FACULTAD DE CIENCIAS

ECONÓMICAS Y EMPRESARIALES

Modelos predictivos en el contexto de la transformación digital del sector automovilístico: una aproximación mediante regresión lineal múltiple

Autor: Paula Díaz Shaw

5º E-3 Analytics

Tutora: María de las Mercedes Barrachina Fernández

Madrid

Abril 2025

Resumen

Este Trabajo de Fin de Grado tiene como principal objetivo llevar a cabo una evaluación sobre la capacidad de un modelo de regresión lineal múltiple como herramienta eficaz para conseguir ayudar a concesionarios de coches a estimar el que debería ser el precio adecuado de un vehículo en función de características fácilmente observables. Para conseguirlo, se parte un conjunto de datos real, y se ha desarrollado un proceso completo de limpieza, transformación y selección de variables, seguido de la implementación y validación del modelo utilizando técnicas de aprendizaje supervisado. Tras dividir la muestra en entrenamiento y test, se ha implementado el modelo, evaluando su rendimiento mediante métricas como el error cuadrático medio y el error absoluto medio.

Aunque el modelo muestra un buen ajuste en el conjunto de entrenamiento, los resultados en test revelan importantes limitaciones predictivas. Las predicciones tienden a centrarse en valores medios, sin captar adecuadamente la variabilidad de los precios. Se concluye que esto puede ser consecuencia de factores clave que no han sido incluidos en el análisis, como el estado real del vehículo, su historial o la percepción subjetiva de marca y modelo. Aun así, el proceso ha permitido entender en profundidad tanto la lógica como las limitaciones de los modelos de regresión en contextos reales, aportando valor metodológico y sirviendo como base para futuros desarrollos más precisos en el ámbito de la digitalización del sector automovilístico.

Palabras clave: regresión lineal múltiple, Cross-validation, desviación típica, normalización, multicolinealidad, sector automovilístico, concesionario, precio, error cuadrático medio, error absoluto medio, variables independientes.

Abstract

The main objective of this Final Degree Project is to evaluate whether a multiple linear regression model can be an effective tool to help dealers and salespeople to estimate the appropriate price of a vehicle based on easily observable characteristics, such as year, mileage, fuel type, transmission, vehicle condition or make. Starting from a real data set, a complete process of cleaning, transformation and variable selection has been carried out, followed by the implementation and validation of the model using supervised learning techniques. After splitting the sample into training and test, the model was implemented, evaluating its performance using metrics such as root mean square error and mean absolute error.

Although the model shows a good fit in the training set, the test results reveal important predictive limitations. Predictions tend to focus on mean values, without adequately capturing price variability. This may be due to key factors that have not been included in the analysis, such as the actual condition of the vehicle, its history or the subjective perception of make and model. Even so, the process has provided an in-depth understanding of both the logic and limitations of regression models in real-life contexts, providing methodological value and serving as a basis for future, more precise developments in the field of of digitalisation in the automotive sector.

Key words: *multiple linear regression, Cross-validation, standard deviation, normalisation, multicollinearity, automotive industry, dealer, price, mean square error, mean absolute error, independent variables.*

ÍNDICE

1. Introducción **6**

1.1. Objetivos de estudio **7**

1.2. Contexto Actual del sector automovilístico **7**

1.2.1. Evolución del vehículo eléctrico **9**

1.2.2. Rol de los concesionarios en la cadena de valor **10**

1.3.1. Modelos de regresión lineal múltiple **11**

1.3.2. Fases del modelo de regresión lineal múltiple empleado **13**

2. Descripción de la base de datos **14**

2.1 Análisis de las variables **14**

2.2 Estadísticos descriptivos **16**

3. Desarrollo del modelo de regresión **17**

3.1. Limpieza y preparación de Datos **18**

3.1.1. Limpieza de datos **18**

3.1.3. Transformación de variables **20**

3.1.3. Revisión de la multicolinealidad **22**

3.2. Selección de las variables relevantes **26**

3.3. División del conjunto de datos: entrenamiento y test **27**

3.4. Implementación del modelo predictivo **30**

3.5. Resultados obtenidos y análisis interpretativo **33**

4. Conclusiones **35**

ÍNDICE DE ILUSTRACIONES

Ilustración 1: fórmula correspondiente a los modelos de regresión lineal múltiple 12
Ilustración 2: código de R correspondiente a la eliminación de duplicados 18
Ilustración 3: código de R correspondiente a la comprobación de valores faltantes 19
Ilustración 4: gráfico destinado a la identificación de outliers 19
Ilustración 5: imagen que demuestra la conversión de las variables categóricas en factor 21
Ilustración 6: imagen que demuestra la normalización de las variables numéricas 22
Ilustración 7: error de multicolinealidad 23
Ilustración 8: resultados del VIF 24
Ilustración 9: matriz de correlaciones 25
Ilustración 10: separación de los datos en entrenamiento y test 27
Ilustración 11: coeficientes del modelo 28
Ilustración 12: comprobación de la capacidad predictiva del modelo 30
Ilustración 13: resultados de la capacidad predictora 31
Ilustración 14: resultados de cross-validation

1. Introducción

A medida que avanzamos en el año 2025, la industria automotriz también lo hace, y no sería exagerado hablar de una evolución a gran escala. Actualmente, nos encontramos en un punto de inflexión en este ámbito, caracterizado indudablemente por una evolución muy rápida en relación con la tecnología de vehículos y una fluctuación importante en las tendencias del mercado global.

El contexto en el que se encuentra el mercado en estos momentos está marcado por continuas fluctuaciones en la demanda, cambios constantes en las regulaciones y ajustes geopolíticos que hacen difícil desarrollar una estrategia de mercado. A todo esto, se suma una clara tendencia a la transición hacia la sostenibilidad con el auge en la producción de vehículos eléctricos en la gran mayoría de ciudades importantes europeas. Cabe recalcar que en Europa, en 2024, se ha experimentado una disminución en la producción de vehículos ligeros, en general. Sin embargo, se espera una recuperación en 2026. (Wall, 2025)

Este Trabajo de Fin de Grado (TFG) tiene como objetivo principal desarrollar un modelo algorítmico que utilice técnicas de aprendizaje supervisado para analizar cómo las características específicas de los automóviles influyen en sus precios de mercado. Para ello, se utilizará un conjunto de datos exhaustivo y, mediante un modelo de regresión lineal múltiple, se analizarán una serie de características detalladas de distintos vehículos con el fin tratar de desarrollar una herramienta analítica que pueda servir como referencia para concesionarios y vendedores privados, facilitándoles la gestión exitosa de las complejidades del mercado automotriz actual y futuro.

Además, con esta investigación se busca proporcionar conocimientos valiosos no solo orientados a la fijación de precios sino también a entender mejor los gustos y preferencias del consumidor general, la viabilidad de los diferentes modelos de vehículos y las estrategias de entrada al mercado en un período de transición energética y tecnológica.

No obstante, dado que el modelo se construye exclusivamente a partir de las variables disponibles, cabe la posibilidad de que los resultados obtenidos no capten completamente la realidad del mercado si existen factores externos o subjetivos que escapan a los datos analizados. Esta limitación, lejos de restar valor al estudio, pondría de manifiesto la importancia de seguir ampliando y refinando las herramientas de análisis en contextos complejos como el del sector automotriz.

1.1. Objetivos de estudio

Si bien ya se ha fijado el principal objetivo de este Trabajo de Fin de Grado, a continuación vamos a establecer una serie de objetivos destinados a encaminarnos hacia la consecución del primordial. Son los siguientes:

1. **Desarrollo y Optimización de un Modelo Predictivo**: Construir y perfeccionar un modelo de regresión lineal múltiple que no solo predice con precisión los precios de los vehículos basándose en sus características, sino que también se adapte a las nuevas tendencias del mercado y las expectativas de los consumidores.
2. **Comprensión Profunda de la Regresión Lineal Múltiple**: Profundizar en el entendimiento del funcionamiento de los modelos de regresión lineal múltiple, explorando cómo se configuran, sus requerimientos estadísticos y las fases de su desarrollo, desde la hipótesis inicial hasta la interpretación de los resultados.
3. **Determinar si un modelo de regresión lineal múltiple es una herramienta eficaz para conseguir nuestro objetivo principal:** predecir el precio de un vehículo en función de una serie de características fácilmente observables y, en caso negativo, comprender las razones por las que, en determinados contextos, la predicción puede no ser tan precisa como cabría esperar.

Estos objetivos están pensados para mejorar la forma en que entendemos y predecimos los cambios en la industria automovilística, mezclando ideas sencillas de economía con técnicas actuales de análisis de datos. El propósito es hacer que el análisis de datos sea práctico y útil, ofreciendo soluciones reales a los problemas cotidianos que enfrenta la industria.

1.2. Contexto Actual del sector automovilístico

El sector automovilístico tiene un papel crucial en nuestro país. Es un sector que tiene una contribución directa al PIB y con una indudable relevancia en términos de empleo de personal. Además, tiene una alta interacción con otros sectores, lo cual se debe principalmente a la larga cadena de producción que implica el desarrollo de un vehículo, y a su alta capacidad innovadora (Heras, s.f.).

España es el segundo mayor productor de vehículos en Europa y en el año 2021 fue el noveno a nivel mundial. Esto se debe, en gran parte, a la existencia de nueve grupos

multinacionales que operan en España a gran escala en la producción de vehículos. Entre estos se encuentran marcas de absoluto renombre en la industria a nivel mundial como son Volkswagen Group, Mercedes-Benz, Renault o Stellantis (incluye marcas como Fiat). Estas desarrollan sus operaciones en 17 plantas de ensamblaje. En 2023, se estima que estas plantas produjeron alrededor de 2.45 millones de vehículos. Además, destacan también por su capacidad de fabricar vehículos de características muy diversas, fabricando más de 40 modelos de automóviles incluyendo 22 modelos electrificados (Heras, s.f.).

En la industria manufacturera española, la automoción ocupa el tercer puesto de importancia siendo únicamente superado por el sector agroalimentario y el de fabricación de productos metálicos. Más específicamente, el sector contribuye con el 6.9% del empleo manufacturero y el 8.6% del Valor Añadido Bruto. Se ha llegado a estimar que por cada euro que produce de forma directa este sector, genera 0.8 euros adicionales en otros sectores con los que trabaja. Sin duda es un sector de una importancia indudablemente grande en la economía (Díaz, 2024).

Sin embargo, el sector automotriz español no solo tiene impacto nacional. En relación con el comercio internacional, se ha establecido por expertos que el 89% de los vehículos y el 60% de los componentes que fueron fabricados en el año 2023 fueron exportados. En cuanto a los destinos, la gran parte de estos se quedaron dentro de la Unión Europea. No obstante, este no fue el único destino ya que mercados en África, América y Asia también son significativos. En el año 2023 el superávit comercial del sector alcanzó los 18.800 millones de euros (Heras, s.f.).

