Análisis de Datos de Viviendas en Bogotá, Colombia

Santiago Niño
Universidad Sergio Arboleda
Bogotá, Colombia
Fecha: 30 de octubre del 2024

***Abstract*—Este documento presenta un análisis de un conjunto de datos que contiene información sobre viviendas en Bogotá, Colombia. Se exploran características clave del mercado inmobiliario, incluyendo precios de venta, distribución geográfica y relaciones entre variables relevantes.**

I. Background

En Bogotá, Colombia, el mercado inmobiliario ha experimentado cambios significativos en la última década, influenciados por factores económicos, sociales y ambientales. Este estudio busca comprender mejor las dinámicas de precios de las propiedades y la distribución geográfica de las viviendas, ofreciendo una visión general del fenómeno habitacional en la ciudad.

II. Descripción de los datos

A. Características del conjunto de datos

El conjunto de datos analizado contiene 44,436 registros y 16 columnas que describen diversas características de las propiedades. Las variables incluyen:

- **tipo_propiedad**: Tipo de propiedad (apartamento, casa, etc.).
- **tipo_operacion**: Tipo de operación (venta, arriendo).
- **precio_venta**: Precio de venta de la propiedad.
- **area**: Área de la propiedad en metros cuadrados.
- **habitaciones**: Número de habitaciones.
- **banos**: Número de baños.
- **latitud** y **longitud**: Coordenadas geográficas de la propiedad.

B. Fuentes complementarias

Además de los datos originales, se consultaron fuentes como el DANE (Departamento Administrativo Nacional de Estadística) y plataformas de bienes raíces para complementar el análisis.

C. Limpieza y transformaciones

Se realizaron diversas transformaciones y limpieza de datos, incluyendo:

- Manejo de valores faltantes en la variable *caracteristicas*.
- Conversión de la variable *antiguedad* a formato adecuado.
- Eliminación de duplicados y datos inconsistentes.

III. Análisis Exploratorio y Preguntas para Contrastar Predictores

El análisis exploratorio se centró en responder a diversas preguntas clave que permiten entender mejor las relaciones entre las variables del conjunto de datos. A continuación, se presentan las preguntas formuladas y las relaciones analizadas:

A. Preguntas para Contrastar Predictores

1) ¿Cuál es la relación entre el área y el precio de venta de las viviendas?: Se analizó la relación entre dos variables numéricas: el área de las propiedades y el precio de venta. Se utilizó el coeficiente de correlación de Pearson para evaluar la fuerza y dirección de esta relación. Un valor cercano a cero indicaría una correlación muy débil. Para más información sobre el coeficiente de correlación de Pearson, consulte [1].

2) ¿Cómo varía el precio de venta según el tipo de propiedad?: Se investigó la relación entre una variable numérica (precio de venta) y una variable categórica (tipo de propiedad). Se aplicó el análisis de varianza (ANOVA) para comparar las medias del precio de venta entre diferentes tipos de propiedad, determinando si existen diferencias significativas. Para detalles sobre el ANOVA, consulte [2].

3) ¿Existe una diferencia significativa en el número de habitaciones entre las diferentes antigüedades de las propiedades?: Se examinó la relación entre el número de habitaciones (variable numérica) y la antigüedad de las propiedades (variable categórica). La prueba de Kruskal-Wallis se utilizó para evaluar las diferencias en el número de habitaciones según la antigüedad, identificando si al menos un grupo difiere significativamente de los demás. Para más información sobre la prueba de Kruskal-Wallis, consulte [3].

4) ¿Cómo se distribuye el precio de venta según el sector de la propiedad?: Se analizó la relación entre el precio de venta (variable numérica) y el sector de la propiedad (variable categórica). Dependiendo de la distribución de los datos, se aplicó ANOVA o la prueba de Kruskal-Wallis para determinar si hay diferencias significativas en los precios de venta entre los distintos sectores. Para un análisis detallado de la prueba de Kruskal-Wallis, consulte [3].

5) ¿Hay alguna relación entre el estado de la propiedad y el estrato socioeconómico?: Se evaluó la relación entre dos variables categóricas: el estado de la propiedad (nuevo o usado) y el estrato socioeconómico. Se realizó un análisis de Chi-cuadrado de independencia para determinar si existe

una relación significativa entre estas variables categóricas. Para información adicional sobre el análisis de Chi-cuadrado, consulte [4].

