Informe del Proyecto

Integrantes

- Giovanny Casas Agudelo
- Carmen Carvajal Gutiérrez
- Camilo Arango Yepes
- Andrés González Restrepo

Análisis de precios de propiedades en Antioquia a partir de características físicas y localización

Resumen Ejecutivo

Este proyecto analiza los determinantes del precio de las viviendas en el departamento de Antioquia (Colombia) utilizando un conjunto de datos inicial de ~1.000.000 de datos de todo el país (Kaggle). Se filtró la información para Antioquia (341.453 registros) y, tras un proceso intensivo de depuración y enriquecimiento (imputación desde campos de texto en *descripción* y *título*), se consolidó un dataset analítico de 21.192 registros con las variables clave completas (precio, área, barrio, municipio, número de habitaciones y baños).

Se documenta la metodología de preparación de datos, exploración, modelado y validación; se desarrollan modelos predictivos para el precio y el precio por metro cuadrado; se evalúa su desempeño con métricas estándar; y se derivan hallazgos sobre patrones de oferta por barrio y el impacto marginal de las características físicas y de localización sobre el valor.

---

1. Planteamiento del Problema

La fijación de precios inmobiliarios es un problema multivariado donde convergen factores físicos (área, número de habitaciones y baños) y localización (barrio, municipio). El objetivo consiste en comprender y cuantificar el efecto de dichas características sobre el precio de las viviendas en Antioquia, y proveer herramientas analíticas para monitorear la oferta por barrio y apoyar el análisis y la toma de decisiones en torno a la valoración de viviendas.

Objetivo general

Analizar el precio de las viviendas en Antioquia con respecto a sus características (área, ubicación, número de habitaciones y baños).

Objetivos específicos

- Identificar patrones de oferta por barrio.
- Analizar el impacto de las características de las viviendas sobre el valor de la propiedad.

---

2. Datos y Preparación

El punto de partida fue un *dataset* de Kaggle de aproximadamente 1 millón de registros a nivel nacional, que se filtró inicialmente a 341.453 registros solo para Antioquia.

Evaluación y Depuración de Calidad de Datos

Se siguieron estrategias clave para asegurar la calidad de la información:

1. **Filtro Geográfico**: Se eliminaron registros con coordenadas geográficas que estaban fuera del área de Antioquia.
2. **Filtro por Tipo de Propiedad:** Se decidió enfocarse únicamente en apartamentos y casas, excluyendo lotes, locales comerciales y fincas para mantener la homogeneidad del análisis de vivienda residencial.
3. **Extracción por Text Mining (Innovación Clave):** Usando Expresiones Regulares se extrajo información crucial como la *superficie* y la *ubicación* (*barrios, municipio*) que estaba oculta en los campos de texto (*descripción* y *título*). Este paso fue fundamental para recuperar datos valiosos y completar las variables clave.

Tras la depuración intensiva, el *dataset* final para el análisis quedó en 21.192 registros.

---

3. Metodología Analítica

La metodología combina análisis exploratorio, modelado predictivo y diagnóstico interpretativo.

Análisis Exploratorio de Datos (EDA)

a) Entendimiento y Estructura del Data
b) Visualización y Depuración de Outliers
c) Eliminación de Valores Atípicos (Outliers)
    a. Precio: Se utilizó el Rango Intercuartílico (IQR) para identificar y eliminar valores de precio que se encontraban por fuera
    b. **Área (surface_total_final):** Se impuso un límite superior fijo de **2000 m²** para eliminar propiedades con áreas excesivamente grandes, que probablemente no corresponden a viviendas residenciales estándar.

4. Desarrollo de Modelos

Entrenamiento de Modelos y Validación

Una vez depurados los datos, se procedió a la etapa de modelado predictivo, cuyo objetivo es entrenar un algoritmo para que aprenda a predecir el precio.
El *dataset* se dividió en conjuntos de **Entrenamiento (80%)** y **Prueba (20%)** para simular la realidad. El modelo aprende solo con los datos de entrenamiento y luego se evalúa con datos que nunca ha visto (conjunto de prueba), lo que garantiza una evaluación imparcial de su rendimiento

Preprocesamiento de Variables

Antes de entrenar, los datos deben transformarse para que los modelos puedan entenderlos. Esto se hizo mediante un *Pipeline* y un *ColumnTransformer*.

- **Variables Numéricas** (surface_total_final, bedrooms_final, bathrooms_final):
- **Variables Categóricas** (l3_final -Ciudad-, l4_final -Barrio-):
    - OneHotEncoder(handle_unknown='ignore'): Convierte las categorías de texto (ej. "Medellín", "Envigado") en **columnas numéricas binarias (0 o 1)**. Esto es necesario porque los modelos solo trabajan con números.

