Implementar un modelo de predicción de precios de vivienda para la vereda Canavita del municipio de Tocancipá, que apoye a los posibles inversores en la decisión de compra y venta de finca raíz

Salomón Palacios Suárez

Asesor

Jorge Luis Quintero

Universidad Nacional Abierta y a Distancia UNAD

Escuela de Ciencias Básicas, Tecnología e Ingeniería ECBTI

Especialización en Ciencia de Datos y Analítica

2025

Resumen

El presente proyecto de grado tiene como objetivo desarrollar un modelo de predicción de precios de vivienda para la vereda Canavita, ubicada en el municipio de Tocancipá, Cundinamarca. En el estudio se consideran los tipos de inmueble (casas y apartamentos) y su estado (nuevo o usado), empleando algoritmos de Machine Learning para la construcción del modelo predictivo. Los datos fueron obtenidos mediante la técnica de web scraping a partir de los portales inmobiliarios FincaRaiz, MetroCuadrado y CienCuadras. La predicción de precios se llevó a cabo aplicando técnicas de aprendizaje automático con el fin de generar un modelo que sirva como apoyo en la toma de decisiones informadas para personas interesadas en invertir o vender inmuebles en esta zona. Los resultados evidencian que variables como el área construida, área privada, tipo de inmueble, estrato, número de habitaciones, precio de administración, número de baños, y otras características propias del inmueble, son factores determinantes en la estimación del precio de una vivienda.

***Palabras clave*:** Algoritmo, predicción, modelo, aprendizaje, automático

Abstract

This undergraduate research project aims to develop a housing price prediction model for the rural area of Canavita, located in the municipality of Tocancipá, Cundinamarca. The study considers different types of properties, including houses and apartments, and their condition (new or used), applying Machine Learning algorithms to build the predictive model. The dataset was collected through web scraping techniques from the real estate platforms FincaRaiz, MetroCuadrado, and CienCuadras. The price prediction process was conducted using machine learning methods to create a model that supports informed decision-making for individuals interested in investing in or selling properties in this area. The results show that variables such as built area, private area, property type, socioeconomic stratum, number of bedrooms, administration fee, number of bathrooms, and other property characteristics are determining factors in estimating the price of a home.

Keywords**:** Algorithm, prediction, model, learning, automatic

Tabla de Contenido

Lista de Tablas

Lista de Figuras

Lista de Apéndices

Introducción

El comportamiento del mercado inmobiliario en Colombia ha mostrado variaciones significativas en los últimos años. Según el DANE, durante 2025 los precios de las casas aumentaron un 4,15% y los de los apartamentos un 2,01%, mientras que la vivienda nueva en general presentó un incremento del 5,63% frente al cierre de 2024. (Galeano, 2025). Contar con estimaciones confiables sobre el valor de los inmuebles resulta esencial para quienes desean comprar, vender o invertir en el sector, ya que facilita la toma de decisiones informadas. A pesar de los avances en el uso de técnicas de Machine Learning para predecir precios de vivienda a nivel global, en Colombia aún existe una limitación en el análisis de datos a escala local, especialmente en municipios y veredas. (Grajales, 2019). Ante esta situación, el presente proyecto tiene como objetivo desarrollar un modelo de predicción de precios de vivienda para la vereda Canavita, del municipio de Tocancipá (Cundinamarca**)**, una zona en expansión donde nuevas obras de infraestructura han incrementado la valorización de los inmuebles. El documento se estructura de la siguiente manera: en la definición del problema se describe la situación que motiva la investigación; en la justificación, se expone la importancia de su abordaje; posteriormente, se presentan los objetivos generales y específicos del estudio. El marco teórico desarrolla los conceptos fundamentales sobre aprendizaje automático y los modelos relevantes para este trabajo. En la metodología se detallan las etapas del proceso de obtención, depuración y análisis de los datos, así como la implementación del modelo predictivo. Finalmente, se presentan los resultados, conclusiones y recomendaciones para futuras investigaciones relacionadas con el sector inmobiliario.

Definición del Problema

¿Vale la pena invertir en bienes raíces en el municipio de Tocancipá?

El municipio de Tocancipá, en el departamento de Cundinamarca, ha experimentado en las últimas décadas un importante proceso de crecimiento industrial y urbano. Este desarrollo se originó con la construcción de la cervecería Leona y se consolidó con la llegada de grandes empresas como Kimberly Colpapel, Bel Star y, más recientemente, Coca-Cola. La presencia de estas industrias generó un notable aumento poblacional, impulsando la demanda de vivienda y la expansión urbanística del municipio.(Rodriguez, 2024) Ante este crecimiento, la administración local otorgó licencias de construcción que facilitaron la llegada de constructoras interesadas en atender la creciente necesidad de vivienda. Entre las ocho veredas que conforman el municipio, la vereda Canavita se destaca por su ubicación estratégica y por ser una de las más beneficiadas con mejoras en infraestructura, como la pavimentación de vías y la construcción de puentes peatonales, lo cual ha incrementado la valorización del suelo y el atractivo para la inversión inmobiliaria. Sin embargo, este mismo dinamismo genera incertidumbre en la fijación de precios de vivienda, ya que no existe una herramienta actualizada que permita estimar con precisión el valor comercial de los inmuebles en la zona. En Colombia, la determinación de los precios inmobiliarios se rige por metodologías técnicas y normativas establecidas por el Instituto Geográfico Agustín Codazzi (IGAC), anteriormente mediante la Resolución 620 de 2008, y actualmente por la Resolución 1137 de 2024, que define los criterios y parámetros para realizar avalúos comerciales dentro del marco legal vigente (Ley 2294 de 1993). Estas metodologías incluyen procedimientos complejos, como los métodos de comparación de mercado, renta-capitalización, valor residual y la definición de zonas homogéneas geoeconómicas, entre otros. Si bien estos procesos garantizan rigor técnico, también resultan lentos, costosos y, en ocasiones,

subjetivos, debido a la intervención de avaluadores que pueden interpretar los valores de forma diferente.(Resolución No 1137 de 2024, 2024). En este contexto, surge la necesidad de diseñar un modelo de predicción de precios de vivienda basado en aprendizaje automático supervisado, que permita estimar los valores de los inmuebles de manera rápida, objetiva y confiable. Este modelo busca convertirse en una herramienta de apoyo para los habitantes e inversionistas del municipio de Tocancipá, facilitando la toma de decisiones informadas en la compra y venta de bienes raíces, especialmente en la vereda Canavita.

Justificación

En un contexto mundial cada vez más orientado hacia el desarrollo tecnológico, resulta fundamental aprovechar las herramientas que ofrece la ciencia de datos para optimizar los procesos de análisis y toma de decisiones. En estudios recientes se ha identificado un enfoque que consiste en emplear atributos internos y externos del inmueble ya que ambos se perciben de forma distinta en el mercado de esta manera los inmuebles con precios similares son diferenciados de los mercados inmobiliarios.(Beto, 2024) El uso de metodologías basadas en Machine Learning permite transformar grandes volúmenes de información en conocimiento útil, promoviendo así la eficiencia y la innovación en distintos sectores, incluido el mercado inmobiliario. En Colombia, el crecimiento demográfico y urbano ha impulsado una mayor demanda de vivienda, especialmente en municipios en expansión como Tocancipá, donde la vereda Canavita se ha convertido en una zona de alto interés para la inversión inmobiliaria. Sin embargo, la falta de herramientas tecnológicas que ofrezcan información confiable y actualizada sobre los precios de los inmuebles genera incertidumbre en las decisiones de compra y venta. Por ello, desarrollar un modelo predictivo de precios de vivienda mediante técnicas de aprendizaje automático constituye una alternativa innovadora y práctica para ofrecer a los ciudadanos una herramienta amigable, precisa y de fácil acceso.(García, 2021) Este modelo permitirá estimar el valor de los inmuebles con base en variables reales y verificables, reduciendo la subjetividad presente en los métodos tradicionales de avalúo y optimizando el tiempo y los recursos requeridos en este proceso. Además, al fomentar decisiones de inversión más informadas, el modelo contribuirá al crecimiento económico local, impulsará la dinamización del mercado inmobiliario y promoverá una mejor calidad de vida para los habitantes de la vereda Canavita del municipio de Tocancipá.

Objetivos

Objetivo General

Implementar modelos de predicción de precios de vivienda para la vereda Canavita del municipio de Tocancipá, utilizando algoritmos de Machine Learning, con el propósito de generar estimaciones del precio de la vivienda que puedan ser consultadas por personas que tengan dominio de machine learning, y que sirvan de apoyo en la toma de decisiones relacionadas con la inversión, compra o venta de bienes inmuebles en la zona rural del municipio.

Objetivos Específicos

Recolectar datos relevantes del mercado inmobiliario en la vereda Canavita, del municipio de Tocancipá, mediante técnicas de web scraping y otras fuentes disponibles.

Realizar un análisis exploratorio de datos (EDA) sobre las variables que influyen en el precio de las viviendas en la vereda Canavita del municipio de Tocancipá.

Entrenar diferentes algoritmos de Machine Learning que permitan el modelamiento y la predicción de los precios de las viviendas en la vereda Canavita del municipio de Tocancipá, comparando métricas de evaluación y seleccionando el enfoque más preciso y eficiente.

Marco de Referencia

Estado del Arte

En los últimos años la aplicación de técnicas de Machine Learning en el sector inmobiliario ha experimentado grandes avances, convirtiéndose en una herramienta valiosa para el análisis de datos, la predicción de precios de vivienda y la detección de oportunidades en el sector inmobiliario. (Bruno, 2023). Según un estudio realizado en el año 2023 se ilustra como el aprendizaje automático en el mercado inmobiliario puede proporcionar predicciones de precios más precisas que la estadística tradicional. Se puede comprobar como modelos como: k-Nearest Neighbors y Random Forest superan los modelos de precios hedónicos en cuanto a minimización de costos y poder explicativo. (Choy y Ho, 2023). Los modelos de Machine Learning se han empleado para detectar oportunidades de inversión como es el caso de Baldominos, quien implementó una aplicación de aprendizaje automático que identificó en tiempo real las oportunidades del mercado, viviendas que aparecían listas para habitar con un precio inferior al del mercado actual. Todo esto logró identificar que las personas interesadas en vender inmuebles no actualizaban los precios o fijaban estos precios deliberadamente sin ningún estudio previo.(González De La Cruz, 2022). Se han realizado estudios empleando métodos simples como la regresión múltiple los cuales logran predecir el precio de los inmuebles con gran eficacia (Zhang, 2021). También se han realizado estudios un poco más complejos como los realizados en España donde se desarrolló una aplicación que proporciona el mejor modelo de predicción de precios para los inmuebles de cada municipio alcanzando los mejores resultados con las técnicas Bagging y Random Forest. (J.-L. Alfaro et al., 2020). Se han realizado estudios empleando técnicas de aprendizaje automático como el realizado en la ciudad de Guayaquil Ecuador donde se identifica la necesidad de implementar un modelo capaz de predecir el precio

de las viviendas de forma precisa y confiable, alimentando el modelo con datos reales obtenidos de los portales inmobiliarios. (Preciado, 2025). Por otro lado, el Machine Learning se ha empleado también como un mecanismo de detección de riesgos, donde se analiza la relación entre los mercados de vivienda y mercados valores centrándose en las burbujas del mercado inmobiliario detectando de forma eficaz los cambios y la volatilidad futura en los precios del mercado. (Park y Ryu, 2021). En el mercado suizo se empleó Machine Learning para analizar la evolución de los precios de las viviendas en todos los distritos suizos donde se identificaron once distritos críticos que mostraban señales de burbujas y siete distritos en los cuales se identificó que la burbuja ya había estallado.(Moraleda, 2023). En los Estados Unidos se han realizado algunos estudios concretamente en La Florida donde se emplearon modelos de predicción de precios de vivienda como Random Forest, Lasso y XGBoost para predecir precios de las viviendas usando datos del sitio web del Condado de Volusia en La Florida. Finalmente se toma un modelo mejorado el cual se emplea para optimizar el mercado inmobiliario. Se identificó que el algoritmo que tiene mejor desempeño es el XGBoost. (Jha et al., 2020). A nivel de Colombia se han realizado implementación de modelos de predicción de precios de vivienda en el municipio de Rionegro Antioquia donde se empleó la técnica de web scraping para obtener datos reales de portales inmobiliarios y se implementaron modelos de predicción de precios de vivienda donde se identificó que variables como: área construida, número de baños, número de parqueaderos y estrato entre otros son importantes para determinar el precio de una vivienda en este municipio. (Grajales, 2019). En el año 2022 se implementó un modelo para predecir precios de viviendas en Bogotá donde se identificó la necesidad de desarrollar herramientas que permitan tomar decisiones efectivas a los agentes inmobiliarios, compradores y vendedores. Se empleó la técnica de web scraping para obtener datos de los portales inmobiliarios concluyendo

que el modelo que mejor se ajusta al estudio es el modelo Light Gradiente Boosting, el cual se sometió a entrenamiento y testeo dando como resultado un error MAPE del 15.58%. En este caso se empleó este estudio para ofrecer a los posibles inversores del sector de finca raíz en Bogotá una herramienta clave para la toma de decisiones.(Nieto, 2022).

A nivel colombiano se encontró la implementación de modelos de predicción de precios en el departamento de Antioquia concretamente en el Valle de San Nicolas donde se buscó explorar las mejores técnicas de aprendizaje supervisado para predecir los precios de las viviendas tomando un conjunto de datos inicial de 2481 registros extraídos del portal de finca raíz. Inicialmente se analizaron las variables a incluir en el conjunto de datos, se realizó un análisis descriptivo previo que permitió un enfoque más adecuado a las necesidades del mercado inmobiliario. Los algoritmos empleados fueron técnicas de ensamble y redes neuronales, regresión lineal y regresión de Ridge. (Soto y David, 2021). En general se puede reconocer que el tema de predicción de precios de vivienda es un tema actual y de gran interés para muchas ciudades y países. Los estudios anteriores muestran como empleando Machine Learning y utilizando específicamente modelos de predicción de precios como: regresión lineal, regresión múltiple, Random Forest, XGBoost, Lasso y Ridge entre otras se pueden obtener resultados deseados para predecir de forma eficiente el precio de los inmuebles en distintas ciudades. Por esta razón en este proyecto se emplearán distintos modelos para predecir el precio de los inmuebles y se determinará cuales tienen mejor desempeño para alcanzar el objetivo. Otro aspecto importante para señalar es que en el estado del arte no se encontraron estudios que demuestren la implementación de modelos de predicción de precios de vivienda en la vereda Canavita del municipio de Tocancipá el cual por su creciente demanda de compra y venta de viviendas necesita un modelo para apoyar la toma de decisiones.

Marco Contextual

Tocancipá es un municipio perteneciente al departamento de Cundinamarca, Colombia. Se encuentra ubicado a 47 kilómetros al norte de Bogotá, sobre la autopista Norte. Como se observa en la figura 1 Tocancipá imita al norte con el municipio de Nemocón; al sur con Sopó y Guasca; al oriente con Gachancipá y Guatavita; y al occidente con Sopó y Zipaquirá. (Rodriguez, 2024)

Figura 1

Límites del Municipio de Tocancipá Cundinamarca

Nota. Tomado de la Página Web de la Alcaldía Municipal de Tocancipá 2011

Como se evidencia en la figura 1 el Municipio de Tocancipá cuenta con las veredas: Tibito, Verganzo, La Fuente, El Porvenir, La Esmeralda y la vereda Canavita. En el año 2014, el municipio contaba con una población de 31146 habitantes. Sin embargo, debido al crecimiento poblacional sostenido en la última década, se proyecta que para el año 2025 la población

alcanzará los 49642 habitantes, según estimaciones del Departamento Administrativo Nacional de Estadística (DANE). Este incremento demográfico se ha visto favorecido por las políticas implementadas por la administración municipal, orientadas a atraer empresas mediante la reducción de impuestos y otros incentivos para la instalación de plantas de producción en el territorio. (Rodriguez, 2024). El cambio territorial del municipio ha estado influenciado por el proceso de metropolización del Distrito Capital de Bogotá. Entre los principales efectos de esta transformación se destacan la migración poblacional e industrial hacia municipios aledaños, como Tocancipá, reconocida hoy como la **"**capital industrial de la Sabana**"**. Asimismo, las políticas públicas de vivienda impulsadas por los últimos gobiernos han sido fundamentales en la evolución urbanística del municipio, orientando las dinámicas urbanas hacia la producción de vivienda de interés social (VIS).(Rodriguez, 2024). En las últimas décadas, Tocancipá ha experimentado transformaciones significativas en la vocación de su suelo en especial la vereda Canavita, donde es importante tener en cuenta que el desarrollo urbano ha crecido de forma exponencial, y la vivienda de interés social ha asumido un papel relevante, impulsada por la expansión de los sectores industrial y productivo característicos del municipio. La acción estatal ha configurado un marco normativo e institucional que ha promovido el crecimiento residencial, consolidando a Tocancipá como un polo de desarrollo dentro de la región de Sabana Centro. (Rodriguez, 2024). Esta región se destaca como una de las zonas económicas más activas del país. La conexión entre Tocancipá y la ciudad de Bogotá ha contribuido significativamente a su desarrollo económico, atrayendo empresas de diversos sectores y reforzando su papel como centro industrial estratégico. La vereda Canavita presenta un alto potencial de crecimiento económico, respaldado por su ubicación estratégica, infraestructura moderna y mano de obra calificada. (Rodriguez, 2024). La economía de Tocancipá es diversificada, con una participación

destacada de los sectores de servicios, industria y construcción. Según datos del DANE (2021), el valor agregado bruto (VAB) del municipio fue de 4,2 billones de pesos, equivalente al 8 % del VAB del departamento de Cundinamarca, ocupando el segundo lugar después de Soacha. Este valor resulta significativo considerando que Tocancipá es el decimosegundo municipio más poblado del departamento. (Rodriguez, 2024). El sector de servicios representa aproximadamente el 65 % del valor agregado, destacándose los servicios financieros y de seguros, el comercio, y los servicios sociales, comunales y personales. Le sigue el sector industrial, con una participación del 20 % del VAB, en el cual sobresalen la industria manufacturera, la construcción y la explotación de minas y canteras. Finalmente, el sector agropecuario aporta el 15 %, con predominio de las actividades agrícolas y ganaderas.(Rodriguez, 2024). El crecimiento poblacional también ha generado nuevos desafíos, como la ampliación de infraestructura y servicios públicos; sin embargo, ha traído consigo oportunidades de desarrollo económico y generación de empleo. Según información del Sisbén (2020), de un total de 10215 hogares encuestados, 6095 viven en arriendo, 1421 en vivienda propia en pago, 1964 en vivienda totalmente pagada y 735 en otras condiciones habitacionales. (Rodriguez, 2024). Es importante resaltar el cambio significativo en el uso del suelo en Tocancipá, que ha pasado de una vocación predominantemente agropecuaria a un panorama dominado por las actividades industriales y residenciales. Esta transformación ha sido impulsada por procesos de urbanización, crecimiento económico e industrialización. Actualmente, la producción de vivienda de interés social constituye la principal oferta inmobiliaria del municipio, destacándose la vereda Canavita como la principal vereda del municipio en auge de construcciones inmobiliarias. Aunque en Bogotá se construyen más unidades habitacionales, la proporción de vivienda de interés social en Tocancipá es considerablemente superior.

(Rodriguez, 2024). Dentro de este contexto, se identifican variables relevantes que inciden en el precio de las viviendas, tales como la ubicación del inmueble, el número de habitaciones y baños, así como las condiciones establecidas por el Plan de Ordenamiento Territorial (POT), entre ellas si el predio se encuentra en zona de reserva o no. No obstante, con una simple búsqueda en páginas web inmobiliarias no es posible establecer con precisión el efecto de cada variable sobre el valor del inmueble. Por esta razón, el presente proyecto propone incorporar dichas variables en un modelo predictivo de precios de vivienda, con el propósito de generar estimaciones confiables y útiles para apoyar la toma de decisiones en el mercado inmobiliario rural del municipio principalmente en la vereda Canavita. Es importante tener en cuenta que en la vereda Canavita se localizan la gran mayoría de proyectos de vivienda del municipio de Tocancipá gracias a su cercanía al casco urbano, debido a la construcción y diseño de vías de acceso que facilitan el transporte, y el acceso a sitios de interés como: colegios, centros comerciales, zonas verdes entre otras características hacen que esta vereda sea principal foco de atención por parte de las firmas constructoras para edificar en esta vereda sus proyectos.

Marco Teórico

Antecedentes

El mercado inmobiliario en Colombia representa uno de los sectores más relevantes para la inversión de recursos, debido a su influencia en el crecimiento económico y urbano del país. Diversos estudios han explorado el uso de técnicas de Machine Learning para la predicción de precios de vivienda, con el fin de mejorar la precisión y eficiencia en la estimación del valor de los inmuebles. Uno de estos estudios fue desarrollado en el departamento de Antioquia, en el municipio de Rionegro, donde se implementó la metodología de web scraping para recopilar información sobre propiedades y comparar sus características. Entre las variables empleadas se

incluyeron el área privada, el área construida, el tipo de vivienda y el estrato socioeconómico. Los resultados mostraron que los modelos de regresión lineal y árboles de decisión ofrecieron los mejores niveles de ajuste y predicción. (Grajales, 2019). De manera similar, en Bogotá se llevó a cabo un estudio en 2022 que utilizó técnicas de aprendizaje automático para la predicción de precios de vivienda. Los datos fueron obtenidos mediante web scraping directamente desde portales inmobiliarios y, posteriormente, se analizaron las variables relevantes para construir distintos modelos. El modelo que presentó el mejor desempeño fue el Light Gradient Boosting Machine (LightGBM), alcanzando un error medio absoluto porcentual (MAPE) del 15,58 %. (Nieto, 2022). A partir de estos antecedentes, se evidencia que los modelos basados en árboles de decisión y sus variantes son los que han mostrado mejores resultados en el contexto colombiano. Estos estudios constituyen una base de referencia para el presente proyecto, enfocado en desarrollar un modelo similar adaptado a las condiciones del municipio de Tocancipá, en particular para la vereda Canavita, donde los datos inmobiliarios presentan un comportamiento más disperso y heterogéneo.

Machine Learning (Aprendizaje Automático)

El aprendizaje automático es una rama de la inteligencia artificial que permite a los sistemas aprender patrones a partir de los datos sin necesidad de programación explícita. Su objetivo es desarrollar modelos capaces de predecir o tomar decisiones basadas en información histórica. En lugar de programar reglas específicas, se entrena un algoritmo con una base de datos, de manera que el modelo aprenda a partir de la experiencia y pueda generar resultados o predicciones precisas. (García, 2021). El aprendizaje supervisado está compuesto por dos subcampos: la clasificación y la regresión. Mientras los modelos de clasificación permiten

categorizar objetos en clases conocidas, el análisis de regresión se utiliza para predecir el resultado continuo de una variable dependiente. (Mirjalili y Raschka, 2020).

