Machine Learning Hackathon

Report

Done By:

Sri Vishnu Lahari Konidena (IMT2017041)
Naga Sri Vaishnavi Dhulipalli (IMT2017514)
Kadiyala Venkata Lasya (IMT2017023)

Problem Statement

A country’s development depends on the health and wealth of the people. Wealth being determined by GDP, the health of a country is judged by the average life expectancy of people in the country. Analysing how different factors affect life expectancy, the country can further work on those factors with an aim of increasing the life expectancy. **The problem statement is to predict the average life expectancy in several countries.**

Exploratory Data Analysis

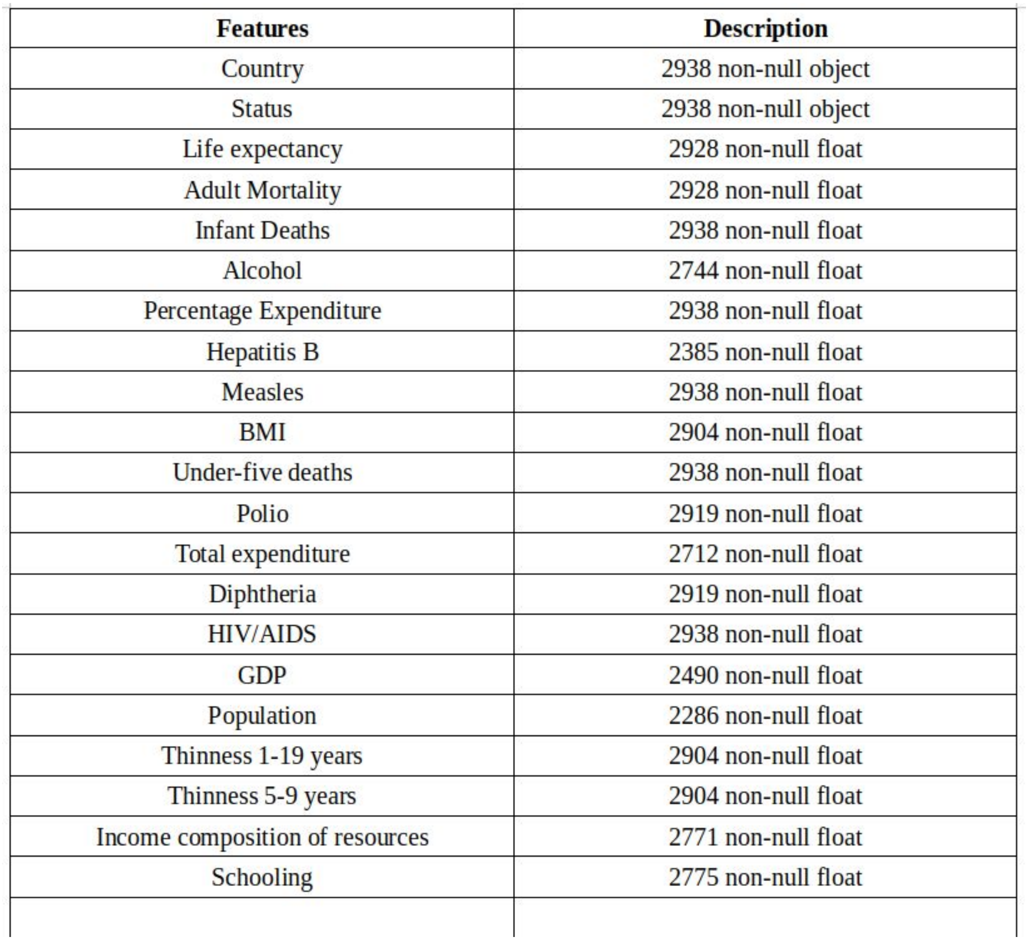
The dataset used is taken from kaggle : https://www.kaggle.com/kumarajarshi/life-expectancy-who

Data Description:

1. The dataset is of the dimension (2938,22)
2. We have 21 input features and 2938 data points.
    - **Country** : Country
    - **Year** : Year
    - **Status** : Developed or Developing status
    - **Life expectancy** : Life Expectancy in age
    - **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
    - **Alcohol** : Alcohol, recorded per capita (15+) consumption (in litres of pure alcohol)
    - **Percentage Expenditure** : Expenditure on health as a percentage of Gross Domestic Product per capita(%)
    - **Hepatitis B** : Hepatitis B (HepB) immunization coverage among 1-year-olds (%)

