

**Estimación de los precios de vehículos en las ciudades de Bogotá, Cali, Medellín y Barranquilla
utilizando Web Scraping, Modelos de Regresión y Machine Learning**

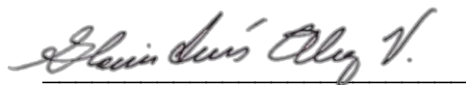
Víctor Hugo Moncayo Ruiz - Autor

Nota de Aceptación

Certificamos que el presente Trabajo de Grado Satisface,
en alcances y calidad, todos los requisitos que demanda
un Trabajo de Grado de Maestría.



David Arango Londoño - Director



Gloria Inés Álvarez - Jurado

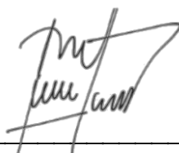


Luis Fernando Macea - Jurado

Aprobado en cumplimiento de los requisitos exigidos por la
Pontificia Universidad Javeriana Cali, para optar el título
de Magister en Ingeniería.



HERNÁN CAMILO ROCHA NIÑO Ph. D.
Decano Facultad de Ingeniería y Ciencias



JUAN CARLOS MARTÍNEZ ARIAS
Director Posgrados de Ingeniería y Ciencias

Santiago de Cali, 29 de Agosto de 2022



Acta de Correcciones al Documento de Trabajo de Grado

Santiago de Cali, 12 de Agosto de 2022

Autor: Víctor Hugo Moncayo Ruiz

Título del Trabajo de Grado: “Estimación de los precios de vehículos en las ciudades de Bogotá, Cali, Medellín y Barranquilla utilizando Web Scraping, Modelos de Regresión y Machine Learning”

Director:

Como indica el artículo 2.13 de las Directrices para Trabajo de Grado de Maestría, he verificado que el estudiante indicado arriba ha implementado todas las correcciones que los Jurados del Proyecto de Trabajo de Grado definieron que se efectuaran, como consta en el Acta de Evaluación correspondiente.

David Arango Londoño

Firma del Director del Trabajo de Grado

DATOS DEL ESTUDIANTE:

NOMBRE COMPLETO: Víctor Hugo Moncayo Ruiz

DIRECCIÓN: Calle 64BN # 10 - 71. Popayán - Cauca

CORREO ELECTRÓNICO: vhmoncayo@javerianacali.edu.co

CELULAR: 3016804852

PROFESIÓN: Ingeniero en Electrónica

EMPRESA DONDE LABORA: Canal 13

CARGO: DBA



**Estimación de los precios de vehículos en las ciudades de Bogotá, Cali,
Medellín y Barranquilla utilizando Web Scraping, Modelos de Regresión y
Machine Learning**

VÍCTOR HUGO MONCAYO RUIZ
Código 8945498

PONTIFICIA UNIVERSIDAD JAVERIANA CALI
FACULTAD DE INGENIERÍA
DEPARTAMENTO DE INGENIERÍA ELÉCTRICA Y CIENCIAS DE LA
COMPUTACIÓN
SEPTIEMBRE 2021



**Estimación de los precios de vehículos en las ciudades de Bogotá, Cali,
Medellín y Barranquilla utilizando Web Scraping, Modelos de Regresión y
Machine Learning**

VÍCTOR HUGO MONCAYO RUIZ
Código 8945498

**Trabajo de grado presentado como requisito para optar al título de Magister
en Ingeniería**

Director
DAVID ARANGO LONDOÑO

PONTIFICIA UNIVERSIDAD JAVERIANA CALI
FACULTAD DE INGENIERÍA
DEPARTAMENTO DE INGENIERÍA ELÉCTRICA Y CIENCIAS DE LA
COMPUTACIÓN
SEPTIEMBRE 2021

Índice

1	Introducción	1
2	DESCRIPCIÓN DEL PROBLEMA	2
2.1	Planteamiento del problema	2
2.2	Justificación	2
2.3	Objetivos	4
2.3.1	Objetivo general	4
2.3.2	Objetivos específicos	4
2.4	Alcance	5
3	DESARROLLO DEL PROYECTO	6
3.1	Antecedentes	6
3.2	Marco Teórico	7
3.2.1	Modelos de Regresión	7
3.2.2	Regresión Logística	7
3.2.3	Modelos lineales generalizados	8
3.2.4	Validación Cruzada	9
3.2.5	Redes Neuronales	10
3.2.6	Arquitectura de las Redes Neuronales	10
3.2.7	Máquinas de Vectores de Soporte	12
3.2.8	Web Scraping	14
3.3	Resultados Obtenidos	15
3.3.1	Web Scraping	15
3.3.2	Registros y Atributos	21
3.3.3	Tratamiento de Datos	23
3.3.4	Resultados finales del tratamiento de datos	28
3.3.5	Selección de Atributos	29
3.3.6	Carga de Datos	37
3.3.7	Preprocesado de Datos	37
3.3.8	Partición de Datos	38
3.3.9	Modelos Propuestos	38
3.3.10	Comparación de Modelos	77
3.3.11	Análisis de Escenarios	86
4	Conclusiones	90

Lista de Figuras

1	Función Sigmoide	8
2	Elementos básicos de una Red Neuronal	10
3	Grafo ciudad de Cali: Web Scraping	20
4	Importancia Variables: Boruta	32
5	Verificación registros Nulos - Todas las ciudades	37
6	Especificaciones - Regresión Lineal	39
7	Box plot Residuos - Regresión Lineal	40
8	Precio Real Vs Precio Estimado - Regresión lineal Bogotá	42
9	Precio Real Vs Precio Estimado - Regresión lineal Cali	43
10	Precio Real Vs Precio Estimado - Regresión lineal Medellín	44
11	Precio Real Vs Precio Estimado - Regresión lineal Barranquilla	45
12	Especificaciones - Regresión Lineal Generalizada	50
13	Precio Real Vs Precio Estimado - Regresión Lineal Generalizada Bogotá	52
14	Precio Real Vs Precio Estimado - Regresión Lineal Generalizada Cali	53
15	Precio Real Vs Precio Estimado - Regresión Lineal Generalizada Medellín	54
16	Precio Real Vs Precio Estimado - Regresión Lineal Generalizada Barranquilla	55
17	Especificaciones - Regresión Lineal Polinómica	56
18	Resultado un sólo predictor - Kilometraje	57
19	Precio Vs Kilometraje - Regresión Lineal Polinómica	58
20	Resultado un sólo predictor - Kilometraje grado 4	59
21	Precio Real Vs Precio Estimado - Regresión Polinomial Grado 4 Bogotá	61
22	Precio Real Vs Precio Estimado - Regresión Polinomial Grado 4 Cali	62
23	Precio Real Vs Precio Estimado - Regresión Polinomial Grado 4 Medellín	63
24	Precio Real Vs Precio Estimado - Regresión Polinomial Grado 4 Barranquilla	64
25	Especificaciones - Árbol de Regresión	65
26	Estructura - Árbol de Regresión	66
27	Precio Real Vs Precio Estimado - Árbol de Regresión Bogotá	67
28	Precio Real Vs Precio Estimado - Árbol de Regresión Cali	68
29	Precio Real Vs Precio Estimado - Árbol de Regresión Medellín	69
30	Precio Real Vs Precio Estimado - Árbol de Regresión Barranquilla	70
31	Especificaciones - Árbol de Regresión	71
32	Error Estimado ntree = 500 - Bosque Aleatorio	72
33	Error Estimado para ntree = 200 y 300 - Bosque Aleatorio	73
34	Resultado modelo - Bosque Aleatorio	73
35	Precio Real Vs Precio Estimado - Bosque Aleatorio Bogotá	74
36	Precio Real Vs Precio Estimado - Bosque Aleatorio Cali	75
37	Precio Real Vs Precio Estimado - Bosque Aleatorio Medellín	76
38	Precio Real Vs Precio Estimado - Bosque Aleatorio Barranquilla	77
39	Comparación Modelos - Presición Bogotá	78
40	Comparación Modelos - Presición Cali	79
41	Comparación Modelos - Presición Medellín	80

42	Comparación Modelos - Presición Barranquilla	81
43	Comparación Modelos - Error Estimado Bogotá	82
44	Comparación Modelos - Error Estimado Cali	83
45	Comparación Modelos - Error Estimado Medellín	84
46	Comparación Modelos - Error Estimado Barranquilla	85

Lista de Tablas

1	Cantidad de registros iniciales por Ciudad	22
2	Cantidad de registros nulos	22
3	Registros resultantes al eliminar nulos	23
4	Registros sin Marca o Línea	23
5	Registros imputados para Combustible, Transmisión y Puertas	24
6	Cantidad de registros en el segundo raspado, por Ciudad	25
7	Cantidad total de registros por Ciudad	26
8	Cantidad de registros nulos finales	26
9	Registros imputados finales para Combustible, Transmisión y Puertas . . .	27
10	Cantidad de registros nulos finales	28
11	Importancia de atributos: Random Forest	33
12	Tabla de resultados: Selección de Atributos	36
13	Cantidad de registros por Ciudad	37
14	Valores de interés - Precio	38
15	Resultados Relevantes - Regresión Lineal	39
16	Comportamiento - Regresión Lineal	41
17	Valores de interés - Predictor Lm	41
18	Tabla de parámetros Regresión Lineal - Bogotá	48
19	Tabla de parámetros Regresión Lineal - Cali	48
20	Tabla de parámetros Regresión Lineal - Barranquilla	49
21	Tabla de parámetros Regresión Lineal - Medellín	49
22	Resultados Relevantes - Regresión Lineal Generalizada	50
23	Comportamiento - Regresión Lineal Generalizada	51
24	Valores mínimos predictores - Regresión Lineal Generalizada	55
25	Resultados Relevantes - Regresión Polinomial de Grado 4	60
26	Comportamiento - Regresión Polinomial de Grado 4	60
27	Comportamiento - Regresión Polinomial de Grado 4	66
28	Comportamiento - Regresión Polinomial de Grado 4	74

ABSTRACT

Knowing the fair price of a vehicle has become a necessity when buying or selling. However, the methods so far used and offered rely heavily on standard prices set uniquely by a few entities and experienced sellers. For this reason, it is necessary to have a tool to estimate in an approximate way the real price of the vehicles according to a particular number of characteristics, which have their contribution in the final price. In this project, we intend to take advantage of the information found in digital media and develop a system that facilitates the user to estimate the price when buying or selling a vehicle.

RESUMEN

Conocer el precio justo de un vehículo, se a convertido en una necesidad a la hora de comprar o vender. Sin embargo, los métodos hasta ahora usados y ofrecidos dependen en gran medida de precios estándar fijados de manera única por unas pocas entidades y de vendedores con experiencia. Por esta razón se hace necesario contar con una herramienta que estime de manera aproximada el precio real de los vehículos de acuerdo a un número particular de características, las cuales tienen su aporte en el precio final. En este proyecto, se pretende aprovechar la información encontrada en medios digitales y desarrollar un sistema que facilite al usuario estimar el precio a la hora de adquirir o vender un vehículo.

Palabras Clave: Estimación, Web Scraping, Tratamiento de datos, Selección de atributos, Modelos de Regresión, Machine Learning, Regresión Lineal, Regresión Lineal Generalizada, Regresión Lineal Polinómica, Árbol de Regresión, Bosque Aleatorio.

1 Introducción

El mercado de venta de vehículos cerró en el 2019 con un total de 263684 unidades para vehículos nuevos y 950 mil traspasos, según la asociación nacional de movilidad sostenible (ANDEMOS) [1]. Según pronósticos de esta misma fuente, el mercado automotor tendrá un excelente impulso en el 2020 debido a la dinámica de la economía colombiana, ofreciendo facilidades para adquirir automotores nuevos o usados. No dejando de lado, claro está, la incertidumbre que en estos momentos presenta el país debido a problemas de salud pública y el desplome del dólar, que en últimas podría no verse tan favorecido este sector.

Cómo se observa en las cifras entregadas por ANDEMOS, el volumen de ventas de vehículos usados es mucho mayor que el volumen de vehículos nuevos. Esto se debe en gran medida a la facilidad de acceso de la mayoría de clases sociales a adquirir vehículos por la diversidad de precios, mantenimiento, repuestos y modelos existentes en este mercado de los usados, dejando en la mayoría de los casos el mercado de los vehículos nuevos a un menor grupo de clase social. En este mercado, uno de los factores más destacados, es sin duda, las diferencias o cambios de precios entre vehículos con características similares. Estas diferencias vienen dadas por diversos factores como la marca, el modelo, las condiciones mecánicas del vehículo, entre otras, provocando que sea difícil estimar de manera precisa o lo más aproximado posible los precios en este mercado por cuenta del comprador.

Es conocido el hecho de que las ventas de los bienes, en especial los vehículos, están influenciadas por variados factores, tanto las características del vehículo como externos a ellos. Dentro de estas características tenemos: marca, modelo, kilometraje, tipo de combustible, número de puertas, entre otros. Entre los elementos externos tenemos: precio del dólar, precios de medios especializados en automotores, único dueño, etc. Cuanto más precisos sean los factores cuantificados en datos de entrada, el pronóstico será más preciso. Estos datos de entrada se están convirtiendo últimamente en activos valiosos para las personas interesadas en comprar o vender, donde gracias al análisis de estos se obtiene información de comportamientos o patrones que puede ayudar en la toma de decisiones.

Para el desarrollo de este trabajo de investigación sobre la estimación de precios de vehículos, se realizará un proceso exploratorio en diversos sitios web especializados en venta de vehículos tanto nuevos como usados analizando las características que se presentan al usuario, reuniendo así la mayor cantidad de información de atributos usados en las ventas. Seguido, se iniciará un proceso de selección de los atributos que más influyen en el precio final mediante herramientas y desarrollo para tal fin. Para finalmente diseñar e implementar un modelo que nos estime el comportamiento de los precios en vehículos, el cual servirá de base para los miles de usuarios que buscan comprar y vender sus automotores con precios justos a partir de sus características.

2 DESCRIPCIÓN DEL PROBLEMA

2.1 Planteamiento del problema

Adquirir un vehículo, tanto nuevo como usado, es una tediosa decisión debido a la cantidad de ofertas presentes en el mercado, y con la cual se lidia casi a diario por diferentes personas. Podemos evidenciar diversos métodos, que van desde lo tradicional, como es recomendación de amigos y conocidos, mirar clasificados en periódicos locales y nacionales e ir a compraventas cercanas, a lo más complejo como son las herramientas y medios tecnológicos. Los medios tecnológicos más conocidos y usados son los sitios web que ya llevan bastante tiempo ofreciendo los servicios de venta de diferentes bienes, entre estos automotores, como lo son Olx, Mercado libre, Carroya, entre otros. En estos sitios web se puede filtrar la información a nivel regional, por marca, por precio, etc, ofreciendo gran diversidad de vehículos dificultando un poco la toma de decisión.

Combinando la parte tradicional y tecnológica, el referente más antiguo y usado por los colombianos a la hora de comprar o vender vehículos nuevos o usados es Motor, información que se puede consultar tanto físicamente mediante revista o por medio digital en un archivo con extensión pdf que se puede descargar desde su sitio web [2]. La información de los precios de los vehículos depende solamente de su marca, su modelo y de su año de salir al mercado, sin importar el estado actual del automotor ni demás características. Esto deja a los vendedores tener la ventaja, injusta a veces, del verdadero precio de un automóvil sobre el comprador, puesto que conoce muy bien el estado, el desgaste y las veces que se le ha hecho mantenimiento. Y a pesar de que en los sitios web se cuenta con más información del vehículo, es difícil por parte del usuario cuantificar el valor de dichas características y saber si el precio ofrecido por el vendedor es justo.

Es por eso que se hace necesario contar con una herramienta que estime de manera aproximada el precio real de los vehículos de acuerdo a un número particular de características, las cuales tienen su aporte en el precio final. Esta herramienta daría un equilibrio a la ventaja que tienen los vendedores, al proporcionar índices de calidad del automotor a adquirir, y así sea más fácil y eficiente la toma de decisiones en el momento de la compra.

2.2 Justificación

A la hora de comprar o vender un vehículo, el cliente necesita estar seguro de que está pagando u ofreciendo el precio justo en cualquier región del país. Y aprovechando que se cuenta con abundantes ofertas de vehículos tanto en medios físicos y digitales, es válido

aprovechar las diferentes herramientas tecnológicas y estadísticas para ayudar a la toma acertada de decisiones. Así, haciendo un correcto análisis de la información consignada en todos los medios, se le facilitará al cliente estimar el precio a la hora de adquirir o vender un vehículo.

Este trabajo de investigación pretende diseñar e implementar un modelo de estimación de precios para la compra y venta de vehículos en distintas regiones del país, esto se logrará utilizando herramientas de regresión y Machine Learning basados en los precios y características que se encuentran en sitios web los cuales serán obtenidos mediante proceso de web scraping, teniendo este trabajo como finalidad principal brindar una información más precisa a los usuarios ante la toma de decisiones de un precio más acorde a la oferta y demanda del mercado.

2.3 Objetivos

2.3.1 Objetivo general

Desarrollar un modelo predictivo sobre precios de vehículos teniendo en cuenta sus características y la ciudad donde se encuentra.

2.3.2 Objetivos específicos

- Determinar y extraer los datos más relevantes de los vehículos mediante la aplicación de web scraping en páginas dedicadas a la venta de vehículos.
- Identificar mediante analítica exploratoria de datos cuales son las variables que mas se relacionan con los precios.
- Entrenar y comparar diferentes modelos paramétricos y de Machine Learning con base en la predicción de los precios de los vehículos.
- Analizar los resultados del modelo aplicado en diversas ciudades capitales como: Bogotá, Cali, Medellín y Barranquilla.

2.4 Alcance

El alcance de este trabajo de grado es el desarrollo e implementación de un modelo que permita predecir el valor de venta de vehículos en distintas ciudades capitales del país a partir de datos extraídos de conocidas páginas dedicadas a la comercialización de estos mismos, mediante web scraping.

3 DESARROLLO DEL PROYECTO

3.1 Antecedentes

En la exploración inicial del estado del arte relacionado con el pronóstico de ventas de vehículos, se encuentran algunos artículos y desarrollos, los cuales se basan principalmente en modelos de regresión y técnicas de aprendizaje automático. En la mayoría de desarrollos, se hace uso de datos ya encontrados en bases de datos oficiales; en cambio, en algunos artículos estos datos son extraídos de manera manual de medios físicos y en pocos casos se extraen de manera automática, mediante herramientas existentes o desarrollos propios, de sitios web especializados en la venta de automotores.

Un trabajo bastante relacionado con el objetivo de nuestro trabajo lo encontramos en [3], donde se construye un modelo para predecir el precio de autos usados en Bosnia y Herzegovina, a partir de datos extraídos mediante web scraper del portal web autopijaca.b. Este modelo se realizó aplicando tres técnicas de aprendizaje automático (Redes neuronales artificiales, Máquina de vectores de soporte y Bosque aleatorio), obteniendo una precisión del 87.38 %.

Otro trabajo relacionado con recopilación de datos desde un sitio web es [4]. Los datos son obtenidos del sitio de Mercado Libre (<http://www.mercadolibre.com.ar>) en el período comprendido desde marzo hasta julio de 2015, con un total de 920 registros. Este modelo hedónico permite detectar las relaciones entre las diversas características de un automóvil y su precio. También se encontraron interesantes relaciones entre el precio y la región donde éste se encuentre.

En [5] se investiga la aplicación de técnicas supervisadas de aprendizaje automático para predecir el precio de vehículos usados en 'Mauritius'. Se usaron cuatro diferentes técnicas para realizar las predicciones, como lo son: Análisis de regresión lineal múltiple, k-vecinos más cercanos, Naives bayes y Árboles de decisión. Estas técnicas seleccionadas proporcionaron un rendimiento comparable. Algo interesante de este trabajo fue la recopilación de los datos de los vehículos, los cuales fueron a partir de periódicos diarios.

Un modelo estadístico para pronosticar los precios de reventa en vehículos usados se presenta en [6]. Se realiza un estudio comparativo de métodos de predicción, donde se evidencia que la regresión forestal aleatoria es más efectiva en la predicción de precios de reventa que la regresión lineal, el método predominante en trabajos anteriores. El estudio también confirmó que los vendedores de automóviles usados poseen ventajas informativas sobre las agencias de investigación de mercado, lo que les permite pronosticar los precios de reventa con mayor precisión.

3.2 Marco Teórico

Esta propuesta pretende solucionar un problema o necesidad perteneciente a un segmento de mercado o población en particular, en este caso la compra y venta de vehículos. Y para tal fin, es necesario conocer conceptos que son relevantes para el sistema.

3.2.1 Modelos de Regresión

Podemos encontrar en el común, muchos problemas donde existe una relación inherente entre una variable independiente y una o más variables independientes. La técnica estadística conocida como análisis de Regresión es usada para definir relaciones que posibiliten predecir una o más variables en términos de otras. Siendo este el objetivo fundamental de diferentes investigaciones.

Estos modelos de regresión son empleados para estimar la mejor relación entre una variable dependiente y una o más variables independientes. Es común considerar que la relación entre estas variables sea descrita mediante una función del tipo

$$y(i) = \phi x_i \tag{1}$$

Sin embargo, el valor de y no será el que se espera en la mayoría de los casos significando que ϕ no será un valor preciso. Es por eso que se introduce un error ϵ , el cual corresponde a una variable independiente y experimental para cada pareja de y y x observadas

$$y(i) = \phi x_i + \epsilon_i \tag{2}$$

3.2.2 Regresión Logística

La Regresión Logística es un método de regresión útil para predecir clases binarias. El objetivo es de naturaleza dicotómica. Estas técnicas de clasificación son parte esencial de Machine Learning, puesto que cerca del 70 % de los problemas corresponden a clasificación.

Este método describe y estima una relación entre la variable binaria dependiente y las variables de entrada independientes. Utiliza la función logística llamada también función Sigmoide, la cual es una curva en forma de S que toma cualquier valor real entre 0 y 1.

Tal como se observa en la Figura 1, si la curva va a $+\infty$ la predicción se será 1, y si la curva va a $-\infty$, la predicción será 0. Si la salida de la función Sigmoide es mayor que 0,5, podemos clasificar el resultado como un SI, y si es menor que 0,5 podemos clasificarlo

como un NO.

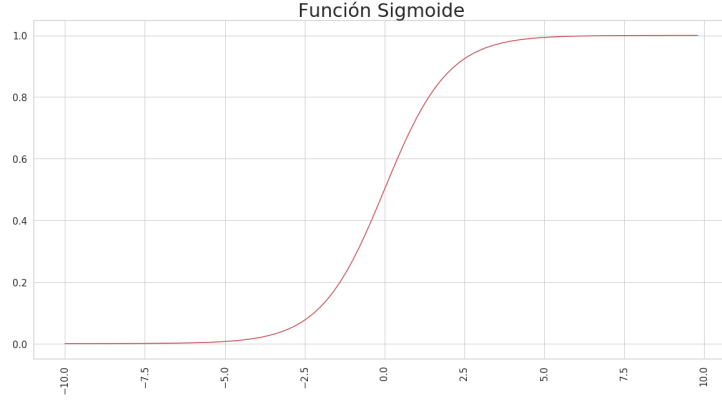


Figura 1: Función Sigmoide

Una Regresión Logística analiza datos distribuidos binomialmente de la forma:

$$Y(i) \sim B(p_i, n_i) \quad (3)$$

Donde n_i son conocidos y corresponden a los ensayos de *Bernoulli* y p_i son las probabilidades de éxito las cuales son desconocidas. Estas probabilidades son modeladas como una función de un vector de entradas obteniendo, para una variable dicotómica, una formulación dada por [8]:

$$p_i = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_{1,i} + \dots + \beta_k x_{k,i})}} \quad (4)$$

Entre los diferentes tipos de Regresión Logística, encontramos los siguientes:

- Binaria: La variable objetivo tiene sólo dos salidas posibles.
- Multinomial: La variable objetivo tiene entre 3 o más salidas nominales.
- Ordinal: La variable objetivo tiene entre 3 o más salidas ordinales.

3.2.3 Modelos lineales generalizados

Los Modelos lineales generalizados (GLM) son una extensión de los modelos lineales que permiten utilizar distribuciones donde existe falta de linealidad y homogeneidad de la varianza. Este modelo fue propuesto por *Nelder* y *Wedderburn* (1972), y puede considerarse ser usado cuando la variable objetivo cae dentro de estos casos [9]:

- Conteo de casos.
- Conteo de casos expresado como proporciones
- Respuesta binaria.

Los elementos de un Modelo Lineal Generalizado vienen dados por [10]: Las variables objetivo $Y_i, i = 1, \dots, n$, un conjunto de variables de entrada X y de parámetros β , y por último una función monótona llamada *link*, $g()$ usada para reparametrización, la cual proporciona el *predictor lineal* η ,

$$g(\mu) = \eta_i = X_i' \beta \quad (5)$$

Con $E(y_i) = \mu_i$.

Los GLM extienden la regresión lineal ordinaria, es decir, el ajuste de curvas de mínimos cuadrados, a situaciones donde las variables de respuesta son binomiales, Poisson, gamma, beta o, de hecho, cualquier forma de familia exponencial [11, pág 116].

3.2.4 Validación Cruzada

La validación cruzada o cross-validation es una técnica usada para evaluar los resultados de un análisis estadístico garantizando la independencia de la partición entre los datos de entrenamiento y de prueba. Esta técnica consiste repetir y calcular la media aritmética obtenida de las medidas de evaluación sobre diferentes particiones. Se utiliza en entornos donde el objetivo principal es la predicción y se quiere estimar cómo de preciso es un modelo que se llevará a cabo a la práctica [12]. Es una técnica muy utilizada en proyectos de inteligencia artificial para validar modelos generados.

El error en esta técnica se calcula como el promedio de los errores cometidos en cada iteración:

$$CV_{(n)} = \frac{1}{n} \sum_{i=1}^n MSE_i \quad (6)$$

Donde

$$MSE_i = (y_i - \hat{y}_i)^2 \quad (7)$$

Siendo $\hat{y}_i$ la predicción para y_i

3.2.5 Redes Neuronales

A mediados de los años 80's se introdujeron las Redes Neuronales (NN, en inglés Neural Networks). Esto con el fin de buscar la forma de imitar algunas funciones propias del ser humano, como la capacidad de memorizar y asociar. Los modelos basados en estas redes buscan de manera artificial y simplificada adquirir conocimiento a través de la experiencia.

Una Red Neuronal está constituida por neuronas interconectadas y organizadas en 3 capas generales, como se aprecia en la Figura 2. Los datos entran por la *capa de entrada*, pasan por la *capa oculta*, la cual puede tener muchas más capas, y finalmente los datos salen por la *capa de salida*.

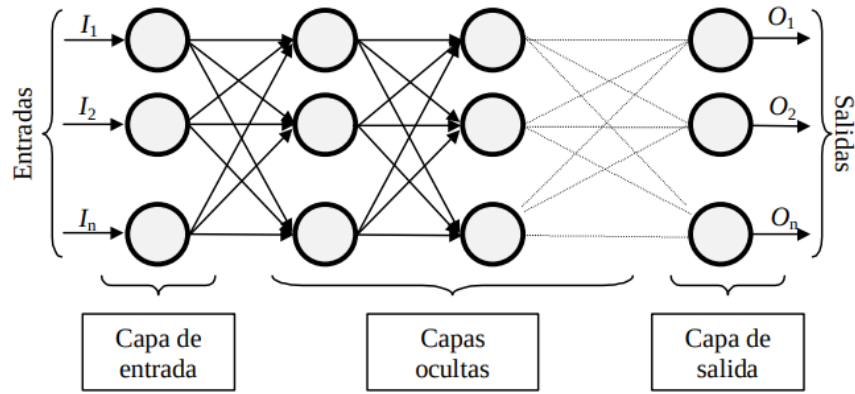


Figura 2: Elementos básicos de una Red Neuronal

Basándose en el modelo abstracto de una neurona artificial propuesto por *McCulloch* y *Pitts* (1943), la expresión de salida viene dada por:

$$y = \gamma \left(\sum_{i=1}^m w_i + x_i + w_0 \right) \quad (8)$$

Donde w_i es el vector de pesos equivalente a las conexiones sinápticas en una red neuronal real, w_0 corresponde al umbral de activación, x_i es el vector de entradas, y finalmente la función no-lineal de activación γ .

