TRABAJO DE GRADO
Opción Seminario-Diplomado.

SISTEMA DE PREDICCION DE VALOR DE AUTOS USADOS, UTILIZANDO ESTRATEGIAS DE MACHINE LEARNING

Corporación Universitaria Remington.
Nombre de la facultad: Ingenierías
Nombre del programa académico: Ingeniería De Sistemas

Estudiante:
Hildegar Camilo Cardenas Aguilar

Tutor: Juan Carlos Briñez de León

Opción de Trabajo de grado Seminario-Diplomado.
2024.

Dedicatoria

A MIS PADRES Y HERMANOS QUIENES SON LOS QUE ME HAN AYUDADO EN ESTE PROGRESO ACADEMICO, Y GRACIAS A ELLOS PUEDO TENER UNA CARRERA UNIVERSITARIA

Tabla de Contenidos

Resumen

El proyecto de predicción de precios de autos usados tiene como objetivo estimar el precio de un automóvil en función de varias características, como el año, kilometraje, tipo de combustible, marca, entre otras. Para ello, se utiliza un modelo de Machine Learning basado en Random Forest, que es un algoritmo de aprendizaje supervisado muy eficaz para tareas de regresión.

El análisis de datos comienza con la recolección de información sobre el vehículo como lo son el modelo, año, kilometraje, tipo de combustible, marca y tipo de transmisión.
A partir de los datos se propone usar algoritmos de machine learning como lo es random forest que nos permitirá entrenar un modelo con la información recopilada, algunos datos tendrán que ser pasados de texto a un valor numérico para que el modelo pueda interpretar y dar una predicción final

Palabras clave

Machine learning, autos usados, entrenamiento, random forest

1. Marco conceptual y contextual

1.1 Contexto: Predicción de venta de carros usados

1.1.1 Sistemas de recomendación.

El sistema automotriz ha venido experimentando un notable crecimiento los últimos años ya que la producción de auto nuevos a disminuido, es necesario desarrollar modelos precisos que nos permitan la toma de decisiones informada tanto para vendedores como compradores

1.2 Descripción de caso de estudio.

Como bien he dicho el mercado de autos usados viene creciendo en los últimos años convirtiéndose en una alternativa de negocio ya que es mas accesible para compradores, uno de los principales desafíos es como determinar un precio justo los precios de lo autos usados. Los precios están influenciados por una gran cantidad de variables las cuales son la marca, el modelo, el kilometraje, el año de fabricación y estado del vehículo, con lo cual se ven en la necesidad de usar programas de inteligencia artificial los cuales predigan un valor teniendo en cuantas las variables (la marca, el modelo, el kilometraje, el año de fabricación y estado del vehículo, etc.)

1.3 Pregunta problema:

¿Cómo predecir con exactitud y precisión el valor de un automóvil usado utilizando las características básicas del vehículo (¿cómo el modelo, el precio y los cambios en los mercados de automóviles?

1.4 Hipótesis:

El año de fabricación tiene una relación negativa en cuanto al precio ya que a medida que aumenta la edad del auto el precio tiende a disminuir.

La marca del auto también influye ya que, si es una marca de lujo como lo es Audi, Mercedes Benz o Bmw su precio será mas alto que a un auto de marca Chevrolet, Renault, entre otras que son marcas más económicas comercialmente.

El tipo de combustible influye mas cuando son autos eléctricos o híbridos ya que estos tienden a tener un valor más alto a comparación de los autos que trabajan con gasolina y Diesel, ya que los autos eléctricos y híbridos tienen una más alta demanda y por la percepción de ser ecológicos.

El kilometraje esta relacionado negativamente con su precio ya que a mayor kilometraje recorrido el precio disminuye debido al desgaste y menor vida útil percibid.

2. Objetivos

2.1 Objetivo general.

Implementar un modelo predictivo usando técnicas de aprendizaje automático, estimando una predicción del precio de los carros usados, con el fin de ayudar a vendedores y compradores en el mercado de carros usados

2.2 Objetivos específicos.

- Utilizar aprendizaje automático como regresión, random forest, o redes neuronales
- Evaluar y analizar el desempeño de los algoritmos implementados para la toma de decisiones.
- Analizar y determinar que variable como el modelo, los km recorridos, la marca etc. Tienen un mayor impacto a la hora de mejorar las predicciones

3. Desarrollo e implementación del aprendizaje

Desarrollar e implementar métodos de aprendizaje para predecir los precios de los automóviles usados implica varios pasos clave, desde la recopilación y el preprocesamiento de datos hasta la capacitación y evaluación de modelos utilizando Random Forest. Una vez entrenado, el modelo se puede utilizar en aplicaciones de pronóstico en tiempo real, proporcionando una poderosa herramienta para compradores y vendedores de automóviles usados. La mejora continua del modelo y su mantenimiento garantizarán su eficacia a largo plazo.

