Predicting Life Expectancy Using Machine Learning

Prepared by: RAJAT VERMA

1. INTRODUCTION

1.1. Overview:

This project “Predicting Life Expectancy using Machine Learning” is a web application that predict the expected average life span of people of a given country based on various features. This project is built using IBM services (Watson studio, Node Red, Watson machine learning).

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: Country, Status, infant deaths, GDP, Population, BMI, other 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.

1.2. Purpose:

The average life Expectancy of a certain country says many things about that particular country. It ultimately helps in predicting the health conditions and the development of the health sector in that particular country. This ultimately helps the nation to find the area which needs attention in an urge to improve its contribution in average lifespan of a human being. The expectancy obviously depends upon the country's population, GDP, the economy of the country and many more factors. It is not enough to have a long life, instead with having a long life one should have a fit life as well.

2. LITERATURE REVIEW

2.1. Existing Problem:

Past studies have revealed a lot of work in the field of predicting life expectancy of a human being. After reviewing existing works and techniques in the prediction of human Life Expectancy, and finally reached a conclusion that it is possible to predict a Average Life Expectancy for individuals using advancing technologies and devices such as big data, AI, machine learning techniques, and PHDs, wearables and mobile health monitoring devices, IOT. It is noticed that the collection of data is a huge challenge due to the privacy and government policy considerations, which will require collaboration of various bodies in the health industry. The interworking of a heterogeneous health network is also a challenge for data collection. Despite these challenges, a possibility of predicting Life by proposing an approach of data collection and application by smartphone, in which users can enter their information to access the cloud server to obtain their own predicted Lifespan based on the given inputs. To verify the accuracy of PLE prediction and validation of data quality, big data techniques and analysis algorithms need to be developed and tested in a real-life situation with several sample groups. As artificial intelligence

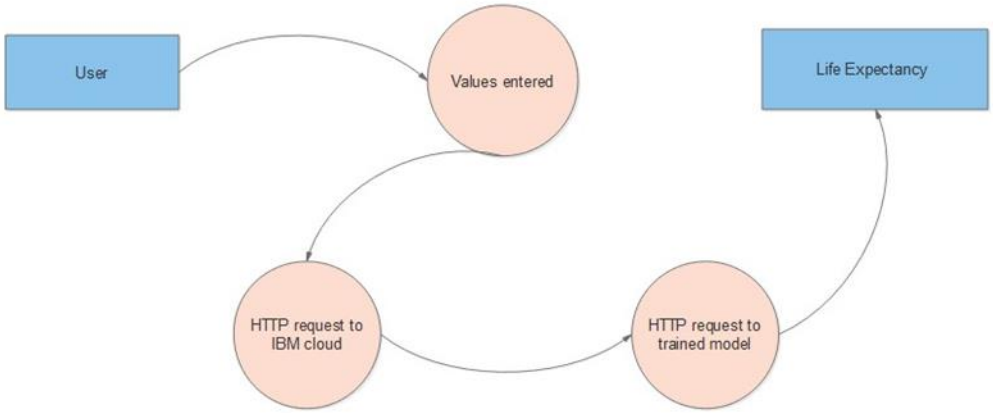
technology is evolving and being applied rapidly, feasibility may be increasing to collect health data from the public as well as existing health agencies such as centralized health servers.

2.2. Proposed Solution:

Although there have been a lot of studies undertaken in the past on factors affecting life expectancy considering demographic variables, income composition and mortality rates. It was found that the effect 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 a 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 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. The model of" Predicting Life Expectancy using Machine Learning" uses IBM Cloud services, which helps to avoid any storage issues. The UI Presented to the users is a website link i.e. on user's fingertips.

