Predicción de precio en venta de vehículos aplicando Machine Learning

Calin Cristian Dinga Pastae
Universidad Carlos III de Madrid
12/12/2023

Resumen: El presente artículo está enfocado en el desarrollo de un modelo de predicción utilizando Machine Learning, con el principal objetivo de estimar el precio de los automóviles. Este modelo busca ofrecer una herramienta precisa y confiable para compradores y vendedores en el mercado de vehículos de segunda mano. Para afrontar el desafío, se han tenido en cuenta una base de datos con un total de más de 300k muestras de automóviles usados, incluyendo las características principales de cada uno.

Abstract: This article is focused on the development of a prediction model using Machine Learning, with the main objective of estimating the price of automobiles. This model seeks to offer a precise and reliable tool for buyers and sellers in the second-hand vehicle market. To meet the challenge, a database with a total of more than 300k samples of used cars has been considered, including the main characteristics of each one.

1 Introducción

En el ámbito actual de la tecnología de la información, el análisis de datos y la inteligencia artificial han cobrado una relevancia significativa en múltiples sectores, incluyendo el mercado automotriz. La compra y venta de vehículos usados es una industria considerable que afecta a una amplia gama de consumidores y negocios. Un aspecto crucial en este mercado es la determinación precisa del valor de un vehículo usado, un desafío complejo debido a la diversidad de factores que pueden influir en el precio de un coche. En este contexto, el presente trabajo se centra en la creación y validación de un modelo de Machine Learning para predecir el precio de coches usados, aportando así una herramienta valiosa tanto para compradores como para vendedores.

La capacidad de predecir con precisión el precio de un vehículo basándose en sus características puede tener un impacto significativo en la transparencia y eficiencia del mercado de vehículos de segunda mano. Tradicionalmente, la valoración de un coche usado se ha basado en la experiencia y en guías de precios estándares, que a menudo no pueden capturar la complejidad y la dinámica del mercado actual. Con la creciente disponibilidad de datos y el avance de las técnicas analíticas, el Machine Learning emerge como una solución prometedora para abordar esta problemática.

Con este trabajo, aspiramos a contribuir al campo del análisis de datos en el mercado automotriz, proporcionando una herramienta innovadora para la estimación objetiva y precisa de precios de vehículos usados, beneficiando así a un amplio espectro de usuarios en este mercado dinámico y en constante evolución.

2 Dataset

2.1 Obtención de datos

El conjunto de datos que se ha utilizado para la creación del modelo proviene de una plataforma de datos abiertos, Data Word [1].

Mediante *web scraping*, se han recolectado un total de 370000 muestras de vehículos usados, obtenidos principalmente de la página principal de anuncios de eBay-Kleinanzeigen.

Este recurso seleccionado ofrece una amplia gama de información sobre vehículos usados, recopilada con el propósito de facilitar el análisis del mercado de automóviles.

El acceso al conjunto de datos se realizó a través de la descarga directa desde la plataforma. La integridad y la calidad de los

datos han sido verificadas para asegurar su fiabilidad.

Cabe destacar que los datos en bruto están en alemán, por lo que se ha manipulado los datos con una traducción al lenguaje castellano.

2.2 Características del conjunto de datos

El conjunto de datos comprende una serie de características clave de los vehículos, fundamentales para la estimación de un precio justo y coherente con el mercado actual.

Estas características incluyen el precio, la marca y modelo, la fecha de venta, etc. Muchos de estos atributos no son representativos o no tienen correlación alguna con el precio, el cual es nuestro objetivo.

Por lo tanto, se ha decidido extraer características de los datos en bruto que tienen que ver directamente con el crawler o con detalles del propio anuncio (nº imágenes, código postal, etc).

Luego, las muestras que se van a analizar y depurar se caracterizan por los siguientes atributos:

- **Precio:** Valoración de venta del vehículo del anuncio, en €.
- **Marca:** Marca del vehículo.
- **Modelo:** Modelo del vehículo.
- **Kilometraje:** Distancia total del vehículo, en km.
- **Potencia:** Potencia del vehículo, en PS (Pferdestärke)
- **Año de registro:** El año en el que el vehículo fue registrado inicialmente.
- **Combustible:** Tipo de combustible que utiliza el vehículo. (Gasolina, Diésel, Electrico, etc)
- **Gearbox:** Tipo de caja de cambios (Automática, Manual)
- **Daños:** Información sobre si el vehículo sufre algún daño significativo (Sin daños, Con daños)

Todos estos atributos escogidos tienen una consecuencia directa sobre el valor del vehículo en el mercado, tal y como vamos a analizar en el siguiente punto.

2.2.1 Impacto de las características sobre el precio final

Vamos a analizar las características mencionadas con el fin de analizar el impacto que puede tener cada una de ellas sobre el precio a estimar.

