DATA SCIENCE III

ENTREGA NUMERO 1

Alumno: Ciorciari Conrado

Comisión: 90415

Índice

Motivación y audiencia:

El análisis se enfoca en el mercado de autos usados y busca identificar los factores que influyen en el precio y el estado del vehículo.

La motivación principal surge de la necesidad de comprender de qué manera, variables como el tipo de combustible, el año de registro, la potencia del motor, el kilometraje y el tipo de vehículo afectan el valor de mercado.

Este tipo de análisis resulta útil para:

- **Compradores y vendedores de vehículos usados**, que podrán contar con más información al momento de evaluar qué atributos impactan en el precio.
- **Concesionarias o plataformas online**, que pueden usar estos insights para optimizar precios, detectar anomalías y mejorar estrategias de venta.
- **Analistas de datos o investigadores**, interesados en construir modelos predictivos que estimen precios o detecten autos con daños no reparados.

En rasgos generales la evaluación es útil para quienes busquen optimizar sus estrategias de fijación de precios o detectar vehículos sobrevaluados o subvaluados.

El proyecto utiliza técnicas de análisis exploratorio (EDA) y Machine Learning para generar una visión integral del comportamiento del mercado automotor.

Resumen de la metadata

El dataset analizado contiene información de vehículos publicados en línea, con aproximadamente 370000 registros y 21 columnas.

Los datos combinan variables numéricas, categóricas y temporales, lo que permite abordar tanto análisis estadísticos como modelos de predicción.

Estructura general:

- **Numéricas:** `price`, `yearOfRegistration`, `powerPS`, `kilometer`, `monthOfRegistration`, estos valores continuos son usados para estimar precios o desgaste.
- **Categóricas:** `fuelType`, `vehicleType`, `gearbox`, `brand`, `model`, `seller`, estas clasifican el tipo o características del vehículo.
- **Fecha/Texto:** `dateCreated`, `dateCrawled`, `lastSeen`, estos son datos temporales y de rastreo.
- **Target:** `target` es de variable binaria: 0 = sin daño, 1 = con daño.

El dataset fue limpiado y procesado, eliminando duplicados y valores nulos para facilitar tanto su análisis como el entrenamiento de modelos.

Hipótesis

1. ¿Qué tipo de vehículo predomina en el mercado de autos usados?

Se busca identificar qué carrocerías son más comunes y cómo se distribuyen las preferencias de los usuarios.

2. ¿Existe una estacionalidad en el registro de vehículos según el mes?

Buscamos analizar si hay meses con mayor cantidad de registros o ventas, lo que puede reflejar patrones de renovación o compra de autos durante el año.

3. ¿Cómo varía la proporción de autos con y sin daño según el tipo de combustible?

Exploramos si ciertos combustibles (nafta, diésel, gas) están más asociados con autos dañados o en buen estado.

4. ¿Cuál es la distribución geográfica aproximada de los registros según el código postal?

Buscamos observar la densidad de publicaciones según zonas, para detectar concentraciones regionales en la venta de autos.

5. ¿Qué relación existe entre el tipo de combustible y el precio?

Evaluamos si los vehículos con ciertos combustibles tienen precios medios más elevados.

6. ¿Cuál es la dispersión de precios en el mercado de autos usados?

Buscamos determinar si existen valores atípicos o una gran diferencia entre autos económicos y de alta gama.

7. ¿Qué variables presentan mayor relación con el precio y con el estado del vehículo?

Identificamos correlaciones entre `price`, `yearOfRegistration`, `powerPS`, `kilometer` y `target`.

8. ¿Qué variables son más relevantes para predecir si un vehículo tiene daños?