- **Measles** : Measles - number of reported cases per 1000 population
- **BMI** : Average Body Mass Index of entire population
- **Under-five deaths** : Number of under-five deaths per 1000 population
- **Polio** : Polio (Pol3) immunization coverage among 1-year-olds (%)
- **Total expenditure** : General government expenditure on health as a percentage of total government expenditure (%)
- **Diphtheria** : Diphtheria tetanus toxoid and pertussis (DTP3) immunization coverage among 1-year-olds (%)
- **HIV/AIDS** : Deaths per 1000 live births HIV/AIDS (0-4 years)
- **GDP** : Gross Domestic Product per capita (in USD)
- **Population** : Population of the country
- **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 Age 5 to 9(%)
- **Income composition of resources** : Human Development Index in terms of income composition of resources (index ranging from 0 to 1)
- **Schooling** : Number of years of Schooling(years)

3. **Dataframe.info():** Gives the information such as columns of the data frame, number of non-null values of the column and the data type of the column, it is observed that every column has the following number of NULL values.



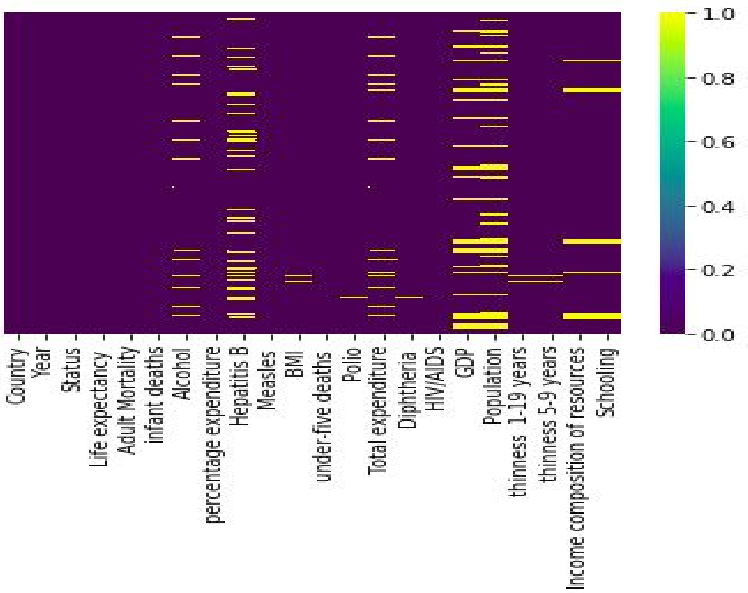
Features	Description
Country	2938 non-null object
Status	2938 non-null object
Life expectancy	2928 non-null float
Adult Mortality	2928 non-null float
Infant Deaths	2938 non-null float
Alcohol	2744 non-null float
Percentage Expenditure	2938 non-null float
Hepatitis B	2385 non-null float
Measles	2938 non-null float
BMI	2904 non-null float
Under-five deaths	2938 non-null float
Polio	2919 non-null float
Total expenditure	2712 non-null float
Diphtheria	2919 non-null float
HIV/AIDS	2938 non-null float
GDP	2490 non-null float
Population	2286 non-null float
Thinness 1-19 years	2904 non-null float
Thinness 5-9 years	2904 non-null float
Income composition of resources	2771 non-null float
Schooling	2775 non-null float

4. **Dataframe.describe():** On checking the statistical details, it is observed that there is a large difference in the value corresponding to 75% of the dataset and the minimum or maximum value of few of the columns. This indicates the presence of outliers in those columns.
5. Our “TARGET” variable is Life expectancy.

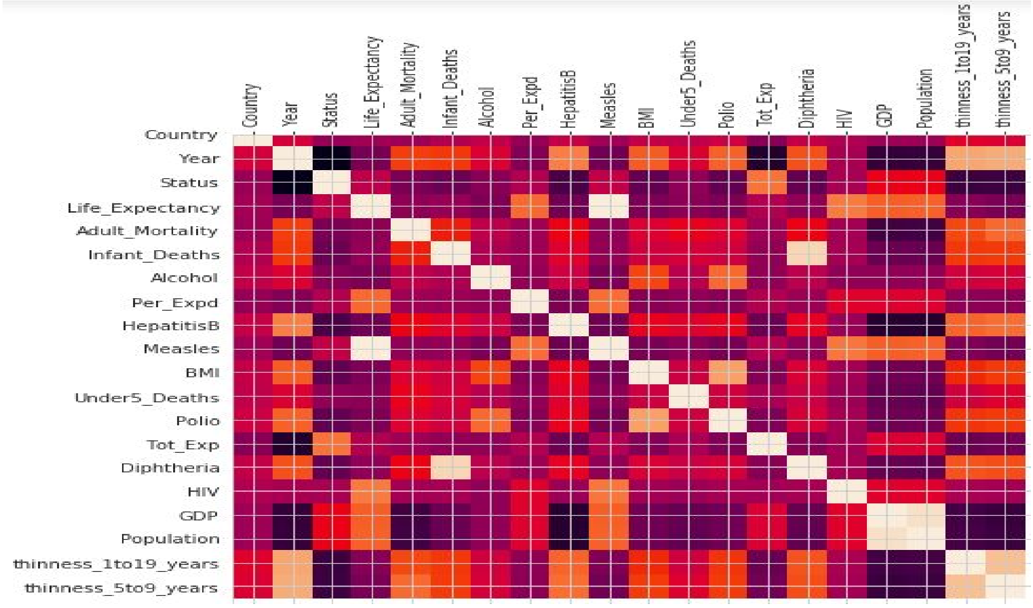
Data Visualization:

1. The below HEATMAP shows the number of null values or missing values in the dataset.

The yellow lines in the heatmap indicate missing values in the dataset.

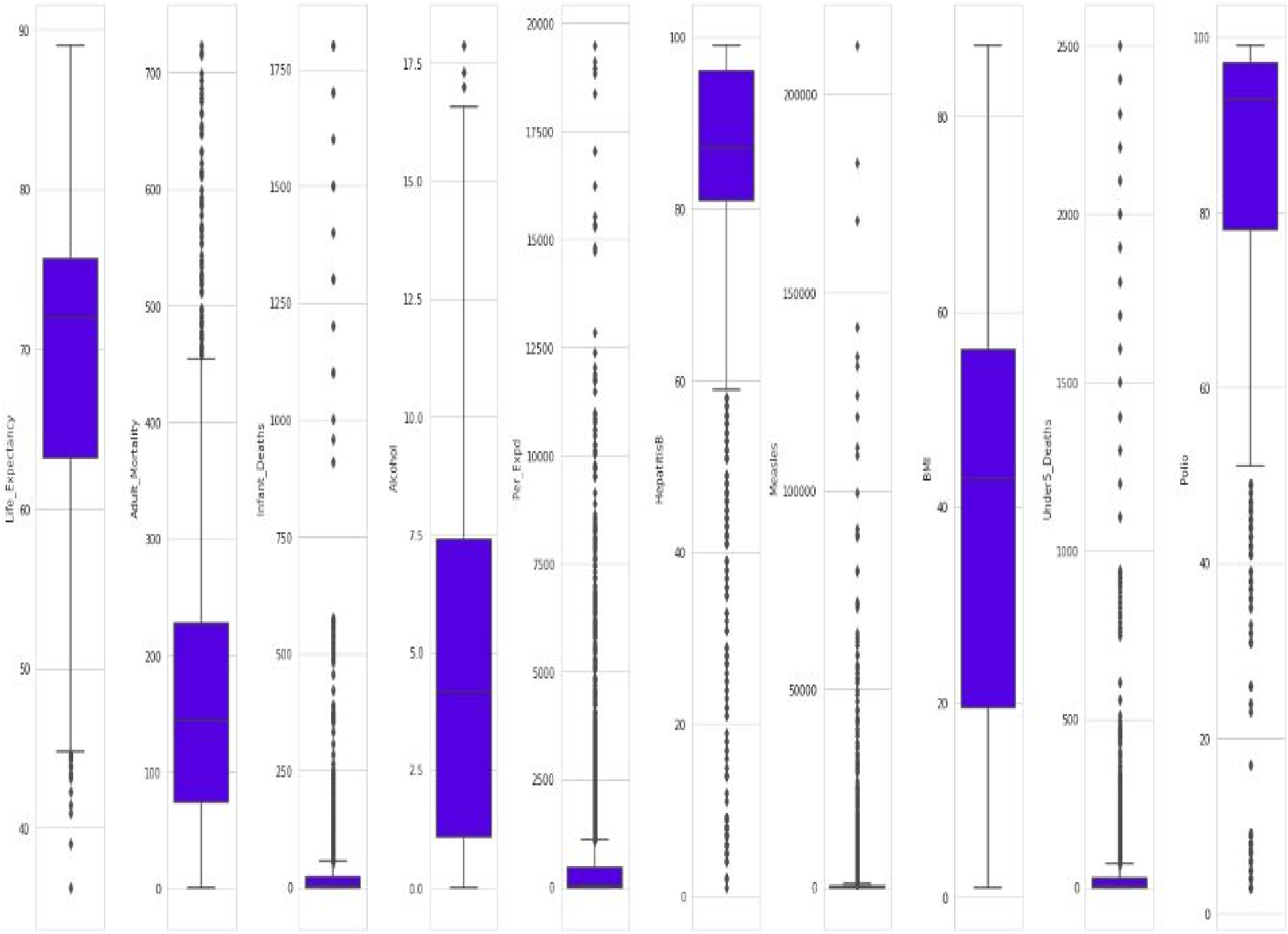


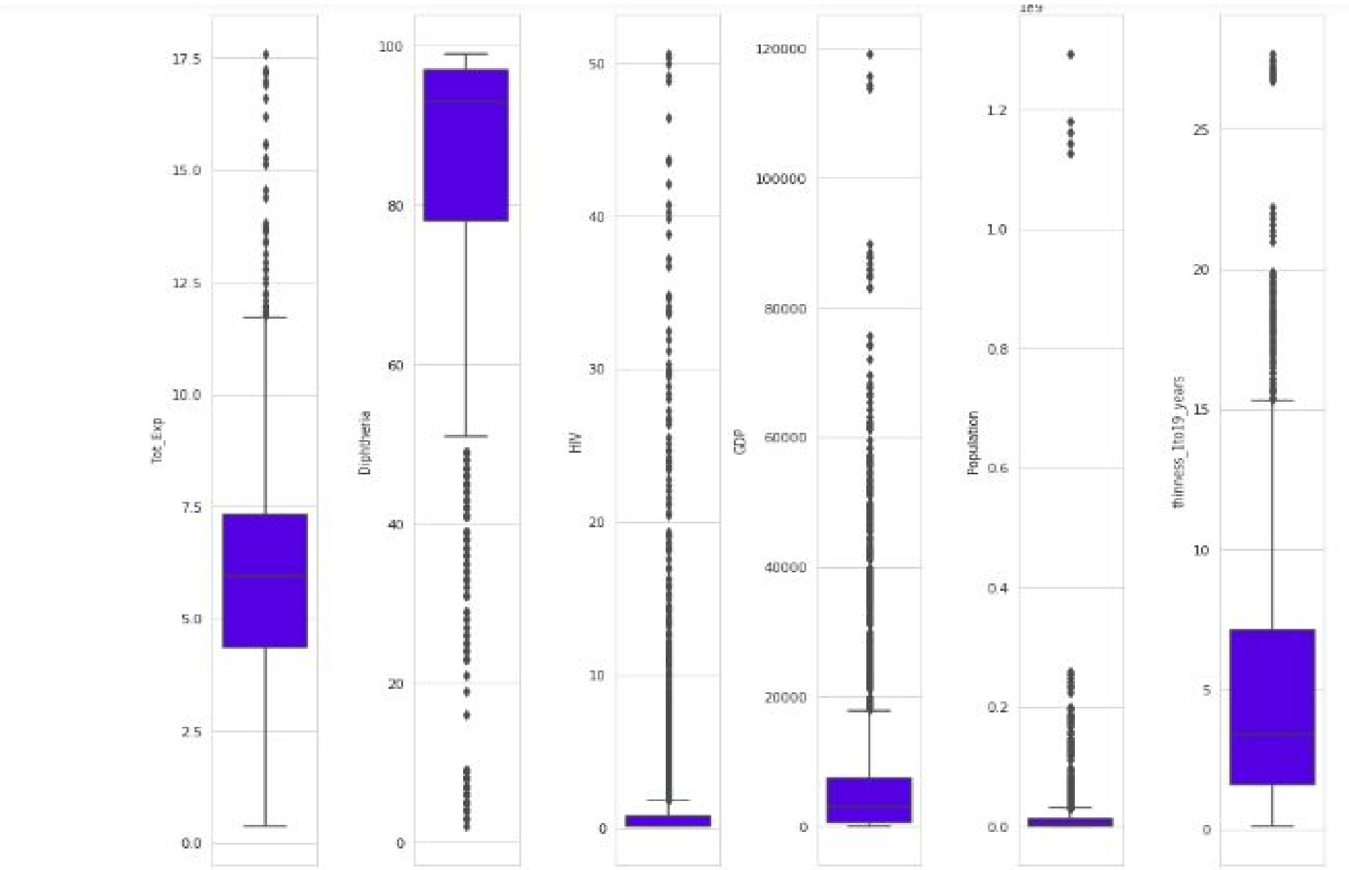
2. The heatmap confirms our earlier observation that all input feature columns apart from (Blood Pressure, Insulin, and Diabetes Pedigree Function) consist of NULL values.
3. **Dataframe.corr():** Gives the pairwise CORRELATION of columns excluding NULL values.