- **Marca y modelo**
  Las diferentes marcas y modelos tienen diferentes reputaciones en términos de calidad, fiabilidad y rendimiento, lo que se refleja en el valor de reventa.

- **Kilometraje**
  Este atributo es indicador directo del uso y desgaste del coche, por lo que es un factor clave a la hora de evaluar un vehículo en el mercado. Generalmente, cuanta más utilidad se le ha dado al vehículo, más bajo es el precio.

- **Año de registro**
  Indica la edad del vehículo. Los coches más nuevos suelen tener un precio más elevado debido a que portan mejor tecnología y tienen un consumo más eficiente, entre otras cosas.

- **Potencia**
  Afecta directamente al rendimiento del vehículo y es un factor importante para muchos compradores.

- **Combustible**
  Afecta al coste posterior a la compra del vehículo y, por tanto, a los compradores. Por ejemplo, los vehículos eléctricos pueden tener un precio más alto debido a la tecnología e impacto medioambiental.

- **Caja de cambios**
  Influye en la experiencia de conducción y por tanto al precio. Aunque las preferencias personales son muy variadas, generalmente los coches automáticos tienen un precio más elevado, ya que se requiere implementar más tecnología

- **Daños**
  La presencia de daños significativos afecta negativamente al precio de venta, por lo que es un factor clave.

2.3 Ingeniería de datos

Se han implementado pasos críticos para refinar, depurar y optimizar nuestro conjunto de datos.

[1]https://data.world/data-society/used-cars-data

2.3.1 Creación del atributo 'Edad'

Se ha manipulado la característica 'Año de Registro', operándolo con el año actual, para obtener el nuevo atributo 'Edad'.

Esto transforma el uso del vehículo en un formato más intuitivo y efectivo para la modelización.

2.3.2 Conversión en potencia

El conjunto de datos original expresa la potencia en PS (Pferdestärke), una unidad común en Alemania, pero no conocida globalmente.

Es por ello por lo que se ha transformado esta medida a Caballos de Vapor (CV)[2], una unidad más estándar en el análisis del mercado automóvil.

2.3.3 Eliminación de valores faltantes

Se deben eliminar las muestras que tengan un valor nulo en alguno de sus atributos.

Evaluando el dataset (utilizando Python), se han encontrado los siguientes valores faltantes:

Atributo	Valores faltantes
Modelo	20498
Gearbox	20223
Combustible	33415

En total, han sido eliminadas más de 50000 muestras.

Otra opción habría sido rellenar los datos faltantes con una predicción, pero sería una tarea más demandante, sobre todo a la hora de rellenar el modelo.

2.3.4 Eliminación de outliers

Para eliminar las muestras sobrantes, vamos a analizar el máximo y mínimo de cada atributo numérico:

Atributo	Min.	Max.
Precio	0	200e9
Año de registro	1000	9999
Potencia	0	20000
Kilometraje	5000	150000

Evidentemente, a excepción del kilometraje, no solo hay datos fuera de lugar, sino que hay datos que no tienen sentido.

Debemos eliminar estos datos sobrantes según un criterio óptimo, que nos ayude a quedarnos con el mayor número de muestras útiles.

2.3.4.1 Precio

Se han eliminado todas las muestras iguales a 0€. Además, se ha puesto un límite superior de 70000€, ya que, a partir de este valor, los precios de las muestras son más dispersos e imprecisos y se pueden considerar outliers.

2.3.4.2 Año de registro

Se han eliminado las muestras que no tendrían sentido, como son las muestras provenientes de años posteriores al año actual o inferiores a la creación del primer vehículo a motor.

Los coches más antiguos que se ha tenido en cuenta han sido fabricados en 1973, mientras que las muestras más nuevas se fabricaron en 2019.

No tendría sentido tener en cuenta coches más antiguos ya que pueden tener el efecto de valorizarse debido a que empiezan a considerar clásicos y, por tanto, outliers que solo añadirían imprecisión a nuestro modelo.

2.3.4.3 Potencia

Generalmente, los coches más comunes llevan una potencia de mercado[3] de entre 100 o 200 CV. El resto de las potencias son de coches muy exclusivos y potentes.

Se ha impuesto un límite inferior de 50CV, que suelen ser de coches pequeños, antiguos y de baja potencia.

En cuanto al límite superior, se ha impuesto un máximo de 400CV, añadiendo marcas cuyos modelos deportivos son posibles (aunque raros) verlos en anuncios, como pueden ser Mercedes o Porsche. Estos no se han tenido en cuenta como outliers ya que suponen un gran número de muestras, y eliminarlas supondría no tener suficientes ejemplos para estas marcas.

Una vez depuradas las muestras, nos hemos quedado con un total del 270k muestras.

2.4 Análisis de datos

Una vez descritas, analizadas y depuradas las muestras, debemos profundizar en el análisis de los datos.

Este paso es crucial para entender patrones, tendencias y posibles correlaciones de las muestras exportadas.

