



Web Scraping y modelación del precio de la oferta de autos usados en venta en la ciudad de Cali en el año 2023

Jose Francisco Botero

Universidad del Valle
Facultad de Ingeniería, Escuela de Estadística
Santiago de Cali, Colombia
2024

Web Scraping y modelación del precio de la oferta de autos usados en venta en la ciudad de Cali en el año 2023

Jose Francisco Botero

Trabajo de grado presentado como requisito parcial para optar al título de:
Estadístico

Director:
Ph.D. Jaime Mosquera Restrepo

Universidad del Valle
Facultad de Ingeniería, Escuela de Estadística
Santiago de Cali, Colombia
2024

Resumen

La compra-venta de autos usados en Colombia es una actividad que posee parte de su cuota de mercado en internet, en donde son ofertados libremente vehículos en plataformas como Mercado Libre, OLX o Carro Ya. En este Trabajo de grado se modela el precio de la oferta de autos usados en la ciudad de Cali mediante los datos obtenidos a través de Web Scraping del sitio web de Mercado Libre Colombia para septiembre de 2023, lo que permitirá tener mejores precios de referencia para comprar o vender un auto usado. Además se estudia el efecto sobre el precio de las características técnicas y de uso de los vehículos como el cilindraje, kilometraje, marca, modelo o antigüedad.

En la modelación del precio de los autos usados se emplearon características del vehículo proporcionadas por los vendedores en el sitio web. Se utilizaron técnicas de preprocesamiento y limpieza de datos para el posterior ajuste de los modelos estadísticos como el Modelo lineal mediante mínimos cuadrados ordinarios y el Modelo lineal generalizado bajo la distribución gamma, en donde se analizaron los efectos de las covariables sobre el precio y se validaron los supuestos del modelo. Como complemento, se emplearon modelos de aprendizaje automático como los Árboles de regresión, Bosques aleatorios y XGBoost, en donde se entrenaron mediante validación cruzada y se encontró el modelo de cada método que obtuviera el menor error de predicción.

Las variables que resultaron tener más efecto positivo sobre el precio son la marca y el cilindraje, mientras que aquellas que más afectan de manera negativa son el kilometraje y la antigüedad de los autos.

El modelo de XGBoost resultó ser el mejor modelo para predecir el precio de los vehículos usados, en donde este logra explicar el 86,7% de la variabilidad del precio mediante las covariables.

Palabras clave: *Autos usados, Web Scraping, Modelo lineal general, Modelo lineal generalizado, Árboles de regresión, Bosques aleatorios, Xgboost*

Abstract

The buying and selling of used cars in Colombia is an activity that has a portion of its market share on the internet, where vehicles are offered on platforms such as Mercado Libre, OLX, or Carro Ya. This thesis models the price of used cars offered in the city of Cali for September 2023 using data obtained through web scraping of the Mercado Libre Colombia website, which will allow for better reference prices when buying or selling a used car. Additionally, the effect on price of the technical and usage characteristics of vehicles such as engine size, mileage, body type, and age is studied.

In modeling the price of used cars, vehicle characteristics provided by sellers on the website such as engine size, mileage, brand, model, or age were employed. Data preprocessing and cleaning techniques were used for the subsequent fitting of statistical models such as linear regression using ordinary least squares and the Generalized Linear Model under the gamma distribution, where the effects of covariates on price were analyzed and the model assumptions were validated. Additionally, machine learning models such as Regression Trees, Random Forests, and XGBoost were employed, where they were trained using cross-validation, and the model of each method that presented the lowest prediction error was found.

The variables that were found to have the most positive effect on price are the brand and engine size, while those that most negatively affect price are mileage and age.

The XGBoost model turned out to be the best model for predicting the price of used vehicles, where it manages to explain 86,7 % of the price variability using the covariates.

keywords: *Used cars, Web Scraping, General linear model, Generalized linear model, Regression trees, Random forests, XGBoost.*

Contenido

1. Introducción	2
1.1. Planteamiento del problema	4
1.2. Justificación	5
1.3. Objetivos	6
1.3.1. Objetivo general	6
1.3.2. Objetivos específicos	6
2. Antecedentes	7
3. Marco teórico	12
3.1. Marco teórico conceptual	12
3.1.1. Industria automotriz	12
3.1.2. Oferta	12
3.1.3. Demanda	12
3.1.4. Depreciación	13
3.1.5. Web Scraping	13
3.2. Marco teórico estadístico	13
3.2.1. Preprocesamiento de datos	13
3.2.2. Modelo lineal general	16
3.2.3. Modelo lineal generalizado	19
3.2.4. Modelo lineal generalizado gamma	21
3.2.5. Modelos de aprendizaje automático	23
3.2.6. Árbol de decisión	23
3.2.7. Validación cruzada	27
3.2.8. Bagging	27
3.2.9. Boosting	28
3.2.10. Bosques aleatorios	29
3.2.11. XGBoost	31
4. Metodología	32
4.1. Entendimiento de los datos y población de estudio	33
4.2. Obtención de datos	34
4.3. Preprocesamiento de datos	36
4.3.1. Transformación de variables	37
4.3.2. Imputación de datos faltantes	39

4.4.	Análisis exploratorio	39
4.5.	Fase de entrenamiento y evaluación de modelos	40
4.5.1.	Modelo lineal general	41
4.5.2.	Modelo lineal generalizado gamma	41
4.5.3.	Árbol de regresión	42
4.5.4.	Bosques aleatorios	43
4.5.5.	XGBoost	44
4.6.	Comparación de modelos entrenados	44
5.	Resultados	46
5.1.	Análisis exploratorio	46
5.1.1.	Análisis de datos atípicos	54
5.2.	Modelación de la oferta de precios	55
5.2.1.	Fase de entrenamiento	56
5.2.2.	Modelo lineal general	56
5.2.3.	Modelo lineal generalizado gamma	57
5.2.4.	Árbol de regresión	64
5.2.5.	Bosque aleatorio	66
5.2.6.	XGBoost	68
5.2.7.	Selección del mejor modelo mediante datos de prueba	70
6.	Conclusiones y recomendaciones	74
6.1.	Conclusiones	74
6.2.	Recomendaciones	75
	Referencias	77
A.	Anexo: Hoja de datos de la oferta de autos usados en Cali en septiembre de 2023.	

Índice de figuras

3.1. Densidad del punto E respecto a sus vecinos	16
3.2. Distribución de densidad de probabilidad gamma	22
3.3. Esquema de la división de un nodo en el Árbol de regresión	25
3.4. Esquema del Árbol de regresión	26
3.5. Esquema de validación cruzada	27
3.6. Esquema de bagging	28
3.7. Esquema de Boosting	29
3.8. Esquema del Bosque aleatorio	30
3.9. Esquema del XGBoost	31
4.1. Flujo de trabajo de investigación	33
4.2. Ofertas de venta de autos usados en Mercado Libre.	34
4.3. Oferta de venta de un auto usado	35
4.4. Error de validación cruzada frente al número nodos terminales del modelo de Árbol de regresión del precio en función de la marca.	38
5.1. Histograma del precio de venta, antigüedad, kilometraje y cilindraje de los autos usados en millones de pesos.	47
5.2. Distribución de los modelos y marcas más ofertadas, color, tipo de combustible, tipo de carrocería, transmisión y número de puertas de los autos usados.	49
5.3. Diagrama de dispersión múltiple de las variables cuantitativas	50
5.4. Distribución del precio de acuerdo a las marcas y modelos más ofertados.	51
5.5. Distribución del precio de acuerdo al color, tipo de combustible, tipo de transmisión y carrocería, y número de puertas.	53
5.6. Diagrama de cajas del precio en millones de pesos.	54
5.7. Distribución del puntaje LOF de las observaciones.	55
5.8. Supuestos sobre los residuos del modelo de regresión lineal mediante mínimos cuadrados ordinarios.	57
5.9. Distribución de densidad y bondad de ajuste bajo la distribución gamma del precio en millones de pesos.	58
5.10. Distribución de los residuos estandarizados del Modelo lineal generalizado gamma.	62
5.11. Valores ajustados frente a los residuos estandarizados del Modelo lineal generalizado gamma.	63
5.12. Factores de inflación de varianza de las variables predictoras del modelo generalizado gamma.	64

5.13. Error de validación cruzada frente al número de nodos terminales del Árbol de regresión.	65
5.14. Esquema del modelo de Árbol de regresión.	66
5.15. Importancia de las variables predictoras del modelo de Bosque aleatorio . . .	68
5.16. Importancia de las variables predictoras del modelo de XGBoost	69
5.17. Coeficientes de determinación ajustado de los modelos frente a datos de prueba y entrenamiento mediante validación cruzada.	71
5.18. Raíz del error cuadrático medio mediante validación cruzada para datos de prueba.	72
5.19. Aplicación del modelo XGBoost del precio de autos usados en R Shiny . . .	73
6.1. Hoja de datos de la oferta de vehículos usados	79

Índice de tablas

4.1.	Variables proporcionadas por los vendedores en la oferta de venta	36
4.2.	Criterios de exclusión de registros	37
5.1.	Resultados de los coeficientes del Modelo lineal generalizado gamma completo.	59
5.2.	Resultados de los coeficientes del Modelo lineal generalizado gamma mediante la técnica stepwise.	61
5.3.	Rangos para búsqueda en cuadrícula de los hiperparámetros del modelo de Bosque aleatorio.	67
5.4.	Rangos para búsqueda en cuadrícula de los hiperparámetros del modelo de XGBoost.	68

1 Introducción

Hoy en día, existe una gran cantidad de sitios web en donde se ofertan autos usados en Colombia, como pueden ser Carro Ya, Olx o Mercado libre, siendo este último uno de los más populares del país. Gracias a estos servicios en línea, las personas con poco conocimiento en el sector pueden hacerse una idea del precio promedio al que pueden vender o comprar un vehículo usado, mediante la comparación de las ofertas existentes del modelo de vehículo en el que se está interesado. Sin embargo, esta referencia del precio de forma empírica no necesariamente puede ser la correcta debido a que si bien, las características técnicas pueden ser las mismas para un modelo de auto en específico, las variables de uso supondrán un cambio significativo en el rango de precios debido a que existen características únicas de cada oferta de venta como el kilometraje, número de dueños, antigüedad, ubicación del vendedor y hasta la paridad de la placa, son variables que tienen un efecto significativo sobre el precio de cada modelo y que de forma intuitiva puede ser difícil de medir.

Gracias al gran número de vehículos usados ofertados en estas plataformas, es posible estimar precios de referencia los cuales se puedan tomar de guía, ya que estas ofertas se pueden asumir como una muestra de la oferta a nivel nacional. El mercado de automóviles usados no está centralizado y se puede acceder mediante múltiples canales como concesionarios, redes sociales, anuncios en periódicos, de persona a persona y por supuesto, en sitios web de compra-venta los cuales se han vuelto muy populares gracias a internet.

En los últimos años, el mercado de autos usados ha tenido un gran crecimiento, ya que solo en 2021 se vendieron 1,7 millones de vehículos, representando un crecimiento del 32 % en comparación con el año 2020,. Además este aumento de la demanda también causó un aumento en los precios llegando a ser hasta del 30 %. Este crecimiento del mercado nacional se produjo debido a la escasez en la oferta de autos nuevos causada por la crisis global en las cadenas de suministros de los fabricantes internacionales de vehículos (A. Gutierrez, 2022).

Durante el año 2023, el precio de los autos usados también se ha ido incrementando debido a la inflación de la economía y la devaluación del peso colombiano, además la especulación de los precios de los autos usados ha generado un aumento significativo en la apreciación de estos, siendo que para el año 2023 se estimó una variación en los precios de los automóviles usados entorno al 20 % y 25 % en el mercado colombiano (Restrepo, 2023).

Teniendo en cuenta este crecimiento del mercado de autos usados y la popularidad de los portales de internet, se propone en este Trabajo de grado recopilar las ofertas de au-

tos usados publicadas por los vendedores ubicados en la ciudad de Cali en el sitio web de Mercado Libre Colombia. Mediante la técnica de Web Scraping se obtendrán las variables técnicas y de uso de los autos ofertados por los vendedores para su limpieza y procesamiento, asegurando la correcta implementación de las metodologías de análisis y modelado de datos, para posteriormente predecir el precio de los vehículos de acuerdo a sus características.

Para lograr esto es necesario tener en cuenta ciertos criterios de inclusión de las ofertas de autos recopiladas mediante la técnica anteriormente mencionada. Es importante conocer que cuando un vendedor decide ofertar su vehículo en la plataforma, debe diligenciar un formulario en donde ingresa las características del vehículo que desea vender. Las variables como la marca, el modelo, el kilometraje, el año del vehículo o el precio son variables obligatorias para realizar la publicación de la oferta. Sin embargo, otras variables como el color, tipo de carrocería y tipo de transmisión no son obligatorias de diligenciar, por lo que el vendedor puede dejarlas vacías en su oferta. Dado que algunos modelos de vehículos tienen versiones con transmisión automática o mecánica, distintos tipos de carrocería como hatchback o sedán, y están disponibles en varios colores, la imputación de valores faltantes en estas variables en específico, basada en características del vehículo como el modelo, precio o el cilindraje, puede generar incertidumbre sobre la confiabilidad de los valores imputados. Por esta razón, es necesario establecer un criterio de completitud de datos que incluya solo aquellas ofertas en las que todas las características hayan sido diligenciadas correctamente por el vendedor.

Además de esto, también es necesario limitar el precio máximo de los vehículos ofertados que se analizarán, ya que precios mayores a \$250 millones de pesos pueden considerarse que entran en otra categoría distinta a los vehículos particulares y de uso familiar, los cuales son vehículos deportivos y de muy alta gamma que no se incluirán en el estudio.

Las metodologías estadísticas, como el Modelo Lineal Generalizado desarrollado en este trabajo de grado, se empleó para explicar el precio de la oferta de autos usados a partir de las características proporcionadas por los vendedores, tales como el kilometraje, la antigüedad, la marca y el cilindraje. Estas metodologías permiten analizar el efecto de cada covariable sobre el precio, identificando cuáles tienen mayor importancia y su impacto positivo o negativo. Además, mediante el análisis exploratorio y de regresión, se podrán estudiar los patrones de comportamiento del precio en función de las características del vehículo.

Por otra parte, se construyeron modelos de aprendizaje automático como XGBoost y Bosques Aleatorios. Aunque estos modelos no permiten medir directamente el efecto de las características del vehículo sobre el precio, sí ofrecen predicciones precisas del precio de los vehículos usados.

El modelo analítico desarrollado en este Trabajo para predecir el precio de la oferta de autos usados en nuevas observaciones permitirá a los usuarios tener una referencia mucho más precisa del precio del automóvil en el que están interesados.

El Trabajo de grado también permitirá sentar un precedente para posteriores investigaciones en donde se analicen los precios de algún bien mediante metodologías estadísticas y de aprendizaje automático, presentándose la posibilidad de emplearlo como referencia para

estudiar el comportamiento del precio de la oferta de autos usados en la ciudad de Cali a través del tiempo.

1.1 Planteamiento del problema

Actualmente en Colombia, la fijación de un precio exacto en el proceso de compra-venta de autos usados puede presentar incertidumbre, existiendo cierta dificultad al establecer un intervalo de precios al que está dispuesto a pagar el usuario, en donde se juega un papel de *tira y afloja* entre el comprador y vendedor debido a la negociación del precio. Agregando a esto también se tiene la falta de conocimiento de la existencia de los factores que pueden influir en los precios, y en los cuales es difícil tener una percepción de su efecto de forma intuitiva, por lo que se añade una capa más de complejidad al proceso de vender o comprar un auto usado.

El interés de las personas al elegir comprar un auto usado radica principalmente en que estos se ofrecen a un menor precio en comparación con los autos nuevos, debido a que comprar un vehículo nuevo supondrá un devalúo del bien una vez que sale del concesionario (KBB, 2022). Es por estas razones que comprar un auto usado puede suponer en una gran mayoría de personas una mejor opción a largo plazo debido a que se posee un menor riesgo financiero, ya que el vehículo ha experimentado su mayor depreciación debido al uso.

Una mayor contextualización del problema se puede evidenciar en el estudio desarrollado por de Toro et al. (2017), quienes lograron obtener la oferta de los autos usados en Mercado Libre Argentina y modelaron el precio de acuerdo a las variables técnicas y de uso de los vehículos. Evidenciaron que variables como el kilometraje, antigüedad o tipo de carrocería, tienen un mayor efecto significativo sobre el precio de los vehículos. La forma en que el comprador evaluaría estos efectos de forma empírica está en realizar la tarea de comparar el precio de los vehículos que más se parezcan a las características en las que está interesado y determinar un precio medio, algo que en la práctica requiere inversión de tiempo y que no suele llegar a ser un método muy preciso debido a que no se sabe con exactitud que tanto puede afectar el uso y desgaste a una determinada marca y modelo de automóvil.

Cómo se ha mencionado, fijar el precio de un auto usado puede variar mucho dependiendo de la percepción que tiene el comprador o vendedor sobre el mercado. Un usuario con mucha experiencia sabrá de forma más precisa el rango de precios al que debería vender su vehículo y no aceptará un valor menor teniendo en cuenta sus características y desgaste.

Actualmente existen varias formas de realizar este proceso, las cuales pueden ser más complejas que otras. Como primera opción y más sencilla de realizar, es la comparación de los precios de los vehículos en internet aprovechando las plataformas como Mercado Libre, OLX u otras más especializadas como Carro Ya. En estos sitios web, el usuario puede filtrar las ofertas de acuerdo a la ciudad en la que se encuentra el vendedor y a las características en las que está interesado. De esta forma, puede hacerse una idea del rango de precios de un modelo de vehículo en específico, siendo una de las opciones para fijar los precios más

fácil y rápida de implementar gracias a la accesibilidad de internet. Es esta alternativa en la que se basará este Trabajo de grado, aprovechando la gran cantidad de ofertas publicadas y explicando el precio de los autos mediante varios modelos como el Modelo lineal general y generalizado, Árbol de regresión, Bosque aleatorio y XGBoost.

Otra opción para fijar los precios, que era muy común en décadas pasadas en donde el internet no era tan popular, es en realizar la comparación de los precios mediante revistas especializadas como pueden ser la revista *Motor* o *Car and Driver*, en las cuales se presenta información del catálogo de los vehículos más populares y se dan referencias de precios de vehículos tanto nuevos como usados.

Otro método usado para realizar la comparación de precios y decidir en qué rango se va a comprar o vender un auto, es mediante la consulta en concesionarios. Esto permite obtener una perspectiva del valor de un vehículo de forma realista y bastante precisa. Sin embargo, es una tarea que puede tomar bastante tiempo y que no todos los usuarios están dispuestos a realizar.

La fijación del precio de un auto usado requiere que el usuario invierta tiempo en realizar una investigación para obtener un contexto de los precios del mercado en donde se encuentra. La variabilidad de estos de acuerdo a las características técnicas y de uso del auto, además de la ubicación geográfica en la que está ubicado el vendedor, puede dificultar la fijación de una forma precisa de los precios. Sin embargo, gracias a la era digital esta tarea se ha facilitado gracias a plataformas en línea, redes sociales, revistas y foros.

Es debido al planteamiento desarrollado en este capítulo que es necesario preguntarse: ¿Cuál es el precio promedio que se oferta en plataformas digitales para un determinado modelo de vehículo usado teniendo en cuenta las características técnicas y de uso en la ciudad de Cali durante el año 2023?

1.2 Justificación

La amplia variedad de plataformas digitales de compra-venta de bienes como los autos usados permiten a su vez tener a disponibilidad un gran número de ofertas de marcas y modelos de vehículos, esto es conveniente para el comprador ya que le permite negociar con muchos vendedores y obtener una mayor oportunidad de ofertas, sin embargo, a su vez puede llegar a provocar tanto en vendedores como en compradores, un posible bloqueo por sobreinformación debido a los rangos de precios que pueden existir para un modelo de auto en específico, llegando a sobrevalorar o devaluar fácilmente el precio de vehículo.

El análisis de las características técnicas y de uso del vehículo es importante para conocer en que medida estas impactan al precio del mismo. Definir mediante métodos estadísticos y de aprendizaje automático, los precios incluyendo las variables más importantes del vehículo como la marca, modelo, cilindraje, tipo de carrocería y transmisión, ayudará a brindar una base de precios más precisos.

Debido a esta situación, es necesaria una profunda investigación que estudie las principales variables que influyen en el precio de los autos usados en la ciudad de Cali, en donde se tendrán en cuenta tanto variables de uso como también, características técnicas del vehículo.

De esta forma se propone modelar el precio de la oferta de autos usados en plataformas digitales ofrecidos en Cali mediante distintos métodos de regresión y de aprendizaje automático, identificando las variables que más puedan influir en el precio y ofreciendo una estimación insesgada, lo que permitirá a las entidades financieras plantear nuevas estrategias para otorgar créditos para la compra de autos usados, o a entidades aseguradoras las cuales podrían determinar una estimación más precisa del valor del vehículo que pueda ser implementada en modelos de actuaría y así obtener primas de seguros más precisas.

1.3 Objetivos

1.3.1 Objetivo general

Modelar el precio de la oferta de autos usados ofrecida en Cali mediante plataformas digitales en el año 2023 considerando variables técnicas y de uso del vehículo.

1.3.2 Objetivos específicos

- Conformar una hoja de datos a través de Web Scraping y técnicas de limpieza de datos con la información actualizada de la oferta de autos usados ofrecida en Cali mediante plataformas digitales.
- Determinar los factores que influyen significativamente en el precio de los autos usados mediante análisis estadístico exploratorio y de regresión.
- Obtener un modelo predictivo para la estimación del precio de la oferta de autos usados utilizando variables técnicas y de uso del vehículo mediante métodos estadísticos y de aprendizaje automático.

2 Antecedentes

En este capítulo se presenta la revisión de antecedentes del Trabajo de grado. Se muestran estudios tanto de Colombia como a nivel internacional sobre la modelación del precio de autos nuevos y usados, en los cuales se evidencian las técnicas implementadas para dar solución a los objetivos de cada investigación. Se identifican las metodologías más comunes usadas para modelar el precio de los vehículos, además del método empleado para la obtención de los datos.

En la primera parte de este capítulo se revisan las investigaciones relacionadas con la modelación de los precios de autos, comenzando por las más recientes y retrocediendo hasta las más antiguas en Argentina, China, Turquía y Alemania realizadas por de Toro et al. (2017), Liu y Li (2022), Koç y Kostak (2021) y Kihm y Vance (2016) respectivamente.

Por último, se analizan los dos trabajos de grado realizados por Gutierrez (2022) y Barrios y Quinto (2016), los cuales están relacionados con la modelación del precio de la oferta de viviendas en la ciudad de Cali. Se presentan las técnicas empleadas como Web Scraping para la obtención de datos y para la modelación de los precios mediante métodos estadísticos y de aprendizaje automático.

Liu et al. (2022)

En su investigación realizaron un conjunto de modelos de redes neuronales empleando los métodos BPNN (Backpropagation Neural Network), GRA (Grey Relational Analysis) y PSO (Particle Swarm Optimization). Dando lugar mediante la combinación de estas técnicas a los modelos BPNN, GRA-BPNN y PSO-GRA-BPNN. Además de esto se calcularon modelos de Regresión lineal múltiple y Bosques aleatorios con el objetivo de predecir los precios de los autos usados de una plataforma de compra-venta en China Oriental. En los modelos analíticos se incluyeron 12 variables predictoras del precio como la marca, caballos de fuerza, millaje, antigüedad, región de venta y consumo de combustible.

El conjunto de datos se extrajo de la plataforma de compra-venta de vehículos líder en China llamada "First Car Network" contando con 10.260 registros de autos usados en China Oriental comprendidos en el periodo de marzo de 2022. Los registros con datos cuantitativos faltantes se imputaron mediante la interpolación de Lagrange, mientras que se excluyeron los registros con datos faltantes en las variables cualitativas. Un supuesto que se tuvo en

cuenta es que todos los autos usados de la plataforma finalmente se negociaron con éxito al precio que determinó cada vendedor en la oferta del vehículo.

Se calcularon varios modelos mediante el modelo BPNN, optimizando el número de neuronas en la capa oculta, definiendo mediante el error cuadrático medio que diez neuronas era el valor óptimo resultando en un modelo con el esquema de capas 12-10-1. Mediante la técnica GRA se identificaron que las covariables más relevantes para predecir el precio de los vehículos son el precio de venta al público (precio del vehículo nuevo), millaje, cilindrada, tipo de transmisión y combustible, de esta forma se seleccionaron solo estas covariables en el modelo. El siguiente paso fue optimizar el modelo mediante PSO, siendo el más óptimo con el esquema de capas de 6-12-1, obteniendo un coeficiente de determinación R^2 de 0,9843. Por otra parte el modelo de regresión múltiple ajustado obtuvo un R^2 de 0,9171 y el de Bosques aleatorios un R^2 de 0,5530 (Liu y Li, 2022).

Koç y Kostak (2022)

Estudiaron los precios de autos usados en Turquía en los periodos de mayo de 2020 y marzo de 2021, examinando los posibles factores detrás del aumento de los precios de los vehículos usados durante la pandemia de COVID-19. Sus datos constaron de vehículos con solamente dos modelos de autos, el Renault Megane y el Volkswagen Passat, siendo estos dos de los diez modelos de autos más vendidos en Turquía durante el año 2020. La hoja de datos fue extraída del portal "Sahibinden", el cual es un sitio web en donde se publicitan ofertas de autos usados en Turquía. Se tomó primero una muestra de 2.416 vehículos usados en mayo de 2020 y otra de 6.527 en marzo de 2021.

Emplearon las variables año de fabricación, kilometraje, tipo de transmisión, tipo de combustible, color y ubicación del vendedor para explicar el precio de los autos mediante el método de mínimos cuadrados ordinarios. Se calculó un modelo de regresión para cada uno de los modelos de autos Megane y Passat, y para cada uno de los años 2020 y 2021 resultando en un total de cuatro modelos de regresión, los cuales estuvieron sujetos a heterocedasticidad por lo que se calcularon las varianzas robustas para corregir los sesgos en las estimaciones.

Las pruebas estadísticas demostraron que las variables color del vehículo y ubicación del vendedor no fueron significativas para explicar el precio de los autos. De acuerdo a los coeficientes de regresión se obtuvo que el precio aumenta de acuerdo al tipo de equipamiento del automóvil, si es motor diesel o si tiene transmisión automática. Por otra parte, el precio disminuye cuando aumenta el kilometraje y los años de antigüedad del vehículo. Así mismo se estudió el efecto de la ubicación del vendedor sobre el precio, resultando en que el precio disminuirá en promedio en 2.343 liras turcas si se encuentra en la ciudad de Estambul, mientras que si se encuentra en Ankara aumentará en promedio en 4.148 liras turcas (Koç y Kostak, 2021).

De Toro et al. (2017)

Su estudio se centró en desarrollar un modelo de precios hedónicos de autos usados en Argentina. Se encontraron con la dificultad de acceder a bases de datos con registros de ventas de autos usados, por lo que optaron en la utilización del método de Web Scraping. La hoja de datos fue extraída del sitio web de Mercado Libre Argentina, contando con 920 registros que incluyen las características del vehículo proporcionadas por los vendedores y fueron complementadas con más de 116 variables técnicas del vehículo.

Se ajustó un modelo de regresión sobre el logaritmo natural de los precios de los vehículos para capturar las relaciones no lineales entre la variable respuesta y las predictoras. En este modelo se incluyeron las variables kilometraje, color y antigüedad, además se analizó el efecto de 11 variables técnicas del vehículo y de esta forma se encontró el mejor modelo evitando la multicolinealidad.

