

---

**Modelado predictivo de precios de venta de automóviles usados para la compañía  
AutoScout24 utilizando R.**

Nelson Mauricio García López

Informe de práctica presentado para optar al título de Ingeniero Electrónico

Modalidad de Práctica  
Modalidad Piloto con materias de maestría

Asesor  
Aníbal José Guerra Soler, Especialista (Esp) en Ingeniería Electrónica y de Computación

Universidad de Antioquia  
Facultad de Ingeniería  
Ingeniería Electrónica  
Medellín, Antioquia, Colombia  
2026

---

| Cita               |     | García López [1]                                                                                                                                                                                                                                  |
|--------------------|-----|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>Referencia</b>  | [1] | N. García López, “Modelado predictivo de precios de venta de automóviles usados para la compañía AutoScout24 utilizando R.”, Trabajo de grado profesional, Ingeniería Electrónica, Universidad de Antioquia, Medellín, Antioquia, Colombia, 2026. |
| Estilo IEEE (2020) |     |                                                                                                                                                                                                                                                   |

---



Centro de Documentación Ingeniería (CENDOI)

**Repositorio Institucional:** <http://bibliotecadigital.udea.edu.co>

Universidad de Antioquia - [www.udea.edu.co](http://www.udea.edu.co)

El contenido de esta obra corresponde al derecho de expresión de los autores y no compromete el pensamiento institucional de la Universidad de Antioquia ni desata su responsabilidad frente a terceros. Los autores asumen la responsabilidad por los derechos de autor y conexos.

## TABLA DE CONTENIDO

|                                                 |    |
|-------------------------------------------------|----|
| RESUMEN.....                                    | 7  |
| ABSTRACT .....                                  | 8  |
| I. INTRODUCCIÓN .....                           | 9  |
| II. OBJETIVOS .....                             | 11 |
| A. Objetivo general .....                       | 11 |
| B. Objetivos específicos .....                  | 11 |
| III. MARCO TEÓRICO .....                        | 12 |
| A. Cálculo de precios en vehículos usados ..... | 12 |
| B. Aprendizaje automático.....                  | 13 |
| 1) Aprendizaje supervisado: .....               | 14 |
| 2) Aprendizaje no supervisado .....             | 15 |
| IV. METODOLOGÍA .....                           | 16 |
| A. Fuente y selección de los datos .....        | 17 |
| B. Análisis descriptivo y multivariado.....     | 18 |
| 1) Análisis exploratorio de datos (EDA) .....   | 18 |
| C. Análisis no supervisado.....                 | 19 |
| 1) PCA .....                                    | 20 |
| 2) FA.....                                      | 20 |
| 3) Clúster .....                                | 20 |
| D. Criterios de limpieza y transformación ..... | 20 |
| 1) Modelo lineal generalizado .....             | 20 |
| 2) Transformación logarítmica: .....            | 21 |
| 3) Análisis Discriminante Cuadrático QDA: ..... | 21 |

---

|                                                                  |    |
|------------------------------------------------------------------|----|
| E. Evaluación de los modelos candidatos y selección final: ..... | 22 |
| F. Indicadores estadísticos e hipótesis de modelado .....        | 22 |
| V. ANÁLISIS DE RESULTADOS .....                                  | 23 |
| A. Presentación de los resultados .....                          | 23 |
| 1) Modelo lineal generalizado MLG .....                          | 23 |
| 2) Análisis discrimine cuadrático QDA.....                       | 24 |
| B. Análisis de los resultados .....                              | 26 |
| 1) Análisis modelo lineal generalizado .....                     | 26 |
| 2) Análisis QDA .....                                            | 27 |
| VI. CONCLUSIONES Y RECOMENDACIONES.....                          | 28 |
| A. CONCLUSIONES .....                                            | 28 |
| B. RECOMENDACIONES .....                                         | 29 |
| REFERENCIAS .....                                                | 30 |
| ANEXOS.....                                                      | 31 |

---

## LISTA DE TABLAS

|                                                      |    |
|------------------------------------------------------|----|
| TABLA I VARIABLES BASE DE DATOS .....                | 17 |
| TABLA II ANALÍSIS EXPLORATORIO DE DATOS .....        | 18 |
| TABLA III VARIABLES SECCIONADAS PARA EL MODELO ..... | 23 |
| TABLA IV MATRIZ DE CONFUSIÓN DEL MODELO QDA .....    | 25 |

---

## LISTA DE FIGURAS

|                                                                   |    |
|-------------------------------------------------------------------|----|
| Fig. 1 Esquema del proceso metodológico.....                      | 16 |
| Fig. 2 Matriz de correlaciones entre las variables continuas..... | 19 |
| Fig. 3 Predicciones del modelo lineal generalizado. ....          | 24 |

---

## RESUMEN

Este trabajo presenta el desarrollo de un modelo predictivo para la estimación del precio de venta de automóviles usados publicados en la plataforma AutoScout24, utilizando el lenguaje estadístico R. El objetivo principal consiste en construir modelos estadísticos y de aprendizaje automático que permitan estimar el precio de los vehículos a partir de sus características técnicas y de uso. La metodología adoptada se basa en un enfoque cuantitativo que incluye procesos de limpieza y transformación de datos, análisis exploratorio, reducción de dimensionalidad mediante técnicas no supervisadas y la implementación de modelos supervisados como el Modelo Lineal Generalizado y el Análisis Discriminante Cuadrático. Los resultados evidencian que variables como el kilometraje, la potencia del motor, el peso y la antigüedad del vehículo tienen un impacto significativo en el precio de venta. El modelo de regresión alcanzó un error relativo aproximado del 20 %, mientras que el modelo QDA obtuvo una exactitud del 83 % en la clasificación de rangos de precios. Se concluye que los modelos propuestos permiten realizar estimaciones objetivas y confiables, aportando herramientas útiles para el análisis del precio de mercado de vehículos usados.