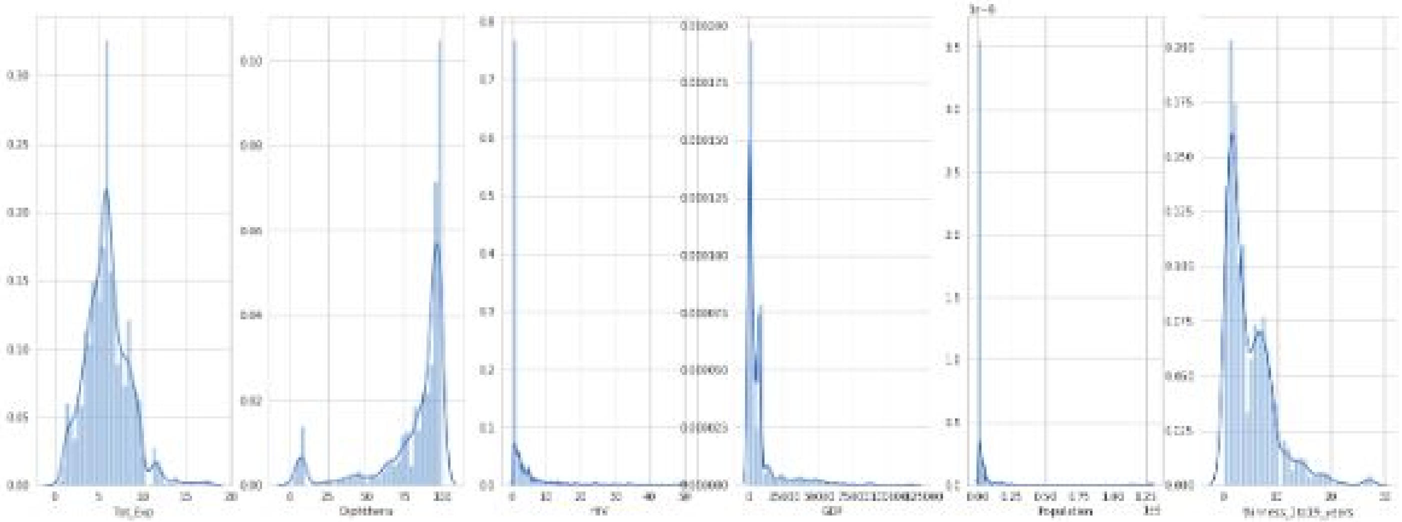
From the above HEATMAP, we can infer that the "Outcome" has a strong positive correlation with "Glucose" and least positive correlation with "BloodPressure".

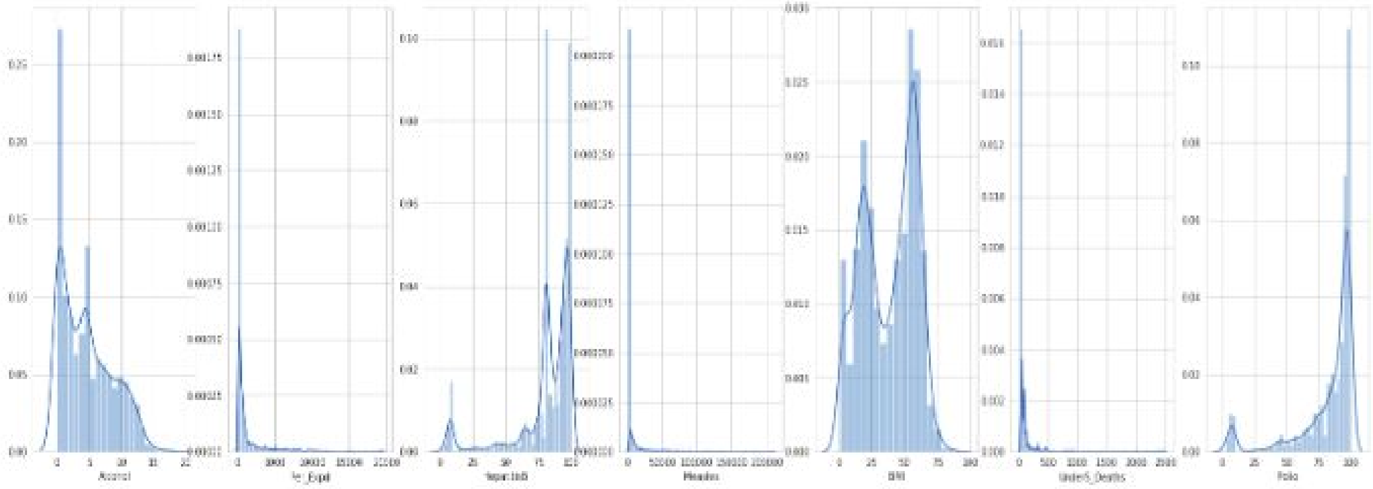
4. Outliers using BOXPLOT:
   An outlier is a value that lies in a data series on its extremes, which is either very small or large and thus can affect the overall observation made from the data series. They are usually treated as **abnormal values** that can affect the overall observation.