El enfoque del modelo ajustado está centrado más en analizar el efecto de las variables sobre el precio de los vehículos que en la predicción de nuevas estimaciones. El estudio evidenció que las variables que más afectan negativamente a los precios de los autos usados fueron la antigüedad del vehículo, el kilometraje y el modelo. Por otra parte, las variables como tipo de combustible diesel, tipo de vehículo de alta gamma y tipo de carrocería pick-up se relacionan con un incremento en el precio del automóvil. La ubicación del vendedor fue significativa para fijar los precios, encontrando que mientras más alejados de la capital Buenos Aires, los precios de los autos usados tienden a aumentar. No se evaluaron supuestos sobre los residuos ni se analiza la bondad de ajuste del modelo. Sin embargo, se estudiaron los factores de inflación de varianza resultando en un modelo libre de multicolinealidad (de Toro et al, 2017).

Kihm y Vance (2016)

En su estudio desarrollaron dos modelos hedónicos para explicar el precio tanto de autos usados como nuevos en venta en Alemania. Se incluyeron variables de uso como el año de modelo del vehículo, millaje, si es nuevo o usado, además de incluir variables técnicas del auto como precio de venta al público (precio de venta cuando salió al mercado), velocidad máxima, cilindraje, caballos de fuerza, consumo de combustible y marca, además se incluyeron interacciones entre variables como por ejemplo, entre la variable indicadora de vehículo usado y el kilometraje.

La base de datos fue extraída en el año 2008 mediante Web Scraping del "ADAD Autokosten", el cual es un club de automovilismo alemán, esta constó de 371.082 registros, siendo 5.378 vehículos nuevos y 365.704 usados.

Evidenciaron que las variables técnicas de los autos no tienen un gran efecto sobre el precio de los autos usados y nuevos mientras no se incluya la variable de precio de venta al público, siendo evidente que los compradores de autos usados dan más prioridad al bajo consumo de combustible que los de autos nuevos. Encontraron grandes diferencias en los

precios de los autos usados dependiendo de la marca y de las versiones de cada modelo.

Las variables que afectan negativamente al precio de los autos usados son los años de antigüedad del vehículo y el kilometraje. De esta forma, por cada año de antigüedad que tenga el vehículo, su precio disminuirá en un 7,4 %, y por cada 1.000 kilómetros recorridos, el vehículo se devaluará en un 0,3 %. Los caballos de fuerza, la velocidad máxima y el cilindraje son factores que afectan positivamente al precio de los autos usados.

Las interacciones entre el tipo de uso y la antigüedad, millaje y consumo de combustible no resultaron ser significativas dentro del modelo. La calidad de ajuste del modelo fue favorable ya que este logra explicar el 94 % de la variabilidad del precio mediante las covariables, sin embargo se presenta un sobreajuste con los datos. La multicolinealidad se evaluó mediante el número de condición evidenciando que se presenta una leve multicolinealidad en el modelo (Kihm y Vance, 2016).

Gutiérrez (2022)

Su trabajo de grado se centra en la predicción y el análisis del comportamiento del precio de las ofertas de viviendas en Cali publicadas en el sitio web de OLX para el periodo prepandemia de COVID-19 en el año 2019, el periodo postpandemia en el año 2020 y a su vez, estudiar el efecto de la pandemia sobre el precio de las viviendas. Para el estudio se contaba con una hoja de datos con la oferta de viviendas para el período prepandemia de 2019. Por otra parte, los datos del periodo postpandemia de 2020 fueron obtenidos empleando Web Scraping en el sitio web de OLX, obteniendo las variables proporcionadas por los vendedores como el número de pisos, habitaciones y parqueaderos, tipo de inmueble, área construida y estrato socioeconómico. A partir de la ubicación de la vivienda se calcularon las variables del entorno como el número de homicidios y la accesibilidad a estaciones de buses MIO, centros comerciales y zonas verdes.

En el preprocesamiento de los datos se excluyeron viviendas que superarían el precio de los \$2.000 millones de pesos, viviendas que no se ubicaran en la ciudad de Cali y se eliminaron registros con datos faltantes en más de dos variables.

Se ajustaron modelos de regresión por mínimos cuadrados parciales para explicar el precio de las viviendas en el periodo prepandemia de 2019 y el periodo postpandemia de 2020 en dónde se incluyeron todas las covariables.

Se evidenció que la pandemia por COVID-19 si tuvo un efecto significativo sobre el precio tanto de casas como de apartamentos. Se determinó que para el periodo postpandemia, las personas preferían ubicaciones de las viviendas alejadas de focos de contagio como estaciones de buses, zonas verdes y centros comerciales. Por otra parte, se evidencia que el precio por metro cuadrado no se vio afectado de forma significativa por la pandemia (D. Gutierrez, 2022).

Barrios y Quinto (2016)

En este trabajo de grado se realiza la modelación del precio de la oferta de viviendas en la ciudad de Cali, Colombia, haciendo uso de variables propias y del entorno del bien. Los datos fueron recopilados mediante la técnica de Web Scraping obteniendo las viviendas ofertadas en el sitio web de Mercado Libre Colombia. Se presentaron variables de las viviendas como el área construida, número de pisos, baños y habitaciones, tipo y ubicación del inmueble. Además de esto, mediante la ubicación proporcionada por los vendedores se calcularon variables del entorno como el precio promedio de las viviendas alrededor del bien, cantidad de homicidios y la accesibilidad a estaciones de buses MIO, centros comerciales y zonas verdes.

Para el preprocesamiento de datos y validación de la información se implementaron algunos criterios de exclusión como la eliminación de viviendas que no se encontraran ubicadas en Cali, viviendas que presentaran datos faltantes en la ubicación debido a que sin esta variable no es posible calcular las covariables de entorno, viviendas que superarán los \$2.000 millones de pesos y la eliminación de variables con presencia de más del 50 % de datos faltantes.

En la modelación del precio de las viviendas se tuvo en cuenta la diferencia de precios y características de casas y apartamentos, por lo que para cada tipo de vivienda se implementaron los métodos de Modelo lineal general y generalizado bajo la distribución gamma, Árbol de regresión, Bosques aleatorios y Redes neuronales. Para el Modelo lineal generalizado se realiza selección de variables mediante la técnica de stepwise y en los modelos de aprendizaje automático se optimizaron los hiperparámetros mediante validación cruzada. Finalmente se compararon los modelos obtenidos mediante la raíz del error cuadrático medio y se eligió el modelo con mejor calidad de ajuste.

El mejor modelo para predecir el precio de la oferta de casas fue el Modelo lineal generalizado teniendo un error promedio de las estimaciones de 94,1 millones de pesos con respecto al valor real. Por otra parte, el mejor modelo para los apartamentos fueron los Bosques aleatorios con un error promedio de 94,1 millones (Barrios y Quinto, 2016).

Después de revisar los antecedentes del Trabajo de grado se destaca la relevancia de estudiar las variables que más afectan al precio de los vehículos en donde se encontró que las más importantes suelen ser la antigüedad, el kilometraje, el cilindraje, la marca y el tipo de transmisión y combustible. Las investigaciones realizadas por de Toro et al. (2017) y Kihm y Vance (2016), sirven de guía metodológica para la obtención de los datos mediante la técnica de Web Scraping y el análisis de los factores que afectan al precio de los autos usados. Por otra parte, el trabajo de grado realizado por Barrios y Quinto (2016) sienta un precedente para la selección de las metodologías para la modelación del precio de los vehículos como pueden ser el Modelo lineal generalizado, el Árbol de regresión y los Bosques aleatorios.

3 Marco teórico

En este capítulo se explican en profundidad los principales conceptos y definiciones necesarias para comprender el Trabajo de grado. En la primera parte, se abordan términos clave relacionados con el mercado de vehículos y definiciones económicas. Mientras que en la segunda parte, se explican los métodos estadísticos y de aprendizaje automático propuestos para dar respuesta a los objetivos de investigación.

3.1 Marco teórico conceptual

En el marco teórico conceptual del trabajo de grado se presentan definiciones sobre la industria automotriz y los conceptos económicos relacionados al contexto del problema de investigación del Trabajo de grado. De esta forma, es un capítulo importante para sentar una base sólida del contexto del estudio dentro de un marco teórico bien definido.

3.1.1 Industria automotriz

Es el sector de la economía que diseña, fabrica y vende vehículos, los cuales van desde motocicletas y automóviles de uso particular hasta camiones, autobuses, tractomulas, excavadoras o cualquier vehículo que se pueda mover por si mismo mediante un motor de combustión interna o eléctrico. A parte de la fabricación y venta, otras de las actividades económicas del mercado automotriz son la fabricación de repuestos, el servicio de talleres de reparación y diagnostico automotor, financiamiento y pólizas de seguros.

3.1.2 Oferta

El concepto de oferta en economía se explica como la cantidad de un bien o servicio que los productores están dispuestos a vender en un intervalo de tiempo específico, en donde cada productor puede fijar un precio al que está dispuesto a vender (Mankiw, 2012).

3.1.3 Demanda

La demanda en economía es la cantidad de un bien o servicio que es adquirida por los consumidores a diferentes precios en un determinado intervalo de tiempo (Mankiw, 2012).

3.1.4 Depreciación

la depreciación de un bien tangible o intangible es la disminución del valor de este con respecto al tiempo provocado por el desgaste, uso u otros factores, como la aparición de nuevos bienes que provoquen la obsolescencia (Brigham y Houston, 2016).

3.1.5 Web Scraping

El Web Scraping o raspado web es una rutina computacional utilizada para extraer datos de sitios web de internet. Esta rutina permite automatizar el proceso de extracción de datos de cualquier página web mediante un algoritmo llamado rastreador web.

Los elementos en una página web están estructurados de tal forma que se pueden identificar mediante una clase o ID, los cuales están programados en lenguajes de marcado como HTML, de esta forma el rastreador identifica estos elementos y extrae texto, imágenes, audio o hipervínculos contenidos en estos, ofreciendo la posibilidad de que se puedan almacenar de forma estructurada en hojas de cálculo o archivos de texto plano. El web scraping es muy utilizado en la minería de datos ya que permite recolectar una gran cantidad de registros de forma automatizada.

El uso del Web Scraping es legal siempre que se cumplan las normas, políticas y términos de uso del sitio web al que se está accediendo para obtener los datos. Es importante no violar los derechos de autor y propiedad intelectual, evitando lucrarse económicamente de forma directa de los datos extraídos. De esta forma es importante cumplir con la Ley de Privacidad del Consumidor de California (CCPA) en Estados Unidos (California State Legislature, 2018), y con el Reglamento General de Protección de Datos (RGPD) en Europa (Parlamento Europeo y Consejo de la Unión Europea, 2016), en los cuales se establecen las normas de privacidad y la protección de datos personales.

3.2 Marco teórico estadístico

En el marco teórico estadístico del Trabajo de grado se presenta el contexto teórico y técnico requerido para entender el problema de investigación. Se explican a profundidad las definiciones, métodos de preprocesamiento de datos, modelos estadísticos y de aprendizaje automático empleados para la modelación del precio de los automóviles usados.

3.2.1 Preprocesamiento de datos

El preprocesamiento de datos es el paso inicial en el análisis de datos en donde se preparan, transforman y depuran los datos recién extraídos para que sean utilizados correctamente en procesos posteriores como el análisis exploratorio y la modelación. A continuación se muestran algunas de las técnicas, métodos y rutinas más utilizadas en el preprocesamiento de datos, y que serán empleadas en el desarrollo del Trabajo de grado.

Minería de datos

La minería de datos es la generación de conocimiento encontrando patrones y relaciones relevantes en grandes conjuntos de datos. Los principales procesos que se realizan en la minería de datos son: la limpieza de datos en donde se depuran inconsistencias y se remueve el ruido de los datos, dando lugar a la validación de la información para su correcta implementación en el análisis de los datos, la integración en donde una vez preprocesadas las distintas fuentes de datos, estas se combinan en un solo conjunto permitiendo cruzar características y variables entre distintas tablas de datos, la selección en la cual se eligen los datos y variables consideradas más relevantes a ser analizados para la toma de decisiones en el análisis exploratorio y en los modelos analíticos, la transformación en la que se transforman y consolidan los datos permitiendo reducir la complejidad de estos, y finalmente la evaluación de patrones en donde se usan técnicas estadísticas, de aprendizaje profundo y automático para modelar los datos con la finalidad de comprender el efecto de los factores asociados, clasificando o prediciendo el comportamiento de datos no observados o las tendencias a través del tiempo, dando lugar a la toma de decisiones mediante la información obtenida (Kambe, Ha, y Pe, 2011).

Un ejemplo de minería de datos es analizar las ventas de una compañía, identificando el comportamiento de compra de los clientes y segmentándolos de acuerdo a sus características demográficas y de consumo, para poder realizar de forma eficiente las campañas de marketing y lograr una optimización en el presupuesto definido para estas.

Identificación de datos atípicos

La identificación de datos atípicos es un paso importante en el preprocesamiento de datos en donde se detectan valores y relaciones atípicas de los datos. Los datos atípicos son identificados debido a que estos se desvían de forma significativa del valor esperado de los datos en la variable o conjunto de variables que se estén analizando. Los datos atípicos pueden presentarse debido a varios factores como errores en la herramienta de medición, malas digitaciones o por la variabilidad natural de la variable en cuestión. Es importante implementar este paso en el preprocesamiento para conocer el comportamiento de los datos atípicos y tomar decisiones sobre como tratarlos, esto debido a que suelen tener bastante influencia en el análisis exploratorio y en los modelos predictivos (Barnett y Lewis, 1994).

Criterio de Tukey

El criterio de Tukey permite detectar valores atípicos en una variable teniendo en cuenta los cuartiles. Para calcular los límites superior e inferior que delimitan si un dato es atípico se utiliza la siguiente fórmula:

$$\begin{aligned} \text{Limite superior} &: Q_3 + RIC * D \\ \text{Limite inferior} &: Q_1 - RIC * D \end{aligned} \tag{3.1}$$

Donde Q_1 es el primer cuartil, Q_3 es el tercer cuartil, RIC es el rango intercuartílico definido por $Q_3 - Q_1$, y D es un escalar. Para evaluar los datos atípicos leves, D toma el valor de 1,5 y para evaluar datos atípicos extremos este toma el valor de 3. Los datos que se encuentren por

fuera de estos límites se consideran datos atípicos leves y extremos respectivamente (Tukey, 1977).

Valor atípico local

El valor atípico local o LOF (Local Outlier Factor) por sus siglas en inglés, es un método muy utilizado para detectar la presencia de datos atípicos mediante un análisis multivariado. Se basa en calcular la densidad local de un punto con respecto a k vecinos más cercanos mediante el cálculo de las distancias entre estos, en donde se suele utilizar la distancia euclidiana y Manhattan.

Así, dados los vectores A y B , que representan las características de estas observaciones, sus distancias en el espacio p -dimensional están dadas por:

$$\text{distancia euclidiana} : \sqrt{\sum_{i=1}^p (A_i - B_i)^2} \quad (3.2)$$

$$\text{distancia Manhattan} : \sum_{i=1}^p |A_i - B_i| \quad (3.3)$$

Mediante las distancias euclidianas se puede calcular la *densidad de alcanzabilidad* de un punto A dada por:

$$\text{lr}d_k(A) : \frac{k}{\sum_{B \in N_k(A)} \text{reachability distance}_k(A, B)} \quad (3.4)$$

En donde $\text{reachability distance}_k(A, B)$ es la distancia de alcanzabilidad entre A y B , que se define como el máximo entre la distancia de A y B , y la distancia euclidiana si B es uno de los k vecinos más cercanos de A .

Por otra parte $N_k(A)$ es el conjunto de los k vecinos más cercanos de A .

De esta forma, se puede calcular el índice del valor atípico local de A de la siguiente manera:

$$\text{LOF}_k(A) : \frac{\sum_{B \in N_k(A)} \frac{\text{lr}d_k(B)}{\text{lr}d_k(A)}}{k} \quad (3.5)$$

Mediante los puntajes de cada observación se pueden identificar regiones con densidades mayores que las demás, de esta forma las observaciones más alejadas de estas regiones pueden ser consideradas atípicas (Breunig, Kriegel, Ng, y Sander, 2000). Esto se puede observar en la Figura 3.1, en donde las observaciones A , B , C y D conforman una región de alta densidad y en donde E se puede considerar una observación atípica dado que se encuentra en un área baja de densidad en observaciones.

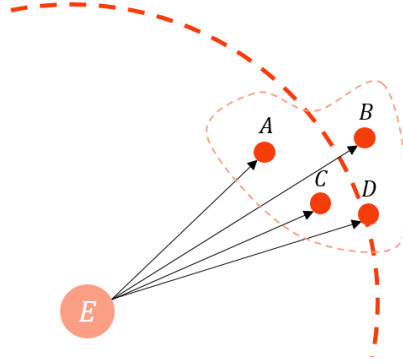


Figura 3.1: Densidad del punto E respecto a sus vecinos

El número k de vecinos a considerar es un parámetro que se debe ingresar al algoritmo, en donde el valor por defecto suele ser de 20. Una vez calculados los puntajes se puede fijar un percentil de orden superior o usar algún otro criterio establecido a priori para establecer un umbral en el cual las observaciones con puntajes que lo sobrepasen sean consideradas atípicas.

3.2.2 Modelo lineal general

Un Modelo lineal general es un modelo matemático que tiene como finalidad aproximar la relación lineal entre una variable dependiente y una o varias variables independientes, por lo que es útil para predecir el comportamiento de la variable dependiente siempre que existan relaciones significativas con las variables independientes.

Según Montgomery y Peck (1994), se pueden estimar los coeficientes de regresión β mediante la técnica de mínimos cuadrados ordinarios (MCO), de esta forma el Modelo lineal general está dado por la siguiente ecuación:

$$Y = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p + \epsilon = X\beta + \epsilon \quad (3.6)$$

En donde Y es la variable dependiente, X_p es la p -ésima variable independiente, β_p es el coeficiente de regresión asociado a la p -ésima variable independiente y ϵ es el error aleatorio de predicción.

El estimador $\hat{\beta}$ de los coeficientes de regresión β mediante mínimos cuadrados ordinarios está dado por:

$$\hat{\beta} = (X'X)^{-1}X'Y \quad (3.7)$$

En donde X es la matriz conformada por las variables explicativas y Y es el vector de la variable respuesta.

Evaluación de supuestos

El Modelo lineal general asume los siguientes supuestos sobre los errores:

- $E(\epsilon_i) = 0$

- $V(\epsilon_i) = \sigma^2$
- $cov(\epsilon_i, \epsilon_j) = 0$, para todo $i \neq j$
- $\epsilon_i \sim Normal(0, \sigma^2)$

Es decir, los errores deben tener media cero, tener igualdad de varianzas, ser independientes entre ellos y tener distribución normal (Behar, 2003). Para que las estimaciones del modelo sean insesgadas debe cumplirse el supuesto de media cero de los errores

Estos supuestos se pueden evaluar mediante pruebas gráficas y estadísticas, haciendo uso de los residuos del modelo ajustado mediante los datos, los cuales son una estimación de los errores del modelo teórico que están dados por $\hat{\epsilon}_i = y_i - \hat{y}_i$, donde y_i es la i -ésima observación y $\hat{y}_i$ es el valor ajustado de la i -ésima observación.

El supuesto de media cero de los errores ($E(\epsilon_i) = 0$) siempre se cumple debido a las propiedades de los estimadores mediante mínimos cuadrados ordinarios.

El supuesto de homogeneidad de varianzas de los errores ($V(\epsilon_i) = \sigma^2$) puede evaluarse mediante un gráfico de dispersión de los valores ajustados frente a los residuos, en donde se espera que la variabilidad de las distancias entre los puntos sea homogénea. Por otra parte, también puede hacerse uso de la prueba de Breusch-Pagan en donde se tiene como hipótesis nula que "existe homogeneidad de varianza en los residuos".

El supuesto de independencia de los errores ($cov(\epsilon_i, \epsilon_j) = 0$) es importante evaluarlo cuando las observaciones son obtenidas a lo largo del tiempo. Este puede evaluarse mediante la prueba de Durbin-Watson en donde su hipótesis nula es que "los residuos son independientes".

El supuesto de normalidad ($\epsilon_i \sim Normal(0, \sigma^2)$), se evalúa mediante la distribución de los residuos en donde se espera mediante un histograma que esta sea aproximadamente normal. Además se puede evaluar mediante las prueba de Shapiro-Wilk en donde la hipótesis nula es que "los residuos tienen distribución normal".

Por otra parte, también es importante analizar la multicolinealidad entre las variables predictoras, es decir, que no exista información dada por una variable que ya pueda ser explicada mejor por otras. La presencia de multicolinealidad en el modelo puede provocar que las estimaciones de los coeficientes de regresión sean poco confiables o que estas no se puedan calcular debido a que no existe inversa de la matriz $(X'X)^{-1}$.

La multicolinealidad se puede evaluar mediante varios métodos como el análisis de la matriz de correlaciones, en donde si existe multicolinealidad, se espera que haya un alto coeficiente de correlación entre las covariables.

Otro método empleado es mediante el cálculo de los factores de inflación de varianza o VIF, por sus siglas en inglés, en donde un factor entre 5 y 10 indica multicolinealidad leve, mientras que un factor mayor a 10 indica alta multicolinealidad.

Métricas de bondad de ajuste

Para evaluar el desempeño y calidad de un Modelo de regresión lineal se utilizan las métricas de bondad de ajuste, las cuales permiten evidenciar la variabilidad explicada por el modelo, o también, evaluar el error medio de las estimaciones con respecto a los valores observados, lo da lugar a realizar comparaciones entre modelos y evaluar cual es el que tiene mejor desempeño. A continuación se muestran las métricas de bondad de ajuste más importantes:

Coefficiente de determinación

De acuerdo a Walpole (2012), el coeficiente de determinación es una medida estadística que expresa la proporción de la variabilidad de una variable respuesta que es explicada por el conjunto de las variables predictoras, de esta forma es una medida que permite evaluar la calidad de ajuste de un modelo de regresión. Su formula matemática está dada por:

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = 1 - \frac{SCE}{SCT} \quad (3.8)$$

Donde y_i es la i -ésima observación de la variable respuesta, $\hat{y}_i$ es la i -ésima predicción ajustada mediante el modelo y $\bar{y}$ es la media de la variable respuesta. Por otra parte SCE es la suma de cuadrados del error o de la regresión y SCT es la suma de cuadrados total.

El coeficiente de determinación es una medida que toma valores entre 0 y 1, en donde un coeficiente igual a cero indica que el modelo ajustado no explica ninguna variabilidad de la variable respuesta mediante las variables predictoras, por otra parte un coeficiente igual a 1 indica que el modelo explica el total de la variabilidad de la variable respuesta, esto se da cuando la suma de cuadrados de los residuos es igual a cero. El coeficiente de determinación es influido por el número de variables regresoras, siendo que cuando éstas aumentan el coeficiente también lo hace, por lo que en ocasiones este aumento no es debido necesariamente a la disminución del error de predicción.

Coefficiente de determinación ajustado

El coeficiente de determinación ajustado es una medida ajustada del coeficiente de determinación en el cual se tiene en cuenta la influencia del número de variables regresoras en el modelo. El coeficiente de determinación ajustado penaliza la inclusión de variables predictoras que no aporten a la disminución de los errores de predicción. Su formula está dada por:

$$R_{adj}^2 = 1 - \left(1 - \frac{SCE}{SCT}\right) \left(\frac{n-1}{n-p-1}\right) = 1 - (1 - R^2) \left(\frac{n-1}{n-p-1}\right) \quad (3.9)$$

Donde R^2 es el coeficiente de determinación, n es el número de observaciones y p es el número de coeficientes de regresión, es decir, el número de covariables más el intercepto.

Error cuadrático medio

El error cuadrático medio o ECM, es una medida que se calcula sobre un estimador de regresión la cual mide el error medio de predicción al cuadrado. La formula del error cuadrático medio está dada por:

$$ECM = \sum_{i=1}^n \frac{(y_i - \hat{y}_i)^2}{n} \quad (3.10)$$

En donde y_i es el i -ésimo valor observado, $\hat{y}_i$ es el i -ésimo valor predicho por el estimador de y_i y n es el número de observaciones utilizadas para ajustar el modelo (Brown y Lehmann, 1984).

En el caso de la regresión, el ECM calcula el promedio de los cuadrados de las diferencias entre los valores observados y los valores ajustados por el modelo. De esta forma, mientras más pequeño sea el ECM, mejor será la calidad y precisión del modelo. Es una medida útil para comparar la calidad de predicción de varios modelos de regresión.

Por otra parte, la raíz del error cuadrático medio o RECM es medida es muy útil ya que esta expresada en la misma escala de la variable respuesta por lo que se facilita su interpretabilidad, permitiendo conocer la diferencia promedio de la estimación con respecto al valor real.

Criterio de información de Akaike

De acuerdo a Akaike (1974), el criterio de información de Akaike o AIC por sus siglas en inglés, es una medida estadística que permite conocer el desempeño relativo de un modelo dado su rendimiento sobre los valores ajustados y la complejidad del modelo. El AIC es utilizado para comparar la calidad de ajuste de modelos estadísticos de regresión como el Modelo lineal general y generalizado, de series temporales como ARIMA o de aprendizaje automático como los Árboles de decisión. Su fórmula está dada por:

$$AIC = 2K - 2Ln(L) \quad (3.11)$$

En donde k es el número de parámetros y L es el valor máximo que toma la función de verosimilitud del modelo de regresión.

El AIC no permite concluir sobre el rendimiento y calidad de un cierto modelo, más bien se puede emplear para realizar la comparación de múltiples modelos teniendo en cuenta la precisión, la complejidad y el sobreajuste, en donde el mejor modelo será el que posea el menor AIC.

3.2.3 Modelo lineal generalizado

Los Modelos lineales generalizados son una extensión del Modelo lineal general que permiten modelar las relaciones entre la variable respuesta y las variables predictoras, en donde no se supone una distribución normal sobre los errores del modelo de acuerdo a McCullagh y Nelder (1986). En un Modelo lineal generalizado se utiliza una función de enlace que permite

linealizar la relación de la media de la variable respuesta condicionada a las variables regresoras, en donde esta distribución condicionada debe pertenecer a la familia exponencial. A continuación se muestra la explicación de lo que es una distribución de probabilidad:

Distribución de probabilidad

Una distribución de probabilidad es una función matemática que asocia a un evento o suceso un número real.

Una variable aleatoria puede ser discreta o continua, en el caso discreto, la variable toma valores enteros, por ejemplo, el número de personas que llegan a un cajero automático. Para el caso continuo, la variable aleatoria toma valores en los números reales, un ejemplo de esta es la temperatura diaria tomada por un termómetro.

La distribución de probabilidad es la función asociada a la variable aleatoria que dará la probabilidad definida entre 0 y 1, de que ocurra un evento del espacio muestral (Evans y Rosenthal, 2003).

Estimación de parámetros del Modelo lineal generalizado

Para estimar los parámetros del Modelo lineal generalizado se suelen utilizar métodos numéricos como Newton-Raphson en donde se minimiza la función de verosimilitud, de forma más intuitiva, se maximiza la probabilidad de que el modelo bajo la distribución genere los datos observados, obteniendo las estimaciones de los coeficientes de regresión.

El Modelo lineal generalizado cuenta con tres definiciones importantes las cuales son:

- La función de distribución asociada a la variable respuesta debe pertenecer a la familia exponencial.
- El predictor lineal está dado por $\eta = X\beta$, en donde X es la matriz de diseño de las variables predictoras y β es el vector de los coeficientes de regresión.
- La función de enlace g que relaciona la función de enlace con el predictor lineal, siendo $E(Y) = \mu = g^{-1}(\eta)$.