Aprendizaje Supervisado

El aprendizaje supervisado es un tipo de aprendizaje automático en el que el modelo recibe datos etiquetados, es decir, ejemplos con una entrada conocida (X) y una salida esperada (Y). Su propósito es encontrar una función que relacione ambas variables, permitiendo predecir la salida correspondiente a nuevas entradas. (García, 2021). El sistema analiza el conjunto de datos y adquiere conocimiento de forma secuencial conforme estos se encuentran disponibles. De esta manera, compara los parámetros de nuevos datos con los parámetros de datos ya existentes, ajustando su estructura para mejorar su capacidad predictiva. (Dueñas, 2020). Este tipo de aprendizaje es especialmente útil en problemas donde el objetivo es estimar valores numéricos, como los precios de vivienda.

Aprendizaje por Refuerzo

El aprendizaje por refuerzo se basa en la retroalimentación que recibe un modelo al interactuar con su entorno. El algoritmo aprende mediante un proceso de ensayo y error, ajustando sus acciones en función de las recompensas o penalizaciones obtenidas. Este enfoque busca maximizar el rendimiento del modelo a lo largo del tiempo. (García, 2021). La ley del efecto, denominada refuerzo, sostiene que el aprendizaje se da en función del refuerzo, es decir, que distintas respuestas ante una situación se asocian con mayor firmeza cuando conllevan una recompensa. Se realizaron proyectos para ciudades inteligentes donde se entrena una agente con el algoritmo para recoger y dejar pasajeros y el segundo entorno optimiza la recolección de residuos, se obtuvieron resultados interesantes ya que se mostró reducción de la ruta en ambos casos. (Escribá Pina, 2021). Aunque este enfoque no es el principal aplicado en el presente

proyecto, resulta importante comprender su base teórica dentro del campo general del Machine Learning.

Algoritmos de Machine Learning

Los algoritmos de aprendizaje automático son conjuntos de operaciones lógicas y matemáticas diseñadas para resolver un problema específico a partir de los datos. En Machine Learning, los algoritmos son esenciales porque determinan la forma en que el modelo aprende, se adapta y toma decisiones. (García, 2021). El Machine Learning es una rama de la inteligencia artificial que se encarga de generar algoritmos que tienen la capacidad de aprender sin ser programados de manera explícita. Las técnicas existentes en la actualidad son evolución de técnicas conocidas las cuales mejoran la competitividad empresarial. (A. Alfaro y Ospina, 2021). A continuación, se describen algunos de los algoritmos más utilizados en la predicción de precios de vivienda:

Regresión Lineal

La regresión lineal es una técnica estadística que permite determinar la relación entre una variable dependiente y una o más variables independientes, con el fin de predecir el comportamiento de la primera en función de la segunda. Este método se utiliza con frecuencia para modelar fenómenos aleatorios. El algoritmo encuentra patrones de datos y los clasifica en grupos, luego compara los nuevos datos y los clasifica en nuevos grupos. (Roque, 2021). Se tomaron trece atributos de un cultivo de vino a los cuales se les realizó una discriminación y luego se agruparon en un conjunto denominado químicos, el óptimo dentro de este grupo fueron los fenoles totales de acuerdo con la regresión lineal. (Torres y Cardenas, 2021).

Árboles de Decisión

Los árboles de decisión son estructuras jerárquicas que dividen los datos en grupos basados en atributos específicos. Cada nodo representa una pregunta sobre una característica, y cada hoja constituye una decisión o predicción final. Los árboles de decisión se consideran pruebas estadísticas de predicción cuya función es interpretar resultados a partir de observaciones. (M. A. Díaz et al., 2021). Además, representan un método potente para el análisis de datos, tanto para efectos de segmentación como para el establecimiento de tipologías. (Alaminos, 2022). Su capacidad para manejar variables categóricas y numéricas los convierte en modelos muy útiles en la valoración inmobiliaria.

Redes Neuronales

Las redes neuronales son modelos inspirados en el funcionamiento del cerebro humano, capaces de aprender y ajustarse mediante iteraciones sucesivas. Son especialmente útiles para detectar patrones complejos en grandes volúmenes de datos. Al realizar estudios donde se integran los grafos para optimizar las redes neuronales se obtuvo una mejora en la eficiencia y la precisión en aplicaciones como procesamiento de lenguaje natural y análisis de imágenes. (Cedeño, 2023). Se realizó un estudio en el cual se emplearon redes neuronales para abordar sistemas complejos para lograrlo se describieron los rasgos distintivos de los sistemas complejos luego se analizan las teorías desarrolladas para abordar estos temas y luego se emplean simuladores computacionales basados en inteligencia artificial para estudiar el plegamiento de las proteínas obteniendo resultados interesantes. (Rubio, 2022).

Máquinas de Soporte Vectorial para Regresión (SVR)

Las máquinas de soporte vectorial para regresión (SVR) son una extensión del algoritmo Support Vector Machine (SVM) orientada a problemas de regresión. Su objetivo es encontrar

una función que aproxime los datos con un margen de error mínimo. Este método pertenece al conjunto de técnicas supervisadas de aprendizaje destinadas tanto a la clasificación como a la regresión. (Porras, 2024). Las máquinas de soporte vectorial facilitan la selección de características y ofrecen grandes ventajas al reducir la complejidad de los problemas, disminuir los costos computacionales y evitar el sobreajuste que puede distorsionar los resultados. (Valero, 2023).

Metodologia CRISP-DM (Cross Industry Standard Process for Data Mining)

Es el marco de trabajo más usado para desarrollar proyectos de minería de datos y ciencia de datos, especialmente cuando se busca aplicar técnicas analíticas o de aprendizaje automático para resolver problemas reales. Esta metodología abarca etapas como: Identificar el problema, determinar objetivos, evaluar situación actual, comprender los datos, prepara los datos, modelado, evaluación del modelo e implementación del modelo. (Espinosa, 2020). La metodología CRISP-DM proporciona una idea clara acerca del ciclo de vida de un proyecto de minería de datos dividiendo el contexto en seis fases las cuales realizan una interacción entre ellas de forma iterativa durante el desarrollo del proyecto. En estas etapas las más importantes son el análisis del problema el cual nos otorga una visión clara, el análisis de los datos, la preparación de los datos, el modelado, la evaluación y el seguimiento, todo con el propósito de planificar, dirigir y dar seguimiento al proyecto. (Sánchez y Pérez, 2023).

A partir de la revisión teórica realizada, se puede concluir que las técnicas de Machine Learning representan una herramienta sólida y flexible para la estimación de precios de vivienda. En el contexto colombiano, los modelos basados en árboles de decisión y sus variantes han demostrado un desempeño destacado, lo que justifica su aplicación en este estudio. Asimismo, la metodología CRISP-DM permitirá garantizar un desarrollo estructurado, replicable y orientado a

resultados, fortaleciendo la calidad y aplicabilidad del modelo predictivo propuesto para la vereda Canavita del municipio de Tocancipá.

Marco Conceptual

El presente marco conceptual tiene como propósito definir los principales conceptos relacionados con la implementación de un modelo de predicción de precios de vivienda en la vereda Canavita, municipio de Tocancipá. Estas definiciones constituyen la base teórica que sustenta el desarrollo metodológico del proyecto, orientado desde un enfoque analítico y predictivo apoyado en técnicas de Machine Learning para el estudio del mercado inmobiliario rural.

Mercado Inmobiliario

El mercado inmobiliario comprende el conjunto de transacciones, actores y dinámicas relacionadas con la producción, comercialización, uso y financiación de bienes inmuebles. A diferencia de otros mercados, se caracteriza porque el bien transado es único e inmóvil, lo que genera un alto grado de heterogeneidad e imprevisibilidad en la formación de precios. (Macías y Osorio, 2020). El mercado inmobiliario no solo incluye las transacciones relacionadas con la oferta y la demanda de los bienes inmuebles, sino que está acompañado por unos factores como: promoción inmobiliaria, inversión en propiedades, financiación tanto para la producción como para la compra y el arriendo. (Botero, 2021). En el contexto de la vereda Canavita, el mercado inmobiliario está condicionado por elementos como la ubicación geográfica, el área construida, el estrato socioeconómico, la valorización del entorno y la cercanía con zonas industriales, factores que inciden directamente en la formación del precio de la vivienda. Este proyecto adopta un enfoque analítico que busca cuantificar y modelar dichos factores mediante herramientas de aprendizaje automático.

Vivienda

De acuerdo con la Constitución Política de Colombia (1991), la vivienda constituye un derecho fundamental que implica no solo el acceso a un techo, sino también a condiciones de seguridad, paz y dignidad. Para considerarse adecuada, debe cumplir criterios de seguridad jurídica, disponibilidad de servicios básicos, habitabilidad, asequibilidad y localización que garantice acceso a equipamientos y servicios sociales. (Jiménez Oliveros y Aguiar Hernández, 2024). En Colombia la vivienda se constituye como un derecho constitucional y como política pública se establece la vivienda de interés prioritario para la población vulnerable cuyo valor máximo alcanza los 90 salarios mínimos. (Gómez et al., 2024).

Bien Inmueble

Según el código civil colombiano un bien inmueble se entiende como aquellos entes de edificación que ocupan un espacio en el mundo y se aprecian por los sentidos, que están dentro del patrimonio de una persona y son susceptibles a ser valorados económicamente. (Toledo, 2023). Un bien inmueble es una propiedad que no puede movilizarse, se caracteriza por ser un bien que está pegado al suelo ya sea de forma permanente o adherido a el. Como ejemplo una casa, edificio o local.(Vilca, 2022). En este proyecto, los bienes inmuebles se constituyen en la unidad básica de análisis para la modelación predictiva de precios.

Tasación Comercial

La tasación comercial es el proceso mediante el cual se determina el valor de mercado de un inmueble en un momento específico, a partir de criterios económicos, técnicos y legales sustentados en el comportamiento del mercado. (Vicarte y Mayo, 2023). Este proceso requiere conocimientos interdisciplinarios en economía, contabilidad y normativa urbanística (Gutierrez et al., 2022). Desde la perspectiva analítica del presente estudio, la tasación comercial constituye

una referencia empírica que permite calibrar los modelos de predicción y validar su precisión frente a estimaciones tradicionales.

Valor Catastral

El valor catastral corresponde a la valoración administrativa de un inmueble, determinada con base en normativas y metodologías oficiales que definen sus características físicas, jurídicas y económicas. Este valor tiene efectos fiscales, especialmente en el cálculo del impuesto predial. (Ramos, 2023). De acuerdo con la Ley 2294 de 2023, el Instituto Geográfico Agustín Codazzi (IGAC) debe adoptar modelos de actualización masiva de avalúos catastrales, con el propósito de reflejar la realidad económica de los predios y reducir rezagos en la información. (Castiblanco y Velandia, 2024).

Valor Metro Cuadrado Integral

El valor metro cuadrado integral es un indicador que expresa el valor comercial de un inmueble dividido entre su área construida, expresado en pesos colombianos. (Escobar, 2024). Este indicador resulta fundamental en los modelos predictivos, ya que permite normalizar los precios y establecer comparaciones entre diferentes propiedades, independientemente de su tamaño o características constructivas.

Avalúo Comercial Rural

El avalúo comercial rural corresponde a la estimación del valor de mercado de un predio ubicado fuera del perímetro urbano, considerando sus características físicas, el uso del suelo y su potencial de desarrollo. (Pineda y Sosa, 2024). Este tipo de avalúo integra tanto el valor del terreno como el de las edificaciones y constituye una referencia clave para el análisis de precios en zonas rurales como la vereda Canavita.

Impuesto Predial

El impuesto predial es un gravamen que recae sobre la propiedad de bienes inmuebles, rurales o urbanos, cuya base gravable se determina a partir del avalúo catastral. (Moncada et al., 2022). Este impuesto representa una de las principales fuentes de ingresos para los municipios, al tiempo que refleja la dinámica económica del mercado inmobiliario.(C. Díaz, 2022).

Municipio

El municipio, conforme a la Constitución Política de Colombia, es la entidad territorial básica del Estado social de derecho, con autonomía administrativa, política y fiscal para la gestión de sus propios intereses. (Charria, 2022). Su estructura comprende una organización central y descentralizada, orientada a la prestación de servicios y al ordenamiento del territorio. En el presente estudio, el municipio de Tocancipá constituye la unidad territorial dentro de la cual se contextualiza la vereda Canavita

Vereda

La vereda es la unidad geográfica rural conformada por un conjunto de predios delimitados por accidentes naturales o vías, donde se establecen comunidades dedicadas principalmente a actividades agropecuarias. (Zapata, 2025). En la vereda Canavita, la expansión urbana y la cercanía a zonas industriales han generado transformaciones en el uso del suelo y en la dinámica de precios de la vivienda, lo que justifica la aplicación de técnicas de análisis predictivo. El enfoque del proyecto es, por tanto, analítico-predictivo con aplicación territorial, donde la ciencia de datos permite transformar información inmobiliaria en conocimiento útil para la toma de decisiones de inversión y gestión en la vereda Canavita ubicada en el municipio de Tocancipá. Es importante destacar la evolución que han tenido las zonas urbanas de este municipio en los últimos años por lo cual estas zonas están en desarrollo constante.

Marco Normativo

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 y se establecen las entidades de certificación y se dictan otras disposiciones.

Ley 1581 de 2012

Protección de datos personales. Por la cual se dictan disposiciones generales para la protección de datos personales.

Ley 1712 de 2014

Por medio de la cual se crea la ley de transparencia y del derecho a la información pública nacional. Además, se dictan otras disposiciones, esta ley tiene como objetivo regular el derecho a la información pública, los procedimientos para el ejercicio y la garantía del derecho y las excepciones a la publicidad de la información.

Metodología

El presente proyecto se desarrolló bajo un enfoque cuantitativo, dado que busca implementar un modelo predictivo basado en el análisis numérico de variables que inciden en el precio de los inmuebles. Se emplearon técnicas estadísticas y de aprendizaje automático (Machine Learning) para analizar los datos y construir el modelo. La metodología seleccionada fue CRISP-DM (Cross Industry Standard Process for Data Mining), la cual establece un proceso estructurado y flexible para el desarrollo de proyectos de minería de datos. Esta metodología comprende seis fases: comprensión del negocio, comprensión de los datos, preparación de los datos, modelado, evaluación e implementación. A continuación, se describe el desarrollo de cada fase adaptada al contexto del proyecto.

Comprensión del Negocio

En esta etapa se definió el propósito principal del estudio: implementar un modelo de predicción de precios de vivienda para la vereda Canavita del municipio de Tocancipá, con el fin de apoyar a posibles inversores en la toma de decisiones de compra y venta de finca raíz.
Se identificó la necesidad de contar con información actualizada y confiable sobre el mercado inmobiliario de la zona, así como la pertinencia de aplicar modelos predictivos que permitieran estimar el valor de los inmuebles con base en sus características físicas y de ubicación.

Comprensión de los Datos

Se realizó un curso de web scraping ofrecido por FreeCodeCamp en la plataforma YouTube, mediante el cual se adquirieron conocimientos sobre el uso de librerías de Python y la interacción con páginas web mediante APIs y estructuras HTML. Este curso aportó los conocimientos necesarios para realizar web scraping en las respectivas plataformas inmobiliarias.

Figura 2

Captura de Pantalla que Muestra Entorno de Canal Free Code Camp en YouTube

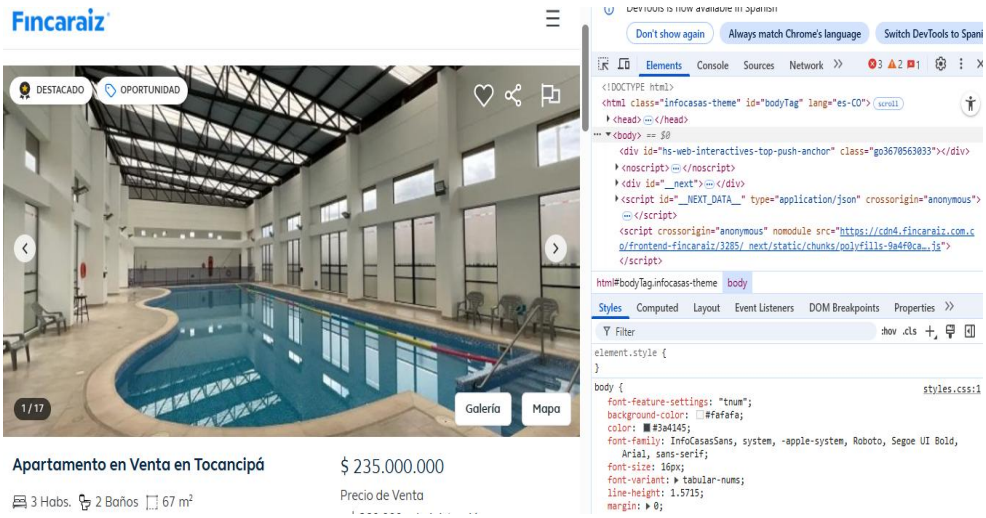
Nota. Tomado del Canal de YouTube FreeCodeCamp

Dado que no existía un conjunto de datos previamente disponible para la zona de estudio, se optó por obtener la información mediante técnicas de web scraping desde portales inmobiliarios de alta cobertura en Colombia: FincaRaiz, MetroCuadrado y CienCuadras. Como

se observa en la figura 3 estos portales tienen disponible información de inmuebles para la venta y se convirtieron en fuente de extracción de los datos.

Figura 3

Captura de Pantalla del Entorno de la Página Web de FincaRaiz Colombia



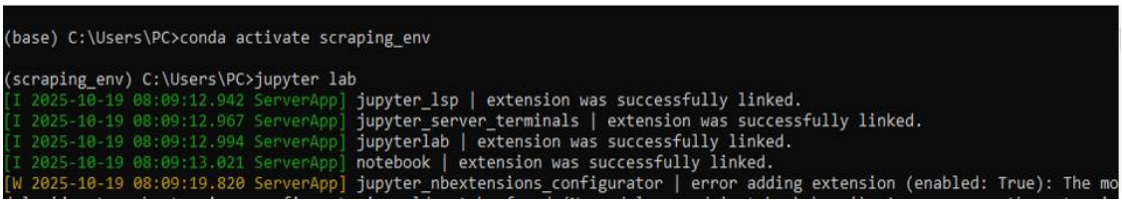
Nota. Tomado de la Página Web de Finca Raíz

Preparación del Entorno de Trabajo

Para la implementación técnica se utilizó Anaconda Jupyter Lab, creando un entorno denominado scraping_env, configurado con Python 3.10. Como se observa en la figura 4.

Figura 4

Captura de Pantalla del Entorno de Anaconda Promt en Windows 10



Se instalaron las librerías y herramientas necesarias para el proceso, entre ellas Selenium (para la automatización de navegación web) y ChromeDriver, asegurando la compatibilidad entre la versión del controlador y el navegador Google Chrome.

Extracción y Comprensión de la Información

En esta fase se accedió a los portales inmobiliarios seleccionados para identificar la estructura de sus páginas y las etiquetas html que contenían la información de interés (precio del inmueble, tipo de propiedad, área, estrato, número de habitaciones, baños, parqueaderos, antigüedad, estado, valor de administración, entre otros). Los pasos realizados fueron los siguientes:

Selección de Portales Web

Se eligieron los sitios FincaRaiz, MetroCuadrado y CienCuadras por su cobertura nacional y relevancia en el mercado.

Inspección del Código HTML

Se analizaron las etiquetas html y los elementos JavaScript donde se encontraban los datos requeridos.

Extracción de Enlaces (URL)

Dado que parte de la información se encontraba en contenido dinámico, se desarrolló código en Python para capturar únicamente los enlaces de las propiedades publicadas.

Extracción de Datos Estructurados

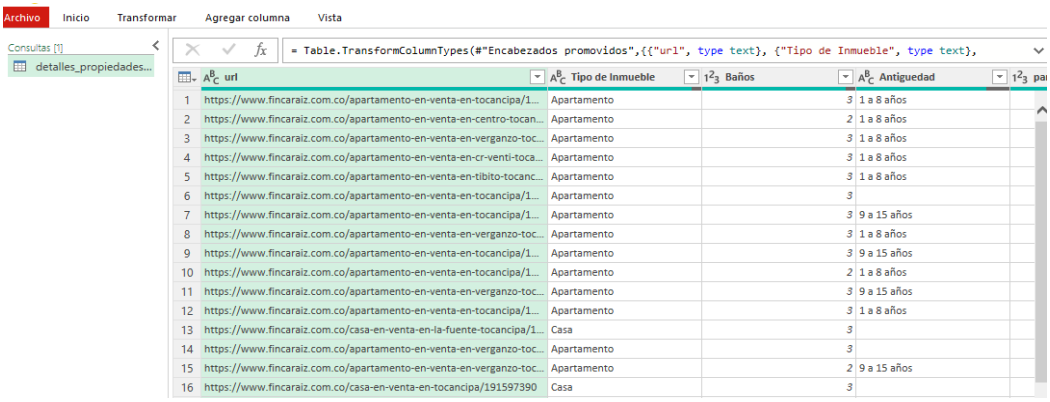
Posteriormente, se ejecutó código para recorrer cada enlace y extraer la información de las variables relevantes, generando tres archivos csv, uno por cada portal consultado. L os resultados iniciales fueron: FincaRaíz*:* 314 inmuebles., MetroCuadrado: 260 inmuebles (tras depurar duplicados), CienCuadras: 96 inmuebles.

Verificación y Limpieza de la Información

Durante la revisión de los datos, se identificaron diferencias en la cantidad y calidad de la información entre los portales. Por ejemplo, el conjunto de FincaRaíz no incluía algunas características del inmueble (como piscina, zonas verdes, ascensor o cercanía a servicios), las cuales fueron complementadas mediante búsqueda manual como se observa en la figura 5. En el portal de MetroCuadrado se observó que inicialmente solo se habían capturado apartamentos; por tanto, se ajustó el código para incluir casas y se eliminaron enlaces duplicados. Asimismo, en el portal de CienCuadras se completaron manualmente campos como el precio de administración y algunas características faltantes.

