TRABAJO DE GRADO II

SMARTPRICE INSIGHTS

SISTEMA BASADO EN INTELIGENCIA ARTIFICIAL PARA LA GENERACIÓN DE INSIGHTS DE PRECIOS EN EL MERCADO INMOBILIARIO DE HOMTY

Autores:

Yesid Humberto Montaño Cuero

Javier Ricardo Muñoz Castillo

Asesor

Ing. Ian Mateo Rodriguez

MAESTRÍA EN CIENCIA DE DATOS

DICIEMBRE 2024

Contenido 2

Introducción

Actualmente, el sector de bienes raíces enfrenta desafíos significativos en la toma de decisiones informadas, principalmente por la escasez de análisis de datos históricos y la falta de comprensión de las tendencias de precios. Este trabajo busca abordar esta problemática mediante el uso de técnicas avanzadas de ciencia de datos. La empresa Homty, dedicada a la recopilación y análisis de información sobre más de 600,000 inmuebles en Colombia, ha identificado una oportunidad crucial para mejorar la rentabilidad de las inversiones en este sector. Este trabajo de grado se propone explorar la implementación de técnicas de ciencia de datos e inteligencia artificial para analizar el comportamiento histórico de los precios de inmuebles residenciales, como apartamentos y casas, en las principales ciudades del país, incluyendo Bogotá, Cali y Medellín.

El objetivo principal de este estudio es desarrollar un modelo que no solo permita entender las fluctuaciones de precios en el pasado, sino que también permita proyectar tendencias futuras, facilitando así decisiones más acertadas para los inversionistas. Además, se busca correlacionar los datos de precios con variables demográficas, como el crecimiento poblacional, el turismo y la capacidad económica, que son fundamentales para evaluar el potencial de inversión en diferentes zonas. A través de este enfoque, se espera contribuir al desarrollo de herramientas analíticas que optimicen la estrategia de inversión en el sector inmobiliario, promoviendo un uso más eficiente de los recursos y una mejor comprensión del mercado.

Este trabajo se fundamentará en la recopilación de datos, el análisis estadístico y la aplicación de algoritmos de inteligencia artificial, con el fin de ofrecer un marco teórico y práctico que respalde la propuesta de Homty. La relevancia de este estudio radica en su potencial para transformar la manera en que se toman decisiones en el sector inmobiliario, alineando la oferta y la demanda de manera más efectiva y, en última instancia, impulsando el crecimiento económico en el país.

Planteamiento del Problema

A pesar de la creciente importancia del sector inmobiliario en Colombia, la industria enfrenta desafíos significativos en la valoración y análisis de precios de inmuebles, especialmente en el contexto de datos históricos y tendencias del mercado. Homty, se encuentra con las siguientes limitaciones:

Inexistencia de Análisis de Datos Históricos: La falta de un análisis sistemático de los datos históricos de precios de inmuebles impide a los inversionistas y a la empresa misma comprender las tendencias del mercado. Sin esta información, las decisiones de inversión se basan en suposiciones y no en datos concretos.

Dificultad en la Predicción de Precios: La incapacidad para predecir el comportamiento futuro de los precios de los inmuebles limita la capacidad de los inversionistas para planificar y maximizar sus retornos. Esto se ve agravado por la falta de herramientas automatizadas que integren variables relevantes para la valoración.

Desconexión con Variables Demográficas: La ausencia de correlación entre los precios de los inmuebles y factores demográficos, como el crecimiento poblacional y la capacidad económica, dificulta la comprensión de la dinámica del mercado. Esta desconexión puede llevar a decisiones de inversión erróneas y a una subestimación de la demanda en ciertas áreas.

Oportunidad de Innovación Tecnológica: A pesar de la disponibilidad de tecnologías avanzadas, como la inteligencia artificial, su implementación en el sector valuatorio es escasa. Esto representa una oportunidad no solo para mejorar los procesos de valoración, sino también para establecer un marco más sólido y actualizado en la industria.

Impacto en el Desarrollo Económico: La ineficiencia en la valoración de inmuebles no solo afecta a los inversionistas individuales, sino que también tiene repercusiones más amplias en el desarrollo económico del país. La modernización de estos procesos podría contribuir a una mayor competitividad y a un crecimiento sostenible en el sector inmobiliario.

Pregunta de Investigación

¿De qué manera pueden las técnicas de ciencia de datos e inteligencia artificial optimizar el análisis y la predicción de tendencias de precios de inmuebles en las principales ciudades de Colombia, y cuáles son los factores demográficos que influyen en estas variaciones?

Justificación

El sector inmobiliario en Colombia enfrenta desafíos significativos relacionados con la falta de herramientas analíticas avanzadas que permitan evaluar de manera precisa las dinámicas de precios y las oportunidades de inversión. En este contexto, Homty, una empresa dedicada a la recopilación y análisis de datos de propiedades inmobiliarias, se encuentra en una posición estratégica para transformar la forma en que se toman decisiones en el mercado.

La implementación de este proyecto no solo responde a las necesidades internas de Homty de optimizar sus procesos de análisis de precios, sino que también posicionará a la empresa como líder en innovación dentro del sector. Al desarrollar un sistema basado en inteligencia artificial y ciencia de datos, Homty podrá:

- Ofrecer Servicios Diferenciados: Proveer análisis predictivos detallados que permitan a los inversionistas tomar decisiones más informadas, incrementando su confianza en los servicios de Homty y fortaleciendo su posicionamiento en el mercado.

- Optimizar Estrategias Comerciales: Identificar tendencias de precios y correlacionarlas con factores demográficos clave permitirá a Homty ajustar su oferta de servicios a las dinámicas cambiantes del mercado, maximizando sus oportunidades de negocio.

- Generar Valor Añadido para los Clientes: La capacidad de proyectar tendencias futuras y analizar variables económicas y sociales otorga a Homty una ventaja competitiva, ya que sus clientes podrán reducir riesgos en sus inversiones y mejorar el rendimiento de sus portafolios.

- Aumentar la Competitividad en el Mercado: En un sector donde la innovación tecnológica es limitada, el uso de algoritmos avanzados y herramientas de visualización interactiva posicionará a Homty como un referente en la aplicación de tecnologías emergentes para resolver problemas reales del mercado inmobiliario.

- Promover la Transparencia del Mercado: Al proporcionar análisis basados en datos confiables y modelos robustos, Homty contribuye a un mercado inmobiliario más transparente, lo que puede atraer a más actores interesados en invertir.

En conclusión, este proyecto representa una oportunidad única para Homty de consolidarse como un líder en análisis de datos inmobiliarios, promoviendo una mejor toma de decisiones en el sector y garantizando una ventaja competitiva sostenible frente a sus competidores. Esto no solo fortalecerá su posición en el mercado, sino que también impactará positivamente en el desarrollo del sector inmobiliario en Colombia.

Objetivos

Objetivo General

Desarrollar un sistema que mediante técnicas de ciencia de datos e inteligencia artificial permite analizar el comportamiento histórico y proyectar tendencias futuras de los precios de inmuebles en las principales ciudades de Colombia, mejorando así la precisión de la valoración y la toma de decisiones en el sector inmobiliario para Homty y sus clientes.

Objetivos Específicos

1. Recopilar, estructurar y seleccionar variables relevantes de la base de datos histórica de Homty de precios de inmuebles residenciales (apartamentos y casas) en Bogotá, Cali, y Medellín.

2. Implementar algoritmos de inteligencia artificial para identificar patrones y tendencias en los datos históricos de precios de inmuebles.

3. Diseñar una herramienta analítica o de visualización que integre las bases de datos de Homty para generar Insights de valor sobre el mercado inmobiliario.

Estado del Arte

Cruz Paredes, Gabriela Pilar (2024). "Predicción de Ventas de Departamentos en el Distrito de Miraflores de una Empresa Inmobiliaria de Lima Utilizando el Modelo de Ensamble por Medias". Trabajo de suficiencia profesional, Universidad Nacional Agraria La Molina. Este trabajo se centra en la predicción de ventas de departamentos en Miraflores, analizando factores económicos y financieros que afectan el mercado inmobiliario en Lima.

Hernandez Parrado, Carlos Adrián & Dávila Martínez, Juan Pablo (2024). "Housing Price Prediction in Colombia using Machine Learning". Este estudio desarrolla un modelo de machine learning para predecir precios de vivienda en Colombia, utilizando técnicas como regresión de árbol y LightGBM, lo que es útil para aplicar métodos de aprendizaje automático en el análisis de precios de propiedades.

Kraus, J. (2019). "Comparación de Métodos de Predicción de Precios de Propiedades". Este trabajo compara la capacidad predictiva de diferentes métodos, incluyendo regresión lineal, árboles de regresión, random forest y bagging, en el contexto colombiano, proporcionando una base para seleccionar métodos adecuados para la predicción de precios.

MV Perception (2022). "Impacto de la Pandemia en el Mercado Inmobiliario". Este trabajo menciona cómo la pandemia del COVID-19 afectó el mercado inmobiliario en Perú, lo que puede ser un contexto importante para tu investigación, especialmente si estás analizando el impacto de eventos externos en el mercado.

Reim, A. (2020). "Estudio de Tendencias del Mercado Inmobiliario". Este trabajo analiza las tendencias del mercado inmobiliario y cómo factores económicos y políticos influyen en la demanda y precios de las viviendas, útil para contextualizar tu investigación dentro de un marco más amplio de tendencias del mercado.

Sicilia Gómez, B. (2024). "Desarrollo de modelos de predicción de precios inmobiliarios utilizando técnicas de machine learning" Trabajo de fin de grado, Universidad Nacional Agraria La Molina. Este trabajo investiga el uso de técnicas de aprendizaje automático en el mercado inmobiliario, enfocándose en la predicción de precios de venta de viviendas en Madrid. Se desarrollaron varios modelos de Machine Learning y Deep Learning utilizando datos del portal inmobiliario Idealista. La metodología incluye la extracción de datos, limpieza, modelado y evaluación de los modelos.

Roig Hernando, J., Gras Alomà, R., & Soriano Llobera, J. M. (2024). "**Análisis y pronóstico del precio de la vivienda en España".** Modelo econométrico desde una perspectiva conductual. Este documento desarrolla un modelo econométrico del ciclo inmobiliario en el mercado residencial español. A diferencia de otros modelos, incorpora un enfoque conductual que considera la influencia de inversores irracionales en los precios. El estudio analiza cómo estos factores afectan la evolución de los precios y proporciona una base para prever tendencias futuras.

Datsko, Artem (2024). "**Análisis y predicción del precio de la vivienda en Madrid"** Este estudio emplea varias técnicas de inteligencia artificial, incluyendo regresión lineal, K-Nearest Neighbors, Gradient Boosting Regressor, Adaptive Boosting, Extreme Gradient Boosting, Random Forest y Ridge, para predecir el precio de la vivienda en Madrid. El trabajo se centra en comparar la efectividad de estos métodos y determinar cuál ofrece mejores resultados en términos de precisión predictiva.

La industria inmobiliaria ha experimentado un cambio significativo en la forma en que se analizan y predicen los precios de las propiedades. Con el avance de la tecnología y la disponibilidad de grandes volúmenes de datos, se han desarrollado diversas metodologías que utilizan técnicas de machine learning y análisis de datos para mejorar la precisión de las predicciones de precios. Este documento presenta un estado del arte que contextualiza la propuesta de Homty, enfocándose en el análisis de datos históricos y la implementación de inteligencia artificial en el sector inmobiliario.

Marco Teórico

En el análisis de datos y el sector inmobiliario, existen diversas teorías y conceptos relevantes que ayudan a entender y modelar el comportamiento del mercado, así como a realizar predicciones precisas sobre precios y tendencias. A continuación, se presentan algunos de los más importantes:

Teoría Hedónica

Esta teoría se basa en la idea de que el precio de un bien (como una propiedad) se puede descomponer en sus características individuales. Por ejemplo, el precio de una vivienda puede depender de factores como el tamaño, la ubicación, el número de habitaciones, y las características del

vecindario. Se utiliza comúnmente en modelos de regresión para estimar el valor de propiedades en función de sus atributos, permitiendo a los analistas entender cómo cada característica influye en el precio final.

Análisis de Series Temporales

Esta técnica se utiliza para analizar datos que se recogen a lo largo del tiempo. Permite identificar patrones, tendencias y estacionalidades en los precios de las propiedades. Es útil para prever cambios en el mercado inmobiliario, como fluctuaciones estacionales en la demanda o cambios a largo plazo en los precios.

Machine Learning y Modelos Predictivos

El uso de algoritmos de machine learning, como árboles de decisión, random forest y redes neuronales, permite crear modelos que pueden aprender de los datos históricos y hacer predicciones sobre precios futuros. Estos modelos pueden manejar grandes volúmenes de datos y encontrar patrones complejos que no son evidentes a través de métodos estadísticos tradicionales.

Big Data

Se refiere al manejo y análisis de grandes volúmenes de datos que pueden incluir información de múltiples fuentes, como datos demográficos, económicos y de comportamiento del consumidor. En el sector inmobiliario, el análisis de big data puede ayudar a identificar tendencias de mercado, segmentar clientes y optimizar estrategias de marketing.

Análisis de Regresión

Esta técnica estadística se utiliza para modelar la relación entre una variable dependiente (por ejemplo, el precio de una propiedad) y una o más variables independientes (como características de la propiedad o condiciones del mercado). Permite a los analistas cuantificar el impacto de diferentes factores en el precio de las propiedades y hacer proyecciones basadas en esos modelos.

Teoría del Comportamiento del Consumidor

Esta teoría estudia cómo los consumidores toman decisiones de compra, incluyendo factores psicológicos, sociales y económicos que influyen en sus elecciones. Comprender el comportamiento del consumidor es crucial para el sector inmobiliario, ya que ayuda a las empresas a diseñar estrategias de marketing efectivas y a anticipar la demanda.

Modelos de Oferta y Demanda

Estos modelos económicos básicos analizan cómo la oferta de propiedades y la demanda de los consumidores interactúan para determinar los precios en el mercado inmobiliario. Ayudan a entender cómo factores como el crecimiento poblacional, la economía local y las tasas de interés afectan el mercado inmobiliario.

Análisis de Sentimiento

Esta técnica implica el uso de datos de redes sociales y otras plataformas para evaluar las percepciones y opiniones del público sobre el mercado inmobiliario. Puede proporcionar información valiosa sobre la confianza del consumidor y las tendencias emergentes en el mercado.

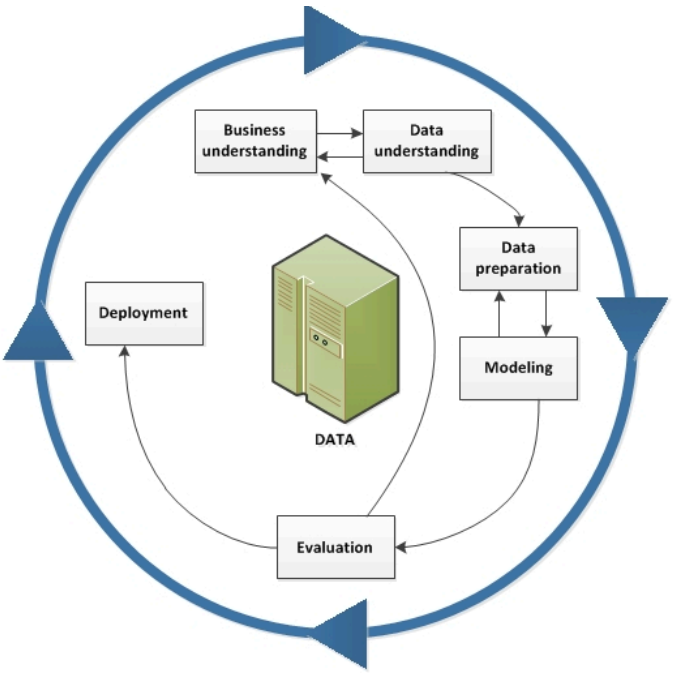
Estos conceptos y teorías son fundamentales para el análisis de datos en el sector inmobiliario, ya que permiten a los profesionales del área tomar decisiones informadas, optimizar estrategias de inversión y anticipar cambios en el mercado. La integración de estas teorías con técnicas avanzadas de análisis de datos, como machine learning y big data, puede mejorar significativamente la precisión de las predicciones y la comprensión del comportamiento del mercado.

Metodología

La metodología (Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. 2000). CRISP-DM (Cross-Industry Standard Process for Data Mining) es elegida para el proyecto debido a su estructura flexible que guía a través de las fases de minería de datos, asegurando que el trabajo esté alineado con los objetivos del negocio y adaptándose a diferentes contextos industriales. Su enfoque en la calidad de los datos es crucial para generar modelos precisos y confiables, mientras que su componente de revisión y aprendizaje permite la mejora continua en proyectos futuros. Además, promueve la documentación y comunicación efectiva de resultados, facilitando la toma de decisiones informadas basadas en datos, lo que maximiza las posibilidades de éxito en la minería de datos.

Figura 1

Flujo de metodología CRISP-DM:



Nota. Tomado de https://portal.amelica.org/ameli/journal/602/6023721001/html/. Jiménez, S., & Merino, A. (2023). Modelos de Aprendizaje Automático basados en CRISP-DM para el Análisis de los niveles de Depresión en los estudiantes de la Escuela Politécnica Nacional. Latin-American Journal of Computing, 10(1).

Aplicación de la Metodología CRISP-DM en el desarrollo de este proyecto:

1. Comprensión del Negocio

1.1 Objetivos del Proyecto

Se definen claramente los objetivos del proyecto en el contexto del negocio, para mejorar la percepción del cliente sobre los productos de Homty y establecer métricas para medir el éxito.

1.2 Contexto del Negocio

Se desarrolla un entendimiento del entorno en el que opera Homty, incluyendo sus productos, clientes y competidores, para alinear los esfuerzos de minería de datos con las necesidades del negocio.

2. Comprensión de los Datos

2.1 Recolección de Datos

Se identifican y recolectan los datos relevantes que se utilizarán en el análisis.

2.2 Exploración Inicial

Se realiza un análisis exploratorio de los datos para entender su estructura, calidad y características, identificando valores faltantes, outliers y correlaciones.

3. Preparación de los Datos

3.1 Limpieza de Datos

Se depuran duplicados, corrigen errores y valores faltantes para asegurar la calidad de los datos en el modelado.

3.2 Transformación de Datos

Se transforman los datos según sea necesario, aplicando normalización y escalado, y añadiendo nuevas variables.

4. Modelado

4.1 Selección de Técnicas de Modelado

Se definen las técnicas de modelado a emplear.

4.2 Construcción del Modelo

Se desarrollan y entrenan distintos modelos utilizando los datos preparados, ajustando los hiperparámetros para optimizar sus rendimientos.

