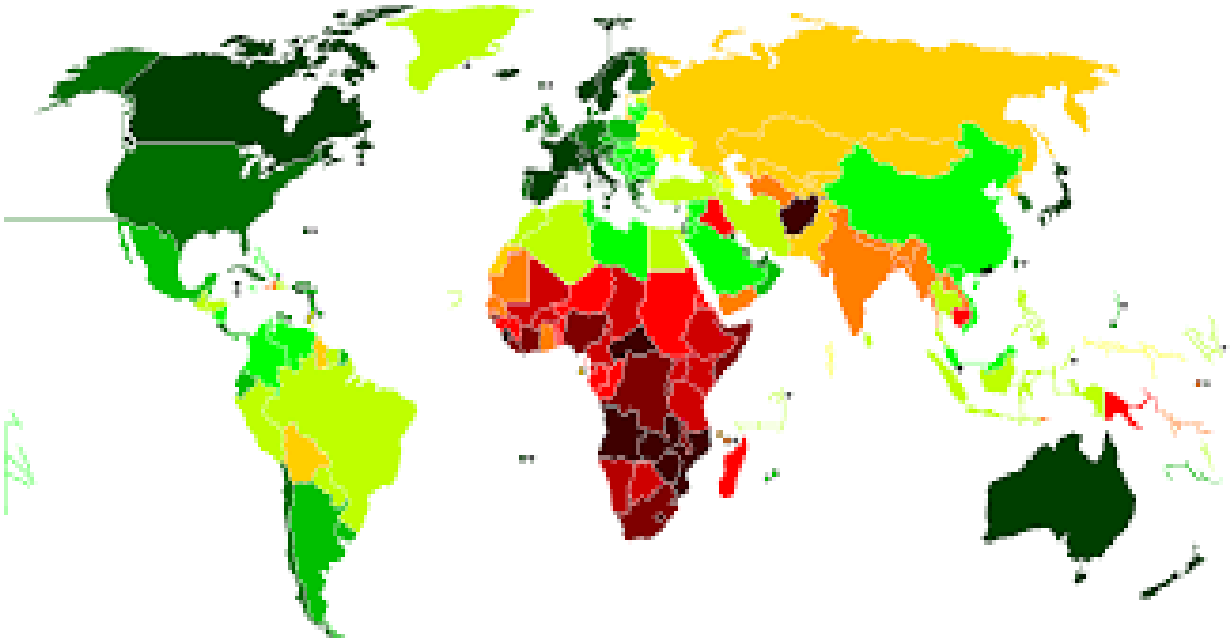
Predicting Life Expectancy using Machine Learning



By

Bhavya singh

Project ID: SPS_PRO_215

GitHub Title: IISPS-INT-1962-Predicting-Life-Expectancy-using-Machine-Learning

Contents

Overview

A lot of studies were done in the past on factors affecting life expectancy considering demographic variables, income composition and mortality rates. It was concluded 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 data set of one year for all the countries. Hence, this gives motivation to resolve both the factors stated previously by formulating a regression model based on mixed effects model and multiple linear regression while considering data from a period of 2000 to 2015 for all the countries. Important immunization like Hepatitis B, Polio and Diphtheria will also be considered. In a nutshell, this study will focus on immunization factors, mortality factors, economic factors, social factors and other health related factors as well. Since the observations this dataset are based on different countries, it will be easier for a country to determine the predicting factor which is contributing to lower value of life expectancy. This will help in suggesting a country which area should be given importance in order to efficiently improve the life expectancy of its population.

Purpose

This project focusses on building a regression model in order to predict the Life Expectancy of a country in a given year based of several other factors such as 'Status', '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'.
With the help of machine learning this project intends to make useful Prediction for a country so that they can work in the correct direction in order to increase their country's Life expectancy.

LITERATURE SURVEY

Existing Problem

In the past various studies have been conducted for predicting the life expectancy, but factors like immunization, and financial factors were not taken into account which led to wrong predictions.

Life expectancy is the average number of years that a person is expected to live. Its 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.

Therefore it is important to correctly predict life expectancy of Country.

