Clasificación y Predicción del Precio de las Propiedades en Bogotá

Natalia Triana Pulido, Sharon Gutierrez Perea, Brandom Alvarez Posada

BIT (Bogota Institute of Techonology)

Ciencia de Datos

Sebastian Moncada

26 de octubre, 2025

[redacted]

1.0. Introducción y Justificación

En este proyecto se busca analizar datos inmobiliarios de la ciudad de Bogotá con el fin de identificar la relación entre la seguridad de las localidades y el valor de las propiedades.

Los datos provienen de tres fuentes principales: dos archivos CSV (properties.csv y localidades.csv) y un archivo Excel convertido a CSV (tasa_seguridad.csv).

El análisis es relevante porque permite comprender cómo los factores sociales, como la tasa de homicidios y hurtos, pueden influir en los precios del mercado inmobiliario.

2.0 Objetivos del Proyecto

2.1 Objetivo General

Analizar y modelar la relación entre la seguridad y los precios de las propiedades en Bogotá.

2.2 Objetivo Específicos

- Unificar las distintas fuentes de datos en un único conjunto.
- Limpiar y procesar los datos para garantizar su calidad.
- Aplicar técnicas de análisis exploratorio para identificar patrones relevantes.
- Entrenar un modelo de clustering para clasificar propiedades según su nivel de lujo.
- Entrenar un modelo de Gradient Boosting (XGBoost) para predecir el precio de las propiedades.
- Entrenar un modelo de Random Forest Regressor para predecir el precio de las propiedades.

3.0 Análisis Exploratorio de Datos (EDA)

[redacted]

- **Dimensión del Dataframe**

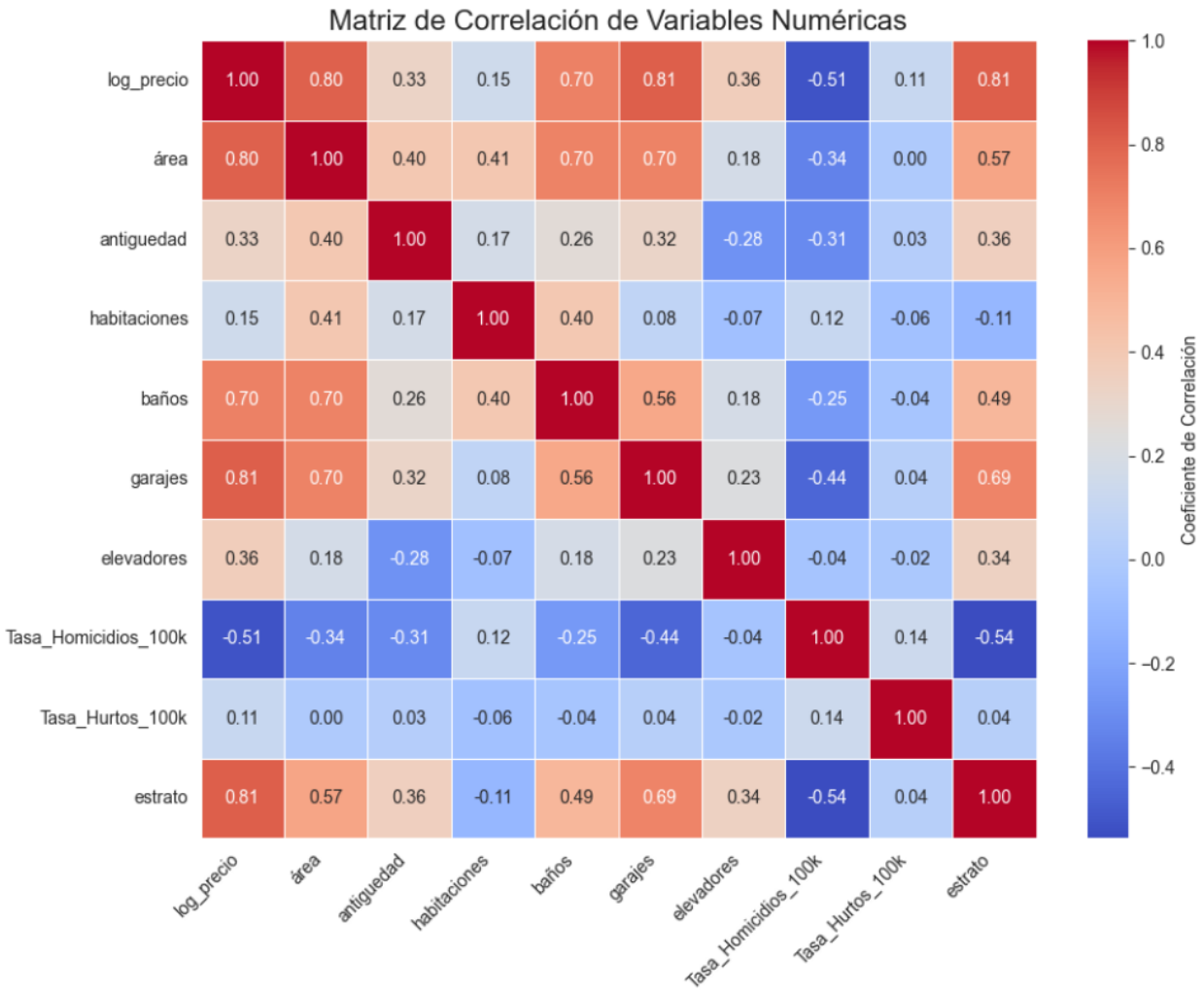
  El conjunto de datos correspondiente a las propiedades (properties.csv) contiene un total de 585 registros y 25 variables, que representan información detallada sobre las características físicas, económicas y de ubicación de los inmuebles analizados.

