# Life Expectancy Prediction Report

## 1.Introduction

This report presents the results of a data science project aimed at predicting life expectancy for various countries. The primary goal of the project is to develop predictive models using different machine learning algorithms, including Linear Regression, Decision Tree, and Random Forest. The models are evaluated to determine which provides the most accurate predictions.

The dataset used for this project contains various features related to socio-economic factors, health indicators, and other relevant attributes that might influence life expectancy.

## 2.Data Understanding

The dataset contains the following columns (features):

- Country

- Year

- Life Expectancy

- Adult Mortality

- Infant Deaths

- Alcohol Consumption

- 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

## 3.Data Exploration

Initial exploration of the dataset included checking for missing values, understanding the distribution of each feature, and identifying any potential outliers. Various visualizations and summary statistics were used to get a comprehensive understanding of the data.

## 4.Data Preparation

Data preparation involved several steps:

### 1.Handling Missing Values:

Missing values were handled by either filling them with appropriate statistics (mean, median, mode) or dropping the rows/columns with a significant number of missing values.

### 2.Feature Engineering:

Additional features were created based on existing ones to enhance the predictive power of the models.

### 3. Data Transformation:

Numerical features were standardized or normalized to ensure they contribute equally to the model. Categorical features were encoded using techniques like one-hot encoding.

### 4. Splitting the Dataset:

The dataset was split into training and testing sets to evaluate the models' performance on unseen data.

## 5.Modeling

Three different machine learning algorithms were implemented to predict life expectancy:

### 1.Linear Regression:

A simple and interpretable model that assumes a linear relationship between the dependent and independent variables.

### 2.Decision Tree:

A non-linear model that splits the data into subsets based on the feature that results in the best split. It is easy to interpret and can capture complex relationships.

### 3.Random Forest:

An ensemble method that builds multiple decision trees and combines their predictions. It reduces overfitting and improves generalization.

## 6.Evaluation

The performance of each model was evaluated using the Mean Squared Error (MSE) metric. The results are as follows:

- Linear Regression: RMSE = 1.65

- Decision Tree: RMSE = 2.39

- Random Forest: RMSE = 3.52

The Random Forest model provided the best performance with the lowest MSE, indicating it was the most accurate in predicting life expectancy.

## 7.Conclusion

This project demonstrated the application of various machine learning algorithms to predict life expectancy based on a range of socio-economic and health-related features. The Random Forest model proved to be the most effective, highlighting the importance of using ensemble methods to improve predictive accuracy.

Future work could involve exploring additional features, fine-tuning model hyperparameters, and applying other advanced algorithms to further enhance the predictions.