Predicción del precio de vehículos

Empresa VALTEL

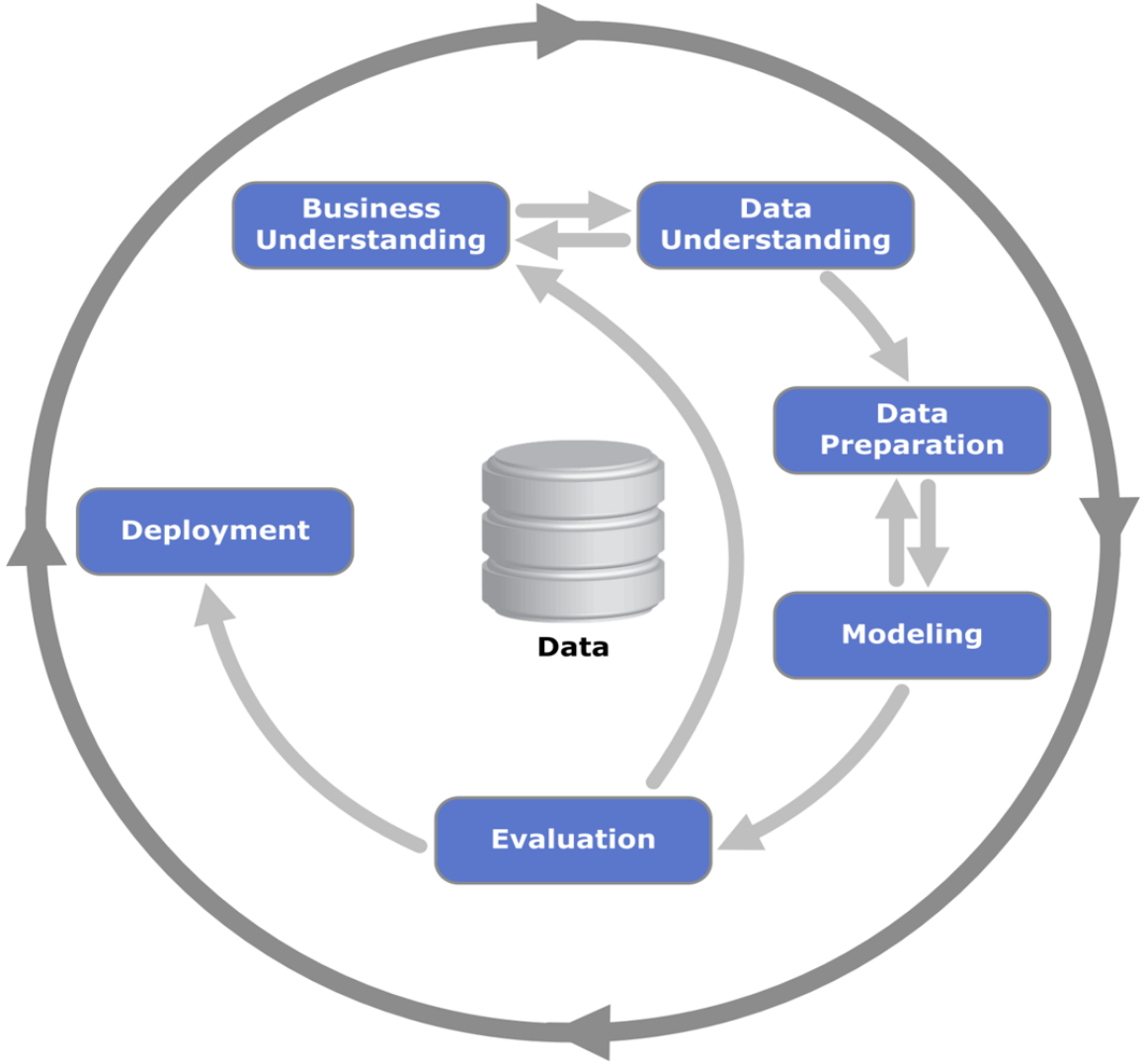
Descripción del proyecto

- **CRISP-DM** (Cross Industry Standard Process for Data Mining) ha sido el marco metodológico para estructurar el proyecto de principio a fin.

