Statistical Analysis on Factors Influencing Life Expectancy

1. Background

Typically, when we look at the factors that contribute to life expectancy, we take into account factors such as population, income, and mortality. However, few predictions and analyses have been conducted for each country; Instead, they were implemented throughout the whole world. Also, there are many other factors such as vaccination that could be a very important factor. Involving those factors allows a country's government to know where to focus to improve life expectancy.

2. Objectives

We would like to conduct multiple linear or non-linear analyses of a large number of factors for different countries to identify factors that contribute to low expected life expectancy. Since the situation for countries might be significantly different, this will help to suggest which areas a specific country should focus on in order to effectively increase the life expectancy of its population. In this project, we will use a combined dataset to explore various factors including the health system, population, social effect, macroeconomic index, etc.

3. Introduction to our dataset

The dataset we use concentrates on immunization factors, mortality factors, economic factors, social factors, and other health-related factors. The dataset related to life expectancy, health factors for 193 countries have been collected from the same WHO data repository website and its corresponding economic data was collected from the United Nations website. Among all categories of health-related factors, only those critical factors were chosen which are more representative. In this project, we have considered data from years 2000-2015 for 193 countries for further analysis. The final dataset consists of 22 Columns and 2938 rows which means 20 predicting variables. All predicting variables were then divided into several broad categories: Immunization related factors, Mortality factors, Economical factors, and Social factors.

4. Data Pre-Processing

Firstly, we used the **describe()** method to generate the basic statistics about the features in the dataset, such as mean, median, maximum, and minimum. We concluded the following findings that could be the possible steps for our data pre-processing(Table 4.1)

1	The minimum value of "Adult Mortality" is 1, which might not be correct since the mortality rate of 1 out of 1000 people is unrealistic.
2	The minimum value of "Infant deaths" is 0. This might be an outlier. Also, the maximum value of "Infant deaths" is 1800 which is much higher than the 75% percentile value(22).
3	"BMI" also has some possible outliers: as low as 1 and as high as 87. It is unrealistic for a person to have such an extreme BMI.

4	Minimum value of “GDP” is 1.68, which might be an outlier.
5	Minimum value of “population” is 34. This might be unrealistic for a country.

Table 4.1

The best way to correctly identify the outliers in a set of variables is a boxplot. We made several boxplots to check for the obvious outliers. Concluding from the boxplots, we decided to replace the data with the following conditions with NULLs. The boxplots for the five variables are as follows(Figure 4.2). Concluding from the boxplots, we decided to replace the data with the following conditions with NULLs(Table 4.3).

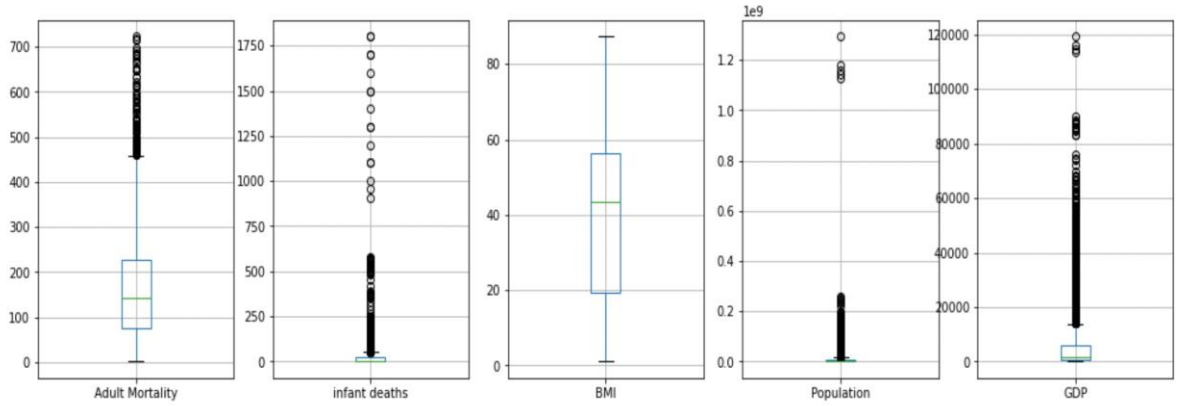

Figure 4.2

We can see from the above boxplots. “Infant Deaths” has extremely small values that equal to zero. “Adult Mortality” has extremely small values that are below 1. “BMI” has extremely small values below 10 and extremely large values above 80. We decided to remove those outliers and impute with NULL values.

