Investigating the Relationship Between Socioeconomic Factors and Life Expectancy: A Cross-Country Analysis

Prabesh N. Kunwar
Dept. of Computer Science
University of Prince Edward Island
Charlottetown, Canada
pnkunwar@upei.ca, prabeshkunwar12@gmail.com

***Abstract*—Abstract-It has always been a concern for all of us about a healthy lifestyle and we are always willing to know about the factors that help us remain healthy. And one of the major health issues in our concern is related to heart. So, for our study, we take different factors like age, sex, Blood Pressure and more into consideration and find the actual cause of heart disease using different machine-learning techniques. This will help us find the factors that we can improve to keep our heart healthy.**

***Index Terms*—heart disease, cardiac disease, lifestyle, health**

I. Introduction

Life expectancy of the country reflects the quality of life within that country. It is also used in determining Human Development Index and Happiness Index. So, the national administration needs to maintain and improve the life expectancy of people.

We can see the governments of various countries decided to spend a certain percentage of their total budget on health. We can also see various immunization campaigns launched to prevent life-threatening diseases. We also see people concerned about their bodies and indexes such as BMI and body fat percentage. We see the government increasing health facilities to decrease Mortality Rates. But how effective are these methods in improving the life expectancy of the whole country?

We must remember that all these values are average for a country, so this data set isn't for individuals to estimate their life expectancy, although it might work sometimes. Although this data was taken from WHO, there might be mistakes in data entry. Some columns have practically impossible values, such as an average BMI higher than 35 and percentage values higher than 100. So, those data were removed before the analysis.

We used a scatter graph for the initial data view and analysis. We used regression techniques to train the Machine Learning model as the target value was continuous. We used Polynomial Regression for further data analysis and then to train the ML model, which was pretty fast and accurate. Then we used Random Forest Regressor to train the more accurate model.

This report will reflect the introduction to the dataset and explain the machine learning models in detail. It contains an analysis section where a scatter graph is used to explain the relation of different attributes with life expectancy, and then ML model training is explained in detail.

II. Background

A. About the data-set

This data-set [1] is taken from The Global Health Observatory(GHO) under World Health Organization(WHO). This data set contains data related to life expectancy from 193 different countries from 2000-2015. The data set also contains missing data related to vaccination and GDP from the least known countries. The Final Merged File consists of 22 Columns and 2938 rows. All predicting variables were divided into broad categories: Immunization Related Factors, Mortality Factors, Economic Factors, and Social Factors.

All the factors are explained below:

Immunization Related Factors:

- HepB: Hepatitis B immunization Coverage among 1-year-olds in percentage.
- Pol3: Polio immunization Coverage among 1-year-olds in percentage
- DTP3: Diphtheria tetanus toxoid and pertussis immunization Coverage among 1-year-olds in percentage

Mortality Factors:

- Adult Mortality: Adult Mortality Rates of both sexes (probability of dying between 15 and 60 years per 1000 population)
- Infant Deaths: Number of Infant Deaths per 1000 population
- Under 5 Deaths: Number of under-five deaths per 1000 population
- HIV/AIDS: Deaths per 1 000 live births HIV/AIDS (0-4 years)

Health and Economical Factors

- Percentage Expenditure: Expenditure on health as a percentage of Gross Domestic Product per capita(
- Total Expenditure: General government expenditure on health as a percentage of total government expenditure (

- GDP: Gross Domestic Product per capita (in USD)
- Thinness 1-19 years: Prevalence of thinness among children and adolescents for Age 10 to 19 (
- Thinness 5-9 years: Prevalence of thinness among children for Ages 5 to 9(
- Income Composition of Resources: Human Development Index in terms of income composition of resources (index ranging from 0 to 1)
- Measles: Measles - number of reported cases per 1000 population
- Alcohol: Alcohol, recorded per capita (15+) consumption (in liters of pure alcohol)
- BMI: Average Body Mass Index of the entire population

Social Factors

- Year:
- Status: Developing or Developed Status
- Population: Population of the country
- Schooling: Number of years of Schooling(years)

target variable:

- Life Expectancy: life expectancy in age

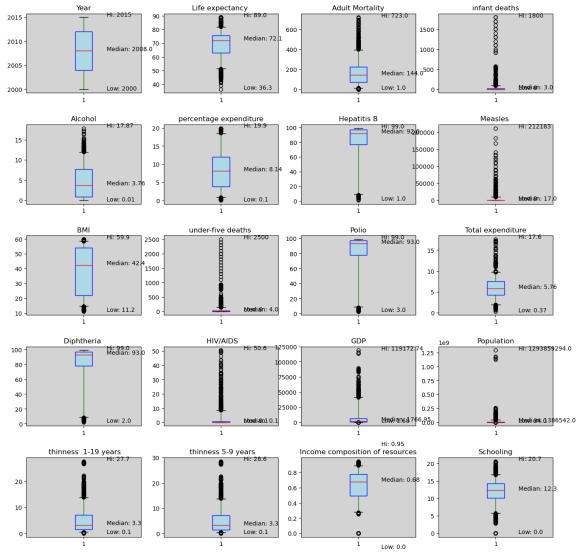

Fig. 1. Fig. 1. Whisker-Box plot diagram of all the attributes with numerical value

B. Machine Learning Technique Used

- Polynomial Regression

  Our target value is continuous, so polynomial regression is the first choice. In addition, our determining values are also mostly continuous.