5. Evaluación y selección del modelo

5.1 Evaluación y selección del Modelo

Se evalúa el rendimiento de los modelos utilizando métricas adecuadas y se selecciona el mejor modelo a partir de un ranking de resultados y conversación con el negocio.

5.2 Revisión de Resultados

Se analizan los resultados y se determina si el modelo proporciona información útil y relevante. Si los resultados no son satisfactorios, se considerará volver a fases anteriores para ajustar el enfoque.

Análisis Exploratorio de Datos

El análisis, procesamiento y limpieza de los datos se desarrolló en colaboración con el dueño del negocio, con el objetivo de asegurar la relevancia y calidad de la información utilizada para el análisis.

En este proyecto se utilizó Google Drive como repositorio de los archivos para integrarlos al entorno de Google Colab, esto permitió un acceso rápido y seguro a los datos necesarios para el análisis. Nuestros datos fueron importados desde un archivo CSV, en la primera etapa se analizaron las estrategias específicas para manejar los valores nulos, dependiendo del impacto en el análisis.

Descripción del Conjunto de Datos

Este conjunto de datos recoge información detallada sobre propiedades inmobiliarias desde el mes de septiembre de 2023 hasta septiembre de 2024 para un total de 38.000 registros. Incluyendo características clave como antigüedad, ubicación, precio, características físicas, y estado de publicación. Cada registro representa una propiedad individual y está asociado a múltiples atributos relevantes que permiten un análisis profundo del mercado inmobiliario.

Las columnas de la base de datos, incluye identificadores únicos, detalles sobre las características físicas del inmueble (número de habitaciones, baños, área en metros cuadrados), coordenadas geográficas, información sobre el precio de venta o arriendo, entre otros. Adicionalmente, se proporciona el estado de la propiedad en cuanto a su publicación y verificación en la base de datos.

Las siguientes secciones describen en detalle las columnas disponibles. Nuestra base de datos contiene un total de 31 columnas (Variables):

1. **id** (Entero): Identificador único numérico de la propiedad, utilizado para diferenciar cada registro en la base de datos.
2. **Antiguedad** (Cadena de texto): Describe el rango de años que tiene el inmueble, clasificado en categorías como "1 a 8 años", "menor a 1 año", etc.
3. **Baños** (Entero): Número total de baños que posee el inmueble.
4. **Barrio** (Cadena de texto): Nombre del barrio o sector donde se encuentra la propiedad.
5. **Colegaje** (Cadena de texto): Indicador de si la propiedad está asociada a un colegaje. Se utiliza una cadena de texto que podría incluir valores como "N" o "Sí".

6. **Descripcion** (Cadena de texto): Texto que proporciona detalles descriptivos adicionales sobre la propiedad. En muchos casos, esta columna está vacía.
7. **Estrato** (Entero): Nivel socioeconómico asignado a la propiedad (usualmente en Colombia se clasifican en un rango de 1 a 6).
8. **Fecha_creacion** (Fecha): Fecha en que se registró la propiedad por primera vez en la base de datos.
9. **Fecha_modificacion** (Fecha): Fecha de la última modificación o actualización de los detalles del registro de la propiedad.
10. **Habitaciones** (Entero): Número total de habitaciones con las que cuenta el inmueble.
11. **Inmueble** (Cadena de texto): Tipo de inmueble al que pertenece la propiedad, como "apartamento", "casa", etc.
12. **Municipio** (Cadena de texto): Municipio o ciudad donde se encuentra la propiedad.
13. **nombre_usuario** (Cadena de texto): Nombre del usuario o vendedor asociado a la propiedad. Esta columna está completamente vacía.
14. **Origen** (Cadena de texto): Describe el origen del inmueble, que podría referirse a una clasificación interna de la base de datos.
15. **Parking** (Entero): Número de espacios de estacionamiento que tiene la propiedad.
16. **Piso** (Entero): Número de piso en el que está ubicada la propiedad (relevante para apartamentos o edificios de varios niveles).
17. **Precio** (Flotante): Precio de venta o arriendo de la propiedad en la moneda local.
18. **Sector** (Cadena de texto): Subdivisión o zona más específica dentro de un barrio o municipio.
19. **Tamaño** (Flotante): Tamaño de la propiedad en metros cuadrados.
20. **Telefono** (Cadena de texto): Número de teléfono de contacto para obtener más información sobre la propiedad. Esta columna está vacía.
21. **URL** (Cadena de texto): Enlace a la página web o anuncio en línea de la propiedad.
22. **Valor_Admin** (Flotante): Valor de la cuota administrativa mensual del inmueble (relevante para propiedades en edificios o conjuntos residenciales).

23. **Vendedor** (Cadena de texto): Nombre del vendedor o empresa encargada de gestionar la venta o arriendo de la propiedad.
24. **Whatsapp** (Entero o Cadena de texto): Número de contacto de WhatsApp del vendedor o agente encargado de la propiedad.
25. **Fecha_adicion** (Fecha): Fecha en la que se añadió algún dato adicional a la información del registro.
26. **Longitud** (Flotante): Coordenada geográfica de longitud de la ubicación de la propiedad, útil para mapeo geoespacial.
27. **Latitud** (Flotante): Coordenada geográfica de latitud de la ubicación de la propiedad.
28. **Duplicado** (Entero): Indicador binario (0 o 1) que señala si la propiedad es un duplicado en la base de datos.
29. **Verificado** (Entero): Indicador binario (0 o 1) que indica si la propiedad ha sido verificada.
30. **Fecha_verificacion** (Fecha): Fecha en la que se verificó la información de la propiedad (si aplica).
31. **Imagenes** (Cadena de texto): Enlaces a las imágenes asociadas con la propiedad.

Análisis de Datos Faltantes

A continuación se muestra un análisis de la calidad de los datos, indicando la cantidad de valores vacíos y el porcentaje de registros faltantes para cada variable en el conjunto de datos.

- **Variables sin valores vacíos (0% faltantes):** id, Antiguedad, Baños, Barrio, Colegaje, Estrato, Fecha_creacion, Habitaciones, Inmueble, Municipio, Origen, Parking, Precio, Sector, Tamaño, URL, Valor_Admin, Longitud, Latitud, Imágenes. Estas variables están completamente pobladas, lo que significa que todos los registros tienen datos para estas columnas. Este es un buen indicador de la calidad de los datos para estas variables.
- **Variables con pocos valores vacíos (menos del 1% faltante):** Fecha_modificacion (0,72%), Sector (0.81%). Aunque algunos datos están ausentes, este porcentaje es bajo y probablemente no afecte significativamente el análisis.
- **Variables con vacíos moderados (entre 1% y 10% faltante):** Estrato (2,38%), Vendedor (99,19%), Whatsapp (7,69%). Estas variables tienen vacíos moderados, lo

que puede influir en los resultados dependiendo de la importancia que tengan en el análisis. Whatsapp y Vendedor son particularmente importantes si el análisis está orientado a la interacción con clientes o ventas.

- **Variables con un alto porcentaje de vacíos (más del 90% faltante):** Descripcion (99,19%), Piso (99,91%), Telefono (100%), Fecha_verificacion (100%), nombre_usuario (100%). Estas variables presentan una cantidad extremadamente alta de valores vacíos. En las secciones siguientes se abordará el tratamiento más adecuado para estas variables.

Manejo de Valores Faltantes

En primer lugar, se realizó un análisis detallado para identificar columnas con altos porcentajes de valores nulos. Aquellas columnas con más del 90% de valores nulos fueron eliminadas, ya que su utilidad para el análisis era limitada. Posteriormente, se eliminaron las filas que contienen valores nulos en las columnas restantes, asegurando un conjunto de datos completo y consistente que sirviera como base para los análisis posteriores.

Filtrado de Columnas y Filas

En esta etapa, se descartaron columnas redundantes o irrelevantes para el análisis, tales como:

- id: Un identificador único que no aportaba al análisis estadístico.
- Fechas de registro (Fecha_creacion y Fecha_modificacion): Estas no influyen en las variables de interés.
- Otras columnas no esenciales (Origen, URL, Whatsapp, entre otras): Eliminadas para reducir la dimensionalidad del conjunto de datos.

Para enfocar aún más el análisis, se filtraron los registros con base en criterios específicos, considerando únicamente:

- Tipo de inmueble: Apartamento o casa.
- Municipios: Registros de las ciudades de Bogotá, Medellín y Cali.

Esta selección permitió trabajar con un subconjunto de datos más relevante y representativo para los objetivos del estudio.

Conversión de Tipos de Datos

Se realizó una conversión de columnas clave a tipo entero, lo que facilitó el análisis y procesamiento eficiente de los datos. Las columnas transformadas incluyeron:

- Baños
- Habitaciones
- Parking
- Precio
- Tamaño
- Valor_Admin

Esta estandarización garantizó la consistencia en el manejo de datos numéricos durante los análisis.

Transformación y Escalamiento de Variables

Se llevaron a cabo transformaciones para normalizar y ajustar las escalas de las columnas monetarias y de tamaño, lo que permitió facilitar su interpretación:

- Las columnas **Precio, Tamaño y Valor_Admin** fueron divididas por 100 para estandarizar sus magnitudes.
- Adicionalmente, el valor de **Precio** fue ajustado restándole el valor de la administración (**Valor_Admin**), con el objetivo de obtener un costo neto más representativo de la propiedad.

Estas transformaciones aseguraron que los datos reflejarán de manera más precisa la realidad y redujeron posibles distorsiones en el análisis.

Identificación y Eliminación de Valores Atípicos

Los valores atípicos fueron detectados y eliminados mediante el método del Rango Intercuartílico (IQR). Este proceso se llevó a cabo en las columnas más relevantes, como Baños, Habitaciones, Parking, Precio y Tamaño.

Selección de Columnas Numéricas

Inicialmente, se identificaron las columnas numéricas relevantes en el conjunto de datos: **Banos, Parking, Habitaciones, Precio, Valor_admin y Tamano.** Estas variables fueron seleccionadas para el análisis de outliers, ya que representan características clave de las propiedades y, en consecuencia, su valor atípico podría impactar de forma significativa en el análisis y en los modelos de predicción a desarrollar en el trabajo de grado.

Cálculo de Límites para Outliers Usando el Método del IQR

Para cada columna numérica, se calcularon el primer cuartil (Q1) y el tercer cuartil (Q3) para determinar el **Rango Intercuartílico (IQR)**, definido como la diferencia entre Q3 y Q1. A partir de este valor, se establecieron los límites inferior y superior para los outliers de acuerdo con la siguiente fórmula:

a. **Límite Inferior** = Q1 - 1.5 * IQR

b. **Límite Superior** = Q3 + 1.5 * IQR

Este enfoque considera como outliers los valores que se encuentran significativamente fuera del rango intercuartílico y ayuda a reducir la influencia de estos en los análisis posteriores.

Filtrado de Outliers

Para cada variable, se aplicó un filtro que selecciona únicamente los datos dentro de los límites establecidos, creando una versión limpia del conjunto de datos. Este proceso garantiza que solo los valores dentro de un rango razonable (es decir, aquellos que no son atípicos) permanezcan en el conjunto de datos, minimizando el sesgo potencial de los valores extremos.

Visualización del Impacto del Filtrado de Outliers

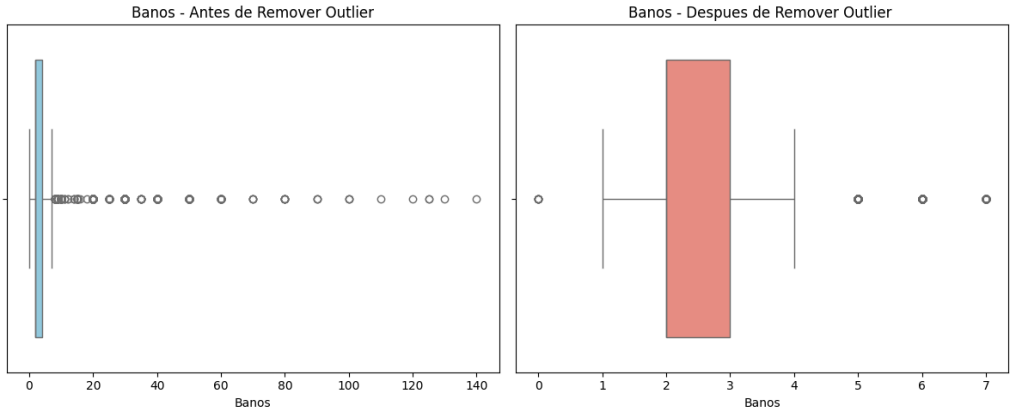
Se generaron gráficos de caja (boxplots) para visualizar la distribución de los datos antes y después de la eliminación de outliers. Los gráficos antes del filtrado permitieron observar la presencia y distribución de valores extremos, mientras que los gráficos después del filtrado reflejaron una representación de los datos más ajustada y sin influencias de valores extremos.

Baños. En el gráfico de la izquierda de la Figura 2 se puede observar una gran cantidad de valores extremos dispersos hacia la derecha, alcanzando hasta 140 baños en un inmueble. La caja central, que representa el rango intercuartil (IQR), es extremadamente pequeña en comparación con la amplitud de los datos, indicando una distribución sesgada. Este comportamiento sugiere que algunos registros contienen errores o valores irreales (e.g., 60, 100, o más baños en una propiedad residencial),

tras aplicar la eliminación de los valores atípicos, la distribución de Baños se concentra en un rango más realista (entre 1 a 4), En el gráfico de la derecha podemos observar la caja más ventral y los bigotes ahora reflejan mejor la dispersión típica esperada para este tipo de variable. Aunque todavía quedan algunos puntos ligeramente alejados, estos no representan una distorsión significativa de los datos.

Figura 2

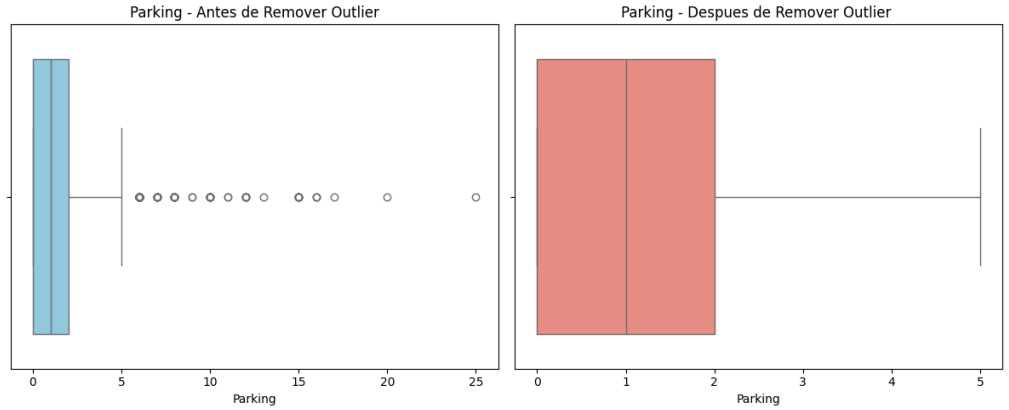
Gráficos de la columna Baños



Parking. En la Figura 3 que representa la variable Parking (plazas de parqueo) antes y después de ajustar los datos eliminando valores atípicos. Al principio, la mayoría de los datos se concentraban entre 0 y 3 plazas, pero también aparecían valores extremos que llegaban hasta 25, algo poco común en propiedades residenciales. Después de eliminar estos outliers, los datos quedaron en un rango más coherente, entre 0 y 5 plazas, con una mediana cercana a 1. Este ajuste hace que la distribución sea más representativa de la realidad, permitiendo un análisis más confiable y preciso.

Figura 3

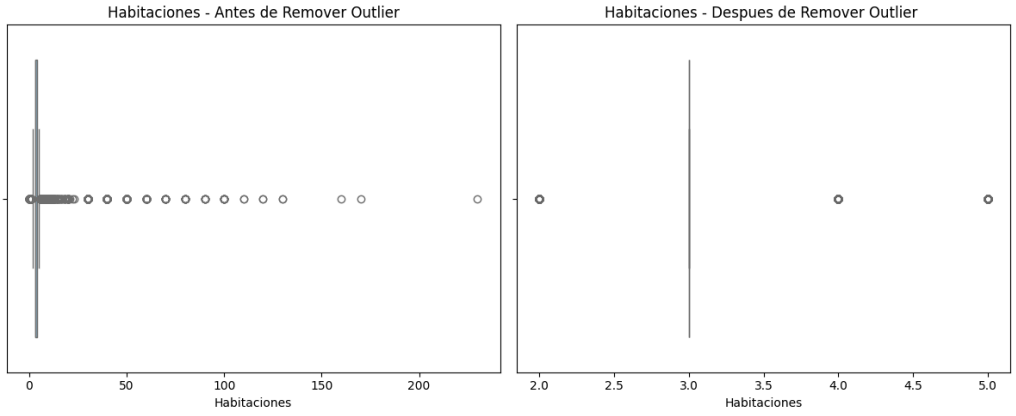
Gráficos de la columna Parking



Habitaciones. En la Figura 4 se muestra un cambio significativo en la distribución después de eliminar los valores atípicos. Antes de la eliminación, la distribución era altamente sesgada hacia la derecha, con varios valores extremadamente altos que afectan significativamente la media y la varianza. Al eliminar estos outliers, la distribución se vuelve mucho más centrada y simétrica, lo que sugiere que la mayoría de las propiedades tienen un número de habitaciones más moderado. Esta transformación en la distribución permitirá realizar análisis estadísticos más precisos y relevantes, ya que los resultados no estarán tan influenciados por unos pocos valores extremos.

Figura 4

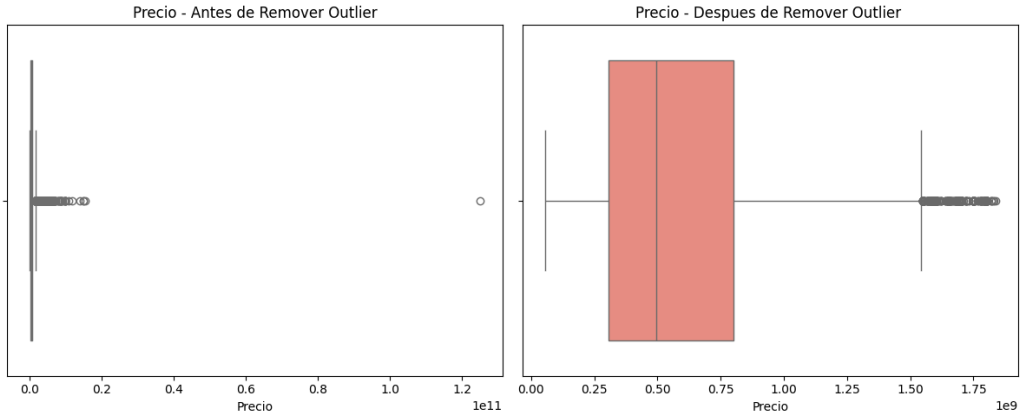
Gráficos de la Columna Habitaciones



Precio. En la Figura 5 se revela que la distribución de precios estaba altamente influenciada por valores extremadamente altos, lo que distorsionaba la interpretación de los datos. Tras remover estos outliers, la distribución se estabiliza, mostrando un rango más representativo y permitiendo identificar con mayor claridad la tendencia central y la variabilidad del conjunto de datos. Esto mejora la calidad del análisis al enfocarse en los valores más relevantes y eliminar el sesgo causado por extremos.