***Palabras clave* — Precio, modelo, clasificación, regresión, discriminante.**

---

## ABSTRACT

This study presents the development of a predictive model to estimate the selling price of used cars published on the AutoScout24 platform using the R statistical environment. The main objective is to build statistical and machine learning models capable of predicting vehicle prices based on technical and usage characteristics. The methodology follows a quantitative approach that includes data cleaning and transformation, exploratory data analysis, dimensionality reduction through unsupervised techniques, and the implementation of supervised models such as the Generalized Linear Model and Quadratic Discriminant Analysis. The results show that variables such as mileage, engine power, vehicle weight, and age significantly influence the selling price. The regression model achieved an approximate relative error of 20 %, while the QDA model reached an accuracy of 83 % in classifying price ranges. It is concluded that the proposed models provide objective and reliable price estimates, contributing valuable tools for used car market analysis.

***Keywords* — Price, Model, Classification, Regression, Discriminant.**

---

## I. INTRODUCCIÓN

El mercado de automóviles usados constituye uno de los sectores más dinámicos dentro de la economía global, impulsado por factores como la rápida depreciación de los vehículos nuevos, el crecimiento del comercio electrónico y la evolución de plataformas digitales especializadas en compraventa de autos [1]. Entre estas plataformas destaca **AutoScout24**, reconocida en Europa por facilitar la conexión entre vendedores y compradores mediante anuncios en línea con información detallada de cada vehículo.

No obstante, este mercado presenta alta variabilidad en los precios, debido a la influencia de múltiples factores como lo son la marca, el modelo, entre muchos otros. La diferencia entre estas características de los autos genera asimetría de información entre compradores y vendedores, lo que, se evidencia en la falencia de la determinación apropiada de los precios de los autos; por ende, muchos anuncios de venta de autos presentan sobrevaloraciones o subvaloraciones significativas, reduciendo la eficiencia del mercado y afectando la satisfacción del usuario final.

En este contexto, las herramientas basadas en modelos predictivos y de aprendizaje automático surgen como alternativas valiosas para estimar precios de forma objetiva y precisa, permitiendo ofrecer recomendaciones automáticas y estandarizadas. El desarrollo de un modelo basado en técnicas de análisis de datos, como son PCA, clúster o análisis factorial, permitiría la precisión del cálculo, lo que no solo mejoraría la transparencia y la confianza en la plataforma, sino que también contribuye a la optimización de procesos internos y a la fidelización de los usuarios.

Diversos estudios han abordado la estimación del precio de vehículos usados mediante técnicas estadísticas y de aprendizaje automático. En particular, trabajos recientes han empleado modelos de regresión lineal múltiple, en los cuales el precio se explica como una combinación ponderada de atributos físicos y técnicos del vehículo [2].

Asimismo, se ha evidenciado que la ingeniería de características, que es la creación y transformación de variables derivadas como la edad del vehículo, el kilometraje marcado o combinaciones entre marca y modelo, puede mejorar significativamente la precisión de los modelos.

En resumen, los avances recientes en análisis de datos y aprendizaje automático han demostrado que la precisión de los modelos depende no solo de los algoritmos empleados, sino también de la calidad y transformación de las variables que los alimentan. Bajo esta perspectiva, el

---

presente trabajo propone un enfoque integral que combina análisis estadístico multivariado, técnicas de reducción de dimensionalidad y modelos predictivos supervisados y no supervisados, con el propósito de estimar el precio de venta de automóviles usados en AutoScout24 con mayor objetividad y confiabilidad.

---

## II. OBJETIVOS

### *A. Objetivo general*

Implementar un modelado predictivo de precios de venta de automóviles usados para la compañía AutoScout24 utilizando R.

### *B. Objetivos específicos*

- Depurar el conjunto de datos proveniente de la plataforma AutoScout24 mediante técnicas propias de la ciencia de datos, identificando valores atípicos, datos faltantes e inconsistencias, que permitan optimizar la representación de la información y mejorar el desempeño de los modelos predictivos de precios de automóviles usados.
- Aplicar técnicas de análisis descriptivo y multivariado para identificar las propiedades de los datos y posibles relaciones entre variables y reducir dimensionalidad.
- Implementar métodos de aprendizaje supervisado y no supervisado para la construcción de modelos predictivos de precios de automóviles usados.
- Evaluar el desempeño de los diferentes algoritmos aplicados, para identificar el modelo más preciso para la predicción del precio y redactar las conclusiones propias de la investigación.

---

### III. MARCO TEÓRICO

Para comenzar a determinar el comportamiento del precio de los autos usados ofertados en la plataforma AutoScout24, es necesario conocer como abordan los modelos otros autores y las definiciones de los modelos utilizados.

#### *A. Cálculo de precios en vehículos usados*

El cálculo del precio de un auto usado es un proceso difícil, debido a todas las variables que influyen en él. Entre las variables cuantitativas más importantes se encuentran el kilometraje recorrido, el cilindraje, la potencia del motor, el consumo de combustible y el peso del vehículo. Estas cualidades o características hacen que se evalúe el nivel de deterioro o desgaste, que posicionen el rendimiento o muestren el costo operativo, lo que infiere directamente en el valor de reventa [2].

De igual forma, variables cualitativas como la marca, el modelo, el año de fabricación, el tipo de combustible o la transmisión influyen en la toma de decisiones, afectando el precio final. Según el informe de mercado “Informe de análisis del tamaño, participación y tendencias del mercado de autos usados” elaborado por la firma Grand View Research [1], el mercado de vehículos usados presenta una alta variabilidad debido a la falta de estandarización en la valoración de automóviles, generando diferencias de información entre compradores y vendedores. Es por ello, que la implementación de modelos predictivos basados en estos escenarios contribuye a establecer precios más objetivos y confiables, reduciendo la incertidumbre del mercado y mejorando la eficiencia de plataformas como AutoScout24.

Algunos autores han abordado el tema de la predicción de precios de automóviles usados, desde enfoques estadísticos y de aprendizaje automático. En su estudio “Using Linear Regression for Used Car Price Prediction”, Muti y Yıldız [2] demostraron que la regresión lineal múltiple puede explicar una proporción significativa de la variabilidad del precio, siempre que se realice una adecuada depuración y selección de variables, y que se controle la multicolinealidad entre ellas.