Estos análisis permiten una mejor comprensión de las dinámicas que rigen el mercado inmobiliario en Bogotá y ofrecen información valiosa para futuras investigaciones y decisiones de inversión.

IV. Resultados

A. Relación entre área y precio de venta

El análisis de correlación entre el área y el precio de venta de las propiedades reveló un coeficiente de Pearson de aproximadamente 0.0031, indicando una correlación prácticamente nula. Esto sugiere que, en el mercado inmobiliario de Bogotá, el área de la propiedad no es un factor determinante del precio de venta. A diferencia de otros mercados, donde una mayor área puede relacionarse directamente con un aumento en el precio, en Bogotá parece que otros factores—como la ubicación, el tipo de propiedad y las características adicionales—tienen un papel más relevante en la determinación del precio.

La mayoría de las áreas de las viviendas se sitúan entre 0 y 1000 metros cuadrados, lo que indica que existe una distribución sesgada en el conjunto de datos. A continuación, se presenta un gráfico que ilustra la relación entre el área y el precio de venta.

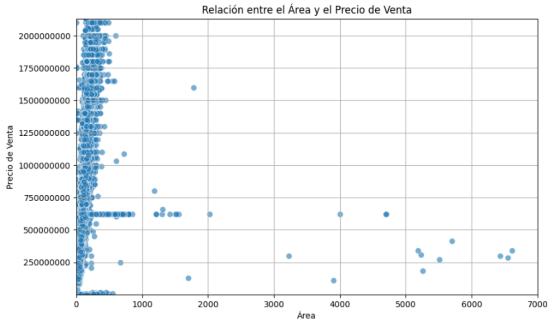

Fig. 1. Relación entre área y precio de venta de las propiedades en Bogotá.

B. Precio de venta según tipo de propiedad

Para analizar la influencia del tipo de propiedad en el precio de venta, se aplicó un análisis de varianza (ANOVA), el cual mostró un estadístico F de 53.94 y un p-valor de 4.00e-24. Estos resultados sugieren que existen diferencias estadísticamente significativas en el precio de venta entre distintos tipos de propiedades. En otras palabras, el tipo de propiedad (ya sea apartamento, casa, entre otros) tiene un impacto importante en la valoración de mercado. La Fig. 2 muestra un boxplot que representa la distribución de precios para cada tipo de propiedad, ilustrando que algunos tipos presentan una mayor variabilidad en los precios que otros.

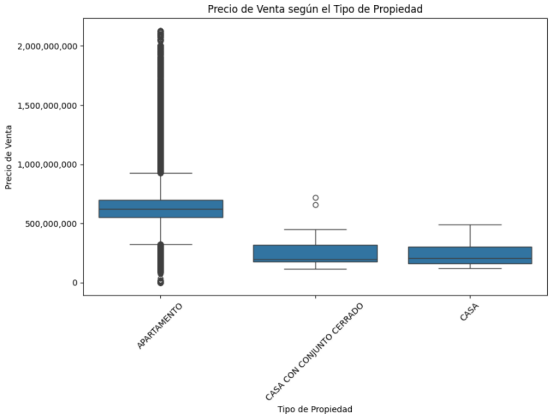

Fig. 2. Distribución del precio de venta según tipo de propiedad.

C. Diferencias en habitaciones según antigüedad

Para evaluar si existen diferencias en el número de habitaciones de acuerdo con la antigüedad de las propiedades, se realizó una prueba de Kruskal-Wallis. Los resultados indicaron un estadístico H de 4248.11 y un p-valor de 0.0, lo cual respalda la hipótesis de que existen diferencias significativas en el número de habitaciones dependiendo de la antigüedad. Este hallazgo sugiere que las propiedades más nuevas tienden a tener un diseño distinto en términos de cantidad de habitaciones comparadas con las propiedades más antiguas. La Fig. 6 muestra un boxplot de esta distribución, evidenciando cómo las propiedades antiguas suelen tener mayor variabilidad en el número de habitaciones.