Figura 5

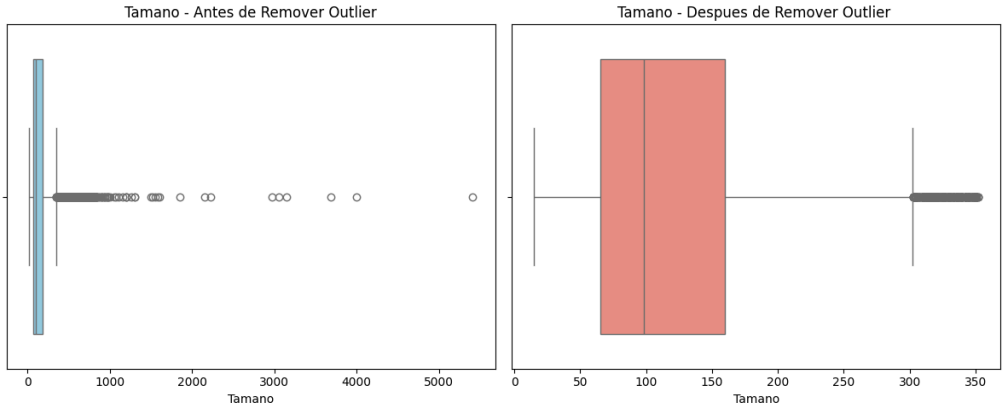
Gráficos Columna Precio



Tamaño. En la Figura 6 se puede observar la presencia de valores extremos en el rango superior con tamaños que van hasta los 5000, mientras que la mayoría de los datos se presentan en un rango, después de eliminar los valores atípicos la distribución se vuelve más compacta, en la gráfica de la derecha, se visualiza el rango intercuartílico (Q1 a Q3) y los valores atípicos restantes son menos influyentes.

Figura 6

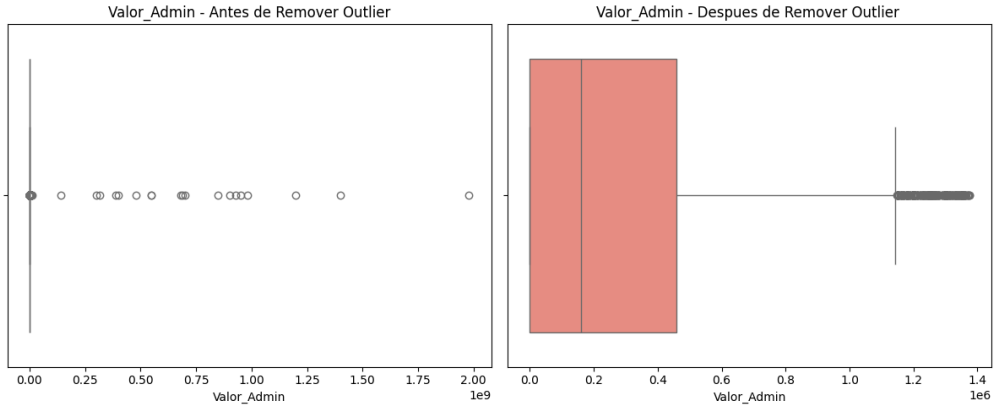
Gráficos Columna Tamano



Valor Administración. En el gráfico de la izquierda de la Figura 7, los datos incluyen valores que alcanzan hasta aproximadamente 2,000 millones, muy por encima del rango típico de la distribución, lo que genera una gran dispersión y posibles distorsiones en las métricas estadísticas, como la media y la desviación estándar. Estos valores extremos indican un sesgo hacia la derecha y dificultan el análisis representativo de los datos. En el gráfico de la derecha, tras eliminar los outliers, la distribución se concentra dentro de un rango más acotado, entre 0 y aproximadamente 1,4 millones, lo que refleja una variabilidad más acorde con el comportamiento central de los datos.

Figura 7

Gráficos Columna Valor Administración



Variables Seleccionadas

Tras el tratamiento y filtrado de los datos, las variables finales seleccionadas para el análisis fueron:

1. Antiguedad
2. Banos
3. Estrato
4. Habitaciones
5. Inmueble
6. Municipio
7. Parking
8. Precio
9. Sector
10. Tamaño
11. Valor_Admin

Estas variables fueron elegidas por su relevancia en los objetivos del estudio y por la calidad alcanzada después del tratamiento de datos.

Justificación del Método Utilizado

El método del Rango Intercuartílico es una técnica estadística robusta y ampliamente utilizada para la detección de outliers. Sin embargo, es importante considerar que esta metodología asume una distribución relativamente simétrica en los datos. Si bien este enfoque es efectivo para este trabajo de grado, en futuros análisis con conjuntos de datos que presenten distribuciones significativamente sesgadas, podrían evaluarse otras técnicas adicionales, como la transformación logarítmica o el escalado robusto, para optimizar la eliminación de outliers.

Este proceso de manejo de outliers es fundamental para garantizar que los análisis y modelos predictivos desarrollados en este estudio sean precisos y representen de manera confiable el comportamiento del mercado inmobiliario, sin la influencia de valores extremos que puedan distorsionar los resultados.

Análisis de Variables

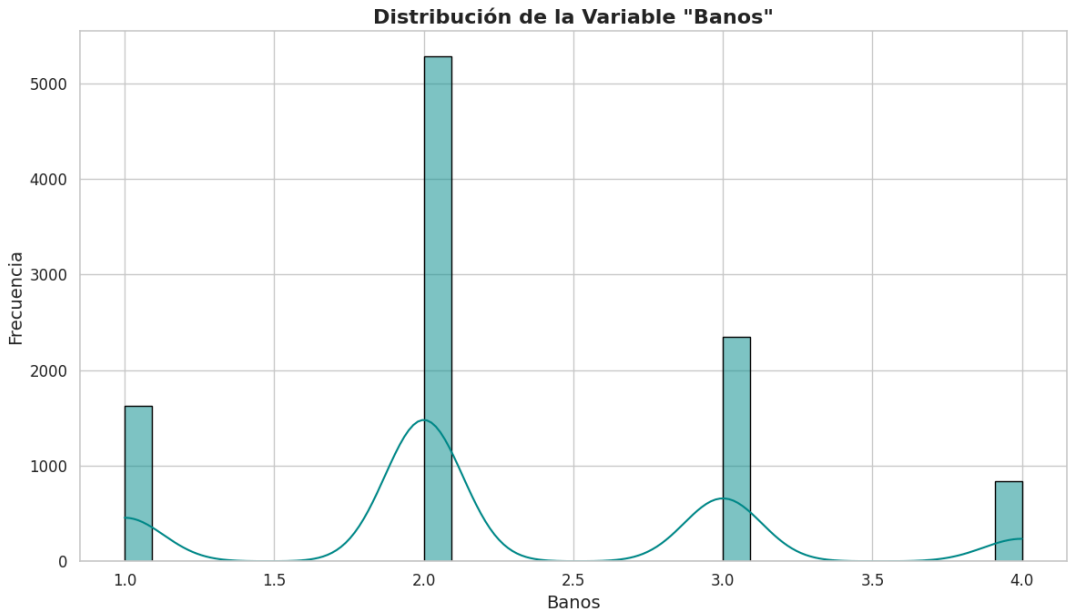
El análisis ayudará a comprender mejor las características predominantes de los inmuebles presentes en este conjunto de datos y cómo se distribuyen según cada una de las variables. Esto proporciona una base sólida para identificar patrones en el mercado y posibles oportunidades de análisis más profundo.

Variables Numéricas

Baños. La mayoría de los inmuebles tienen entre 1 y 3 baños. Esto indica que la mayor parte de los datos son de inmuebles de tamaño moderado, adecuadas para familias pequeñas o medianas. Existe una menor proporción de inmuebles con 4 o más baños, lo cual sugiere que las propiedades más grandes y lujosas son menos comunes en este conjunto de datos. La concentración principal en 2 baños es consistente con la oferta de apartamentos y casas familiares que buscan un balance entre comodidad y costo. Ver Figura 8.

Figura 8

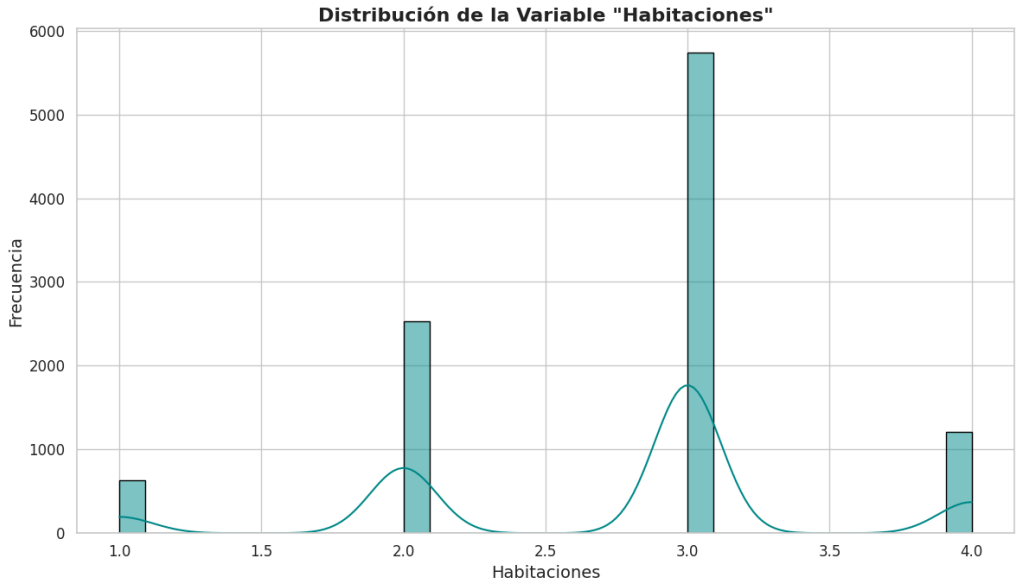
Histograma de Distribución de Variable



Habitaciones. La mayoría de los inmuebles tienen entre 2 y 3 habitaciones, lo cual es típico para apartamentos o casas familiares orientadas a familias pequeñas o parejas. Existe una menor cantidad de inmuebles con sólo 1 habitación o con más de 4 habitaciones. Los inmuebles con una sola habitación podrían corresponder a estudios o apartaestudios, generalmente enfocados a personas solteras o estudiantes. Los inmuebles con 4 o más habitaciones están destinados a familias numerosas o representan propiedades más grandes, a menudo vinculadas a un estrato socioeconómico más alto. Ver Figura 9.

Figura 9

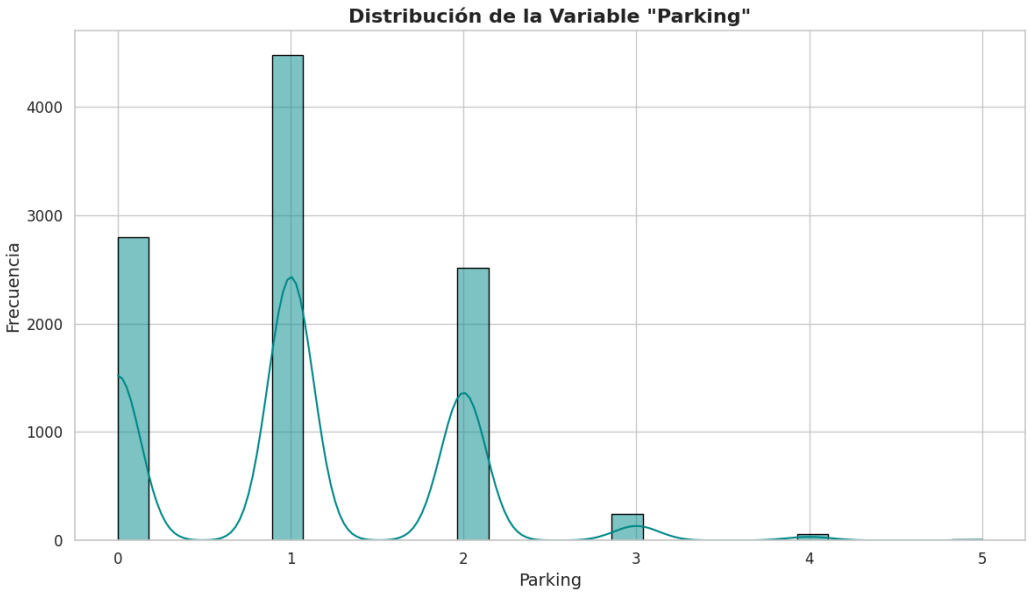
Histograma de Distribución de Variable



Parking. La mayoría de los inmuebles cuentan con 1 o 2 plazas de estacionamiento, lo cual indica que la mayor parte de las propiedades están orientadas a familias con uno o dos vehículos. Hay una proporción significativa de inmuebles sin plazas de estacionamiento, lo cual puede indicar que se trata de apartamentos ubicados en zonas urbanas densamente pobladas, donde el espacio para estacionamiento es limitado y es más común el uso de transporte público. Los inmuebles con más de 2 plazas de estacionamiento son menos comunes y probablemente representan propiedades más exclusivas o de lujo. Ver Figura 10.

Figura 10

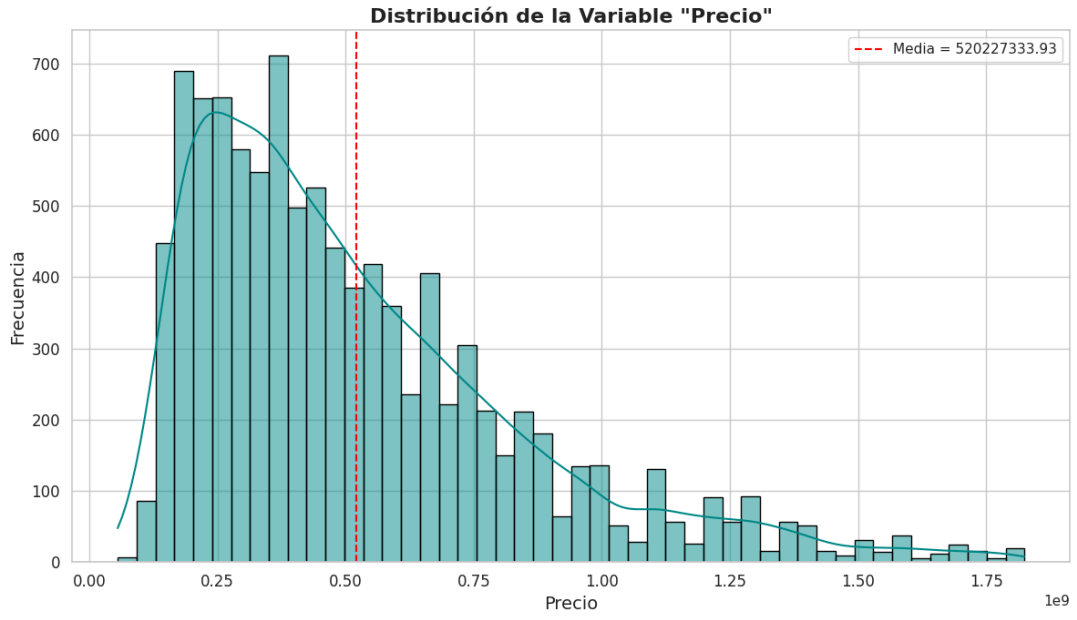
Histograma de Distribución de Variable



Precio. El precio de los inmuebles presenta una gran variabilidad, pero se observa una concentración significativa en el rango más bajo (menos de 400 millones de pesos). Esto sugiere que una buena parte del mercado está compuesto por propiedades accesibles, orientadas a la clase media. A medida que el precio aumenta, la frecuencia de inmuebles disminuye, indicando que las propiedades de alto valor son menos comunes. Esto refleja la estructura del mercado inmobiliario, donde las viviendas de lujo representan un segmento más reducido. Ver Figura 11.

Figura 11

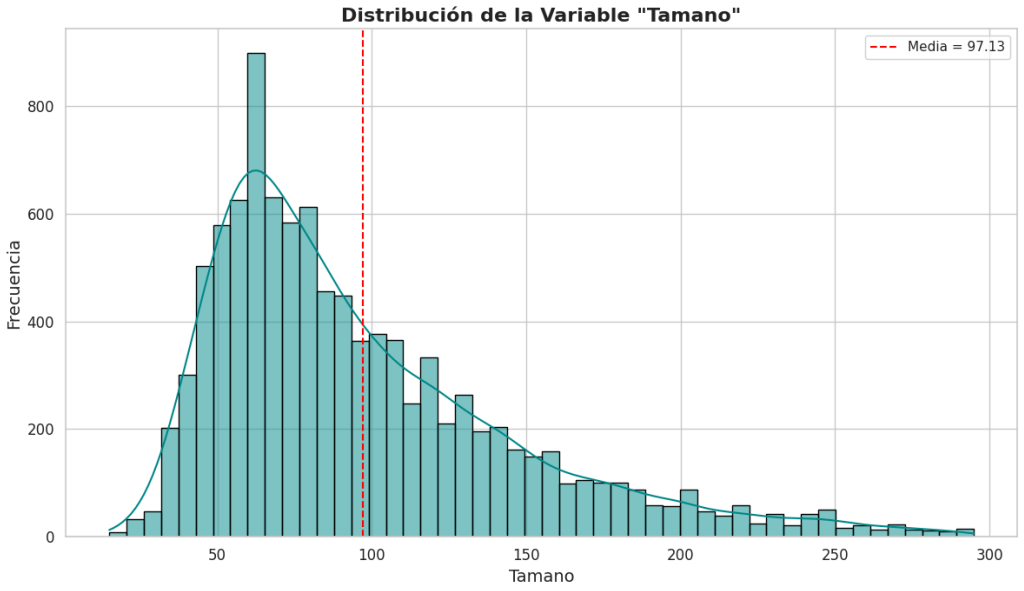
Histograma de Distribución de Variable



Tamaño. La mayoría de los inmuebles tienen un tamaño entre 50 y 150 metros cuadrados. Esto es consistente con el tamaño promedio de apartamentos y casas en zonas urbanas de Colombia, adaptado a las necesidades de familias pequeñas y medianas. Los inmuebles de menor tamaño (menos de 50 m2) podrían ser apartaestudios o apartamentos pequeños, generalmente enfocados a personas solteras, estudiantes o parejas sin hijos. Los inmuebles con tamaños significativamente mayores a 150 m2 son menos comunes y probablemente corresponden a casas grandes o apartamentos de lujo, destinados a familias de mayor poder adquisitivo. Ver Figura 12.