Figura 5

Datos de FincaRaiz



	url	Tipo de Inmueble	Baños	Antiguedad
1	https://www.fincaraiz.com.co/apartamento-en-venta-en-tocancipa/1...	Apartamento	3	1 a 8 años
2	https://www.fincaraiz.com.co/apartamento-en-venta-en-centro-tocan...	Apartamento	2	1 a 8 años
3	https://www.fincaraiz.com.co/apartamento-en-venta-en-verganzo-toc...	Apartamento	3	1 a 8 años
4	https://www.fincaraiz.com.co/apartamento-en-venta-en-cr-venti-toca...	Apartamento	3	1 a 8 años
5	https://www.fincaraiz.com.co/apartamento-en-venta-en-tibito-tocanc...	Apartamento	3	1 a 8 años
6	https://www.fincaraiz.com.co/apartamento-en-venta-en-tocancipa/1...	Apartamento	3	
7	https://www.fincaraiz.com.co/apartamento-en-venta-en-tocancipa/1...	Apartamento	3	9 a 15 años
8	https://www.fincaraiz.com.co/apartamento-en-venta-en-verganzo-toc...	Apartamento	3	1 a 8 años
9	https://www.fincaraiz.com.co/apartamento-en-venta-en-tocancipa/1...	Apartamento	3	9 a 15 años
10	https://www.fincaraiz.com.co/apartamento-en-venta-en-tocancipa/1...	Apartamento	2	1 a 8 años
11	https://www.fincaraiz.com.co/apartamento-en-venta-en-verganzo-toc...	Apartamento	3	9 a 15 años
12	https://www.fincaraiz.com.co/apartamento-en-venta-en-tocancipa/1...	Apartamento	3	1 a 8 años
13	https://www.fincaraiz.com.co/casa-en-venta-en-la-fuente-tocancipa/1...	Casa	3	
14	https://www.fincaraiz.com.co/apartamento-en-venta-en-verganzo-toc...	Apartamento	3	
15	https://www.fincaraiz.com.co/apartamento-en-venta-en-verganzo-toc...	Apartamento	2	9 a 15 años
16	https://www.fincaraiz.com.co/casa-en-venta-en-tocancipa/191597390	Casa	3	

Consolidación de la Información

Una vez verificados y corregidos los tres conjuntos de datos, se procedió a unificarlos en un único dataset consolidado. Se definió una estructura uniforme de columnas con los siguientes campos: URL, Tipo_Inmueble, Precio_Inmueble, Área_Construida_m2, Área_Privada_m2, Estrato, Habitaciones, Baños, Parqueaderos, Precio_Administración, Antigüedad_Años, Estado, Piso, Características_Inmueble. Es importante señalar que antes de unir los conjuntos de datos se

asignaron tipos de datos adecuados: float64**:** variables numéricas continuas (precio, área construida, área privada, administración), int64**:** variables enteras (estrato, habitaciones, baños, parqueaderos, antigüedad, piso), **o**bject**:** variables categóricas (tipo de inmueble, estado, características, URL).

Tabla 1

Variables Conjunto de Datos Consolidado

Variables	Descripción
URL	Enlace de la página web
Tipo_Inmueble	Si es casa o apartamento
Precio_Inmueble,	Precio de venta del inmueble
Área_Construida_ m2	Área construida del inmueble
Área_Privada_m2	Área privada del inmueble
Estrato	Estrato socio económico donde se encuentra el inmueble
Habitaciones	Número de habitaciones del inmueble en venta
Baños	Número de baños disponibles del inmueble en venta
Parqueaderos	Cantidad de parqueaderos disponibles del inmueble
Precio_Administración	Valor de pago de administración del inmueble
Antigüedad_Años	Años de antigüedad del inmueble
Estado	Nuevo: si antigüedad es menor o igual a 1 año Usado: si antigüedad es mayor a 1 año
Piso	El número de piso donde se encuentra el apartamento en venta
Características_Inmueble	Características que definen el tipo de inmueble en venta como: ascensor, tipo de cocina, zonas verdes entre otras.

Preparación de los Datos

Para garantizar la calidad y pertinencia de la información empleada en el modelo de predicción de precios de vivienda, se desarrolló un proceso riguroso de preparación y depuración del conjunto de datos, partiendo de la estructura del conjunto de datos consolidado que se observa en la figura 6. Es importante que esta preparación de los datos se basó en la guía empleada de acuerdo con la metodología en la cual se tuvieron en cuenta aspectos importantes que buscan garantizar la calidad y consistencia de los datos a la hora de implementar los respectivos modelos de predicción.

Figura 6

Tipo de Datos del Conjunto de Datos Consolidado

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 452 entries, 0 to 451
Data columns (total 14 columns):
```

Toggle output scrolling

		Non-Null Count	Dtype
0	URL	452 non-null	object
1	Tipo_Inmueble	452 non-null	object
2	Precio_Inmueble	452 non-null	float64
3	Area_Construida_m2	452 non-null	float64
4	Area_Privada_m2	452 non-null	float64
5	Estrato	452 non-null	int64
6	Habitaciones	452 non-null	int64
7	Baños	452 non-null	int64
8	Parqueaderos	452 non-null	int64
9	Precio_Administracion	452 non-null	float64
10	Antiguedad_Años	452 non-null	int64
11	Estado	452 non-null	object
12	Piso	452 non-null	int64
13	Caracteristicas_Inmuebles	452 non-null	object

```
dtypes: float64(4), int64(6), object(4)
memory usage: 49.6+ KB
```

Análisis de Coherencia y Validación de Datos

Se verificó la coherencia interna de las variables numéricas (como Área Construida, Estrato o Precio de Administración), asegurando que los valores se encontraran dentro de rangos razonables para el mercado inmobiliario colombiano. Asimismo, se inspeccionaron los valores categóricos de columnas como Tipo_Inmueble y Estado, confirmando que presentaran únicamente categorías válidas (Apartamento, Casa y Nuevo, Usado respectivamente).

Identificación y Tratamiento de Valores Atípicos (Outliers)

Para identificar posibles valores outliers se decidió mostrar las estadísticas para las variables numéricas con el método. describe (), se dibujó un boxplot para detectar outliers, se dibujó un histograma para ver la distribución, se calculó y mostró el número de outliers detectados por IQR.

Figura 7

Estadísticas de la Variable Precio_Inmueble

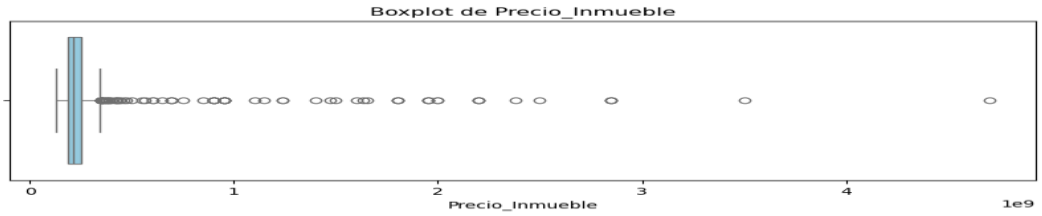
```
📊 Estadísticas:
count    4.520000e+02
mean     3.580877e+08
std      4.949695e+08
min      1.289000e+08
25%      1.837500e+08
50%      2.100000e+08
75%      2.500000e+08
max      4.700000e+09
Name: Precio_Inmueble, dtype: float64
🚨 Outliers detectados por IQR: 76
  - Límite inferior: 84375000.00
```

Como se observa en la figura 7 la variable Precio_Inmueble, correspondiente al valor comercial de los predios ubicados en la vereda Canavita del municipio de Tocancipá, presenta un total de 452 registros válidos. De acuerdo con las estadísticas descriptivas obtenidas mediante el

método describe, el precio promedio de los inmuebles es de $ 358.087.700, con una desviación estándar de $ 494.969.500, lo cual sugiere una alta dispersión respecto a la media. El valor mínimo registrado es de $ 128.900.000, mientras que el máximo alcanza los $ 4.700.000.000, evidenciando una marcada asimetría hacia la derecha y la presencia potencial de valores extremos.

Figura 8

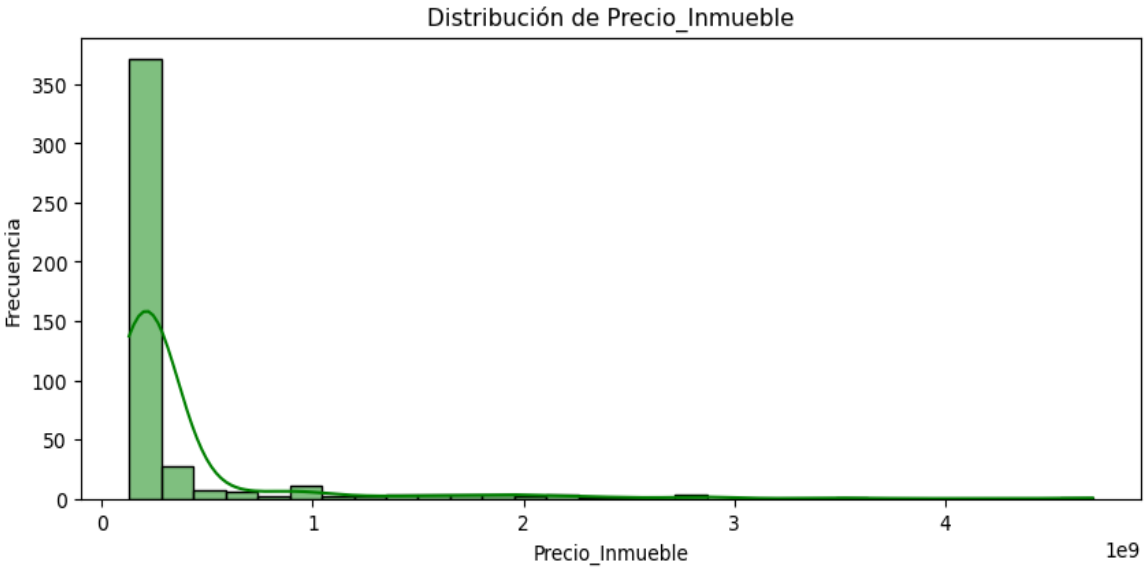
Boxplot de la Variable Precio_Inmueble



Como se observa en la figura 8 el diagrama de caja evidencia una distribución marcadamente asimétrica hacia la derecha, con una concentración central de los valores dentro del rango intercuartílico comprendido entre aproximadamente 183.8 y 250 millones de pesos. No obstante, el boxplot muestra un número considerable de valores atípicos por encima del límite superior, lo que indica la presencia de inmuebles con precios significativamente más altos que el comportamiento típico del mercado. Estos outliers coinciden con la elevada dispersión observada en la variable y sugieren la existencia de propiedades con características o extensiones excepcionales dentro de la zona analizada. Al consultar valores reales en las páginas web se verificó que estos datos son reales por lo cual se conservarlos en el conjunto de datos.

Figura 9

Distribución de la Variable Precio_Inmueble



El histograma de la figura 9 confirma una distribución sesgada positivamente, caracterizada por una alta concentración de propiedades en rangos de precio medio-bajo y una cola larga hacia valores elevados. Esta forma de distribución es habitual en mercados inmobiliarios donde una minoría de inmuebles presenta precios muy superiores al resto, lo cual incrementa la variabilidad global. En nuestro caso para esta variable se decide emplear np.log1p() debido a que los precios no están distribuidos de forma normal se evidencia que hay muchos inmuebles con precios normales y algunos con precios altos como las casas, esto genera una distribución sesgada a la derecha donde los outliers afectan el modelo al usar log se mejora el ajuste del modelo en algoritmos como regresión lineal. Esto hace que el modelo no esté tan dominado por precios de viviendas muy altos a su vez mejoran las métricas y la generalización.

Figura 10

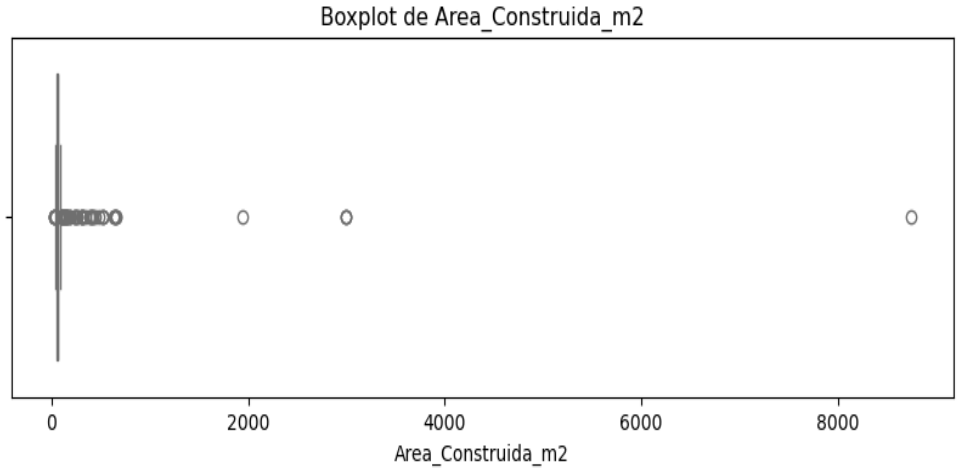
Estadísticas de la Variable Area_Construida_m2

```
========================================
📊 Estadísticas:
count     452.000000
mean      131.557876
std       487.825907
min        27.000000
25%        56.000000
50%        61.000000
75%        68.000000
max      8743.000000
Name: Area_Construida_m2, dtype: float64
🚨 Outliers detectados por IQR: 75
  - Límite inferior: 38.00
  - Límite superior: 86.00
```

Como se observa en la figura 10 la variable área construida (m²) presenta 452 registros válidos, con un promedio de 131,56 m² y una desviación estándar de 487,83 m², lo que evidencia una alta dispersión respecto al valor central. La mediana (61 m²) y los cuartiles (Q1 = 56 m², Q3 = 68 m²) muestran que la mayoría de los inmuebles poseen áreas construidas relativamente pequeñas, mientras que el valor máximo (8743 m²) indica la presencia de edificaciones atípicamente grandes, generando una distribución altamente asimétrica. Se identificaron 75 outliers mediante el método IQR, establecidos fuera del rango definido por los límites de 38 m² y 86 m², lo que confirma la existencia de unidades con áreas inusualmente elevadas respecto al conjunto principal. Se identificó que los valores elevados son reales pues corresponden al tipo de inmueble casa por lo cual se decide conservarlos en el conjunto de datos. Es importante mencionar que se realiza la verificación para garantizar la calidad de los datos.

Figura 11

Boxplot de la Variable Area_Construida_m2

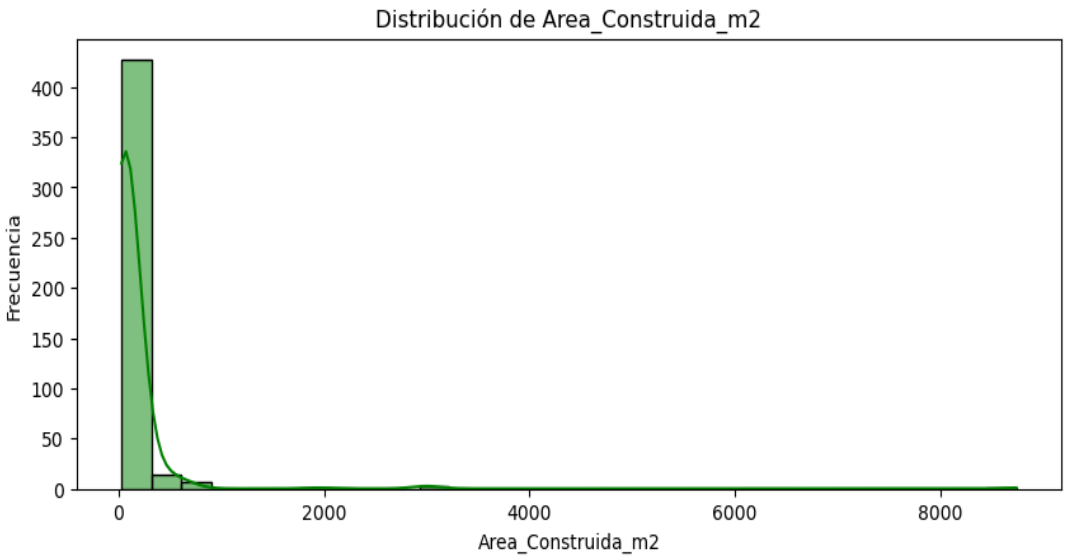


El boxplot de la figura 11 revela una distribución fuertemente sesgada hacia la derecha, con un rango intercuartílico estrecho y numerosos valores extremadamente altos situados por encima del límite superior. Esto indica que, aunque la mayoría de los inmuebles presentan áreas construidas similares, existe un número significativo de propiedades con superficies excepcionalmente amplias que distorsionan la dispersión global de la variable. Se verificó que los valores elevados son reales y corresponden al tipo de inmueble casa por lo cual se conservan estos valores en el conjunto de datos. Es importante señalar que en una zona rural como lo es la vereda Canavita el área construida para un tipo de inmueble casa es mayor comparado con el área construida para el tipo de inmueble apartamento ya que las casas en sector rural disponen de mayor espacio y por las características culturales de las familias ocupantes suele tener mayor disponibilidad para número de habitaciones y tipos de cocina más amplios ya que son inmuebles ocupados por un número mayor de personas en comparación con los apartamentos. Esta

característica que mencionamos es común en inmuebles ubicados en zonas rurales donde por lo general el área construida es extensa.

Figura 12

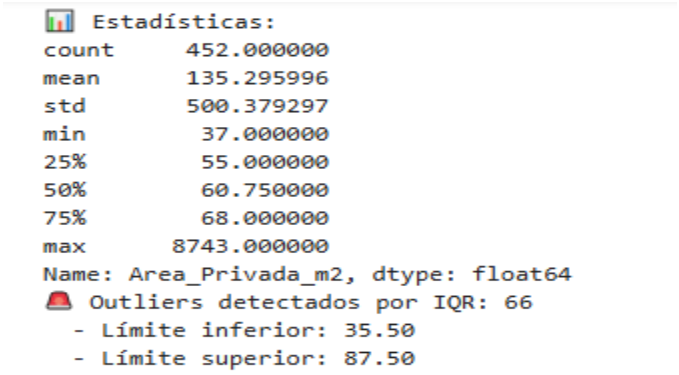
Distribución Normal de la Variable Area_Construida_m2



El histograma de la figura 12 confirma una distribución altamente asimétrica y con cola larga, donde la mayoría de las observaciones se concentra entre los 50 y 70 m². La presencia de pocos valores muy grandes genera una cola extendida, característica típica de territorios rurales o semirrurales en los que coexisten viviendas pequeñas y predios con construcciones de gran tamaño.

Figura 13

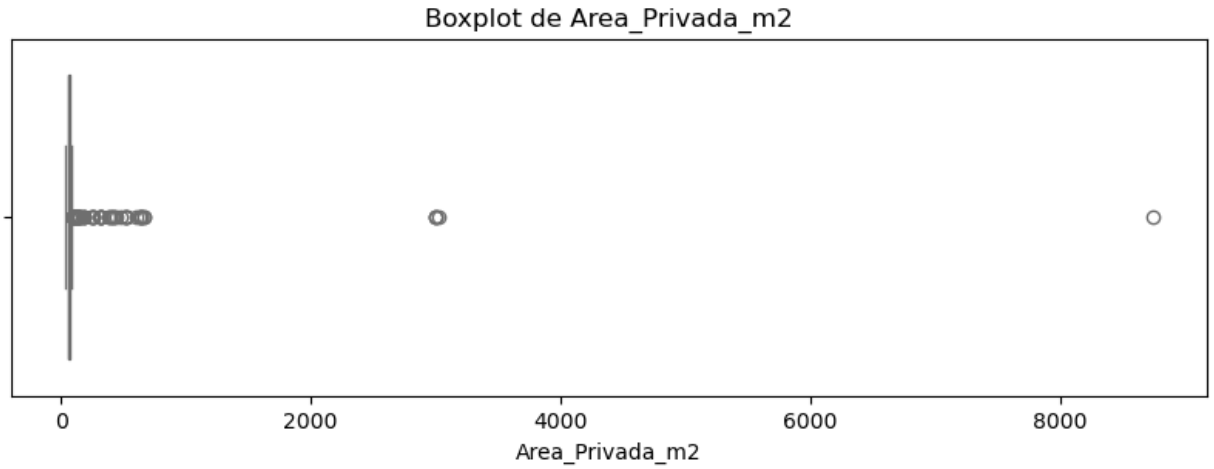
Estadísticas de la Variable Area_Privada_m2



De acuerdo con lo observado en la figura 13, la variable área privada (m²) cuenta con 452 registros válidos y presenta un promedio de 135,30 m², acompañado de una desviación estándar de 500,38 m², lo que evidencia una alta variabilidad y la presencia de valores extremos. La mediana (60,75 m²) y los cuartiles (Q1 = 55 m², Q3 = 68 m²) muestran que la mayoría de los inmuebles poseen áreas privadas reducidas y similares entre sí. Sin embargo, el valor máximo (8743 m²) indica la existencia de predios con superficies privadas extraordinariamente grandes, generando una marcada asimetría en la distribución. Mediante el método IQR se identificaron 66 outliers, ubicados fuera de los límites de 35,5 m² y 87,5 m². Al verificar la información en los respectivos portales se comprobó que son valores reales por lo cual se conservan en el conjunto de datos.

Figura 14

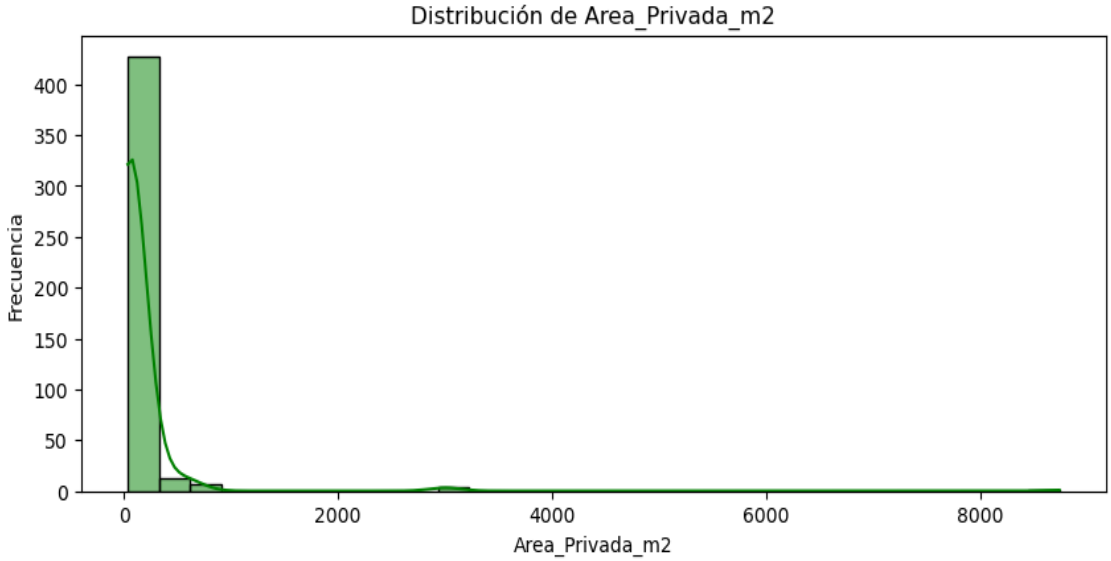
Boxplot de la Variable Area_Privada_m2



El boxplot de la figura 14 muestra una distribución claramente sesgada hacia la derecha, con un rango intercuartílico estrecho y múltiples observaciones que superan significativamente el límite superior. Estos valores extremos reflejan la coexistencia de propiedades con áreas privadas típicas y otras excepcionales que influyen notablemente en la dispersión general de la variable.

Figura 15

Distribución de la Variable Area_Privada_m2



El histograma de la figura 15 evidencia una distribución altamente asimétrica, donde la mayoría de los inmuebles se concentran entre los 50 y 70 m². La presencia de una cola larga hacia valores elevados confirma que existen pocas propiedades con áreas privadas muy superiores al promedio. Se verificaron los datos en los respectivos portales y se evidenció que son reales por lo cual se conservan en el conjunto de datos.

Figura 16

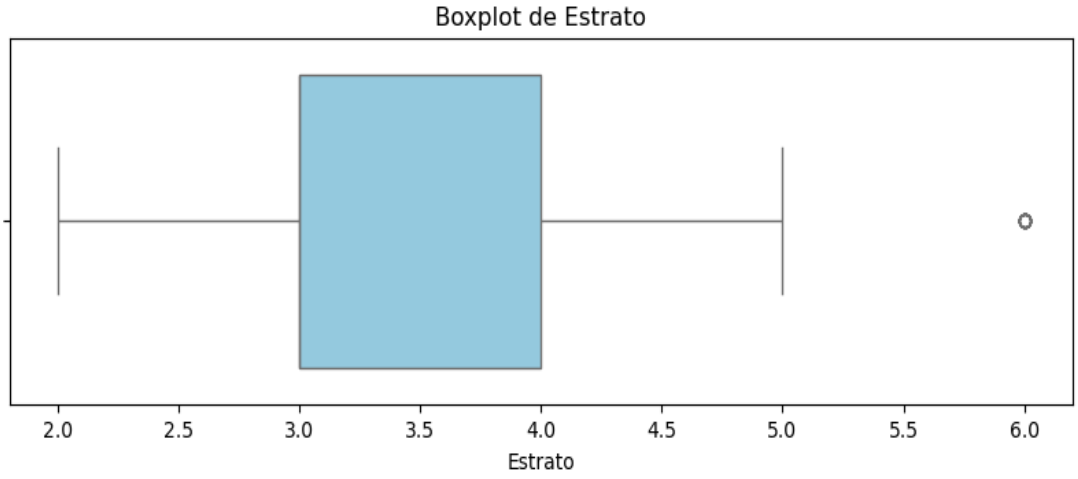
Estadísticas de la Variable Estrato