- Permite seguir un **enfoque iterativo**, **flexible** y **replicable** que se adapta perfectamente a problemas de predicción como el que abordamos.

- Facilita una **clara división de etapas**: desde el entendimiento del negocio y los datos, hasta la preparación, modelado, evaluación y despliegue.

- Su enfoque centrado en el **valor del negocio** nos asegura que las decisiones técnicas estén siempre alienadas con los objetivos del análisis.

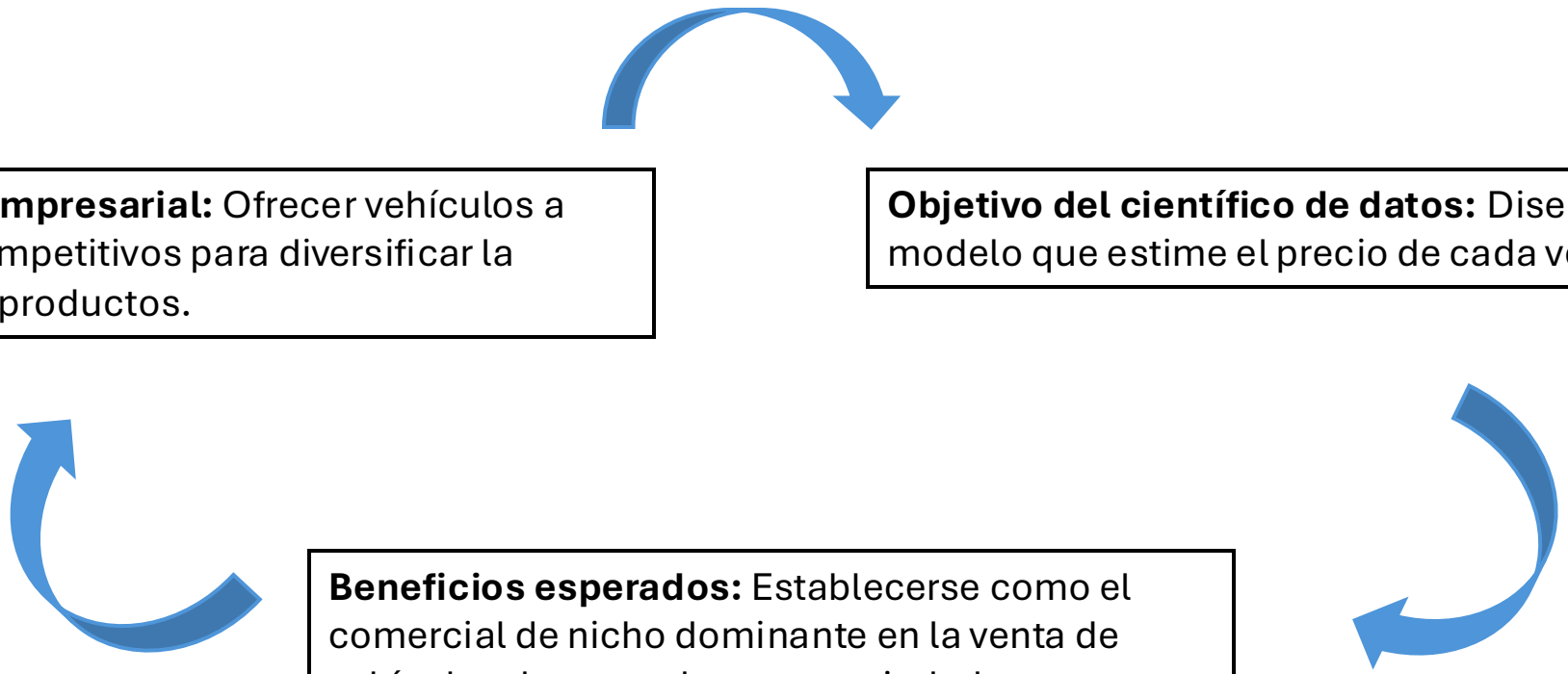


Entendimiento del negocio

Objetivo empresarial: Ofrecer vehículos a precios competitivos para diversificar la cartera de productos.