3. THEORITICAL ANALYSIS

3.1 Block/Flow Diagram:

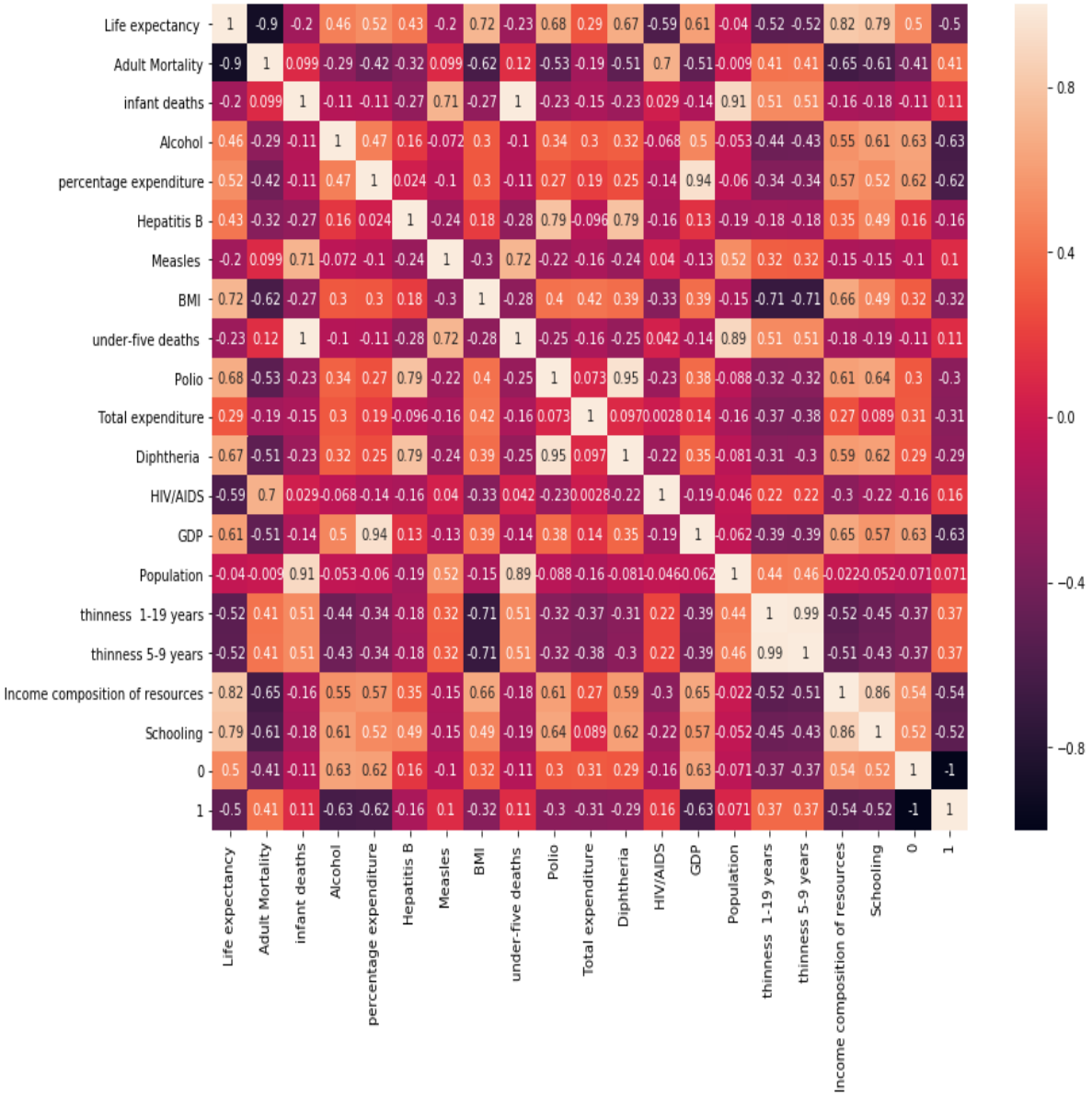


3.2 Hardware / Software designing:

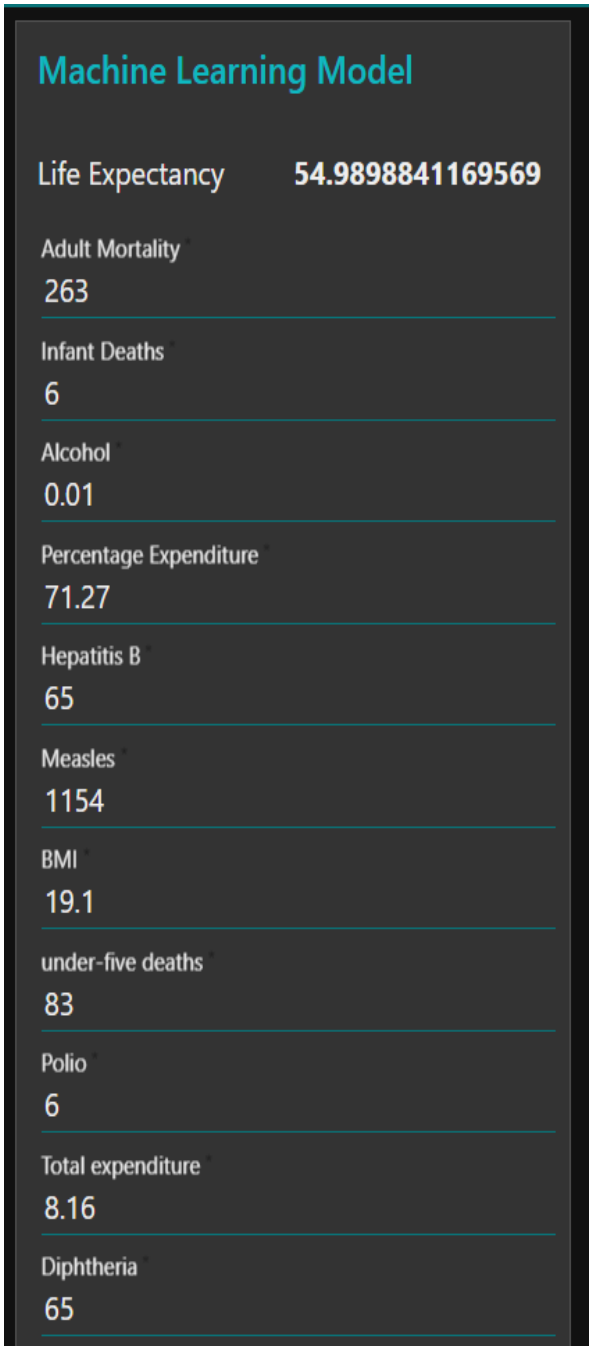
1. Create necessary IBM Cloud services
2. Create Watson studio project
3. Configure Watson Studio
4. Create IBM Machine Learning instance
5. Create machine learning model in Jupyter notebook
6. Deploy the machine learning model
7. Create flow and configure node
8. Integrate node red with machine learning model
9. Deploy and run Node Red app.

