



UNIVERSIDAD EAN

3 de junio del 2026.

Implementación de Modelos Predictivos en Automóviles Nescar S.A.S.

ESTUDIANTES

Luis Ernesto Caro Barrera.

Juan José Jiménez Pulido.

Alejandro Rodríguez Ulloa.

CURSO

Proyecto de Grado

PROFESOR

Lina María Chacón Rivera.

Tabla de contenido

Resumen.....	2
Abstract	3
Introducción.....	4
Marco Teórico.....	5
Marco Conceptual.....	5
Marco Contextual	9
Marco Legal.....	11
Estado del arte	13
Metodología de revisión sistemática (PRISMA)	13
Estrategias de búsqueda.....	14
Criterios de inclusión	14
Criterios de exclusión	15
Proceso de selección.....	15
Resultados de la revisión sistemática.....	16
Modelos predictivos	16
Recolección de datos	17
Variables predictoras.....	18

Desafíos y brechas	19
PRISMA	25
Metodología	26
Resultados	29
Resultados del desarrollo del sistema	29
Resultados de la recolección y el procesamiento de datos	32
Resultados del análisis del mercado	33
Discusión	35
Plan de Divulgación.....	37
Posibles mejoras	39
Conclusiones	40
Referencias	42

Resumen

El presente proyecto de investigación tiene como objetivo desarrollar e implementar un modelo predictivo basado en técnicas de aprendizaje automático para mejorar la estimación del valor de vehículos usados en la empresa Automóviles Nescar S.A.S. Actualmente, la valoración de estos vehículos suele realizarse mediante criterios empíricos y referencias generales del mercado, lo que puede generar inconsistencias y limitar la precisión en los procesos de negociación. Ante esta problemática, se propone el uso de herramientas de analítica de datos e inteligencia artificial que permitan realizar estimaciones más objetivas y basadas en información real del mercado.

La investigación adopta un enfoque cuantitativo, mediante la recolección y análisis de datos obtenidos a través de técnicas de web scraping y registros históricos del sector automotor. Posteriormente, la información es procesada y analizada mediante modelos de aprendizaje automático como regresión lineal y algoritmos de machine learning, los cuales permiten identificar patrones entre variables como marca, modelo, año de fabricación, kilometraje y estado del vehículo. Los resultados esperados buscan contribuir a la construcción de un sistema automatizado de valoración vehicular que mejore la precisión en la fijación de precios, optimice la toma de decisiones comerciales y fortalezca la competitividad de Automóviles Nescar S.A.S. en el mercado automotor colombiano.

Palabras clave: machine learning, valoración de vehículos, vehículos usados, modelos predictivos, web scraping, análisis de datos.

Abstract

This research project aims to develop and implement a predictive model based on machine learning techniques to improve the estimation of used vehicle values at Automóviles Nescar S.A.S. Currently, the valuation of these vehicles is usually carried out through empirical criteria and general market references, which can generate inconsistencies and limit precision in negotiation processes. To address this problem, the use of data analytics and artificial intelligence tools is proposed to enable more objective estimations based on real market information.

The research adopts a quantitative approach, through the collection and analysis of data obtained via web scraping techniques and historical records from the automotive sector. Subsequently, the information is processed and analyzed using machine learning models such as linear regression and machine learning algorithms, which allow the identification of patterns among variables such as brand, model, year of manufacture, mileage and vehicle condition. The expected results seek to contribute to the construction of an automated vehicle valuation system that improves pricing accuracy, optimizes commercial decision-making and strengthens the competitiveness of Automóviles Nescar S.A.S. in the Colombian automotive market.

Keywords: machine learning, vehicle valuation, used vehicles, predictive models, web scraping, data analytics.

Introducción

El mercado de vehículos usados en Colombia presenta una dinámica compleja, donde el precio de los automóviles varía considerablemente debido a factores como sus características técnicas, su historial de uso y las condiciones del mercado. En organizaciones como Automóviles Nescar S.A.S., la determinación del valor comercial suele apoyarse en criterios empíricos y referencias generales, lo que puede generar inconsistencias y limitar la precisión en los procesos de negociación. Esta situación plantea la necesidad de incorporar herramientas analíticas que permitan una valoración más objetiva y basada en datos.

En el ámbito académico y empresarial, diferentes investigaciones han evidenciado que la aplicación de técnicas de inteligencia artificial y aprendizaje automático contribuye a mejorar la estimación de precios en mercados complejos. Modelos como Random Forest, XGBoost y redes neuronales han demostrado ser efectivos al analizar grandes volúmenes de información y detectar patrones relevantes a partir de datos obtenidos mediante procesos automatizados como el web scraping. Estos enfoques permiten superar las limitaciones de los métodos tradicionales, proporcionando resultados más consistentes y confiables.

Bajo este panorama, el presente proyecto tiene como finalidad diseñar e implementar un modelo predictivo orientado a estimar el valor de los vehículos usados en Automóviles Nescar S.A.S., utilizando datos históricos y variables representativas del mercado. Para ello, se adopta un enfoque cuantitativo que integra técnicas de análisis de datos y algoritmos de aprendizaje automático. En este sentido, se plantea la siguiente pregunta de investigación:

¿De qué manera un modelo predictivo basado en machine learning puede mejorar la precisión en la valoración de vehículos usados en Automóviles Nescar S.A.S.?

Marco Teórico

Marco Conceptual

En este trabajo se exploran varios conceptos clave que son esenciales para desarrollar e implementar un modelo de inteligencia artificial. Este modelo tiene como objetivo automatizar y mejorar el proceso de valoración de vehículos usados. A continuación, se presentan los conceptos y enfoques más importantes que se abordarán:

- **Valoración de Vehículos:** La valoración de vehículos consiste en estimar el valor monetario de un automóvil, tomando en cuenta sus características físicas, mecánicas, estéticas y de mercado. Tradicionalmente, este proceso ha dependido de la experiencia del vendedor o de referencias generales del mercado, lo cual puede generar resultados subjetivos y poco prácticos o verídicos. Sin embargo, el uso de modelos automatizados basados en datos permite calcular precios más exactos y alineados con las condiciones reales del mercado (Repositorios Uniremington, 2023).
- **IA (Inteligencia Artificial):** En este entorno se aborda el campo de la informática que busca desarrollar sistemas capaces de imitar procesos cognitivos humanos, como el razonamiento, el aprendizaje y la toma de decisiones. En el ámbito empresarial, la IA permite automatizar tareas complejas y mejorar la eficiencia de los procesos, entre ellos la tasación de vehículos (Infobae, 2024).

- **Machine Learning:** El machine learning o aprendizaje automático es una rama que se desprende de la IA la cual permite aprender a partir de datos históricos a mejorar su rendimiento con el tiempo sin ser programados explícitamente. Este enfoque es clave para la predicción de precios en el caso de negocio planteado en este proyecto, ya que los modelos pueden analizar variables como año, kilometraje, marca y estado del vehículo para estimar su valor de forma acertada (Arxiv, 2017).
- **Big Data:** Big Data se refiere a la gestión y análisis de grandes volúmenes de información que, debido a su tamaño y complejidad, no pueden ser procesados con herramientas convencionales. En el ámbito de la valoración de vehículos, el análisis de Big Data es fundamental para identificar patrones en transacciones pasadas y establecer tendencias del mercado con mayor precisión.
- **Minería de Datos:** La minería de datos es un proceso para explorar grandes conjuntos de datos mediante los cuales permite descubrir patrones significativos. Cuando se aplica al mercado automotriz, esta técnica permite identificar variables las cuales afectan en la toma de decisiones del cliente y el costo final del vehículo, lo que facilita la creación de modelos predictivos más sólidos.
- **Regresión:** La regresión es una técnica estadística que ayuda a predecir valores numéricos basándonos en una o más variables independientes. En este estudio, se utiliza para estimar el precio de los vehículos, teniendo en cuenta factores como la marca, el modelo, el kilometraje y el año de fabricación.
- **Árboles de Decisión:** Los árboles de decisión son modelos predictivos que dividen los datos en diferentes ramas según reglas específicas, lo que facilita tanto la clasificación

como la regresión. Su habilidad para interpretar datos no lineales y su fácil visualización los convierte en herramientas especialmente útiles para estimar precios en el sector automotriz (Jin, 2021; Yilmaz & Selvi, 2023).