Objetivo del científico de datos: Diseñar un modelo que estime el precio de cada vehículo.

Beneficios esperados: Establecerse como el comercial de nicho dominante en la venta de vehículos de segunda mano a ciudadanos no nacionales.

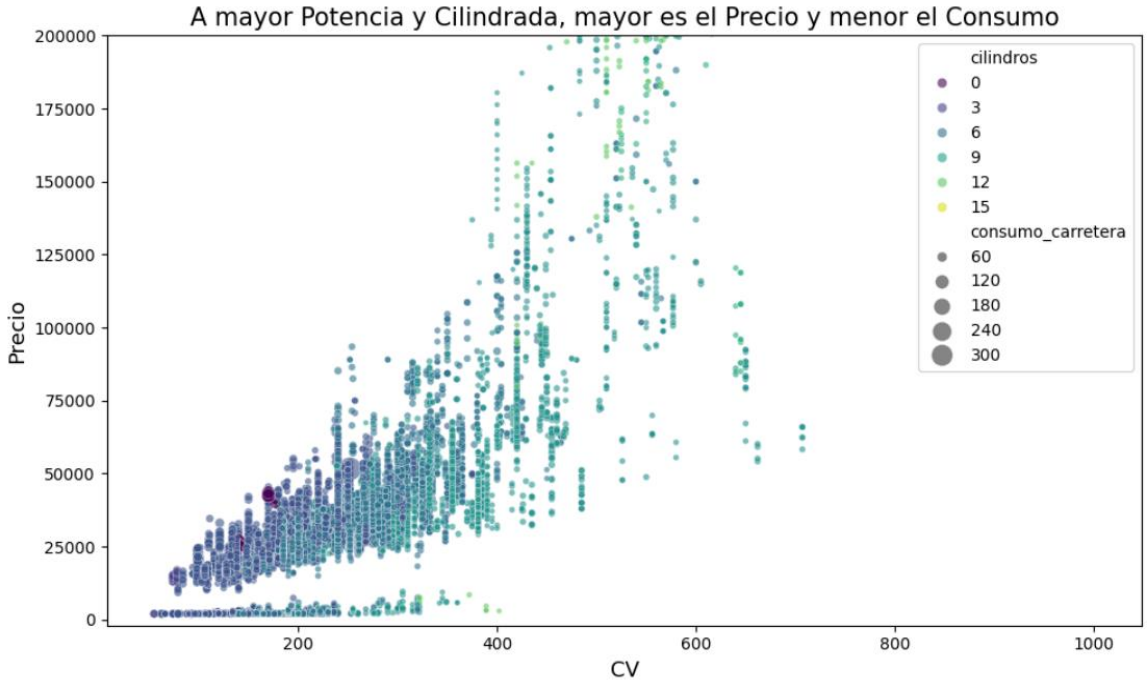


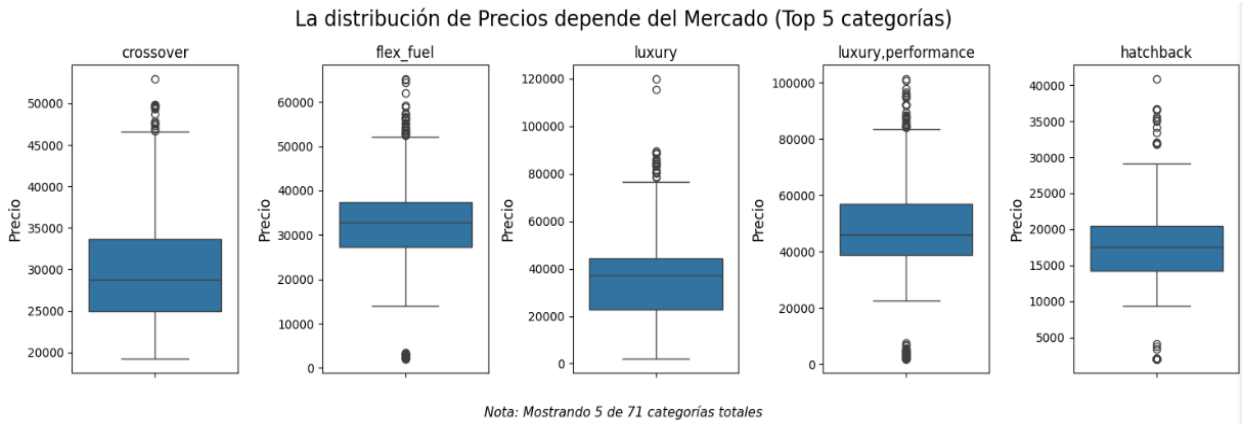
Entendimiento de los datos

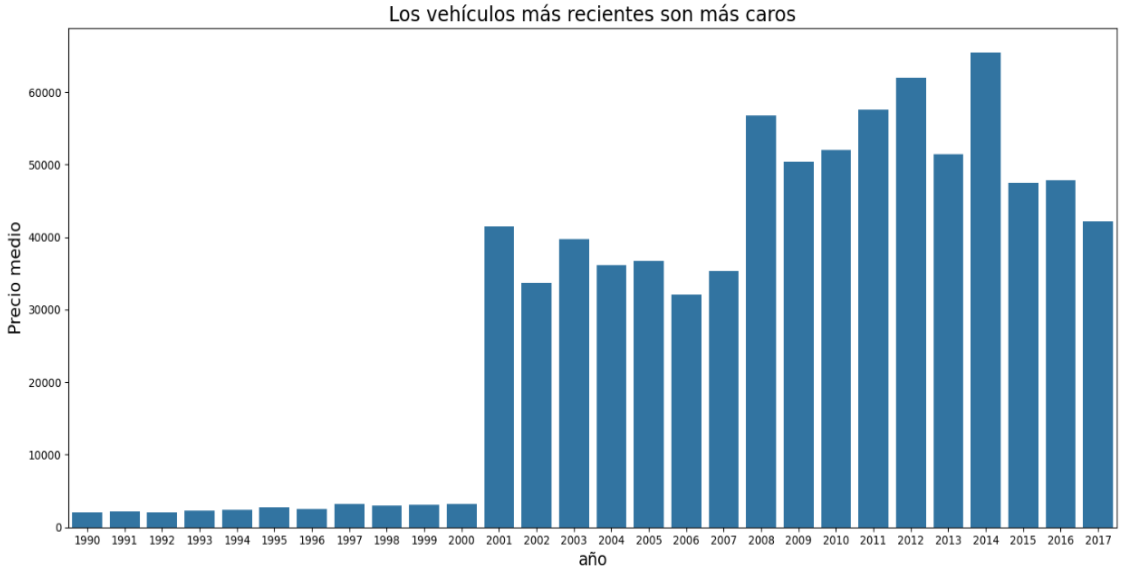
El departamento de I+D nos ha proporcionado un dataset con **11199 registros**.