Por otra parte, Kumar y Sinha [3] compararon distintas técnicas de regresión, incluyendo la regresión polinómica y la regresión Ridge (técnica de regularización estadística para modelos de aprendizaje automático que corrige el sobreajuste y la multicolinealidad), concluyendo que la

combinación de ingeniería de características y validación cruzada permite obtener mejores niveles de precisión.

Mientras que, Huang y otros autores [4], en “Analysis of Circular Price Prediction Strategy for Used Electric Vehicles”, propusieron un enfoque alternativo de predicción circular aplicado a vehículos eléctricos, integrando variables dinámicas de batería y eficiencia energética, destacando la importancia de la selección de variables y la optimización iterativa de los modelos.

De manera complementaria, el estudio “Machine Learning for Predicting Used Car Resale Prices Using Granular Vehicle Equipment Information” [5], utilizó técnicas avanzadas de aprendizaje automático como árboles de decisión, máquinas de soporte vectorial (SVM) y redes neuronales, demostrando que el nivel de granularidad en la información técnica y de equipamiento mejora notablemente la exactitud de las predicciones.

Estas investigaciones muestran una tendencia hacia la utilización de modelos híbridos y multivariados que combinan métodos estadísticos y algoritmos de machine learning. Una buena selección de las variables para el modelo, hacen que las herramientas de predicción sean más confiables y es de vital importancia conocer el grado de contribuciones de estas variables en el comportamiento del precio de los autos usados.

Según los planteamientos de estos autores como apertura de la investigación, se buscará ilustrar las definiciones formales de los modelos utilizados en la predicción del precio de automóviles.

### *B. Aprendizaje automático*

El aprendizaje automático se divide en dos grandes categorías: supervisado y no supervisado. En el aprendizaje supervisado, los modelos son entrenados con datos etiquetados (por ejemplo, el precio conocido del vehículo) para predecir resultados futuros; mientras que, en el aprendizaje no supervisado, como en el análisis de componentes principales (PCA), el análisis factorial (FA) o el análisis de clúster, se busca descubrir estructuras internas o patrones sin una variable objetivo [6].

### 1) *Aprendizaje supervisado:*

Es una rama del aprendizaje automático en la cual el modelo se entrena a partir de un conjunto de datos etiquetados, es decir, observaciones para las cuales se conoce previamente la variable respuesta o salida. Su objetivo principal es aprender una función que relacione las variables explicativas con la variable objetivo, de modo que el modelo pueda realizar predicciones precisas sobre nuevos datos no observados [7].

Entre las técnicas de aprendizaje supervisado más utilizadas para la predicción de precios tenemos:

La **regresión lineal** es una técnica estadística que modela la relación entre una variable dependiente y una sola variable independiente usando una línea recta. La regresión lineal múltiple es una extensión de la regresión simple que utiliza múltiples variables independientes para predecir una única variable dependiente, permitiendo un análisis más complejo de sus influencias conjuntas.

**Modelos Lineales Generalizados** unifica modelos lineales y no lineales, con variables respuesta (Y) normales y no normales. Es básicamente un modelo de regresión, donde se pueden utilizar variables categóricas y continuas:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p + \varepsilon \quad (1)$$

En un MLG la variable respuesta Y puede tener cualquier distribución que sea miembro de la familia de la distribución exponencial: Normal, Poisson, Binomial, Exponencial, Gamma.

Se validan los mismos supuestos de la regresión múltiple, con la diferencia de que en Homoscedasticidad se deben probar para las categóricas el análisis ANOVA y para las continuas el de Regresión.

Por otra parte, en los modelos de aprendizaje supervisado, se encuentran los siguientes modelos de clasificación:

El Análisis Discriminante Lineal **Discriminant Analysis (LDA)** es un método de clasificación supervisado de variables cualitativas en el que dos o más grupos son conocidos por su comportamiento y nuevas observaciones se clasifican en uno de ellos en función de sus características.

El clasificador cuadrático o **Quadratic Discriminat Analysis QDA** asume que cada clase sigue una distribución normal multivariante, pero a diferencia de LDA, permite que cada clase tenga su propia matriz de covarianza.

Además, de estos modelos utilizados, la base de datos fue analizada mediante técnicas de aprendizaje no supervisado, como son PCA, análisis factorial y análisis clúster, donde se evidencian relaciones importantes entre las variables utilizadas y que dan validez al modelo propuesto [8].

## 2) *Aprendizaje no supervisado:*

El aprendizaje no supervisado constituye un complemento analítico relevante dentro del estudio, ya que permite explorar la estructura interna de los datos sin la necesidad de una variable respuesta previamente definida. A través de técnicas como el análisis de componentes principales, el análisis factorial y los métodos de agrupamiento, es posible identificar patrones latentes, reducir la complejidad de la información y reconocer perfiles o segmentos naturales en la población analizada. No obstante, los resultados derivados de estos procedimientos deben interpretarse con cautela, pues su carácter exploratorio no permite establecer conclusiones definitivas por sí mismo, sino que orienta hipótesis y líneas de análisis posteriores. [8].

Con lo anterior se da cierre al marco teórico, habiendo revisado los principales fundamentos conceptuales y metodológicos que sustentan el estudio, dejando claro que los hallazgos teóricos y metodológicos aquí descritos apoyan el análisis, pero no constituyen, en esta etapa, resultados concluyentes.

#### IV. METODOLOGÍA

La presente investigación se desarrolla bajo un enfoque cuantitativo y predictivo, basado en los principios del aprendizaje supervisado y no supervisado en estadística multivariada. Se empleó el software R como entorno de análisis estadístico, dado que este permite de forma eficiente procesos de limpieza, análisis exploratorio, reducción de dimensionalidad, modelado predictivo mediante librerías especializadas y su capacidad para manejar grandes volúmenes de datos. El entorno R permite además la automatización de flujos de trabajo mediante scripts reproducibles, garantizando la trazabilidad de los análisis y la validación de los resultados, elementos fundamentales en la investigación estadística multivariada [9].