- **Redes Neuronales Artificiales:** Las redes neuronales artificiales se inspiran en el funcionamiento del cerebro humano y son capaces de reconocer patrones complejos a través de capas de procesamiento. En la tasación de vehículos, se pueden utilizar para procesar tanto datos estructurados como imágenes del automóvil, lo que mejora la precisión del modelo (Arxiv, 2018).
- **Web scraping:** El web scraping es una técnica que permite extraer información de páginas web de manera automatizada. Esta herramienta es muy útil para recopilar datos de portales de compra y venta de vehículos, creando bases de datos actualizadas con información real del mercado.
- **Tendencia del Mercado Automotor:** La tendencia del mercado automotor se refiere a los cambios o variables que se presentan en precios, demanda y preferencias de los consumidores frente a los vehículos por los cuales desean ofertar. Estos factores pueden variar según el contexto económico, la ubicación geográfica del cliente y las condiciones del mercado actual. Analizar estas tendencias ayuda a ajustar las predicciones de los modelos a la realidad del momento de mercado (Infobae, 2024).
- **Random Forest:** Es una técnica de machine learning que combina varios árboles de decisión para mejorar la precisión de las predicciones. Se utiliza ampliamente en la predicción de precios de vehículos gracias a su capacidad para manejar grandes volúmenes de datos y reducir el sobreajuste (Arxiv, 2017).

- **Boosting:** El boosting es una técnica de ensamblaje de modelos que se encarga de corregir los errores de predicción de modelos anteriores, lo que mejora su rendimiento de manera gradual. Cuando se utiliza junto con árboles de decisión, ha demostrado ser muy efectiva en la estimación de precios y en otras tareas de predicción.
- **Computer Vision:** La computer vision, o visión por computadora, es una subdisciplina de la IA que permite a las máquinas interpretar y analizar contenido visual. En el sector automotor, esta tecnología puede utilizarse para evaluar automáticamente el estado estético de un vehículo, detectando daños o imperfecciones que pueden afectar su valor (Infobae, 2024).

Marco Contextual

El sector automotriz en Colombia ha estado creciendo de manera constante, gracias a una combinación de factores económicos, tecnológicos y sociales que han cambiado la manera en que las empresas venden vehículos y manejan la información de sus clientes. En este escenario, las organizaciones del sector se enfrentan al reto de adaptarse a la transformación digital, incorporando herramientas tecnológicas que les ayuden a optimizar sus procesos comerciales y a tomar decisiones estratégicas más informadas.

Según la Asociación Nacional de Movilidad Sostenible (ANDEMOS, 2023), el mercado automotor colombiano ha experimentado cambios significativos en el comportamiento del consumidor, evidenciando una mayor demanda de servicios personalizados, análisis de financiamiento y procesos de compra digitalizados. Esto ha generado la necesidad de que las empresas implementen sistemas de analítica de datos que permitan comprender las tendencias del mercado y anticipar los requerimientos de los clientes.

Automóviles Nescar S.A.S. es una empresa dedicada a la comercialización de vehículos, la cual desarrolla sus actividades dentro de un entorno competitivo caracterizado por la constante evolución tecnológica y la creciente demanda de información estratégica. La organización maneja grandes volúmenes de datos relacionados con ventas, clientes, inventarios y comportamiento del mercado, información que, si es correctamente analizada, puede convertirse en una ventaja competitiva frente a los demás concesionarios.

Actualmente, muchas empresas del sector automotriz han comenzado a implementar soluciones basadas en inteligencia de negocios y análisis predictivo para mejorar la eficiencia operativa, reducir riesgos comerciales y optimizar la gestión de inventarios. De acuerdo con Davenport y Harris (2017), las organizaciones que utilizan análisis avanzado de datos logran mejorar significativamente sus procesos de toma de decisiones, generando mayores niveles de competitividad en el mercado.

En el contexto organizacional de Automóviles Nescar S.A.S; se pueden identificar oportunidades para mejorar el uso estratégico de la información, ya que gran parte de los datos generados por las operaciones comerciales no se están aprovechando con herramientas analíticas avanzadas. Esta situación limita la capacidad de la empresa para anticipar comportamientos del mercado, proyectar ventas y optimizar la gestión de clientes.

Desde el contexto tecnológico, el crecimiento del Big Data, la analítica avanzada y el aprendizaje automático ha permitido que las empresas desarrollen modelos predictivos que facilitan la proyección de escenarios futuros y el diseño de estrategias comerciales más eficientes. Según Mayer-Schönberger y Cukier (2013), el uso de grandes volúmenes de datos

permite a las organizaciones transformar la información en conocimiento estratégico para la toma de decisiones.

Por otra parte, el contexto económico colombiano también influye en el comportamiento del sector automotriz. Factores como la inflación, las tasas de interés, el acceso al crédito y las condiciones del mercado influyen directamente en la demanda de vehículos. El Departamento Administrativo Nacional de Estadística (DANE, 2023) señala que las variaciones económicas afectan el comportamiento del consumo, generando cambios en las dinámicas comerciales de las empresas automotrices.

Asimismo, el avance de la transformación digital ha impulsado la adopción de sistemas de gestión empresarial, plataformas de análisis de datos y herramientas tecnológicas que facilitan la administración de la información. La implementación de modelos predictivos dentro de Automóviles Nescar S.A.S. permitirá fortalecer sus procesos comerciales, mejorar la planificación estratégica y optimizar la toma de decisiones basadas en datos.

En este sentido, el desarrollo del presente proyecto se enmarca dentro de la necesidad de fortalecer la capacidad analítica de la empresa, aprovechando los datos históricos para generar proyecciones que permitan mejorar la eficiencia operativa, incrementar la competitividad y responder a las exigencias del mercado actual.

Marco Legal

El desarrollo de soluciones tecnológicas basadas en análisis de datos y modelos predictivos dentro de organizaciones empresariales debe regirse por la normatividad vigente en

Colombia, especialmente en materia de protección de datos, comercio electrónico, seguridad de la información y uso responsable de tecnologías digitales.

En primer lugar, la Ley 1581 de 2012 establece el régimen general de protección de datos personales en Colombia. Esta ley regula el tratamiento de información personal, garantizando el derecho constitucional al habeas data, el cual permite a las personas conocer, actualizar y rectificar la información que sobre ellas repose en bases de datos. En el caso de Automóviles Nescar S.A.S., esta normativa es fundamental, ya que el análisis predictivo implica el almacenamiento, procesamiento y análisis de información relacionada con clientes, ventas y comportamiento del mercado, lo cual exige garantizar la confidencialidad, integridad y disponibilidad de los datos.

De igual forma, el Decreto 1377 de 2013 reglamenta parcialmente la Ley 1581, estableciendo los procedimientos para la autorización del uso de datos personales, la política de tratamiento de información y las responsabilidades de los responsables y encargados del manejo de datos. Esta regulación obliga a las empresas a adoptar mecanismos claros de protección y gestión de la información.

Por otra parte, la Ley 1266 de 2008, conocida como Ley de Habeas Data financiero y comercial, regula el manejo de información crediticia y comercial. Esta ley resulta relevante en el contexto del sector automotriz, debido a que muchas operaciones de venta de vehículos involucran análisis de comportamiento financiero de los clientes, evaluación de riesgo y consulta de historial crediticio.

Asimismo, la Ley 527 de 1999 reconoce la validez jurídica de los mensajes de datos, el comercio electrónico y las firmas digitales en Colombia. Esta normativa respalda el uso de

sistemas informáticos para la gestión de información comercial, contratos digitales y transacciones electrónicas, elementos que pueden integrarse dentro de sistemas de inteligencia de negocios y análisis predictivo.

Adicionalmente, la Ley 1712 de 2014, relacionada con la transparencia y acceso a la información pública, promueve la gestión adecuada de la información y fomenta prácticas de gobierno de datos que fortalecen la toma de decisiones basada en información confiable.

Finalmente, desde el ámbito internacional, se considera como referente la norma ISO/IEC 27001, la cual establece estándares para la gestión de la seguridad de la información. Esta norma proporciona lineamientos para proteger los datos frente a accesos no autorizados, pérdidas o alteraciones, siendo un marco importante para el desarrollo de sistemas analíticos empresariales.