3.1 Preparación y análisis de los datos

	Car_Name	Year	Selling_Price	Present_Price	Kms_Driven	Fuel_Type	Seller_Type	Transmission	Owner
0	ritz	2014	3.35	5.59	27000	Petrol	Dealer	Manual	0
1	sx4	2013	4.75	9.54	43000	Diesel	Dealer	Manual	0
2	ciaz	2017	7.25	9.85	6900	Petrol	Dealer	Manual	0
3	wagon r	2011	2.85	4.15	5200	Petrol	Dealer	Manual	0
4	swift	2014	4.60	6.87	42450	Diesel	Dealer	Manual	0
5	vitara brezza	2018	9.25	9.83	2071	Diesel	Dealer	Manual	0
6	ciaz	2015	6.75	8.12	18796	Petrol	Dealer	Manual	0
7	s cross	2015	6.50	8.61	33429	Diesel	Dealer	Manual	0
8	ciaz	2016	8.75	8.89	20273	Diesel	Dealer	Manual	0
9	ciaz	2015	7.45	8.92	42367	Diesel	Dealer	Manual	0

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 541 entries, 0 to 540
Data columns (total 9 columns):
 #   Column         Non-Null Count  Dtype  
---  ------         --------------  -----  
 0   Car_Name       541 non-null    object 
 1   Year           541 non-null    int64  
 2   Selling_Price  541 non-null    float64
 3   Present_Price  541 non-null    float64
 4   Kms_Driven     541 non-null    int64  
 5   Fuel_Type      541 non-null    object 
 6   Seller_Type    541 non-null    object 
 7   Transmission   541 non-null    object 
 8   Owner          541 non-null    int64  
dtypes: float64(2), int64(3), object(4)
memory usage: 38.2+ KB
```

Figura 1 de las variables

