PROJECT TITLE

Predicting Life Expectancy using Machine Learning

FINAL PROJECT REPORT

NAME : ESHIKA AGARWAL
COMPLETED UNDER SMARTBRIDGE
DATE : 12/06/2020

1.INTRODUCTION

1.1 OVERVIEW

Life expectancy is a statistical measure of the average time a human being is expected to live. Life expectancy depends on various factors including: financial status, regional variations, economic circumstances, gender differences, mental illnesses, physical illnesses, education, year of their birth, demographic factors and many more.
Prediction of life expectancy is difficult for humans. Research shows that machine learning and natural language processing techniques offer a feasible and promising approach to predicting life expectancy.

1.2 PURPOSE

A typical Regression Machine Learning project leverages historical data to predict insights into the future.
In this project we collect historical data regarding the GDP, year, development status, education, alcohol intake of people in the country, expenditure on healthcare system and some specific disease related deaths that happened in the country and utilize this data to develop a Machine Learning model which is then used to predict the life expectancy.
This project is aimed at predicting Life Expectancy rate of a country, given various factors, using Machine Learning algorithms.

2.LITERATURE SURVEY

2.1 EXISTING PROBLEM

Life Expectancy rate is the average time a human being lives. This is based on various health, education and economic factors. If life expectancy rate is low, the time a person lives is less, which is not desired.

2.2 PROPOSED SOLUTION

This project develops a Regression Machine Learning model which makes use of historical data to train and develop the model. This model is then used to predict the average life expectancy of people living in a country provided, given conditions. Through this we can get to know the various factors affecting life expectancy and the ranges or values the factors must have to get a desired life expectancy.

Following are the features of the data:

- *'Country',*
- *'Year',*
- *'Status'*
- *'Life expectancy '*
- *'Adult Mortality'*
- *'infant deaths'*
- *'Alcohol'*
- *'percentage expenditure'*
- *'Hepatitis B'*
- *'Measles '*
- *' BMI '*
- *'under-five deaths '*
- *'Polio'*
- *'Total expenditure'*

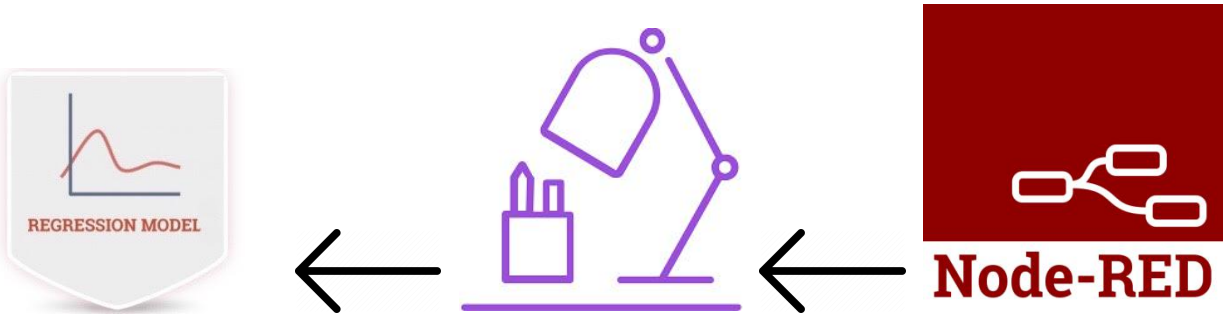
- *'Diphtheria '*
- *' HIV/AIDS'*
- *'GDP'*
- *'Population'*
- *' thinness 1-19 years'*
- *' thinness 5-9 years'*
- *'Income composition of resources'*
- *'Schooling'*

Target is Life Expectancy, measured in number of years. The assumptions are:

1. *These are country level average*
2. *There is no distinction between male and female*

3.THEORETICAL ANALYSIS

3.1 BLOCK DIAGRAM



Regression Model IBM Watson Studio Node-RED

User Interface

3.2 HARDWARE/SOFTWARE DESIGNING

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 a new Jupyter notebook is created to train

