

Aplicación del Ciclo de Vida CRISP-DM:

Predicción de Eficiencia de Combustible (MPG)

Carlos Andrés Cantor Castaño

Institución Universitaria Pascual Bravo

MACHINE LEARNING

Wilson Andres Ramirez Rios

18/11/2025

## **Aplicación del Ciclo de Vida CRISP-DM: Predicción de Eficiencia de Combustible (MPG)**

1. Business Understanding (Comprensión del Negocio)
2. Data Understanding (Comprensión de los Datos)
3. Data Preparation (Preparación de los Datos)
4. Modelado
5. Evaluación y Selección
6. Análisis de Ética y Sesgos
7. Conclusiones

## **1. Business Understanding**

### **Definición del Problema**

El problema central de este estudio es el desarrollo de un modelo de aprendizaje supervisado, específicamente un modelo de regresión. El objetivo principal que se busca es predecir o estimar la eficiencia de combustible de un vehículo, medida en Millas por Galón (MPG). La variable objetivo, por lo tanto, es mpg, una variable numérica continua. Esta predicción se realizará utilizando un conjunto de características (variables predictoras) que describen las especificaciones técnicas de los automóviles, tales como el número de cilindros, la potencia (horsepower), el peso del vehículo, la aceleración y su origen de fabricación.

### **Justificación**

La capacidad de predecir la eficiencia del combustible es un problema de alta relevancia por dos razones principales. Primero, desde la perspectiva del consumidor, un modelo preciso permite estimar los costos operativos futuros (gasto en combustible) asociados a un vehículo, influyendo directamente en la decisión de compra. Segundo, desde el punto de vista de la industria automotriz, este tipo de análisis es crucial. Permite a los fabricantes e ingenieros entender qué características de diseño (como la reducción de peso o los cambios en el motor) tienen el mayor impacto en el consumo. Esto es vital para optimizar el diseño de nuevos modelos, cumplir con las crecientes regulaciones ambientales y satisfacer la demanda del mercado de vehículos más eficientes.

## Criterios de Éxito

Dado que este es un problema de regresión y no de clasificación, la métrica de evaluación clave seleccionada es el Coeficiente de Determinación, comúnmente conocido como  $R^2$  (R-cuadrado).

Se elige  $R^2$  porque ofrece una interpretación muy clara: mide el porcentaje de la variación en la variable objetivo (mpg) que es explicado por las variables predictoras del modelo. A diferencia de las métricas de error, el  $R^2$  nos dice qué tan bien el modelo comprende el problema en su totalidad. Un modelo que no entiende nada tiene un  $R^2$  de 0, mientras que un modelo perfecto tiene un  $R^2$  de 1. Por lo tanto, el criterio de éxito para este proyecto se define como alcanzar un valor de  $R^2$  igual o superior a 0.80 (o 80%) en el conjunto de datos de prueba. Lograr este umbral indicaría que el modelo es capaz de explicar la gran mayoría de la variabilidad en el consumo de combustible.

## 2. Data Understanding (Comprensión de los Datos)

### 2.1. Descripción Inicial

Se cargó el conjunto de datos "mpg" mediante la librería Seaborn. El dataset consta de 398 registros y 9 columnas. Las variables incluyen datos continuos (como peso y desplazamiento), discretos (cilindros, año) y categóricos (origen, nombre).

### 2.2. Análisis Exploratorio

Durante la inspección inicial, detectamos que la variable horsepower (caballos de fuerza) presentaba 6 valores nulos, lo cual requirió tratamiento posterior. Al analizar la variable objetivo mpg mediante histogramas, observamos una distribución sesgada a la derecha, indicando que la mayoría de los autos tienen un rendimiento medio-bajo.

### 2.3. Visualización y Relaciones

Utilizamos mapas de calor (heatmaps) y gráficos de dispersión para entender las relaciones:

- **Correlaciones Negativas:** Confirmamos que a mayor peso (`weight`) y potencia (`horsepower`), menor es el rendimiento (`mpg`).
- **Influencia del Origen:** Mediante diagramas de caja (boxplots), observamos que los vehículos de origen japonés tienden a ser más eficientes que los estadounidenses en este conjunto de datos.
- **Valores Atípicos:** Se identificaron valores atípicos (outliers) en la potencia y aceleración, lo cual influyó en nuestra estrategia de imputación.