3.2.6 Arquitectura de las Redes Neuronales

Una Red Neuronal puede dar lugar a diversas estructuras o modelos que se podrían clasificar de diferentes maneras, y una de ellas es según el número de capas, las cuales se dividen en 2 grupos.

3.2.6.1 Redes Neuronales Monocapas

Es la Red Neuronal más sencilla. Consta de una sola capa de neuronas ubicadas en la salida que se encarga de ejecutar los diferentes cálculos. Existe una capa de entrada, pero esta no realiza ningún cálculo, sólo se encarga de recibir los datos de entrada.

La primera estructura neuronal en aparecer y la más sencilla es el *Perceptrón*. Su funcionamiento se basa en comparar la salida con la señal deseada del sistema mediante una función de activación.

Inicialmente, se inicializan los pesos W_k de manera aleatoria y se determina la salida y de acuerdo a esos pesos y el vector de entrada X_k [13, pág 32]

$$y = \sum_{k=1}^N W_k^n . X_k^n \quad (9)$$

Luego se compara la salida con el umbral predeterminado

$$v = y - umbral \quad (10)$$

Seguidamente se aplica la función signo

$$o = signo(v) \quad (11)$$

Una vez hecha la comparación con la señal deseada, se actualizan los coeficientes de los pesos

$$w_k = w_k + \alpha . (d - o) . x_k \quad (12)$$

3.2.6.2 Redes Neuronales Multicapa

Corresponde a una generalización de la red anterior, donde se encuentran un conjunto de capas intermedias entre la entrada y la salida, llamadas capas ocultas. Este tipo de red

puede estar completamente conectada o parcialmente conectada.

Ejemplo de este tipo de red es el *Perceptrón multicapa* (MLP), siendo también la más conocida y con mayor número de aplicaciones. Cómo se mencionó al inicio de esta sección, esta tipo de red tiene al menos 3 capas como se muestra en la Figura 2.

Dentro de las características más destacables de esta arquitectura están [13, pág 57]: Altamente no lineal, presenta tolerancia a fallos, es capaz de establecer una relación entre dos conjuntos de datos, y es posible realizar una implementación hardware.

La relación entre las variables de entrada y salida se logra propagando hacia adelante los valores de las variables de entrada. Esto se logra procesando la información proveniente de las entradas y propagando el resultado hacia las neuronas de la siguiente capa. Esta función de activación viene dada por:

$$y_i = a_i^C = f \left(\sum_{j=1}^{n_{c-1}} w_{ji}^{c-1} a_j^{c-1} + u_i^c \right) \quad (13)$$

para $i = 1, 2, \dots, n_c$ y $c = 2, 3, \dots, C - 1$, y donde $Y = (y_1, y_2, \dots, y_{nC})$ es el vector de salida.

Siendo un Perceptrón multicapa con C capas, $-C - 2$ capas ocultas, n_c neuronas en la capa c , w_{ji}^c representa el peso de la conexión de la neurona i de la capa c , a_i^c la activación de la neurona i de la capa c , y f la función de activación.

3.2.7 Máquinas de Vectores de Soporte

Las Máquinas de Vectores Soporte (SVM, en inglés Support Vector Machine) tienen su origen en los años 90. Y aunque originalmente fueron pensadas para problemas de clasificación binaria, actualmente se utilizan para resolver otros tipos de problemas, como son los de: Regresión, Agrupamiento y Multiclasificación.

El objetivo de SVM es encontrar un hiperplano óptimo como la solución al problema de aprendizaje. La formulación más simple es la lineal, donde el hiperplano se encuentra en el espacio de los datos de entrada x . En este caso, el espacio de hipótesis es un subconjunto de todos los hiperplanos de la forma:

$$f(x) = w.x + b \quad (14)$$

SVM encuentra un hiperplano de características en un espacio diferente al de los datos de entrada x , inducido por un núcleo K . A través de este núcleo, el espacio de hipótesis se define como un conjunto de *hiperplanos* en el espacio de características inducido por K . Esto también puede verse como un conjunto de funciones en un espacio de reproducción de Kerbert Hilbert (RKHS) definido por K (Wahba, 1990).

Este espacio se puede escribir formalmente como

$$\{f : \|f\|_K^2 < \infty\} \quad (15)$$

donde K es el núcleo que define el RKHS, y $\|f\|_K^2$ es la norma de la función RKHS.

Lo que hace que SVM sea un modelo de aprendizaje automático recomendable se puede resumir en [14]:

- **Es una técnica de dispersión.** Requiere que todos los datos de entrenamiento estén disponibles. Sin embargo, una vez que se identifican los parámetros del modelo, la SVM depende sólo de un subconjunto de estas instancias de entrenamiento, llamadas vectores de apoyo, para la predicción futura.
- **Es una técnica núcleo.** Utiliza el truco del núcleo para mapear los datos en un espacio de dimensiones superiores antes de resolver la tarea de aprendizaje de la máquina. SVM mapea los datos, usando funciones predefinidas del núcleo, en un espacio nuevo pero de mayor dimensión, donde un separador lineal sería capaz de discriminar entre las diferentes clases.
- **Es un separador de margen máximo.** Impone una restricción adicional al problema de la optimización: el hiperplano debe estar situado de tal manera que esté a una distancia máxima de las diferentes clases. Tal término obliga a que el paso de optimización encuentre el hiperplano que eventualmente se generalizaría mejor porque está situado a una distancia igual y máxima de las clases.
- **Utiliza la minimización del riesgo estructural (SRM) y satisface los requisitos de dualidad y convexidad.** SRM es un principio inductivo que selecciona un modelo de aprendizaje a partir de un conjunto de datos de entrenamiento finito. A medida que el índice de complejidad aumenta, el error alcanza su mínimo para el modelo óptimo antes de que empiece a aumentar de nuevo.

3.2.8 Web Scraping

El Web Scraping, conocido también como extracción o recolección de la web, es una técnica automatizada que encuentra y extrae datos con estructuras repetitivas en sitios web, para posteriormente guardarlos en un sistema de archivos o base de datos para su posterior recuperación o análisis. Esta técnica es reconocida como eficiente y poderosa cuando se necesita recolectar grandes cantidades de información siendo capaces de convertir sitios web completos en un conjunto de datos bien organizados.

Este proceso de extracción de datos, se puede dividir en dos grandes pasos secuenciales: En primer lugar se selecciona, de manera ordenada, la información que se desea extraer, siguiendo una serie de jerarquías que tiene el sitio web seleccionado. En segundo lugar, y una vez seleccionado lo que requiere, se procede a la extracción como tal mediante la emulación de un usuario navegando en el sitio.

El resultado final se puede almacenar en múltiples formatos, como lo son las páginas web, fuentes de datos en XML o JSON, o también como datos multimedia.

3.3 Resultados Obtenidos

3.3.1 Web Scraping

En esta primera etapa del proyecto, se exploraron 7 diferentes sitios web que se dedican, o parte de sus servicios es la venta de vehículos nuevos y/o usados. Se analizó las características de cada automóvil y la estructura de la información, es decir, que estuviera ordenada y en lo posible completa para la mayoría de los registros. Cabe aclarar que muchos usuarios que ponen en venta su vehículo, dejan muchos campos incompletos, o no registran de manera correcta la información.

Se listan a continuación los sitios web explorados y una breve descripción de lo encontrado.

- Mercado Libre [16]

En este sitio web se encontraron las siguientes características por cada vehículo:

- Marca
- Modelo
- Versión
- Año
- Kilometraje
- Puertas
- Tipo de combustible
- Placa
- Precio
- Descripción de la publicación
- Adicionales:
Único dueño, Color, Transmisión, Dirección, Control de tracción, Motor.

No todos los autos tienen la misma estructura en la información, es decir, no todos contienen las mismas características. Presentando corrimiento en los registros cuando se realiza el proceso de Scraping. Además, tampoco todos los autos contienen la misma información adicional.

- OLX [7]

En este sitio web se encontraron las siguientes características por cada vehículo:

- Marca
- Modelo
- Año
- Kilometraje
- Condición
- Combustible
- Color
- Transmisión
- Puertas
- Único dueño
- Placa
- Tipo de vendedor
- Precio
- Descripción

Al igual que en el caso anterior, no todos los autos tienen la misma estructura en la información, pero en una cantidad de registro mucho menores.

- Carroya [17]

En este sitio web las características de cada vehículo se dividen en 3 partes:

1. Información

- Estado del vehículo
- Pico y placa
- Kilometraje
- Ciudad placa
- Cilindraje
- Ciudad venta
- Tipo caja
- Color
- Tipo dirección
- Código
- Tipo combustible

- Precio

2. Ficha técnica

- Motor
Capacidad del tanque de combustible (Galones), cilindrada (c.c), cantidad de válvulas por cilindro, tracción, configuración árbol de levas, torque (Kg/Rev. x min), cantidad y configuración de cilindros, potencia (H.P), tipo de alimentación Combustible, tipo de alimentación.
- Transmisión
Cantidad de cambios adelante, tipo de transmisión.
- Suspensión
Capacidad del tanque de combustible (Galones), cilindrada (c.c), cantidad de válvulas por cilindro, tracción, configuración árbol de levas, torque (Kg/Rev. x min), cantidad y configuración de cilindros, potencia (H.P), tipo de alimentación combustible, tipo de alimentación, cantidad de cambios adelante, tipo de transmisión, suspensión delantera, suspensión trasera, tipo de frenos traseros, tipo de frenos delanteros, material rines, dimensiones rin, peso (kgrs), largo/ancho/alto (mts)
- Frenos
Tipo de frenos traseros, tipo de frenos delanteros.
- Ruedas
Material rines, rimensiones rin
- Dimensiones
Peso (kgrs), largo/ancho/alto (mts)

3. Comentarios del vendedor

En este apartado se encuentran algunos comentarios que hacen los venderes acerca de información adicional que no se encuentra especificada en los campos anteriores.

A pesar de tener información mucho más completa, se puede observar que está desorganizada en especial en la parte de Ficha técnica/suspensión, donde esta se repite para los demás campos. También se encontró que su estructura varía dependiendo si se tratan de autos usados o nuevos. Y por último, algunos

automóviles no tenían la misma cantidad de campos de información, lo cual no se logró encontrar la causa de esta situación.

- Los Coches [18]

En este sitio web no se realizó una exploración muy detenida, puesto que no hay filtro para buscar por región. Además, se debe filtrar por autos usados y nuevos, haciendo una descarga por separado. Y una vez habiendo seleccionado el tipo de auto a comprar, toca elegir la marca, convirtiendo en algo tedioso la descarga de la mayor cantidad de registros posibles. Y por último, todos los autos no tienen la misma información, dependiendo esto si el auto es usado, nuevo o de una marca en particular.

- Tu Carro [19]

En este sitio web se encontraron las siguientes características por cada vehículo:

- Marca
- Modelo
- Versión
- Año
- Cilindrada
- Kilometraje
- Puertas
- Tipo de combustible
- Placa
- Precio
- Descripción
- Adicionales:
Color, Transmisión, Dirección, Control de tracción, Motor, tipo.

No todos los autos tienen la misma estructura en la información, presentando corrimiento en los registros al descargar la información. Además, no todos los vehículos tienen la misma información en *Adicionales*.

- Auto Expo [20]

En este sitio web tampoco se realizó una exploración muy detenida, puesto que sólo aplica para la ciudad de bogotá. Además, los autos no tienen la misma información, dependiendo de la marca del mismo.

- Auto cosmos [21]

En este sitio web se encontraron las siguientes características por cada vehículo:

- Precio
- Motor
Combustible, cilindrada, potencia, torque, alimentación, cilindros, válvulas, sistema start/stop.
- Performance
Aceleración 0-100 km/h, consumo en ciudad, consumo en ruta, consumo mixto, velocidad máxima.
- Transmisión y chasis
Motor - tracción, transmisión, llantas, frenos (del. - tras), suspensión delantera, suspensión trasera
- Medidas y capacidades
Largo, ancho sin espejos, ancho con espejos, alto, distancia entre ejes, baúl, tanque de combustible, peso, capacidad de carga, altura de piso, capacidad de vadeo, ángulo de ataque, ángulo de partida, ángulo ventral, remolque con frenos, remolque sin frenos, capacidad de pendiente, escalonamiento vertical, inclinación lateral.
- Confort
Aire acondicionado, alarma de luces encendidas, asientos delanteros, asientos traseros, tapizados, cierre de puertas, vidrios (del. - tras.), espejos exteriores, espejo interior, faros delanteros, faros antiniebla, computadora de a bordo, control de velocidad crucero, dirección asistida, rines, encendido del motor, limpiaparabrisas posterior, techo solar, timón, cámara de visión.
- Seguridad
ABS, distribución electrónica de frenado, airbags, alarma e inmovilizador de motor, anclaje para asientos infantiles, cinturones de seguridad, tercer Stop, autobloqueo de puertas con velocidad.

- Comunicación y entretenimiento
Equipo de música, parlantes, conexión auxiliar (iPod y Mp3), conexión USB, interfaz bluetooth, sistema de video, pantalla, sistema de navegación, puertas con velocidad.

Las características varían dependiendo si se tratan de vehículos nuevos o usados. Depende particularmente de la marca del mismo, puesto que no todas las marcas poseen los mismo accesorios.

Selección

Se seleccionó el sitio web de **OLX** al contener información con menos registros incompletos y corridos al momento de descargar los datos al realizar el proceso de Web Scraping. Además, cuenta con la información justa y necesaria, pero no tan detallada para todo vehículo sin importar la marca ni el estado del mismo: Usado o Nuevo.

En la siguiente Figura 3 se muestra el Selector graph del Sitemap usado para ejecutar el Scraping para vehículos de la ciudad de Cali, el cual es similar para el resto de ciudades:

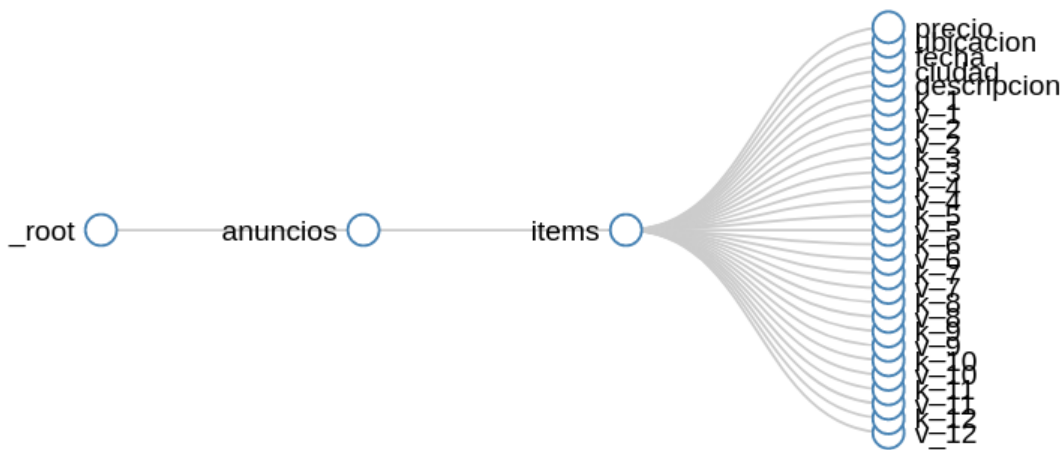


Figura 3: Grafo ciudad de Cali: Web Scraping

3.3.2 Registros y Atributos

En total se recopilaron de Olx 17 características (aquí en adelante serán llamados atributos) en cada vehículo, las cuales se redujeron a 13, y que se describen a continuación:

1. **Kilometraje:** Valor numérico. Cantidad de kilómetros recorridos.
2. **Precio:** Valor numérico. Precio del vehículo.
3. **Puertas:** Valor categórico. Cantidad de puertas.
4. **Placa:** Valor categórico. Número en que termina la placa.
5. **Marca:** Valor categórico. Marca del vehículo.
6. **Linea:** Valor categórico. Nombre del modelo según la marca.
7. **Modelo:** Valor categórico. Año en que salió el vehículo del concesionario.
8. **Condición:** Valor categórico. Si es nuevo o usado.
9. **Combustible:** Valor categórico. Tipo de combustible usado por el vehículo.
10. **Color:** Valor categórico. Color del vehículo.
11. **Transmisión:** Valor categórico. Tipo de transmisión del vehículo.
12. **Único dueño:** Valor categórico. Si ha tenido un sólo dueño o no.
13. **Tipo de vendedor:** Valor categórico. Si es vendido por persona natural o por un vendedor profesional.

Se eliminó el atributo *Ubicación*, el cual se encontraba en el anteproyecto, por tener un valor muy abierto y no aportar información relevante en el proyecto.

El atributo objetivo es *Precio*, y los demás atributos son los predictores.

Dando continuidad a lo mostrado en el anteproyecto, donde se realizó la importación en el sitio web de compra y venta de vehículos tanto nuevos como usados Olx [7], se retoma la siguiente tabla con la cantidad de registros obtenidos inicialmente:

Ciudad	Registros
Bogotá	717
Cali	860
Medellín	719
Barranquilla	495
Total Registros	2791

Tabla 1: Cantidad de registros iniciales por Ciudad

A partir de estos registros, se analiza la cantidad de datos nulos de cada atributo en cada conjunto de datos para cada ciudad, obteniendo la siguiente tabla:

Atributo	Bogotá	Cali	Medellín	Barranquilla
Marca	0	1	0	0
Línea	6	10	20	6
Modelo	0	0	0	0
Condición	20	22	20	8
Combustible	9	18	18	10
Color	4	8	7	9
Transmisión	19	32	38	19
Único dueño	34	47	40	22
Tipo vendedor	16	24	30	11
Kilometraje	0	0	0	0
Puertas	29	28	35	27
Placa	83	196	149	149

Tabla 2: Cantidad de registros nulos

En la tabla anterior se puede observar la gran cantidad de registros nulos, evidenciándose la falta de compromiso por parte del usuario al momento de registrar de manera completa, en el sitio web, la información del vehículo a vender.

Un resultado interesante de observar es la cantidad de registros resultantes al eliminar filas de acuerdo a la cantidad de nulos presentes por cada registro, obteniendo la siguiente tabla:

Ciudad	Registros iniciales	Nulos: 1	Nulos: 2	Nulos: 3
Bogotá	717	553	707	703
Cali	860	581	838	852
Medellín	719	483	694	708
Barranquilla	495	303	482	488

Tabla 3: Registros resultantes al eliminar nulos

Las columnas 'Nulos: X', hace referencia a los registros resultantes al momento de eliminar filas con X atributos nulos.

La gran cantidad de registro eliminados se debe en gran medida al atributo 'Placa', puesto que tiene gran cantidad de datos nulos.

3.3.3 Tratamiento de Datos

Se procesan los datos de cada atributo de diferente manera, de acuerdo a sus diferentes características siguiendo ciertos criterios que se explican en cada apartado.

- Marca y Línea

Estos atributos tienen marcada importancia para el comprador, puesto que muchas veces se basan en la marca y su línea para tomar la decisión de adquirir un vehículo. Además, está el hecho de que la marca influye mucho en el precio del mismo, y sin esta información es difícil aproximar el valor del mismo. Así que, como primera medida se eliminaron los registros que no tuvieran esta información, obteniendo la siguiente tabla con los registros eliminados que no tuvieran marca o línea, por ciudad:

Ciudad	Registros eliminados
Bogotá	6
Cali	10
Medellín	20
Barranquilla	6

Tabla 4: Registros sin Marca o Línea

- Combustible, Transmisión y puertas

Se implementó un método propio para imputar la mayor cantidad posible de datos faltantes en estos atributos. Se eligieron al ser dependientes de la Marca y la Línea, y que además se cuentan con registros dentro del set de datos con la información completa para estos vehículos.

Procedimiento de imputación para cada atributo:

1. Se creó un set de datos con los registros nulos del atributo a imputar.
2. Se recorrió el set de datos anterior a partir de su indexación con las siguientes acciones:
 - (a) Se creó un set de datos auxiliar a partir del inicial con la marca y línea del registro indexado.
 - (b) Con el set de datos auxiliar se generó un listado con los elementos del atributo a imputar.
 - (c) Se calculó la moda del listado anterior.
 - (d) Se imputó el dato faltante con el resultado obtenido en el paso anterior.

Culminado este proceso, se muestra a continuación el resultado de implementar el proceso anterior. En la siguiente tabla se observa la cantidad de datos nulos antes y después de implementar el proceso descrito, y su correspondiente porcentaje de imputación, por cada atributo.

Ciudad	Combustible			Transmisión			Puertas		
	Antes	Después	%	Antes	Después	%	Antes	Después	%
Bogotá	9	1	88.8	19	2	89.47	29	7	75.86
Cali	18	1	94.4	32	5	84.37	28	3	89.28
Medellín	18	1	94.4	38	4	89.47	35	3	91.42
Barranquilla	10	0	100	19	5	73.68	27	6	77.77

Tabla 5: Registros imputados para Combustible, Transmisión y Puertas

- Kilometraje

Para este atributo, los datos obtenidos en el sitio web no parecen consistentes. Es decir, existen vehículos con un valor de kilometraje en cero con valores de modelo de años anteriores, también valores de kilometraje menores de 100km con modelos que superan los 10 años atrás, entre otros.

En este orden de ideas, se procede a filtrar datos que no parecen tener sentido de acuerdo al siguiente criterio:

- Kilometraje menor de 100km con modelos menores a 2019.

En este primer análisis, los datos no consistentes con el atributo de kilometraje fueron al rededor del 40% en los set de datos de las ciudades, reduciéndose de manera drástica el número de registros para cada ciudad, teniendo muy poca información para la siguiente etapa de *Selección de atributos*.

Por tal razón se decidió realizar una vez más el proceso de Web Scraping en el sitio web para concatenarlos a los datos iniciales y tener un mayor número de registros para su posterior análisis.

En este segundo raspado se obtuvieron los siguientes registros:

Ciudad	Registros
Bogotá	970
Cali	518
Medellín	736
Barranquilla	422
Total Registros	2646

Tabla 6: Cantidad de registros en el segundo raspado, por Ciudad

Con estos nuevos datos, se procedió a adecuarlos para que estuvieran organizados de la misma forma que los anteriores. Una vez concatenados y eliminando los registros repetidos, se obtuvieron las siguientes cantidades de registros totales por ciudad:

Ciudad	Registros
Bogotá	1676
Cali	1356
Medellín	1426
Barranquilla	895
Total Registros	5353

Tabla 7: Cantidad total de registros por Ciudad

Es importante mencionar que el concatenado se realizó con el registro inicial antes de haber eliminado los datos nulos de los atributos *Marca* y *Línea*.

Se muestran una vez más la cantidad de registros nulos por ciudad del nuevo conjunto de datos:

Atributo	Bogotá	Cali	Medellín	Barranquilla
Marca	0	1	0	0
Línea	6	10	20	6
Modelo	0	0	0	0
Condición	42	36	34	17
Combustible	23	29	35	15
Color	10	14	11	12
Transmisión	46	55	65	28
Único dueño	58	67	90	40
Tipo vendedor	44	37	54	25
Kilometraje	0	0	0	0
Puertas	53	43	57	39
Placa	221	319	321	277

Tabla 8: Cantidad de registros nulos finales

Realizando una vez más la imputación a los 3 atributos mencionados anteriormente (Combustible, Transmisión y Puertas), se obtuvieron los siguientes resultados:

Ciudad	Combustible			Transmisión			Puertas		
	Antes	Después	%	Antes	Después	%	Antes	Después	%
Bogotá	23	0	100	46	3	93.47	53	5	90.56
Cali	29	2	93.1	55	8	85.45	43	3	93.02
Medellín	35	1	97.14	65	4	93.84	57	4	92.98
Barranquilla	15	0	100	28	2	92.85	39	6	84.65

Tabla 9: Registros imputados finales para Combustible, Transmisión y Puertas

Es importante notar la gran mejora que se obtuvo agregando más registros. Esto se debe a la nueva información en los nuevos registros, que sirvió de insumo en el momento de aplicar el proceso de imputación, mejorando así el porcentaje del mismo.

Teniendo mayor número de registros por ciudad, procedemos a generar más criterios para el filtrado del atributo Kilometraje:

- Eliminar registros con kilometraje 0km, si el modelo no es 2020 y el vendedor no sea profesional.
- Eliminar registros con kilometraje menor o igual a 10m si el modelo es mayor o igual a 2018 y menor a 2020.
- Eliminar registros con kilometraje menor o igual a 40m si el modelo es menor a 2018.
- Eliminar registros con kilometraje menor o igual a 10000m si el modelo es menor a 2015.
- Multiplicar por 1000 si el kilometraje es menor a 100m en modelos 2020.
- Multiplicar por 1000 si el kilometraje es menor a 1000m en modelos menores a 2020.

Es importante aclarar que estos criterios no se usaron para imputar datos en este atributo, puesto como se mostró en la tabla de datos nulos, kilometraje no tenía información faltante. Estos criterios se usaron para editar algunos datos que no tenían lógica, y que al parecer fueron generados por la no interpretación correcta por parte del usuario a la hora de ingresar esta información.

- Condición

Para el atributo Condición, se implementó un único criterio tanto para imputar información faltante, como para editar datos que tampoco parecían tener lógica:

- Todo vehículo tendrá un valor de *usado* en este atributo, excepto los vehículos con 0km, modelo 2020 y vendedor profesional.

Este mismo criterio se aplicó para los datos nulos.

- Demás atributos

Para el resto de atributos: Color, Único_dueño, Tipo_vendedor y Placa, no se encontraron criterios que nos ayudaran a la imputación o edición de los mismos, esto debido a que no dependen de otros atributos y tampoco se cuenta con información que ayuden a analizarlos.

3.3.4 Resultados finales del tratamiento de datos

Una vez aplicados los filtros mencionados en el ítem anterior, se obtienen los resultados mostrados en la siguiente tabla:

Atributo	Bogotá	Cali	Medellín	Barranquilla
Marca	0	0	0	0
Linea	0	0	0	0
Modelo	0	0	0	0
Condición	0	0	0	0
Combustible	1	3	1	0
Color	10	11	7	8
Transmisión	6	6	4	2
Único dueño	54	60	74	36
Tipo vendedor	39	33	44	22
Kilometraje	0	0	0	0
Puertas	6	6	3	5
Placa	196	259	269	227

Tabla 10: Cantidad de registros nulos finales

Comparando estos resultados con la tabla 8, se observa una notable reducción de los registros nulos en atributos analizados anteriormente, y por ende una gran mejora.

3.3.5 Selección de Atributos

Cuando se trabaja con Aprendizaje Automático, la Selección de Atributos se vuelve un factor muy importante, el cual se especializa en la tarea de seleccionar el conjunto de atributos que aportan más relevancia en las tareas posteriores de clasificación o regresión.

Entre los múltiples beneficios que pueden aportar la Selección de Atributos, están: reducción de overfitting, mejora de la precisión de las predicciones, eliminación de atributos redundantes, simplificación de los modelos y reducción del tiempo de proceso.

El objetivo que persiguen estas técnicas es obtener la lista de atributos más relevantes de un conjunto de datos iniciales. Si dicho proceso se realiza con éxito, como resultado se obtendrán atributos que nos aportarán la misma información que el conjunto original, eliminando así los atributos irrelevantes o redundantes de los datos originales.

Existe una gran cantidad de trabajos publicados sobre técnicas de selección de atributos. Para nuestro proyecto, se decidió probar diferentes técnicas para seleccionar los atributos que tengan una mayor contribución en el pronóstico del precio del vehículo.