Proposed Solution

On performing Exploratory data analysis it was concluded that random forest regression would be the apt choice as it give the highest level of accuracy. Random forest is a flexible, easy to use machine learning algorithm that produces, even without hyper-parameter tuning, a great result most of the time. It is also one of the most used algorithms, because of its simplicity and diversity.

Theoretical Analysis And Experimental Investigations

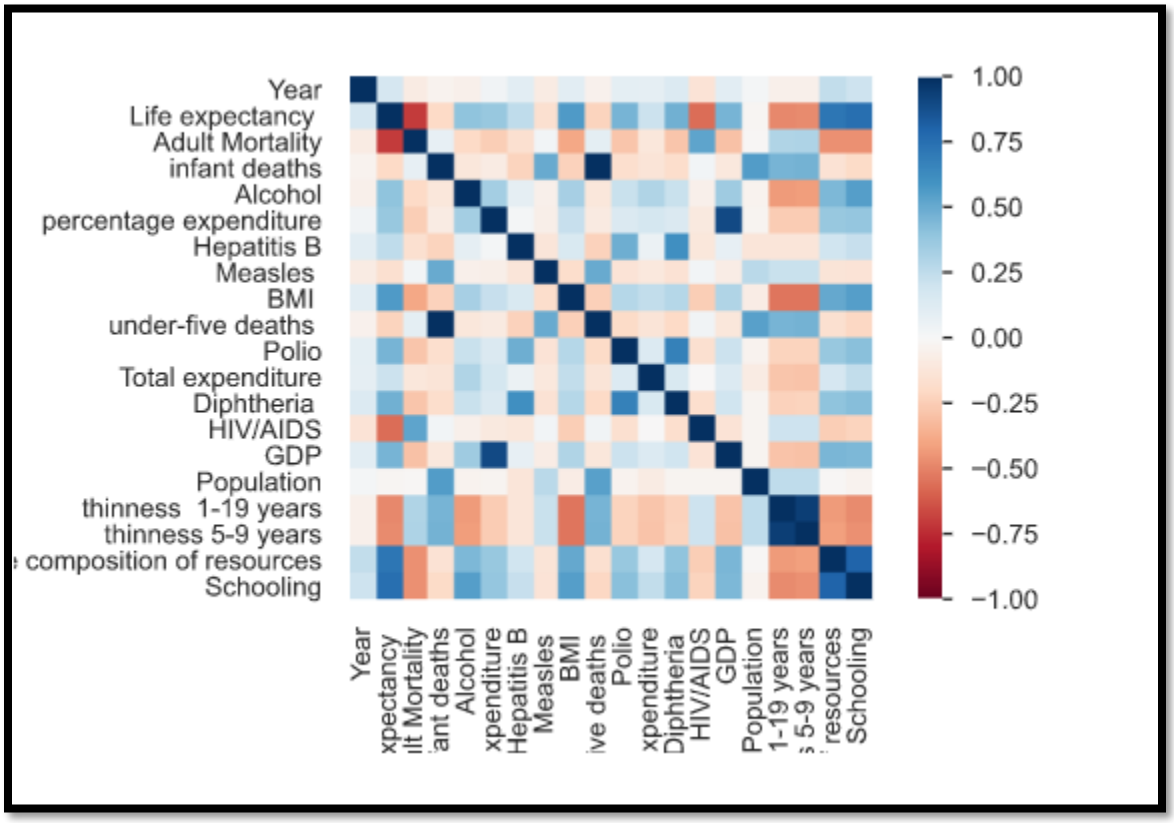
Dataset

Country- Country
Year- Year
Status- Developed or Developing status
Life Expectancy- Age(years)
Adult Mortality- Adult Mortality Rates of both sexes(probability of dying between 15&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)
Percent Expenditure- Expenditure on health as a percentage of Gross Domestic Product per capita(%)
Hep B- Hepatitis B (HepB) immunization coverage among 1-year-olds(%)
Measles- number of reported measles cases per 1000 population
BMI- Average Body Mass Index of entire population
U-5 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
Thinness 10-19- Prevalence of thinness among children and adolescents for Age 10 to 19
Thinness 5-9(%)- Prevalence of thinness among children for Age 5 to 9(%)
Income Composition-Human Development Index in terms of income composition of resources(0-1)
Schooling- Number of years of Schooling

