PROJECT REPORT

Life Expectancy Prediction

Project Group Number 5

Group Members :

Teja (19341)

Abhay Ajith (19301)

Gokul Krishnan G (19321)

Karthik Narayanan (19327)

M Ranjith (19331)

TABLE OF CONTENTS

TITLE PAGE

Abstract

Longevity is affected by a variety of factors, including the country's economic growth and regional health breakthroughs. Along with the prophecy of existence, we also determine how susceptible a specific landmass is to a few chronic ailments. These factors have a significant influence on the population's prospective life span. We investigate the biological and economic elements of continents and their countries in order to estimate population life expectancy and the likelihood of the continent having long-standing illnesses such as measles, HIV/AIDS, and so on.

Our project is based on a hypothesis that shows how life expectancy is linked to a variety of elements, including both health and economic issues.

Introduction

Machine learning is a branch of computer science that has exploded in popularity in recent years. Big data and machine learning are affecting almost every area of life. The field of health informatics presents a significant challenge to this sector. The ultimate goal of machine learning is to create algorithms that can be well-trained and improve over time. Many policymakers and academics use life expectancy as an indicator to supplement economic metrics of wealth such as GDP. The average age of a demographic group's members when they die is depicted by prognosis of life. Life expectancy differs between industrialised and developing nations, as well as the ratio of birth to death, mortality rates in various countries, and the ratio of literate to illiterate populations, all of which have an impact on survival time in some manner. The pace of population increase is influenced by the country's growth, advances, and resource accessibility. The life expectancy is computed as the average survival time, which represents the median age of the population. Some people may live longer, while others may live shorter, but on average, the expected figure is the continent's lifespan.

Life prognosis is important not only for projecting survival rates, but also for determining if a region has a high risk of illness. Another part of study is disease categorization, which is in addition to life prediction. Disease prediction is done by taking into account the economic and sociological aspects of several countries within a continent, and then combining that information to forecast disease throughout the continent. The occurrence of illness in a country is influenced by its growth. GDP, population awareness, illiteracy rate to literacy rate, birth-to-death ratio, and other variables all have a combined influence on the onset of a disease. As a result, machine learning is the most appropriate approach for predicting and classifying. To accomplish the desired result, regression, classification, and prediction algorithms can be applied in a variety of ways. To acquire the necessary findings, regression methods such as linear regression a are used to estimate life expectancy, whilst classification techniques such as decision tree, k-nearest neighbour algorithm are used to classify illness incidence.

Study System And Methods

Although some demographic traits, income composition, and mortality rates have been studied in relation to variables impacting life expectancy, many other factors such as HIV/AIDS, polio, measles have not yet been studied properly.

This project is primarily concerned with factors such as mortality rates, economic factors, social characteristics, and other health-related data.

Countries can better identify which factors contribute to poor life expectancy in particular, and then the continent as a whole.

Linear Regression

Linear regression is one of the most easy and straightforward algorithms to comprehend and implement. It is a predictive machine learning technique that is both statistical and machine learning in nature. The linear relationship between the label and one or more characteristics is determined using linear regression.

As a result, the traits have a role in predicting the label. A regression line is created by plotting data points. The best fit line is the highest approximated value with the smallest space between the expected and actual value. It is mostly used to determine the strength of predictors, forecast an impact, and anticipate trends.

Multi Linear Regression

There is just one dependent variable in multiple linear regression, but there are many independent factors. The basic goal is to anticipate the result by modelling the linear connection between the dependent and independent variables. All of the characteristics are taken into account while drawing the regression line. Multiple Linear Regression functions are based on the following assumptions:

- Between the independent and dependent variables, there is a linear connection.
- There shouldn't be too much connection between the independent variables.

- Observations are chosen from the population at random and independently.
- The residuals should have a normal distribution with a mean of 0 and a variance of sigma.

Decision Tree

Decision tree is a supervised machine learning algorithm. This approach may be used for both classification and regression. The goal of a decision tree is to use decision rules that have been applied to the model by producing some training data to forecast the value of a target variable. In comparison to other categorization methods, this one is simple to comprehend. It solves the problem by using a tree representation, in which each internal node corresponds to an attribute and each leaf node corresponds to a class label.

The selection of an attribute for the root node from the dataset is the major source of concern in decision trees. Attribute selection is a term used to describe the process of addressing this issue. There are two approaches for doing so: one is the information gain method, and the other is the Gini impurity method.

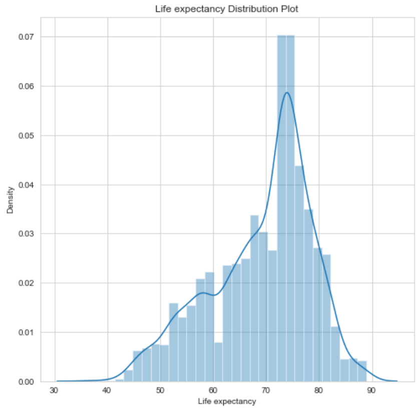
Furthermore, we have also used a decision tree to predict the expected life expectancy.

