PROJECT REPORT

Predicting Life Expectancy Using Machine Learning

Category: Machine Learning

Abstract

This project analyses the health across several countries worldwide. Data was collected from the World Health Organization as well as the World Bank. The data sets collected contain variables for nation population, GNI per capita (PPP), poverty headcount ratio at $1.00, life expectancy at birth for males and females as well as the averages between the two, the expenditure on health per capita, the completion rate of secondary education, physicians per 1000 individuals as well as the number of hospital beds per 1000, and the adequacy of social protection (Social Security). Regressions on between these variables show whether or not the variables are correlated as well as what the degree of correlation. This regression will then give insight as to how strongly health is affected by the world's varying societal factors. The motivation for seeking this information is that we are interested in how different elements affect the health of a population.

Introduction

As the current era of relative peace presses on, quite a number of people have become increasingly interested in maintaining it for the entire global community. Thus came the inception of the United Nations' Sustainable Development Goals (SDGs) in an attempt to secure this future for the world (or at least its member nations). One of such goals is that of the third one, good health and well-being, the pursuit of seeking relatively healthy lives and wellbeing for every person at any age. However in order to ensure these things, one must first understand what factors may affect them and to what degree. Although a great many factors can be said to affect the health and wellbeing of a population, it is only realistic to cover a comparatively small number of such factors for the sake of statistical analysis. Thus the question arises: what exactly are the factors that have such an effect on a population? In order to account for this, variables must be chosen and tested for correlation against other variables that will be used as a relative measure of health and/or wellbeing. The variables we will be testing for the sake of finding this information will be: Gross National Income per capita as Purchasing Power Parity, because there is the belief that the number of goods an individual can buy will undoubtedly affect their health and wellbeing; the poverty headcount ratio at $1.00, due to the understanding that poverty can be detrimental to a person's well-being; life expectancy, because this value serves as a good measure of how healthy the members of a population are at a given time; the expenditure on health per capita, because it is important to know how a nation's government is making an attempt to remedy its citizens' health in this test; completion rate of secondary education, because schools are a place where one can learn about health; the number of physicians per 1000 individuals as well as the number of hospital beds per 1000 individuals, because it takes into account the proportion of the population can access medical services at any given time; and the adequacy of social protection (Social Security), because it is a good measure of how much the government is putting into

supporting its citizens with no or inadequate income. If one tests these variables for correlation with a multiple linear regression one may believe that one will come closer to understanding what factors affect the health and wellbeing of a population

Problem Description

A typical Regression Machine Learning project leverages historical data to predict insights into the future. This problem statement is aimed at predicting Life Expectancy rate of a country given various features.

Life expectancy is a statistical measure of the average time a human being is expected to live, Life expectancy depends on various factors: Regional variations, Economic Circumstances, Sex Differences, Mental Illnesses, Physical Illnesses, Education, Year of their birth and other demographic factors. This problem statement provides a way to predict average life expectancy of people living in a country when various factors such as year, GDP, education, alcohol intake of people in the country, expenditure on healthcare system and some specific disease related deaths that happened in the country are given.

SOLUTION

After learning deep data exploration and many other tools on Python, now time to have a further step on Regression. Machine learning helps us to have a lot of models with different degrees and choices.

In order to make regression models we need to use a lot of libraries and tools like statsmodels, Linear Regression and train test split from sklearn besides Pandas, Numpy, Matplotlib, etc. in Python.

I will use some Variance Analysis in Regression models in order to determine whether regression models are accurate or misleading.

Following a flawed model is a bad idea, so it is important that we can quantify how accurate our model is.

I made this research based on Life Expectancy data set which is published by The Global Health Observatory (GHO) data repository under World Health Organization (WHO) keeps track of the health status, as well as many other related factors for all countries. The datasets are made available to the public for the purpose of health data analysis.

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 Nation website.

We will see our values from the year 2000 till 2015 for 193 countries

In order to have an accurate filling on missing values and a clear view of the data external data set has been merged together into a single dataset.

I examined adjusted R squared value of test group due to having different number of variables on each model and, plus I consider MSE on test group for my decisions. I kept 'random_state = 0' to have the same number for each step Firstly, I check out for information on our values by checking the missing values for each column Before starting for data exploration and filling, I preferred to merge external data frame for further steps it is very helpful to use left_on and right on codes to connect countries to the correct regions.