El objetivo metodológico consistió en construir un modelo predictivo capaz de estimar el precio de venta de automóviles usados publicados en la plataforma AutoScout24, utilizando variables técnicas cuantitativas y cualitativas disponibles públicamente en la base de datos de Kaggle [10].

Esta investigación se sustenta en que las variables técnicas y de rendimiento de los vehículos permiten estimar de forma significativa el precio de venta de los automóviles usados en la plataforma AutoScout24.



Fig. 1 Esquema del proceso metodológico.

### A. Fuente y selección de los datos

El análisis se realizó a partir de una base de datos de vehículos usados que contiene información técnica, de uso y características comerciales de automóviles ofrecidos en el mercado [10]. Este dataset contiene información detallada de miles de anuncios de automóviles usados publicados en el portal AutoScout24, compuesta por 23 variables con 15.915 observaciones; la base de datos ofrece una representación significativa del mercado automotor de segunda mano en Europa, lo que la convierte en una fuente adecuada para la exploración, el análisis estadístico y la modelación predictiva. Las variables de la base de datos se muestran en la **TABLA I**:

TABLA I  
VARIABLES BASE DE DATOS

| VARIABLE               | UNIDAD DE MEDIDA                                                                  | TIPO DE VARIABLE |
|------------------------|-----------------------------------------------------------------------------------|------------------|
| make_model             | La marca y el modelo del vehículo                                                 | Cualitativa      |
| body_type              | El estilo de carrocería del vehículo.                                             | Cualitativa      |
| <b>price</b>           | <b>Precio del auto en dólares</b>                                                 | <b>Continua</b>  |
| vat                    | Indica el estado del IVA para el precio del vehículo.                             | Cualitativa      |
| <b>km</b>              | <b>Kilómetros recorridos</b>                                                      | <b>Continua</b>  |
| Type                   | Estado del vehículo, ya sea “Usado” o “Nuevo”.                                    | Cualitativa      |
| Fuel                   | Tipo de combustible que utiliza el vehículo                                       | Cualitativa      |
| Gears                  | El número de marchas en la transmisión del vehículo                               | Cualitativa      |
| Comfort_Convenience    | Características de comodidad y conveniencia.                                      | Cualitativa      |
| Entertainment_Media    | Funciones multimedia disponibles en el vehículo.                                  | Cualitativa      |
| Extras                 | Características adicionales                                                       | Cualitativa      |
| Safety_Security        | Elementos de seguridad como 'ABS'.                                                | Cualitativa      |
| age                    | Edad del vehículo (calculada en función del año del modelo).                      | Cualitativa      |
| Previous_Owners        | El número de propietarios anteriores que ha tenido el coche.                      | Cualitativa      |
| <b>hp_kW</b>           | <b>Potencia del motor en kilowatts (kW)</b>                                       | <b>Continua</b>  |
| Inspection_new         | Indica si el vehículo ha sido inspeccionado recientemente (1 para sí, 0 para no). | Cualitativa      |
| Paint_Type             | El tipo de pintura del coche, como 'Metálico', 'Mate', etc.                       | Cualitativa      |
| Upholstery_type        | El material utilizado para la tapicería interior, como 'Tela', 'Cuero', etc.      | Cualitativa      |
| Gearing_Type           | El tipo de transmisión que utiliza el coche ya sea “Automática” o “Manual”.       | Cualitativa      |
| <b>Displacement_cc</b> | <b>El motor en centímetros cúbicos (cc), indica el tamaño del motor.</b>          | <b>Continua</b>  |
| <b>Weight_kg</b>       | <b>Kilogramos</b>                                                                 | <b>Continua</b>  |
| Drive_chain            | El tipo de transmisión, indicando si es tracción “delantera” o “trasera”.         | Cualitativa      |
| <b>cons_comb</b>       | <b>El consumo de combustible en litros por cada 100 kilómetros.</b>               | <b>Continua</b>  |

## B. Análisis descriptivo y multivariado

### 1) Análisis exploratorio de datos (EDA):

El análisis descriptivo se realizó mediante la generación de histogramas y diagramas de dispersión de las variables continuas, arrojando los siguientes resultados en las medidas de tendencia central, variabilidad y de forma (**TABLA II**):

TABLA II  
ANÁLISIS EXPLORATORIO DE DATOS

| Medidas descriptivas    |                   |         |       |              |          |         |           |          |
|-------------------------|-------------------|---------|-------|--------------|----------|---------|-----------|----------|
| Variables cuantitativas |                   |         |       |              |          |         |           |          |
| Variable                | Tendencia central |         |       | Variabilidad |          |         | Forma     |          |
|                         | Media             | Mediana | Moda  | Varianza     | DesvEst  | CoefVar | Asimetría | Curtosis |
| price                   | 18024.38          | 16900   | 14990 | 54489189.55  | 7381.68  | 40.95   | 1.236     | 5.964    |
| km                      | 32090             | 20413   | 10    | 1367314350   | 36977.21 | 115.23  | 1.623     | 5.824    |
| hp_kW                   | 88.5              | 85      | 85    | 711.52       | 26.67    | 30.14   | 1.292     | 7.385    |
| Displacement_cc         | 1428.66           | 1461    | 1598  | 76068        | 275.8    | 19.31   | -0.097    | 2.798    |
| Weight_kg               | 1337.7            | 1295    | 1163  | 39873.05     | 199.68   | 14.93   | 1.082     | 4.594    |
| cons_comb               | 4.83              | 4.8     | 4     | 0.75         | 0.87     | 17.95   | 0.419     | 3.115    |

Donde se evidencia que la variable de intereses de estudio, Price presenta valores atípicos hacia la derecha, lo que significa que, en la base de datos, hay vehículos muy costosos, que se alejan demasiado de las medidas de tendencia central; igualmente se presente este tipo de comportamiento en las variables de km y potencia.