En este caso no se cuenta con indicadores de bondad de ajuste como en el Modelo lineal general que se puedan calcular directamente con los coeficientes de regresión, como puede ser el coeficiente de determinación, sin embargo se puede calcular la devianza, la cuál está dada por:

$$D = -2[\ln L(\text{modelo ajustado}) - \ln L(\text{modelo nulo})] \quad (3.12)$$

En donde $\ln L(\text{modelo ajustado})$ es el valor obtenido de la función de log-verosimilitud haciendo uso de los datos observados del modelo ajustado y $\ln L(\text{modelo nulo})$ es el valor obtenido de la función de log-verosimilitud del modelo sin ninguna variable predictora.

La devianza se basa en la diferencia entre los valores ajustados por el modelo y los observados, por lo que puede utilizarse para evaluar la bondad de ajuste, siendo un modelo mejor

en comparación a otro si su devianza es menor.

Además de la devianza, también puede ser calculado el criterio de información de Akaike mediante la formulación dada en la ecuación (3.11), en donde su uso y definición es igual que en el Modelo lineal general. También puede emplearse el error cuadrático medio y su raíz para realizar la comparación de modelos.

3.2.4 Modelo lineal generalizado gamma

El Modelo lineal generalizado gamma puede emplearse cuando la variable respuesta sigue una distribución gamma. Comúnmente la variable respuesta tiende a tener asimetría positiva y debe tomar valores positivos, como puede ser por ejemplo el tiempo que transcurre hasta que muere un paciente. A continuación se muestra la definición de la distribución de probabilidad gamma:

Distribución gamma

La distribución gamma es una distribución de probabilidad continua con parámetro de forma α , de escala λ , además de poder contar con un tercer parámetro de umbral el cual define el valor en el que se mueve la asimetría de la distribución.

Sea $\mathbf{Y}$ una variable aleatoria con $y > 0$ la cual distribuye $gamma(\alpha, \beta)$ con $\alpha > 0$ y $\beta > 0$, esta tendrá la función de densidad de probabilidad:

$$f_y(y|\alpha, \beta) = \frac{1}{\Gamma(\alpha)\beta^\alpha} y^{\alpha-1} e^{-\frac{y}{\beta}} \quad (3.13)$$

Con $E(x) = \alpha\beta$ y $Var(x) = \alpha\beta^2$.

La distribución gamma es útil para modelar el comportamiento de variables aleatorias continuas que cuenten con una asimetría positiva y que tome valores mayores a cero, como por ejemplo el tiempo de vida de una colonia de bacterias en una placa de Petri (Reinsel, Hogg, y Craig, 1980).

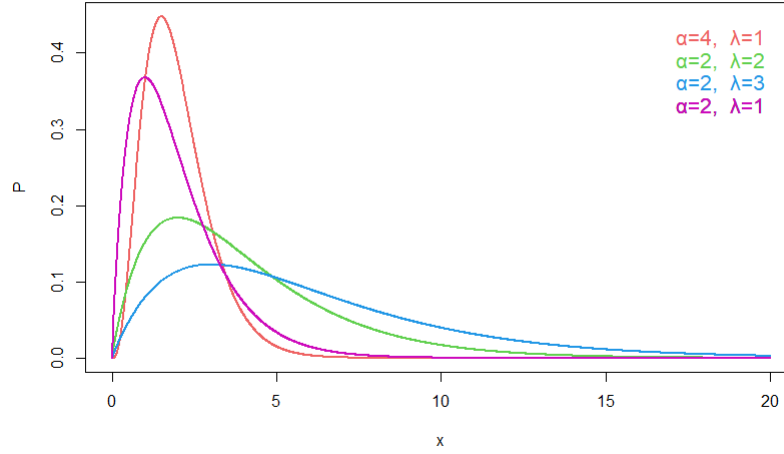


Figura 3.2: Distribución de densidad de probabilidad gamma

Formulación del Modelo lineal generalizado gamma

Sea la variable respuesta con distribución gamma con media $\mu = \alpha\beta$ y varianza $\sigma^2 = \alpha\beta^2 = \mu\beta$, en donde se puede evidenciar que la varianza aumenta con la media de la variable.

Con la transformación $\beta = \mu/\alpha$ aplicada en la ecuación (3.14), se puede reparametrizar la función de densidad de la variable respuesta de la siguiente forma:

$$f_y(y|\alpha, \mu) = \frac{1}{\Gamma(\alpha)} \left(\frac{\alpha}{\mu}\right)^\alpha y^{\alpha-1} e^{-\frac{y\alpha}{\mu}} \quad (3.14)$$

Con parámetros α y μ mayores a cero. Donde $E(Y) = \mu$ y $Var(Y) = \mu^2/\alpha$. De esta forma, la varianza es una función de la media, por lo que el modelo es heterocedastico por naturaleza.

Las funciones de enlace que se utilizan son comúnmente la logarítmica $\eta = \log(\mu)$, la identidad $\eta = \mu$ y la inversa $\eta = 1/\mu$, las cuales permiten linealizar las relaciones entre la variable respuesta y las predictoras

En el Modelo lineal generalizado, al igual que en el Modelo lineal general, los residuales, que son una estimación de los errores del modelo, están dados por $\hat{\epsilon}_i = y_i - \hat{y}_i$, donde y_i es la i -ésima observación y $\hat{y}_i$ es el valor ajustado por el modelo de la i -ésima observación.

Los supuestos de este modelo son:

- La variable respuesta condicionada a las covariables debe tener distribución gamma.
- Las observaciones deben ser independientes.

Es importante evaluar y gestionar la presencia de multicolinealidad del modelo, ya que su presencia puede afectar la interpretación de los coeficientes.

Los parámetros del modelo se calculan maximizando la función de verosimilitud. Al ser un Modelo lineal generalizado, los parámetros se obtienen mediante métodos iterativos como Newton-Rapshon, en el cual se utilizan derivadas de la función de verosimilitud para actualizar los parámetros en cada iteración. En este método se calcula la matriz hessiana que contiene las segundas derivadas respecto a los parámetros del modelo, y la matriz de gradientes, que contiene las primeras derivadas. A partir de estas, se ajustan los parámetros hasta que la función de verosimilitud converja a un máximo. (McCullagh y Nelder, 1986).

Bondad de ajuste y comparación de modelos

La calidad de ajuste de un Modelo lineal generalizado gamma se evalúa mediante distintas métricas de bondad de ajuste como:

- **Devianza:** esta medida evalúa la diferencia entre los valores ajustados por el modelo y los observados. No permite analizar el desempeño relativo de un modelo en específico, pero da lugar a realizar la comparación de la calidad de distintos modelos, en donde una menor devianza indica un mejor modelo.
- **AIC:** al igual que las anteriores métricas, solo permite la comparación de distintos modelos.

Para medir la capacidad predictiva del modelo, puede emplearse el error cuadrático medio que permite conocer el error medio de predicción de las estimaciones con respecto a los valores observados. Para mejorar la interpretabilidad, suele utilizarse la raíz del error cuadrático medio ya que evalúa el error en la misma escala de la variable respuesta.

3.2.5 Modelos de aprendizaje automático

Un modelo de aprendizaje automático es un algoritmo que permite encontrar patrones en los datos, que puedan ser utilizados para predecir o clasificar una variable respuesta mediante el conjunto de variables predictoras.

Al igual que los modelos de regresión estadísticos, los modelos de aprendizaje automático permiten predecir una variable mediante una o más covariables predictoras, que pueden ser continuas o categóricas. Los modelos más populares de aprendizaje automático son los Árboles de regresión, Bosques aleatorios, XGBoost, Máquina de soporte vectorial y las Redes neuronales (Hastie y Tibshirani, 2008).

3.2.6 Árbol de decisión

Los Árboles de decisión son un método de aprendizaje supervisado que tiene como objetivo predecir o clasificar valores de la variable respuesta mediante las variables predictoras relacionadas. Un Árbol de decisión empieza con un nodo inicial en donde se selecciona una variable o característica específica de las observaciones, la cual suele ser la que tiene más peso para predecir los valores de la variable respuesta. Seguido de esto se presentan los nodos

intermedios de división en donde de acuerdo a las reglas de decisión, se irá dividiendo rectangularmente el conjunto de predictores original X y se asocia a cada espacio la predicción más probable de la variable respuesta Y . Por último, al dejar de dividir los nodos de acuerdo al criterio de parada establecido a priori, se da lugar a los nodos terminales, los cuales son empleados para la predicción o clasificación, en donde para el caso de la regresión, se calcula la media de la variable respuesta de las observaciones pertenecientes al nodo terminal, mientras que para la clasificación se elige la clase mayoritaria de las observaciones.

La idea principal de los Árboles de decisión se basa en seleccionar la mejor variable y punto de corte para dividir las observaciones en cada paso, dando lugar a nuevos nodos. Para el caso de la regresión, se emplea el error cuadrático medio para realizar las divisiones.

División de nodos

Para realizar la división de un nodo en dos nodos hijos, se emplea el error cuadrático medio para calcular la varianza de un nodo padre y de los dos nodos hijos resultantes de su división, de esta forma para la creación de una nueva división, se elige la variable predictora X_j y el punto de corte t_k que maximicen la reducción de varianza entre el nodo padre y los hijos.

La varianza de un conjunto de observaciones Q está dada por:

$$Var(Q) = \frac{\sum_{i \in N_Q} (y_i - \bar{y})^2}{N_Q} \quad (3.15)$$

En donde N_Q es el número de observaciones pertenecientes al subconjunto Q .

La reducción de varianza entre el nodo padre S_i y los dos nodos hijos $S_{i,1}$ y $S_{i,2}$ respectivamente, está dada por:

$$\Delta Var_k = Var(S_i) - \left(\frac{N_{S_{i,1}}}{N_{S_i}} Var(S_{i,1}) + \frac{N_{S_{i,2}}}{N_{S_i}} Var(S_{i,2}) \right) \quad (3.16)$$

En donde N_{S_i} , $N_{S_{i,1}}$ y $N_{S_{i,2}}$ es el número de observaciones pertenecientes al nodo padre S_i y a los dos nodos hijos $S_{i,1}$ y $S_{i,2}$ respectivamente.

De esta forma, dada una variable predictora X y cada valor dado del punto de corte t_k , se dividen las observaciones pertenecientes al nodo padre en dos subconjuntos $S_{i,1}$ y $S_{i,2}$, en donde después se calcula la varianza de cada uno. Luego de esto, se procede a calcular la reducción de varianza de la división. Este proceso se muestra en la Figura 3.4:

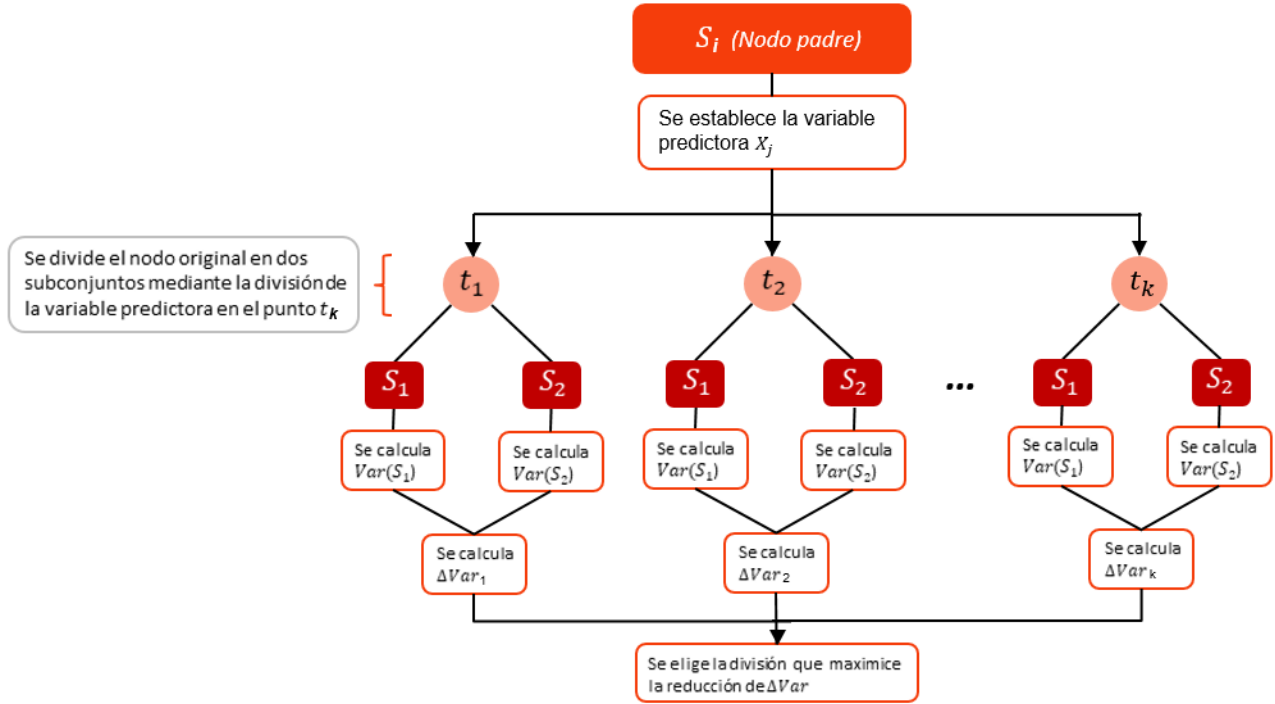


Figura 3.3: Esquema de la división de un nodo en el Árbol de regresión

El cálculo se repite para cada variable predictora X_j , donde finalmente se elige el valor de t y la variable que logran maximizar la reducción de varianza ΔVar , dando lugar a una nueva división del nodo padre (Hastie y Tibshirani, 2008).

Criterio de parada

La creación de nuevos nodos debe estar limitada a un criterio de parada, el cual es un hiperparámetro que debe optimizarse, este paso se llama "poda" del árbol y permite evitar el sobreajuste del modelo y que no se creen nuevos nodos indefinidamente. Para el árbol de decisión existen varios criterios de parada como son:

- **Valor mínimo en la reducción del error:** controla el valor mínimo de la reducción del error de predicción (ecuación 3.16) que se obtiene al dividir un nodo. Mientras que la reducción del error en la división sea mayor al valor especificado, se irán creando nuevos nodos.
- **Profundidad máxima:** controla el número máximo de niveles que puede tener el árbol, evitando el sobreajuste y reduciendo la complejidad del modelo.

Además del criterio de parada, también existen otros hiperparámetros que permiten tener más control en el entrenamiento del árbol, como el número mínimo de observaciones para dividir un nodo, el número mínimo de observaciones en cada hoja del árbol y el número máximo de covariables que se deben tener en cuenta para dividir un nodo.

El Árbol de regresión se puede representar de forma gráfica como se muestra en la Figura 3.4. En cada nodo del árbol se ha dividido anteriormente la variable independiente en dos

grupos. De acuerdo a los valores que tomen las observaciones en sus variables predictoras se irán siguiendo ciertas ramas y hojas hasta llegar al final del árbol, llegando a un nodo terminal.

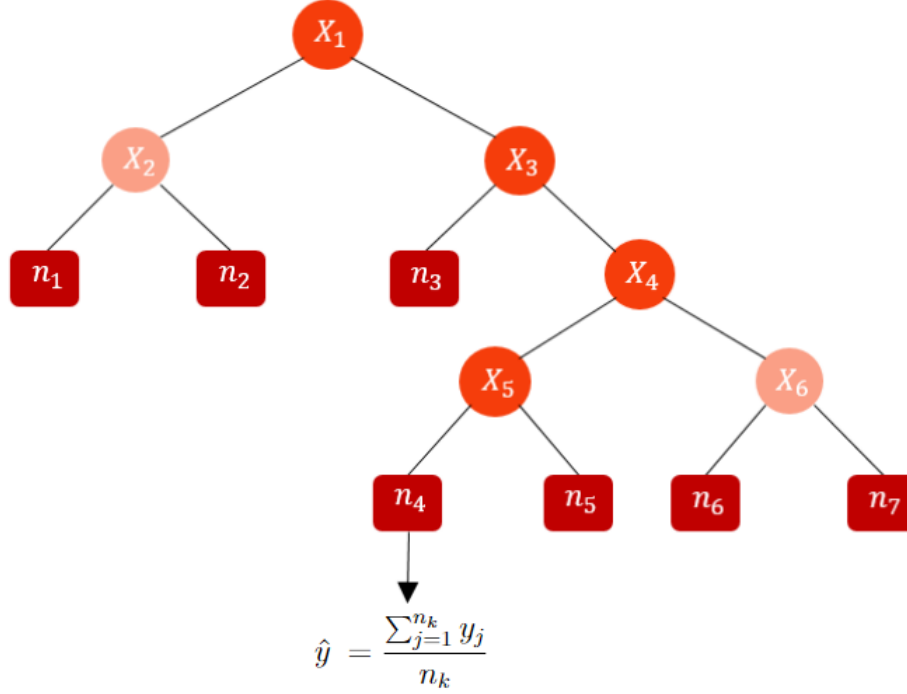


Figura 3.4: Esquema del Árbol de regresión

Para obtener la predicción de una observación se calcula el promedio de la variable respuesta de las observaciones pertenecientes al nodo, el cual está dado por:

$$\hat{y}_i = \frac{\sum_{j=1}^{n_k} y_j}{n_k} \quad (3.17)$$

En donde y_i son los valores que toma la variable respuesta para el conjunto de observaciones pertenecientes al k -ésimo nodo terminal y n_k es el número de observaciones pertenecientes a este. De esta forma, la predicción es el promedio de la variable respuesta del grupo de individuos pertenecientes al nodo.

Para evaluar la calidad de ajuste de un árbol de regresión, se emplean las mismas métricas de bondad de ajuste que suelen ser usadas en los modelos de regresión lineal, como el error cuadrático medio y el coeficiente de determinación ajustado.

3.2.7 Validación cruzada

La validación cruzada es un método utilizado para evaluar el desempeño de un modelo de clasificación o de regresión. Se basa en dividir el conjunto de datos en subconjuntos o 'folds' como se puede observar en la Figura 3.5, de esta forma en cada iteración se toma un subconjunto diferente para evaluar el modelo y el resto para ajustarlo (Refaeilzadeh, Tang, y Liu, 2009).

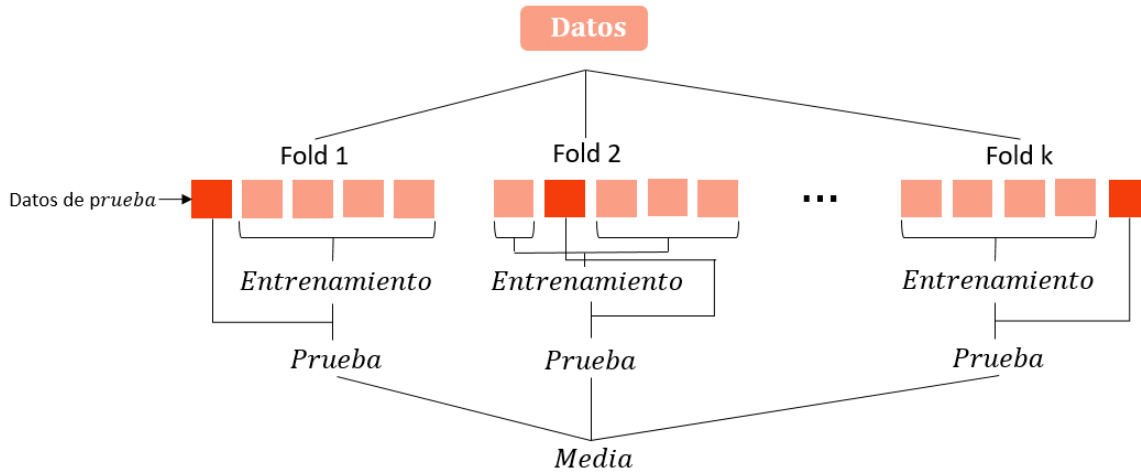


Figura 3.5: Esquema de validación cruzada

La utilidad de la validación cruzada radica en que permite evaluar el rendimiento del modelo de una manera más robusta en datos no observados, además de analizar el sobreajuste sobre los datos de entrenamiento.

La validación cruzada se emplea en la optimización de los hiperparámetros de los modelos de aprendizaje automático. Mediante la técnica de búsqueda en cuadrícula se entrena cada modelo con las diferentes combinaciones de hiperparámetros, y se evalúa el desempeño del modelo con métricas como el error cuadrático medio de predicción, en donde en cada iteración se utiliza un subconjunto diferente de datos de entrenamiento y prueba.

3.2.8 Bagging

El bagging o agregación de bootstrap, es una técnica de aprendizaje automático en donde su principal objetivo es disminuir la varianza de las estimaciones de nuevas predicciones y evitar el sobreajuste de un modelo de clasificación o regresión.

El bagging se basa en realizar remuestreo bootstrap sobre un conjunto de datos de entrenamiento, es decir se toman m muestras aleatorias con reemplazo de igual tamaño con las cuales se procede a entrenar un modelo con cada una de estas muestras en donde al final se tendrán m estimaciones de una sola observación. Para obtener la estimación final de la variable respuesta se promedian los valores ajustados de estos modelos para el caso de la

regresión, mientras que para la clasificación se toma la categoría mayoritaria predicha por los modelos (Breiman, 1996). En la Figura 3.6 se ilustra el esquema de este procedimiento:

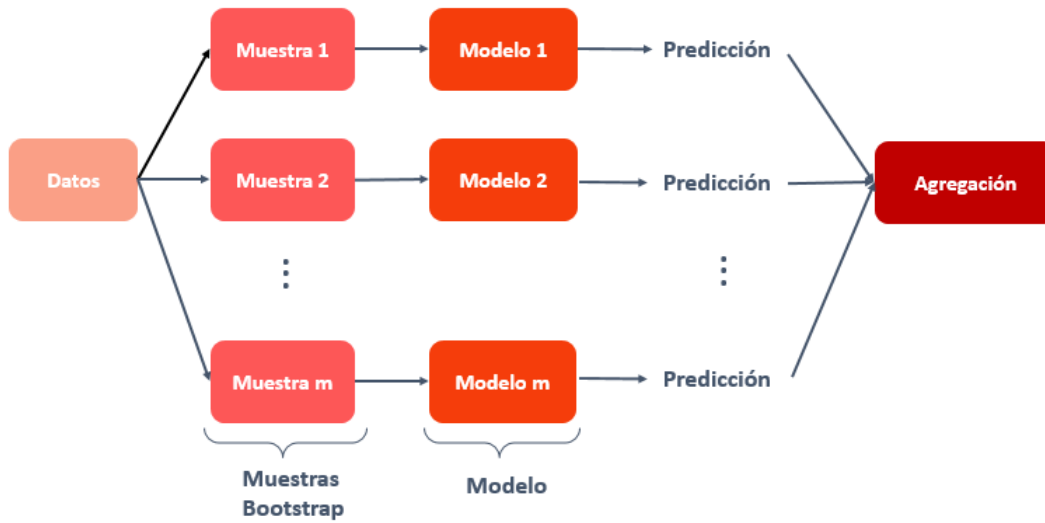


Figura 3.6: Esquema de bagging

Esta técnica puede aplicarse a cualquier modelo de aprendizaje automático siendo lo más común que se aplique a los Árboles de decisión, dando lugar al modelo más robusto de Bosques aleatorios, ayudando a mejorar la confiabilidad y la precisión de las estimaciones ante conjuntos de datos no observados o con presencia de ruido.

3.2.9 Boosting

El boosting es una técnica que se emplea en modelos de aprendizaje automático como XGBoost, la cual consiste en entrenar de forma iterativa modelos simples de aprendizaje automático como los Árboles de decisión, en donde al final se obtendrá un modelo mucho más robusto. El primer algoritmo de boosting fue introducido por Freund y Shapire (1995), el cual fue bautizado como "AdaBoost", abreviatura de "Adaptative Boosting", y fue aplicado sobre arboles de decisión.

El boosting se centra en utilizar el mismo conjunto de datos de entrenamiento para cada iteración, es decir, se entrena un modelo el cual puede ser un Árbol de decisión y se calculan los pesos de las observaciones de acuerdo a los errores de predicción o las malas clasificaciones de la iteración, en el caso de la regresión y clasificación respectivamente, de esta forma las observaciones con mayor error de predicción tendrán una mayor ponderación. Así, en cada nueva iteración se calculará un modelo que tendrá como objetivo corregir los errores cometidos por el modelo anterior, como se puede observar en la Figura 3.7, en donde se presenta el esquema de boosting con m modelos o iteraciones.

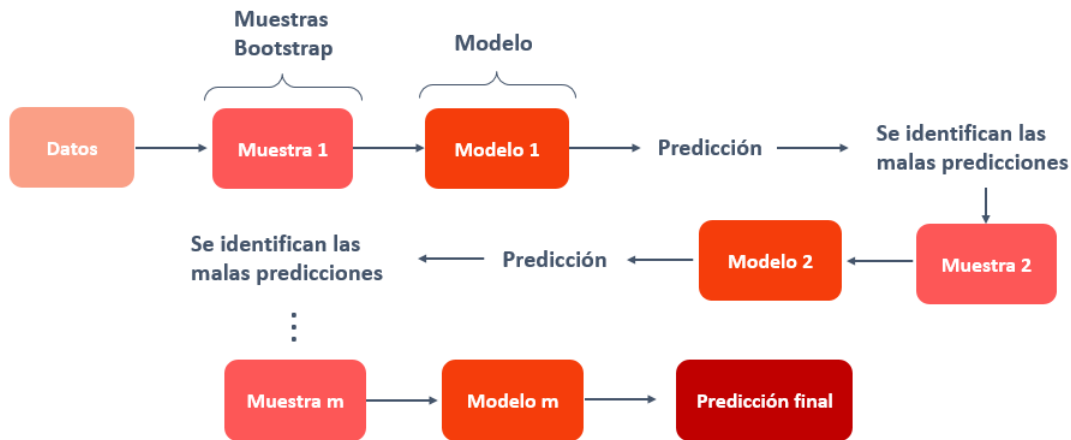


Figura 3.7: Esquema de Boosting

El algoritmo termina una vez que se haya alcanzado un número máximo de iteraciones o cuando el error de estimación converja de acuerdo a un criterio establecido a priori. De esta forma se da lugar a que cada árbol dependa del anterior, la cual es la característica principal que diferencia al boosting del bagging, en donde en este último todos los árboles de decisión son independientes entre sí.

3.2.10 Bosques aleatorios

Los Bosques aleatorios son una técnica de aprendizaje automático la cual hace el uso del método de bagging. Un bosque está conformado por un conjunto de predictores de árboles aleatorios sin podar e independientes entre sí, de tal forma que cada uno de estos es entrenado con un subconjunto de datos muestreados de forma independiente mediante bootstrap a partir del conjunto de datos original para cada árbol del bosque (Breiman, 2001).

Los bosques pueden contar con decenas o cientos de árboles en donde para el caso de la regresión, en cada uno de estos se calcula una predicción de la observación en la variable respuesta. Para calcular la predicción final del bosque, se calcula el promedio de las predicciones de cada árbol entrenado, este paso se le conoce como "ensamblado".

Los Árboles de regresión son entrenados con una muestra aleatoria con reemplazo mediante bootstrapping proveniente del conjunto de datos original. El número de observaciones que se toma en cada muestra es una fracción del conjunto de datos, este número de observaciones elegido para entrenar cada árbol es un hiperparámetro que debe optimizarse.

Cada árbol es entrenado con un número definido de variables predictoras, el cual está dado generalmente por $\frac{p}{3}$, en donde p es el número total de variables empleadas en el modelo, sin embargo, este también es un hiperparámetro que debe optimizarse para entrenar el bosque, en donde en cada árbol se selecciona un subconjunto de variables predictoras de forma aleatoria.