Estado del arte

La estimación de activos vehiculares ha impactado la forma tradicional de la valoración a ojo y la consulta de guías de precios estáticas, tanto así que ha evolucionado hacia un campo de estudio mayor en la analítica de datos y la inteligencia artificial.

Hoy en día, en el mercado la capacidad de predecir el valor comercial de un vehículo usado con alta precisión no es solo una ventaja competitiva sino también un factor clave para la sostenibilidad de las organizaciones como Automoviles Nescar S.A.S; la falta de información clara y las variaciones constantes en el mercado de carros usados generan riesgos financieros tanto para compradores como para vendedores. Según estudios recientes, el uso de algoritmos en

plataformas digitales permite estandarizar los precios y evitar pérdidas económicas causadas por una mala valoración (Pias et al., 2025).

El objetivo de esta revisión sistemática es analizar la literatura reciente relacionada con la aplicación de modelos de aprendizaje automático en la predicción de precios de vehículos usados, con el fin de identificar los algoritmos más utilizados, las variables predictoras relevantes y las principales limitaciones metodológicas presentes en los estudios existentes.

Metodología de revisión sistemática (PRISMA)

Estrategias de búsqueda

La estrategia de búsqueda se diseñó con el propósito de identificar estudios académicos relevantes que abordaran la predicción de precios de vehículos usados mediante técnicas de aprendizaje automático. Para ello, se consultaron bases de datos reconocidas como Google Scholar, SciSpace, Scopus y ArXiv, las cuales ofrecen acceso a literatura científica actualizada.

Se utilizaron combinaciones de palabras clave en inglés, tales como “used car price prediction”, “machine learning”, “random forest”, “XGBoost” y “web scraping”. Estas fueron combinadas mediante operadores booleanos para mejorar la precisión de los resultados y asegurar la recuperación de estudios pertinentes. La ecuación de búsqueda utilizada fue la siguiente: (‘used car price prediction’ OR “vehicle price estimation”) AND (“machine learning” OR “artificial intelligence”) AND (“random forest” OR “XGBoost”) AND (“web scraping”) Asimismo, se estableció un rango temporal entre 2020 y 2026 con el fin de enfocar la revisión en desarrollos recientes del área.

Criterios de inclusión

Para garantizar la relevancia de los estudios seleccionados se definieron criterios de inclusión orientados a filtrar investigaciones con evidencia empírica sólida. En primer lugar, se consideraron únicamente aquellos estudios que implementaran modelos de aprendizaje automático aplicados a la predicción de precios de vehículos usados.

Adicionalmente, se exigió que los artículos reportaran métricas de desempeño cuantitativas como el coeficiente de determinación (R^2), el error cuadrático medio (RMSE) o el error absoluto medio (MAE), lo cual permite evaluar objetivamente la efectividad de los modelos. Finalmente, se incluyeron publicaciones en inglés y español para ampliar el alcance de la revisión.

Criterios de exclusión

De manera complementaria, se establecieron criterios de exclusión para descartar estudios que no aportaran evidencia relevante al análisis. En este sentido, se excluyeron investigaciones que no presentaran validación empírica o que carecieran de resultados cuantitativos verificables.

Asimismo, se descartaron documentos no académicos como blogs o reportes sin revisión por pares debido a su limitada rigurosidad científica, así como estudios publicados antes del 2020 con el fin de garantizar la actualidad de la información garantizada. Además, se excluyeron

aquellos estudios centrados exclusivamente en vehículos nuevos dado que el enfoque de la investigación se orienta al mercado de segunda mano.

Proceso de selección

El proceso de selección se llevó a cabo siguiendo las fases establecidas por la metodología PRISMA 2020. En una primera etapa se identificaron un total de 32 documentos a partir de las búsquedas realizadas en las bases de datos seleccionadas.

Posteriormente, se procedió a la eliminación de duplicados y a la revisión de títulos y resúmenes lo que permitió descartar aquellos estudios que no cumplían con los criterios temáticos definidos. En la fase de elegibilidad se realizó una lectura completa de los artículos preseleccionados para verificar su pertinencia metodológica.

Como resultado de este proceso se incluyeron finalmente 9 estudios en la síntesis cualitativa los cuales constituyen la base del análisis presentado en el estado del arte.

Resultados de la revisión sistemática

Modelos predictivos

Los estudios analizados evidencian un uso predominante de modelos de aprendizaje automático en la predicción de precios de vehículos usados destacándose especialmente los

métodos de ensamble. Estos modelos han ganado popularidad debido a su capacidad para manejar grandes volúmenes de datos y capturar relaciones complejas entre variables.

Dentro de este grupo, algoritmos como XGBoost y LightGBM han mostrado resultados sobresalientes en términos de precisión superando consistentemente a modelos tradicionales. Su funcionamiento basado en la optimización iterativa del error permite mejorar progresivamente las predicciones lo que explica su alto desempeño en diferentes contextos de aplicación (Yilmaz & Selvi, 2023).

Por otra parte, el algoritmo Random Forest continúa siendo una alternativa relevante particularmente en escenarios donde se requiere un equilibrio entre precisión e interpretabilidad. Este modelo es ampliamente valorado por su capacidad para reducir el sobreajuste y manejar variables categóricas sin necesidad de transformaciones complejas (Jin, 2021).

En estudios más recientes se han explorado enfoques híbridos como el stacking, los cuales combinan múltiples algoritmos con el objetivo de mejorar el rendimiento predictivo. Aunque estos modelos suelen ofrecer resultados superiores, también implican un mayor costo computacional y complejidad en su implementación (Fayyaz et al., 2025).

Recolección de datos

La recolección de datos constituye un elemento fundamental en el desarrollo de modelos predictivos y en el caso del mercado de vehículos usados, el web scraping se ha consolidado

como la principal herramienta para este propósito. Esta técnica permite extraer información directamente desde plataformas digitales facilitando la construcción de datasets a gran escala.

Sin embargo, los datos obtenidos mediante scraping suelen presentar diversos problemas de calidad, entre los más comunes se encuentran la presencia de registros duplicados, valores atípicos y datos incompletos lo que puede afectar significativamente el desempeño de los modelos si no se abordan adecuadamente. Por esta razón, los estudios revisados enfatizan la importancia de implementar procesos rigurosos de limpieza y preprocesamiento (Yilmaz & Selvi, 2023).

Adicionalmente, algunos trabajos destacan que la integración de datos provenientes de múltiples fuentes puede generar inconsistencias debido a diferencias en los formatos y estructuras de la información. Este aspecto representa un desafío adicional que debe ser considerado durante la construcción del dataset (Cui et al., 2022).

Variables predictoras

En relación con las variables utilizadas en los modelos existe un consenso generalizado sobre la importancia de ciertos factores clave en la determinación del precio de los vehículos usados. Entre ellos, el año de fabricación y el kilometraje se destacan como los principales determinantes ya que están directamente relacionados con el desgaste y la depreciación del vehículo.

Estas variables suelen presentar relaciones no lineales con el precio lo que justifica el uso de modelos de aprendizaje automático capaces de capturar este tipo de patrones. Asimismo, factores como la marca y el modelo influyen significativamente en la valoración al reflejar aspectos relacionados con la reputación y la demanda del mercado (Jin, 2021).

Por otro lado, variables como el tipo de combustible, el color o el número de propietarios previos tienen un impacto menor, aunque contribuyen a mejorar la precisión del modelo al proporcionar información adicional. No obstante, se ha identificado que muchos estudios no consideran variables macroeconómicas, lo cual puede limitar la aplicabilidad de los modelos en ciertos contextos (El-Amin, 2023).

Desafíos y brechas

A pesar de los avances en el uso de modelos predictivos la literatura presenta diversas limitaciones que deben ser abordadas en futuras investigaciones. Uno de los principales desafíos es la falta de validación temporal lo que implica que los modelos son evaluados con datos históricos sin considerar cambios en el tiempo.

Esta limitación es particularmente relevante en mercados dinámicos donde los precios pueden variar significativamente en función de factores externos. La ausencia de este tipo de validación puede generar una sobreestimación del desempeño de los modelos (Fayyaz et al., 2025).