4. EXPERIMENTAL INVESTIGATIONS:

These are some graphs from the refined data analysis, that make s us understand collinearity.

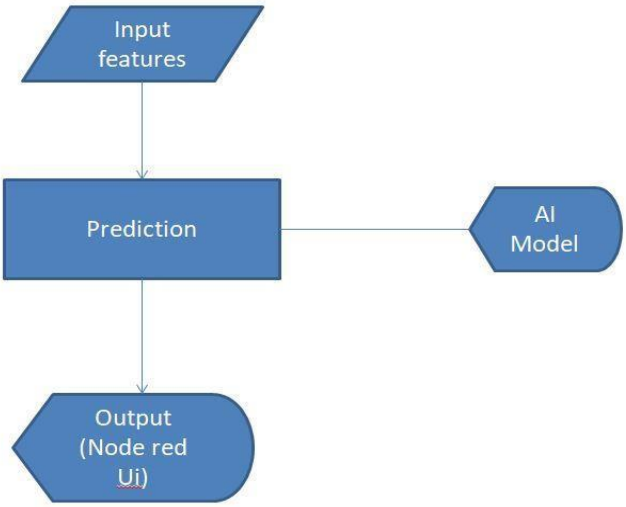


Some random inputs are given to the deployed machine learning model and output is here:



4. FLOWCHART

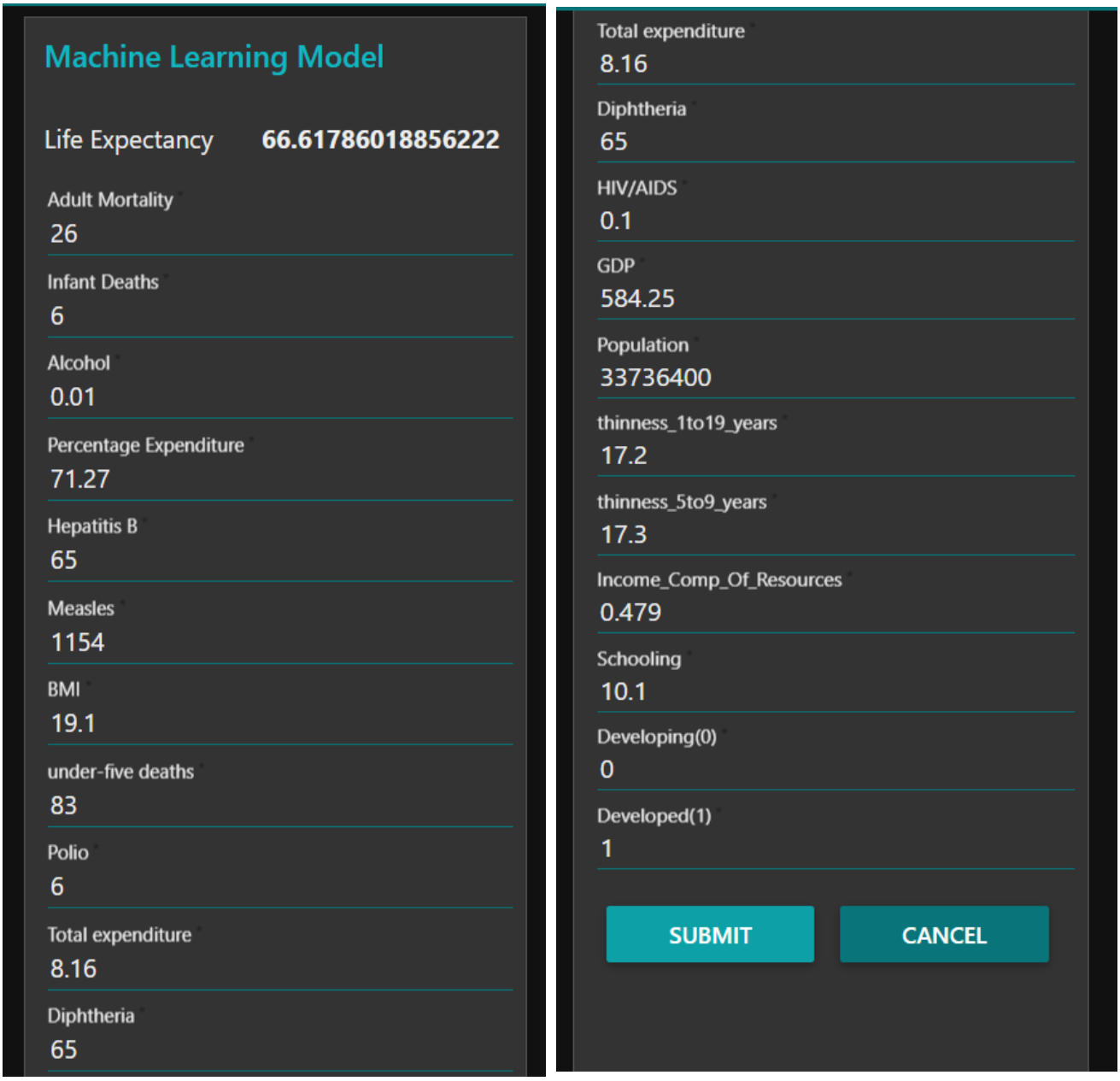
A flowchart is a diagram that depicts a flow of process, system or computer algorithm. They are widely used in multiple fields to document, study, plan, improve and communicate complex processes in clear, easy-to-understand diagrams. Flowcharts, sometimes spelled as flow charts, use rectangles, ovals, diamonds and potentially numerous other shapes to define the type of step, along with connecting arrows to define flow and sequence.



5. RESULTS

Finally, our Node-RED dash board integrates all the components and displayed in the Dashboard UI by typing URL https://node-red-hvcxr.eu-gb.mybluemix.net/ui/#!/0?socketid=BZrAL4-qmCnPY1ajAAAk

The output is the prediction of average life span of a person in years.



7. ADVANTAGES & DISADVANTAGES

7.1 Advantages:

1) 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.

2) The data-sets are made available to public for the purpose of health data analysis.

3) Can be used in any organization to analyze the data.

4) This will help in suggesting a country, which area should be given importance in order to efficiently improve the life expectancy of its population.

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

6) The application is easy and simple to use.

7.2 Disadvantages:

1. Error in data can result in wrong prediction
2. Accuracy is not 100%
3. Error may occur due to inappropriate analysis of data.

8. APPLICATIONS

1. This will help in suggesting a country which area should be given importance in order to efficiently improve the life expectancy of its population.

2. 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 organization to improve the quality of service.

3. The project can be used as a basis to develop personalized health applications.

4. The governments can plan and develop their health infrastructures by keeping the most correlated factors in mind.

5. The project can help governments to keep track of their country's health status so they can plan for the future accordingly.

9. CONCLUSION

Thus, we have developed a model that will predict the life expectancy of a specific demographic region based on the inputs provided. The potential use of project is not limited to health care in practice, but could also be useful in other clinical applications such as clinical trials. Users can interact with the system via a simple Graphical user interface which is in the form of a form with input spaces which the user needs to fill the inputs into and then press the "predict" button.

10. FUTURE SCOPE

As future scope, we can connect the model to the database which can predict the life Expectancy of not only human beings but also of the plants and different animals present on the earth. This will help us analyze the trends in the life span. A model with country wise bifurcation can be made, which will help to segregate the data demographically.

APPENDIX

A. Dataset Link:

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

B. Demonstration Video Link:
   https://drive.google.com/file/d/1nquzWwXxLVVHyx1zKhgnSDGw3YjQGG1w/view?usp=sharing

C. GUI url link:

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

D. Source Code (GitHub link):

https://github.com/SmartPracticeschool/llSPS-INT-2600-Predicting-Life-Expectancy-using-Machine-Learning/blob/master/Predicting%20Life%20Expectancy%20Model.ipynb