Capstone 2: Life Expectancy Predictors

1.Introduction:

Life expectancy is the average number of years that a person is expected to live. It can be calculated for different ages (at birth, at the age of 65…) and it is carried out assuming that mortality rates by age are being maintained (which can also be called age-specific mortality pattern). It is a measure that summarizes the mortality of a country, allowing us to compare it by generations and analyze trends. Its interpretation and meaning is even richer and can provide us with key information on the level of development of a country's welfare state.
In fact, this indicator is so important for describing population conditions that, together with the education index and the Gross Domestic Product (GDP) index, it forms the Human Development Index used by the United Nations Development Programme (UNDP). There is no better indicator of a country's social development than having a long and healthy life.

Life expectancy expansion is a result of, among other things, improvements in nutrition, health and, above all, a decrease in mortality, but also of other various reasons that lead to those mentioned previously. As an indicator, it allows us to extrapolate at a glance the living conditions of populations in general. Yet we must not forget that the life expectancy lengthening results from the reduction of mortality at all ages and not just among those who find themselves "at the top" of the population pyramid. The life expectancy increases not only when the older ones live longer, but also when less young people die.

2. Data acquisition:

The data set has been obtained from Kaggle website. The data-sets are made available to public for the purpose of health data analysis. The data-set related to life expectancy, health factors for 193 countries has been collected from the same WHO data repository website and its corresponding economic data was collected from United Nation website.

There have been a lot of studies undertaken in the past on factors affecting life expectancy considering demographic variables, income composition and mortality rates. It was found that affect of immunization and human development index was not taken into account in the past. Also, some of the past research was done considering multiple linear regression based on a data set of one year for all the countries. In a nutshell, this study will focus on mortality factors, economic factors, social factors, and other health related factors as well.

link:https://www.kaggle.com/kumarajarshi/life-expectancy-who?select=Life+Expectancy+Data.csv

3. Aim of the Project:

To predict the life expectancy by formulating a regression model based on multiple linear regression, Gradient boosting regressor, and Random Forest Regressor while considering data from a period of 2000 to 2015 for all the countries.

4. Data Exploration and Cleaning:

The data set consists of 2938 records and 22 fields. The dependent variables are 'Country', 'Year', 'Status', 'Life expectancy ', 'Adult Mortality', 'infant deaths', 'Alcohol', 'percentage expenditure', 'Hepatitis B', 'Measles ', ' BMI ', 'under-five deaths ', 'Polio', 'Total expenditure', 'Diphtheria ', ' HIV/AIDS', 'GDP', 'Population',' thinness 1-19 years', ' thinness 5-9 years' 'Income composition of resources', and 'Schooling' and the target variable is 'Life_expectancy'.
The data set needs to be get rid of missing values and outliers before performing any analysis to see the relationship between variables and draw conclusions. The missing values in this project were imputed using the median value of that particular field. In addition, data type of each field was checked and some of the fields that did not provide any meaningful information were removed. The rows having the 'Null' values in the target variable were also dropped.

5. Data Analysis:

After the data was cleaned data visualization was performed using libraries such as matplotlib and seaborn to see the relationship between target and the independent variables. Following are some of the important visualizations that provided informative insights:

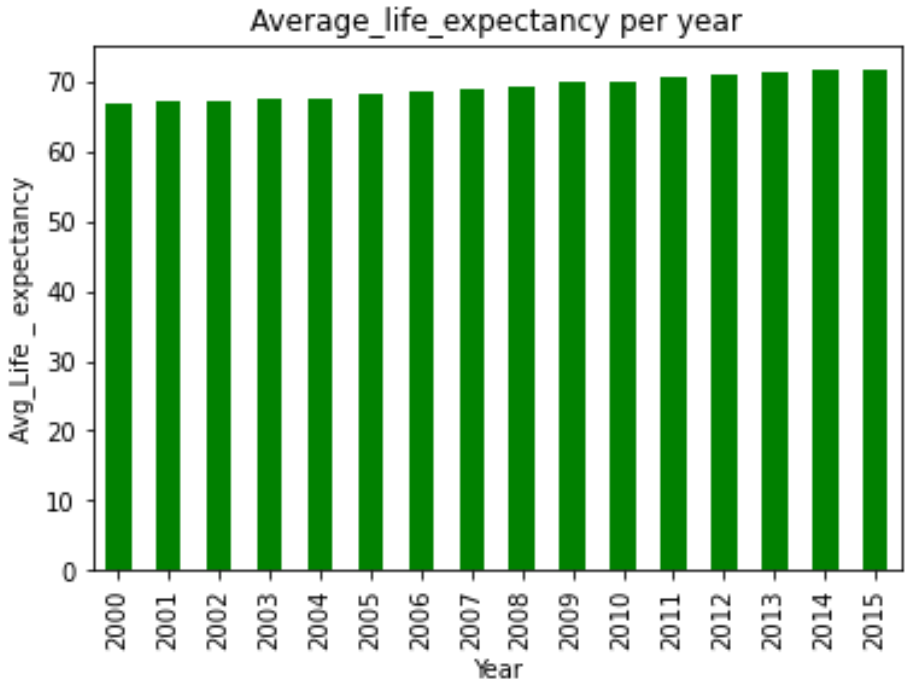

Fig.1. Average life expectancy Vs Year

The plot in Fig.1. shows that average life expectancy has slightly increased over 15 years.From the fFig.2 below we can infer that the average life expectancy is higher for the developed

countries compared to developing countries. This could be due to various factors such as economic conditions, health conditions and income etc.

