IISPS_INT_3299_Predicting Life Expectancy using Machine Learning

Project Report On:

Prediction of Life Expectancy Using Machine Learning

Project Done By:

Nikhil Kumar Cherukuri

B.Tech(Third Year)

e-mail id: chnikhilkumar0906@gmail.com

Mobile Number: +91 83099 67932

linkedin id: https://www.linkedin.com/in/nikhil-0906/

IN

SUMMER INTERNSHIP PROGRAM 2020 by SMART INTERNZ

1. INTRODUCTION

The term Life Expectancy refers to the number of years a person can expect to live. By definition, life expectancy is based on the estimate of the average age that members of a particular population group will be when they die. Numbers of factors will come into picture while predicting the life expectancy of a country like schooling, status, adult mortality, total expenditure, and other demographic factors. This project aims to predict life expectancy using Machine Learning.

1.1 Overview:

Life expectancy at birth is an estimate of how long a person born today would live, on average, if current mortality rates in every age group remained constant throughout the person's life. It is a way to summarize current mortality rates in an easily understood measure to which most people can directly relate.

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

We can also define life expectancy as a statistical measure of the average time a human being is expected to live, it 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 the 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 the healthcare system, and some specific disease-related deaths that happened in the country are given.

The data set is called Life Expectancy(WHO) collected from the Kaggle website. The project relies on the accuracy of the data. The Global Health Observatory (GHO) data repository under the World Health Organization (WHO) keeps track of the health status as well as many other related factors for all countries. The data sets are made available to the public for health data analysis, this project aims to use the same for predicting the Life Expectancy. The data set 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. 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 the improvement of

human mortality rates especially in the developing nations in comparison to the past 30 years. Therefore, this proposed project has considered data from the year 2000-2015 for 193 countries for further analysis. The individual data files have been merged into a single data set. On initial visual inspection of the data showed some missing values. As the data sets were from WHO, they haven't found any evident errors. The results indicated that most of the missing data were for population, Hepatitis B, and GDP. The missing data were from less known countries like Vanuatu, Tonga, Togo, Cabo Verde, etc. Finding all data for these countries was difficult and hence, it was decided to exclude these countries from the final model data-set. The final merged file(final dataset) consists of 22 Columns and 2938 rows which meant 20 predicting variables. All predicting variables were then divided into several broad categories: Immunization related factors, Mortality factors, Economical factors, and Social factors.

Here in this project, a different type of python functions like interpolate, dropna, more are planned to use manage and to fill or drop the missed or empty data in the dataset and also other normal statistical methods like mean, std, more. After checking the accuracy of the model, the method which helps for increasing the model will be fixed for missing data in the overall project.

1.2 Purpose:

The purpose of this project is to predict the life expectancy of a country using the Life Expectancy(WHO) dataset for a country. The dataset consists of data for nearly 193 countries including factors like Schooling, Immunization factors, HDI, Expenditure, Mortality rates, and more for the years 2000 to 2015. The life expectancy of a country can help in estimating the health conditions, Immunization factors, Mortality rates to the other country's people without knowing all the factors of their country in detail. Various factors that affect life expectancy are used as features in this prediction. So if the life expectancy is low, then the nation should plan for managing or increasing the respective factors for better life expectancy. Some of them are like GDP per capita increases the life expectancy at birth through increasing the economic growth and development in a country and thus leads to prolongation of longevity. Immunization factors matter a lot for the life expectancy of the people, if they are immune power is less in people of their country, the nation's life expectancy will be very low as it related to health conditions of the people. Higher income also implies better access to housing, education, health services, and other items which tend

to lead to improved health, lower rates of mortality, and higher life expectancy. It is not surprising, therefore, that aggregate income has been a pretty good predictor of life expectancy historically. Hence one can conclude life expectancy is affected by many factors such as socioeconomic status, including employment, income, education and economic wellbeing; the quality of the health system and the ability of people to access it; health behaviors such as tobacco and excessive alcohol consumption, poor nutrition and lack of exercise; social factors; genetic factors; and environmental factors including overcrowded housing, lack of clean drinking water and adequate sanitation, and hence it is a way to summarise current mortality rates, and other factors which mentioned above in an easily understood measure to which most people can directly relate. This project is planned to predict life expectancy through a user interface in which all people can access it from their places.