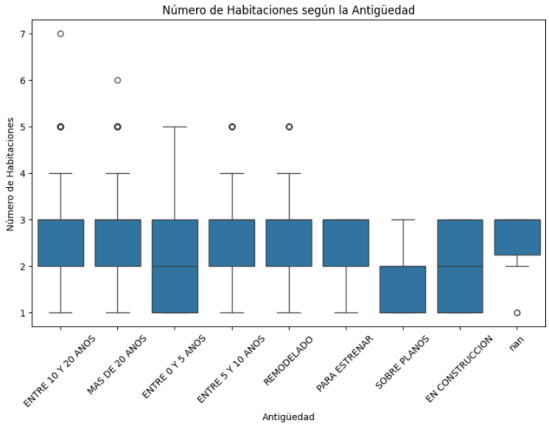

Fig. 3. Distribución del número de habitaciones según antigüedad de la propiedad.

D. Precio de venta según número de parqueaderos

Se llevó a cabo una prueba de Kruskal-Wallis para investigar si existen diferencias en el precio de VENTA según el número de parqueaderos disponibles en cada propiedad. Con un estadístico H de 10,766.96 y un p-valor de 0.0, los resultados indican diferencias significativas en el precio de arriendo entre las propiedades de acuerdo con la cantidad de parqueaderos.

Este hallazgo es indicativo de que las propiedades con más parqueaderos suelen tener un precio de venta superior, probablemente debido a la conveniencia adicional que representan para los vendedores.

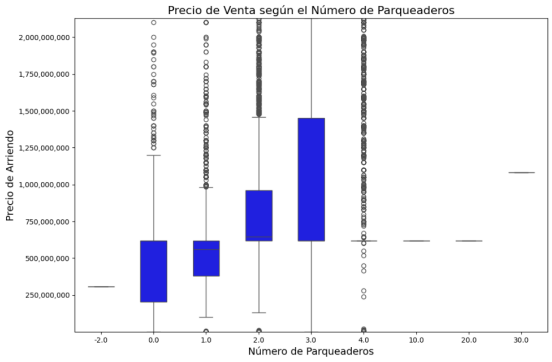

Fig. 4. Precio de venta según número de parqueaderos.

E. Relación entre estado y estrato socioeconómico

Para evaluar la relación entre el estado de la propiedad (nuevo o usado) y el estrato socioeconómico, se realizó un análisis de Chi-cuadrado. Los resultados mostraron un estadístico de 3,879.10 y un p-valor de 0.0, indicando una asociación significativa entre ambas variables. Esto sugiere que el estado de una propiedad influye en el estrato al que pertenece, evidenciando que las propiedades nuevas tienden a encontrarse en estratos específicos, posiblemente debido a la oferta y demanda en zonas en desarrollo o renovación urbana.

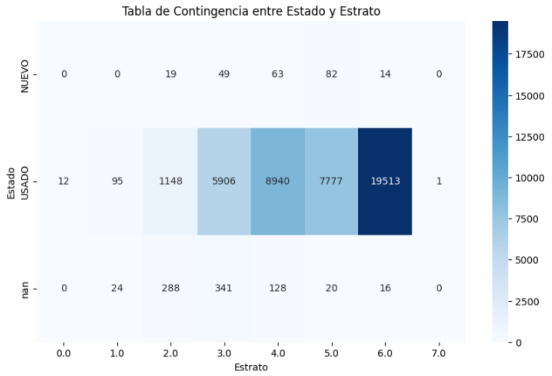

Fig. 5. Relación entre estado y estrato socioeconómico

V. Visualización de la Distribución de Precios en Bogotá

Para comprender la distribución espacial de los precios de venta de las viviendas en Bogotá, se creó un mapa de calor centrado en la ciudad, utilizando las coordenadas de cada propiedad. Este mapa ofrece una representación visual de las variaciones de precios en diferentes zonas de la capital. A continuación, se detalla el procedimiento y los hallazgos obtenidos.

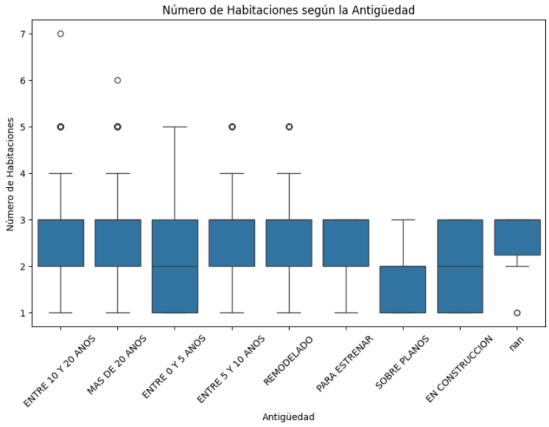

Fig. 6. Distribución del número de habitaciones según antigüedad de la propiedad.

