Internship Report

Predicting Life Expectancy Using Machine Learning

Submitted By:-

Shrutika Kharat

shrutikharat267@gmail.com

Web Page Link

https://shrutikas-first-app.mybluemix.net/ui/#!/0?socketid=Jv_8PLAt1MUsIA34AAAH

Github Link

https://github.com/SmartPracticeschool/llSPS-INT-2108-Predicting-Life-Expectancy-using-Machine-Learning

Video Demonstration Link

https://drive.google.com/file/d/1A5YTocnQ2qZ5wxfveL_HaaGDuTaWc_Nm/view?usp=sharing

Internship Title:- **Predicting Life Expectancy using Machine Learning - SB15612**

Category:- **Machine Learning.**

Table Of Contents

1. INTRODUCTION

1.1. Overview 3

1.2. Purpose 4

2. LITERATURE SURVEY

2.1 Existing Problem 5

2.2 Proposed solution 7

3. THEORETICAL ANALYSIS

3.1 Block Diagram 8

3.2 Hardware/Software design 8

4. EXPERIMENTAL INVESTIGATION 9

SCREENSHOTS 10

5. FLOWCHART 13

6. RESULT 14

7. ADVANTAGES & DISADVANTAGES 15

8. APPLICATION 16

9. CONCLUSION 16

10. FUTURE SCOPE 16

11. BIBLIOGRAPHY 17

APPENDIX

A. SOURCE CODE 18

1. INTRODUCTION

1.1 Overview

This Project consists of a machine learning model which predicts life expectancy based on various factors.This project is developed using IBM Machine Learning and deployed on IBM Cloud and is integrated with Node Red Application which provides Web Interface for the project. Life expectancy is a statistical measure of the average time a human being is expected to live. Life expectancy depends on various factors such as Regional variations, Economic Circumstances, Sex Differences, Mental Illnesses, Physical Illnesses, Education, Year of their birth and other demographic factors.

This project tries to create a model based on data provided by the World Health Organization (WHO) to evaluate the life expectancy for different countries in years. The data offers a timeframe from 2000 to 2015. The data originates from here https://www.kaggle.com/kumarajarshi/life-expectancy-who/data.

This project 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.

•Project Requirements: IBM Cloud, IBM Watson Studio,Jupyter Notebook, Node-RED.

•Functional Requirements: IBM cloud, Machine Learning Libraries, Algorithms(Regression).

•Technical Requirements: WATSON Machine Learning, Python.

•Software Requirements: Python, Watson studio,Node-Red

• Project Team: Shrutika Kharat.

1.2 Purpose and Scope

Life expectancy is the most significant aspect for decision making. Good projection for example helps to decide the course of treatment and helps to anticipate the procurement of health care services and facilities, or more broadly: facilitates Advance Care Planning. Advance Care Planning improves the quality of the final phase of life by stimulating doctors to explore the preferences for end-of-life care with their patients, and people close to the patients.

This project can be used in **hospitals** and the **doctors** can use it to predict the life expectancy of a patient with the underlying disease or a new born baby. It can be used by the government to predict the life expectancy of the economically backward people due to poverty.With the help of this project it will be easy for **governments of the countries** with less life expectancies to improve their medical and healthcare services.

This Project will give overall prediction about the life expectancy of people living in various countries who have various diseases like Diphtheria, HIV, Hepatitis, Polio, Measles and also people taking alcohol based on Body Mass Index(BMI), GDP, Population, Mortality Rate of a particular country.

2. Literature Survey

2.1 Existing Solution

We have reviewed existing works and techniques used in the prediction of the human life expectancy, and reached a conclusion that it is feasible for individuals using evolving technologies and devices wearables and mobile health monitoring devices. We have also identified that the factors used for predicting were just personal causes and not related to the surrounding, healthcare facilities, demographic, social, regional and economical factors of the country he resides. These country dependent factors can also be an important feature to predict the life expectancy of an individual. So we need more data to predict more accurately.

2.2 Proposed Solution

• For the above problem to get solved we have a dataset consisting of various factors .In this system we have taken all the correlated features into consideration. So the target output variable i.e expected life span of the people depends upon a variety of factors and not factors of particular fields.

• Important immunization like Hepatitis B, Polio and Diphtheria are also considered.

