: Project Report on :

Predicting Life Expectancy using Machine Learning – SB45406

By

MD SAJJAD ANSARI

Project ID : SPS_PRO_215

Github : IISPS-INT-1519-Predicting-Life-Expectancy-using-Machine-Learning

Table of Content

1. INTRODUCTION

1.1 Project Overview

Life expectancy is a key feature that plays an active as well as passive role in industrial growth all around the world. Health forecasts and alternative future scenarios can serve as vital inputs into long- term planning and investments in health, particularly in terms of framing different choices, their potential effects, and the relative certainty associated with each option. It can help the companies to hire employees and well as governments to look after the factors and try to bring improvement to extend the life span of people of their countries. Understanding potential trajectories in health and drivers of health is crucial to guiding long-term investments and policy implementation. At this situation we must need someone who can tell us the prediction about how more years a person can live and what's their regional life span. As the technology is growing faster than ever and computers are learning better and understanding the case scenarios, we can use its power to expect Life span of a human based on some values of the area/country where they lives such as: whether he/she lives in a Developed country or in a developing one, what's the Adult Morality of that country, what amount of alcohol they take, how many infant dies there and many more.

1.2 Purpose:

Here, this project aims at comparing such above mentioned factors and exploring the relationship between them by using machine learning algorithms like linear regression, random forest, and choose the best or optimum model which will predicts based on the input of various factors an age which will be the expected year people of that particular area can live or the estimate of the average age that members of a that area will be when they die. And that will be his/her "life expectancy".

It is intended to create a Life Expectancy prediction model with a User Interface where user can input the values of dependent features. And Machine Learning model will leverages historical data to predict Life Expectancy. We can use IBM cloud for development and its services like Watson Studio, Machine Learning Service, and Cloudant app Node-RED to design user interface and to deliver prediction.

2. LITERATURE SURVEY

2.1 Existing Problem:

Predicting a human's life expectancy has been a long-term question to humankind. Past work to generate health-focused forecasts includes that from the UN Population Division, and the Austrian Wittgenstein Center, which produces life expectancy forecasts with different scenarios to the end of the 21st century. There are so many organizations that are making research in the prediction of life expectancy. Many calculations and research papers have been done and published to create an equation despite it being impractical to simplify these variables into one equation.

Currently there are various smart devices and applications such as smartphone apps and wearable devices that provide wellness and fitness tracking. Some apps provide health related data such as sleep monitoring, heart rate measuring, and calorie expenditure collected and processed by the devices and servers in the cloud. However no existing works provide the Personalized Life expectancy.
The World Health Organization (WHO) used to produce annual life tables for the countries but after 2011 it said to shift for two year cycle for the updating of life tables and even still the model is not really updated in every fields. WHO applies standard methods to the analysis of Member State data to ensure comparability of estimates across countries. This will inevitably result in differences for some Member States with official estimates for quantities such as life expectancy, where a variety of different projection methods and other methods are used.

2.2 Purpose Solution:

Some of the past research was done considering multiple linear regression based on dataset of one year or two years for all the countries. We can resolve it by formulating a regression model while considering data from a period of year 2000 to 2015 for all the countries. Important immunization like Hepatitis B, HIV/AIDS, Polio and Diphtheria will also be considered. We will also focus on Adult Mortality, Alcohol intake, percentage expenditure, Measles, BMI, Death of under 5 years, Schooling, thinness in 1-19 years and 5-9 years and Population related factors as well. Since the dataset is 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.

Foe the solution, First we will examine the dataset provided by World Health Organization (WHO) and find patterns in a dataset and we will attempt to fit various algorithms such as linear Regression and Random Forest Algorithm, and see what gives less error and good prediction score.

And to pick a good machine learning algorithm we must need to consider to first check the data and solve the problem within it such as : whether if any value is missing or not? If missing can we remove that feature or can we put the mean value there? Is all values are of same type? Does object value pay any contribution? If no then can we remove it or if yes then can we change object values into categorical value such as integer? And other problematic factors.