Otro aspecto crítico es la reproducibilidad de los estudios ya que en muchos casos la falta de información detallada sobre los parámetros utilizados y los procesos de entrenamiento dificulta la replicación de los resultados lo que afecta la transparencia científica.

Finalmente, se identifica una brecha importante en términos geográficos ya que la mayoría de las investigaciones se han desarrollado en Asia y Europa y esto limita la aplicabilidad de los modelos en regiones como Latinoamérica donde las condiciones del mercado presentan características particulares (Pias et al., 2025).

Tabla 1. Matriz del estado del arte

Autor(es) y Año	Título del Estudio	Objetivo Principal	Metodología Empleada	Variables Analizadas	Principales Hallazgos	Limitaciones Reportadas	Relación con el Proyecto Actual
(Pias et al., 2025)	<i>Optimizing Used Car Valuation with AI: A Predictive Modeling Approach</i>	Desarrollar un modelo preciso para estimar el precio de vehículos usados.	Uso de múltiples modelos de machine learning: Linear Regression, Decision Trees, Random Forest, KNN, XGBoost y Stacking. Incluye preprocesamiento, feature engineering y ajuste de hiperparámetros.	Edad del vehículo, kilometraje, tipo de combustible y factores de volatilidad del mercado.	. XGBoost obtuvo el mejor desempeño con un RMSE de 0.240, demostrando alta precisión en la predicción de precios.	No se mencionan limitaciones explícitas en el abstract.	La metodología permite desarrollar un sistema automatizado de valoración de vehículos en Automóviles Nescar S.A.S., mejorando la precisión en la fijación de precios, reduciendo la subjetividad y fortaleciendo la toma de decisiones comerciales con base en datos reales del mercado.

Jin, C. (2021)	<i>Price Prediction of Used Cars Using Machine Learning</i>	Construir un modelo para predecir precios razonables de vehículos usados considerando múltiples características.	Uso de técnicas de machine learning y data science. Aplicación de modelos de regresión: linear, polynomial, support vector regression (SVR), decision tree y random forest. Incluye visualización de datos, preprocesamiento y evaluación con R ² .	Kilometraje, año de fabricación, consumo de combustible, tipo de transmisión, impuesto vehicular (road tax), tipo de combustible y tamaño del motor.	Random Forest obtuvo el mejor desempeño con un R ² de 0.90416, logrando alta precisión. El modelo mejora estudios previos al incluir más variables.	No se mencionan limitaciones explícitas en el abstract.	Se relaciona con el proyecto al reforzar el uso de múltiples variables y modelos predictivos para mejorar la precisión en la estimación de precios. Apoya la implementación de un sistema automatizado en el mercado colombiano, validando el uso de técnicas de machine learning y análisis de datos para optimizar la toma de decisiones en la valoración de vehículos.
Yilmaz, S., & Selvi, İ. H. (2023).	<i>Price Prediction Using Web Scraping and Machine Learning Algorithms in the Used Car Market</i>	Analizar datos de ventas de vehículos y desarrollar un modelo de estimación de precios utilizando machine learning.	Recolección de datos mediante web scraping (Selenium y BeautifulSoup), preprocesamiento de datos, selección de características con Lasso y reducción de dimensionalidad con PCA. Ajuste de hiperparámetros con GridSearchCV. Uso de modelos: Random Forest, KNN, Gradient Boost, AdaBoost, SVR y XGBoost.	No se especifican explícitamente, pero corresponden a características de vehículos extraídas de datos de ventas (implícitamente variables como precio, características técnicas y comerciales).	XGBoost obtuvo el mejor resultado en el dataset 1 (R ² = 0.973, RMSE = 0.161). KNN fue el mejor en el dataset 2 (R ² = 0.978, RMSE = 0.145), mostrando alta precisión en ambos casos.	No se mencionan limitaciones explícitas en el texto.	Se relaciona directamente con el proyecto al validar el uso de web scraping, preprocesamiento avanzado y múltiples algoritmos para mejorar la precisión en la predicción de precios. Refuerza la importancia de la selección de variables y optimización de modelos para construir un sistema robusto aplicable al mercado automotor colombiano.

Isaev, R.,	<i>Predicting Passenger</i>	Desarrollar un	Recolección de datos	Modelo,	CatBoost fue el	No se mencionan	Se relaciona con el
Nuradilova, B.,	<i>Car Prices with</i>	modelo preciso para	mediante web scraping	kilometraje, año,	modelo con mejor	limitaciones	proyecto al validar
&	<i>Machine Learning</i>	predecir precios de	en Python desde un	color, condición	desempeño,	explícitas en el	el uso de web
Gimaletdinova,	<i>Models</i>	reventa de vehículos	sitio de anuncios.	del vehículo y tipo	alcanzando una	texto.	scraping y modelos
G. (2024).		y evitar sobrecostos	Aplicación de modelos	de carrocería	precisión del		avanzados como
		para los	de machine learning:	(extraídos de	86.05%, siendo		CatBoost para
		compradores.	Random Forest,	descripciones de	útil para la toma		mejorar la precisión
			CatBoost Regressor y	anuncios).	de decisiones de		en la predicción de
			Support Vector		clientes e		precios. Refuerza la
			Machine (SVM).		instituciones		importancia de
			Evaluación		financieras.		incluir variables
			comparativa de				extraídas de
			desempeño.				descripciones
							(texto) y de
							comparar modelos
							para seleccionar el
							más adecuado en la
							valoración de
							vehículos en el
							mercado
							colombiano.

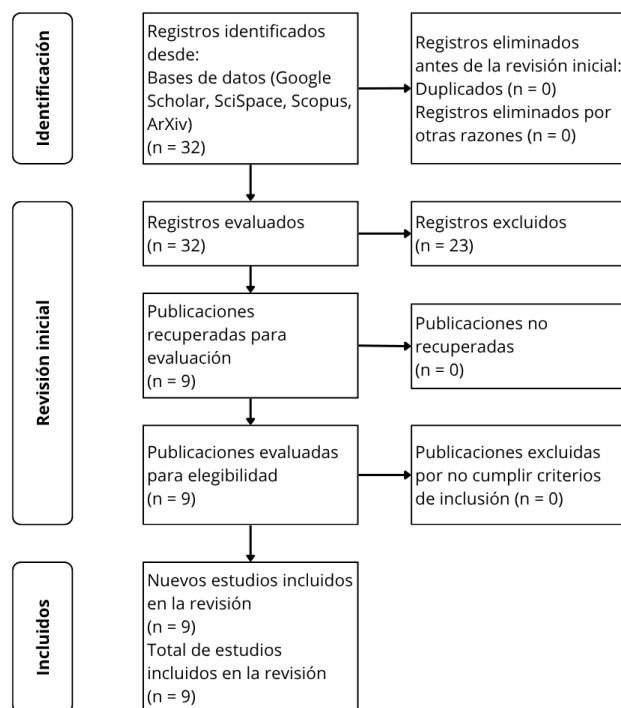
El-Amin, M. F. (2023).	<i>Price Estimation of Used Cars Using Machine Learning Algorithms</i>	Encontrar métodos eficientes para predecir con mayor precisión el precio real de vehículos usados.	Aplicación de modelos de machine learning: Support Vector Regression (SVR), Random Forest y CatBoost. Comparación con métodos existentes mediante métricas como MAE y accuracy.	No se especifican explícitamente, pero corresponden a características relevantes de vehículos usadas para estimar su valor de mercado.	Random Forest obtuvo el mejor desempeño con menor MAE (-0.1317) y mayor precisión (87.16%). Los modelos propuestos mejoran significativamente frente a métodos anteriores (Decision Tree MAE 0.6711).	No se mencionan limitaciones explícitas en el texto.	Se relaciona con el proyecto al reforzar la importancia de comparar múltiples modelos para seleccionar el más preciso. Evidencia que algoritmos como Random Forest y CatBoost pueden ofrecer mejores resultados que modelos tradicionales, aportando una base sólida para optimizar el sistema de valoración vehicular en el contexto colombiano.
Satarkar, P. A., Shitole, S. B., Kagade, S. S., et al. (2022)	<i>Car Price Predictor</i>	Desarrollar un modelo confiable y preciso para la predicción del precio de vehículos usados.	Recolección de datos mediante web scraping en PHP desde portales web. Aplicación y comparación de modelos de machine learning: Artificial Neural Network (ANN), Support Vector Machine (SVM) y Random Forest. Integración del modelo final en una aplicación Java.	No se especifican explícitamente, pero corresponden a múltiples atributos de vehículos usados necesarios para la predicción del precio.	La comparación de modelos permitió identificar el algoritmo con mejor desempeño para el dataset, logrando una predicción más confiable. Se destaca la integración del modelo en una aplicación funcional.	No se mencionan limitaciones explícitas en el abstract.	Se relaciona con el proyecto al reforzar el uso de web scraping, comparación de modelos y desarrollo de soluciones aplicadas. Aporta valor al evidenciar la importancia de implementar el modelo en un sistema funcional, alineándose con el objetivo de construir una herramienta automatizada de valoración de vehículos en el

contexto
empresarial.