Figura 12

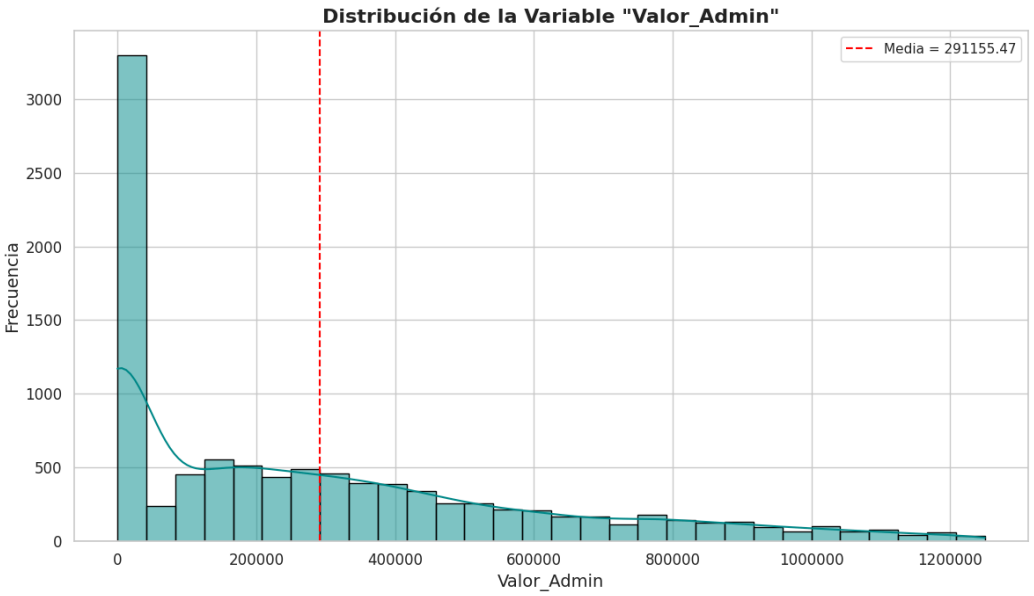
Histograma de Distribución de Variable



Valor Administración. La mayoría de los valores de administración se concentran cerca del 0, lo cual sugiere que hay muchos inmuebles sin costo de administración. Esto podría indicar que son casas independientes o inmuebles en conjuntos donde no se aplica este costo. Para los inmuebles que tienen administración, los valores se distribuyen en un amplio rango, lo que indica la presencia tanto de propiedades con costos bajos (posiblemente en conjuntos básicos) como con costos elevados (en propiedades más exclusivas con mayores servicios y comodidades).

Figura 13

Histograma de Distribución de Variable

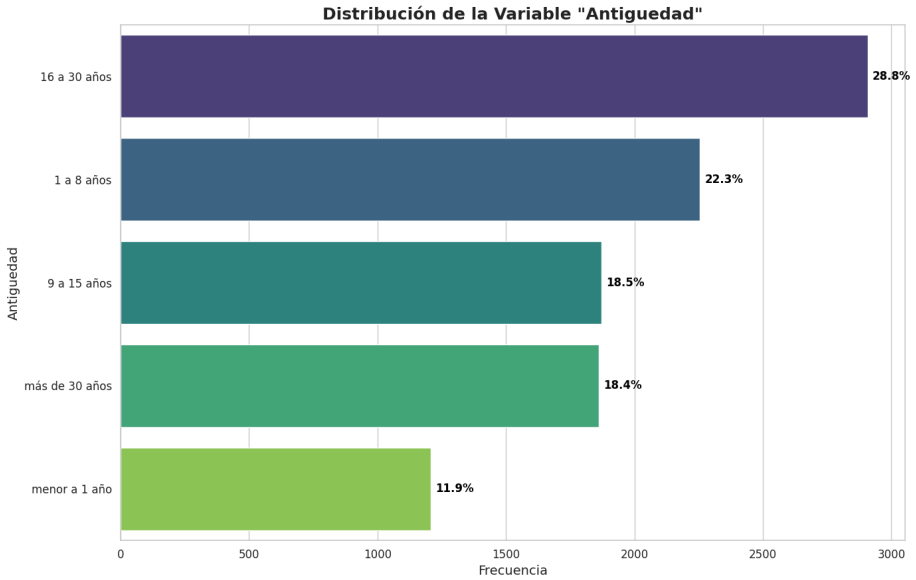


Variables Categóricas

Antiguedad. Los inmuebles se distribuyen en diferentes rangos de antigüedad: "1 a 8 años", "9 a 15 años" y "16 a 30 años". La mayor proporción de inmuebles se encuentra en el rango de "16 a 30 años", lo que indica que una parte importante de la oferta está compuesta por viviendas que ya tienen varios años de construcción, lo cual podría influir en el precio y las condiciones del inmueble. Los inmuebles más nuevos ("1 a 8 años") representan una proporción significativa, lo cual indica una buena cantidad de propiedades relativamente recientes en el mercado, posiblemente con mejores condiciones y características modernas.

Figura 14

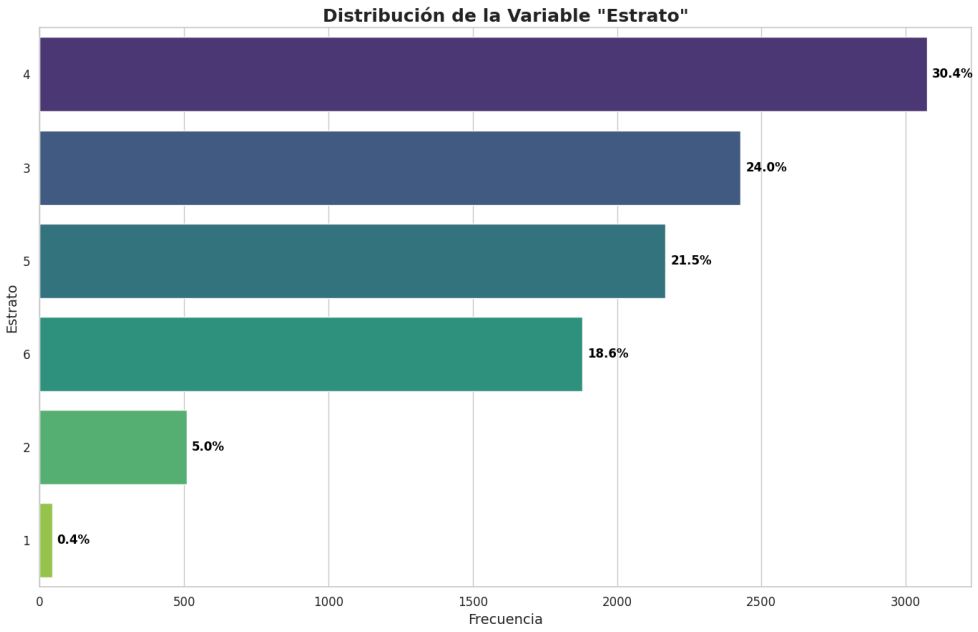
Gráfico de Distribución de Variable



Estrato. La distribución del estrato muestra una tendencia hacia los valores intermedios. Los estratos 3, 4 y 5 son los más comunes, lo cual refleja la composición socioeconómica de muchas áreas urbanas en Colombia, especialmente en las zonas de clase media. Los estratos bajos (1 y 2) tienen una menor representación, lo que podría indicar una baja oferta de viviendas en estos segmentos, posiblemente debido a una menor densidad de construcciones formales en estas áreas. De manera similar, el estrato 6 también presenta menos propiedades, lo cual es común ya que los inmuebles de lujo suelen ser más escasos y exclusivos.

Figura 15

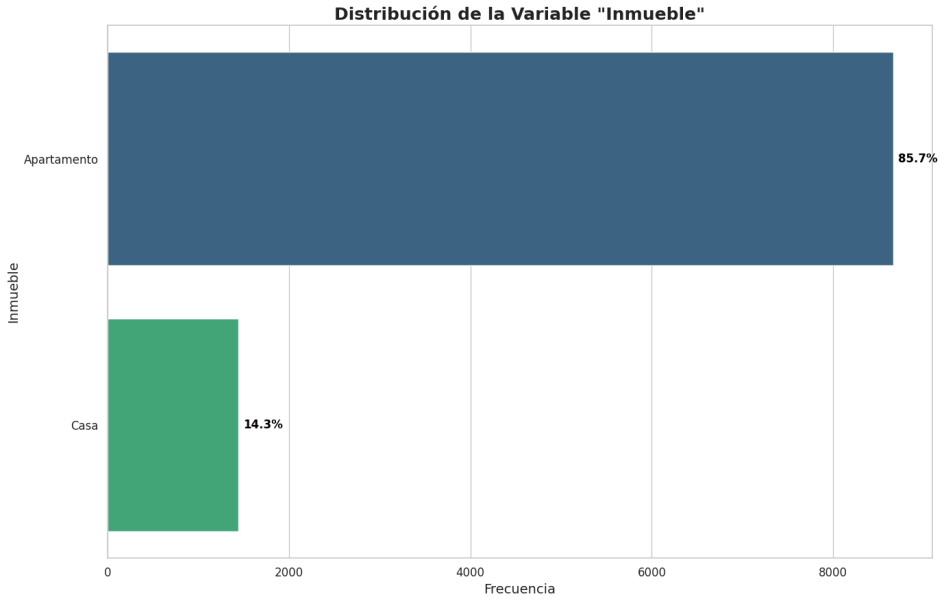
Gráfico de Distribución de Variable



Inmueble. La categoría de tipo de inmueble está dominada por los "Apartamentos", que representan la mayor parte de los datos. Esto sugiere que el mercado inmobiliario está mayormente orientado hacia viviendas verticales, probablemente debido a la densidad urbana y la falta de espacio para desarrollos horizontales.

Figura 16

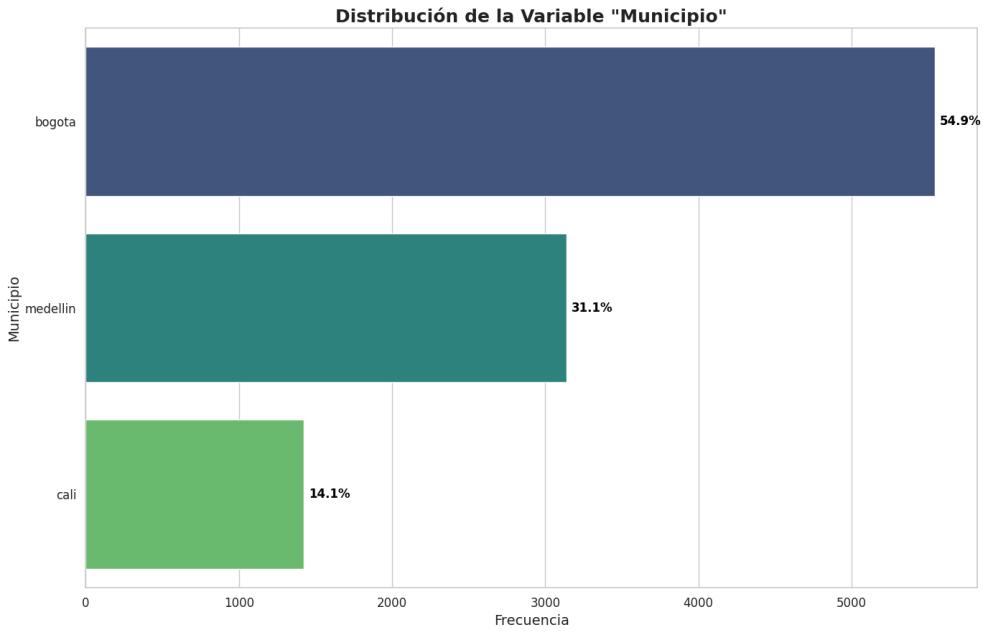
Gráfico de Distribución de Variable



Municipio. El municipio más representado en el conjunto de datos es "Bogotá", lo cual refleja que la muestra está centrada en la capital del país. Esto es relevante ya que el comportamiento del mercado inmobiliario en Bogotá puede diferir considerablemente del de otras regiones de Colombia.

Figura 17

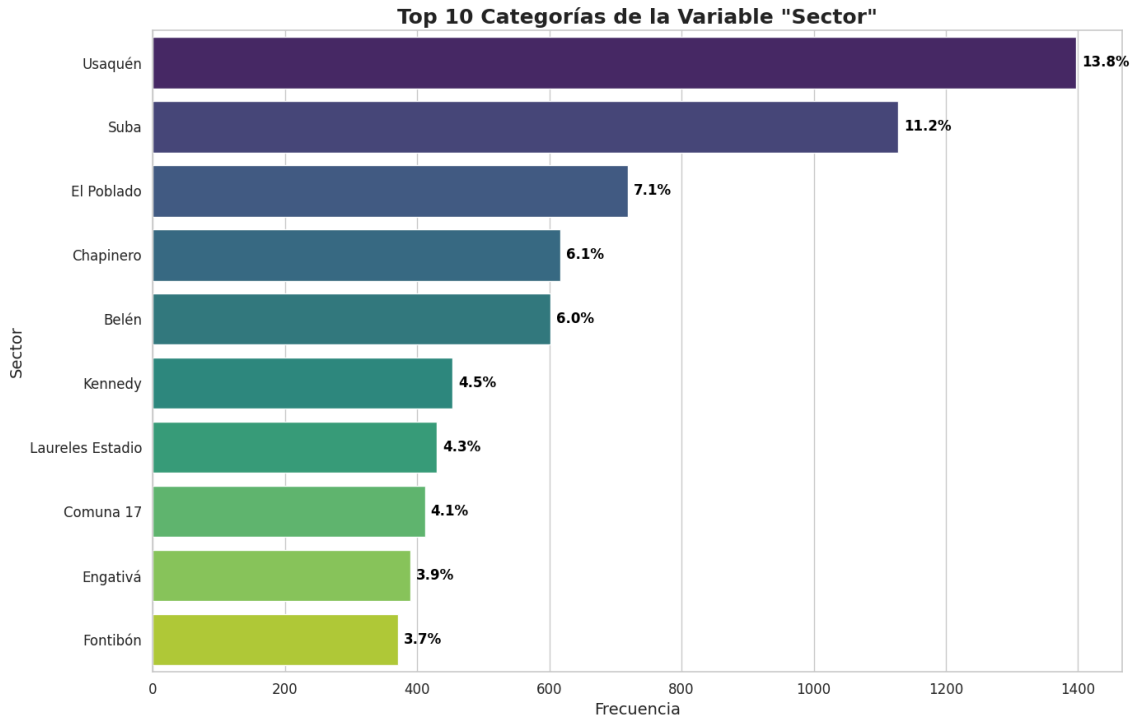
Gráfico de Distribución de Variable



En la Figura 18 se muestra la distribución de frecuencias de los principales sectores en el conjunto de datos, con Usaquén a la cabeza, alcanzando alrededor de 1,400 observaciones (13.8%), seguida por Suba con aproximadamente 1,200 observaciones (11.2%). A partir de ahí se observa una disminución gradual, lo que genera una clara jerarquía entre los sectores. Los dos primeros, Usaquén y Suba, agrupan el 25% del total, mientras que los cinco principales (Usaquén, Suba, El Poblado, Chapinero y Belén) concentran cerca del 46.3%. Esta diferencia entre sectores más y menos frecuentes es notable, con un rango que va del 13.8% al 3.7% (Fontibón), lo cual sugiere una distribución no uniforme y una fuerte concentración en determinadas áreas, particularmente las de mayor frecuencia. En conjunto, el análisis indica que ciertos sectores, sobre todo Usaquén y Suba, podrían ser de especial interés para estudios posteriores o para la toma de decisiones estratégicas.

Figura 18

Gráfico de Distribución de Variable



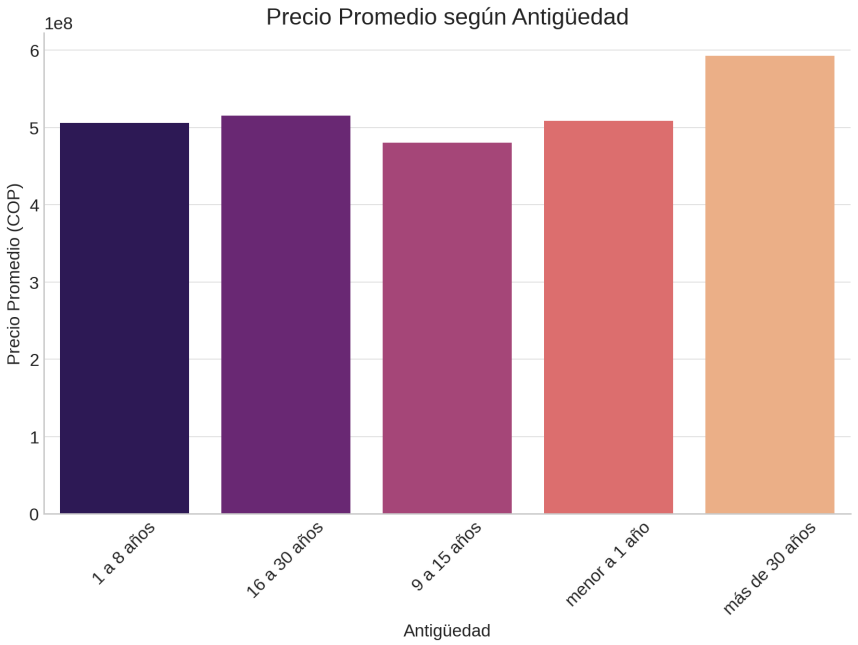
Análisis Multivariado

Relación Antigüedad-Precio Promedio

La Figura 19 muestra cómo, en promedio, el precio de las propiedades varía según su antigüedad, con una tendencia general al alza a medida que aumenta el número de años. Las propiedades con más de 50 años alcanzan precios promedio cercanos a los 600 millones de pesos, superando incluso a las más nuevas. Por otro lado, las más económicas, con alrededor de 480 millones de pesos, se encuentran en el rango de 9 a 15 años de antigüedad. La segmentación temporal en cinco categorías (1-8 años, 16-59 años, 9-15 años, menos de 1 año y más de 50 años) revela patrones no lineales: mientras las propiedades muy antiguas parecen gozar de una prima por factores como su valor histórico o ubicación, las más recientes también exhiben precios por encima del promedio. Esto contrasta con una ligera depresión en las propiedades con una antigüedad intermedia (9-15 años), sugiriendo que la influencia de la antigüedad no es lineal y que existen otros elementos determinantes en la dinámica de precios del mercado inmobiliario.