Posteriormente, se construyó una matriz de correlaciones que reveló las relaciones entre las variables continuas (**Fig. 2**):

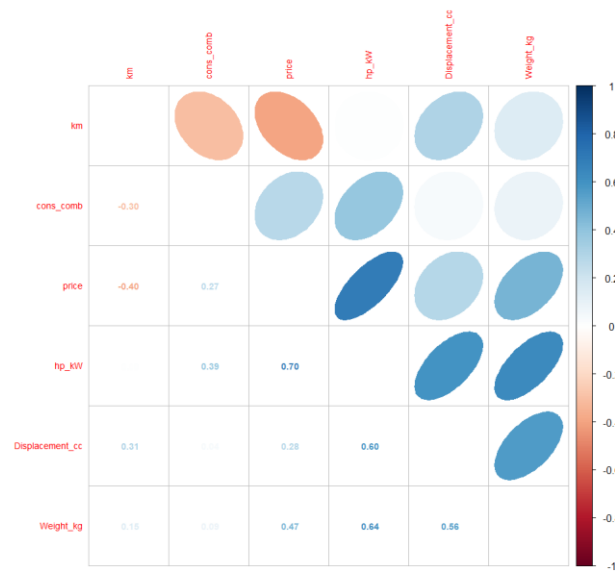


Fig. 2 Matriz de correlaciones entre las variables continuas

De la matriz de correlaciones podemos observar del precio(price), las siguientes relaciones:

- Fuerte correlación con potencia hp\_kW (0.70): Autos con más potencia suelen ser más caros.
- Correlación media con el peso Weight\_kg (0.47): Los autos más pesados tienden a tener mayor precio.
- Correlación baja con Displacement\_cc (0.28): Cilindraje ayuda, pero no explica tanto como la potencia.
- Correlación negativa con kilometraje (-0.40): A más kilometraje, menor precio (esperado en autos usados).
- Este análisis permitió confirmar una correlación moderada entre variables técnicas, sugiriendo multicolinealidad potencial que debe ser abordar en fases posteriores en el análisis de supuestos del modelo seleccionado.

### C. *Análisis no supervisado*

Este análisis se realiza con el objetivo de comprobar relaciones entre las variables de la base de datos; se realizó disminución de dimensionalidad mediante PCA, análisis factorial y clasificación mediante clúster no jerárquico, como ejercicio de comportamiento de las variables.

### 1) *PCA*

Se aplicó un Análisis de Componentes Principales (PCA) con todas las variables continuas, obteniéndose una reducción significativa de la dimensionalidad. Los dos primeros componentes explicaron un porcentaje de 72.1%, de la varianza total, destacando el peso y la potencia como variables con mayor carga factorial.

### 2) *FA*

Asimismo, se ejecutó un Análisis Factorial (FA) tanto con las variables originales como con sus transformaciones logarítmicas, logrando estructuras factoriales más estables tras la transformación. Este procedimiento permitió identificar factores latentes asociados al rendimiento técnico y al consumo del vehículo, respaldando la reducción de dimensionalidad antes del modelado supervisado.

### 3) *Clúster*

Se optó por un enfoque no jerárquico mediante k-means, centrando el análisis en la variable precio y en los componentes principales obtenidos previamente. Se determinaron los centroides mediante el método del codo (Elbow Method), encontrando una solución óptima cercana a  $k = 3$  clústeres, que representaron rangos bajos, medios y altos de precios. Esta segmentación permitió evidenciar agrupamientos lógicos de vehículos con características técnicas y de consumo similares, aportando información relevante para las fases de predicción y regresión.

## *D. Criterios de limpieza y transformación*

### 1) *Modelo lineal generalizado:*

En la base de datos se encontraron 3.033 vehículos que presentaban iguales características en las variables seleccionadas por ende se tomó la decisión de eliminar estos valores duplicados con el objetivo de que no genere sesgo en el análisis y, además, se eliminaron los datos atípicos de vehículos que se alejaban demasiado de la media del precio de los autos.

Se aplicaron criterios estadísticos para la detección de valores atípicos en variables continuas mediante intervalos intercuartílicos, descartando aquellos puntos que distorsionaban la distribución.

### 2) *Transformaciones MLG:*

Después, se aplicó una transformación logarítmica a las variables continuas altamente asimétricas, especialmente al precio. Este tipo de transformación es ampliamente utilizada cuando la distribución de una variable presenta una cola larga hacia la derecha, grandes diferencias entre valores mínimos y máximos o la presencia de valores extremos que pueden influir de manera desproporcionada en los modelos. La transformación logarítmica tiene como finalidad reducir la asimetría, estabilizar la varianza y atenuar el efecto de los valores atípicos, permitiendo que las relaciones estadísticas entre variables sean más lineales y que los modelos predictivos basados en regresión o distancias operen de manera más estable y confiable [11].

Por último, se transformaron como factores las variables Fuel (tipo de combustible), age (edad del vehículo) y type (estado del auto). Esta transformación fue necesaria porque los algoritmos estadísticos y de aprendizaje automático no pueden procesar directamente categorías en formato textual. Lo que permite evaluar su influencia en el precio del vehículo, ya que estas variables son un factor determinante en aspectos como eficiencia y mantenimiento.

En conjunto, ambas transformaciones contribuyen a mejorar la calidad de los datos, la estabilidad de los análisis y la capacidad predictiva de los modelos, facilitando interpretaciones más robustas y coherentes con el comportamiento real del mercado.

### 3) *Análisis Discriminante Cuadrático QDA:*

Para el modelo de análisis discriminante cuadrático **QDA**, considerando las observaciones del experto. Primero se procede a clasificar el precio de los autos de la base de datos original en tres categorías:

- Autos de precios bajos: hasta los 30.000 euros.
- Autos de precios medios: entre 30.001 y los 50.000 euros
- Autos de precios altos: mayores de 50.000

Luego, se realizó el balance de datos, puesto que, en la base de datos, había demasiados autos clasificados con precios bajos, y se debían equiparar las categorías para evitar sesgos del modelo.

*E. Evaluación de los modelos candidatos y selección final:*

En primera instancia se implementó una regresión lineal simple, donde cada variable independiente explicaba menos del 20% del precio, por lo cual este modelo fue descartado.