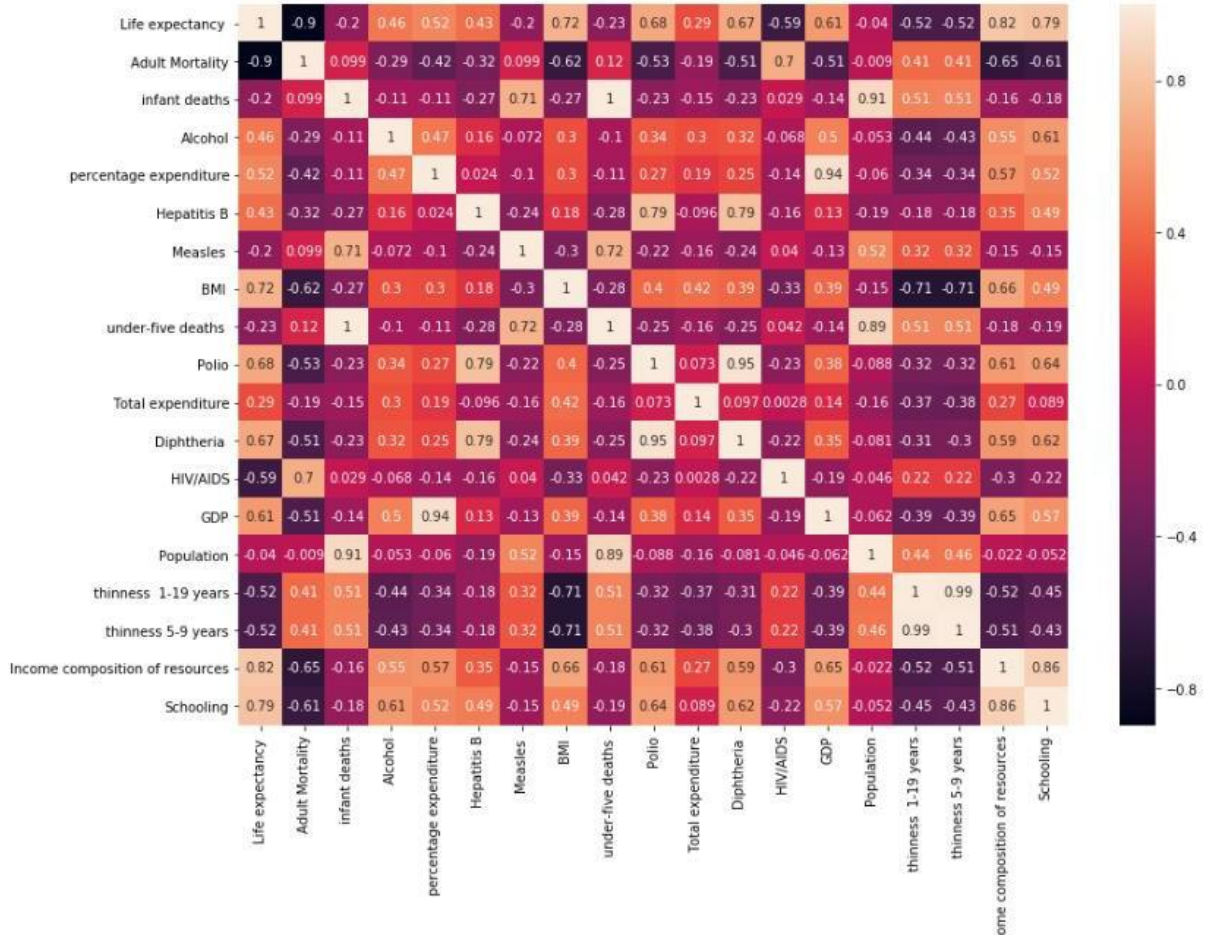
the model. The best algorithm of the many algorithms is saved as a model. This model can be opened in the Watson Studio project and tested giving the inputs.

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.