Mediante la aleatorización usando el método de bootstrapping y ensamble, se logra reducir la varianza de las predicciones del bosque, por lo que se convierte en un modelo mucho más preciso y robusto comparado a usar un solo árbol de regresión.

En la Figura 3.8 se muestra la estructura de un Bosque aleatorio enfocado a la regresión en donde la predicción de una nueva observación está dada por la siguiente formula:

$$\hat{y}_{bosque} = \frac{\sum_{i=1}^{n_k} \hat{y}_i}{n_k} \quad (3.18)$$

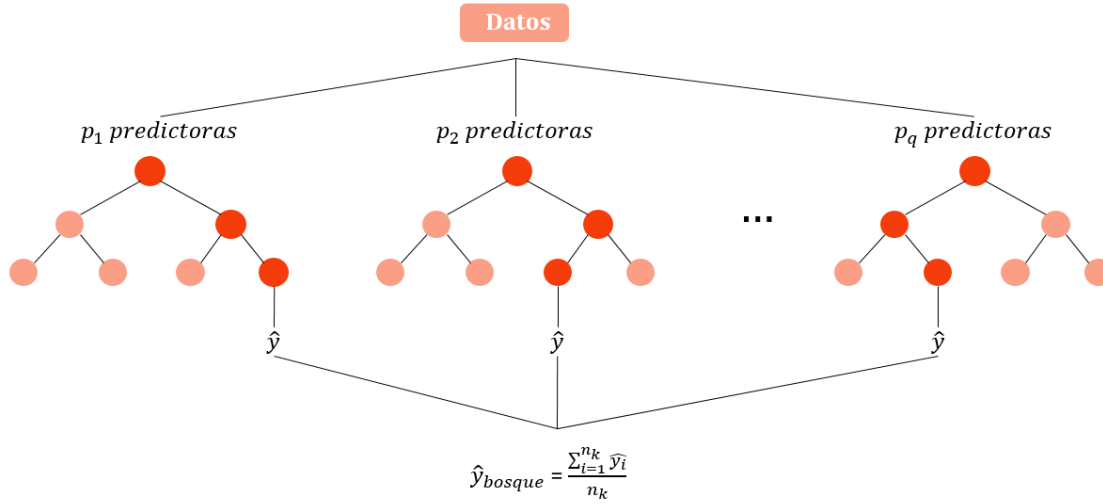


Figura 3.8: Esquema del Bosque aleatorio

Optimización de hiperparámetros

El modelo de bosque aleatorio cuenta con varios hiperparámetros que deben ser optimizados los cuales son: el número de árboles del bosque, la fracción de muestreo y el número de variables predictoras utilizadas para entrenar cada árbol, el número mínimo de observaciones requeridas para dividir un nodo del árbol y el número mínimo de observaciones requeridas en un nodo terminal.

La optimización de los hiperparámetros se basa en encontrar la combinación que permita obtener el mejor bosque aleatorio, en donde se define a priori un rango de valores para cada hiperparámetro que será evaluado, técnica llamada "búsqueda en cuadrícula", y mediante validación cruzada se evalúa el desempeño de cada modelo. El modelo que se selecciona será el que mejor se desempeñe de acuerdo a la métrica de desempeño utilizada.

Las métricas de bondad de ajuste utilizadas en los bosques aleatorios son las mismas explicadas anteriormente para el modelo de árbol de decisión.

3.2.11 XGBoost

El XGBoost o Extreme Gradient Boosting, por sus siglas en inglés, es un modelo ensamblado de aprendizaje automático que hace uso del boosting y los árboles de decisión. En este se toma un modelo de predicción o clasificación considerado débil o con poca fiabilidad y lo va mejorando de forma secuencial reduciendo los errores del modelo entrenado en la iteración anterior, resultando finalmente en un modelo más preciso y robusto (Chen y Guestrin, 2016).

El procedimiento que realiza XGBoost es entrenar un Árbol de decisión mediante el conjunto de datos original y se calculan los residuos de este modelo, luego un segundo árbol es entrenado a partir de estos residuos, este modelo se enfoca en corregir estos errores de predicción. Este paso se repite hasta lograr reducir significativamente una función de coste como puede ser el error cuadrático medio para el caso de la regresión y la entropía para clasificación.

En la Figura 3.9 se muestra la estructura de un modelo de XGBoost en la cual se puede observar su naturaleza secuencial y de refuerzo del modelo inicial.

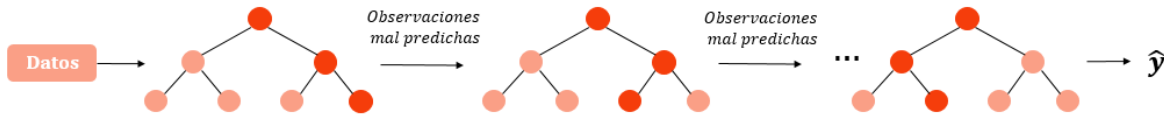


Figura 3.9: Esquema del XGBoost

Optimización de hiperparámetros

En todos los árboles del modelo existen hiperparámetros que controlan la poda de los árboles o el número de variables predictoras a emplear, además de hiperparámetros generales que controlan la regularización para evitar el sobreajuste o que definen el número de árboles a utilizar en el modelo. Así, la predicción final del modelo será la ponderación de la predicción de cada árbol que será mayor en los arboles que mejor rendimiento tuvieron.

Un paso muy importante al implementar este modelo es la optimización de los hiperparámetros lo cual se puede lograr definiendo un cálculo de búsqueda en cuadrícula al igual que en los Bosques aleatorios, en donde se entrena un modelo con las combinaciones de hiperparámetros dados unos rangos establecidos a priori. Mediante validación cruzada se encuentra la combinación de hiperparámetros que tiene el menor error de predicción.

Para evaluar el desempeño y calidad de un árbol se utilizan las métricas de bondad de ajuste empleadas en los árboles de regresión.

4 Metodología

En este capítulo se explica la metodología que se empleó en el Trabajo de grado. En la Figura 4.1 se muestra el flujo de trabajo que se planteado para dar cumplimiento tanto al objetivo general como a los específicos.

Se inicia entendiendo la población de estudio y el método utilizado para obtener los datos, seguido se explica como se procesarán los datos teniendo en cuenta ciertas condiciones de exclusión y completitud de registros para su correcta implementación en el análisis exploratorio y analítico, también se explican las transformaciones que se realizan sobre las variables, la forma en como se imputarán los datos faltantes y los métodos utilizados en el análisis exploratorio para conocer las características de los vehículos usados.

Por último, se explica la metodología empleada en los modelos estadísticos y de aprendizaje automático para conocer el efecto de las covariables sobre el precio de los vehículos y su predicción en nuevas observaciones. Seguido de esto se termina con el procedimiento utilizado para la selección del mejor modelo.

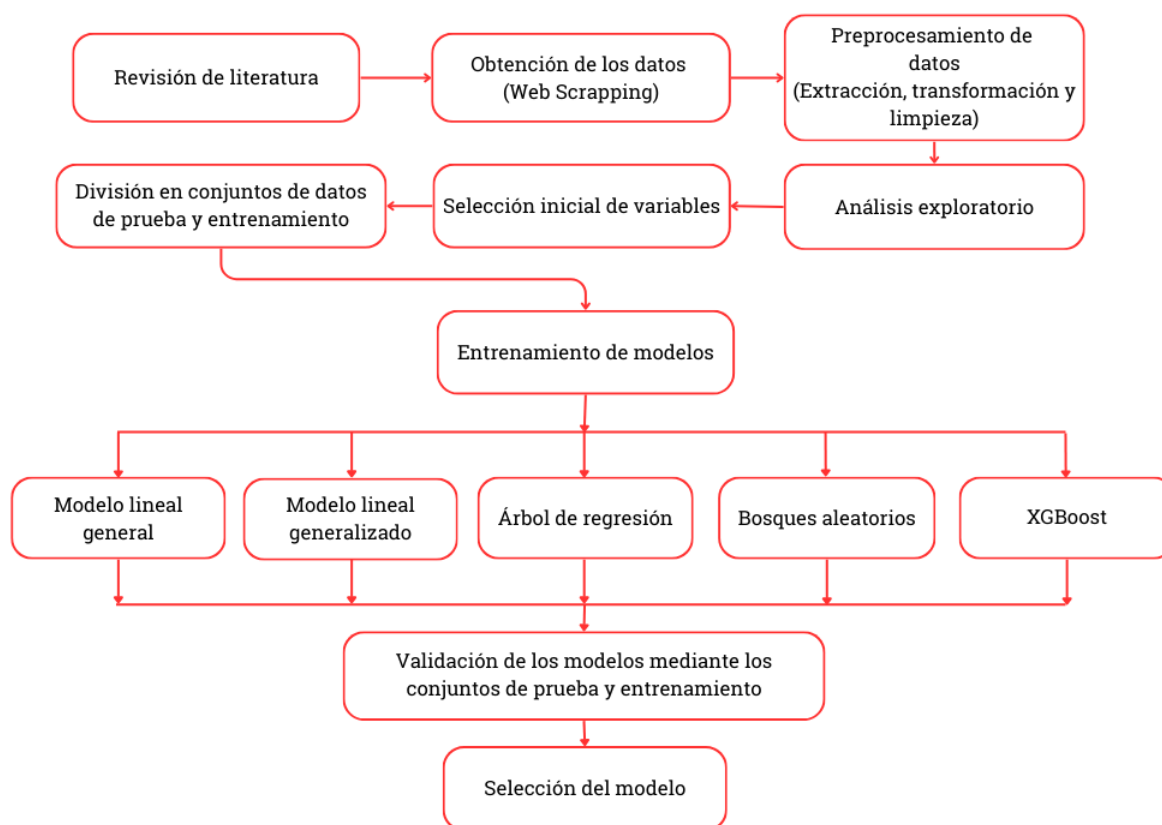


Figura 4.1: Flujo de trabajo de investigación

4.1 Entendimiento de los datos y población de estudio

El trabajo de grado se desarrolla para la población de autos usados en venta en la plataforma Mercado Libre Colombia, que estaban ofertados en la ciudad de Santiago de Cali para la fecha del 24 de septiembre de 2023.

En el estudio solo se incluirán automóviles de cuatro ruedas de uso particular, entiéndase esto como vehículos empleados con fines personales y familiares. Se incluyen solo vehículos en donde el tipo de carrocería es sedán, hatchback, coupé, camioneta, todo terreno, SUV, convertible, van y minivan. Los vehículos como camiones, buses de uso comercial o público no se incluyen, además de vehículos con precios superiores a los 250 millones de pesos, los cuales su mercado potencial no se encuentra en el sitio web de Mercado Libre, ni hacen parte de los objetivos del Trabajo de grado.

Se cuenta con diversas variables proporcionadas por los vendedores cuando ofertan su vehículo en el sitio web como son la marca, modelo, kilometraje, cilindraje o antigüedad del automóvil. Además de esto se tiene que algunas variables no son obligatorias de diligenciar

como son el tipo de carrocería, el tipo de transmisión o el color del vehículo, por lo que es necesario evaluar el tratamiento que se le dará a los registros con datos faltantes. La variable de versión del modelo de vehículo es digitada por el usuario, por lo que su tratamiento es complejo cuando se homologan las versiones de cada modelo, por lo que esta no se incluirá en el estudio.

Mediante la variable de ubicación del vendedor se seleccionan solo las ofertas que hayan sido publicadas por vendedores ubicados en la ciudad de Santiago de Cali.

4.2 Obtención de datos

Mediante las librerías de R, *rvest* y *dplyr* se realiza la recopilación de la oferta de autos usados en venta haciendo uso de la técnica de web scraping en el sitio web de Mercado Libre Colombia en la categoría de automóviles usados en donde el vendedor se encuentra ubicado en la ciudad de Santiago de Cali.

Mediante *rvest* es posible acceder a los datos estructurados en lenguaje HTML de cada página del sitio web y seguido se extraen las etiquetas de interés de la oferta en donde se encuentran las variables de cada vehículo.

En la Figura 4.2 se muestran tres ofertas de vehículos las cuales están ubicadas en una página de la categoría de vehículos usados, cada una de estas ofertas es un contenedor programado en HTML y CSS las cuales contienen el enlace que dirige a la publicación que la información completa de las variables de cada oferta, mediante el algoritmo llamado rastreador web o crawler se accede a cada una para obtener las variables específicas. El principio de este algoritmo para obtener todas las ofertas de autos usados está en rastrear todas las paginas del sitio web que contienen ofertas de este tipo y que el vendedor esté ubicado en la ciudad de Cali. De esta forma primero se almacenan los enlaces web de las ofertas para después acceder en cada una de estas y extraer las variables de cada oferta.

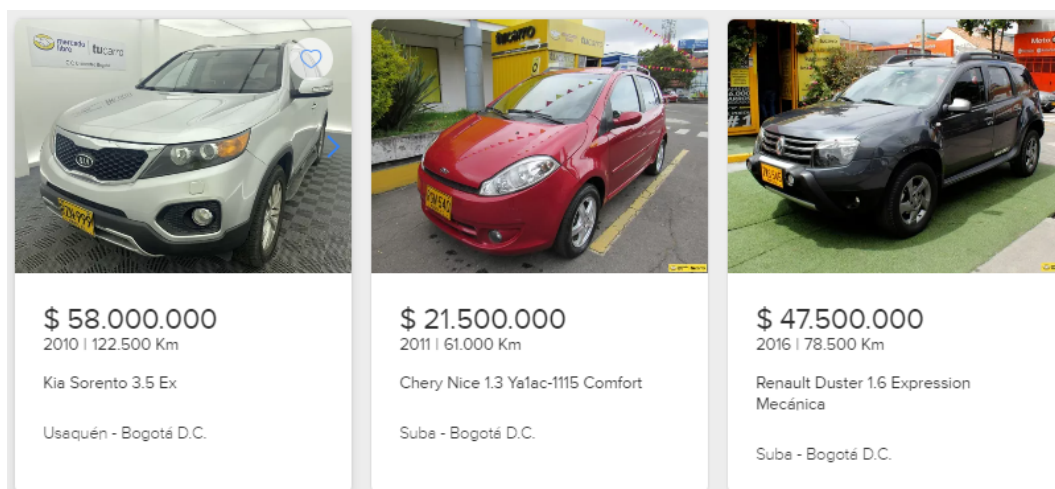


Figura 4.2: Ofertas de venta de autos usados en Mercado Libre.

En la Figura 4.3 se muestra un ejemplo de oferta autom3vil la cual cuenta con las caracter3sticas del veh3culo proporcionadas por el vendedor. Las variables de inter3s poseen un identificador 3nico el cual es un ID o clase de HTML, por lo que el algoritmo puede extraer el texto que est3 contenido en cada uno de estos. Este proceso debe repetirse para cada una de las ofertas encontradas al rastrear todas las paginas de la categor3a. El Web Scraping es una tarea que puede tardar ya que suele realizarse de forma secuencial, sin embargo puede optimizarse para que la obtenci3n de los datos sea de forma paralela mediante la librer3a *parallel* de R.

2010 | 122.500 km - Publicado hace 1 d3a

Kia Sorento Ex

\$ 58.000.000

[Preguntar](#) [WhatsApp](#)

¿Tuviste un problema con la publicaci3n? [Av3sanos.](#)

Principales

Marca	Kia
Modelo	Sorento
A3o	2010
Versi3n	3.5 Ex
Color	Plateado
Tipo de combustible	Gasolina
Puertas	5
Transmisi3n	Autom3tica
Motor	3500
Tipo de carrocer3a	Camioneta
Kil3metros	122500 km

Confort y conveniencia

Aire acondicionado	S3
V3drios el3ctricos	S3
Comando remoto para radio en el volante	S3
Porta vasos	S3
Cierre autom3tico de vidrios	S3

Informaci3n general

Direcci3n	Hidr3ulica
-----------	------------

Rendimiento y dimensiones

Control de tracci3n	4x4
---------------------	-----

¿La referencia de precios fue 3til? [Si](#) [No](#)

Informaci3n del particular

Ubicaci3n del veh3culo
Santa B3rbara - Usaqu3n - Bogot3 D.C.

[Ver tel3fono](#)

Figura 4.3: Oferta de venta de un auto usado

Finalmente las variables extra3das de cada publicaci3n se pueden observar en la Tabla 4.1 en donde se muestra su tipo y lo que significa cada una de estas:

Variable	Tipo	Descripción de la variable
Año	Continua	Año de fabricación
Kilometraje	Continua	Cantidad de kilómetros recorridos
Precio	Continua	Precio en pesos colombianos
Cilindraje	Continua	Cilindrada o desplazamiento del motor
Puertas	Discreta	Número de puertas
Color	Categórica	Color de la carrocería
Ciudad	Categórica	Ciudad en la que se ubica el vendedor
Marca	Categórica	Empresa fabricante
Modelo	Categórica	Modelo o serie
Nombre	Categórica	Nombre de la ficha de producto
Tipo de combustible	Categórica	Tipo de combustible que utiliza el motor
Tipo de transmisión	Categórica	Tipo de transmisión
Tipo de carrocería	Categórica	Tipo de carrocería
Fecha	Categórica	Fecha en la que fue publicada la oferta
Versión	Categórica	Versión del modelo

Tabla 4.1: Variables proporcionadas por los vendedores en la oferta de venta

Se cuenta con un total de 12 variables técnicas y de uso del vehículo, además de estas se incluyen la ciudad y barrio del vendedor, la fecha de publicación, código de la oferta del vehículo y el enlace web. Los registros se extraen y se almacenan en formato .csv. Una vista previa de los datos se puede observar en la Figura 6.1 ubicada en el Anexo A.

La oferta total cuenta con 4.265 registros de vehículos usados en Cali para el 24 de septiembre de 2023.

4.3 Preprocesamiento de datos

Para garantizar la validación de la información para su correcta implementación se realiza la depuración y limpieza de los datos.

No se encontraron ofertas de vehículos duplicadas considerando las variables dadas por los vendedores, de esta forma no se presentaron ofertas repetidas de un mismo automóvil que hubieran sido publicadas por los mismos vendedores en distintas fechas.

Para el estudio se presentan algunos criterios de exclusión de registros para dar sentido al concepto de vehículo usado, además de esto solo se tendrá en cuenta vehículos de uso personal y familiar, por lo que se excluye cierto tipo de carrocería. Además los vehículos de tipo eléctrico también se excluyen, debido a que actualmente su uso no está tan extendido en Colombia y porque además no es tan fácil acceder a ofertas de venta de este tipo de vehículos. Por otra parte tampoco se tendrán en cuenta vehículos en donde su precio sea mayor a los 250 millones de pesos ya que estos suelen pertenecer a las categorías de deportivos, súper deportivos, uso especial o de colección, por lo que no se consideran vehículos particulares de

uso diario.

Por último, solo se incluyen los registros en donde las ofertas de venta hayan sido diligenciadas por completo en todas sus variables ya que para realizar la publicación de la oferta de venta, las variables como el tipo de carrocería, el tipo de transmisión y el color no son obligatorias de diligenciar. Esta situación puede crear ambigüedades cuando se fija el precio de un vehículo ya que por ejemplo, un mismo modelo de vehículo puede variar su precio en gran medida si este tiene cierto tipo de transmisión o carrocería.

Los registros con datos faltantes en el cilindraje no se excluirán ya que en algunos casos se puede generar una aproximación mediante imputación aprovechando la disponibilidad de las demás variables en el registro, relacionadas con la variable del cilindraje que presenta el dato faltante.

De acuerdo a estos criterios se eliminan los registros que cumplan las condiciones mostradas en la Tabla 4.3:

Criterio de exclusión	Número de autos	Porcentaje de autos
Datos faltantes en carrocería, transmisión y color	1.479	34,6 %
Autos con cero kilómetros	202	4,7 %
Autos con precio mayor a 250 millones de pesos	95	2,2 %
Autos con tipo de combustible eléctrico	20	0,5 %
Autos fabricados en el año 2024	11	0,2 %
Autos con tipo de carrocería furgón	9	0,2 %
Total general	1.816	42,5 %

Tabla 4.2: Criterios de exclusión de registros

Aplicando estos criterios se eliminan 337 registros con cero kilómetros, autos fabricados en el año 2024, autos con precios mayores a 250 millones de pesos, tipo de motor eléctrico, tipo de carrocería furgón. Además se eliminan 1.479 registros que presentaron datos faltantes en las variables tipo de carrocería, tipo de transmisión y color, las cuales son variables difícil de imputar.

Finalmente se obtiene una hoja de datos con 2.449 ofertas de vehículos que se empleará en el análisis exploratorio y en el desarrollo de los modelos analíticos.

4.3.1 Transformación de variables

Para facilitar el análisis de la información, la variable respuesta y algunas covariables se transformaron y depuraron de la siguiente manera:

- **Precio:** se explica en millones de pesos.
- **Kilometraje:** se explica en una escala de 10.000 kilómetros.

- **Cilindraje:** se estandariza la escala a litros (por ejemplo 1,0) se depuraron valores como 1000, 1200L, 1100 L, 1,1 L, 1,5l, o 0,8 CC.
- **Número de puertas:** se agruparon las categorías 2, 3 y 4, 5 puertas, por lo que esta variable pasa de ser discreta a categórica.
- **Color:** Se homologaron los colores de los vehículos, por ejemplo el color gris se denominó plateado y el azul oscuro como azul.

Mediante el año de fabricación del modelo de vehículo se calculó la antigüedad con respecto al año 2023.

Por otra parte, debido al alto número de categorías en la variable marca, en la cual se cuentan con un total de 31 marcas de vehículos, se puede presentar un problema al ajustar los modelos con tan pocos grados de libertad, es por esto que se decide realizar un agrupamiento de las marcas, apoyados en el ajuste de un modelo de Árbol de decisión entre la variable precio y la marca. Este ajuste sugerirá agrupaciones naturales de marcas con similar precio medio.

En la Figura 4.4 se muestra como es el comportamiento del error de validación cruzada de acuerdo al número de nodos terminales en el Árbol de regresión. Se puede observar que con cuatro nodos terminales se logra la mayor disminución del error de validación cruzada, sin embargo con más de cuatro nodos se disminuye solo un poco el error a costa de aumentar el número de grupos de las marcas, por lo que finalmente se usan cuatro nodos terminales en el modelo.

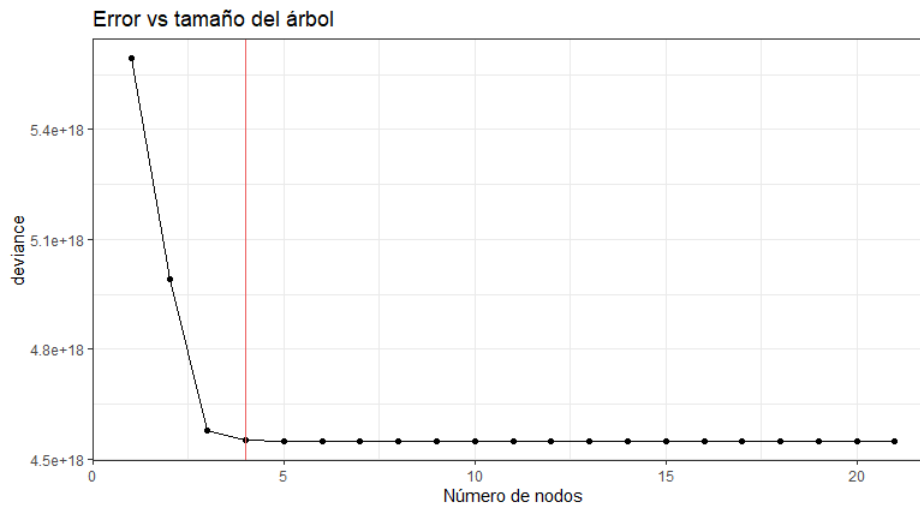


Figura 4.4: Error de validación cruzada frente al número nodos terminales del modelo de Árbol de regresión del precio en función de la marca.

De esta forma se obtienen las marcas que condicionan a cada uno de los cuatro nodos terminales por lo que las agrupaciones de las marcas sugeridas por el árbol son las siguientes:

- **Grupo A:** Chevrolet, FIAT, Hyundai, MG, Renault, Skoda, Suzuki, Zotye.
- **Grupo B:** Citroën, Dodge, Ford, Honda, JAC, KIA, Mazda, Nissan, Peugeot, RAM, Ssangyong, Subaru, Volkswagen.
- **Grupo C:** Audi, BMW, Jeep, Mercedes-Benz, MINI, Mitsubishi.
- **Grupo D:** Land Rover, Porsche, Toyota, Volvo.

Por otra parte, se incluirá una nueva variable calculada la cual es la razón entre el número de kilómetros y la antigüedad del vehículo. Este indicador permite evaluar cuántos kilómetros ha recorrido un vehículo en promedio por cada año de antigüedad, lo que puede ser importante para comprender el precio de acuerdo al desgaste y uso del vehículo.

4.3.2 Imputación de datos faltantes

Debido al criterio de completitud de registros mencionado anteriormente solo se presentan 200 (8,2 %) registros con datos faltantes en la variable cilindraje los cuales serán incluidos en el análisis.

Los datos faltantes en esta variable son fácilmente de imputar debido a que el cilindraje de un vehículo está dado principalmente por su modelo, tipo de carrocería y transmisión, presentándose muy poca variación entre distintas versiones de cada modelo, por lo que se imputarán los datos faltantes mediante regresión empleando como covariables el precio, marca, antigüedad, modelo, número de puertas, kilómetros, tipo de carrocería, transmisión y combustible.

4.4 Análisis exploratorio

Se realiza el análisis exploratorio para cada una de las variables de la hoja de datos. Para las variables cuantitativas, se analiza su distribución identificando la naturaleza de estas y la presencia de datos extremos. Mientras que para las variables cualitativas se analizan las categorías que destaquen por su mayor frecuencia.

El análisis individual de cada variable se complementa con la evaluación de su comportamiento frente a la variable del precio y las demás covariables, evaluando la correlación entre estas mediante los coeficientes de correlación de Pearson y de Spearman.

Para la variable del precio se analizan los datos atípicos mediante el criterio de Tukey y para complementar, se analizan los datos atípicos de forma multivariada mediante el análisis múltiple de las relaciones entre las variables mediante el índice del valor atípico local.

Datos atípicos

Los datos atípicos del precio de los autos usados se analizan de forma univariada mediante el método de Tukey el cual permite detectar valores atípicos teniendo en cuenta el

primer y tercer cuartil. Se usará una desviación de 3 veces el rango intercuartilico para evaluar datos atípicos extremos, en donde el cálculo de los límites está dado por la ecuación (3.1).

También se analizan los datos atípicos de forma multivariada mediante los puntajes de valor atípico local o LOF, este método se centra en analizar las relaciones entre las variables evidenciando ciertos patrones que no se presentarían en un análisis univariado. Para este método se usan todas las variables cuantitativas y se convierten las variables categóricas a variables dummy y se usarán 20 vecinos cercanos para calcular los puntajes LOF de cada observación como se define en la ecuación (3.5). De esta forma se logra analizar de una forma más completa el impacto de las relaciones entre las variables. Se consideran valores atípicos las observaciones que superen el percentil del 99 % en sus puntajes.

Se evaluará la inclusión los datos atípicos y se medirá el efecto de estos sobre la estabilidad y calidad de ajuste de los modelos.

4.5 Fase de entrenamiento y evaluación de modelos

Para el análisis de las relaciones de la variable respuesta y las predictoras, además de la predicción del precio de autos usados se emplearán modelos estadísticos y de aprendizaje automático con los que se identificarán las características de los vehículos que afectan al precio.