Posterior se procede a realizar regresión lineal múltiple, donde las variables continuas explicaban el 56% del comportamiento del precio de autos. Aunque era una cifra relevante el modelo no cumplió con el supuesto de linealidad del modelo, lo que lo invalida.

Por último, se analiza un análisis del modelo lineal generalizado, que permite incluir variables categóricas, las cuales fueron tipo de combustible, tipo de auto (nuevo o usado) y edad del auto; permitiendo que el análisis y los supuestos se cumplieran.

Además, se evaluó un modelo de clasificación, en este caso el análisis discriminatorio lineal, pero no cumplió con los supuestos que plantea el modelo; se decide implementar el análisis discriminatorio cuadrático, ya que este hace que el comportamiento del precio de los autos permita representar mejor la distribución de los datos.

*F. Indicadores estadísticos e hipótesis de modelado*

Para la evaluación de los modelos predictivos se emplean diversos indicadores estadísticos. Entre ellos destacan el coeficiente de determinación ( $R^2$ ), que mide la proporción de la variabilidad explicada por el modelo; el error cuadrático medio (RMSE), que evalúa la desviación promedio entre los valores reales y los estimados. En modelos de clasificación se utilizan métricas como accuracy, y la precisión, que reflejan la capacidad del modelo para predecir correctamente las categorías de precios y que se puede observar mediante la matriz de confusión del análisis discriminante cuadrático.

## V. ANÁLISIS DE RESULTADOS

### A. Presentación de los resultados

La variable respuesta principal corresponde al precio del vehículo, mientras que las variables explicativas seleccionadas incluyen características continuas como kilometraje, potencia del motor, cilindraje, peso y consumo combinado de combustible, así como variables categóricas asociadas al tipo de combustible, tipo de vehículo y antigüedad (TABLA III).

TABLA III  
VARIABLES SECCIONADAS PARA EL MODELO

| VARIABLE               | UNIDAD DE MEDIDA                                                         | TIPO DE VARIABLE |
|------------------------|--------------------------------------------------------------------------|------------------|
| <b>price</b>           | <b>Dólares</b>                                                           | <b>Continua</b>  |
| <b>km</b>              | <b>Kilómetros</b>                                                        | <b>Continua</b>  |
| Type                   | Estado del vehículo, ya sea “Usado” o “Nuevo”.                           | Cualitativa      |
| Fuel                   | Tipo de combustible que utiliza el vehículo                              | Cualitativa      |
| age                    | Edad del vehículo (calculada en función del año del modelo).             | Cualitativa      |
| <b>hp_kW</b>           | <b>Potencia del motor en kilowatts (kW)</b>                              | <b>Continua</b>  |
| <b>Displacement_cc</b> | <b>El motor en centímetros cúbicos (cc), indica el tamaño del motor.</b> | <b>Continua</b>  |
| <b>Weight_kg</b>       | <b>Kilogramos</b>                                                        | <b>Continua</b>  |
| <b>cons_comb</b>       | <b>El consumo de combustible en litros por cada 100 kilómetros.</b>      | <b>Continua</b>  |

Se evaluaron dos modelos, un de regresión lineal generalizado MLG donde se incluyó las variables cualitativas de edad, tipo de vehículo y tipo de combustible utilizado; y por otro lado se realizó un análisis discriminante cuadrático, donde se clasificaron los vehículos en tres categorías de autos, con precios bajos, precios medianos y por último autos de precios altos.

#### 1) Modelo lineal generalizado MLG

El Modelo Lineal Generalizado fue estimado utilizando como variable respuesta el logaritmo del precio del vehículo usado (L\_price), con el fin de estabilizar la varianza y aproximar la normalidad de los errores. Como variables explicativas se incluyeron características técnicas y de uso del vehículo, específicamente: kilometraje (L\_km), potencia (L\_hp\_kW), cilindraje

(L\_Displacement\_cc), peso (L\_Weight\_kg), consumo combinado (L\_cons\_comb), tipo de combustible (Fuel), antigüedad del vehículo (age) Y tipo de vehículo (Type).

En la siguiente gráfica (**Fig. 3**) podemos observar el comportamiento del modelo:

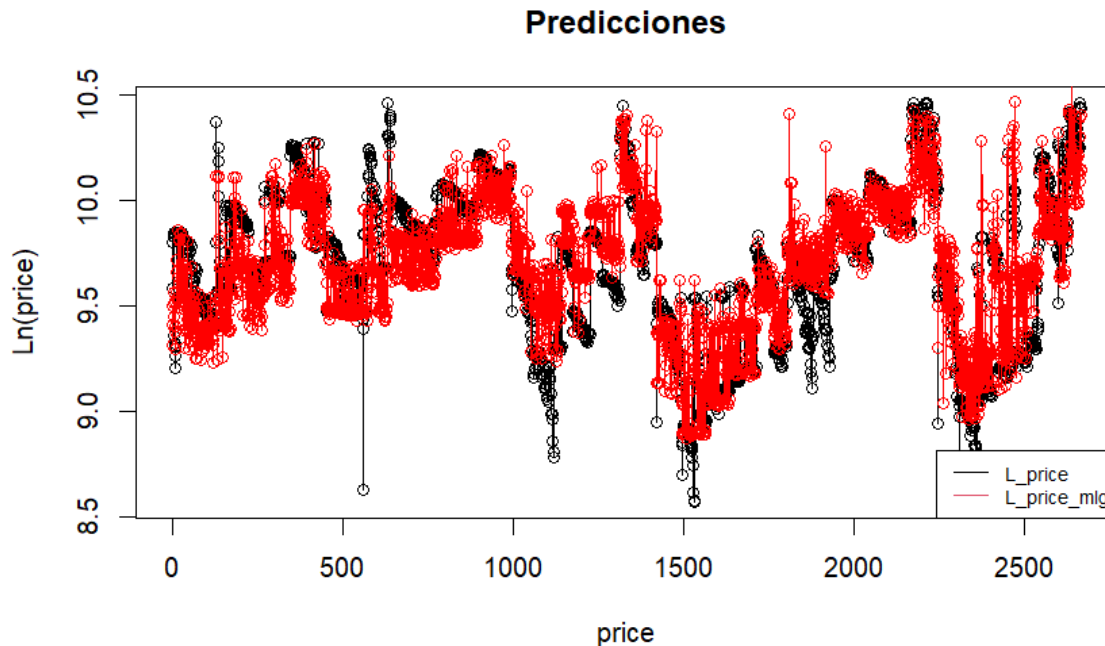


Fig. 3 Predicciones del modelo lineal generalizado.

