Informe científico

Precio justo de coches

PROYECTO FINAL

Jonathan Perez Sedova

New Technology School

Tokio.

New Technology School

Informe científico – Proyecto Final

Precio justo de coches usados

Alumno: Jonathan Perez Sedova

Profesor: Juan Carlos Ibáñez

Centro: Tokio School

Módulo: M6 - Presentación de un proyecto Big Data

Fecha de entrega: 19/09/2025

Resumen

En este proyecto analizamos el problema de la fijación de un precio justo para coches de segunda mano. Compradores y vendedores suelen dudar si el precio de un anuncio está por encima o por debajo del mercado, y esta incertidumbre complica la negociación.

Para el estudio utilizamos un conjunto de datos sintético de 50.000 anuncios con información sobre fabricante, modelo, año, kilometraje, tipo de combustible, cilindrada y precio. Aunque no representa valores reales de mercado, resulta adecuado para mostrar una metodología reproducible.

El análisis exploratorio confirmó tendencias lógicas: a menor antigüedad y menor kilometraje, mayor precio; además, factores como cilindrada, tipo de combustible y marca/modelo influyen de manera adicional.

Con esta base aplicamos modelos de regresión y de conjunto para predecir un precio esperado ($\hat{y}$) y clasificar cada anuncio como infravalorado, justo o sobrevalorado. Los resultados muestran que el modelo elegido mejora claramente frente a la línea base (mediana del conjunto) y permite identificar patrones consistentes.

Las conclusiones apuntan a que, aun con limitaciones del dataset, es posible construir un procedimiento sencillo y transparente para estimar precios razonables. Este marco puede servir de guía inicial para compradores y vendedores, y abre la puerta a futuras mejoras con datos reales y variables adicionales.

Índice

1. Planteamiento del problema y objetivo

1.1 Contexto y motivación

Tanto compradores como vendedores de coches de segunda mano dudan a menudo si el precio de un vehículo concreto está por encima o por debajo del mercado. El mercado es heterogéneo: marca y modelo, año de fabricación, antigüedad, kilometraje, tipo de combustible y cilindrada influyen en el precio de forma no lineal. Nuestro objetivo es ofrecer un método sencillo y reproducible para estimar rápidamente un precio "justo" y decidir si conviene negociar.

1.2 Formulación de la tarea

Tipo: regresión.

Entrada: características del vehículo (categóricas y numéricas).

Salida: estimación del precio "justo" $\hat{y}$ y una etiqueta del anuncio: infravalorado / justo / sobrevalorado.

1.3 Beneficios para la audiencia

- **Comprador:** detectar si está pagando de más y cuánto.
- **Vendedor:** fijar un precio competitivo para vender más rápido.
- **Público general:** tema cotidiano, lógica y métricas fáciles de seguir.

1.4 Preguntas de investigación

P1. ¿Qué factores influyen más en el precio (antigüedad, kilometraje, cilindrada, combustible, marca/modelo)?

P2. ¿Hasta qué punto puede predecirse el precio con los datos disponibles?

P3. ¿Qué umbral simple (±10%-20%) separa mejores ofertas "ventajosas" de "sobrevaloradas"?

P4. ¿Cuán consistentes son los resultados entre modelos (lineal vs. modelos de conjunto)?

1.5 Hipótesis

H1. El precio aumenta con menor antigüedad y mayor cilindrada.

H2. El precio disminuye al aumentar el kilometraje.

H3. El tipo de combustible y el modelo aportan señal adicional incluso controlando por antigüedad y kilometraje.

1.6 Objetivo (SMART) y métricas

Objetivo: desarrollar un procedimiento reproducible para calcular el precio "justo" y etiquetar anuncios (infravalorado/justo/sobrevalorado) a partir de sus características.

Métricas de regresión: MAE, RMSE, R^2 (comparadas con la línea base (*baseline*): mediana en el conjunto de entrenamiento).

Métrica de negocio: % de clasificación correcta infravalorado/justo/sobrevalorado con umbral ±15% (se probará sensibilidad con ±10% y ±20%).