2. Literature Survey

2.1 Existing Problem:

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 the motivation to resolve both the factors stated previously by formulating a regression model based on the 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 this dataset is based on different countries, it will be easier for a country to determine the predicting factor which is contributing to a lower value of life expectancy. This will help in suggesting a country in which areas should be given importance to efficiently improve the life expectancy of its population.

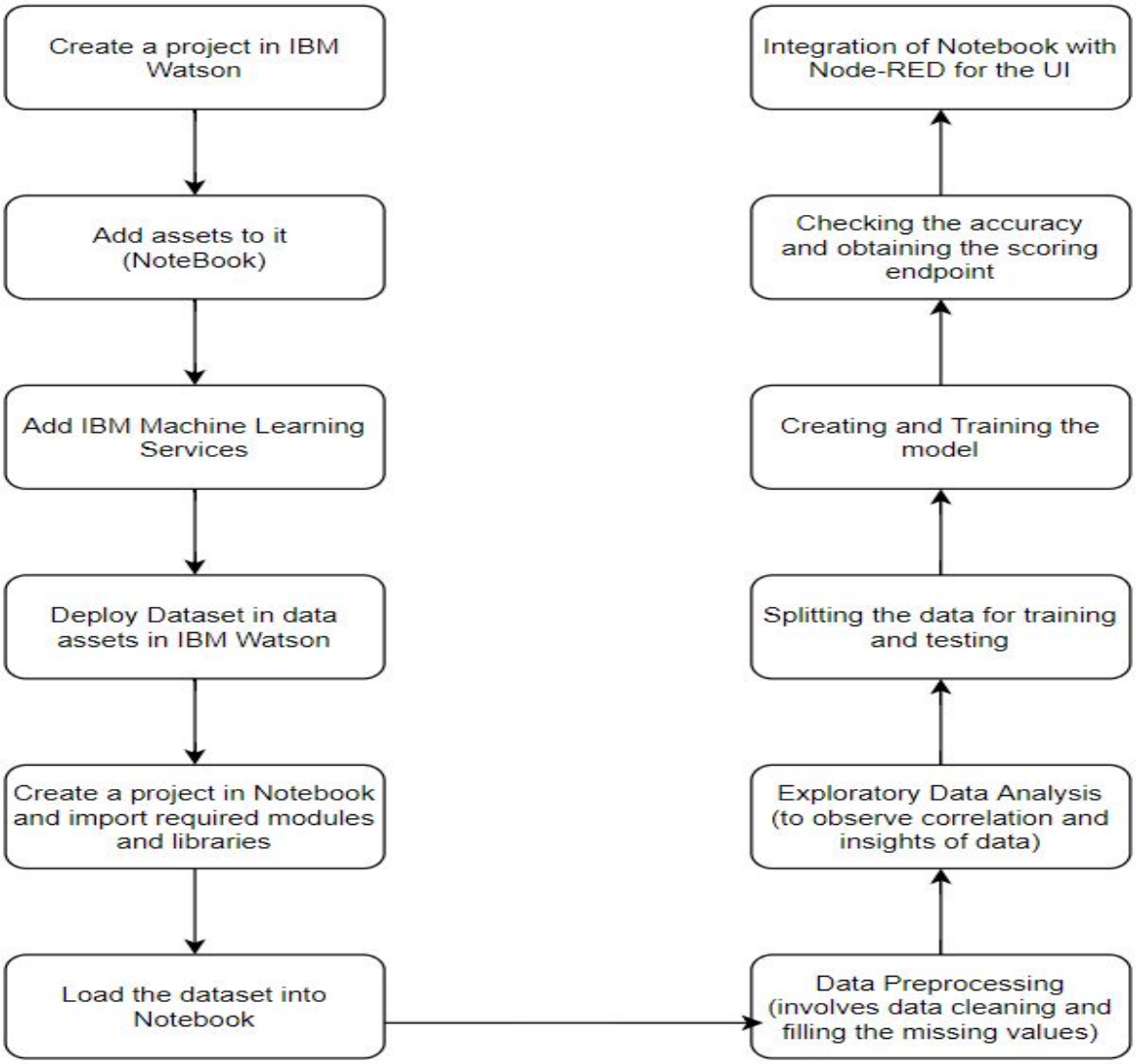
2.2 Proposed Solution :

In this proposed solution the project considers the immunization and human development index of the people in the countries. As life expectancy is a continuous value,

the regression analysis is suitable for the prediction. So, this project aims at using the different regression learning algorithms like Support Vector Regression, Random Forest Regressor for predicting the life expectancy of a country using Machine Learning. And also creating a user interface for the prediction of life expectancy which any user can access through the link to predict their country's life expectancy by giving the required inputs such as Schooling, Immunization factors, HDI, Expenditure, Mortality rates and more.

