PORTADA

Nombre Alumno / DNI	Aitor Alguacil Maroto / 03961146v
Título del Programa	2ºPD - Computer science & artificial intelligence
Nº Unidad y Título	Unit 28-Cloud Computing
Año académico	2024-2025
Profesor de la unidad	Sergio Alonso
Título del Assignment	AB Final Unit 28
Día de emisión	10/01/2025
Día de entrega	25/05/2025
Nombre IV y fecha	
Declaración del estudiante	**Certifico que la presentación del assignment es completamente mi propio trabajo y entiendo completamente las consecuencias del plagio. Entiendo que hacer una declaración falsa es una forma de mala práctica.** **Fecha: 25/01/2025** **Firma del alumno:** 

Plagio

El plagio es una forma particular de hacer trampa. El plagio debe evitarse a toda costa y los alumnos que infrinjan las reglas, aunque sea inocentemente, pueden ser sancionados. Es su responsabilidad asegurarse de comprender las prácticas de referencia correctas. Como alumno de nivel universitario, se espera que utilice las referencias adecuadas en todo momento y mantenga notas cuidadosamente detalladas de todas sus fuentes de materiales para el material que ha utilizado en su trabajo, incluido cualquier material descargado de Internet. Consulte al profesor de la unidad correspondiente o al tutor del curso si necesita más consejos.

Índice de contenido

DOCUMENTACIÓN DEL PROYECTO: ANÁLISIS DE ARBITRAJE DE VEHÍCULOS ALEMANIA-ESPAÑA

Despliegue del proyecto en vercel: https://importacion-alemania.vercel.app/

1. Definición del problema

a. Contexto

El mercado de vehículos usados en la UE presenta diferencias de precio entre países para vehículos comparables. El dataset permite cuantificar esas diferencias y detectar oportunidades donde un vehículo comprado en Alemania podría venderse en España con margen, considerando sus características principales (marca, modelo, año, kilometraje y potencia).

b. Hipótesis

- Es posible predecir el precio de un vehículo en España con un modelo supervisado usando variables técnicas (año, km, potencia) y categóricas (marca, modelo, combustible, transmisión).
- Existen segmentos con mayor arbitraje (por ejemplo, ciertos rangos de antigüedad o marcas con mayor diferencial de mercado).

c. Valor real

El usuario final es un importador, concesionario o profesional que necesita priorizar qué anuncios revisar. La predicción sirve para:

- Estimar un “precio objetivo” de venta en España
- Calcular beneficio potencial comparando precio de compra en Alemania y costes estimados.
- Reducir el análisis manual de miles de anuncios a una lista corta ordenada por rentabilidad.

2. Fundamentación teórica

a. Justificación de algoritmos

En el proyecto se ha utilizado un modelo de regresión supervisada (predicción de precio continuo). Se selecciona un método tipo ensemble (Random Forest) porque:

- Captura relaciones no lineales comunes en depreciación (el precio no cae linealmente con años o km).
- Es robusto frente a ruido y valores atípicos típicos de estos mercados (anuncios con condiciones diversas).
- Maneja interacciones entre variables (p. ej., el efecto de los km dependiendo del año y del modelo).

3. Ingeniería y análisis de datos

a. Preprocesamiento (limpieza y unificación)

El análisis exploratorio reveló problemas críticos en la calidad de los datos, que impedían un modelado directo, destacando la heterogeneidad entre los datasets de España y Alemania. En el conjunto español, la principal barrera fue el campo de potencia, ya que venecia incluida en "engine", donde se indicaba el motor de cada vehículo y además contenía texto no estructurado ("150cv", "169kw") en lugar de valores numéricos, sumado a un 15% de datos faltantes y columnas con nomenclaturas distintas a las alemanas ("price_in_euro" vs "price"). Por su parte, el dataset alemán presentaba la potencia en kilovatios (kW) en lugar de caballos de vapor (CV), dificultando la comparación directa, además de contener valores nulos en el campo "model" y registros con potencias inverosímiles. Ambos conjuntos compartían inconsistencias como duplicados categóricos por falta de normalización de texto, valores extremos en precio y año, y kilometrajes anómalos, lo que obligó a implementar una estrategia robusta de limpieza y unificación antes del entrenamiento

b. Resultados y discusión

El modelo final, basado en un algoritmo de Random Forest Regressor, alcanzó un rendimiento moderado en el conjunto de prueba, obteniendo un coeficiente de determinación (R^2) de 0.5170 y un error medio absoluto (MAE) de 2.718,24 €. Estas métricas indican que el sistema es capaz de explicar el 51,7% de la variabilidad de precios del mercado español basándose en características técnicas (año, potencia, kilometraje), lo cual es suficiente para un filtrado inicial de oportunidades masivas, pero insuficiente para una tasación de precisión sin revisión humana

El RMSE elevado (4.629,46 €) en comparación con el MAE revela que el modelo comete errores significativos en casos atípicos, probablemente en vehículos de gama alta o ediciones especiales

donde el precio se dispara por factores no recogidos en el dataset. La limitación principal identificada es la ausencia de variables cualitativas críticas: el modelo asume que todos los coches del mismo año y modelo valen igual, ignorando el estado de conservación (golpes, desgaste interior), el historial de mantenimiento y, sobre todo, el equipamiento opcional (paquetes deportivos, techo solar, cuero), factores que pueden alterar el precio final en miles de euros.