Cabe destacar que España no es solo un país relevante mundialmente en términos de producción de vehículos, sino también en términos de inversión en I+D+i en este ámbito. Se invierten una media de 4.000 millones al año en modernización e innovación de plantas. Esta inversión va principalmente destinada a la automatización de las fábricas españolas y a la mejora de su eficiencia. Actualmente se ha fijado una cifra de 1.199 robots industriales por cada 10.000 empleados, cifra sólo ligeramente inferior a la de Alemania, país referente en el sector (Heras, s.f.).

Cabe destacar también en este ámbito del sector automotriz a nivel mundial que a pesar de los problemas sufridos en los últimos años en relación con la pandemia y las tensiones de suministros globales, el sector ha conseguido demostrar que es resiliente y que

tiene capacidad de recuperación. Esto se refleja en que, si bien es cierto que el sector aún no ha recuperado las cifras y niveles previos a la pandemia, principalmente debido a la debilidad de los mercados europeos, sí ha experimentado una crecimiento de producción del 5,8% en 2022 y del 10,4% en el 2023 (Díaz, 2024).

Todo lo hasta ahora mencionado evidencia la importancia estratégica de este sector para la economía española, no solo en el ámbito económico y laboral, sino también en el campo de la innovación.

1.2.1. Evolución del vehículo eléctrico

En relación con los vehículos eléctricos, un informe de la ANFAC sobre la electromovilidad en España durante el 4º trimestre de 2024 establece que nuestro país ha progresado significativamente hacia la electrificación. Diferenciamos los vehículos eléctricos puros y enchufables (BEV, PHEV y E-REV) (ANFAC, 2025).

Sin embargo, la Comisión Europea ha fijado unos objetivos en el paquete denominado Fit for 55 y España, aunque está realizando avances importantes en los últimos tiempos, aún debe realizar avances mucho más significativos con el fin de alcanzar los mismos. En cuanto a estos avances podemos recalcar que, al cierre de 2023, la infraestructura de recarga estaba compuesta por 29.301 puntos, y durante el año pasado se instalaron 9.424 puntos más de recarga. Como ya se ha dicho, se trata de un crecimiento sustancial en términos de movilidad sostenible pero no suficientes a ojos de la UE (ANFAC, 2025).

Sin duda, este desarrollo a gran escala de nuevos puntos de recarga pública ha sido indispensable para apoyar el crecimiento ya que las estaciones de recarga rápida son esenciales para una transición eficiente hacia la electromovilidad. Ahora bien, a pesar de esta mejora y avances en esta área, el ritmo de instalación de estos puntos de recarga debe acelerarse si se pretende cumplir con los objetivos antes mencionados. Para ello, se establecen como necesarios el apoyo gubernamental y la implementación de políticas más agresivas en este contexto con el fin de fomentar tanto la demanda de vehículos eléctricos, como la expansión de accesible y rápida infraestructura de recarga en todo el país (ANFAC, 2025).

1.2.2. Rol de los concesionarios en la cadena de valor

En el ámbito del sector automotriz, los costos externos e internos del distribuidor serán los que configuren el precio final de un vehículo. Estos pueden variar mucho dependiendo de las políticas del fabricante y las dinámicas del mercado. Los concesionarios comienzan adquiriendo sus vehículos a los fabricantes a un precio que se conoce como Precio Franco de Fábrica. Este precio suele tener ciertos descuentos respecto al precio real que establece el fabricante (Ramos Penabad, 2010).

Los concesionarios, como es lógico, venden el coche a un precio mayor al de su compra y generalmente, su margen de beneficio fluctúa entre 8% y 12%. Este porcentaje va a depender principalmente del volumen de ventas y de las promociones específicas que hayan pactado con el fabricante. En cuanto a los gastos de transporte del vehículo hasta el punto de venta desde la fábrica, y otros costes logísticos, estos son invariables y se trasladan directamente al comprador (Botin, 2012)

A este precio pagado por el concesionario, se deben añadir los impuestos aplicables y otros cargos adicionales. Ejemplo de estos últimos son los servicios administrativos, que pueden variar de un concesionario a otro, y los gastos de matriculación, entre otros. La suma de todos estos y el precio pagado por el concesionario conforman la base imponible sobre la que se calcula el IVA del vehículo y todos aquellos otros impuestos que sean de aplicación. De esta última resultará el precio final de venta a los consumidores (Ramos Penabad, 2010).

En este contexto, cabe referirse de nuevo a la transformación significativa que está teniendo lugar en el sector automotriz hacia la electrificación. Esta está siendo impulsada también por cambios legislativos, lo que está llevando a los fabricantes a realizar grandes inversiones en este tipo de vehículos, que influyen en las estrategias de precios y distribución de todo el sector, en general (Botin, 2012).

Además, también se está llevando a cabo una evolución de los modelos de distribución. Son bastantes los fabricantes que a día de hoy optan por modelos de agencia o venta directa y ofrecen por tanto experiencias de compra que pueden estar más alineadas con las expectativas modernas de los consumidores. Esto se debe a la simplicidad, eficiencia y transparencia en el proceso de compra, tanto en entornos online como offline, que ofrecen este tipo de modelos (Alfonso Peña, 2022).

Este cambio de paradigma está, sin duda, redefiniendo la relación entre fabricantes, distribuidores y consumidores. Como consecuencia, es muy importante que los concesionarios se adapten a estas nuevas realidades para poder mantener su competitividad en un mercado que, como ya hemos mencionado, no para de evolucionar. Además, esta dinámica también afecta cómo los concesionarios gestionan sus inventarios y definen sus estrategias de precios, en un esfuerzo por equilibrar la necesidad de atraer a los consumidores con la de asegurar una operación rentable y sostenible (Alfonso Peña, 2022).

Gracias a este modelo se permite a los concesionarios comprobar si el precio de fábrica que les ofrecen por un vehículo se ajusta realmente al valor que debería tener según sus características. De este modo, no solo les ayuda a predecir precios de mercado, sino también a tomar decisiones más informadas al evaluar si el precio de partida es coherente o está por encima o por debajo de lo que sería razonable.

1.3. Metodología empleada

El enfoque metodológico va destinado al desarrollo del algoritmo que consiga ser eficaz y responder a los objetivos de este trabajo, que es entender qué características influyen más en el precio de un coche para así conseguir crear una herramienta para concesionarios y vendedores. Para ello, se ha decidido utilizar Rstudio y se pretende crear un modelo de regresión lineal múltiple.

1.3.1. Modelos de regresión lineal múltiple

La regresión es una técnica estadística que se utiliza para llevar a cabo el análisis de la relación entre distintas variables. En función del número de variables implicadas, o la forma en la que estas se relacionan entre sí, podemos diferenciar varios tipos. El tipo de regresión por excelencia es la regresión lineal, ya que se trata de la forma más básica e informativa de implementar esta técnica. La regresión lineal parte de la idea de que existe una relación lineal, o que puede transformarse en lineal, entre las variables. Dentro de esta categoría, encontramos la regresión lineal simple, que analiza la relación entre solo dos variables, y la regresión lineal múltiple, que amplía el análisis incorporando varias variables independientes que pueden influir de forma conjunta sobre una variable dependiente. Esta última es especialmente útil cuando se estudian fenómenos complejos en los que intervienen múltiples factores (Montero Granados, 2016).

En un modelo de regresión lineal múltiple, la variable que queremos predecir o analizar se denomina habitualmente variable dependiente, aunque también puede recibir otros nombres como endógena, explicada o variable respuesta. Por otro lado, las variables que se utilizan para explicar o predecir el comportamiento de la variable dependiente se llaman variables independientes, exógenas, explicativas o regresores. Aunque este tipo de modelos no permiten afirmar con certeza la existencia de relaciones causales, ya que la correlación no implica causalidad, sí que establecen una dirección clara de análisis: se parte de las variables independientes (X) para intentar comprender o estimar el valor de la variable dependiente (Y). Normalmente, en este tipo de modelos se trabaja con una sola variable dependiente y varias independientes, ya que incluir más de una variable dependiente complica considerablemente el modelo y su interpretación (Montero Granados, 2016).

Esta relación se expresa a través de una fórmula como la que aparece en la imagen inferior, donde se representa que el valor de la variable dependiente (representada por la "Y") se calcula como una combinación lineal de todas las variables independientes, representadas por las betas, más un término de error que será el que recoge el error del modelo. Por lo tanto, en este modelo se asume que cada variable explicativa tiene una influencia sobre el resultado final y trata de explicar cuál es el peso de cada variable exógena en esa combinación. Así, se obtiene una visión más completa de cómo distintas características contribuyen conjuntamente a determinar el comportamiento de una variable de interés (Wooldridge, 2012).

$$y_j = b_o + b_1 x_{1j} + b_2 x_{2j} + \cdots + b_k x_{kj} + u_j$$

Ilustración 1: fórmula correspondiente a los modelos de regresión lineal múltiple

Fuente: elaboración propia

En este momento, es relevante concretar que existen distintos tipos de modelos dentro de la regresión lineal múltiple. Por un lado tenemos el modelo explicativo, que se centra en comprender las relaciones entre variables, utilizando todos los datos disponibles para estimar el modelo y contrastar hipótesis. En este enfoque, se da prioridad a la interpretación de los coeficientes y a la medición del impacto que cada variable independiente tiene sobre la variable dependiente (MSMK, 2024).

Por el contrario, el enfoque predictivo, que es el adoptado en este trabajo, se orienta a construir modelos capaces de estimar con precisión el valor de la variable de interés cuando se aplican a nuevos datos. Para ello, es necesario dividir el conjunto de datos original en dos subconjuntos: uno de entrenamiento, con el que se ajusta el modelo, y otro de test, con el que se evalúa su capacidad de generalización (MSMK, 2024).

Mientras que en los modelos explicativos la evaluación suele basarse en medidas como el coeficiente de determinación (R^2) obtenido en el mismo conjunto de entrenamiento, en los modelos predictivos el rendimiento se mide mediante el error cometido en el conjunto de test. (Wooldridge, 2012). Para ello, se emplean métricas como el MAE (Mean Absolute Error), que calcula el promedio de las diferencias absolutas entre los valores reales y los predichos (MSMK, 2024), y el RMSE (Root Mean Squared Error), que mide la raíz cuadrada del promedio de los errores al cuadrado, penalizando en mayor medida aquellos errores más grandes (IBM, 2024). Esta estrategia permite validar si el modelo es realmente útil para realizar predicciones en contextos reales, que es precisamente el objetivo principal.