Antes de continuar con este análisis, es importante mencionar que se eliminó el atributo *Placa*, debido a que contenía gran cantidad de registros nulos, y en las técnicas probadas para la Selección de Atributos, en su gran mayoría, no trabajaban con datos faltantes, obligando a eliminar los registros correspondientes.

Cabe mencionar también que algunas de estas técnicas no se pudieron correr debido a que sólo trabajan con datos de tipo numérico, y nuestro set de datos sólo 1 de los 11 atributos es continuo (No se cuenta la variable objetivo *Precio*).

Las técnicas se implementaron y configuraron en la herramienta estadística R. Esto debido a la mayor flexibilidad que posee esta herramienta para trabajar con datos categóricos, comparado con Python donde se trabajó el tratamiento de los datos.

A continuación se muestran los resultados de las técnicas que arrojaron resultados.

◇ Earth

El paquete Earth implementa la importancia de las variables basándose en la validación cruzada generalizada (GCV), el número de modelos de subconjuntos en que se produce la variable (nsubconjuntos) y la suma residual de cuadrados (RSS).

Suprimiendo los registros nulos, y corriendo el script se obtuvieron los siguientes atributos resultantes, sin ningún orden en especial:

- ✓ modelo
- ✓ linea
- ✓ transmisión
- ✓ marca
- ✓ combustible
- ✓ kilometraje

◇ Step-wise Regression

Con esta técnica se analizan qué atributos tienen más peso. Esta técnica da como resultado varios niveles dentro de las variables categóricas.

Descartando los sub-niveles, se tienen los siguientes atributos resultantes, sin ningún orden en especial:

- ✓ linea
- ✓ marca
- ✓ transmisión
- ✓ combustible
- ✓ kilometraje
- ✓ unico_dueño

◇ Boruta

Este método puede ser usado para decidir si una variable es importante o no en la variable objetivo.

Esta técnica nos arroja los atributos que contribuyen a nuestro objetivo, sin ningún orden en especial:

- ✓ combustible
- ✓ modelo
- ✓ transmisión
- ✓ kilometraje
- ✓ marca
- ✓ linea
- ✓ puertas
- ✓ único_dueño

Esta técnica también nos arroja una gráfica de la importancia de las variables. En la Figura 4 se puede observar atributos de color verde, amarillo y rojo. Siendo los de color verde los confirmados, los de color amarillo que podrían llegar a contribuir y los rojos que definitivamente no contribuyen. Los bloques de color azul solo sirven de separador.

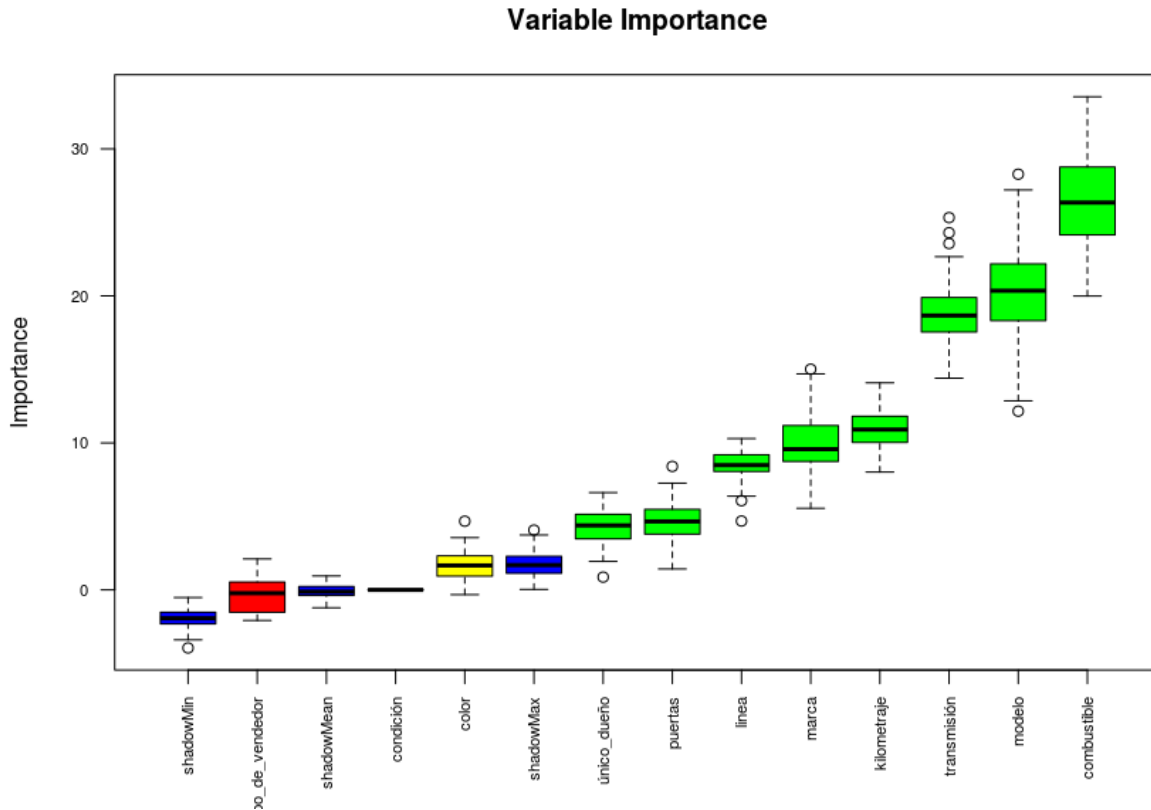


Figura 4: Importancia Variables: Boruta

Según la anterior figura se extraen los atributos con un valor de importancia sobre 8 que contribuyen a nuestro objetivo, sin ningún orden en especial:

- ✓ combustible
- ✓ modelo
- ✓ transmisión
- ✓ kilometraje
- ✓ marca
- ✓ línea

◇ Random Forest

Randomforest puede ser muy efectivo en el momento de desear encontrar un conjunto de predictores que expliquen mejor la varianza en la variable de respuesta.

Imprimiendo los valores de la importancia, tenemos:

Atributo	%IncMSE	IncNodePurity
marca	11.0065759	5.535792e+16
linea	11.3047260	7.536781e+16
modelo	16.6294758	1.172390e+17
condición	0.0000000	6.874086e+13
combustible	25.8405069	3.179358e+16
color	2.8314037	2.449459e+16
transmisión	18.0229026	5.796589e+16
único_dueño	5.8335817	1.959688e+16
tipo_de_vendedor	-0.5861978	1.553394e+15
kilometraje	11.2356688	1.113799e+17
puertas	5.1341405	1.294239e+16

Tabla 11: Importancia de atributos: Random Forest

Ordenando y extrayendo los atributos con un mayor valor de IncNode Purity se muestran los que contribuyen nuestro objetivo:

- ✓ modelo
- ✓ kilometraje
- ✓ linea
- ✓ transmisión
- ✓ marca
- ✓ combustible

◇ XGBoost

Xgboost realiza un entrenamiento aditivo y controla la complejidad del modelo mediante la regularización. El algoritmo XGboost ordena las características más importantes por *Ganancia*, *Cobertura* y *Frecuencia*. La ganancia nos indica la importancia de una característica para hacer más pura una rama de un árbol de decisión. La cobertura mide la cantidad relativa de observaciones afectadas por un atributo y la frecuencia cuenta el número de veces que se utiliza un atributo en todos los árboles generados [15].

Esta técnica nos arroja los atributos que contribuyen a nuestro objetivo, sin ningún orden en especial:

- ✓ modelo
- ✓ linea
- ✓ kilometraje
- ✓ transmisión
- ✓ marca
- ✓ combustible

◇ Caret 1.0

Los métodos de selección automática de atributos pueden utilizarse para construir muchos modelos con diferentes subconjuntos de un conjunto de datos e identificar aquellos atributos que son y no son necesarios para construir un modelo preciso.

Un método automático muy popular para la selección de atributos que ofrece el paquete caret R se denomina Eliminación Recursiva de Características o RFE.

En cada iteración se utiliza un algoritmo de Random Forest para evaluar el modelo. El algoritmo está configurado para explorar todos los posibles subconjuntos de los atributos.

Esta técnica nos da como resultado un top 5 de atributos que contribuyen a nuestro objetivo:

- ✓ combustible
- ✓ modelo
- ✓ transmisión
- ✓ kilometraje
- ✓ marca

◇ Caret 2.0

Para este segundo conjunto de técnicas, se hace uso de diversos algoritmos de Machine learning.

El paquete Caret R proporciona herramientas que informan automáticamente sobre la relevancia e importancia de los atributos en los datos e incluso seleccionan los atributos más importantes.

Esta técnica proporciona findCorrelation, que analiza una matriz de correlación de atributos de los datos sobre los atributos que se pueden eliminar. Para este análisis, se usa el método 'rpart'.

Reuniendo los datos con atributos en común, nos arroja como resultado los que contribuyen a nuestro objetivo, sin ningún orden en especial:

- ✓ transmisión
- ✓ modelo
- ✓ kilometraje
- ✓ único_dueño
- ✓ marca

◇ Simulated Annealing

Corresponde a un algoritmo de búsqueda global que permite aceptar una solución subóptima con la esperanza de que una solución mejor aparezca eventualmente.

Funciona haciendo pequeños cambios aleatorios a una solución inicial y ve si el rendimiento mejora. El cambio es aceptado si mejora, o puede ser aceptado si la diferencia de rendimiento cumple con un criterio de aceptación.

Esta técnica nos da de manera automática 7 atributos:

- ✓ linea
- ✓ modelo
- ✓ transmisión
- ✓ único_dueño
- ✓ tipo_de_vendedor

✓ kilometraje

◇ Tabla de resultados

En la siguiente tabla se muestran los resultados, de los atributos elegidos, de las diferentes técnicas.

Técnica	marca	linea	modelo	condición	combustible	color	transmisión	único_dueño	tipo_de_vendedor	kilometraje	puertas
Earth	X	X	X		X		X			X	
Step-wise Regression	X	X			X		X	X		X	
Boruta	X	X	X		X		X	X		X	X
Random Forest	X	X	X		X		X			X	
XGBoost	X	X	X		X		X			X	
Caret 1.0	X		X		X		X			X	
Caret 2.0	X		X				X	X		X	
Simulated Annealing		X	X				X	X	X	X	X

Tabla 12: Tabla de resultados: Selección de Atributos

Los resultados mostrados en la tabla anterior, no deja en evidencia que hay 6 atributos constantes en las diferentes técnicas:

- ♣ kilometraje
- ♣ transmisión
- ♣ modelo
- ♣ marca
- ♣ combustible
- ♣ linea

Otro atributo que se podría considerar, sería: unico_dueño. Puesto que está presente en 4 de 7 técnicas.

3.3.6 Carga de Datos

Una vez seleccionados los atributos más representativos para el cálculo aproximado del precio de los vehículos, se cargan los registros por ciudad con los cuales se generarán los diferentes modelos para su posterior comparación y selección del que mejor se aproxime al precio justo.

En la siguiente tabla muestran la cantidad de registros por ciudad.

Ciudad	Registros
Bogotá	1492
Cali	1148
Medellín	1235
Barranquilla	785

Tabla 13: Cantidad de registros por Ciudad

Cabe resaltar que el número de registros por ciudad, es el resultado del tratamiento de datos por atributo.

3.3.7 Preprocesado de Datos

Para esta fase, se decidió hacer un proceso mínimo sobre los datos, puesto que en la sección *Tratamiento de datos* se hizo un previo tratamiento de los mismos.

Se eliminaron todas las filas que tuvieran algún valor nulo, de una forma sencilla, procediendo a su correspondiente verificación, como lo muestra la Figura 5:

marca 0 línea 0 modelo 0 combustible 0 transmisión 0 único_dueño 0 kilometraje 0 precio 0

Figura 5: Verificación registros Nulos - Todas las ciudades

Como nuestra variable objetivo es el atributo *Precio*, se muestra a continuación algunos datos de interés por ciudad:

	Bogotá	Cali	Medellín	Bquilla
Valor mínimo	1234567	1000000	2000000	1000000
Valor máximo	350500000	219000000	430000000	240000000
Valor medio	24556676	26980472	28646339	29762660

Tabla 14: Valores de interés - Precio

3.3.8 Partición de Datos

El porcentaje a usar para particionar el conjunto de datos, fue de 80 – 20. 80 % para entrenar los modelos y un conjunto de validación de 20 % para probarlos.

El mayor inconveniente presentado fue lograr la existencia de cada muestra en el set de entrenamiento de las diferentes opciones en algunos atributos especiales, tales como: marca y línea. Este problema conlleva a no poder predecir algunos registros en el set de pruebas, al no haber estado presente en el set de entrenamiento.

Para resolver este percance, se procedió a diseñar e implementar un código en la herramienta R, el cual aseguraba crear un set de datos de entrenamiento con al menos una de las diferentes opciones de las tuplas (marca, linea), y luego completar este set de manera aleatoria con el resto de registros hasta completar la cuota del 80 % para entrenamiento.

3.3.9 Modelos Propuestos

En esta sección se procedió a generar y configurar 5 diferentes modelos. Se analizaron los resultados de cada uno para finalmente comparar la precisión y el error estimado de todos y así conocer el modelo con el enfoque ganador.

- **Regresión Lineal - Modelo 1**

En Figura 6 se encuentran las especificaciones correspondientes al modelo de Regresión Lineal.

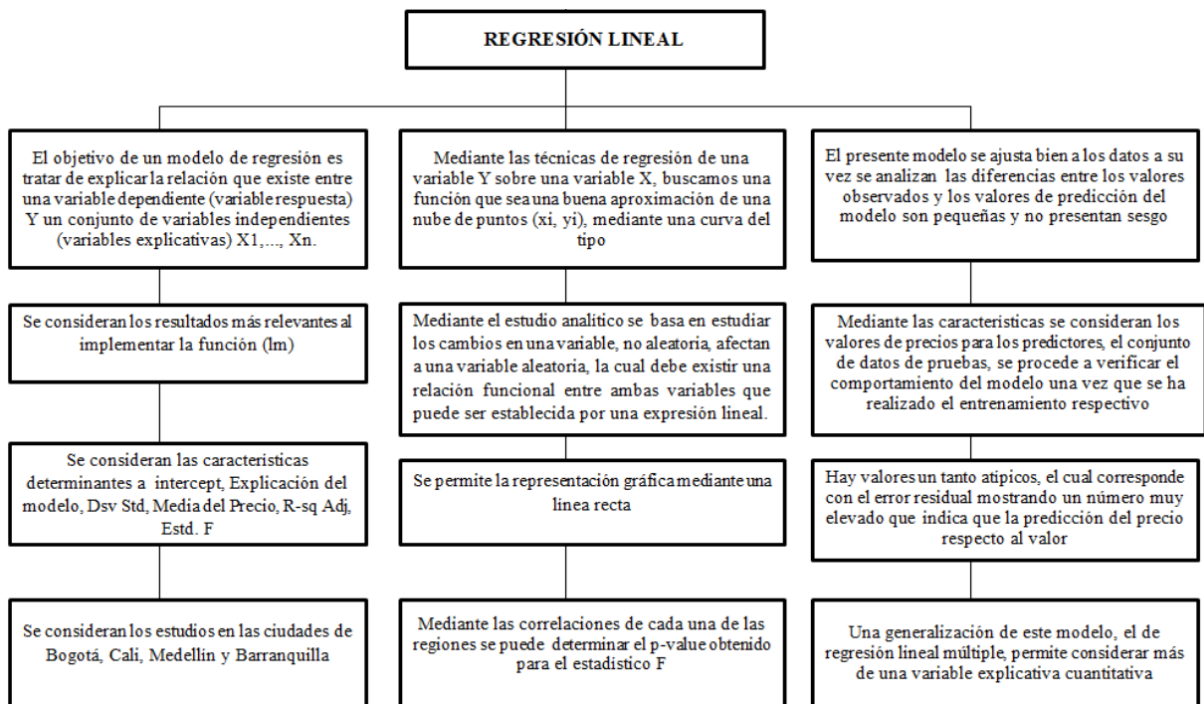


Figura 6: Especificaciones - Regresión Lineal

En la siguiente tabla se muestran los resultados más relevantes al implementar este modelo usando la función *lm*:

Característica	Bogotá	Cali	Medellín	Barranquilla
Intercept	-2.695364e+09	-3.020652e+09	-3.406217e+09	-4.462850e+09
Explicación del modelo	83.23 %	90.57 %	92.10 %	92.25 %
Dsv Std	10090726.74	7134218.40	7533347.06	8128002.94
Media del Precio	24617833.77	27172235.57	28385737.86	29671004.04
R-sq Adj	0.77	0.87	0.90	0.89
Estd. F	12.69	27.28	37.88	25.03

Tabla 15: Resultados Relevantes - Regresión Lineal

Los valores del intercepto, para este modelo, se ubican por debajo de $-2,69^{09}$. Una revisión de los estadísticos **F**, **RSE** y **R²**, nos permiten emitir las siguientes conclusiones:

- Residual standar error (RSE): En promedio RSE es de 41 % para Bogotá, 26,25 % para Cali, 26,5 % para Medellín y 27,39 % para Baranquilla. Esto nos señala que el error aproximado que el modelo comete en la predicción es

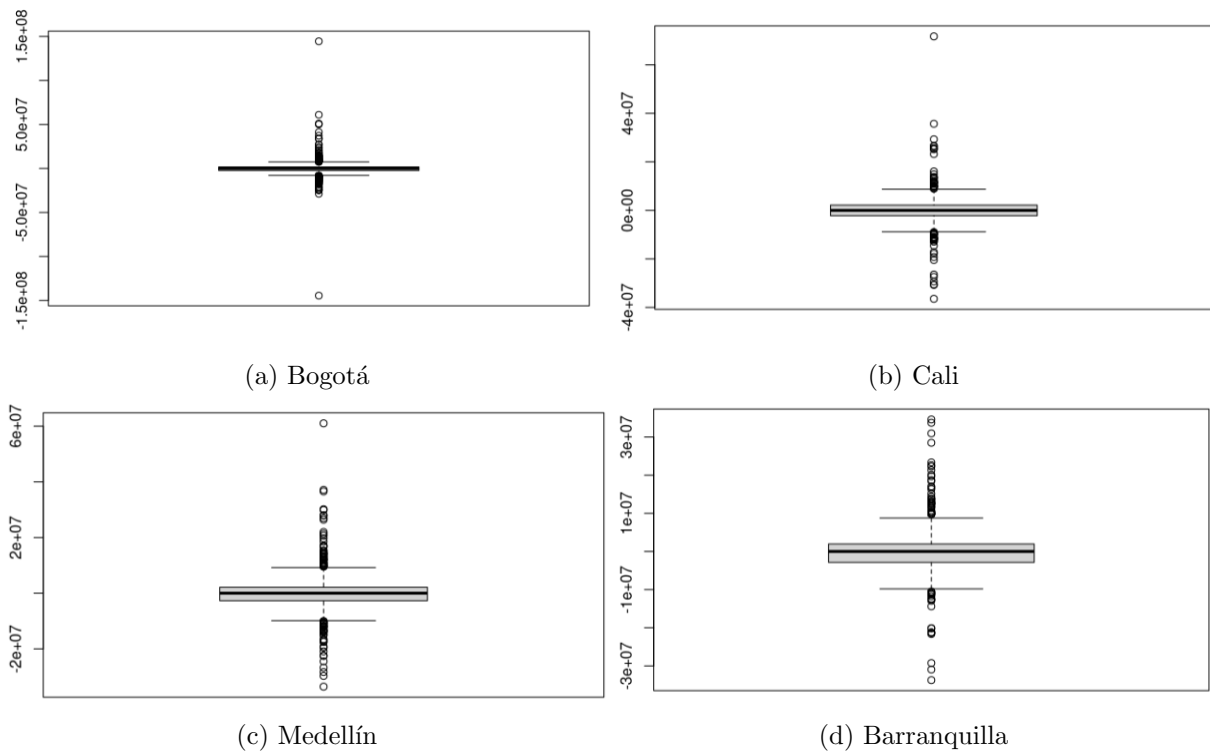


Figura 7: Box plot Residuos - Regresión Lineal

bastante elevado.

- R^2 : Los predictores empleados por el modelo son capaces de explicar, mínimo el 83,23 % para Bogotá y un máximo de 92,25 % para Barranquilla, de la variabilidad observada en el precio de los vehículos.
- El p-value obtenido para el estadístico F es muy pequeño ($< 2,2^{-16}$) para todas las ciudades. Lo que indica que al menos uno de los predictores introducidos en el modelo está relacionado con la variable respuesta precio.

Podemos revisar los residuos del modelo, para cada ciudad, en la Figura 7:

La mayoría de residuos están en torno al 0, así que parece correcto.

Valores de precios para los predictores

Empleando el conjunto de datos de pruebas, se procede a verificar el comportamiento

del modelo una vez que se ha realizado el entrenamiento respectivo. Dando como resultado lo mostrado en la tabla 16:

	Bogotá	Cali	Medellín	Barranquilla
RMSE	9481338	9006246	17868751	9633236
R²	83.98	90.57	92.1	92.25

Tabla 16: Comportamiento - Regresión Lineal

Hay valores un tanto atípicos, el cual corresponde con el error residual mostrando un número muy elevado que nos indica que la predicción del precio respecto al valor real puede tener una diferencia de mínimo ± 9 millones.

En la siguiente tabla, se muestran algunos resultado relevantes de la predicción:

	Bogotá	Cali	Medellín	Bquilla
Valor mínimo	-20526740.98	-7248583.25	-11651298.37	-27995814.16
Valor máximo	120788045.54	119285041.53	186098268.77	109726340.25
Valor medio	24140319.80	28294680.39	27652169.58	27259451.35

Tabla 17: Valores de interés - Predictor Lm

En las sigientes figuras se pueden observar los comportamientos de los precios estimados de la Regresión Lineal para cada ciudad.

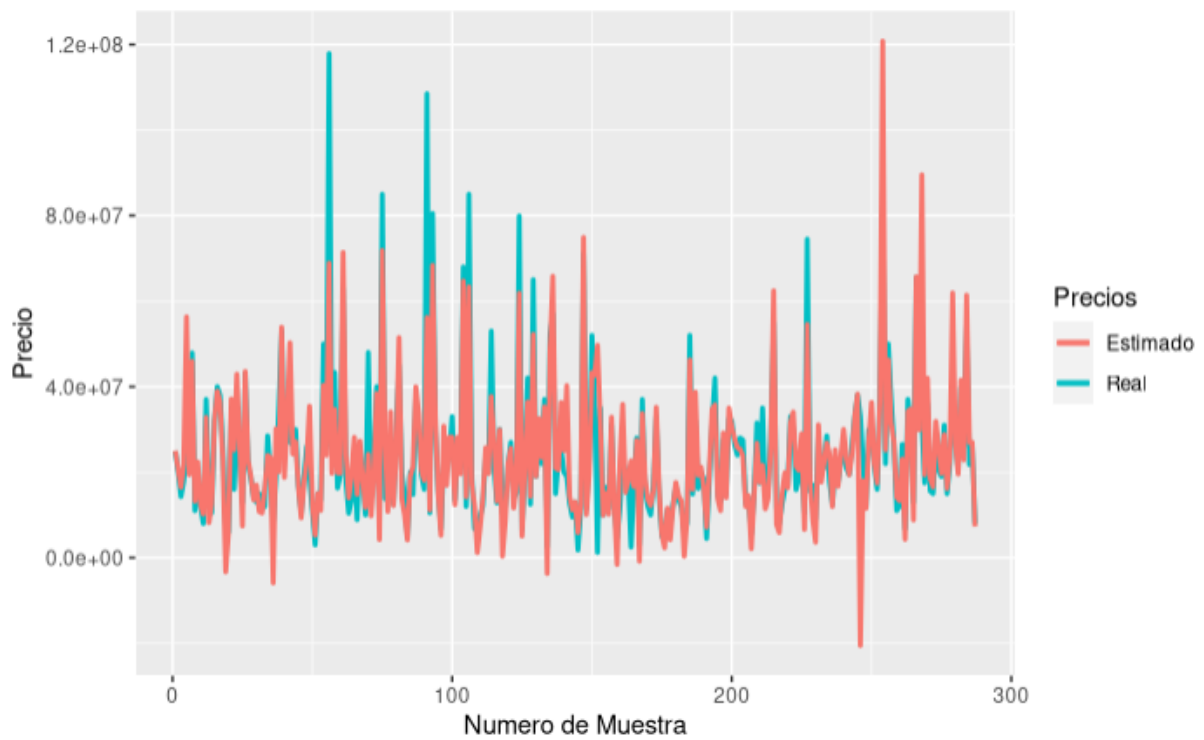


Figura 8: Precio Real Vs Precio Estimado - Regresión lineal Bogotá

En el caso de la Figura 8 se puede observar los precios reales vs los precios estimados en la ciudad de Bogotá. Según el comportamiento de la gráfica existe un nivel entre ambos precios, aunque en la mayoría de los picos más altos el precio real supera al estimado.

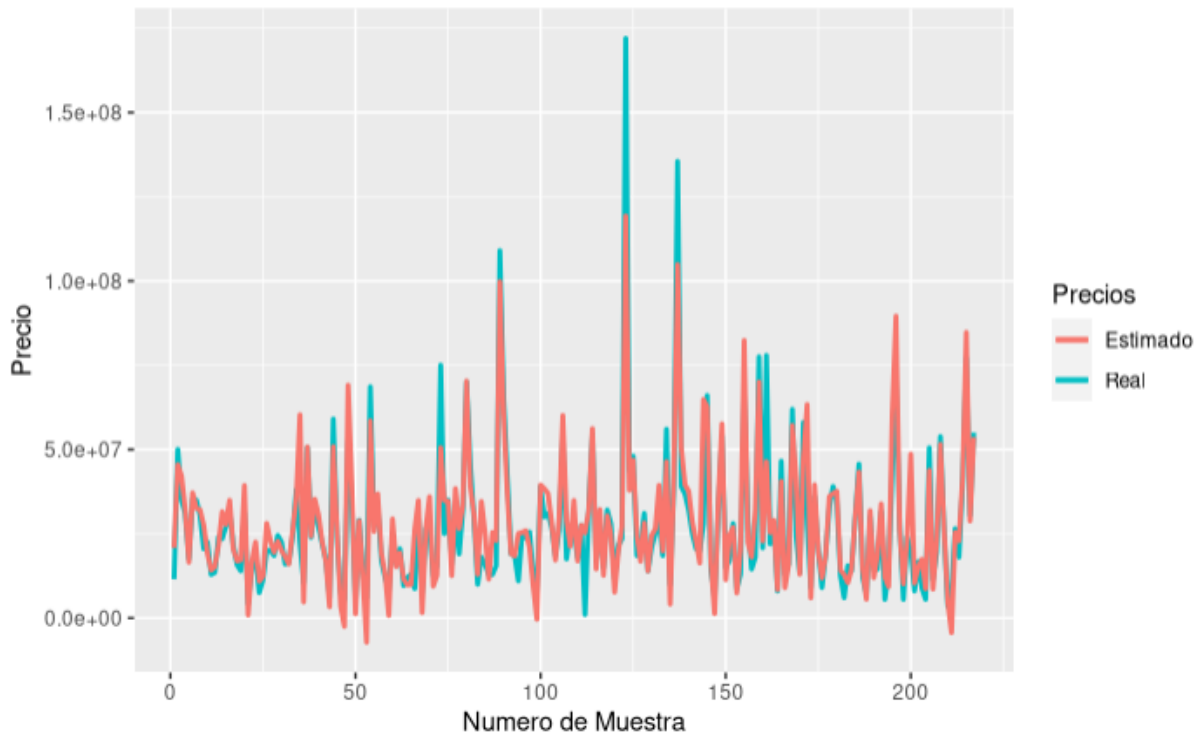


Figura 9: Precio Real Vs Precio Estimado - Regresión lineal Cali

Al visualizar la Figura 9 se puede observar los precios reales vs los precios estimados en la ciudad de Cali. Aunque se observa en el comportamiento de la gráfica que existe similitud entre ambos precios, en la mayoría de los picos más altos el precio real supera al estimado.