3. THEORITICAL ANALYSIS:

3.1 Block Diagram:



- Creating accounts in Github, slack app, and the Zoho writer.
- Cloning GitHub repository in the local system by using the GIT tool.
- Installing the right IDE for coding and other requirements.

- Creating an account in the IBM cloud.
- Creating the project in the IBM Watson studio services and also other IBM services.
- Adding the required Watson services to the project(Notebook, machine learning).
- Creating a Node-RED starter Application.
- Exploring IBM Watson studio and learning how to Integrate nodes in the Node-red Starter application.
- Downloading the dataset and Starting the data cleaning and processing.
- Writing the code in the Notebook for the entire project up to score endpoint.
- Linking the services to Node-RED and Preparing a webpage for the entire project using the Node-RED Starter application.
- Uploading the Source codes into Github and also project documentation.
- Creating the video of project demonstration and uploading it to drive or GitHub.

3.2 SOFTWARE/HARDWARE REQUIREMENTS:

Project Requirements :

- IBM Cloud account
- IBM Watson Services
    - IBM Watson Studio
    - IBM python Notebook
    - IBM Machine Learning
- Node-Red Starter application linked with IBM cloud
- Life Expectancy(WHO) Dataset.

Functional Requirements :

- Data loading into Notebook.
- Data manipulation and cleaning
- Exploratory Data Analysis(EDA)
- Preparing a Machine Learning model.
- Linking API and instance-id keys (Machine Learning Services) to the Notebook and scoring endpoint to the Node-Red Starter application.
- Specific inputs like BMI, adult mortality, thinness, schooling, and more are must be required and should be given by the user for the prediction of Life Expectancy rate.
- Connecting the required nodes appropriately for the required web page.

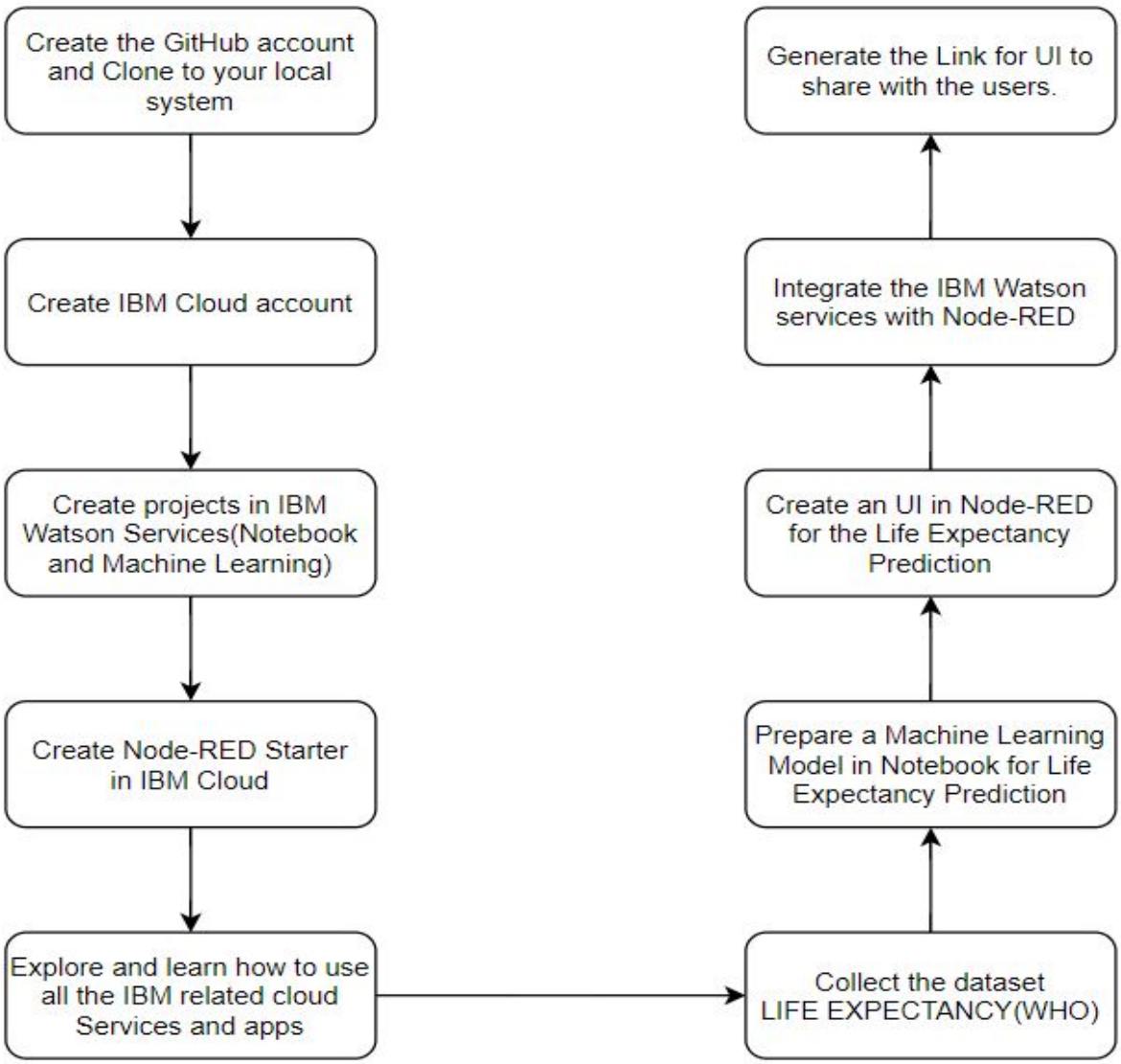
Technical/Software Requirements :