```
📊 Estadísticas:
count    452.000000
mean       3.340708
std        1.026836
min        2.000000
25%        3.000000
50%        3.000000
75%        4.000000
max        6.000000
Name: Estrato, dtype: float64
🚨 Outliers detectados por IQR: 12
  - Límite inferior: 1.50
  - Límite superior: 5.50
```

Como se evidencia en la figura 16, la variable estrato presenta 452 registros válidos y muestra un promedio de 3,34, con una desviación estándar de 1,03, lo que indica una variabilidad moderada en los niveles socioeconómicos de los inmuebles. Los cuartiles (Q1 = 3, Q2 = 3, Q3 = 4) evidencian que la mayoría de las propiedades pertenecen a estratos medios. El rango observado va desde estrato 2 hasta estrato 6. A través del método IQR se identificaron 12 valores atípicos, definidos fuera de los límites de 1,5 y 5,5, los cuales corresponden principalmente a inmuebles ubicados en estrato 6.

Figura 17

Boxplot de la Variable Estrato

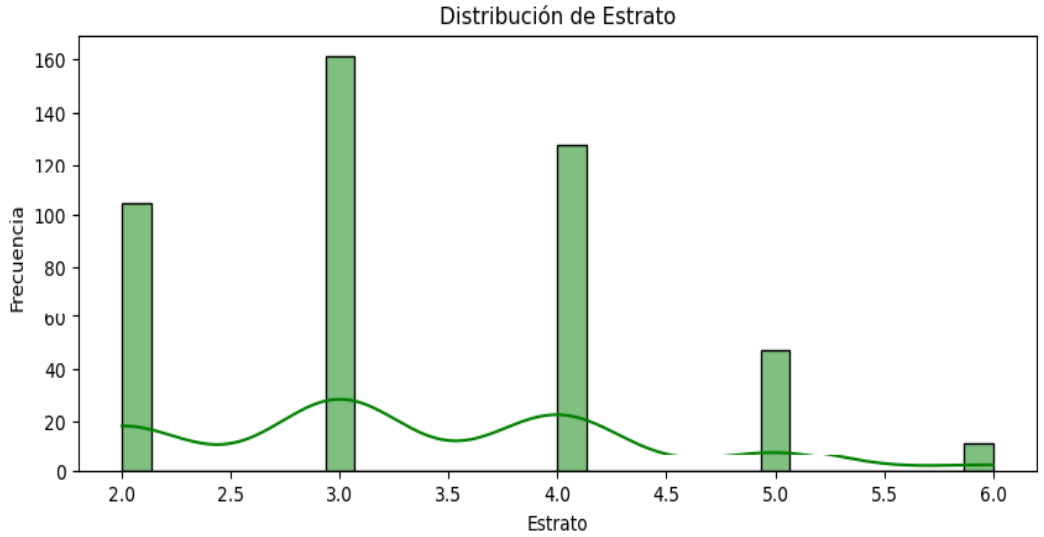


El boxplot de la figura 17 muestra una distribución ligeramente sesgada hacia los estratos superiores, con una mayor concentración de datos en estratos 3 y 4. Los valores atípicos se observan, asociados a inmuebles de estrato 6, cuya presencia resulta menos frecuente en comparación con los estratos centrales. En el boxplot de estrato se detectaron 12 valores outliers lo que explica que en zonas rurales puede haber diferencia de estratificación de acuerdo con las características de la vivienda y otros factores importantes como lo son los ingresos y otras

características de cada inmueble, es factible que personas con mayores recursos inviertan más en mejorar sus viviendas y generar valorización.

Figura 18

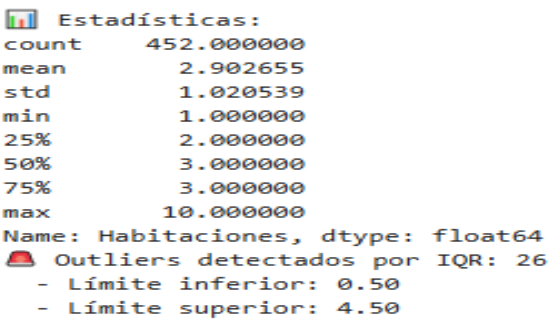
Distribución Normal de la Variable Estrato



El histograma de la figura 18 evidencia una distribución discreta y moderadamente asimétrica, con un pico claramente definido en estrato 3, seguido del estrato 4. La frecuencia de los estratos 2 y 6 es considerablemente menor, lo cual es coherente con la distribución socioeconómica típica de zonas rurales y periurbanas donde predominan los estratos medios. Al verificar la información en los portales web y comprobar que es real se decide conservar los valores en el conjunto de datos.

Figura 19

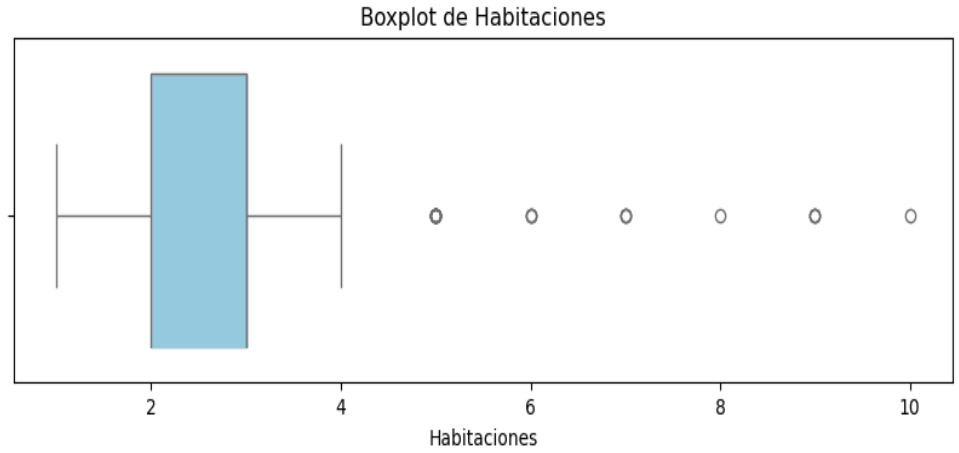
Estadísticas de la Variable Habitaciones



Como se observa en la figura 19, la variable habitaciones cuenta con 452 registros válidos y presenta un promedio de 2,90 habitaciones por inmueble, con una desviación estándar de 1,02, lo que indica una variabilidad moderada entre las propiedades analizadas. Los cuartiles (Q1 = 2, Q2 = 3, Q3 = 3) muestran que la mayoría de las viviendas poseen entre dos y tres habitaciones, lo que es consistente con construcciones típicas de uso residencial familiar. El valor máximo registrado es de 10 habitaciones, mientras que el método IQR permitió identificar 26 outliers, ubicados fuera de los límites de 0,5 y 4,5, asociados a inmuebles con un número atípicamente alto de habitaciones. Es importante señalar que después de verificar los datos en los respectivos portales se comprobó que son valores reales por lo cual se conservan en el conjunto de datos.

Figura 20

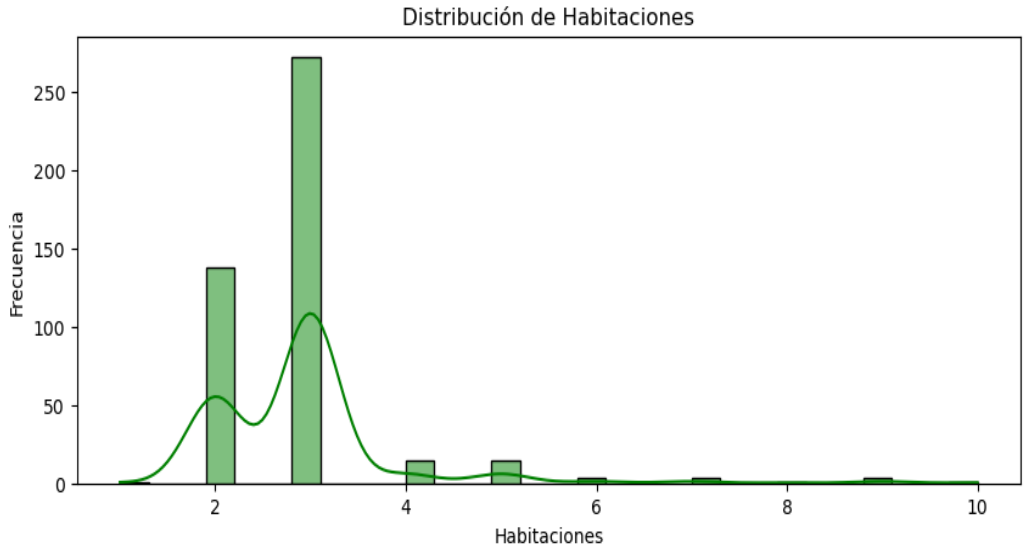
Boxplot de la Variable Habitaciones



El boxplot de la figura 20, evidencia una distribución relativamente concentrada en torno a los valores típicos (2 y 3 habitaciones), con una ligera asimetría hacia la derecha. Los outliers se observan principalmente en la parte superior del diagrama, correspondientes a propiedades con cinco o más habitaciones, las cuales representan casos poco frecuentes dentro del conjunto de datos. Se decide conservar los valores en el conjunto de datos ya que se verificó que son reales.

Figura 21

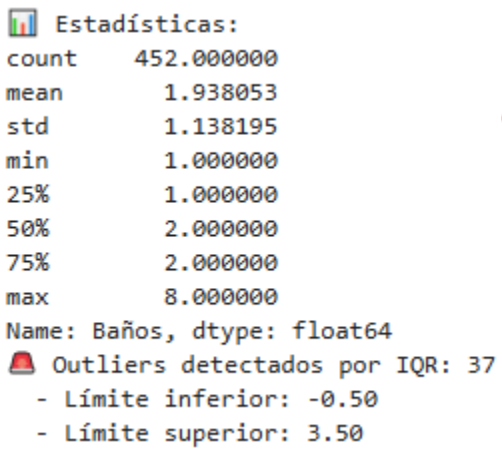
Distribución Normal de la Variable Habitaciones



El histograma de la figura 21, muestra una distribución discreta y moderadamente sesgada hacia valores mayores, con una marcada concentración en tres habitaciones, seguida de dos habitaciones. Las frecuencias disminuyen progresivamente para valores superiores, confirmando que las viviendas con muchas habitaciones constituyen excepciones dentro del mercado analizado. Es importante señalar que los tipos de inmueble casa en zonas rurales tienen un mayor número de habitaciones disponibles ya que por lo general este tipo de inmueble es ocupado por familias que tienen mayor número de habitantes y las parejas suelen tener más de un hijo. Al realizar verificación en los portales inmobiliarios se comprobó que son valores reales por lo cual se decide conservarlos en el conjunto de datos ya que son una muestra de la característica del tipo de inmueble analizado y que muestra una diferencia con respecto al otro tipo de inmueble analizado que es el apartamento donde por lo general el número de habitaciones es más reducido.

Figura 22

Estadísticas de la Variable Baños

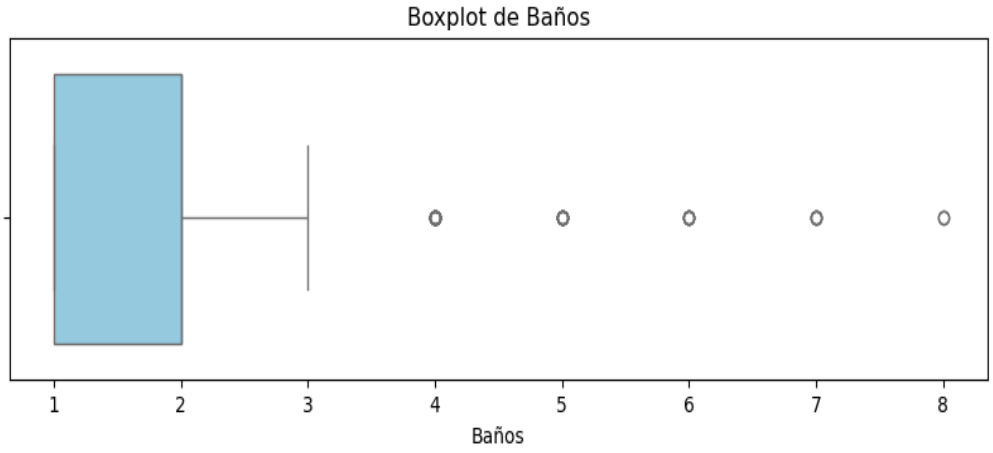


Se puede observar en la figura 22 que la variable baños incluye 452 registros válidos y presenta un promedio de 1,94 baños por inmueble, con una desviación estándar de 1,14, lo cual indica cierta variabilidad entre las viviendas. Los cuartiles (Q1 = 1, Q2 = 2, Q3 = 2) muestran que la mayoría de los inmuebles cuentan entre uno y dos baños, lo que coincide con las características típicas de viviendas residenciales estándar. El valor máximo registrado es de 8 baños, y mediante el método IQR se identificaron 37 valores atípicos, ubicados fuera de los límites de –0,5 y 3,5, correspondientes principalmente a inmuebles con un número inusualmente alto de baños. Se puede evidenciar que algunas casas tienen mayor cantidad de baños debido a que al estar en zona cercana a la parte central del municipio estas casas se construyeron para ofrecer servicio de arriendo de habitaciones y esto explica porque tienen ese número de baños disponibles, así como número alto de habitaciones. Debido al crecimiento poblacional del municipio y teniendo en cuenta que muchos de los proyectos de vivienda actuales del municipio no se habían construido aun muchas familias construyeron casas para satisfacer la demanda de

vivienda y así tener ingresos adicionales. Después de verificar la información en los respectivos portales inmobiliarios se decide conservar los valores en el conjunto de datos.

Figura 23

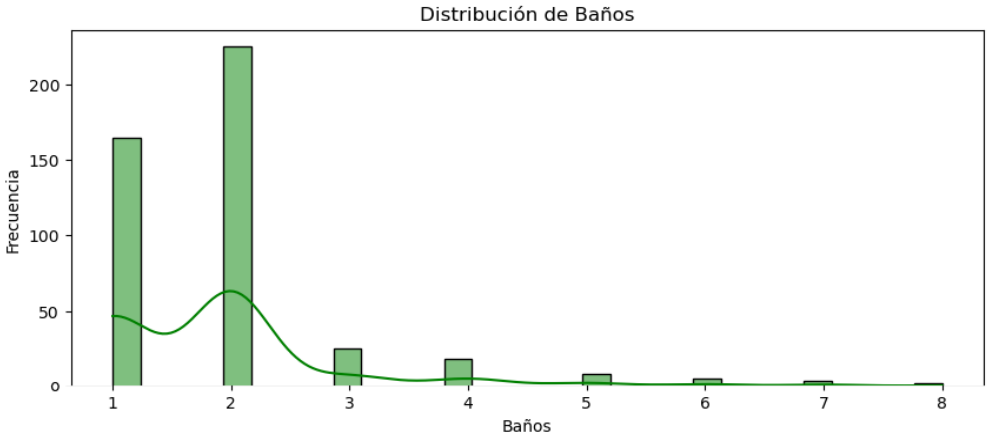
Boxplot de la Variable Baños



El boxplot de la figura 23 refleja una distribución concentrada en torno a los valores de uno y dos baños, con presencia de valores extremos hacia la parte superior. Estos outliers representan propiedades de mayor tamaño o con configuraciones internas especiales, cuya frecuencia es considerablemente menor dentro del conjunto analizado.

Figura 24

Distribución de la Variable Baños



El histograma de la figura 24 muestra una distribución discreta y moderadamente sesgada hacia la derecha, con picos claros en uno y dos baños. Las frecuencias disminuyen progresivamente para valores superiores, evidenciando que los inmuebles con tres o más baños conforman una proporción menor del mercado. Esta estructura confirma la presencia de pocos valores elevados que generan una cola larga hacia la derecha.

Figura 25

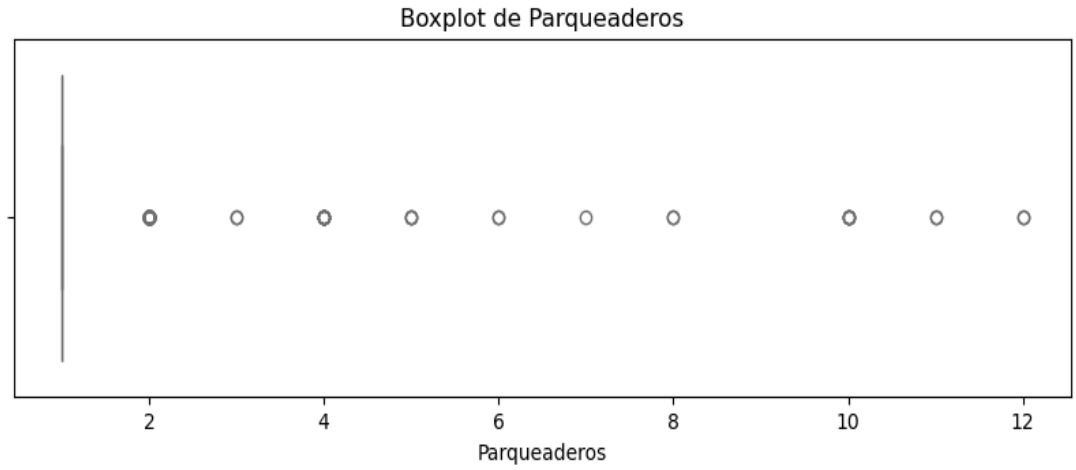
Estadísticas de la Variable Parqueaderos