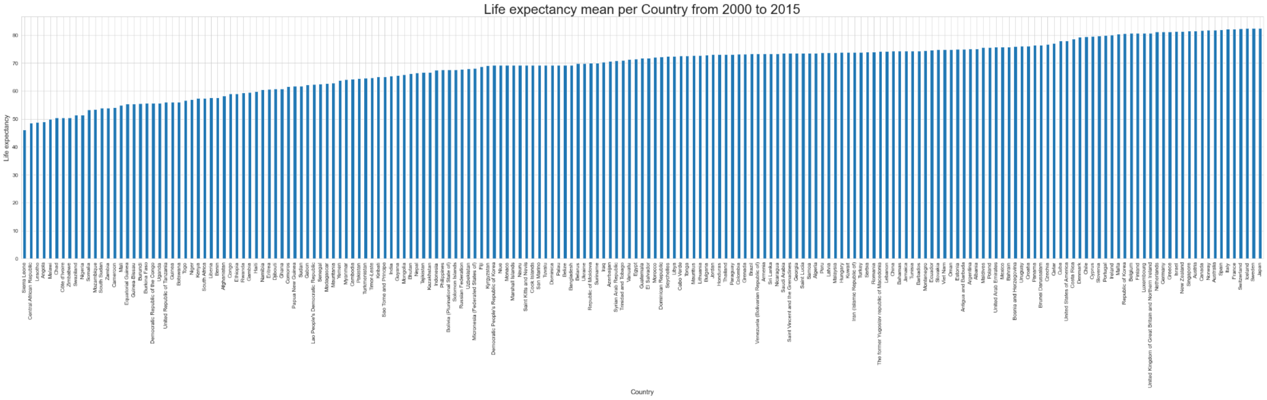
SVM

Support Vector Machine (SVM) is supervised machine learning that can be used for both classification or regression challenges. However, it is mostly used in classification problems. In the SVM algorithm, we plot each data item as a point in n-dimensional space (where n is a number of features you have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiates the two classes very well.

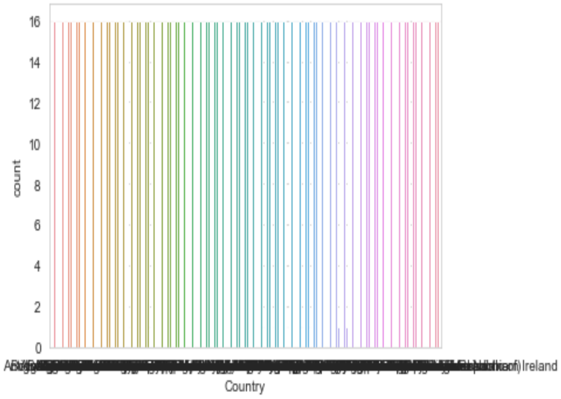
Results And Discussion

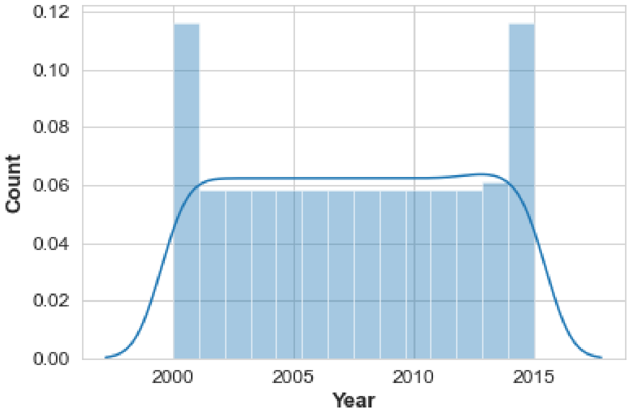
Analysis was done on the target variable life expectancy and plotted the graph which is given below:

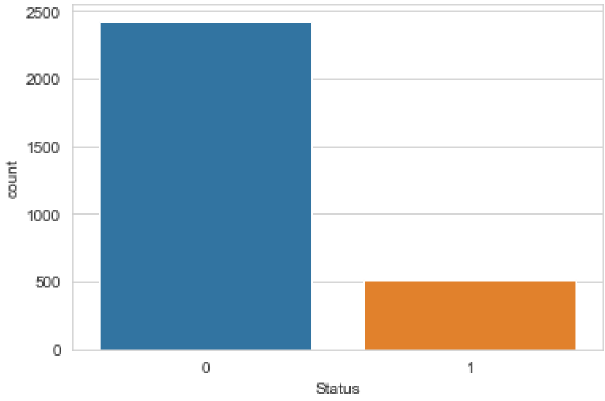


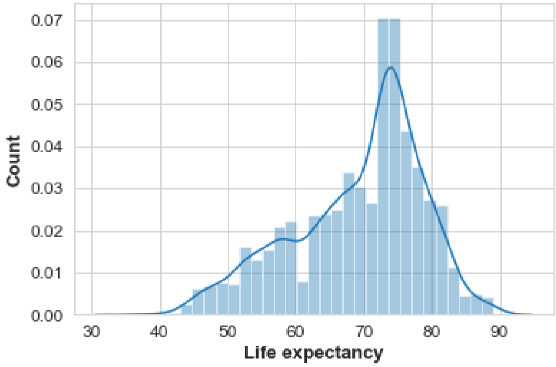


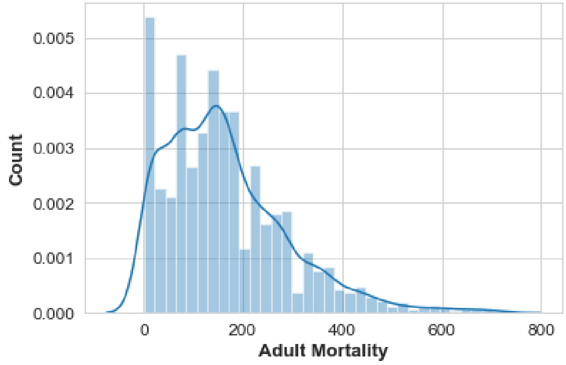
Furthermore, several statistics were done by plotting graphs for each attribute.

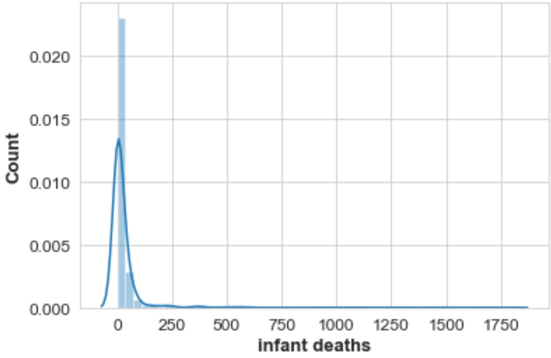


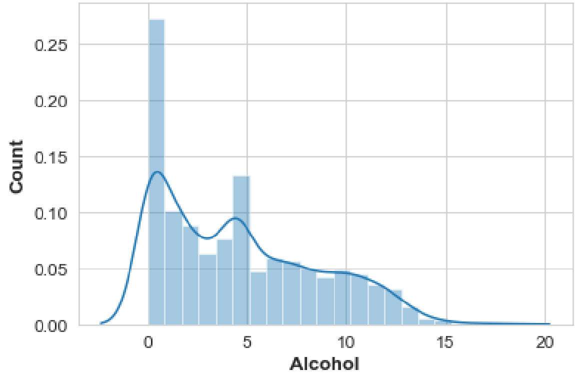


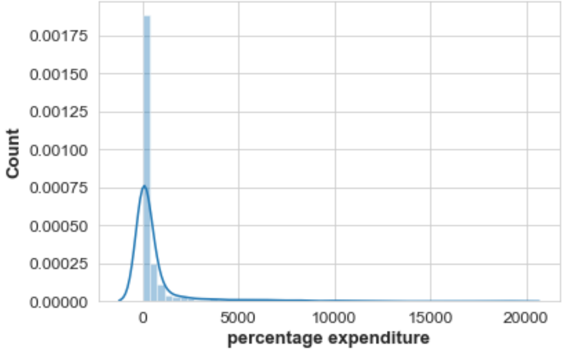

2500
2000
1500
1000
500
0
count
0
1
Status

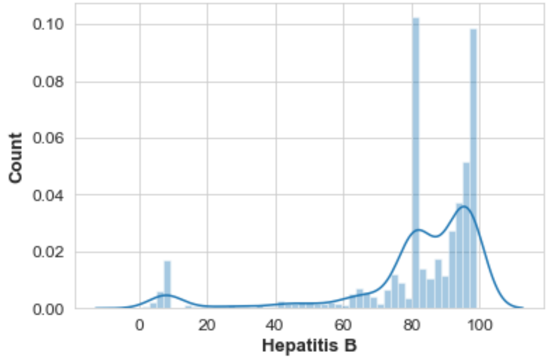

0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
Count
30
40
50
60
70
80
90
Life expectancy

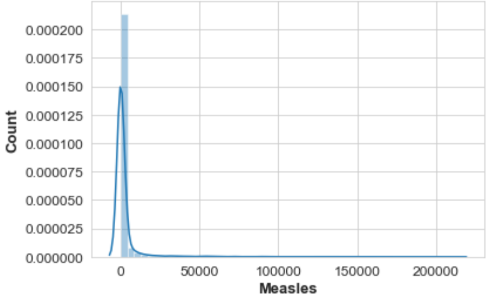

0.005
0.004
0.003
0.002
0.001
0.000
Count
0
200
400
600
800
Adult Mortality

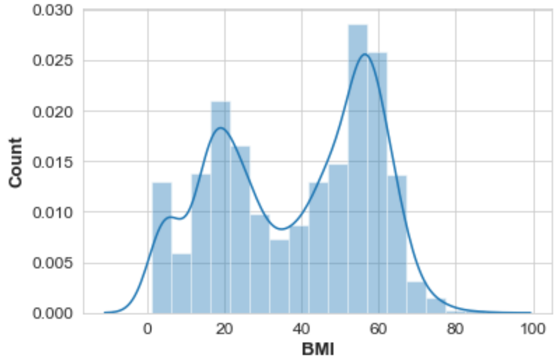

0.020
0.015
0.010
0.005
0.000
Count
0
250
500
750
1000
1250
1500
1750
infant deaths