- **Variables más relevantes**

  Durante el análisis exploratorio se identificaron como variables más influyentes las de área, estrato y garajes, debido a su alta correlación con la variable objetivo precio.

  Figura 1

  Matriz de correlación de variables del dataset



Nota. La mayor correlación entre el precio y otras variables se encuentra centralizada en tres características principales: el área, el estrato y la cantidad de garajes, lo cual sugiere que estas variables impactan positivamente en la valoración de las propiedades.

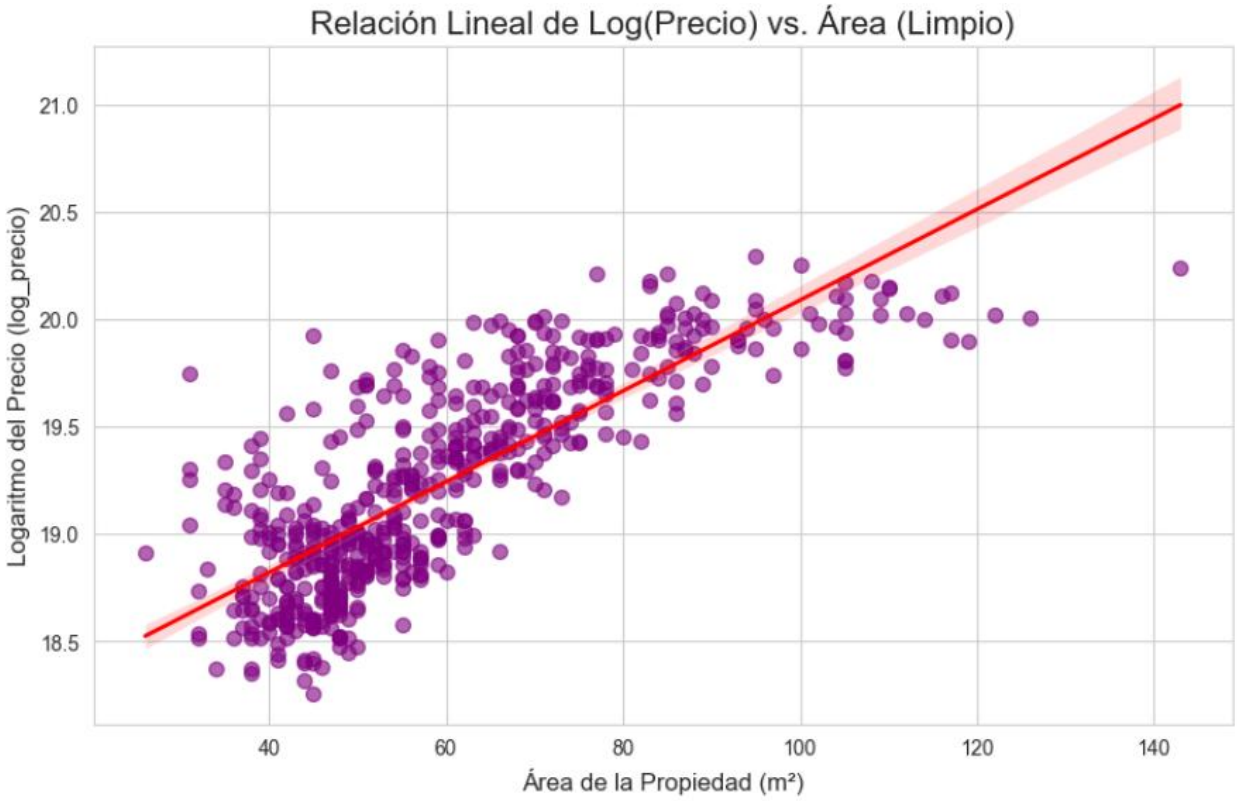
- **Patrones notables**

  Durante el desarrollo del EDA, se observó una relación lineal positiva entre las variables precio y área. Esto indica que, en general, a mayor área construida, mayor es el valor de la propiedad. Este comportamiento se muestra en la Figura 2, donde la tendencia ascendente refuerza la hipótesis de que el tamaño del inmueble es un factor determinante del precio.