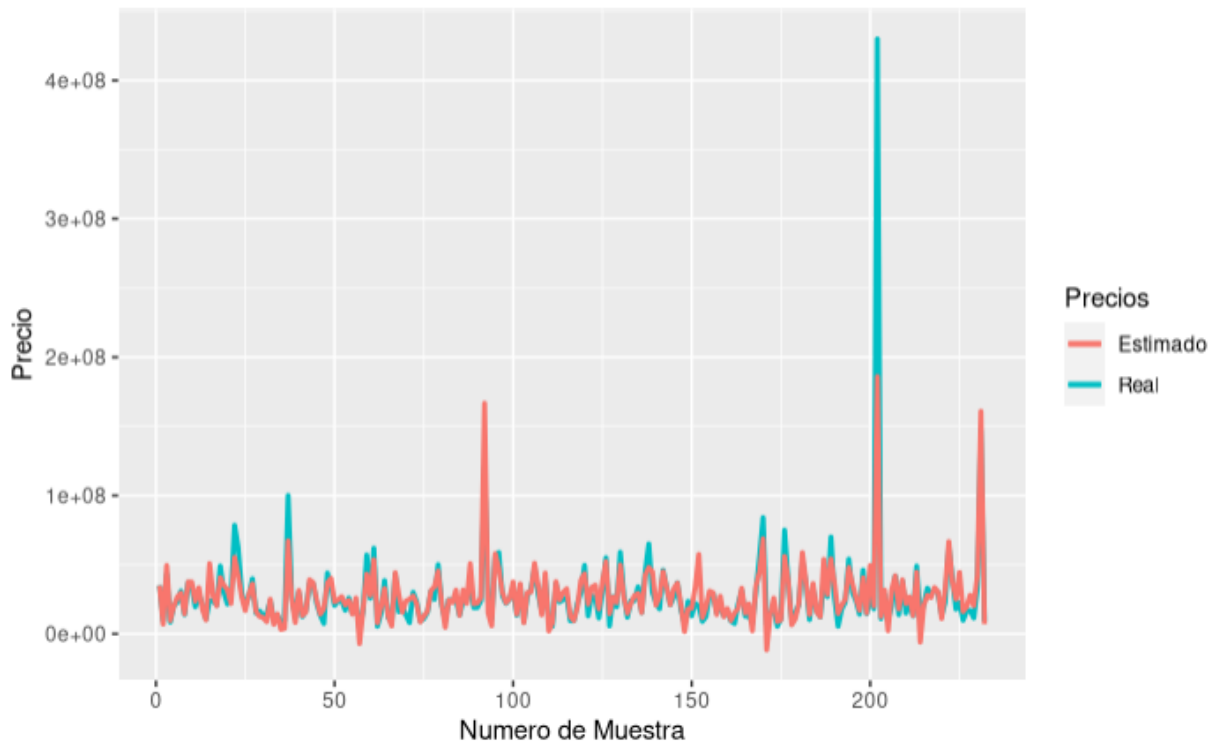


Figura 10: Precio Real Vs Precio Estimado - Regresión lineal Medellín

Con relación a la Figura 11 en la ciudad de Medellín, la regresión lineal tiene como comportamiento que los precios reales superan en su gran mayoría a los precios estimados, excepto que se pueden observar dos picos considerables en la gráfica donde los precios estimados superaron los reales.

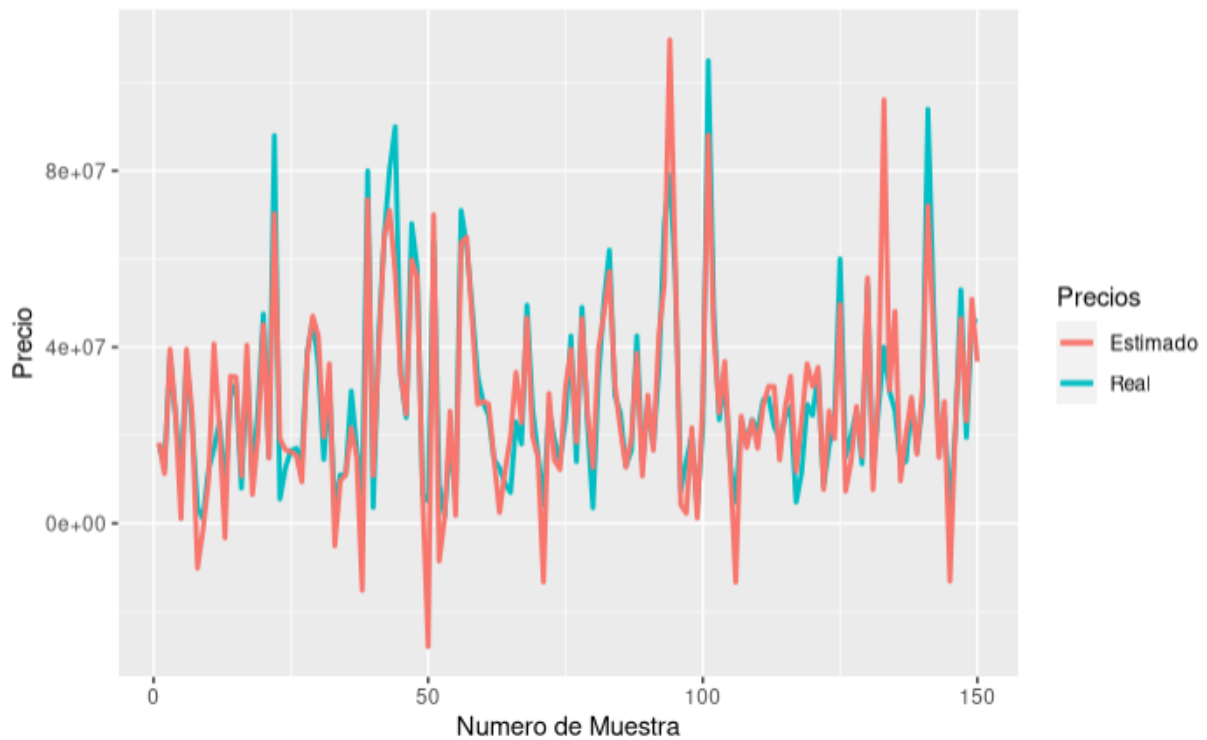


Figura 11: Precio Real Vs Precio Estimado - Regresión lineal Barranquilla

En la Figura 11 de Regresión Lineal para la ciudad de Barranquilla se pueden observar valores fuera del rango ya que son menores que 0.

Para la obtención de las ecuaciones correspondientes al modelo de Regresión para cada ciudad, se procedió a agrupar los registros de acuerdo al precio de los vehículos. Se decidió dejar 20 grupos para obtener la mayor separación de rangos en los precios.

La siguiente es la ecuación correspondiente al modelo de Regresión Lineal para la ciudad de Bogotá, obtenida mediante la herramienta R.

$$\begin{aligned}
\text{Precio} = & 5,108192e^7 + 1,493318e^7 * G10 + 1,634252e^7 * G11 + 1,831160e^7 * G12 \\
& 2,067095e^7 * G13 + 2,293897e^7 * G14 + 2,535965e^7 * G15 + 2,832365e^7 * G16 \\
& 3,197196e^7 * G17 + 3,797692e^7 * G18 + 4,953846e^7 * G19 + 2,906111e^6 * G2 \\
& 8,702769e^7 * G20 + 4,811979e^6 * G3 + 6,620761e^6 * G4 + 7,903426e^6 * G5 \\
& 9,281335e^7 * G6 + 1,060929e^7 * G7 + 1,185846e^7 * G8 + 1,346371e^7 * G9 \\
& - 2,461038e^4 * \text{modelo} + 3,456875e^6 * \text{combustibleEléctrico} \\
& 1,884371e^6 * \text{combustibleGasolina} + 1,560334e^6 * \text{combustibleGasolinayGas} \\
& 9,304853e^4 * \text{transmisiónAutomáticaSecuencial} + 4,662874e^5 * \text{transmisiónMecánica} \\
& 4,204735e^4 * \text{único_dueñoSí} + 1,459728 * \text{kilometraje}
\end{aligned}
\tag{16}$$

La siguiente es la ecuación correspondiente al modelo de Regresión Lineal para la ciudad de Cali, obtenida mediante la herramienta R.

$$\begin{aligned}
\text{Precio} = & - 5,943113e^7 + 1,717684e^7 * G10 + 1,849968e^7 * G11 + 2,042673e^7 * G12 \\
& 2,204913e^7 * G13 + 2,377406e^7 * G14 + 2,611480e^7 * G15 + 2,918907e^7 * G16 \\
& 3,286441e^7 * G17 + 4,009742e^7 * G18 + 5,171075e^7 * G19 + 3,257120e^6 * G2 \\
& 8,557065e^7 * G20 + 6,009911e^6 * G3 + 7,760270e^6 * G4 + 9,144562e^6 * G5 \\
& 1,056935e^7 * G6 + 1,227896e^7 * G7 + 1,436163e^7 * G8 + 1,580501e^7 * G9 \\
& 3,453981e^4 * \text{modelo} - 3,077148e^6 * \text{combustibleGasolina} \\
& - 2,791027e^6 * \text{combustibleGasolinayGas} - 3,584598e^6 * \text{combustibleHíbrido} \\
& - 1,833069e^6 * \text{transmisiónAutomáticaSecuencial} - 1,524558e^6 * \text{transmisiónMecánica} \\
& 5,528315e^5 * \text{único_dueñoSí} - 4,711380 * \text{kilometraje}
\end{aligned}
\tag{17}$$

La siguiente es la ecuación correspondiente al modelo de Regresión Lineal para la ciudad de Barranquilla, obtenida mediante la herramienta R.

$$\begin{aligned}
\text{Precio} = & -4,304236e^7 + 1,839687e^7 * G10 + 1,967241e^7 * G11 + 2,165847e^7 * G12 \\
& 2,340929e^7 * G13 + 2,632474e^7 * G14 + 3,023848e^7 * G15 + 3,549354e^7 * G16 \\
& 4,286160e^7 * G17 + 5,360898e^7 * G18 + 6,709763e^7 * G19 + 3,456766e^6 * G2 \\
& 1,102490e^8 * G20 + 6,595241e^6 * G3 + 8,586137e^6 * G4 + 1,005873e^7 * G5 \\
& 1,147145e^7 * G6 + 1,327653e^6 * G7 + 1,502412e^7 * G8 + 1,656773e^7 * G9 \\
& 2,450209e^4 * \text{modelo} - 7,735966e^5 * \text{combustibleGasolina} \\
& - 4,469895e^5 * \text{combustibleGasolinayGas} \\
& - 2,446416e^6 * \text{transmisiónAutomáticaSecuencial} - 4,584447e^5 * \text{transmisiónMecánica} \\
& 1,008000e^5 * \text{único_dueñoSí} - 4,276219 * \text{kilometraje}
\end{aligned}
\tag{18}$$

La siguiente es la ecuación correspondiente al modelo de Regresión Lineal para la ciudad de Medellín, obtenida mediante la herramienta R.

$$\begin{aligned}
\text{Precio} = & -1,418855e^8 + 1,413327e^7 * G10 + 1,467008e^7 * G11 + 1,682849e^7 * G12 \\
& 1,867628e^7 * G13 + 2,008398e^7 * G14 + 2,289971e^7 * G15 + 2,705211e^7 * G16 \\
& 3,117810e^7 * G17 + 3,985491e^7 * G18 + 5,161395e^7 * G19 + 2,393820e^6 * G2 \\
& 1,021961e^8 * G20 + 3,788840e^6 * G3 + 5,390846e^6 * G4 + 6,40955e^6 * G5 \\
& 7,798967e^7 * G6 + 9,073414e^6 * G7 + 1,094089e^7 * G8 + 1,287553e^7 * G9 \\
& 8,077636e^4 * \text{modelo} - 9,321071e^6 * \text{combustibleGasolina} \\
& - 8,069898e^6 * \text{combustibleGasolinayGas} \\
& - 3,709290e^6 * \text{transmisiónAutomáticaSecuencial} - 2,455701e^6 * \text{transmisiónMecánica} \\
& 6,526921e^5 * \text{único_dueñoSí} - 1,321847e^1 * \text{kilometraje}
\end{aligned}
\tag{19}$$

Para los otros modelos, las ecuaciones no son explícitas porque son modelos basados en un enfoque diferente.

En la siguiente tabla se pueden observar los parámetros estimados de la Regresión Lineal para la ciudad de Bogotá.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
marca	54	9.1627e+16	1.6968e+15	16.6058	$< 2,2e - 16$	***
linea	261	2.8350e+17	1.0862e+15	10.6302	$< 2,2e - 16$	***
modelo	1	3.6914e+16	3.6914e+16	361.2647	$< 2,2e - 16$	***
combustible	2	1.2080e+15	6.0401e+14	5.9112	0.002825	**
transmisión	2	2.0298e+15	1.0149e+15	9.9323	5.466e-05	***
único_dueño	1	8.9833e+13	8.9833e+13	0.8792	0.348707	
kilometraje	1	1.0201e+150	1.0201e+15	9.9837	0.001637	**
Residuals	823	88.4094e+16	1.0218e+14			

Tabla 18: Tabla de parámetros Regresión Lineal - Bogotá

En la Tabla 18, se puede observar los atributos que son estadísticamente significativos para la predicción del precio. Esto se corrobora por los asteriscos en la última columna de la tabla. Basicamente todos los atributos explican con mucha significancia, excepto *único_dueño* que es la única variable que no influye sobre el precio.

En la siguiente tabla se pueden observar los parámetros estimados de la Regresión Lineal para la ciudad de Cali.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
marca	38	1.0918e+17	2.8732e+15	58.8376	$< 2,2e - 16$	***
linea	179	1.5773e+17	8.8117e+14	18.0446	$< 2,2e - 16$	***
modelo	1	3.2563e+16	3.2563e+16	666.8328	$< 2,2e - 16$	***
combustible	3	8.0879e+14	2.6960e+14	5.5208	0.0009548	***
transmisión	2	1.2072e+15	6.0359e+14	12.3605	5.409e-06	***
único_dueño	1	6.8550e+14	6.8550e+14	14.0377	0.0001954	***
kilometraje	1	6.6096e+14	6.6096e+14	13.5353	0.0002539	***
Residuals	639	3.1204e+16	4.8833e+13			

Tabla 19: Tabla de parámetros Regresión Lineal - Cali

En la Tabla 19, se puede observar que basicamente todos los atributos explican con mucha significancia el precio del vehículo.

En la siguiente tabla se pueden observar los parámetros estimados de la Regresión Lineal para la ciudad de Baranquilla.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
marca	31	1.2851e+17	4.1454e+15	61.9718	$< 2,2e - 16$	***
linea	154	1.5226e+17	9.8873e+14	14.7810	$< 2,2e - 16$	***
modelo	1	3.4298e+16	3.4298e+16	512.7320	$< 2,2e - 16$	***
combustible	2	5.7654e+14	2.8827e+14	4.3095	0.014064	*
transmisión	2	3.4558e+14	1.7279e+14	2.5831	0.076783	.
único_dueño	1	1.8522e+14	1.8522e+14	2.7690	0.096884	.
kilometraje	1	7.0238e+14	7.0238e+14	10.5002	0.001292	**
Residuals	404	2.7024e+16	6.6892e+13			

Tabla 20: Tabla de parámetros Regresión Lineal - Barranquilla

En la Tabla 20, se puede observar que los atributos estadísticamente más significativos para la predicción del precio son marca, línea, modelo y kilometraje.

En la siguiente tabla se pueden observar los parámetros estimados de la Regresión Lineal para la ciudad de Medellín.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
marca	35	1.6699e+17	4.7712e+15	83.5482	$< 2,2e - 16$	***
linea	175	2.5788e+17	1.4736e+15	25.8035	$< 2,2e - 16$	***
modelo	1	4.7754e+16	4.7754e+16	836.2040	$< 2,2e - 16$	***
combustible	2	8.4287e+14	4.2143e+14	7.3796	0.0006730	***
transmisión	3	1.0794e+15	3.5980e+14	6.3004	0.0003204	***
único_dueño	1	1.4819e+15	1.4819e+15	25.9489	4.501e-07	***
kilometraje	1	4.8365e+14	4.8365e+14	8.4690	0.0037258	**
Residuals	708	4.0432e+16	5.7108e+13			

Tabla 21: Tabla de parámetros Regresión Lineal - Medellín

En la Tabla 21, se puede observar que todos los atributos son estadísticamente significativos para la predicción del precio.

• Regresión Lineal Generalizada - Modelo 2

En Figura 12 se encuentran las especificaciones correspondientes al modelo de Regresión Lineal Generalizada.

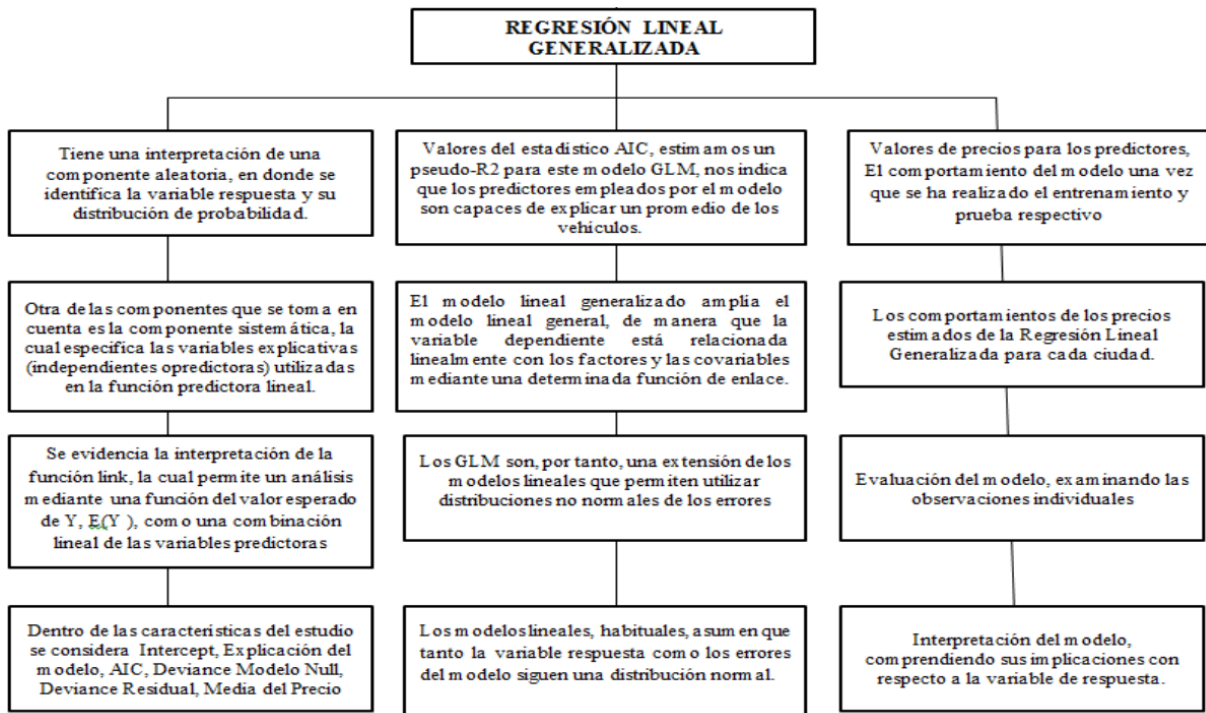


Figura 12: Especificaciones - Regresión Lineal Generalizada

Aplicando un modelo lineal generalizado con respuesta *Poisson*, nuestro objetivo fue modelizar los conjuntos de datos utilizando este tipo de distribución.

En la siguiente tabla se muestran los resultados estadísticos más relevantes al implementar este modelo usando la función *glm*:

Característica	Bogotá	Cali	Medellín	Barranquilla
Intercept	-2.695364e+09	-3.020652e+09	-3.385716e+09	-4.462850e+09
Explicación del modelo	93.53 %	95.05 %	96.44 %	96.93 %
AIC	907095750.30	503148908.84	482064098.03	272681578.93
Deviance Modelo Null	14013389467.27	10165084079.55	13554405518.14	8895209687.37
Deviance Residual	907073768.81	503132261.00	482046286.76	272669977.04
Media del Precio	24617833.77	27172235.57	28300312.96	29671004.04

Tabla 22: Resultados Relevantes - Regresión Lineal Generalizada

Según resultados de la tabla anterior, podemos afirmar lo siguiente:

- En primer lugar, vemos como los valores del estadístico AIC, parece bastante

elevados.

- Sin embargo, si estimamos un pseudo- R^2 para este modelo GLM, nos indica que los predictores empleados por el modelo son capaces de explicar un promedio por encima del 93,53 % de la variabilidad observada en el precio de los vehículos. Lo cual es superior con respecto al modelo LM, aplicado anteriormente.

Valores de precios para los predictores

El comportamiento del modelo una vez que se ha realizado el entrenamiento y prueba respectivo, dan los siguientes resultado para cada ciudad:

	Bogotá	Cali	Medellín	Barranquilla
RMSE	6907320	5387073	4775375	9879942
pseudo – R^2	93.53	95.05	96.44	96.93

Tabla 23: Comportamiento - Regresión Lineal Generalizada

A continuación, en las figuras siguientes se pueden observar los comportamientos de los precios estimados de la Regresión Lineal Generalizada para cada ciudad.

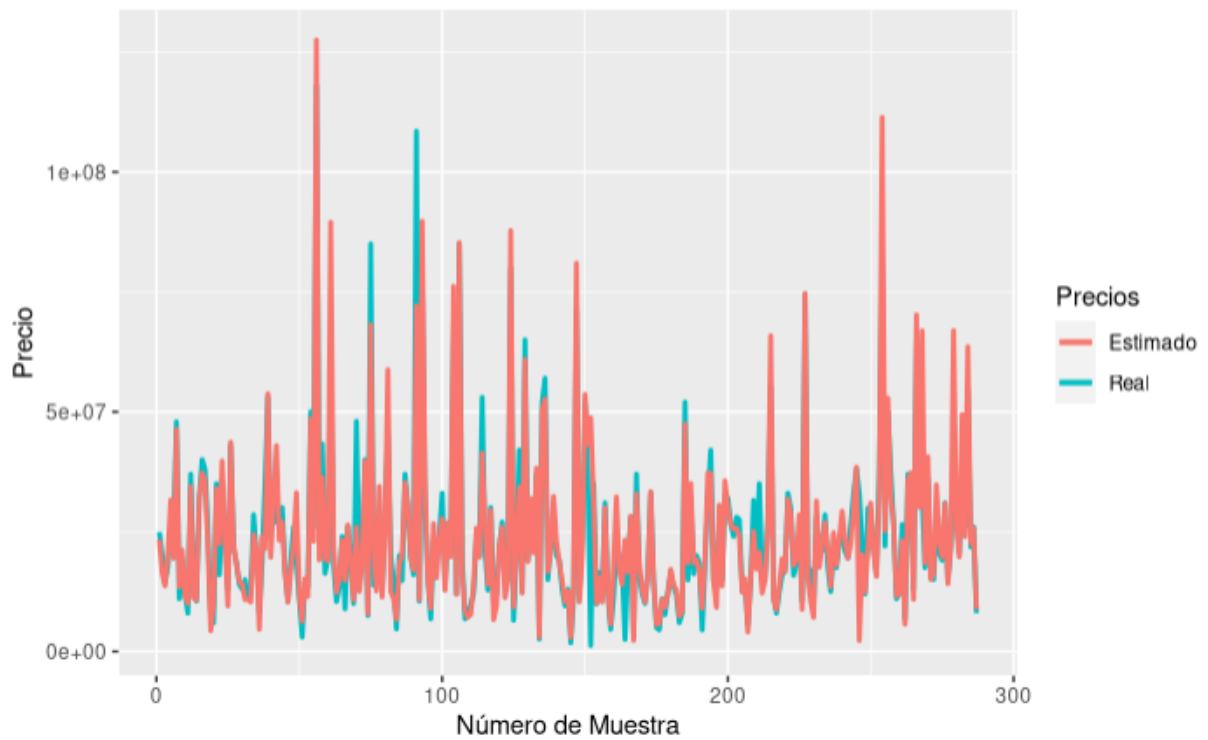


Figura 13: Precio Real Vs Precio Estimado - Regresión Lineal Generalizada Bogotá

En la Figura 13 se puede observar el precio Real Vs Precio Estimado por la Regresión Lineal Generalizada en la ciudad de Bogotá, en este caso varios precios estimados superan los precios reales del modelo.

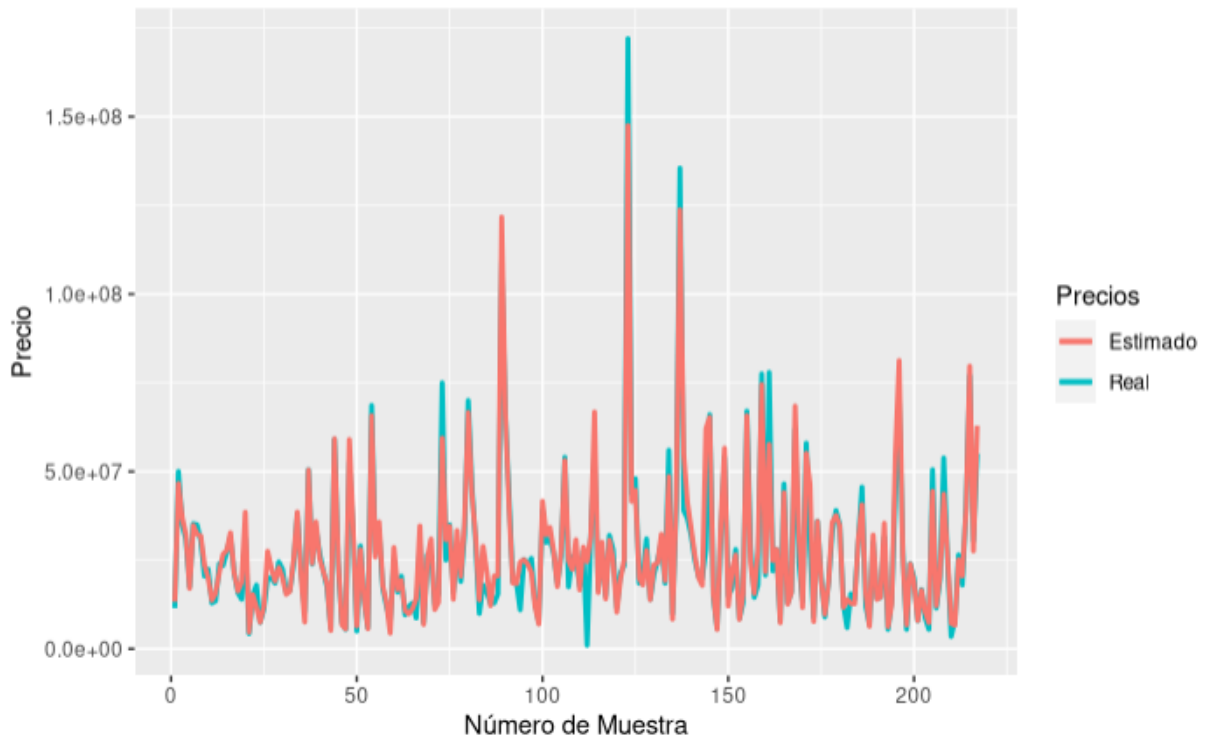


Figura 14: Precio Real Vs Precio Estimado - Regresión Lineal Generalizada Cali

En la Figura 14 se tiene la Regresión Lineal Generalizada para la ciudad de Cali en donde se puede observar que los precios estimados en su gran mayoría superan los reales.