Figura 19

Gráfico de Distribución de Variable



Distribución del Precio según Municipio

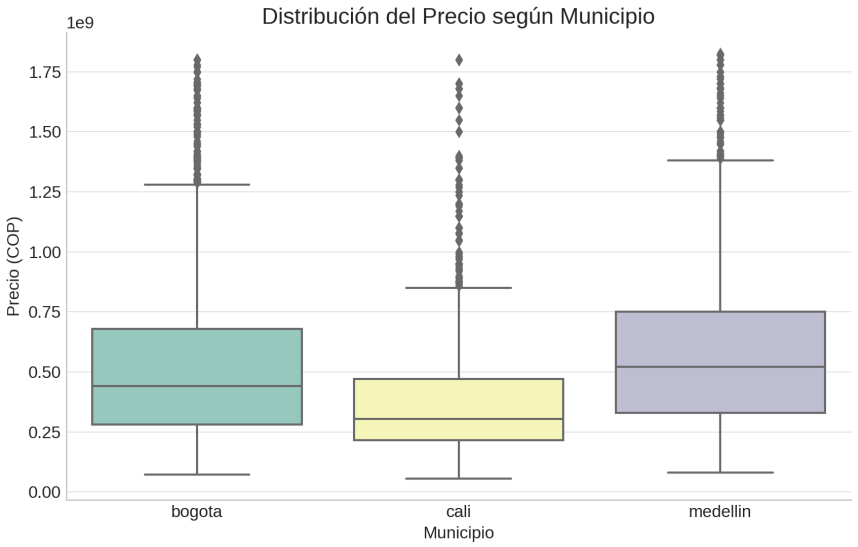
La gráfica de cajas y bigotes de la Figura 20 que compara la distribución de precios de propiedades en Bogotá, Cali y Medellín muestra diferencias notables entre estos tres municipios. Medellín presenta la mediana más alta, cercana a los 500 millones de pesos, seguida muy de cerca por Bogotá, mientras que Cali exhibe una mediana más baja, alrededor de 300 millones. Las tres ciudades poseen valores atípicos que superan los 1.500 millones, con Medellín mostrando la mayor dispersión en su rango intercuartílico y Cali, por el contrario, el conjunto más compacto.

En las tres distribuciones se observa asimetría positiva, lo que implica una concentración mayor de precios en el rango de 200 a 750 millones de pesos, acompañada de numerosos valores atípicos en la parte superior. Aunque Medellín y Bogotá comparten patrones similares de precios, con mayor dinamismo y valor, Cali mantiene una estructura más homogénea y reducida. A pesar de estas diferencias, los valores atípicos más elevados son comparables en las tres ciudades, llegando a alrededor de 1.750 millones de pesos.

En definitiva, este análisis sugiere que Medellín y Bogotá presentan mercados inmobiliarios más intensos y de mayor valor, mientras que Cali se caracteriza por una distribución de precios más concentrada y relativamente baja.

Figura 20

Boxplot de Distribución de Variable

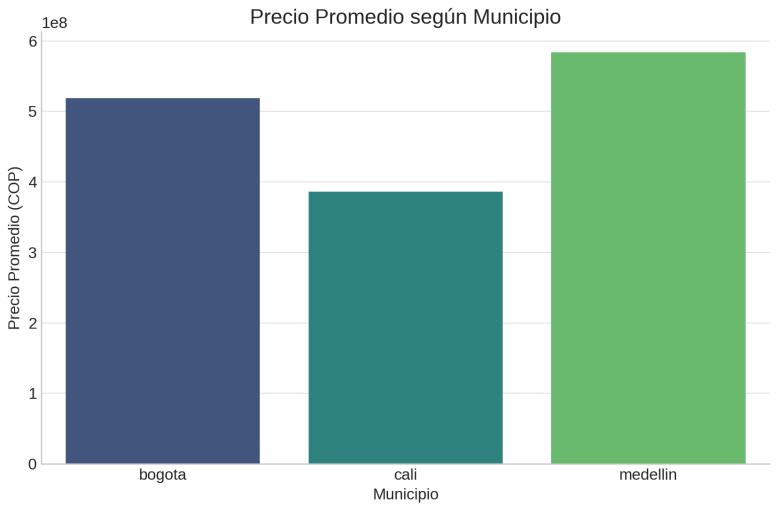


Precio Promedio según Municipio

La Figura 21 , que compara los precios promedio de propiedades en Bogotá, Cali y Medellín revela una clara segmentación del mercado inmobiliario entre las tres ciudades. Medellín lidera con cerca de 580 millones de pesos, seguida por Bogotá con aproximadamente 520 millones, mientras que Cali se queda atrás con un promedio cercano a 380 millones. Estas diferencias generan una brecha de alrededor de 200 millones entre Medellín y Cali, y de 60 millones entre Medellín y Bogotá. Estos resultados sugieren la existencia de tres niveles de mercado: uno más “premium” en Medellín, otro intermedio-alto en Bogotá y un nivel más accesible en Cali. Las marcadas variaciones en los precios promedio podrían estar relacionadas con factores económicos, sociales y urbanos específicos de cada municipio, así como con el funcionamiento particular de la oferta y la demanda inmobiliaria en cada región.

Figura 21

Gráfico de Distribución de Variable

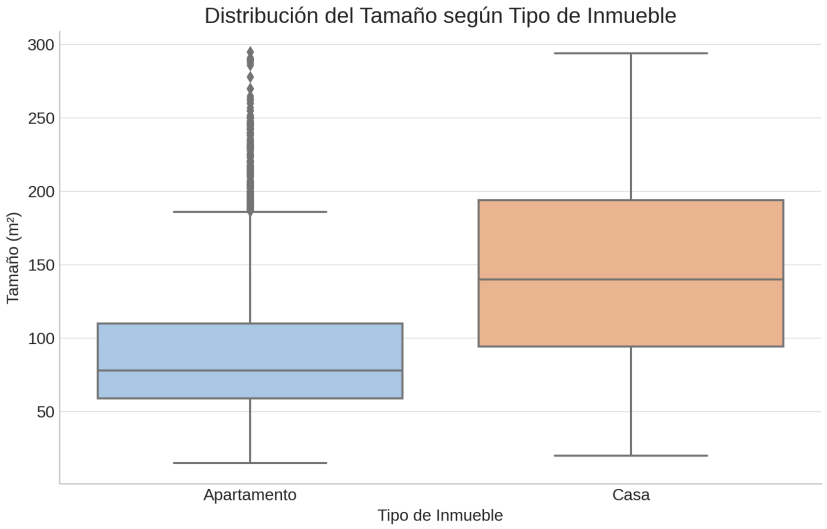


Distribución del Tamaño según Tipo de Inmueble

La comparación entre apartamentos y casas en el diagrama de cajas y bigotes (Figura 22) revela diferencias notables en la distribución de tamaños. Las casas muestran una mediana en torno a los 140 m², mientras que la de los apartamentos ronda los 75 m², marcando así una diferencia de aproximadamente 65 m². Además, las casas presentan una mayor dispersión, con un rango intercuartílico de alrededor de 100 a 200 m², frente a los 60 a 110 m² de los apartamentos. En ambas categorías se observan valores atípicos que alcanzan hasta 300 m², indicando la existencia de inmuebles de lujo o características especiales. Tanto casas como apartamentos presentan asimetría positiva y un valor mínimo similar, cercano a los 25 m². Sin embargo, las casas exhiben una variabilidad más amplia en sus tamaños, mientras que los apartamentos mantienen una distribución más compacta. Estos hallazgos concuerdan con las diferencias típicas entre ambos tipos de inmuebles en el mercado inmobiliario.

Figura 22

Boxplot de Distribución de Variable



Distribución del Precio según Número de Baños

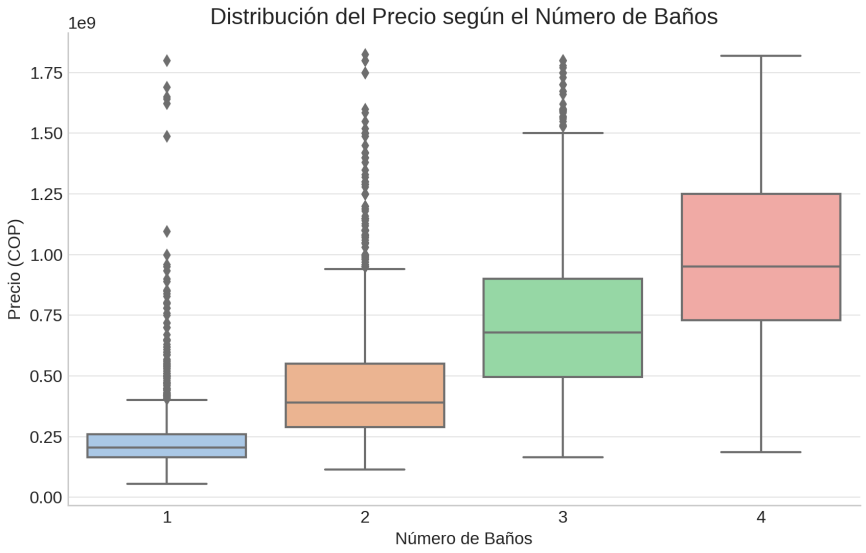
La distribución de precios según la cantidad de baños en los inmuebles, representada con diagramas de caja y bigotes de la Figura 23, muestra una clara relación ascendente: a medida que aumenta el número de baños, también lo hace el valor de la propiedad. Las propiedades con 4 baños alcanzan una mediana cercana a los 1,000 millones de pesos, mientras que aquellas con solo 1 baño rondan los 200 millones, evidenciando una diferencia sustancial. Además, la dispersión de precios aumenta con cada categoría, y todas presentan valores atípicos por encima de los 1,750 millones de pesos.

Al observar cada grupo, la categoría con 1 baño tiene un rango intercuartílico bastante estrecho, en contraste con las propiedades de 4 baños, que muestran una gran variabilidad entre 750 y 1,250 millones de pesos. Entre las categorías intermedias, pasar de 2 a 3 baños marca un salto notable en la mediana (de 400 a 700 millones de pesos) y un incremento en la dispersión.

Estos hallazgos sugieren que el número de baños es un factor relevante para la valoración de una propiedad, aunque no explica por sí solo la complejidad del mercado. La relación no lineal y la creciente variabilidad a medida que aumenta el número de baños indican que también intervienen otros elementos, como la ubicación, el tamaño o la calidad del inmueble, en la determinación de su precio final.

Figura 23

Boxplot de Distribución de Variable

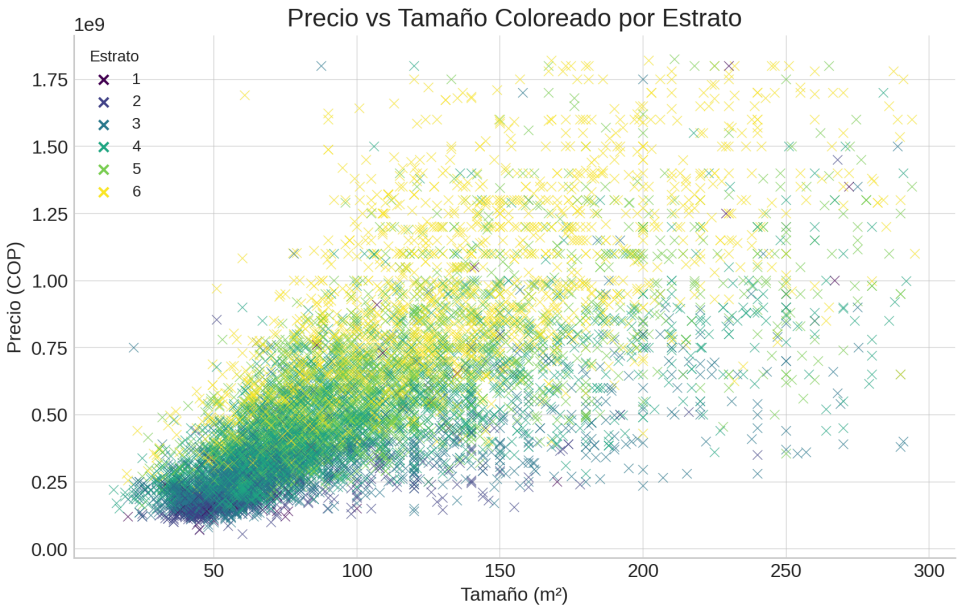


Relación Precio vs Tamaño según Estrato

La Figura 24 muestra una relación claramente positiva entre el precio y el tamaño de los inmuebles, donde a medida que ambos aumentan, también lo hace la dispersión de los datos. Los precios varían entre 0 y aproximadamente 1.750 millones de pesos, mientras que los tamaños oscilan entre los 25 y 300 metros cuadrados. Al diferenciar por estratos socioeconómicos mediante colores, se observa que los estratos más bajos (1-2) tienden a concentrarse en la parte inferior izquierda del gráfico, mientras que los estratos altos (5-6) se ubican en la zona superior derecha. Los estratos medios (3-4) presentan una distribución más amplia en el centro, con cierto solapamiento entre las distintas categorías. La mayor densidad de inmuebles se encuentra en rangos de 50 a 150 metros cuadrados y precios entre 250 y 750 millones de pesos, notándose que, a mayor estrato, la pendiente de la relación precio-tamaño se hace más pronunciada. La variabilidad en los precios crece con el tamaño, y si bien se aprecia una clara segmentación por estrato, esta no es rígida, pues hay áreas donde se superponen distintas categorías. Además, los valores atípicos suelen aparecer en los estratos más altos, creando una compleja interacción entre precio, tamaño y estrato socioeconómico, que refleja la diversidad del mercado inmobiliario.

Figura 24

Gráfico de dispersión multivariado

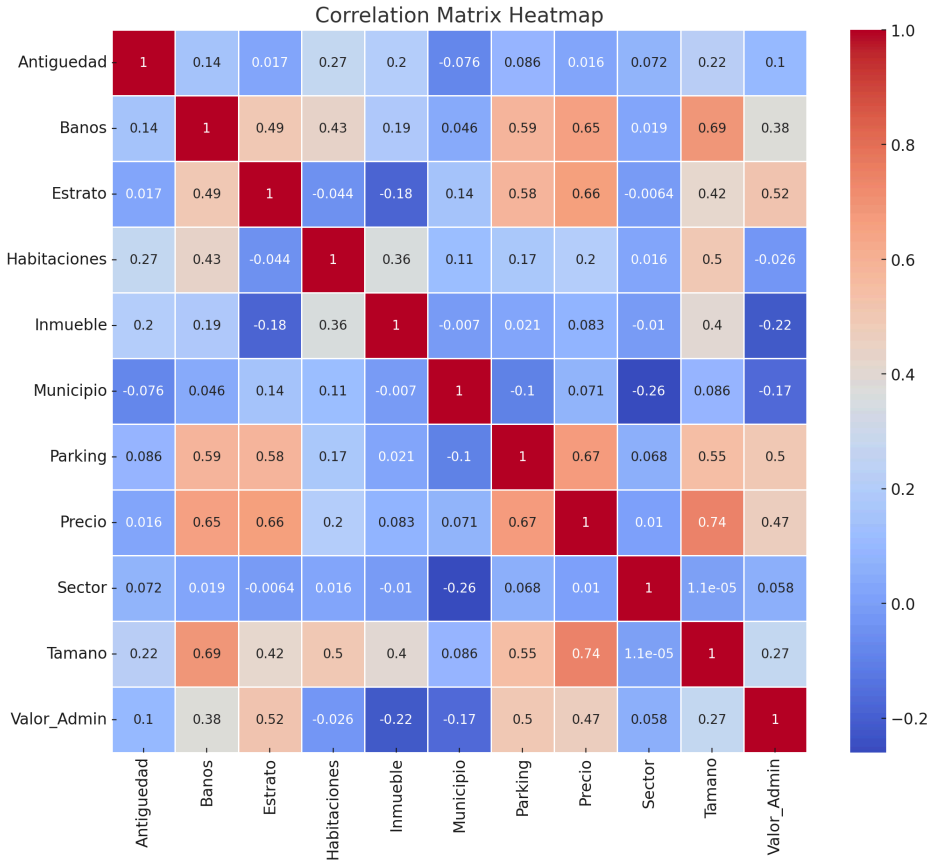


Matriz de Correlación

La matriz de correlación (Figura 25) muestra las relaciones entre diferentes variables del mercado inmobiliario, donde los valores oscilan entre -1 (correlación negativa perfecta) y 1 (correlación positiva perfecta).

Figura 25

Matriz de correlación



Correlaciones más significativas

Las correlaciones más significativas muestran que el precio de un inmueble está fuertemente asociado con su tamaño (0.74), la disponibilidad de estacionamiento (0.67) y el estrato socioeconómico (0.66), además de que las propiedades más grandes tienden a contar con más baños (0.69). En cuanto a las correlaciones moderadas, los estratos más altos suelen tener mayor disponibilidad de estacionamiento (0.58), el tamaño presenta una relación moderada con los costos administrativos (0.27) y el número de habitaciones aumenta a medida que crece el área del inmueble (0.50). Por otro lado, la antigüedad y el estrato prácticamente no guardan relación (0.017), y las correlaciones con el municipio y el sector resultan en su mayoría débiles o negativas, lo que sugiere que factores como el tipo de inmueble, la ubicación o la antigüedad tienen menor incidencia. Este análisis evidencia que las características físicas del inmueble, junto con el nivel socioeconómico del entorno, son los factores más influyentes en el precio, mientras que la antigüedad y la ubicación parecen desempeñar un papel más limitado en la determinación de su valor.

Selección y Justificación de las Métricas de Evaluación de Desempeño

Las métricas de desempeño seleccionadas para evaluar el rendimiento del modelo incluyen el R^2 Score, el Mean Absolute Error (MAE), el Mean Squared Error (MSE) y el Root Mean Squared

Error (RMSE). Cada una de estas métricas aporta una visión complementaria sobre la calidad de las predicciones:

- **R^2 Score:** Mide qué tan bien las predicciones se ajustan a los datos reales, indicando la proporción de la varianza en la variable objetivo que es explicada por el modelo. Es esencial porque muestra la capacidad global del modelo para capturar patrones en los datos. Un valor cercano a 1 significa un buen ajuste, mientras que un valor bajo indica que el modelo no explica bien la variabilidad observada.