For the further step, I decided to drop the population column as having a lot of missing population values in many countries. However, having GDP values from the population for each country can help us as well. We also have a status (Developed or Developing) for each country. Therefore, I preferred to drop a column from the data frame . We have a lot of missing population values in many countries. However, having GDP values from the population for each country can help us as well. We also have a status (Developed or

Developing) for each country. Therefore, I preferred to drop a column from the data frame.

Cleaning Row Data:

In order to fill missing values, it is important to check column types. Some data needs to be filled by mean as time series. On the other hand, this data set has missing values on the country base. Therefore, using the interpolate method helped me to keep the trend of values. Interpolate method provides many different options to deal with missing valuesApplying “limit direction= both” Interpolate method with grouping by Country information, did not help on missing values. It only helped to decrease the number of missing values.On those rows, there is no previous information for relevant countries. Thus, I used the interpolate method with grouping by sub-region and Year information. In the end, my data set is ready for investigation and regression process! The first thing I always do is to check correlations between variables. It gives me the strength to evaluate urn values and meaning behind it. While checking correlations, I advise you to use “abs()” code to have both negative and positive correlations.

‘Schooling’, ‘Income_composition_of_resources’, and ‘Adult Mortality’ have a high correlation between Life Expectancy.

‘HIV/AIDS’, ‘BMI’, ‘Diphtheria’, ‘thinness_1_19_years’, ‘thinness_5_9_years’, ‘Polio’, ‘GDP’, and ‘Alcohol’ have medium correlation between Life Expectancy.

And the rest of our columns; ‘percentage_expenditure’, ’Hepatitis_B’, ‘Total_Expenditure’, ‘under_five_deaths’, ‘infant_deaths’, ‘Year’, and ‘Measles’ have low correlation between Life Expectancy.

As a last step before applying machine learning tools, I also would like to have a general idea about Life Expectancy in years.

BUILDING REGRESSION MODELS

Linear Regression:

Firstly, I divided all numerical variables by splitting the data in two, so out of 100 rows, 80 rows will go into the training set, and 20 rows will go into the testing set. This data frame will be called as "LifeExpectancyData_num" in further steps.

As we see above, R squared value is only %66. Before testing MSE, RMSE values, I also checked for residual distribution as an example.We can also do Jarque Bera Test to be sure about the outcome.

Adding Polynomial Feature:ß0 + ß1 X1 + ß2 X22 + ... + ßi Xni

In case a linear model is not appropriate, and a polynomial may do better, we do use models within Polynomial Degree.

For curiosity, I checked model performance with polynomial degrees to check values again. In the end, we do see test and train values are performing much way better than the OLS model.

Before going further, you need to import PolynomialFeatures from sklearn"from sklearn.preprocessing import PolynomialFeatures"

I recommend you to use 'def' function to have all the lists and return what we need in one row. I added necessary lists of 'number_of_variables', 'R_list', "adj_R_test'' , 'R_train_list', 'adj_R_train', 'MSE_list_test', 'MSE_train_list', 'MAE_list', 'RMSE_list' and 'MAPE_list' for polynomial models. After applying 3 degrees of polynomials here are the results:

MSE Test Values in pltplot:Showing MSE Test and Train values with barplot on Python:Here we see adjusted R squared values for each model as well.I also checked distributions of MSE Train and Test values with a scatter plot.There are two critical characteristics of

estimators to be considered: the bias and the variance. The bias is the difference between the true population parameter and the expected estimator. It measures the accuracy of the estimates. Variance, on the other hand, measures the spread, or uncertainty, in these estimates.

So, setting λ to 0 is the same as using the OLS, while the larger its value, the stronger is the coefficients' size penalized as λ becomes larger, the variance decreases, and the bias increases. A more traditional approach would be to choose λ such that some information criterion, Akaike or Bayesian(AIC or BIC), is the smallest. A more machine learning-like approach is to perform cross-validation and select the value of λ that minimizes the cross-validated sum of squared residuals.

As we see on scatter plots, True values of Poly 2 model are distributed better than Poly 3 Model on test and train group. Poly 3 Model is not enough to explain some of higher values.