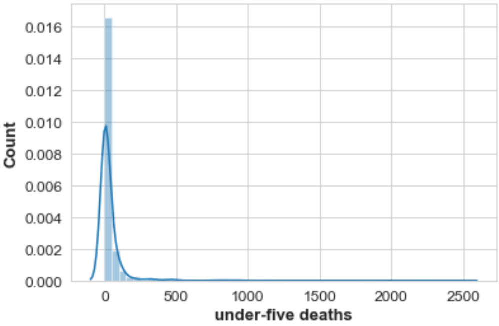

0.25
0.20
0.15
0.10
0.05
0.00
Count
0
5
10
15
20
Alcohol

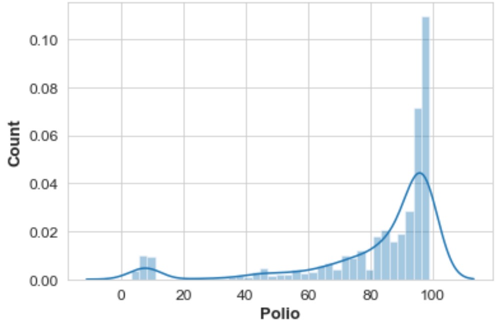

0.00175
0.00150
0.00125
0.00100
0.00075
0.00050
0.00025
0.00000
Count
0
5000
10000
15000
20000
percentage expenditure

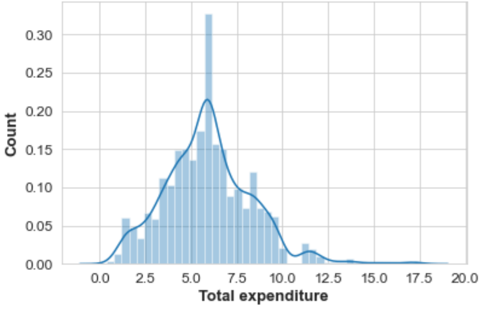

0.10
0.08
0.06
0.04
0.02
0.00
Count
0
20
40
60
80
100
Hepatitis B

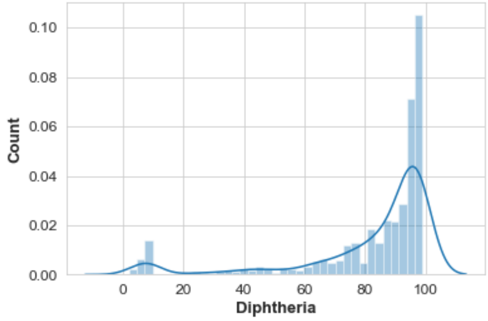

0.000200
0.000175
0.000150
0.000125
0.000100
0.000075
0.000050
0.000025
0.000000
Count
0
50000
100000
150000
200000
Measles

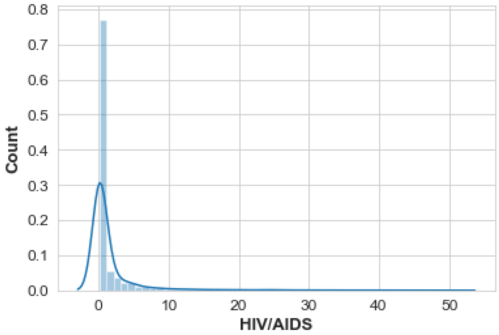

Count
0.030
0.025
0.020
0.015
0.010
0.005
0.000
0
20
40
60
80
100
BMI

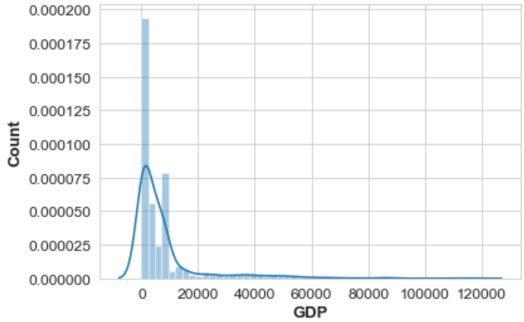

Count
0.016
0.014
0.012
0.010
0.008
0.006
0.004
0.002
0.000
0
500
1000
1500
2000
2500
under-five deaths