  As shown in the figure 2, there are two types of variable. Independent variables (x), which are not dependent upon each other, and dependent variable (y), whose value we have to find with the help of x.

  The simplest form of polynomial regression is linear regression. It only has one independent variable and has a simple linear equation. If we add multiple independent variables, it becomes multiple linear regression. The independent variables should not be related to each other. If the straight line cannot define the relation, we increase the degree of x. We should be careful not to over-fit or under-fit the data. [3]

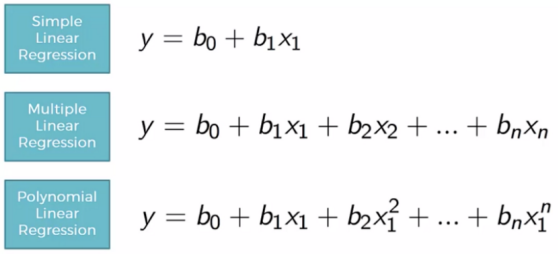

Fig. 2. Polynomial regression

To define how well the model fits the data, there are two methods, RMSE (Root Mean Squared Error) and R-squared(R^2). RMSE shows how far are the predicted values from observed value. Lower the RMSE, better the model. It is calculated as: [4]

$$RMSE = \sqrt{\frac{\sum(P_i - O_i)^2}{n}}$$

where:

- sum is a sum symbol
- P_i is the predicted value for the i^th observation
- O_i is the observed value for the i^th observation
- n is the sample size

Whereas r^2 defines the proportion of variance in the response variable of a model that predictor variables can explain. The higher the r^2 value, the better the model, and its value lies between 0 and 1. It is calculated as: [4]

$$R^2 = 1 - \frac{RSS}{TSS}$$

where:

- RSS represents the sum of squares of residuals
- TSS represents the total sum of squares

- Random Forest Regressor

  Random Forest Classifier is the improved version of Decision Trees. We train data to have multiple Decision trees with random features and random samples selected. This will allow the algorithm to accept the new samples and not over-fit in the test samples.

  RFR is very simple and very easy to train. It uses the Bagging technique (Bootstrapping and aggregation). It bootstraps all the samples and data samples and selects them randomly to form a given number of trees. It uses all those trees to predict results independently and aggregates all the results from all the trees, and uses them to form the result.