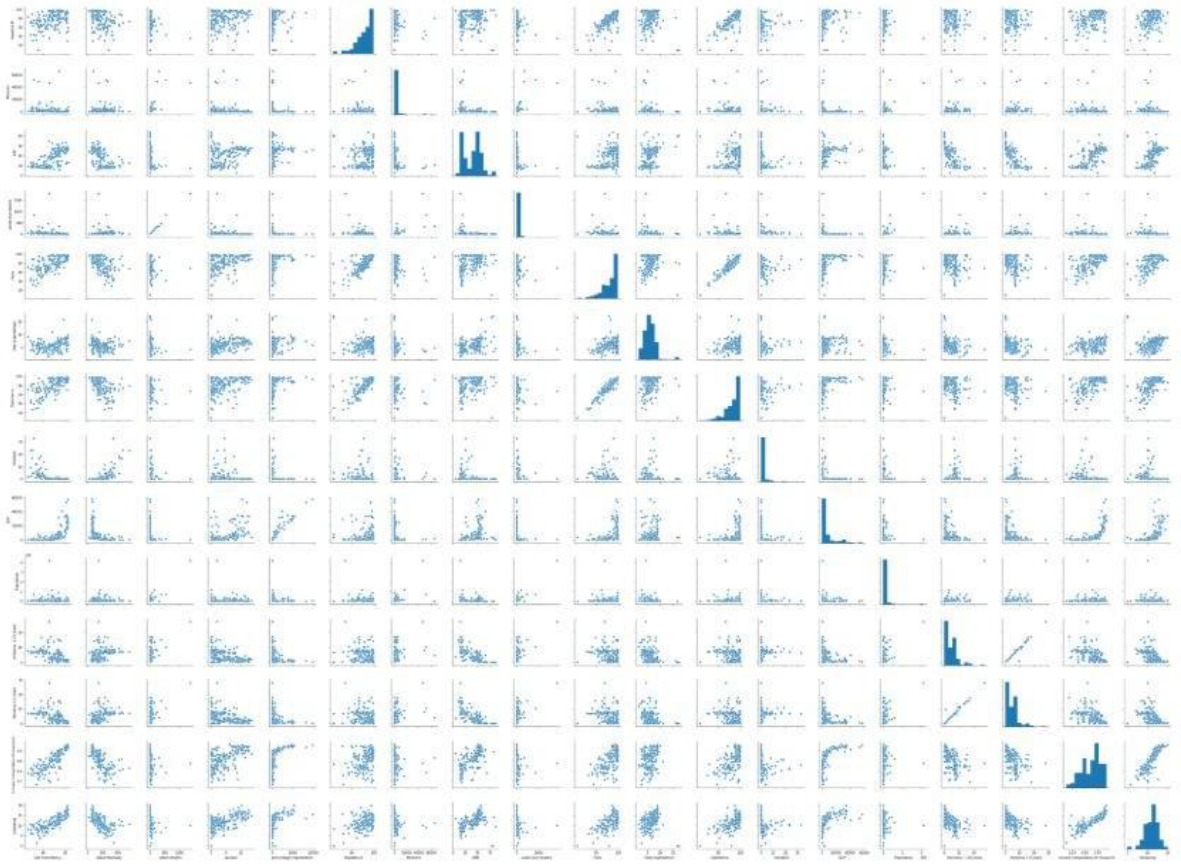
DATA ANALYSIS

Following are some graphs generated using the refine data set .

