Project Proposal

Project team details: Shangfeng Huang, Yash Alpesh Rajpuriya

Title: Correlation Analysis of Life Expectancy

Problem statement. With the progress in global healthcare and social development, life expectancy is steadily on the rise. It serves not only as a crucial indicator of the health status of a country or region but also as a key marker reflecting social development and welfare conditions. Longevity implies extended periods during which individuals can contribute to societal and economic activities, while also indicating the potential for prolonged family life and community engagement. However, life expectancy is subject to various influences, including cultural practices and the region or country of residence. In this project, our objective is to investigate the primary factors influencing changes in life expectancy. These factors include, but are not limited to, government healthcare expenditure, government GDP, vaccination rates, body weight, alcohol consumption, educational attainment, and geographical region. By examining the correlation between life expectancy and these factors, our summarized analysis aims to serve as a reminder for governments and individuals to pay increased attention to their lifestyle habits and physical health.

Objectives. The database utilized in this project, comprising information on Life Expectancy and 18 potential influencing factors for the years 2000 and 2015, was compiled from 193 countries sourced from the World Health Organization (WHO)'s Global Health Observatory (GHO). Conducting Exploratory Data Analysis (EDA) on the dataset and visually analyzing the impact of various factors on life expectancy can yield valuable insights. Examining the results enables us to comprehend the extent to which each factor influences life expectancy. For instance, we may discover that government expenditure on healthcare significantly affects life expectancy. Furthermore, our objective is to investigate the combined effects of multiple factors on life expectancy, such as the impact of population size and healthcare expenditure, along with the influence of per capita income and per capita healthcare expenditure on life expectancy. Finally, based on the provided data, we will develop a predictive model to estimate life expectancy values based on various inputs, including healthcare expenditure. This model serves as a tool for governments to monitor the life expectancy levels of their population and reminds individuals to pay attention to their personal health and well-being.

Possible solutions.

EDA: (not alternative methods)

a. Boxplot: Boxplots are utilized for visualizing data and identifying outliers. These outliers may arise from noise or possibly due to an insufficient sample size. To address this, we replace outliers with the upper or lower limit values of the boxplot accordingly.

b. Histogram: Histograms are employed to visualize the distribution of data for each influencing factor.
c. Line Plot: Line plots are utilized to observe the changes in each influencing factor from 2000 to 2015.
d. Heatmap: A heatmap is employed to illustrate the correlation between various influencing factors.
e. GeoMap: World maps are utilized to offer a more intuitive representation of the variations in life expectancy among different countries or regions.

Prediction Module:

Solution 1 (Best model): Fitting the data involves employing various classifiers or regressors, and selecting the optimal model based on performance metrics or cross-validation techniques. The models intended for use in this project include Decision Tree (DB tree), Support Vector Machine (SVM), Linear Regression, Random Forest Regressor, Extra Trees Regressor, Gradient Boosting Regressor, and Hist Gradient Boosting Regressor. The selection process will be based on their performance and suitability for the given dataset.

Solution 2 (Bagging): Improving accuracy involves aggregating predictions from the classifiers through an averaging mechanism.

Solution 3 (Boosting): We will employ a weighted voting strategy with the classifiers, wherein each classifier contributes to the final prediction based on its performance and relevance. This approach combines the strengths of multiple models for more accurate predictions.

Solution 4 (Deep Learning): While we acknowledge the effectiveness of existing solutions for this task and have reviewed relevant literature exploring them, we have not found methods specifically applying deep learning to this dataset. Consequently, we prefer to employ deep learning-based approaches for our analysis.

Current project progress

Having completed the Exploratory Data Analysis (EDA), we are now in search of a lightweight and suitable deep learning model. Here are some observational findings, presented from left to right: Life expectancy distribution, healthcare expenditure distribution, Boxplot (outliers removed), heatmap of correlation, and trends of factors changing over time.