Tras esta explicación, resulta sencillo comprender porque se ha seleccionado el modelo de regresión lineal múltiple predictivo como el adecuado para la investigación que se está llevando a cabo. Gracias al mismo conseguiremos entender y estimar cómo las distintas variables de nuestra base de datos, influyen a la variable endógena, el precio y evaluaremos su capacidad de predicción.

1.3.2. Fases del modelo de regresión lineal múltiple empleado

A continuación, vamos a fijar las distintas fases necesarias para desarrollar un modelo de regresión lineal múltiple para así dejar clara la metodología del modelo.

1. **Preparación y Limpieza de Datos**: empezaremos con una fase crucial para el correcto funcionamiento del modelo estadístico. Esta tiene como fin asegurar que los datos estén en el formato adecuado para poder tratarlos. Entre sus acciones principales se encuentran, por ejemplo, la conversión de variables categóricas y numéricas, el manejo de valores perdidos, la eliminación de outliers y duplicados y el análisis de la multicolinealidad. Esta etapa sentará las bases para que todos los análisis subsiguientes sean sobre datos precisos y representativos (Faster Capital, 2019).
2. **Entrenamiento:** en esta fase se identificarán y seleccionarán las variables independientes que formarán parte del modelo y se establecerá la relación funcional

que estas mantienen con la variable dependiente (el precio del coche), definiendo así la estructura matemática del modelo de regresión. El entrenamiento se realizará utilizando una partición del conjunto de datos original, lo que permitirá al modelo ajustar sus parámetros y aprender los patrones subyacentes presentes en los datos (Faster Capital, 2019).

3. **Test**: en esta fase comenzaremos a ver ya los resultados de nuestro modelo. Se evaluará el mismo contra una parte del conjunto de datos que no haya sido utilizada durante el entrenamiento con el fin de conseguir así probar su efectividad y precisión. Se verifican indicadores clave como el R-cuadrado, el F-statístico y los p-valores para asegurar que el modelo es robusto y válido para hacer predicciones o inferencias (Faster Capital, 2019).
4. **Análisis del Modelo**: por último, debemos analizar los resultados obtenidos y ver si es posible que estos se correspondan con la realidad. Esta interpretación incluirá la evaluación de la importancia y el impacto de cada variable dentro del modelo y cómo estas podrán ser utilizadas para tomar decisiones informadas en el sector automotriz (Faster Capital, 2019).

2. Descripción de la base de datos

La base de datos seleccionada para el desarrollo de esta investigación, denominada "car_price_prediction_.csv", es una herramienta de carácter fundamental, ya que será la principal fuente de información.

La elección de esta base de datos se fundamenta en su riqueza en datos y su indudable relación con el objetivo perseguido. La base de datos cuenta con 10 variables, que reflejan las características de distintos vehículos que se presume que tienen un impacto relevante en el precio final del mismo. Con esta base de datos conseguimos extraer una muestra de representación del mercado actual en relación con vehículos de combustión tradicional y vehículos eléctricos.

2.1 Análisis de las variables

A continuación, las clasificaremos en cuantitativas o cualitativas, daremos una breve explicación de cada una y haremos hincapié en por qué consideramos que pueden ser relevantes para nuestra investigación.

1. **ID del coche:** esta variable sirve para identificar de forma única cada vehículo dentro del conjunto de datos, diferenciándolo del resto aunque compartan las mismas características. Es una variable cuantitativa ordinal.
2. **Año del modelo**: el año en el que se fabricó el coche es una variable categórica y nominal. Los vehículos más recientes se espera que tengan un precio más alto que los antiguos, debido principalmente a la depreciación del estado de los mismos. Va desde el año 2000 hasta el 2023.
3. **Marca del vehículo**: como variable categórica, identifica el fabricante del vehículo y es una variable nominal. La marca del vehículo es un factor clave por motivos de fiabilidad, popularidad, prestigio, calidad, lujo y los factores específicos que pueda tener asociados una marca en concreto. Entre las marcas incluidas en la base de datos encontramos: Tesla, BMW, Audi, Ford, Honda, Mercedes y Toyota, que representan tanto marcas de lujo como estándar. Se espera que marcas como Tesla o Mercedes puedan hacer que el precio del coche sea más elevado.
4. **Tipo de combustible**: de nuevo nos encontramos ante una variable categórica y nominal ya que clasifica los vehículos según el tipo de combustible que deben utilizar. De nuevo esta clasificación tendrá una influencia directa en el precio del vehículo debido a las diferencias de costos de fabricación, operativos y ambientales. Las distintas categorías en este ámbito son: gasolina ('Petrol'), eléctrico ('Electric'), diésel ('Diesel') e híbrido ('Hybrid'). Estas son de mucha relevancia, principalmente de cara a resolver el objetivo secundario de esta investigación, que está relacionado con los vehículos eléctricos.
5. **Kilometraje**: el kilometraje, una variable cuantitativa y continua. Los vehículos que presentan una menor carga de kilómetros en principio tienden a ser más caros pues esto va a implicar que presentan un menor desgaste. Varía desde 15 km hasta 299.967 km.
6. **Transmisión**: esta variable categórica y nominal indica si el coche es de conducción automática o manual. Un coche con transmisión automática puede ser más costoso que el mismo coche, pero con transmisión manual.
7. **Condición**: La variable de condición se refiere al estado en el que se encuentra el vehículo cuando se realiza la venta. Es una variable categórica y nominal que diferencia tres categorías: Nuevo ('New'), usado ('Used') y como nuevo ('Like New').
8. **Desplazamiento del motor:** es una variable que se mide en litros y, por lo tanto, es cuantitativa y continua. Gracias a ella conocemos el volumen de todos los cilindros

del motor del vehículo combinados. Este volumen tendrá efecto directo en el precio del coche pues es un indicador de su potencia y rendimiento. En nuestra base de datos encontramos volúmenes de entre 1 y 6 litros.

9. **Modelo de coche:** dentro de cada marca de coche existen distintos modelos que tendrán influencia en el precio del vehículo por presentar distintas características. Es una variable categórica y nominal.
10. **Precio**: La variable objetivo, cuantitativa y continua, que queremos predecir. El objetivo es analizar su relación con el resto de las variables con el fin de conocer mejor qué factores son más influyentes en la determinación del mismo. Es importante recalcar que el precio está calculado en dólares y va desde $5,011.27 a $99,982.59.

2.2 Estadísticos descriptivos

Empezaremos este apartado con el análisis descriptivo de las variables numéricas de nuestra base de datos. Con el mismo conseguiremos obtener características básicas de estas variables mediante la utilización de varias medidas estadísticas fundamentales. Además, gracias a estas medidas podremos tener una visión más clara de la distribución de los datos, e incluso identificar posibles valores atípicos. De esta forma, conseguiremos entender más adecuadamente las características fundamentales de las variables involucradas en nuestra investigación.

En esta tabla vemos los resultados:

	Desplazamiento del motor	Kilometraje	Precio
MEDIA	3,47	149.750	52.638
MEDIANA	3,4	149.085	53.485
DESVIACIÓN TÍPICA	1,43	87.920	27.296
MÁXIMO	6	299.967	99.982,59
MÍNIMO	1	15	5.011,27

Fuente: elaboración propia

Al analizar los resultados obtenidos de este primer análisis descriptivo de las variables numéricas, los patrones obtenidos son bastante interesantes. Por ejemplo, en cuanto a la variable de desplazamiento del motor, podemos observar que los valores de la media y la mediana son muy cercanos. Esto nos indica que la distribución de esta variable es bastante equilibrada, lo cual también se ve en el valor su desviación típica que indica variaciones moderadas alrededor de la media. Sin embargo, dada la naturaleza de la variable y sus valores máximos y mínimos entendemos que esta desviación típica baja también se debe a que el rango de valores de esta característica no puede ser mucho más amplio.

En cuanto al kilometraje, mientras que la media y la mediana son muy cercanas, en este caso la desviación típica es muy superior. Esto se debe a que, a diferencia de la variable anterior, en este caso el rango de los valores de la variable es mucho más amplio. De estos valores podemos sacar conclusiones como que la base de datos incluye tanto vehículos nuevos como vehículos muy usados. Prueba de esto último son también sus valores máximos y mínimos.

En cuanto al análisis de la variable objetivo, la más importante, nos revela también una variabilidad considerable. Una vez más la media y la mediana muestran valores muy parecidos, lo que nos indica que, aunque con el resto de los resultados observamos que hay coches con precios muy diversos en nuestra base de datos, realmente la gran mayoría se agrupa en un rango medio. Esto es un indicador de una distribución no sesgada del precio.

Gracias a este análisis conseguimos un mejor entendimiento de las variables y valores de nuestra base de datos y podemos seguir con nuestra investigación teniendo un conocimiento más profundo de la misma de cara a enfocar mejor el trabajo. Volveremos a referirnos a las mismas más adelante cuando sea conveniente.

3. Desarrollo del modelo de regresión

Como ya se ha indicado previamente en el apartado de metodología, para desarrollar el modelo algorítmico se utilizará la herramienta Rstudio. La finalidad de este modelo de regresión lineal múltiple, dividido en las fases ya especificadas y que se desarrollarán a continuación, será conseguir un mejor entendimiento de la relación e influencia existente de cada una de las variables de la base de datos, con nuestra variable objetivo, el precio. Para ello debemos desarrollar los pasos convenientes.

3.1. Limpieza y preparación de Datos

3.1.1. Limpieza de datos

Una vez seleccionada la base de datos "car_price_prediction_.csv" considerada adecuada para nuestra investigación, el siguiente paso será la limpieza y preprocesamiento de los datos. El preprocesamiento y la limpieza de los datos tienen como finalidad asegurar que el modelo que estamos creando sea lo más preciso y eficiente posible.

Para empezar con esta fase del proceso, la primera decisión que se ha tomado ha sido la de eliminar la columna *Car ID*. Esto se debe a que, al analizar las diversas columnas de la base de datos, entendemos que esta en concreto no es útil para el análisis pues no aporta ningún tipo de información que pueda tener influencia directa en el precio del vehículo. Además una sobrecarga de variables puede entorpecer el funcionamiento del modelo.

El siguiente paso de esta fase será eliminar aquellos valores que estén duplicados. Es posible que la base de datos contenga información sobre un mismo modelo de coche más de una vez y esta información duplicada puede tener consecuencias negativas a la hora de llevar a cabo nuestro estudio. Al hacer esto hemos descubierto que no existen valores duplicados en nuestra base de datos por lo que, una vez tenemos esto claro podemos seguir con nuestra investigación.

```r
# Contamos las filas originales antes de eliminar duplicados
original_rows <- nrow(datos)

# Eliminamos duplicados
datos <- datos %>% distinct()

# Contamos las filas después de eliminar duplicados
distinct_rows <- nrow(datos)

# Calculamos el número de filas duplicadas eliminadas
removed_rows <- original_rows - distinct_rows

# Imprimir el número de duplicados eliminados
print(paste("Número de registros duplicados eliminados:", removed_rows))
```

Ilustración 2: código de R correspondiente a la eliminación de duplicados

Fuente: elaboración propia

A continuación, antes de construir el modelo, es necesario llevar a cabo la eliminación de aquellas filas que puedan tener valores en blanco o faltantes que puedan perjudicar al funcionamiento del modelo. Los modelos de regresión no pueden funcionar correctamente si

hay datos que faltan, y trabajar con información incompleta puede hacer que los resultados sean poco fiables.