Statistical Analysis

	year	life_expectancy	adult_mortality	alcohol	bmi	polio	diphtheria	hiv/aids	gdp	thinness__1_19_years
count	2938.000000	2938.000000	2938.000000	2938.000000	2938.000000	2938.000000	2938.000000	2938.000000	2938.000000	2938.000000
mean	2007.518720	69.224932	164.796448	4.602861	38.321247	82.550188	82.324084	1.742103	7483.158469	4.839704
std	4.613841	9.507640	124.080302	3.916288	19.927677	23.352143	23.640073	5.077785	13136.800417	4.394535
min	2000.000000	36.300000	1.000000	0.010000	1.000000	3.000000	2.000000	0.100000	1.681350	0.100000
25%	2004.000000	63.200000	74.000000	1.092500	19.400000	78.000000	78.000000	0.100000	580.486996	1.600000
50%	2008.000000	72.000000	144.000000	4.160000	43.000000	93.000000	93.000000	0.100000	3116.561755	3.400000
75%	2012.000000	75.600000	227.000000	7.390000	56.100000	97.000000	97.000000	0.800000	7483.158469	7.100000
max	2015.000000	89.000000	723.000000	17.870000	87.300000	99.000000	99.000000	50.600000	119172.741800	27.700000

Correlations



Overview of the Dataset

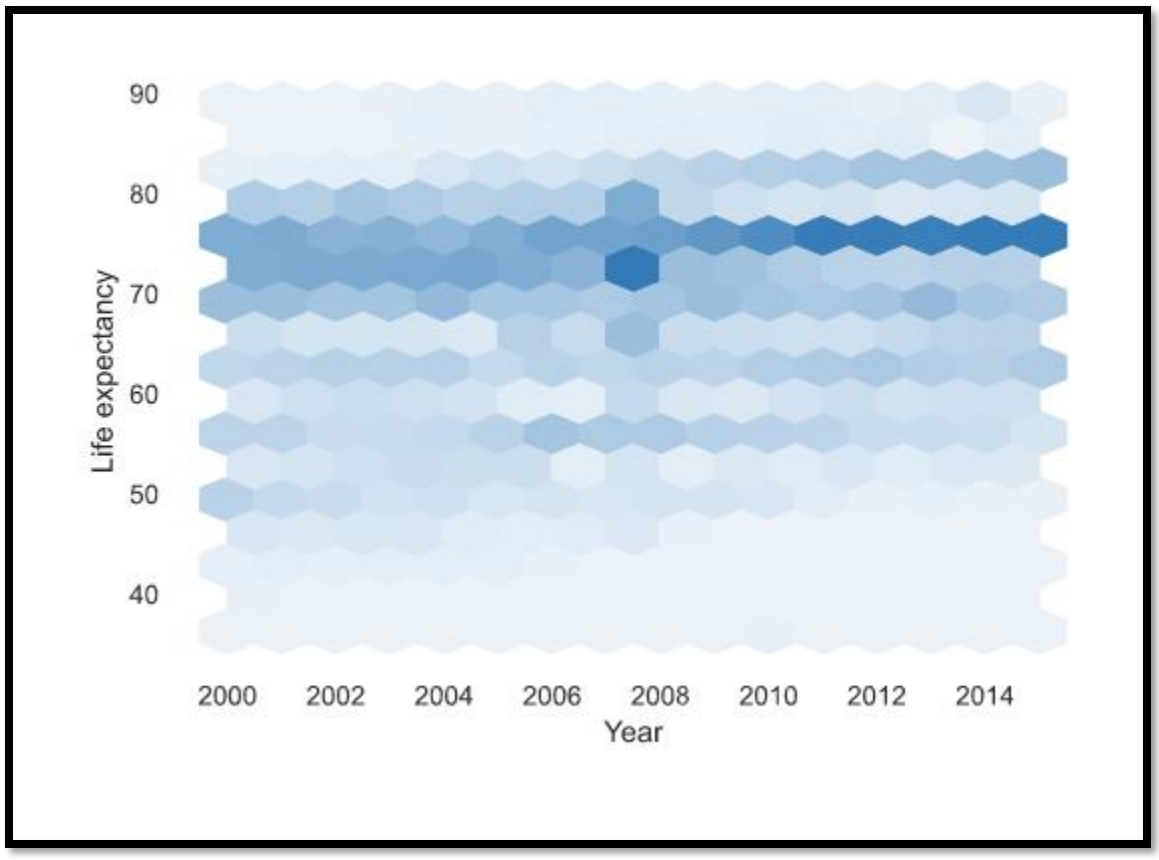
Dataset statistics

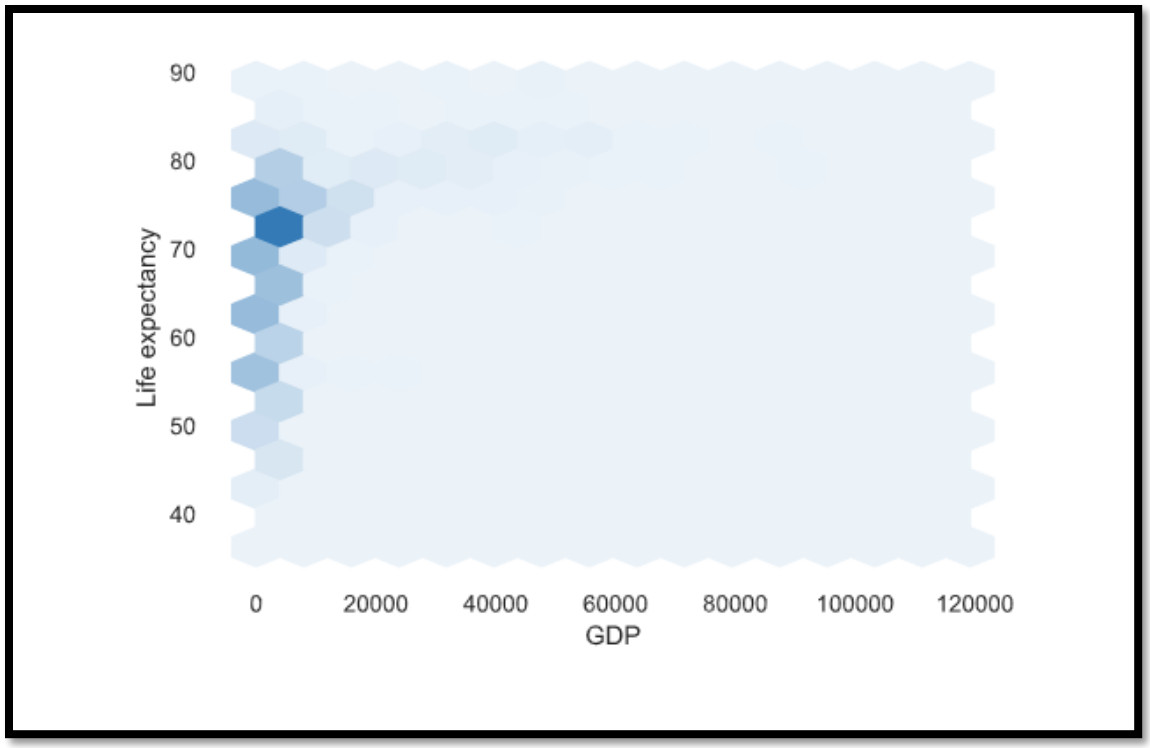
Number of variables	22
Number of observations	2938
Missing cells	2563
Missing cells (%)	4.0%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	505.1 KiB
Average record size in memory	176.0 B