```
📊 Estadísticas:
count    452.000000
mean       1.497788
std        1.580787
min        1.000000
25%        1.000000
50%        1.000000
75%        1.000000
max       12.000000
Name: Parqueaderos, dtype: float64
🚨 Outliers detectados por IQR: 91
  - Límite inferior: 1.00
  - Límite superior: 1.00
```

Como se muestra en la figura 25 la variable parqueaderos contiene 452 registros válidos y presenta un promedio de 1,50 parqueaderos por inmueble, con una desviación estándar de 1,58, lo cual indica una alta dispersión en comparación con el valor central. Los cuartiles (Q1 = 1, Q2 = 1, Q3 = 1) muestran que la mayoría de los inmuebles disponen de un solo parqueadero, mientras que el valor máximo (12 parqueaderos) revela la existencia de propiedades con capacidades de estacionamiento excepcionalmente elevadas. Mediante el método IQR se identificaron 91 outliers, dado que los límites inferior y superior coincidieron en 1, lo que ubica como valores atípicos a todos los inmuebles con dos o más parqueaderos. El alto número de parqueaderos se explica en el tipo de inmueble casa donde los habitantes del sector rural aprovechan sus inmuebles para ofrecer servicio de parqueadero y obtener ingreso adicional.

Figura 26

Boxplot de la Variable Parqueaderos

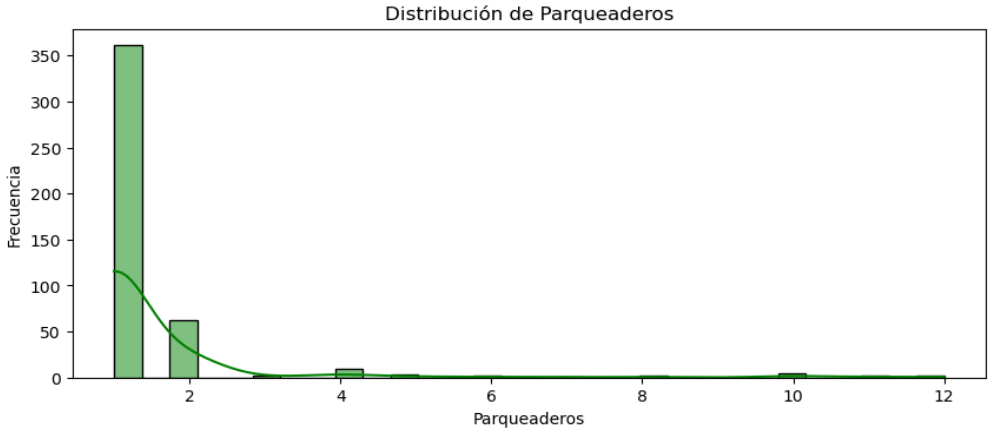


El boxplot de la figura 26 evidencia una distribución altamente concentrada en torno al valor de un parqueadero, con un número considerable de valores extremos hacia la parte superior. Estos outliers representan propiedades con dos o más parqueaderos. Esto indica que los

propietarios del tipo de inmueble casa en la vereda Canavita adecuar parte de sus casas para ofrecer servicio de parqueadero y así obtener ingresos adicionales. Se verificó la información en los portales inmobiliarios y se decide conservar los valores en el conjunto de datos.

Figura 27

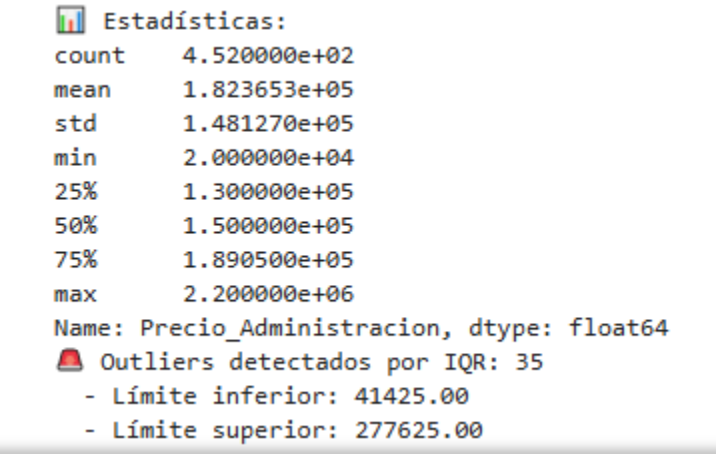
Distribución de la Variable Parqueaderos



El histograma de la figura 27 muestra una distribución marcadamente asimétrica, con un pico dominante en un parqueadero y una cola larga hacia valores superiores. La frecuencia de inmuebles con múltiples parqueaderos es baja, pero suficiente para generar una notable cantidad de valores atípicos. Esta estructura confirma la diferenciación entre viviendas estándar y propiedades con mayor capacidad vehicular. Al verificar los datos en los respectivos portales inmobiliarios se comprobó que es información real por lo cual se conservan estos valores en el conjunto de datos.

Figura 28

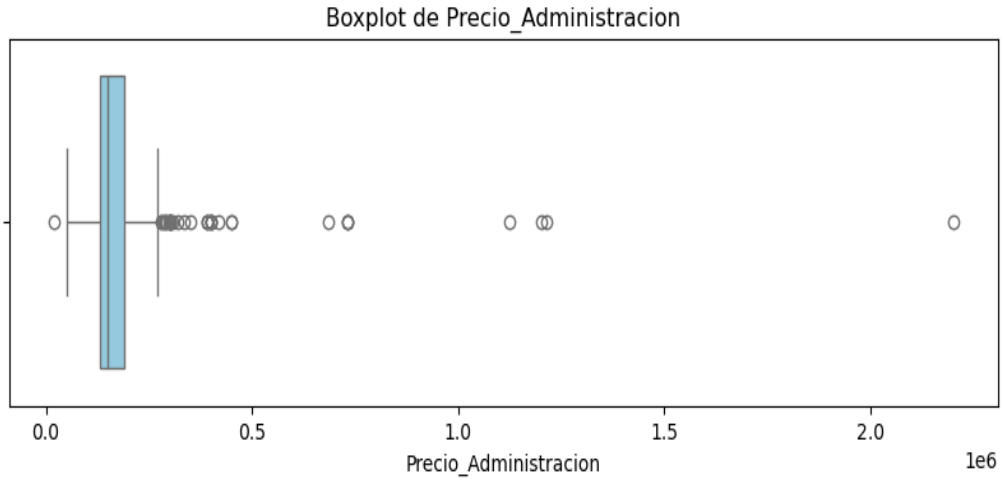
Estadísticas de la Variable Precio_Administración



Como se evidencia en la figura 28 la variable precio de administración contiene 452 registros válidos y presenta un promedio de 182.365, con una desviación estándar de 148.127, lo que refleja una variabilidad considerable en los costos de administración de los inmuebles. Los cuartiles (Q1 = $130.000, Q2 = $150.000, Q3 = $189.050) indican que la mayoría de los valores se concentran en un rango relativamente acotado, característico de las tarifas de administración habituales. No obstante, el valor máximo (2.200.000) evidencia la presencia de casos atípicamente altos. A través del método IQR se identificaron 35 outliers, situados fuera de los límites de $41.425 y $277.625. Los valores elevados corresponden al tipo de inmueble casa y después de verificar la información en los respectivos portales inmobiliarios se comprobó que son valores reales por lo cual se decide conservarlos en el conjunto de datos.

Figura 29

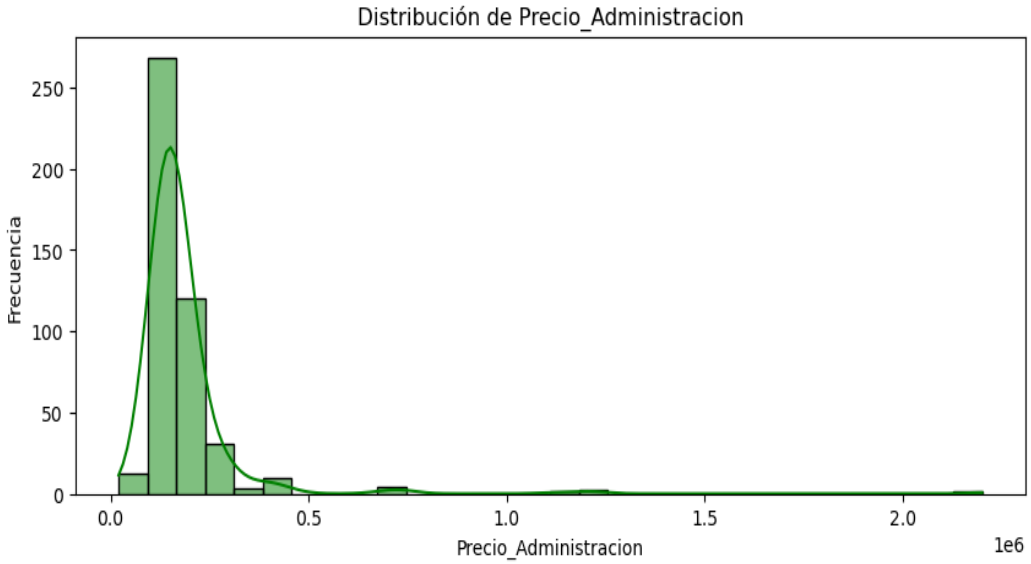
Boxplot de la Variable Precio_Administración



El boxplot de la figura 29 revela una concentración de valores alrededor del rango típico de costos administrativos, con una cola superior alargada que representa propiedades con tarifas de administración notablemente elevadas. Estos valores extremos son responsables de la dispersión observada y corresponden a inmuebles con características o servicios diferenciado en este caso al tipo de inmueble casa. Al realizar la respectiva verificación de los valores en los portales inmobiliarios se comprobó que es información real y que explica la característica del tipo de inmueble casa en la cual tiene un costo superior en comparación con el tipo de inmueble apartamento por lo cual se decide conservar los valores en el conjunto de datos ya que son datos reales y que muestran claramente una diferencia marcada entre el tipo de inmueble casa y el tipo de inmueble apartamento en esta zona rural del municipio de Tocancipá. Es adecuado tener en cuenta que muchos de los tipos de inmueble casa que presentan costos de administración elevados se debe a características que lo diferencian de los apartamentos.

Figura 30

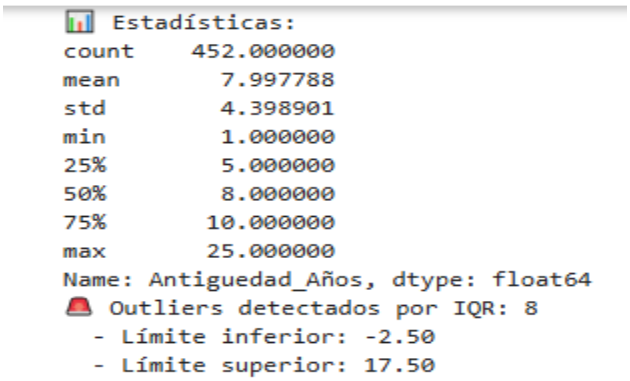
Distribución de la Variable Precio_Administración



El histograma de la figura 30 muestra una distribución asimétrica hacia la derecha, con una clara concentración en valores entre 100.000 y 200.000. La presencia de una cola larga hacia montos más altos confirma que existen pocos inmuebles con costos de administración significativamente superiores al promedio, que al ser identificados como valores reales se decide mantenerlos en el conjunto de datos.

Figura 31

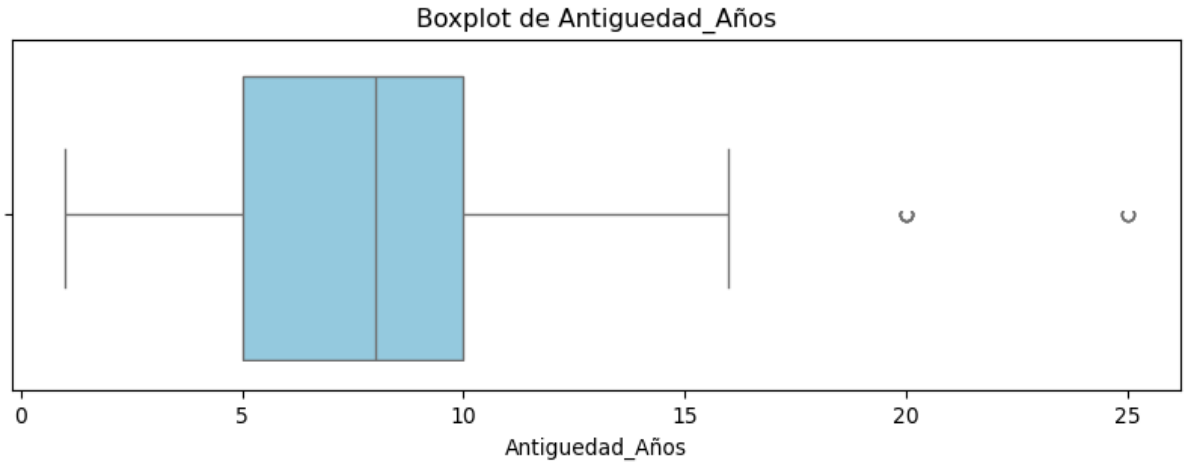
Estadísticas de la Variable Antigüedad_Años



Se observa en la figura 31 que la variable antigüedad cuenta con 452 registros válidos y presenta un promedio de 8 años, con una desviación estándar de 4,40 años, lo cual refleja una variabilidad moderada en la edad de los inmuebles. Los cuartiles (Q1 = 5 años, Q2 = 8 años, Q3 = 10 años) indican que la mayoría de las propiedades tienen entre 5 y 10 años de construcción, lo que corresponde a edificaciones relativamente recientes. El rango total oscila entre 1 año como mínimo y 25 años como máximo. Mediante el método IQR se identificaron 8 outliers, ubicados fuera de los límites de –2,5 y 17,5 años, correspondientes principalmente a inmuebles con mayor antigüedad.

Figura 32

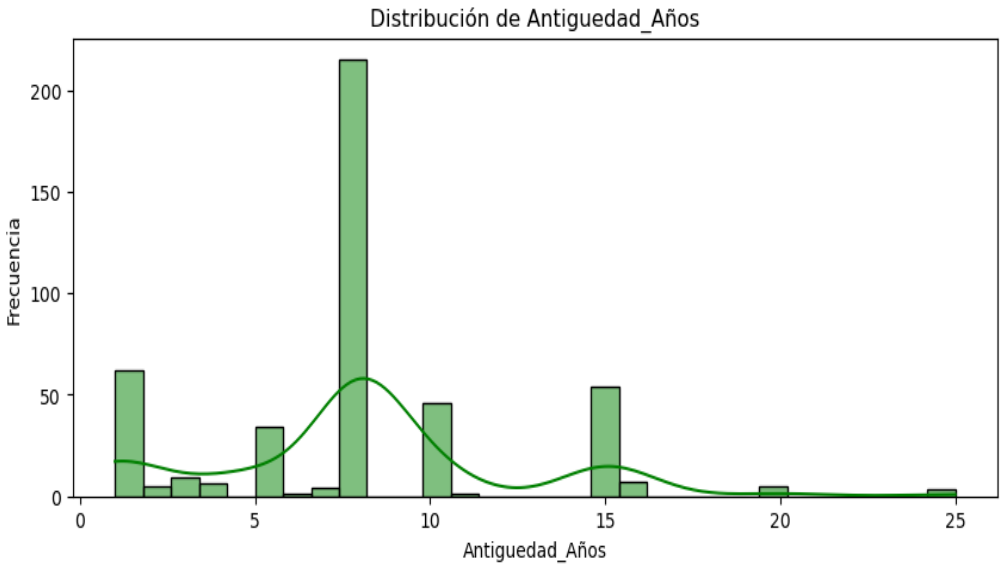
Boxplot de la Variable Antigüedad_Años



El boxplot de la figura 32 muestra una distribución relativamente compacta, con la mayor parte de los valores concentrados entre 5 y 10 años. Los pocos valores atípicos se ubican en la parte superior del diagrama y representan propiedades significativamente más antiguas que el promedio del conjunto. Este comportamiento se explica particularmente en la vereda Canavita donde se evidencia que algunos inmuebles tienen una antigüedad mayor a 10 años y corresponde a casas lo cual es habitual para este tipo de inmuebles.

Figura 33

Distribución de la Variable Antigüedad_Años



El histograma de la figura 33 evidencia una distribución ligeramente sesgada hacia la derecha, donde predominan inmuebles de antigüedad media (entre 5 y 10 años). Las frecuencias disminuyen progresivamente para edades más altas, lo cual es coherente con un entorno donde predomina la construcción relativamente reciente y pocas propiedades presentan antigüedades elevadas.

Figura 34

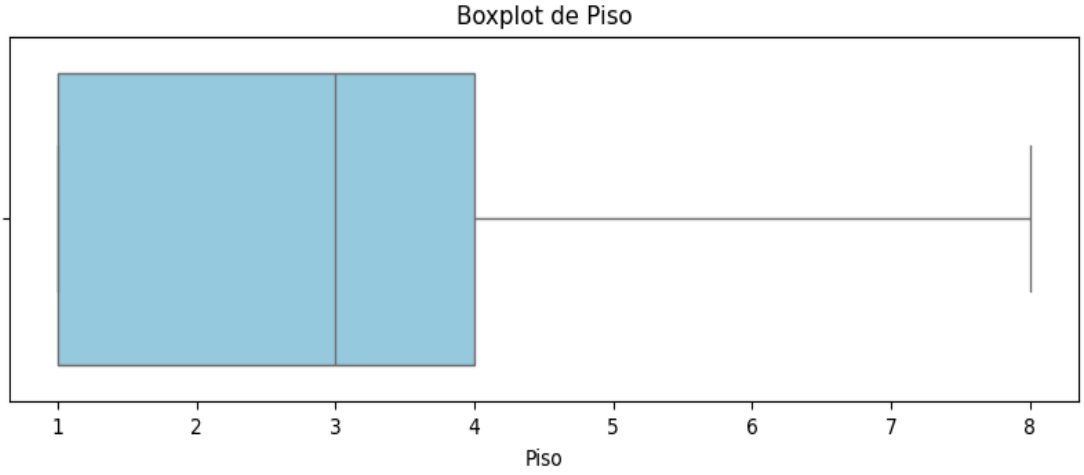
Estadísticas de la Variable Piso