5. Distributed Skewness:
   Skewness is a measure of the asymmetry of the probability distribution of a real-valued random variable about the mean.



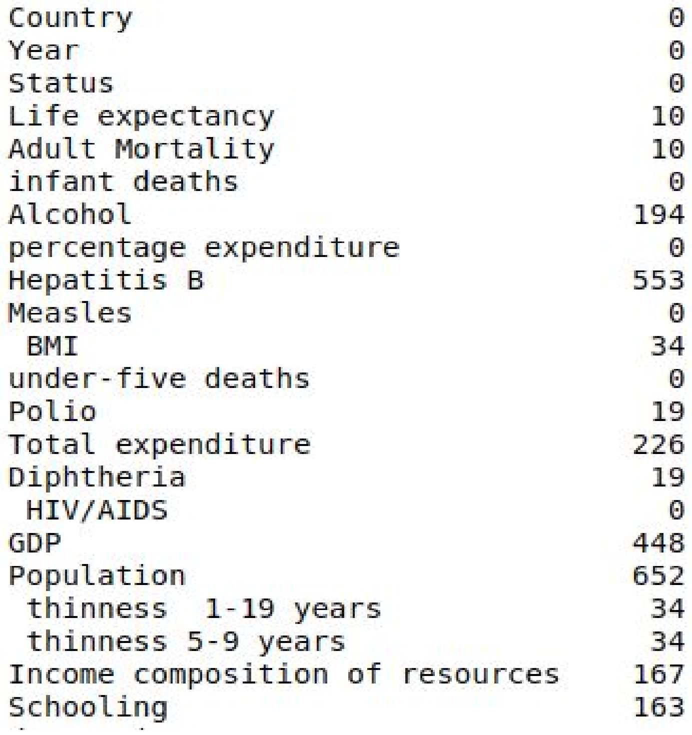


Data Preprocessing

Data Handling:

It includes handling null values or missing values, zero values and negative values.

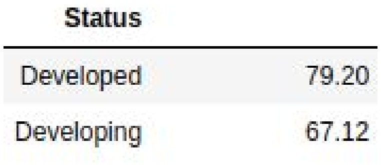
Attribute **number of NULL values**



```
Country                              0
Year                                 0
Status                               0
Life expectancy                     10
Adult Mortality                     10
infant deaths                        0
Alcohol                            194
percentage expenditure               0
Hepatitis B                        553
Measles                              0
 BMI                                34
under-five deaths                    0
Polio                               19
Total expenditure                  226
Diphtheria                          19
 HIV/AIDS                            0
GDP                                448
Population                         652
 thinness  1-19 years               34
 thinness 5-9 years                 34
Income composition of resources    167
Schooling                          163
```

On checking for null values, zero values and negative values, we got

1. **Country** : This is an object type of value.
2. **Status** : This is an object type of value. From t-test, we analysed that the status attribute also plays an important role in the life expectancy prediction. We used one-hot encoding for status attribute having two unique values “developed” and “developing”.



Status	
Developed	79.20
Developing	67.12

3. **Adult Mortality:** All the negative values, zero values and null values are replaced with average adult mortality rate of that particular country over several years.
4. **Infant Deaths:** Only negative values and null values are replaced with average deaths of infants of that particular country, number of deaths of infants per 1000 is allowed to be zero.
5. **Alcohol**: All the negative values, zero values and null values are replaced with average alcohol recorded per capita of that particular country over several years.
6. **Percentage Expenditure**: All the negative values, zero values and null values are replaced with average Percentage Expenditure of that particular country over several years.
7. **Hepatitis B**:All the negative values, zero values and null values are replaced with the average Hepatitis B(HepB) immunization coverage among 1-year olds of that particular country over several years.
8. **Measles**:Only negative values and null values are replaced with average reported cases of measles of that particular country, reported cases per 1000 is allowed to be zero.
9. **BMI**: All the negative values, zero values and null values are replaced with average BMI of that particular country over several years.
10. **Under-five deaths**: Only negative values and null values are replaced with average deaths of children under the age of five of that particular country, number of deaths because of HIV/AIDS per 1000 is allowed to be zero.