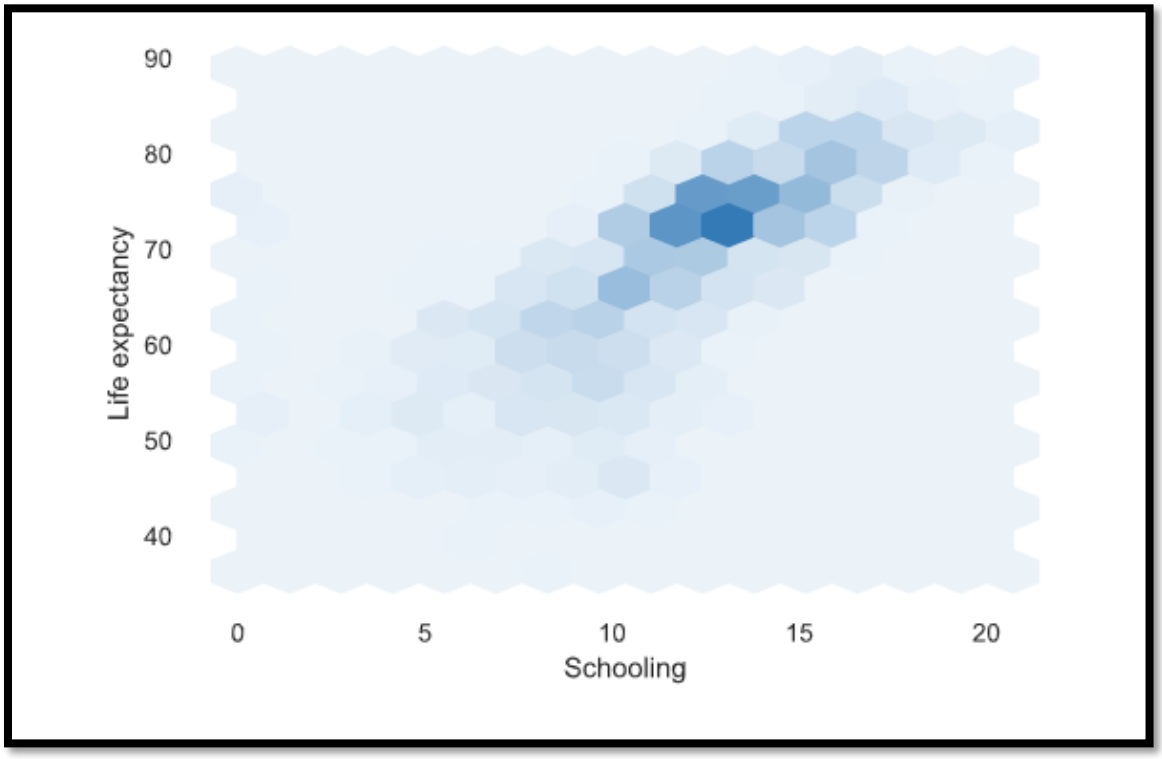
Variable types

NUM	20
CAT	2

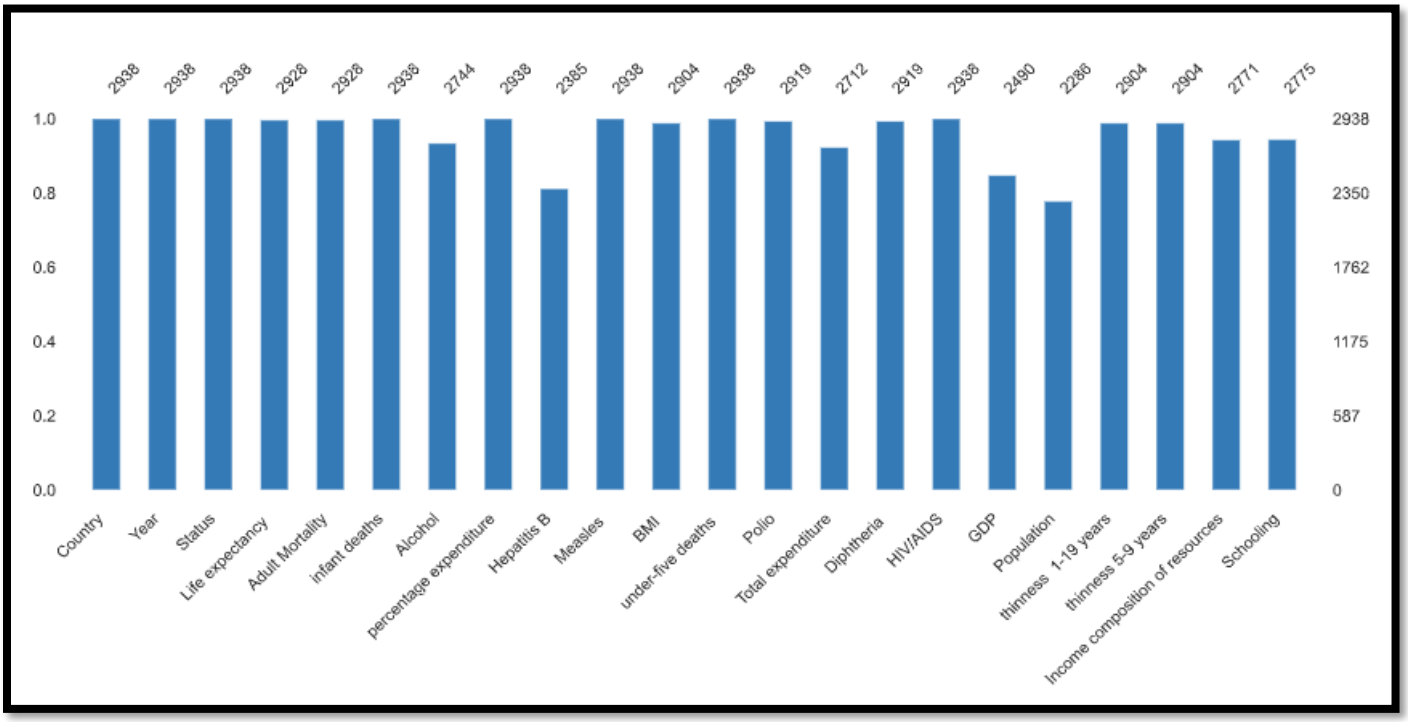
Some Important Plots



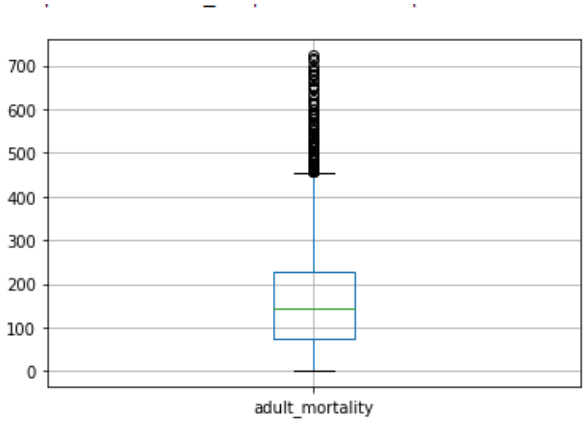

90
80
70
60
50
40
Life expectancy
0
20000
40000
60000
80000
100000
120000
GDP

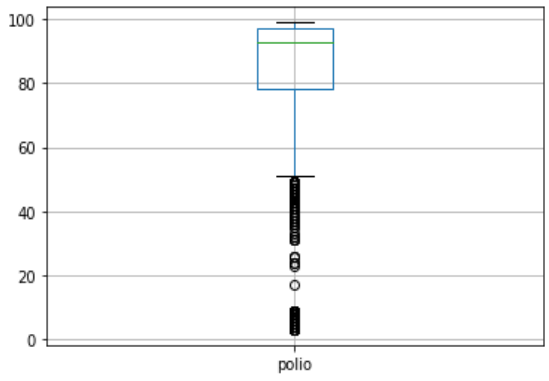

90
80
70
60
50
40
Life expectancy
0
5
10
15
20
Schooling

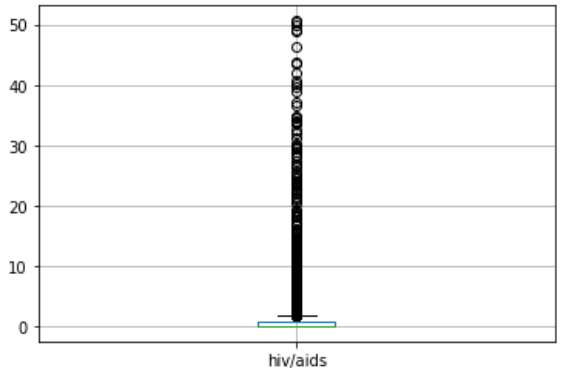
Missing value Analysis

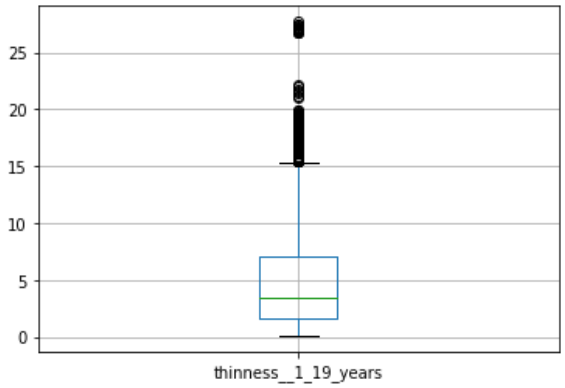


Outlier analysis