  Figura 2

[redacted]

Relación lineal entre precio y área



Nota. La dispersión de los puntos muestra una tendencia lineal clara, confirmando la influencia directa del área sobre el precio final.

- **Hallazgos**

1. Existe una correlación positiva entre el área y el precio de las propiedades.
2. El estrato socioeconómico influye directamente en el valor del inmueble.
3. La cantidad de garajes impacta de forma moderada en el precio final.
4. Se detectaron outliers en área y precio, asociados a propiedades de lujo.
5. Algunas localidades similares presentan diferencias de precio por factores externos como ubicación o seguridad.
6. Las variables área, estrato y garajes se identificaron como las más relevantes para el modelado.

[redacted]

4.0 Procesamiento y Limpieza de Datos

Durante el proceso de limpieza y procesamiento de los datos se llevaron a cabo diversas transformaciones con el objetivo de garantizar la coherencia, integridad y calidad del conjunto de datos utilizado. A continuación, se detallan las principales acciones realizadas:

4.1 Tratamiento de Valores Nulos

Se reemplazaron los valores faltantes en las columnas gas y remodelado por la categoría 'no', bajo el supuesto de que la ausencia de información indica que las propiedades no cuentan con dichos servicios.

4.2 Imputación de Estrato

Los valores nulos en la columna estrato fueron reemplazados con el estrato correspondiente al barrio asociado a cada registro, asegurando consistencia con la información del contexto urbano.

4.3 Transformación de Variables Categóricas

Las columnas remodelado, depósito, zona de lavandería, gas y parqueadero, cuyos valores posibles eran “sí” o “no”, fueron convertidas a tipo booleano con el fin de facilitar su interpretación y posterior análisis en el modelado.

4.4 Conversión de Tipo de Dato

La columna estrato fue transformada a tipo entero (int) para permitir su uso en cálculos numéricos y modelos predictivos.

4.5 Eliminación de Variables Irrelevantes

Se eliminaron las columnas descripción, conjunto, nombre y dirección, ya que no aportaban información relevante para el análisis ni para el modelo a desarrollar.

Finalmente, una vez completada la limpieza de los datos, se realizó una integración con el dataset de localidades, mediante un proceso de merge, que permitió asociar cada barrio del conjunto principal con su respectiva localidad. Esta unión resultó fundamental para enriquecer el análisis con variables territoriales y sociales adicionales.

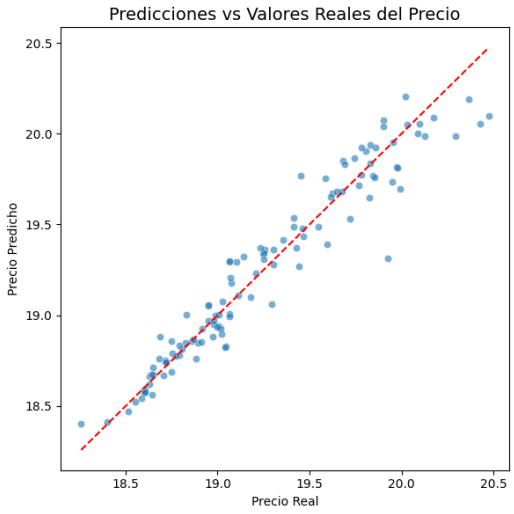
5.0 Modelo

5.1 Random Forest Regressor

Para predecir el precio de las viviendas se utilizó Random Forest Regressor, un modelo que combina múltiples árboles de decisión para generar predicciones más estables y precisas.

Su capacidad para manejar variables numéricas y categóricas, junto con su resistencia al sobreajuste, lo convierte en una herramienta ideal para este tipo de problemas.