- Python Programming Language.
- JS Basics.
- Machine Learning.
- Can be used/deployed in any OS.(Laptop)
- Minimum Net requirents for running the web page.
- User Interface is integrated with the backend of the trained model.

4. FLOWCHART:

A flowchart is simply a graphical representation of steps. A flowchart can also be used to define a process or project to be implemented.

The following flowchart explains the entire work flow of this project(not in detail), but the major things are as followed to proceed in the project.



Node-RED Flow:

The following figure shows the flow of Node-RED application for the User Interface.

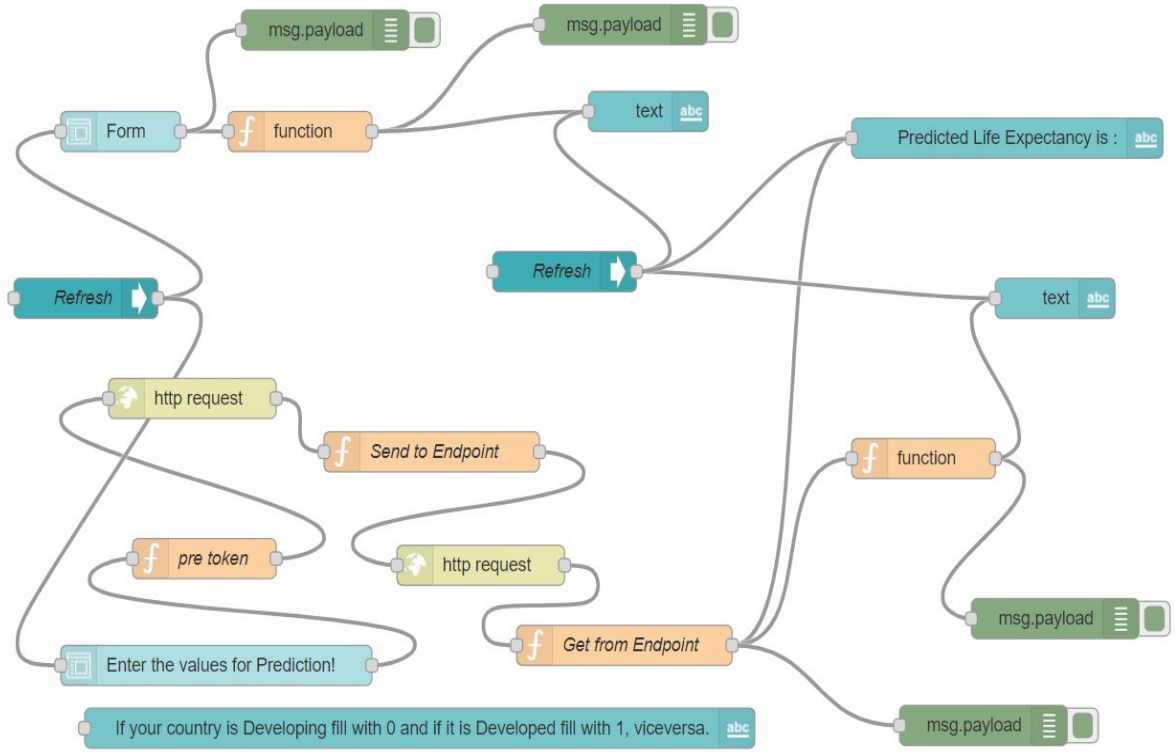


Figure-1

5. RESULT

The Life Expectancy prediction using Machine Learning got 96.1% of accuracy using the Random Forest Regressor in scikit learn with the estimators of 100. Different