**Assignment 2 – Life Expectancy**

**By: S V Prissha – 2313128**

**Steps:**

* The data set provided is Life Expectancy.
* Analyse the data set provided: It is evident that the dataset has a lot of missing data points hence needs data cleaning and more specifically data preprocessing.
* Initially upload the data set in Google Colab and then import the necessary libraries like pandas, numpy, matplotlib, sklearn etc.
* Read the file.
* Data Preprocessing: Minmax Scaling, Label encoding, Z-Score, Feature engineering and scaling
* After preprocessing the data then decide the dependent and independent variable from the data set.
* Do train test split, 80% of the data set is used to train the model and the rest 20% is used to the test the model and see its accuracy.
* Import the pickle library for serialization.

**Simple Linear Regression:**

**Interpretation and result:**

1. **Variables:**

**Dependent variable:** Life Expectancy

**Independent variable:** BMI

1. **R² Score: 0.351**

Inference: The R Square is 35.1%, which shows that the model is not a good fit. 35.1% of the variation in Life Expectancy is explained by the variation in BMI.

1. **Regression Equation:**

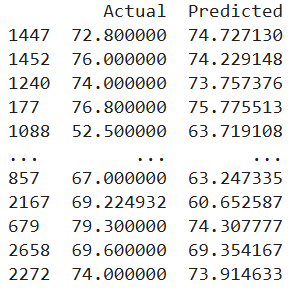
**Intercept:** 59.290

**Coefficient:** 0.262

**Life Expectancy = 59.290 + 0.262\* (BMI)**

Inference: Which show that if the BMI increases the Life Expectancy will also increase by 0.262

1. **Comparing the actual and predicted Life Expectancy:**



**Multiple Linear Regression:**

**Interpretation and Result:**

1. **Variables:**

**Dependent Variable**: Life Expectancy

**Independent Variables:** 'Adult Mortality', 'infant deaths', 'Alcohol', 'percentage expenditure', 'Hepatitis B', 'Measles ', ' BMI ', 'under-five deaths ', 'Polio', 'Total expenditure', 'Diphtheria’, ' HIV/AIDS', 'GDP', 'Population', ' thinness 1-19 years', ' thinness 5-9 years', 'Income composition of resources', 'Schooling

1. **R Square = 0.843**

Inference: The R Square is 84.3%, which shows that the model is a good fit. 84.3% of the variation in Life Expectancy is explained by the variation in all the independent variables.

1. **Mean Squared Error: 11.449**

The MSE is low which shows there is low deviation from the actual Life Expectancy.

1. **Regression Equation:**

**Intercept:** 56.971

**Coefficients:** [-8.33471703 27.81201012 0.38618859 2.47817654 -2.81604449 0.05806632 -0.37302702 -30.40114899 2.7187973 2.6584802 5.15746618 -11.30306909 -0.99778646 1.40731999 2.08159109 -4.0180467 22.54295921 -4.78988747]

**Life Expectancy = 56.971 – 8.334 \*(Adult Mortality) + 27.812\*(Infant Deaths) + 0.386\*(Alcohol) + 2.478\*(Percentage Exp) -2.816 \*(Hepatitis B)** **+ 0.058\* (Measles) - 0.373\*(BMI) - 30.402\*(Under 5 deaths) + 2.719\*(Polio) + 2.658\*(Total Exp) + 5.157\*(Diptheria) - 11.303\*(HIV/AIDS) - 0.998\*(GDP) + 1.407\*(Population) + 2.082\*(Thinness 1-19 yrs) - 4.018\*(Thinness 5-9 yrs) + 22.543\*(Income Composition of resources) - 4.790\*(Schooling)**

This shows the impact of each feature on Life Expectancy,

There is an inverse relationship between features like adult mortality, Hepatitis B, BMI, Under 5 deaths, HIV/AIDS, GDP, Schooling and Life expectancy.

Other features or variables seem to have a direct relationship with Life Expectancy.

1. **Comparing the actual profits and predicted profits:**

A screenshot of a number

Description automatically generated