En esta imagen podemos observar que este dataset este hecho para 541 autos, mide las

Las variables nombre del carro, año, precio de venta, precio actual, kilómetros conducidos, tipo de combustible, tipo de vendedor, transmisión y dueño.

	Year	Selling_Price	Present_Price	Kms_Driven	Owner
count	383.000000	383.000000	383.000000	383.000000	383.000000
mean	2013.738903	4.916736	7.856397	36268.214099	0.036554
std	2.742859	4.668873	7.787693	35472.286988	0.225868
min	2003.000000	0.100000	0.320000	500.000000	0.000000
25%	2012.000000	1.250000	1.900000	15570.500000	0.000000
50%	2014.000000	4.400000	6.950000	32000.000000	0.000000
75%	2016.000000	6.400000	9.900000	48000.000000	0.000000
max	2018.000000	35.000000	92.600000	500000.000000	3.000000

Figura 2 del análisis de los datos

En los análisis de los datos podemos observar que el año mínimo de los datos en la base de datos es 2003 y la máxima 2018, que los kilómetros recorridos el mínimo es 500km y el máximo 500000km

	Car_Name	Year	Selling_Price	Present_Price	Kms_Driven	Fuel_Type	Seller_Type	Transmission	Owner
0	ritz	2014	3.35	5.59	27000	Petrol	Dealer	Manual	0
1	sx4	2013	4.75	9.54	43000	Diesel	Dealer	Manual	0
2	ciaz	2017	7.25	9.85	6900	Petrol	Dealer	Manual	0
3	wagon r	2011	2.85	4.15	5200	Petrol	Dealer	Manual	0
4	swift	2014	4.60	6.87	42450	Diesel	Dealer	Manual	0

