Entrega 2 – S5 PROJECT E2 (Semana 5)

Proyecto: Valoración de inmuebles en Bogotá

Integrantes del equipo

- Diego Alejandro Lemus Guzman
- Valeria Iglesias Miranda
- Sergio Andres Perdomo Murcia
- Danilo Suarez Vargas

1. Problema y contexto (Bogotá)

La estimación del valor comercial de vivienda en **Bogotá** suele apoyarse en comparaciones manuales y criterios subjetivos. Esto introduce variabilidad y tiempos de respuesta elevados para compradores, vendedores y entidades financieras. Con el crecimiento de fuentes abiertas confiables a nivel **ciudad**, es factible construir una **solución analítica** que estandarice y acelere la valoración, con métricas de precisión trazables.

Esta situación reduce la transparencia y la comparabilidad de los avalúos e impacta la toma de decisiones de hogares, inmobiliarias, aseguradoras y banca hipotecaria. Con este proyecto buscamos disminuir tiempos y sesgos, entregando estimaciones consistentes y explicables para inmuebles en Bogotá, soportadas en datos abiertos verificables.

2. Pregunta de negocio y alcance

Pregunta de negocio. ¿Cómo desarrollar e implementar un **modelo de predicción** (aprendizaje supervisado) que estime con precisión y rapidez el **valor** de un inmueble en **Bogotá** usando variables físicas, de localización y socioeconómicas?

Alcance (MVP de esta entrega).

- Entradas: *área cubierta*, *número de cuartos*, *tipo de inmueble*, *localidad/barrio*, y otras disponibles en el dataset.
- Salida: *precio estimado* y bandas de error (±MAE).
- Métricas objetivo: **RMSE** y **MAE** en validación; reporte de **R^2**.
- **Supuestos de la entrega:** enfoque en vivienda residencial; valores en COP; uso de datos abiertos consolidados para Bogotá.
- Fuera de alcance: Integración con APIs externas, actualización en tiempo real y despliegue productivo.

Cambios respecto a la Entrega 1

- **Ámbito geográfico:** de *Colombia* → **Bogotá**, por disponibilidad y confiabilidad de datasets abiertos a nivel ciudad.
- **Datos:** se sustituyen fuentes generales por un corte consolidado exclusivo de Bogotá; se priorizan variables robustas y disponibles (precio, área, habitaciones, baños, tipo, barrio/UPZ).
- **Alcance del MVP:** se mantiene el prototipo con predicción y métricas (RMSE/MAE), sin APIs públicas ni actualización en tiempo real; se incorporan experimentos trazables en MLflow en EC2.

3. Conjuntos de datos a emplear (Bogotá)

Archivo base: `inmuebles_bogota.csv` (9,520 registros, 8 columnas).
Breve descripción: datos de anuncios de inmuebles en Bogotá consolidados desde fuentes abiertas; variables principales: `valor` (precio), `área`, `habitaciones`, `baños`, `tipo`, `barrio` y `upz`.
Exploración breve (EDA mínima): a continuación se incluye un resumen con hallazgos rápidos relevantes para el modelado.

3.1 Exploración breve (EDA mínima)

- **Tamaño:** 9,520 registros / 8 columnas.
- **Variables clave:** valor (precio), área, habitaciones, baños, tipo, barrio y UPZ.
- **Hallazgos rápidos:** predominan apartamentos sobre casas; la oferta se concentra en zonas del norte; se recomienda tratar atípicos en área/precio antes del entrenamiento.

Top barrios por número de registros

	registros
Usaquén	1105
Zona Noroccidental	877
Bosa	589
Kennedy	589
Cedritos	554
Barrios Unidos	473
Engativa	462
Suba	443
Santa Barbara	438
Chapinero	332
Fontibón	270
Chico Reservado	225
Teusaquillo	180

Distribución por tipo de inmueble

	registros
Apartamento	7327
Casa	2043
Oficina/Consultorio	60
Local	38
Edificio	22
Bodega	13
Finca	11
Lote	6

	registros
El Batán	133
Puente Aranda	116

5. Modelos desarrollados y evaluación

Se consideran al menos dos familias: (i) **baselines** (Regresión Lineal, Ridge/Lasso) y (ii) **ensambles** (Random Forest, opcional Gradient Boosting). Se comparan por RMSE/MAE/R^2 en validación.

Modelo	Features	Hiperparámetros	RMSE	MAE	R^2
Regresión lineal	[área, cuartos, tipo, localidad/barrio]	log(target)=NO/SÍ			
Ridge/Lasso	idem	α=			
Random Forest	idem + interacciones simples	n_estimators=, max_depth=			
(Opcional) GB/XGB	idem	learning_rate=, n_estimators=			

> **Criterio de selección:** mejor RMSE/MAE, estabilidad y simplicidad del modelo.

6. Experimentos (MLflow en EC2)

- Registrar parámetros, métricas y artefactos (firma del modelo).
- Incluir en `docs/evidencias_mlflow/` pantallazos con **IP pública de la EC2** y **usuario** visibles.
- Conclusiones: hiperparámetros con mayor impacto y run “champion” seleccionado.

7. Prototipo / Tablero

- **Entradas:** área, cuartos, tipo, localidad/barrio.
- **Salida:** precio estimado y bandas ±MAE; importancias de variables (feature importance/SHAP) si aplica.
- **Estado:** se mantiene la maqueta aprobada; capturas en `docs/`.

8. Reporte de trabajo en equipo (resumen)

> **Integrantes:** Diego Alejandro Lemus Guzman; Valeria Iglesias Miranda; Sergio Andres Perdomo Murcia; Danilo Suarez Vargas.

- **Datos/EDA:** preparación de cortes Bogotá, diccionario, limpieza básica.
- **Modelado:** experimentos en MLflow, comparación de modelos y selección.
- **Tablero:** implementación de la maqueta y conexión al modelo.
- **Infra/DevOps:** configuración de entorno (EC2/venv), tracking MLflow.
- **Documentación:** armado de este reporte y evidencias.

9. Observaciones y siguientes pasos

- La ubicación (localidad/barrio) y el tipo de inmueble son determinantes del precio; el área presenta efecto no lineal.
- Siguientes pasos: enriquecer con variables geoespaciales (distancia a vías/zonas de interés), robustecer detección de atípicos y evaluar validación cruzada por localidad.