HEATMAP :

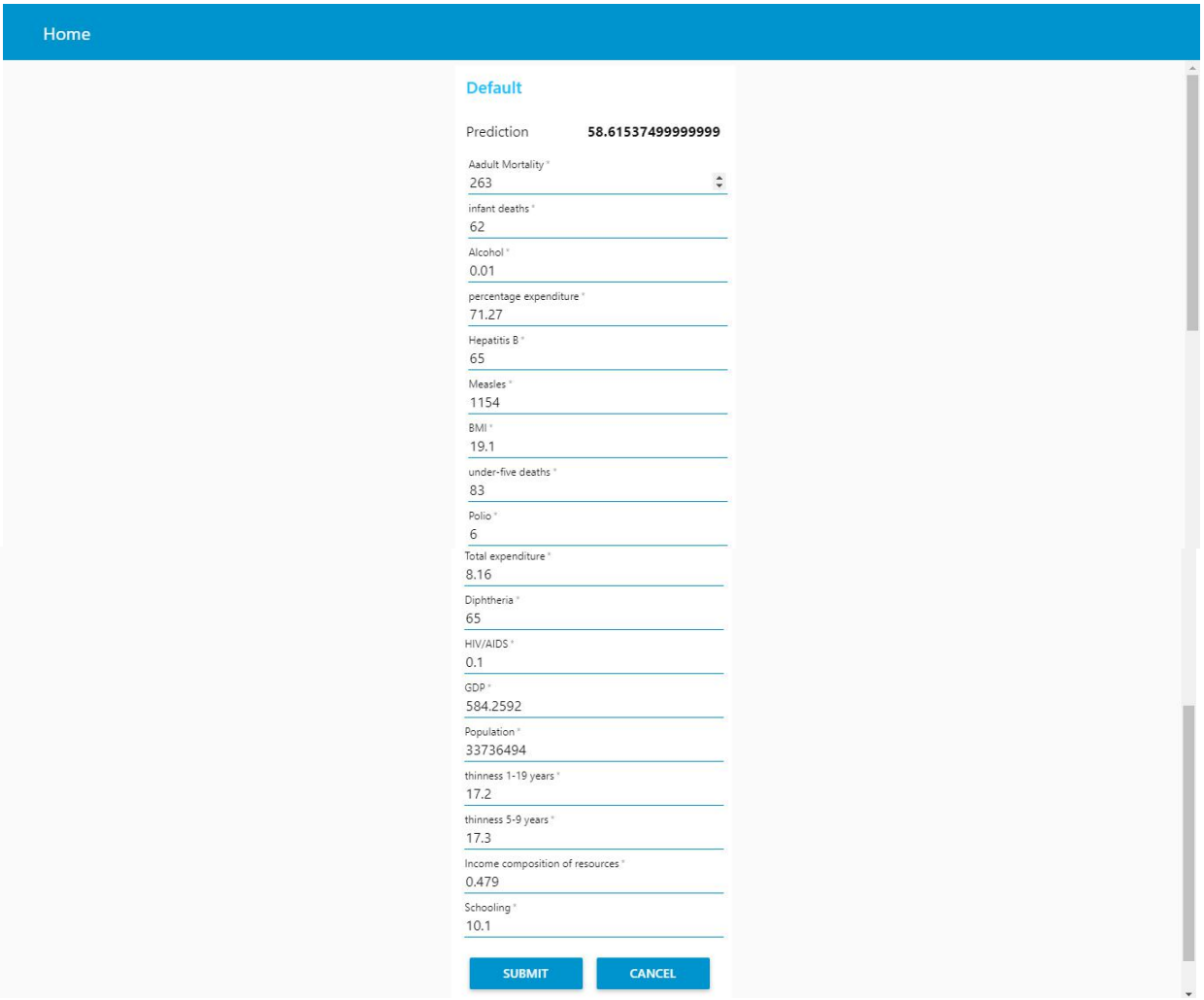


SNS PAIRPLOT :