Building Ridge Regression Models

While Least Squares determines values for the parameters in an equation, it minimizes the sum of the squared residuals. On the other hand, Ridge Regression minimizes the sum of the squared residuals plus lambda and squaring the slope of the regression lineMoreover, to ensure I don't overfit my training data, I checked Ridge Model as well.In order to create a Ridge Regression Model, we need to import"from sklearn.linear_model import Ridge" on our notebookI used different alphas to see breakpoint in MSE_list values for both 2nd and 3rd degrees.

The Best Model option with minimum MSE_test value with maximum Adjusted R Squared on Alpha 10^{-5} and polynomial 2nd degreeSame function repeated for the 3rd degree Polynomial and I get results as below

The Best Model option with minimum MSE_test Value on Alpha 1^{03} and polynomial 3rd degree.Here we can see the change between test and train values.

While having the same trend until 125th variable on the Poly 2 MSE results, Poly 3 MSE results shows us that after the 200th variable trend is not good any moreBecause having the low MSE value, I will continue with 2nd polynomial degree ridge Model. Later on, I also compare adjusted R squared values as well.

Building Lasso (least absolute shrinkage and selection operator) Regression Models

While Ridge Regression minimizes the sum of the squared residuals plus lambda and squaring the slope of the regression line, Lasso Regression minimizes the sum of the squared residuals, plus lambda and absolute value of slope of the regression line.While Ridge shrink the parameters by keeping all of them, Lasso Regression eliminates and creates a simpler model to explain.Therefore, I would like to have results of this model as well to have a wide range of elements for my predictionAgain, we need a new library to import in our noteboo“from sklearn.linear_model import Lasso”.

I continued Lasso Model with the same alpha range of [0.000001, 0.00001, 0.0001, 0.001, 0.01, 1, 10, 100, 1000] and two different polynomial degreesThe Best Model option with minimum MSE_test Value on Alpha 10^{-5} and polynomial 2 degree.Hereby, we do see that adjusted R squared values are lower than Ridge Regression, while MSE test values are getting higher.

After searching different type of regression models, we have the minimum MSE and the better adjusted R^2 values from Linear Regression and Ridge Regression on two polynomial degrees test group. Polynomial degree does not affect values on different type of regression models.

Low MSE values and highest adjusted R^2 came from two polynomial degree models. Applying other type of regressions with three polynomial degree only increased MSE Test values to a higher level.

Therefore, I agree to choose the Ridge Regression with two polynomial degree.

After selecting the best model of Ridge Regression with 2 polynomial degree on alpha 0.000001, here we will see the results of our model by applying coefficients on each variable as an example to check our model performance.

OUTPUT OF SOLUTION

```
sns.heatmap(df.corr(), square=True, cmap='RdYlGn')
```

```
Out[151]:
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x17a60265988>
```