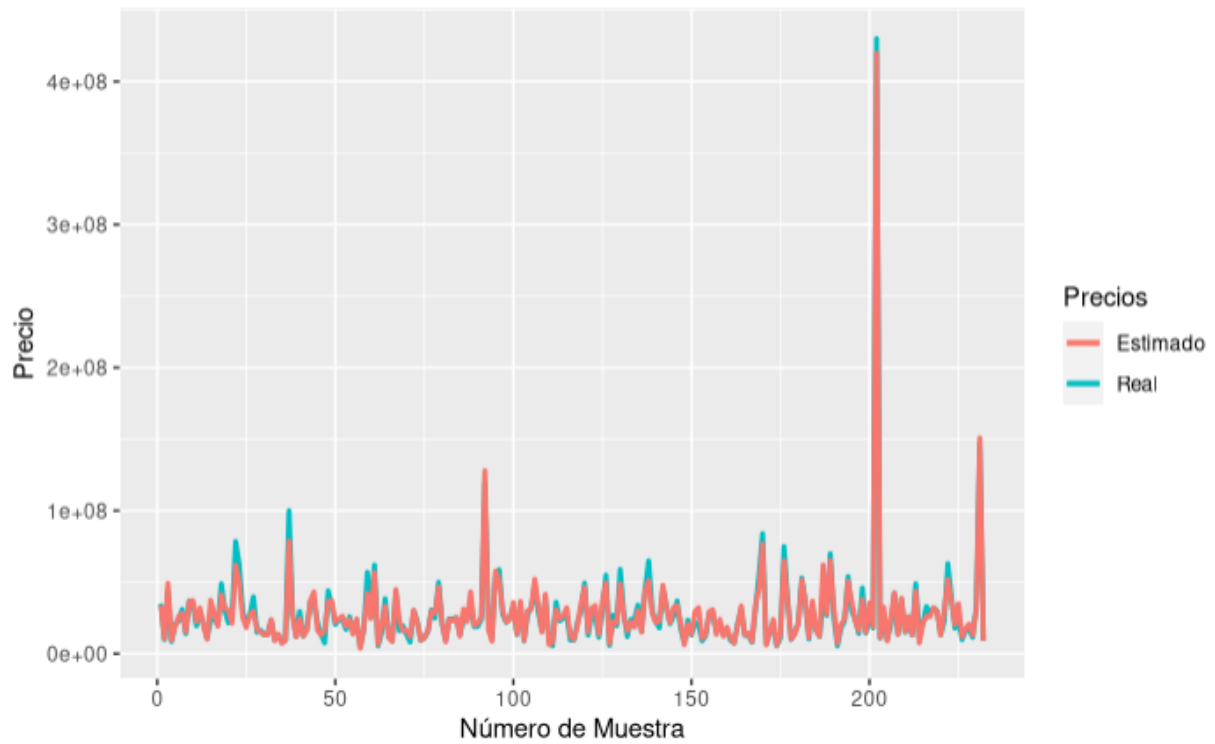


Figura 15: Precio Real Vs Precio Estimado - Regresión Lineal Generalizada Medellín

En la figura 15 para la ciudad de Medellín, existe una similitud considerable entre el precio real y el precio estimado en el modelo de Regresión Lineal Generalizada con un pico alto para una muestra de 200 con precio mayor a $4e+08$.

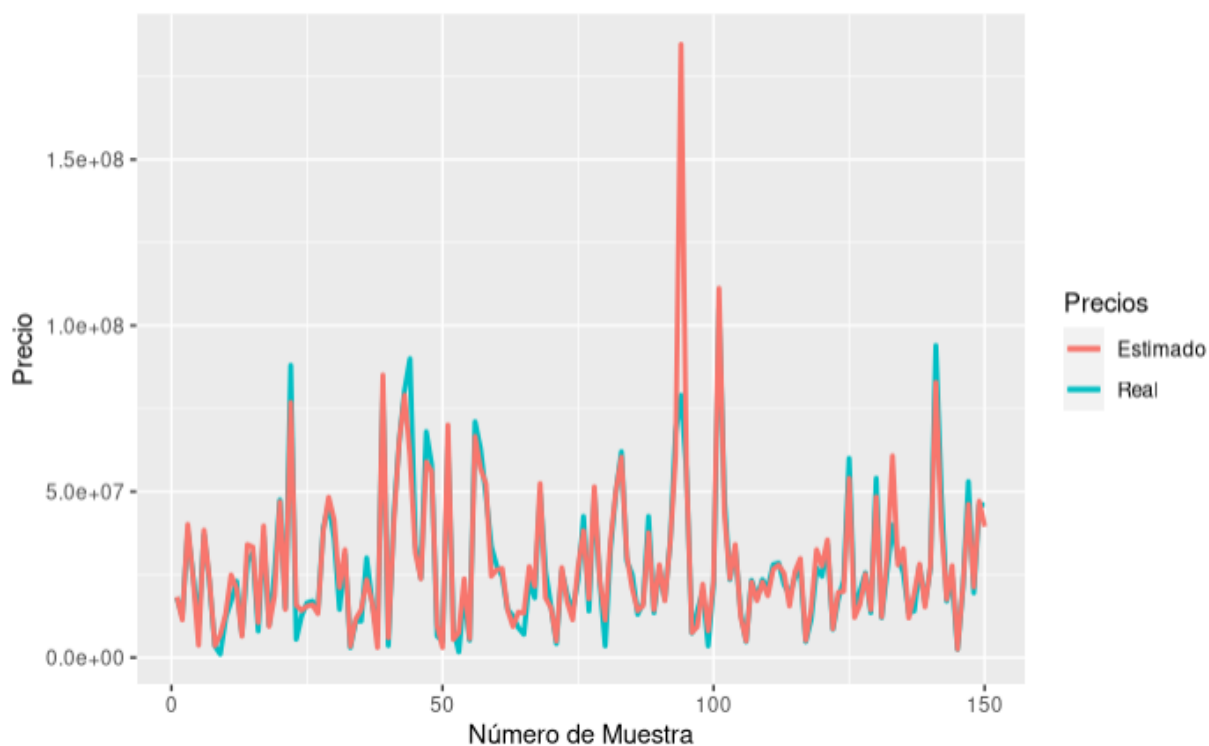


Figura 16: Precio Real Vs Precio Estimado - Regresión Lineal Generalizada Barranquilla

En la figura 16 para la ciudad de Barranquilla se observa que el precio estimado tiene mayor picos que los precios reales según la Regresión Lineal Generalizada. Además se observa que los precios estimados parecen estar en el rango correcto, es decir son mayores que 0.

Esto se puede corroborar calculando los valores mínimos de los predictores:

	Bogotá	Cali	Medellín	Barranquilla
Mínimos	2264445.69	4435788.44	3792255.19	2526357.97

Tabla 24: Valores mínimos predictores - Regresión Lineal Generalizada

• Regresión Lineal Polinómica - Modelo 3

En Figura 17 se encuentran las especificaciones correspondientes al modelo de Regresión Lineal Polinómica.

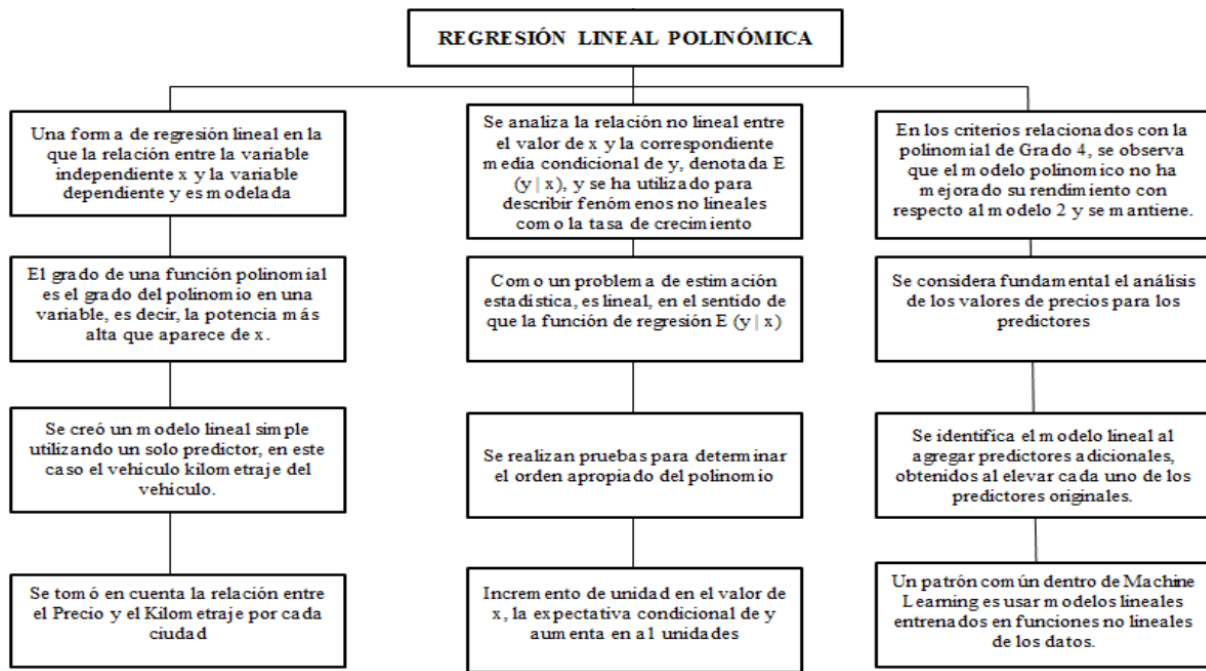


Figura 17: Especificaciones - Regresión Lineal Polinómica

Análisis Preliminar

Se creó un modelo lineal simple utilizando un sólo predictor, en este caso el kilometraje del vehículo. Los resultados se evidencian en la figura 18

Cómo se observa, el modelo aplicado es muy simple y explica muy poco porcentaje de variabilidad, pero el objetivo de este modelo será verificar si es posible aplicar un modelo polinomial sobre la variable kilometraje.

En la Figura 19 se observa de manera más clara la relación entre el *Precio* y el *Kilometraje* por cada ciudad.

Sin embargo, si se aplica un modelo basado en regresión lineal polinomial, se observa una leve mejora respecto del modelo simple, cómo se observa en la Figura 20.

Basados en este análisis se puede considerar aplicar una regresión polinomial, para el predictor *Kilometraje*.

```
Call:
lm(formula = precio ~ kilometraje, data = autos_df_train)

Residuals:
    Min       1Q   Median       3Q      Max
-31620791 -9698135 -4737690  4511731 321737309

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.724e+07  1.077e+06   34.58  <2e-16 ***
kilometraje -1.060e+02  7.664e+00  -13.83  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 19350000 on 1144 degrees of freedom
Multiple R-squared:  0.1432,    Adjusted R-squared:  0.1425
F-statistic: 191.2 on 1 and 1144 DF,  p-value: < 2.2e-16
```

(a) Bogotá

```
Call:
lm(formula = precio ~ kilometraje, data = autos_df_train)

Residuals:
    Min       1Q   Median       3Q      Max
-34937952 -11773569 -6663614  4651904 205740339

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.000e+07  1.280e+06   31.25  <2e-16 ***
kilometraje -1.019e+02  9.161e+00  -11.13  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 22230000 on 925 degrees of freedom
Multiple R-squared:  0.118,    Adjusted R-squared:  0.1171
F-statistic: 123.8 on 1 and 925 DF,  p-value: < 2.2e-16
```

(c) Medellín

```
Call:
lm(formula = precio ~ kilometraje, data = autos_df_train)

Residuals:
    Min       1Q   Median       3Q      Max
-34963725 -9194930 -4216692  4170779 182548805

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.731e+07  1.109e+06   33.63  <2e-16 ***
kilometraje -9.557e+01  8.571e+00  -11.15  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 18690000 on 863 degrees of freedom
Multiple R-squared:  0.1259,    Adjusted R-squared:  0.1249
F-statistic: 124.3 on 1 and 863 DF,  p-value: < 2.2e-16
```

(b) Cali

```
Call:
lm(formula = precio ~ kilometraje, data = autos_df_train)

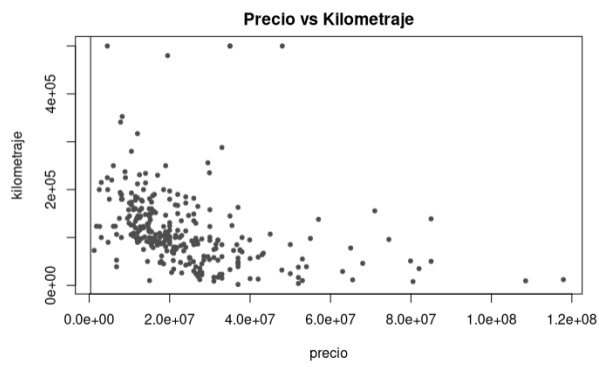
Residuals:
    Min       1Q   Median       3Q      Max
-36054870 -12524802 -5977383  4573597 201172804

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.088e+07  1.627e+06   25.12  < 2e-16 ***
kilometraje -1.028e+02  1.224e+01  -8.398  3.32e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

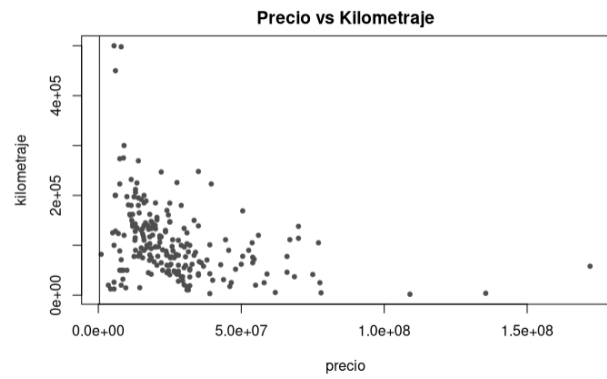
Residual standard error: 22740000 on 595 degrees of freedom
Multiple R-squared:  0.106,    Adjusted R-squared:  0.1045
F-statistic: 70.53 on 1 and 595 DF,  p-value: 3.323e-16
```

(d) Barranquilla

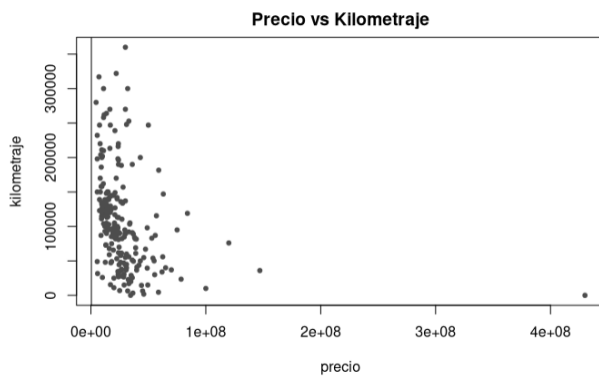
Figura 18: Resultado un sólo predictor - Kilometraje



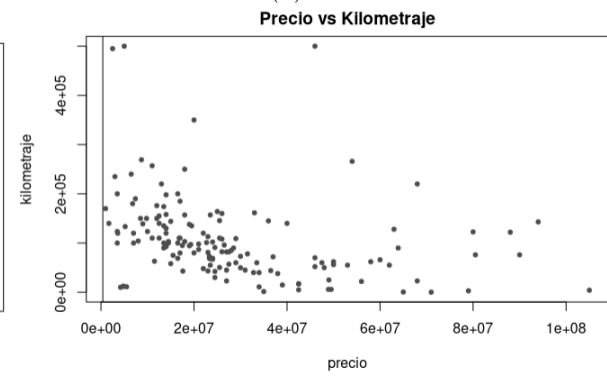
(a) Bogotá



(b) Cali



(c) Medellín



(d) Barranquilla

Figura 19: Precio Vs Kilometraje - Regresión Lineal Polinómica

```
Call:
lm(formula = precio ~ poly(kilometraje, 4), data = autos_df_train)

Residuals:
    Min       1Q   Median       3Q      Max
-40638661 -8615959 -4270272  4663534 323193118

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    24511560      552737  44.346 < 2e-16 ***
poly(kilometraje, 4)1 -267559185    18711599 -14.299 < 2e-16 ***
poly(kilometraje, 4)2  149624549    18711599  7.996 3.12e-15 ***
poly(kilometraje, 4)3 -54516013    18711599 -2.913 0.00364 **
poly(kilometraje, 4)4  14888087    18711599  0.796 0.42639
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 18710000 on 1141 degrees of freedom
Multiple R-squared:  0.1956,    Adjusted R-squared:  0.1928
F-statistic: 69.38 on 4 and 1141 DF,  p-value: < 2.2e-16
```

(a) Bogotá

```
Call:
lm(formula = precio ~ poly(kilometraje, 4), data = autos_df_train)

Residuals:
    Min       1Q   Median       3Q      Max
-39639488 -8661825 -3993579  3923503 177557429

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    27172236      627339  43.313 < 2e-16 ***
poly(kilometraje, 4)1 -208413752    18450586 -11.296 < 2e-16 ***
poly(kilometraje, 4)2  91876600    18450586  4.980 7.7e-07 ***
poly(kilometraje, 4)3 -15110783    18450586 -0.819 0.413
poly(kilometraje, 4)4 -6679691    18450586 -0.362 0.717
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 18450000 on 860 degrees of freedom
Multiple R-squared:  0.1512,    Adjusted R-squared:  0.1473
F-statistic: 38.3 on 4 and 860 DF,  p-value: < 2.2e-16
```

(b) Cali

```
Call:
lm(formula = precio ~ poly(kilometraje, 4), data = autos_df_train)

Residuals:
    Min       1Q   Median       3Q      Max
-48221600 -10162713 -4977003  4224613 195048260

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    28300313      702651  40.277 < 2e-16 ***
poly(kilometraje, 4)1 -247337096    21393376 -11.561 < 2e-16 ***
poly(kilometraje, 4)2  158038595    21393376  7.387 3.35e-13 ***
poly(kilometraje, 4)3 -93924367    21393376 -4.390 1.26e-05 ***
poly(kilometraje, 4)4  37228245    21393376  1.740 0.0822 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 21390000 on 922 degrees of freedom
Multiple R-squared:  0.1859,    Adjusted R-squared:  0.1824
F-statistic: 52.64 on 4 and 922 DF,  p-value: < 2.2e-16
```

(c) Medellín

```
Call:
lm(formula = precio ~ poly(kilometraje, 4), data = autos_df_train)

Residuals:
    Min       1Q   Median       3Q      Max
-45362304 -11393811 -4951640  5002846 193755342

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    29671004      908530  32.658 < 2e-16 ***
poly(kilometraje, 4)1 -190986755    22198639 -8.604 < 2e-16 ***
poly(kilometraje, 4)2  118334633    22198639  5.331 1.39e-07 ***
poly(kilometraje, 4)3 -43207102    22198639 -1.946 0.0521 .
poly(kilometraje, 4)4 -10118245    22198639 -0.456 0.6487
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 22200000 on 592 degrees of freedom
Multiple R-squared:  0.1524,    Adjusted R-squared:  0.1467
F-statistic: 26.61 on 4 and 592 DF,  p-value: < 2.2e-16
```

(d) Barranquilla

Figura 20: Resultado un sólo predictor - Kilometraje grado 4

Polinomial de Grado 4

En la siguiente tabla se muestran los resultados más relevantes al implementar la Regresión Polinomial de Grado 4 usando la función *lm*:

Característica	Bogotá	Cali	Medellín	Barranquilla
Intercept	-2.695364e+09	-3.020652e+09	-3.385716e+09	-4.462850e+09
Explicación del modelo	83.51 %	90.78 %	92.74 %	93.00 %
Dsv Std	10018304.85	7067361.05	7298908.66	7738946.52
Media del Precio	24617833.77	27172235.57	28300312.96	29671004.04
R-sq Adj	0.77	0.87	0.90	0.90
Estd. F	12.84	27.61	41.02	27.55

Tabla 25: Resultados Relevantes - Regresión Polinomial de Grado 4

Analizando los resultados obtenidos, se observa que el modelo polinómico no ha mejorado su rendimiento con respecto al modelo 2 y se mantiene constante con respecto al modelo 1.

Valores de precios para los predictores

Las métricas obtenidas de este modelo se muestran a continuación:

	Bogotá	Cali	Medellín	Barranquilla
RMSE	9143613	8691854	17592926	8998966
R²	83.51	90.78	92.74	93

Tabla 26: Comportamiento - Regresión Polinomial de Grado 4

A continuación, en las siguientes figuras se puede observar los comportamientos de los precios estimados de la Regresión Polinomial Grado 4.

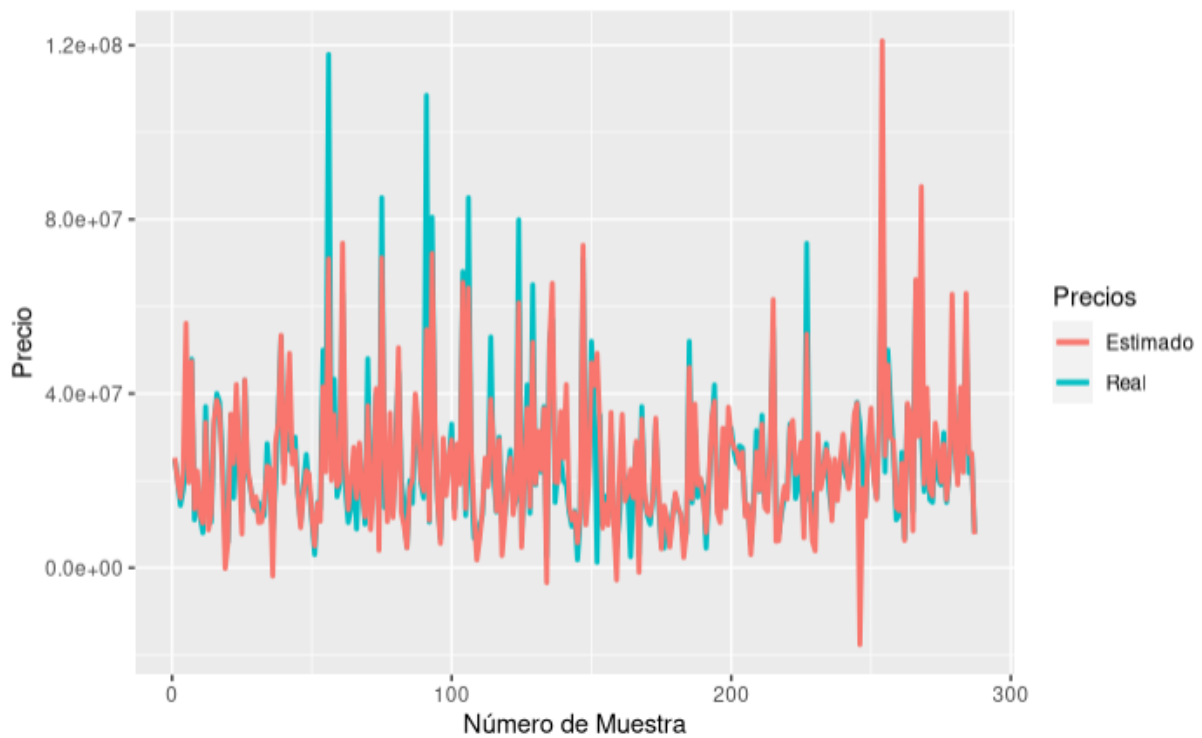


Figura 21: Precio Real Vs Precio Estimado - Regresión Polinomial Grado 4 Bogotá

En la Figura 21 se puede observar que en la ciudad de Bogotá para la Regresión Polinomial grado 4 existen valores por debajo de 0.

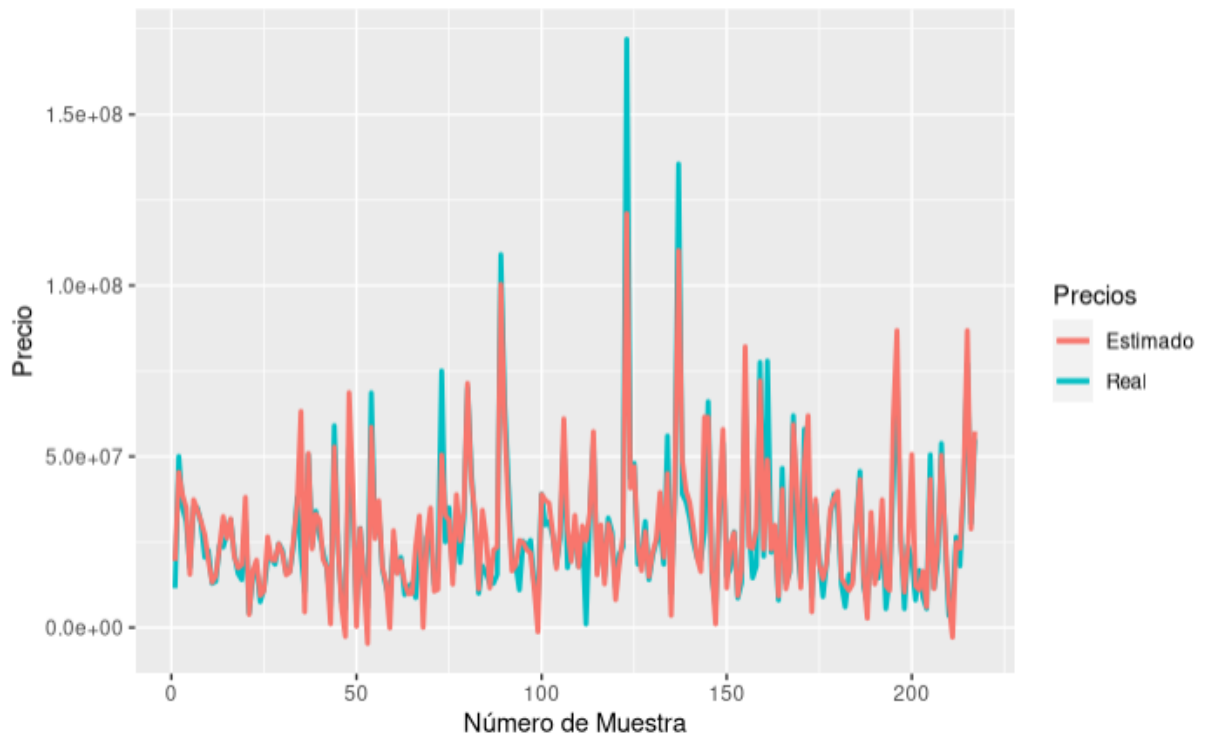


Figura 22: Precio Real Vs Precio Estimado - Regresión Polinomial Grado 4 Cali

En la Figura 22 se puede observar que en la ciudad de Cali para la Regresión Polinomial grado 4 existen valores por debajo de 0.