	Car_Name	Year	Selling_Price	Present_Price	Kms_Driven	Fuel_Type	Seller_Type	Transmission
0	ritz	2014	3.35	5.59	27000	Petrol	Dealer	Manual
1	sx4	2013	4.75	9.54	43000	Diesel	Dealer	Manual
2	ciaz	2017	7.25	9.85	6900	Petrol	Dealer	Manual
3	wagon r	2011	2.85	4.15	5200	Petrol	Dealer	Manual
4	swift	2014	4.60	6.87	42450	Diesel	Dealer	Manual

Figura 3. Eliminar variables obsoletas

Luego decidimos quitar la variable dueño (Owner) ya que la variable dueño como tal no es tan representativa de manera aislada sin un previo contexto.

Tabla de Frecuencia:

	Intervalo	Frecuencia Absoluta	Frecuencia Acumulada
0	[2015.0, 2018.015)	251	251
1	[2012.0, 2015.0)	177	428
2	[2009.0, 2012.0)	73	501
3	[2006.0, 2009.0)	26	527
4	[2003.0, 2006.0)	14	541

Figura 4 tabla de frecuencia de la variable año

En la tabla de frecuencia podemos observar que lo dividimos en 5 intervalos, siendo el primer intervalo de 2015 a 2018 donde hay 251, de 2012 a 2015 y así sucesivamente como observamos en la imagen

Tabla de Frecuencia:

	Intervalo	Frecuencia Absoluta	Frecuencia Acumulada
0	[500.0, 83750.0)	514	514
1	[83750.0, 167000.0)	21	535
2	[167000.0, 250250.0)	4	539
3	[416750.0, 500499.5)	2	541

Figura 5 tabla de frecuencia de la variable kilómetros recorridos

En la tabla podemos observar que la mayor cantidad de autos esta entre 500km a 83750km correspondientes a 514 autos