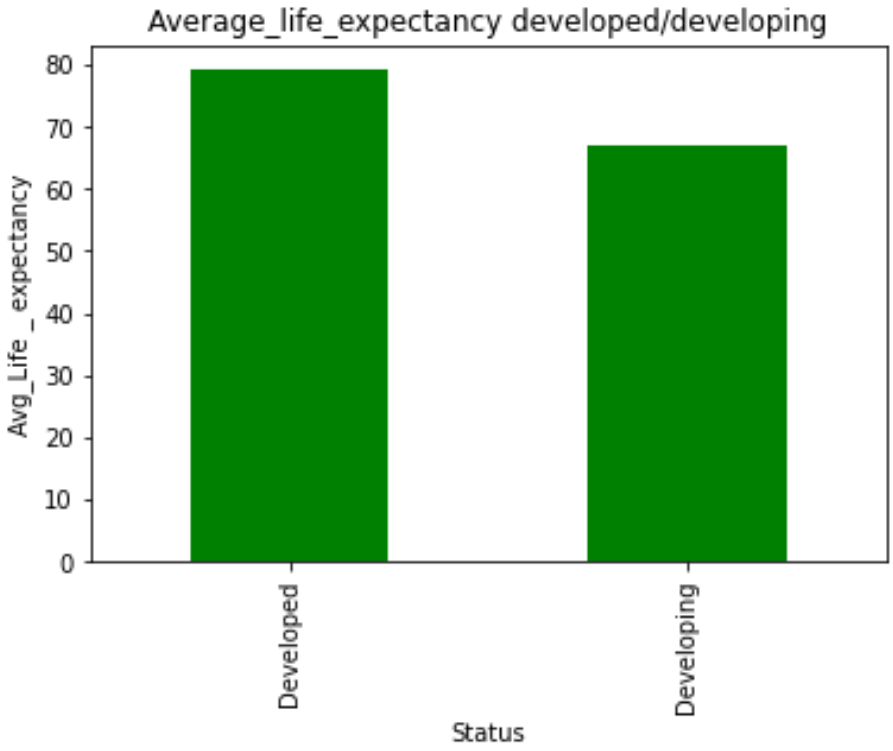

Fig.2. Average life expectancy Vs Year

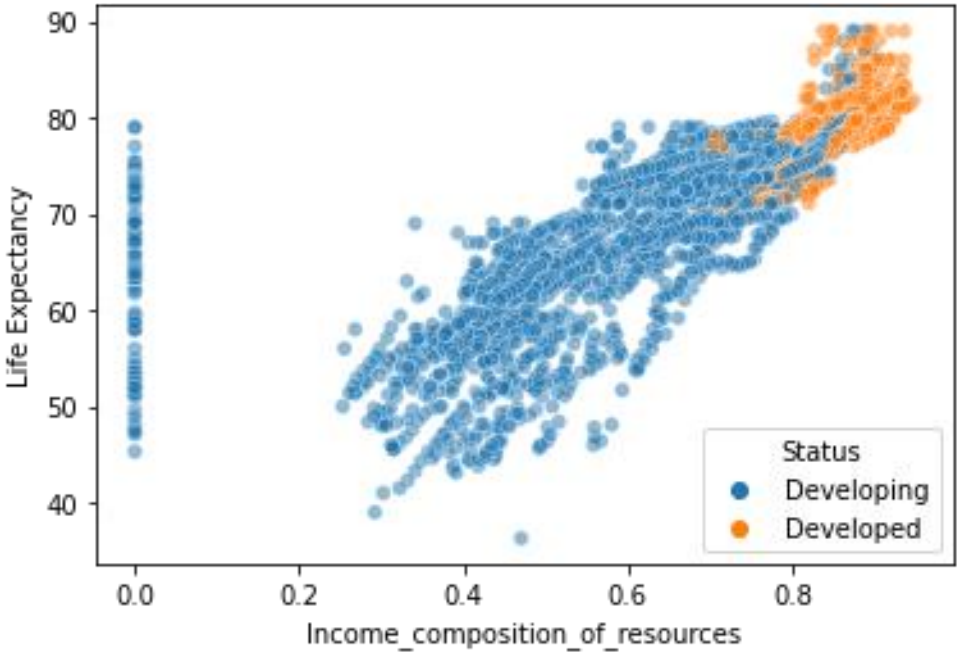

Fig.3. Average life expectancy Vs Income composition/GDP

The scatter plot in Fig.3 shows that there is a linear pattern between life expectancy and income composition of resources/ GDP. This indicates that it could be one of the important factor that can drastically change life expectancy. The countries having higher GDP have higher life expectancy and standard of living.

