**Project Report: Analysis of Life Expectancy Factors**

**1. Introduction**

Life expectancy is a crucial indicator of a country's overall health and development. Various socio-economic and health-related factors influence life expectancy, making it essential to identify key determinants. This study employs **Ordinary Least Squares (OLS) Regression** and **Sequential Feature Selection (SFS)** to analyze the impact of different variables on life expectancy.

**2. Objective**

The primary objective of this study is to determine the most significant factors influencing life expectancy using statistical modeling techniques. The focus is on understanding the impact of **Adult Mortality, Infant Deaths, and Schooling** on life expectancy.

**3. Methodology**

**3.1 Dataset**

The dataset consists of multiple health and socio-economic indicators for different countries. The dependent variable is **Life Expectancy**, while the independent variables include factors such as mortality rates, education levels, and economic conditions.

**3.2 Data Preprocessing**

* **Null Value Treatment:** Checked for missing values and handled them appropriately.
* **Duplicates Removal:** Ensured no duplicate records were present.
* **Outlier Treatment:** Used IQR method to remove extreme values.
* **Encoding:** Converted categorical variables like 'Status' into numerical values.

**3.3 Statistical Techniques Used**

* **Train-Test Split:** Divided data into training and testing sets.
* **OLS Regression:** Identifies relationships between independent variables and life expectancy.
* **Resolving Multicollinearity:** Used Variance Inflation Factor (VIF) to eliminate highly correlated predictors.
* **Feature Selection:** Used **Sequential Feature Selection (SFS)** to retain only the most significant variables.

**4. Results and Discussion**

**4.1 OLS Regression Findings**

* **R-squared:** 0.866, indicating that **86.6% of the variation in life expectancy** is explained by the model.
* **Adjusted R-squared:** 0.732, suggesting room for improvement by refining the model.
* **Significant variables:**
  + **Measles (-0.0062, p = 0.039):** Higher measles rates negatively impact life expectancy.
  + **HIV (-10.6502, p = 0.055):** HIV prevalence tends to lower life expectancy significantly.
* **Non-significant variables:** Variables like **GDP, Polio, BMI, and Population** did not show a strong statistical relationship.

**4.2 Key Features Selected by SFS**

1. **Adult Mortality (-0.0246, p = 0.131):** Higher adult mortality decreases life expectancy.
2. **Infant Deaths (0.0001, p = 0.999):** While statistically insignificant in OLS, SFS suggests an important role in predicting life expectancy.
3. **Schooling (0.9569, p = 0.136):** Higher education levels positively correlate with life expectancy.

**4.3 Model Evaluation**

* **Linear Regression Performance:**
  + **Train R²:** 0.866
  + **Test R²:** 0.740
  + **Train MSE:** 4.97
  + **Test MSE:** 9.60
* **Ridge Regression:**
  + **Train R²:** 0.838
  + **Test R²:** 0.703
  + **Train MSE:** 5.99
  + **Test MSE:** 10.97
* **Conclusion:** Linear Regression performed the best, making it the final model.

**5. Conclusion and Recommendations**

**5.1 Conclusion**

* Life expectancy is significantly influenced by **mortality rates and education levels.**
* **Healthcare policies** should focus on reducing **adult and infant mortality** to improve life expectancy.
* **Investment in education** can lead to better health outcomes and increased life expectancy.

**5.2 Recommendations**

* Governments should **implement better healthcare infrastructure** to lower mortality rates.
* **Educational programs** should be enhanced to promote awareness of health and well-being.
* Further studies can **incorporate non-linear models** to capture complex relationships between variables.

**6. Future Scope**

* Exploring additional factors like **nutrition, environmental conditions, and lifestyle habits.**
* Applying **Machine Learning models** (Random Forest, Gradient Boosting) to improve predictive accuracy.
* Conducting a **country-wise comparison** to tailor policy recommendations more effectively.