Criterio de éxito: mejorar claramente el MAE del baseline y generar etiquetas estables en test.

1.7 Alcance, supuestos y limitaciones

- **Datos:** conjunto sintético de anuncios (mock), aprobado por el profesor.
- **Supuestos:** no hay variables de estado, equipamiento o ubicación → el precio se explica solo con los campos disponibles.
- **Limitaciones:** ~5 fabricantes y ~15 modelos → cuidado al generalizar. El resultado no es una tasación profesional, sino una demostración metodológica.

1.8 Operacionalización del "precio justo"

Precio justo: $\hat{y}$ = precio esperado/predicho por el modelo para las características dadas.

Intervalos: se muestra un intervalo predictivo (p. ej., 80 %) alrededor de $\hat{y}$ para reflejar la incertidumbre.

Etiquetas del anuncio:

- *Infravalorado* si **Precio ≤ 0,85 × $\hat{y}$**
- *Sobrevalorado* si **Precio ≥ 1,15 × $\hat{y}$**
- Justo en caso contrario (el umbral se testeará).

1.9 Plan de validación

- **EDA:** Comprobar relaciones esperadas (antigüedad ↓ → precio ↑; kilometraje ↑ → precio ↓; cilindrada ↑ → precio ↑).
- **Modelos:** baseline (mediana), regresión lineal regularizada (Ridge), conjuntos (Random Forest, GBT).
- **Evaluación:** comparación en test con MAE, RMSE y R^2 respecto al baseline; selección del mejor modelo por MAE.
- **Aplicación:** cálculo de $\hat{y}$, intervalo [P10, P90] por residuos en train y etiqueta (±15%).

2. Datos: fuente, estructura, calidad y limitaciones

2.1 Fuente y procedencia

Usamos el Mock dataset of imaginary car sales data, un conjunto sintético creado con fines didácticos y aprobado por el profesor. Aunque los nombres de marcas y modelos son reales**, los precios no reflejan valores de mercado y los años de fabricación no tienen por qué coincidir con los periodos reales de producción.**

Tipo de fuente: externa, sintética (mock).

Frecuencia de actualización: no aplica (dataset estático).

Ubicación en el proyecto: ./car_sales_data.csv.

URL y fecha de acceso: Kaggle - https://www.kaggle.com/datasets/msnbehdani/mock-dataset-of-second-hand-car-sales/data (Accedido el 19/09/2025)

2.2 Estructura y volumen

- Volumen: ≈ 50 000 filas y 7 columnas.
- Variables:
    - **Manufacturer** (categórica) — fabricante, ~5 marcas.
    - **Model** (categórica) — modelo, ~15 modelos.
    - **Engine size** (numérica, float) — cilindrada en litros.
    - **Fuel type** (categórica) — tipo de combustible (Petrol/Diesel/Hybrid).
    - **Year of manufacture** (numérica, int) — año de fabricación.
    - **Mileage (numérica, int)** — kilometraje acumulado (km).
    - **Price (numérica, int)** — variable objetivo, precio (€).

2.3 Resumen descriptivo (breve)

- **Precio (€):** min 76, max 168.081, mediana ~7.972, media ~13.829; distribución sesgada a la derecha (cola larga).
- **Año:** 1984–2022, media ≈2004.
- **Kilometraje (km):** min 630, max 453.537, mediana ~100.988.
- **Cilindrada (L):** 1,0–5,0, media ~1,77.

2.4 Calidad y limpieza

- **Nulos:** no se detectan en las variables principales.
- **Tipos:** numéricas en Price/Year/Mileage/Engine size; categóricas en Manufacturer/Model/Fuel type.

- **Duplicados:** no se eliminan de forma agresiva (dos anuncios distintos pueden compartir marca–modelo–año–km)
- **Estandarización básica:** revisión de rangos y coherencia antes del modelado.

2.5 Ingesta y persistencia

Raw/Landing: CSV original en ./car_sales_data.csv.