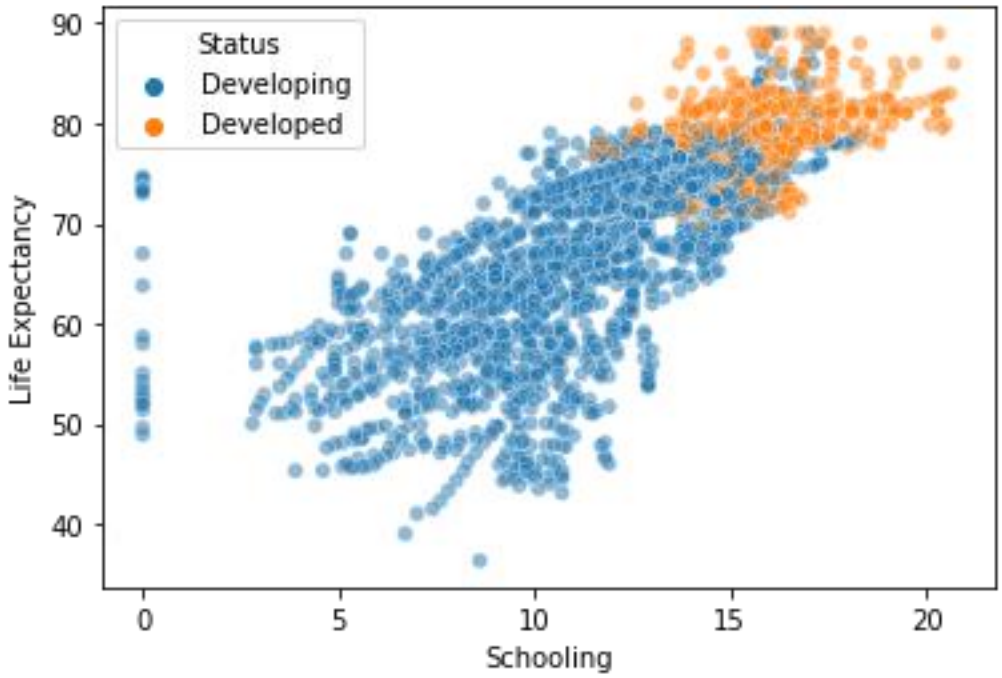

Fig.4. Average life expectancy Vs Schooling

Similarly, the plot above shows a linear pattern between schooling and life expectancy indicating that education or higher schooling helps people lead a better life with increased life expectancy. Furthermore, boxplots were also plotted for the variables to check for the outliers.
The heatmap in the Fig.5 below shows the correlation between variables. Apart from looking for correlation between independent variables and the target variables it can be used to detect collinearity between independent variables.

From the correlation matrix below, Life expectancy has a strong relationship with schooling and income composition of resourcesThe correlation matrix above also indicates muticolinearity between percentage_expenditure and GDP, infant_deaths and under-five_deaths, thinness1-19 and thinness5-9 years, schooling and income_composition
To overcome this we removed some variables that are highly correlated to others and leave the more significant ones in the set.