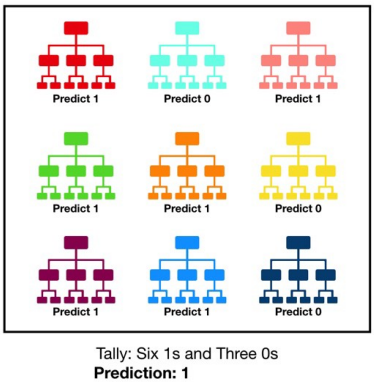

Fig. 3. Random Forest Regressor

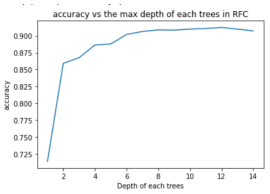

Fig. 4. Accuracy vs. Maximum depth of each tree

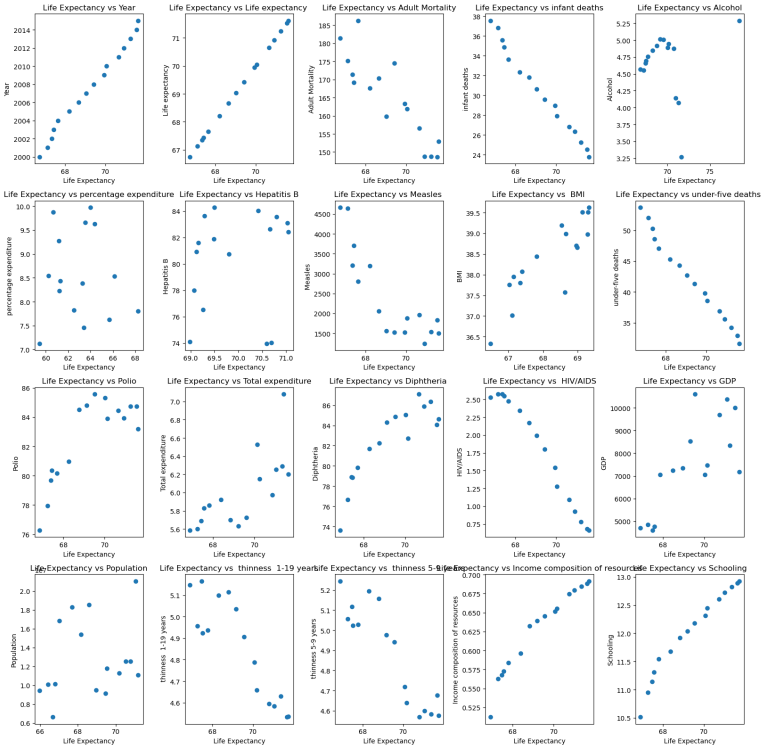

Fig. 5. Scatter graph showing the relation between Life Expectancy vs. rest of the attributes(grouped values concerning the year and averaged and Life Expectancy)

As this technique uses multiple trees to predict the result, predicting takes longer than training. Memory efficiency depends upon the number and depth of trees. But after the significant increase of the number of trees and depth of trees, the increase in accuracy slows down. So, we can tune the algorithm by tuning those hyper-parameters. Other hyper-parameters include the number of features in a tree which can be set either to square root or the log2 of the total number of features as a standard practice.

III. Analysis and Relults

A. Relationship Between all the attributes (grouped values with respect to year and averaged and Life Expectancy)

As most of the data is continuous and our target data itself is continuous, we should use a line or scatter graph. We took the values of each column with numerical values and grouped them by year. Then we averaged the values concerning the year. Then we plot a scatter graph, Life Expectancy vs. rest of the attributes as shown in figure 5.

There is a clear relation between life expectancy and year due to the increase in health facilities and reach to it. If we consider the mortality factors (adult mortality, infant deaths, under-five deaths, and HIV/AIDS), there is an inverse relation, as expected. Although adult mortality (deaths under 60) doesn't show a strong relation, the rest of the mortality factors significantly decrease life expectancy.

If we look at the immunization factors, we can find the increase in life expectancy with the increase in immunization, too, except for some outliers. This data also makes sense, as immunization prevents deaths from severe diseases.

There are some health-related attributes too. Measles have an inverse relation with life expectancy. BMI positively relates to LE, although medically high BMI (obesity) is also a leading cause of death. But the BMI data didn't seem accurate because the lowest was 1 for the entire country, and the highest was 87. Moreover, a BMI of over 35 is considered obese, except for those incredibly muscular people. BMI 1 means that the person is only a few kilos, whereas BMI 87 means weighing hundreds of kg, even for a person of 5 ft. So, we filtered out BMI less than ten and higher than 60 to make it more realistic. Both thinness attributes show an inverse relation to LE.

Most of the economic factors don't show strong relationships with LE. There is no strong relation between GDP with LE, and percentage expenditure shows no relation. Percentage expenditure had many data that were not valid. As per statistica [2], the US is the highest spender in health by percentage, almost 20 percent. So, values greater than 0 and less than or equal to 20 are considered. Total expenditure also has weak positive relation, whereas income composition has a strong relation with LE.

Talking about social factors, the population doesn't seem to have any relation with LE. In contrast, there is strong positive relation with schooling which seems reasonable as education helps people to live a healthy and long life.

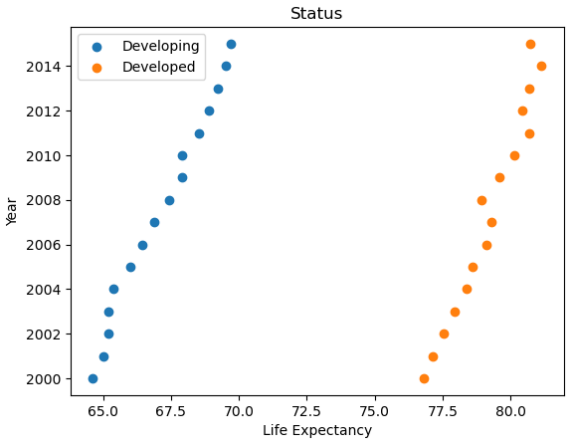

Fig. 6. Increase in LE in developed vs. developing countries

B. Applying Polynomial Regression

Regression techniques only accept continuous values, so we changed the status of columns to 0 and 1 for 'Developing' and 'Developed' countries, respectively. We didn't consider the 'Country' column. We used masks in variables like BMI, Total Expenditure, and Percentage Expenditure to validate values. We also used the KFold method with k=5 for cross-validation. For the first use of multiple linear regression to train the model with all the variables with numerical values. We got an average RMSE of 4.75 and an average R^2 of 0.53. These values showed that this model wasn't that accurate in predicting the accuracy.