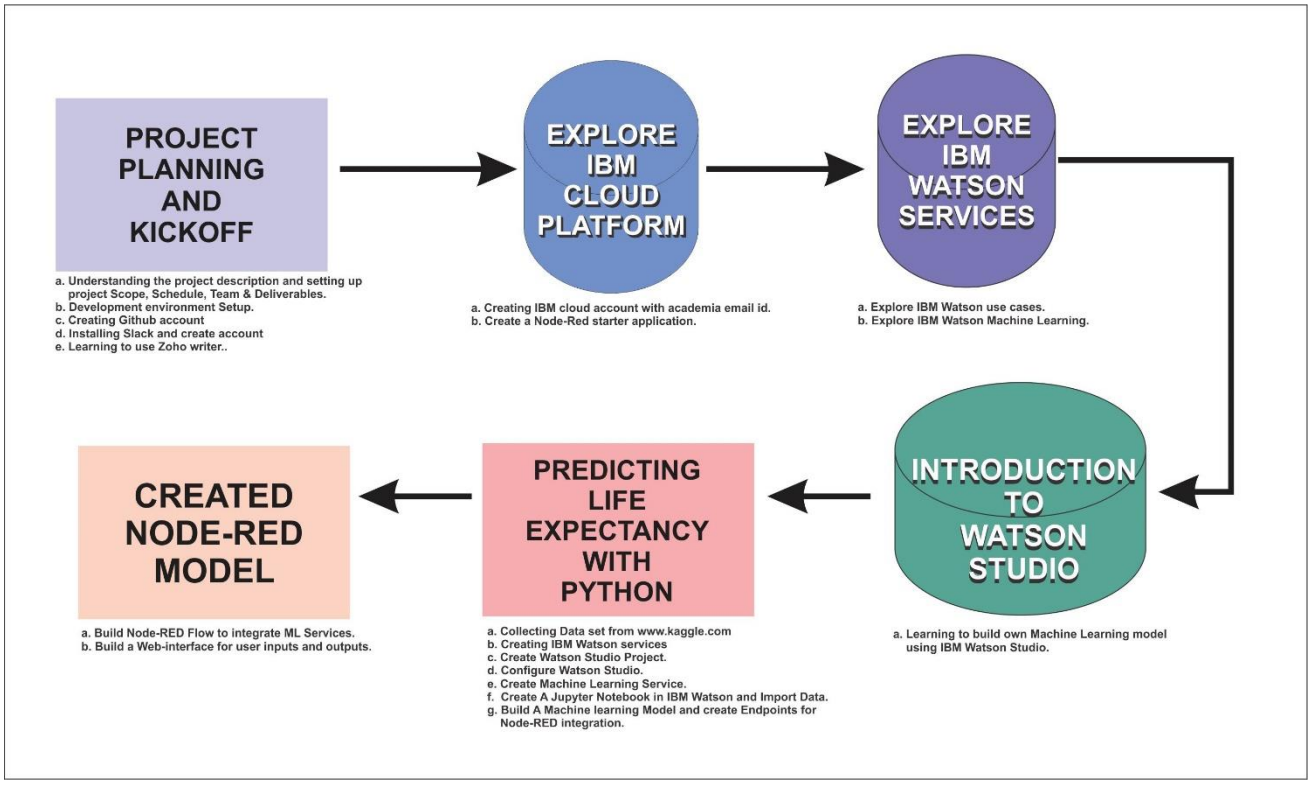
We must need to remember, the algorithm must not attempt to infer the function that exactly matches all the data. Being not careful in fitting the data can cause over-fitting, after which the model will answer perfectly for all training examples but will have a very high error for Unseen samples.

Only after considering all these factors can we pick a Machine learning algorithm that will work perfectly and In order to predict we will be using that Model to draw inferences from the given dataset and give an output. For better usability and input operations we are also going to create a UI with the help of Node-RED for users.

Front-end: A web page taking the necessary inputs from the user to implement the designed model.
Back-end: User given input gets processed according to the trained Model and finally gives the desired output of life expectancy.

3. THEORETICAL ANALYSIS

3.1 Block diagram



3.2 Hardware / Software designing

IBM cloud offers limited resources for its free/personal accounts. Smartinternz provided an academic initiative account to use IBM cloud's academic feature which played top role while completing this project.

The academic initiative account provide us Hardware features like:

- Cloud Object Storage
- 2 vCPU
- 8GB RAM

And in the field of software, it provided us many Resources like:

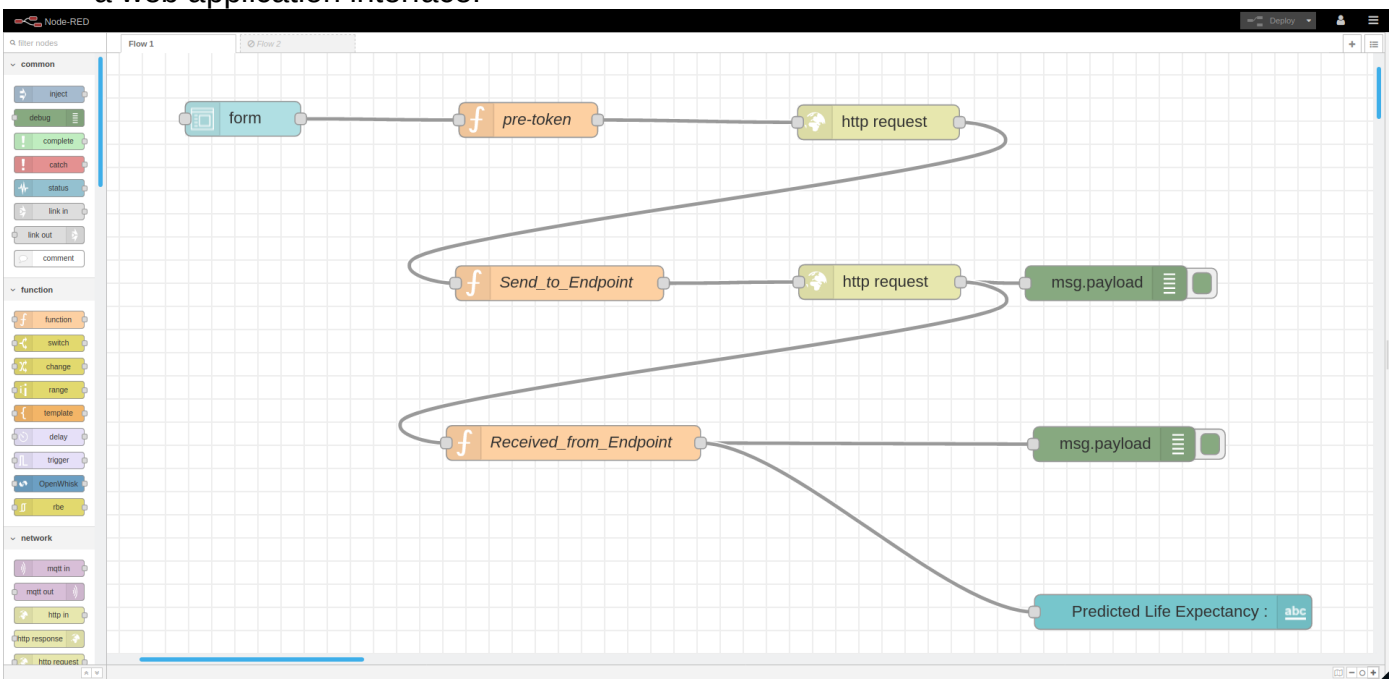
- Python 3.6 programming language.
- Watson Studio.
  - ✓ Notebook asset.