Principales dimensiones:
- **CV (potencia)**
- **Cilindros**
- **Mercado**
- **Año**
- **Consumo**

Indicador principal:
- **Precio**



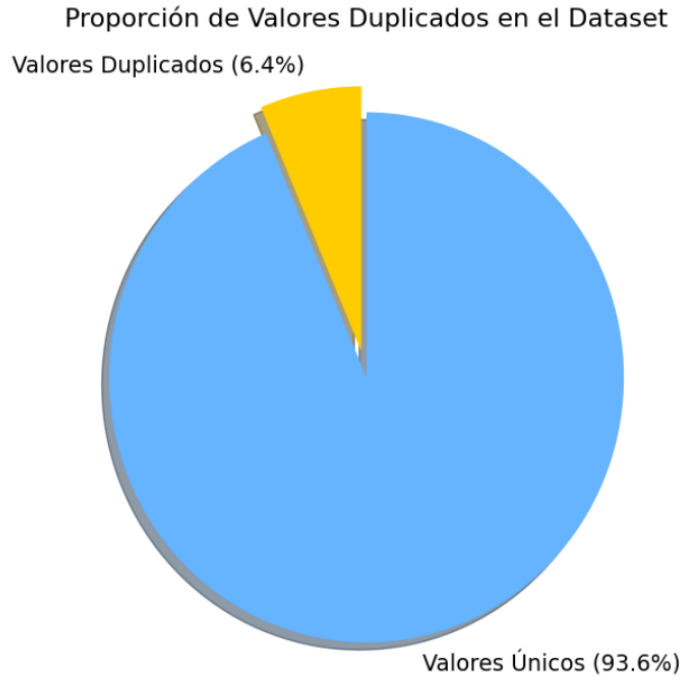


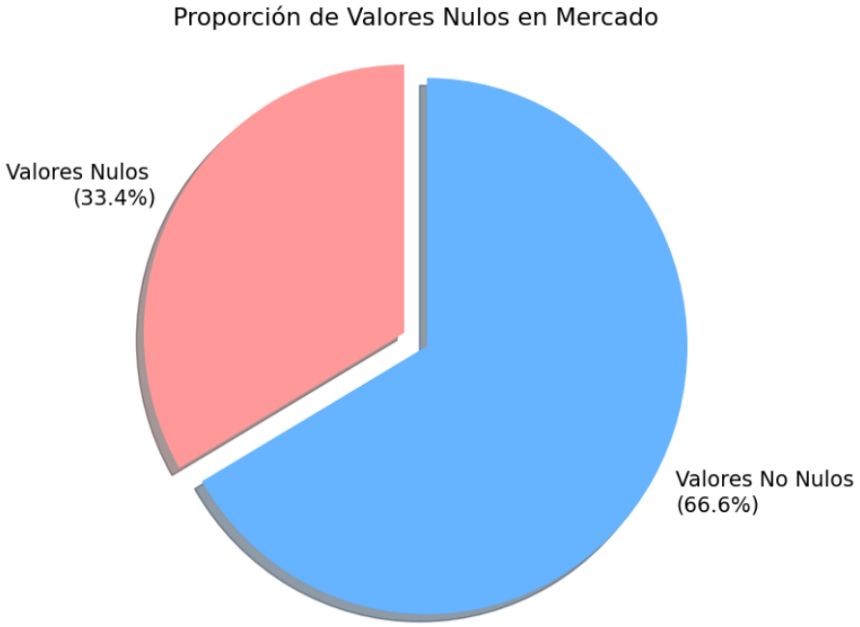


Preparación de los datos

Problemas principales:
- **Mercado**: 3742 nulos (31.41%).
- **Duplicados**: 715 duplicados (6.4%).
- **Outliers**: Gran cantidad de valores atípicos en todas las variables.
- **Variables** correlacionadas **no linealmente**.
- **Variables redundantes** -> **Consumo en Ciudad**

Soluciones propuestas:
- **Imputar** los **nulos** según la moda/mediana/media en **variables similares**.
- **Eliminar** los **duplicados**.
- Utilizar **modelos basados en árboles**, pues soportan **outliers** y la **no linealidad.**
- **Eliminar** la variable **Consumo en Ciudad** y **utilizar** únicamente el **Consumo en Carretera**.