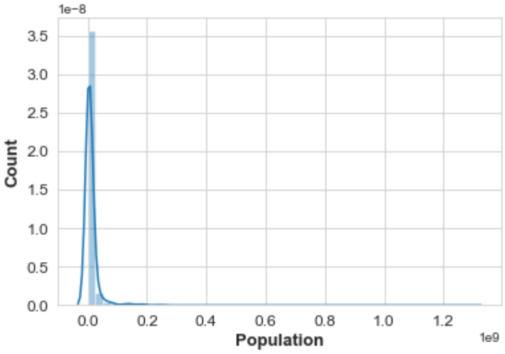

Count
0.10
0.08
0.06
0.04
0.02
0.00
0
20
40
60
80
100
Polio

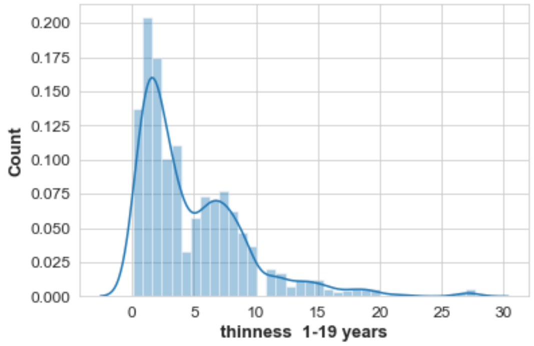

Count
0.30
0.25
0.20
0.15
0.10
0.05
0.00
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
Total expenditure

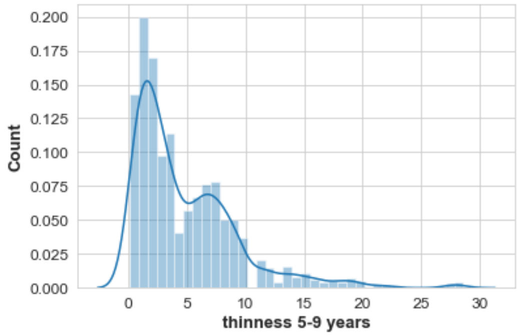

Count
0.10
0.08
0.06
0.04
0.02
0.00
0
20
40
60
80
100
Diphtheria

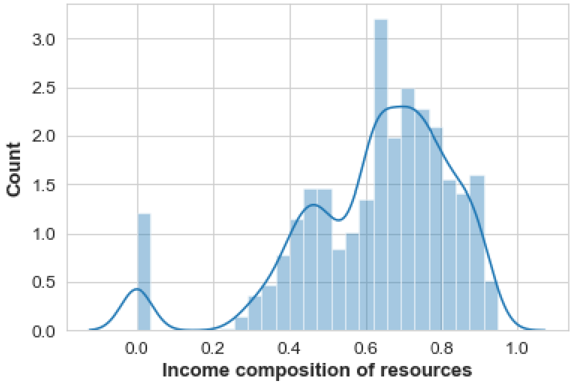

Count
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0
10
20
30
40
50
HIV/AIDS

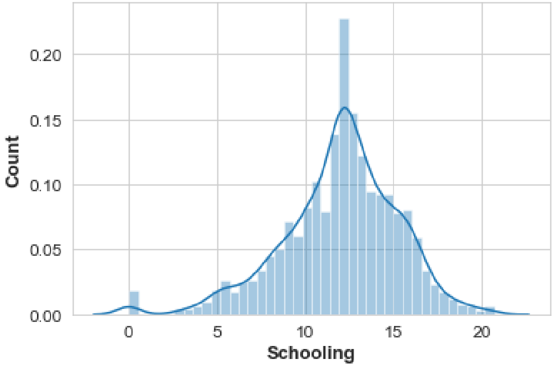

Count
0.000200
0.000175
0.000150
0.000125
0.000100
0.000075
0.000050
0.000025
0.000000
0
20000
40000
60000
80000
100000
120000
GDP