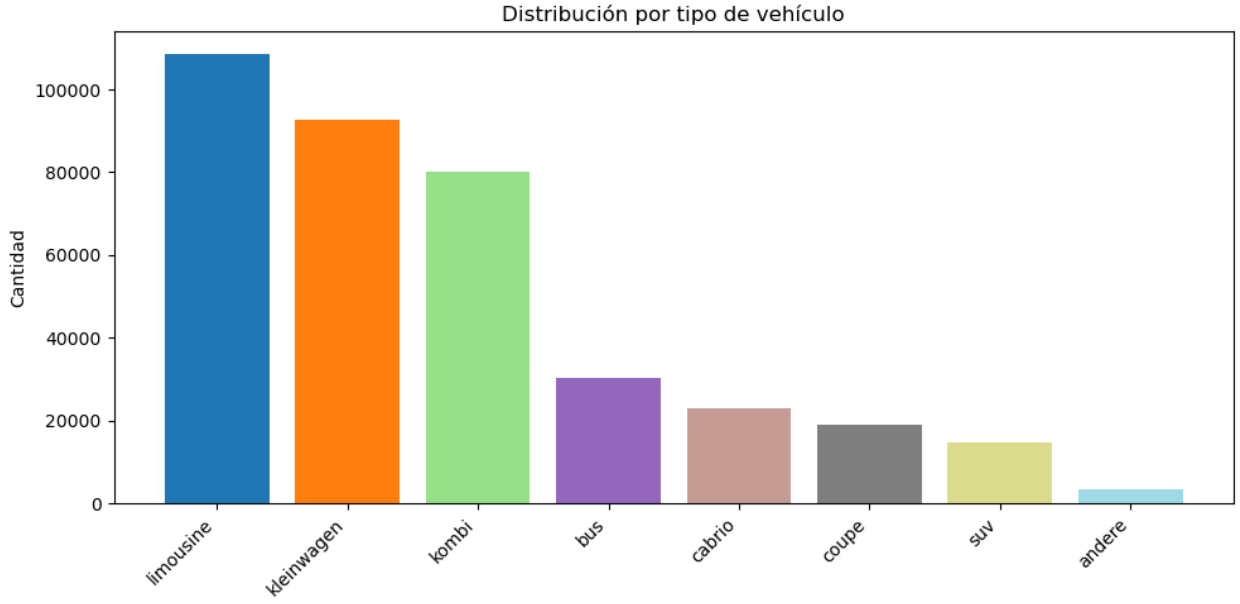
Evaluamos qué características (por ejemplo, `price`, `powerPS`, `kilometer`, `yearOfRegistration`) aportan más información al modelo predictivo.

9. ¿Qué tan bien clasifica el modelo Random Forest los vehículos dañados y no dañados?

Analizamos el rendimiento del modelo: precisión, recall y balance entre clases.