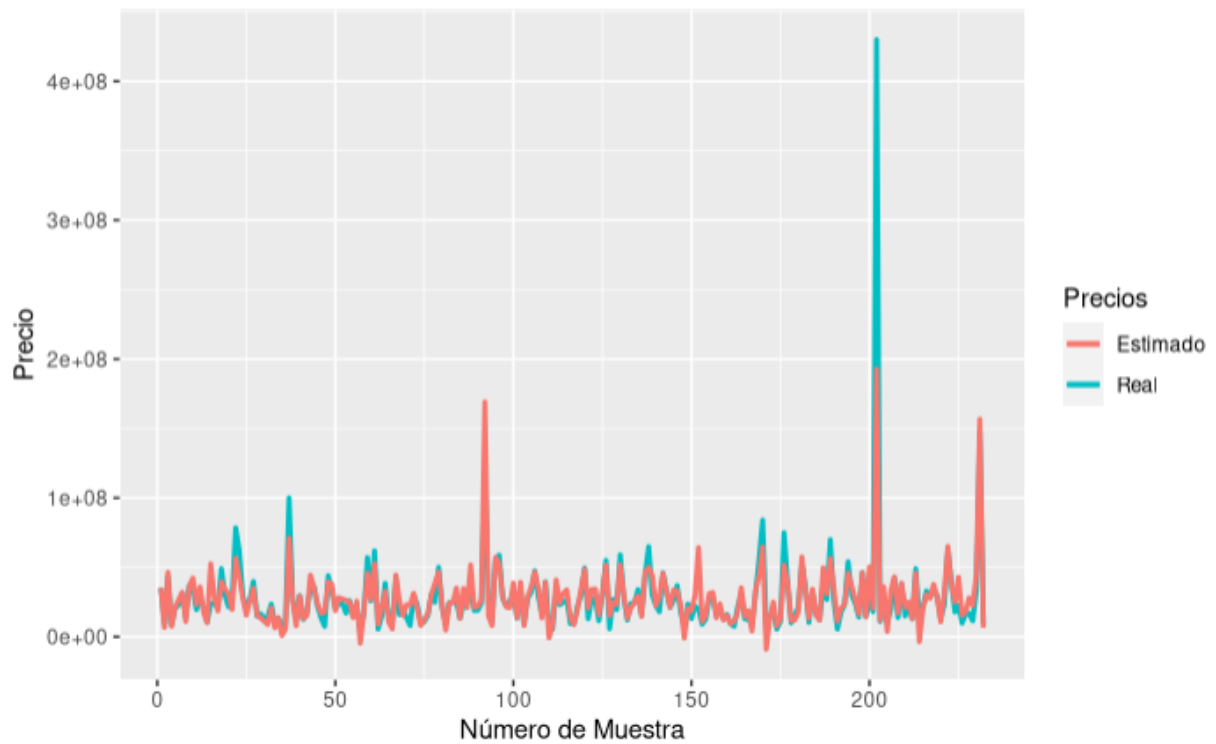


Figura 23: Precio Real Vs Precio Estimado - Regresión Polinomial Grado 4 Medellín

En la Figura 23 se puede observar que en la ciudad de Medellín para la Regresión Polinomial grado 4 existen valores por debajo de 0.

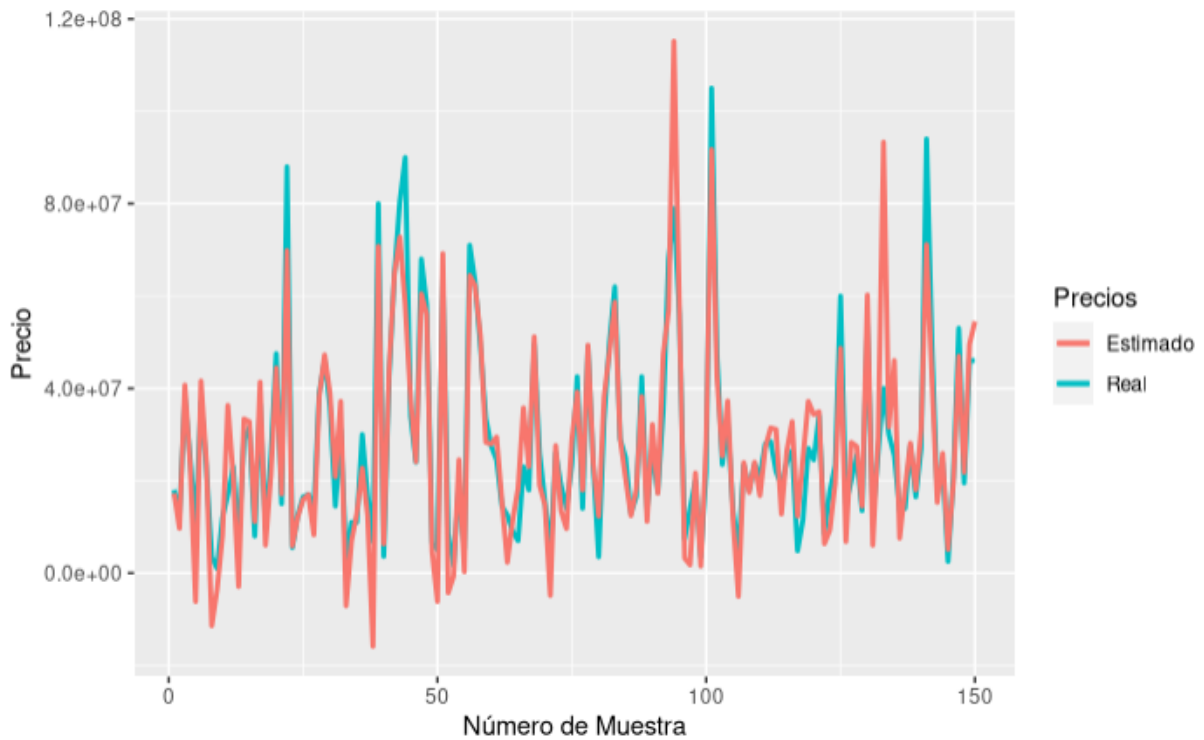


Figura 24: Precio Real Vs Precio Estimado - Regresión Polinomial Grado 4 Barranquilla

En la Figura 24 se puede observar que en la ciudad de Barranquilla para la Regresión Polinomial grado 4 existen valores por debajo de 0.

- **Árbol de Regresión - Modelo 4**

En Figura 25 se encuentran las especificaciones correspondientes al modelo de Árbol de Regresión.

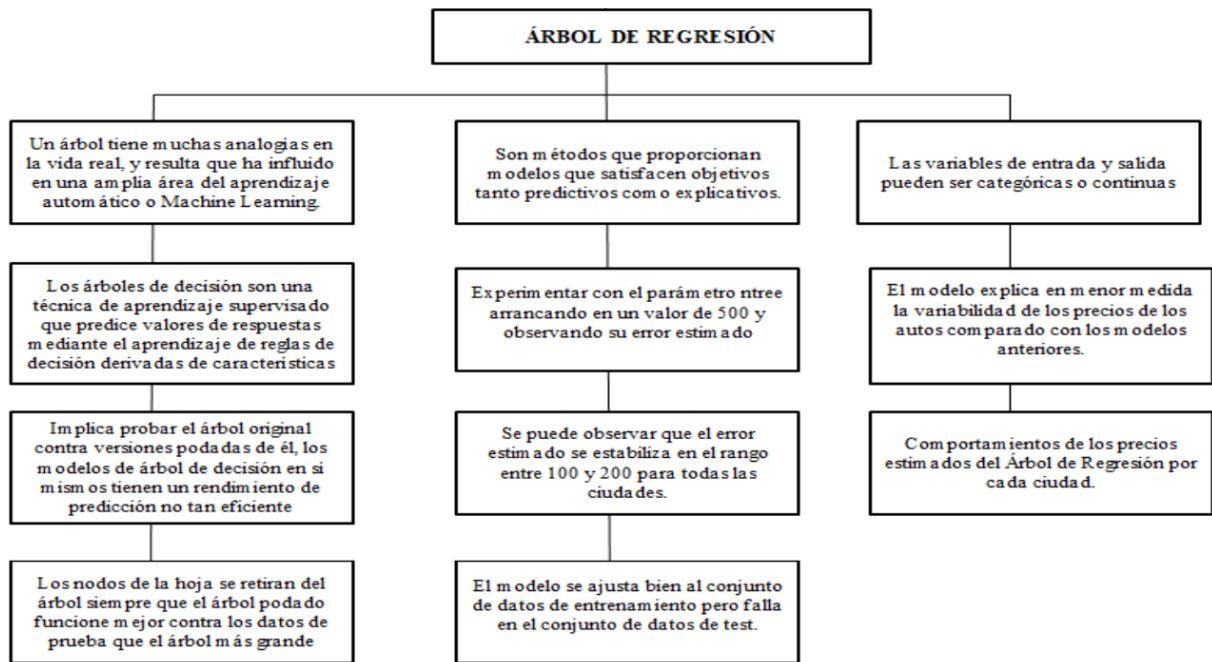


Figura 25: Especificaciones - Árbol de Regresión

Un árbol de regresión consiste en hacer preguntas de tipo $x_k \leq c$ que involucre cada una de las variables utilizadas en el modelo, de esta forma el espacio de las covariables es dividido en hiper-rectángulos y todas las observaciones que queden dentro de un hiper-rectángulo tendrán el mismo valor estimado $\hat{y}$.

En la Figura 26, se muestra la estructura del árbol de la ciudad de Bogotá. Los árboles de las otras ciudades no se muestran ya que la estructura es muy similar entre ellas.

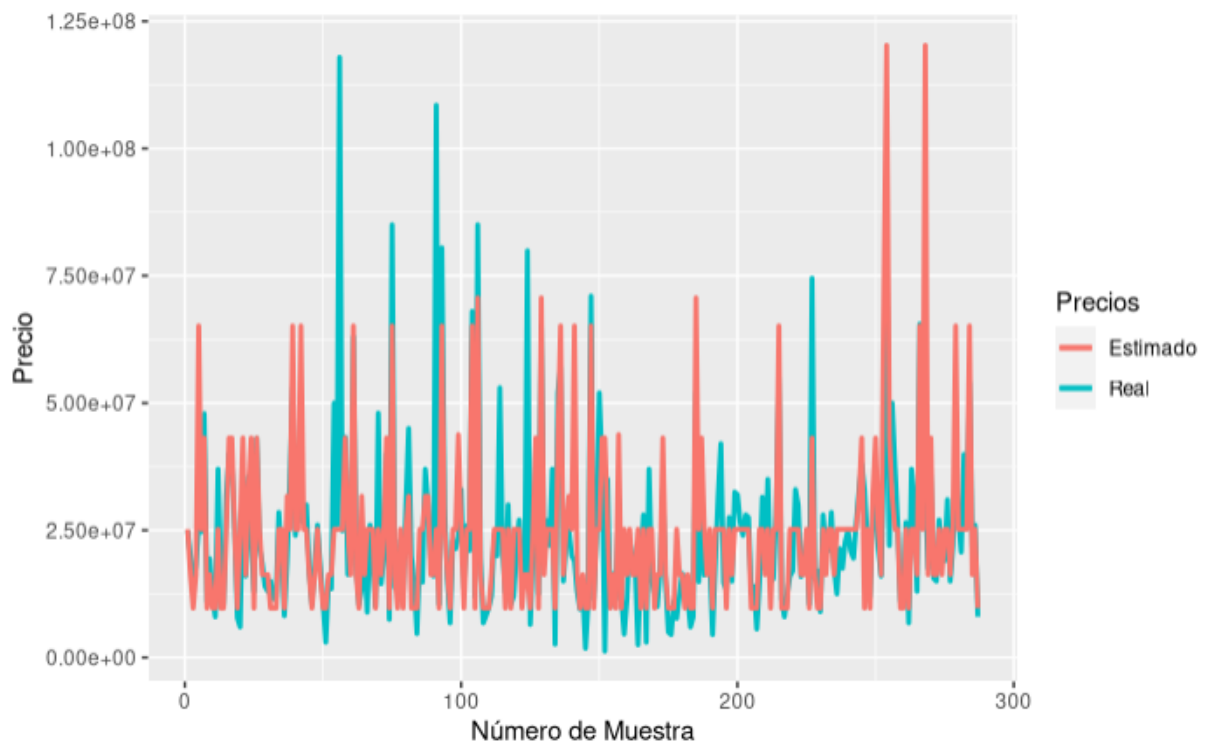


Figura 27: Precio Real Vs Precio Estimado - Árbol de Regresión Bogotá

En la Figura 27 se puede observar que en la ciudad de Bogotá para el árbol de Regresión existe una gran diferencia entre los precios reales y los estimados.

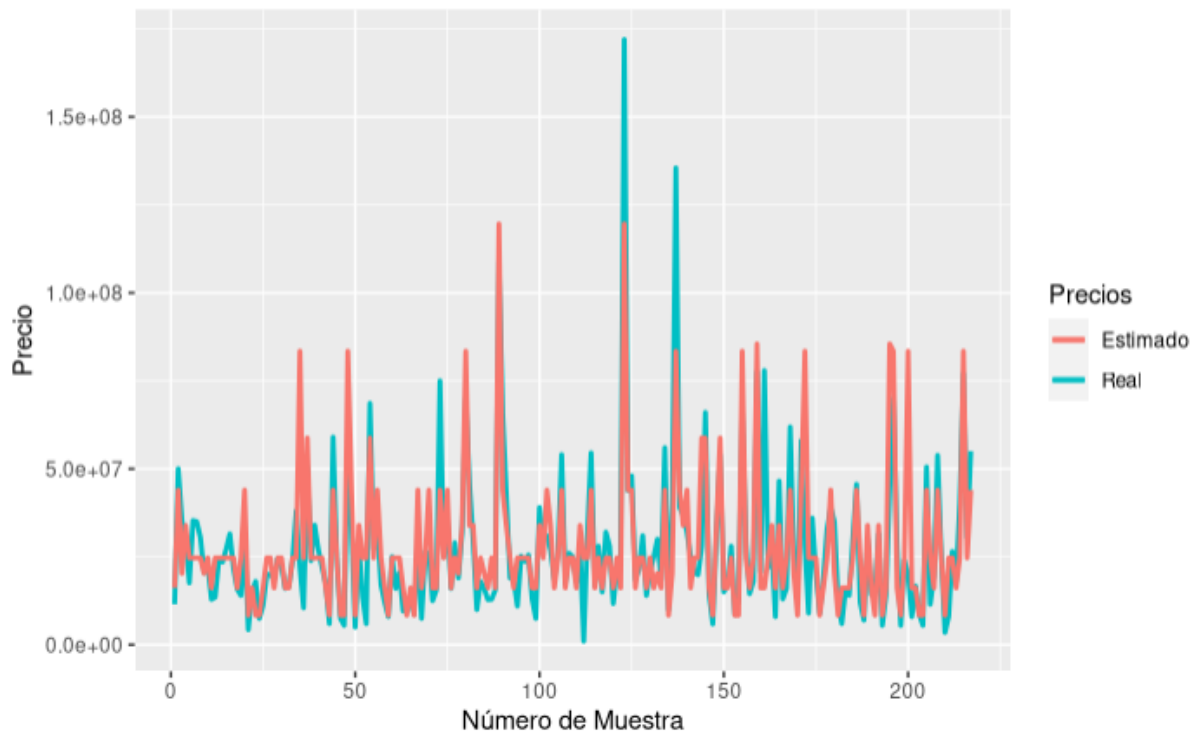


Figura 28: Precio Real Vs Precio Estimado - Árbol de Regresión Cali

En la Figura 28 se puede observar que en la ciudad de Cali para el árbol de Regresión existe similitud entre los precios reales y los precios estimados.

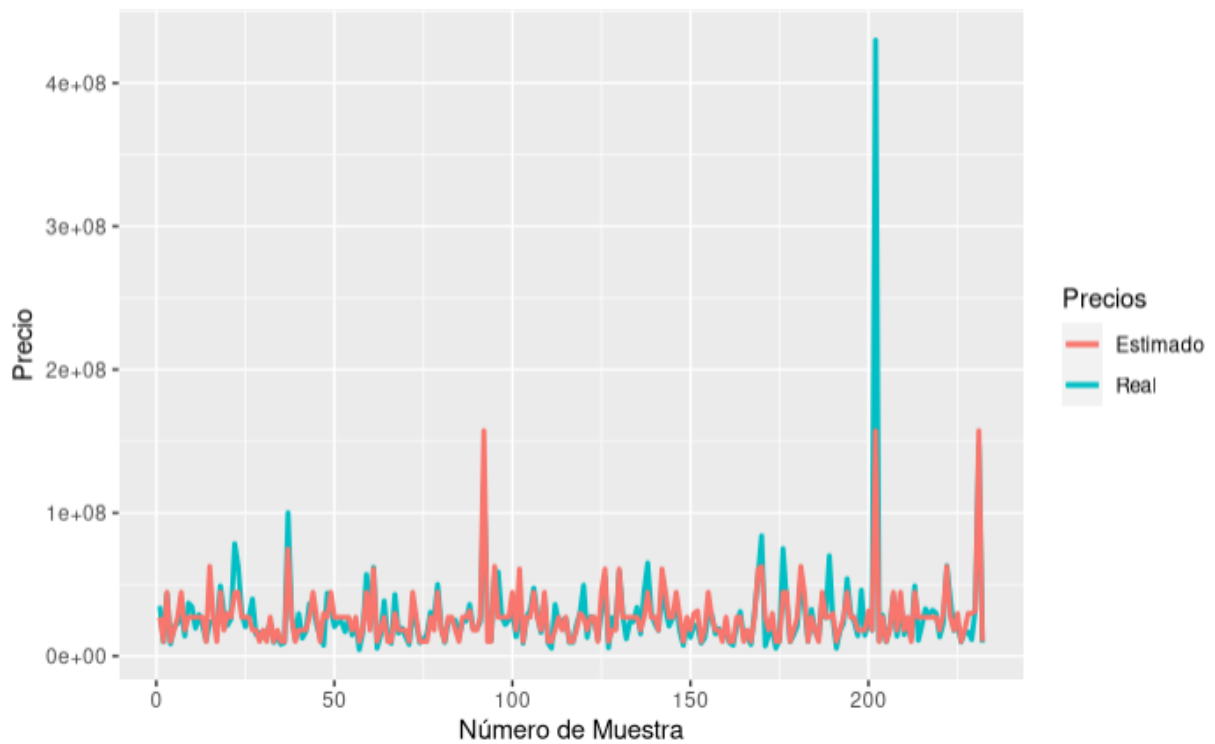


Figura 29: Precio Real Vs Precio Estimado - Árbol de Regresión Medellín

En la Figura 29 se puede observar que en la ciudad de Medellín para el árbol de Regresión existe similitud entre los precios reales y los precios estimados, excepto para una muestra mayor de 200 y un precio superior a $4e+08$ que existe un pico considerable en el precio real.

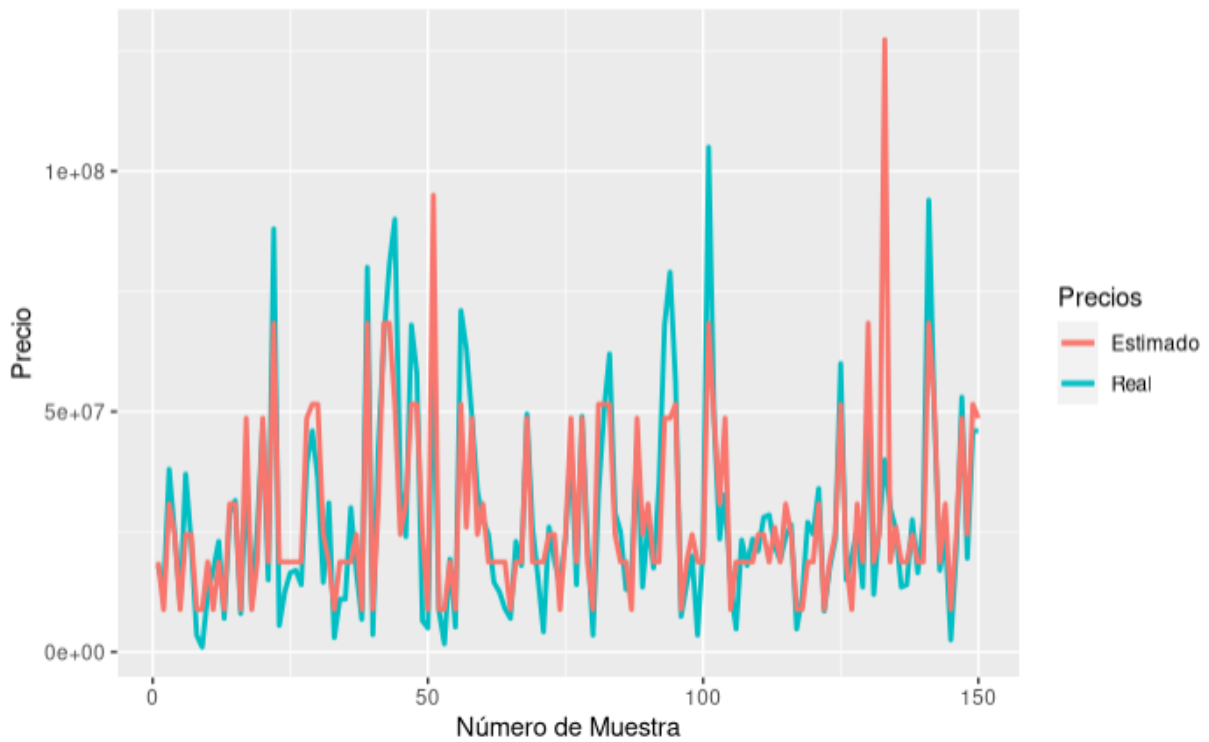


Figura 30: Precio Real Vs Precio Estimado - Árbol de Regresión Barranquilla

En la Figura 30 se observa que a pesar de no presentar valores atípicos en los precios, es decir, no se ubican por debajo de cero, los precios estimados no siguen muy bien a los precios reales.

- **Bosque Aleatorio - Modelo 5**

En Figura 31 se encuentran las especificaciones correspondientes al modelo de Bosque Aleatorio.

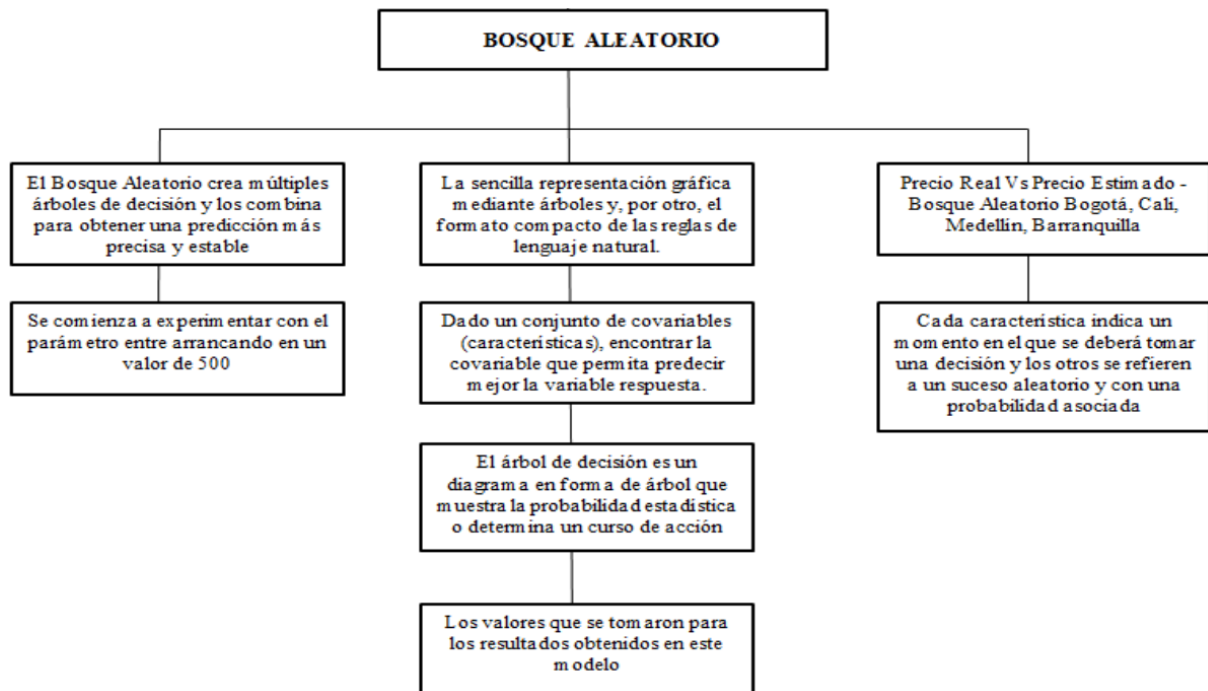


Figura 31: Especificaciones - Árbol de Regresión

Dado que los árboles de decisión tiende a padecer de *overfitting*, se empleó la técnica de Bosque Aleatorio para abordar el problema del proyecto, y para ello se recurrió a crear, no sólo un árbol sino, un grupo de árboles que puedan trabajar en conjunto y de esta forma mejorar la propuesta inicial.

Se comenzó a experimentar con el parámetro *ntree* arrancando en un valor de 500 y observando su error estimado. El resultado se muestra en la Figura 32.

Se puede observar que el error estimado se estabiliza en el rango entre 100 y 200 para todas las ciudades. Para verlo de una manera más clara, se tiene la Figura 33.

A partir de la Figura 33, se observa que el error estimado comienza a estabilizarse a partir de un valor de *ntree* entre 100 y 200. El cual fueron los valores que se tomaron para los resultados obtenidos en este modelo, los cuales se muestran en la Figura 34.