1e−8
Count
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1e9
Population

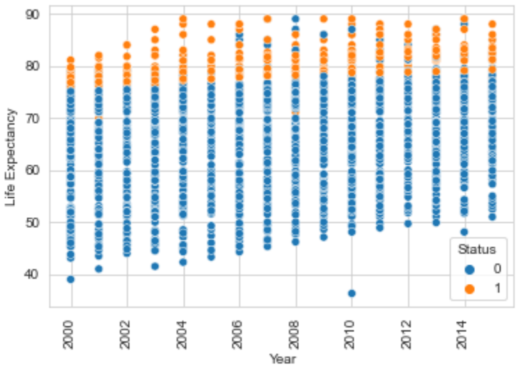


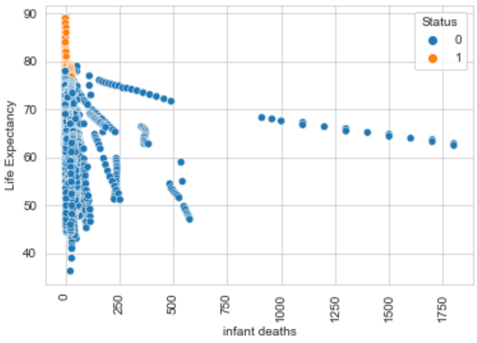


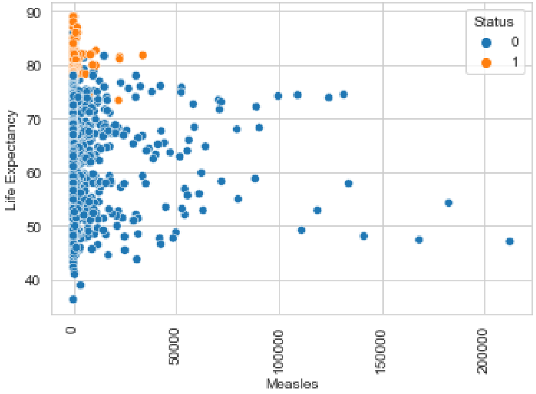


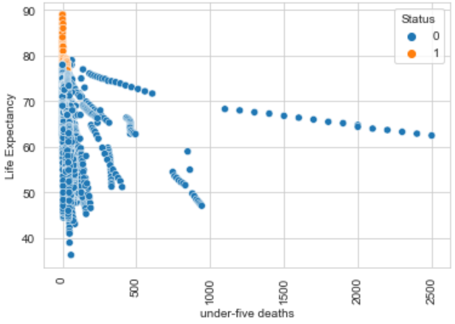


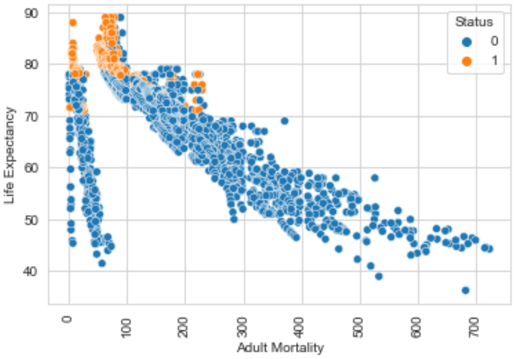
In this project, we have plotted graphs between various attributes and the target. Analysis was done on the graph plots given below.

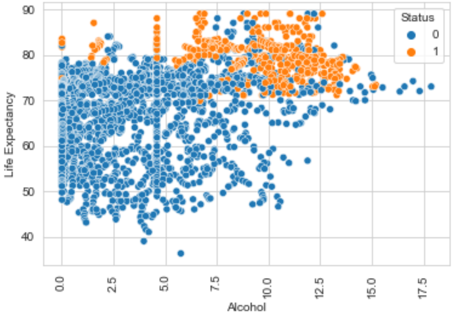


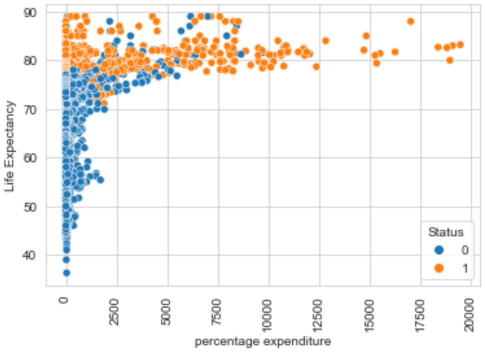


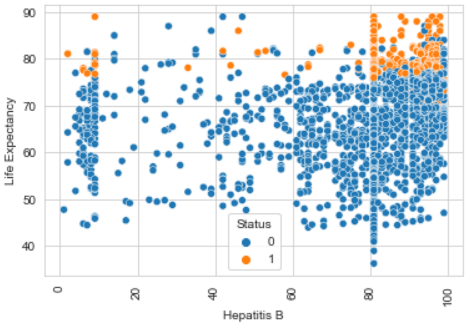


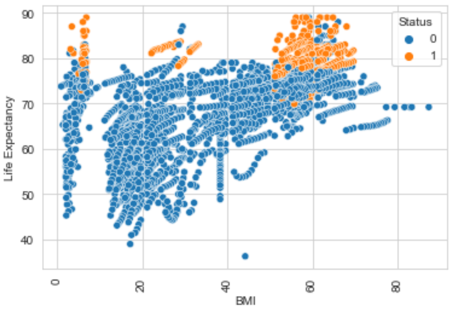


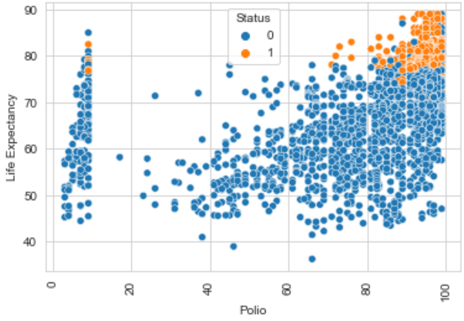

Status
0
1
Life Expectancy
90
80
70
60
50
40
0
100
200
300
400
500
600
700
Adult Mortality

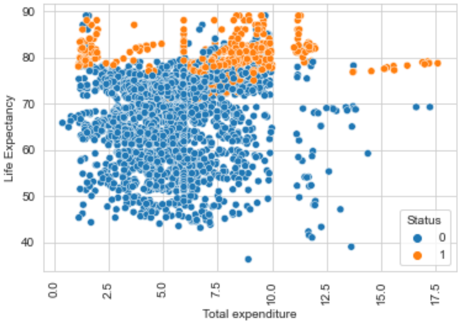

Status
0
1
Life Expectancy
90
80
70
60
50
40
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
Alcohol