[2] 1PS = 0.98632CV

[3] la potencia de mercado se define como la potencia que se le atribuye a un coche a la hora de venta, y no es completamente representativo de la potencia que tiene el coche realmente, aunque el margen de error sea pequeño.

2.4.1 Análisis de cada atributo

En primer lugar, vamos a explorar algunos de los atributos por separado, con el fin de entender mejor nuestro dataset.

2.4.1.1 Precio

Análisis estadístico básico:

- **Rango:** 300 a 70000€
- **Media:** 6382,89€
- **Mediana:** 3900€
- **Deviación estándar:** 7272€

Vemos que la mediana es considerablemente inferior a la media, lo que sugiere que la mayoría de las muestras están por debajo de la media. Esto se traduce en que las muestras con un elevado precio suben considerablemente la media. Veámoslo gráficamente:

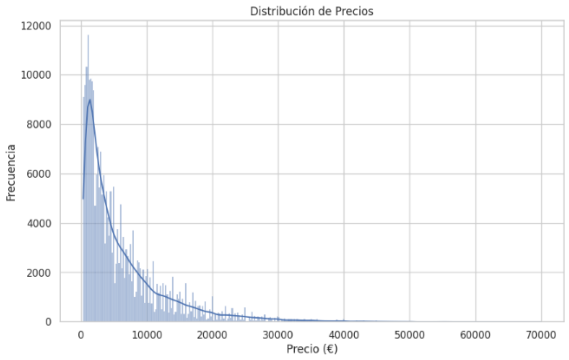

Figura 1: Distribución de precios

Observamos que la mayoría de los vehículos están valorados entre 300 y 5000€, y la distribución va disminuyendo conforme aumenta el precio.

Se podría considerar la posibilidad de eliminar también los vehículos de, por ejemplo, más de 40000€, pero hay que tener en cuenta que la mayoría de estos coches suelen ser marcas y modelos altamente valorizados en el mercado.

Será nuestro trabajo implementar un modelo que se adapte a esta característica.

2.4.1.2 Marca

La distribución es la siguiente:

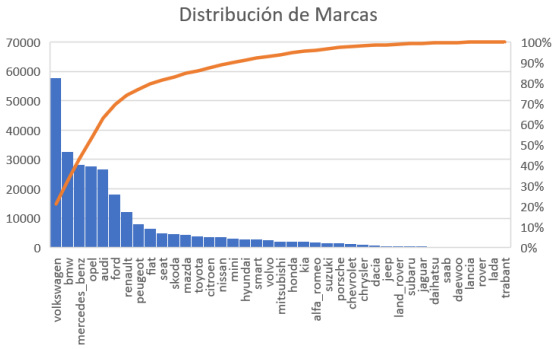

Figura 2: Distribución de Marcas

Como característica principal, observamos que Volkswagen, BMW, Mercedes, Opel y Audi suponen algo más de un 50% del total de muestras, dado posiblemente a que el origen de la base de datos es alemán.

Aunque no es lo ideal, ya que el modelo será más preciso para las marcas con más datos, el conjunto de datos es suficientemente grande para abarcar de manera precisa el resto de las marcas.

2.4.1.3 Kilometraje

El kilometraje en esta base de datos esta recolectada de forma discreta, es decir, no hay valores continuos, sino rangos:

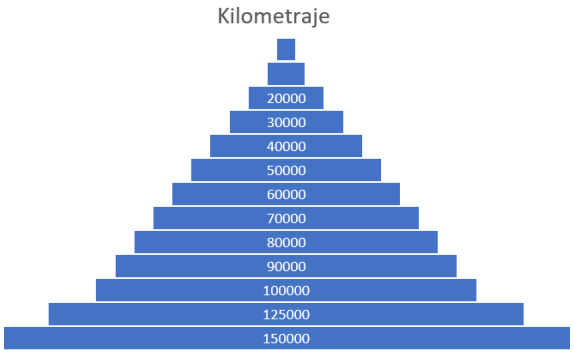

Figura 3: Distribución del kilometraje

Se observa que la mayoría de las muestras están en el rango de 150000km. Esto no es lo ideal, ya que a partir de este valor se pierde precisión en los datos, ya que podemos estar hablando de, por ejemplo, 200000km y 400000km a la vez.

Sin embargo, si se supera este kilometraje, se intuye que el coche está bastante usado y desgastado.

2.4.1.4 Año de Registro

Análisis estadístico básico:

[1]https://data.world/data-society/used-cars-data

- **Rango:** 1973 a 2019
- **Media:** 2003
- **Mediana:** 2004

Veamos la distribución:

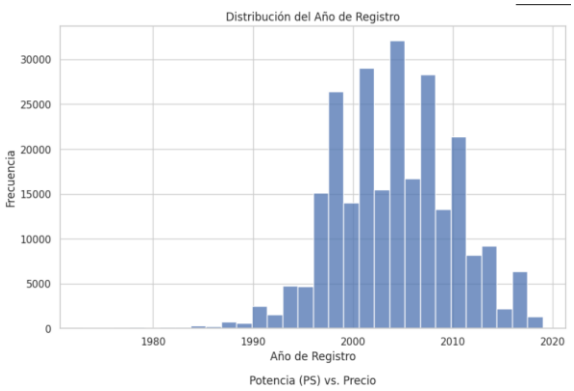

Figura 4: Distribución del Año de Registro

Observamos que los coches son relativamente modernos.