Learning algorithms are used to train the Life Expectancy model like Linear Regression, Support Vector Regression(SVR) but the accuracy is around 82.2% with the linear kernel of SVR and almost the same for Linear Regression. When you add countries as extra columns by using the dummies, the accuracy of both SVR and linear regression was increased to 94.1% (approx). But Random Forest Regressor can give better accuracy without dummies of the country feature. So the model which is trained with Random Forest Regressor is successfully integrated with the Node-RED Starter application as back end for the User Interface. The following figures show the User Interface page.

- Figure-2 shows the Home page of User Interface page.

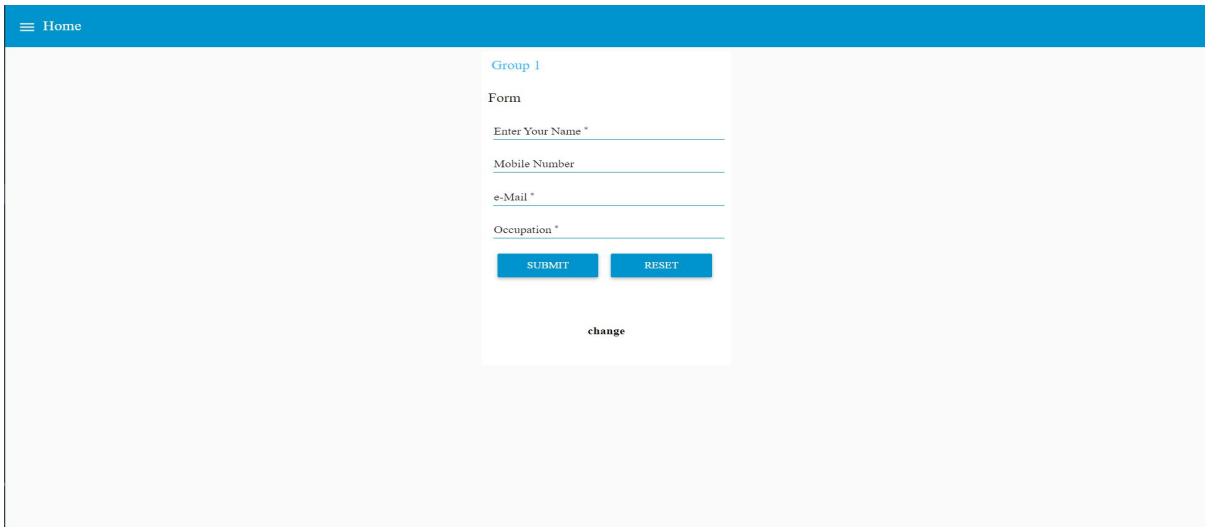


Figure-2

- Figures-3,4 shows the Life Expectancy Prediction page of User Interface.