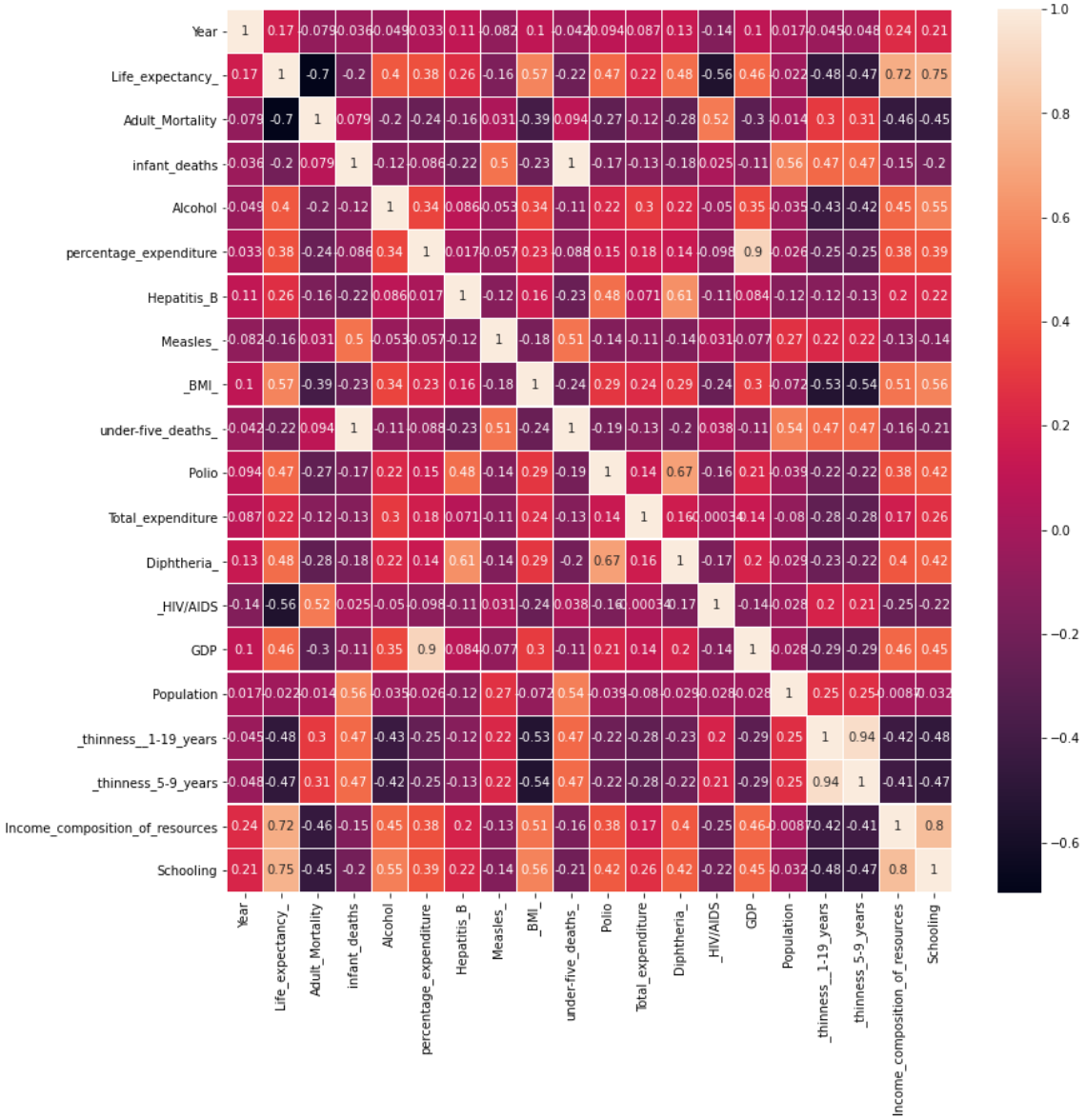

Fig.5. Heat map showing correlation between variables.

6. Modeling methods:

Since, the target variable was continuous numeric regression analysis/models were used to predict the target variable . Regression analysis is a way of mathematically sorting out which of those variables does indeed have an impact. It answers the questions: Which factors matter most? Which can we ignore? How do those factors interact with each other? And, perhaps most importantly, how certain are we about all of these factors? The regression models used from scikit-learn library in the project are:

- **Linear Regression :** Linear regression is a linear model, e.g. a model that assumes a linear relationship between the input variables (x) and the single output variable (y). More specifically, that y can be calculated from a linear combination of the input variables (x).
- **Random Forest Regressor:** Random Forest is a type of supervised learning algorithm that uses ensemble methods (bagging) to solve both regression and classification problems. The algorithm operates by constructing a multitude of decision trees at training time and outputting the mean/mode of prediction of the individual trees.
- **KNN Regressor:** The KNN algorithm assumes that similar things exist in close proximity. In other words, similar things are near to each other.
- **Decision Tree Regressor:** The decision trees is used to fit a sine curve with addition noisy observation. As a result, it learns local linear regressions approximating the sine curve.
- **Gradient Boosting Regressor:** GB builds an additive model in a forward stage-wise fashion; it allows for the optimization of arbitrary differentiable loss functions. In each stage a regression tree is fit on the negative gradient of the given loss function.

6.1 Pipelines:

A machine learning pipeline is a way to codify and automate the workflow it takes to produce a machine learning model. Machine learning pipelines consist of multiple sequential steps that do everything from data extraction and preprocessing to model training. The data was split into train and test groups with a test_size of 0.3 In this project pipelines were made for different regressors used and it included SimpleImputer() to impute the missing values, StandardScaler() to scale the values and finally the type of regressor.

To avoid overfitting of the model the cross-validation method was applied. In K Fold cross validation, the data is divided into k subsets. Now the holdout method is repeated k times, such that each time, one of the k subsets is used as the test set/ validation set and the other k-1 subsets are put together to form a training set.

Further, to optimize the R square score of the models some of the hyper parameters were tuned using GridSearchCV(), where, Grid-search is a way to select the best of a family of models, parametrized by a grid of parameters. Grid search exercise can save us time, effort ,resources, and most importantly can improve efficiency of the model.

7. Model performance and Result:

After the pipelines were made they were fit and trained on the training sets, followed by cross-validation and grid search cv and using the test sets prediction were made. Then, the best scores were obtained using '.best_scores_' method on the instance created using grid search.

R-squared (R2) is a statistical measure that represents the proportion of the variance for a dependent variable that's explained by an independent variable or variables in a regression model. Whereas correlation explains the strength of the relationship between an independent and dependent variable, R-squared explains to what extent the variance of one variable explains the variance of the second variable.
The following table shows the R square scores of the models:

R square Scores	
Random Forest Regressor	0.95
KNN regressor	0.89
Decision Tree regressor	0.74
Gradient boosting regression	0.93
Linear regression	0.81

Table.1. R square scores of the models.