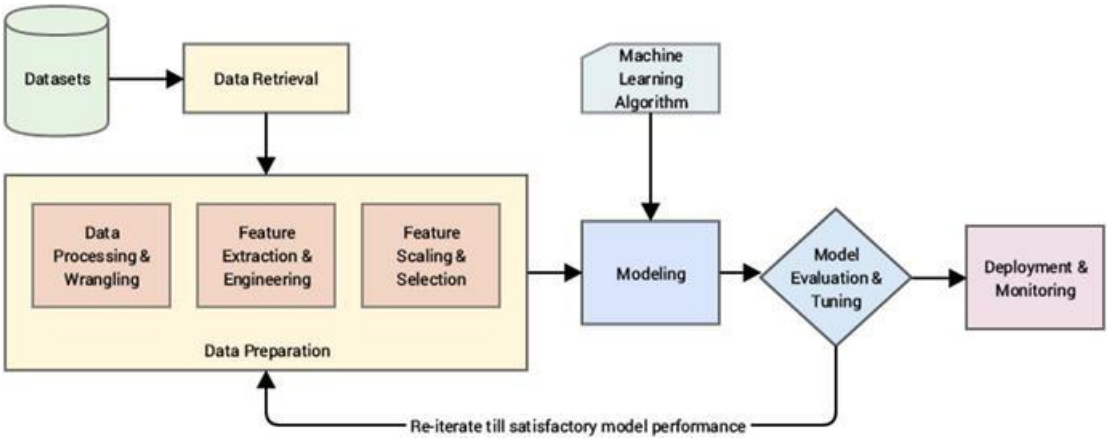


4.EXPERIMENTAL INVESTIGATIONS

Some random inputs are given to the deployed machine learning model. We got the following output.



5.FLOWCHART



6.RESULT

Based on the given data, the machine learning model understands the data and cross reference the data to watch what are the factors that are affecting the results we require ilife expectancy. Then when we give any input, it has already run algorithm to get the output based on previously given data. So the results we get are approximations, they are not definitely true, but it works in maximum number of cases.
Observing the performance, the Random Forest Regression model provided the best output among the models tested with the lowest error rates. The accuracy was 0.98 on the training data and 0.91 on the testing data.

7.ADVANTAGES AND DISADVANTAGES

- *Since we can predict the life span, we can know what factors are influencing the expectancy on life span in what ways.*
- *So, therefore by trying to change those factors in the real world we can increase the life span.*

8.APPLICATIONS

This project tells the average age a person of a country is expected to live. This provides insights in various factors and their levels required to keep the life expectancy rate as high as expected.

To conclude, here are some interesting insights:

1.Japan has the highest life expectancy (83.7 years). Central African Republic (49.5 years) and many countries in the African continent are at the bottom of scale. Singapore is ranked #5 (82.7 years).

2. Take good care of the environment. It has the largest coefficient (impact) on the country's life expectancy.

9.CONCLUSION

The outcome of the project is the prediction of the Life expectancy of an individual based on certain factors such as Schooling, diseases, GDP of the country, population, Alcohol consumption, percentage expenditure etc.
It was found that factors such as the GDP, illnesses such as Diptheria, Alcohol consumption and Schooling, significantly impacted the overall regression output. Other factors that less impacted the outcome were Hepatitis B, Adult Mortality and Status i.e. whether a country was developing or developed.
The model successfully calculated the Life expectancy with a mean absolute error as small as 2.14 .

10. FUTURE SCOPE

The problem of processing datasets such as electronic medical records(EMR) and their integration with genomics, environmental factors, socioeconomic factor and patient behavior variations have posed a problem for researchers in the health industry. Due to rapid innovations in machine learning field such as big data, analytics,visualization, deep learning, health workers now have improved way of processing, and developing meaningful information from huge datasets that have been accumulated over many years .

Big data and machine learning can benefit public health researchers with analyzing thousands of variables to obtain data regarding life expectancy. We can use demographics of selected regional areas and multiple behavioral health disorders across regions to find correlation between individual behavior indicators and behavioral health outcomes.

I could possibly collect more data by expanding the scope to cities instead of countries, and to explore other features (factors) affecting life expectancy. Also, I could split the data to male and female categories for such life expectancy regression analysis.

11.BIBILOGRAPHY

https://smartinternz.com

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

https://cloud.ibm.com

APPENDIX

WEBPAGE

https://node-red-jhjip.eu-gb.mybluemix.net/ui/#!/0?socketid=mW1AeVB5OSfp4eIIAAAG

WHO data for life expectancy of different country.

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