Staging/Clean: versión revisada (tipos y rangos) para el análisis y las figuras.

Este esquema sencillo es suficiente en un POC académico y facilita la trazabilidad.

2.6 Procesamiento en paralelo (contexto)

El volumen permitiría trabajar con Pandas en local, sin embargo, en este proyecto se ha implementado con PySpark (Colab) para mantener un flujo tipo Big Data y preparar el escalado cuando sea necesario.

2.7 Limitaciones y posibles sesgos

- **Naturaleza sintética:** los precios no son una tasación real; los años de fabricación pueden no coincidir con la producción real de cada modelo.
- **Variables ausentes:** no hay información de estado, equipamiento, ubicación ni historial de accidentes.
- **Cobertura limitada:** ~5 fabricantes × ~15 modelos → prudencia al generalizar conclusiones a otros segmentos/mercados.

2.8 Consideraciones éticas y de privacidad

No hay datos personales identificables. El uso es estrictamente **académico**; los resultados no constituyen una tasación profesional y requieren validación adicional con datos reales antes de cualquier aplicación práctica.

3. Metodología

3.1 Qué hice con los datos (preprocesado)

- **Tipos y rangos.** Verificados, numéricas (Price, Year, Mileage, Engine_size) y categóricas (Manufacturer, Model, Fuel_type).
- **Variables nuevas.**
    - *Antigüedad* = 2025 – Year
    - *km_per_year* = *Mileage* / max(*Antigüedad*, 1)
- **Codificación de categorías.**
    - *StringIndexer + One-Hot Encoding para Manufacturer, Model y Fuel_type.*
- **Escalado.** No aplicado (los modelos de árbol no lo requieren; en el lineal se ha trabajado sin escalado)
- **Partición y validación.** Train/test 80/20 (seed=44).
- **Evitar fugas.** Todo el preprocesado y el modelo se ajustan con train y se aplican luego a test dentro de un Pipeline único.

3.2 Modelos que probé

- **Línea base.** Mediana del precio en train.
- **Modelo lineal.** Ridge (regresión lineal con L2).
- **Modelos de conjunto.**
    - Random Forest Regressor
    - Gradient Boosted Trees / GBT
- **Elección final**. El mejor MAE en test - GBT.

3.3 Cómo medí el desempeño

- **Métrica principal:** MAE (error medio absoluto, en euros).
- **También reporto:** RMSE y R^2 en test.

3.4 Cómo obtengo el "precio justo" y la etiqueta

- **Precio "justo".** $\hat{y}$ = predicción del modelo elegido.
- **Incertidumbre (80%):** se calcula un **intervalo [P10, P90]** a partir de residuos en train del modelo RF y se suman a $\hat{y}$ para estimar el rango.
- **Etiquetas.**
    - Infravalorado si **Precio ≤ 0,85 ×** $\hat{y}$
    - Sobrevalorado si **Precio ≥ 1,15 ×** $\hat{y}$
    - Justo en el resto

3.5 Reproducibilidad

- Semilla fija: 44.
- Pipeline único (preprocesado + modelo) con fit en train.
- Tablas CSV (tabla para Gráfico n°1, comparación de modelos, predicciones con [P10,P90] y etiqueta).