Nota. Elaboración propia.

PRISMA

Figura 1. *Diagrama de flujo del proceso de selección de estudios según PRISMA 2020*



Nota. Elaboración propia.

Metodología

La investigación se sitúa dentro de un enfoque cuantitativo, ya que se centra en la recolección, tratamiento y análisis de datos objetivos relacionados con las características de los vehículos usados y su conexión con el valor de mercado. Este enfoque facilita la identificación de patrones estadísticos y la creación de relaciones entre variables a través de técnicas de análisis de datos y modelos de inteligencia artificial, lo que resulta esencial para desarrollar un sistema automatizado que pueda estimar el precio de los vehículos de manera precisa y confiable. El estudio tiene un alcance descriptivo y correlacional. Es descriptivo porque busca caracterizar las

principales variables que afectan la valoración de vehículos usados, como la marca, modelo, año de fabricación, kilometraje, estado general del automóvil, historial de accidentes, número de propietarios anteriores y color del vehículo. Además, es correlacional porque intenta establecer la relación entre estas variables y el precio del vehículo, identificando cuáles factores tienen un impacto más significativo en la determinación del valor comercial en el mercado automotor. En cuanto al diseño de la investigación, se opta por un diseño no experimental de tipo transversal. Se considera no experimental porque las variables de estudio no se manipulan intencionadamente, sino que se analizan datos reales obtenidos del mercado automotor. Asimismo, es transversal porque la recolección de información se lleva a cabo en un momento específico, utilizando registros históricos y actuales de diversas fuentes, lo que permite examinar el comportamiento del mercado sin necesidad de un seguimiento a largo plazo.

En este estudio, se definen tanto variables dependientes como independientes. La variable dependiente se refiere al valor del vehículo, que se entiende como el precio comercial al que se puede comprar o vender un automóvil usado, expresado en pesos colombianos. Por otro lado, las variables independientes abarcan aspectos como la marca, el modelo, el año de fabricación, el kilometraje, el estado general del vehículo, el historial de accidentes, el número de propietarios anteriores y el color del automóvil. Estas variables son clave para establecer relaciones estadísticas que ayudan a construir el modelo predictivo. La población que se estudia está compuesta por vehículos usados que se comercializan en el mercado automotor colombiano.

Dada la diversidad y amplitud de esta población, se utiliza un muestreo probabilístico estratificado, que permite dividir la población en grupos homogéneos según características

relevantes como el tipo de vehículo, el año de fabricación, el rango de precios, el tipo de combustible y el kilometraje. Este enfoque de muestreo asegura que se represente adecuadamente a los diferentes segmentos del mercado, minimizando sesgos en la recolección de datos. El tamaño mínimo de la muestra se calcula utilizando la fórmula de Cochran, estableciendo aproximadamente 385 unidades de análisis con un nivel de confianza del 95 % y un margen de error del 5 %.

Para recolectar información, se emplea la técnica de web scraping, que se implementa a través del lenguaje de programación Python. Esta técnica permite extraer datos reales del mercado automotor desde plataformas digitales especializadas en la compra y venta de vehículos. El proceso de extracción automatizada se lleva a cabo utilizando herramientas tecnológicas como Playwright, que facilita la automatización del navegador y la interacción dinámica con los sitios web; Polars, que se utiliza para limpiar, estructurar y analizar datos; y Nest Asyncio, que optimiza la ejecución de procesos asíncronos, mejorando así la eficiencia en la recolección de información. El script se ejecuta en el entorno de desarrollo Google Colab, lo que hace más sencillo el procesamiento y almacenamiento de la información recolectada. Además, se incorporan fuentes complementarias de información, como registros internos de la empresa relacionados con históricos de compra y venta de vehículos, formatos de inspección vehicular que permiten evaluar el estado físico y mecánico de los automóviles, y entrevistas estructuradas con expertos del sector automotor. Estas fuentes ayudan a complementar la información obtenida mediante web scraping y aseguran una mayor confiabilidad en los datos analizados.

Una vez que se ha reunido toda la información, se lleva a cabo un proceso de limpieza, depuración y organización de los datos. Esto se hace para eliminar inconsistencias, normalizar formatos y preparar la base de datos para su análisis. Luego, se aplica un modelo estadístico de regresión lineal múltiple, que permite estimar el precio de los vehículos en función de varias variables independientes. Este modelo es útil para identificar cómo cada característica del vehículo afecta su valor comercial, lo que permite generar estimaciones automatizadas y precisas. Finalmente, los resultados obtenidos del modelo predictivo permiten establecer patrones en el comportamiento del mercado automotor. Esto contribuye a optimizar el proceso de valoración de vehículos usados en la empresa Automóviles Nescar S.A.S., fortaleciendo la toma de decisiones comerciales, mejorando la competitividad y aumentando la transparencia en la fijación de precios.

Resultados

Resultados del desarrollo del sistema

El producto de este trabajo es un sistema funcional de extracción y análisis automatizado de precios de vehículos usados, capaz de procesar cualquier consulta de marca o modelo que el usuario solicite. Está compuesto por tres elementos integrados: un servidor de aplicación construido sobre Flask, un módulo de extracción de datos basado en Playwright y una interfaz web de consulta. El servidor expone una interfaz de programación de aplicaciones (API) con cuatro puntos de acceso inicio de búsqueda, consulta de estado, entrega de datos procesados y

descarga de resultados en formato Excel que desacoplan la capa de presentación de la lógica de extracción. Esta separación permitió que cada operación de extracción, que es prolongada, se ejecutara en un hilo en segundo plano sin bloquear la interfaz, y que el estado del proceso se persistiera en disco para ser consultado de forma asíncrona por el cliente.

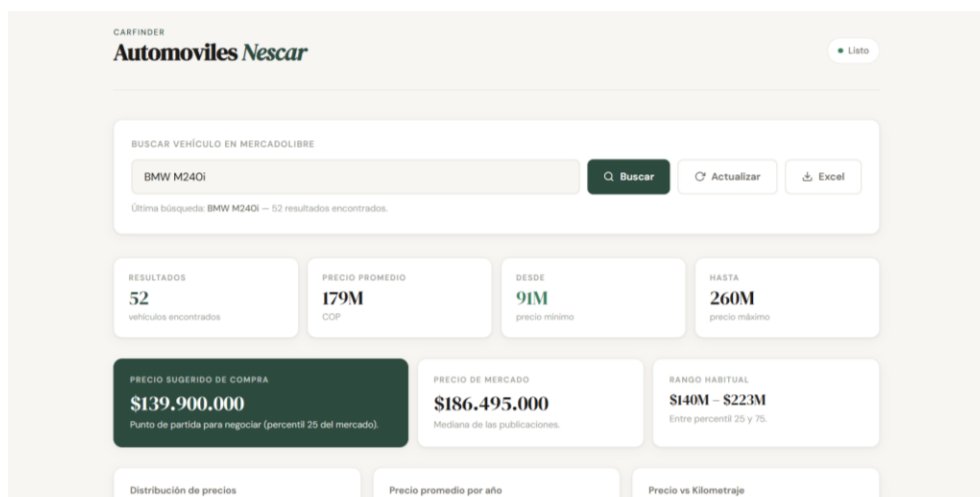
El sistema fue concebido para operar de manera repetida sobre búsquedas distintas: el usuario introduce un término, el sistema realiza la extracción, almacena los resultados de la última consulta y genera tanto las estadísticas como el archivo descargable correspondiente. De este modo, la herramienta no produce un único informe estático, sino que funciona como un instrumento de consulta recurrente aplicable a cualquier segmento del mercado de vehículos usados.