11. **Polio**: All the negative values, zero values and null values are replaced with the average Polio immunization coverage among 1-year olds of that particular country over several years.

12. **Total Expenditure**: All the negative values, zero values and null values are replaced with the average total expenditure of that particular country.

13. **Diphtheria**:All the negative values, zero values and null values are replaced with the average Diphtheria tetanus toxoid immunization coverage among 1-year olds of that particular country over several years.

14. **HIV/AIDS**:Only negative values and null values are replaced with average deaths due to HIV/AIDS of that particular country, number of deaths because of HIV/AIDS per 1000 is allowed to be zero.

15. **GDP**: All the negative values, zero values and null values are replaced with the average GDP of that particular country over several years.

16. **Population:** All the negative values, zero values and null values are replaced with the average population that particular country over several years.

17. **Thinness 1-19 years:** All the negative values, zero values and null values are replaced with the average average thinness among kids from 1-19 years that particular country over several years

18. **Thinness 5-9 years:** All the negative values, zero values and null values are replaced with the average schooling rate thinness among kids from 5-9 years that particular country over several years

19. **Income Composition of resources:** All the negative values, zero values and null values are replaced with average income composition of resources of that particular country.

20. **Schooling:** All the negative values, zero values and null values are replaced with the average schooling rate that particular country over several years.

Outliers Handling:

- BOXPLOT has been used to plot the outliers in the features.
- The plot is typically depicted by quartiles and inter quartiles. It helps in defining the upper limit and lower limit beyond which any data lying will be considered as outliers.
- Lower Quartile(Q1) = (n+1)/4 th term
- Upper Quartile(Q3) = 3(n+1)/4 th term
- Interquartile Range(IQR) is the spread of the middle 50% of the data values.
- IQR = Q3 - Q1
- Lower Limit = Q1 - 1.5*IQR
- Upper Limit = Q3 + 1.5*IQR
- The outliers in the attributes here are too many, they are not less than 10% of the data. If we try removing outliers from the data, nearly 1000 data points are being removed which is nearly 30% of the data. So it is not a good idea to remove outliers in this case.

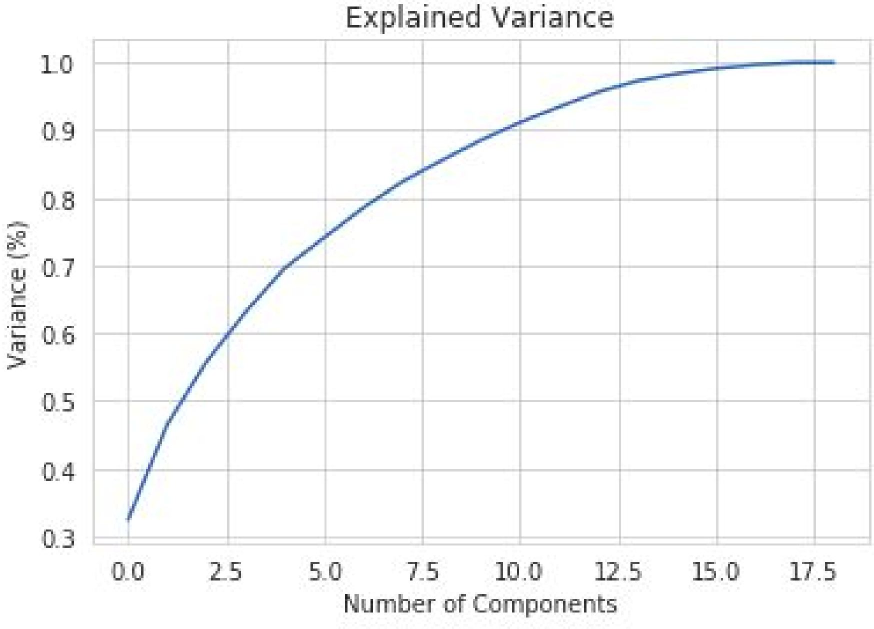
Normalization:

- Normalization is to change the values of numeric columns in the dataset to a common scale, without distorting differences in the ranges of values.
- As our data columns have different ranges. We should normalize the data.
- Using StandardScalar, the data columns have been normalized to a range [0,1].