Como variable respuesta en todos los modelos se tiene el precio de venta del vehículo en pesos colombianos expresado en millones el cual fue proporcionado por el vendedor. Como variables explicativas del precio se tiene la marca, antigüedad, color, tipo de combustible, tipo de transmisión, tipo de carrocería, cilindraje, kilometraje y número de puertas del vehículo. Todas las variables categóricas son convertidas a variable dummy en donde la variable dummy correspondiente a la categoría más frecuente es retirada para evitar la multicolinealidad. De esta forma se tendrán un total de 18 variables explicativas del precio.

Se emplean tanto modelos estadísticos, como el Modelo lineal general y el generalizado bajo la distribución gamma mediante función de enlace logarítmica, como también modelos de aprendizaje automático como el Árbol de regresión, Bosques aleatorios y XGBoost. Estos modelos se desarrollan en los lenguajes de programación R y Python mediante distintas librerías y se optimizan los hiperparámetros de cada modelo mediante validación cruzada para obtener el mejor de cada método.

Se emplea el 80 % de los datos para entrenar los modelos y el 20 % restante para evaluar el rendimiento de estos en datos no observados.

Mediante el coeficiente de determinación ajustado definido en la ecuación (3.9), se compara el desempeño de cada modelo tanto para los datos de prueba como para los de entrenamiento, evaluando el rendimiento con cada conjunto de datos detectando posibles sobreajustes y finalmente eligiendo el mejor modelo analítico encontrado para ajustar el precio de los vehículos. Haciendo uso de la raíz del error cuadrático medio definido en la ecuación (3.10),

se evalúa la diferencia promedio en escala de millones de pesos de los valores ajustados por cada modelo frente al valor real.

4.5.1 Modelo lineal general

Se desarrolla un Modelo lineal general mediante la función *lm* del lenguaje de programación R. Se ajusta el modelo completo del precio transformado mediante su logaritmo natural para capturar las relaciones no lineales entre este y las covariables, además de también lograr estabilizar la varianza de los errores. El modelo está dado por:

$$\begin{aligned} \hat{precio}_i = & \beta_0 + \beta_1(antigüedad_i) + \beta_2(cilindraje_i) + \beta_3(kilometros_i) + \beta_4(km/antigüedad_i) + \\ & \beta_5(marcaA_i) + \beta_6(marcaC_i) + \beta_7(marca_D_i) + \beta_8(combustible_Diesel_i) + \\ & \beta_9(combustible_GasolinaGas_i) + \beta_{10}(combustible_Híbrido_i) + \\ & \beta_{11}(2_y_3_puertas_i) + \beta_{12}(Convertible_i) + \beta_{13}(Coupé_i) + \beta_{14}(Hatchback_i) + \\ & \beta_{15}(Minivan_i) + \beta_{16}(Pick_Up_i) + \beta_{17}(Sedán_i) + \beta_{18}(Transmisión_Mecánica_i) \end{aligned} \quad (4.1)$$

Donde $\hat{y}_i$ es el precio ajustado por el modelo de la i -ésima observación. Las variables categóricas son transformadas a variables dummy que tomarán valores de 1 y 0 de acuerdo a la característica que tenga el vehículo. Por otra parte, $\beta_0, \beta_1, \beta_2, \dots, \beta_{18}$ son los coeficientes de regresión asociados al intercepto y a las covariables.

Se evalúan los supuestos del modelos mediante los residuos haciendo uso tanto de pruebas gráficas como estadísticas. Ninguno de los supuestos se cumple debido a la naturaleza de la relación del precio y las covariables.

Además se transforman las covariables mediante sus logaritmos naturales, y se prueba la transformación de Box-Cox sin lograr mejora alguna en el cumplimiento de los supuestos.

4.5.2 Modelo lineal generalizado gamma

Como alternativa al incumplimiento de los supuestos del Modelo lineal general, se ajusta un Modelo lineal generalizado bajo la distribución gamma en donde se usa la función de enlace logarítmica debido a que es la mejor opción ante la presencia de asimetría en la distribución del precio, esto antes de probar la bondad de ajuste del precio bajo la distribución gamma mediante un gráfico cuantil-cuantil en donde se calculan los cuantiles observados y los cuantiles teóricos. Como en este caso se usa la función de enlace logarítmica, la forma explícita del modelo para obtener las estimaciones del precio es:

$$\begin{aligned}
g^{-1}(\eta) = \hat{precio}_i = & \exp(\beta_0 + \beta_1(antigüedad_i) + \beta_2(cilindraje_i) + \beta_3(kilometros_i) + \\
& \beta_4(km/antigüedad_i) + \beta_5(marcaA_i) + \beta_6(marcaC_i) + \\
& \beta_7(marca_D_i) + \beta_8(combustible_Diesel_i) + \\
& \beta_9(combustible_GasolinaGas_i) + \beta_{10}(combustible_Híbrido_i) + \\
& \beta_{11}(2_y_3_puertas_i) + \beta_{12}(Convertible_i) + \beta_{13}(Coupé_i) + \\
& \beta_{14}(Hatchback_i) + \beta_{15}(Minivan_i) + \beta_{16}(Pick_Up_i) + \beta_{17}(Sedán_i) + \\
& \beta_{18}(Transmisión_Mecánica_i))
\end{aligned}
\tag{4.2}$$

Debido a que uno de los objetivos del trabajo es estudiar las variables que afectan al precio de los vehículos usados se emplearon desde un inicio todas las covariables, mediante los coeficientes y sus exponenciales se estudian las covariables que más influencia tienen en el precio tanto positiva como negativamente. Mediante las pruebas estadísticas de cada coeficiente del modelo se evalúa si las variables son significativas en el modelo.

Se analiza la distribución de los residuales y su ajuste bajo la distribución gamma, además se estudia la multicolinealidad mediante los factores de inflación de varianza verificando si existen covariables que puedan estar proporcionando información redundante que sea mejor explicada por otras covariables en el modelo.

Para encontrar la mejor combinación de variables predictoras en el modelo se emplea la técnica "forward-backward stepwise" la cual se basa en dos algoritmos. El primero es llamado "paso hacia adelante" en donde se empieza con el modelo nulo en donde no se tiene ninguna variable predictora y se va agregando una variable en cada iteración, así se agrega al modelo la variable que logre reducir más el criterio de selección elegido como el AIC o BIC. El otro algoritmo es llamado "paso hacia atrás" en donde se empieza con el modelo con todas las variables predictoras y en cada paso se va eliminando una a la vez, la variable que se elimina en cada paso será la que proporciona la menor disminución en el criterio de selección. Estas dos técnicas se combinan para obtener el modelo con mayor rendimiento en el criterio de selección elegido, en este caso se empleará el criterio de información de Akaike o AIC.

4.5.3 Árbol de regresión

Mediante la librería *tree* de R y la función con el mismo nombre, se encuentra el Árbol de regresión más complejo posible teniendo como variable respuesta el precio del vehículo y todas las variables predictoras.

Como hiperparámetros del modelo de Árbol de regresión se tienen:

- **mincut:** es la cantidad mínima de observaciones necesarias para realizar la división de un nodo interno.
- **mindev:** es la reducción mínima en la devianza al dividir un nodo.

- **minsize:** es el número mínimo de observaciones requeridas en un nodo terminal.

Mediante la evaluación de estos hiperparámetros se puede obtener el número de nodos terminales del modelo. En el proceso de optimización del modelo, se calcula un Árbol de regresión para cada una de las combinaciones de estos hiperparámetros establecidas a priori. De acuerdo a estos hiperparámetros se analiza el número de nodos terminales que logran reducir al mínimo el error de predicción cruzada "xerror" sin que se llegue a un árbol tan complejo. Después de esto se puede actualizar el árbol, proceso llamado "poda", encontrando así el mejor modelo mediante esta técnica de aprendizaje automático.

El modelo de Árbol de regresión es meramente predictivo por lo que mediante este no se podrá evaluar el efecto de las covariables sobre el precio, sin embargo, se puede estudiar la importancia de estas al analizar las covariables que fueron seleccionadas en el proceso de poda.

4.5.4 Bosques aleatorios

Para el entrenamiento del modelo de Bosque aleatorio se utiliza la librería *scikit-learn* de python y sus funciones *RandomForestRegressor* y *GridSearchCV*, debido a su facilidad para la optimización de los hiperparámetros del modelo los cuales se muestran a continuación:

- **n_estimators:** es el número de árboles que tendrá el modelo.
- **max_feature:** es el número máximo de variables predictoras aleatorias que se usarán para entrenar cada árbol.
- **bootstrap:** indica si la muestra bootstrap que se utiliza para entrenar cada árbol se realizará con o sin reemplazo.
- **max_depth:** es el número máximo de niveles permitido en cada árbol.
- **min_samples_leaf:** es el número mínimo de observaciones que debe contener cada hoja del árbol.
- **min_samples_split:** es el número mínimo de observaciones necesarias para dividir un nodo.

Mediante búsqueda en cuadrícula se dan a priori los rangos a evaluar de cada hiperparámetro y se ajusta un modelo para cada combinación de estos. Empleando validación cruzada con cinco subconjuntos se encuentra el modelo que tenga el menor error de predicción.

El modelo de Bosque aleatorio le da un puntaje a las covariables cuando su inclusión logra reducir la impureza por lo que se puede evaluar la importancia de cada una, ofreciendo más interpretabilidad al análisis de los efectos de las variables en el precio en comparación a usar un solo Árbol de regresión.

4.5.5 XGBoost

Para encontrar el mejor modelo de XGBoost se utilizan las librerías *xgboost* y *scikit-learn* de python. Con la función *GridSearchCV* se realiza búsqueda en cuadrícula de los hiperparámetros que mejor rendimiento obtengan en el error de predicción de validación cruzada. Estos hiperparámetros son los siguientes:

- **n_estimators:** es el número de árboles del modelo.
- **max_depth:** es la profundidad máxima de cada árbol. Controla el número de nodos que estos tendrán.
- **eta:** controla el aporte de cada árbol al modelo. Cada salida del árbol se multiplica por este factor, permitiendo evitar el sobreajuste.
- **gamma:** penaliza la creación de nuevos nodos si estos no aportan en reducir la pérdida esperada en la función de pérdida seleccionada la cual puede ser el error cuadrático medio.
- **lambda:** controla la regularización ridge por lo que ayuda a evitar el sobreajuste reduciendo en cierta medida la complejidad del árbol.
- **alpha:** controla la regularización LASSO, también ayuda a evitar el sobreajuste centrando el entrenamiento de los árboles en ciertas variables predictoras.
- **colsample_bytree:** es la proporción de variables predictoras aleatorias que se usarán para entrenar cada árbol.
- **min_child_weight:** es el número de observaciones mínimo requerido en cada hoja del árbol.
- **subsample:** es la proporción de la muestra que se usará para entrenar cada árbol.

Al igual que los Bosques aleatorios, el modelo de XGBoost también calcula la importancia de las variables predictoras, permitiendo evaluar cuales son más relevantes para predecir el precio.

Como se puede evidenciar, el método de optimización es el mismo para los tres modelos de aprendizaje automático en donde se centra en encontrar mediante validación cruzada los hiperparámetros que tengan el menor error de predicción y a la vez eviten el sobreajuste.

4.6 Comparación de modelos entrenados

Una vez entrenado el mejor modelo de cada método se utiliza validación cruzada con cinco subconjuntos de datos para evaluar el rendimiento del coeficiente de determinación ajustado y la raíz del error cuadrático medio presentados en el capítulo de Marco teórico en las ecuaciones (3.9) y (3.10) respectivamente, determinando la varianza de las estimaciones de cada modelo y mediante este análisis, evaluar el mejor modelo, para la predicción del

precio de los vehículos usados, analizando las ventajas y desventajas de cada modelo tanto para los datos de entrenamiento como de prueba.

5 Resultados

En este capítulo se presentan los resultados del Trabajo de grado. Para abordar el análisis exploratorio de las variables cuantitativas se emplearon histogramas para analizar sus distribuciones, acompañados del cálculo de las medidas de tendencia central más relevantes. En el caso de las variables cualitativas, el número de ofertas de autos por cada una de sus categorías se representaron gráficamente mediante gráficos de barras.

Posteriormente se explora la existencia de relaciones entre las variables cuantitativas y el precio de los vehículos mediante un diagrama de dispersión múltiple. Además, se analiza la distribución del precio en función de las variables cualitativas mediante diagramas de cajas. También se identifican los datos atípicos presentes en la variable del precio mediante un diagrama de caja y la utilización del método de Tukey, en donde también se evalúan los datos atípicos de forma multivariada, calculando los puntajes LOF de cada registro y representándolos gráficamente.

Por último, se analizan los resultados de cada uno de los modelos ajustados y se realiza la comparación para elegir el modelo con mejor calidad de ajuste.

5.1 Análisis exploratorio

En la Figura 5.1 se muestra la distribución de las variables cuantitativas. Empezando con el análisis del precio de los autos usados, este ronda en un rango entre 5 y 247 millones de pesos, en donde el 50 % de los vehículos tienen precios entre los 44,5 y 105 millones de pesos. Además la media del precio es de 80,6 millones, en este caso debido a la asimetría de la distribución, el mejor indicador de tendencia central es la mediana la cual es de 68,7 millones de pesos.

Para la antigüedad del vehículo, se evidencia que los automóviles tienen una antigüedad entre 1 y 71 años, en donde el 50 % de los automóviles tienen una antigüedad entre 3 y 8 años, una media de 6,1 y mediana de 5 años.

En el kilometraje se presenta que el 50 % de los vehículos tienen un kilometraje entre los 32 mil y 92 mil kilómetros, observándose un máximo de 805,6 mil kilómetros recorridos. Además se obtuvo una media de 72 mil y mediana de 58,2 mil kilómetros.

Por último, en la variable cilindraje se evidencia que los vehículos tienen un cilindraje entre los 0,8 y 6,4 litros, en donde el 50 % poseen un cilindraje entre 1,6 y 2 litros, además se presenta una media de 1,9 y mediana de 1,8 litros.

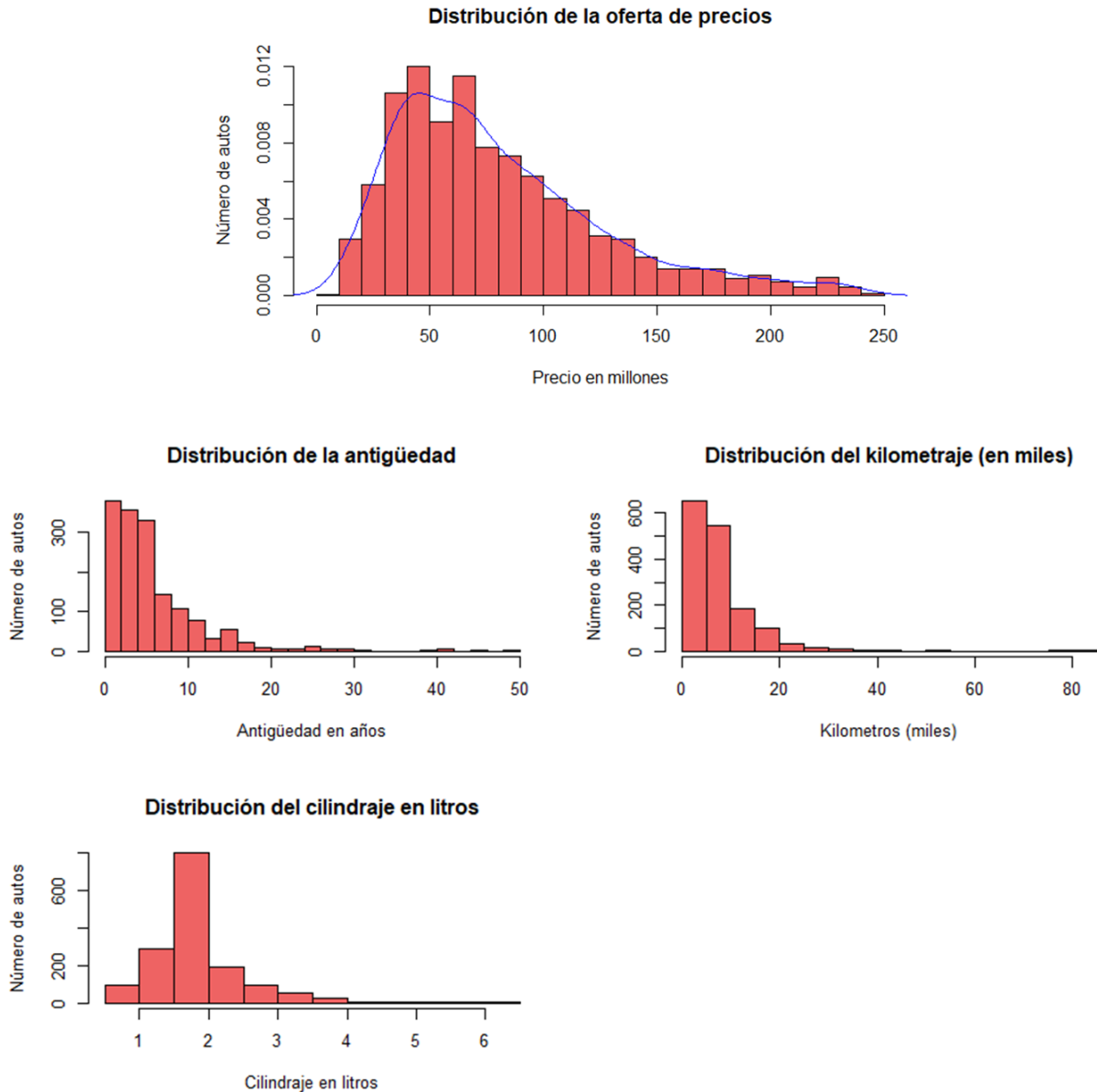


Figura 5.1: Histograma del precio de venta, antigüedad, kilometraje y cilindraje de los autos usados en millones de pesos.

En la Figura 5.2 se muestran las marcas y modelos más ofertados, además de la distribución de frecuencias de las demás variables cualitativas. Se evidencia que el 80,8 % de las marcas más ofertadas son Mazda, Chevrolet, Renault, Nissan, Ford, Toyota y Kia. Por otra parte, el 84 % de los modelos de automóviles más ofertados está conformado por los modelos Spark de Chevrolet, Mazda 3 y CX-5 de Mazda, Duster, Kwid y Sandero de Renault, y los modelos Ford Escape, Honda CR-V, Nissan March y BMW Serie X.

En los siguientes gráficos se analizan las características técnicas de los vehículos. Se evidencia que los vehículos con colores plateado y blanco conforman el 80,1 % de la muestra. Los vehículos con motor a gasolina son los más frecuentes siendo un 92,1 % del total. Por otra parte, para el tipo de carrocería se tiene que el 92 % de los automóviles poseen carrocerías de camioneta, sedán y hatchback. El tipo de transmisión más frecuente es la automática con el 72 % de los vehículos. Por último, los vehículos con 4 y 5 puertas son los más ofertados con un 94,3 % de la muestra.

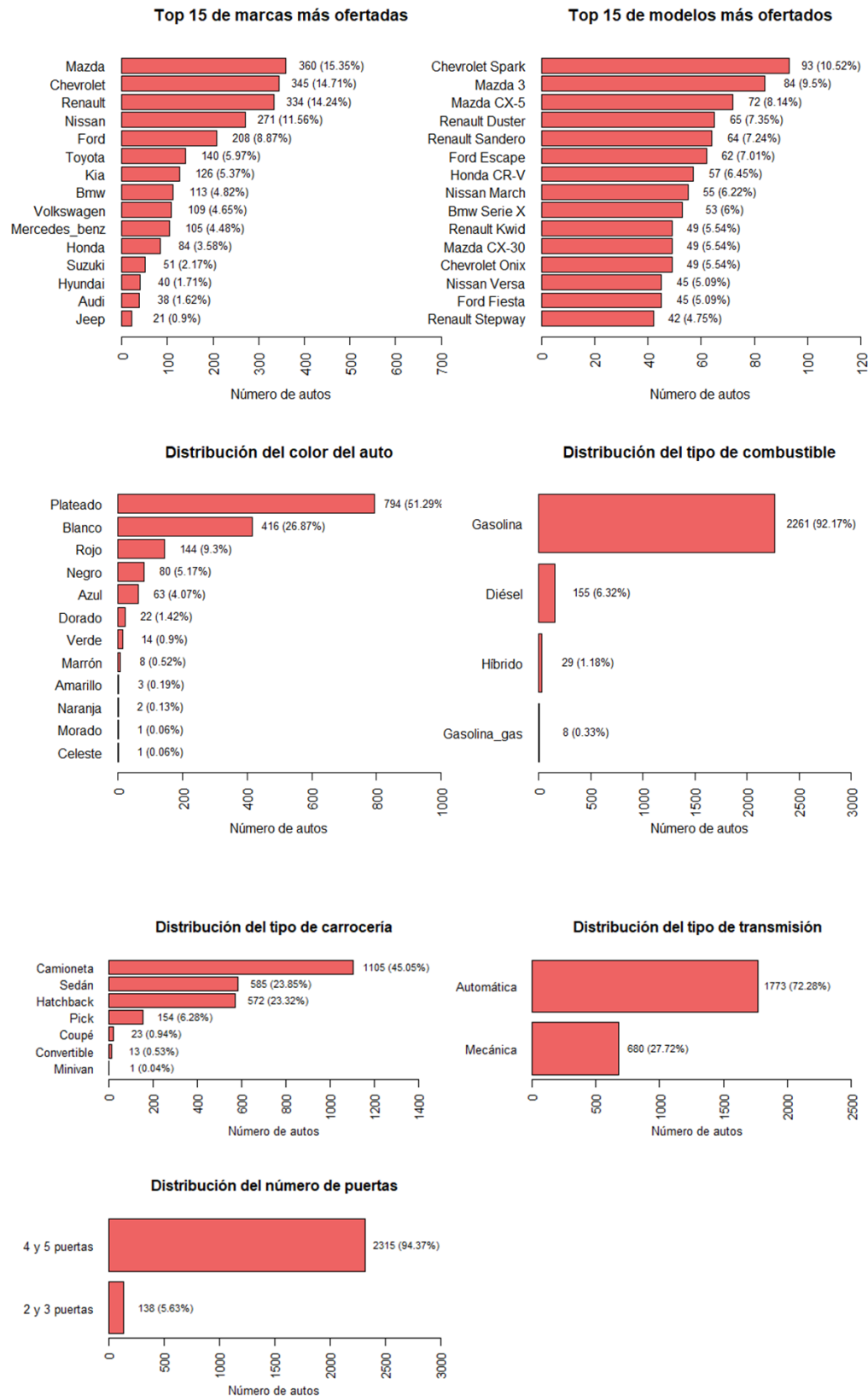


Figura 5.2: Distribución de los modelos y marcas más ofertadas, color, tipo de combustible, tipo de carrocería, transmisión y número de puertas de los autos usados.

En la Figura 5.3 se analiza la distribución bivariada de las cinco variables cuantitativas que se evalúan en el estudio. Se obtuvo que la variable más correlacionada linealmente con el precio fue el cilindraje, presentándose una relación positiva entre estas (Pearson: 0,4424, Spearman: 0,5352), seguida de la antigüedad del vehículo con una relación negativa en donde se observa un comportamiento no lineal (Pearson: -0,4037, Spearman: -0,4869). La variable menos correlacionada con el precio fueron los kilómetros recorridos por año en donde se observa una correlación lineal casi nula (Pearson: 0,0813, Spearman: 0,1628).

Por otra parte la antigüedad y el kilometraje del vehículo fueron las variables más correlacionadas en donde si aumenta la antigüedad, en general también lo hará el kilometraje (Pearson: 0,6562, Spearman: 0,8171). Mientras que las variables menos correlacionadas entre sí fueron los kilómetros recorridos por año y el cilindraje. Hay que recordar que la variable de los kilómetros recorridos por año es una razón entre los kilómetros y la antigüedad del vehículo, por lo tanto su relación con otras variables no será lineal, sino que adoptará una forma de L.

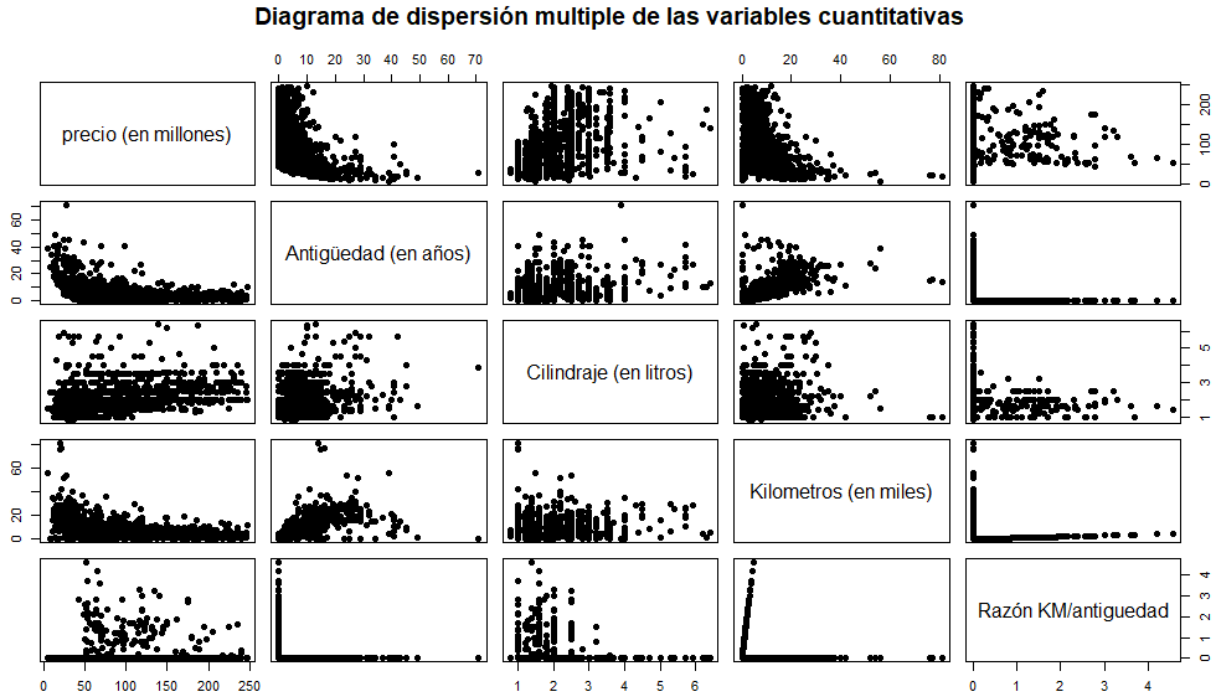


Figura 5.3: Diagrama de dispersión múltiple de las variables cuantitativas

En la Figura 5.4 se muestra la distribución del precio de acuerdo a las marcas y modelos de los vehículos más ofertados mostrados en la Figura 5.2. Las ocho marcas más ofertadas presentan precios entre los 11,5 y 245, con una mediana de 66,5 millones de pesos. Las que presentan un mayor precio medio son Toyota, Mercedes Benz y Mazda con medianas de 140, 98 y 83 millones respectivamente. Por otra parte, las marcas más económicas son Chevrolet, Renault y Kia con medianas de 41, 46,5 y 63 millones respectivamente.

En la siguiente gráfica se evidencia que los ocho modelos de automóviles más ofertados tienen un precio entre los 16 y 239 millones, con una mediana de 67,5 millones de pesos. Los modelos con mayor precio medio son el BMW Serie X, el Honda Cr-V y Mazda CX-5 con medianas de 145, 114 y 108,5 millones de pesos. Además los modelos con menor precio medio son el Chevrolet Spark, Renault Sandero y Renault Duster con medianas de 31,9, 44 y 60,9 millones respectivamente.