Mientras que el Modelo Lineal Generalizado (MLG) tiene como objetivo predecir el valor continuo del precio del vehículo, el Análisis Discriminante Cuadrático (QDA) se emplea para clasificar los vehículos en categorías de precio (Alto, Medio y Bajo), proporcionando una perspectiva complementaria del problema.

El QDA se justifica frente al LDA debido a la violación del supuesto de homogeneidad de matrices de covarianza, evidenciada por la prueba de Box's M, la cual resultó estadísticamente significativa ( $p\text{-value} \approx 0$ ), tanto en variables originales como transformadas logarítmicamente. Esto indica que las matrices de covarianza difieren entre grupos, condición bajo la cual QDA es más adecuado.

## 2) Análisis discriminante cuadrático QDA

El modelo QDA fue estimado utilizando las mismas variables continuas empleadas en el MLG, garantizando coherencia entre ambos enfoques. Los resultados principales indican:

Las funciones discriminantes están dominadas por variables relacionadas con:

- Kilometraje
- Consumo de combustible
- Potencia del motor

La primera función discriminante concentra la mayor proporción de la variabilidad entre grupos ( $> 94 \%$ ), lo que sugiere una clara separación entre las categorías de precio.

Los centroides de grupo muestran diferencias sustanciales, especialmente entre las categorías Alta y Baja, confirmando la capacidad del modelo para discriminar entre segmentos del mercado.

Desde el punto de vista de supuestos:

- La normalidad multivariada no se cumple estrictamente, lo cual es común en bases de datos grandes y heterogéneas.
- La ausencia de multicolinealidad severa fue verificada mediante VIF y correlaciones, sin evidencia de problemas críticos, todos los valores fueron inferiores a 5.
- La heterogeneidad de covarianzas respalda la elección de QDA sobre LDA.

Al ejecutar el modelo, se halla la matriz de confusión:

TABLA IV  
MATRIZ DE CONFUSIÓN DEL MODELO QDA

| Desempeño en test |      |      |       |
|-------------------|------|------|-------|
|                   | Real |      |       |
| Predicción        | Alto | Baja | Media |
| Alto              | 2485 | 3    | 402   |
| Baja              | 0    | 2525 | 139   |
| Media             | 503  | 463  | 2374  |
| Accuracy : 0.830  |      |      |       |

| LOOCV en train          |      |       |       |
|-------------------------|------|-------|-------|
|                         | Real |       |       |
| Predicción              | Alto | Baja  | Media |
| Alto                    | 9595 | 33    | 1495  |
| Baja                    | 0    | 10211 | 445   |
| Media                   | 2357 | 1724  | 9721  |
| Accuracy (LOOCV): 0.830 |      |       |       |

## *B. Análisis de los resultados*

### *1) Análisis modelo lineal generalizado*

Los resultados del modelo Lineal generalizado indican que la mayoría de las variables continuas resultaron estadísticamente significativas ( $p < 0.05$ ), lo que evidencia su contribución explicativa sobre el precio del vehículo. En particular:

- El kilometraje y la antigüedad presentan coeficientes negativos, confirmando que un mayor uso y mayor edad del vehículo se asocian con una reducción del precio.
- La potencia del motor y el peso muestran coeficientes positivos, indicando que vehículos con mejores valores altos en estas características, tienden a tener precios más altos.
- El consumo de combustible y el cilindraje presentan efectos negativos, sugiriendo una penalización en el precio para vehículos menos eficientes.

En cuanto a las variables categóricas, el tipo de combustible diésel y LPG/CNG resultaron significativos frente a la categoría de gasolina, mientras que la categoría eléctrica no mostró un efecto estadísticamente significativo.

Desde el punto de vista de la especificación del modelo, la prueba RESET de Ramsey no evidenció problemas de mala especificación funcional ( $p > 0.05$ ), lo cual sugiere que la forma funcional del modelo es adecuada.

#### Evaluación del error y capacidad predictiva

La capacidad predictiva del modelo fue evaluada sobre un conjunto de prueba independiente, empleando métricas de error ampliamente utilizadas en modelos de regresión:

Error Cuadrático Medio (MSE): 0.0404

Raíz del Error Cuadrático Medio (RMSE): 0.2010

Dado que la variable respuesta se encuentra en escala logarítmica, el valor de  $RMSE \approx 0.20$  puede interpretarse como un error relativo aproximado del 20–22 % en el precio, lo cual es consistente con resultados reportados en la literatura para modelo esperados de precios de vehículos usados. Este nivel de error se considera aceptable para aplicaciones de predicción económica y

análisis de mercado, especialmente considerando las diferencias en las características presentadas en el mercado automotriz.

## 2) *Análisis QDA*

De la TABLA IV podemos inferir la interpretación del modelo:

- El modelo distingue bien la categoría Alto, aunque existe confusión con Media, lo cual es esperable por la cercanía de precios entre ambos grupos.
- El modelo presenta un desempeño sobresaliente en la identificación de vehículos de bajo precio, con una separación clara respecto a las demás categorías.
- La categoría Media es la más difícil de clasificar, ya que actúa como un grupo intermedio, y presenta solapamiento parcialmente con los niveles Altos y Bajos.
- Presenta un indicador de Accuracy =0.830, Esto indica que el modelo clasifica correctamente el 83 % de los vehículos en datos nuevos; En problemas multiclase reales, un accuracy superior al 80 % se considera alto, especialmente cuando las clases presentan solapamiento natural [6].