```
📊 Estadísticas:
count    452.000000
mean       2.984513
std        1.820609
min        1.000000
25%        1.000000
50%        3.000000
75%        4.000000
max        8.000000
Name: Piso, dtype: float64
🚨 Outliers detectados por IQR: 0
  - Límite inferior: -3.50
  - Límite superior: 8.50
```

Se puede observar en la figura 34 que la variable piso contiene 452 registros válidos y presenta un promedio de 2,98 pisos, con una desviación estándar de 1,82, lo que refleja una variabilidad moderada en la altura de los inmuebles. Los cuartiles (Q1 = 1, Q2 = 3, Q3 = 4) indican que la mayoría de las propiedades se ubican entre el primer y cuarto piso. El valor mínimo registrado es el primer piso y el máximo corresponde al octavo piso. De acuerdo con el método IQR, no se identificaron valores atípicos, dado que todos los registros se encuentran dentro de los límites establecidos (–3,5 y 8,5).

Figura 35

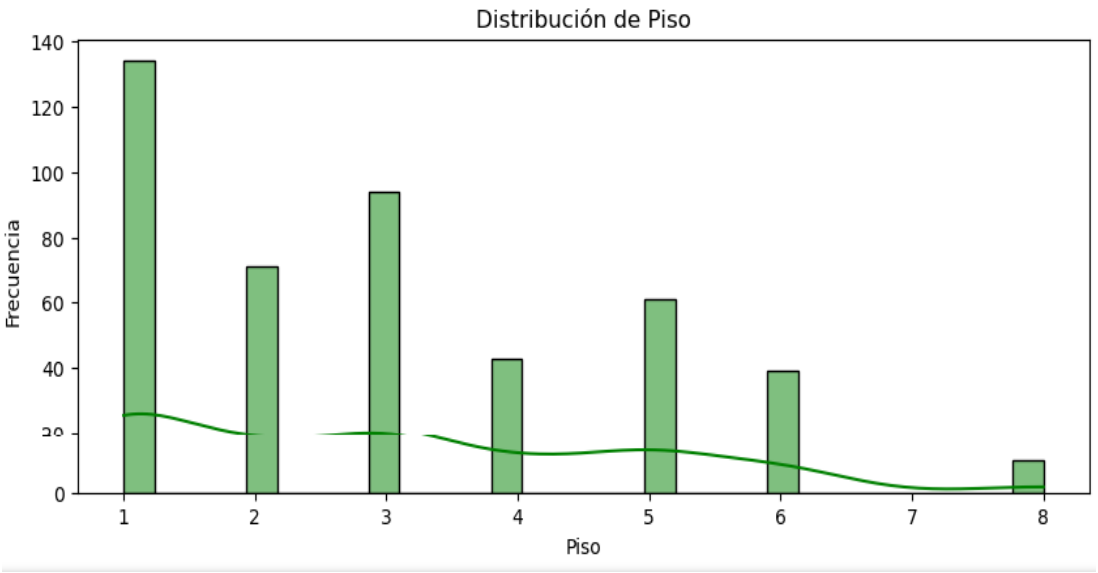
Boxplot de la Variable Piso



El boxplot de la figura 35 muestra una distribución bien delimitada, sin presencia de outliers, con una mayor concentración de valores entre los pisos 1 y 4. La ausencia de valores atípicos refleja una estructura de altura homogénea entre las edificaciones evaluadas. Es importante señalar que por lo general de los inmuebles analizados en la vereda Canavita el tipo de inmueble casa tiene un predominio de ubicación en el piso 1 y el otro tipo de inmueble que es apartamento tiene una ubicación en distintos pisos lo cual es completamente normal en este tipo de inmuebles, aunque es importante señalar que algunos apartamentos analizados se localizaron en el piso 1.

Figura 36

Distribución de la Variable Piso



El histograma de la figura 36 se evidencia una distribución discreta y moderadamente sesgada hacia valores superiores, con picos en los pisos 1 y 3. A medida que aumenta la altura del piso, la frecuencia disminuye, aunque aún se observan valores representativos hasta el octavo piso. Esta distribución sugiere una predominancia de edificaciones de baja a media altura en la vereda Canavita.

Normalización y Escalamiento

Dado que el conjunto de datos presenta variables numéricas con alta dispersión y valores atípicos significativos, el método de escalamiento más adecuado es RobustScaler, debido a su robustez frente a outliers. Este enfoque utiliza la mediana y el rango intercuartílico (IQR) para transformar las variables, evitando que valores extremos distorsionen la escala. Esta técnica mejora la estabilidad de modelos sensibles a la magnitud de los datos, como la regresión lineal, sin afectar el desempeño de algoritmos basados en árboles como Random Forest o XGBoost.

Selección de Variables

Se seleccionó como variable objetivo el precio del inmueble, la variable url se decide eliminar pues no aporta valor predictivo, se tiene en cuenta la variable tipo de inmueble que es categórica porque compara casas o apartamentos y afecta el precio, se tiene en cuenta el área construida y el área privada ya que estas dos variables afectan el precio, también se tienen en cuenta las variables: estrato, habitaciones, baños, parqueaderos, precio de administración, antigüedad, estado y piso. No se tiene en cuenta la variable característica de inmueble ya que tiene varios valores cuya cantidad no es la misma en todos los inmuebles. Después de seleccionar las variables se aplica np.log1p() a la variable precio de inmueble con el objetivo de suavizar los precios más altos, a las variables tipo de inmueble y estado se aplicó el método OneHotEncoding Para lograr que se asignen valores numéricos ya que son columnas categóricas.

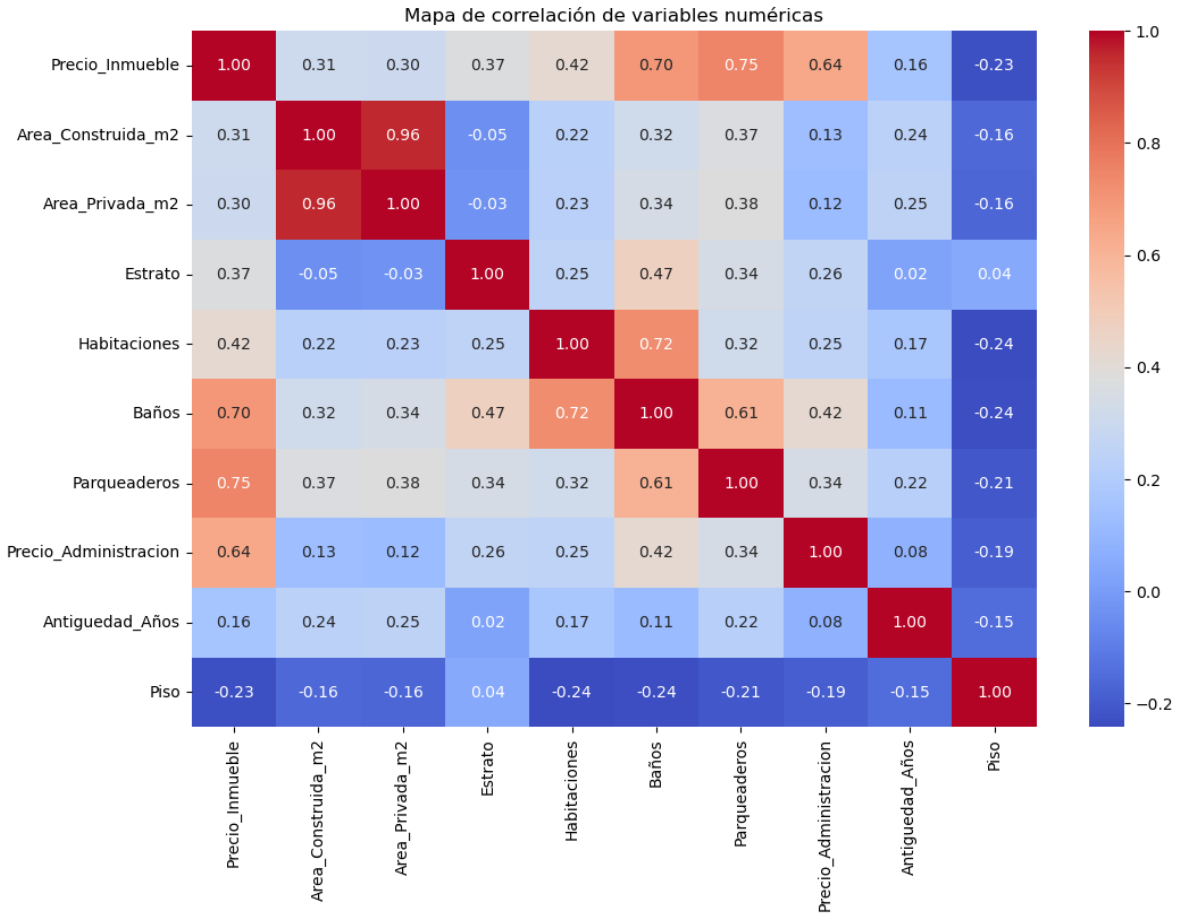
Correlación de Variables Numéricas

Se realiza un análisis de la correlación de las variables seleccionadas, primero realizando un mapa de correlación de variables empleando la correlación de Pearson figura 37. Al observar el mapa de correlación de variables se puede evidenciar con color rojo oscuro las correlaciones positivas donde se encuentra que hay una alta correlación entre el área construida y el área privada con el precio, lo cual es una interpretación lógica ya que en este tipo de inmuebles como lo son casas y apartamentos el valor del precio aumenta a medida que se dispone de un área construida y un área privada mayor, también se destacan las habitaciones con respecto al precio pues en inmuebles con mayor número de habitaciones el precio suele ser mayor, característica que se puede contemplar también con respecto a la variable número de baños pues es comprensible que inmuebles con mayor número de baños disponibles el precio suele ser mayor,

esta característica de correlación positiva se destaca también en las variables parqueaderos y precio de administración donde a mayor disponibilidad de una variable aumenta la otra.

Figura 37

Mapa de Correlación de Pearson para las Variables Seleccionadas



Un aspecto importante observado en el mapa de correlaciones de la figura 37 es las correlaciones negativas de color azul donde se evidencia que existen correlaciones débiles donde por ejemplo La variable Piso presenta correlaciones negativas débiles con la mayoría de las variables del conjunto de datos, incluyendo precio del inmueble, área construida, área privada, número de habitaciones, baños, parqueaderos, precio de administración y antigüedad, mientras

que su relación con el estrato es prácticamente nula. Estos valores, que se observa oscilan entre un - 0.15 y -0.24, indican que la altura del inmueble ejerce una influencia muy limitada y no determinante en las características físicas de la propiedad ni en su valor de mercado dentro de la vereda Canavita, lo cual es coherente con contextos rurales o semirrurales donde la verticalización es baja y la ubicación en pisos superiores no constituye un atributo relevante de apreciación económica. En consecuencia, aunque la variable puede ser incluida en los modelos de predicción, su aporte explicativo será reducido, especialmente en enfoques lineales, desempeñando un papel secundario frente a otras variables estructurales más relevantes. Se decide conservar esta variable ya que para los modelos que se emplearan se pueden capturar interacciones más complejas.

Tabla 2

Correlación de Variables con el Precio del Inmueble

Variables	Correlación
Precio_Inmueble	1.00
Parqueaderos	0.74
Baños	0.69
Precio_Administracion	0.64
Habitaciones	0.42
Estrato	0.37
Area_Construida_m2	0.30
Area_Privada_m2	0.30
Antigüedad_Años	0.15
Piso	-0.22

De acuerdo con lo observado en la tabla 2, En los inmuebles de la vereda Canavita, las características que mayor relación tienen con el precio son la cantidad de parqueaderos (0.74), la cantidad de baños (0.69) y el precio de la administración (0.64). Esto indica que las propiedades más costosas corresponden principalmente a casas amplias y de mayor nivel, más que a apartamentos de pisos altos. Variables como área construida (0.31), área privada (0.30) y antigüedad (0.16) muestran relaciones positivas, pero más bajas, mientras que el piso presenta una correlación negativa (- 0.23), debido a que las casas ubicadas en piso 1 suelen tener mayor valor que los apartamentos ubicados en pisos superiores.

Correlación de Variables Categóricas

Para determinar la relación entre las variables categóricas del conjunto de datos y el precio de los inmuebles en la vereda Canavita, se empleó un método estadístico apropiado para este tipo de información. Las variables cualitativas no pueden analizarse mediante coeficientes de correlación tradicionales, debido a que estos requieren variables numéricas continuas. Por esta razón, se calculó un indicador que permite cuantificar la proporción de variación del precio atribuible a las diferencias entre categorías. Este indicador se fundamenta en el análisis de la variabilidad total del precio y en la parte de dicha variabilidad que puede explicarse por las categorías de cada variable. El procedimiento se aplicó a las variables Tipo_Inmueble y Estado. Los resultados obtenidos fueron los siguientes: el valor para Tipo_Inmueble fue de 0.3592, mientras que el valor correspondiente a Estado fue de 0.0100. La interpretación de estos resultados permite concluir que el tipo de inmueble explica alrededor del 35,9% de la variación observada en el precio. Esto evidencia que las diferencias entre casas, apartamentos, lotes u otras tipologías tienen un impacto importante en el valor de los inmuebles ofertados en la zona de estudio. En consecuencia, esta variable aporta información relevante y debe mantenerse dentro

del conjunto de predictores considerados para el modelo. En contraste, el valor obtenido para la variable Estado indica que solo explica aproximadamente el 1% de la variabilidad del precio, lo que sugiere una relación muy débil. Esta baja influencia podría deberse a una escasa variación entre los estados reportados, a una falta de precisión en la información o a que este factor tiene poca incidencia real en el mercado inmobiliario específico de la vereda Canavita. A pesar de ello, se decidió conservar esta variable en la base de datos, tanto por su pertinencia conceptual en los procesos de valoración de inmuebles como por la posibilidad de que ciertos modelos predictivos avanzados identifiquen patrones complejos o interacciones que no se evidencian mediante un análisis univariado.

Modelado

Selección de los Modelos

La siguiente actividad consistió en seleccionar los modelos de predicción de precios de vivienda en este caso se decide iniciar con un modelo sencillo como la regresión lineal y después se decide continuar con el modelo de Random Forest, el modelo XGBoost y el modelo CatBoost para realizar comparación de resultados obtenidos. La regresión lineal se incluyó como modelo base por su simplicidad y facilidad de interpretación, sirviendo como referencia para evaluar mejoras con modelos más complejos. Random Forest se seleccionó por su capacidad de manejar relaciones no lineales y detectar interacciones entre variables, ofreciendo mayor robustez frente a valores atípicos. XGBoost fue elegido por su algoritmo de Boosting, que optimiza iterativamente la predicción corrigiendo errores de modelos anteriores y capturando patrones complejos con alta precisión. Finalmente, CatBoost se incorporó por su manejo nativo de variables categóricas y su algoritmo de Boosting ordenado, lo que permite conservar la información original de las

categorías y reducir el riesgo de sobreajuste, mostrando superioridad en todas las métricas evaluadas y justificando su elección como modelo final.

División de los Datos de Entrenamiento y Prueba

La división de los datos en un 80% para entrenamiento y 20% para prueba es un enfoque común y equilibrado para el presente dataset de tamaño moderado (452 registros). Esta proporción permite que el modelo aprenda de una cantidad suficiente de datos mientras se reserva un subconjunto representativo para evaluación independiente, asegurando que las métricas calculadas reflejen de manera confiable su desempeño en datos no vistos.

Entrenamiento del Modelo de Regresión Lineal

El modelo de regresión lineal se entrenó utilizando como variable objetivo el precio del inmueble, transformado mediante el logaritmo natural para reducir la asimetría y mejorar la estabilidad de las predicciones. Las variables predictoras incluyeron características numéricas y categóricas, siendo estas últimas codificadas con one-hot encoding y las numéricas escaladas mediante un método robusto para mitigar el efecto de valores atípicos. Los datos se dividieron en entrenamiento y prueba (80/20) para asegurar una evaluación confiable sobre datos no vistos. Las predicciones se transformaron nuevamente a la escala original de precios y la calidad del modelo se evaluó mediante métricas de error absoluto y cuadrático, errores porcentuales y el coeficiente de determinación, proporcionando una visión integral de la precisión y capacidad explicativa del modelo sobre propiedades de distintos rangos de precio.

Entrenamiento del Modelo Random Forest

Para entrenar el modelo de Random Forest se utilizó la misma preparación aplicada en la regresión lineal, transformando la variable objetivo mediante logaritmo natural, codificando las variables categóricas mediante one-hot encoding y escalando las variables numéricas para

reducir el efecto de valores atípicos. El conjunto de datos se dividió en entrenamiento y prueba en proporción 80/20, asegurando una evaluación confiable sobre datos no vistos. El modelo consistió en un conjunto de árboles de decisión entrenados sobre el conjunto de entrenamiento para capturar relaciones no lineales y posibles interacciones entre variables. Las predicciones se transformaron nuevamente a la escala original de precios y se evaluaron mediante el error cuadrático medio, el error absoluto medio, el coeficiente de determinación y las métricas de error porcentual, proporcionando una visión completa de la precisión, la magnitud de los errores y la capacidad explicativa del modelo sobre los precios de los inmuebles.

Entrenamiento del Modelo XGBoost

El modelo XGBoost se entrenó siguiendo la misma preparación de datos utilizada en los modelos anteriores, incluyendo la transformación logarítmica de la variable objetivo, codificación de variables categóricas y escalado de las variables numéricas. Se dividió el conjunto de datos en entrenamiento y prueba en proporción 80/20, garantizando una evaluación confiable sobre datos no vistos. XGBoost, mediante un algoritmo de Boosting, ajusta iterativamente los errores de predicciones anteriores para optimizar la precisión y capturar patrones complejos en los datos. Las predicciones se transformaron nuevamente a la escala original de precios y se evaluaron usando error cuadrático medio, error absoluto medio, coeficiente de determinación y métricas de error porcentual, permitiendo medir la exactitud de las predicciones, su capacidad para explicar la variabilidad del precio y la magnitud relativa de los errores respecto a los valores reales.

Entrenamiento del Modelo CatBoost

Para el entrenamiento del modelo CatBoost se empleó la misma preparación de datos aplicada en los demás modelos, transformando la variable objetivo mediante logaritmo natural y

escalando las variables numéricas, mientras que las variables categóricas fueron indicadas directamente al modelo sin necesidad de codificación adicional, aprovechando su manejo nativo de categorías. La división de los datos en entrenamiento y prueba se realizó en proporción 80/20 para garantizar una evaluación confiable. CatBoost utiliza un algoritmo de Boosting ordenado que reduce el riesgo de sobreajuste y optimiza la precisión de las predicciones, capturando relaciones complejas y patrones no lineales en los datos. Las predicciones se transformaron nuevamente a la escala original y se evaluaron mediante error cuadrático medio, error absoluto medio, coeficiente de determinación y métricas de error porcentual, ofreciendo una visión completa de la exactitud, capacidad explicativa y errores relativos del modelo sobre los precios de los inmuebles. Al observar los resultados de las métricas obtenidas con estos modelos en la tabla 3. Podemos evidenciar que el modelo CatBoost obtiene un menor RMSE lo cual indica que predice mejor los precios altos, tiene un menor MAE lo cual indica una mejor precisión promedio, el R^2 es mayor lo cual indica que este modelo realiza una mejor explicación del precio, su valor de MAPE menor indica que el error relativo es menor, su valor de SMAPE indica que este modelo tiene una mejor estabilidad. Este modelo funciona bien porque en el conjunto de datos hay categorías no ordinales como lo son el tipo de inmueble y el estado pues no necesita one-hot encoding porque aprende directamente de las iteraciones con otras categorías, también es importante señalar que este tipo de modelos se desempeña muy bien cuando los dataset no son muy grandes en nuestro caso menos de 600 datos. Se puede identificar que de acuerdo con las características del conjunto de datos donde los precios rurales tienen grandes saltos dependiendo del tipo de inmueble, el estado, el área, la antigüedad entre otros el modelo CatBoost captura esto mejor que otros modelos. Otra característica de este modelo es que maneja muy bien los valores atípicos especialmente en sectores como la finca raíz, donde se

pudo identificar que existen precios altos de un inmueble con respecto a otro teniendo en cuenta las características de cada tipo de propiedad.

Tabla 3

Métricas de los Modelos Implementados

Métrica	Catboost	Xgboost	random forest	regresión lineal
Rmse	95,575,716.02	102,628,122.88	106,937,755.85	144,538,988.95
Mae	48,386,039.39	54,838,163.89	55,186,654.13	69,566,681.75
r^2	0.9155	0.9026	0.8943	0.8068
Mape	14.89%	17.89%	16.15%	18.82%
Smape	14.36%	15.91%	15.46%	18.38%

El mejor desempeño se obtuvo con el modelo CatBoost, alcanzando un R^2 de 0.915, un MAE de 48 millones y un MAPE de 14.89%. Estos valores indican un excelente ajuste para un mercado inmobiliario heterogéneo como el de la vereda Canavita, siendo este el modelo recomendado para el apoyo a la toma de decisiones de inversión en finca raíz.

Aplicación de Hiperparametrización en los Modelos Seleccionados

La hiperparametrización es el proceso de ajustar los parámetros de un modelo que no se aprenden directamente durante el entrenamiento, con el objetivo de optimizar su rendimiento y capacidad de generalización. En los modelos CatBoost y XGBoost, se empleó la hiperparametrización para equilibrar precisión y robustez frente a la complejidad del dataset, que incluía variables mixtas, presencia de valores atípicos y un amplio rango de precios, evitando así el sobreajuste. En CatBoost, la hiperparametrización se realizó ajustando el número de iteraciones, la tasa de aprendizaje, la profundidad de los árboles, la regularización L2 y

parámetros específicos como Bagging temperature y border count, aprovechando además su manejo nativo de variables categóricas. En XGBoost se configuraron parámetros equivalentes, incluyendo número de estimadores, profundidad máxima, tasa de aprendizaje, submuestreo de filas y columnas y regularización, asegurando robustez y control del sobreajuste. En ambos casos, la estrategia buscó un equilibrio entre aprendizaje detallado y generalización del modelo para lograr predicciones precisas y estables. Los resultados de la hiperparametrización se observan en la tabla 4.

Tabla 4

Resultados de la Hiperparametrización de los Modelos

Métrica	xgboost optimizado	catboost optimizado
Rmse	113,155,313.24	90,058,774.36
Mae	57,863,946.75	49,861,569.49
r^2	0.8816	0.9250
Mape	17.24%	15.39%
Smape	15.70%	14.86%

Tras realizar la optimización de hiperparámetros con RandomizedSearchCV tanto para CatBoost como para XGBoost, se observa que el rendimiento del modelo CatBoost supera ampliamente al de XGBoost, según los valores de la tabla 4. El modelo XGBoost optimizado obtuvo un $R^2 = 0.8816$ y un RMSE de 113 millones, mientras que el modelo CatBoost Optimizado alcanzó un $R^2 = 0.925$ y un RMSE de 90 millones, lo cual representa una reducción del 20% en el error medio cuadrático. Esto indica que CatBoost es significativamente más adecuado para este problema, probablemente debido a su manejo nativo de variables categóricas,

su robustez frente a outliers y su mejor capacidad para capturar relaciones complejas entre las variables del mercado inmobiliario. Sin embargo, se va a realizar validación cruzada en ambos modelos para elegir de acuerdo con las métricas obtenidas el mejor modelo.