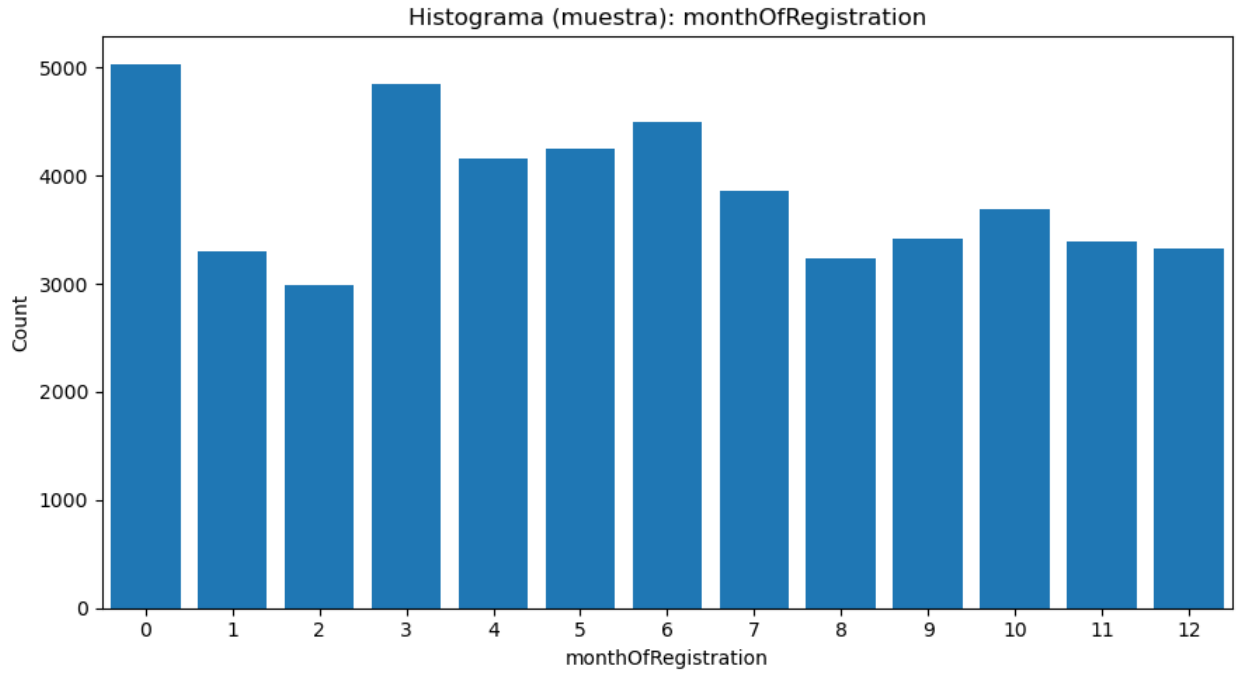
Visualizaciones

1-*Gráfico: Distribución por tipo de vehículo*



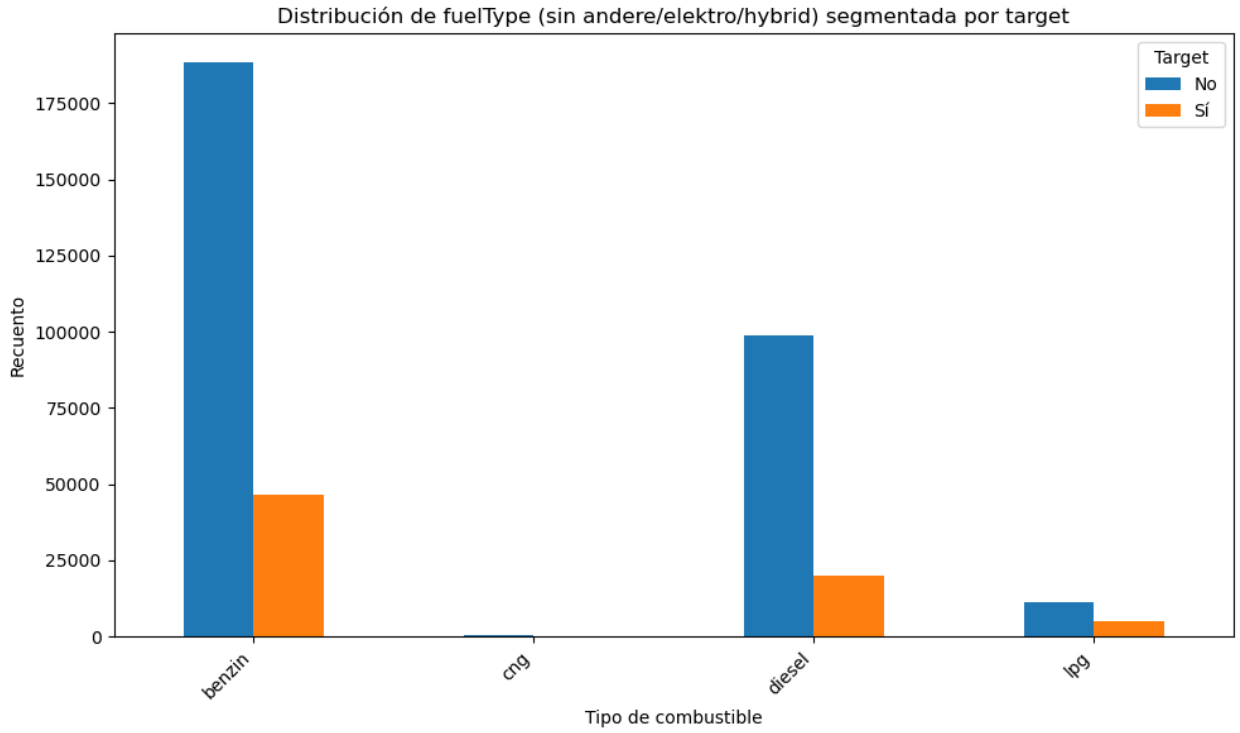
Los datos muestran que los vehículos tipo “limousine”, seguidos por “kleinwagen” y “kombi” son los más comunes en el mercado.
Esto sugiere que la mayor parte de las publicaciones corresponden a autos de uso urbano o familiar, lo que concuerda con la demanda de transporte personal en ciudades.