2.4.1.5 Potencia

Para la potencia se tiene una distribución bastante estable entre los 50CV y 200CV. Potencias más altas son más raras.

2.4.1.6 Combustible

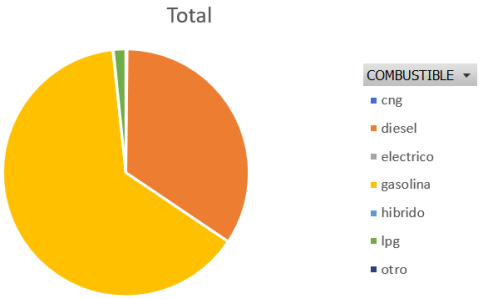

Figura 5: Distribución de Combustible

Observamos que el diésel y gasolina predominan sobre el resto. Esto está directamente relacionado con la distribución del año de registro de las muestras, ya que los eléctricos o híbridos son más comunes en tiempos más modernos.

2.4.1.7 Caja de cambios

Por la misma razón que con el combustible, las muestras de coches automáticos solo representan el 23% del total.

2.4.1.8 Daños

Apenas el 8% de las muestras recogidas sufren daños.

2.4.2 Atributo vs Precio

Ahora vamos a analizar la influencia de los atributos con el valor objetivo, el precio.

2.4.2.1 Kilometraje vs precio

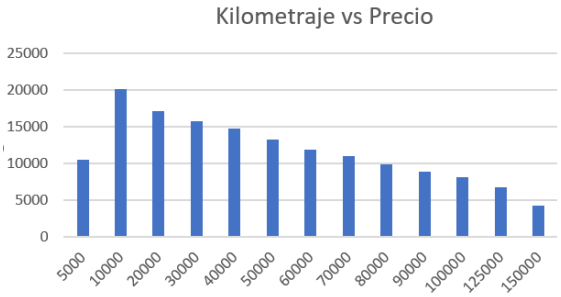

Figura 6: Kilometraje vs Promedio Precio

Vemos que, salvo el rango de 5000km, el resto de las muestras es bastante coherente e intuitivo. Cuanto más kilometraje, más bajo es el promedio del precio.

Se ha analizado los datos con kilometraje de 5000km y se ha llegado a la conclusión de que el promedio del precio se debe a que la gran mayoría de estas muestras sufren daños.

2.4.2.2 Marca vs Precio

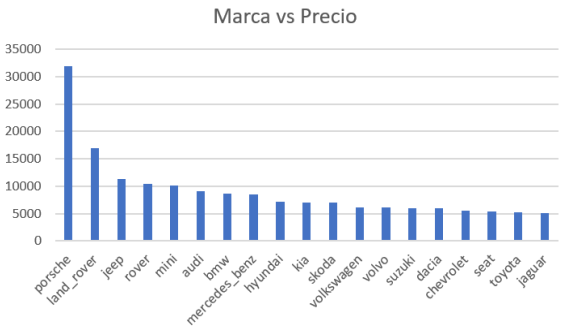

Figura 7: Marca vs Precio

Vemos que la marca tiene impacto directo en el precio. Marcas consideradas valiosas como Porsche o Land Rover tienen un promedio del precio bastante más elevado.

Sin embargo, también debemos tener en cuenta que el modelo también tiene un impacto directo en el precio. Por ejemplo, para la marca Audi, su distribución de precios según el modelo es el siguiente.

[2] 1PS = 0.98632CV

[3] la potencia de mercado se define como la potencia que se le atribuye a un coche a la hora de venta, y no es completamente representativo de la potencia que tiene el coche realmente, aunque el margen de error sea pequeño.

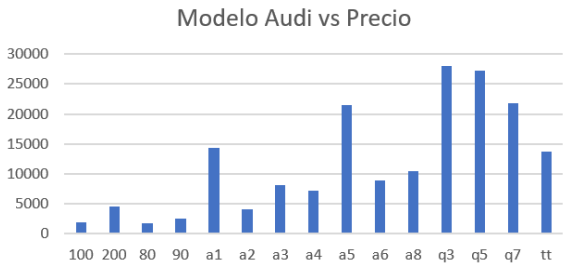

Figura 8: Modelo Audi vs Precio

2.4.2.3 *Potencia vs Precio*

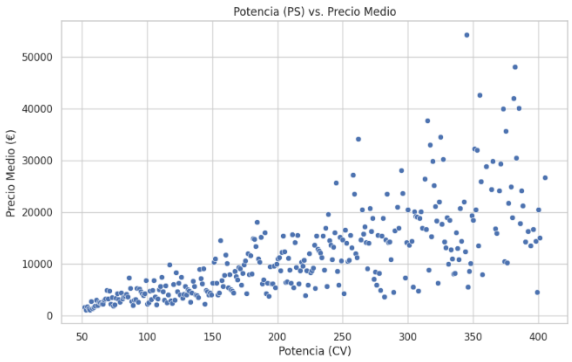

Figura 9: Potencia vs Precio

No solo los precios suben conforme sube la potencia, sino que los precios de los coches más potentes son más dispersos.