Aplicación de Validación Cruzada a los Modelos Seleccionados

La validación cruzada es una técnica de evaluación de modelos que permite estimar su desempeño de manera robusta utilizando distintos subconjuntos del dataset, evitando el sesgo de evaluar sobre un único conjunto de prueba y maximizando el uso de los datos disponibles. En los modelos CatBoost y XGBoost se empleó validación cruzada de tipo K-Fold, en la que el conjunto de datos se dividió en cinco pliegues, entrenando el modelo en cuatro de ellos y evaluando en el restante, repitiendo este proceso de manera rotativa hasta cubrir todos los pliegues. Este enfoque se utilizó para obtener métricas de rendimiento más confiables y estables, asegurando que las evaluaciones reflejaran la capacidad de generalización de los modelos sobre datos no vistos, y permitiendo comparar de manera objetiva la precisión y robustez de ambos algoritmos antes de seleccionar el mejor para el entrenamiento final.

Tabla 5

Métricas Aplicando Validación Cruzada a los Modelos Seleccionados

métrica	modelo xgboost	desviación xgboost	modelo catboost	desviación catboost
rmse	214,768,359.45	± 350,182.44	204,484,577.40	± 121,231,176.90
mae	71,048,209.22	± 24,864,803.26	68,909,369.84	± 28,300,085.83
r^2	0.8037	± 0.1101	0.8200	± 0.1181
mape	14.39%	± 2.63%	13.82%	± 2.39%
smape	13.98%	± 1.89%	13.64%	± 1.99%

La validación cruzada de los modelos XGBoost y CatBoost muestra que, si bien ambos algoritmos son robustos y capturan de manera consistente la estructura de los datos, CatBoost presenta un rendimiento ligeramente superior en todas las métricas evaluadas. Los errores absolutos (RMSE y MAE) son elevados en ambos modelos debido a la presencia de outliers en los precios, pero CatBoost mantiene valores ligeramente menores y más consistentes entre los folds, indicando una menor sensibilidad a la variabilidad de las particiones. En cuanto a los errores relativos (MAPE y SMAPE), CatBoost también exhibe valores más bajos y estables, lo que refleja un desempeño proporcional más confiable en comparación con XGBoost. El coeficiente de determinación promedio (R^2) es mayor en CatBoost (0.82 frente a 0.804), mostrando que captura algo mejor la varianza de los datos. Considerando tanto las métricas absolutas como las relativas, y teniendo en cuenta la estabilidad y capacidad de generalización, se seleccionó CatBoost como el mejor modelo para el entrenamiento final, ya que combina precisión y consistencia frente a la dispersión y los valores atípicos presentes en el dataset.

Selección y Entrenamiento del Mejor Modelo

El modelo CatBoost fue seleccionado como la mejor alternativa debido a que, tras el proceso comparativo con regresión lineal, Random Forest y XGBoost, demostró el mayor poder predictivo y la mejor capacidad de generalización, alcanzando consistentemente las métricas más sólidas, especialmente después de la hiperparametrización. Una vez identificado como el modelo óptimo, fue entrenado nuevamente desde cero utilizando el conjunto de entrenamiento completo (80% de los datos) y los mejores hiperparámetros obtenidos mediante RandomizedSearchCV como iterations = 1500, learning_rate = 0.03, depth= 4, l2_leaf_reg =3, border_count = 64 y bagging_temperature = 0.5. Este reentrenamiento es necesario porque, durante la validación cruzada y la búsqueda de hiperparámetros, se entrenan múltiples modelos parciales con

diferentes particiones del dataset; ninguno de ellos representa un modelo definitivo listo para producción. Por ello, tras identificar la configuración óptima, se entrena un modelo final único, con todos los datos disponibles para aprender patrones de forma completa y entregar las métricas finales sobre un conjunto de prueba independiente, garantizando así un rendimiento robusto, reproducible y preparado para su uso real en predicción de precios inmobiliarios. Es importante tener en cuenta que al entrenar nuevamente el modelo elegido se definen la totalidad de las métricas para ser analizadas y así entender de una manera más precisa la calidad del modelo.

Resultados

Después de entrenar el modelo CatBoost que mostró un mejor resultado comparado con los otros modelos de predicción implementados se obtuvieron los siguientes resultados mostrados en la tabla 6.

Tabla 6

Métricas CatBoost Final (Test)

Métrica	catboost final (test)
Rmse	90,058,774.36
Mae	49,861,569.49
r^2	0.9250
Mape	15.39%
Smape	14.86%

Las métricas que se reportan como resultado final del proyecto en la tabla 6, corresponden al modelo definitivo entrenado, y no a las métricas de la validación cruzada, porque solo el modelo final refleja el desempeño real de la versión que efectivamente será utilizada en producción. La validación cruzada cumple un propósito distinto: permite evaluar la estabilidad del algoritmo, comparar modelos y seleccionar los hiperparámetros óptimos; sin embargo, cada iteración de la validación cruzada entrena modelos distintos sobre diferentes particiones, por lo que sus métricas representan un promedio de desempeño, no el rendimiento del modelo final que se implementará. Una vez seleccionados los mejores hiperparámetros, se entrena un único modelo con el 80% de los datos y se evalúa sobre un 20% totalmente independiente, lo que proporciona una estimación más fiel del rendimiento esperado con datos

nuevos. En cuanto a los resultados finales, el modelo CatBoost optimizado obtuvo un RMSE de aproximadamente 90 millones, lo cual indica el error cuadrático promedio en pesos y, aunque elevado por la magnitud natural de los precios inmobiliarios, muestra una mejora significativa frente a otros modelos evaluados. El MAE cercano a 50 millones sugiere que, en promedio, el modelo se desvía esa cantidad respecto al precio real, un nivel aceptable considerando la variabilidad de los inmuebles en la vereda Canavita. El R^2 de 0.925 evidencia que el modelo explica el 92.5% de la variabilidad del precio, un desempeño muy alto para datos inmobiliarios, donde el ruido y heterogeneidad suelen ser considerables. Finalmente, los valores de MAPE (15.39%) y SMAPE (14.86%) indican que el modelo tiene un error porcentual moderado, relativamente bajo para mercados con alta dispersión de precios, lo que confirma que el modelo es robusto y adecuado para apoyar decisiones de compra y venta. En conjunto, estas métricas muestran un modelo con buen poder predictivo, estabilidad y precisión suficiente para uso aplicado en el contexto inmobiliario.

Gráfico de Residuales

Los gráficos de residuales son herramientas fundamentales en la validación de modelos de regresión porque permiten evaluar no solo qué tan bien predice el modelo, sino cómo se comportan los errores (residuales) a lo largo del rango de valores predichos. En este proyecto, donde se predicen precios de vivienda un problema típicamente ruidoso y con alta variabilidad estos gráficos ayudan a identificar patrones, sesgos y posibles problemas estructurales en el modelo CatBoost final. En este proyecto, los residuales se calcularon como la diferencia entre el valor real del inmueble y el valor predicho por el modelo CatBoost, ambos transformados de regreso a su escala original mediante expm1() para asegurar que el análisis del error fuera

interpretable en pesos colombianos. Es decir, para cada observación del conjunto de prueba se aplicó: Residual = Valor Real – Predicción

En este caso se realizó el cálculo de los primeros diez residuales de nuestro modelo los cuales se muestran en la tabla 7.

Tabla 7

Primeros 10 Residuales del Modelo CatBoost

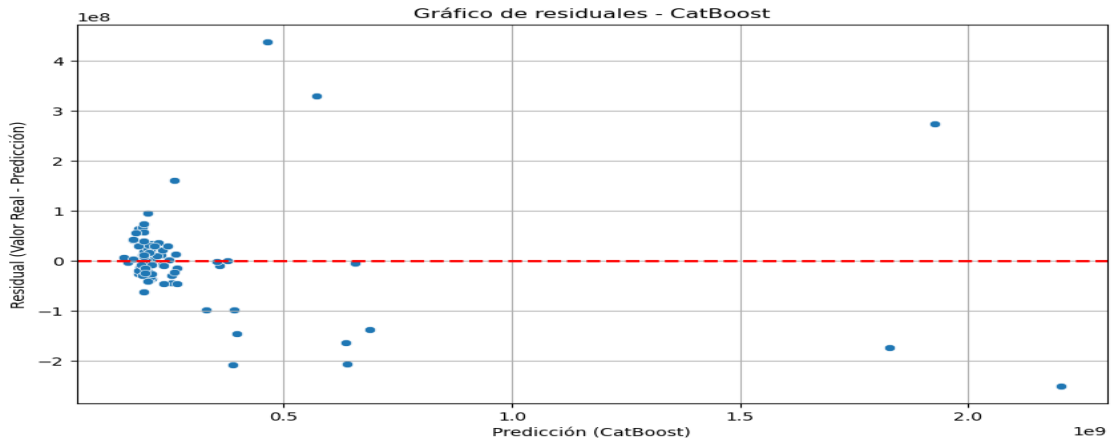
valor real	Predicción	Residual
219,000,000	197,478,750	21,521,250
245,000,000	181,966,298	63,033,702
225,000,000	204,081,099	20,918,901
350,000,000	359,935,840	-9,935,840
250,000,000	266,146,857	-16,146,857
210,000,000	205,267,539	4,732,461
235,000,000	203,852,364	31,147,636
190,000,000	195,154,740	-5,154,740
180,000,000	203,867,511	-23,867,511
179,999,999	193,776,973	-13,776,974

La tabla 7 muestra los primeros 10 residuales del modelo CatBoost entrenado para predecir precios de vivienda. Cada fila indica el valor real de la vivienda, la predicción del modelo y el Residual, calculado como la diferencia entre el valor real y el predicho. Los residuales positivos indican que el modelo subestimó el precio, mientras que los negativos indican que sobreestimó el precio. Observando los valores, la mayoría de los residuales se

encuentran dentro de un rango razonable comparado con los valores de los inmuebles, con algunos casos puntuales de desviaciones más grandes (por ejemplo, un residual de 63 millones). Esto sugiere que, aunque el modelo puede presentar errores mayores en propiedades con características extremas o atípicas, en general captura correctamente la tendencia y la magnitud de los precios de la mayoría de las viviendas. El hecho de que los residuales no muestren un patrón sistemático (como subestimaciones o sobreestimaciones constantes) y que su magnitud relativa sea moderada respalda la alta capacidad predictiva del modelo, coherente con métricas de desempeño como RMSE de 90 millones, $R^2 = 0.925$ y MAPE = 15%. Esto indica que el modelo es confiable para estimar precios de vivienda y tiene un buen equilibrio entre precisión y generalización, siendo útil para tareas de valoración inmobiliaria o análisis de mercado.

Figura 38

Residuales del Modelo CatBoost



De acuerdo con lo observado en la figura 38, la combinación de una concentración fuerte de residuales alrededor de cero, dispersión limitada y ausencia de patrones sistemáticos respalda que el modelo CatBoost ofrece predicciones precisas y confiables para la mayoría de las propiedades, con errores significativos solo en casos atípicos. Esto es coherente con las métricas

de desempeño obtenidas (RMSE, MAE y R^2), confirmando la alta calidad del modelo para la predicción de precios de vivienda.

Importancia de las Características (Features)

En análisis de modelos de machine learning, la importancia de las características (feature importance) se refiere a la medida del impacto relativo que tiene cada variable predictora sobre las predicciones del modelo, permitiendo identificar cuáles son los factores más determinantes y facilitando la interpretación del modelo. En el caso del modelo CatBoost entrenado para predecir precios de vivienda, la importancia se calculó mediante el método PredictionValuesChange, que evalúa el cambio promedio en las predicciones cuando cada característica se utiliza en las decisiones de los árboles del modelo; es decir, mide cuánto contribuye cada feature a alterar el valor predicho. Estos valores se normalizaron a porcentaje, permitiendo comparar de manera clara el peso relativo de cada variable. Este análisis permite interpretar cómo el modelo utiliza cada feature para generar sus predicciones, facilitando la comprensión de relaciones dentro de los datos y la toma de decisiones basadas en el modelo. Además, ayuda a detectar features irrelevantes o poco influyentes, optimizando la eficiencia del modelo y su capacidad de generalización.

Tabla 8

Características para el Modelo CatBoost

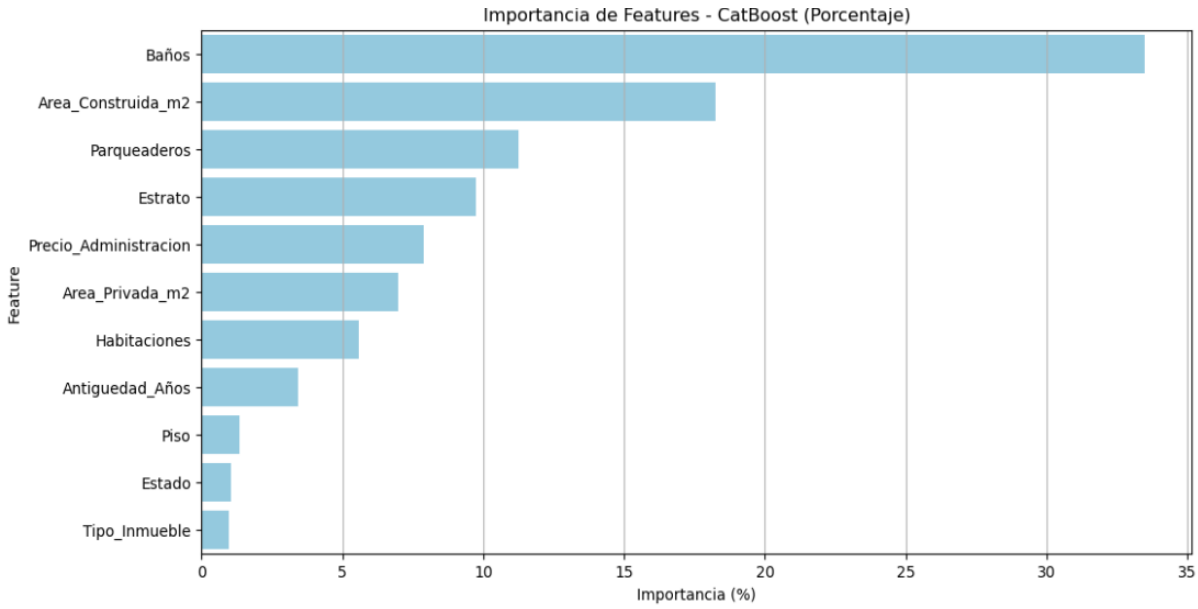
Variable	importancia de features en %
Baños	33.48%
Area_Construida_m2	18.26%
Parqueaderos	11.24%
Estrato	9.73%

Variable	importancia de features en %
Precio_Administracion	7.90%
Area_Privada_m2	7.00%
Habitaciones	5.58%
Antigüedad_Años	3.43%
Piso	1.35%
Estado	1.04%
Tipo_Inmueble	0.98%

De acuerdo con lo observado en la tabla 8, el análisis de importancia de características del modelo CatBoost revela qué variables influyen de manera más significativa en la predicción de precios de vivienda. En este caso, se observa que “Baños” es la característica con mayor aporte (33,48%), seguida por “Área Construida” (18,26%) y “Parqueaderos” (11,24%), lo que indica que estos factores son determinantes para estimar correctamente el valor de las propiedades. Variables como “Estrato”, “Precio de Administración” y “Área Privada” también aportan de manera moderada, mientras que “Tipo de Inmueble”, “Estado” y “Piso” tienen un impacto relativamente menor. Esta distribución confirma que el modelo prioriza características físicas y funcionales de la vivienda sobre variables categóricas menos influyentes, lo que fortalece la interpretabilidad del modelo y respalda su capacidad para realizar predicciones precisas basadas en los factores más relevantes del mercado inmobiliario.

Figura 39

Características del Modelo CatBoost



En la figura 39 se observa la clasificación por importancia de cada una de las variables para el modelo CatBoost donde se evidencia la supremacía que tiene la variable “Baños” por encima de las otras variables y donde se destacan la importancia de las variables “Area_Construida_m2”, “Parqueaderos” y “Estrato”. Es de destacar que la variable que menor peso en importancia tiene para el modelo es la variable “Tipo_Inmueble”, debido a que se evidencia una diferencia entre los dos inmuebles del estudio que son casas y apartamentos.

Comportamiento del Modelo CatBoost por Segmentos y Rangos de Precios Diferentes

El análisis del comportamiento del modelo CatBoost por rangos de precios y tipo de inmueble es fundamental para evaluar su desempeño en distintos segmentos del mercado y detectar posibles sesgos o limitaciones en la predicción. En este estudio, se creó un DataFrame de evaluación con los valores reales y predichos, agrupando las propiedades por tipo de inmueble y asignando cada una a un rango de precio (“Bajo”, “Medio” y “Alto”) según los percentiles de

cada grupo. A continuación, se calcularon métricas de desempeño RMSE, MAE, MAPE y SMAPE de manera segura por cada combinación de tipo y rango de precio, permitiendo observar cómo varía la precisión del modelo en distintos segmentos.

Tabla 9

Modelo CatBoost por Segmentos y Rangos de Precios

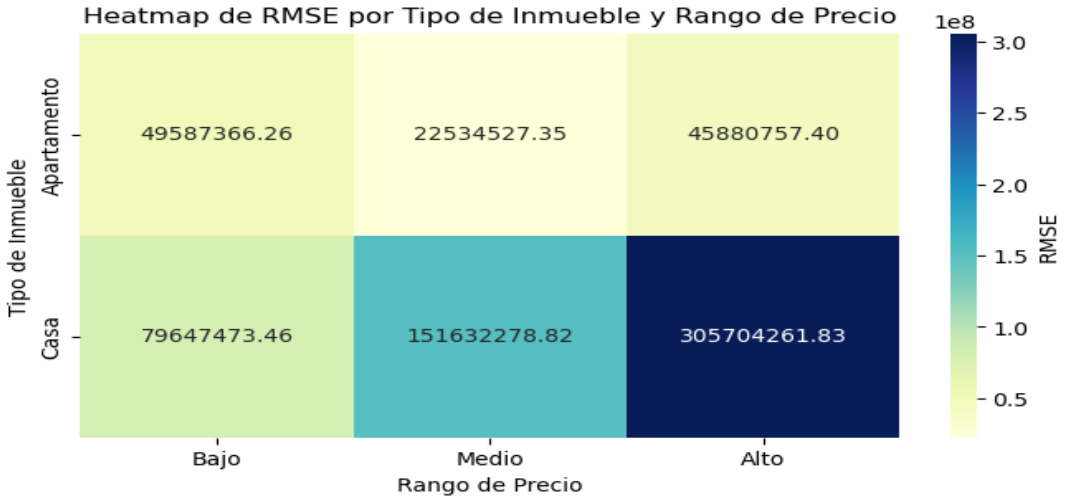
tipo de inmueble	rango de precio	rmse (cop)	mae (cop)	mape (%)	smape (%)
Apartamento	Bajo	49,587,366.26	30,077,154.16	17.73%	14.79%
Apartamento	Medio	22,534,527.35	18,187,771.13	8.44%	8.61%
Apartamento	Alto	45,880,757.40	37,548,378.85	15.03%	16.08%
Casa	Bajo	79,647,473.46	54,172,516.29	21.41%	17.60%
Casa	Medio	151,632,278.82	135,366,893.30	29.50%	27.80%
Casa	Alto	305,704,261.83	292,981,323.27	24.19%	28.84%

Analizando los resultados de la tabla 9, por segmentos y rangos de precios, se observa que el modelo CatBoost presenta un comportamiento diferencial según el tipo de inmueble y el rango de precio, lo cual es común en modelos de predicción inmobiliaria debido a la heterogeneidad de los datos. En términos de fortalezas, el modelo muestra su mayor precisión al predecir apartamentos de rango de precio medio, con un RMSE de 22,534,527.35 y un MAPE de 8.44%, lo que indica que los errores absolutos y relativos son más bajos en este segmento, reflejando una capacidad destacada del modelo para capturar la variabilidad de precios en propiedades con valores intermedios. Por tipo de inmueble, los apartamentos en general tienen errores más contenidos en comparación con las casas, especialmente en rangos medios y bajos,

lo que sugiere que el modelo aprende mejores patrones de precios más homogéneos, como los apartamentos, mientras que las casas presentan mayor dispersión de precios y, por ende, errores más altos. En cuanto a rangos de precios, el modelo es más confiable en rangos medios tanto para apartamentos como para casas, mientras que los rangos altos muestran un incremento significativo en RMSE y MAE (por ejemplo, casas altas con RMSE de 305,704,261.83), reflejando la dificultad de predecir propiedades muy costosas debido a la alta variabilidad y menor representatividad en los datos de entrenamiento. En resumen, la interpretación indica que el modelo tiene fortalezas claras para predecir apartamentos y propiedades de rango medio, mientras que su desempeño disminuye en propiedades de alto valor y casas de rango alto, lo que proporciona información estratégica para enfocar futuras mejoras del modelo o la toma de decisiones basadas en predicciones. Se emplearon mapas de calor para visualizar el comportamiento del modelo, y así ofrecer una interpretación más adecuado sobre la calidad de las predicciones realizadas por el modelo, de esta manera permitiendo una facilidad de la lectura de los resultados.

Figura 40

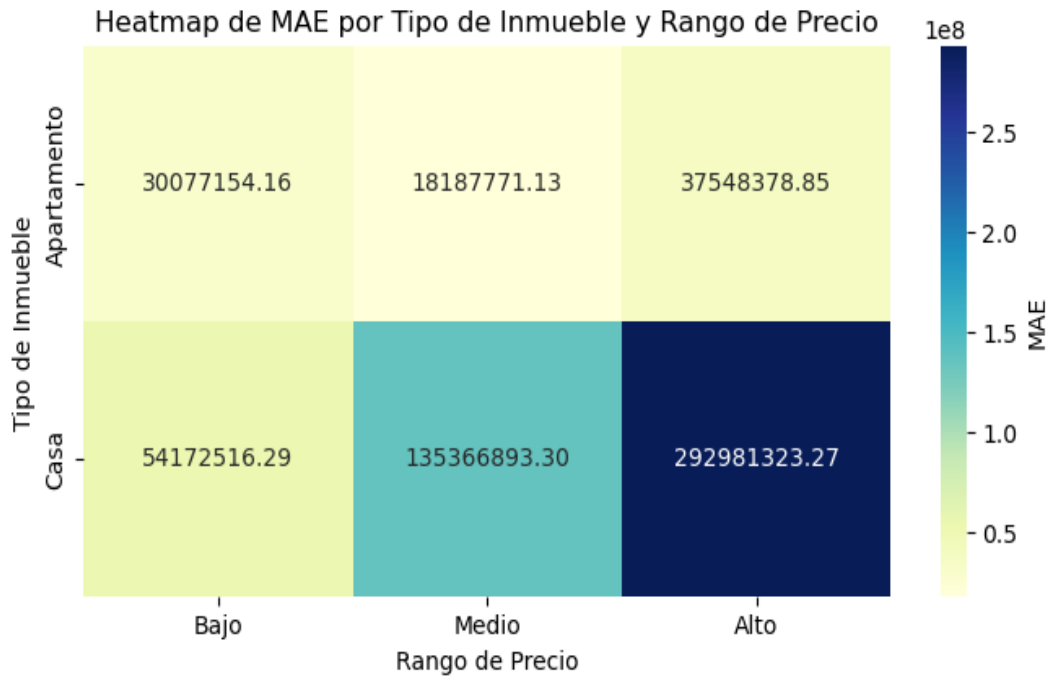
Mapa de Calor de RMSE por Tipo de Inmueble y Rango de Precio



El mapa de calor de RMSE por tipo de inmueble y rango de precio de la figura 40 refleja visualmente la magnitud de los errores del modelo a través de la intensidad del color: los valores en azul oscuro, como las casas de rango alto y medio, indican errores mayores y mayor dificultad en la predicción debido a mayor variabilidad o menor disponibilidad de datos, mientras que los segmentos con azul más claro, como las casas de rango bajo y algunos apartamentos, representan errores moderados y predicciones aceptables. Los casilleros sin color, como los apartamentos de rango medio, señalan los errores más bajos, destacando que el modelo predice con mayor precisión este grupo. En conjunto, la intensidad del color permite identificar rápidamente dónde el modelo tiene un desempeño fuerte o débil, evidenciando que el tipo de inmueble y el rango de precio influyen significativamente en la calidad de la predicción. Es adecuado señalar que en la implementación de modelo para predecir precios en la vereda Canavita el tipo de inmueble casa se diferencia claramente del tipo de inmueble apartamento, ya que las casas en esta zona tienden a tener un precio más elevado el cual sin duda es también otorgado por las características del suelo del municipio ya que es zona industrial.

Figura 41

Mapa de Calor de MAE por Tipo de Inmueble y Rango de Precio

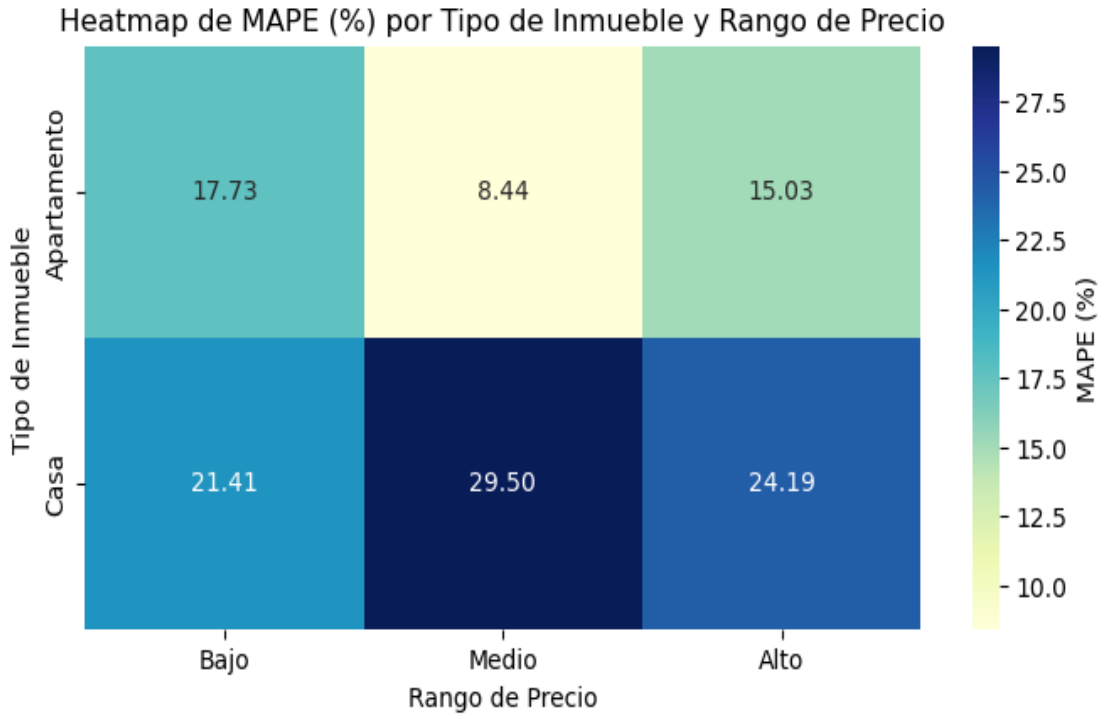


El mapa de calor de MAE por tipo de inmueble y rango de precio de la figura 41 muestra la intensidad de los errores absolutos del modelo mediante la intensidad del color, donde los valores más altos, como las casas de rango alto y medio con MAE de 292,981,323.27 y 135,366,893.30, aparecen en azul oscuro indicando errores significativos y mayor dificultad de predicción. Los errores moderados, correspondientes a casas de rango bajo y apartamentos de rango alto y bajo, con MAE entre 30,077,154.16 y 54,172,516.29, se representan con un azul más tenue, reflejando un desempeño aceptable, pero con cierto margen de error. Los segmentos sin color, como los apartamentos de rango medio con MAE de 18,187,771.13, indican los errores más bajos, destacando que el modelo predice con mayor precisión este grupo. En conjunto, la

visualización permite identificar de manera clara qué combinaciones de tipo de inmueble y rango de precio presentan mayor o menor calidad de predicción.

Figura 42

Mapa de Calor de MAPE por Tipo de Inmueble y Rango de Precio

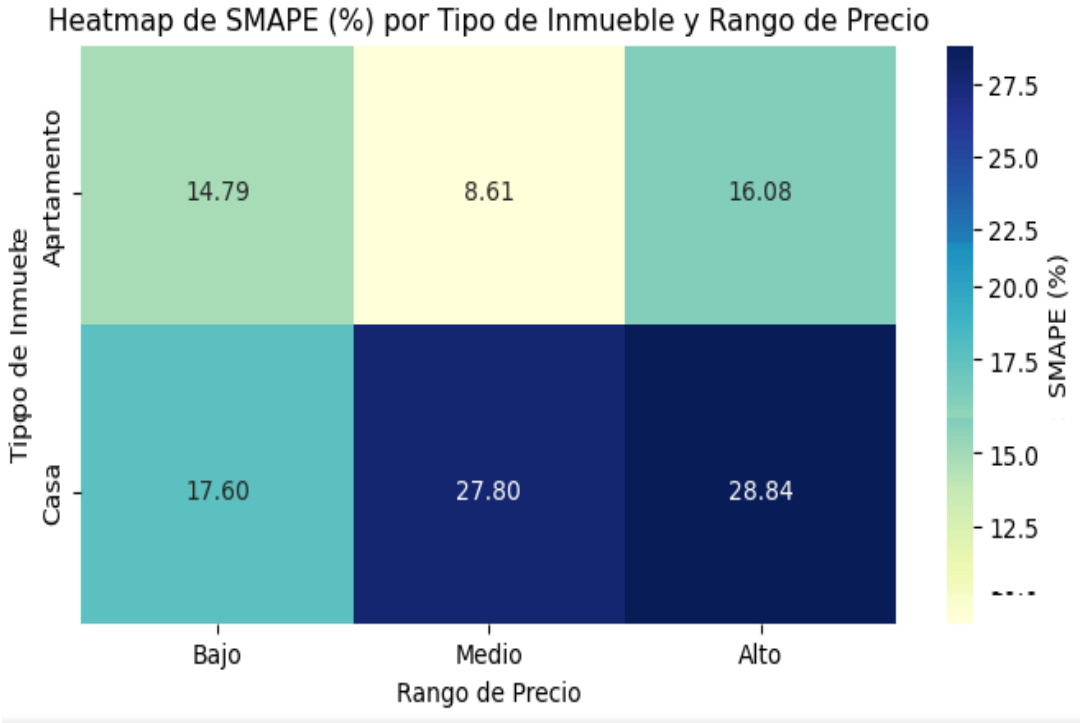


El mapa de calor de MAPE por tipo de inmueble y rango de precio de la figura 42 indica que los errores porcentuales más altos, como en las casas de rango medio y alto con 29.50% y 24.19%, se muestran con color azul intenso, reflejando menor precisión en estos segmentos. Los errores moderados, como en casas de rango bajo y apartamentos de rango bajo y alto con valores entre 15.03% y 21.41%, se representan con azul más tenue, mientras que los apartamentos de rango medio con 8.44% no presentan color, destacando la mayor exactitud del modelo en este grupo. La gráfica anterior nos permite reflexionar en la forma como el modelo es más preciso cuando los segmentos del mercado convergen, es decir cuando el precio de las casas que en su

