Project Title:

Life Expectancy Prediction based on Health and Demographics Dataset.

Dataset Name:

Health And Demographics

Source: Kaggle

https://www.kaggle.com/datasets/uom190346a/health-and-demographics-dataset

Data Preprocessing:

- **Data Cleaning:**
  - Handled missing values
  - Renamed the variables for better readability
  - Created a new column ‘Continent’ using ‘countrycode’ package
  - Reordered the columns using ‘select’ function from the ‘dplyr’ package
  - Converted qualitative variables into factors

- **Exploratory Data Analysis (EDA):** Conducted EDA to understand the distribution of variables, identify trends, and explore relationships between variables.

- **Statistical Analysis:** Utilized statistical techniques such as correlation analysis and regression analysis to examine associations between health indicators and life expectancy.

- **Visualization:** Created visualizations such as scatterplots, histograms, and boxplots to illustrate key findings and insights.

Model Building:

- The initial regression model created has an R-squared value of 0.851 which explains about 85% of variability in predicting the life expectancy.

- Upon performing manual elimination based on p-test and F-test statistics, the final R-squared value is 0.8512 which is ideal to the initial model.

- The backward elimination and forward selection models also resulted in the same adjusted R-squared value.

- Overall, the model suggested that various factors such as health status, mortality rates, lifestyle factors (like alcohol consumption and BMI), immunization coverage, disease prevalence, and socioeconomic factors (like income composition and schooling) influence life expectancy.

Multicollinearity:

- Created a correlation matrix and a plot to identify the correlation between variables.
- Also, calculated the VIF of the model using vif(model) in order to identify the variables that are highly correlated in the model.
- The VIF of two variables are found to be >10 (average threshold value of correlation). These variables must be removed from the model since they depict a high correlation which might lead to potential outliers.

Residual Analysis:

- Calculated the residuals and identified the distribution by using a histogram which resulted in a normal distribution.
- The statistical normality tests are conducted using a QQ plot which projects the relationship between the residual and the fitted values which are found to be positively linear.

N- Fold Cross Validation:

- The train and test models were created in order to analyze the predicted average life expectancy for each row by comparing them with the actual values provided in the dataset.
- A 3-fold cross validation is then performed upon the models to verify the Mean Square Error of the train model.

Advanced Regression Analysis:

- Performed Ridge, LASSO and ElasticNet regression models to identify the number of variables used in the model building and observed the predictors' efficiency.
- Also, I have tried applying logistic regression on the dataset in order to predict the outcome of the average life expectancy to be 'high' or 'low' for every continent.