2.4.2.4 *Caja de cambios vs Precio*

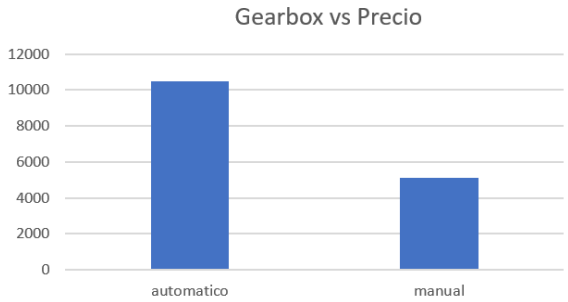

Figura 10: Caja de cambios vs Precio

Se confirma que, aunque la preferencia del conductor puede ser variada, los coches automáticos tienen más valor debido a su demanda tecnológica.

2.4.2.5 *Combustible vs Precio*

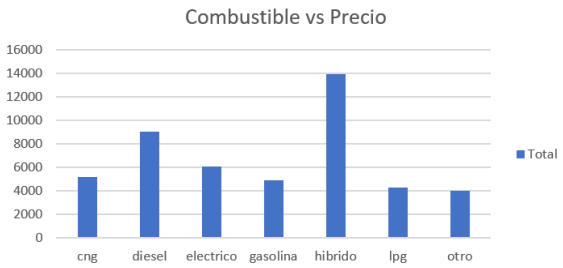

Figura 11: Potencia vs Precio

Vemos que, en nuestro conjunto de datos, los híbridos son los coches que más valor tienen de media. No obstante, sabemos del apartado 2.4.1.6 que se tienen pocas muestras respecto a gasolina y diésel.

2.4.2.6 *Daños vs Precio*

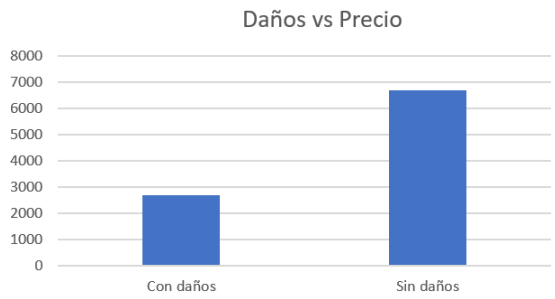

Figura 12: Potencia vs Precio

Evidentemente, podemos confirmar que los coches con daños están menos valorizados

3 *Implementación del modelo*

Una vez analizado y entendido el comportamiento de nuestro conjunto de datos, vamos a implementar nuestro modelo.

La elección del modelo a utilizar es un paso crucial para la precisión y eficacia de las predicciones.

Se ha optado por la elección de un modelo de regresión. Esto es debido a la naturaleza de nuestro valor objetivo, el precio, una variable continua y cuantitativa.

3.1 Preparación del dataset

Debemos manipular nuestro conjunto de dataset, con el fin de traducir las muestras a un lenguaje numérico que un modelo pueda entender.

Los atributos que no están en formato numérico son:

[1]https://data.world/data-society/used-cars-data

- **Marca**
- **Modelo**
- **Combustible**
- **Gearbox**
- **Daños**

Como primera medida, se han conjuntado los atributos de marca y modelo en un nuevo atributo 'Marca_Modelo'. Esto tiene sus ventajas y desventajas.

El inconveniente será que el modelo dejará de distinguir las marcas de los modelos, tratando a todos los coches como diferentes independientemente de la marca, perdiendo posible información útil, ya que hay marcas que son valorizadas independientemente del modelo.

La ventaja será que no se dará ningún margen a posibles confusiones, es decir, aplicar un modelo a una marca no existente (por ejemplo, mercedes_a4), además de reducir la dimensionalidad.

Se ha aplicado *Label Encoding* para codificar este y los demás atributos. El dataset ha quedado modificado de la siguiente manera:

Marca_Modelo	Cod.
alfa_romeo_145	0
audi_a4	12
citroen_c1	39
fiat_punto	64
etc.	etc.
volvo_xc_reihe	256

Combustible	Cod.
cng	0
Diésel	1
Eléctrico	2
Gasolina	3
Híbrido	4
lpg	5
Otro	6

Gearbox	Cod.
Automático	0
Manual	1

Daños	Cod.
Sin daños	0
Con daños	1

De esta manera, ya podemos entrenar los modelos.