Después de haber desarrollado el código adecuado para identificar aquellas filas con valores faltantes y eliminarlas, hemos identificado que el número de filas restantes en nuestra base de datos es 2.500, qué es exactamente el mismo que antes de implementar esta función. Como consecuencia entendemos que no existían filas con variables faltantes y podemos continuar con la investigación.

```
# Eliminamos filas con valores faltantes
datos <- na.omit(datos)

# Imprimimos el nuevo número de filas para verificar
print(paste("Número de filas después de eliminar valores faltantes:", nrow(datos)))
```

Ilustración 3: código de R correspondiente a la comprobación de valores faltantes

Fuente: elaboración propia

Por otro lado, en este momento de preparación de los datos suele ser interesante llevar a cabo la identificación de posibles outliers o datos que puedan no ser representativos con el objetivo de evitar el sesgo de los resultados. Sin embargo, en este caso consideramos que no debemos eliminar los outliers ya que, teniendo en cuenta que el objetivo del trabajo, la existencia de outliers puede ser inherente y significativa debido a la existencia de una diversidad natural en el mercado automotriz. Lo visualizamos en el gráfico de abajo:

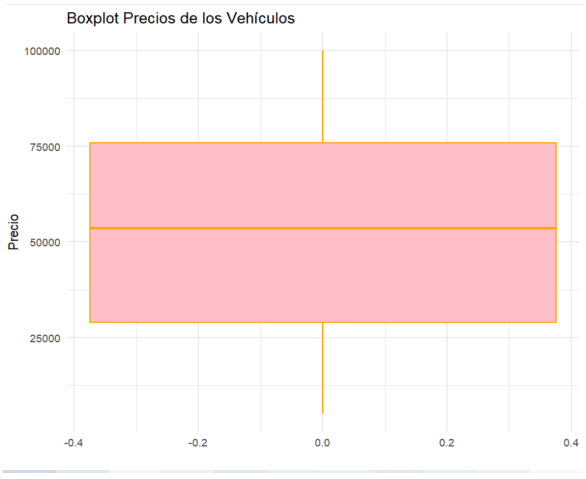

Ilustración 4: gráfico destinado a la identificación de outliers

Fuente: elaboración propia

Podemos observar que el precio medio de los vehículos está en alrededor de 50.000 euros y que se consideran *outliers* aquellos con un precio superior a 76.000 e inferior a 24.000, aproximadamente.

Los *outliers* pueden indicar precios extremadamente altos que podrían ser vehículos de lujo o modelos con características especiales o, por el contrario, tratarse de vehículos muy básicos, antiguos o usados. Por lo tanto, consideramos que son información relevante y no los vamos a eliminar.

3.1.3. Transformación de variables

Comenzamos ahora con una subfase un poco más compleja dentro de esta primera fase de limpieza y preprocesamiento de los datos. Nos referimos a la necesidad existente de transformar las variables categóricas en factores. Esta transformación es necesaria ya que estas variables van a tener que ser relacionadas con una variable numérica, el precio, para poder conseguir el objetivo establecido. Por lo tanto, es imprescindible transformarlas para que el modelo pueda entenderlas. Lo que se consigue con este proceso es asignar a cada variable categórica un valor numérico y de esta forma se consigue que el modelo pueda establecer comparaciones y asignar un coeficiente específico a cada una.

En resumen, al tratar las variables como factores el modelo puede saber el efecto que cada una de ella tiene sobre el precio del vehículo mediante su comparación con una categoría de referencia, que suele ser seleccionada automáticamente por el software estadístico. Esta transformación no solo simplifica el análisis estadístico, sino que también enriquece los resultados, proporcionando conocimientos claros sobre cómo cada característica categórica afecta el precio del vehículo. De este modo, podemos entender mejor las preferencias del mercado y cómo ciertas características como la marca o el tipo de combustible influyen en la valoración de un coche.

```
tibble [2,500 × 11] (S3: tbl_df/tbl/data.frame)
 $ Car ID      : num [1:2500] 1 2 3 4 5 6 7 8 9 10 ...
 $ Brand       : Factor w/ 7 levels "Audi","BMW","Ford",..: 6 2 1 6 3 1 1 6 4 3 ...
 $ Year        : Factor w/ 24 levels "2000","2001",..: 17 19 14 12 10 20 21 18 24 11 ...
 $ Engine Size : num [1:2500] 2.3 4.4 4.5 4.1 2.6 2.4 4 5.3 5.7 1.5 ...
 $ Fuel Type   : Factor w/ 4 levels "Diesel","Electric",..: 4 2 2 1 1 1 2 3 2 2 ...
 $ Transmission: Factor w/ 2 levels "Automatic","Manual": 2 2 2 1 2 1 1 1 2 1 ...
 $ Mileage     : num [1:2500] 114832 143190 181601 68682 223009 ...
 $ Condition   : Factor w/ 3 levels "Like New","New",..: 2 3 2 2 1 1 3 2 1 1 ...
 $ Price       : num [1:2500] 26614 14680 44403 86374 73577 ...
 $ Model       : Factor w/ 28 levels "3 Series","5 Series",..: 20 2 4 21 22 25 24 21 8 12 ...
 $ FuelType    : Factor w/ 4 levels "Diesel","Electric",..: 4 2 2 1 1 1 2 3 2 2 ...
```

Ilustración 5: imagen que demuestra la conversión de las variables categóricas en factor
Fuente: elaboración propia

En la imagen superior observamos que tras haber implementado el código adecuado, efectivamente todas las variables categóricas han sido convertidas en factores.

Como podemos ver, cada fila tiene al final una serie de valores numéricos. Por ejemplo, a la variable “Brand” se le asignan valores del 1 al 7 en función de las distintas marcas existentes facilitando el manejo y análisis de estos datos como categorías. Así, en este caso las marcas de coches como "Audi", "BMW", "Ford", etc., se representan, por ejemplo, con números como 1, 2 o 3.

Se han convertido las variables categóricas en factores porque, al hacerlo así, el programa R las entiende directamente como categorías distintas y se encarga él solo de tratarlas correctamente en el modelo. Esto hace que el proceso sea más sencillo y ayuda a interpretar mejor los resultados sin necesidad de transformar manualmente cada categoría en una nueva columna. Además, se evitan errores comunes que pueden aparecer cuando se crean estas variables de forma manual (Grolemund, 2014).

Una vez hecho lo oportuno con las variables categóricas, las variables numéricas también tienen que ser preparadas para desarrollar el modelo de forma más eficiente. En este caso, el proceso consiste en ajustar las escalas de las variables numéricas (precio, tamaño del motor y kilometraje) para conseguir que estas tengan un peso justo y comparable de cara a análisis posteriores. Este proceso es el que se conoce como normalización de variables numéricas. En resumen, de lo que se trata es de que todas las variables se ajusten a una escala estándar para poder realizar una comparación directa entre las variables y poder así entender cual tiene una mayor influencia en el precio del coche. Así, eliminamos distorsiones que

puedan surgir a raíz de simples diferencias en la forma de medir o en la magnitud de los números.

En la imagen inferior podemos ver el resultado de llevar a cabo esta normalización de las variables numéricas.

```
# A tibble: 6 × 11
  `Car ID` Brand Year  `Engine Size`[,1] `Fuel Type` Transmission Mileage[,1] Condition Price[,1] Model  FuelType
     <dbl> <fct> <fct>              <dbl> <fct>       <fct>              <dbl> <fct>         <dbl> <fct>  <fct>
1        1 Tesla 2016              -0.814 Petrol      Manual           -0.397  New          -0.953 Model… Petrol
2        2 BMW   2018               0.653 Electric    Manual           -0.0746 Used         -1.39  5 Ser… Electric
3        3 Audi  2013               0.723 Electric    Manual            0.362  New          -0.302 A4     Electric
4        4 Tesla 2011               0.443 Diesel      Automatic        -0.922  New           1.24  Model… Diesel
5        5 Ford  2009              -0.604 Diesel      Manual            0.833  Like New      0.767 Musta… Diesel
6        6 Audi  2019              -0.744 Diesel      Automatic         1.10   Like New      1.33  Q7     Diesel
> |
```

Ilustración 6: imagen que demuestra la normalización de las variables numéricas

Fuente: elaboración propia

Como se puede ver en la imagen, la normalización de datos ajusta la media de los datos a cero y estandariza la desviación estándar a uno. Observamos en la tabla que los valores se encuentran entre -1, 0 y 1. Estos valores indican cuantos se desvían los valores originales del promedio. Gracias a este ajuste podemos comparar de forma equitativa cuánta influencia tiene cada variable sobre la variable dependiente, en este caso el precio, asegurándonos de que ninguna variable domine el modelo debido a la magnitud de su escala. (UC3M, s.f.)

A modo de ejemplo, si consideramos la columna "Engine Size" (tamaño del motor), podemos observar que los valores transformados van desde -0.814 hasta 0.723. De esto deducimos que en 2016 que el tamaño del motor del Tesla concreto está muy por debajo de la media del conjunto de datos. Al contrario, el Audi de 2013 muestra un valor de 0.723, esto nos indica que el tamaño de su motor está significativamente por encima de la media.

En conclusión, al estandarizar los valores, nos aseguramos de que el modelo de regresión trate las variables en pie de igualdad y elimina sesgos que podrían llegar a surgir de diferencias en las unidades de medida.

3.1.3. Revisión de la multicolinealidad

Cuando las variables independientes están muy relacionadas entre sí en un modelo de regresión lineal, surge lo que se conoce como la multicolinealidad. Se trata de un problema que nace cuando existen variables que, en lugar de aportar información diferente y útil, nos están proporcionando prácticamente la misma información. La consecuencia principal de este

problema es la dificultad que surge a la hora de identificar cual es la variable que está influyendo realmente en el resultado y, por ende, sucede que los coeficientes de la regresión se convierten en inestables o poco fiables. Es decir, aunque el modelo pueda seguir funcionando, las conclusiones que se saquen sobre la importancia de cada variable pueden ser erróneas (Amat Rodrigo, 2018).