Feature Extraction

Principal Components Analysis:

- PCA has been used for feature extraction.
- PCA has been implemented using sci-kit-learn on the normalized data after removing the outliers.
- **Explained Variance** measures the discrepancy between the model and actual data.
- Higher percentages of explained variance indicate a stronger strength of association. It also means that the model can make better predictions.
    - On plotting a graph between variance(%) and the number of components, we observed that for 15 components we get a maximum variance, that is, better prediction.
    - After 15 components, there is overfitting in training.
    - Therefore, for PCA, we used 15 as the number of principal components.

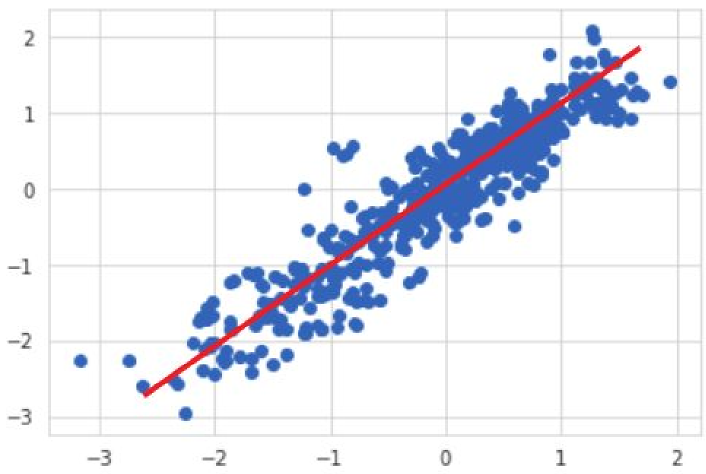


Model Building

Regression Approach has been used as the dependent variable(target) in our problem is a numerical value prediction. In our problem, the target variable is “Life Expectancy”. So, we used **Linear Regression, Polynomial Regression, Random-forest Regression** and **Decision Tree Regression** for the problem.
We have used test_train_split function to split our dataset into test and train sets. 20% of the data has been used as test data and the remaining 80% as the train data.
The below function from sci-kit learn has been used to split the data into train and test data.

x_train, x_test, y_train, y_test = train_test_split(X,Y,test_size = 0.2,random_state = n)

Linear Regression:
Linear regression is a linear approach to modeling the relationship between a scalar response (or dependent variable) and one or more explanatory variables (or independent variables). In the train_test_spilt function, we changed the random_state value from 0 to 1000 in a loop and measured the average accuracy and maximum accuracy for both train and test data. Along with accuracy, average Root mean squared error, average Mean Absolute error, average Mean squared error has been calculated.
ON TRAIN DATA:
Avg train accuracy: **0.894**
max train accuracy: **0.903**
Root Mean Squared Error: 0.325
Mean Absolute Error: 0.242
Mean Squared Error: 0.105
ON TEST DATA:
Avg test accuracy: **0.892**
max test accuracy: **0.922**
Root Mean Squared Error: 0.328
Mean Absolute Error: 0.244
Mean Squared Error: 0.108



Polynomial Regression:

Polynomial provides the best approximation of the relationship between the dependent and independent variable.A Broad range of function can be fit under it.Polynomial basically fits a wide range of curvature.

The presence of one or two outliers in the data can seriously affect the results of the nonlinear analysis.

These are too sensitive to the outliers.In addition, there are unfortunately fewer model validation tools for the detection of outliers in nonlinear regression than there are for linear regression. This dataset contains a lot of outliers which polynomial regression cannot handle. The tools couldn't fit the most of the values into a curve and hence the negative value of R_square score.

Mean squared error: 4.50

Mean absolute error: 0.39

R_square score: **-3.56**

Random Forest Regression:

Random forest is a Supervised Learning algorithm which uses ensemble learning method for classification and regression. It is one of the most accurate learning algorithms available. It can handle large number of attributes or input variables. As a significant part is missing in our case, random forest regression is an effective method for estimating data and it maintains good accuracy.

But there are chances of overfitting based on the hyperparameters taken.
Cross Validation score and r_2 score is used to measure the performance of the model. ‘cv’ value is used as the default value of 5. In the train_test_spilt function, we changed the random_state value from 0 to 50 in a loop and measured the average accuracy on test data.
Average mean cross validation score: **0.93**
Average r_2 score on the test data: **0.93**

Decision Tree Regression:

Decision tree regression observes features of an object and trains a model in the structure of a tree to predict data in the future to produce meaningful continuous output. Cross Validation score and r_2 score is used to measure the performance of the model. ‘cv’ value is used as the default value of 5. In the train_test_spilt function, we changed the random_state value from 0 to 50 in a loop and measured the average accuracy on test data.
Average mean cross validation score: **0.86**
Average r_2 score on the test data: **0.87**