Un resultado relevante desde el punto de vista de la ingeniería fue la necesidad de reemplazar la estrategia inicial de adquisición de datos. El enfoque original contemplaba el consumo de la interfaz oficial de búsqueda del marketplace mediante credenciales de aplicación y autenticación por token, según se documenta en el módulo de diagnóstico desarrollado durante el proyecto. Sin embargo, las pruebas demostraron que dicha interfaz devolvía respuestas de acceso denegado, lo que obligó a redefinir la arquitectura de adquisición. La solución adoptada fue la extracción directa de los listados públicos del portal sin requerir inicio de sesión, complementada con técnicas de mitigación de detección automatizada emulación de las cabeceras y propiedades de un navegador real y enmascaramiento de los indicadores típicos de automatización y con un mecanismo de reconocimiento de la barrera de seguridad del sitio que interrumpe la extracción de forma controlada cuando esta aparece. Este cambio de estrategia constituye en sí mismo un hallazgo del proyecto: confirma que la disponibilidad de fuentes de

datos comerciales no es estable en el tiempo y que un sistema de recolección robusto debe contemplar rutas de adquisición alternativas.

El módulo de extracción demostró capacidad para recorrer múltiples páginas de resultados de manera paginada, eliminar registros duplicados mediante una clave compuesta por título y precio, y normalizar cada vehículo en una estructura uniforme de atributos: título, marca, modelo, año, kilometraje, precio y enlace. La extracción del kilometraje exigió un tratamiento particular, ya que en las tarjetas de los listados este dato aparece contiguo al año de fabricación; se diseñó una rutina de reconocimiento que aísla la cifra de kilómetros sin confundirla con el año, garantizando la integridad de ambos campos. Sobre esta estructura, el servidor calcula automáticamente, para cada búsqueda, un conjunto de estadísticas descriptivas y una distribución de precio por año de fabricación, que constituyen la base del componente analítico del sistema.

Figura 2. *Interfaz principal del programa.*

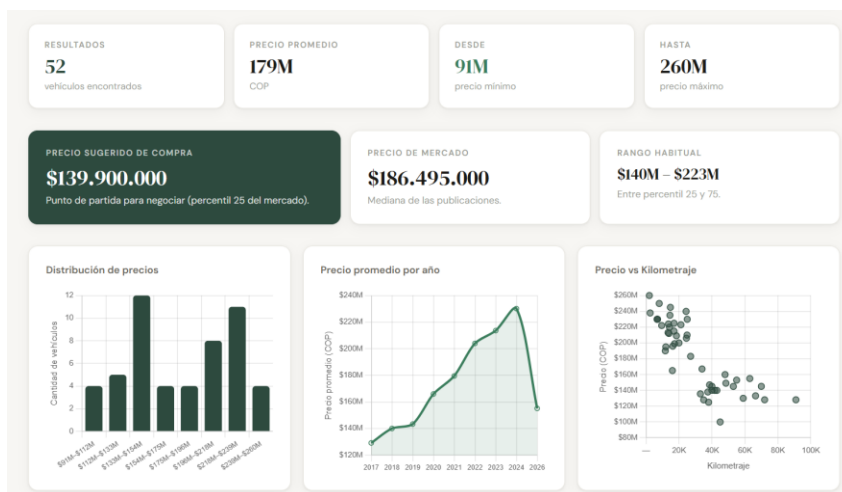


Nota. Elaboración propia.

Resultados de la recolección y el procesamiento de datos

A lo largo de las distintas búsquedas ejecutadas, el sistema demostró un comportamiento consistente en la calidad de la captura. Para cada consulta recolecta un volumen significativo de publicaciones del orden de un centenar de vehículos por búsqueda y obtiene de forma fiable los campos relevantes para el análisis: el precio, el año de fabricación y el kilometraje. Las rutinas de extracción y normalización se mostraron robustas frente a la variabilidad de formato de las publicaciones, entregando registros estructurados aptos para el cálculo estadístico inmediato. En el caso del kilometraje, la captura depende de que la publicación incluya el dato; cuando no está presente, el registro se conserva con el valor ausente debidamente señalado, de modo que su falta no contamina los promedios ni las gráficas.

Figura 3. *Interfaz principal del sistema con los indicadores de mercado*



Nota. Elaboración propia.

La inspección de los datos también permitió identificar una limitación recurrente en la precisión del filtrado. Junto a los vehículos que corresponden estrictamente al término

consultado, la fuente incorpora ocasionalmente resultados relacionados de otras marcas o modelos, que el sistema captura como parte del listado. Esta proporción de ruido es reducida y no compromete la validez general de los indicadores, pero evidencia la dependencia del sistema respecto de la relevancia de los resultados que entrega el portal, un aspecto que se retoma en las conclusiones y en las mejoras propuestas.

Figura 4. *Interfaz de resultados encontrados en MercadoLibre*

Mostrando 52 resultados

VEHÍCULO	MARCA	AÑO	KILOMETRAJE	PRECIO	ENLACE
Bmw Serie 2 3.0 M240i F22 Coupe	Bmw	2026	2.000 km	COP 260.000.000	Ver ↗
Bmw Serie 2 M240i	Bmw	2024	8.000 km	COP 249.900.000	Ver ↗
Bmw Serie 2 M240i Coupé 3.0	Bmw	2024	14.775 km	COP 245.000.000	Ver ↗
Bmw Serie 2 3.0 M240i	Bmw	2024	24.300 km	COP 239.900.000	Ver ↗
Bmw M2 3.0 M240i Xdrive Coupe	Bmw	2026	2.500 km	COP 237.900.000	Ver ↗
Bmw Serie 2 3.0 M240i	Bmw	2024	14.500 km	COP 235.000.000	Ver ↗
Bmw Serie M 3.0 M240i Xdrive Coupe	Bmw	2024	6.800 km	COP 230.000.000	Ver ↗
Bmw Serie 2 3.0 M240i G42	Bmw	2024	25.000 km	COP 229.900.000	Ver ↗
Bmw Serie 2 M240i	Bmw	2024	7.000 km	COP 229.900.000	Ver ↗

Nota. Elaboración propia.

Resultados del análisis del mercado

El componente analítico del sistema genera, para cada búsqueda, una caracterización cuantitativa del segmento consultado mediante un conjunto coherente de indicadores: precio promedio, mediana, valores mínimo y máximo, y los cuartiles primero y tercero. La combinación

de estas medidas permite describir no solo el nivel de precios de un segmento, sino también su dispersión y la forma de su distribución. De manera sistemática se observa que la diferencia entre el promedio y la mediana revela distribuciones sesgadas, en las que un grupo reducido de unidades de precio elevado desplaza el promedio por encima del valor central, y que el coeficiente de variación dentro de un mismo modelo puede ser considerable, atribuible a las diferencias de año, versión, kilometraje y estado de conservación de cada unidad.

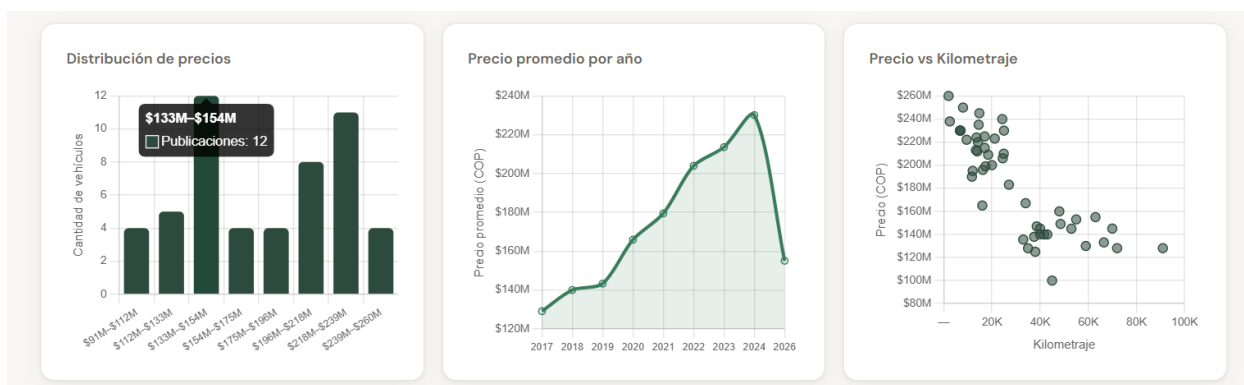
El primer eje de análisis es la distribución de precio promedio según el año de fabricación. De forma consistente entre búsquedas, esta distribución exhibe una relación monótona y creciente: los modelos más antiguos concentran los precios más bajos y los más recientes los más altos, con una progresión sostenida que cuantifica el efecto de la depreciación sobre el valor de reventa. A modo de ilustración, en una de las búsquedas realizadas correspondiente a un modelo de gama media-alta los precios promedio variaron desde alrededor de 38 millones de pesos para las unidades más antiguas hasta cerca de 141 millones para las más recientes, a lo largo de más de una década de años de fabricación. Esta clase de resultado proporciona una referencia objetiva para estimar el precio esperado de una unidad en función de su antigüedad, replicable para cualquier modelo que el usuario consulte.