- **Mean Absolute Error (MAE)**: Representa el error promedio entre las predicciones y los valores reales sin considerar la dirección del error (si es positivo o negativo). Esta métrica es útil porque proporciona una interpretación directa del error medio en las unidades de la variable objetivo, lo cual facilita entender cuánto se desvía en promedio el modelo de los valores reales. El MAE es menos sensible a los valores atípicos en comparación con el MSE.

- **Mean Squared Error (MSE)**: Calcula el promedio del error al cuadrado. El MSE penaliza los errores grandes, lo cual es útil cuando se desea que el modelo minimice los errores grandes y se vuelva más robusto frente a variaciones grandes en los datos. Sin embargo, debido a la penalización al cuadrado, no es tan intuitivo como el MAE para interpretaciones directas.

- **Root Mean Squared Error (RMSE)**: Es la raíz cuadrada del MSE y se utiliza para tener una medida de error en la misma escala que la variable objetivo. El RMSE es útil para medir la variabilidad de las predicciones alrededor de los valores reales y ayuda a entender la magnitud de los errores del modelo.

La combinación de estas métricas proporciona una evaluación integral del rendimiento del modelo. Mientras que el R^2 Score indica la capacidad global de explicación del modelo, el MAE y el RMSE ayudan a entender la precisión de las predicciones. Usar múltiples métricas permite tomar decisiones más informadas sobre la efectividad del modelo, evaluar su estabilidad y comparar su rendimiento con otros enfoques. En este caso, el Stacking Regressor fue capaz de lograr un equilibrio sólido, mostrando buenos valores en cada una de estas métricas, lo cual refuerza su selección como el modelo óptimo en esta etapa del análisis.

Selección del Modelo

La selección del modelo dependerá de factores como el tamaño del conjunto de datos, la necesidad de interpretabilidad y la complejidad de las relaciones entre las variables. Para asegurar la selección del modelo más adecuado, se utilizarán métricas como el error cuadrático medio (MSE), error absoluto medio (MAE) y el coeficiente de determinación (R^2). Cada una de estas métricas

permite evaluar la precisión y efectividad de los modelos de manera objetiva, lo cual es esencial para garantizar un rendimiento óptimo en la predicción de precios de propiedades.

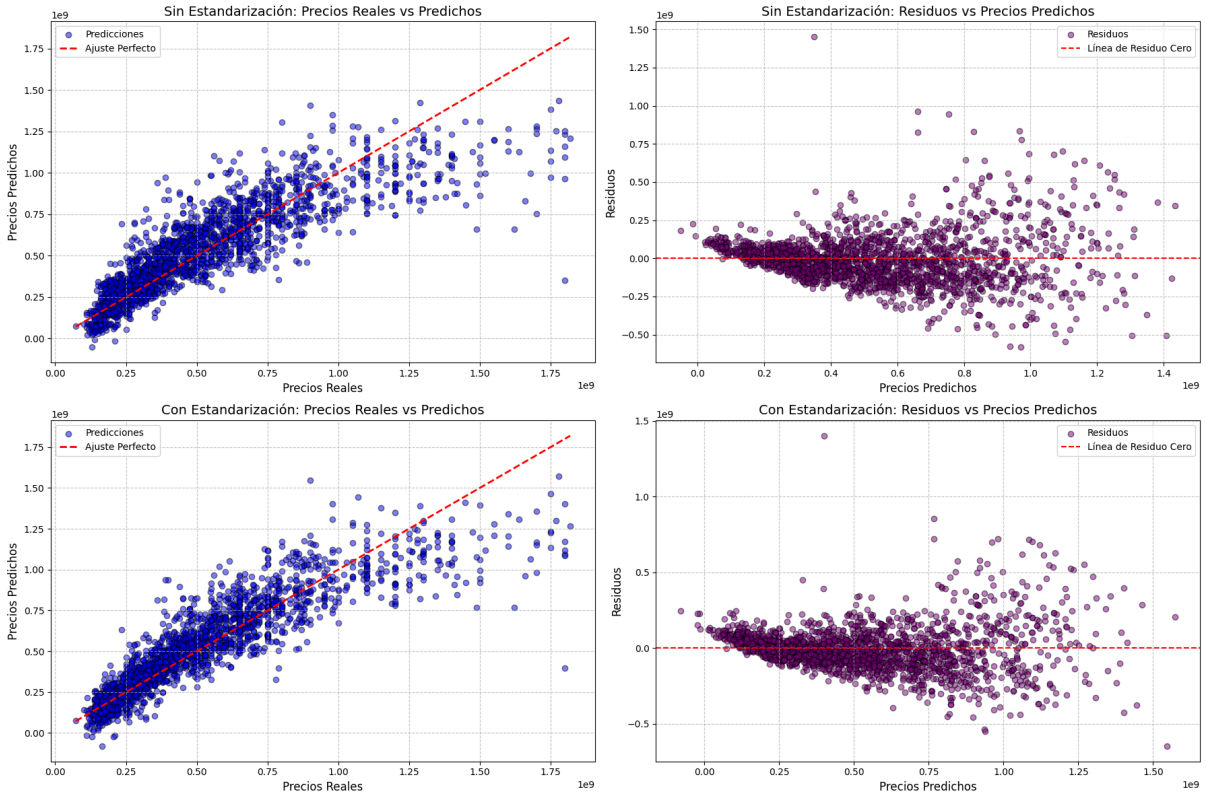
Regresión Lineal Múltiple

Este modelo permite una interpretación clara de las relaciones lineales entre las variables predictoras y el precio. Es especialmente útil cuando se quiere comprender la influencia de cada variable en el precio de una propiedad. Según Montgomery, Peck y Vining (2012), la regresión lineal es adecuada para entender las relaciones simples en los datos, proporcionando una base sólida para modelos más complejos (Introduction to Linear Regression Analysis, John Wiley & Sons).

En este proyecto el modelo de regresión lineal múltiple muestra un **R^2** de **0.7427**, lo que indica que el **74.27%** de la variabilidad en los precios de las propiedades es explicada por las características incluidas, sugiriendo un ajuste moderadamente bueno. Sin embargo, el **MSE** es muy alto (2.71e+16), lo que indica que hay una considerable desviación entre los valores reales y las predicciones, señalando la necesidad de mejorar el modelo.

Figura 26

Gráfico de valores reales vs predichos



Para mejorar el modelo de regresión lineal múltiple, se estandarizaron las características numéricas, para asegurar que todas tuvieran la misma escala. Estas mejoras resultaron en un aumento

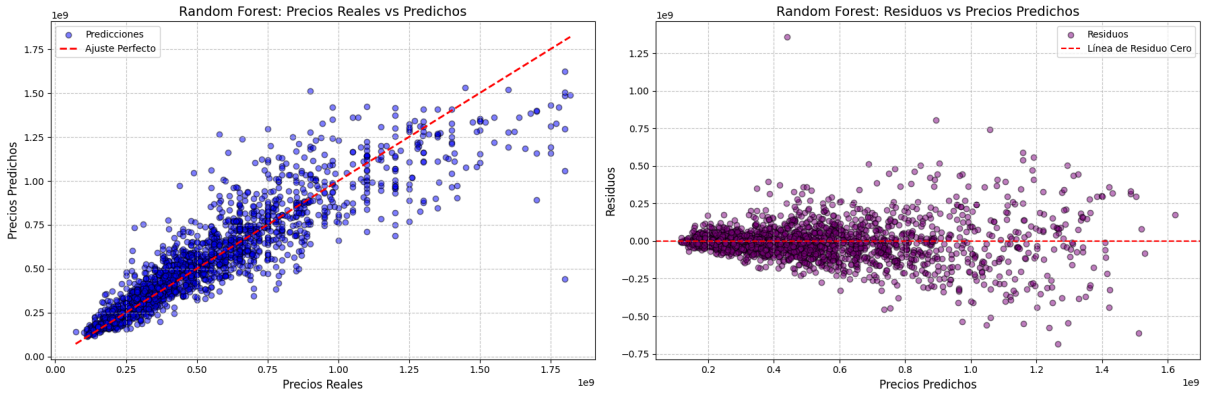
del **R²** de **0.7427** a **0.7956**, lo cual indica que el modelo ahora explica un **79.56%** de la variabilidad en los precios de las propiedades. Además, el MSE se redujo de 2.71e+16 a 2.16e+16, lo que muestra una mejora en la precisión de las predicciones, aunque aún hay margen para reducir los errores.

Árboles de Decisión y Random Forest

Los árboles de decisión son modelos no lineales que permiten capturar relaciones complejas entre características. El método Random Forest mejora la precisión mediante la combinación de múltiples árboles, reduciendo la varianza y evitando el sobreajuste. Breiman (2001) destacó la robustez de Random Forest al combinar múltiples árboles, mejorando tanto la precisión como la generalización del modelo (Random Forests, Machine Learning, 45(1), 5-32).

Figura 27

Gráfico de valores reales vs predichos



En este modelo podemos decir que el valor de **R²** ha mejorado significativamente a **0.8342**, lo que indica que el modelo explica el **83.42%** de la variabilidad en los precios, lo cual es mejor que los resultados anteriores con el modelo de regresión Lineal Múltiple. Además, el **MSE** también se redujo a **1.75e+16**, mostrando una mejora en la precisión de las predicciones.

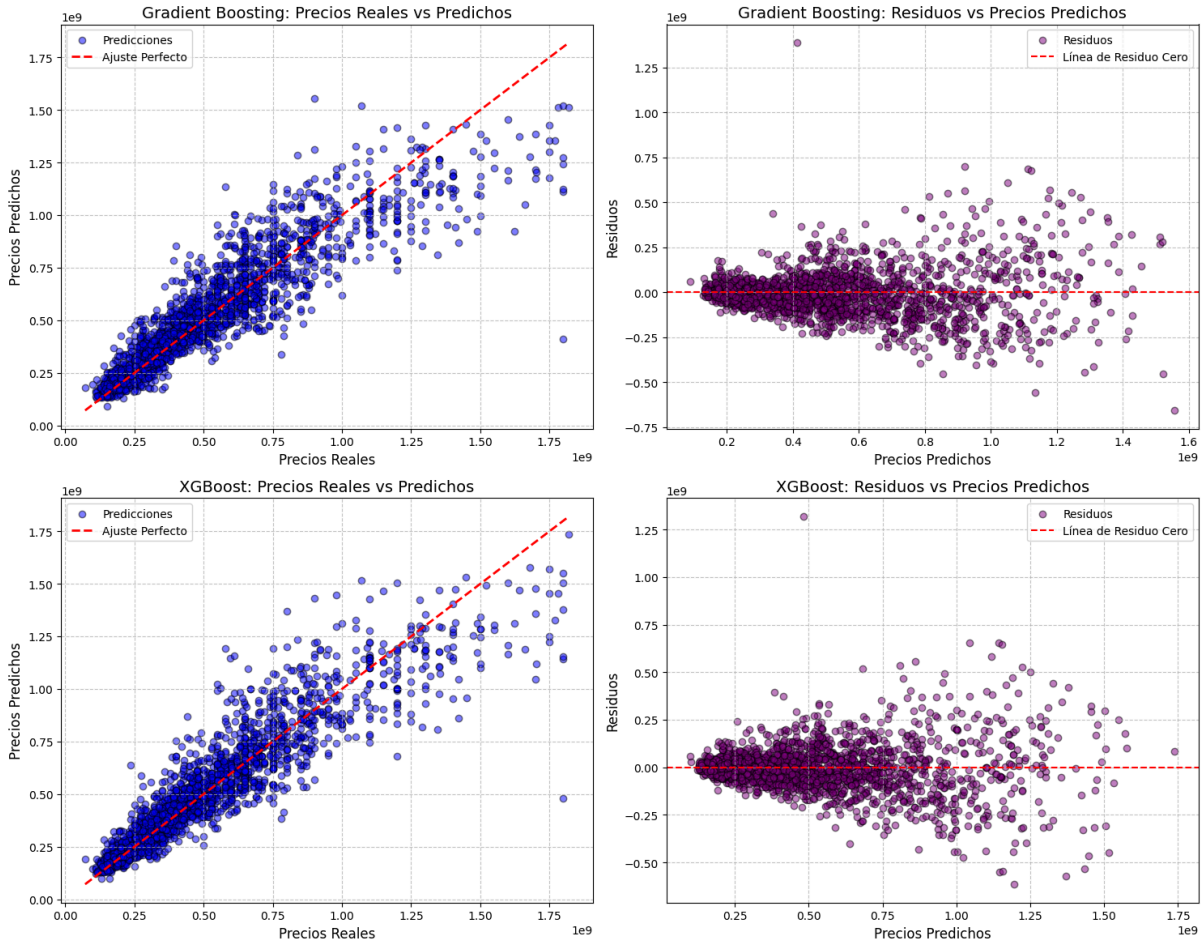
Estos resultados indican que el **Random Forest Regressor** está capturando mejor la complejidad de los datos en comparación con el modelo de regresión lineal múltiple.

Gradient Boosting Machines (GBM) y XGBoost

Estos modelos utilizan el enfoque de boosting para mejorar el rendimiento general. XGBoost es particularmente eficiente y se ha destacado en muchas competiciones de predicción. Chen y Guestrin (2016) explican que XGBoost es una opción escalable y precisa para problemas complejos de predicción, lo que lo hace muy popular en el ámbito del análisis de datos (XGBoost: A Scalable Tree Boosting System, Proceedings of the 22nd ACM SIGKDD International Conference).

Figura 28

Gráfico de valores reales vs predichos



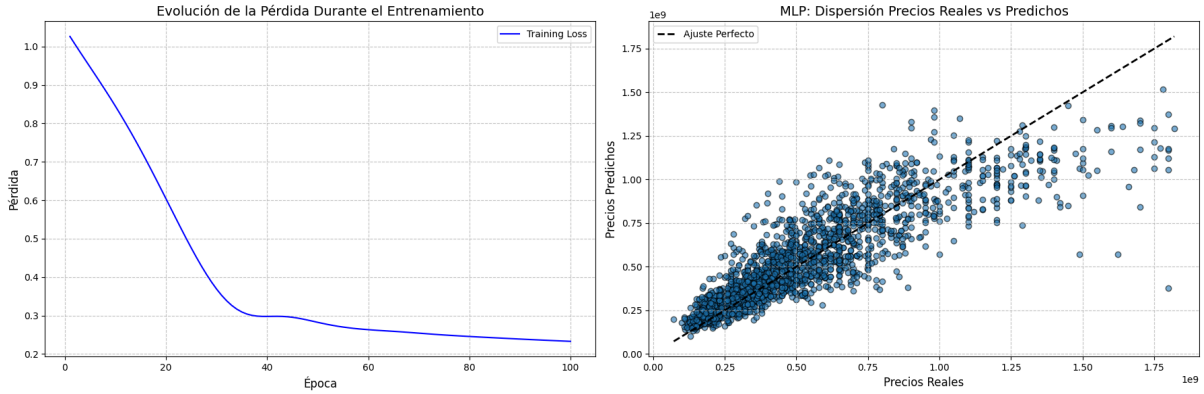
Con base en los resultados obtenidos de los modelos Gradient Boosting Regressor (GBM) y XGBoost Regressor, podemos concluir que el modelo XGBoost tuvo un rendimiento superior en comparación con GBM. El MSE de XGBoost es menor (1.64e+16 frente a 1.78e+16), lo que indica que las predicciones de XGBoost se desvían menos de los valores reales, lo cual sugiere una mejor precisión. Además, el **R^2** de XGBoost es mayor (**0.8440 frente a 0.8308**), lo que significa que este modelo explica mejor la variabilidad en los precios de las propiedades, alcanzando un **84.4%** de capacidad explicativa. Estos resultados indican que, entre los dos, XGBoost es más adecuado para capturar la complejidad de los datos y proporcionar mejores predicciones en este problema.

Redes Neuronales (Perceptrón Multicapa)

Las redes neuronales son útiles para capturar patrones complejos y no lineales en grandes volúmenes de datos. Según Goodfellow, Bengio y Courville (2016), estos modelos son capaces de identificar relaciones que otros métodos no pueden modelar debido a su capacidad de representar funciones complejas (Deep Learning, MIT Press).

Figura 29

Evaluación de pérdida durante entrenamiento y Gráfico de valores reales vs predichos



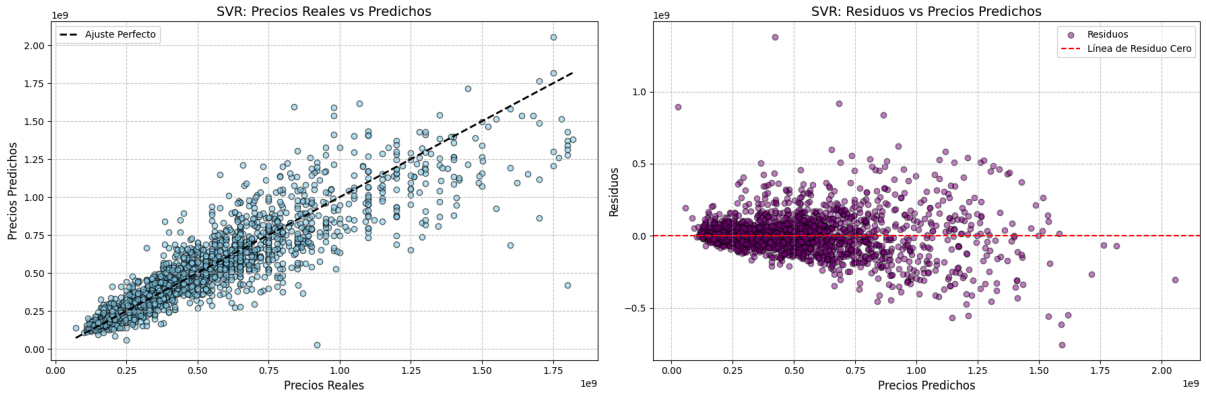
Este modelo de red neuronal (Perceptrón Multicapa) utilizada para la predicción de precios inmobiliarios logró un **R^2 de 0.75**, lo que indica que el modelo explica un **75%** de la variabilidad en los precios, sugiriendo un ajuste razonable pero con margen de mejora. Sin embargo, los errores, representados por un **MAE de 111 millones** y un **RMSE de 162 millones**, muestran que aún existen desviaciones significativas entre las predicciones y los valores reales

Support Vector Regressor (SVR)

Este modelo es adecuado para problemas de regresión donde se desea encontrar una función de margen óptimo para los precios. Puede capturar relaciones no lineales al usar diferentes tipos de kernels, lo que lo hace versátil. Smola y Schölkopf (2004) proporcionan una guía detallada sobre cómo aplicar SVR para modelar datos complejos de manera efectiva (A Tutorial on Support Vector Regression, Statistics and Computing, 14(3), 199-222).