Then, we check the missing values for each feature in the dataset. Clearly, we should have multiple missing values, so we firstly generate a statistical table that gives us a general look about the missing values.

Feature Name	Number of Missing Values	Percentage of Missing Values
Life expectancy	10	0.34%
Adult Mortality	22	0.75%
Infant Deaths	848	28.86%
Alcohol	194	6.6%
Hepatitis	553	18.82%
BMI	523	17.8%
Polio	19	0.65%
Total expenditure	226	7.69%
Diphtheria	19	0.65%
GDP	448	15.25%

Population	652	22.19%
Thinness 1-19 years	34	1.16%
Thinness 5-9 years	34	1.16%
Income composition of resources	167	5.68%
Schooling	163	5.55%

Table 4.3

As a whole, 15 features contain null values. From the table above, we can see that there are numbers of null values and for each feature, and the percentage of null values is not too high (less than 30%). According to this threshold, we do not remove any columns of data; instead, we decided to impute the missing values with the mean of all the other values for the features.

5. Data Visualization

We initially generated the below correlation heatmap(Figure 5.1) and tried to identify the potential collinearity among features. We found that "under-five deaths" and "Infant deaths" have a strong positive correlation of 0.98. Also, "Percentage Expenditure" and "GDP" have a strong positive correlation of 0.89. “Population”, “Infant Death” and “Under-five death” have strong correlations among each other. The strong correlation among features could potentially imply that our model might have collinearity, which would underestimate the significance of a variable.