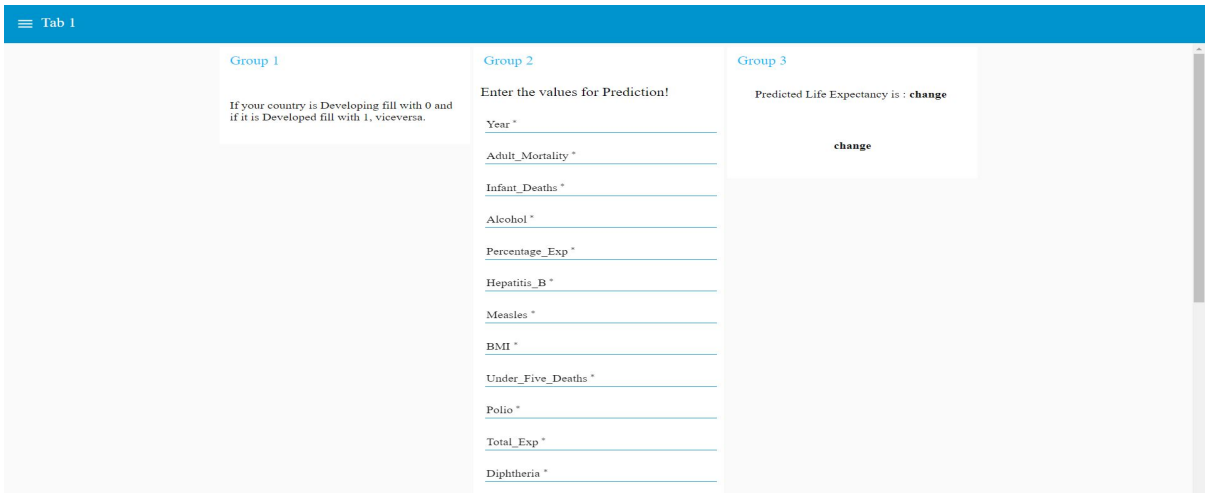


Figure-3

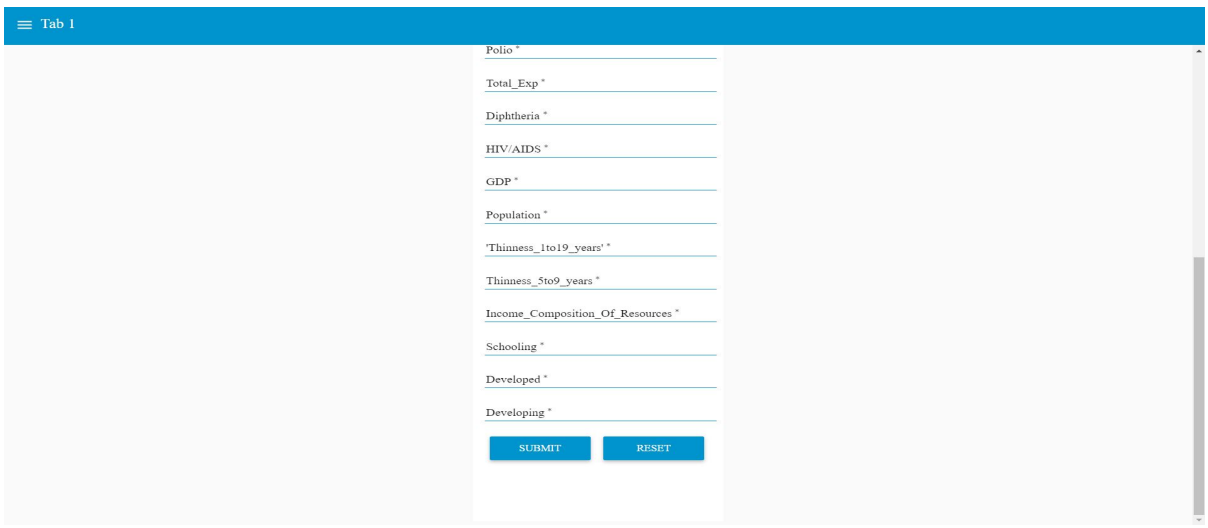


Figure-4

- Figures-5,6,7 shows the User Interface Page after filling with predicted results.