Una revisión de los valores anteriores nos señala que el modelo es capaz de explicar un máximo de 69% (correspondiente a la ciudad de Cali) de la variabilidad observada en el precio de los vehículos. No siendo este valor muy bueno comparado

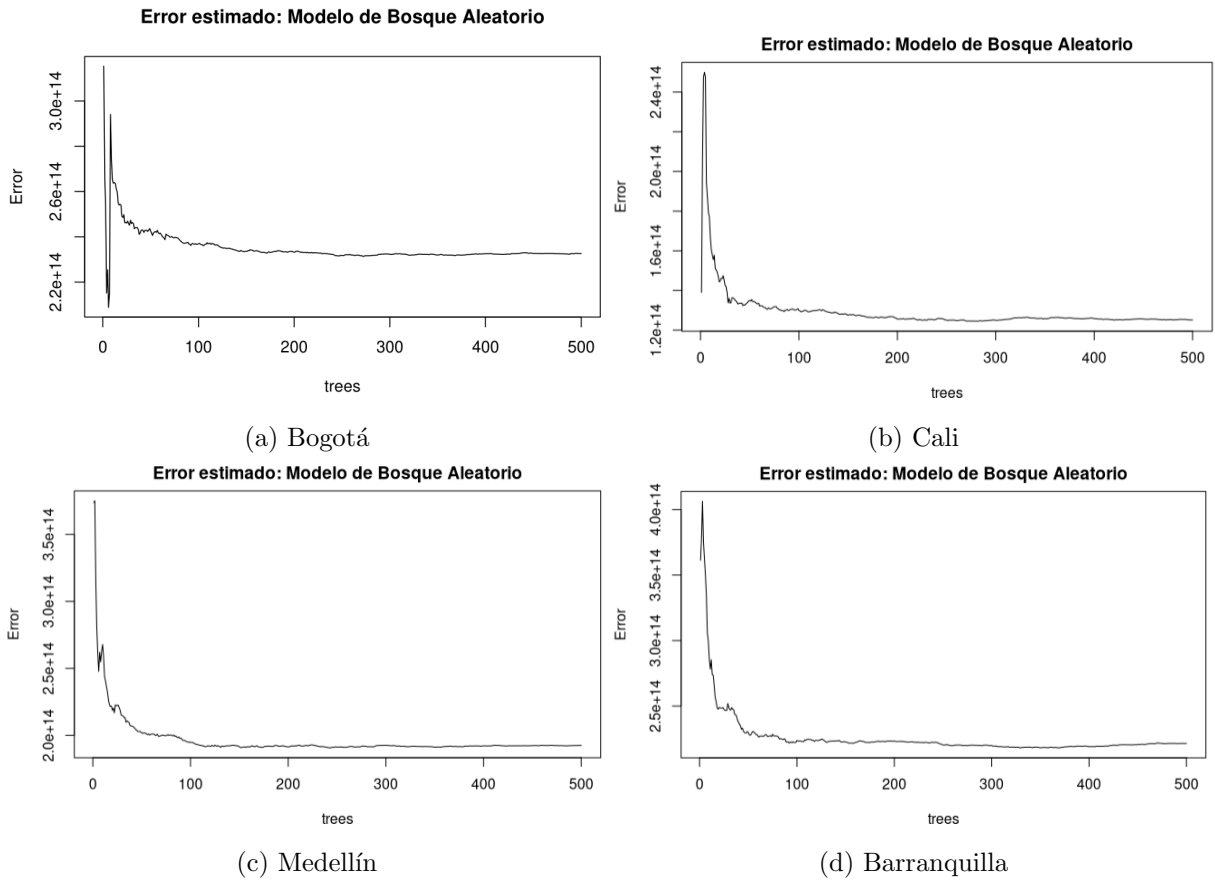


Figura 32: Error Estimado ntree = 500 - Bosque Aleatorio

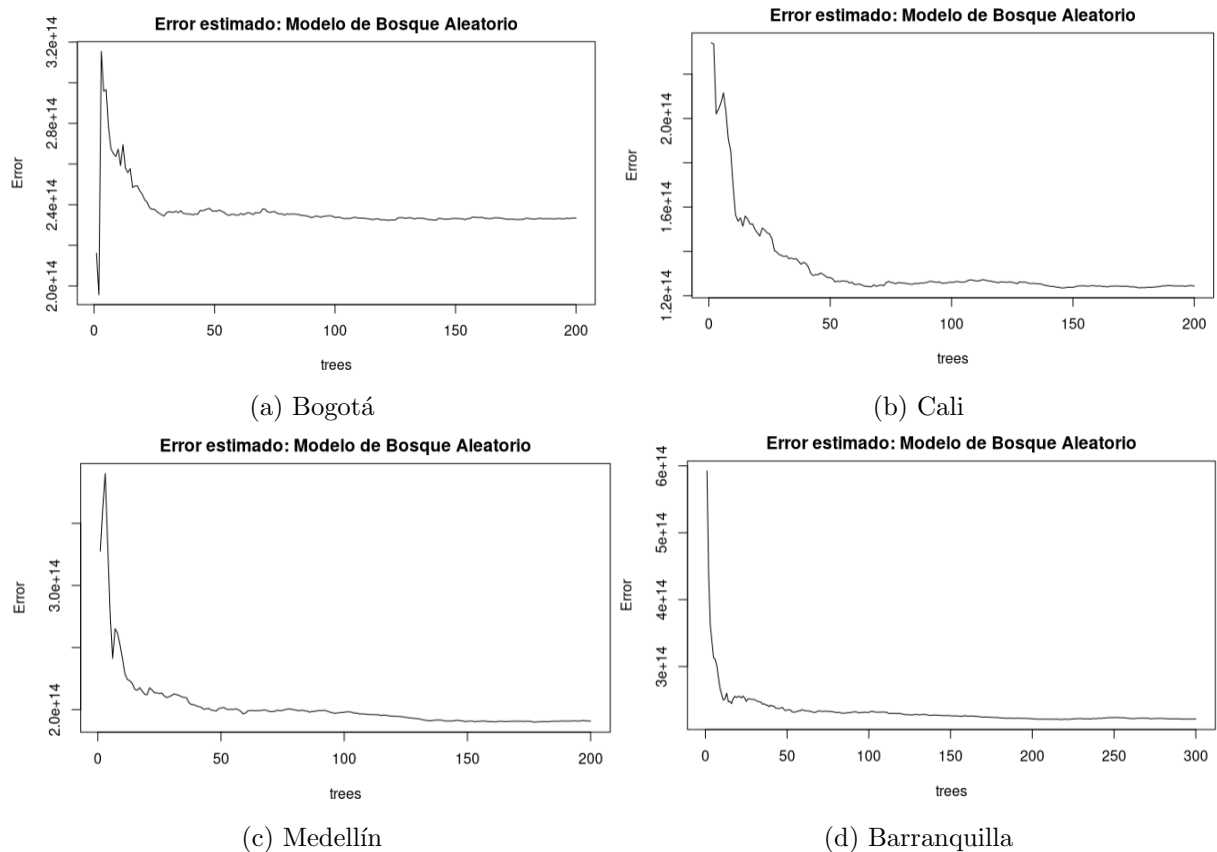


Figura 33: Error Estimado para ntree = 200 y 300 - Bosque Aleatorio

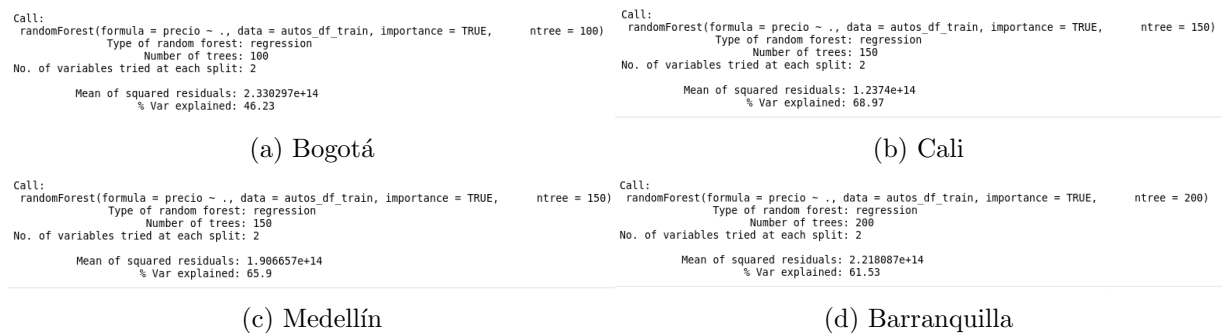


Figura 34: Resultado modelo - Bosque Aleatorio

con los anteriores modelos.

Las métricas obtenidas de este modelo se muestran a continuación:

	Bogotá	Cali	Medellín	Barranquilla
RMSE	20559209	15770591	26819024	13391418
Precisión	60.25	68.51	55.72	77.58

Tabla 28: Comportamiento - Regresión Polinomial de Grado 4

Podemos ver que el error residual del Bosque Aleatorio es mayor que el error presentado por el Árbol de Regresión.

A continuación, en las siguientes figuras se puede observar los comportamientos de los precios estimados del Bosque Aleatorio por cada ciudad.

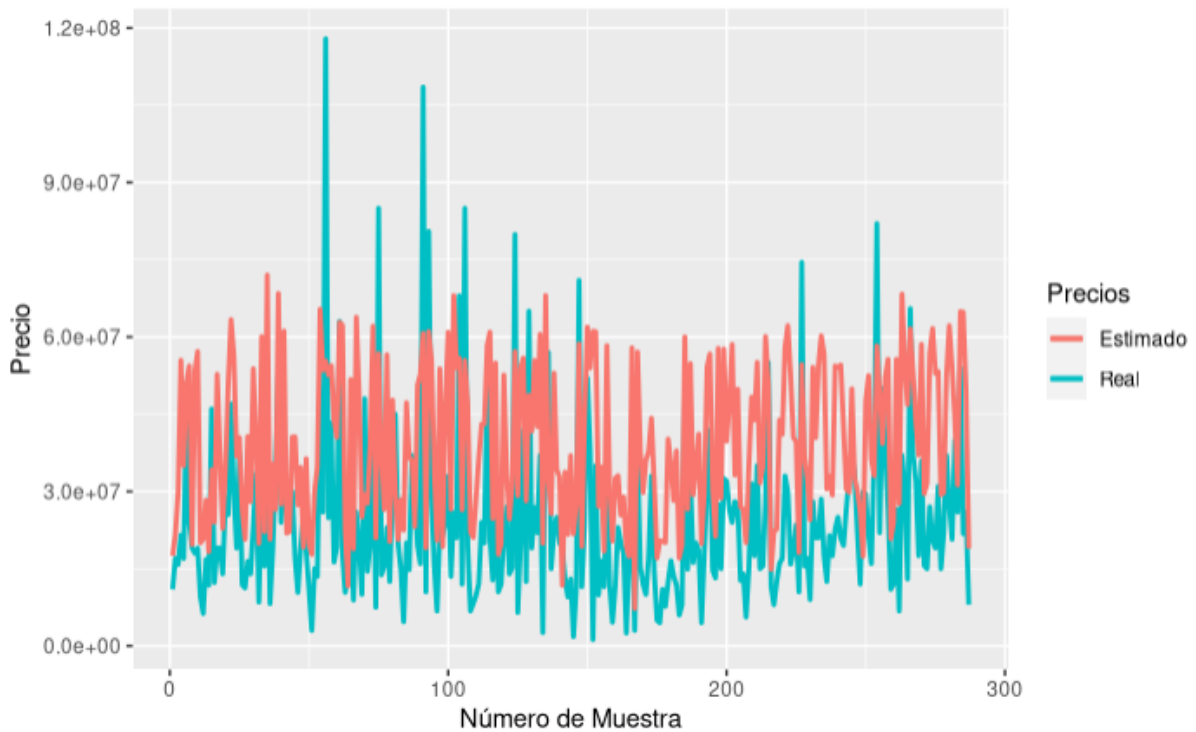


Figura 35: Precio Real Vs Precio Estimado - Bosque Aleatorio Bogotá

En la Figura 35 se puede visualizar que el Precio Real Vs Precio Estimado en el modelo Bosque Aleatorio en la ciudad de Bogotá siguen un comportamiento

disparejo con poca similitud.

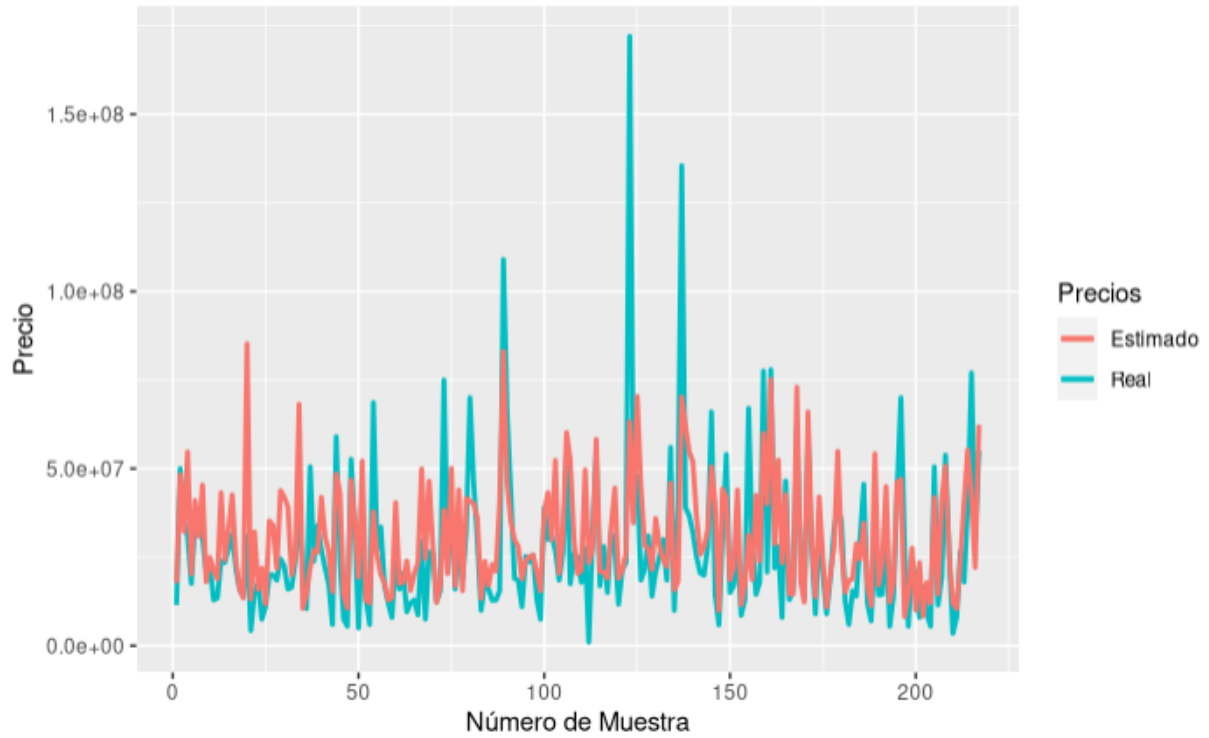


Figura 36: Precio Real Vs Precio Estimado - Bosque Aleatorio Cali

En la Figura 36 se puede visualizar que el Precio Real Vs Precio Estimado en el modelo Bosque Aleatorio en la ciudad de Cali siguen un comportamiento disparejo con poca similitud.

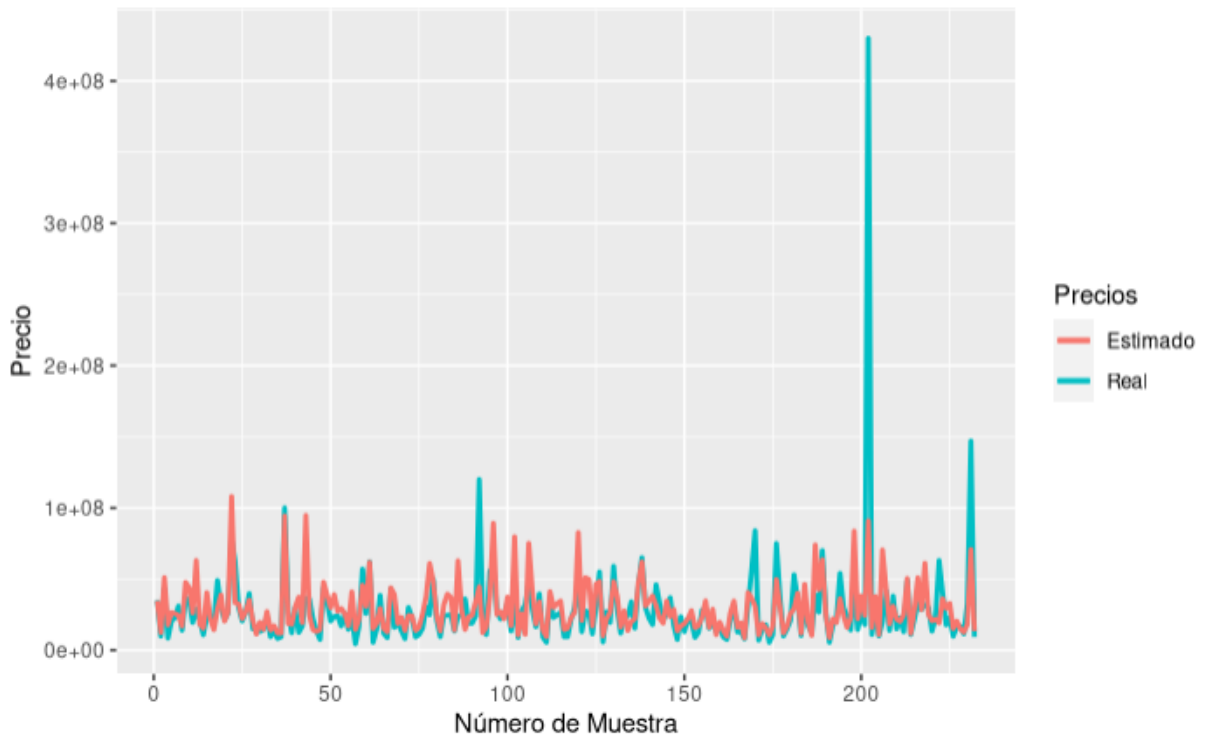


Figura 37: Precio Real Vs Precio Estimado - Bosque Aleatorio Medellín

En la Figura 37 se puede visualizar que el Precio Real Vs Precio Estimado en el modelo Bosque Aleatorio en la ciudad de Medellín siguen un comportamiento con mediana similitud.

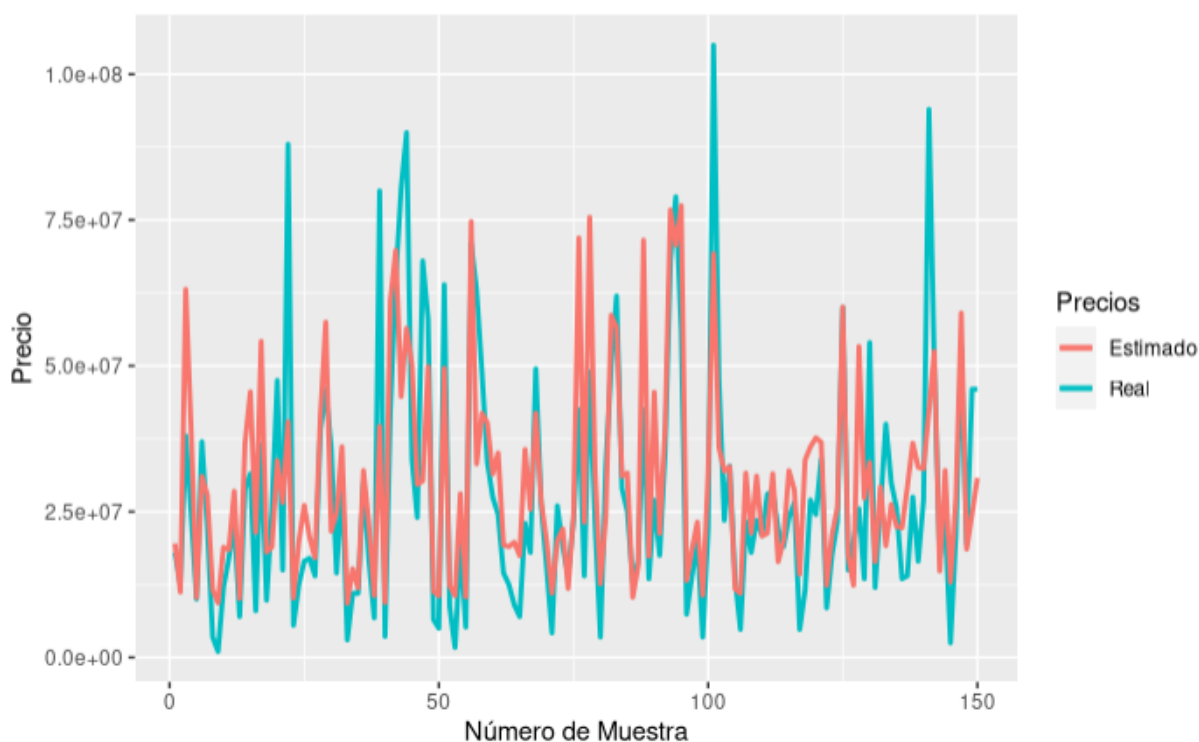


Figura 38: Precio Real Vs Precio Estimado - Bosque Aleatorio Barranquilla

En la Figura 38 se puede visualizar que el Precio Real Vs Precio Estimado en el modelo Bosque Aleatorio en la ciudad de Barranquilla siguen un comportamiento con poca similitud.

3.3.10 Comparación de Modelos

- **Precisión**

Como primer parámetro de comparación de los modelos propuestos en la anterior sección, se graficó el porcentaje de precisión. Para el cálculo de este parámetro, se obtuvo a partir de la correlación entre los valores reales y los estimados del atributo *Precio*.

El resultado se muestra en las siguientes figuras.

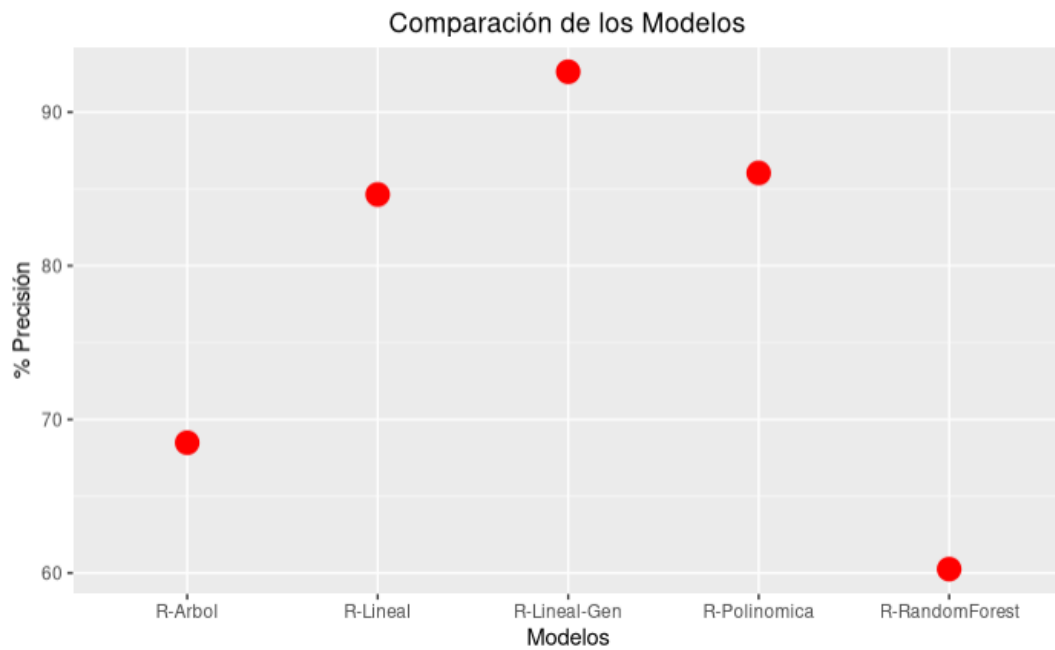


Figura 39: Comparación Modelos - Presición Bogotá

En la Figura 39 donde se compara los modelos para la ciudad de Bogotá se tiene que donde existe mayor porcentaje de similitud es en los modelos R-Lineal-Gen con un valor mayor a 90 %, seguidamente se tiene R-Polinómica, con un porcentaje aproximado de 86, R-Lineal con un porcentaje de 85 %, R-Árbol porcentaje de 68 % y R-RamdomForest 60 %

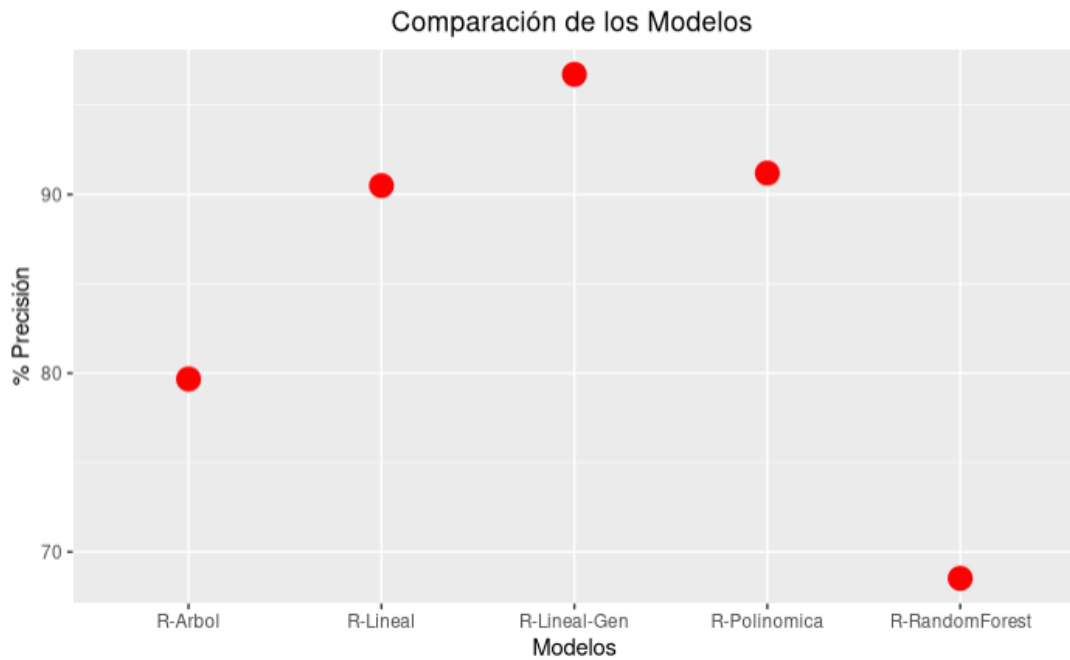


Figura 40: Comparación Modelos - Presición Cali

En la Figura 40 donde se compara los modelos para la ciudad de Cali se tiene que donde existe mayor porcentaje de similitud es en los modelos R-Lineal-Gen con un valor mayor a 95 %, seguidamente se tiene R-Polinómica, con un porcentaje aproximado de 92 R-Lineal con un porcentaje de 90 %, R-Árbol porcentaje de 80 % y R-RandomForest 68 %.

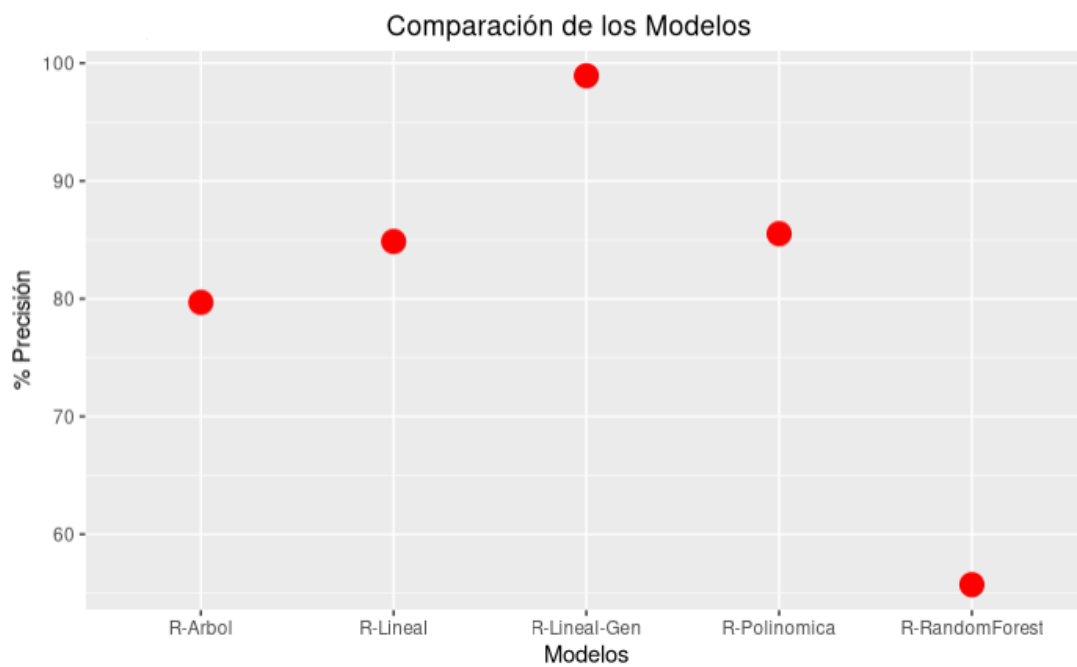


Figura 41: Comparación Modelos - Presición Medellín

En la Figura 41 donde se compara los modelos para la ciudad de Medellín se tiene que donde existe mayor porcentaje de similitud es en los modelos R-Lineal-Gen con un valor casi del 100 %, seguidamente se tiene R-Polinómica, con un porcentaje aproximado de 85 %, R-Lineal con un porcentaje de 84 %, R-Árbol porcentaje de 80 % y R-RandomForest 56 %.

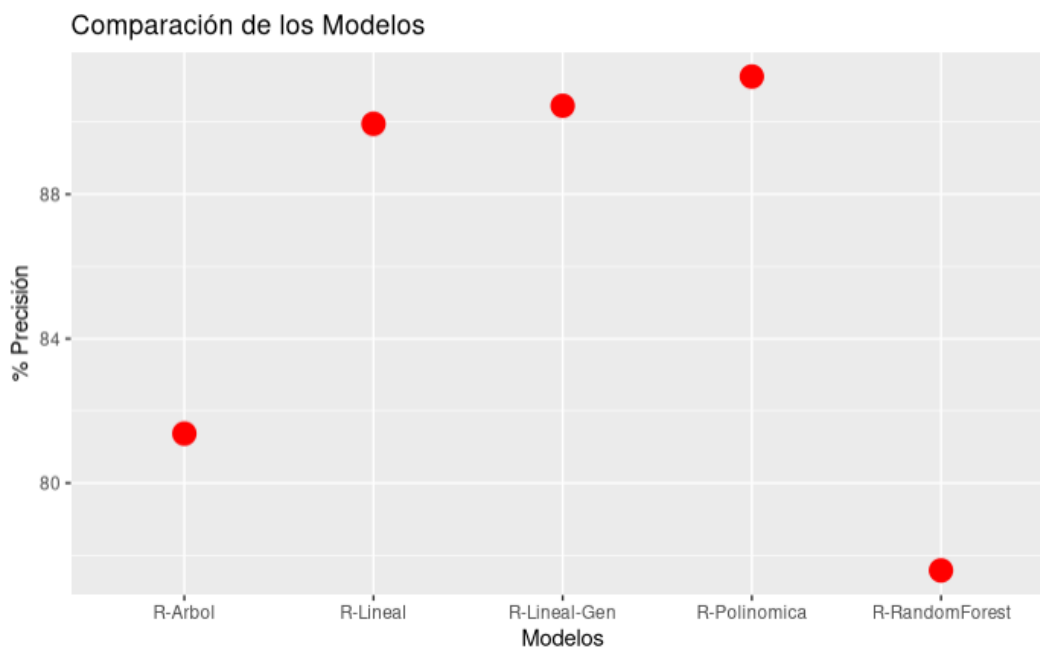


Figura 42: Comparación Modelos - Presición Barranquilla

En la Figura 42 donde se compara los modelos para la ciudad de Barranquilla se tiene que donde existe mayor porcentaje de similitud es en los modelos R-Polinómica, con un porcentaje aproximado de 91, seguidamente R-Lineal-Gen con un valor casi del 90 %, posteriormente se tiene, R-Lineal con un porcentaje de 89 %, R-Árbol porcentaje de 81 % y R-RamdomForest 78 %.