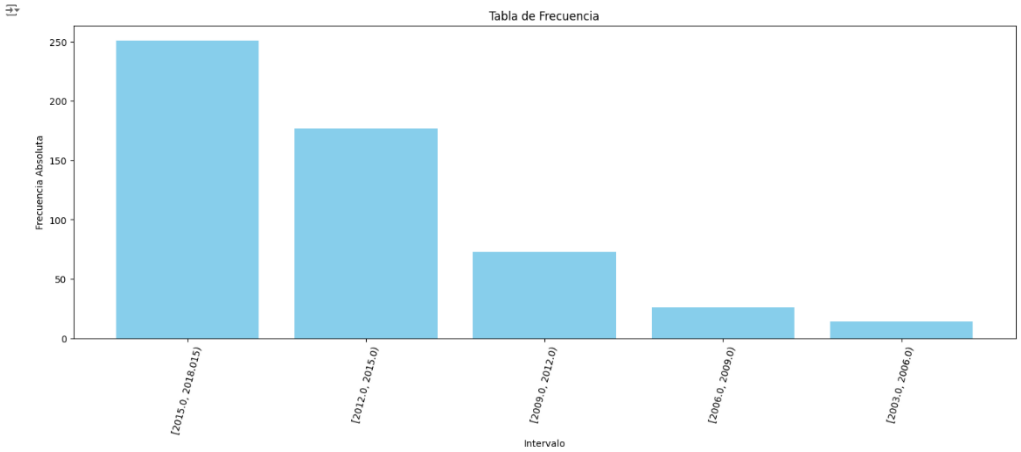

Figura 6 distribución de los autos según el año

Acá podemos observar la grafica de la tabla de frecuencia que nos proporciona una forma más clara de cómo se distribuyen los datos a través de los intervalos

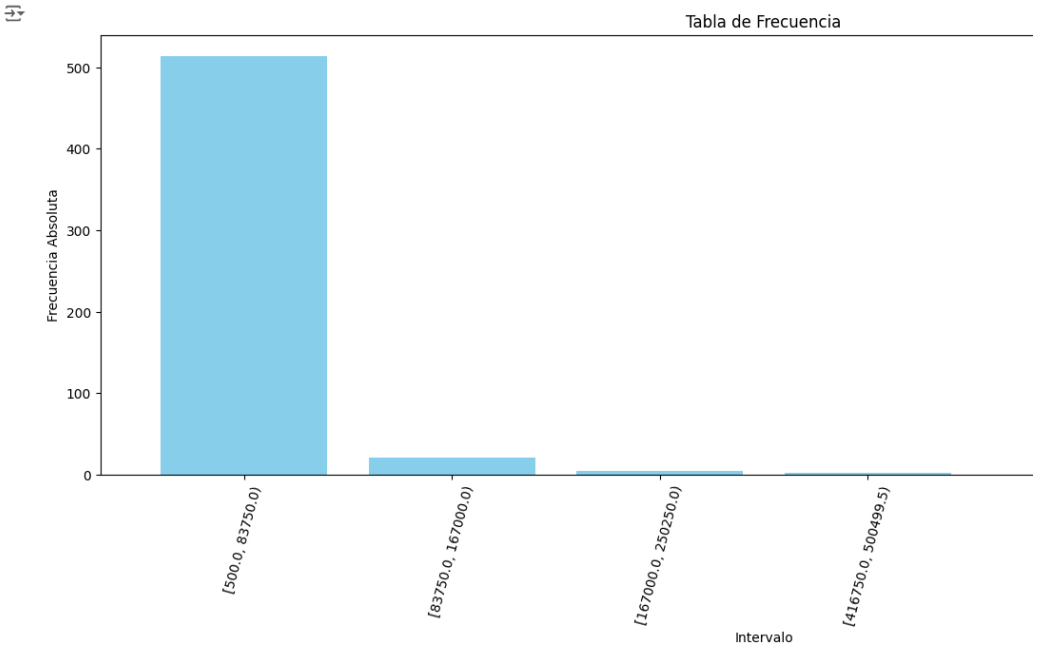

Figura 7 grafica de la variable kilometro recorridos

En la imagen podemos divisar de como es la diferencia de rangos de una manera más entendible.

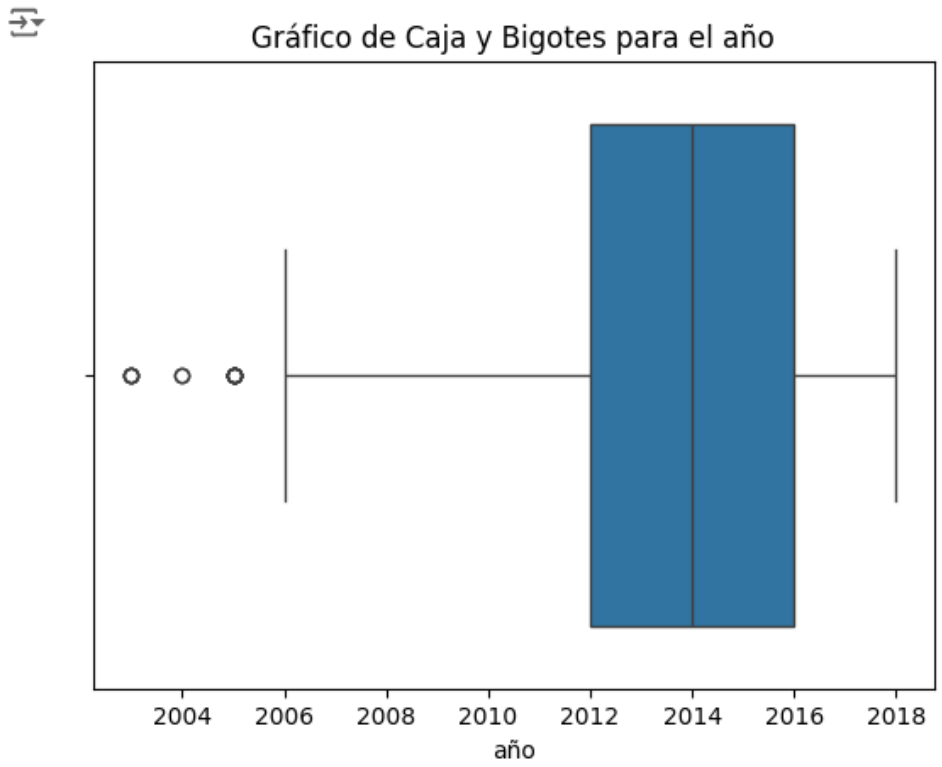

Figura 8 grafico de cajas y bigotes

Ente este grafico nos dice que la mayor parte de auto que tenemos disponibles esta en el rango de 2012 a 2016

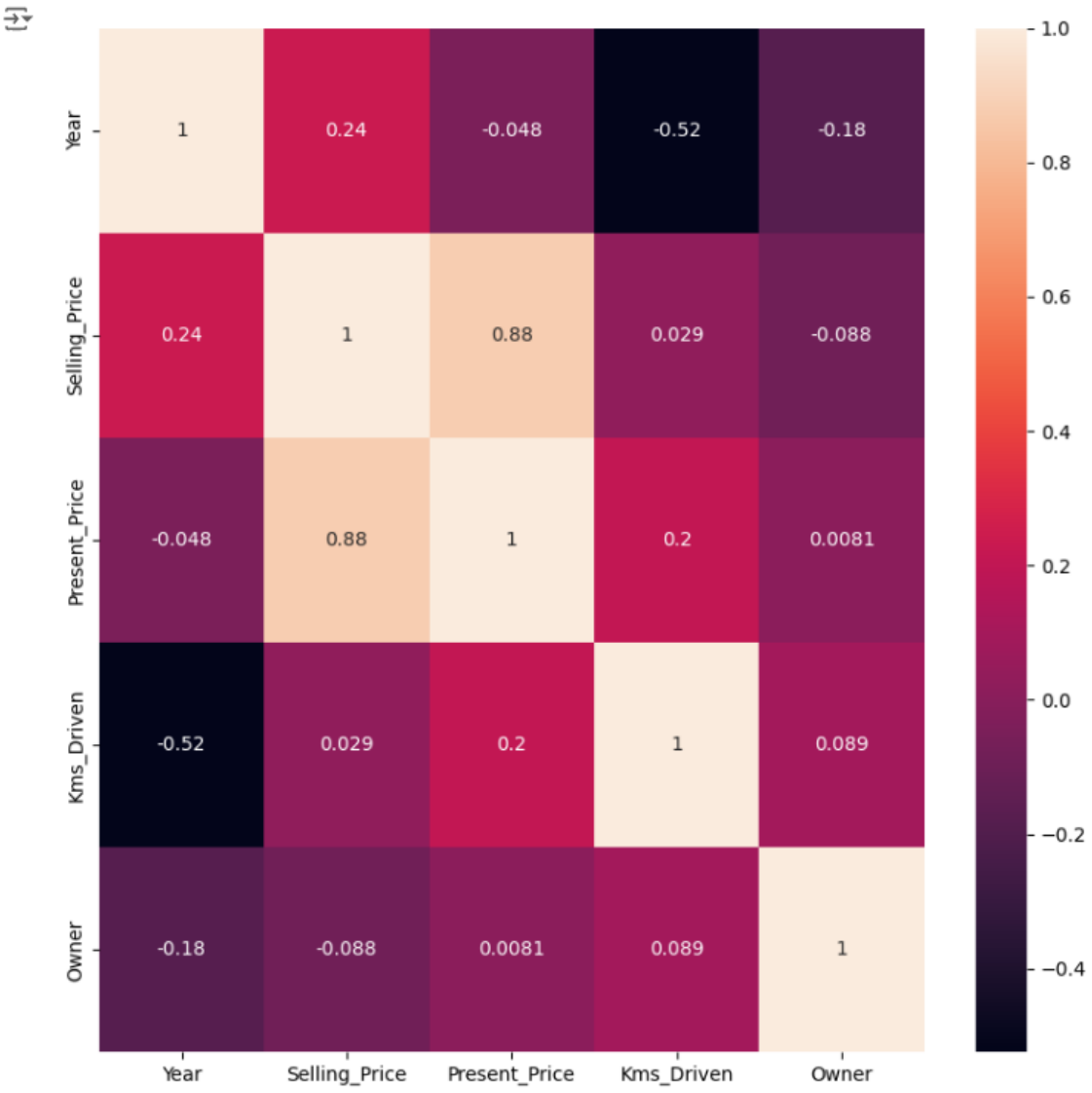

Figura 9 matriz de correlación

En la matriz podemos observar que kilómetros recorridos y año tienen una correlación negativa directa lo cual es normal por que entres más nuevo el vehículo menos kilómetros tendrá

Todo lo que hicimos en la primera clase, se debe explicar desde el punto de vista de ustedes, pegando gráficos y pantallazos de resultados.

3.2 Modelo de toma de decisiones

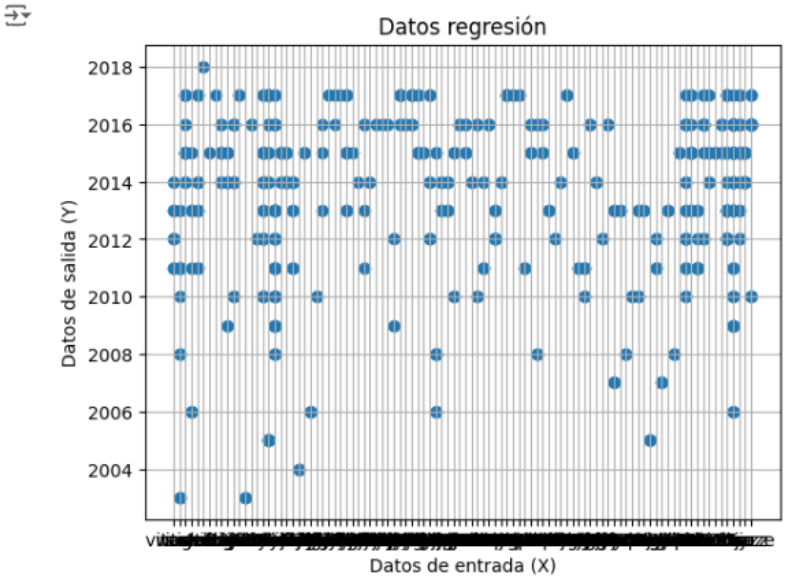

Figura 10 datos de regresión

En la imagen podemos observar que los datos están dispersos y agrupados en el mismo año viendo que el año 2013 a 2017 tienen mayor agrupación de datos

3.3 Análisis de desempeño