3.2 Evaluación de modelos

Vamos a evaluar para poder elegir con criterio que modelo se adapta mejor a nuestra aplicación.

Para cada uno de los modelos que se van a examinar a continuación se llevan a cabo los siguientes pasos: Entrenamiento, Predicción y Evaluación.

Se ha separado el conjunto de datos en conjunto de entrenamiento y test para la evaluación del modelo. Los parámetros que se van a tener en cuenta para la examinación del modelo son los siguientes.

-R^2 (coeficiente de Determinación)

Indica la proporción de la varianza de la variable dependiente (precio) respecto al de las variables independientes. Rango de 0 a 1.

-RMSE (Raíz del error cuadrático medio)

Indicará cual es el error medio en € que comete el modelo

-Median-AE (Mediana del Error Absoluto)

Similar a RMSE, pero es más robusto frente a outliers. Nos será útil ya que no se verá afectado el modelo negativamente frente a predicciones de coches caros.

-Gráfica Predicted vs Real

Con esta gráfica se podrá visualizar el rendimiento del modelo, comparando los valores reales con los que predice el modelo. Nos puede servir de igual modo para ver si el modelo se equivoca con precios bajos o altos de igual manera.

3.2.1 Modelos lineales

Los modelos lineales trabajan bajo el supuesto de que existe una relación lineal entre las características del coche y el precio.

Aunque esto no es cierto para todos los atributos, no sirve como base para comparar el rendimiento de modelos más complejos.

Cabe destacar que se han normalizado los datos para probar estos modelos.

3.2.1.1 Linear Regresor

Modelo básico de regresor Lineal.

- **R^2:** 0.65
- **RMSE:** 4124,38
- **Median-AE:** 2065,13

[2] 1PS = 0.98632CV

[3] la potencia de mercado se define como la potencia que se le atribuye a un coche a la hora de venta, y no es completamente representativo de la potencia que tiene el coche realmente, aunque el margen de error sea pequeño.

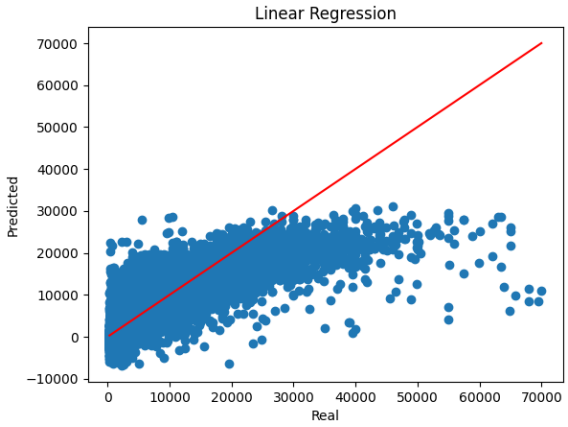

Figura 13: Rendimiento Linear Regressor

3.2.1.2 Otros

Se han evaluado de igual manera los modelos *Ridge Regressor* y *Lasso Regressor*. Los resultados son extremadamente similares a los de *Linear Regressor*.

3.2.2 Modelos no lineales

Procedemos a incluir modelos no lineales. Estos modelos se caracterizan por detectar relaciones entre características no lineales, lo que puede incrementar el rendimiento de nuestro sistema.

Se ha hecho uso de validación cruzada para encontrar los hiperparámetros óptimos de los modelos que lo requieran.

3.2.2.1 Kernel Ridge Regression

Combina la regularización de Ridge Regressor con el uso de kernels, permitiendo calcular relaciones no lineales en los datos.

- **R^2:** 0.84
- **RMSE:** 2749,29
- **Median-AE:** 886,22

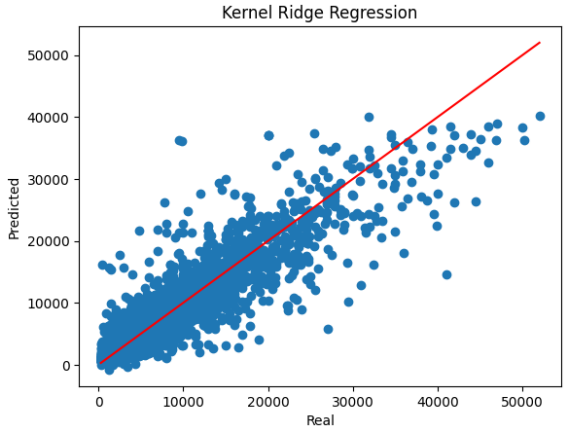

Figura 14: Rendimiento Kernel Ridge Regressor

NOTA: Se debe tener en cuenta que para Kernel Ridge se ha tenido que reducir la dimensionalidad del modelo, debido a la memoria RAM que se a utilizado para el estudio.