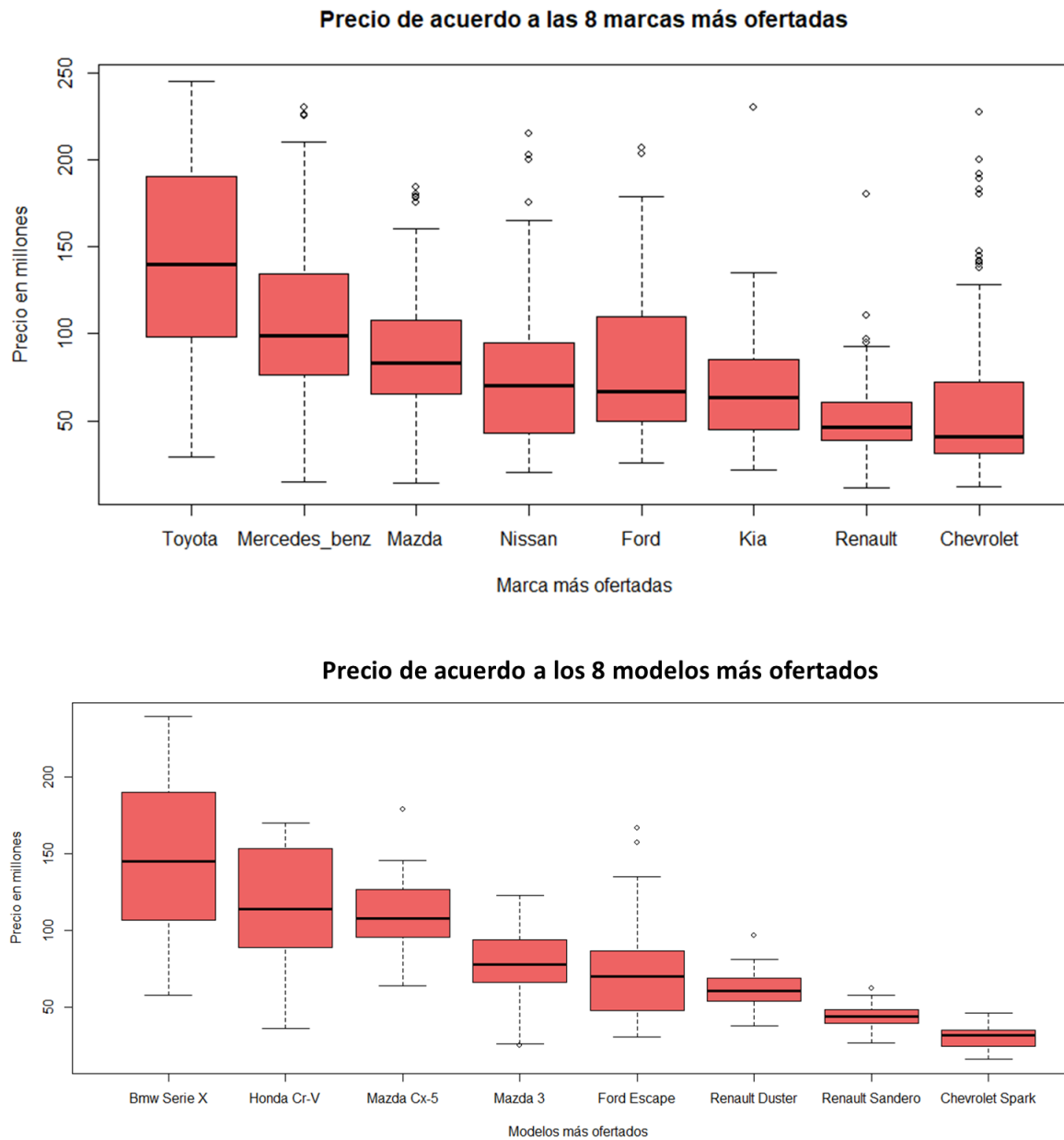


Figura 5.4: Distribución del precio de acuerdo a las marcas y modelos más ofertados.

Por otra parte, en la Figura 5.5 se muestra la distribución del precio de acuerdo al color, tipo de combustible, transmisión, y carrocería, y número de puertas. Se observa que los vehículos con colores blanco, amarillo y plateado son los que tienden a tener un mayor precio medio, teniendo una mediana de 79, 75 y 67,9 millones de pesos respectivamente. Por otra parte, los vehículos con color verde y dorado son los más económicos en promedio con medianas de 38,5 y 41,8 millones de pesos.

Para el tipo de combustible se evidencia que los vehículos con motor a diesel son los que tienen un precio más elevado con una mediana de 148 millones, seguido de los híbridos con 126 millones y a gasolina con 66,5 millones de pesos.

Seguido se muestra la distribución del precio dado el tipo de transmisión y de carrocería. Se presenta que los autos con tipo de transmisión automática tienen en promedio un mayor precio, con una mediana de 83 millones, mientras que los que tienen transmisión mecánica presentan una mediana de 39,7 millones de pesos.

Los autos con tipo de carrocería con mayor precio medio es el convertible con una mediana de 134,9 millones, seguido de los pick-up con 116 millones. Además el tipo de carrocería camioneta, el cual fue el más ofertado por los vendedores, presentó una mediana de 93 millones. Por último, los hatchback y sedán son los vehículos con el tipo de carrocería más económico con una mediana de 42,3 y 57,9 millones de pesos respectivamente.

Para finalizar se presenta la distribución del número de puertas en donde los vehículos con dos o tres puertas resultaron ser los que presentaron un mayor precio medio, evidenciando una mediana de 95,5 millones, seguido de los vehículos con cuatro o cinco puertas con 67,9 millones.

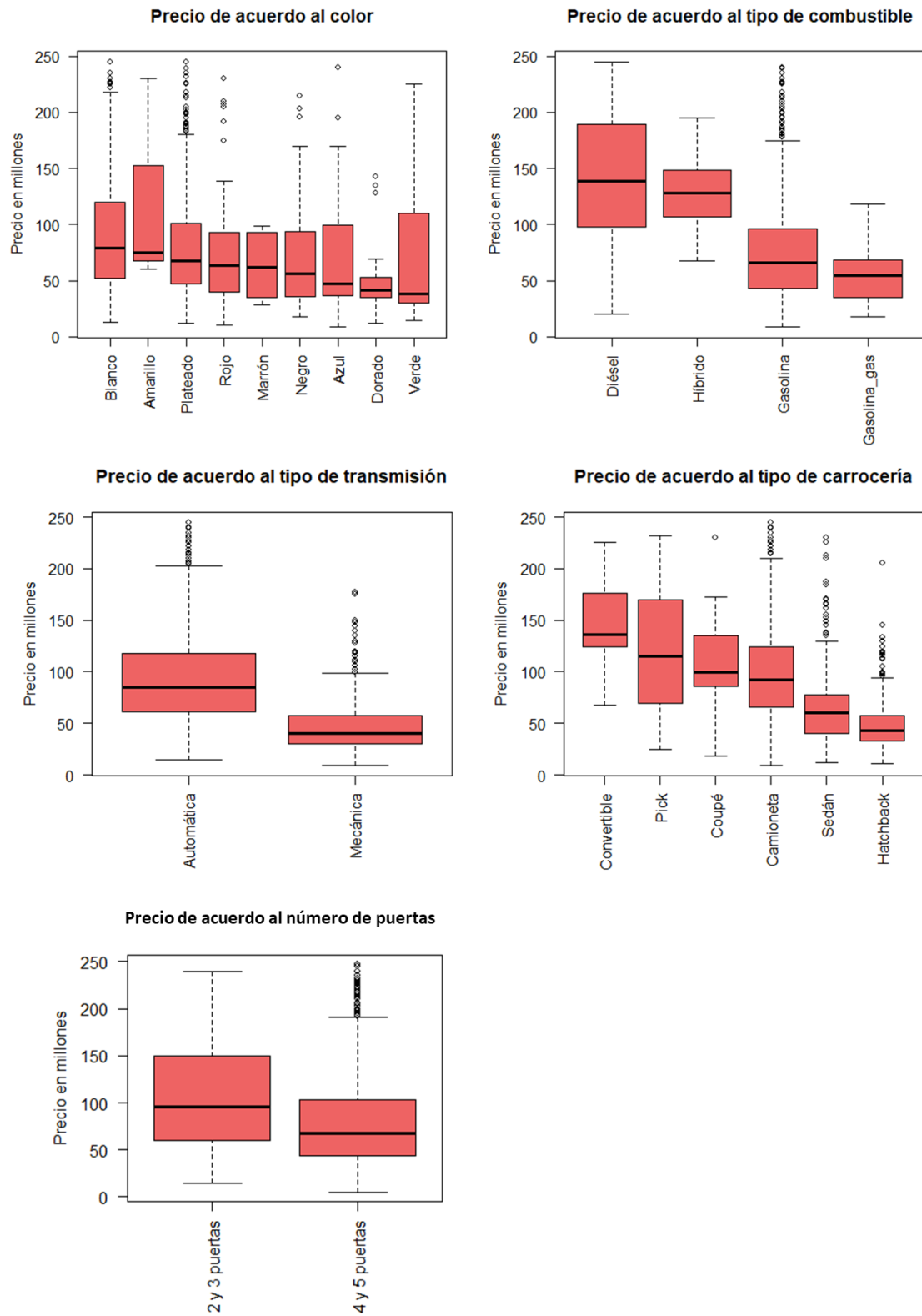


Figura 5.5: Distribución del precio de acuerdo al color, tipo de combustible, tipo de transmisión y carrocería, y número de puertas.

5.1.1 Análisis de datos atípicos

Se evalúa la presencia de datos atípicos para el precio de los vehículos de forma univariada empleando el criterio de Tukey y de forma multivariada mediante los puntajes de valor atípico local o LOF.

En la Figura 5.6 se muestra la distribución del precio de los vehículos usados mediante un diagrama de cajas. Haciendo uso del criterio de Tukey, fijando los límites como tres veces el rango intercuartílico siendo el límite inferior de 5 y el superior de 240 millones, no se identificaron datos atípicos en la variable de interés mediante este método.

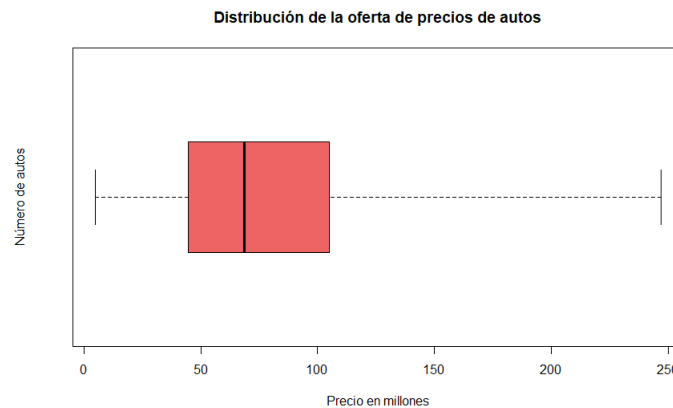


Figura 5.6: Diagrama de cajas del precio en millones de pesos.

Por otra parte, se calcularon los puntajes LOF de cada observación los cuales se muestran en la Figura 5.7. Se consideraron valores atípicos los registros que en sus puntajes superarán el percentil 99 siendo en este caso el valor de 39.900, de esta forma se identificaron un total de 25 registros los cuales están ubicados encima de la línea roja.

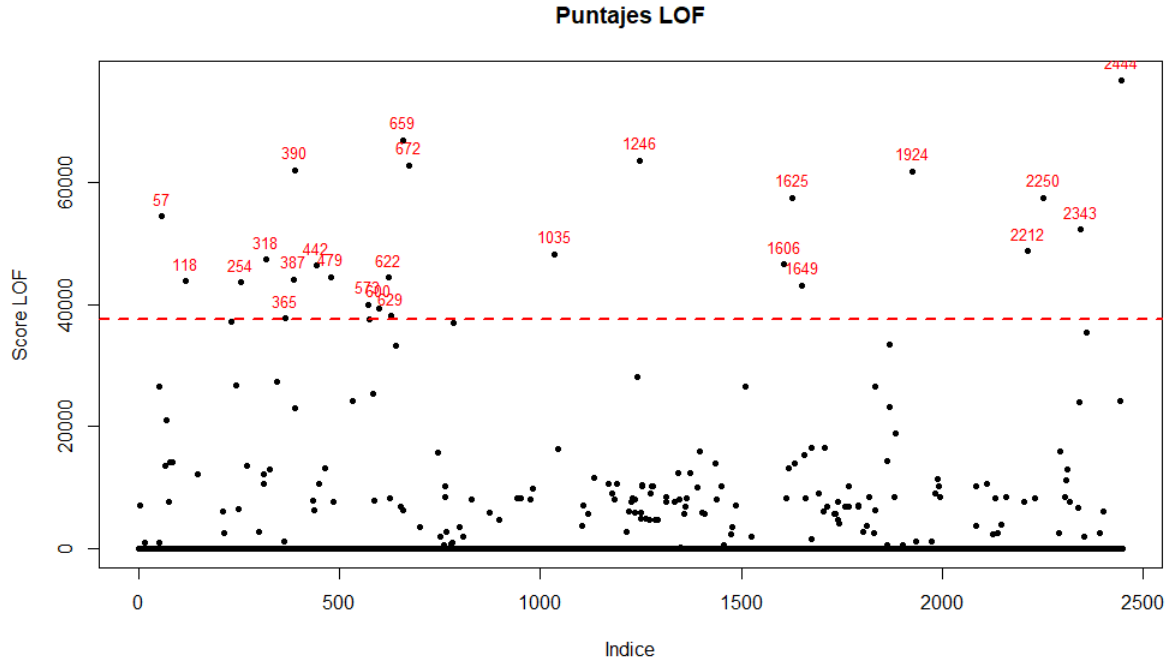


Figura 5.7: Distribución del puntaje LOF de las observaciones.

Estos datos atípicos se analizarán antes de ajustar el modelo de regresión y así poder concluir sobre su influencia en el desempeño de este.

El análisis exploratorio permitió obtener una visión integral del comportamiento de las características de los vehículos, logrando identificar patrones y relaciones entre las variables, lo que permitió conocer las variables más relacionadas con el precio. Se observó que las variables más influyentes en el precio son la antigüedad, el kilometraje, la marca y el tipo de carrocería. Esto es crucial para entender cómo las variables impactan el valor de los vehículos y lograr desarrollar modelos predictivos más precisos.

5.2 Modelación de la oferta de precios

Para el desarrollo de la modelación del precio de los vehículos usados se dividen los datos en dos subconjuntos, el 80% de estos hacen parte de los datos de entrenamiento que se emplean para ajustar los modelos estadísticos y de aprendizaje automático, mientras que el 20% restante serán los datos de prueba con los que se calculan las predicciones de los modelos y se evalúa su rendimiento para predecir nuevas observaciones.

El conjunto de datos final consta de 2,449 registros sin datos faltantes en donde 1,959 registros conforman el conjunto de entrenamiento y 490 registros el conjunto de prueba.

5.2.1 Fase de entrenamiento

Para el entrenamiento de los modelos se cuenta con un total de nueve variables predictoras del precio, las continuas son el kilometraje, cilindraje, antigüedad del vehículo y la variable calculada del número de kilómetros recorridos por año. Como variables categóricas se tienen la marca agrupada, tipo de carrocería, combustible y transmisión, y el número de puertas. Las variables de este tipo se convierten a variables dummy para su correcta implementación en los modelos, en donde se toma como categoría de referencia a la de mayor frecuencia. De esta forma quedan un total de 18 variables que serán empleadas para predecir el precio de los vehículos usados.

Hay que recordar que la variable marca fue agrupada para entrenar todos los modelos, estos grupos son:

- **Grupo A:** Chevrolet, FIAT, Hyundai, MG, Renault, Skoda, Suzuki, Zotye.
- **Grupo B:** Citroën, Dodge, Ford, Honda, JAC, KIA, Mazda, Nissan, Peugeot, RAM, Ssangyong, Subaru, Volkswagen.
- **Grupo C:** Audi, BMW, Jeep, Mercedes-Benz, MINI, Mitshubishi.
- **Grupo D:** Land Rover, Porsche, Toyota, Volvo.

El coeficiente de determinación y la raíz del error cuadrático medio serán calculados mediante el conjunto de datos de prueba.

5.2.2 Modelo lineal general

Se ajusta un Modelo lineal general mediante mínimos cuadrados ordinarios haciendo uso de todas las variables predictoras.

En la Figura 5.8, se evalúan los supuestos sobre los residuos del modelo en donde el precio fue transformado a priori mediante su logaritmo natural para estabilizar la varianza de los residuos y capturar las relaciones no lineales entre el precio y las covariables.

En la gráfica superior izquierda se evidencia que la distribución de los residuos estandarizados es simétrica con presencia de colas pesadas. Realizando el test de Shapiro-Wilk se rechaza la hipótesis de que los residuos tienen distribución normal. Por otra parte, en la gráfica inferior izquierda se muestran los valores ajustados del modelo frente a los residuos estandarizados, se observa un patrón cónico en donde hay una diferencia significativa de varianzas entre los residuos. De esta forma no se cumple la normalidad y homocedasticidad en los residuos.

Los supuestos no se cumplieron aun realizando transformaciones sobre las variables.

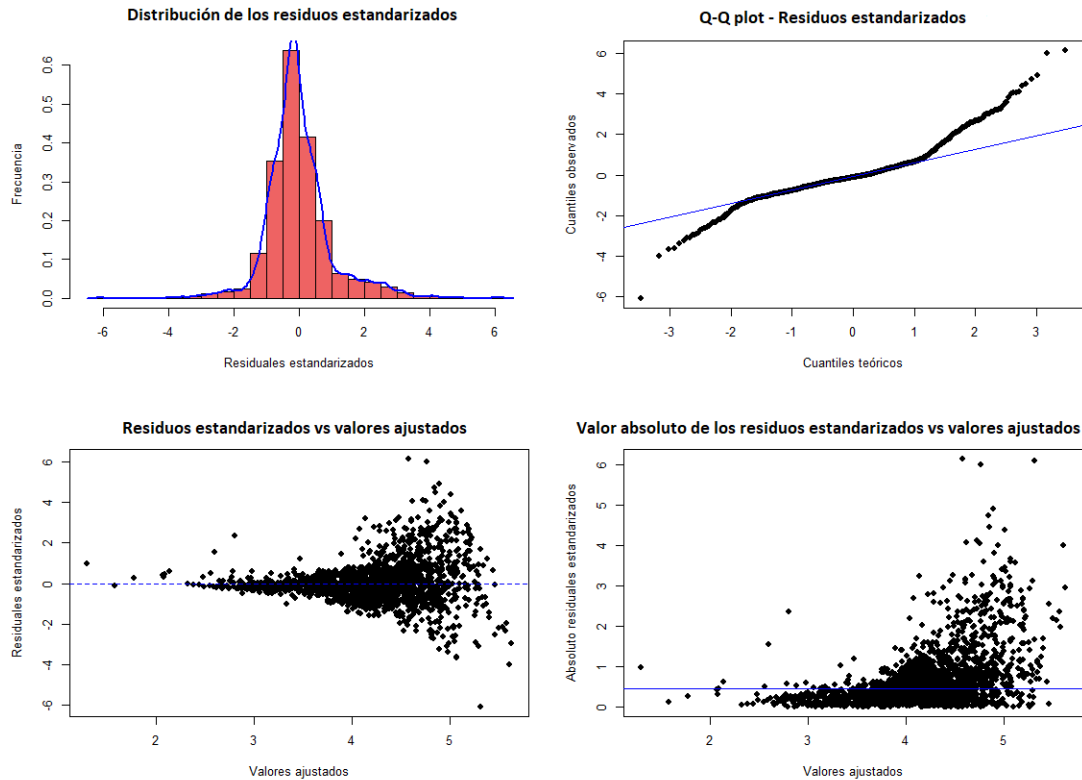


Figura 5.8: Supuestos sobre los residuos del modelo de regresión lineal mediante mínimos cuadrados ordinarios.

Cabe detallar que para este modelo, el coeficiente de determinación ajustado de predicción para los datos de prueba es de 0,7772 con un RECM de 22,1 millones de pesos.

Debido al incumplimiento de los supuestos de regresión se procede a evaluar la distribución del precio y su ajuste mediante el modelo lineal generalizado.

5.2.3 Modelo lineal generalizado gamma

Para el Modelo lineal generalizado gamma se incluirán en un inicio todas las covariables, lo que permitirá estudiar el efecto de estas sobre el precio, y posteriormente se realizará una selección de variables para encontrar el mejor modelo.

Antes de ajustar el modelo se evalúa la distribución del precio de los vehículos usados. En la Figura 5.9, en el gráfico izquierdo se muestra el histograma del precio y en el gráfico derecho el Q-Q plot del ajuste del precio bajo la distribución gamma.

Se evidencia que el precio se ajusta bien a la distribución gamma cuando los precios observados son menores a 150 millones. Para los precios entre 140 y 245 millones, los valores observados son un poco mayores a los teóricos bajo la distribución. Para los valores extremos

mayores a 245 millones, los valores observados son mucho menores a los teóricos. En general, el precio tiene un ajuste aceptable bajo la distribución gamma, sin embargo para precios mayores a 140 millones pueden presentarse errores de predicción un poco mayores al promedio cuando se evalúe el desempeño del modelo.

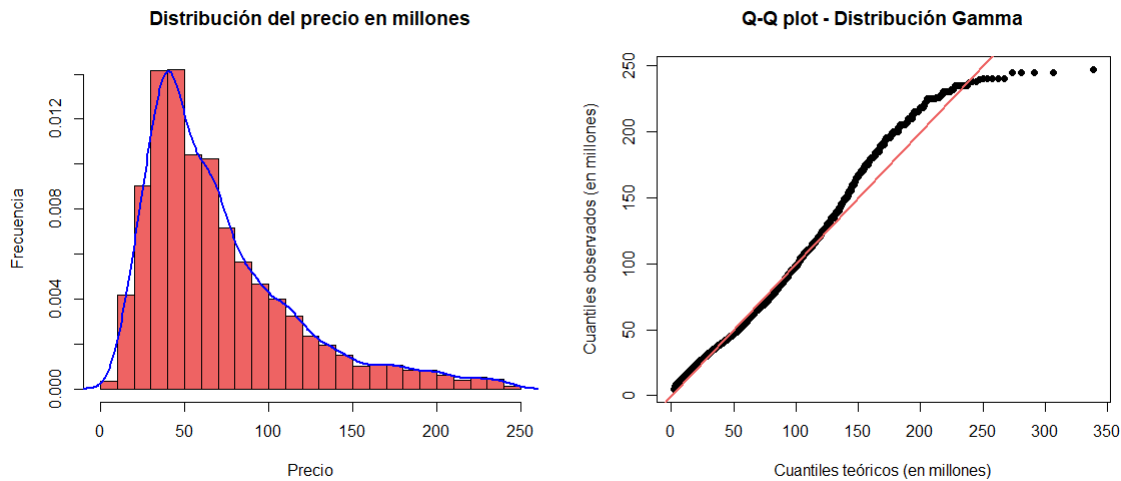


Figura 5.9: Distribución de densidad y bondad de ajuste bajo la distribución gamma del precio en millones de pesos.

Se ajusta un Modelo lineal generalizado bajo la distribución gamma con función de enlace logarítmica para capturar todas las relaciones entre el precio y las variables predictoras, además de que es la más adecuada ya que como se evidenció en el análisis de los residuales del Modelo lineal general, la media del precio aumenta con la media. El ajuste se desarrolla con los datos de entrenamiento y empleando desde un principio todas las covariables. En la Tabla 5.1 se muestran las estimaciones de los coeficientes del modelo con sus respectivas significancias:

Variable	Estimación	exp(estimación)	Error std.	P-valor
Intercepto	18.2948	88169499	0.0599	***
antigüedad	-0.0429	0.9580	0.0014	***
cilindraje	0.2210	1.2474	0.0119	***
kilómetros	-0.0156	0.9845	0.0012	***
razon_kil_antig	0.0000	1.0000	0.0000	0.0781
marca_Grupo_A	-0.1940	0.8237	0.0153	***
marca_Grupo_C	0.4151	1.5145	0.0200	***
marca_Grupo_D	0.4353	1.5455	0.0267	***
tipo_de_combustible_Diesel	0.2051	1.2256	0.0292	***
tipo_de_combustible_Gasolina_gas	-0.2558	0.7742	0.1350	***
tipo_de_combustible_Híbrido	-0.1232	0.8841	0.0587	0.0360
puertas_2_3	0.1173	1.1245	0.0404	0.0037
tipo_de_carrocería_Convertible	-0.1687	0.8448	0.0916	0.0655
tipo_de_carrocería_Coupé	-0.2691	0.7641	0.0755	***
tipo_de_carrocería_Hatchback	-0.3764	0.6863	0.0173	***
tipo_de_carrocería_Minivan	-0.2476	0.7807	0.0545	0.0854
tipo_de_carrocería_Pick	0.0431	1.044	0.0378	0.2543
tipo_de_carrocería_Sedán	-0.2872	0.7503	0.0161	***
transmisión_Mecánica	-0.1439	0.866	0.0171	***

Tabla 5.1: Resultados de los coeficientes del Modelo lineal generalizado gamma completo.

¹ Los códigos de significancia del p-valor están dados por: *** es un valor cercano a cero, ** un valor de 0,001, * un valor de 0,01 y . un valor de 0,05.

Como se puede evidenciar, empleando un nivel de significancia del 5 % se concluye que individualmente, las covariables kilómetros recorridos por año, los tipos de carrocería convertible, minivan y pick-up no resultan ser significativas dentro del modelo. Por otra parte se obtuvo un coeficiente de determinación ajustado de predicción de 0,7626, un RECM de 22,3 millones y un AIC de 70.710.

Selección de variables

Mediante el modelo completo se emplea el método "stepwise" para realizar la selección de variables, tanto empleando la técnica de paso hacia adelante como también paso hacia atrás. El modelo resultante tiene un AIC de 70.708, dos unidades menos con respecto al modelo completo, sin embargo en este modelo final no se incluyen las covariables del tipo de carrocería minivan y pick-up, las cuales resultaron no ser significativas en el modelo completo, por lo que la categoría de referencia del tipo de carrocería queda agrupada en camioneta, pick-up y minivan, siendo un total de 15 variables predictoras empleadas para explicar el precio de los autos.

En la Tabla 5.2 se muestran las variables predictoras y las estimaciones de los coeficientes del modelo final.

Debido a que en el modelo se utilizó la función de enlace logarítmica, es necesario exponenciar los coeficientes de regresión para interpretar correctamente el efecto de cada covariable sobre el precio.

Se evidencia que la marca y el cilindraje del automóvil son las covariables que presentan un mayor efecto positivo sobre el precio, debido a que sus coeficientes de regresión exponenciados son los más altos de todos. De esta forma si un vehículo es de las marcas que hacen parte del grupo D como Land Rover, Toyota y Volvo, el precio del auto siempre que las demás covariables se mantengan constantes, aumentará en un 54,1 % con respecto a las marcas del grupo B como Citroën, Dodge, Ford o Mazda. A diferencia de esto, las marcas del grupo A como Chevrolet, FIAT o Renault son las que menos efecto tienen sobre el precio, disminuyendo este en un 21,48 % con respecto a las marcas del grupo B.