• The data-set related to life expectancy, health factors for 193 countries has been collected from WHO data repository website and its corresponding economic data was collected from the United Nations website. Among all categories of health-related factors only those critical factors were chosen which are more representative. It has been observed that in the past 15 years, there has been a huge development in the health sector resulting in

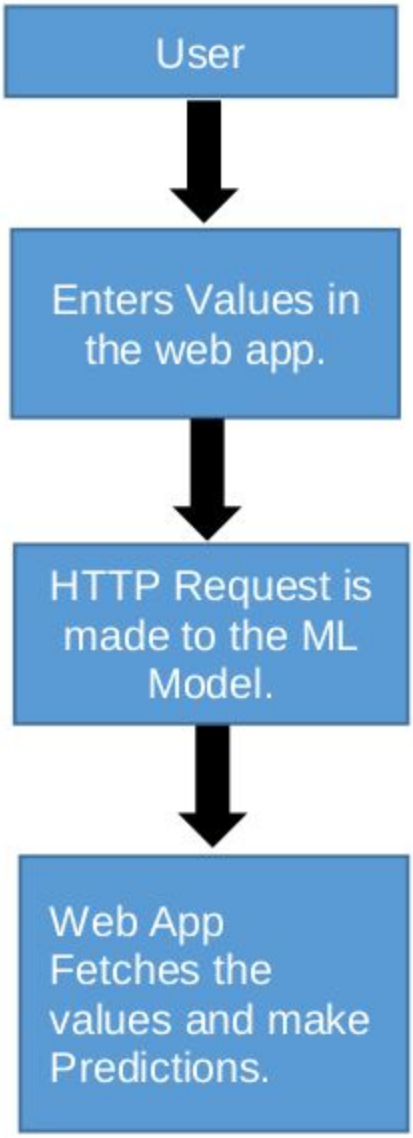
improvement of human mortality rates especially in the developing nations in comparison to the past 30 years. Therefore, in this project we have considered data from the year 2000-2015 for 193 countries for further analysis. The individual data files have been merged together into a single data-set.

• The project uses immunization factors, mortality factors, economic factors, social factors and other health related factors to predict life expectancy of a country for a given year using a machine learning model.

• Since the observations in 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.

3. Theoretical Analysis

3.1 Block Diagram



3.2 Hardware / Software Designing

- Create necessary IBM Cloud services
- Create Watson studio project
- Configure Watson Studio
- Create IBM Machine Learning instance
- Create machine learning model in Jupyter notebook
- Deploy the machine learning model
- Create flow and configure node
- Integrate node red with machine learning model
- Deploy and run Node Red app.
- Input is taken from the user using a “Form” element in Node-Red. Then, an HTTP request is made to the IBM cloud that further makes an HTTP request to the deployed model using model’s instance id.
- After verification of id, the model sends an HTTP response which is finally parsed by the Node-Red application and the result is displayed on the user screen.

4. Experimental Investigations

Following factors are taken into account for predicting the life expectancy of a country.

1. Country
2. Status: Developed or Developing status of the country.
3. Year
4. Adult mortality: Adult Mortality Rates of both sexes (probability of dying between 15 and 60 years per 1000 population).
5. Infant deaths: Number of Infant Deaths per 1000 population.
6. Alcohol: Alcohol, recorded per capita (15+) consumption.
7. Percentage Expenditure: Expenditure on health as a percentage of Gross Domestic Product per capita (%).
8. Hepatitis B: Hepatitis B =immunization coverage among 1-year-olds (%).
9. Measles: Measles - number of reported cases per 1000 population.
10. BMI: Average Body Mass Index of entire population.
11. Under-five deaths: Number of under-five deaths per 1000 population.
12. Polio: Polio (Pol3) immunization coverage among 1-year-olds (%).
13. Total expenditure: General government expenditure on health as a percentage of total government expenditure (%).
14. Diphtheria: Diphtheria tetanus toxoid and pertussis (DTP3) immunization coverage among 1-year- olds (%).
15. HIV/AIDS: Deaths per 1 000 live births HIV/AIDS (0-4 years).
16. GDP: Gross Domestic Product per capita (in USD).
17. Population: Population of the country.
18. Thinness 10-19 years: Prevalence of thinness among children and adolescents for Age 10 to 19(%).
19. Thinness 5-9 years: Prevalence of thinness among children for Age 5 to 9(%).
20. Income composition of resources: Human Development Index in terms of income composition of resources (index ranging from 0 to 1).
21. Schooling: Number of years of schooling.