Para evitar este problema y poder detectarlo a tiempo se utiliza la función *VIF(),* que se corresponde con el análisis de los valores del factor de inflación de la varianza. Se considera que cuando existe un VIF superior a 5, nos encontramos ante variables que están muy correlacionadas entre sí y que pueden estar generando multicolinealidad (Amat Rodrigo, 2018).

Para conseguir sobrepasar este problema, existen varias estrategias distintas. La más común sería eliminar una de las variables que está causando la multicolinealidad, pues si no aporta información relevante consideramos que no es necesaria en nuestro modelo. Otra opción más compleja sería combinar las variables que están provocando el problema en una sola. Veremos más adelante que en el desarrollo de nuestro modelo se ha optado por la primera opción (Ferrero, 2022).

En este ámbito, durante la construcción de nuestro modelo de regresión, se ha identificado un problema de colinealidad en el modelo ya que al implementar la función VIF obtuvimos el siguiente aviso:

```
Error in vif.default(modelo) :
  there are aliased coefficients in the model
```

Ilustración 7: error de multicolinealidad

Fuente: elaboración propia

El mismo refleja un problema de colinealidad perfecta o casi perfecta entre algunas de nuestras variables y por tanto, algunos coeficientes no podrían calcularse de forma concreta. Por lo tanto, el siguiente paso ha sido identificar cuáles eran estas variables y tomar las medidas adecuadas para seguir adelante con el modelo. Para conseguir esto utilizamos la función *alias()* que nos muestra aquellas variables que están entorpeciendo el modelo. Gracias a la misma se ha descubierto que se trata de las variables “Marca” y “Modelo”.

Para evitar la multicolinealidad y poder desarrollar nuestro modelo eliminanos una de ellas. La variable “Modelo” presenta un número muy elevado de categorías distintas, muchas de ellas con muy pocos registros, lo que puede introducir ruido, sobreajuste y pérdida de generalización. Sin embargo, consideramos que la variable “Marca” puede ser una mejor opción ya que nos ofrece una representación de la muestra más agregada y fácil de manejar a la vez que puede permitir captar las diferencias existentes entre fabricantes sin comprometer la estabilidad del modelo. Como consecuencia, se ha optado por conservar la variable “Marca” como predictora y sacamos la variable “Modelo” de nuestro modelo garantizando así un equilibrio adecuado entre detalle, interpretabilidad y robustez estadística.

Tras esta corrección hemos obtenido los siguientes resultados:

	GVIF	Df	GVIF^(1/(2*Df))
Year	1.137332	23	1.002801
`Engine Size`	1.011890	1	1.005927
Mileage	1.011146	1	1.005558
`Fuel Type`	1.041676	3	1.006828
Transmission	1.012174	1	1.006069
Condition	1.032337	2	1.007988
Brand	1.074406	6	1.005999

Ilustración 8: resultados del VIF

Fuente: elaboración propia

Una vez hemos obtenido los resultados de implementar la función VIF, nos damos cuenta que al eliminar la variable “Modelo” de nuestra función lm(), ha desaparecido el problema significativo de multicolinealidad.

En concreto, todos los valores de la columna GVIF^(1/(2*Df)) se sitúan en torno a 1, esto es indicador de que cada variable aporta información individual al modelo y que no está significativamente relacionada con el resto de variables. Como consecuencia de estos resultados podemos garantizar que los coeficientes obtenidos son estables y fiables y, por tanto, podremos identificar la influencia de cada variable sobre el precio de forma individualizada. De este modo, se refuerza la validez del modelo de regresión múltiple utilizado, al cumplir con una de sus condiciones fundamentales: la independencia entre predictores (Amat Rodrigo, 2018).

Por otro lado, hemos considerado importante la realización de una matriz de correlación de las variables numéricas del modelo como análisis complementario al VIF. La

matriz de correlaciones es una herramienta muy útil para visualizar de forma rápida e intuitiva la intensidad y el sentido de las relaciones lineales entre pares de variables. A pesar de que el VIF se puede considerar más adecuado para detectar multicolinealidad en el contexto de un modelo de regresión múltiple, la matriz de correlaciones nos ayuda a detectar posibles redundancias o relaciones muy estrechas entre las distintas variables que se van a utilizar como predictores incluso antes de construir el modelo. Gracias a la misma se podría llegar a considerar necesario eliminar alguna variable más con el objetivo de asegurar una mayor robustez del modelo. Vemos en la gráfica de abajo la matriz de correlaciones correspondiente a nuestro modelo.

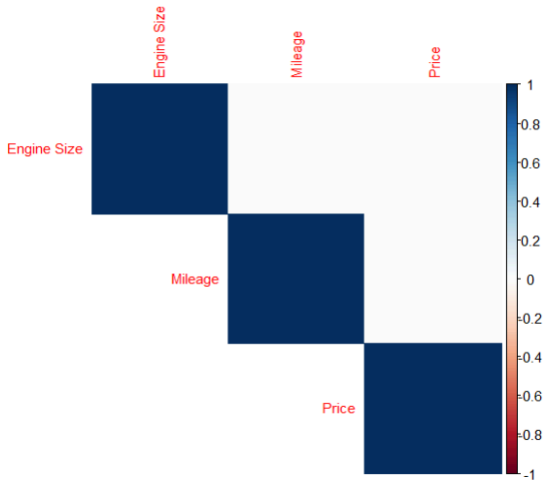

Ilustración 9: matriz de correlaciones

Fuente: elaboración propia

Esta matriz permite observar el grado de relación lineal entre las variables numéricas del conjunto de datos. En nuestro caso, analizamos el "Desplazamiento del motor", el "Kilometraje" y el "Precio" de los vehículos. Si bien es cierto que en este caso no tiene mucho sentido incluir la variable price en esta matriz, pues la multicolinealidad realmente examina la relación entre las variables independientes, hemos considerado oportuno incluirla con el objetivo de ir obteniendo información previa al modelo que pueda ser útil.

Como se puede observar, la matriz no muestra correlaciones relevantes entre ninguna de las tres variables. Esto es señal de que no existe redundancia entre las variables independientes, lo cual favorece nuestro modelo pues muestra que no existe una relación lineal significativa entre ellas.

En cuanto a la relación de la variable “Precio” con las dos variables independientes, se aprecia una ligera relación negativa entre el "Kilometraje" y el "Precio", lo cual es lógico: los coches con mayor kilometraje tienden a tener un valor de mercado más bajo. Por otro lado, el "Desplazamiento del motor" presenta una correlación muy baja con el "Precio", lo que sugiere que su impacto puede depender de otros factores combinados.

En conclusión, la matriz de correlaciones ha sido una herramienta util que confirma que, tras haber eliminado las variables oportunas, no hay multicolinealidad preocupante antes de aplicar la regresión (Ferrero, 2019).

3.2. Selección de las variables relevantes

Una vez completada la fase de limpieza y transformación de los datos, es relevante dejar claro cuales son las variables que se han seleccionado como independientes para proceder con el desarrollo del modelo. Como ya se ha mencionado anteriormente, se han eliminado del modelo las variables "Identificador del coche" y "Modelo", ya que se ha considerado que no aportan valor predictivo relevante, la primera por ser simplemente un identificador único de cada vehículo, y la segunda para evitar problemas de colinealidad al estar estrechamente relacionada con la variable "Marca", que sí se ha conservado.

Por lo tanto, las variables seleccionadas que si hemos conservado son finalmente: "Año", "Desplazamiento del motor", "Kilometraje", "Tipo de combustible", "Transmisión", "Condición" y "Marca". Se ha considerado que todas estas variables, en principio, tienen una justificación teórica para influir en el precio y, además, se ha verificado que no presentan problemas de multicolinealidad significativos. Como consecuencia, el razonamiento lógico nos lleva a pensar que podrían ser relevantes a la hora de determinar el precio de un coche y por lo tanto deberían cumplir con las expectativas de nuestro objetivo.

Con todo lo mencionado hasta ahora, es posible construir un modelo fiable y útil que permita cumplir el objetivo principal de este trabajo: entender qué factores influyen más en la fijación del precio de los vehículos. A continuación, procedemos a desarrollar el modelo de regresión lineal múltiple. Para ello, será necesario dividir el modelo en dos particiones, el set de entrenamiento y el set de test.

3.3. División del conjunto de datos: entrenamiento y test

Con el objetivo de construir un modelo de regresión lineal múltiple robusto y capaz de generalizar bien sobre datos nuevos, se ha dividido el conjunto de datos en dos subconjuntos: uno de entrenamiento y otro de test. Esta separación busca evitar el sobreajuste, que tendría lugar si el modelo se adaptará en exceso a los datos con los que se ha entrenado. Como consecuencia del sobreajuste, el modelo perdería precisión a la hora de predecir situaciones reales (IBM, 2021). En este caso, se ha asignado el 80% de los datos al conjunto de entrenamiento, que será utilizado para ajustar el modelo, y el 20% restante al conjunto de test, que será utilizado para evaluar el rendimiento del modelo. La división se ha realizado de forma aleatoria para asegurar que ambas muestras sean representativas de la base de datos original. Esta metodología es comúnmente utilizada en modelos de aprendizaje supervisado, ya que la misma nos permite estimar de forma más realista la capacidad predictiva del modelo al utilizar observaciones que no ha visto durante su entrenamiento.