Seguido de esto, el cilindraje también tiene un efecto positivo sobre el precio en donde por cada litro de cilindrada de más que posea el vehículo su precio aumentará en un 24,9 %

Por otra parte, el desgaste del vehículo tiene un efecto negativo sobre el precio en donde por cada diez mil kilómetros recorridos, se evidencia que el vehículo se devaluará en un 1,6 %. Además de esto se tiene que por cada año de antigüedad que tenga el vehículo su precio disminuirá en un 4,3 %

El hatchback es el tipo de carrocería que tiene mayor efecto negativo, disminuyendo el precio en un 31,5 % si el vehículo es de este tipo con respecto al tipo de carrocería camioneta.

En este modelo final se evidencia que los kilómetros recorridos por año es la única covariable que resulta no ser significativa para predecir el precio de los vehículos usados.

Variable	Estimación	exp(estimación)	Error std.	P-valor
Intercepto	18,3334	91638207,7200	0,0492	***
antigüedad	-0,0431	0,9578	0,0014	***
cilindraje	0,2229	1,2497	0,0118	***
kilómetros	-0,0156	0,9846	0,0012	***
razon_kil_antig	0,0000	1,0000	0,0000	0,0735
marca_Grupo_A	-0,1946	0,8231	0,0153	***
marca_Grupo_C	0,4131	1,5114	0,0198	***
marca_Grupo_D	0,4327	1,5413	0,0265	***
tipo_de_combustible_Diésel	0,2195	1,2454	0,0269	***
tipo_de_combustible_Gasolina_gas	-0,2497	1,2836	0,0268	***
tipo_de_combustible_Híbrido	0,0838	1,0874	0,0528	0,1117
puertas_2_3	0,1443	0,8656	0,0314	***
tipo_de_carrocería_Convertible	-0,1943	0,8234	0,0885	0,0283
tipo_de_carrocería_Coupé	-0,2956	0,7441	0,0712	***
tipo_de_carrocería_Hatchback	-0,3782	0,6851	0,0172	***
tipo_de_carrocería_Sedán	-0,2880	0,7497	0,0161	***
transmisión_Mecánica	-0,1409	0,8686	0,0169	***

Tabla 5.2: Resultados de los coeficientes del Modelo lineal generalizado gamma mediante la técnica stepwise.

¹ Los códigos de significancia del p-valor están dados por: *** es un valor cercano a cero, ** un valor de 0,001, * un valor de 0,01 y . un valor de 0,05.

Por último, se tiene que el modelo obtuvo un coeficiente de determinación ajustado de predicción para el conjunto de datos de prueba de 0,7621 y el RECM de 22 millones. De esta forma se tiene una reducción en la variabilidad explicada por este modelo del 0,04 % con respecto al modelo completo, una disminución que no resulta ser tan significativa y que por el contrario, resulta en un modelo con menos variables y más sencillo de implementar.

Diagnóstico del modelo

Para evaluar la confiabilidad de las estimaciones del Modelo lineal generalizado se realiza el diagnóstico y evaluación de los supuestos. En la Figura 5.12 se puede observar la distribución de los residuos estandarizados del modelo, estos cuentan con valores tanto negativos como positivos, en donde se evidencia cierta simetría teniendo una media de 0,56 millones. Además se calcula el coeficiente de asimetría de los residuos dando un valor de 0,71, muy cercano a cero por lo que se asume una simetría de en la distribución.

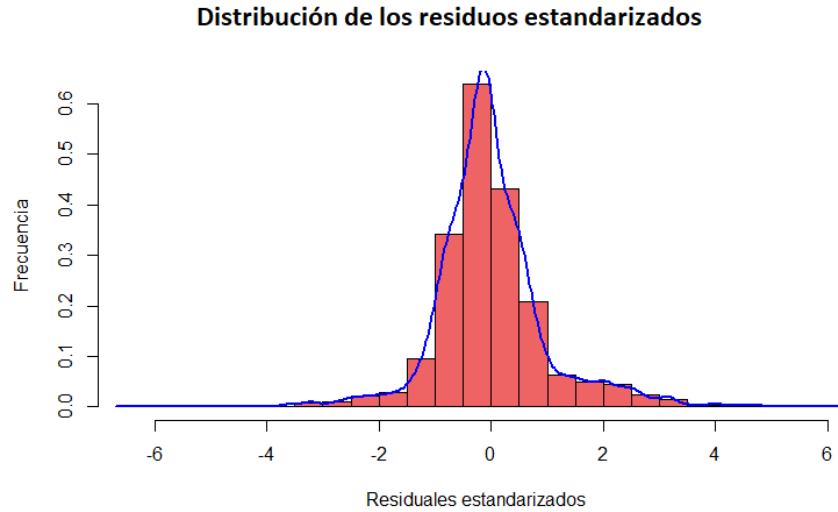


Figura 5.10: Distribución de los residuos estandarizados del Modelo lineal generalizado gamma.

Para evaluar el comportamiento de la varianza de los residuos, en la Figura 5.11 se muestran los valores ajustados frente a los residuos estandarizados. Se evidencia un patrón cónico de los residuos en donde a medida que aumenta la media de los valores ajustados también lo hace la varianza, algo esperado de acuerdo a la ecuación (3.14) en la cual se define la distribución gamma empleada en el modelo generalizado en donde la varianza de la variable respuesta también aumenta con la media.

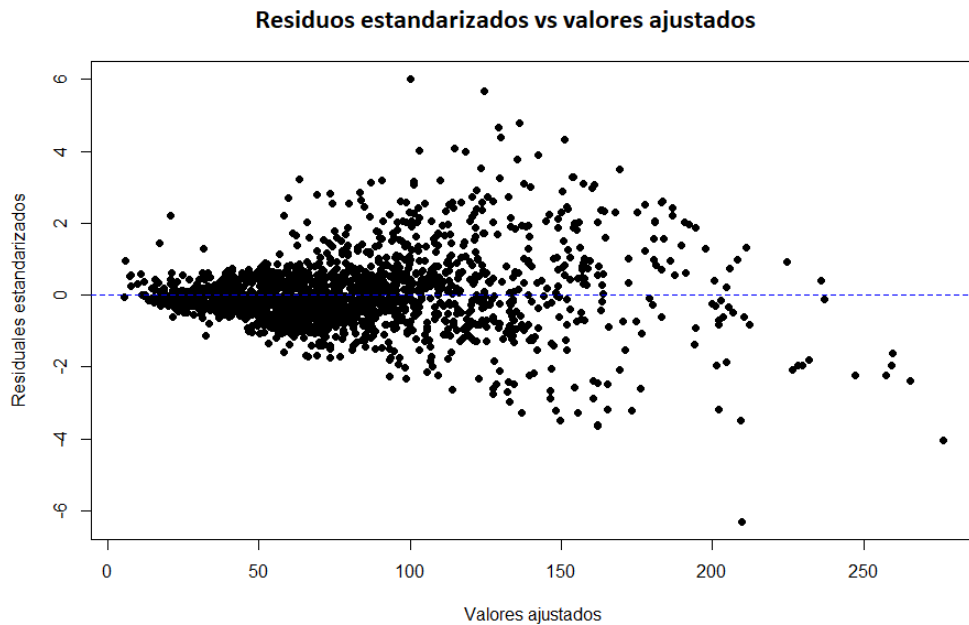


Figura 5.11: Valores ajustados frente a los residuos estandarizados del Modelo lineal generalizado gamma.

Para evaluar la presencia de multicolinealidad en el modelo, se calcularon los factores de inflación de varianza para cada variable. Se define que si el valor del factor está entre 5 y 10, la multicolinealidad que presenta dicha variable en el modelo es leve. Si el factor es mayor a 10, la variable presenta una multicolinealidad grave la cual es necesario examinar.

En la Figura 5.12 se muestran estos factores para cada variable, se puede observar que ninguna de las variables tiene un factor mayor a cinco, por lo que se concluye que no existe presencia de multicolinealidad en el modelo.

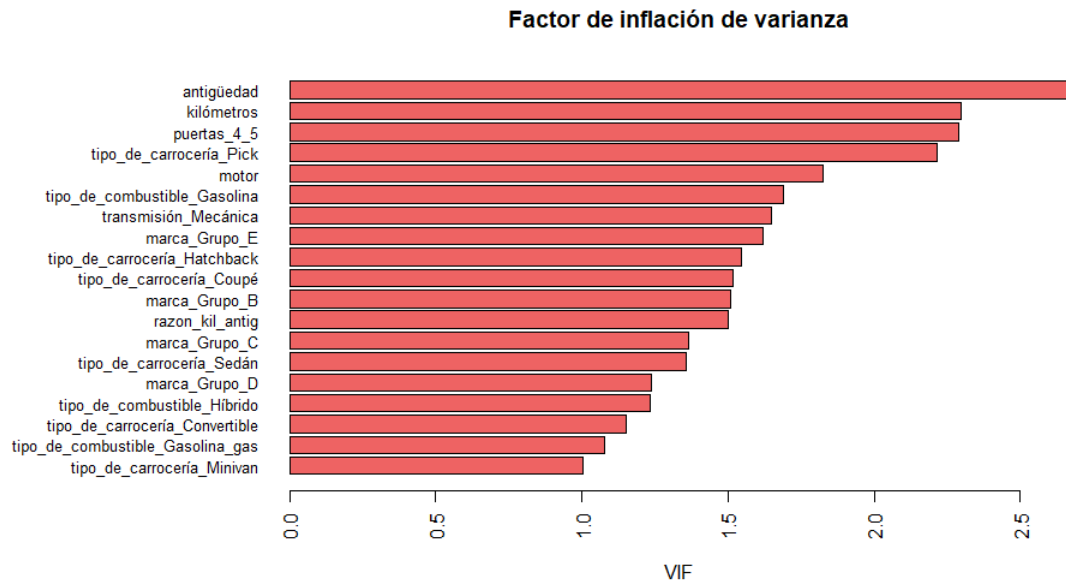


Figura 5.12: Factores de inflación de varianza de las variables predictoras del modelo generalizado gamma.

Por otra parte, también se evalúa el ajuste del modelo excluyendo los datos atípicos que se identificaron en el capítulo de análisis exploratorio. Se excluyeron los 25 registros atípicos encontrados y se ajustó el mejor modelo mediante el método "stepwise" en donde resultaron seleccionadas 15 covariables, excluyendo el tipo de combustible gasolina-gas y los tipos de carrocería convertible y minivan, por lo que en este modelo, las categorías de referencia para el combustible serán gasolina y gasolina-gas y para el tipo de carrocería será camioneta, convertible y minivan.

Se obtuvo un coeficiente de determinación de ajustado de predicción de 0,7636 con un RECM de 22,1 millones. Como se puede evidenciar se obtuvo un aumento del coeficiente de determinación ajustado de solo el 1,14%, por lo que la inclusión de estos datos atípicos no afecta de forma significativa al rendimiento del modelo.

5.2.4 Árbol de regresión

El modelo de Árbol de regresión se entrena haciendo uso de todas las covariables y se ajusta el árbol más complejo posible para su posterior poda, de esta forma se obtiene un modelo final el cual presenta el menor error de predicción teniendo en cuenta la complejidad del árbol. Se hace uso de la librería *Tree* de R y empleando validación cruzada en la cuál se usarán 5 subconjuntos de datos.

Para este modelo inicial se cuenta con 1.427 nodos terminales el cual obtuvo un coeficiente de determinación ajustado de predicción de 0,7257 y un RECM de 24 millones.

En la Figura 5.13 se muestra el comportamiento del error de validación cruzada frente al número de nodos terminales. En el modelo se tomará el valor de 1 en los hiperparámetros del número mínimo de registros en un nodo terminal y el número mínimo de registros necesarios para realizar la división de un nodo.

Se evidencia que con 16 nodos terminales se logra reducir el error de predicción sin incrementar tanto la complejidad del modelo, por lo que se elige este número de nodos para entrenar el árbol final. Haciendo uso de un mayor número de nodos no se lograría una reducción significativa del error y se incrementaría aun más la complejidad del árbol.

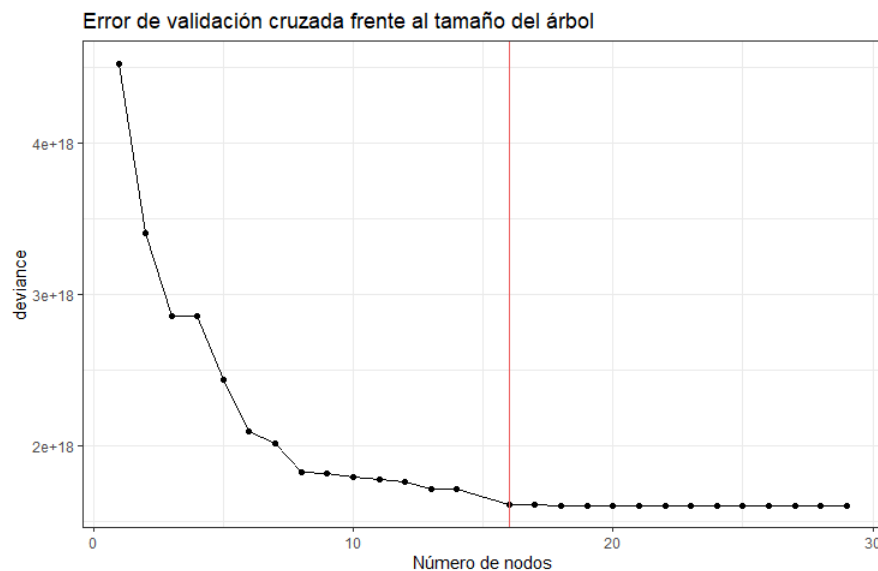


Figura 5.13: Error de validación cruzada frente al número de nodos terminales del Árbol de regresión.

En la Figura 5.14 se muestra la estructura del Árbol de regresión final. Se puede observar que la variable cilindraje es la que mayor importancia obtuvo ya que es la primera que condiciona los demás nodos de acuerdo al punto de corte de 1,675 litros. Este árbol solo incluye ocho variables para predecir el precio de los vehículos usados las cuales son el cilindraje, tipo de transmisión mecánica, las marcas del grupo B, grupo C, grupo D, antigüedad, tipo de carrocería hatchback y tipo de carrocería sedán.

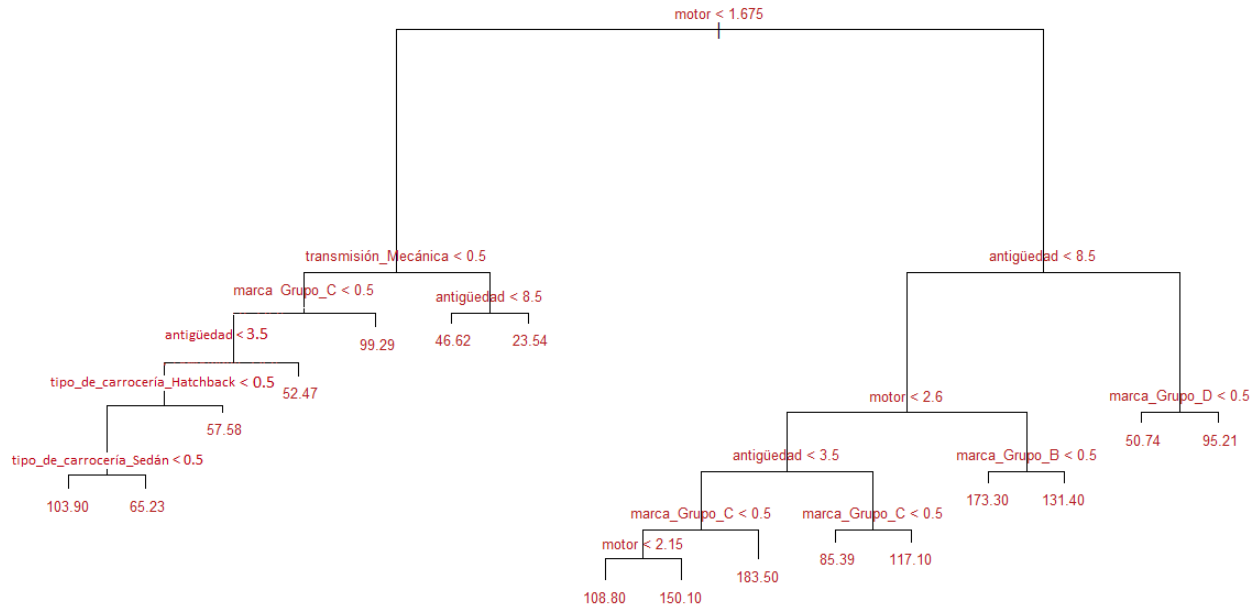


Figura 5.14: Esquema del modelo de Árbol de regresión.

En este modelo final se presenta un coeficiente de determinación ajustado de predicción de 0,6880, algo menor al conseguido con el modelo inicial, además se obtuvo un RECM de 25,9 millones.

Mediante la poda del árbol se obtiene una disminución del 5,1% en el coeficiente de determinación ajustado de predicción a costa de disminuir la complejidad del modelo.

5.2.5 Bosque aleatorio

En el entrenamiento del modelo de Bosque aleatorio se utiliza el método de validación cruzada haciendo uso de 5 subconjuntos y búsqueda en cuadrícula para optimizar los hiperparámetros mostrados en la Tabla 5.3 y los cuales fueron explicados en el capítulo de metodología. De esta forma el algoritmo entrena un modelo con cada combinación de hiperparámetros y evalúa su rendimiento mediante los cinco subconjuntos de validación cruzada, calculando el error de predicción de cada modelo. Se evaluaron los siguientes valores de hiperparámetros:

Hiperparámetro	Rango evaluado
bootstrap	True, False
max_depth	5, 10, 25, 27, 30, 35, 40, 45, 50, 60, 100
max_features	2, 5, 6, 8, 10, 12, 15, 18
min_samples_leaf	2, 3, 4, 5, 8, 10, 15
min_samples_split	2, 3, 4, 5, 8, 10, 15
n_estimators	500, 750, 1000, 1200, 1450, 1480, 1500, 1700, 2000

Tabla 5.3: Rangos para búsqueda en cuadrícula de los hiperparámetros del modelo de Bosque aleatorio.

El mejor modelo que logra minimizar el error de validación cruzada tiene los siguientes hiperparámetros:

- **bootstrap:** True.
- **max_depth:** 30.
- **max_features:** 8.
- **min_samples_leaf:** 2.
- **min_samples_split:** 2.
- **n_estimators:** 1.480.

El bosque consta entonces de un total de 1.480 árboles de regresión en donde cada uno tiene un máximo de 8 variables predictoras aleatorias. El número mínimo requerido de observaciones para dividir un nodo es de 2 y el número mínimo de observaciones que contiene cada hoja del árbol también es de 2. Además de esto, el bootstrap que se realiza para obtener la muestra para entrenar cada árbol se hace con reemplazo.

Este modelo obtuvo un rendimiento superior a los anteriores métodos empleados, obteniendo un coeficiente de determinación ajustado de predicción de 0,8607, además el RECM es de 17,1 millones.

En la Figura 5.15 se presenta la importancia que le da el modelo a cada una de las variables siendo la principal el cilindraje, seguida de la antigüedad del vehículo, los kilómetros recorridos y el tipo de transmisión mecánica. El tipo de carrocería más importante es el hatchback y para el tipo de combustible, el motor a gasolina. El grupo de marcas más importante es el C en donde hay marcas de gamma alta como BMW, Audi y Mercedes-Benz. Las variables menos relevantes son el tipo de carrocería minivan, seguida del tipo de combustible gasolina-gas y el tipo de carrocería convertible.

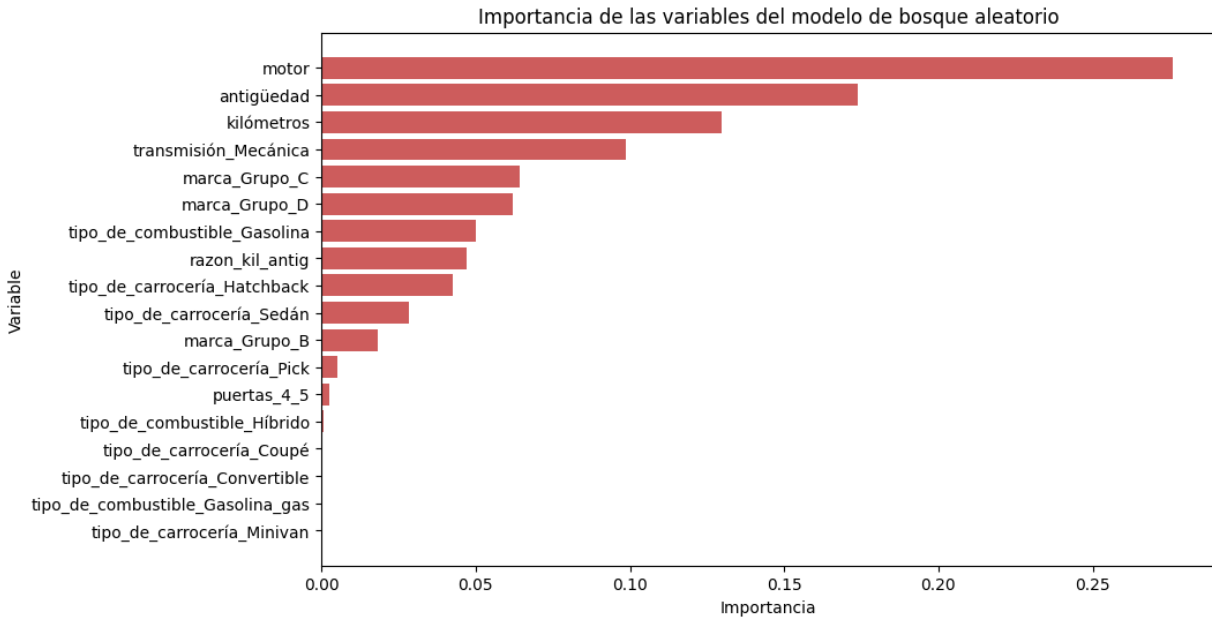


Figura 5.15: Importancia de las variables predictoras del modelo de Bosque aleatorio

5.2.6 XGBoost

En el ajuste del modelo de XGBoost también se utilizaron cinco subconjuntos de datos para validación cruzada y la búsqueda de los hiperparámetros del mejor modelo que logren reducir el error de predicción y el sobreajuste, el rango de búsqueda de los hiperparámetros se muestran en la Tabla 5.4:

Hiperparámetro	Rango evaluado
max_depth	3, 5, 6, 7, 8, 9, 15
n_estimators	50, 100, 500, 600, 650, 700, 750, 900, 1500
min_child_weight	2, 3, 4, 8, 15, 30
eta	0, 0,01, 0,02, 0,05, 0,1, 0,2, 0,4, 0,5, 0,6, 0,8, 1
gamma	0, 0,01, 0,02, 0,05, 0,1, 0,2, 0,4, 0,5, 0,6, 0,8, 1
lambda	0, 0,01, 0,02, 0,05, 0,1, 0,2, 0,4, 0,5, 0,6, 0,8, 1
alpha	0, 0,01, 0,02, 0,05, 0,1, 0,2, 0,4, 0,5, 0,6, 0,8, 1
subsample	0,4, 0,5, 0,55, 0,6, 0,8, 0,9, 1
colsample_bytree	0,2, 0,5, 0,6, 0,75, 0,68, 0,8, 0,82, 0,85

Tabla 5.4: Rangos para búsqueda en cuadrícula de los hiperparámetros del modelo de XG-Boost.

De esta forma se encuentra que los hiperparámetros del modelo que tuvo el menor error de validación cruzada fueron los siguientes:

- **n_estimators:** 750.
- **max_depth:** 7.

- **eta:** 1.
- **gamma:** 1.
- **alpha:** 0,66.
- **lambda:** 1.
- **colsample_bytree:** 0,78.
- **min_child_weight:** 2.
- **subsample:** 0,8.

Así, el modelo de XGBoost consta de 750 iteraciones o árboles de acuerdo al valor tomado por el hiperparámetro $n_estimators$, en donde el número máximo de niveles en estos es de 7, cada árbol consta de 14 covariables aleatorias y el 80 % de los datos fueron empleados para entrenar cada árbol. No se aplicaron penalizaciones sobre las salidas de los árboles mediante los hiperparámetros $gamma$, eta y $lambda$ ya que estos toman el valor de 1, sin embargo si se presenta regularización *LASSO* mediante el valor de $alpha$ igual a 0,66.

Con el modelo de XGBoost se logra el mejor ajuste con los datos de prueba obteniendo un coeficiente de determinación ajustado de predicción de 0,8675, además de tener el menor RECM de todos los modelos el cual es de 16,7 millones.

Por otra parte, en la Figura 5.16 se muestra la importancia de cada covariable en el modelo, siendo las marcas del grupo D como Land Rover, Toyota o Volvo, seguida de la transmisión mecánica, las marcas del grupo C como Audi, BMW o Jeep, y el tipo de combustible gasolina.

Las covariables que tuvieron un efecto menor o casi nulo sobre el precio son los tipos de carrocería minivan, convertible y el tipo de combustible gasolina-gas. Como se puede evidenciar, estos resultados son muy parecidos a los obtenidos en el modelo de Bosque aleatorio.

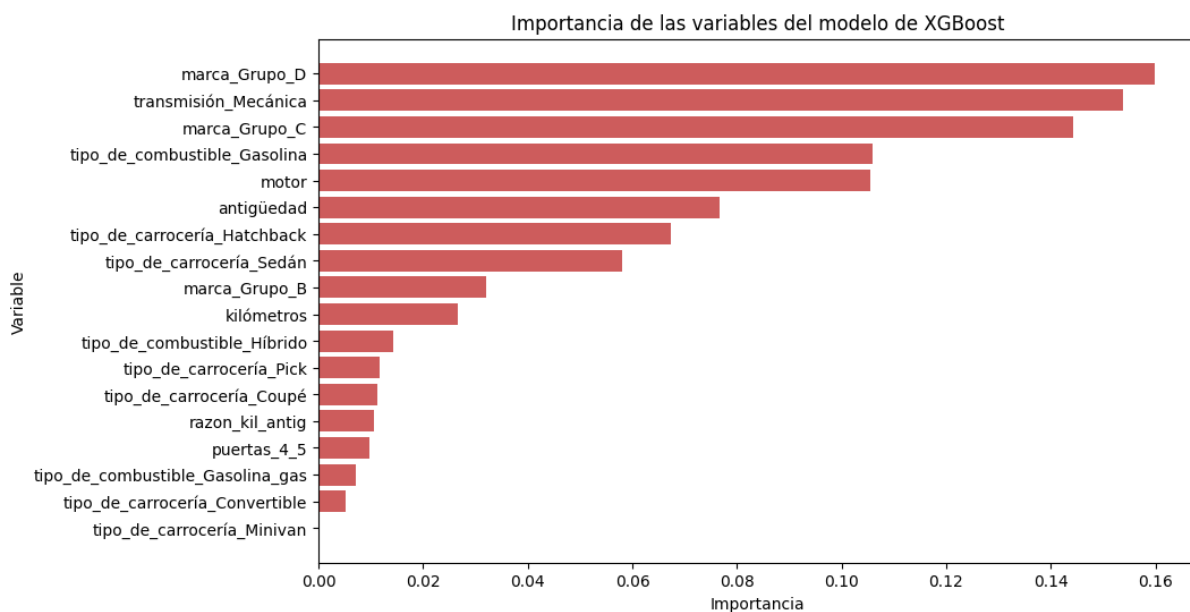


Figura 5.16: Importancia de las variables predictoras del modelo de XGBoost

5.2.7 Selección del mejor modelo mediante datos de prueba

Para seleccionar el mejor modelo predictivo que logre ofrecer el menor error de predicción se evalúa cada uno mediante validación cruzada con cinco subconjuntos de datos y se calcula el coeficiente de determinación ajustado promedio de predicción de los subconjuntos de prueba y entrenamiento.