A. Construcción del Mapa de Calor

Se utilizó la biblioteca `folium` de Python para generar un mapa centrado en las coordenadas de Bogotá. A partir de los datos geográficos de cada propiedad—latitud, longitud y precio de venta—se prepararon los puntos de datos que fueron posteriormente añadidos al mapa en forma de capas de calor. El código principal para esta visualización fue el siguiente:

El mapa de calor generado proporciona una visualización directa de la intensidad de los precios de las viviendas a lo largo de la ciudad. El *radius* se estableció en 15 para ajustar la densidad de las zonas con mayor o menor concentración de precios.

B. Interpretación del Mapa de Calor

El análisis visual del mapa de calor revela patrones geográficos importantes en los precios de las propiedades en Bogotá:

- **Zona Nororiental**: Esta área se caracteriza por una alta concentración de precios elevados. Las áreas de Chapinero y Usaquén, que abarcan sectores como La Cabrera, Chicó y Rosales, son conocidas por sus propiedades de lujo y su cercanía a centros financieros y comerciales, lo que contribuye a los altos valores observados en esta región.
- **Zona Media de la Ciudad**: Las zonas cercanas al centro de Bogotá muestran una variabilidad en los precios, con áreas que presentan tanto precios elevados como moderados. Esto puede deberse a la combinación de propiedades antiguas, instituciones gubernamentales y desarrollos habitacionales recientes que caracterizan esta región.
- **Zona Sur**: En contraste con el norte, la zona sur presenta una concentración de precios más bajos. Este fenómeno es común en sectores con una menor densidad de desarrollos residenciales de alta gama y en zonas más alejadas de los principales centros económicos de la ciudad.

C. Analisis

La visualización a través del mapa de calor sugiere una correlación clara entre la ubicación geográfica y los precios de las viviendas en Bogotá. Los precios más altos se concentran en la zona nororiental, lo que destaca la desigualdad en la distribución de los valores inmobiliarios en la ciudad. Este hallazgo puede ser útil tanto para inversionistas como para las autoridades, quienes podrían orientar estrategias de desarrollo urbano considerando estos patrones.

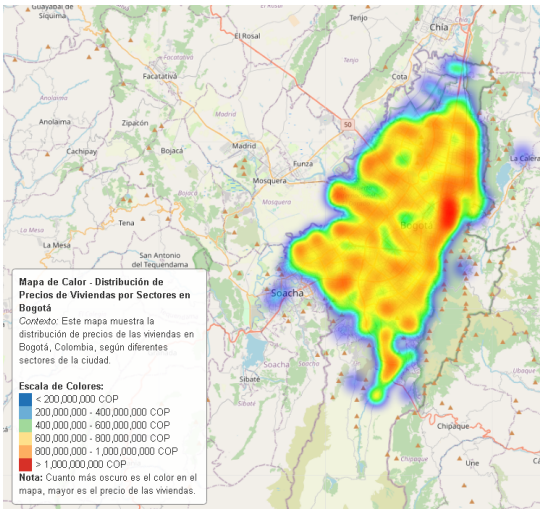

Fig. 7. Mapa de calor de los precios de las viviendas en Bogotá.

D. Construcción del Mapa de Concentración de Áreas de Viviendas

Se utilizó la biblioteca `folium` de Python para generar un mapa que representa la concentración de áreas de viviendas en Bogotá. Este mapa visualiza las áreas pequeñas, medianas y grandes, con los colores verde, amarillo y rojo, respectivamente. El código principal para esta visualización fue el siguiente:

El mapa de concentración de áreas de viviendas permite observar de manera clara la distribución de las diferentes categorías de áreas a lo largo de la ciudad.