From the table above , the highest R square score, of the Random Forest Regressor is model is 0.95, which means that 95% half of the observed variation can be explained by the model's inputs.

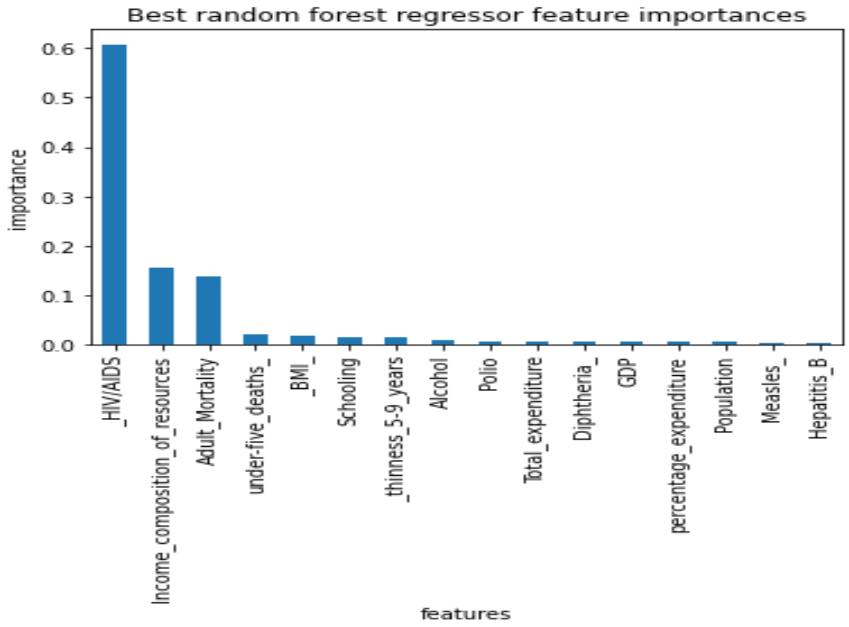

Fig.6. Important features

So, from the importance plot above it can be seen that the top important predictors that cause an impact on the life expectancy are Income composition of resources, HIV/AIDS, Adult mortality, schooling, BMI and under-five_ deaths.
The results in Fig.6 are analogous to the correlation coefficients obtained from linear regression model. In regression with a single independent variable, the coefficient tells you how much the dependent variable is expected to increase (if the coefficient is positive) or decrease (if the coefficient is negative) when that independent variable increases by one.

```
Schooling                          2.383978
Income_composition_of_resources    1.227533
Diphtheria_                        1.009965
_BMI_                              0.873260
Polio                              0.744830
GDP                                0.401361
Alcohol                            0.330759
percentage_expenditure             0.327036
Population                         0.318730
Total_expenditure                  0.267146
_thinness_5-9_years               -0.147974
Measles_                          -0.317868
Hepatitis_B                       -0.364793
under-five_deaths_                -0.507803
_HIV/AIDS                         -2.386217
Adult_Mortality                   -2.749388
```

Fig.7. snippet showing the correlation coefficient of each variable.

Information about important features can help organizations such as The US Department of Health and Human Services (HHS) in examining social inequities in health, disease, and mortality and developing programs to reduce health inequities among populations that experience increased risk of poor health based on race/ethnicity, gender and socioeconomic status.

Future Applications:

Further work can be done on building a model that combines the income and gender data set and which can give insights of income groups can impact life expectancy in developed and developing countries. It can also inform us about the quality of life of those last years and life expectancy in health (we will analyze it) but to understand the aging and the importance of the vital stage of the old age.
It's not surprising that those with more wealth tend to live longer than those with less. If you have more money, you probably have access to better health care as well as more nutritious

foods. You also have less stress from worrying about money, and stress is a factor in mortality, as well.

References:

- **https://www.kaggle.com/**
- **https://scikit-learn.org/stable/**
- **https://www.cdc.gov/nchs/fastats/life-expectancy.htm**