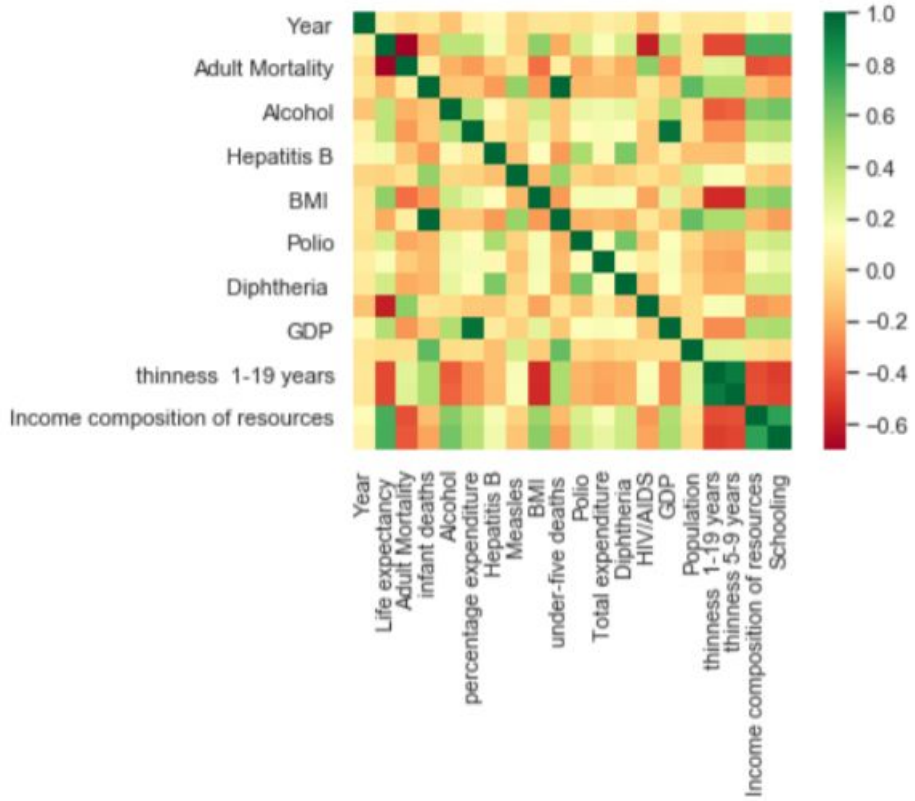


FIG 1: HEAT-MAP OF CSV FILE

```
df.head()
```

Out[141]:

	Country	Year	Status	Life expectancy	Adult Mortality	infant deaths	Alcohol	percentage expenditure	Hepatitis B
0	Afghanistan	2015	Developing	65.0	263.0	62	0.01	71.279624	65.0
1	Afghanistan	2014	Developing	59.9	271.0	64	0.01	73.523582	62.0
2	Afghanistan	2013	Developing	59.9	268.0	66	0.01	73.219243	64.0
3	Afghanistan	2012	Developing	59.5	272.0	69	0.01	78.184215	67.0
4	Afghanistan	2011	Developing	59.2	275.0	71	0.01	7.097109	68.0

5 rows × 22 columns

FIG 2 :- DATA AFTER REMOVING NULL VALUES

```
plt.figure(figsize=(20,8))
plt.scatter(X_gdp, y)
plt.plot(prediction_space, y_pred, color='black', linewidth=3)

plt.xlabel('GDP')
plt.ylabel('Life Expectancy')
plt.show()
```

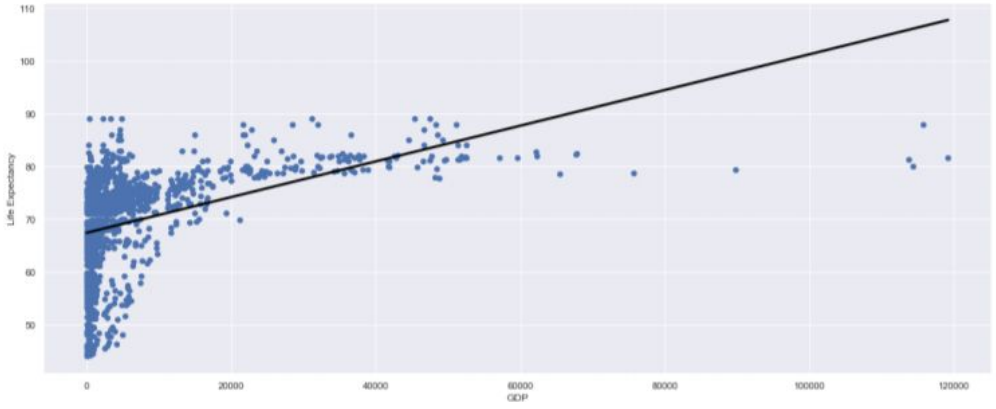


FIG 3 :- LINER REGRESSION OF LIFE -EXPECTANCY AGAINST GDP

In [74]:

```
plt.figure(figsize=(20,8))
plt.scatter(X_POP, y)
plt.plot(prediction_space1, y_pred, color='red', linewidth=3)
plt.xlabel('POPLUTION')
plt.ylabel('Life Expectancy')
plt.show()
```