## VI. CONCLUSIONES Y RECOMENDACIONES

### A. CONCLUSIONES

El objetivo general de la investigación, consistente en realizar un modelado predictivo de precios de venta de automóviles usados para la plataforma AutoScout24 utilizando R, se cumplió mediante la preparación del conjunto de datos, la aplicación de técnicas de análisis estadístico y la implementación de modelos de aprendizaje supervisado y no supervisado. Se logró construir y evaluar modelos predictivos que permiten estimar el precio de los automóviles usados en función de sus características observadas.

En relación con los objetivos específicos, se verificó, en primer lugar, que el proceso de depuración y preparación de los datos permitió identificar valores atípicos, así como aplicar las transformaciones requeridas para garantizar la calidad del conjunto de datos utilizado en los análisis posteriores. En segundo lugar, se implementaron técnicas de ingeniería de características que permitieron representar de manera más adecuada la información y observar mejoras en el desempeño de los modelos ajustados.

Respecto al análisis descriptivo y multivariado, se observaron patrones de comportamiento en las variables y relaciones entre ellas que permitieron comprender mejor la estructura de los datos. Las técnicas de reducción de dimensionalidad facilitaron la simplificación del conjunto de variables sin pérdida sustancial de información relevante. En cuanto al aprendizaje supervisado y no supervisado, se implementaron diferentes algoritmos, se ajustaron sus parámetros y se verificó su comportamiento frente al conjunto de datos disponible.

Finalmente, al evaluar el desempeño de los modelos, fue posible identificar aquellos que presentaron mayor precisión en la predicción del precio de los vehículos usados, de acuerdo con los indicadores de error utilizados. Los resultados obtenidos permiten observar tendencias y comportamientos de los datos que aportan insumos relevantes para la toma de decisiones.

Dentro del aprendizaje supervisado, los modelos de regresión lineal y múltiple son los más utilizados en problemas de predicción continua como el precio, mientras que los métodos logísticos, discriminantes (LDA y QDA) permiten identificar categorías o niveles de precios. Por otro lado, los métodos no supervisados como PCA y análisis factorial (FA) son útiles para reducir

la dimensionalidad y eliminar redundancia entre variables correlacionadas, mejorando la estabilidad de los modelos posteriores; estos últimos no fueron utilizados ya que los variables no presentaban multicolinealidad.

El modelo de Análisis Discriminante Cuadrático (QDA) mostró un desempeño sólido en la clasificación de vehículos en categorías de precio, alcanzando una exactitud del 83 %. La matriz de confusión evidencia una alta capacidad para identificar correctamente los vehículos de bajo precio, mientras que la mayor fuente de error se concentra en la categoría media, debido al solapamiento natural con las clases adyacentes. La consistencia entre los resultados de prueba y LOOCV sugiere una adecuada capacidad de generalización y ausencia de sobreajuste.

La hipótesis planteada se verificó en el sentido de que fue posible construir modelos predictivos de precios de automóviles usados con un nivel de desempeño medible y evaluable mediante métricas estadísticas. Se observó que las variables seleccionadas tienen capacidad explicativa sobre el precio y que los algoritmos aplicados logran aproximaciones consistentes de este. No obstante, la verificación realizada corresponde al ámbito de los datos y métodos utilizados, sin extenderse a generalizaciones más allá del alcance del estudio.

## *B. RECOMENDACIONES*

A partir del desarrollo de la investigación se formulan las siguientes recomendaciones:

Se recomienda modular el estudio en subsecciones temáticas para futuras investigaciones, considerando líneas como: comportamiento por marca, por país, por tipo de combustible o por transmisión, que no fueron analizadas de manera específica en el presente trabajo.

Es aconsejable documentar nuevas iteraciones metodológicas a partir de los procesos aplicados aquí, aprovechando la experiencia adquirida y evitando caminos que no aportaron mejoras significativas en el desempeño de los modelos.

## REFERENCIAS

- [1] Grand View Research, “Used Car Market (2025 - 2030)”. 2025. [En línea] Disponible en: <https://www.grandviewresearch.com/industry-analysis/used-car-market>
- [2] S. Muti y K. Yıldız, “Using Linear Regression for Used Car Price Prediction,” *International Journal of Computational and Experimental Science and ENgineering*, vol.9, no. 1, pp. 11-16, 2023. [En línea]. Disponible en: <https://dergipark.org.tr/en/download/article-file/2242032>
- [3] S. Kumar y A. Sinha, “Predicting Used Car Prices with Regression Techniques,” *International Journal of Computer Trends and Technology (IJCTT)*, vol. 72, no. 6, pp. 132–141, 2024. [En línea]. Disponible en: <https://doi.org/10.14445/22312803/IJCTT-V72I6P118>
- [4] S. Huang, Y. Zhu, J. Huang, E. Zhang y T. Xu, “Analysis of Circular Price Prediction Strategy for Used Electric Vehicles,” *Sustainability*, vol. 16, no. 13, 5761. 2023. [En línea]. Disponible en: <https://www.mdpi.com/2071-1050/16/13/5761>
- [5] S. Bergmann y S. Feuerriegel, “Machine learning for predicting used car resale prices using granular vehicle equipment information,” *Expert Systems with Applications*, vol. 263, 2024. [En línea]. Disponible en: <https://www.sciencedirect.com/science/article/pii/S0957417424025077>
- [6] D. N. Gujarati y D. C. Porter, *Econometría*, 5th ed. México, D.F.: McGraw-Hill/Interamericana Editores, 2010.
- [7] C. M. Bishop, *Pattern Recognition and Machine Learning*, New York, NY, USA: Springer, 2006.
- [8] T. Hastie, R. Tibshirani, y J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* 2nd ed. New York, NY, USA: Springer, 2009.
- [9] R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing, 2024. [En línea]. Disponible en: <https://www.R-project.org/>
- [10] Y. Iitturan, “Auto Scout Car Price [Dataset]. Kaggle,” 2019. [En línea]. Disponible en: <https://www.kaggle.com/datasets/yaaryiitturan/auto-scout-car-price>
- [11] R. J. Larsen y M. L. Marx, *An Introduction to Mathematical Statistics and Its Applications*, 5th ed. Pearson, 2012.

## ANEXOS

## Anexo A. Imagen de poster

