---1---

Linear Regression Challenge

## **Predicting Life Expectancy**

### **Overview of the Task**

Life expectancy is a critical indicator of a country's overall health and development. In this challenge, you are provided with a dataset from the World Health Organization (WHO) that includes various features related to health and socio-economic factors. Your task is to clean and structure the data, perform exploratory data analysis (EDA), and develop a regression model to predict life expectancy based on these features.

### **Problem Statement**

The dataset includes a mix of structured and potentially unstructured data that can affect the quality of your regression model. The challenge is to preprocess the data, handle any missing or inconsistent values, and build a regression model that accurately predicts life expectancy using the available features.

### **Objective**

Your main objectives are:

1. **Data Cleaning and Structuring**: Address and correct issues such as missing values, inconsistencies, and potential unstructured data.
2. **Feature Engineering**: Identify and create features that will improve the performance of the regression model.
3. **Model Building**: Develop and train a regression model to predict life expectancy using the cleaned dataset.
4. **Evaluation**: Assess the performance of the model using appropriate metrics and validate its accuracy.
5. **Deployment**: Create an application or tool that allows users to input health and socio-economic data and receive life expectancy predictions.

### **Data Description**

**Data Source**: The dataset for this challenge is sourced from the Kaggle dataset [Life Expectancy (WHO)](https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who). The dataset includes:

* **Features**:
  + Country: Name of the country.
  + Year: Year of the recorded data.
  + Status: Country's income group classification (e.g., High income, Low income).
  + Life Expectancy: Life expectancy at birth (target variable), measured in years.
  + Adult Mortality: Number of adult deaths per 1,000 adults aged 15-60.
  + infant deaths: Number of infant deaths per 1,000 live births.
  + Alcohol: Per capita alcohol consumption in liters.
  + Percentage Expenditure: Percentage of GDP spent on health care.
  + Hepatitis B: Percentage of people vaccinated against Hepatitis B.
  + Measles: Number of reported cases of measles per 1,000 population.
  + BMI: Average Body Mass Index of the population.
  + under-five deaths: Number of deaths per 1,000 live births for children under five years old.
  + Polio: Percentage of children vaccinated against Polio.
  + Total expenditure: Total health expenditure as a percentage of GDP.
  + Diphtheria: Percentage of children vaccinated against Diphtheria.
  + HIV/AIDS: Number of people living with HIV/AIDS per 1,000 population.
  + GDP: Gross Domestic Product per capita.
  + Population: Population of the country.
  + thinness 1-19 years: Percentage of children aged 10-19 years who are underweight.
  + thinness 5-9 years: Percentage of children aged 5-9 years who are underweight.
  + Income composition of resources: Measures the educational attainment and economic resources.

### **Evaluation Criteria**

|  |  |
| --- | --- |
| **Data Collection, Connectivity, and Cleaning** | **30%** |
| **Logic and Results** | **40%** |
| **Presentation** | **5%** |
| **Code Quality** | **5%** |
| **Deployment / Running App** | **20%** |