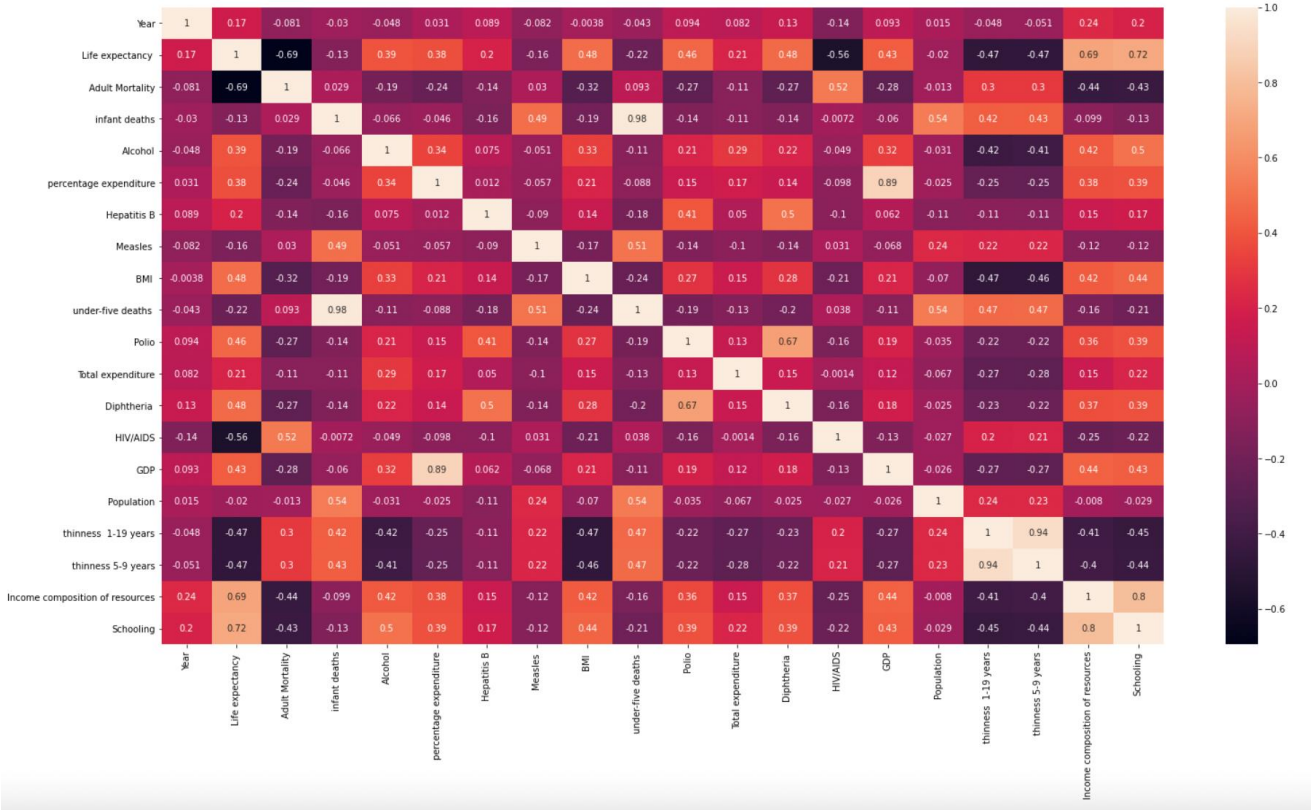

Figure 5.1

Moreover, we calculated the relationship for each feature with target variable "Life expectancy" stored coefficients, P-value, level of relationship and significance into the table below. We would like to check for the P-value to see if we can drop some of the insignificant features which also have

collinearity. From the table(Table 5.2) below, we can conclude that "Population" is an insignificant feature since its P-value is greater than 0.05. Also, for the two pairs of collinear features, we decided to drop the ones which have weaker correlations with the target variable. Therefore, we decided to drop "Population", "Percentage Expenditure", and "Infant deaths" to form a sparse model.

"Status" is a categorical variable with values of "developed" and "developing" of a country. We decided to encode the text data with numerical data: 1 for developing countries and 2 for developed countries.

	Features	Coefficient	P-value	Relation with Target Variable	Significance
0	Life expectancy	1.000000	0.000000e+00	Strong Positive	Significant
1	Adult Mortality	-0.694875	0.000000e+00	Strong Negative	Significant
2	infant deaths	-0.131308	8.981158e-13	Weak Negative	Significant
3	Alcohol	0.391598	2.781464e-108	Weak Positive	Significant
4	percentage expenditure	0.381791	1.384449e-102	Weak Positive	Significant
5	Hepatitis B	0.203771	6.570486e-29	Weak Positive	Significant
6	Measles	-0.157574	8.613245e-18	Weak Negative	Significant
7	BMI	0.480402	1.597173e-169	Weak Positive	Significant
8	under-five deaths	-0.222503	2.795862e-34	Weak Negative	Significant
9	Polio	0.461574	5.667050e-155	Weak Positive	Significant
10	Total expenditure	0.207981	4.518444e-30	Weak Positive	Significant
11	Diphtheria	0.475418	1.391634e-165	Weak Positive	Significant
12	HIV/AIDS	-0.556457	1.518471e-238	Strong Negative	Significant
13	GDP	0.430493	7.229205e-133	Weak Positive	Significant
14	Population	-0.019638	2.872932e-01	Weak Negative	Insignificant
15	thinness 1-19 years	-0.472162	4.822782e-163	Weak Negative	Significant
16	thinness 5-9 years	-0.466629	8.623362e-159	Weak Negative	Significant
17	Income composition of resources	0.692483	0.000000e+00	Strong Positive	Significant
18	Schooling	0.715066	0.000000e+00	Strong Positive	Significant

Table 5.2

We also generate histograms to explore the basic statistical features of data. From the figure below, we can see that many features like infant deaths, percentage expenditure and measles, etc, do not vary much in the whole dataset and remain at a low level. Besides, some features like adult mortality and thinness concentrate on a certain level. It remains to be seen whether it has a specific relationship to life expectancy.