- ✓ Machine Learning Service.
- ✓ Token.

➢ Cloud Foundry App

- ✓ Node-RED Application.

Although there is nothing to design in hardware, we have to design software for the user interface where user inputs value of his/her country and get the life expectancy as a prediction. We will use Cloud Foundry App Node-RED for this purpose and design a web application interface.



In the above shown figure, there are:
1 form to take inputs from the user and deliver life expectancy output,
3 functions to process inputs and outputs,
2 http requests,
2 debug nodes,
And 1 string output.

4. EXPERIMENTAL INVESTIGATION

The data is collected from https://www.kaggle.com/kumarajarshi/life-expectancy-who and saved as a csv file, first we will investigate if there is any null/missing value then fill it with mean value then make every features of same type i.e integer or floats then the Year column is dropped as it will not be used in the analysis. By observing the data we came to know the data contains 21 columns and 2938 rows with the header row and concluded that there are various Factors affecting Life Expectancy of a country such as:

1. Country's Adult Mortality
2. Number of Infant Deaths
3. Alcohol consumption
4. Expenditure on health

5. Hepatitis B immunization
6. Measles reported cases
7. Average Body Mass Index of the entire population
8. Number of dead under-five years
9. Polio immunization coverage
10. Government expenditure on health a
11. Diphtheria immunization coverage
12. HIV/AIDS cases
13. GDP (Gross Domestic Product of the country)
14. Population of the country
15. Thinness among children for Age 5 to 9
16. Thinness among children for Age 10 to 19
17. Income composition of resources
18. Schooling

```
corr_matrix = df.corr()
corr_matrix ["Life expectancy"].sort_values(ascending=False)
```