2-*Gráfico: Histograma de monthOfRegistration*



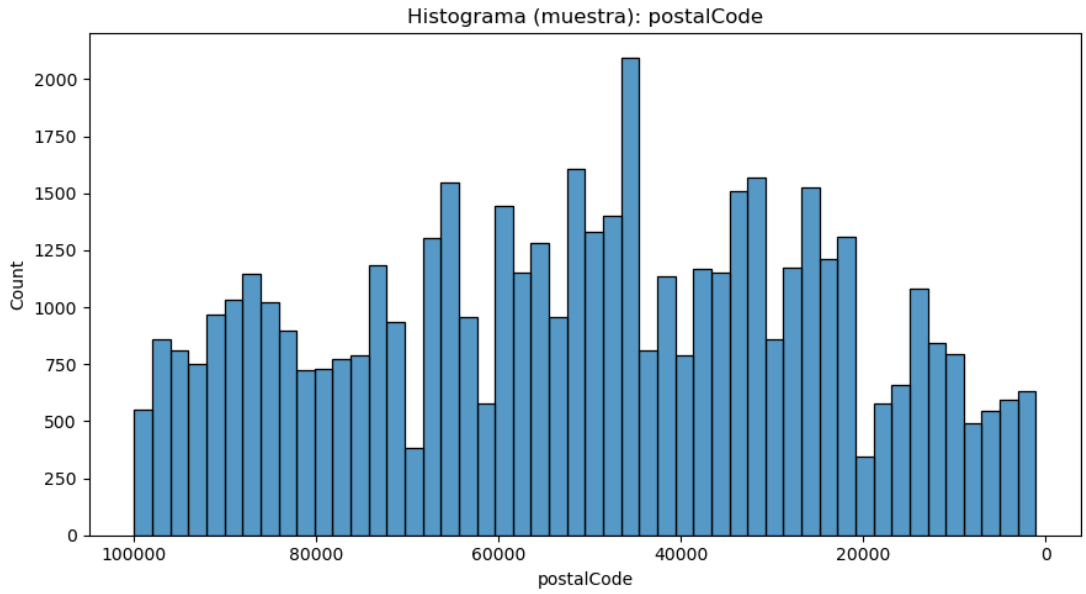
El histograma muestra una distribución bastante uniforme durante el año, aunque con picos en los meses 0, 3 y 6, lo que puede deberse a períodos de renovación o lanzamientos de nuevos modelos.