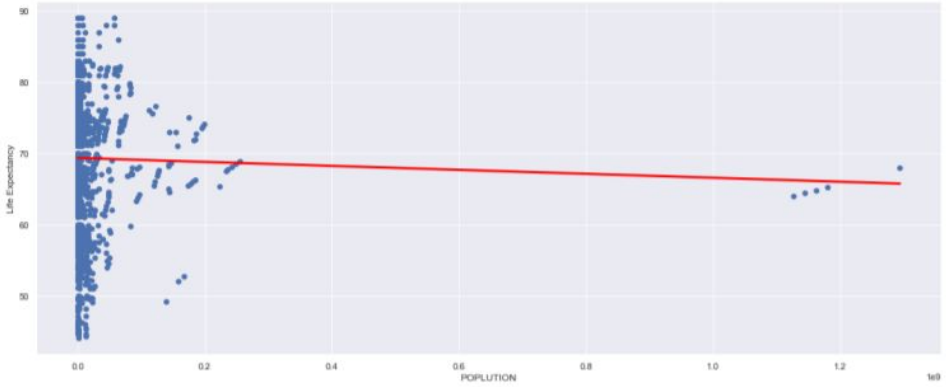


FIG 4 :- LINER REGRESSION OF LIFE -EXPECTANCY AGAINST POPULATION

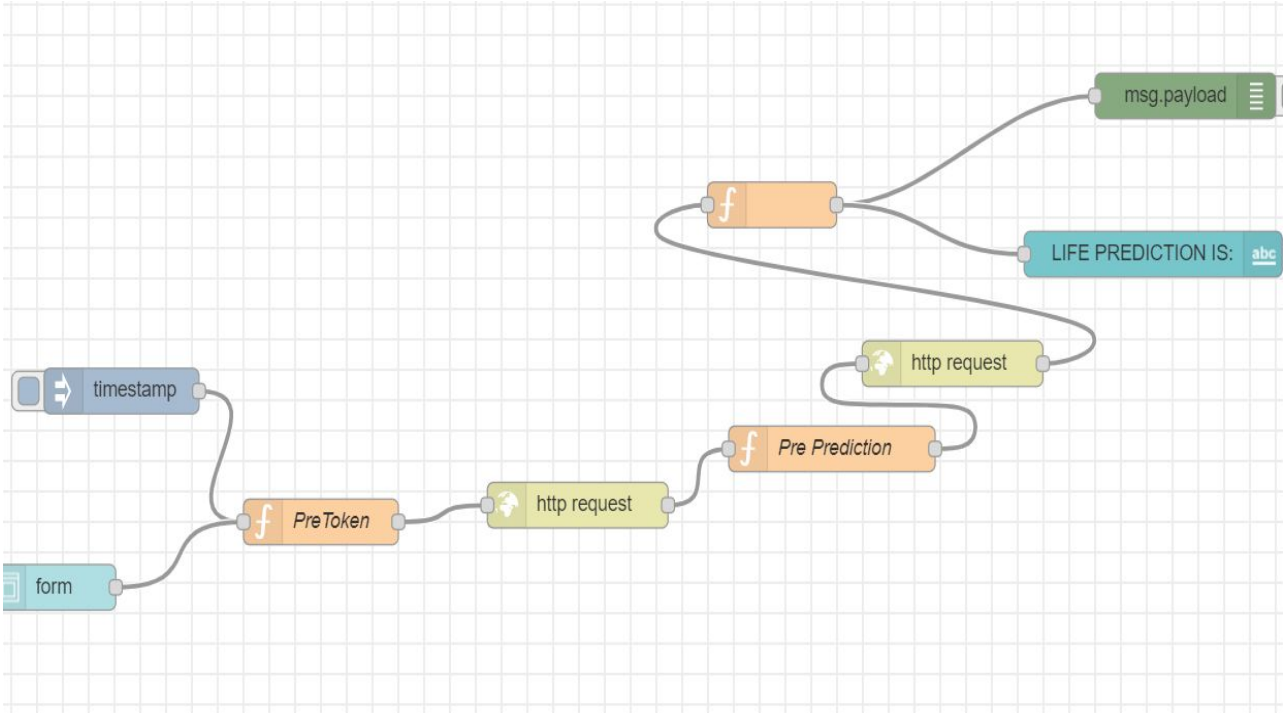