E. Interpretación del Mapa de Concentración de Áreas de Viviendas

El análisis visual del mapa revela patrones importantes en la distribución de las áreas de viviendas en Bogotá:

- **Zona Nororiental**: Esta área destaca por una notable concentración de áreas de viviendas. Se observa que muchas de las propiedades en esta región pertenecen a las categorías de áreas medianas y pequeñas.
- **Distribución General**: La mayoría de las viviendas en el conjunto de datos estudiado son de áreas medianas, con algunas pequeñas, y estos grupos están presentes en gran parte de la ciudad de Bogotá. Esta distribución sugiere una tendencia hacia propiedades de tamaño moderado en la ciudad.

F. Análisis

La visualización a través del mapa de concentración de áreas de viviendas indica que la zona nororiental de Bogotá es un área clave en términos de densidad habitacional. La predominancia de áreas medianas y pequeñas refleja una preferencia por propiedades más accesibles en comparación con áreas más grandes. Este hallazgo puede ser relevante para desarrolladores y planificadores urbanos al considerar la oferta de viviendas en la ciudad.

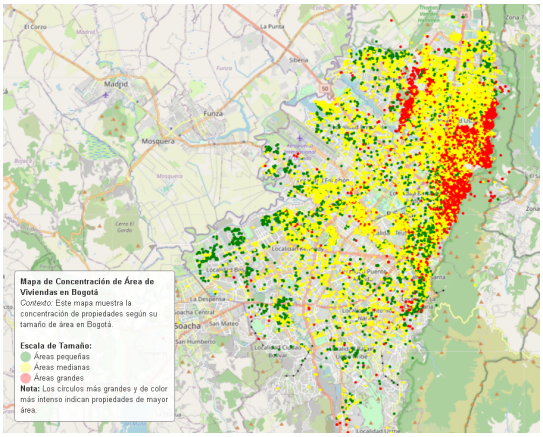

Fig. 8. Mapa de concentración de áreas de viviendas en Bogotá.

VI. Resultados de Modelos

En este estudio, se desarrolló un modelo predictivo para estimar el precio de las viviendas en Bogotá, Colombia. Para ello, se implementaron procesos de preprocesamiento que incluyeron la agrupación de categorías poco frecuentes y la eliminación de valores atípicos, con el objetivo de mejorar la calidad del conjunto de datos y, en consecuencia, la precisión de los modelos.

A. Procesos de Preprocesamiento

Se utilizó un enfoque para agrupar las categorías que presentaban una frecuencia inferior al umbral del 5% en una categoría denominada “Otros”. Este procedimiento es fundamental, ya que reduce la complejidad del modelo y minimiza el riesgo de sobreajuste, lo que puede mejorar el rendimiento general en la predicción de precios.

Además, se implementó un método para eliminar valores atípicos en las columnas numéricas utilizando el rango intercuartílico (IQR). Este método asegura que los valores extremos que pueden distorsionar las predicciones se eliminen, permitiendo que el modelo se enfoque en datos representativos. Esta estrategia es crucial, dado que los valores atípicos pueden tener un impacto desproporcionado en métricas como el error cuadrático medio (MSE).

B. Modelos Evaluados

Se evaluaron diferentes modelos de regresión, incluyendo KNN, Regresión Lineal, Regresión Polinómica y Random Forest. Cada uno de estos modelos fue analizado en términos de su capacidad para predecir el precio de las viviendas.

C. Comparación de Modelos

Los resultados evidencian que el modelo de Random Forest se destaca como la mejor opción para predecir el precio de las viviendas en Bogotá. Este modelo presenta un mejor rendimiento en comparación con KNN, destacándose en métricas clave que indican precisión y capacidad de generalización.

El coeficiente de determinación (R^2) de Random Forest muestra una mayor capacidad para explicar la variabilidad de los precios de las viviendas en comparación con KNN. Esto significa que Random Forest es capaz de capturar patrones complejos en los datos, lo que se traduce en predicciones más precisas.