```
encoded_df = df.copy()

from sklearn.preprocessing import LabelEncoder
encoder = LabelEncoder()

columns_to_encode = ['Car_Name', 'Fuel_Type', 'Seller_Type', 'Transmission']

for column in columns_to_encode:
    encoded_df[column] = encoder.fit_transform(encoded_df[column])
```

```
[ ] encoded_df.head(10)
```

	Car_Name	Year	Selling_Price	Present_Price	Kms_Driven	Fuel_Type	Seller_Type	Transmission	Owner
0	90	2014	3.35	5.59	27000	2	0	1	0
1	93	2013	4.75	9.54	43000	1	0	1	0
2	68	2017	7.25	9.85	6900	2	0	1	0
3	96	2011	2.85	4.15	5200	2	0	1	0
4	92	2014	4.60	6.87	42450	1	0	1	0
5	95	2018	9.25	9.83	2071	1	0	1	0
6	68	2015	6.75	8.12	18796	2	0	1	0
7	91	2015	6.50	8.61	33429	1	0	1	0
8	68	2016	8.75	8.89	20273	1	0	1	0
9	68	2015	7.45	8.92	42367	1	0	1	0

Figura 11 cambiando variables carro, tipo de combustible, tipo de vendedor y transmisión a datos numéricos.

Como podemos observar unas variables fueron cambiadas a datos numéricos para poder facilitar el uso de datos en los modelos machine learning

```
from sklearn.ensemble import RandomForestRegressor
from sklearn.model_selection import train_test_split

X = encoded_df.drop('Selling_Price', axis = 1)
y = encoded_df['Selling_Price']

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.2, random_state = 42)

forest1 = RandomForestRegressor()

forest1.fit(X_train, y_train)
forest1.score(X_test, y_test)
```

```
0.9622700745468055
```

Figura 12 entrenamiento

Esto prepara un modelo de regresión usando random forest para predecir el espacio de venta de un automóvil primero dividiendo los datos de entrada (x) y la variable objetivo (y), luego divide y entrena el modelo random forest, una vez entrenado esta listo para hacer predicciones. Lo cual el modelo esta entrenado con un 96% de efectividad

3.4 Validación del modelo

```
sns.scatterplot(check1)
plt.grid(True, alpha = 0.3)

plt.title('Actual VS Predicted ["Selling_Price"]', fontweight = 'bold')
plt.ylabel('Price')
plt.xlabel('Count')

plt.tight_layout()
plt.show()