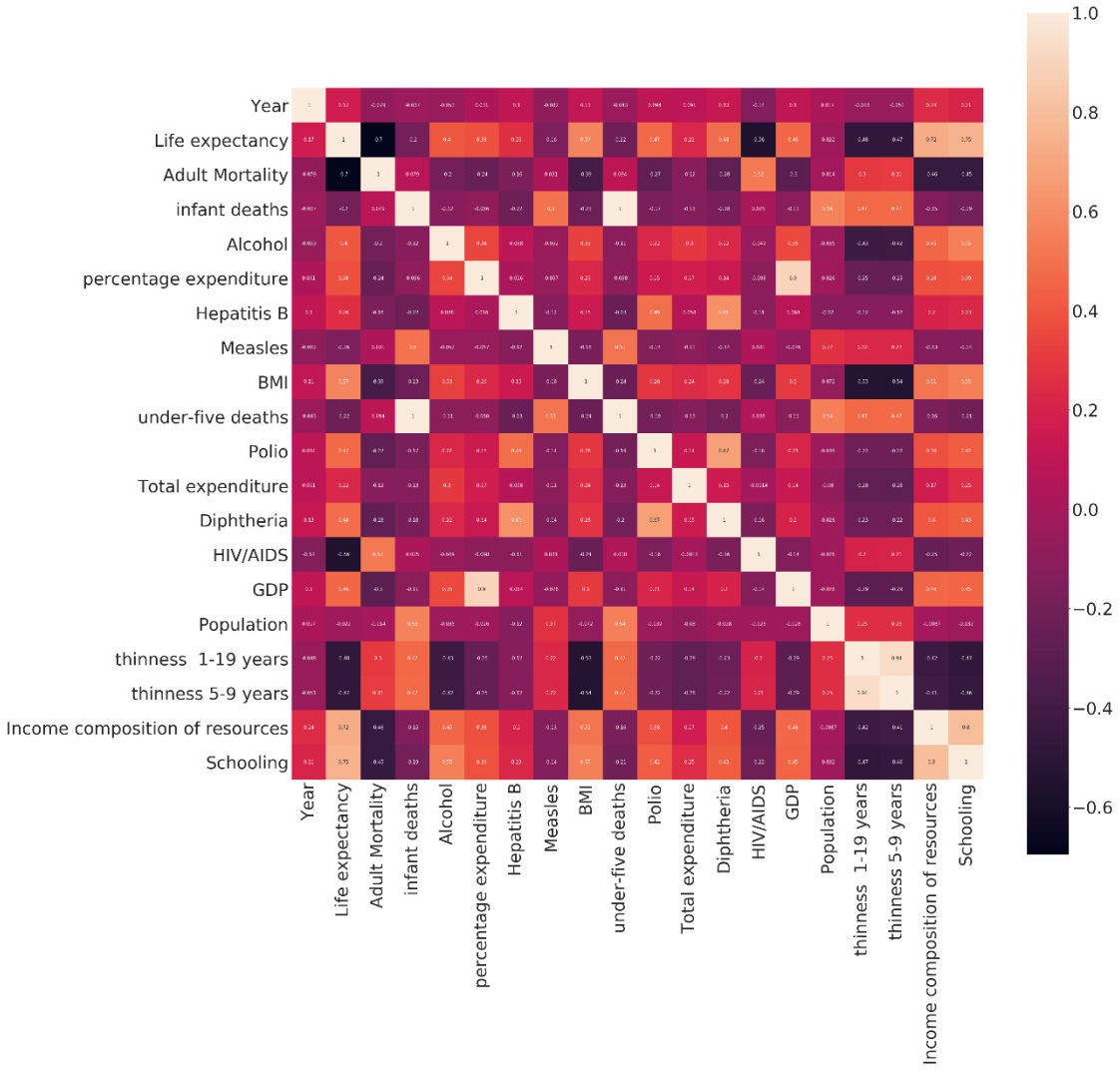
```
Life expectancy                  1.000000
Schooling                        0.715066
Income composition of resources  0.692483
BMI                              0.559255
Status                           0.481962
Diphtheria                       0.475418
Polio                            0.461574
GDP                              0.430493
Alcohol                          0.391598
percentage expenditure           0.381791
Total expenditure                0.207981
Hepatitis B                      0.203771
Year                             0.169623
Population                      -0.019638
Measles                         -0.157574
infant deaths                   -0.196535
under-five deaths               -0.222503
thinness 5-9 years              -0.466629
thinness  1-19 years            -0.472162
HIV/AIDS                        -0.556457
Adult Mortality                 -0.696359
Name: Life expectancy, dtype: float64
```

And By applying correlation function of Panda library we can see above the individual average relation between the features and Life expectancy which also tell us about the factors of the countries on which the life expectancy of its citizen depends.

It is observable that Schooling, Income composition of resources, BMI, country's Status (whether its Developing or Developed), Dipheria, Polio, Alcohol, Percentage expenditure, Total expenditure, Hepatitis B and Year plays and highly positive correlated with Life expectancy of that country and improving in it may leads to extend the life span of the people of that areas as well as we can also see some feature like Population, Measles, infant deaths, thinness 5-9 years, thinness 1-19 years, HIV/AIDS and Adult Mortality are negatively related to the life expectancy of

the people, means in order to increase life span we must need to decrease/improve these features.

Below with the help of correlation heat map of the data-set we can a compact understanding of relation between the individual features and how there are inter-related in the scale of 0.5 to 1.0. Where the shades of light represents high co-relation while the shades of dark represents poor correlations.



Here we got the round understanding for the factoring on which life expectancy depend and how much government must need to improve. As well as the relation between different features.

Now, as we look at data and understand it not a classification problem because the target variable is not categorical (i.e. the output cannot be classified into classes — like it belongs to either Class A or B or something else). Here we have to find the value of life expectancy that is why we will Implement Regression Models for the prediction of life expectancy.

Here I have applied two regression models on the data-set.

1. Linear Regression
2. Random Forest Regression

Linear Regression:

Linear regression is used for finding linear relationship between target and one or more predictors. There are two types of linear regression- Simple and Multiple. As Simple linear regression is useful for finding relationship between only two continuous variables while Multiple Regression is a statistical technique that uses several explanatory variables to predict the outcome of a response variable. The goal of multiple linear regression (MLR) is to model the linear relationship between the explanatory (independent) variables and response (dependent) variable i.e. Life Expectancy. That is why I have used here Multiple Linear Regression.