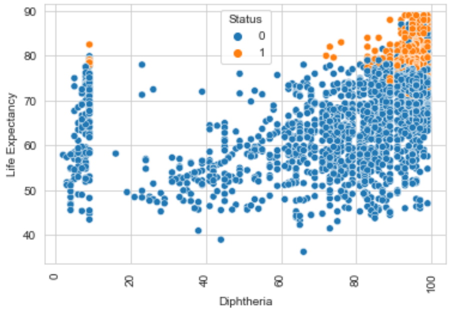

Status
0
1
Life Expectancy
90
80
70
60
50
40
0
2500
5000
7500
10000
12500
15000
17500
20000
percentage expenditure

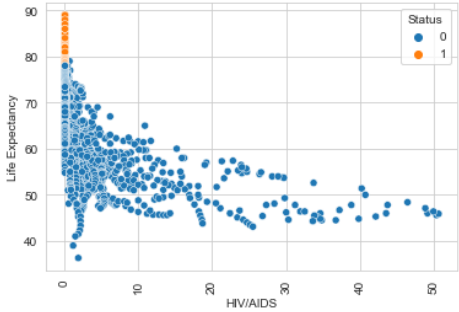

Status
0
1
Life Expectancy
90
80
70
60
50
40
0
20
40
60
80
100
Hepatitis B

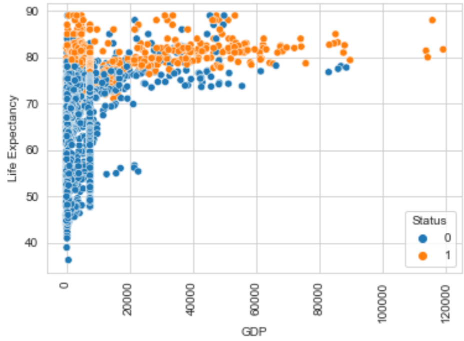

Status
0
1
Life Expectancy
90
80
70
60
50
40
0
20
40
60
80
BMI

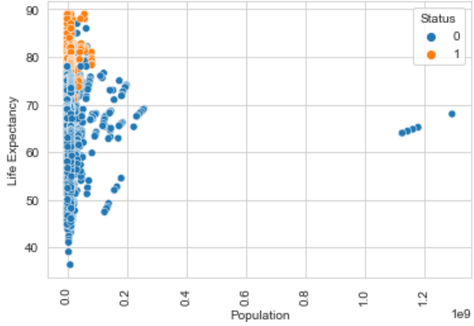

Status
0
1
Life Expectancy
90
80
70
60
50
40
0
20
40
60
80
100
Polio

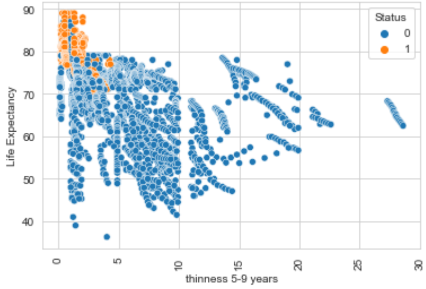

Status
0
1
Life Expectancy
90
80
70
60
50
40
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
Total expenditure