mayoría es más alto en alguna medida trata de acercarse al valor medio de los inmuebles donde encontramos que el tipo de inmueble apartamento tiene valores, acá en este punto el modelo mejora la precisión.

Figura 43

Mapa de Calor de SMAPE por Tipo de Inmueble y Rango de Precio



El mapa de calor de SMAPE por tipo de inmueble y rango de precio de la figura 43 muestra que los errores relativos más altos se concentran en las casas de rango medio y alto, con 27.80% y 28.84%, representados con azul intenso, indicando menor precisión en estos segmentos. Los errores moderados, como en casas de rango bajo y apartamentos de rango alto y bajo con valores entre 14.79% y 17.60%, se reflejan con azul más tenue, mientras que los apartamentos de rango medio con 8.61% no presentan color, destacando la mayor exactitud del modelo en este grupo.

Fortalezas del Modelo CatBoost

El uso de CatBoost en este contexto ha demostrado un desempeño superior frente a otros modelos evaluados (regresión lineal, Random Forest, XGBoost), especialmente por su capacidad para manejar características categóricas (como tipo de inmueble y estado) sin necesidad de un one-hot encoding exhaustivo, lo cual mantiene la eficiencia y la interpretabilidad. Gracias a la optimización de hiperparámetros (iterations, learning_rate, profundidad, l2_leaf_reg, temperatura de bagging, border_count), el modelo ha alcanzado una elevada precisión (RMSE más bajo, alto R^2) lo que sugiere que es capaz de capturar relaciones complejas y no lineales entre las variables estructurales de las viviendas y su precio. Además, el análisis de importancia de características reveló que variables intuitivamente relevantes como número de baños, área construida y parqueaderos son de hecho las que más contribuyen a la predicción, lo que aporta transparencia y justificación al modelo desde una perspectiva de negocio.

Limitaciones del Modelo CatBoost

Sin embargo, el modelo también presenta limitaciones. El análisis por rangos de precio y tipo de inmueble evidenció que tiene errores muy grandes en casas de rango alto y medio, lo que indica que su desempeño se deteriora cuando las propiedades son costosas o menos representadas en el conjunto de datos. Esta variabilidad sugiere que podría haber una subrepresentación de viviendas de alto valor en tu muestra, o que faltan variables cualitativas (acabados, localización precisa, vista, terreno) que explican una parte significativa del precio para esas casas. Adicionalmente, aunque la validación cruzada mostró una medida aceptable de desempeño, la dispersión de los residuales y la presencia de outliers en las predicciones indican que el modelo no es completamente robusto en todos los escenarios y puede sobreajustarse a ciertos datos.

Implicaciones Practicas del Modelo CatBoost para Inversores

Desde el punto de vista de un inversor en finca raíz en la vereda Canavita, el modelo CatBoost representa una herramienta muy útil para la valoración rápida y eficiente de propiedades, especialmente apartamentos o viviendas de precio medio, donde el error es más bajo. Esto facilita la toma de decisiones sobre compras o ventas, ya que los inversores pueden estimar el valor de mercado con un margen razonable de error, optimizando su análisis de coste-beneficio. No obstante, para propiedades de alto valor, los inversores deben ser cautelosos: el modelo puede subestimar o sobreestimar significativamente, por lo que es recomendable combinar la predicción automatizada con inspecciones físicas o evaluaciones manuales más detalladas. También puede orientarse a usar otras estrategias y diligencia previa: por ejemplo, si el modelo predice un precio muy alto para una casa, el inversor podría requerir un análisis más profundo de las características no cuantitadas (vista, terreno, potencial de revalorización). Es fundamental tener en cuenta que otro factor importante a tener en cuenta por parte de los posibles inversores es la evolución y desarrollo constante que tiene el municipio, pues en muchas zonas rurales donde los tipos de inmuebles ya construidos se identifica que son inmuebles que no se construyeron con una visión a futuro, es decir el propietario en su memento construyó el inmueble para ser habitado por una familia sin tener en cuenta que con el paso del tiempo y la llegada de nuevos residentes al municipio se tendría que pensar en construir zonas para parqueaderos y adecuar espacios internos del inmueble para construir cuartos de estudio, en algunas casas también hay demanda de inmuebles con chimenea. Todas estas características hacen que cada día el desarrollo este transformando las características de los inmuebles ubicados en las zonas rurales y por consiguiente aumentado sus precios. Es por ello que el inversor debe tener en cuenta incluir en su análisis previo el acceso a información de planes de desarrollo, es

decir estar informado si en esta vereda se planean construir nuevas vías, nuevos colegios, o si por parte de alguna empresa se planea ubicar su planta en sus cercanías, ya que estos elementos son importantes también a la hora de incrementar el valor de los inmuebles y pueden ayudar a tomar una decisión correcta a la hora de invertir sus recursos.

Conclusiones

CatBoost se identificó como el modelo más preciso para predecir precios de inmuebles, gracias a su capacidad de manejar variables categóricas y relaciones no lineales, logrando un alto R^2 y bajos errores en general

El análisis de importancia de características reveló que el número de baños, el área construida y los parqueaderos son los factores que más impactan en el precio, lo que permite enfocar decisiones de inversión y valoración en estas propiedades clave.

El modelo CatBoost predice con mayor precisión los apartamentos, especialmente de rango medio, mientras que las casas de alto valor presentan mayores errores, indicando que la dispersión de precios y la menor representación de estas propiedades afectan la confiabilidad del modelo.

Los rangos de precio medio presentan los errores más bajos en todas las métricas evaluadas (RMSE, MAE, MAPE, SMAPE), mientras que los rangos altos, particularmente en casas, muestran errores significativamente mayores, lo que refleja la necesidad de cautela al estimar propiedades de mayor valor.

El análisis de residuales indicó que la mayoría de las predicciones se concentran cerca de la línea cero, evidenciando un buen ajuste general, aunque existen algunos outliers que representan propiedades cuyo precio real difiere considerablemente de la predicción, señalando limitaciones en casos extremos.

El modelo ofrece una herramienta confiable para la valoración rápida de propiedades en Canavita, facilitando decisiones de compra, venta o inversión en el segmento de apartamentos y viviendas de precio medio, mientras que para inmuebles de alto valor se sugiere complementar la predicción con evaluaciones cualitativas y análisis de mercado más detallados.

Recomendaciones

Ampliar la base de datos y mejorar la representatividad de casas de alto valor**,** dado que las casas, especialmente en rangos altos, presentan los mayores errores, es recomendable incorporar más observaciones de este segmento para mejorar la estabilidad del modelo y reducir el efecto de la dispersión de precios.

Incluir nuevas variables estructurales y geoespaciales, factores como proximidad a vías principales, equipamientos cercanos, calidad del entorno, pendiente del terreno o valorización histórica pueden mejorar significativamente la precisión del modelo, especialmente en inmuebles de precio alto.

Explorar modelos avanzados o híbridos**,** aunque CatBoost mostró el mejor desempeño, modelos híbridos (CatBoost + regresión lineal ajustada para residuales), técnicas de stacking o redes neuronales podrían capturar mejores comportamientos complejos presentes en propiedades de mayor valor.

Implementar calibración del modelo por segmentos**,** debido a que el desempeño varía entre apartamentos y casas, y entre rangos de precio, sería útil entrenar modelos específicos para cada segmento, lo que puede reducir errores y generar estimaciones más precisas en contextos heterogéneos.

Desarrollar una herramienta interactiva para usuarios finales, implementar un dashboard o aplicación web donde agentes inmobiliarios o inversionistas puedan ingresar características de un inmueble y recibir una predicción confiable, junto con intervalos de confianza, aumentaría la aplicabilidad práctica del modelo

Referencias

Alaminos, A. F. (2022). *Árboles de decisión en R con Random Forest*. https://rua.ua.es/entities/publication/77ce0f1e-5872-4edb-a70f-0038bd77bb6b

Alfaro, A., y Ospina, J. V. (2021). Revisión sistemática de literatura: Técnicas de aprendizaje automático (Machine Learning). *Cuaderno activa*, *13*(1), 113-121.

Alfaro, J.-L., Cano, E. L., Alfaro-Cortés, E., García, N., Gámez, M., y Larraz, B. (2020). A Fully Automated Adjustment of Ensemble Methods in Machine Learning for Modeling Complex Real Estate Systems. *Complexity*, *2020*(1), 5287263. https://doi.org/10.1155/2020/5287263

Beto, B. (2024). *Predicción del precio de la vivienda mediante aprendizaje automático*. https://uvadoc.uva.es/handle/10324/71534

Botero, A. (2021). *Propuesta para involucrar las consideraciones relacionadas con el mercado inmobiliario en el análisis del ordenamiento territorial.* https://repositorio.unal.edu.co/handle/unal/80082

Bruno, A. (2023). *Análisis Predictivo Del Precio De La Vivienda En Los Distritos De Ciudad Lineal Y La Latina Con Modelos De Machine Learning—Bruno Cueto, Ana*. https://repositorio.comillas.edu/xmlui/handle/11531/68911

Castiblanco, E. M., y Velandia, S. V. (2024). *Análisis integral de avalúos catastrales rurales y su impacto en la gestión socioeconómica en el municipio de Ovejas, Sucre*. http://hdl.handle.net/11349/42672

Cedeño, A. (2023). ---APLICACIÓN DE LA TEORÍA DE GRAFOS A LA OPTIMIZACIÓN DE REDES NEURONALES ARTIFICIALES. *Revista SOCIENCYTEC*, *1*(2), 51-69. https://doi.org/10.61396/276jga94

Charria, F. (2022). Los municipios y la protección constitucional del patrimonio ecológico y cultural en Colombia. *Revista de Derecho de la UNED (RDUNED)*, *30*, 79-108. https://doi.org/10.5944/rduned.30.2022.36842

Choy, L. H. T., y Ho, W. K. O. (2023). *El uso del aprendizaje automático en la investigación inmobiliaria*. https://www.mdpi.com/2073-445x/12/4/740

Díaz, C. (2022). *Efectividad y optimización del recaudo del impuesto predial unificado en el municipio de Úmbita Boyacá*. http://repository.unilibre.edu.co/handle/10901/23158

Díaz, M. A., Ahumada-Cervantes, M. de los A., Melo-Morín, J. P., Díaz-Martínez, M. A., Ahumada-Cervantes, M. de los A., y Melo-Morín, J. P. (2021). Arboles de Decisión como Metodología para Determinar el Rendimiento Académico en Educación Superior. *Revista Lasallista de Investigación*, *18*(2), 94-104. https://doi.org/10.22507/rli.v18n2a8

Dueñas, J. M. (2020). *Aplicación de técnicas de machine learning a la ciberseguridad: Aprendizaje supervisado para la detección de amenazas web mediante clasificación basada en árboles de decisión*. https://hdl.handle.net/10609/118166

Escobar, V. A. (2024). *Estudio multitemporal de la dinámica inmobiliaria de predios comerciales en el sector de Veracruz de la localidad de Santa Fe, Bogotá D.C. (2017-2023)*. http://hdl.handle.net/11349/39156

Escribá Pina, E. (2021). *Aprendizaje por refuerzo mediante Deep Learning para las ciudades inteligentes* [Masters, E.T.S. de Ingenieros Informáticos (UPM)]. https://oa.upm.es/68733/

Espinosa, J. J. (2020). Aplicación de metodología CRISP-DM para segmentación geográfica de una base de datos pública. *Ingeniería, investigación y tecnología*, *21*(1). https://doi.org/10.22201/fi.25940732e.2020.21n1.008

Galeano, P. (2025). *Precios de casas y apartamentos en Colombia para 2025: Así cambian por ciudades*. Portafolio.co. https://www.portafolio.co/mis-finanzas/vivienda/comprar-casa-o-apartamento-en-2025-comparacion-de-precios-por-ciudades-en-colombia-638443

García, P. A. (2021). *Implementación de un modelo machine learning para la estimación del valor del metro cuadrado de un inmueble ubicado en Cundinamarca*. http://hdl.handle.net/1992/55114

Gómez, L. H., Restrepo-Yepes, A. M., Herrán, C., Hernández-Calle, J. A., Gómez-Ospina, L. H., Restrepo-Yepes, A. M., Herrán, C., y Hernández-Calle, J. A. (2024). Bienestar y habitar en la vivienda de interés prioritario en Medellín-Colombia. *Revista INVI*, *39*(110), 138-163. https://doi.org/10.5354/0718-8358.2024.68816

González De La Cruz, A. D. (2022). *Sistema de costos y la fijación de precios en la fábrica de bloques E.J.R. del cantón Salinas, provincia de Santa Elena, año 2021.* https://repositorio.upse.edu.ec/handle/46000/8490

Grajales, Y. V. (2019). *Modelo de predicción de precios de viviendas en el Municipio de Rionegro para apoyar la toma de decisiones de compra y venta de propiedad raíz* [masterThesis, Escuela de Ingenierías]. https://repository.upb.edu.co/handle/20.500.11912/5285

Gutierrez, S., Milla Ramirez, T. S., Rodriguez Ramos, O. O., y Salazar Vera, L. D. (2022). *Análisis global de las metodologías de valorización y guía de criterios para valorizar activos inmobiliarios rentistas*. https://hdl.handle.net/20.500.12640/3061

Jha, S. B., Babiceanu, R. F., Pandey, V., y Jha, R. K. (2020). *Housing Market Prediction Problem using Different Machine Learning Algorithms: A Case Study* (No. arXiv:2006.10092). arXiv. https://doi.org/10.48550/arXiv.2006.10092

Jiménez Oliveros, P. A., y Aguiar Hernández, D. F. (2024). Por la materialización del concepto de vivienda digna. *Bitácora Urbano-Territorial*, *34*(3), 124-140.

Macías, E., y Osorio, N. (2020). *Efectos del confinamiento por pandemia en el comportamiento del mercado inmobiliario colombiano*. http://repository.unilibre.edu.co/handle/10901/22959

Mirjalili, V., y Raschka, S. (2020). *Python Machine Learning*. Marcombo.

Moncada, L. V., Garcés Jaramillo, J. D., y Jiménez Sánchez, M. (2022). *Repercusiones que presenta la actualización catastral en el impuesto predial*. https://hdl.handle.net/10495/30728

Moraleda, F. (2023, mayo). *Análisis de datos socioeconómicos mundiales de economics @intelligence* [Info:eu-repo/semantics/bachelorThesis]. E.T.S. de Ingenieros Informáticos (UPM). https://oa.upm.es/75475/

Nieto, M. A. (2022). Modelo de Aprendizaje Automático para la Predicción de Precios de Vivienda en la Ciudad de Bogotá. *Artículo de Revista*, 68-76. https://doi.org/10.37511/apuntesci.v1n1a6

Park, D., y Ryu, D. (2021). A Machine Learning-Based Early Warning System for the Housing and Stock Markets. *IEEE Access*, *9*, 85566-85572. https://doi.org/10.1109/ACCESS.2021.3077962

Pineda, J. M., y Sosa, I. M. (2024). *Elaboración de zonas de valor similar mediante el cálculo del Índice de Valoración Predial (IVP) en las ciudades de Cartagena, Florencia y Montería y estructuración de información catastral para los avalúos de la Unidad Administrativa Especial de Gestión de Restitución de Tierras Despojadas -URT como contribución a la Subdirección de avalúos del Igac.* http://hdl.handle.net/11349/36395

Porras, W. I. (2024). Implementación de una máquina de soporte vectorial en la clasificación de calidad de ladrillos en la región de Huancayo. *Universidad Peruana Los Andes*. http://repositorio.upla.edu.pe/handle/20.500.12848/8796

Preciado, A. D. (2025). *Modelo predictivo para precios de viviendas en la ciudad de Guayaquil mediante Machine Learning* [B.S. thesis]. https://dspace.ups.edu.ec/handle/123456789/31102

Ramos, D. (2023). *Aplicación del valor de referencia en el sistema impositivo español: Origen, evolución y presente*. https://uvadoc.uva.es/handle/10324/68166

Resolución Nº 1137 de 2024, 11 (2024). https://www.igac.gov.co/sites/default/files/transparencia/normograma/RESOLUCION%201137%20DE%202024.pdf

Rodriguez, D. M. (2024). *El papel de la administración pública en la transformación de los territorios: Metropolización y expansión urbana. Estudio de caso, Tocancipá* [Escuela Superior de Administración Pública - ESAP]. https://hdl.handle.net/20.500.14471/28056

Roque, J. (2021). *Técnicas de selección de variables en regresión lineal múltiple* [masterThesis, Universidad Internacional de Andalucía]. https://dspace.unia.es/handle/10334/6557

Rubio, E. M. (2022). *Complejidad, redes neuronales artificiales y simulación computacional en la investigación científica*. http://repositorio.filo.uba.ar/handle/filodigital/16219

Sánchez, M. G., y Pérez, J. Á. (2023). *Metodología CRISP-DM en la gestión de proyecto de Data Mining. Caso enfermedades dermatológicas*. http://hdl.handle.net/10882/13087

Soto, R. A., y David, E. (2021). *Modelo de Predicción del Precio de la Vivienda en el Valle de San Nicolás*. https://bibliotecadigital.udea.edu.co/server/api/core/bitstreams/d54fe203-c457-45be-909a-8e4a9956bb78/content

Toledo, M. M. (2023). La restitución de tierras en Colombia y su incidencia en el derecho de propiedad sobre bienes inmuebles. *Màster Oficial - Dret de l'Empresa i els Negocis*. https://diposit.ub.edu/dspace/handle/2445/200548

Torres, J. I. S., y Cardenas, E. G. (2021). Análisis y aplicación de algoritmos de minería de datos. *Perspectivas*, *6*(21), 71-88. https://doi.org/10.26620/uniminuto.perspectivas.6.21.2021.71-88

Valero, D. (2023). *Nueva metodología, avances computacionales y aplicaciones a través de Support Vector Machines*. http://dspace.umh.es/handle/11000/31715

Vicarte, E. A. R., y Mayo, A. A. F. (2023). Importancia de la justificación del factor comercial en tasación. *Acreditas*, *9*, 22-26. https://doi.org/10.61752/acd.vi9.136

Vilca, W. A. (2022). *Análisis para la predicción del valor de un bien inmueble en la Ciudad de Quito post pandemia*. https://repositorio.uisek.edu.ec/handle/123456789/4677

Zapata, S. A. (2025). *Ruralidad conurbada. De Vereda a «Barrio Rural». Transformación socio-territorial pos reglamentación del Distrito Rural Campesino*. https://repositorio.unal.edu.co/handle/unal/88857

Zhang, Q. (2021). Housing Price Prediction Based on Multiple Linear Regression. *Scientific Programming*, *2021*(1), 7678931. https://doi.org/10.1155/2021/7678931

Apéndices

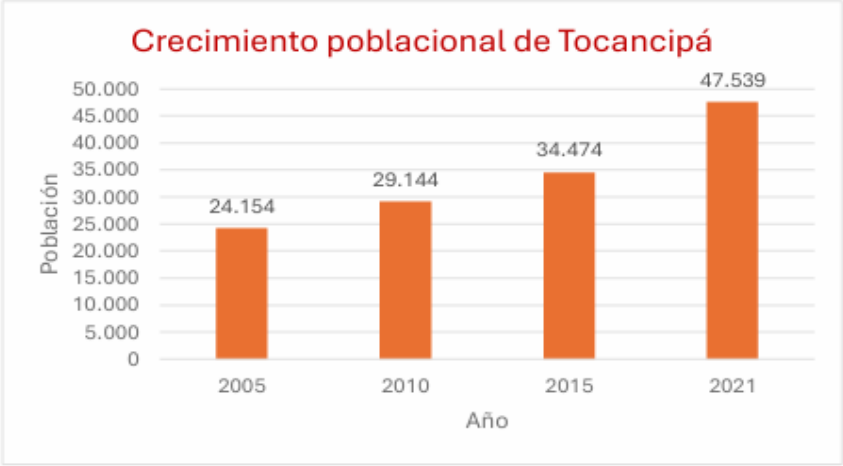
Apéndice A

Histórico de Licencias de Construcción por Destino en Tocancipá

Nota. Tomado de El papel de la administración pública en la transformación de los territorios: metropolización y expansión urbana. Estudio de caso, Tocancipá. (Rodriguez, 2024)

Apéndice B

Crecimiento Poblacional de Tocancipá 2005-2021



Nota. Tomado de El papel de la administración pública en la transformación de los territorios: metropolización y expansión urbana. Estudio de caso, Tocancipá. (Rodriguez, 2024)

Apéndice C

Precio vs Vivienda del Conjunto de Datos de Finca Raíz

Apéndice D

Enlace al Conjunto de Datos del Proyecto

https://unadvirtualedu-my.sharepoint.com/:x:/g/personal/spalacioss_unadvirtual_edu_co/EYYLjIq0Jl5Aixp55sQst6QB1v9BaBziF16m1lNxQrSNOQ?e=AhugxX

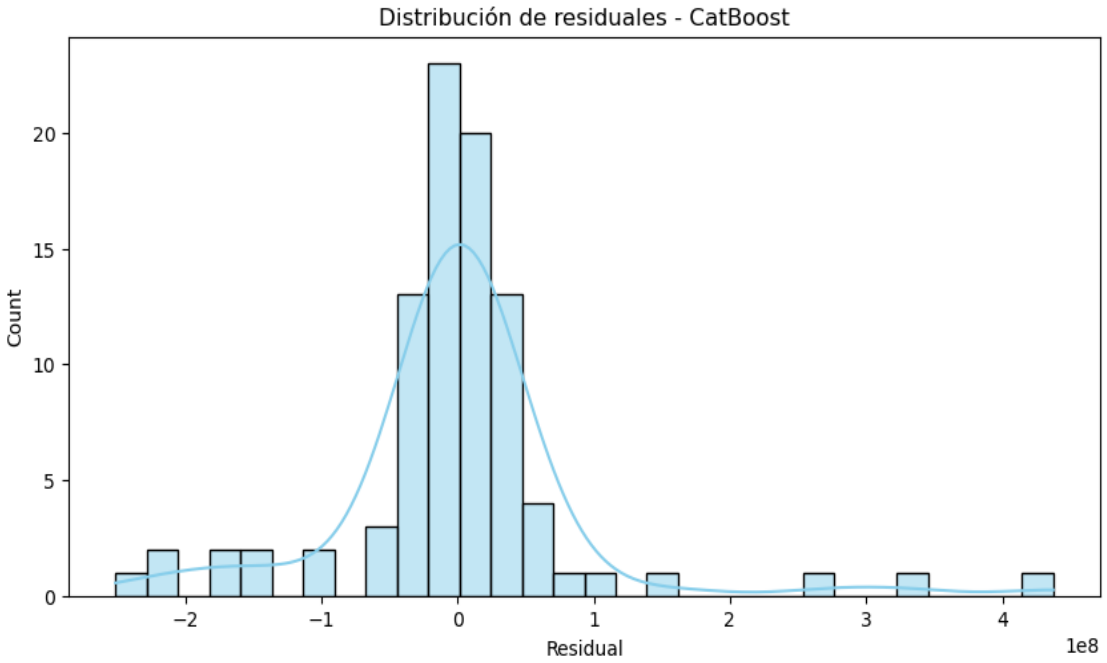
Apéndice E

Enlace al Código Utilizado para Implementar Modelos de Predicción

https://github.com/salopalacios/Mis-Proyectos/blob/2e5a58af55a218254f27939d317be9342ddb1aa6/Implementacion_de_modelos.ipynb

Apéndice F

Histograma de los Residuales del Modelo CatBoost



Apéndice G

Enlace al Video de la Presentación del Proyecto

https://youtu.be/gxDkYkVRjoE