3.2.2.2 Random Forest Regressor

Los árboles de decisión son capaces de capturar relaciones no lineales. Random Forest ensambla varios árboles de decisión, haciéndolo más robusto al sobreajuste.

Los hiperparámetros óptimos que se han encontrado son:

-n_estimators: 150 (puede ser más)
-max_depth: 20
-max_features: 'sqrt'

- **R^2:** 0.91
- **RMSE:** 2042,75
- **Median-AE:** 637,85

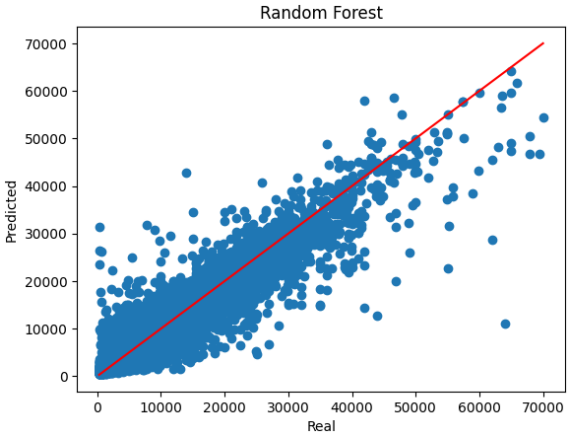

Figura 15: Rendimiento Random Forest

3.2.2.3 Gradient Boosting Regressor

Similar a Random Forest, pero intenta corregir los errores de árboles anteriores. Puede ser propenso al sobreajuste. Se han utilizado los mismo hiperparámetros que en Random Forest

- **R^2:** 0.89
- **RMSE:** 2316,64
- **Median-AE:** 623,31

[1]https://data.world/data-society/used-cars-data

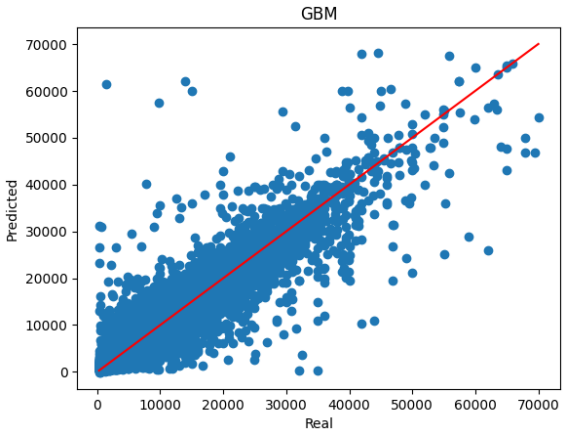

Figura 16: Rendimiento GBM

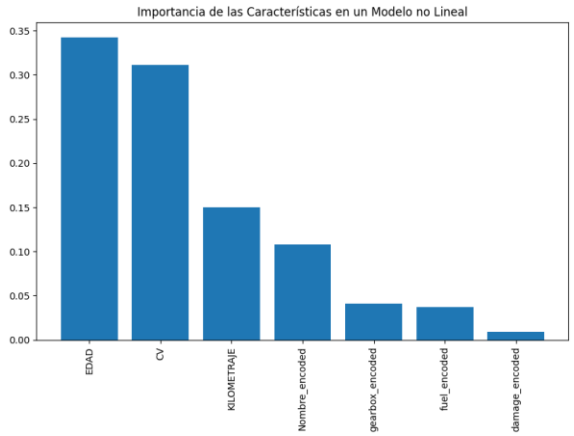

Figura 18: Coeficientes Random Forest

3.3 Elección del modelo

Una vez evaluada la variedad de modelos tanto lineales como no lineales, podemos elegir con criterio el modelo para nuestra aplicación.

Hemos observado que los modelos no lineales obtienen mejores resultados que los lineales. Aunque muchas de las características son lineales con el precio, como hemos estudiado en el aparatado 2.4.2, la relación de marca_modelo vs precio no es tan intuitiva y lineal para muchas de las marcas.

Es por ello por lo que los modelos lineales son más efectivos para nuestro problema, ya que son capaces de relacionar la marca y el modelo con el precio de una forma no lineal.

Para entenderlo mejor, vamos a comprobar el peso o coeficientes pertenecientes a cada una de las características en un modelo lineal y otro no lineal.

Estos coeficientes están entre 0 y 1, siendo 1 la suma de los coeficientes de cada atributo.

