

**INGENIERIA EN DESARROLLO Y GESTIÓN DE SOFTWARE**

Logotipo

Descripción generada automáticamente con confianza media

**ACTIVIDAD**

M2-PROYECTO DE MES

**MATRICULA**:

20221267

**ALUMNO**:

HERNÁNDEZ TAPIA HAILIE

**CUATRIMESTRE: GRUPO:**

9 B

**DOCENTE:** DR. EFRÉN JUÁREZ CASTILLO

**ASIGNATURA:** EXTRACCIÓN DE CONOCIMIENTO EN BASE DE DATOS

**HUEJUTLA DE REYES, HIDALGO 27 JUNIO 2025**

**Introducción:**

El conjunto de datos Auto MPG proviene del Repositorio de Aprendizaje Automático de la UCI y contiene especificaciones técnicas de vehículos relacionadas con el consumo de combustible en ciclo urbano, medido en millas por galón (mpg). Es una versión modificada del conjunto original de la biblioteca StatLib de la Universidad Carnegie Mellon, utilizado en la Exposición de la Asociación Estadística Estadounidense de 1983.

**Número de instancias:** 398

**Número de atributos:** 9 incluido el atributo de clase

1. **mpg**: Consumo de combustible en millas por galón (continuo, variable objetivo).
2. **cylinders**: Número de cilindros (discreto multivalor).
3. **displacement**: Desplazamiento del motor en pulgadas cúbicas (continuo).
4. **horsepower**: Caballos de fuerza (continuo, con 6 valores faltantes).
5. **weight**: Peso del vehículo en libras (continuo).
6. **acceleration**: Tiempo en segundos para acelerar de 0 a 60 mph (continuo).
7. **model year**: Año del modelo (discreto multivalor).
8. **origin**: Origen del vehículo (discreto multivalor, e.g., 1: EE.UU., 2: Europa, 3: Japón).
9. **car name**: Nombre del vehículo (cadena, único para cada instancia).

**Objetivo de la predicción:**

* El objetivo principal es predecir el consumo de combustible (variable mpg).

**Contexto y relevancia del problema**

El conjunto de datos Auto MPG se centra en el consumo de combustible en ciclo urbano, un tema relevante tanto en el pasado como en la actualidad debido a preocupaciones económicas, ambientales y de eficiencia energética. En la década de 1970, durante la crisis del petróleo, la eficiencia del combustible se convirtió en un factor crítico para los fabricantes de automóviles y los consumidores. Este conjunto de datos, recopilado en 1983, refleja el interés en comprender cómo las características técnicas de un vehículo (como el número de cilindros, el peso o los caballos de fuerza) influyen en su rendimiento de combustible.

**Desafíos del Conjunto de Datos**

* **Valores Faltantes**: Los 6 valores faltantes en **horsepower** requieren técnicas de imputación o manejo específico para evitar sesgos en los modelos.
* **Naturaleza Mixta de Atributos**: La combinación de atributos continuos y discretos multivalor exige un preprocesamiento cuidadoso (e.g., codificación de variables categóricas como "origin").
* **Tamaño Moderado**: Con 398 instancias, el conjunto es manejable pero requiere validación cruzada (como se hizo en estudios previos, e.g., Quinlan, 1993) para garantizar modelos robustos.

**Entendimiento de los Datos:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Dato** | **Definición y significado de cada atributo** | **Tipo de atributo** | **Tipo de dato en python** | **Presencia de valores nulos o faltantes** |
| mpg | Consumo de combustible en millas por galón. | Ratio | float64 | 0 |
| cylinders | Número de cilindros. | Ordinal | int64 | 0 |
| displacement | Desplazamiento del motor en pulgadas cúbicas. | Ratio | float64 | 0 |
| horsepower | Caballos de fuerza | Ratio | object | 0 |
| weight | Peso del vehículo en libras. |  | int64 | 0 |
| acceleration | Tiempo en segundos para acelerar de 0 a 60 mph. | Ratio | float64 | 0 |
| model year | Año del modelo. | Ordinal | int64 | 0 |
| origin | Origen del vehículo. | Nominal | int64 | 0 |
| car name | Nombre del vehículo. | Nominal | object | 0 |

**Preparación de los Datos:**

* + - **Tratamiento de Valores Faltantes:** Generar y tratar valores nulos si no existen en el DataSet. Explicar y justificar la técnica utilizada.
    - **Conversión de Atributos:** Convertir atributos de texto a numéricos si es necesario. Explicar la necesidad de esta conversión.
    - **Escalado de Atributos:** Normalizar los valores de X si es necesario, justificando la decisión.
    - **Selección de Atributos:** Seleccionar entre 4 y 6 atributos utilizando RFE, matriz de correlación y árboles de decisión. Justificar, con base en los tres métodos la elección y exclusión de atributos.
  + **Modelado:**
    - **Entrenamiento de Modelos:** Entrenar dos modelos de regresión, uno de los cuales debe ser una red neuronal (el otro modelo no debe ser red neuronal). Utilizar k-fold cross-validation.
  + **Evaluación de Modelos:**
    - **Métricas de Evaluación:** Evaluar los modelos utilizando R² y alguna otra métrica adecuada (RMSE, MAE), realizar validación cruzada y reportar la media, desviación estándar y varianza.
    - **Gráfica de Dispersión:** Incluir y analizar gráficas de dispersión de los resultados de ambos modelos.
    - **Gráfica de Residuos:** Incluir y analizar gráficas de residuos para cada modelo (Para las gráficas realiza una ejecución empleando Train Test Split ).
  + **Evaluación:**
    - **Conclusión:** Indicar cuál modelo fue mejor y explicar el motivo.
  + **Despliegue:**
    - **Implementación en Render:** Describir brevemente la implementación y despliegue del modelo utilizando Flask y Render.
  + **Bibliografía:** Incluir referencias en formato APA o IEEE.
  + **Generalidades:** Incluir explicación, código y evidencia de ejecución del código copiados de la libreta de Jupyter.

1. **Despliegue del Modelo en Render:**
   * **Servicio Web:** Crear un servicio web utilizando Flask.
   * **Página Web:** Desarrollar una página web que muestre el modelo de regresión con un formulario para ingresar valores.
   * **Despliegue en Render:** Desplegar el modelo en la plataforma Render.
   * **Compartir URL:** Compartir la URL del formulario y colocar el link en la libreta de Jupyter.