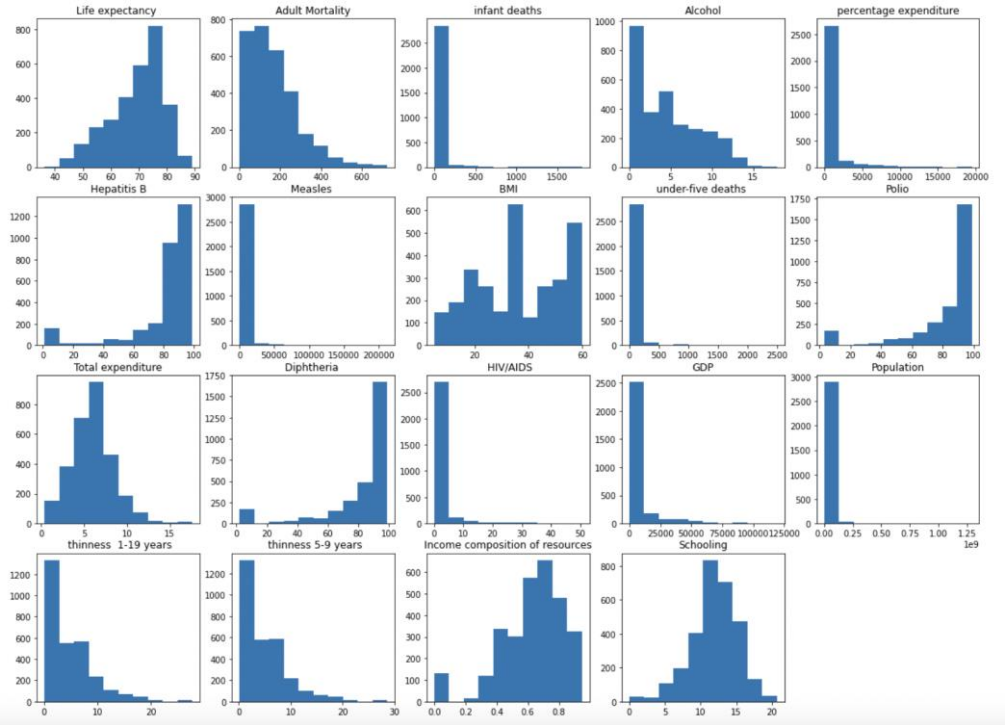

Figure 5.3

6. Models and Methodology

6.1 Linear Regression

Firstly, we split our dataset into a training set and a testing set. We start our model with simple linear regression and we get the following summary of the model.

Dep. Variable:	Life expectancy	**R-squared (uncentered):**	0.996
Model:	OLS	**Adj. R-squared (uncentered):**	0.996
Method:	Least Squares	**F-statistic:**	3.399e+04
Date:	Sun, 05 Dec 2021	**Prob (F-statistic):**	0.00
Time:	13:47:48	**Log-Likelihood:**	-5844.6
No. Observations:	2056	**AIC:**	1.172e+04
Df Residuals:	2039	**BIC:**	1.182e+04
Df Model:	17		
Covariance Type:	nonrobust		

Figure 6.1.1

	coef	std err	t	P>|t|	[0.025	0.975]
Year	0.0264	0.000	69.877	0.000	0.026	0.027
Status	1.6229	0.329	4.938	0.000	0.978	2.268
Adult Mortality	-0.0206	0.001	-20.901	0.000	-0.023	-0.019
Alcohol	0.0194	0.032	0.609	0.543	-0.043	0.082
Hepatitis B	-0.0197	0.005	-4.117	0.000	-0.029	-0.010
Measles	-3.469e-05	9.2e-06	-3.769	0.000	-5.27e-05	-1.66e-05
BMI	0.0340	0.007	4.756	0.000	0.020	0.048
under-five deaths	-0.0009	0.001	-1.122	0.262	-0.002	0.001
Polio	0.0265	0.006	4.778	0.000	0.016	0.037
Total expenditure	0.0602	0.041	1.465	0.143	-0.020	0.141
Diphtheria	0.0448	0.006	7.706	0.000	0.033	0.056
HIV/AIDS	-0.4907	0.021	-23.464	0.000	-0.532	-0.450
GDP	4.369e-05	8.24e-06	5.301	0.000	2.75e-05	5.99e-05
thinness 1-19 years	-0.0880	0.060	-1.465	0.143	-0.206	0.030
thinness 5-9 years	0.0053	0.059	0.089	0.929	-0.111	0.121
Income composition of resources	6.8281	0.798	8.560	0.000	5.264	8.392
Schooling	0.7177	0.052	13.763	0.000	0.615	0.820

Figure 6.1.2

The linear model fits the training data well, but after we fit the model with testing data, the test MSE is much higher than the training MSE. This is an overfit and it implies that our true model is not linear. We will try to use Lasso regularization to penalize the complexity of the model since we include too many features.