Una vez dividido el conjunto de datos, se procede al entrenamiento del modelo de regresión lineal múltiple utilizando únicamente el conjunto de entrenamiento, como se puede ver en la imagen superior. Este proceso tiene como finalidad conseguir que el modelo aprenda las relaciones que existen, en caso de que las haya, entre las variables independientes seleccionadas (que como sabemos son año del vehículo, marca, tipo de combustible, etc.) y la variable dependiente, que en este caso es el precio del coche. Como resultado del modelo obtendremos coeficientes que describan estas relaciones y que permitan predecir con precisión el precio en nuevos datos. En las imagen de debajo podemos ver tanto la fórmula como los coeficientes obtenidos.

```r
# Establecemos semilla para reproducibilidad
set.seed(478)

# Número de filas total
n <- nrow(datos)

# Crear índices aleatorios para el conjunto de entrenamiento (80%)
indices_entrenamiento <- sample(1:n, size = 0.8 * n)

# Crear los conjuntos
datos_entrenamiento <- datos[indices_entrenamiento, ]
datos_validacion <- datos[-indices_entrenamiento, ]
modelo <- lm(Price ~ + `Engine Size` + Mileage + `Fuel Type` + Transmission + Condition + Brand , data = datos_entrenamiento)
print(modelo)
summary(modelo)
```

```
lm(formula = Price ~ +`Engine Size` + Mileage + `Fuel Type` +
    Transmission + Condition + Brand, data = datos_entrenamiento)

Residuals:
     Min       1Q   Median       3Q      Max
-1.90056 -0.86063  0.01898  0.85253  1.87967

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          0.05945    0.08028   0.741   0.4590
`Engine Size`       -0.02805    0.02260  -1.242   0.2145
Mileage             -0.01320    0.02246  -0.588   0.5569
`Fuel Type`Electric -0.12378    0.06349  -1.950   0.0514 .
`Fuel Type`Hybrid   -0.10070    0.06319  -1.594   0.1112
`Fuel Type`Petrol   -0.13606    0.06334  -2.148   0.0318 *
TransmissionManual   0.01895    0.04496   0.422   0.6734
ConditionNew        -0.05914    0.05505  -1.074   0.2828
ConditionUsed       -0.05860    0.05496  -1.066   0.2864
BrandBMW             0.10372    0.08159   1.271   0.2038
BrandFord           -0.02263    0.08482  -0.267   0.7896
BrandHonda           0.06293    0.08219   0.766   0.4440
BrandMercedes        0.07473    0.08315   0.899   0.3689
BrandTesla           0.08665    0.08314   1.042   0.2974
BrandToyota          0.12312    0.08247   1.493   0.1356
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.002 on 1985 degrees of freedom
Multiple R-squared:  0.00729,   Adjusted R-squared:  0.0002889
F-statistic: 1.041 on 14 and 1985 DF,  p-value: 0.4083
```

Ilustración 11: coeficientes del modelo

Fuente: elaboración propia

Al implementar la función *lm()* a nuestro 80% seleccionando de los datos, la columna *estimate* de la imagen nos muestra los coeficientes estimados (que se corresponden con las betas de la fórmula vista al principio de la investigación) que nos permiten interpretar cómo se relaciona cada una de estas variables con el precio de los vehículos, asumiendo que todas las demás permanecen constantes (Pizarro, 2020).

Los coeficientes calculados para cada variable indican la dirección y magnitud del efecto que tiene cada variable sobre el precio. A modo de ejemplo, podemos observar que el coeficiente para “Fuel Type Electric” es de -0,12378, esto sugiere que si lo comparamos con la categoría de referencia (probablemente “Diesel”), entendemos que los coches eléctricos tienden a tener un precio medio más bajo, teniendo en cuenta que mantenemos el resto de factores constantes. Además, cabe resaltar que los datos estadísticos relacionados con esta variable específica son especialmente importantes como consecuencia de su valor p-valor.

El p-valor de una variable nos indica la probabilidad de que el efecto observado sea debido al azar. En términos prácticos, un p-valor bajo (normalmente inferior a 0,05) sugiere que la variable tiene un efecto estadísticamente significativo sobre el precio, mientras que un valor elevado puede indicar que la relación no es lo suficientemente robusta como para ser considerada fiable (Grolemund, 2014). Por lo tanto, teniendo en cuenta que en este caso el

p-valor es 0,0514, muy cercano al umbral clásico, podemos considerar que esta diferencia es estadísticamente significativa.

Lo mismo ocurre con “Fuel Type Petrol”, cuyo coeficiente es de -0,13606 y su p-valor es de 0,0318, lo que permite afirmar con un grado de confianza razonable que esta variable sí tiene un efecto relevante en el precio. En ambos casos, la interpretación es coherente con las dinámicas actuales del mercado, donde los coches de gasolina y eléctricos pueden tener precios más bajos que otras alternativas por razones de antigüedad o incentivos.

Por el contrario, la realidad es que si seguimos explorando los resultados, nos damos cuenta de que no han sido los esperados y mucho menos los deseados. Podemos observar que la mayoría de los coeficientes asociados a otras variables no presentan significación estadística, es decir, sus p-valores son superiores a 0,05. Esto es reflejo de que, a pesar de que en un principio hemos considerado que debían ser incluidas en el modelo por su relevancia teórica, no se ha encontrado una relación clara y consistente entre esas variables y el precio en los datos utilizados. Esto puede deberse a múltiples factores, como la variabilidad dentro de cada categoría o el tamaño limitado de la muestra.

A nivel general, el modelo presenta un R^2 ajustado de 0,0002889, lo que significa que es capaz de explicar una proporción muy baja de la variabilidad total del precio. Este resultado refleja que, a pesar de que algunas variables como el tipo de combustible parecen tener un cierto efecto, la realidad es que el modelo en su conjunto no logra capturar de forma precisa los patrones que determinan el precio de los vehículos en esta base de datos y, por lo tanto, parece que no cumple al nivel que esperábamos con el objetivo fijado (Amat Rodrigo, 2016).

Por otro lado, con el objetivo de obtener más información y poder complementar el análisis del modelo, se han calculado las métricas de error sobre el conjunto de entrenamiento. En concreto, el error absoluto medio (MAE) obtenido en el conjunto de entrenamiento ha sido de 0,866 aproximadamente, mientras que el error cuadrático medio (RMSE) se ha situado en torno a 0,998. Estos valores reflejan la desviación media del modelo respecto a los valores reales, expresadas en unidades normalizadas. A priori, estos resultados son indicadores de que el modelo no está funcionando de la manera esperada, más adelante los analizaremos en detalle. y los compararemos con los errores del conjunto de test para poder valorar si el modelo está sobreajustado.

Antes de pasar a la siguiente fase, es importante destacar, que este modelo se ha construido con fines exploratorios y educativos, y que parte del proceso consiste precisamente en identificar las limitaciones y entender por qué un modelo no ofrece los resultados esperados. Por tanto, este análisis sigue siendo útil y permite extraer conclusiones sobre la relación entre variables, aunque el ajuste global del modelo no haya sido óptimo. Además, el hecho de que algunas variables sí resulten significativas nos da una pista clara sobre qué factores deberían explorarse más a fondo. En definitiva, aunque los resultados no son tan buenos como se esperaba inicialmente, el análisis ha servido para poner en valor el tipo de variables que sí podrían influir en el precio, y sentar las bases para desarrollar modelos más potentes y precisos.

A continuación, se procede a validar el modelo utilizando el conjunto de datos de test (20% de los datos no utilizados en el entrenamiento), con el objetivo de evaluar su capacidad predictiva sobre nuevos datos y obtener métricas como el error medio absoluto (MAE), el error cuadrático medio (RMSE) y el coeficiente de determinación (R^2) aplicado sobre datos no vistos.

3.4. Implementación del modelo predictivo

Una vez entrenado el modelo de regresión lineal múltiple, se ha procedido a comprobar su capacidad predictora utilizando el 20% de los datos previamente reservados como conjunto de test. Una buena capacidad predictora sería indispensable de cara a ayudar a los concesionarios a tomar sus decisiones, que como ya sabemos es el objetivo principal de esta investigación.

```
# Predicciones sobre el conjunto de validación
predicciones <- predict(modelo, newdata = datos_validacion)

# Cálculo de los errores de predicción
mae <- mean(abs(predicciones - datos_validacion$Price))
rmse <- sqrt(mean((predicciones - datos_validacion$Price)^2))
```

Ilustración 12: comprobación de la capacidad predictiva del modelo

Fuente: elaboración propia

Este paso es esencial para comprobar si el modelo, más allá de ajustarse a los datos de entrenamiento, es capaz de generalizar correctamente a nuevos casos y ofrecer predicciones fiables sobre observaciones no vistas anteriormente.

Para evaluar el rendimiento del modelo, se han calculado dos métricas clave:

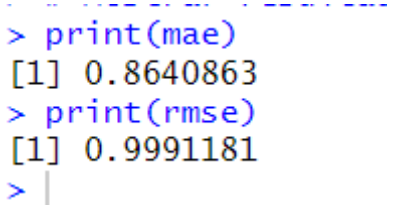

Ilustración 13: resultados de la capacidad predictora

Fuente: elaboración propia

Ambas métricas, al igual que en el caso anterior, están expresadas en los términos normalizados, ya que en la etapa de preparación de los datos hemos llevado a cabo la normalización de los mismos con fines de eficiencia. Con esto nos referimos a que los resultados no están en unidades monetarias (dólares) directamente, sino que se refieren a desviaciones estándar respecto a la variable dependiente normalizada (el precio del coche). A modo orientativo, si tenemos en cuenta que la desviación estándar del precio original era de aproximadamente 27.295 dólares (podemos verlo en la tabla de arriba), estos errores equivalen a una desviación media de unos 23.600 dólares en el MAE y 27.270 dólares en el RMSE, que representa el error máximo que puede cometer. Estos resultados se calculan multiplicando los valores obtenidos (0,8640863 y 0,9991181) expresados en escala normalizada por el valor de la desviación típica inicial.

Aunque estas cifras pueden parecer elevadas, es importante ponerlas en contexto. Por un lado, el mercado del automóvil presenta una elevada variabilidad de precios, incluso dentro de una misma marca, por lo que en ocasiones un error de 25.000 dólares en la predicción no tendría porqué considerarse escandaloso. Aún así, haber obtenido estos valores tan altos de RMSE y MAE es un claro indicador de que el modelo no está entendiendo bien el funcionamiento de la imposición de los precios y, en la gran mayoría de casos, podría suponer errores muy graves en las predicciones.

Debemos tener en cuenta que el modelo ha sido construido únicamente a partir de características visibles y disponibles en la base de datos (sin incluir otros elementos clave como equipamiento, año de matriculación exacto, historial de uso o mantenimiento, entre otros), lo que entendemos que puede estar limitando su capacidad para entender y capturar la complejidad que realmente influye en el precio final.

Aun así, cabe destacar que el modelo demuestra un comportamiento estable en la predicción, con un rendimiento en test muy similar al observado en entrenamiento, ya que los valores de MAE y RMSE en ambas fases son muy similares. Como consecuencia, podemos considerar que el modelo no está sobreajustado y que mantiene cierta coherencia al aplicarse a datos nuevos, a pesar de que su capacidad predictora sea muy baja (IBM, 2021).

En resumen, a pesar de que los resultados de precisión predictiva no alcanzan los niveles deseables para las aplicaciones previstas, se puede considerar que este modelo ha permitido establecer una primera base interpretativa sólida. Se ha conseguido identificar qué variables podrían llegar a tener una influencia mayor en el precio (como el tipo de combustible), se ha verificado la robustez del modelo con datos externos y se ha demostrado que la regresión lineal múltiple, aun con limitaciones, puede ofrecer un punto de partida para estudios más avanzados o modelos complementarios.

3.4.1. Cross-Validation

Tras comprobar que los resultados obtenidos al dividir los datos en un conjunto de entrenamiento y otro de test no alcanzaban el nivel de precisión deseado, se ha optado por tratar de mejorar los resultado mediante la implementación de un método más robusto y extendido en el ámbito del aprendizaje supervisado: la validación cruzada (cross-validation).