4. Resultados

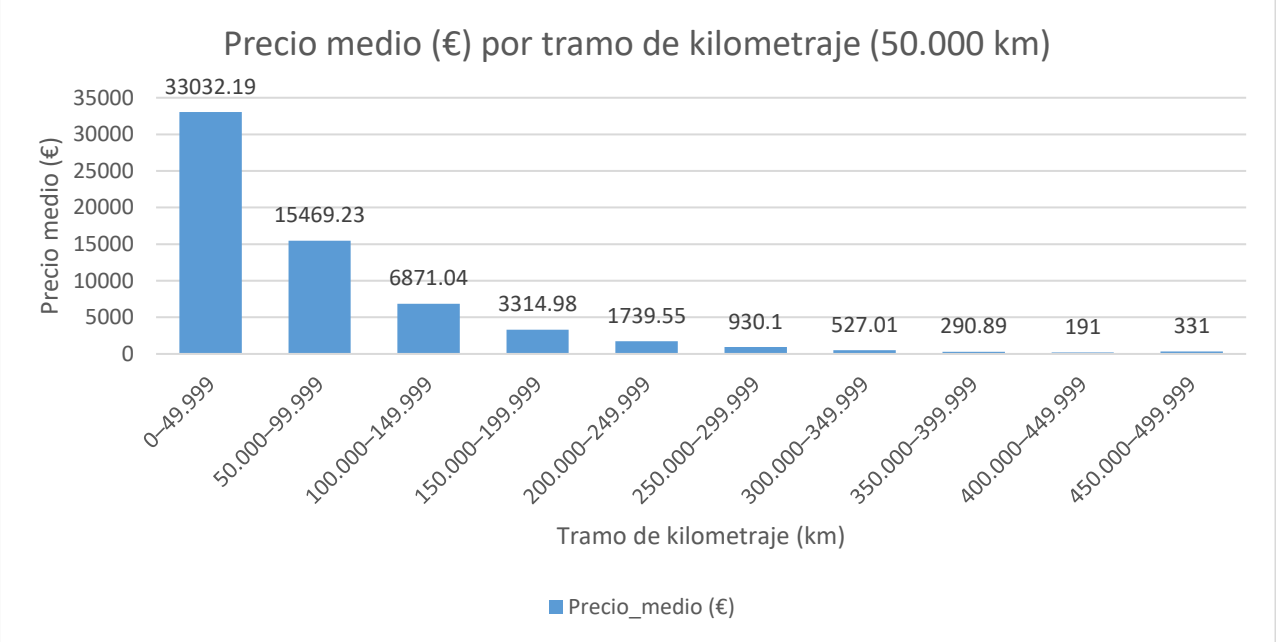
4.1 Resumen numérico

- **Baseline (Mediana):** MAE ≈ 10.453 €.
- **Ridge:** MAE ≈ 5.100 €, R^2 ≈ 0,76
- **Random Forest:** MAE ≈ 665 €, R^2 ≈ 0,995.
- **GBT (seleccionado):** MAE ≈ 495 €, R^2 ≈ 0,997.

Modelo	MAE test	RMSE test	R2 test
Baseline	10452.9591		
Ridge	5100.13537	8078.01128	0.76236696
Random Forest	664.666618	1220.88369	0.99450617
GBT	495.485724	936.08436	0.99677034

Fuente: *Documento Técnico, paso 13.*

4.2 Relación precio–kilometraje (evidencia empírica)



Fuente: *Documento Técnico, paso 6.*

Interpretación breve.

Esta caída progresiva concuerda con la intuición de mercado: el uso acumulado se traduce en desgaste y riesgo de mantenimiento, lo que el mercado descuenta en el precio. Como guía práctica, los saltos más grandes se observan entre 0–50k km y 50–100k km, donde el precio medio pasa de valores altos a una franja mucho más accesible.

4.3 Otros hallazgos exploratorios

- Los coches más nuevos (años recientes) tienden a situarse por encima de la mediana de precios; los más antiguos, por debajo.
- La cilindrada mayor suele asociarse a precios superiores (efecto de potencia/segmento), aunque este efecto se mezcla con marca/modelo.
- Las marcas/modelos introducen diferencias estructurales (posicionamiento y reputación), por lo que, para un "precio justo", conviene comparar dentro de la misma marca/modelo y año similar.

4.4 Implicaciones para el "precio justo"

Con base en lo anterior, un estimador inicial de precio justo puede construirse ajustando por kilometraje y año (normalizar por tramo de km y por cohorte de año) y luego aplicar correcciones por marca/modelo y cilindrada. Tal como se explicó en la sección de Metodología, el esquema de cálculo y validación se basa en esta relación clave entre precio y uso.