3- ***Gráfico: Distribución de fuelType segmentada por target***



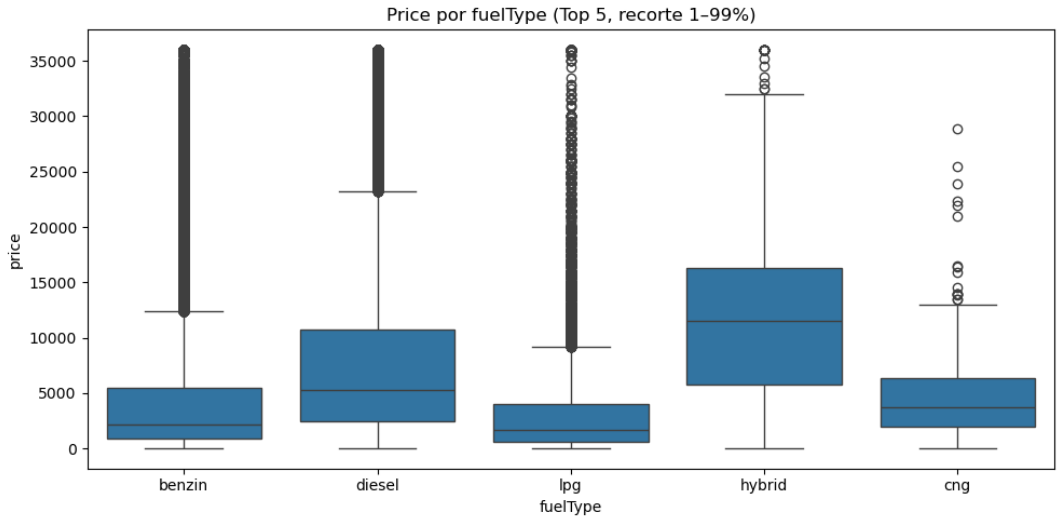
Los vehículos a nafta (benzin) son los más comunes tanto en autos dañados como no dañados.
Los diésel presentan una menor proporción de daño, mientras que los de gas (LPG) y CNG tienen menor representación general.
Esto podría indicar que los autos a nafta son los más populares, pero también los que más tiempo permanecen en el mercado, aumentando su exposición a fallas o reparaciones.

4-*Gráfico: Histograma de postalCode*



El histograma revela una dispersión amplia, sin concentraciones muy marcadas, aunque con mayor densidad en rangos intermedios (30 000–60 000).
Esto sugiere que las publicaciones provienen de distintas regiones, lo que hace del dataset una muestra representativa del mercado alemán en general.

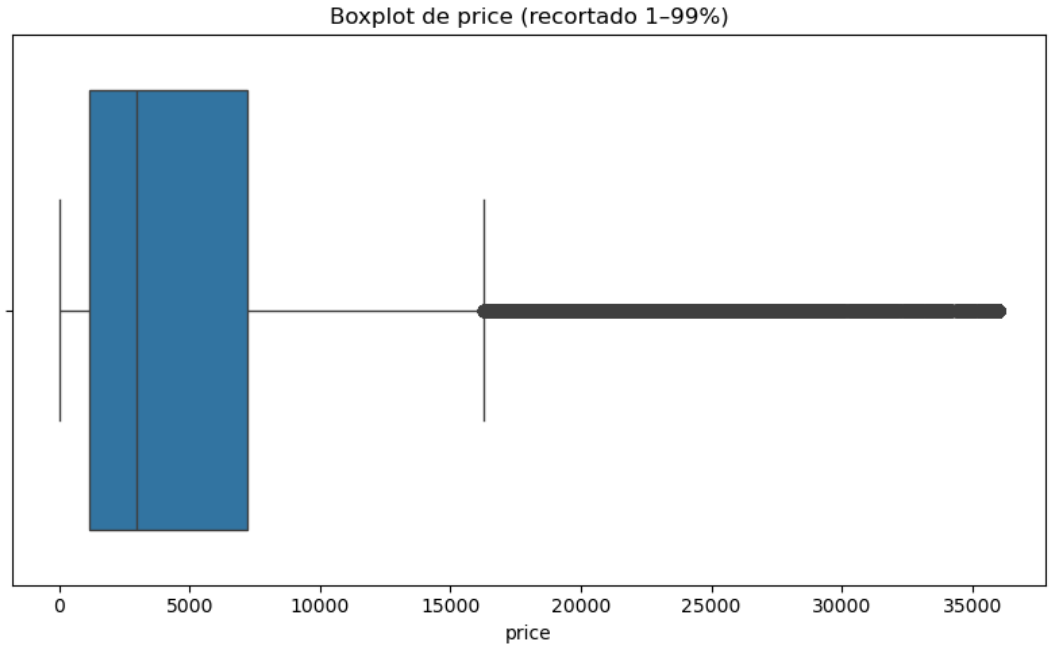
5- *Gráfico: Boxplot de price por fuelType*



El boxplot muestra que los vehículos diésel e híbridos tienden a tener precios medianos más altos, mientras que los de nafta (benzin) y gas (LPG) presentan mayor variabilidad y precios más bajos.