FIG -6: NODE -RED FLOW AND CONNECTION WITH NODE RED

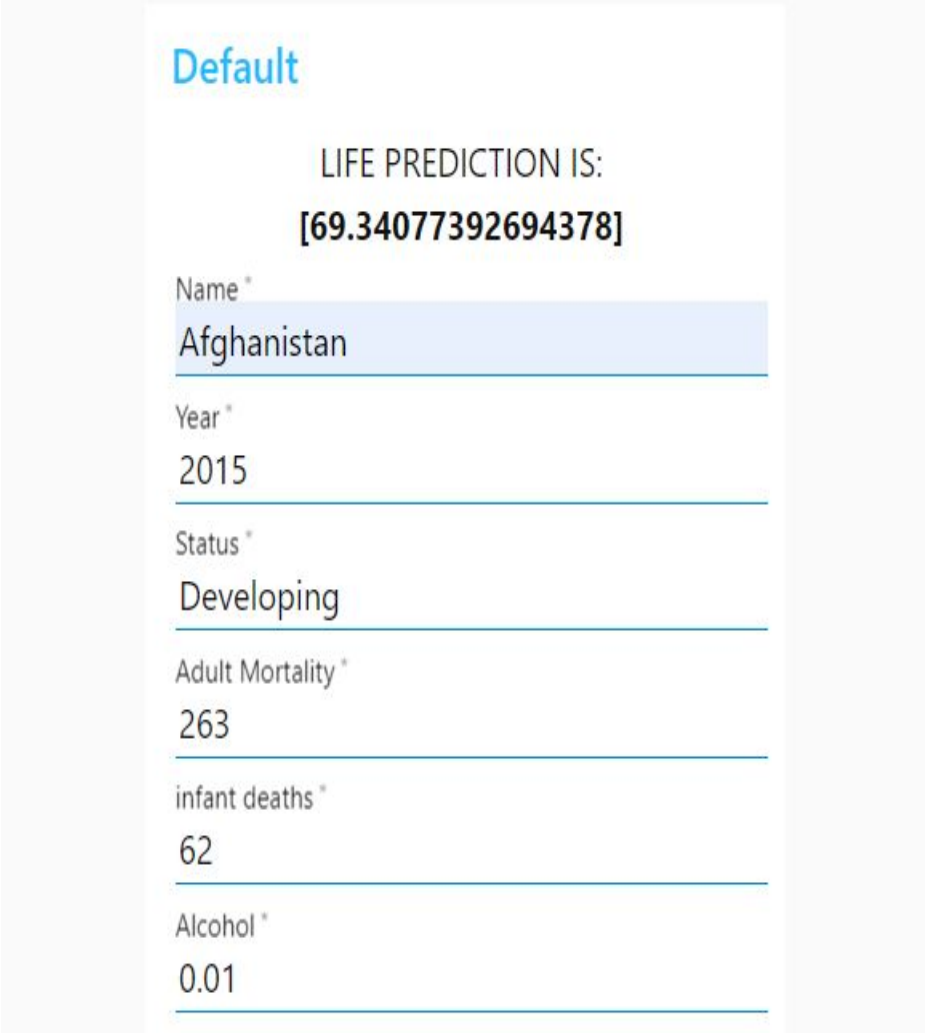
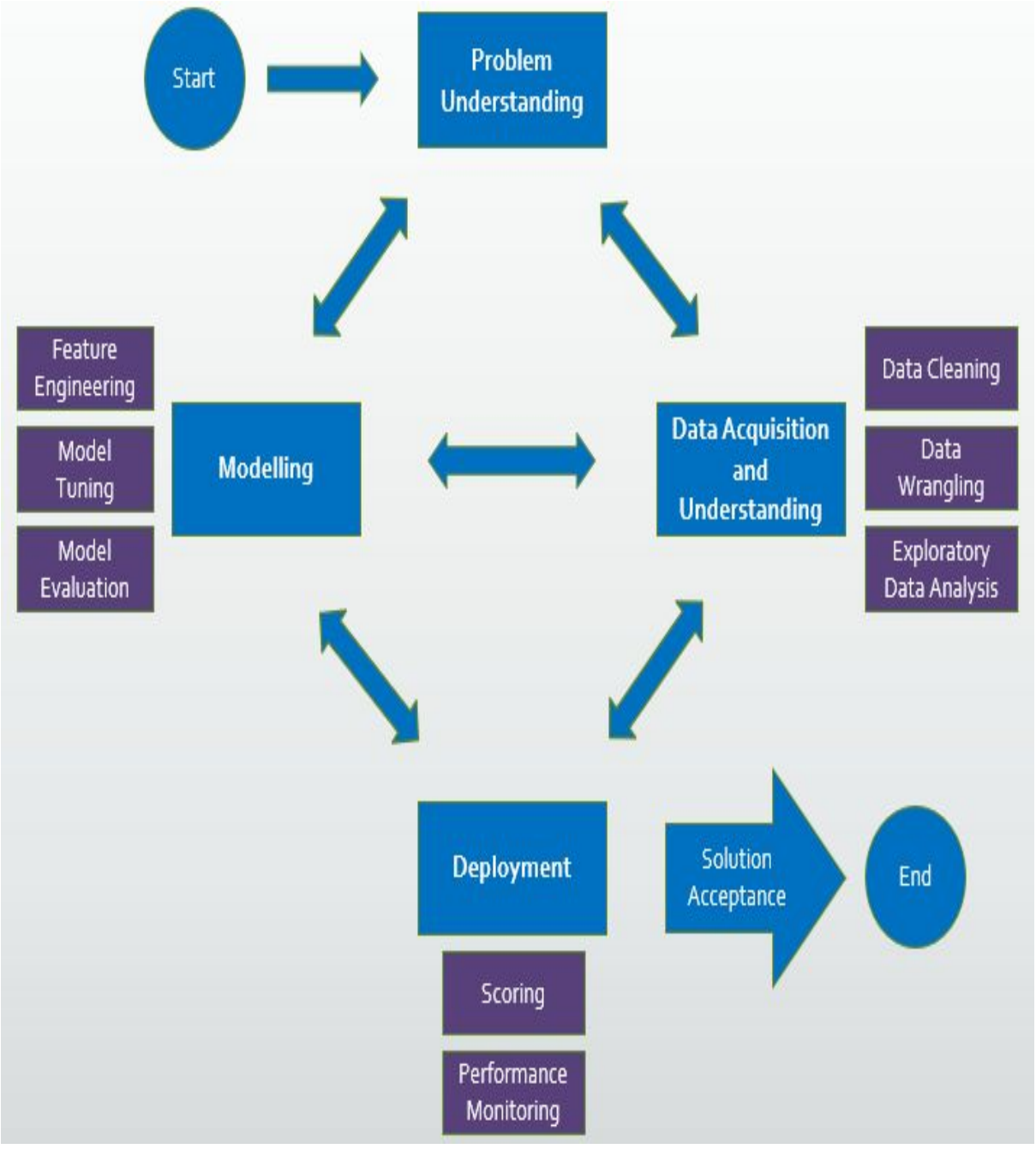