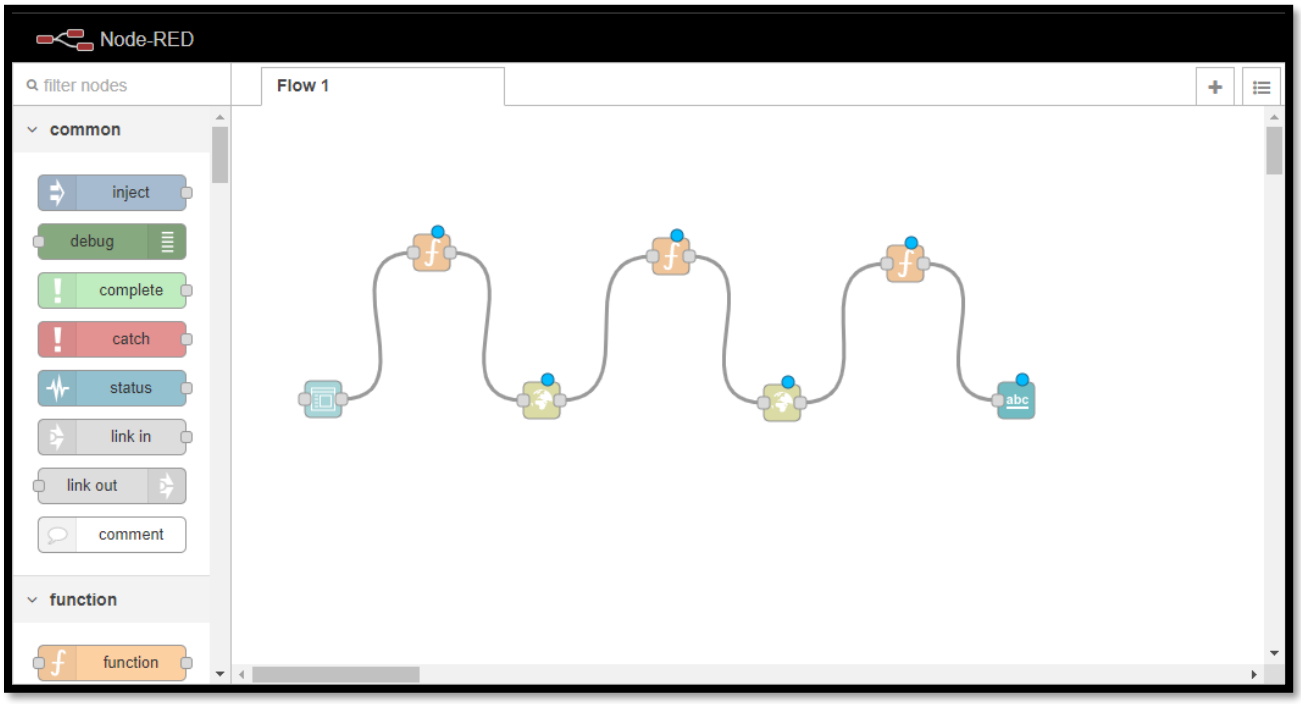






Flowcharts

Node Red flow



Results

After removing all the missing values and dropping less correlated columns and standardizing the data

The scores were

Mean Squared Error: 3.8521721182038737
R2 Score: 95.7687707126783

Future Scope

- Look at class within a particular country and see if these same factors are sa me in determining life expectancy for an individual.
- Use the Twitter API to incorporate NLP analysis for a country to see how it r elates to Life Expectancy.
- Increase the dataset size with continuing UN and Global Data to incorporate new added features like population, GDP, environmental, and etc in order to test and clarify country groupings.
- Mental Health versus Life Expectancy

Bibliography

https://machinelearningmastery.com/

https://www.kaggle.com/

https://towardsdatascience.com/machine-learning/home

https://www.geeksforgeeks.org/machine-learning/

https://scikit-learn.org

Appendix

Source Code