display(check1.reset_index(drop=True))
```

Figura 13 validación del entrenamiento

Como observamos esta parte de código nos creara un gráfico de dispersión donde mostrara el valor actual y el valor que predijo.

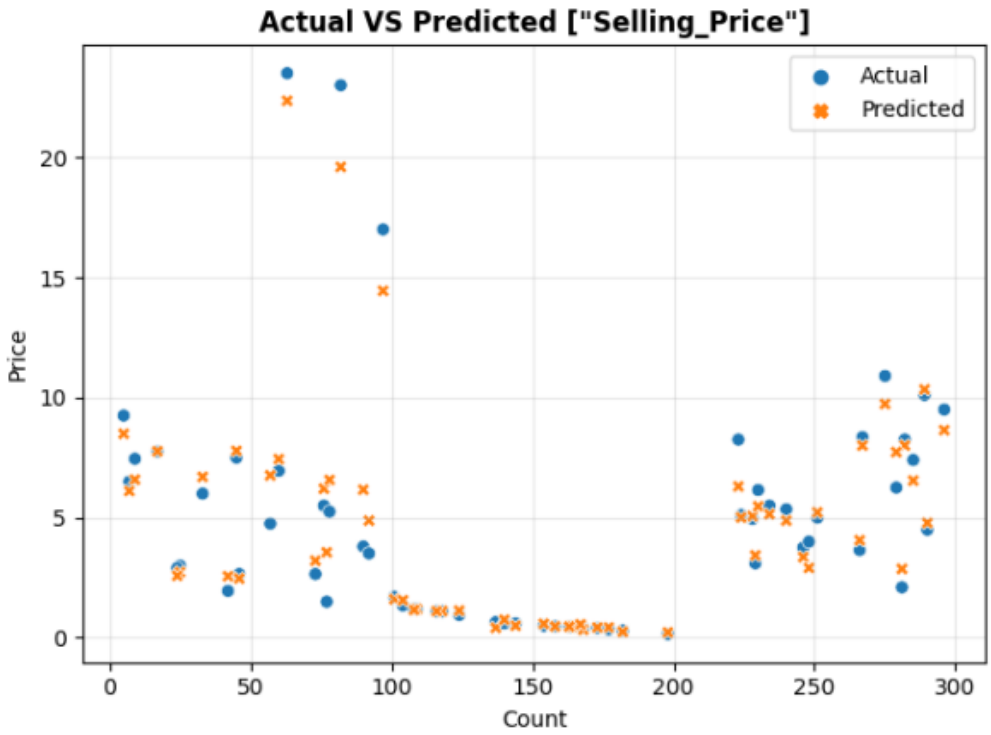

Figura 14 grafica del precio actual y el que predijo

Como podemos observar la predicción en algunos vehículos fue mayor al precio actual que estaba y en otros es menor al precio actual quizás por tener mucho kilometraje o por ser de un año 2003 haciendo que su precio bajara.

56	5.50	5.1405
57	9.50	8.6445
58	2.10	2.8664
59	7.40	6.5435
60	0.30	0.2524

Y muestra los valores actuales y el que predice los valores son en miles de dólares, como se puede apreciar algunos bajan y otros suben de valor respecto al valar actual que tenían

4. Conclusiones y trabajos futuros

En conclusión, el uso de Random Forest ha demostrado ser altamente efectivo para la predicción del precio de autos usados, principalmente por su habilidad para manejar relaciones complejas y no lineales entre las características de los datos.

La predicción precisa de los precios de los autos usados tiene un alto valor comercial, ya que permite a los vendedores fijar precios más competitivos y a los compradores tomar decisiones más informadas. También puede ser de utilidad para plataformas de compraventa de autos, aseguradoras y concesionarios de vehículos.

Para trabajos futuro seria bueno implementar la condición en la que se encuentra el auto, tener información sobre la pintura, el estado de las llantas. Implementar también un historial de accidentes ya que eso también implica en el valor del auto

Referencias bibliográficas

Tyagi, S., Sirohi, S., Singh, Y., & Vishwakarma, A. (2024, July). Hybrid Model for Predicting Used Car Prices: Integrating Natural Language Processing with Random Forest Regressor. In *2024 Second International Conference on Advances in Information Technology (ICAIT)* (Vol. 1, pp. 1-6). IEEE.

Caro Martínez, M. (2017). Sistemas de Recomendación basados en técnicas de predicción de enlaces para jueces en línea.

Rigatti, S. J. (2017). Random forest. *Journal of Insurance Medicine*, *47*(1), 31-39.

Speiser, J. L., Miller, M. E., Tooze, J., & Ip, E. (2019). A comparison of random forest variable selection methods for classification prediction modeling. *Expert systems with applications*, *134*, 93-101.

Asghar, M., Mehmood, K., Yasin, S., & Khan, Z. M. (2021). Used cars price prediction using machine learning with optimal features. *Pakistan Journal of Engineering and Technology*, *4*(2), 113-119.

Alhowaity, A., Alatawi, A. A., & Alsaadi, H. (2023). ¿Are Used Cars More Sustainable? Price Prediction Based on Linear Regression. *Sustainability*, *15*(2), 911.