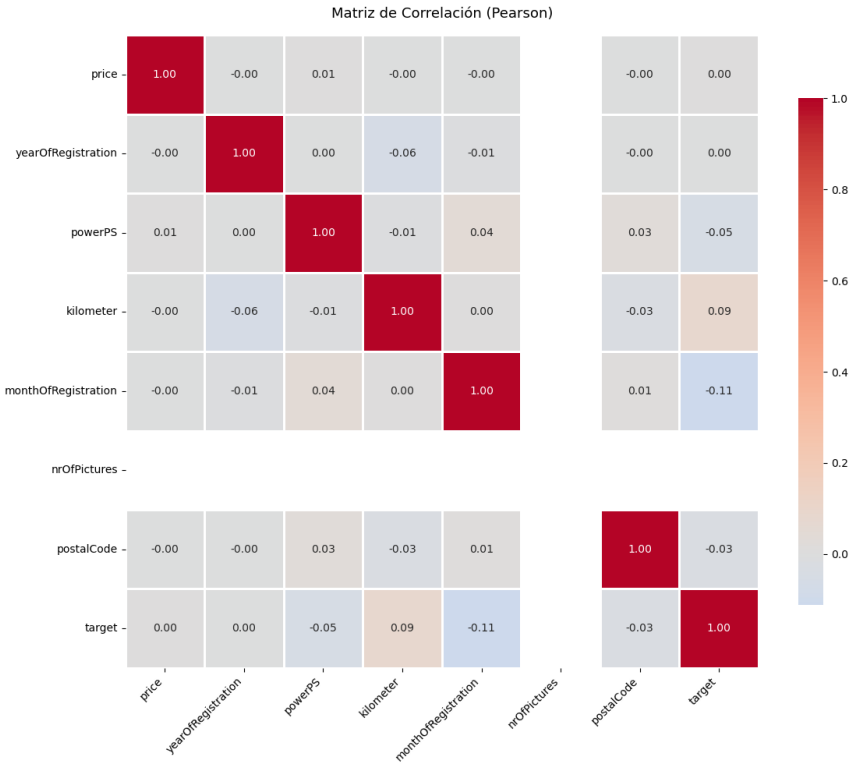
Estos resultados apoyan la percepción de que los autos diésel e híbridos, sin bien son menos comunes, a su vez son más costosos por su tecnología.

6-Gráfico: Boxplot general de Price



El boxplot general evidencia una fuerte asimetría en la distribución de precios.
La mayoría de los autos se encuentran entre 500 y 15 000 euros, con presencia de valores atípicos que superan los 30 000 euros.