Figura 30

Gráfico de valores reales vs predichos



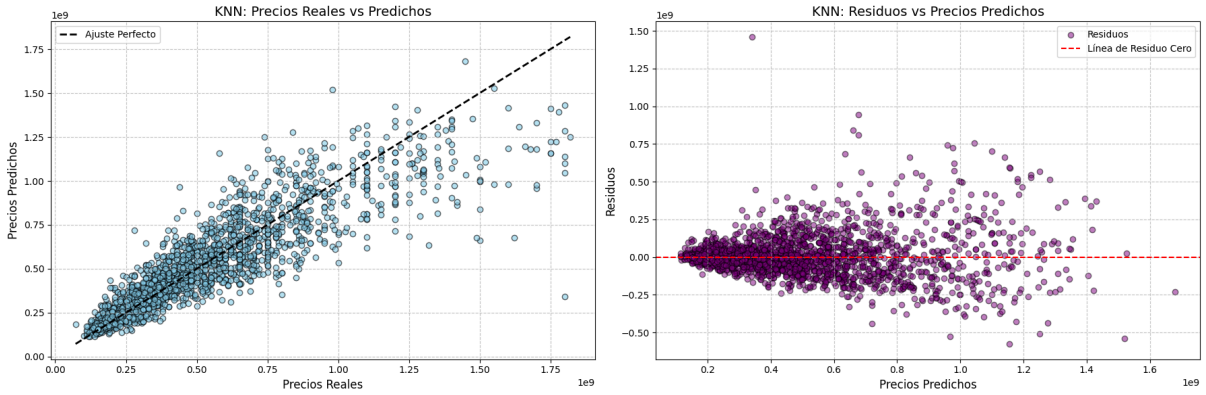
El modelo SVR logró un **R²** de 0.80, lo que indica que explica el **80.43%** de la variabilidad en los precios, sugiriendo un ajuste bueno en comparación con los otros modelos utilizados. El MAE es de aproximadamente 93 millones, y el RMSE es de 143.7 millones, lo cual muestra que todavía hay una desviación significativa entre los valores predichos y los reales.

K-Nearest Neighbors (KNN) para Regresión

Este modelo es útil para estimar los precios basándose en la similitud con propiedades vecinas en el espacio de características. Cover y Hart (1967) introdujeron el algoritmo KNN, que se ha utilizado tanto para clasificación como para regresión debido a su simplicidad y eficacia (Nearest Neighbor Pattern Classification, IEEE Transactions on Information Theory, 13(1), 21-27).

Figura 31

Gráfico de valores reales vs predichos



El modelo K-Nearest Neighbors (KNN) optimizado logró un mejor rendimiento en comparación con el modelo no ajustado, alcanzando un **R² de 0.82**, lo que indica que explica el **82%** de la variabilidad en los precios inmobiliarios. El Error Absoluto Medio (MAE) se redujo a 89 millones, y la Raíz del Error Cuadrático Medio (RMSE) también disminuyó a 139 millones, reflejando una mejora en la precisión de las predicciones. Los mejores hiperparámetros encontrados fueron n_neighbors=11, weights='distance', y p=1 (Manhattan distance), lo cual permitió que el modelo ajustado realizará mejores predicciones en comparación con la versión inicial.

Comparación de Métricas de Desempeño

Tabla 1

Métricas de desempeño de modelos

Model	Mean Absolute Error (MAE)	R2 Score
Linear Regression	$ 111.408.984	0,7427
Random Forest Regressor	$ 93.151.491	0,8342
Gradient Boosting Regressor	$ 111.408.984	0,8308
XGBoost Regressor	$ 89.076.116	0,844
Neural Network	$ 111.408.984	0,75
SVR	$ 93.151.491	0,8
KNN (Optimized)	$ 89.076.116	0,82

En el cuadro comparativo se muestra el rendimiento de todos los modelos utilizados.

- XGBoost Regressor tiene el mejor R^2 Score (0.844), lo que indica que explica el 84.4% de la variabilidad en los precios de las propiedades, el valor más alto entre todos los modelos.
- También tiene uno de los menores Mean Absolute Error (MAE) lo cual indica que, en promedio, las predicciones de XGBoost están más cerca de los valores reales, y que los errores grandes se han minimizado en comparación con otros modelos.
- KNN Optimizado también muestra un buen rendimiento, con un R^2 Score de 0.82 y un RMSE razonablemente bajo, pero no alcanza los resultados del XGBoost.

Por lo tanto, XGBoost Regressor sería la mejor opción para este proyecto, ya que ofrece la mejor combinación de precisión y capacidad de explicar la variabilidad de los datos, con menores errores de predicción.

Optimización

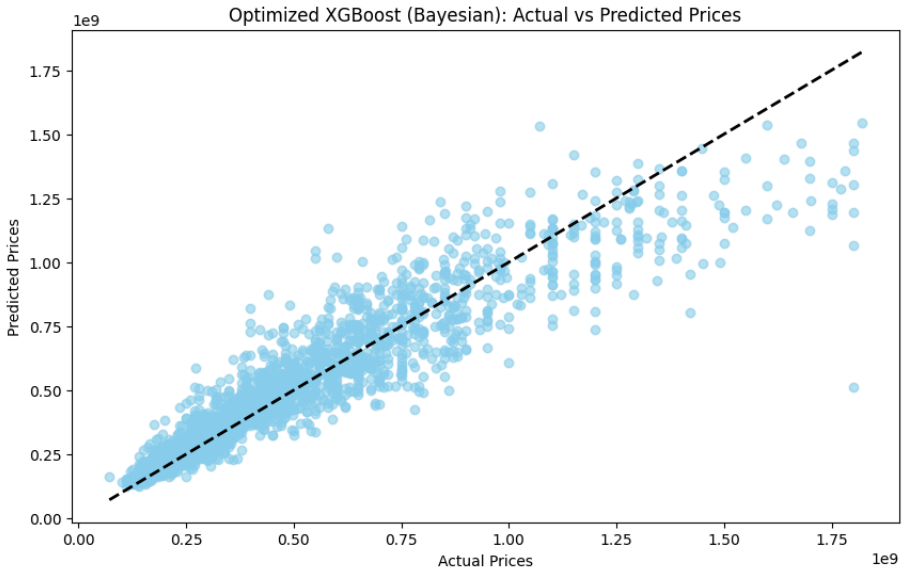
Para mejorar el rendimiento del modelo **XGBoost Regressor**, que ya tiene el mejor desempeño entre los modelos evaluados, implementaremos las siguientes estrategias:

1. Optimización de Hiperparámetros

El gráfico y los resultados muestran que la optimización bayesiana del modelo **XGBoost** mejoró significativamente el rendimiento:

Figura 32

Gráfico de valores reales vs predichos



- **Mean Absolute Error (MAE)**: 81,903,064, que representa una mejora en comparación con los modelos anteriores, lo cual indica que las predicciones tienen una menor desviación en promedio.
- **Mean Squared Error (MSE)**: 1.55e+16, y un **Root Mean Squared Error (RMSE)** de 124,524,740, lo cual muestra una reducción de los errores más grandes.
- **R^2 Score**: 0.85, lo que indica que el modelo explica el 85% de la variabilidad en los precios, mejorando respecto a las versiones anteriores.

Los **mejores hiperparámetros** encontrados fueron:

- subsample: 1.0
- reg_lambda: 0.001
- reg_alpha: 0.001
- n_estimators: 500
- max_depth: 10
- learning_rate: 0.01
- colsample_bytree: 0.5

Estos hiperparámetros ayudaron a mejorar el rendimiento del modelo al permitir una mejor generalización y capturar de manera más precisa la estructura de los datos.

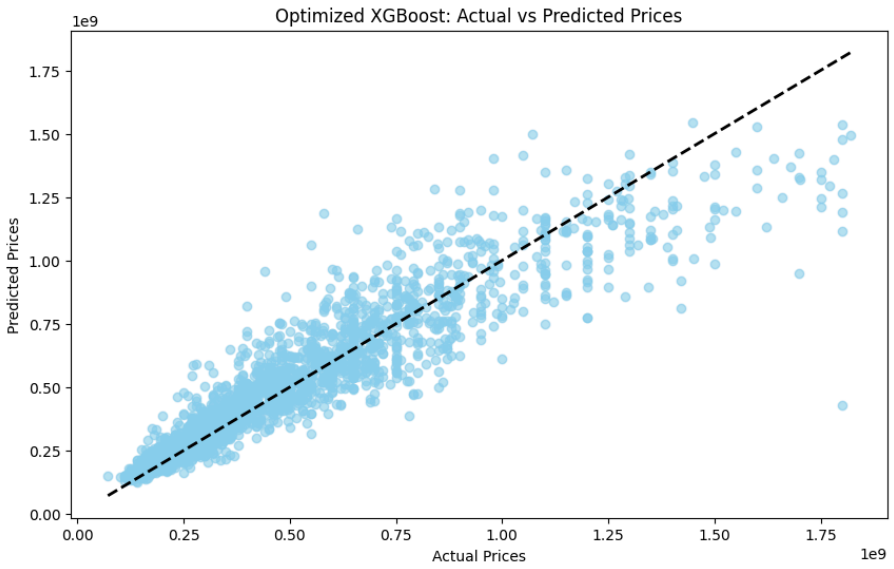
La optimización bayesiana logró mejorar el rendimiento del modelo **XGBoost**, logrando un **R^2** más alto y errores menores en comparación con otros enfoques.

2. Feature Engineering

El modelo **XGBoost** mejorado, con ingeniería de características y ajuste de hiperparámetros específicos, alcanzó un **R^2 de 0.85**, indicando que explica el 85% de la variabilidad en los precios, con un **MAE** de **81 millones** y un **RMSE** de **125.7 millones**, lo que refleja una mejora significativa en la precisión de las predicciones respecto a versiones anteriores. Estos resultados destacan el impacto positivo de las características derivadas y los ajustes óptimos en la mejora de la capacidad predictiva del modelo.

Figura 33

Gráfico de valores reales vs predichos



Las nuevas características creadas durante el proceso de ingeniería de características fueron las siguientes:

1. **Estrato_Tamano**: Esta característica se creó multiplicando el valor del **Estrato** por el **Tamano** de la propiedad. Esta característica captura cómo varía el tamaño de la propiedad en función del estrato socioeconómico, proporcionando una idea de las diferencias de tamaño según el nivel de vivienda.

2. **Banos_Habitaciones_Ratio**: Es la relación entre el número de **Baños** y el número de **Habitaciones**.
   Se añadió un valor pequeño (**1e-5**) al denominador para evitar la división por cero en casos en los que no hay habitaciones registradas. Esta característica ayuda a entender la disponibilidad de baños por habitación, lo cual puede influir en la comodidad y el valor percibido de la propiedad.

3. **Antiguedad_Municipio**: Se creó multiplicando la **Antigüedad** de la propiedad por el **Municipio**.
   Esta característica refleja cómo la antigüedad de las propiedades puede estar influenciada por la ubicación, permitiendo al modelo capturar mejor las diferencias de valor según la región y el estado de las propiedades.

Estas nuevas características fueron diseñadas para mejorar la capacidad del modelo de capturar relaciones importantes y complejas dentro de los datos, lo cual contribuyó al incremento del rendimiento predictivo.

3. Regularización

Para esta estrategia se probaron tres configuraciones de regularización diferentes para el modelo **XGBoost** con los hiperparámetros: reg_lambda (regularización L2) y reg_alpha (regularización L1). Estas regularizaciones controlan la magnitud de los coeficientes del modelo para evitar el sobreajuste.

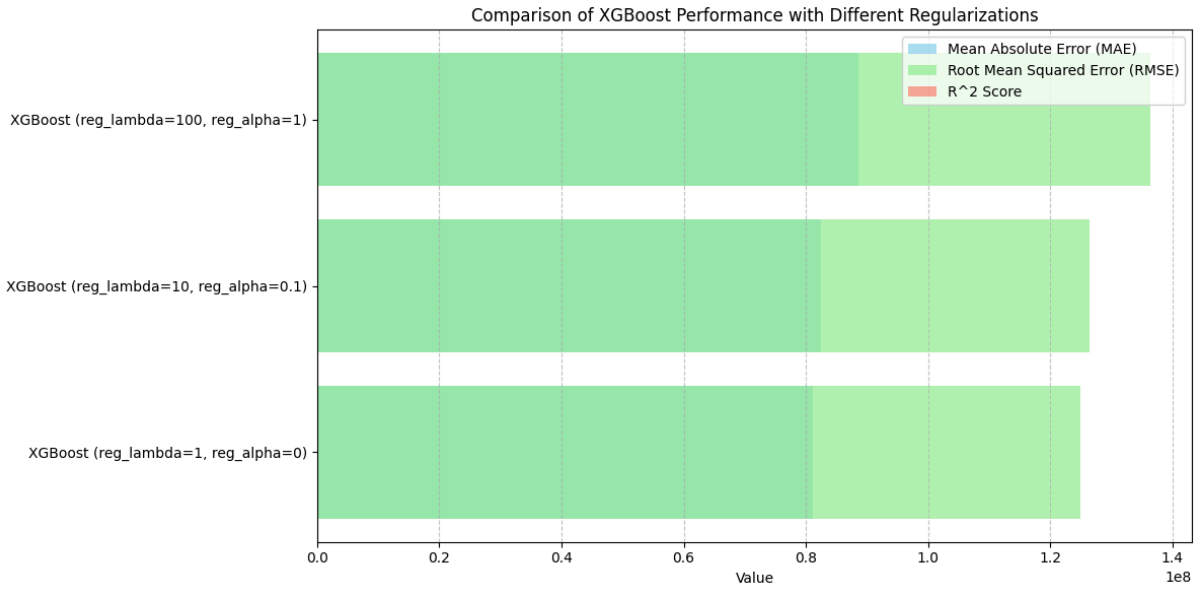
Las configuraciones evaluadas fueron:

1. reg_lambda=1, reg_alpha=0
2. reg_lambda=10, reg_alpha=0.1
3. reg_lambda=100, reg_alpha=1

El mejor rendimiento lo obtuvo la primera configuración (reg_lambda=1, reg_alpha=0), con un **R 2 Score** de **0.85**, **MAE** de **81,127,464**, y **RMSE** de **124,980,500**. Esta configuración presentó el menor error, indicando que el uso de una baja regularización permitió al modelo capturar las relaciones en los datos sin sobre ajustarse.

Figura 34

Gráfico de comparación de desempeño con diferentes hiperparametros



Las estrategias de optimización aplicadas al modelo XGBoost han demostrado ser efectivas para mejorar el rendimiento del modelo en la predicción de precios. La combinación de la optimización de hiperparametros, regularización adecuada y la ingeniería de características, y ajuste dinámico de la tasa de aprendizaje resultó en un modelo XGBoost optimizado con un **R^2 Score de 0.86** y un **MAE de 80,990,676**. Estos resultados reflejan un rendimiento preciso y confiable, mostrando que las estrategias de optimización aplicadas han mejorado significativamente la capacidad del modelo para capturar patrones en los datos, logrando una predicción más precisa sin caer en el sobreajuste. La estrategia de regularización con reg_lambda=1 y reg_alpha=0 fue la más efectiva para este conjunto de datos.

Dashboard

Con el objetivo de diseñar una herramienta analítica o de visualización que integre las bases de datos de Homty para generar Insights de valor sobre el mercado inmobiliario, se desarrolló un Dashboard en Power BI.

Planteamiento

Se siguió un proceso estructurado de 5 pasos:

1. ¿Para quién se va a comunicar?

   Para el equipo interno de Homty, ejecutivos con amplio conocimiento en el sector inmobiliario.

2. ¿Qué se va a comunicar?

   El Dashboard debe responder a las siguientes preguntas:

- ¿Cuál es el valor de metro cuadrado más económico por ciudad?.
- ¿Cuál es la tendencia del valor del metro cuadrado por ciudad?.
- ¿Cual es la tendencia de precios en apartamentos por ciudad ?.
- ¿A qué tasa han crecido los precios de apartamentos por ciudad?.
- ¿Cuál es el barrio que tiene mejor tasa de arriendo vs precio del inmueble?

3. ¿Por qué se va a comunicar?

   La visual analítica permitirá potenciar la estrategia de Homty y aperturar un nuevo modelo de negocio basado en el aporte de insights de mercado.

4. ¿Para qué se va a comunicar?

   Para llevar a los clientes de Homty información valiosa que se comercializará como activo, y simultáneamente establecer a Homty como fuente confiable de información del mercado inmobiliario.

5. ¿Cómo se va a comunicar?

   Se mostrarán las tendencias por tipo de inmueble, ciudades, barrios y cómo varía el precio por metro cuadrado en estas segmentaciones.