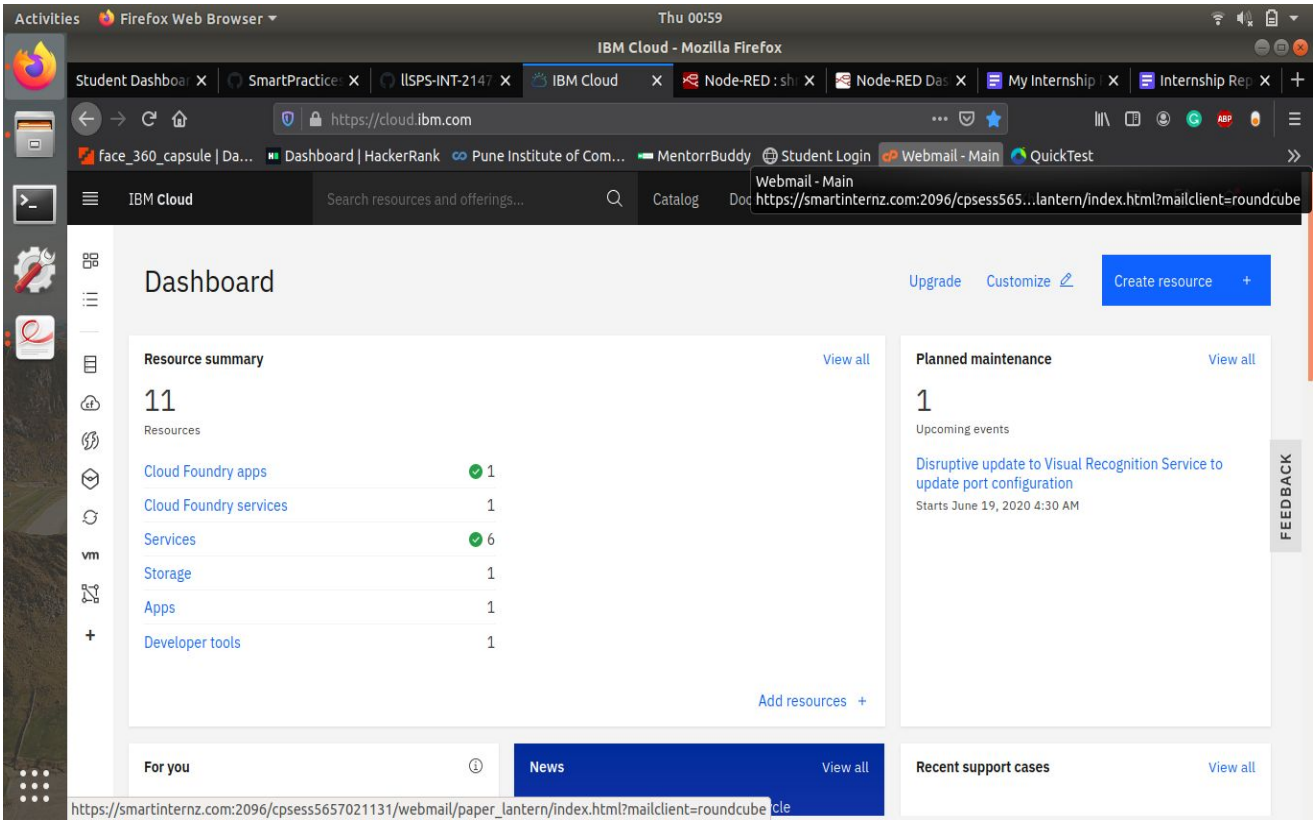
Finding The Best Algorithm:-

Random Forest Regression Provides Best Accuracy Of – 95.25%

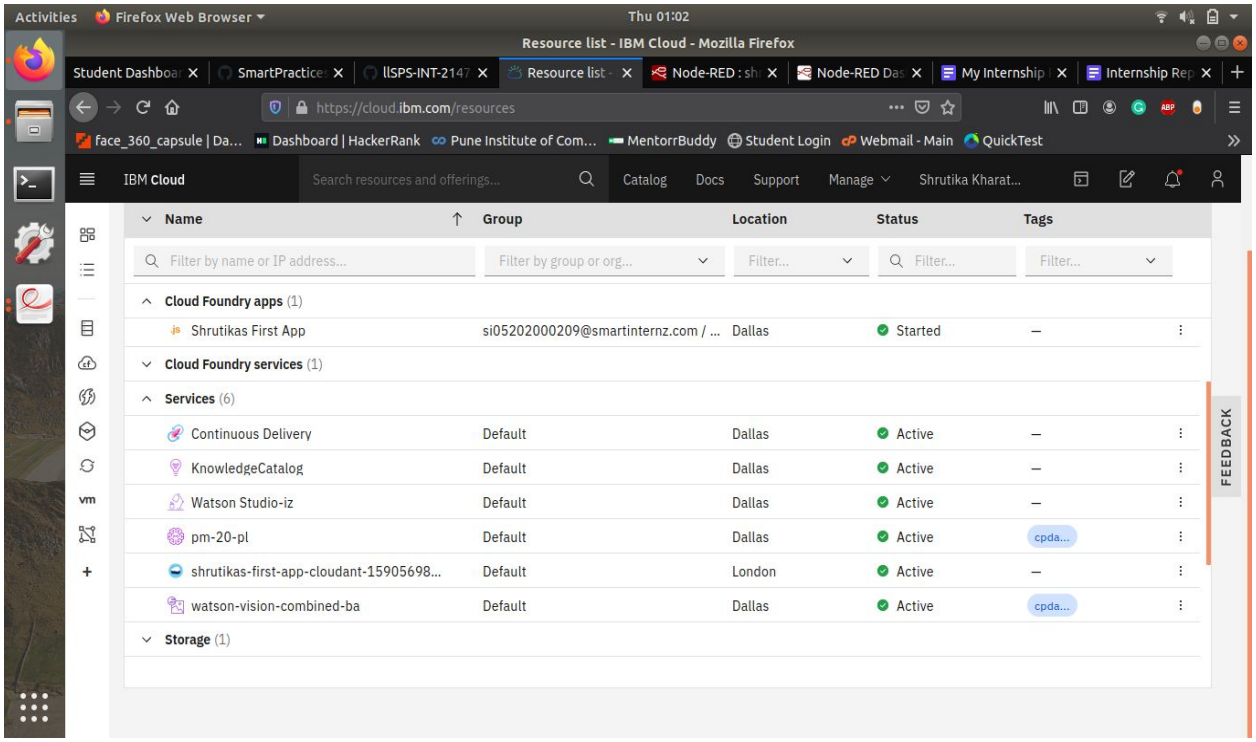
OrderedDict([('Decision Tree Regressor', 0.9189785430104337),
('Linear Regression', 0.8188082000368992), ('Random Forest
Regressor', 0.9525555281025279)])