Los modelos utilizados fueron:

- "Linear Regression": LinearRegression()
- "Decision Tree": DecisionTreeRegressor(random_state=42)
- "Random Forest": RandomForestRegressor(random_state=42)
- "Gradient Boosting": GradientBoostingRegressor(random_state=42)
- "Support Vector Regressor": SVR()

Variables de salida evaluadas: **Precio** (COP)

---

5. Optimización

Se entrenaron cinco modelos de Regresión (predicción de un valor numérico) y se evaluaron usando **Validación Cruzada con (cv=5)**.

- **Validación Cruzada:** Divide el conjunto de entrenamiento en 5 partes, entrena 5 veces y evalúa la parte restante. Esto garantiza que la métrica de rendimiento sea **robusta** y no dependa de una única división de los datos.

Resultados de la Validación Cruzada

Modelo	R2 Promedio
Random Forest	0.8431
Decision Tree	0.7956
Gradient Boosting	0.7817
Linear Regression	0.7016
SVR	-0.0581

Métricas de Evaluación

- **R^2 (Coeficiente de Determinación):** Mide la proporción de la varianza en la variable objetivo (precio) que es predecible a partir de las variables de entrada. **Mayor es mejor**. Un R^2 de **0.84** significa que el modelo explica el 84% de la variabilidad del precio.
- **MAE (Error Absoluto Medio):** El error promedio, en unidades de precio, que el modelo comete. **Menor es mejor**.
- **RMSE (Raíz del Error Cuadrático Medio):** Similar al MAE, pero penaliza más fuertemente los errores grandes. **Menor es mejor**.

Conclusión de CV: El modelo **Random Forest Regressor** demostró ser el **mejor** predictor, con el **R^2** más alto y el MAE/RMSE más bajo.

---

6. Optimización con Hiperparámetros

Se usó **GridSearchCV** para encontrar la combinación ideal de hiperparámetros del modelo Random Forest y mejorar su rendimiento.

- **Hiperparámetros Optimizados:**
  - n_estimators: Número de árboles en el bosque (bosque = conjunto de árboles de decisión).
  - max_depth: Profundidad máxima de cada árbol.

- min_samples_split: Número mínimo de muestras requeridas para dividir un nodo.

- **Resultados de la Búsqueda:**
  - **Mejores Hiperparámetros:** {'model__max_depth': None, 'model__min_samples_split': 2, 'model__n_estimators': 300}.
  - **Mejor Score (R^2): 0.8434** (Una ligera mejora respecto al 0.8431 inicial).

Evaluación Final del Modelo Optimizado

El modelo optimizado se evaluó en el conjunto de **Prueba (test set)**, datos que nunca se utilizaron en el entrenamiento o la validación.

Métrica	Valor en Conjunto de Prueba
R^2	0.8436

Conclusión de la Optimización: El rendimiento del modelo optimizado en datos no vistos (test) es **excelente** (0.8436), confirmando su capacidad para generalizar la predicción de precios.

---

7. Integración

Se implementó una función para usar el modelo optimizado (**best_rf**) y predecir el precio de una vivienda, simulando la entrada de datos de un usuario.

Caso de uso (ejemplo)
Entrada:
`area_m2=192,`
`habitaciones=5,`
`banos=2,`
`municipio=Medellín,`
`barrio=Doce de Octubre.`

Salida: **$312,496,162** (≈ **$1,627,584/m²**).

---

8. Conclusiones

- **Alto Poder Predictivo:** El proyecto logró desarrollar un modelo (Random Forest) con un R^2 de **0.8436**, lo que demuestra una alta capacidad para predecir los precios de las propiedades en Antioquia basándose en las características físicas y de ubicación.

- **Valor de la Preparación de Datos:** El proceso de limpieza, especialmente la **Extracción por Text Mining** para recuperar datos de área y ubicación, fue crucial. Demuestra que el 80% del éxito en Ciencia de Datos radica en tener información completa y de alta calidad.

- **Impacto de la Ubicación:** El uso de las variables categóricas **l3_final** (municipio) y **l4_final** (barrio) dentro del modelo, mediante *One-Hot Encoding*, permitió al algoritmo capturar el valor marginal de la localización, el cual es un factor determinante en el precio inmobiliario.

- **Modelo Robusto:** La **Validación Cruzada** confirmó la solidez del modelo Random Forest frente a otros, asegurando que el rendimiento reportado no es un golpe de suerte, sino una métrica estable y confiable.

El modelo resultante es una **herramienta analítica sólida** para monitorear la oferta y soportar decisiones de valoración inmobiliaria en Antioquia.