Remote Summer Internship Program 2020

Machine Learning, Career Basic Program

Smartinternz, SmartBridge



Predicting Life Expectancy using

Machine Learning

Internship Report

By

Aditya Reddy Lellapati



From:

SmartBridge Educational Services Pvt Ltd. Plot No 132, Above DCB bank, 2nd floor, Bapuji Nagar, Habsiguda,

Nacharam Main Road, Hyderabad – 500076

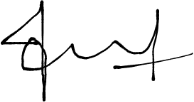
Date: 17/05/2020.

Dear **Lellapati Aditya Reddy**

**SmartBridge Educational Services Pvt Ltd**, is pleased to offer a training cum internship opportunity. During this period you would be associated with our mentors and The Smart Practice School Platform.

For further details you can contact us on +91 8499004200.

Thanks and Regards,



Ch. Jaya Prakash

Program Manager – SIP2020, Date: 17/05/2020.

1 **INTRODUCTION**

1.1 Overview

1.2 Purpose

**2 LITERATURE SURVEY**

2.1 Existing problem

2.2 Proposed solution

**3 THEORITICAL ANALYSIS**

3.1 Block diagram

3.2 Hardware / Software designing

**4 EXPERIMENTAL INVESTIGATIONS**

**5 FLOWCHART**

**6 RESULT**

**7 ADVANTAGES & DISADVANTAGES**

**8 CONCLUSION**

**9 BIBILOGRAPHY**

**APPENDIX**

A. Source code

**INTRODUCTION**

**1.1 Overview**

**Program Information**

Summer Internship Program by Smartbridge in collaboration with IBM is an annual initiative taken up by them to teach and prepare students across the globe for industry experience. They believe that experiential learning and development in a professional like environment can only bridge the gap between students and industries opening ways for both of them to achieve better results. This initiative enables students to better their resume to embark upon a successful industrial journey. They provide various roles for internship according to the possible aptitude of the students such as Artificial Intelligence, Machine Learning, Internet of Things etc.

Remote Summer Internship Program 2020 is a 4-Weeks Industry Driven, Project based, Mentor supported Program. Selected interns will be joined to a project team, where he will learn the technology and complete the assigned tasks. SmartInternz platform provides the company like collaborative environment to build professional skills along with technical skills.

A project is assigned to students individually with access to the platform wherein the students code and deploy their model just like in industries. This not only introduces them to environments like IBM cloud but also helps them in understanding the industrial environment better. Classes are organised according to the technologies to be taught and doubt sessions are taken up by the mentors to help students complete their projects too.

**1.2 Purpose**

**Project title :** Predicting Life Expectancy using Machine Learning

**Category :** Machine Learning

**Skills required :** Python, IBM Cloud, IBM Watson

**Problem Description**

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 measue of the average time a human being is expected to live, Life expectancy depends on various factors : Regional variations, Economic Circumstacnes, Sex Differences, Mental Illnesses ,Physical Illnesses , Education, Year of the ir birth ans other demographic 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.

The purpose of this project is that the people from various places can easily predict their life expectancy by providing the inputs asked by the model. This software can be used by all people in the world because the training part of this model contains inputs and predictions of more number of countries.

**Economic growth**

Predicting life expectancy would play a vital role in judging the growth and development of the economy. Across countries, high life expectancy is associated with high income per capita. Increase in life expectancy also leads to an increase in the “manpower” of a country. The knowledge asset of a country increases with the number of individuals in a country.

**Population Growth**

Helps the government bodies take appropriate measures to control the population growth and also direct the utilization of the increase in human resources and skillset acquired by people over many years.

**Personal growth**

This project would also help an individual assess his/her lifestyle choices and alter them accordingly to lead a longer and healthier life. It would make them more aware of their general health and its improvement or deterioration over time.

**Growth in Health Sector**

Based on the factors used to calculate life expectancy of an individual and the outcome, health care will be able to fund and provide better services to those with greater need.

**Insurance Companies**

Insurance sector will be able to provide individualized services to people based on the life expectancy outcomes and factors.

**LITERATURE SURVEY**

**2.1 Existing Problem**

Although there have been lot of studies undertaken in the past on factors affecting life expectancy considering demographic variables, income composition and mortality rates. It was found that affect 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 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 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.