SCREENSHOTS:

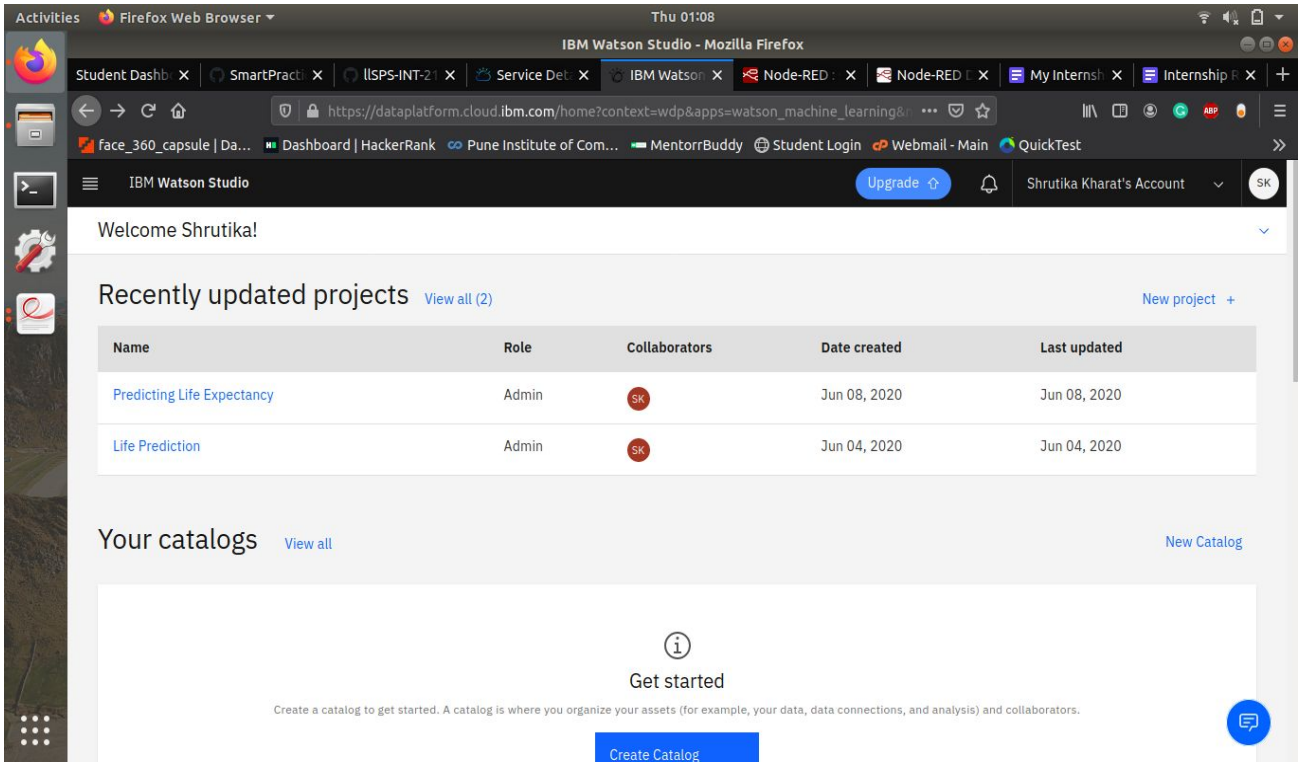
IBM CLOUD DASHBOARD:



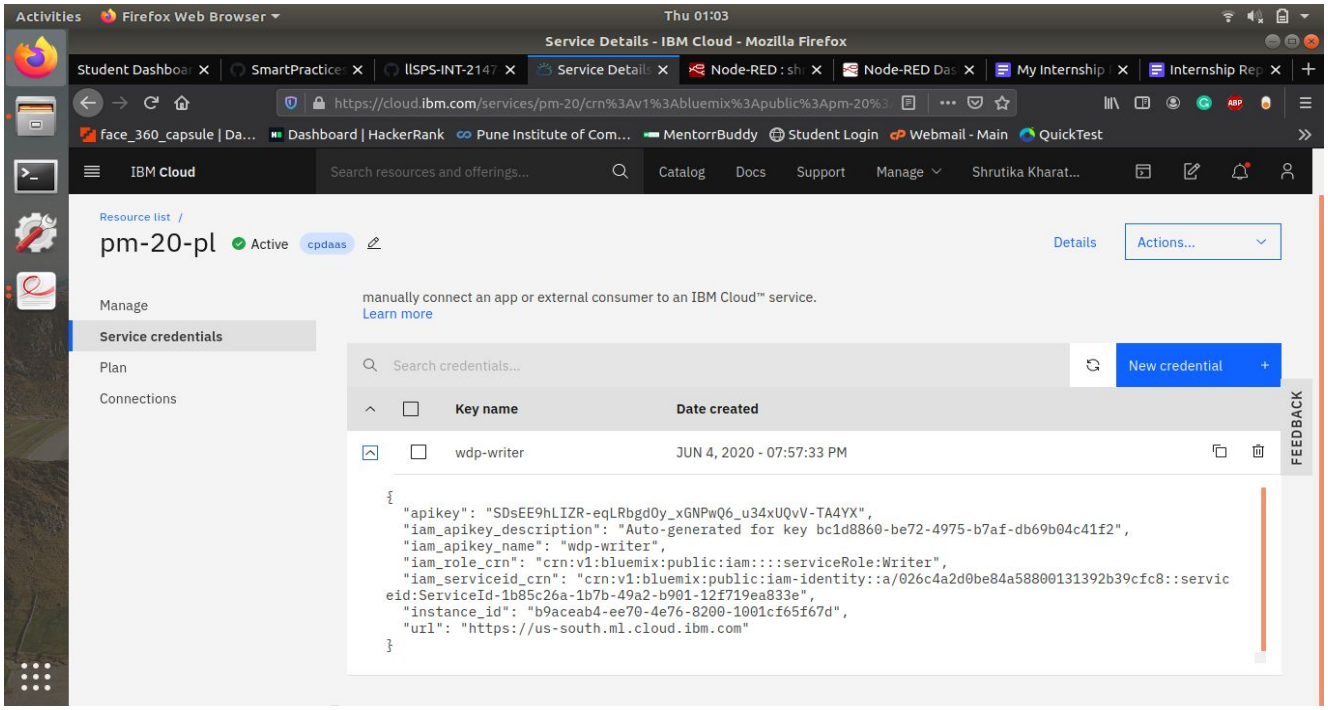
Resources List:



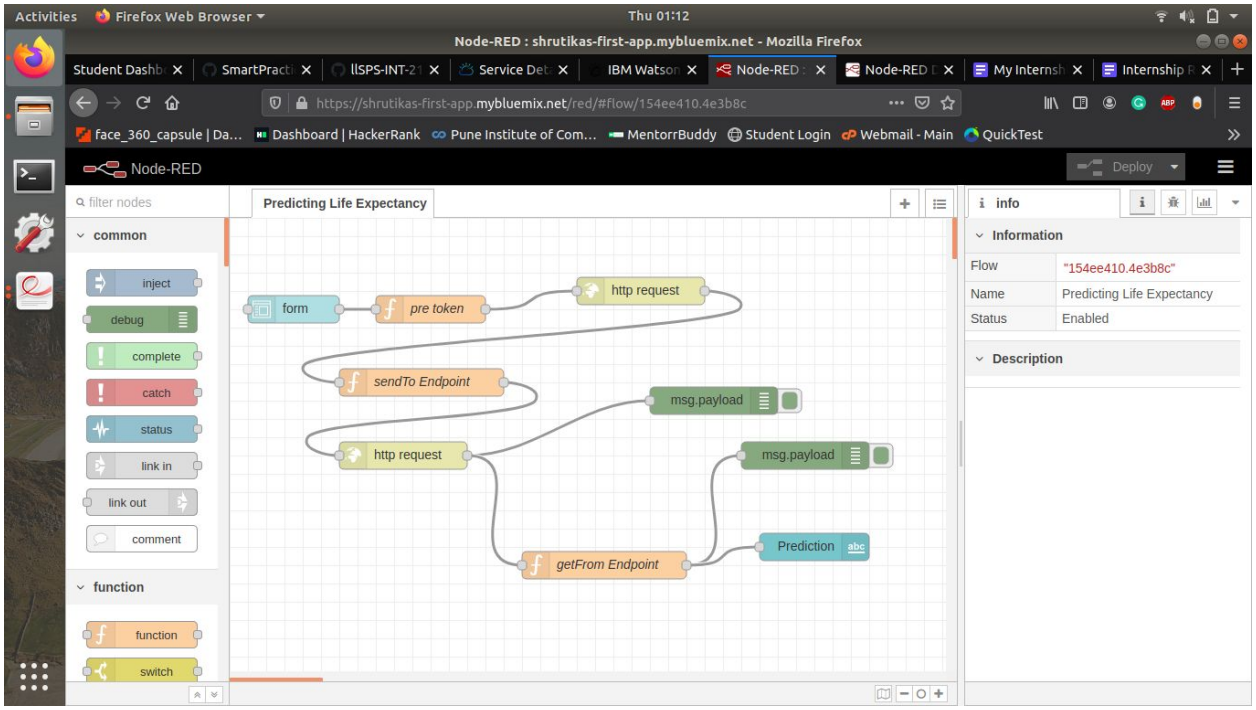
WATSON STUDIO :