6.2 Lasso Regression

Lasso regularization is the technique we can use to penalize the model with too many features, which results in potential overfitting. Also, since Lasso can force some coefficients to be zero, we can get a sparse model based on that, which is better for us to interpret the results.

One of the tuning parameters in Lasso regularization is λ, which is the constraint parameter. As we decrease the value of λ, the model will resemble linear regression. The objective function of Lasso regularization is shown below.

$$minimize \sum_{i=1}^{n} (y_i - w^T x_i)^2 + \lambda \sum_{i=1}^{n} |w_i|$$

To find the best value of λ, we created a grid below for us to search for the best value using 3-fold cross validation.

$\lambda \in$ {0.1, 0.2, 0.3, 0.4, 0.5, 1, 2, 3, 4, 5, 10, 50, 100, 200, 300, 400, 500, 600, 700, 800, 1000}

Below is the line chart(Figure 6.2.1) of the cross validation score vs alpha. We can see that the score decreases as the value of alpha increases.

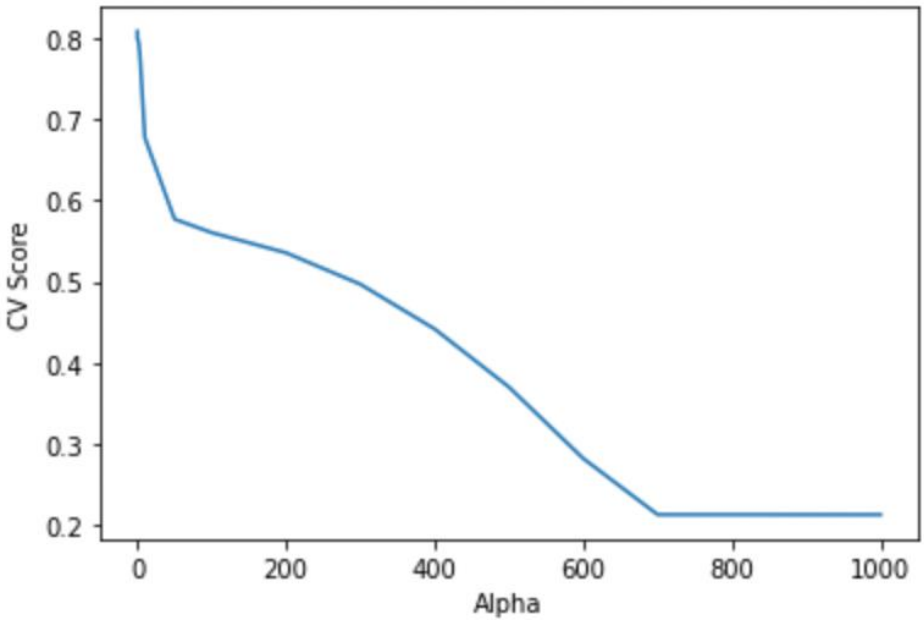

Figure 6.2.1

Below is the bar plot of the coefficients we got from our Lasso model(6.2.2).