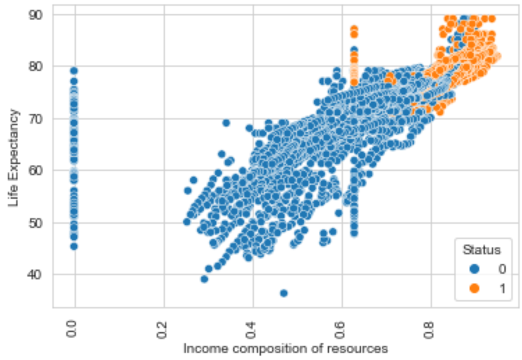

Status
0
1
Life Expectancy
90
80
70
60
50
40
0
20
40
60
80
100
Diphtheria

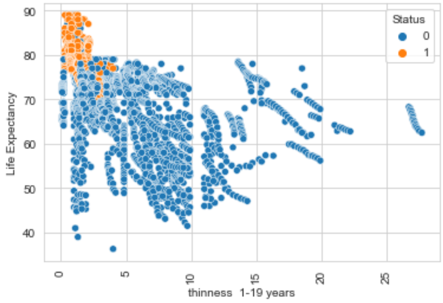


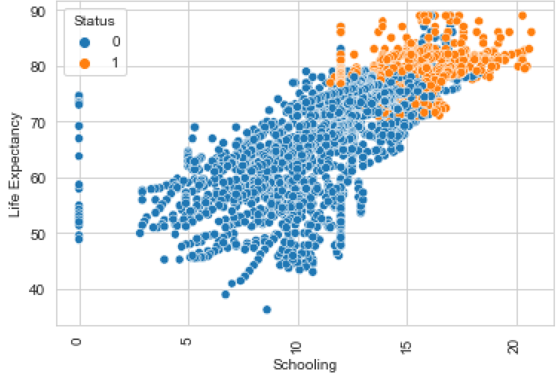




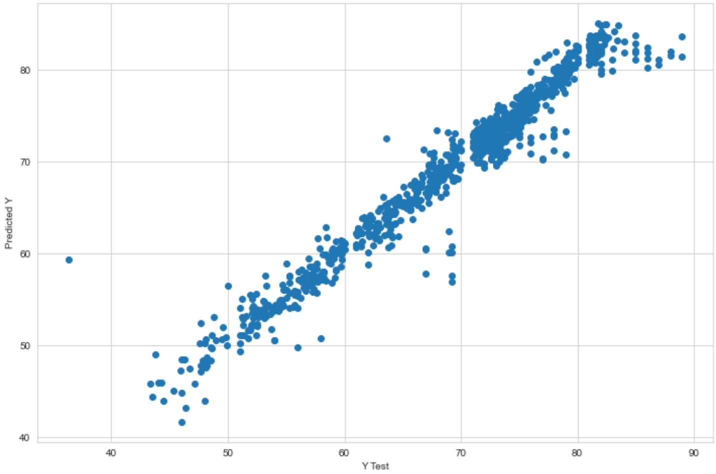




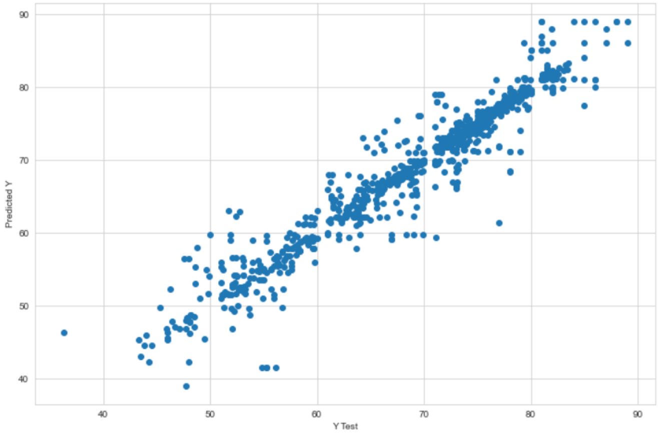




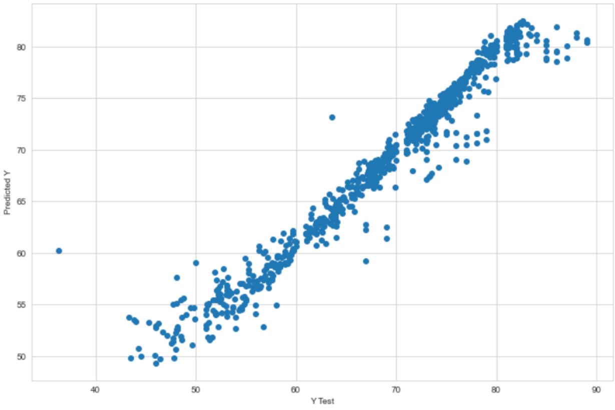
The linear regression model was applied to the dataset, and the characteristics were altered while leaving the label of life expectancy fixed, resulting in the graph below.



The Decision Tree was applied to the dataset, and a graph between predicted and actual values were found.



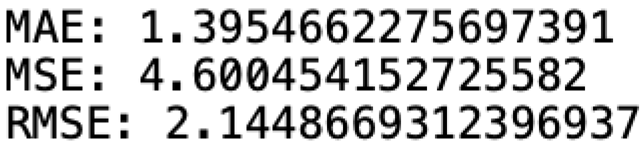
SVM was applied to the dataset, and a graph between predicted and actual values were found



Conclusion

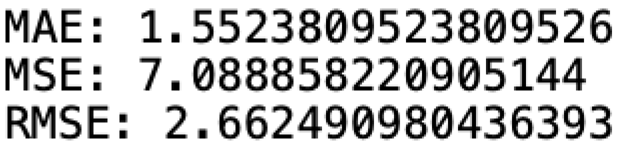
We conclude that the features that have the greatest impact on life expectancy are Adult mortality rate, Percentage expenditure and total expenditure on healthcare and treatments, Hepatitis B, Polio, Under 5 death rate, Measles, Population, GDP, HIV/AIDS, Schooling, Income composition, BMI, and Alcohol consumption rate, based on the given dataset. The use of multiple linear regression on the dataset yielded good results in forecasting the life expectancy of the global population as well as the continents.

Using Linear Regression, an R2 score of **95.503%** was found.



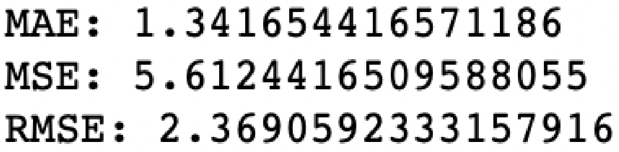
```
MAE: 1.3954662275697391
MSE: 4.600454152725582
RMSE: 2.1448669312396937
```

Using the Decision tree, an R2 score of **92.141%** was found.



```
MAE: 1.5523809523809526
MSE: 7.088858220905144
RMSE: 2.662490980436393
```

Using the SVM, an R2 score of **93.994%** was found.



```
MAE: 1.341654416571186
MSE: 5.6124416509588055
RMSE: 2.3690592333157916
```

These high accuracies indicate that the expected outcome was almost correct.