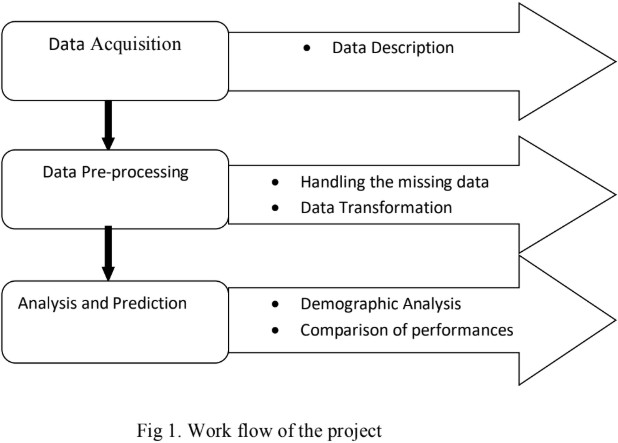
The project relies on accuracy of data. The Global Health Observatory (GHO) data repository under World Health Organization (WHO) keeps track of the health status as well as many other related factors for all countries The data-sets are made available to public for the purpose of health data analysis. The data-set related to life expectancy, health factors for 193 countries has been collected from the same WHO data repository website and its corresponding economic data was collected from 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 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 year 2000-2015 for 193 countries for further analysis. The individual data files have been merged together into a single data-set. On initial visual inspection of the data showed some missing values. As the data-sets were from WHO, we found no evident errors. Missing data was handled in R software by using Missmap command. The result indicated that most of the missing data was 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 that we 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 was then divided into several broad categories:​Immunization related factors, Mortality factors, Economical factors and Social factors.

**2.2 Proposed Solution**

The 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](https://www.kaggle.com/kumarajarshi/life-expectancy-who/data%20). The output algorithms have been used to test if they can maintain their accuracy in predicting the life expectancy for data they haven’t been trained.

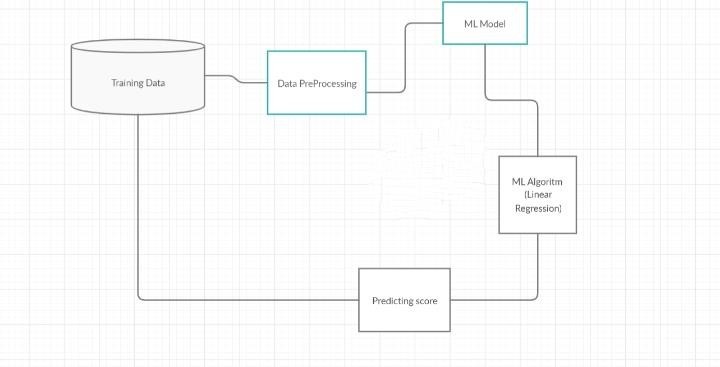
So many people were expecting to use a model of life expectancy prediction. In order to that, many institutions and companies are leading their team to build that model in this regard. In my project, I have proposed a solution to predict the life expectancy using machine learning. Machine Learning is the process of training the computer to think and decide solutions like human. The reason why I have chosen this architecture was only with the help of Machine Learning, deep understanding of the data and an ability to create a model can be done. Design a Regression model to predict life expectancy ratio of a given country based on some features provided such as year, GDP (gross domestic product), education, alcohol intake of people in the country, expenditure on healthcare system and some specific disease related deaths that happened in the country.

**Methodology used**

Work flow for this project can be divided into three sub-tasks. These include acquiring the data and understanding various features of the data, preprocessing the data set to align it with our requirements and remove any inconsistency, and finally analyzing the data using regression based prediction algorithm with the key performance index being accuracy of prediction.

Work Flow of the project

**THEORITICAL ANALYSIS**

**3.1 Block diagram**

**3.2 Hardware / Software designing**

Project Planning and Kickoff

* Understanding the project description and analyze the data and attributes in the given dataset.
* Creating GitHub account
* Installing Slack and create account with the mail id
* Learning to use Zoho writer.

Explore IBM Cloud Platform

* Creating IBM cloud account with the mail id
* Creating IBM academic initiative account with the mail id
* a Node-Red starter application.

Explore IBM Watson services

* Exploring IBM Watson use cases.
* Learning about IBM Watson Machine Learning.

Introduction to Watson Studio

* Learning to build own Machine Learning model using IBM Watson.
* Automate the Machine Learning Model

Predicting Life Expectancy with Python

* Collecting Data set from [www.kaggle.com](http://www.kaggle.com/)
* Creating IBM Watson services
* Create a jupyter notebook and import data from Object storage.

Predicting Life Expectancy without Python

* Import data set and create AutoAI experiment
* Created Node-Red model and integrated with AutoAI

**EXPERIMENTAL INVESTIGATIONS**

# Machine Learning Model

**Algorithm Formulated to solve the given Problem Statement**

Algorithm steps:

Step 1 : Import the Data set

Step 2 : Read and Understand the data

Step3 : Explore the Data set

Step4 : Decide the amount of data for training data and testing data

Step5 : Give 70 percent data for training and remaining data for testing.