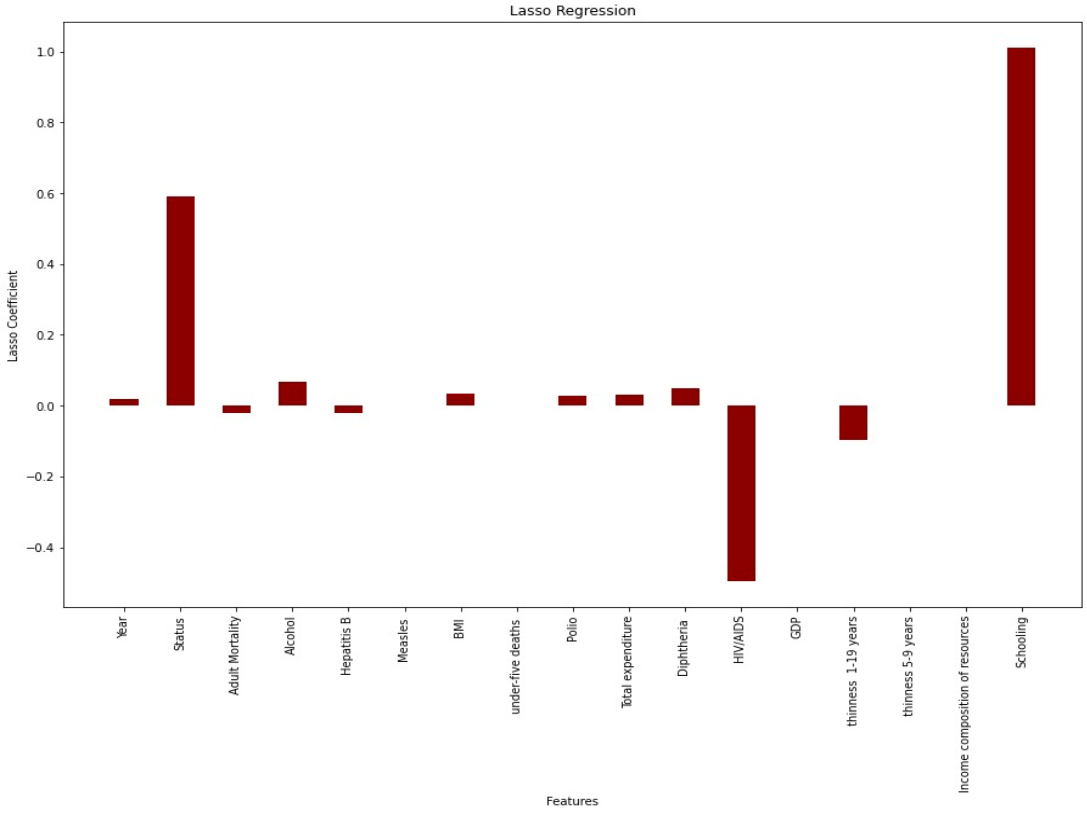

Figure 6.2.2

We can see that “Measles”, “under-five deaths” and “thinness between 5-9 years” have zero coeficient. “Status” and “schooling” have a significant positive influence on the expected life expectancy. This is reasonable since a more developed country will result in a longer life expectancy. Also, people with higher education levels would have better-living quality. Thus, they could have a longer life expectancy. “HIV/AIDS” has a negative influence on life expectancy, meaning a higher death rate of HIV/AIDS results in shorter life expectancy.

6.3 Random Forest

We also fit a random forest regressor by starting with a baseline model, with a number of estimator 500, and random state equals 42. Then we created the following set of grids to randomly choose different combinations and use 3-fold cross-validation to evaluate. Then we got our best parameter based on the search. And our final test MSE is 2.976, which improved by about 6%.

We make a general procedure below to show our steps to improve as below(Figure 6.3.1). As we think the number of estimators, the max features extracted each time, and the maximum depth are three important parameters that could impact the results, we generate a dictionary of the random grid. In each iteration of the search, we randomly choose the combination of these parameters and apply 3-fold cross-validation to evaluate.

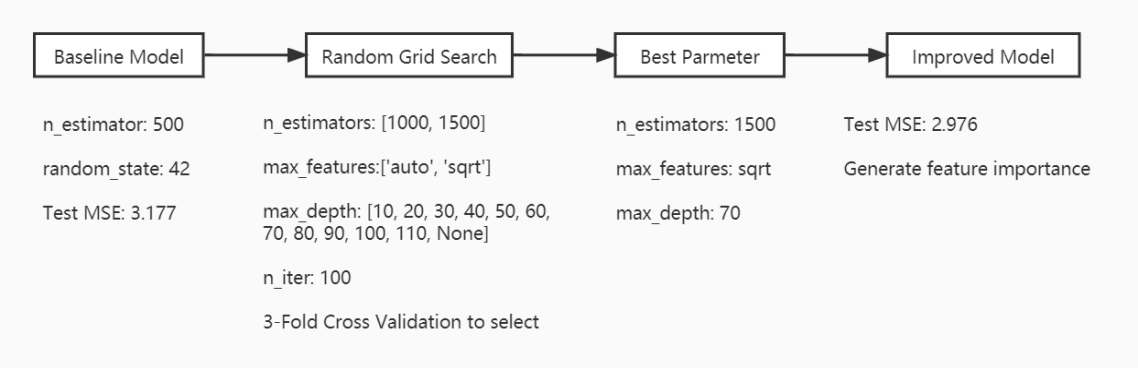

Figure 6.3.1

And we also generate a plot for feature importance to better for us to interpret as below(6.3.2).