FIG 7:- NODE RED FINAL OUTPUT

FLOW CHART


Start
Problem Understanding
Feature Engineering
Model Tuning
Model Evaluation
Modelling
Data Acquisition and Understanding
Data Cleaning
Data Wrangling
Exploratory Data Analysis
Deployment
Scoring
Performance Monitoring
Solution Acceptance
End

Conclusion

As we see on the graph of this model, best performance is starting after 125th variable.Here are the real values from 5th row belongs to 2010.Let's have a look at our model performance!

Functions shows us that having values as following 'Year': 2010, 'Adult_Mortality': 279.0, 'infant_deaths': 74.0, 'Alcohol': 0.01, 'percentage_expenditure': 79.67936736, 'Hepatitis_B': 66.0, 'Measles': 1989.0, 'BMI': 16.7, 'under_five_deaths': 102.0, 'Polio': 66.0, 'Total_Expenditure': 9.2, 'Diphtheria': 66.0, 'HIV/AIDS': 0.1, 'GDP': 553.32894, 'thinness_1_19_years': 16.6, 'thinness_5_9_years': 6.9, 'Income_composition_of_resources': 0.45, 'Schooling': 9.2, gives the result of Life Expectancy as '61' gives us the result of Life Expectancy value as 61.14.The original value of Life Expectancy was 58.8 in 2010. MSE Test value is 6.367 with average 2.52 of RMSE value. Simply 61 minus 2.52 gives the result with +-2.52 from the real value of Life Expectancy.

Regression models is luckily helping us to predict our dependent variable with using many parameters. In order to have an accurate result, we need to check as many as regression models. Having the lowest MSE and highest adjusted R squared are helping us on our way.

After comparing all the algorithms we can conclude the Lasso Regression offer the best:

Best Parameters: {'alpha': 0, 'max_iter': 10}

R square on the test data of 92%

MAE of 1.83

MSE of 6.05

ADVANTAGES ,DISADVANTAGES AND APPLICATION

- IT IS USED FOR MAKING THE POPULATION HEALTHER.
- HELPING SICK PEOPLE TO LIVE LONGER WITH THEIR SICKNESS.
- HELP TO PREDICT AVERAGE LIFE OF ANY COUNTRY.

Resources

The following sources have been used:

- https://www.kaggle.com/kumarajarshi/life-expectancy-who