The Formula for Multiple Linear Regression Is

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} + \epsilon$$

where, for $i = n$ observations:

y_i = dependent variable

x_i = expanatory variables

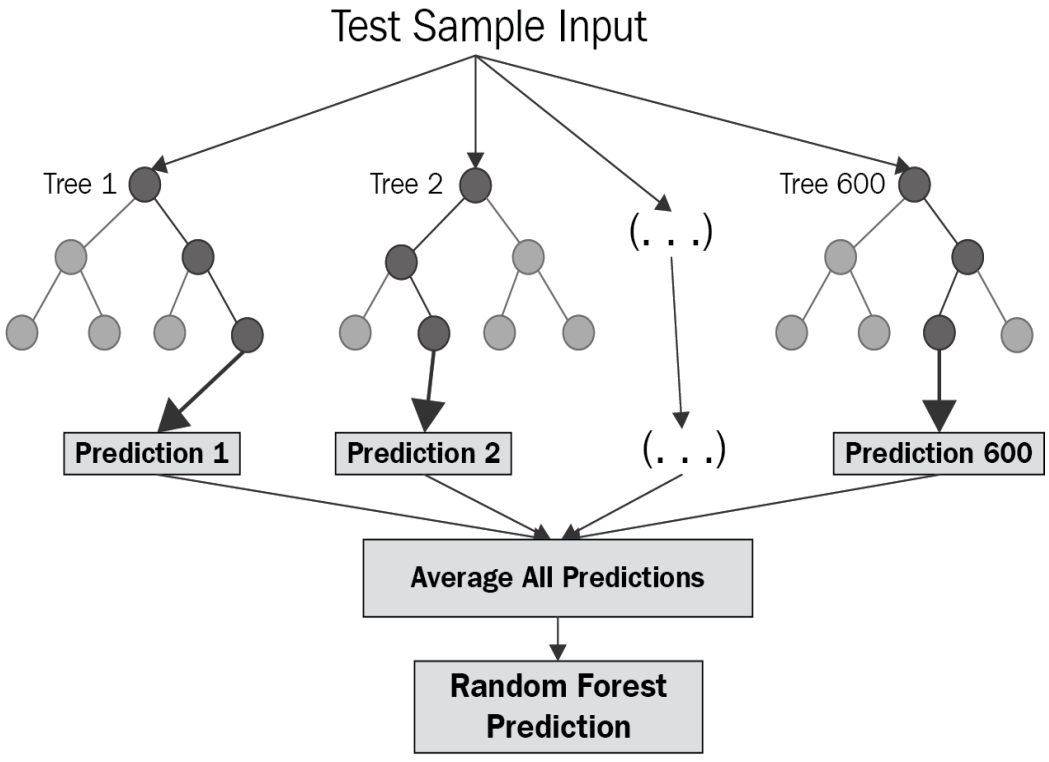
β_0 = y-intercept (constant term)

β_p = slope coefficients for each explanatory variable

ϵ = the model's error term (also known as the residuals)

Random Forest Regression:

Random forest is a Supervised Learning algorithm which uses ensemble learning method for classification and regression**.** Random forest bagging not a boosting random forests. It operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees.



A random forest is a meta-estimator (i.e. it combines the result of multiple predictions) which aggregates many decision trees, with some helpful modifications:

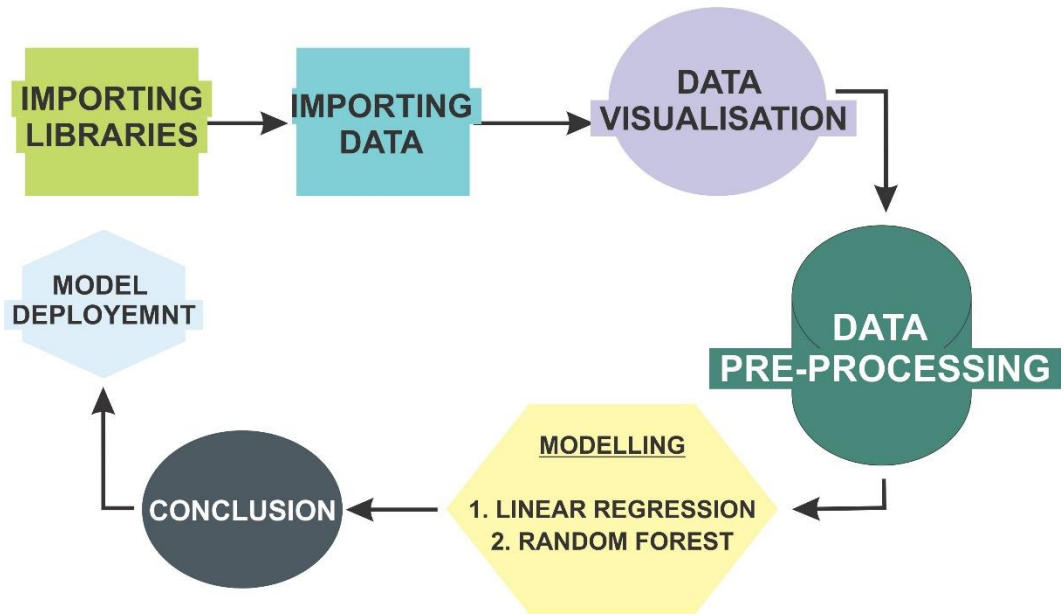
1. The number of features that can be split on at each node is limited to some percentage of the total (which is known as the hyper parameter). This ensures that the ensemble model does not rely too heavily on any individual feature, and makes fair use of all potentially predictive features.
2. Each tree draws a random sample from the original data set when generating its splits, adding a further element of randomness that prevents over fitting.