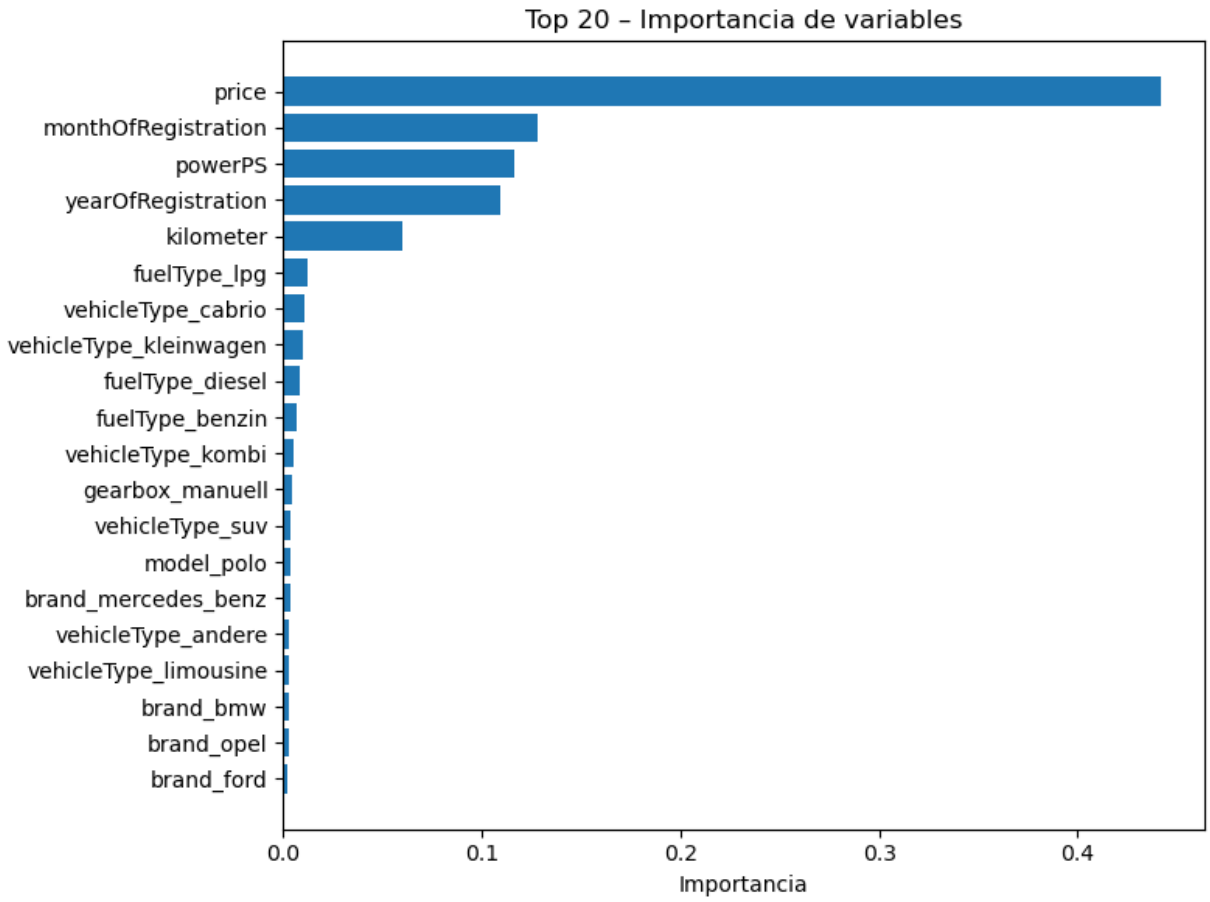
7-Gráfico: Matriz de correlación Pearson



La matriz de correlación muestra relaciones débiles entre las variables numéricas, aunque se destacan:

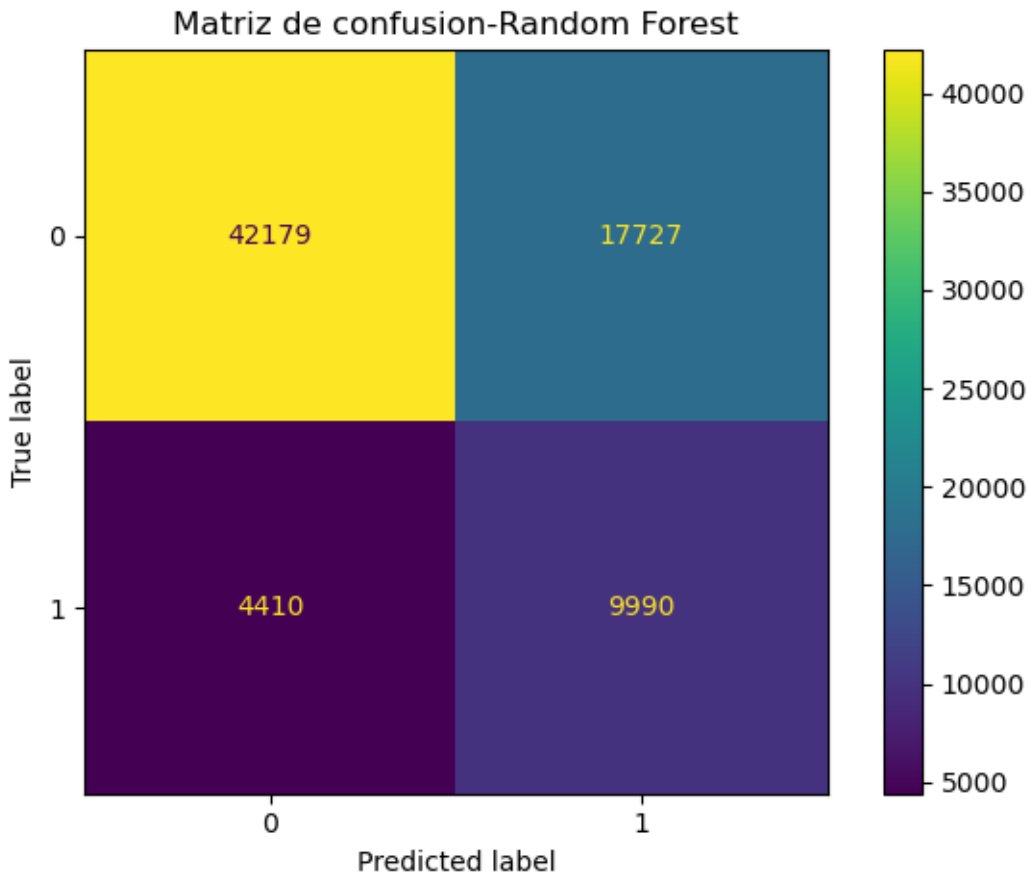
- Correlación positiva entre powerPS y price es decir, a mayor potencia, mayor valor.
- Correlación negativa entre kilometer y price es decir, a mayor kilometraje, menor valor.
- Baja correlación entre yearOfRegistration y target, lo que indica que el año no determina directamente el estado del vehículo.

8-Gráfico: Importancia de variables - Random Forest



El modelo de Random Forest identifica a price, monthOfRegistration, powerPS, yearOfRegistration y kilometer como las variables con mayor peso predictivo.
Esto muestra que el valor económico, la antigüedad, la potencia y el uso del vehículo son factores claves para inferir su estado.

9-Gráfico: Matriz de confusión del modelo Random Forest



El modelo logra una precisión global del 70%, con un mejor desempeño en la clase “sin daño” (0) que en “con daño” (1).
Esto indica que el modelo identifica correctamente la mayoría de los autos en buen estado, pero aún confunde algunos dañados.

Insights finales

El precio del vehículo es la variable más importante para explicar tanto el valor como la probabilidad de daño. Los vehículos con precios más bajos tienden a presentar más daños no reparados.

Las variables `yearOfRegistration` y `kilometer` también influyen significativamente: los autos más antiguos y con mayor kilometraje suelen ser más propensos a daños.

Con respecto al tipo de combustible, los vehículos diésel y de gas muestran precios medios más altos, posiblemente debido a su durabilidad y eficiencia, mientras que los de nafta son los más comunes, pero con mayor variabilidad de precios.

El modelo Random Forest, alcanzó una precisión del 70%, mostrando un desempeño razonable para un primer enfoque.
Esto indica que los patrones en los datos son capturables, pero existe margen para mejorar mediante modelos más complejos o ajuste de hiper parámetros.

Como conclusión, el análisis demuestra que el precio y las características técnicas del vehículo (potencia, kilometraje, antigüedad) son determinantes en su estado y valoración, ofreciendo insights útiles para el mercado automotor y una base sólida para futuros modelos predictivos.