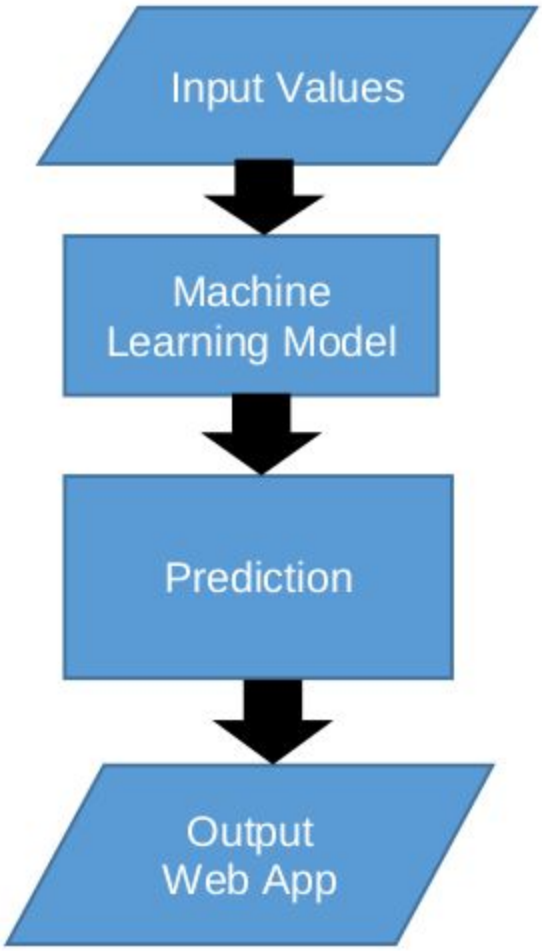
WATSON MACHINE LEARNING SERVICE:



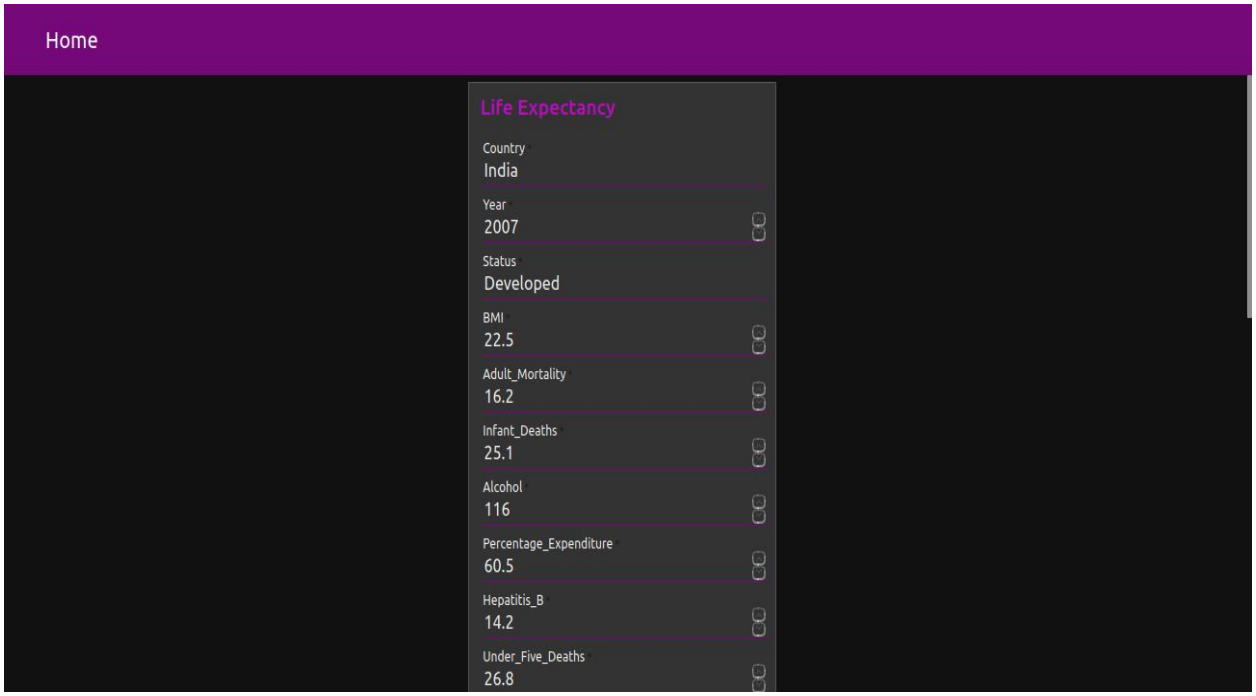
Node-Red Flow:

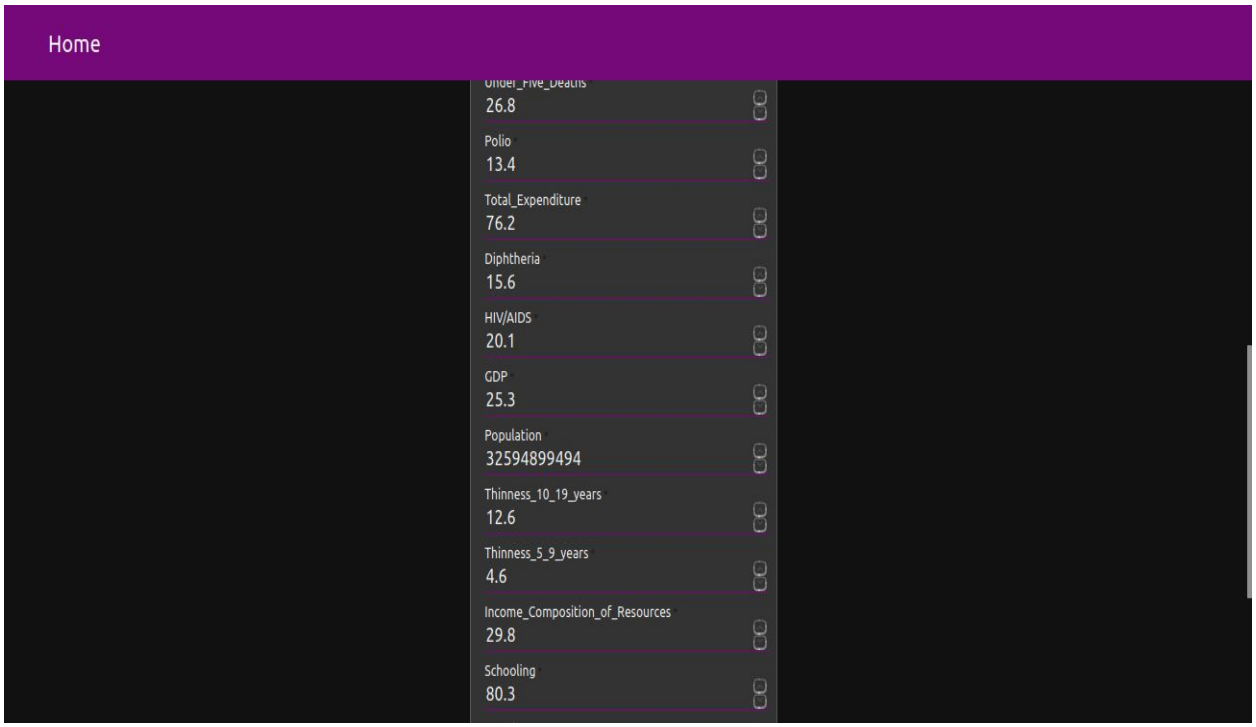


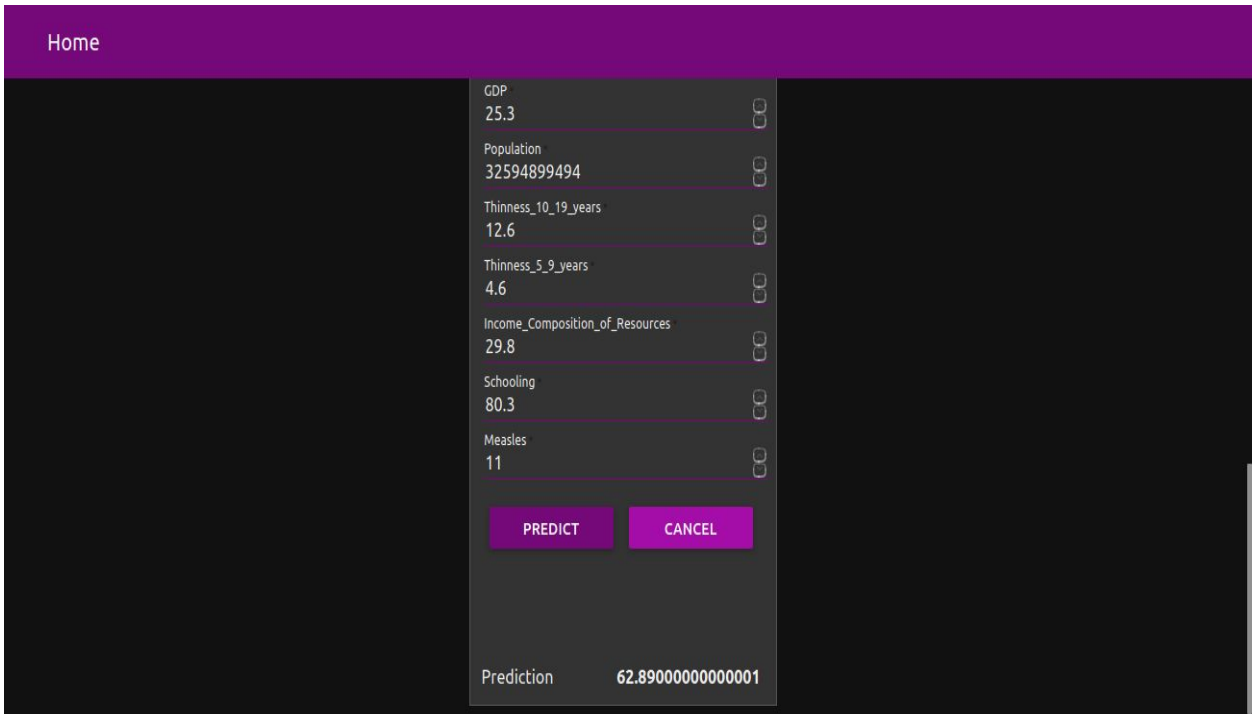
5. FLOWCHART


Input Values
Machine Learning Model
Prediction
Output Web App

6. RESULTS:


Home
Life Expectancy
Country
India
Year
2007
Status
Developed
BMI
22.5
Adult_Mortality
16.2
Infant_Deaths
25.1
Alcohol
116
Percentage_Expenditure
60.5
Hepatitis_B
14.2
Under_Five_Deaths
26.8


Home
26.8
Polio
13.4
Total_Expenditure
76.2
Diphtheria
15.6
HIV/AIDS
20.1
GDP
25.3
Population
32594899494
Thinness_10_19_years
12.6
Thinness_5_9_years
4.6
Income_Composition_of_Resources
29.8
Schooling
80.3


Home
GDP
25.3
Population
32594899494
Thinness_10_19_years
12.6
Thinness_5_9_years
4.6
Income_Composition_of_Resources
29.8
Schooling
80.3
Measles
11
PREDICT
CANCEL
Prediction
62.89000000000001

7. Advantages And Disadvantages:

Advantages:-

• Can be used by any organization to predict.

• The observations in the dataset used 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.

• Some of the past research was done considering multiple linear regression based on a data set of one year for all the countries. But the dataset used for training the model contained data of the past 15 years to give a fairly better prediction.

• The application is easy and simple to use.

• The machine learning algorithm used in the project is Random Forest regression which is based on the bagging algorithm and uses Ensemble Learning technique. It creates as many trees on the subset of the data and combines the output of all the trees. In this way it reduces overfitting problems in decision trees and also reduces the variance and therefore improves the accuracy.

Disadvantages:

• Can be only used by the people having the knowledge of data analysis.

• As the model is deployed on cloud, so one requires good internet connection to use the application.

• The Node-Red application needs to make an HTTP request to IBM cloud and then another HTTP request to the model before providing the prediction. That

makes the app a bit slow.

8. Applications:

- This will help in suggesting a country which area should be given importance in order to efficiently improve the life expectancy of its population.
