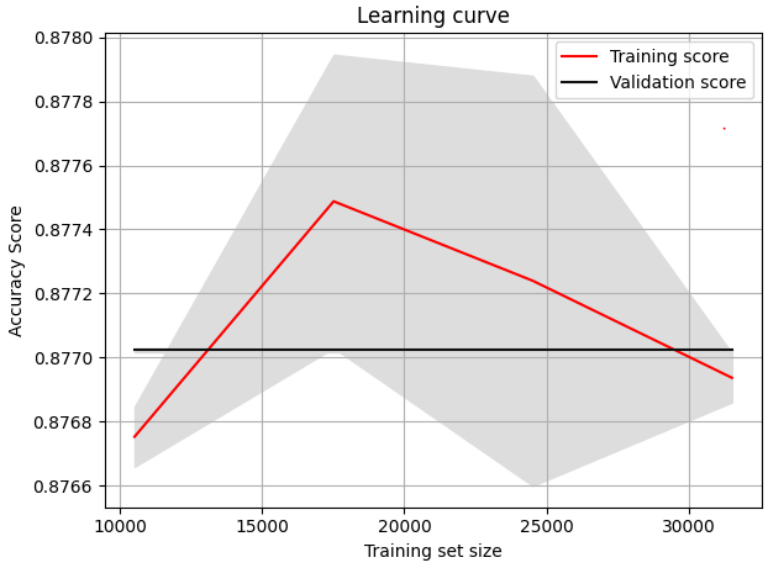

Figura 8. Curva de aprendizaje PCA + Regresión logística.

4. Retos y consideraciones de despliegue

Para el despliegue de los modelos aplicados a este proyecto se tiene varios retos y consideraciones:

- Lograr una precisión y un accuracy mayor buscando lograr mejores predicciones para saber si un vehículo en venta es una estafa o no. Para mejorar estos resultados, se pueden incorporar características adicionales o incluso aplicar nuevos modelos.

- Adecuar los modelos puede ser un gran reto cuando se tienen grandes volúmenes de datos. Debido a lo anteriormente mencionado, se deben realizar pruebas de rendimiento y escalabilidad antes de la implementación de cada uno de los modelos con volúmenes de datos muy altos.

- Los modelos pueden degradarse con el tiempo debido a cambios en el entorno. Por esta razón presentada, se debe establecer un sistema de monitoreo continuo para evaluar el rendimiento de cada uno de los modelos, realizando ajustes o actualizaciones cuando sea necesario.

- Los modelos pueden exponer datos de algunas empresas que pueden afectar la imagen de esta, por lo tanto, se deben implementar medidas para garantizar la privacidad de los datos y cumplir con las regulaciones y políticas pertinentes.

5. Conclusiones

- En este proyecto se obtuvo que los modelos supervisados presentan mejores resultados que el modelo no supervisado.

- El manejo de datos faltantes en el conjunto de datos es esencial. La imputación de valores para datos numéricos y categóricos, así como la eliminación de características que no contribuyen significativamente, son pasos importantes en el preprocesamiento de datos.

- En el dataset utilizado para este proyecto, la cantidad de datos etiquetados como posibles estafas es pequeña en comparación con la cantidad total de los datos. Lo anterior puede afectar la capacidad del modelo para aprender patrones específicos de estafas.

- La elección de las características que se consideraron importantes puede variar significativamente los resultados obtenidos, por lo tanto, la elección y la calidad de las características son fundamentales para el rendimiento del modelo.

6. Referencias

[1] https://www.kaggle.com/competitions/DontGetKicked/overview

[2] https://www.themachinelearners.com/metricas-de-clasificacion/

[3] https://rramosp.github.io/ai4eng.v1/content/LAB%2001.02%20-%20METRICS.html

[4] https://www.ibm.com/es-es/topics/decision-trees

[5] https://www.aprendemachinelearning.com/comprende-principal-component-analysis/

[6] https://aws.amazon.com/es/what-is/logistic-regression/#:~:text=La%20regresi%C3%B3n%20log%C3%ADstica%20es%20una%20t%C3%A9cnica%20importante%20en%20el%20campo,de%20datos%20sin%20intervenci%C3%B3n%20humana.