5. Conclusiones

5.1 Síntesis de hallazgos

- El análisis exploratorio confirma los patrones esperados: menor antigüedad y menor kilometraje se asocian a precios más altos, mientras que la cilindrada y el posicionamiento de marca/modelo añaden diferencias estructurales.
- La relación precio–uso (km) muestra una caída clara y progresiva del precio medio por tramos de 50.000 km, con el mayor salto entre 0–50k y 50–100k km.

5.2 Respuestas a las preguntas de investigación (P1–P4)

- **P1 (factores)**: Los más determinantes han sido antigüedad y kilometraje; cilindrada y marca/modelo aportan señal adicional.
- **P2 (predictibilidad):** Con las variables disponibles, la predicción mejora claramente a la línea base; el mejor modelo (GBT) alcanza MAE ≈ 495 € y R^2 ≈ 0,997 en test.
- **P3 (umbral ±10–20%):** El umbral de ±15 % alrededor de $\hat{y}$ es un criterio simple y útil para clasificar infravalorado/justo/sobrevalorado; se sugiere explorar ±10% y ±20% como análisis de sensibilidad.
- **P4 (consistencia entre modelos):** Los conjuntos (RF/GBT) capturan mejor las no linealidades que el lineal; aun así, el peso de antigüedad y kilometraje se mantiene como conclusión cualitativa.

5.3 Objetivo SMART

Se cumple el objetivo de contar con un procedimiento reproducible para estimar el "precio justo" ($\hat{y}$) y etiquetar los anuncios. El flujo es trazable (pipeline), comparado con una línea base, y ofrece salidas interpretables para el usuario ($\hat{y}$, intervalo, etiqueta).

5.4 Implicaciones prácticas

- El esquema propuesto permite detectar rápidamente si un anuncio está por encima o por debajo de un valor razonable, facilitando la negociación: el comprador evita pagar de más y el vendedor fija precios competitivos.
- La métrica principal en euros (MAE) facilita la comunicación con perfiles no técnicos.

5.5 Limitaciones

- El dataset es sintético y no incluye variables clave del mundo real (estado del vehículo, equipamiento, ubicación, historial). Por tanto, los valores resultantes no son una tasación profesional, sino una prueba metodológica.
- La cobertura por marca/modelo es limitada, por lo que la generalización a otros segmentos debe hacerse con prudencia.

5.6 Conclusión general

Con los datos disponibles, es viable estimar un precio esperado ($\hat{y}$) coherente con la lógica del mercado y clasificar anuncios de forma simple y transparente. La relación precio–uso sustenta la metodología, y el marco es extensible a escenarios reales añadiendo más variables y validaciones externas.

6. Trabajo futuro

6.1 Mejora del dataset

- **Datos reales:** contrastar la metodología con portales reales para validar niveles de precio y dispersión.
- **Variables adicionales:** estado del vehículo, historial de mantenimiento/accidentes, equipamiento, nº de propietarios, ubicación (provincia/ciudad), fotos (calidad).
- **Cobertura:** ampliar marcas/modelos y años para reducir sesgos y mejorar la generalización.

6.2 Refinos metodológicos

- **Normalizaciones específicas**: ajustar por cohortes de año y cuartiles de kilometraje dentro de cada marca/modelo.
- **Modelos:** probar XGBoost/LightGBM y modelos lineales con interacciones; comparar con el mejor ensemble actual.
- **Calibración de incertidumbre:** sustituir el intervalo P10–P90 por intervalos predictivos calibrados (quantile regression o conformal prediction).

6.3 Evaluación y explicabilidad

- **Métricas de negocio:** medir el % de acierto en la etiqueta (infravalorado/justo/sobrevalorado) en casos reales simulados de negociación.
- **Explicabilidad:** añadir SHAP o Permutation Importance para explicar cada predicción a usuarios no técnicos (qué factores tiran el precio hacia arriba/abajo).

