BIG DATA

PROYECTO FINAL - Precio justo de coches usados

Jonathan Perez Sedova

Tokio.

Planteamiento del problema

- Compradores: ¿estoy pagando de más?
- Vendedores: ¿precio competitivo?
- Mercado: mucha variabilidad por marca, año, km.

Objetivo y alcance

- Objetivo: predecir ŷ y clasificar: infra/justo/sobre (±15%).
- Métrica clave: MAE en test (comparativa de modelos).
- Alcance: dataset CSV (Drive), sin PII.

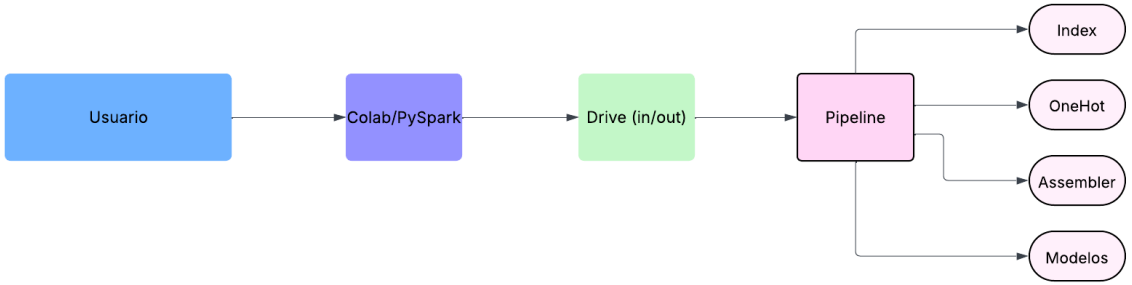
Fuente de datos

- Origen: car_sales_data.csv (Drive).
- Tamaño: ~50k registros, 7 columnas.
- Tipo: numéricas + categóricas (marca/modelo/combustible).

Manufacturer	Model	Engine size	Fuel type	Year of manufacture	Mileage	Price
Ford	Fiesta	1	Petrol	2002	127300	3074
Porsche	718 Cayman	4	Petrol	2016	57850	49704
Ford	Mondeo	1.6	Diesel	2014	39190	24072
Toyota	RAV4	1.8	Hybrid	1988	210814	1705
VW	Polo	1	Petrol	2006	127869	4101
Ford	Focus	1.4	Petrol	2018	33603	29204
Ford	Mondeo	1.8	Diesel	2010	86686	14350

Arquitectura

- Colab (PySpark) → Pipeline ML.
- Ingesta/persistencia en Drive.
- Artefactos: grafico_1_binss/, resultados_modelos_test.csv, predicciones_intervalos_etiquetas.csv.



Limpieza y features

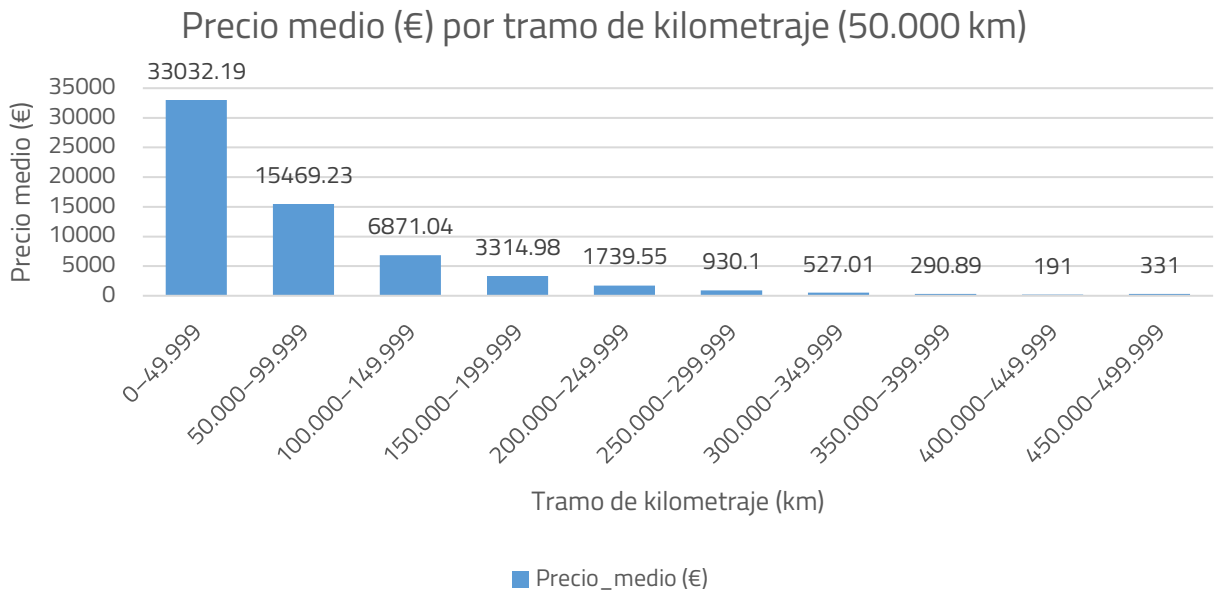
- Renombrado sin espacios (Engine_size, Fuel_type, Year).
- Derivadas: Antigüedad = 2025−Year, km_per_year.
- Nulos: no detectados.