Hasta ahora, el modelo se había construido mediante la división de los datos en dos conjuntos, el conjunto de entrenamiento y el conjunto destinado a testear el modelo y evaluar su rendimiento objetivamente, calculando el error sobre observaciones no utilizadas durante el ajuste. Este enfoque, si bien puede ser útil para estimar la capacidad de generalización del modelo, pues es aleatorio y evita sesgos, puede ser sensible a cómo se distribuyen los datos en una única partición. Por eso, se ha considerado interesante enfocar el objetivo de otra manera e intentar mejorar la estabilidad de la evaluación, recurriendo a la técnica del k-fold cross-validation. (Bagnato, 2020)

Este método de cross-validation consiste en dividir el conjunto de datos en k subconjuntos, utilizando en cada iteración uno de ellos como conjunto de validación y los restantes para el entrenamiento. La consecuencia del valor de k es el número de veces que se repite el proceso, en este caso serán 5. Cada una de estas 5 repeticiones es utilizada tanto para entrenar como para validar el modelo, y el rendimiento final se calcula promediando los errores obtenidos en cada fold. Con esta técnica se busca, no solo eliminar el riesgo de

overfitting, sino también conseguir una estimación más fiable en relación con el error generalizado (DataScientest, 2018).

Al implementar este modelo, los resultados obtenidos han sido coherentes con los valores que habían sido calculados previamente sobre el conjunto test, lo cual podemos considerar que aporta una mayor confianza sobre la estabilidad y fiabilidad del modelo. Esto se debe a que la coincidencia de estos valores es representativa de que el modelo no está aprendiendo patrones específicos, como ya habíamos comprobado en la comparación de resultados entre el conjunto de entrenamiento y el conjunto de test, sino que es capaz de generalizar razonablemente bien a nuevos datos, evitando así el problema del sobreajuste. Podemos ver los resultados en la imagen inferior:

```
Linear Regression

2500 samples
   7 predictor

No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 2000, 2000, 2000, 2000, 2000
Resampling results:

  RMSE      Rsquared      MAE
  1.008073  0.0002897405  0.872646
```

Ilustración 14: resultados de cross-validation

Fuente: elaboración propia

La similitud entre los errores de test y los de validación cruzada también indica que la partición inicial de los datos fue representativa del conjunto completo y no introduce sesgos, lo que valida la metodología empleada. Por último, la ausencia de grandes diferencias entre errores sugiere que el conjunto de test no incluía datos atípicos o significativamente distintos del conjunto de entrenamiento, reforzando así la conclusión de que el modelo, pese a sus limitaciones, ha llevado a cabo un control adecuado del error y que, probablemente, sus limitaciones se deben a una falta de información relevante en las variables disponibles. Aun así, la aplicación de esta técnica ha resultado valiosa desde un punto de vista metodológico, aportando rigor al proceso de evaluación y sentando las bases para futuros modelos más robustos y completos (DataScientest, 2018).

3.5. Resultados obtenidos y análisis interpretativo

Como ya hemos mencionado y deducimos de apartados anteriores, los resultados que hemos obtenido tras aplicar el modelo de regresión múltiple a nuestra base de datos reflejan una importante limitación en su capacidad predictiva. A pesar de haber llevado a cabo un procedimiento riguroso en la preparación y división de los datos, y de haber ajustado correctamente el modelo siguiendo las prácticas básicas del aprendizaje supervisado, los resultados obtenidos muestran que el modelo no ha sido capaz de estimar con precisión el precio de los vehículos y confirma que la variabilidad en el precio apenas ha podido ser explicada por las variables seleccionadas.

Todo apunta a que el modelo desarrollado presenta indicios evidentes de underfitting, un fenómeno que se produce cuando el modelo resulta ser demasiado simple para captar adecuadamente las relaciones existentes entre las variables del conjunto de datos. En este caso, la causa más probable radica en la falta de información suficiente: las variables disponibles, aunque relevantes, no abarcan toda la complejidad que influye en la formación del precio de un vehículo. Esto sugiere que sería necesario incorporar un mayor número de características que reflejen mejor las particularidades del mercado automovilístico para lograr una predicción más precisa (Mora Caballero, 2023).

A pesar de todo esto, consideramos relevante destacar como resultado positivo que el modelo presenta un comportamiento estable al aplicarse a datos nuevos. Como hemos visto, los errores asociados al conjunto de entrenamiento, de test y cross- validation son muy similares, lo que indica que el modelo no ha caído en sobreajuste. Esto es muestra de que aunque no hemos conseguido que el modelo prediga con la exactitud adecuada, al menos es consistente en sus errores y generaliza de forma equilibrada. Es fundamental entender que un modelo puede tener un rendimiento predictivo pobre y aun así no estar sobreentrenado. Esto señala que el problema no está tanto en cómo aprende el modelo, sino en lo que tiene disponible para aprender.

Consideramos que es relevante llevar a cabo un análisis de cuáles son las posibles causas que pueden estar causando este bajo rendimiento de un modelo que, a priori, consideramos que debería haber proporcionando mejores resultados. En este sentido, está claro que las variables han sido insuficientes por ser poco representativas de los factores que realmente terminan por influir en el precio de un coche. Entendemos que la falta de

información sobre atributos como el color del coche, el tipo de carrocería, el estado del interior del vehículo o incluso la disponibilidad de ciertos extras, entre muchas otras características que no están presentes en nuestra base de datos, serían claves para explicar el porqué de los precios.

Además, también debe considerarse que factores intangibles como pueden ser el prestigio de la marca, su atractivo visual o la percepción del consumidor sobre el mismo, suelen tener un papel importante en la decisión de compra y, por tanto, en el precio de mercado, pero son extremadamente difíciles de cuantificar y modelizar. Estos aspectos, aunque no se incluyan directamente, explican por qué dos vehículos con especificaciones técnicas similares pueden tener precios radicalmente distintos y por ende dificultan el funcionamiento de nuestro modelo.

A pesar de estos resultados, el proceso ha resultado valioso. Ha permitido aplicar de manera realista las fases de preparación, modelado y evaluación propias del aprendizaje supervisado. Además, ha sido útil para detectar las limitaciones del enfoque lineal en contextos complejos como el mercado automovilístico. En este sentido, los resultados obtenidos, aunque alejados del ideal, no suponen un fracaso sino una oportunidad para comprender mejor el problema y orientar futuras investigaciones hacia modelos más flexibles o conjuntos de datos más ricos.

4. Conclusiones

Esta investigación tenía como objetivo inicial el desarrollar una herramienta, mediante un modelo de regresión lineal múltiple, que pudieran utilizar los concesionarios como apoyo en la estimación de precios de vehículos, basándose en características objetivas como el kilometraje del vehículo o su marca, y así favorecer la transformación digital del sector.

Sin embargo, los resultados han sido que, si bien el modelo ha permitido identificar ciertas variables que, efectivamente, tienen influencia en el precio, como el tipo de combustible, y ha demostrado un comportamiento relativamente estable, sin indicios claros de sobreajuste, su capacidad predictiva ha resultado ser limitada, con errores medios de predicción demasiado elevados como para ser utilizados con plena confianza en la práctica diaria de un concesionario.

Esta falta de precisión es indicadora de que, a pesar de que el modelo puede ayudar a captar tendencias generales, no resulta ser suficiente para estimar con fiabilidad el precio concreto de un coche individual. Entre las posibles causas se encuentra la ausencia de variables que, aunque son más difíciles de cuantificar, influyen decisivamente en el precio y son aspectos que escapan al alcance de un modelo puramente basado en características técnicas.

En este sentido, entendemos que el modelo creado no es lo suficientemente preciso como para servir de herramienta final que sustituya la experiencia y el conocimiento del mercado que poseen los profesionales del sector, y que por lo tanto aún no está preparada para ser implementada. A pesar de ello, el modelo puede entenderse como una primera aproximación, útil para identificar patrones, detectar valores atípicos o complementar el juicio profesional.

Este trabajo ha permitido explorar en profundidad el proceso de diseño, entrenamiento y validación de un modelo predictivo, y ha servido para evidenciar los retos reales a los que se enfrenta la digitalización del sector automovilístico. Para convertir este tipo de herramientas en recursos verdaderamente útiles para concesionarios, será necesario avanzar en la calidad y el tipo de datos empleados, así como en la exploración de modelos más complejos y adaptativos.

Declaración de Uso de Herramientas de Inteligencia Artificial Generativa en Trabajos Fin de Grado

ADVERTENCIA: Desde la Universidad consideramos que ChatGPT u otras herramientas similares son herramientas muy útiles en la vida académica, aunque su uso queda siempre bajo la responsabilidad del alumno, puesto que las respuestas que proporciona pueden no ser veraces. En este sentido, NO está permitido su uso en la elaboración del Trabajo fin de Grado para generar código porque estas herramientas no son fiables en esa tarea. Aunque el código funcione, no hay garantías de que metodológicamente sea correcto, y es altamente probable que no lo sea.

Por la presente, yo, Paula Díaz Shaw, estudiante de E3- Analytics de la Universidad Pontificia Comillas al presentar mi Trabajo Fin de Grado titulado "Modelos predictivos en el contexto de la transformación digital del sector automovilístico: una aproximación mediante regresión lineal múltiple", declaro que he utilizado la herramienta de Inteligencia Artificial Generativa ChatGPT u otras similares de IAG de código sólo en el contexto de las actividades descritas a continuación [el alumno debe mantener solo aquellas en las que se ha usado ChatGPT o similares y borrar el resto. Si no se ha usado ninguna, borrar todas y escribir “no he usado ninguna”]

1. Metodólogo: Para descubrir métodos aplicables a problemas específicos de investigación.
2. Interpretador de código: Para realizar análisis de datos preliminares.
3. Corrector de estilo literario y de lenguaje: Para mejorar la calidad lingüística y estilística del texto.
4. Generador previo de diagramas de flujo y contenido: Para esbozar diagramas iniciales.
5. Sintetizador y divulgador de libros complicados: Para resumir y comprender literatura compleja.
6. Revisor: Para recibir sugerencias sobre cómo mejorar y perfeccionar el trabajo con diferentes niveles de exigencia.

Afirmo que toda la información y contenido presentados en este trabajo son producto de mi investigación y esfuerzo individual, excepto donde se ha indicado lo contrario y se han dado

los créditos correspondientes (he incluido las referencias adecuadas en el TFG y he explicitado para que se ha usado ChatGPT u otras herramientas similares). Soy consciente de las implicaciones académicas y éticas de presentar un trabajo no original y acepto las consecuencias de cualquier violación a esta declaración.