### 3. Data Preparation (Preparación de los Datos)

Esta fase fue crítica para asegurar la calidad del modelo. Se tomaron las siguientes decisiones:

#### 3.1. Imputación de Datos

Para los valores faltantes en horsepower, optamos por imputar utilizando la mediana en lugar del promedio. Esta decisión se basó en la presencia de valores atípicos detectados en el análisis anterior, ya que la mediana es más robusta y no se deja distorsionar por valores extremos.

#### 3.2. Transformación de Variables (Feature Engineering)

- **Eliminación:** Se eliminó la variable `name` por ser un identificador único sin valor predictivo generalizable.
- **Codificación Categórica:** La variable `origin` se transformó mediante *One-Hot Encoding* para que el modelo pudiera interpretar matemáticamente las regiones (USA, Europa, Japón).
- **Tratamiento de Discretas:** Las variables `cylinders` y `model_year`, aunque numéricas, se trataron como categóricas. Esto se debe a que la relación entre el número de cilindros y el consumo no es perfectamente lineal; tratarlas como categorías permite al modelo capturar mejor las diferencias específicas entre tener 4, 6 u 8 cilindros.

#### 3.3. División y Estandarización

Se dividieron los datos en un conjunto de entrenamiento (80%) y prueba (20%). Posteriormente, aplicamos una estandarización (StandardScaler) a las variables numéricas para ponerlas en la misma escala, evitando que variables con magnitudes grandes (como el peso en miles de libras) dominaran a las pequeñas (como la aceleración).

#### 4. Modelado

Seleccionamos dos algoritmos supervisados de regresión para comparar su desempeño:

1. **Regresión Lineal:** Elegida como modelo base por su simplicidad y capacidad de interpretación.
2. **Árbol de Decisión (Regressor):** Elegido por su capacidad para capturar relaciones no lineales en los datos.

Ambos modelos fueron entrenados con el mismo conjunto de datos procesado.

## 5. Evaluación y Selección

Para medir el rendimiento en el conjunto de prueba (datos desconocidos para el modelo), utilizamos las métricas  $R^2$ , MAE y RMSE.

### 5.1. Discusión de Resultados

Al comparar los modelos, la Regresión Lineal demostró un desempeño sólido y consistente. El modelo logró un  $R^2$  superior al 0.80, cumpliendo con nuestro criterio de éxito inicial. El Error Absoluto Medio (MAE) se mantuvo en un rango bajo, lo que significa que las predicciones del modelo suelen desviarse pocas millas por galón del valor real, ofreciendo una precisión útil para el usuario final.

### 5.2. Selección Final

Se seleccionó el modelo que ofreció el mejor equilibrio entre precisión ( $R^2$  alto) y generalización (menor error RMSE). Este modelo es capaz de explicar satisfactoriamente el comportamiento del consumo de combustible basándose en las especificaciones técnicas.

## 6. Análisis de Ética y Sesgos

Es fundamental reconocer las limitaciones éticas y prácticas del modelo desarrollado:

- **Sesgo Temporal:** El dataset pertenece a las décadas de 1970 y 1980. Usar este modelo para predecir la eficiencia de un auto moderno (2024) sería incorrecto y éticamente cuestionable, ya que la tecnología automotriz ha cambiado radicalmente.
- **Sesgo de Origen:** Los datos reflejan una época donde los autos americanos eran notablemente menos eficientes debido a contextos históricos (precios de gasolina). El modelo podría perpetuar el estereotipo de que "los autos de USA consumen mucho", lo cual puede no ser cierto en el mercado actual.

## 7. Conclusiones

El proyecto ha completado exitosamente el ciclo CRISP-DM. Se logró desarrollar un modelo predictivo que cumple con los estándares de calidad definidos. Sin embargo, su aplicación debe limitarse al contexto histórico de los datos o utilizarse con fines educativos, evitando su despliegue en entornos comerciales actuales sin una actualización de los datos.