El modelo alcanzó un R^2 de 0.93 y un RMSE de 0.14, demostrando una alta exactitud en la estimación de precios según las características físicas y de entorno




```
MAE (Error Absoluto Medio): 0.10
RMSE (Raíz del Error Cuadrático Medio): 0.14
R² (Coeficiente de Determinación): 0.931
```

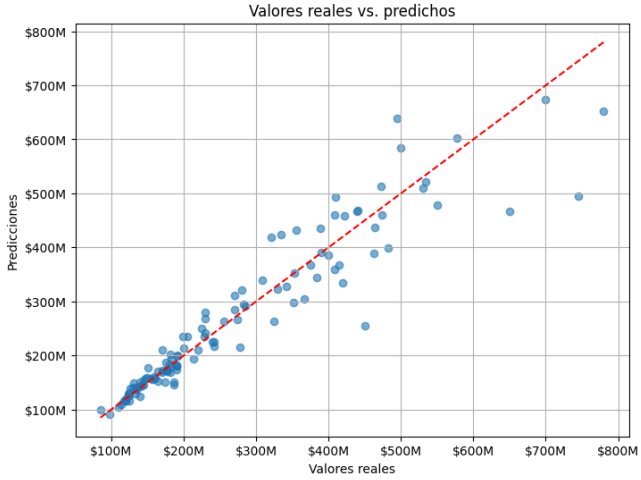
[redacted]

5.2 Gradient Boosting (XGBoost)

También se empleó XGBoost, un modelo basado en el enfoque de Gradient Boosting, que mejora iterativamente los errores de predicción.

Este algoritmo destaca por su eficiencia, precisión y capacidad para capturar relaciones no lineales entre variables, además de incluir regularización que evita el sobreajuste.

Su flexibilidad y robustez lo hacen ideal para problemas complejos de predicción de precios inmobiliarios.



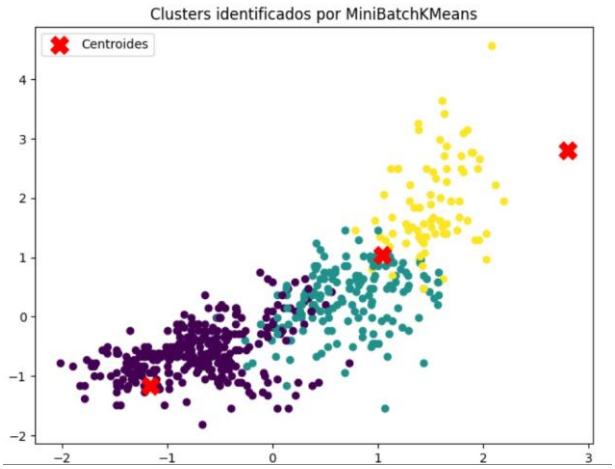
5.3 K-Means Clustering

Como modelo de agrupación, se aplicó K-Means Clustering para segmentar las propiedades según sus características y precios.

Mediante el método del codo y el índice de silueta, se determinó el número óptimo de grupos, identificando diferentes niveles de lujo y patrones de valor.

Este análisis complementó la predicción, permitiendo una mejor comprensión de las tendencias del mercado inmobiliario.

[redacted]




```
Calinski-Harabasz: 1018.164
Davies-Bouldin: 0.816
Coeficiente de Silueta: 0.55
```

6.0 Resultados y Conclusiones

Los modelos implementados demostraron un desempeño sólido y coherente con los objetivos planteados.

El Random Forest Regressor obtuvo un R^2 de 0.93, evidenciando una excelente capacidad predictiva sobre el precio de las viviendas.

El XGBoost reforzó estos resultados al ofrecer un balance entre precisión y eficiencia, confirmando su utilidad para datos complejos y heterogéneos.

Por su parte, el K-Means Clustering permitió identificar agrupaciones de propiedades con distintos niveles de lujo, aportando una visión complementaria al análisis predictivo.

En conjunto, los hallazgos evidencian que las variables área, estrato y garajes tienen un impacto significativo en el valor de las propiedades, y que los modelos aplicados logran capturar estas relaciones con alta precisión.

El proyecto demostró la viabilidad del uso de técnicas de machine learning en el análisis inmobiliario, combinando predicción y segmentación para obtener información de valor.

[redacted]

7.0 Siguientes Pasos / Futuro del Proyecto

Como continuación del trabajo, se plantea:

- Incorporar más variables contextuales, como accesibilidad, transporte o cercanía a servicios.
- Optimizar los hiperparámetros de los modelos para mejorar su desempeño.
- Implementar un panel interactivo que permita visualizar las predicciones y clusters en tiempo real
- Explorar modelos más avanzados, como redes neuronales o ensambles híbridos, para comparar su rendimiento.

En futuras versiones, este enfoque podría convertirse en una herramienta de apoyo a la toma de decisiones inmobiliarias, orientada tanto a inversionistas como a entidades urbanísticas.