En la Figura 5.17 se muestra el coeficiente de determinación ajustado de cada modelo tanto para los datos de entrenamiento como para los datos de prueba. Se puede evidenciar que el modelo de XGBoost es el que mejor rendimiento presenta con los datos de prueba ya que logra explicar el 86,7% de la variabilidad del precio de los autos usados mediante todas las covariables empleadas. Por otra parte, el coeficiente de determinación ajustado calculado mediante el conjunto de datos de entrenamiento es de 0,9690, por lo que se podría presentar un sobreajuste del modelo, sin embargo con datos no observados el XGBoost es el que presenta el mejor rendimiento.

El segundo mejor modelo es el Bosque aleatorio el cual presenta un coeficiente de determinación ajustado de 0,8619 con los datos de prueba, mientras que con los datos de entrenamiento es de 0,9490, en este también se aprecia un sobreajuste evidente.

El tercer modelo es el Modelo lineal generalizado bajo la distribución gamma mediante función de enlace logarítmica presentando un coeficiente de determinación ajustado de 0,7621 con datos de prueba y de 0,7976 con datos de entrenamiento. A pesar de tener un coeficiente menor en un 13,8% con respecto al modelo de XGBoost en los datos de prueba, presenta un coeficiente de determinación mucho menor con los datos de entrenamiento, por lo que su sobreajuste es mínimo y su desempeño con datos nuevos es bastante aceptable.

Por último, el modelo que obtuvo el menor desempeño es el Árbol de regresión con un coeficiente de determinación ajustado de 0,6880 con datos de prueba y 0,6956 con los de entrenamiento. A pesar de esto, es un modelo que no tiene un sobreajuste muy alto y se desempeña bien con datos no observados.

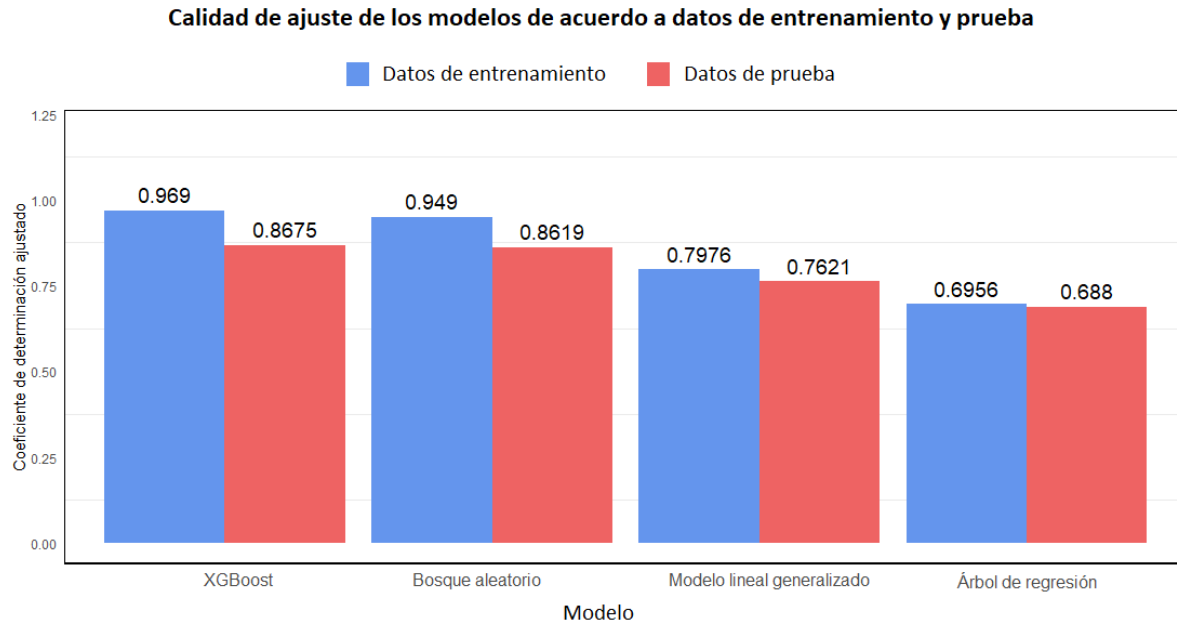


Figura 5.17: Coeficientes de determinación ajustado de los modelos frente a datos de prueba y entrenamiento mediante validación cruzada.

Por otra parte, empleando también el mismo método de validación cruzada se calcula la raíz del error cuadrático medio para el conjunto de datos de prueba en donde su distribución por cada modelo se puede ver en la Figura 5.18. En esta se muestra el mínimo, máximo, promedio y el comportamiento de la varianza del RECM de cada modelo. Se puede concluir que el modelo que obtuvo el menor RECM promedio fue el XGBoost, en dónde la varianza de este no es tan grande y la predicción de este modelo tiene una diferencia promedio de 16,7 millones con respecto al valor real. Por otra parte, el Árbol de regresión es el que tiene peor rendimiento con un RECM promedio de 27,5 millones de pesos, una diferencia del 64 % con respecto al XGBoost.

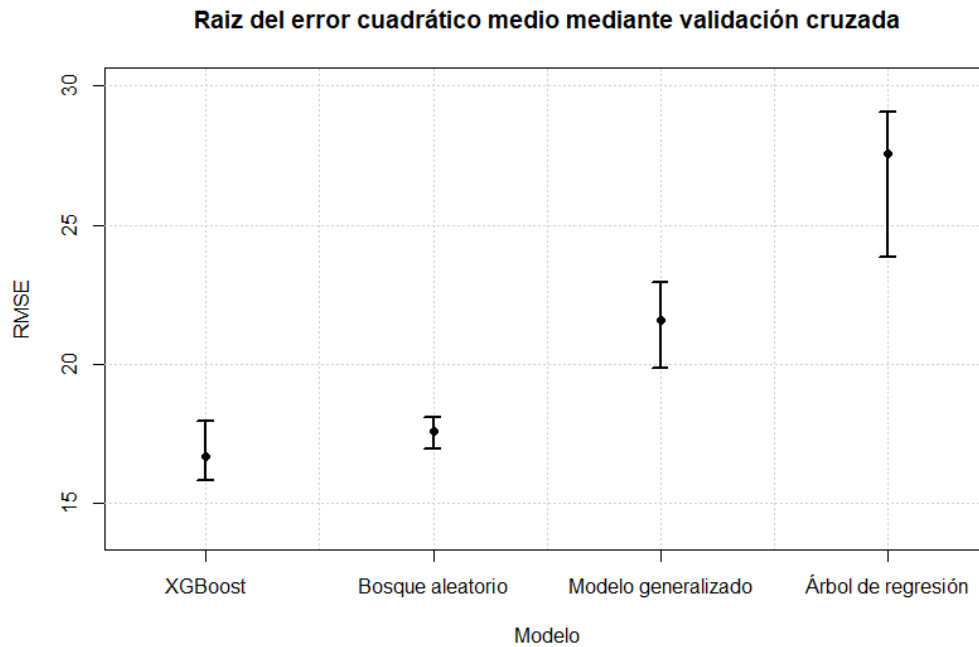


Figura 5.18: Raíz del error cuadrático medio mediante validación cruzada para datos de prueba.

Como se puede evidenciar mediante todos los modelos entrenados, las covariables que más afectan al precio son el kilometraje, cilindraje, antigüedad y la marca del vehículo. Estas variables muestran una influencia significativa y consistente en el valor del automóvil. Por otra parte, existe una correlación notable entre la antigüedad y el kilometraje, donde se espera que a medida que aumente la antigüedad, también lo haga el kilometraje. Esta relación es lógica, ya que los vehículos más antiguos han tenido más tiempo para acumular kilómetros. Comprender estas relaciones es fundamental para mejorar la precisión de los modelos predictivos y para tomar decisiones en la valoración de los vehículos usados en la ciudad de Cali.

Para complementar, el modelo de XGboost se implementó en una aplicación de R Shiny en donde cualquier persona puede consultar el precio promedio que estima el modelo para comprar o vender un vehículo proporcionando las características del auto en el que se está interesado como se puede ver en la Figura 5.19. Este modelo puede consultarse desde el siguiente enlace web.

Predicción del precio de autos usados en Cali



Figura 5.19: Aplicación del modelo XGBoost del precio de autos usados en R Shiny

Por último, la rutina de Web Scrapping y el modelo XGBoost se dispusieron en un repositorio de Github en donde se puede acceder libremente al código implementado en el desarrollo del trabajo de grado, los cuales se pueden acceder desde el siguiente enlace.

6 Conclusiones y recomendaciones

6.1 Conclusiones

En este Trabajo de grado se realizó la modelación del precio de los vehículos usados en la ciudad de Cali para el año 2023 mediante los datos proporcionados por los vendedores en la plataforma de compra-venta Mercado Libre Colombia los cuales fueron obtenidos mediante la técnica de Web Scraping desarrollada en el lenguaje de programación R.

La modelación se llevó a cabo mediante modelos estadísticos y de aprendizaje automático, siendo un total de cinco modelos ajustados. Se encontró el mejor Modelo lineal generalizado mediante la técnica de stepwise usando como criterio de selección al AIC. Por otra parte, para encontrar el mejor modelo de cada método de aprendizaje automático, se optimizaron los hiperparámetros y se evaluaron mediante validación cruzada tomando como medidas de desempeño la raíz del error cuadrático medio y el coeficiente de determinación ajustado.

Se determinó que el mejor modelo para predecir el precio de los autos usados en Cali es el XGBoost, ya que tiene un error promedio de las estimaciones de 16,7 millones de pesos con respecto al valor real y en donde el modelo explica un 86,7 % de la variabilidad del precio mediante las características del vehículo.

De acuerdo con el Modelo lineal generalizado, se concluyó que la marca agrupada mediante Árbol de decisión es la variable que tiene el mayor efecto en el precio del vehículo. Las marcas del grupo D, como Land Rover, Toyota o Volvo, aumentan su precio en un 54,1 % en comparación con las marcas del grupo B, que incluyen Citroën, Ford, Nissan y Volkswagen. Por otro lado, las marcas del grupo A, como Chevrolet, Renault, Skoda y Suzuki, tienden a tener una disminución del precio en un 21.5 % en referencia a las marcas del grupo B.

Además, se evidenció que, manteniendo las demás variables constantes, el precio del vehículo disminuirá en un 1.6 % por cada 10.000 kilómetros recorridos y en un 4.6 % por cada año de antigüedad del vehículo.

En general, mediante los cuatro modelos ajustados se evidencia que las variables más importantes para predecir el precio de un vehículo usado son el kilometraje, cilindraje, marca, tipo de carrocería y tipo de transmisión, siendo la marca la que obtuvo el mayor peso en las predicciones.

La estimación de los precios de los vehículos realizado en este estudio proporciona una referencia tanto a vendedores como compradores con respecto al precio promedio al que se vende un vehículo usado con ciertas características en la ciudad de Cali. Sin embargo, el precio final de venta puede ser menor debido a las negociaciones que se realicen entre comprador y vendedor.

La técnica de Web Scraping fue de gran ayuda para la obtención de los datos estructurados del sitio web de Mercado Libre. Mediante técnicas de preprocesamiento y limpieza de datos se transformaron y homologaron las variables para su posterior análisis. Algo que cabe destacar del Web Scraping es que gracias a su facilidad de implementación, permite recolectar grandes cantidades de datos sin la necesidad de realizar una inversión monetaria adicional en la adquisición de estos, lo que abre las puertas a estudios que no se hayan podido realizar anteriormente debido a la dificultad para obtener los datos.

6.2 Recomendaciones

Para posteriores estudios o proyectos relacionados, no solo con los precios de autos usados, si no también con precios de otro tipo de bien en donde se emplee el Web Scraping como técnica de obtención de datos, se recomienda desarrollar la automatización del preprocesamiento y limpieza de datos mediante lenguajes de programación como Python o R. En R, existen librerías especializadas para el tratamiento de datos, como *dplyr*, que permite la manipulación, filtrado, selección y mutación de datos. La librería *mice*, la cual se emplea para la imputación de datos faltantes, en donde para este caso, se realizó la imputación en la variable cilindraje mediante regresión. *Stringr* ofrece la manipulación de cadenas de texto, lo que permite limpiar y homogeneizar correctamente las categorías de las variables como marca y modelo. Por último, *lubridate* permite definir correctamente el formato de fecha de publicación de cada oferta.

También se tiene la posibilidad de actualizar a tiempo real los datos mediante el algoritmo para así poder tener un mayor número de ofertas que puedan ser estudiadas e implementadas en el modelo. Esto permitiría tener conocimiento exacto de como es el comportamiento de los precios diariamente, además también se podrían estudiar los precios con respecto al tiempo y evaluar como es su comportamiento de acuerdo a situaciones políticas y sociales, tanto a nivel nacional como internacionalmente. Automatizar estos procesos no solo mejora la eficiencia y la reproducibilidad de la hoja de datos, sino que también facilita la actualización y el mantenimiento del modelo predictivo.

Por otra parte, se recomienda entrenar más métodos de aprendizaje automático como las redes neuronales o la maquina de soporte vectorial con las cuales se pueda lograr un mayor desempeño en el ajuste del precio de los vehículos.

Además, si se desea predecir el precio para cada ciudad del país, se recomienda agrupar las ciudades por departamentos o regiones. Dado el elevado número de ciudades y corregimientos, es posible que en muchas localidades haya pocas ofertas de vehículos. Esto puede

llevar a que la muestra para ciertos modelos de vehículos sea insuficiente, lo que podría afectar la precisión de las predicciones. En casos donde no haya ofertas de un modelo específico en una localidad, podría ser imposible estimar su precio de manera confiable.

Mediante la técnica de Web Scraping se obtuvo de forma libre la oferta de vehículos usados en Cali desde un solo sitio web el cual es Mercado Libre, por lo que se propone ampliar el número de plataformas desde donde obtener más ofertas de vehículos, pero evaluando las posibles ofertas duplicadas ya que los usuarios suelen realizarlas en varios sitios web a la vez. Es importante identificar estos registros estudiando la similitud de las variables de los vehículos, lo que es posible empleando métodos de similitud que hacen uso de distancias entre vectores, clustering o modelos de aprendizaje automático. Implementando estas técnicas, es posible mejorar la calidad de los datos y minimizar el impacto de los registros duplicados.

Finalmente, se puede enriquecer la información de las ofertas de vehículos incluyendo características técnicas específicas de cada modelo, como el peso en vacío, el número de pasajeros y los caballos de fuerza, las cuales pueden ser identificadas mediante la marca, el modelo, el cilindraje y el año del vehículo. Para acceder a estas nuevas variables, se pueden consultar sitios web especializados que proporcionan las características técnicas detalladas de cada modelo.

Referencias

- Barnett, V., y Lewis, T. (1994). *Outliers in statistical data* (Vol. 3). John Wiley Sonsg.
- Barrios, S., y Quinto, K. (2016). Modelación del precio de la oferta de vivienda en venta de la ciudad de cali, considerando variables propias del activo y covariables de su entorno.
- Behar, R. (2003). *Validación de supuestos en el modelo de regresión* (Vol. 1). Universidad del Valle, Escuela de Ingeniería Industrial y Estadística, Facultad de Ingeniería.
- Breiman, L. (1996). Bagging predictors: Machine learning. *Kluwer Academic Publishers*. doi: 10.1007/BF00058655.
- Breiman, L. (2001). Random forests: Machine learning. *Kluwer Academic Publishers*, 45. doi: 10.1023/A:1010933404324
- Breunig, M. M., Kriegel, H. P., Ng, R. T., y Sander, J. (2000). Lof: Identifying density-based local outliers. *SIGMOD Record (ACM Special Interest Group on Management of Data)*, 29. doi: 10.1145/335191.335388
- Brigham, E., y Houston, J. (2016). *Fundamentals of financial management*. Cengage Learning.
- Brown, P. J., y Lehmann, E. L. (1984). Theory of point estimation. *Journal of the Royal Statistical Society. Series A (General)*, 147. doi: 10.2307/2981857
- California State Legislature. (2018). *California consumer privacy act of 2018, cal. civ. code 1798.100 et seq.* Descargado de https://leginfo.legislature.ca.gov/faces/codes_displayText.xhtml?division=3.&part=4.&lawCode=CIV&title=1.81.5 (Accedido: 2024-05-25)
- Chen, T., y Guestrin, C. (2016). Xgboost: A scalable tree boosting system. En (Vol. 13-17-August-2016). doi: 10.1145/2939672.2939785
- de Toro et al, G. (2017). Un modelo hedónico de precios en línea de automóviles usados en argentina. *Revista de Metodos Cuantitativos para la Economia y la Empresa*, 24.
- Evans, M., y Rosenthal, J. (2003). *Probability and statistics* (Second Edition ed.). University of Toronto.
- Gutierrez, A. (2022, 3). Industria reporta que los precios de los vehículos usados crecieron entre 15 % y 30 %. *La Republica*.
- Gutierrez, D. (2022). Evaluación del efecto de la pandemia covid-19 sobre el precio de la vivienda en la ciudad de santiago de cali.
- Hastie, T., y Tibshirani, R. (2008). *The elements of statistical learning: data mining, inference, and prediction* (2.^a ed.). Springer.
- Kambe, M., Ha, J., y Pe, J. (2011). *Data mining: Concepts and techniques* (Third edition ed.). Morgan Kaufmann. doi: 10.1016/C2009-0-61819-5

- KBB. (2022). How to beat car depreciation. *Kelley Blue Book*. Descargado de <https://www.kbb.com/car-advice/how-to-beat-car-depreciation/> (Accedido: 2024-07-26)
- Kihm, A., y Vance, C. (2016). The determinants of equity transmission between the new and used car markets: A hedonic analysis. *Journal of the Operational Research Society*, 67. doi: 10.1057/jors.2016.16
- Koç, M., y Kostak, N. (2021). What is happening to used car prices in turkey? evidence from hedonic price model (hpm). *Selçuk Üniversitesi Sosyal Bilimler Enstitüsü Dergisi*. doi: 10.52642/susbed.909818
- Liu, E., y Li, J. (2022). Research on the prediction model of the used car price in view of the pso-gra-bp neural network. *Sustainability (Switzerland)*, 14. doi: 10.3390/su14158993
- Mankiw, N. (2012). *Principles of economics* (6.^a ed.). Cengage Learning.
- Mccullagh, P., y Nelder, J. A. (1986). *Generalized linear models* (Second edition ed., Vol. 28). Chapman and Hall. doi: 10.1137/1028043
- Parlamento Europeo y Consejo de la Unión Europea. (2016). *Reglamento (ue) 2016/679 del parlamento europeo y del consejo de 27 de abril de 2016 en los que se deroga la directiva 95/46/ce (reglamento general de protección de datos)*. Descargado de <https://eur-lex.europa.eu/eli/reg/2016/679/oj> (Accedido: 2024-05-25)
- Refaeilzadeh, P., Tang, L., y Liu, H. (2009). Cross-validation.. doi: 10.1007/978-0-387-39940-9_565
- Reinsel, G., Hogg, R. V., y Craig, A. T. (1980, 6). Introduction to mathematical statistics (4th ed.). *Journal of the American Statistical Association*, 75, 474. doi: 10.2307/2287488
- Restrepo, O. (2023, 5). Venta de carros usados: ¿negocio jugoso ante escasez de autos nuevos? *El Carro Colombiano*.
- Tukey, J. W. (1977). *Exploratory data analysis* (Vol. 33).

Anexo

A. Anexo: Hoja de datos de la oferta de vehículos usados en Cali en septiembre de 2023.

marca	modelo	nombre	precio	ciudad	año	antigüedad	versión	color	tipo_de_combustible	puertas	transmisión	motor	tipo_de_carroceria	kilómetros	id	fecha	enlace
Mazda	3	Mazda 3 2.0 Tou	110000000	Cali - Valle Del Cau	2023	0	2.0 Touring		Gasolina	4	Automática	2	Sedán	8000	#1341030429	2023	8.000 kr https://carro.m
Mazda	3	Mazda 3 2.0 Tou	115000000	La Hacienda - Cali -	2023	0	2.0 Touring	Gris	Gasolina	4	Automática	2	Sedán	11830	#1337041259	2023	11.830 l https://carro.m
Mazda	3	Mazda 3 Tourin	99000000	Valle Del Lili - Cali	2023	0	2.0 Touring	Gris	Híbrido	4	Automática	2	Sedán	11650	#1339817931	2023	11.650 l https://carro.m
Peugeot	3008	Peugeot 3008 1.	117000000	La Flora - Cali - Val	2023	0	1.6 Active	Plateado	Gasolina	5		1,6	Camioneta	11387	#1787096360	2023	11.387 l https://carro.m
Peugeot	3008	Peugeot 3008 1.	132000000	Cali - Valle Del Cau	2023	0	1.6 Allure		Gasolina	5	Automática	1,6	Camioneta	9500	#1861033620	2023	9.500 kr https://carro.m
Peugeot	3008	Peugeot 3008 1.	129990000	Valle Del Lili - Cali	2023	0	1.6 Active	Gris	Gasolina	5		1,6	Camioneta	9300	#1338262755	2023	9.300 kr https://carro.m
RAM	700	Ram 700 Bighor	83000000	El Refugio - Cali - Vi	2023	0	Mecanica	Gris	Gasolina	4	Mecánica	1,4	Pick-Up	595	#1324041033	2023	595 km https://carro.m
Renault	Alaskan	Renault Alaskan	179990000	Salomia - Cali - Val	2023	0	2.5 Intens	Plateado	Diesel	2 y 3	Automática	2,5	Pick-Up	9900	#1338102399	2023	9.900 kr https://carro.m
Citroen	C4	Citroen C4 Sheir	92000000	Cali - Valle Del Cau	2023	0	SHEIN		Gasolina	5		1,76	Camioneta	11000	#1869228652	2023	11.000 l https://carro.m
Chevrolet	Captiva	Chevrolet Captiv	125000000	El Refugio - Cali - Vi	2023	0	Automatico	Plateado	Gasolina	5	Automática	1,77	Camioneta	5800	#1905249214	2023	5.800 kr https://carro.m
Renault	Captur	Renault Captur	88990000	Salomia - Cali - Val	2023	0	2.0 Intens Autotri	Plateado	Gasolina	5	Automática	2	Camioneta	12100	#1940240222	2023	12.100 l https://carro.m
Honda	City	Honda City Ex Se	94000000	La Alianza - Cali - V	2023	0	EKL	Plateado	Gasolina	4	Automática	1,5	Sedán	8746	#1913472532	2023	8.746 kr https://carro.m
Mercedes-B	Clase CLA	Mercedes Benz C	184000000	Capri - Cali - Valle	2023	0	200	Plateado	Gasolina	4	Automática	1,25	Sedán	5000	#1325404307	2023	5.000 kr https://carro.m
Mercedes-B	Clase CLA	Mercedes-benz C	185000000	Cali - Valle Del Cau	2023	0	200 AMG LINE		Gasolina	5	Automática	1,77		5700	#1333398311	2023	5.700 kr https://carro.m
Chevrolet	Colorado	Chevrolet Colori	189000000	Capri - Cali - Valle	2023	0	2.8 Ltz	Plateado	Diesel	4	Automática	2,8	Pick-Up	4671	#1320964069	2023	4.671 kr https://carro.m
MINI	Cooper	MINI Cooper 2.0	167000000	Cali - Valle Del Cau	2023	0	2.0 F55 S Cooper Salt		Gasolina	5	Automática	2	Hatchback	2000	#1331863073	2023	2.000 kr https://carro.m
MINI	Cooper	Mini Cooper S ic	172900000	Pance - Cali - Valle	2023	0	ICONIC	Gris	Gasolina	2 y 3	Automática	2	Coupé	1236	#1330537697	2023	1.236 kr https://carro.m
Toyota	Corolla	Toyota Corolla C	137000000	Cali - Valle Del Cau	2023	0	Cross 2.0 GR	Rojo	Gasolina	4	Automática	2	Sedán	4100	#1846290268	2023	4.100 kr https://carro.m
Toyota	Corolla	Toyota Corolla C	118000000	Cali - Valle Del Cau	2023	0	GR SPORT		Gasolina	4		1,83	Sedán	11000	#1902111744	2023	11.000 l https://carro.m
Toyota	Corolla	Toyota Corolla i	156200000	Cali - Valle Del Cau	2023	0	hibrida 2023		Híbrido	2		1,76	Sedán	6500	#1331535251	2023	6.500 kr https://carro.m
Toyota	Corolla	Toyota Corolla i	140000000	Cali - Valle Del Cau	2023	0	1.8 Se-g		Gasolina	4	Automática	1,8	Sedán	30000	#1340309445	2023	30.000 l https://carro.m
Toyota	Corolla	Toyota Corolla i	120500000	Cali - Valle Del Cau	2023	0	1.8 Xe-i		Gasolina	4	Automática	1,8	Sedán	5999	#1333138799	2023	5.999 kr https://carro.m
Toyota	Corolla	Toyota Corolla i	175000000	Cali - Valle Del Cau	2023	0	1.8 Xe-i		Gasolina	4	Automática	1,8	Sedán	60	#183911510	2023	60 km - https://carro.m
Toyota	Corolla	Toyota Corolla i	118500000	Cali - Valle Del Cau	2023	0	1.8 Se-g	Plateado	Híbrido	5	Automática	1,8	Sedán	7700	#1913251888	2023	7.700 kr https://carro.m
Toyota	Corolla	Toyota Corolla i	115000000	Cali - Valle Del Cau	2023	0	1.8 Xe-i		Gasolina	4	Automática	1,8	Sedán	2000	#1340726769	2023	2.000 kr https://carro.m
Toyota	Corolla	Toyota Corolla i	148700000	Cali - Valle Del Cau	2023	0	1.8 Híbrido SEG	Blanco	Híbrido	5	Automática	1,8	Sedán	16500	#1335890013	2023	16.500 l https://carro.m
Mazda	CX-30	Mazda Cx-30 2.0	122000000	Cali - Valle Del Cau	2023	0	2.0 GRAN TOURING HIB		Híbrido	5	Automática	2	Camioneta	13000	#1341429705	2023	13.000 l https://carro.m
Mazda	CX-30	Mazda Cx-30 Gr	119800000	Cali - Valle Del Cau	2023	0	Grand Touring 2.0 AT		Híbrido	4	Automática	2	Camioneta	16000	#1910955844	2023	16.000 l https://carro.m
Mazda	CX-30	Mazda Cx-30 2.0	119000000	Cali - Valle Del Cau	2023	0	2.0 GRAN TOURING		Híbrido	5	Automática	2	Camioneta	30000	#1341430359	2023	30.000 l https://carro.m
Mazda	CX-30	Mazda Cx-30 Gr	126000000	Cali - Valle Del Cau	2023	0	Grand Touring hibrid		Híbrido	5	Automática	2	Camioneta	19350	#1910979574	2023	19.350 l https://carro.m
Mazda	CX-30	Mazda Cx30 20i	132000000	San Bosco - Cali - V	2023	0	GRAND TOURING	Plateado	Híbrido	5	Automática	2	Camioneta	14000	#1925563200	2023	14.000 l https://carro.m
Mazda	CX-30	Mazda Cx-30 HII	117000000	Cali - Valle Del Cau	2023	0	hibrida		Híbrido	5	Automática	2	Camioneta	33100	#1834421152	2023	33.100 l https://carro.m
Mazda	CX-30	Mazda Cx 30 20	149000000	Valle Del Lili - Cali	2023	0	2.0 Grand Touri	Blanco	Híbrido	5	Automática	2	Camioneta	17650	#133660245	2023	17.650 l https://carro.m

Figura 6.1: Hoja de datos de la oferta de vehículos usados