El segundo eje de análisis, incorporado mediante la captura del kilometraje, es la relación entre el precio y el recorrido de cada vehículo. El sistema representa esta relación en una gráfica de dispersión que enfrenta kilometraje y precio para todas las unidades de la búsqueda. Esta visualización tiene un valor práctico directo: permite identificar oportunidades de compra al detectar vehículos que se ubican por debajo de la nube general de puntos, es decir, unidades con precio comparativamente bajo para su nivel de kilometraje, así como reconocer publicaciones

sobrevaloradas. De esta forma, el kilometraje complementa al año como segunda dimensión objetiva de valoración, acercando el análisis a la lógica real con la que un comprador evalúa el estado y el desgaste de un vehículo.

En conjunto, las estadísticas descriptivas, la distribución por año y la relación precio-kilometraje constituyen una herramienta de valoración de utilidad práctica y de aplicación general. La mediana ofrece un precio de mercado de referencia; el primer cuartil señala un objetivo de negociación a la baja; y el tercer cuartil delimita la zona alta del mercado. De este modo, para cualquier búsqueda, el sistema transforma un conjunto disperso de publicaciones en información estructurada que respalda la toma de decisiones de compra, venta o tasación.

Figura 5. *Interfaz de estadísticas de búsquedas encontradas.*



Nota. Elaboración propia.

Discusión

A partir de los resultados obtenidos y la investigación realizada, se logra identificar un gran potencial en la implementación de técnicas de web scraping y analítica de datos para

fortalecer la gestión de valoración de vehículos usados en empresas del sector automotor colombiano. La capacidad del sistema para recolectar, estructurar y analizar información de mercado de manera automatizada demuestra que es posible construir una herramienta de apoyo a la decisión funcional sin depender exclusivamente de modelos predictivos complejos de aprendizaje automático, lo cual representa una ventaja significativa en términos de implementación y escalabilidad para una organización como Automóviles Nescar S.A.S.

Un aspecto por destacar de esta investigación es que se logró generar valor analítico real utilizando únicamente datos públicos disponibles en un marketplace, sin requerir acceso a fuentes propietarias ni infraestructura computacional avanzada. La combinación de indicadores estadísticos descriptivos —precio promedio, mediana, cuartiles— con la distribución de precios por año de fabricación y la relación precio-kilometraje ofrece un marco de valoración objetivo que supera directamente las estimaciones empíricas y subjetivas que predominan actualmente en el sector.

Sin embargo, los resultados también evidencian limitaciones que deben considerarse para mejorar la robustez del sistema. La dependencia de una única fuente de datos externa representa el principal riesgo de sostenibilidad del modelo, tal como se observó durante el desarrollo cuando el cierre del acceso a la interfaz oficial obligó a redefinir la estrategia de adquisición. Este hallazgo es consistente con lo señalado por Yilmaz y Selvi (2023), quienes identifican la inconsistencia entre fuentes como uno de los desafíos centrales en sistemas de recolección de datos para mercados automotores. La incorporación de múltiples portales de venta y mecanismos de adquisición alternativos podría mitigar este riesgo de manera considerable.

Adicionalmente, la presencia recurrente de registros no pertenecientes al término consultado —ruido derivado de los resultados relacionados que entrega el portal— limita parcialmente la precisión de los indicadores calculados. Aunque esta proporción es reducida y no compromete la validez general del análisis, su corrección mediante filtrado posterior más estricto contribuiría a mejorar la confiabilidad de las estimaciones, en especial en búsquedas de modelos con nombres que generan ambigüedad en la plataforma.

Desde una perspectiva más amplia, los resultados de este proyecto se alinean con la tendencia identificada en la revisión sistemática de literatura, donde algoritmos como Random Forest y XGBoost han demostrado consistentemente su capacidad para capturar relaciones no lineales entre variables como el año de fabricación, el kilometraje y el precio (Jin, 2021; Pias et al., 2025). El sistema desarrollado sienta las bases de datos y la estructura necesaria para que, en una fase posterior, se implemente un modelo predictivo de este tipo, convirtiendo la herramienta actual en un verdadero tasador automático con mayor alcance y precisión.

Finalmente, cabe resaltar la relevancia geográfica de esta propuesta. Como lo señala la revisión de literatura, la mayoría de los estudios sobre predicción de precios de vehículos usados se han desarrollado en Asia y Europa, dejando una brecha importante para mercados latinoamericanos con dinámicas particulares (Pias et al., 2025). Este proyecto constituye un paso concreto hacia la reducción de esa brecha en el contexto colombiano, aportando una solución adaptada a las condiciones reales del mercado automotor local.

Plan de Divulgación

Con la finalidad de dar a conocer y promover los resultados de este proyecto, se propone un plan de divulgación orientado al ámbito académico, tecnológico y empresarial. El objetivo es visibilizar el potencial de las técnicas de analítica de datos y aprendizaje automático en la valoración de vehículos usados, y fomentar su adopción como herramienta de apoyo a la toma de decisiones en el sector automotor colombiano.

En el ámbito académico, se propone la socialización del proyecto mediante presentaciones en espacios institucionales, semilleros de investigación y eventos relacionados con ciencia de datos e inteligencia artificial. De igual forma, se busca la publicación del trabajo en repositorios académicos y plataformas de acceso abierto, con el fin de facilitar su consulta y contribuir al desarrollo del conocimiento en un área con escasa representación para el contexto latinoamericano.

En el ámbito tecnológico, se plantea la elaboración de documentación detallada del sistema desarrollado, incluyendo la arquitectura del aplicativo, el flujo de extracción y procesamiento de datos, y las decisiones metodológicas adoptadas. Adicionalmente, se propone la publicación del código fuente en plataformas colaborativas como GitHub, promoviendo así la visibilidad del proyecto y permitiendo que otros desarrolladores o investigadores puedan replicarlo, adaptarlo o mejorarlo.

En el ámbito empresarial, se busca presentar los hallazgos a empresas del sector automotor colombiano, mostrando cómo un sistema de valoración basado en datos de mercado puede reemplazar criterios subjetivos, mejorar la fijación de precios y fortalecer la competitividad comercial. En particular, se propone socializar los resultados directamente con

Automóviles Nescar S.A.S., evidenciando las oportunidades concretas de implementación y escalabilidad de la solución desarrollada.

Finalmente, este plan de divulgación tiene como propósito incentivar futuras investigaciones en la intersección entre inteligencia artificial y el mercado automotor, promover la transferencia tecnológica hacia organizaciones del sector y consolidar el uso de herramientas de analítica de datos como base para la toma de decisiones más objetivas, eficientes y competitivas.

Posibles mejoras

- **Mejoras a corto plazo**

En el corto plazo, la prioridad es elevar la precisión y la fiabilidad de lo que el sistema ya hace. La primera mejora consiste en incorporar un filtrado posterior a la extracción que descarte de forma automática los registros cuya marca o modelo no coincide con el término consultado, eliminando así el ruido recurrente que introduce la fuente. Una segunda mejora es ampliar el conjunto de atributos capturados por publicación de manera destacada la transmisión, el tipo de combustible y la ubicación, aprovechando que parte de esa información ya aparece en las tarjetas de los listados, lo que enriquecería el análisis sin cambiar la arquitectura. En tercer lugar, conviene reforzar el manejo de errores y reintentos ante la aparición de la barrera de seguridad del portal, de modo que una búsqueda interrumpida pueda reanudarse o reportar con claridad el motivo del corte. Finalmente, resulta de bajo costo y alto valor añadir validaciones a los datos recolectados descartando precios o kilometrajes atípicos evidentes y años inconsistentes para mejorar la confiabilidad de las estadísticas.