La comparación de las métricas de R^2 entre el conjunto de entrenamiento y el conjunto de prueba para KNN y Random Forest se muestra en la siguiente figura:

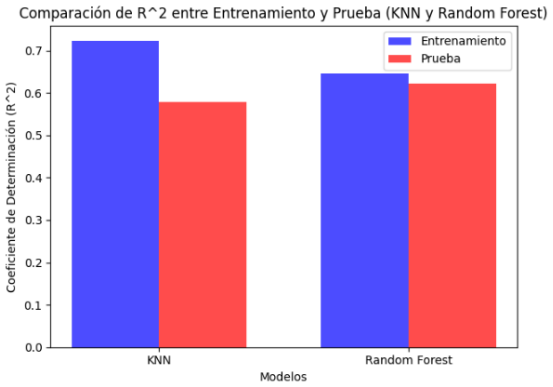

Fig. 9. Comparación de R^2 entre KNN y Random Forest

En conclusión, el modelo de Random Forest se elige como la opción preferida debido a sus mejores métricas de rendimiento en comparación con otros modelos, lo que sugiere que proporciona una mayor consistencia y fiabilidad en la predicción de precios. Esta capacidad es especialmente relevante en el contexto de la predicción de precios de viviendas en un mercado dinámico como el de Bogotá.

El modelo de Random Forest ha generado un conjunto de predicciones para los precios de las propiedades en Bogotá, evidenciando un rango diverso de resultados en comparación con los valores reales. En ciertos casos, las predicciones se acercan notablemente a los precios reales, lo que sugiere que el modelo es capaz de identificar patrones significativos en los datos. No obstante, también se observan varias instancias donde las predicciones difieren considerablemente de los valores reales. Estas discrepancias indican que, a pesar de la utilidad del modelo, aún existe un margen de mejora.

Es probable que las características utilizadas en el modelo no abarquen toda la complejidad del mercado inmobiliario en Bogotá. Esto podría deberse a variables no consideradas que podrían influir en los precios o a la necesidad de realizar ajustes adicionales en el modelo, como la optimización de hiperparámetros o la incorporación de nuevas variables.

En resumen, los resultados obtenidos enfatizan la necesidad de continuar el proceso de refinamiento del modelo y de explorar diferentes enfoques y características. Esto permitirá aumentar la precisión de las predicciones y, por ende, facilitar una comprensión más profunda del mercado inmobiliario.

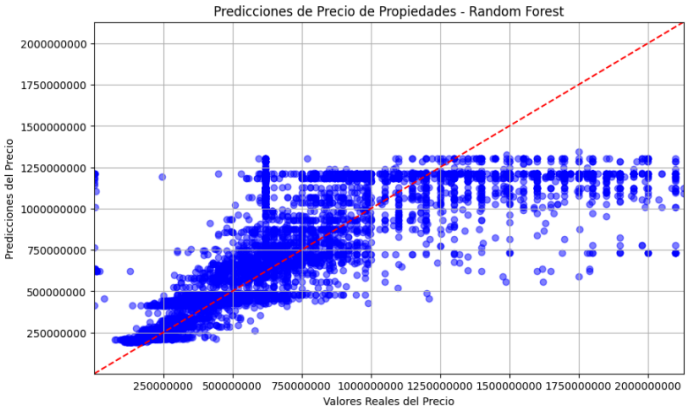

Fig. 10. Predicciones de precio de propiedades vs. valores reales usando Random Forest

VII. Conclusiones

El análisis revela patrones significativos en el mercado inmobiliario de Bogotá, destacando la importancia de variables como el tipo de propiedad y la ubicación geográfica en la determinación de precios. A diferencia de mercados tradicionales, donde el área suele ser un factor determinante en el precio, en Bogotá el impacto del área es mínimo, mientras que características como el tipo de propiedad y la disponibilidad de parqueaderos resultan determinantes en los precios de venta y arriendo. Este estudio proporciona una base para futuras investigaciones en el campo de la vivienda en la capital colombiana y puede contribuir al desarrollo de estrategias de inversión y políticas habitacionales más informadas.

References

[1] K. Pearson, "Mathematical contributions to the theory of evolution. On a form of spurious correlation which may arise when indices are used in the measurement of organs," *Proceedings of the Royal Society of London*, vol. 66, no. 428-439, pp. 489-498, 1900.

[2] R. A. Fisher, "Statistical methods for research workers," *Oliver and Boyd*, 1925.

[3] W. H. Kruskal and W. A. Wallis, "Use of ranks in one-criterion variance analysis," *Journal of the American Statistical Association*, vol. 47, no. 260, pp. 583-621, 1952.

[4] K. Pearson, "On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling," *Philosophical Magazine*, vol. 50, no. 302, pp. 157-175, 1900.