Then we removed all the outliers (all the values out of inter quantile range). Then trained model with individual attributes separately, every three times for the first, second, and third-degree respectively. We calculated the average RMSE and R^2 for each model and found the following attributes to have the strongest relation with LE:

- Schooling
- Adult Mortality
- Income Composition of Resources
- BMI
- HIV/AIDS

Some features such as GDP and percentage expenditure had relatively high R^2 but also high RMSE and vice versa, Measles and Population. The RMSE and R^2 for the second degree were significantly better than that of the first, but the third degree didn't improve much. So, the second degree was considered optimal.

So, we only took the variables having the strongest relation with LE to train the model. We trained three models with degrees one, two, and three, respectively, and calculated the average. Using all the filters and masks, we had 1638 data sets remaining. These models performed better than the previous model. The average RMSE was 2.79, 2.34, and 2.28; the average R^2 was 0.81, 0.86, and 0.87 for the first, second, and third degree. The

C. Applying Random Forest Regressor

For this model, too, we removed outliers using InterQuantile Range, removed data with null values and masked the data set with valid values. We only have 84 remaining data after this process. We first trained the model using all the columns with numerical values and default hyperparameters. After using KFold with k=5, we got an average RMSE of 2.61 and R^2 of 0.70, which is better than the linear regression model, and also we only had 84 total data sets.

Then we created a list of 'the number of trees' with values 100, 150, 200, 250, 300, 350, and 400 and a list of 'max depth' with values 5, 6, 7, 8, 9, and 10. Then we ran a nested loop to iterate the values of these lists for hyperparameter tuning. We trained the forty-two RFR models. Taking RMSE, R^2, time is taken, and over and under fitting into consideration, we found a model with 250 trees and a max depth of 9 to be the optimum. It had an RMSE of 2.5 and R^2 of 0.73, which was still worse than the best polynomial regressor we trained before.

Then, we trained the models by using five variables that performed well in polynomial regression. We only removed the data sets with null values for the five variables, life expectancy and masked BMI values. We had 1638 remaining values. Then we again used the same hyperparameter tuning method to get 42 models. The lowest average RMSE was 0.87, which was still better than the previous model. Taking RMSE, R^2, time taken to train, and over and under fitting into consideration, we found a model with 200 trees and a max depth of 10 to be the optimum. It had an average RMSE of 1.85 and an average R^2 of 0.92.

We took the optimum hyperparameter to train the model again using five variables, which can be considered the best model we trained using RFF and the given datasets.

IV. Conclusion

Scatter graph explained the relationship between different factors and LE. However, we created it by grouping the values according to the year, taking the mean, and visualizing the data, but we found the best indicators using Polynomial Regression.

Before the analysis, we thought increasing government expenditure on health would increase Life expectancy. But expenditure only had R^2 of 0.00028 whereas schooling had 0.61. This means not only educated people live longer, but they also create a healthy society. They share the information regarding health to the remaining population which ultimately results in high LE. So, the education sector should be considered more than investing in the health sector.

Similarly, even immunization was also found to be an important indicator. Polio, Hepatitis and DPT had R^2 of 0.32, 0.17 and 0.31. So, people should be encouraged to be vaccinated.

Mortality factors Adult Mortality, Infant Mortality, Under Five Deaths, and HIV deaths had R^2 of 0.48, 0.29, 0.26, and 0.33 resp. This also shows that mortality factors also are very important indicators of LE. The obvious solution to this problem is to increase health services. However, education

could be the better solution to decrease the mortality factors. The health factors such as BMI, thinness(5-9) and thinness(1-19) had R^2 of 0.50, 0.37 and 0.36 resp. From the graph, we can see that more BMI more LE. It doesn't mean that obese people live longer. What it means is in the county where people have enough to eat and have healthy bodies live longer. Thinness in children is mostly caused due to malnutrition, which leads to lower LE.

Income composition of resources had the highest R^2 of 0.76. It is the HDI of people with respect to the ability of a nation to use its resources.

For the ML models, we got two models: Polynomial regression and RFR. Polynomial Regression is fast to train and also fast to predict. It only takes 0.2 sec to train and test the 3 PR models with KFold k value 5. Conversely, it takes 5.5 sec to train and predict 1 model with k=5. But the R^2 increases from 0.87 to 0.92, and RMSE decreases from 2.28 to 1.85.

REFERENCES

[1] https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who: Life Expectancy (WHO): KUMARRAJARSHI

[2] https://www.statista.com/statistics/236541/per-capita-health-expenditure-by-country/: :text=health

[3] https://www.analyticsvidhya.com/blog/2021/10/understanding-polynomial-regression-model/: :text=Polynomial

[4] https://www.statology.org/rmse-vs-r-squared/: :text=Both