Optimisation and Machine Learning Project 17 June 2024,
titled “Life Expectancy Prediction”

Introduction: The project involves the development of a Machine Learning model to predict the ‘Life Expectancy’ of a person based on the relevant features in the “life” Dataset.

1. Data Exploration and preprocessing

- The “life” dataset is a csv file with ***2928 Rows and 22 Columns***. After loading the Dataset into the Pandas Dataframe the Statistical characterisation of the different features of the Dataset has been observed with the **.describe()** function.

- The most important step in this stage was to identify any Missing Data in many columns(features) using **isnull().sum()** function.
  Except for a few features like Hepatitis_B, Population, GDP, Alcohol all the other features had minimal or zero missing values.

- Before filling the missing values the task of analysing the Distribution of the Data of each individual feature was analysed using the Distribution Plot, Histogram from seaborn, matplotlib modules.

- It was observed that the data was not constant, there were considerable Outliers in the Dataset for each feature, it is expected because of the clear contrast between the Economic and Social Indicators of the Developing and Developing countries.

- Regardless, after the going through the Distribution of the Data, the further approach to fill the missing values was done using many techniques like **Mean, Median, Mode, linear and polynomial Interpolation, Forward and Backward fill etc**

- After filling the missing values, very few missing values that couldn’t be filled were dropped **dropna()**

2. Feature Selection and Engineering

- The data set had many features that could play no role in the prediction of the Life Expectancy of a person, like Country Name, Year of collection Data, Development status of the country. These columns have been dropped from the Dataset before preparing standardisation and training.

- The next step was to look for any **CORRELATION** among the features of the Dataset and this was done accordingly with the help of **.corr()** function and a **Heat Map** of all the Features with their respective correlations has been plotted.

3. Model Development and Evaluation

- At this stage the Data has been analysed and features that are important have been identified. But the task of Standardising the Data and Splitting the Data into Training and the Testing data have to be performed.

- First, I choose to do the Training and Splitting of the Data then go for the Standardisation.

- All the above tasks were performed using the **train_test_split()** and **StandardScalar()** modules of the sklearn library, Standardisation of Training and Test Data plays an important role in Normalising Data so that No feature can dominate the Training Model.

- For choosing a good Machine learning Model, the choice was fairly straightforward since the task was to Predict the Life Expectancy of a person given some information about the living conditions of the country he lives in.

- My choice for this task was the **LinearRegression()** module which performs the Regression Task i.e approximating the best possible value under certain conditions with minimal Loss. *Since the Dataset was not that large I didn't choose SGDRegressor().*

4. Prediction and Interpretation:

- The Training and Test Data were in 4:1 ratio. The model was trained using the training Data and the Model was evaluated using the two important metrics MEAN SQUARED ERROR, and R2_SCORE.

- Both the metrics were in the acceptable ranges, MSE: `15.664568853169072 and` R2_SCORE: `0.8264685389240712`

- There were no signs of over-fitting of the Model, Interestingly after the first training of the model, I went back to step-2 to alter the way the Missing values were filled using the techniques to fill the missing values, In this process I have randomly tried few changes and found out that OUTLIERS in the data had significant impact in filling the missing values as the disparity was too large the mean, median, mode values couldn't do justice to the task completely.

5. Conclusion and Recommendations:

- The Trained Model gave decent predictions and has performed well over some randomly chosen solo values which can be seen in the code.

- The Model is very subjective and the designer can choose the features he desires, the impact of all the features provided is even though relevant from prediction point of view, But grouping the Developed and Underdeveloped Countries definitely introduced many discrepancies and outliers in the Data. Therefore no straight forward approach which was sufficient to fill the missing Data.

- Even more pre-processing could be done where the Data Set could be further divided in more manageable subsections based on the development index to make more relevant predictions.