- **Mejoras a largo plazo**

En el largo plazo, las mejoras apuntan a transformar la herramienta de un instrumento de consulta puntual a una plataforma de inteligencia de mercado. Una primera línea es la persistencia histórica de los resultados: almacenar cada búsqueda en una base de datos con marca temporal permitiría analizar la evolución de los precios en el tiempo, detectar tendencias y estacionalidades, y ofrecer comparaciones entre periodos. Sobre esa base de datos histórica sería posible, en una segunda línea, desarrollar un modelo de estimación de precios, por ejemplo, de aprendizaje automático que prediga el valor esperado de un vehículo a partir de su modelo, año, kilometraje y demás atributos, convirtiendo el sistema en un verdadero tasador automático; la incorporación del kilometraje como variable ya disponible constituye un primer insumo natural para dicho modelo. Una tercera línea consiste en independizar la solución de una única fuente, integrando varios portales de venta y consolidando sus resultados, lo que reduciría el riesgo asociado a la dependencia de un solo sitio y ampliaría la cobertura del mercado. Por último, se proponen mejoras de despliegue y accesibilidad como la ejecución del sistema en la nube y la programación de búsquedas automáticas periódicas, junto con un panel de visualización más completo que presente la evolución de los indicadores de manera gráfica e interactiva.

Conclusiones

El presente trabajo logró diseñar, implementar y validar un sistema funcional capaz de recolectar de forma automatizada las publicaciones de vehículos usados de un marketplace y de transformarlas en información analítica de soporte a la decisión, aplicable de manera recurrente a cualquier marca o modelo. El sistema cumplió su objetivo a lo largo de las búsquedas realizadas:

para cada consulta obtuvo un volumen representativo de publicaciones con datos de precio, año y kilometraje, calculó indicadores estadísticos de mercado y generó un reporte descargable, demostrando la viabilidad de la propuesta de principio a fin.

Desde la perspectiva de la ingeniería de sistemas, se concluye que la arquitectura desacoplada en tres capas: Extracción, servicio y presentación resultó adecuada para una tarea de naturaleza prolongada y asíncrona, al permitir que la interfaz permaneciera disponible mientras la extracción se ejecutaba en segundo plano. Asimismo, se concluye que la dependencia de fuentes de datos comerciales externas representa el principal riesgo de un sistema de esta clase: el cierre del acceso a la interfaz oficial durante el desarrollo obligó a una redefinición de la estrategia de adquisición y demuestra que la sostenibilidad de la solución exige prever rutas alternativas y mecanismos de adaptación ante los cambios y las defensas anti-automatización de la fuente.

Desde la perspectiva del análisis de mercado, se concluye que los segmentos de vehículos usados presentan una dispersión de precios elevada incluso dentro de un mismo modelo, y que tanto el año de fabricación como el kilometraje son determinantes claros y cuantificables de su valor. El año exhibe una relación creciente y sostenida con el precio, observada de forma consistente entre búsquedas, mientras que el kilometraje aporta una segunda dimensión que permite distinguir, dentro de un mismo año, las unidades más o menos desgastadas. La combinación de ambos factores valida la utilidad del sistema como instrumento de valoración objetiva, en contraste con la estimación subjetiva que predomina habitualmente en las transacciones de vehículos usados.

Entre las limitaciones del trabajo se reconoce, en primer lugar, la dependencia de la herramienta respecto de la estructura del sitio web de la fuente, lo que la hace sensible a cambios en el diseño de sus páginas. En segundo lugar, se identificó una pequeña proporción recurrente de registros no pertenecientes al término consultado, originada en los resultados relacionados que entrega la propia fuente, lo que señala la conveniencia de incorporar un filtrado posterior más estricto. Por último, la disponibilidad del kilometraje depende de que cada publicación lo informe, por lo que el análisis de esta variable se limita a las unidades que lo declaran.

Referencias

- Infobae (2024). *La inteligencia artificial y su aporte a mejorar el mercado de autos usados*. Recuperado de: <https://www.infobae.com/autos/2024/08/27/>
- Repositorio Uniremington (2023). *Modelo de predicción del precio de vehículos usados usando Random Forest*.
- Arxiv (2017). *Used car price prediction using machine learning*. Recuperado de: <https://arxiv.org/abs/1711.06970>
- Infobae. (2024, 27 de agosto). *La inteligencia artificial y su aporte a mejorar el mercado de autos usados*. <https://www.infobae.com/autos/2024/08/27/>
- Arxiv (2018). *Deep Learning based Visual Analysis for Vehicle Price Estimation*. Recuperado de: <https://arxiv.org/abs/1803.11227>
- Congreso de la República de Colombia. (1999). Ley 527 de 1999. Por medio de la cual se define y reglamenta el acceso y uso de los mensajes de datos, del comercio electrónico y de las firmas digitales.
- Congreso de la República de Colombia. (2008). Ley 1266 de 2008. Por la cual se dictan disposiciones generales del hábeas data y se regula el manejo de la información financiera.
- Congreso de la República de Colombia. (2012). Ley 1581 de 2012. Por la cual se dictan disposiciones generales para la protección de datos personales.
- Congreso de la República de Colombia. (2014). Ley 1712 de 2014. Ley de transparencia y del derecho de acceso a la información pública nacional.

- ANDEMOS. (2023). Informe del sector automotor colombiano. Asociación Nacional de Movilidad Sostenible.
- DANE. (2023). Indicadores económicos del sector comercial en Colombia. Departamento Administrativo Nacional de Estadística.
- Davenport, T. H., & Harris, J. G. (2017). Competing on analytics: The new science of winning. Harvard Business Review Press.
- Mayer-Schönberger, V., & Cukier, K. (2013). Big data: A revolution that will transform how we live, work, and think. Houghton Mifflin Harcourt.
- Pias, M., Sartaz, M. S., Mazumder, A., et al. (2025). Optimizing Used Car Valuation with AI: A Predictive Modeling Approach. ICMI 2025.
<https://ieeexplore.ieee.org/document/11141166/>
- Jin, C. (2021). Price Prediction of Used Cars Using Machine Learning. ICESIT 2021.
<https://ieeexplore.ieee.org/document/9696839>
- Yilmaz, S., & Selvi, İ. H. (2023). Price Prediction Using Web Scraping and Machine Learning Algorithms in the Used Car Market. Sakarya University Journal of Computer and Information Sciences. <https://dergipark.org.tr/en/download/article-file/3185615>
- Cui, B., Ye, Z., Zhao, H., et al. (2022). Used Car Price Prediction Based on the Iterative Framework of XGBoost+LightGBM. Electronics, 11(18), 2932.
<https://doi.org/10.3390/electronics11182932>
- Fayyaz, I., Nawaz Ali, G. G. Md., & Khairunnesa, S. S. (2025). Advanced Feature Engineering and Machine Learning Techniques for High Accurate Price Prediction of Heterogeneous Pre-Owned Cars. Vehicles, 7(3). <https://www.mdpi.com/2624-8921/7/3/94>

- Isaev, R., Nuradilova, B., & Gimaletdinova, G. (2024). Predicting Passenger Car Prices with Machine Learning Models. Preprints.
<https://www.preprints.org/manuscript/202412.0998>
- El-Amin, M. F. (2023). Price Estimation of Used Cars Using Machine Learning Algorithms. Lecture Notes in Computer Science. https://doi.org/10.1007/978-3-031-28975-0_3
- Satarkar, P. A., Shitole, S. B., Kagade, S. S., et al. (2022). Car Price Predictor. International Journal of Advanced Research in Science, Communication and Technology.
<https://doi.org/10.48175/ijarsct-4654>
- Fayyaz, I., Nawaz Ali, G. G. Md., & Khairunnesa, S. S. (2025). Advanced Feature Engineering and Machine Learning Techniques for High Accurate Price Prediction of Heterogeneous Pre-own Cars. Preprints.
<https://www.preprints.org/manuscript/202507.1150>