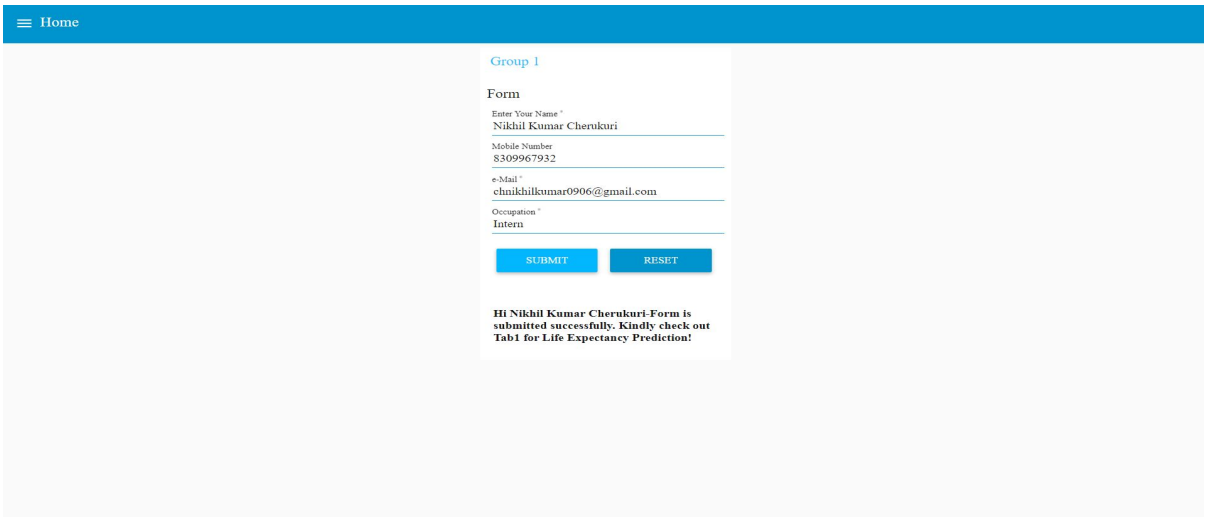


Figure-5

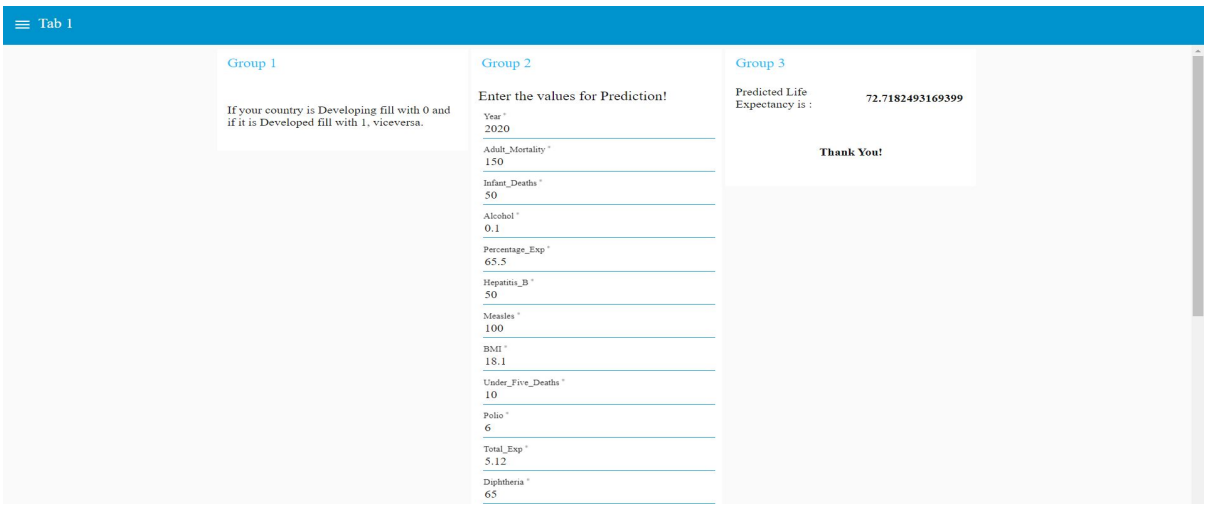


Figure-6

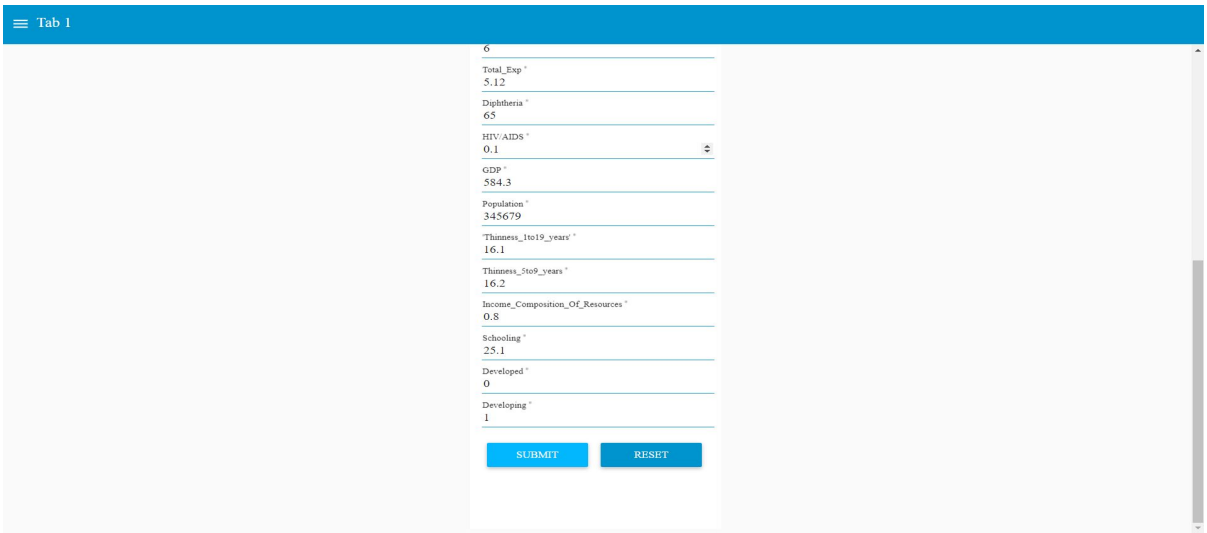


Figure-7

6. ADVANTAGES and DISADVANTAGES

6.1 Advantages:

Advantages of using major services of IBM like IBM cloud, Watson Studio, Node-RED starter:

- Easy to build and make integrations.
- Processes unstructured data.
- Fills human limitations.
- Acts as a decision support system don’t replace humans.
- Improves performance + abilities by giving the best available data.
- Improve and transform customer service.
- Handle enormous quantities of data.
- Sustainable Competitive Advantage.
- Node-RED is flow-based visual programming, all the functions you need from prototype to production will be available, perfect for using on low-cost hardware, built on Node.js, a mature technology stack with the largest ecosystem of open source components and finally, it makes life on the edge.

Predicting Life Expectancy using Machine Learning will help us to know the life expectancy of a country every year in a short period using UI provided. As life Expectancy allows to estimate so many things of a country like the features which are used for training the model. This is one of the advantages that anyone can know the life expectancy of their country using the user interface with 96.1%(approx.) of accuracy.

6.2 Disadvantages:

Disadvantages of using IBM services:

- Only in English (Limits areas of use).
- It is seen as a disruptive technology.
- It Needs Maintenance.
- Doesn't process structured data directly.
- The increasing rate of data, with limited resources.
- Barriers of adoption are high switching costs, takes time to integrate Watson and it's services into a company takes time and effort to teach Watson to use it to its full potential.

7. APPLICATIONS:

- One important method of assessing the health of a population is to ask how long people can expect to live. Life expectancy, usually reported at birth although it can be applied to other ages as well, is a commonly used summary measure that can also be used to compare against countries.