Visuales Analíticas Desarrolladas

Figura 35

Visual analítica de "Overview"

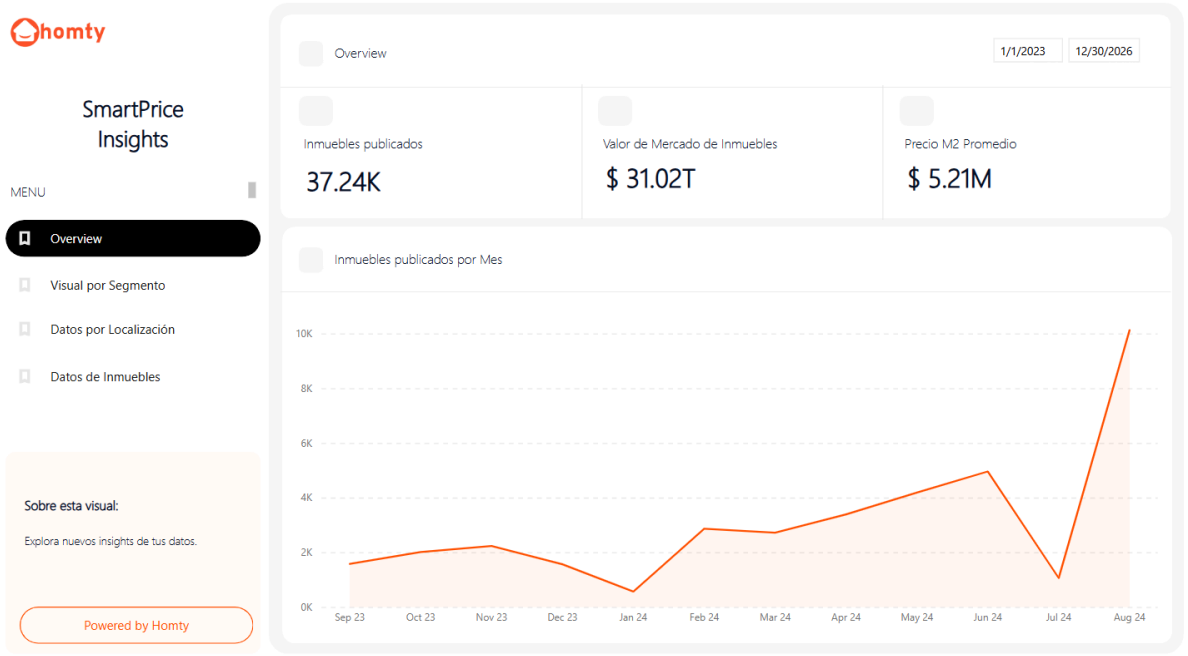


Figura 36

Visual analítica de "Visual por Segmento"

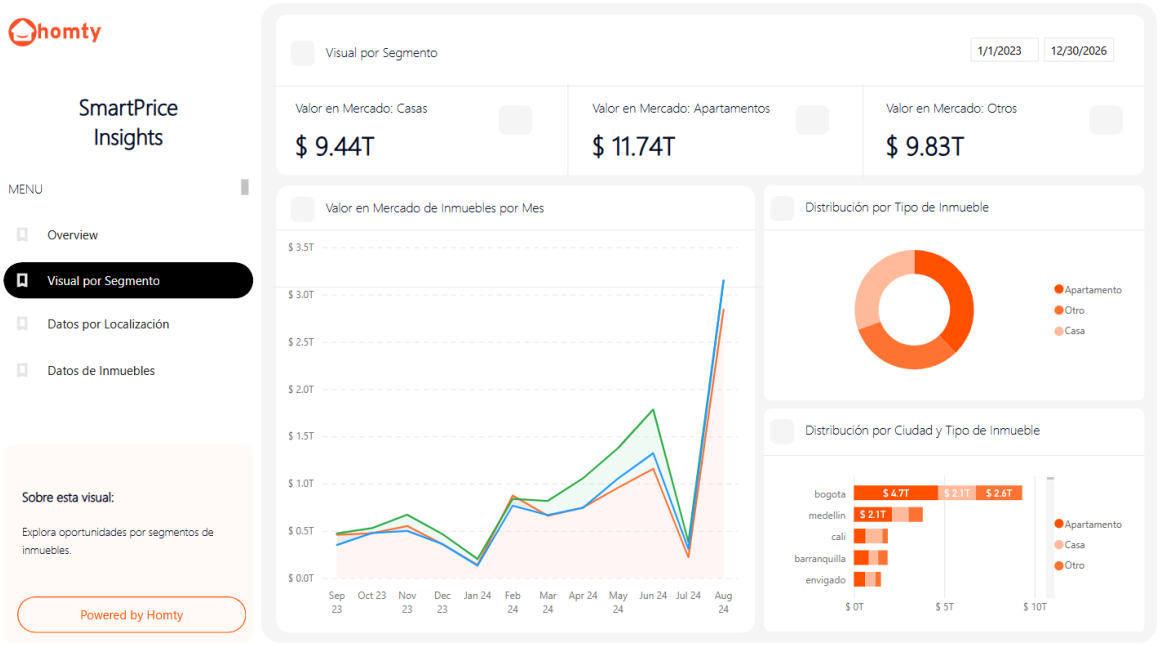


Figura 37

Visual analítica "Datos por Localización"

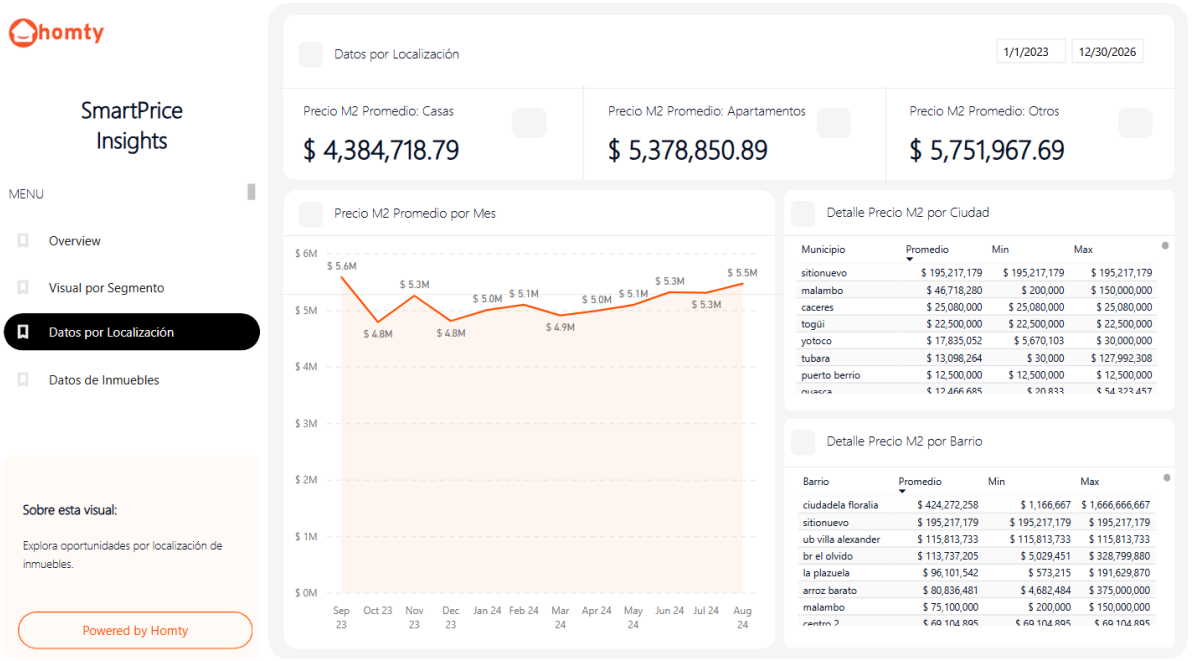
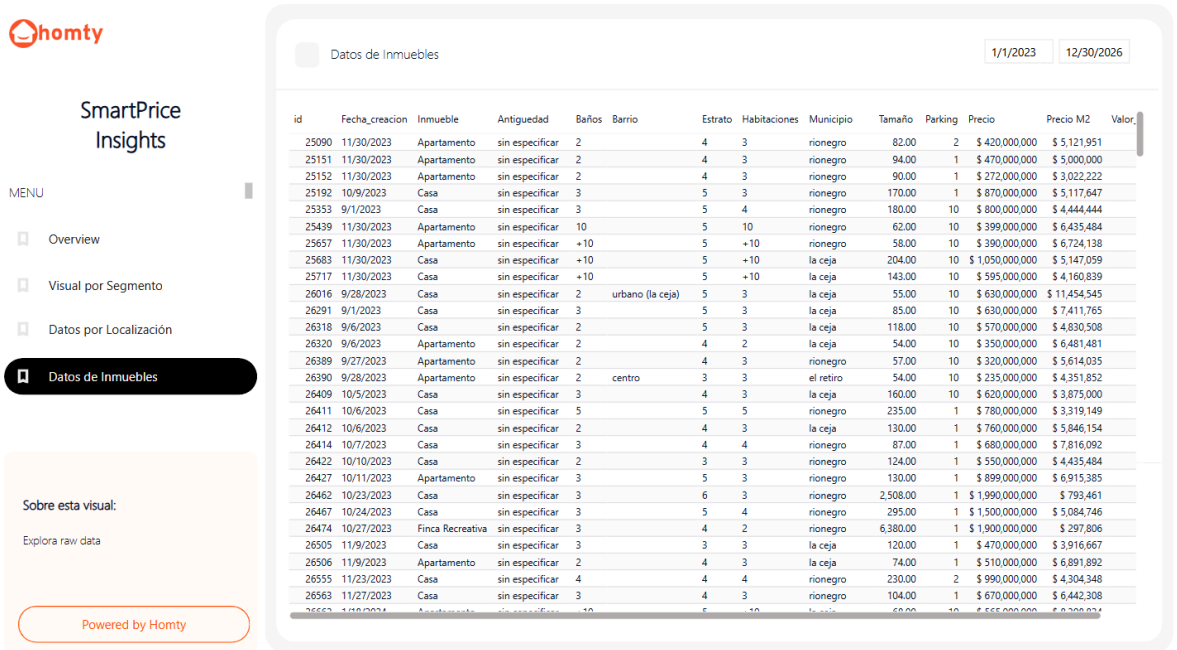


Figura 38

Visual analítica "Datos de Inmuebles"



Consideraciones Éticas

La ética en la valoración de inmuebles se refiere a la responsabilidad de los valuadores y profesionales del sector inmobiliario de actuar con integridad y transparencia (IVSC, 2020). La utilización de algoritmos y modelos predictivos plantea preguntas éticas sobre la equidad y la justicia en la

valoración. Si los modelos utilizados no son transparentes, pueden perpetuar sesgos y discriminación, afectando desproporcionadamente a ciertos grupos sociales (López, 2022).

Predicciones incorrectas de los modelos desarrollados en el proyecto pueden llevar a prácticas fraudulentas, como la sobrevaloración o subvaloración de propiedades, lo que no solo perjudica a los inversionistas, sino que también afecta la estabilidad del mercado inmobiliario.

Desde la perspectiva moral, la valoración también implica un compromiso con la justicia social. Las decisiones de valoración que priorizan el beneficio económico sobre el bienestar de la comunidad pueden ser vistas como inmorales. Por ejemplo, la sobrevaloración de la propiedad puede desplazar a comunidades vulnerables, impulsando la desigualdad social.

Las reflexiones de Adela Cortina (2014) en su obra sobre ética resaltan la importancia de la responsabilidad social y la interconexión entre individuos y comunidades. Cortina argumenta que no existe un individuo aislado, y que nuestras decisiones deben tener en cuenta el bienestar de los demás. Esto es especialmente relevante en el contexto de la valoración de inmuebles, donde las decisiones pueden tener un impacto significativo en la vida de las personas, especialmente en comunidades vulnerables o sectores de alto atractivo para inversionistas.

Resultados

Presentación de los Hallazgos

Desempeño de los Modelos Predictivos

El modelo seleccionado, XGBoost Regressor optimizado, mostró un desempeño superior en comparación con otras técnicas evaluadas:

- R^2 Score: 0.85, indicando que el modelo explica el 85% de la variabilidad en los precios de los inmuebles.
- Mean Absolute Error (MAE): 81 millones COP, evidenciando una menor desviación promedio entre los precios reales y los predichos.
- Root Mean Squared Error (RMSE): 125 millones COP, reflejando una mejora significativa en la precisión respecto a los modelos iniciales.

Principales Variables Explicativas

El análisis de importancia de las características reveló que las variables más influyentes en la predicción de precios fueron:

- Tamaño del inmueble: Contribuye significativamente al precio, mostrando una correlación positiva (r=0.74).
- Estrato socioeconómico: Altamente correlacionado con el precio, especialmente en inmuebles de estratos altos.
- Número de baños: Presenta una correlación directa, con incrementos en el precio a medida que aumenta la cantidad de baños.

Hallazgos Geográficos y Demográficos

- Municipios: Medellín lidera con los precios promedio más altos (580 millones COP), seguido de Bogotá (520 millones COP) y Cali (380 millones COP).
- Estratificación: Los estratos altos (5 y 6) muestran una variabilidad de precios más amplia, reflejando mercados menos homogéneos.
- Tendencias de tamaño: Las casas tienden a ser más grandes y costosas que los apartamentos, especialmente en Medellín.

Visualización Analítica

El dashboard desarrollado en Power BI permitió identificar tendencias clave como:

- La evolución del precio por metro cuadrado en las principales ciudades.
- Los barrios con mejor tasa de retorno en arriendos respecto al precio de compra.

Interpretación de Resultados

Comparación con las Expectativas Iniciales

- **Concordancia con hipótesis iniciales**:
  La relación positiva esperada entre el tamaño, estrato y el precio del inmueble fue confirmada por las correlaciones observadas (r>0.6). También se validó la hipótesis de que Medellín y Bogotá presentan dinámicas de mercado más competitivas en comparación con Cali.
- **Resultados inesperados**:
  - Propiedades con mayor antigüedad (>50 años) mostraron precios promedio superiores a inmuebles más nuevos (9-15 años), posiblemente debido a su ubicación estratégica o valor histórico.

- La variabilidad del precio por estrato no fue lineal; en algunos casos, inmuebles de estrato 4 presentaron precios más altos que los de estrato 5, sugiriendo factores adicionales, como amenidades o ubicación, que no fueron completamente captados por el modelo.

Impacto en las Decisiones Estratégicas

- La capacidad predictiva del modelo permite a Homty ofrecer recomendaciones más confiables a los inversionistas, optimizando la selección de inmuebles en función de los retornos esperados.
- La visualización de datos facilita identificar oportunidades en zonas emergentes y mejorar estrategias de comercialización en mercados subrepresentados, como Cali.

Relevancia de los Hallazgos

- Los insights generados no solo posicionan a Homty como un líder en análisis inmobiliario, sino que también abren oportunidades para integrar servicios de consultoría basados en datos en su portafolio.
- El modelo proporciona una base sólida para escalar el análisis a otras ciudades o mercados con características similares.

Discusión

Implicaciones de los Resultados

Para el problema de negocio o investigación los resultados muestran que la precisión alcanzada permite una valoración automatizada confiable de propiedades, considerando que el margen de error promedio es de $89 - $111 millones, lo cual es considerado aceptable teniendo en cuenta el rango de precios del mercado. Por último, cabe destacar que se pudo crear herramientas de apoyo para agentes inmobiliarios-

Comparación con Trabajos Similares

A diferencia de Sicilia Gómez (2024), incluyó variables de entorno urbano, nuestro estudio se limitó a características propias del inmueble. MV Perception (2022) enfatizó en la importancia de considerar eventos externos como la pandemia, este aspecto no fue incluido en nuestro análisis. A diferencia de Reim (2020), nuestro enfoque se centra más en características físicas de los inmuebles que en tendencias macroeconómicas. Similar al trabajo de Datsko (2024), encontramos que los modelos más complejos como XGBoost y Random Forest superan a los métodos más simples. A

diferencia de la mayoría de trabajos, este incorpora un análisis más detallado del impacto socioeconómico. El uso de múltiples métricas de evaluación (MAE, RMSE, R2) permiten una comparación más precisa entre modelos.

Limitaciones del Estudio

Ausencia de variables importantes como el estado de la propiedad, renovaciones, está en un conjunto residencial, entre otras. que pudieron aportar al modelo. Tamaño limitado de la muestra que puede afectar la generalización de los modelos. Desbalance en la representación de los tipos de propiedades y zonas geográficas- Falta de datos históricos que permitan analizar tendencias temporales del mercado.

Sugerencias para Trabajos Futuros

Ampliar el tamaño de la muestra mediante la inclusión de datos históricos, basarse en más de dos o tres fuentes de datos, y mejorar la cobertura geográfica. Desarrollar un sistema de recolección más robusto que permita incluir una mayor cantidad de características. Trabajar con datos como indicadores económicos, datos geográficos y socioeconómicos.

Conclusiones

Síntesis de los Principales Hallazgos y su Relevancia

El proyecto "Smart Price Insights" logró demostrar la capacidad de las técnicas de ciencia de datos e inteligencia artificial para optimizar el análisis y la predicción de precios en el mercado inmobiliario colombiano. La implementación de modelos avanzados, como XGBoost, permitió identificar patrones complejos y proyectar tendencias futuras con gran precisión.

Además, el desarrollo de dashboards interactivos en Power BI facilitó la visualización de datos y la generación de insights de valor, convirtiendo los resultados del análisis en herramientas útiles para la toma de decisiones estratégicas. Estos hallazgos son fundamentales para reducir la incertidumbre en la valoración de propiedades, promover la transparencia y mejorar la competitividad en el sector inmobiliario.

Cumplimiento de los Objetivos Planteados

El principal objetivo de desarrollar un sistema basado en ciencia de datos e inteligencia artificial para analizar y proyectar el comportamiento de precios en el mercado inmobiliario fue alcanzado exitosamente.

Se logró recopilar, estructurar y analizar los datos históricos de Homty, implementando técnicas avanzadas para identificar patrones y tendencias significativas. La correlación entre precios y

variables demográficas permitió enriquecer el análisis, proporcionando una visión integral del mercado.

Por último, la implementación de herramientas visuales, como los dashboards, cumplió con el objetivo específico de presentar de manera clara y accesible los resultados del análisis, lo cual fue clave para facilitar el entendimiento y uso de los insights generados.

Recomendaciones Prácticas Basadas en los Resultados del Análisis

Es necesario incrementar el volumen de datos disponibles para el análisis, ya que un mayor conjunto de datos contribuirá a la robustez y precisión de las predicciones. También se sugiere implementar herramientas que permitan capturar datos de los usuarios de manera más eficiente y precisa, así como optimizar el proceso de digitación para garantizar que las coordenadas geográficas de las propiedades sean un factor clave en la predicción de precios, mejorando así la calidad de los resultados.

Finalmente, se sugiere ofrecer a los clientes de Homty los insights generados a través de una plataforma abierta, como parte de una estrategia de diferenciación en el mercado, que permita a los inversionistas acceder a información valiosa para la toma de decisiones, fortaleciendo así el posicionamiento de Homty como un referente en el análisis de datos del mercado inmobiliario colombiano.

Referencias

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). *Introduction to Linear Regression Analysis* (5th ed.). John Wiley & Sons.

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5-32. https://doi.org/10.1023/A:1010933404324

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

Smola, A. J., & Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing, 14*(3), 199-222. https://doi.org/10.1023/B:STCO.0000035301.49549.88

Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory, 13*(1), 21–27. https://doi.org/10.1109/TIT.1967.1053964

Cruz Paredes, G. P. (2024). *Predicción de ventas de departamentos en el distrito de Miraflores de una empresa inmobiliaria de Lima utilizando el modelo de ensamble por medias*. Universidad Nacional Agraria La Molina.

Hernández Parrado, C. A., & Dávila Martínez, J. P. (2024). *Housing price prediction in Colombia using machine learning*. Trabajo académico, Universidad Nacional Agraria La Molina.

Sicilia Gómez, B. (2024). *Desarrollo de modelos de predicción de precios inmobiliarios utilizando técnicas de machine learning*. Universidad Nacional Agraria La Molina.

Roig Hernando, J., Gras Alomà, R., & Soriano Llobera, J. M. (2024). *Análisis y pronóstico del precio de la vivienda en España: Modelo econométrico desde una perspectiva conductual*. Trabajo académico, Universidad Nacional de España.

Datsko, A. (2024). *Análisis y predicción del precio de la vivienda en Madrid*. Universidad Politécnica de Madrid.

International Valuation Standards Council (IVSC). (2020). *Code of Ethics*. IVSC Publications.

López, P. (2022). *Desafíos éticos en el uso de inteligencia artificial para la valoración inmobiliaria*. Fundación Ética Digital.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). *CRISP-DM 1.0: Step-by-step data mining guide*. The CRISP-DM Consortium.