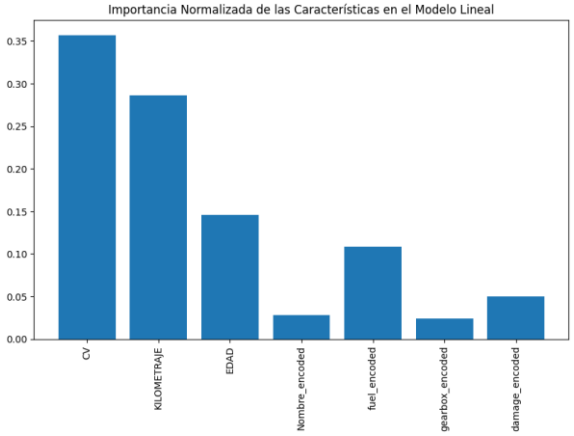

Figura 17: Coeficientes Linear Regressor

Si se comparan las figuras 17 y 18, se puede apreciar que la Edad, el Kilometraje y el Nombre (Marca y Modelo) son atributos más tenidos en cuenta por parte de los modelos no lineales, resultando directamente en predicciones más precisas y fiables.

Es por ello por lo que Random Forest ha sido el modelo por el que se ha optado, debido a que es el que mejores resultados ha obtenido dentro de los modelos no lineales, con una mediana de error de 600€.

4 Simulación y resultados

A modo de ejemplo, vamos a simular interactivamente con nuestro modelo, viendo como el cambio de parámetros impacta a nuestro modelo.

4.1 Desgaste y antigüedad

Para un mismo modelo (Audi_a4) de 170CV, automático y sin daños:

Nombre	KM	Edad	Precio
Audi_a4	10k	**1**	**33048,34**
Audi_a4	100k	4	**18653,08**
Audi_a4	200k	8	**8057,75**

Observamos que efectivamente, cuanto más uso se le ha dado al coche, más decrece el precio.

4.2 Potencia

Para un Ford Focus de 4 años y 50000km, con gasolina, manual y sin daños:

CV	Precio (€)
100	**8322,34**
150	**11869,26**
170	**14496,91**

[2] 1PS = 0.98632CV

[3] la potencia de mercado se define como la potencia que se le atribuye a un coche a la hora de venta, y no es completamente representativo de la potencia que tiene el coche realmente, aunque el margen de error sea pequeño.

El modelo se comporta como esperamos. Más potencia, más caro.

4.3 Combustible

Para el mismo coche (Citroen C5), con una potencia de 170 CV, 50000km de 4 años, con cambio manual y sin daños:

Fuel	Precio (€)
cng	**16657,14**
Diésel	**18347,70**
Eléctrico	**17819,25**
Gasolina	**14487,20**
Híbrido	**14847,32**
lpg	**14579,65**

Vemos que no afecta mucho el tipo de combustible. Esto no es muy preciso, ya que en el mercado actual los motores híbridos y eléctricos están más valorizados.

4.4 Cambio

Para un mismo modelo Honda Civic de 180 CV, con 80000km de 6 años, diésel y sin daños:

Fuel	Precio (€)
Manual	**17288,31**
Automático	**20209,28**

Como esperábamos, cuando se le añade el cambio automático se eleva el precio.

4.5 Daños

Para un mismo vehículo:

Daños	Precio (€)
Sin daños	**14595,96**
Con daños	**12423,18**

Como es lógico, baja el precio

4.6 Marca

Para un mismo coche con 200 CV, 50000km de 4 años, gasolina, automático y sin daños:

Marca_Modelo	Precio (€)
Seat Ibiza	**18817,85**
Volkswagen Golf	**19744,95**
Mercedes-Benz Clase E	**22056.41**
Porsche 911	**29563.17**

Vemos que cuanto más lujosa es la marca, más elevado es el precio.

5 Conclusión

En la presente investigación nos hemos propuesto crear un modelo robusto para la predicción de precios en vehículos usados, haciendo uso de una amplia base de datos de venta de coches en Alemania, obtenidos por un web crawler.

Se ha realizado un análisis extenso y preciso para comprender el comportamiento del conjunto de datos seleccionado.

Posteriormente, se ha realizado un estudio para comparar y estudiar diferentes modelos, tanto lineales como no lineales.

Random Forest resultó ser el modelo que mejor se adaptaba a las características de nuestro conjunto de datos, siendo esta elección respaldada por unos resultado sólidos y coherentes.

Este modelo no solo cumple con nuestras expectativas iniciales, sino que también establece una base firme para futuras investigaciones y mejoras.

6 Posibles mejoras y trabajo futuro

Hemos visto que pese a tener unos resultados coherentes, el modelo tiene algunas impresiones, como la invariabilidad del combustible. Esto es debido a la antigüedad de los datos, ya que predominan los coches diésel y gasolina.

Además, se ha podido comprobar que, para algunos vehículos, aumentar la edad del coche supone un aumento en el precio. Esto es debido a la falta de muestras de esa respectiva marca y modelo.

Sería de gran ayuda obtener una base de datos más actualizada, y aunque no necesariamente con más datos, sí una con mejor distribución de muestras para todas las marcas y modelos. Esto mejoraría nuestro modelo y por tanto tendríamos una aplicación más precisa.

[1]https://data.world/data-society/used-cars-data