En 3 de las 4 ciudades, se observó que el modelo con mejores resultados en precisión fue el *Regresión Lineal Generalizada*, mientras que la *Regresión Lineal Polinómica* obtuvo un mejor resultado para una ciudad con un margen muy estrecho sobre el modelo considerado ganador en esta primera comparación.

• Error Estimado

Como segundo parámetro de comparación de los modelos propuestos, se graficó la fluctuación del *RMSE*. Este valor fue extraído directamente del análisis que se realizó en cada modelo.

El resultado se muestra en las siguientes figuras.

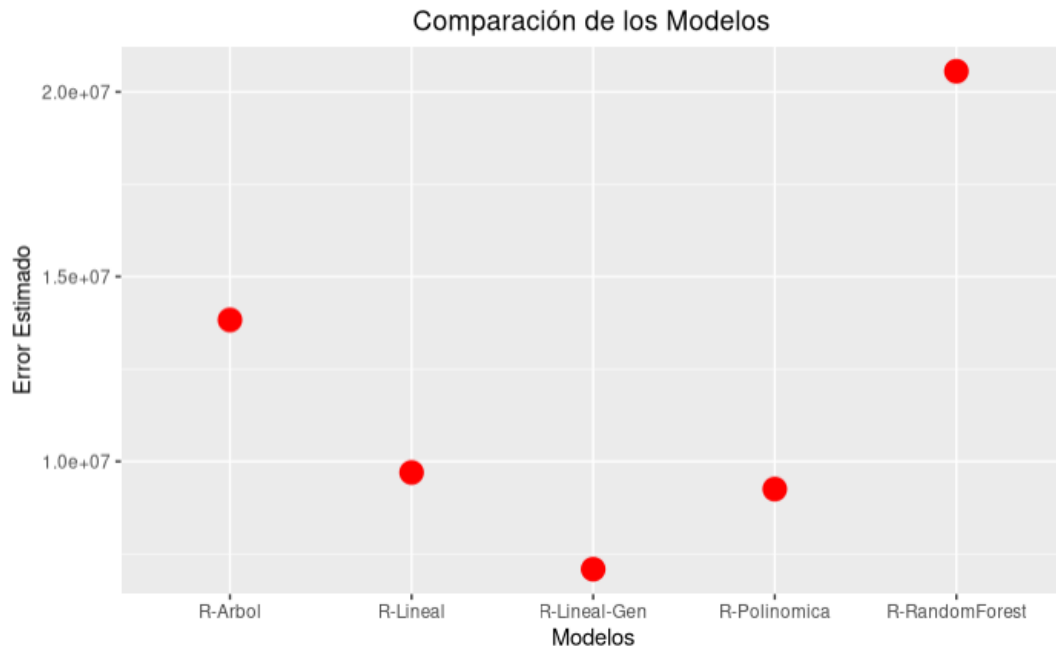


Figura 43: Comparación Modelos - Error Estimado Bogotá

En la Figura 43 donde se presenta el error estimado para la ciudad de Bogotá, el modelo con menor error es R-Lineal Generalizada, posteriormente el modelo R-Polinómica, sigue R-Lineal, R-Árbol y el modelo con mayor error es R-RandomForest.

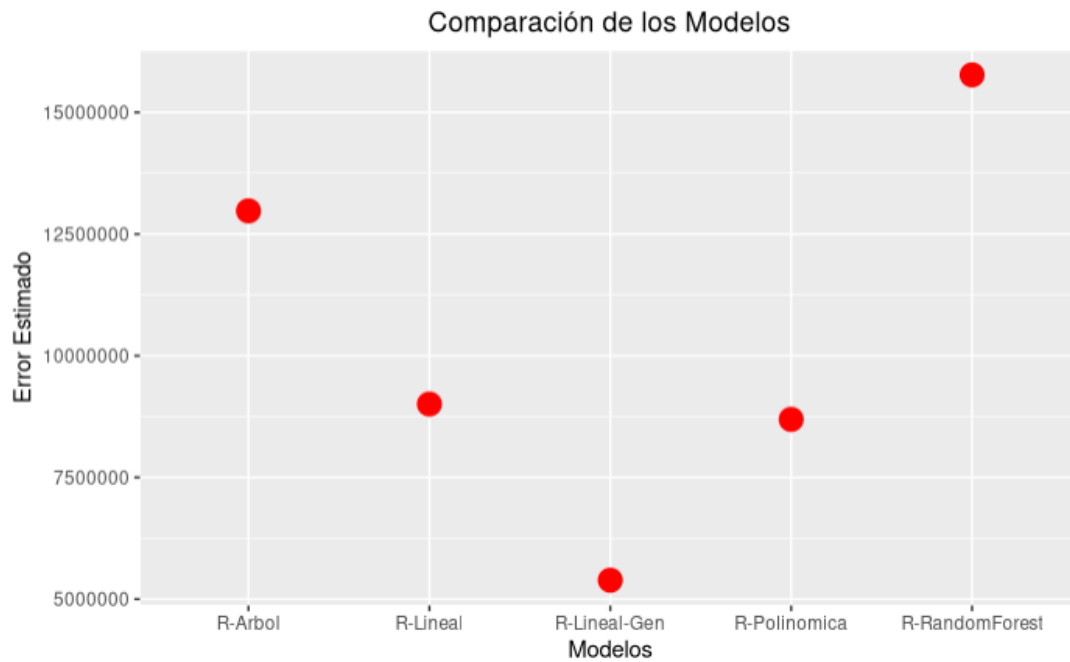


Figura 44: Comparación Modelos - Error Estimado Cali

En la Figura 44 donde se presenta el error estimado para la ciudad de Cali, el modelo con menor error es R-Lineal Generalizada, posteriormente el modelo R-Polinómica, sigue R-Lineal, R-Árbol y el modelo con mayor error es R-RandomForest.

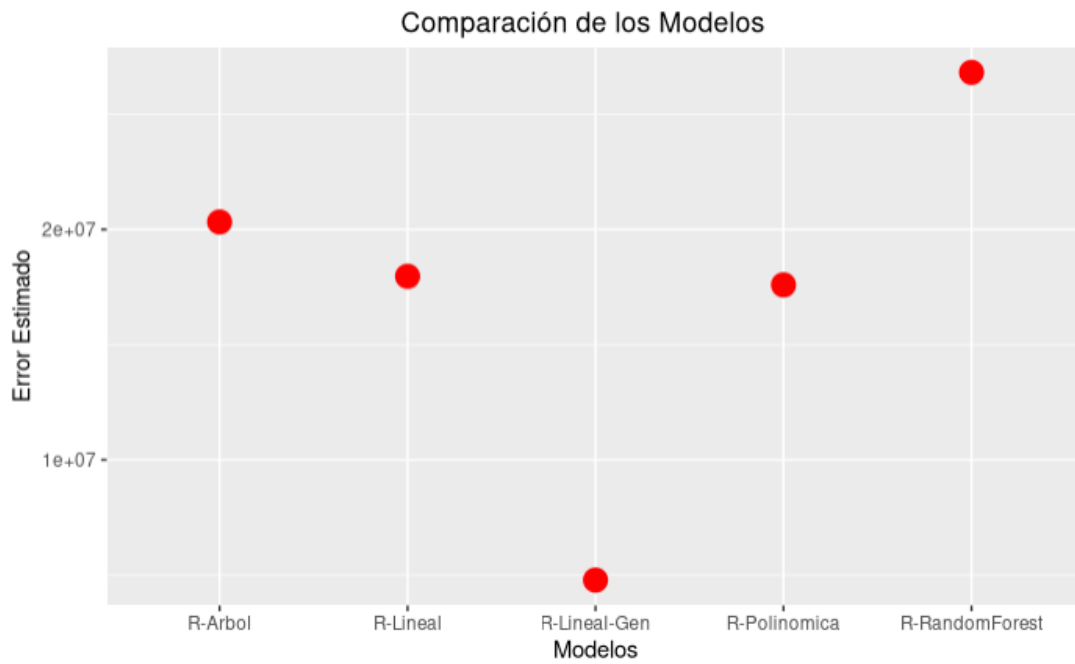


Figura 45: Comparación Modelos - Error Estimado Medellín

En la Figura 45 donde se presenta el error estimado para la ciudad de Medellín, el modelo con menor error es R-Lineal Generalizada, posteriormente el modelo R-Polinómica, sigue R-Lineal, R-Árbol y el modelo con mayor error es R-RandomForest.

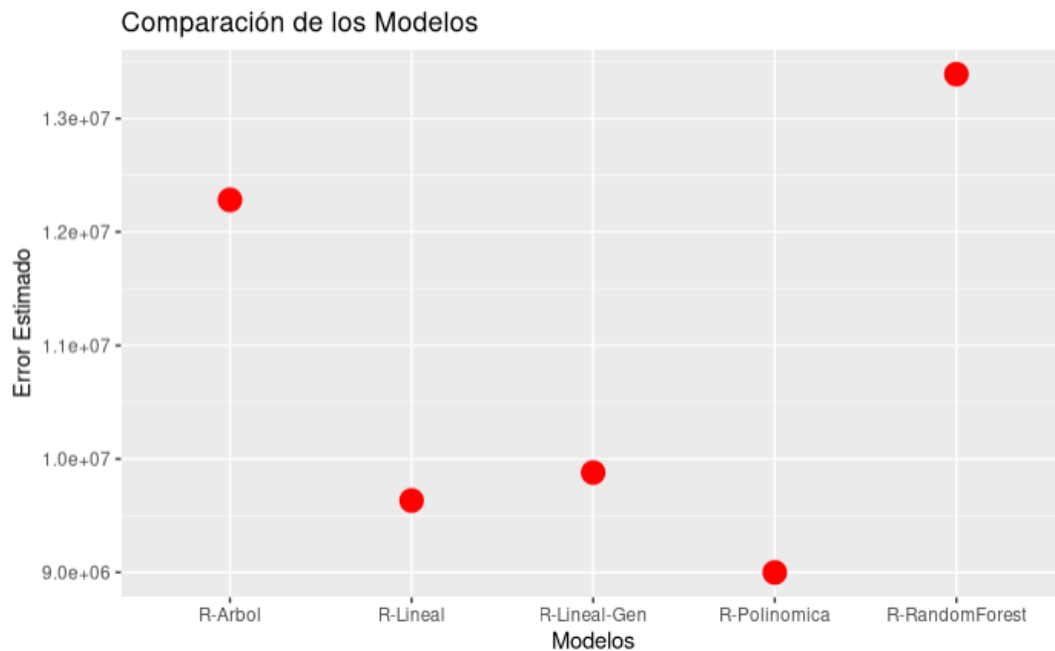


Figura 46: Comparación Modelos - Error Estimado Barranquilla

En la Figura 46 donde se presenta el error estimado para la ciudad de Barranquilla, el modelo con menor error es R-Polinómica, posteriormente el modelo R-Lineal, sigue R-Lineal Generalizada, R-Árbol y el modelo con mayor error es R-RandomForest.

Al igual que en caso anterior, en 3 de las 4 ciudades, el modelo con mejores resultados fue el *Regresión Lineal Generalizada*, mientras que la *Regresión Lineal Polinómica* obtuvo un mejor resultado en una de las ciudades.

Mediante cada uno de los estudios realizados y los análisis considerados en el programa, se tiene una visión general de cómo se encuentra el mercado de los automóviles y los precios reales de estos. Conociendo además las bases estadísticas fundamentales para establecer un modelo predictivo basado en un modelo de regresión múltiple y considerando los 5 modelos se aprecia en los resultados que el proyecto es factible ya que se cumplen cada uno de los objetivos específicos planteados. Mediante el análisis teórico y práctico se hace necesario poder contar con una herramienta que estime de manera aproximada el precio real de los vehículos de acuerdo a un número particular de características, las cuales tienen su aporte en el precio final.

En otras palabras la herramienta utilizada considera un equilibrio a la ventaja que tienen los vendedores, al proporcionar índices de calidad del automotor a adquirir, de esta manera se analiza que mediante la exploración de una serie de elementos y características claves

en cuanto a la información de los precios de los vehículos depende solamente de su marca. Esto deja a los vendedores tener la ventaja, injusta a veces, del verdadero precio de un automóvil sobre el comprador, puesto que conoce muy bien el estado. Mediante cada uno de los análisis realizados apreciados se torna fundamental cada uno de los estudios que logren el cumplimiento de los criterios, en donde mediante los factores que se denominan como importantes persiguen técnicas en obtener la lista de atributos más relevantes de un conjunto de datos iniciales. Mediante los análisis realizados se puede apreciar un proceso que se realiza con éxito, como resultado se obtienen una serie de atributos que aportarán la información que el conjunto original.

3.3.11 Análisis de Escenarios

- **Regresión Lineal**

Para realizar la formulación aplicado en el modelo de regresión lineal múltiple primero se debe seleccionar la variable dependiente Y y las variables independientes X , a continuación, se muestra la selección de las variables y la ecuación.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \dots \quad (20)$$

Dónde:

Y : Ventas

X_n : Atributos de nuestro proyecto.

Estadísticamente luego de realizar la identificación de las variables se procedió a realizar un análisis de causa - efecto entre la variable dependiente e independiente con el fin de determinar el grado de relación entre las variables. Con esta ecuación se puede determinar que la variable *kilometraje* tiene un peso significativo sobre las variables restantes. El p -value obtenido para el estadístico F es muy pequeño ($< 2,2 \cdot 10^{-16}$) para todas las ciudades. Lo que indica que al menos uno de los predictores introducidos en el modelo está relacionado con la variable respuesta precio.

- **Regresión Lineal Generalizada**

En el análisis de este escenario se consideró aplicar un modelo lineal generalizado con respuesta Poisson. Esto debido a la distribución que tiene la variable objetivo, la cual asemeja a este tipo de distribución.

De la misma manera se analizó el número de eventos que se producen en un conjunto de ensayos, en este caso considerados en las regiones Bogotá, Cali,

Medellín, Barranquilla, cuando la respuesta es un número de eventos que ocurren en un conjunto de ensayos, la variable dependiente contiene el número de eventos y puede seleccionar una variable adicional que contenga el número de ensayos. Otra posibilidad, si el número de ensayos es el mismo en todos los sujetos, consiste en especificar los ensayos mediante un valor fijo. El número de ensayos debe ser mayor o igual que el número de eventos para cada caso. Los eventos deben ser enteros no negativos y los ensayos deben ser enteros positivos.

- **Regresión Lineal Polinómica**

En este escenario se considera importante una serie de estrategias las cuales deben transformar el entorno, en este caso se debe tomar en cuenta la dependencia entre la variable de respuesta y la regresora que frecuentemente no es lineal. No obstante, para ello se consideró como fórmula importante el análisis de:

$$Y = A + BX + CX_2 + DX_3 + \dots \quad (21)$$

En este caso se consideró una variable explicativa continua X , pero esta se puede ajustar potencias mayores, como X^2 , X^3 y añadirlas al modelo, junto a X , para describir diversos tipos de curvatura en la relación Y y X . Las funciones polinomiales presentan gran flexibilidad de formas, incluso al añadir un solo término cuadrático, dependiendo de los signos de los términos lineales y cuadráticos. Sin embargo, si se aplica un modelo basado en regresión lineal polinomial, en donde se observa una leve mejora respecto del modelo simple. Analizando los resultados obtenidos, se observa que el modelo polinómico no ha mejorado su rendimiento con respecto al modelo 2 y se mantiene constante con respecto al modelo 1.

- **Árbol de Regresión**

En este análisis se identifican las variables factores y dependen mucho de la interpretación que se les pueda dar y lo estables que sean en el tiempo. Las relaciones causales o teóricas entre las variables económicas no siempre se cumplen y estas cambian en el tiempo. Resulta difícil cuantificar en que momento el mercado considera que la persistencia de uno o varios indicadores económicos vislumbra un escenario o el otro. Resulta difícil delimitar el rango en que las variables pasan a estar en otro escenario. Es difícil delimitar o discriminar entre un escenario y otro. A pesar de no presentar valores atípicos en los precios, es decir, no se ubican por debajo de cero, los precios estimados no siguen muy bien a los precios reales.

- **Bosque Aleatorio**

En este escenario se da a conocer el algoritmo para entrenar el proceso de desarrollo del modelo, con el fin de poder analizar cómo se desempeña y es

difícil construir un mal modelo con este algoritmo debido a su simplicidad, en este caso se considera como importante para el desarrollo del presente modelo analizando un determinado periodo de tiempo, además de eso, proporciona un buen indicador de la importancia que asigna a sus características. Una revisión de los valores anteriores señala que el modelo es capaz de explicar un máximo de 69 % (correspondiente a la ciudad de Cali) de la variabilidad observada en el precio de los vehículos. No siendo este valor muy bueno comparado con los anteriores modelos.

Mediante el análisis de cada uno de los escenarios propuestos, con el fin de llevar a cabo el cumplimiento del objetivo general, denominado *desarrollar un modelo predictivo sobre precios de vehículos teniendo en cuenta sus características y la ciudad donde se encuentra*, se realizó la etapa de Web Scraping, el cual es visto como un proceso dentro del Data Science que se utiliza para la extracción de datos de sitios web, normalmente en el lenguaje de programación Python y mediante machine learning, simulando cómo navegaría un ser humano en determinadas páginas web, en este caso se consideraron una serie de características de los vehículos. Luego del análisis realizado partir de estos registros, se analiza la cantidad de datos nulos de cada atributo en cada conjunto de datos para cada ciudad.

Considerando el análisis de los precios de automóviles como un ambiente competitivo, como lo es la producción y ventas de automóviles ensamblados en el país, es necesario conocer el negocio a profundidad, todo esto para no encontrar problemas, sino oportunidades de mejora, que hagan crecer y mantenerse, que es lo más importante en el mercado, en esta fase en la interpretación de un gran número de registros por ciudad, se puede llegar a generar más criterios para el filtrado del atributo Kilometraje.

Mediante los procesamientos de datos utilizados se aprecia una notable reducción de los registros nulos en atributos analizados y por ende una gran mejora. De la misma manera se pudo obtener mediante la supresión de registros nulos, y corriendo el script atributos resultantes, sin ningún orden en especial, modelo, línea, transmisión, marca, combustible, kilometraje, en donde mediante la utilización características correctas de los atributos obtenidos se genera que dentro de los valores agregados en donde actualmente dicho comprador exige calidad, garantía, respaldo, conveniencia, ecología, seguridad y diseño. Cabe resaltar que dichos factores de decisión varían dependiendo de ciertas características comprendidas por su segmentación geográfica, demográfica, psicológica, estilo de vida, uso del producto y beneficios del mismo. Para ciertos ciudadanos, el diseño puede ser el factor de decisión principal, pero para otra persona puede ser el precio (Martínez, M. 2011).

En relación al presente estudio Gamero, M (2017) en su estudio titulado factores de decisión para la compra de vehículos particulares en la ciudad de Cartagena. En este caso se analizó que los gustos y preferencias de las personas son factores decisivos que influyen en la compra de vehículos particulares, ya que los compradores se dejan llevar

por sus gustos personales y por ciertas preferencias en autos que se acomoden a su estilo de vida, personalidad, ocupación y demás factores personales. Los compradores tienden a adquirir cierto tipo de vehículo con un estilo, marca, modelo y color que les resulte atractivo. A continuación se analizarán estos gustos y preferencias que tienen las personas a la hora de comprar un vehículo. En otras palabras el actual estudio monográfico da con el cumplimiento al objetivo general, ya que se observó que los conjuntos de datos de las 4 ciudades tienen un comportamiento similar, esto se evidencio en los resultados obtenidos en los diferentes modelos propuestos. Exceptuando para la ciudad de Barranquilla, donde arrojó resultados cercanos a las demás ciudades.

En donde mediante una revisión de los valores anteriores se señala que el modelo es capaz de explicar un máximo de 69 % (correspondiente a la ciudad de Cali) de la variabilidad observada en el precio de los vehículos. No siendo este valor muy bueno comparado con los anteriores modelos. En relación con el presente resultado está la investigación realizada por la multinacional Nielsen, en donde se reveló que las personas que compraban vehículos lo hacían porque les encantaba manejar, otra gran porción se enfocó en que ese acto comercial daba estatus al final, se concluyó que el 44 % lo hacía por movilidad. En otras palabras, la firma Raddar hizo esta misma investigación en Colombia, en donde se pudo concretar que “un 52 % de los colombianos compra carro por necesidad, el 41 % por gusto, 5 % por antojo y un 2 % por temas de mejora en la calidad de vida”. En otras perspectivas los valores agregados que los concesionarios entregaban a los compradores se encontraba el entregar el vehículo full equipo (radio, accesorios, soat) no se realizaba un estudio tan a fondo; hoy en día, este aspecto forma parte de lo que conoce como vehículo básico por tanto el comprador exige o espera otros pluses.

4 Conclusiones

- El contar con datos limpios y organizados, esto gracias al diseño e implementación de técnicas propias desarrolladas en el proyecto, se facilita el proceso de selección de los atributos más importantes del conjunto de datos descargados y el análisis de los modelos seleccionados. Por medio de la comparación directa del análisis de los datos con el objetivo general, se tiene diversidad de técnicas que cumplen con las características planteadas inicialmente, lo inicial que se realiza es la etapa de selección, una vez identificado el conocimiento relevante y prioritario, las cuales se encuentran definidas las metas del proceso en cuanto a los datos analizados de cada una de las ciudades, desde el punto de vista del usuario final, se crea un conjunto de datos objetivo, seleccionando todo el conjunto de datos mediante una muestra representativa el cual cumple con el modelo general que se quiere trabajar, conociendo los precios correctos de los vehículos en el mercado de cada una de las ciudades estudiadas.
- En el primer objetivo específico se pudo determinar y extraer los datos más relevantes de los vehículos mediante la aplicación de web scraping en páginas dedicadas a la venta de vehículos. Con esta página, se representa una técnica para extraer información de sitios web, en donde se pudo obtener información clave de la venta de automóviles y los precios correctos que se manejan en cada una de las ciudades consideradas como estudio para la muestra. Usualmente, estos programas simulan la navegación ya sea utilizando el protocolo HTTP manualmente, o incrustando un navegador en una aplicación.
- En el segundo objetivo específico se identificó mediante criterios exploratorios en variedad de datos diferentes variables, las cuales son interpretadas como aspectos que se relacionan con los precios. En este caso se analizaron los atributos que son significantes en la presentación de los resultados. Consecuentemente, se estableció diferentes atributos que aportan información relevante en la compra de un vehículo.
- Como tercer objetivo específico se quiere lograr entrenar y comparar diferentes modelos paramétricos y de Machine Learning con base en la predicción de los precios de los vehículos, se tomó en cuenta nuevos análisis sistemáticos e interpretativos, como aprendizaje supervisado, en donde se visualiza los criterios que son vistos para el análisis de aprender con un patrón a partir de un conjunto de datos en cuestión de entrenamiento, y que permitirá realizar predicciones de conjuntos de datos no observados previamente. De esta forma, se puede indicar que el concepto puede analizar una serie de datos que son usados para la interpretación de los datos mediante el modelo interpretativo.
- Para el cuarto objetivo específico, en relación con el objetivo general se tiene que dentro del desarrollo propuesto se observó que los conjuntos de datos de las 4

ciudades tienen un comportamiento similar, esto se evidencio en los resultados obtenidos en los diferentes modelos propuestos. Exceptuando para la ciudad de Barranquilla, donde arrojó resultados cercanos a las demás ciudades.

- Se identifica que la *Regresión Lineal Generalizada* es el enfoque con los mejores resultados en precisión y error estimado en 3 de las 4 ciudades, solo denotando un comportamiento inferior en uno de los casos debido a la poca cantidad de registros en este último conjunto de datos. Concluyendo así que el modelo de *Regresión Lineal Generalizada* es el ideal para este proyecto ya que presenta un mejor comportamiento.

Referencias

- [1] <http://www.andemos.org/index.php/2020/01/02/diciembre-2019-sector-cierra-el-ano-con-un-crecimiento-del-2-7/>
- [2] <https://www.motor.com.co/precios>
- [3] Enis Gegic, Becir Isakovic, Dino Keco, Zerina Masetic, Jasmin Kevric. Car Price Prediction using Machine Learning Techniques. TEM Journal. Volume 8, Issue 1, Pages 113-118, ISSN 2217-8309, DOI: 10.18421/TEM81-16, February 2019.
- [4] Ramírez Muñoz de Toro, Gonzalo R., Uriarte, Juan I., Delbianco, Fernando, Larrosa, Juan M.C. Un modelo hedónico de precios en línea de automóviles usados en Argentina. REVISTA DE MÉTODOS CUANTITATIVOS PARA LA ECONOMÍA Y LA EMPRESA (24). Páginas 25–53. Diciembre de 2017. ISSN: 1886-516X. D.L: SE-2927-06.
- [5] Sameerchand Pudaruth. Predicting the Price of Used Cars using Machine Learning Techniques. International Journal of Information & Computation Technology. ISSN 0974-2239 Volume 4, Number 7 (2014), pp. 753-764. © International Research Publications House.
- [6] Stefan Lessmann, Stefan Voss. Car Resale Price Forecasting: The Impact of Regression Method, Private Information, and Heterogeneity on Forecast Accuracy. ResearchGate, International Journal of Forecasting, October 2017.
- [7] https://www.olx.com.co/carros_c378
- [8] https://es.wikipedia.org/wiki/Regresión_logística
- [9] Luis Cayuela. Modelos lineales generalizados (GLM). EcoLab, Centro Andaluz de Medio Ambiente, Universidad de Granada – Junta de Andalucía, Avenida del Mediterráneo s/n, E-18006, Granada.
- [10] Ma Asunción Martínez, Javier Morales. Modelos lineales generalizados. Dpto Estadística y Matemática Aplicada, Centro de Investigación Operativa. Universidad Miguel Hernández
- [11] Bradley Efron, Trevor Hastie. Computer Age Statistical Inference. Stanford University, 2017.
- [12] https://es.wikipedia.org/wiki/Validación_cruzada
- [13] Antonio Serrano, Emilio Soria, José Martín. Redes Neuronales Artificiales. Escola Tècnica Superior d' Enginyeria, Departament d' Enginyeria Electrònica. Curso 2009-2010

- [14] Mariette Awad, Rahul Khanna. Support Vector Machines for Classification. Chapter 3. DOI: 10.1007/978-1-4302-5990-9_3
- [15] <https://mlampros.github.io/2016/02/14/feature-selection/>
- [16] <https://carros.mercadolibre.com.co/>
- [17] <https://www.carroya.com>
- [18] <https://loscoches.com/>
- [19] <https://carros.tucarro.com.co>
- [20] <https://autoexpo.com.co/concesionario-carros-usados-bogota/>
- [21] <https://www.autocosmos.com.co/auto>