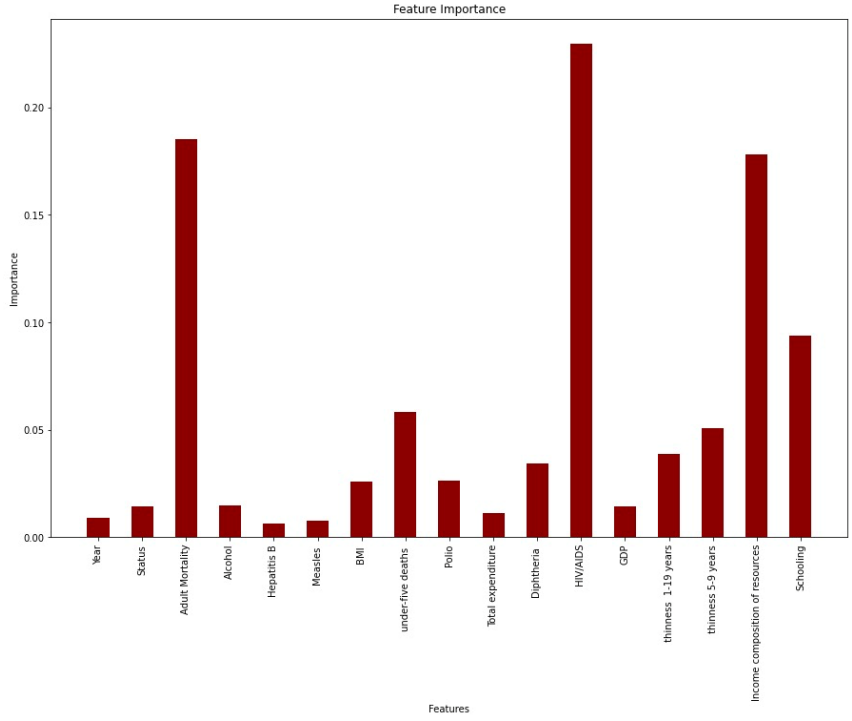

Figure 6.3.2

From this plot of feature importance, we can see that Adult Mortality, HIV/AIDS death rate, Income composition of resources and Schooling are the most important factors.

7. Actions Summary

According to the model we conducted, Status(whether a country is in developing or developed), adult mortality rate, HIV death rate, income composition of resources play important roles in affecting life expectancy. Thus, we can say that the government should make efforts on improving the whole people's quality including being knowledgeable and having a decent standard of living. To be more specific, the health system should be improved and people should have more chances to get a higher and longer education. Government should also advocate people the importance of body weight management and AIDS prevention.

8. Fairness Analysis

In order to find the general model of predicting life expectancy and exclude the bias of different countries, we decide to use the unawareness to just exclude the country column. Though there may exist proxies that may correlate with countries, we think that it will still result in a relatively subjective and general model of the question we concentrate on.

9. Direction for Future Work

For the direction of the future that could improve the model, we may collate in the following way. Since we did not take into account different countries into the model, every country was fit with the same model. Ideally, we should not do this since different countries would have different situations. Before our modelling, we would better use unsupervised learning such as K-means to cluster similar countries into the same group. (WU, 2011)

Also, we used the mean value to input the missing value in this project. We could use PCA to impute the missing data according to a research paper from IEEE International Conference named "PCA-guided k-Means clustering with incomplete data"(HONDA, 2011)

Reference

HONDA, K. *et al.* PCA-guided k-Means clustering with incomplete data. 2011 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE 2011), Fuzzy Systems (FUZZ), 2011 IEEE International Conference on, *[s. l.]*, p. 1710–1714, 2011.

WU, B. K-means clustering algorithm and Python implementation. 2021 IEEE International Conference on Computer Science, Artificial Intelligence and Electronic Engineering (CSAIEE), Computer Science, Artificial Intelligence and Electronic Engineering (CSAIEE), 2021 IEEE International Conference on, *[s. l.]*, p. 55–59, 2021.