- It will be easier for a country to determine the predicting factor which is contributing to lower value of life expectancy and can be used in various organizations to improve the quality of service.
- The project can be used as a basis to develop personalized health applications.
- The governments can plan and develop their health infrastructures by keeping the most correlated factors in mind.
- The project can help governments to keep track of their country's health status so they can plan for the future accordingly.

9. Conclusion:

By doing this project I have successfully created a Life expectancy prediction system using IBM Watson studio, Watson machine learning and Node-RED service. Gained knowledge about IBM Cloud, IBM Watson with its services, Node-Red and how to integrate Node-Red with Machine Learning Model. The project makes a good use of machine learning in predicting the life expectancy of a country that can help the respective government in making policies that will serve for the benefit of the nation and entire humankind.

10.Future Scope:

• 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.

• Currently, the project is just a web application. It can be developed to support other platforms like Android, IOS and Windows Mobile.

• Other regression models can also be used for prediction and later the best among them should be chosen.

11.Bibliography

[1] https://cloud.ibm.com/

[2] https://www.ibm.com/cloud/get-started

[3] https://developer.ibm.com/tutorials/how-to-create-a-node-red-starter-application/

[4] https://www.ibm.com/watson/products-services

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

[6] https://www.researchgate.net/publication/45868855_Predicting_Human_Lifespan_Limits

[7] https://www.thelancet.com/action/showPdf?pii=S0140-6736%2818%2931694-5

APPENDIX

A. Source Code

Life_expectancy_prediction.ipynb

https://dataplatform.cloud.ibm.com/analytics/notebooks/v2/360063d2-a010-45fb-b80a-f22facd8ef1a/view?access_token=f45fb48f48b1cef6e7f3c7e1d7c66dac207bd989c468d90f0c44dc8118417b38