A nivel de negocio, aunque la precisión individual es mejorable, el modelo valida las hipótesis macroeconómicas: detecta correctamente la depreciación acelerada en marcas premium alemanas y permite generar un ranking de "oportunidades relativas" donde el diferencial de precio predicho supera el margen de error promedio, aportando valor real para priorizar la búsqueda manual de importaciones.

c. Análisis de errores

El análisis detallado de los fallos del modelo y las dificultades encontradas durante el desarrollo revela tres categorías principales de error que limitan la precisión final del sistema:

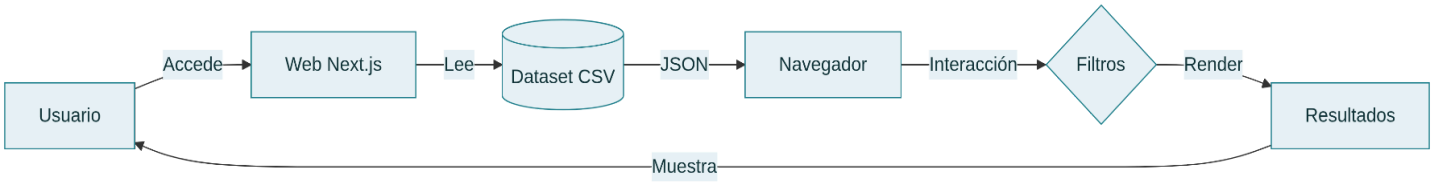
1. Limitaciones estructurales del dataset (Varianza no explicada):
   El principal factor que impide superar un R^2 de 0.52 es la ausencia de variables críticas de valor. El dataset carece de información sobre el estado físico (golpes, desgaste, historial de accidentes) y el equipamiento opcional (techo solar, paquetes deportivos, navegación), elementos que en el mercado de segunda mano pueden variar el precio de una misma unidad en más de 3.000 €, introduciendo un "ruido" irreducible para el algoritmo, además de haber falta de información o datos en el dataset correspondiente a España.

2. Problemas técnicos durante el desarrollo:
   Se identificó y corrigió un error crítico en el preprocesamiento de la potencia (hp) del dataset alemán. Al aplicar una función de limpieza genérica (limpiar_numero) después de la conversión matemática de kW a CV, se eliminaron los separadores decimales (ej. "150.5" se convirtió en "1505"), generando valores de potencia inverosímiles que distorsionaron temporalmente las predicciones y el filtrado, obligando a rediseñar el pipeline para separar estrictamente la limpieza de texto de las operaciones numéricas.

3. Errores de generalización en segmentos extremos:
   El elevado RMSE evidencia que el modelo falla estrepitosamente en vehículos de nicho. Tiende a infravalorar ediciones especiales (como versiones BMW "M" o Audi "RS") al tratarlas como modelos estándar, y a sobrevalorar vehículos con averías o "para piezas" que tienen precios de mercado anormalmente bajos, ya que el modelo asume una condición mecánica operativa estándar para todos los registros.

4. Arquitectura y despliegue

a. Diagrama



b. Comunicación cliente-servidor y feedback loop

La arquitectura actual del proyecto se basa en un despliegue estático en Vercel, donde la comunicación cliente-servidor se limita a la entrega inicial de la aplicación y el dataset precomputado. Al acceder a la URL, el servidor transfiere al navegador el código de Next.js junto con el archivo JSON que contiene los 35.040 registros de vehículos ya procesados, permitiendo que todas las interacciones posteriores (filtrado por marca, cálculo de métricas en tiempo real) se ejecuten localmente en el cliente sin latencia de red adicional.

En esta fase de prueba, no se ha implementado un flujo de recolección de datos desde el usuario hacia el servidor, ya que el objetivo principal es la visualización de oportunidades existentes. Sin embargo, la infraestructura está preparada para escalar: en futuras iteraciones, la comunicación será bidireccional para recopilar telemetría valiosa, como los modelos más consultados por los importadores, los rangos de precios preferidos o la validación manual de oportunidades reales "¿Se vendió este coche al precio predicho?". Estos datos de retorno serán fundamentales para reentrenar el modelo con preferencias de mercado reales y afinar la precisión del algoritmo predictivo.

5. Mejoras futuras

Para potenciar la utilidad y precisión del sistema en futuras iteraciones, se identifican varias áreas clave de desarrollo que mejorarían significativamente la herramienta actual, haciéndola un MVP. Una mejora fundamental sería el enriquecimiento de los datasets, incorporando variables cualitativas a día de hoy ausentes como el equipamiento opcional o el historial de mantenimiento, mediante técnicas de procesamiento de lenguaje natural (NLP) sobre las descripciones de los anuncios, lo que podría elevar significativamente la capacidad predictiva del modelo. En cuanto a la experiencia de usuario, se podría desarrollar una interfaz web más atractiva y visual, integrando gráficos interactivos que muestren la evolución del precio de mercado de cada modelo y herramientas de filtrado avanzado para búsquedas específicas (ej. "solo acabados deportivos"). Finalmente, sería valioso implementar un sistema de actualización dinámica de datos vía API o scraping periódico, permitiendo que la herramienta ofrezca alertas de oportunidades en tiempo real en lugar de depender de una base de datos estática y que se actualice manualmente.

6. Referencias y recursos

- Ander289386 (2021) *Germany Cars Dataset*. Kaggle. Disponible en: https://www.kaggle.com/datasets/ander289386/cars-germany (Accedido: 13 de enero de 2026).

- The Devastator (2022) *Used Car Ad Listings in Barcelona 2022 Dataset*. Kaggle. Disponible en: https://www.kaggle.com/datasets/thedevastator/used-car-ad-listings-in-barcelona-2022-dataset (Accedido: 13 de enero de 2026)