The above modifications help prevent the trees from being too highly correlated.

Based on our study and information gathered we will
Predict the expected age of a person when he will die. We have to Construct Machine Learning Models and finally selecting the model with maximum accuracy and less error.
To perform operation o data such as cleaning, splitting and normalizing we will use Python Libraries such as:
Pandas: it is used for data manipulation and analysis through operations and data structures on numerical tables and time series.
Numpy: It adds support as well as contains high-level mathematical functions to operate on large multidimensional arrays and matrices.
Matplotlib : It is a plotting library that that enables 2d diagramming and begetting of bar charts, histograms and so forth.
Sci-kit learn: It is a free software machine learning library that features various regression, clustering and classification algorithms. It works in conjunction with numPy and python scientific library sciPy.

5. FLOWCHART

Firstly, the required data set is collected from https://www.kaggle.com/kumarajarshi/life-expectancy-who. Secondly, an IBM Watson studio project is created in the Watson Studio service, provided by IBM Cloud. Then, the data set is imported into this project and latter an AutoAI experiment is performed. The best algorithm of the many algorithms run during the AutoAI experiment is saved as a model. This model can be opened in the Watson Studio project and tested giving the inputs.



6. RESULT

After performing various operation on data and splitting it then training and testing the model we got two different scores and the error functions values,

From Linear Regression Model:

```
from sklearn.metrics import mean_squared_error, mean_absolute_error
print("Mean Square Error (MSE) : " +str(mean_squared_error(Y_test, lpred)))
print("Mean absolute error(MAE) : " +str(mean_absolute_error(Y_test, lpred)))
print("Quadratic mean (RMSE):  " +str(np.sqrt(mean_squared_error(Y_test, lpred))))
```

```
Mean Square Error (MSE) : 12.419195468621215
Mean absolute error(MAE) : 2.6871992362841883
Quadratic mean (RMSE):  3.524087891727619
```

Printing Training accuracy and Testing accuracy

```
print("Training accuracy: " +str(lreg.score(X_train, Y_train)))
print("Testing accuracy:  " +str(lreg.score(X_test, Y_test)))
```

```
Training accuracy: 0.8604056256890611
Testing accuracy:  0.8655194249468139
```

Score values

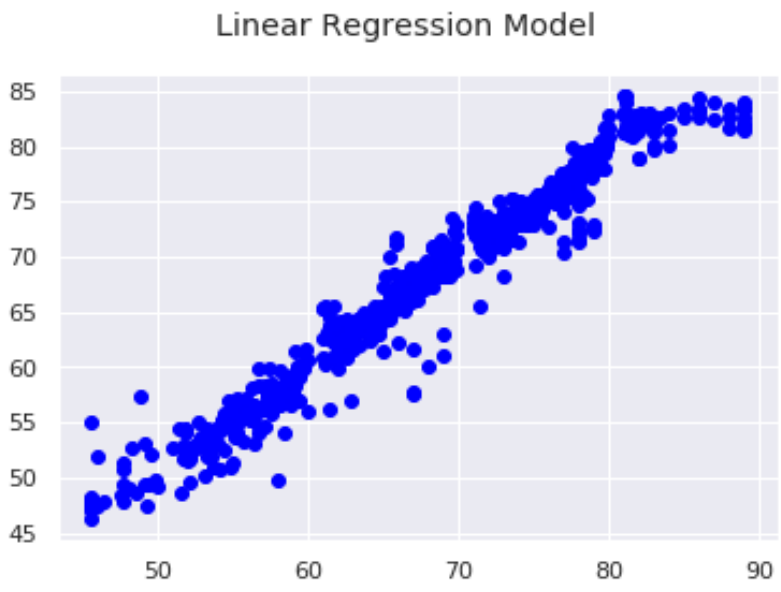
```
lreg.score(X_test,Y_test)
```

```
0.8655194249468139
```

```
from sklearn.metrics import r2_score
lscore = r2_score(Y_test, lpred)
print("R2 score: " +str(lscore))
```

```
R2 score: 0.8655194249468139
```

Prediction Graph



From Random Forest Model:

```
print('Mean absolute error(MAE) : ' +str(mean_absolute_error(Y_test, rf_pred)))
print('Mean Square Error (MSE) :  ' +str(mean_squared_error(Y_test, rf_pred)))
print('Quadratic mean (RMSE): ' +str(np.sqrt(mean_squared_error(Y_test,rf_pred))))
```

```
Mean absolute error(MAE) : 1.2095251031560181
Mean Square Error (MSE) :  3.7825079406040514
Quadratic mean (RMSE): 1.944867075304647
```

Printing Training accuracy and Testing accuracy

```
print("Training accuracy: " +str(rf.score(X_train, Y_train)))
print("Testing accuracy:  " +str(rf.score(X_test, Y_test)))
```

```
Training accuracy: 0.9946064761225994
Testing accuracy:  0.9590413208101194
```

