Data Minning and Machine Learning
Project Proposal

“Average Life Expectancy Prediction Model”

Mahroz Abbas: BSDSF22M037

Hashir Nasir: BSDSF22M026

Abdul Raqeeb Khan: BSDSF22M038

1. Problem Statement

Life expectancy is a crucial indicator of a country's overall development, health infrastructure, and socioeconomic status. Accurately predicting life expectancy based on public health metrics can inform policy decisions and targeted interventions. However, with numerous contributing factors, selecting the most impactful variables becomes challenging. This project aims to develop a robust and interpretable machine learning model to predict life expectancy using a refined subset of features derived through Recursive Feature Elimination (RFE).

2. Objectives

- To build a machine learning regression model that predicts life expectancy.
- To implement Recursive Feature Elimination (RFE) for optimal feature selection.
- To evaluate the model performance using R^2 and Adjusted R^2 metrics.
- To deploy a user-interactive interface for life expectancy prediction based on selected health statistics.

3. Dataset Description

The dataset comprises global life expectancy and health indicators from multiple countries over several years. Features include infant mortality, under-five deaths, adult mortality, alcohol consumption, and more. After preprocessing, only numerical variables are retained. The final model uses 4 selected features with the highest correlation to life expectancy.

4. Proposed Methodology

- **Data Preprocessing**: Clean the dataset by handling missing values and focusing on numerical features.
- **Feature Selection**: Use RFE with a **Linear Regression model** to identify the top 4 features influencing life expectancy.
- **Model Training**: Split data into training and testing sets (80/20), scale features using StandardScaler, and train using Linear Regression.
- **Evaluation Metrics**: Measure performance using R^2 and Adjusted R^2 scores.
- **User Interface**: Integrate with a Streamlit-based frontend for interactive predictions using the selected features.

5. Expected Outcomes

- A trained model capable of predicting life expectancy with high accuracy ($R^2 > 0.85$).
- A reduced, interpretable feature set that offers actionable insights.
- A responsive and informative prediction dashboard for academic and policy use.
- Visualizations such as regression plots and a correlation matrix for model explainability.