Fecha: 10 de abril de 2025
Firma: Paula Díaz Shaw

Referencias bibliográficas

- Wall, M. (2025). March 2025 Light Vehicle Production Forecast. S&P Global. https://www.spglobal.com/automotive-insights/en/blogs/2025/01/2025-light-vehicle-production-forecast
- Díaz, S. (2024). El sector de automoción en España: el reto de mantener la competitividad. *CaixaBank Research.* *https://www.caixabankresearch.com/es/analisis-sectorial/observatorio-sectorial/sector-automocion-espana-reto-mantener-competitividad*
- Heras, A. E. (s.f.). Automoción. *Investin Spain*. https://www.investinspain.org/es/sectores/automocion-movilidad
- ANFAC. (2025). Barómetro de la Electromovilidad. https://anfac.com/categorias_publicaciones/barometro-electro-movilidad/
- Alfonso Peña, F. J. (2022). ¿Concesionarios o agencias?. *Revista CESVIMAP.* https://www.revistacesvimap.com/concesionarios-o-agencias/
- Botin, R. (2012). ¿Cómo se configura el precio de un coche nuevo?. *Wanderer75.* https://w-75.com/2012/11/29/tip-como-se-configura-el-precio-de-un-coche-nuevo/
- Ramos Penabad, L. Conoce todos los términos de la factura de tu coche nuevo. *Cochescom.* https://www.coches.com/noticias/consejos/todos-los-terminos-de-la-factura-de-tu-coche-nuevo/32679
- Montero Granados, R. (2016): Modelos de regresión lineal múltiple. Documentos de Trabajo en Economía Aplicada. *Universidad de Granada*. España. https://www.ugr.es/~montero/matematicas/regresion_lineal.pdf
- Wooldridge, J. M. (2012). *Introductory Econometrics: A Modern Approach.* https://cbpbu.ac.in/userfiles/file/2020/STUDY_MAT/ECO/2.pdf
- MSMK. (2024). Mean Absolute Error. https://msmk.university/mean-absolute-error/
- Subirats Maté, L., Oswaldo Pérez Trenard, D., Calvo González, M. (2019). Introducción a la limpieza y análisis de los datos. Universidad Abierta de Cataluña. https://openaccess.uoc.edu/bitstream/10609/148647/1/IntroduccionALaLimpiezaYAnalisisDeLosDatos.pdf

- IBM. (2024). Raíz del error cuadrático medio en las métricas de calidad de Watson OpenScale. https://dataplatform.cloud.ibm.com/docs/content/wsj/model/wos-quality-root-of-mean-sq-error.html?locale=es&context=cpdaas

- Amat Rodrigo, J. (2016). Introducción a la Regresión Lineal Múltiple. https://cienciadedatos.net/documentos/25_regresion_lineal_multiple

- Pizarro, R. (2020). Modelo de Regresión Lineal Multiple: Datos Esperanza de Vida. *Rpubs*. https://rpubs.com/rpizarro/615274

- Ferrero, R., (2022). Qué es la multicolinealidad y por qué es un problema. https://www.maximaformacion.es/blog-ciencia-datos/que-es-la-multicolinealidad-y-por-que-es-un-problema/
- UC3M. (s.f.). Análisis de regresión lineal: el procedimiento Regresión lineal. p.341. https://halweb.uc3m.es/esp/personal/personas/jmmarin/esp/guiaspss/18reglin.pdf
- Ferrero, R. (2019). Analisis de correlacion: Guía rápida en R. *Máxima formación* https://www.maximaformacion.es/blog-dat/analisis-de-correlacion-guia-rapida-en-r/
- Faster Capital. (2019). Preparación Y Limpieza De Datos Para El Análisis De Regresión. https://fastercapital.com/es/tema/preparaci%C3%B3n-y-limpieza-de-datos-para-el-an%C3%A1lisis-de-regresi%C3%B3n.html
- IBM. (2021). ¿Qué es el sobreajuste?. https://www.ibm.com/es-es/think/topics/overfitting
- https://www.aprendemachinelearning.com/sets-de-entrenamiento-test-validacion-cruzada/
- Grolemund, G. (2014) *Hands-On Programming with R*. https://web.itu.edu.tr/~tokerem/Hands-On_R.pdf
- Bagnato, J. I. (2020). Sets de Entrenamiento, Test y Validación. Aprende Machine Learning.

- DataScientest. (2018). Cross-Validation. Definición e importancia en Machine Learning. https://datascientest.com/es/cross-validation-definicion-e-importancia

- Mora Caballero, M. (2023). Parte 7. Regrepedia: Retos y Limitaciones de las Regresiones en el Análisis de Datos. *Linkedin.* *https://www.linkedin.com/pulse/parte-7-regrepedia-retos-y-limitaciones-de-las-en-el-mora-caballero-cz5de/*

ANEXO 1: CÓDIGO DE R UTILIZADO PARA EL DESARROLLO DE LA INVESTIGACIÓN

```
# libreria necesaria para poder manipular los datos de una base de datos

if(!require(readr)) install.packages("readr")

library(readr)

getwd()

setwd("C:/Users/paudi/Desktop/ICADE/tfg analytics")

datos<- read_csv("car_price_prediction_.csv")

head(datos)

# Cargamos más librerias que pueden ser necesarias

library(dplyr)

library(tidyr)

# 3.1.2. Limpieza de datos

# Contamos las filas originales antes de eliminar duplicados

original_rows <- nrow(datos)

# Eliminamos duplicados

datos <- datos %>% distinct()

# Contamos las filas después de eliminar duplicados

distinct_rows <- nrow(datos)

# Calculamos el número de filas duplicadas eliminadas

removed_rows <- original_rows - distinct_rows

# Imprimir el número de duplicados eliminados
```

```
print(paste("Número de registros duplicados eliminados:", removed_rows))

# Eliminamos filas con valores faltantes

datos <- na.omit(datos)

# Imprimimos el nuevo número de filas para verificar

print(paste("Número de filas después de eliminar valores faltantes:", nrow(datos)))

#aparatdo de anilisis descriptivo de las variables

if(!require(knitr)) install.packages("knitr")

library(knitr)  # Para formatear y visualizar tablas

kable(descriptive_stats, caption = "Estadísticas Descriptivas de las Variables Numéricas")

# Crear un dataframe con medias y medianas

calcmedi2<- datos %>%

  summarise(across(where(is.numeric), list(

    mean = ~mean(., na.rm = TRUE),

    median = ~median(., na.rm = TRUE)

  ))) %>%

  pivot_longer(everything(), names_to = c(".value", "stat"), names_pattern = "(.*)_(.*)")

print(calcmedi2)

# Calcular la desviación típica para cada columna numérica

desviacion<- datos %>%

  summarise(across(where(is.numeric), sd, na.rm = TRUE))

# Ver las desviaciones típicas
```

```
print(desviacion)

if (!require(modeest)) install.packages("modeest")

library(modeest)

# Calcular el máximo y el mínimo para cada columna numérica

extreme_values <- data %>%

  summarise(across(where(is.numeric), list(

    maximo = max,

    minimo = min

  ), .names = "{col}_{fn}"))  # Personaliza los nombres de las columnas para claridad

# Ver los valores extremos

print(extreme_values)

glimpse(extreme_values)

# 3.1.3. Transformación de variables

# Convertimos cada variable categórica en un factor

datos$Brand <- factor(datos$Brand)

datos$Year <- factor(datos$Year)

datos$`Fuel Type` <- factor(datos$`Fuel Type`)

datos$Transmission <- factor(datos$Transmission)

datos$Condition <- factor(datos$Condition)

datos$Model <- factor(datos$Model)

str(datos)
```

```
sd(datos$Price)

# Normalizando las variables numéricas en el dataframe 'datos'

datos$Price <- scale(datos$Price)

datos$`Engine Size` <- scale(datos$`Engine Size`)

datos$Mileage <- scale(datos$Mileage)

# Verificar las primeras filas para confirmar las transformaciones

head(datos)

# Calcular multicolinealidad

modelo <- lm(Price ~ Year + `Engine Size` + Mileage + `Fuel Type` + Transmission + Condition + Model + Brand, data = datos)

library(car)

alias(modelo)

alias(lm(Price ~ Year + `Engine Size` + Mileage + `Fuel Type` + Transmission + Condition + Brand, data = datos))

modelo2 <- lm(Price ~ Year + `Engine Size` + Mileage + `Fuel Type` + Transmission + Condition + Brand, data = datos)

vif(modelo2)

datos <- datos[, !names(datos) %in% "Car ID"]

# Cargar las librerías

library(dplyr)

library(corrplot)

# Seleccionar solo las variables numéricas del data frame

datos_numericos <- datos %>% select(where(is.numeric)
```

```
# Calculamos la matriz de correlaciones

matriz_correlaciones <- cor(datos_numericos, use = "complete.obs")

# Visualizar la matriz con corrplot

corrplot(matriz_correlaciones, method = "color", type = "upper", tl.cex = 0.8)

# 3.3. División del conjunto de datos: entrenamiento y validación

# Establecemos semilla para reproducibilidad

set.seed(478)

# Número de filas total

n <- nrow(datos)

# Crear índices aleatorios para el conjunto de entrenamiento (80%)

indices_entrenamiento <- sample(1:n, size = 0.8 * n)

# Crear los conjuntos

datos_entrenamiento <- datos[indices_entrenamiento, ]

datos_validacion <- datos[-indices_entrenamiento, ]

modelo <- lm(Price ~ + `Engine Size` + Mileage + `Fuel Type` + Transmission + Condition +
Brand , data = datos_entrenamiento)

print(modelo)

summary(modelo)

# Predicciones del modelo sobre el conjunto de entrenamiento

pred_entrenamiento <- predict(modelo, newdata = datos_entrenamiento)



# Cálculo del RMSE (Error cuadrático medio)
```

```
rmse_entrenamiento <- sqrt(mean((datos_entrenamiento$Price - pred_entrenamiento)^2))

# Cálculo del MAE (Error absoluto medio)

mae_entrenamiento <- mean(abs(datos_entrenamiento$Price - pred_entrenamiento))

# Mostrar resultados

print(paste("RMSE (entrenamiento):", rmse_entrenamiento))

print(paste("MAE (entrenamiento):", mae_entrenamiento))

# Predicciones sobre el conjunto de validación

predicciones <- predict(modelo, newdata = datos_validacion)

# Cálculo de los errores de predicción

mae <- mean(abs(predicciones - datos_validacion$Price))

rmse <- sqrt(mean((predicciones - datos_validacion$Price)^2))

# Mostrar resultados

print(mae)

print(rmse)

# para el cross validation

install.packages("caret")

library(caret)

# Establecer semilla para reproducibilidad

set.seed(123)

# Configurar validación cruzada con k = 5

control <- trainControl(method = "cv", number = 5)
```

```
# Entrenar el modelo con validación cruzada
modelo_cv <- train(
  Price ~ Year + `Engine Size` + Mileage + `Fuel Type` + Transmission + Condition + Brand,
  data = datos,
  method = "lm",
  trControl = control)
# Ver resultados
print(modelo_cv)
```