Score values

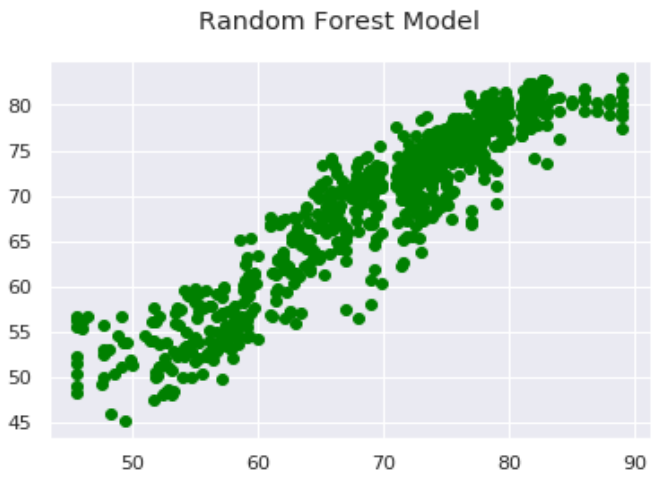
```
rf.score(X_test,Y_test)
```

```
0.9590413208101194
```

```
rfscore = r2_score(Y_test, rf_pred)
print("R-2 score is: " +str(rfscore))
```

```
R-2 score is: 0.9590413208101194
```

Prediction Graph

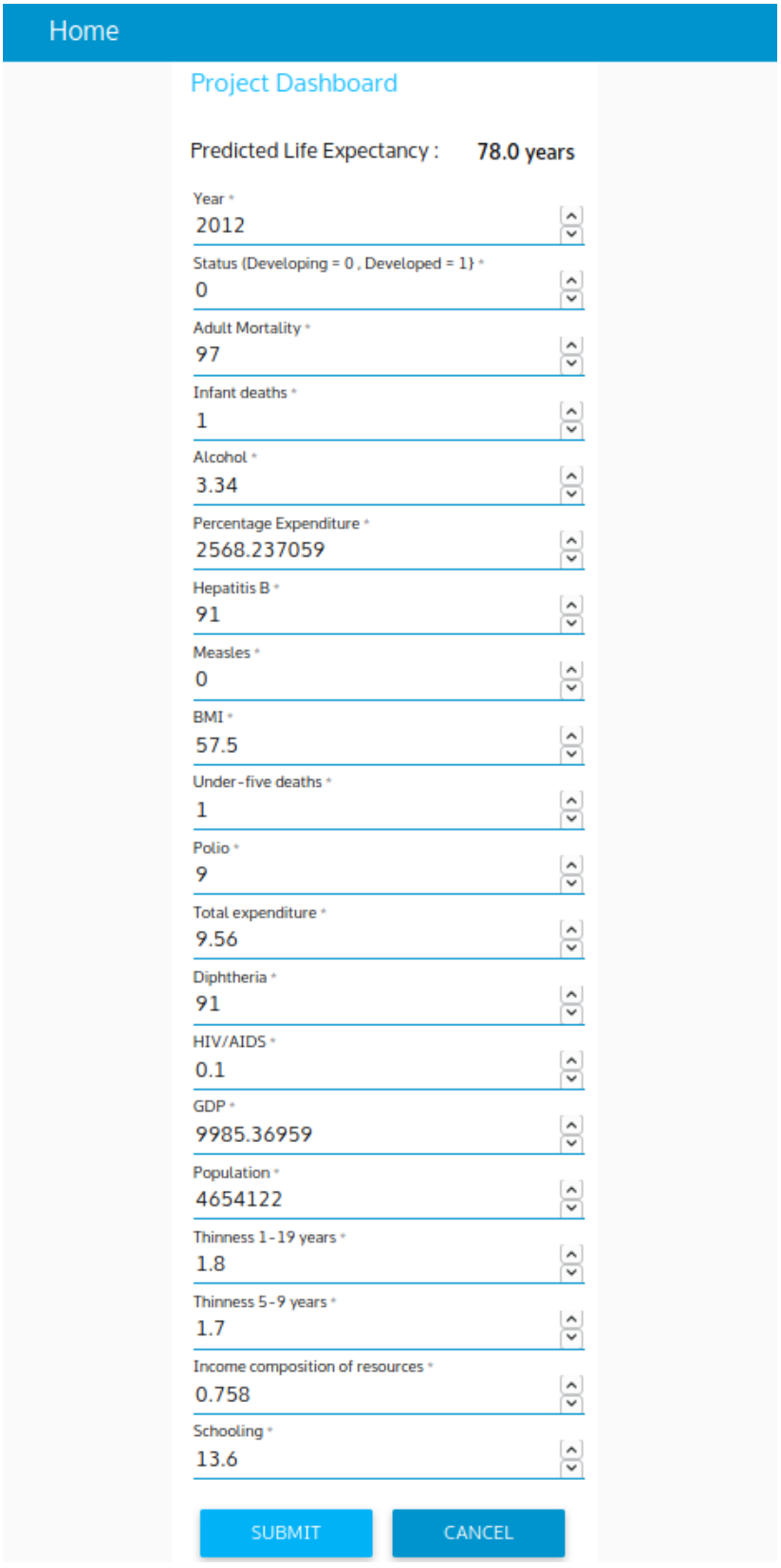


Comparing Both Models:

	Linear Regression Model	Random Forest Model
Prediction score	0.865	0.959
Mean absolute error (MAE)	2.687	1.209
Mean Square Error (MSE)	12.419	3.782
Quadratic mean (RMSE)	3.524	1.944

On comparing Both the models we came to conclusion that we got an near to accurate prediction in Random Forest Model gives less error and better prediction rate which has an accuracy of 95% with compare to linear Regression whose accuracy is 86%. Random Forest model can be used for welfare of human society and increasing the life expectancy.

