Factors influencing life expectancy

Raj Aryan Singh

In the past, we have seen that the effect of immunization was not included in the studies that determine the factors affecting life expectancy rates considering various factors such as composition of income, demographic variables and even the mortality rates. Thus, the goal of this project is to focus on economic, mortality, social immunization, and various factors to predict life expectancy, which will help suggesting countries which area should be given importance in order to efficiently improve the life expectancy of its population.

The data used for this project was collected from WHO and United Nations websites and can be found on Kaggle: https://www.kaggle.com/kumarajarshi/life-expectancy-who. This dataset contains 22 columns that include Country, status(developing/developed) and 19 other features like GDP, alcohol consumption, immunization coverage for diseases, mortality rates, education factors and more.

For analysis I selected all of the 19 features to evaluate their effects on life expectancy, 1) year, 2) adult mortality, 3) infant deaths, 4) alcohol consumption, 5) expenditure on health as a percentage of GDP, 6) HepB immunization coverage, 7) measles reported cases (per 1000 population), 8) BMI, 9) under-five deaths, 10) Polio immunization coverage, 11) total expenditure on health, 12) DTP3 immunization coverage, 13) Deaths per 1000 live births HIV/AIDS, 14) GDP, 15) population, 16) thinness 1-19 years, 17) thinness 5-9 years, 18) income composition of resources, and 19) no. of years of schooling.

For Figure 1**,** I compared three features 1) no. of years of schooling, 2) Human Development Index in terms of income composition of resources, and 3) GDP to show how they affect life expectancy in a country, we can see that countries with higher education rate, higher-income composition and GDP have a greater life expectancy. Other features like immunization coverage also show a similar trend, however, other features like alcohol consumption, mortality rate, and disease cases represent a negative trend i.e. if they are high, life expectancy is lower.

For Figure 2**,** I used a principal component analysis plot of the transformed data. I applied standard scaler on every numerical column to standardize the values, as they were not comparable unscaled with a very high explained variance ratio of about 1.0. After, scaling it showed an explained variance of about 0.39.

In order to predict the life expectancy, I created a sklearn PipeLine with StandardScalar and LinearRegression, and used all 19 features to predict it. I performed a train/test split on the data, then fitted the model on the training data. The model achieved a r2 score of 82.5%, score mean of 80.9% and variance of 0.00068, making it a perfect accurate model for what we plan to achieve from it.

Figure 3 plots the coefficient weights for each of the features used by the model and infers that the major factor for predicting life expectancy is mortality rate(under-five deaths, infant deaths), schooling, and income composition. We can also infer that prevalence of thinness among children and adolescents, diseases like HIV/AIDS have a negative correlation with life expectancy.

In conclusion, we can say that all of the features given in the dataset are important and the model can be helpful in order to accurately predict and work on efficiently improve the life expectancy of a country's population by determining these factors.

Figure 1: Comparison of Schooling, Income Composition and GDP's effect on life expectancy

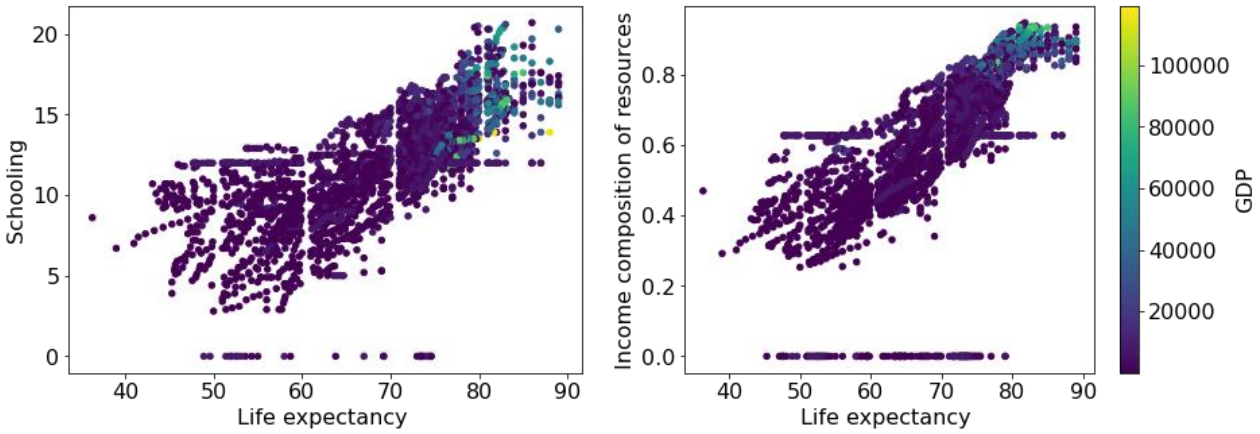

Figure 2: Principal Component Analysis

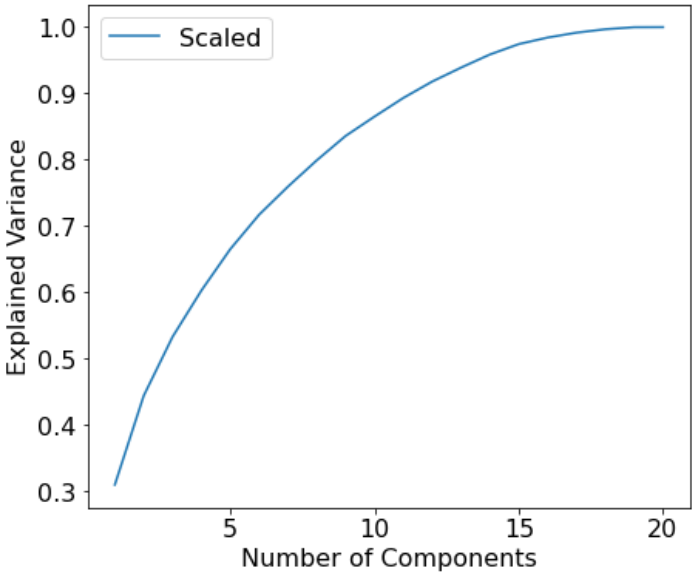

Figure 3: Linear regression Coefficients