Ingesta y persistencia

- Lectura: spark.read.csv(header=True, inferSchema=True).
- Persistencia EDA: grafico_1_binss (CSV).
- Persistencia modelos/predicciones: resultados_modelos_test.csv, predicciones_intervalos_etiquetas.csv.

EDA - Exploratory Data Analysis

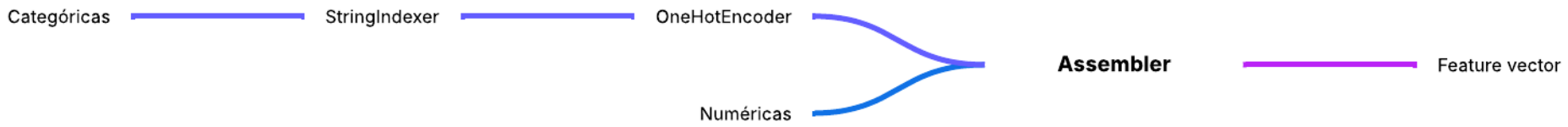
- Precio: min/Q1/mediana/Q3/max/media.
- Precio por tramos de km (P10–P90).
- Sesgo esperado a la derecha.



Tokio.

Pipeline de preprocesado

- Categóricas → StringIndexer + OneHotEncoder.
- Numéricas → Engine_size, Year, Mileage, Antigüedad, km_per_year.
- Ensamblado → feature.



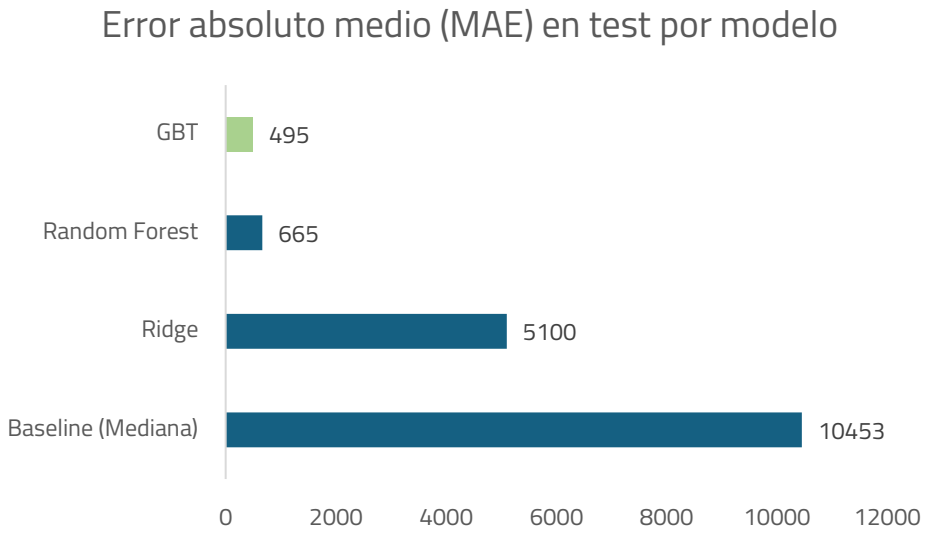
Tokio.

Modelos y criterio

- Baseline (mediana).
- Lineal Ridge.
- Random Forest.
- GBT (elección final por menor MAE).

Resultados

- Baseline: MAE ~10.453 €.
- Ridge: MAE ~5.100 €.
- RF: MAE ~665 €.
- GBT: MAE ~495 €, R^2 ~0,997 (mejor).



Tokio.

Interpretación de ŷ + etiquetas

- Incertidumbre: [ŷ+P10, ŷ+P90] con residuos de RF (train).
- Etiquetas (±15%): Infra ≤ 0,85·ŷ, Sobre ≥ 1,15·ŷ, Justo resto.
- Ejemplo: tarjeta con 1 coche.

BMW Z4 (2022) – 7.052 km, gasolina, motor 3.0

Precio real: 71.296 €
Precio justo (ŷ): 66.246 €
Intervalo [P10–P90]: 65.384 € – 67.141 €
Etiqueta: ✅ Justo

Conclusiones y próximos pasos

- GBT ofrece el mejor MAE y buena explicabilidad vía etiquetas.
- Paquete reproducible (artefactos) y visualizaciones simples.
- Siguientes pasos: variables extra (ubicación/equipamiento), monitorizar drift.

Q&A

- ¿Por qué Spark con 50k? Escalabilidad y requisito académico.
- ¿Year + Antigüedad juntas? Cohortes + depreciación.
- ¿Por qué residuos RF y $\hat{y}$ de GBT? Elección conservadora, alternativa: residuos GBT.