Prediction of Life Expectancy Using Socio-Economic Indicators
A Statistical and Machine Learning Approach

1. Introduction

Life expectancy is one of the most important indicators of a nation's overall health and socio-economic development. It reflects the combined effects of healthcare facilities, economic growth, education, and lifestyle factors. Understanding the determinants of life expectancy helps governments and policymakers design effective public health strategies.

With the availability of large publicly accessible datasets, secondary data analysis has become an essential tool in modern statistical research. Advances in statistical modelling and machine learning techniques enable accurate prediction and deeper understanding of complex relationships among variables.

This project uses secondary data collected by the World Health Organization (WHO) to analyse and predict life expectancy using classical statistical models and machine learning algorithms implemented in Python.

2. Objectives of the Study

The main objectives of this project are:
- To analyse the relationship between life expectancy and socio-economic indicators.
- To build predictive models for life expectancy using statistical and machine learning techniques.
- To compare the performance of classical regression and machine learning models.
- To identify the most significant factors influencing life expectancy.

3. Source and Nature of Data

3.1. Type of Data

The data used in this study are secondary data, meaning they were collected by an external agency for purposes other than the current research.

3.2. Data Source

The dataset was obtained from the World Health Organization (WHO) and made available through Kaggle – Life Expectancy Dataset.

3.3. Variables Used

Variable	Description
Life expectancy	Average number of years a person is expected to live
GDP	Gross Domestic Product per capita
Adult Mortality	Probability of dying between ages 15 and 60
Schooling	Average years of education
BMI	Body Mass Index
Alcohol	Alcohol consumption per capita

4. Methodology

This study follows a quantitative analytical research design using statistical modelling and machine learning. Python was used as the primary software with libraries including pandas, NumPy, matplotlib, seaborn, and scikit-learn.

The steps followed include data cleaning, exploratory data analysis, feature selection, model building, model evaluation, and interpretation of results.

- **Exploratory Data Analysis (EDA)**
  Descriptive statistics were computed to understand the central tendency and dispersion of variables. Correlation analysis was performed to identify relationships between life expectancy and explanatory variables. GDP and schooling showed positive correlation, adult mortality showed negative correlation, and BMI and alcohol consumption showed moderate association.

- **Statistical Models**
  Multiple Linear Regression was used to model the linear relationship between life expectancy and predictors. The model is expressed as $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \varepsilon$.
  To address multicollinearity, Ridge and Lasso regression techniques were applied. Ridge uses L2 penalty, while Lasso uses L1 penalty and performs feature selection.

- **Machine Learning Model**
  Random Forest Regressor is an ensemble learning method that constructs multiple decision trees and averages their predictions. It captures non-linear relationships, handles interactions between variables, and provides feature importance.

- **Model Evaluation**
  The dataset was divided into training (80%) and testing (20%) samples. Model performance was evaluated using R^2 and RMSE. Random Forest achieved the highest predictive accuracy among the models.

- **Feature Importance Analysis**
  Feature importance analysis revealed adult mortality as the most influential predictor, followed by schooling, GDP, BMI, and alcohol consumption.

5. Findings

- ✓ results confirm that socio-economic and health factors significantly affect life expectancy. Classical regression models provide interpretability, while machine learning models offer superior accuracy.
- ✓ Random Forest achieved the highest predictive accuracy, indicating that life expectancy depends on complex, non-linear interactions among predictors.

- ✓ **Feature Importance Analysis**
  Feature importance analysis from the Random Forest model revealed:
  - Adult Mortality – Most influential
  - Schooling
  - GDP
  - BMI

- Alcohol consumption

This highlights the critical role of healthcare and education in improving life expectancy.

- ✓ Adult mortality and schooling are the most influential predictors.
- ✓ GDP and schooling show positive correlation, while adult mortality shows negative correlation with life expectancy.
- ✓ GDP and schooling show positive correlation with life expectancy.
- ✓ Adult mortality shows negative correlation.
- ✓ BMI and alcohol consumption show moderate association.

6. Conclusion

The project successfully predicted life expectancy using secondary data and advanced statistical techniques. Improving education, reducing adult mortality, and enhancing economic conditions can substantially increase life expectancy.

7. References

World Health Organization (WHO), Global Health Observatory.
James et al. (2013), An Introduction to Statistical Learning.
Hastie, Tibshirani & Friedman (2009), The Elements of Statistical Learning.
Kaggle Life Expectancy Dataset.