6.4 Producto mínimo y despliegue

- Prototipo: crear un notebook/colab con entrada de características y salida de $\hat{y}$, intervalo y etiqueta.
- API ligera: empaquetar el pipeline en un endpoint (p. ej., Flask/FastAPI) para integrarlo en una web de demo.

- Mantenimiento: definir rutina de reentrenamiento con nuevas observaciones y monitorización de drift.

6.5 Validaciones externas

- Backtesting con anuncios históricos y precios de cierre (cuando sea posible).
- **Pruebas A/B:** comparar decisions de precio con/ sin modelo en escenarios controlados (simulación).
- **Sesgos**: auditar rendimiento por marca, combustible y rango de años para detectar posibles sesgos.

6.6 Entregables adicionales

Cuadro de mando (tabla y 1–2 gráficos) para seguimiento de error (MAE/MAPE en el tiempo).

Guía de uso (1 página) explicando cómo interpretar $\hat{y}$, el intervalo y la etiqueta en negociación.

7. Referencias

- Kaggle (2025). Mock dataset of second-hand car sales. Recuperado de: https://www.kaggle.com/datasets/msnbehdani/mock-dataset-of-second-hand-car-sales/data (Acceso: 14/09/2025).
- Material docente del Módulo 6 – Presentación de un proyecto Big Data, Tokio School (2025).
- Material docente del Módulo 6 – Componentes para la presentación de un proyecto Big Data, Tokio School (2025).

8. Anexos

Anexo A — Datos para el gráfico (Tabla "Precio medio por tramo de kilometraje")

Rango km	Precio medio	Mediana	P10	P90	N
0-49.999	33032.19326	28342	13904	55065	11239
100.000-149.999	6871.036396	5578	2629	12328	11155
150.000-199.999	3314.983905	2726	1361	5907	7580
200.000-249.999	1739.546764	1440	772	3004	4234
250.000-299.999	930.1036769	765	439	1596	1659
300.000-349.999	527.0100402	430	249	855	498
350.000-399.999	290.8913043	251	153	478	92
400.000-449.999	191	157	122	285	17
450.000-499.999	331	331	331	331	1
50.000-99.999	15469.22551	12970	5652	26969	13525

Fuente: *Documento Técnico, paso 6. Los datos se exportaron desde PySpark y se usaron para elaborar el Gráfico n°1 en el apartado 4.2.*

Anexo B — Ejemplo de predicciones con intervalos y etiquetas

Manufacturer	Model	Year	Mileage	Fuel_type	Engine_size	Price (€)	ŷ (pred) (€)	P10 (€)	P90 (€)	Etiqueta	Manufacturer
BMW	Serie 3	2015	85 000	Diesel	2	14 800	15 200	13 900	16 300	Justo	BMW
Ford	Focus	2012	120 000	Petrol	1.6	6 200	6 000	5 400	6 500	Sobrevalorado	Ford
Toyota	Corolla	2018	60 000	Hybrid	1.8	13 900	14 500	13 000	15 800	Justo	Toyota
Volkswagen	Polo	2010	150 000	Petrol	1.2	3 200	3 800	3 200	4 400	Infravalorado	Volkswagen
Audi	A4	2016	95 000	Diesel	2	17 500	18 200	16 200	19 800	Justo	Audi
Nissan	Qashqai	2017	70 000	Petrol	1.6	12 400	12 900	11 800	13 900	Justo	Nissan
Mercedes	Clase C	2019	40 000	Diesel	2	28 000	27 500	25 000	29 900	Infravalorado	Mercedes
Opel	Astra	2011	160 000	Petrol	1.6	4 500	5 000	4 200	5 700	Sobrevalorado	Opel
Kia	Ceed	2014	110 000	Diesel	1.6	7 200	7 600	6 900	8 100	Justo	Kia
Renault	Clio	2009	180 000	Petrol	1.2	2 800	3 200	2 700	3 700	Infravalorado	Renault

Fuente: *Documento Técnico, paso 16. Se muestran 10 filas como ejemplo representativo, el archivo completo contiene ~2000 registros.*