- Other characteristics can also be used to distinguish different risk factors for life expectancy, such as smoking status, occupation, socio-economic class, and others. More complex analyses for assessing cancer survival, that involves comparisons between two populations or a population in two points in time can also be undertaken.
- In addition to public health domains, it is also used by insurance companies and actuary departments.
- When used in biology, age-specific fertility rates are also included in the calculations.
- When data have not been available, such as in low-income countries, life tables have been modeled using what data are available, usually childhood mortality data.

8.CONCLUSION:

Thus in this project, a model and the user interface were developed for the prediction of life expectancy using the various factors as features that affect life expectancy rates of countries such as Schooling, Immunization factors, Mortality rates, Alcohol rate, and more. So, a user can use UI for predicting the life expectancy of their country by providing the accurate inputs(which are shown above).

9. FUTURE SCOPE:

As this project mainly focuses on the prediction of life expectancy of human beings in a country, the future scope is on predicting the life expectancy of animals as they are becoming extinct for long years. It will be very helpful if we know the factors which are affecting the animal's life span and allowing the national officials to concentrate on managing the factors or requirements that are required for increasing the life expectancy of animals. As every living thing has equal rights on this earth it is important to consider and concentrate on the factors which affect the life expectancy of the animals.

APPENDIX:

1. DATA SET can be found here:

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

2. GitHub repository Link:

https://github.com/SmartPracticeschool/llSPS-INT-3299-Predicting-Life-Expectancy-using-Machine-Learning

3. Google Drive Link

https://drive.google.com/drive/folders/1d_9JmhM0KNxeg8NqGRO91FuPG_DWET-E?usp=sharing