A User Interface (UI) is designed for the project using Node-RED. In this User Interface the users can type in data regarding their country's GDP, alcohol intake, diseased count, deaths count, income and expenditure, education level, etc. and get the output which is the predicted life expectancy of the country. [Check Live site.]



7.1 Advantages :

1. It is one of the most accurate learning algorithms available. For many data sets, it produces a **highly accurate classifier**.
2. It runs efficiently on large databases.
3. It can **handle thousands of input variables** without variable deletion.
4. It gives estimates of what variables that are important in the classification.
5. It generates an internal **unbiased estimate of the generalization error** as the forest building progresses.
6. It has an **effective method for estimating missing data** and maintains accuracy when a large proportion of the data are missing.

7.2 Disadvantages :

1. Random forests have been observed to **overfit for some datasets** with noisy classification/regression tasks.
2. For data including categorical variables with different number of levels, **random forests are biased in favour of those attributes with more levels**. Therefore, the variable importance scores from random forest are not reliable for this type of data.

8. APPLICATIONS

We can apply this predicting model in various field to predict the life expectancy of humans with minimal human labor. This can also be used by the insurance companies to have a predication about its customers as well as other companies working on products affecting human life or to maintain the work cycle running with the expected life span of its workers.
This can be used by doctors to understand the biological difference and cure measure for the affected patients.

This can also be used by the government to take decisions for human welfare. And to take appropriate measures to control the population growth and other factors which negatively effects life span of the people of the country. It also direct the utilization of the increase in human resources and skillset acquired by people over many years.

This could help make common people more aware of their general health, and its improvement or deterioration over time. This may motivate them to make healthier lifestyle choices.

9. Conclusion

While completing the project we came to the conclusion that the life expectancy of people of different country's is being affect in a negative by most of the same cause like Alcohol intake, HIV/AIDS, Population, Adult Morality and much more. And it is degrading every year. People as well as government must need to b aware about it and look for the solutions.

And Machine learning is a promising field and with new researches publishing every day. By this study, open the scope of ecologist, medical scientists which allows common people to not to depend on expert and know their biological status and condition of human being in desired region.

10. Future Scope

This application has very bright future and large number of applicable cases, it must need to be upgraded and taken into account in school as well as college syllabus for the welfare and knowledge purpose. This provides insights in various factors and their levels required to keep the life expectancy rate as high as expected. It can be used to suggest good health practices and life style to the users based on their daily activities and provide suggestions for exercises for improving their health. Pharmaceutical companies can check which diseases impact more people and therefore impact life expectancy and based on this manufacture medicine.
We can also say it is a time machine which predict the life of someone who haven't born yet on the factors of his/her country's Adult Mortality, Population, Under 5 Deaths, Thinness 1-5 Years, Alcohol, HIV, Hepatitis B, GDP, Percentage Expenditure, and others.
As the technology is growing faster than ever, as the world is leaning towards more man power need it necessary to improve the factor by which it can extend the life expectancy of its people.

11. BIBLIOGRAPHY

Book: Free and Open Machine Learning Release 1.0.1
by Maikel Mardjan

Online Resources:

Getting start with IBM Cloud
https://www.ibm.com/cloud/get-started
Watson Studio workshop.
https://bookdown.org/caoying4work/watsonstudio-workshop/jn.html
Watson Studio notebook introduction.
https://www.youtube.com/watch?v=Jtej3Y6uUng
About IBM Cloud Services
https://www.youtube.com/watch?v=NmdjtezQMSM
Node-RED Application tutorial
https://developer.ibm.com/tutorials/how-to-create-a-node-red-starter-application/
https://github.com/watson-developer-cloud/node-red-labs
About API
https://www.youtube.com/watch?v=s7wmiS2mSXY&feature=youtu.be
Introduction to Machine Learning
https://developer.ibm.com/technologies/machine-learning/series/learning-path-machine-learning-for-developers/
Data Collection
https://www.kaggle.com/kumarajarshi/life-expectancy-who

End Point creation Reference
https://bookdown.org/caoying4work/watsonstudio-workshop/jn.html#deploy-model-as-web-service

APPENDIX

Source Code:

https://github.com/SmartPracticeschool/llSPS-INT-1519-Predicting-Life-Expectancy-using-Machine-Learning/blob/master/Predicting%20Life%20Expectancy.ipynb