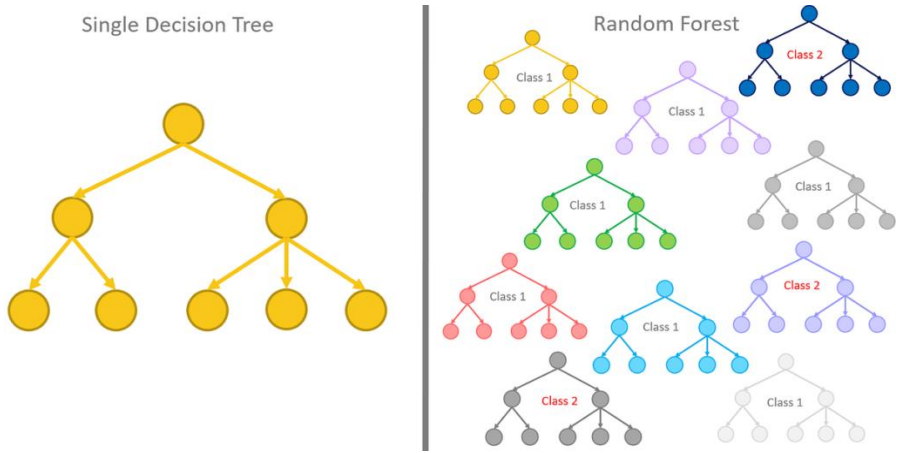


Modelado

¿Qué modelos escogemos?

Los modelos escogidos:
- **Árboles de Decisión**.
- **Random Forest**.
- **XGBoost,**

reúnen las siguientes características:
- No requieren escalado ni estandarización.
- Tolerantes a outliers.
- Manejan eficientemente múltiples variables.
- Admiten relaciones no lineales entre variables.



¿Sobre qué datos los entrenamos?

Por redundancia, crearemos **2 datasets**:
- Uno con las filas **nulas imputadas**.
- Otro con las filas **nulas eliminados**.

¿Cómo los entrenamos?

- Escogeremos un **80%** de datos para el **entrenamiento** y un **20%** para la **evaluación**.
- Aplicaremos **validación cruzada**.
- Generaremos: 3 árboles x 2 datasets = **6 modelos.**
- Seleccionaremos el modelo, con los datos, que **mejor desempeño** obtenga.

Evaluación

Random Forest

Dataset **sin nulos**:
- MAE: 4677.43
- RMSE: 21846.95
- R^2: 0.9137

Dataset **con nulos imputados**:
- MAE: 3060.03
- RMSE: 6332.32
- R^2: 0.9826

Árbol de Decisión

Dataset **sin nulos**:
- MAE: 5124.49
- RMSE: 23423.72
- R^2: 0.9008

Dataset **con nulos imputados**:
- MAE: 3215.62
- RMSE: 6707.06
- R^2: 0.9805

Escogemos esta configuración

XGBoost

Dataset **sin nulos**:
- MAE: 4971.11
- RMSE: 15680.18
- R^2: 0.9556

Dataset **con nulos imputados**:
- MAE: 3404.06
- RMSE: 6277.15
- R^2: 0.9829

- Los datasets con **nulos imputados** obtienen un **mayor desempeño**.
- Los 3 modelos obtienen un **rendimiento** muy similar y **excepcionalmente alto**.
- Escogemos el **Arbol de Decisión** por tener una mayor explicabilidad para el negocio.

Despliegue

¿Qué acciones debemos tomar?

Piloto	Monitorización	Operativizar
Realizar una **prueba piloto** en un conjunto reducido de clientes.	**Medir** los **resultados** del modelo en los clientes.	**Desplegar** el modelo y **automatizar** sus acciones.
Decidir si se debe **ampliar** a otros clientes.	Establecer **KPI's** clave: - Satisfacción del cliente - Porcentaje de compras reales en interesados	**Monitorizar** métricas de rendimiento. Si el modelo pierde rendimiento, considerar reentrenamiento (MLOps).

Muchas gracias por su atención