Step6 : Assign train data set to the models

Step7 : Choose the algorithm and create the model

Step8 : Make predictions for test data set.

Step9 : Calculate accuracy for the algorithm

Step 10 : Apply the model for further predictions.

**Setting Up the Environment**

An IBM cloud account was set up to access various services to create and deploy the model. The following services have been used in the project:

1. Watson Studio - This is where the notebook has been created in a project to write the regression code along with the data set.

1. Node Red - Node Red is the front end application that uses interconnecting nodes to interact with machine learning services of the cloud and the model to show predictions when inputs are given

## Data Acquisition

### The data set on Statistical Analysis on factors influencing Life Expectancy is provided in the smart internz platform . The below link helps in downloading the data set required to complete this project

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

## Model Requirements

**Python**

Python is a multi-paradigm, general purpose, high level programming language, which focuses on code read- ability. It has a large library, which provides tools for many tasks and has a wide support base. This project uses python 3.5.

**Python Libraries**

**Pandas :** Pandas 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.

**Data Preprocessing**

Data preprocessing is an essential step in order to increase the accuracy of machine learning models. It involves handling inaccurate and missing data, noisy data in the form of outliers, and inconsistent data in the form of duplication and others.

**Data Cleaning**

Data was often not consistent; missing values or values out of range was common. The methods used for cleaning is to replace the missing or noisy values by forward filling them using mean of the feature.

**Analysis and Prediction**

Linear Regression and Random Forest algorithm has been applied to the data set to train the model and increase the accuracy for prediction of the the life expectancy of any given country.

**Data Transformation**

**Scaling :** Scaling is required to standardize the independent feature in the dataset to a fixed range. Primarily, two types of feature scaling methods:

1. Min-max scaling (Normalization) (value - min)/(max - min) Sklearn provides a class called MinMaxScaler for this
2. Standardization (value - mean)/std Sklearn provides a class called StandardScaler for this

**Evaluation Metrics**

Evaluation metrics calculated are:

1. Mean Cross validation score
2. Score without cv
3. R2*score*

**Linear Regression**

Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on – the kind of relationship between dependent and independent variables, they are considering and the number of independent variables being used. Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y(output). Hence, the name is Linear Regression.

**Random Forest Regression**

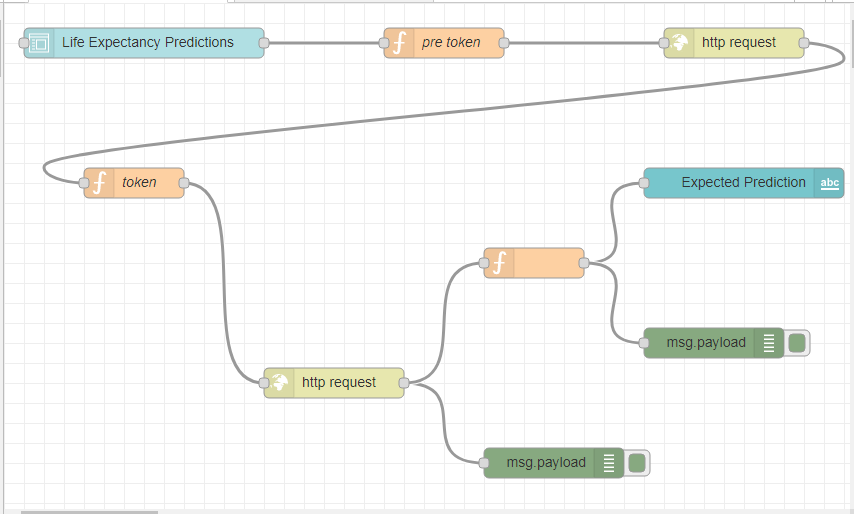
Random forests is a supervised learning algorithm. It can be used both for classification and regression. It is also the most flexible and easy to use algorithm. A forest is comprised of trees. It is said that the more trees it has, the more robust a forest is. Random forests creates decision trees on randomly selected data samples, gets prediction from each tree and selects the best solution by means of voting. It also provides a pretty good indicator of the feature importance.

It technically is an ensemble method (based on the divide-and-conquer approach) of decision trees gener- ated on a randomly split dataset. This collection of decision tree classifiers is also known as the forest. The individual decision trees are generated using an attribute selection indicator such as information gain, gain ra- tio, and Gini index for each attribute. Each tree depends on an independent random sample. In a classification problem, each tree votes and the most popular class is chosen as the final result. In the case of regression, the average of all the tree outputs is considered as the final result. It is simpler and more powerful compared to the other non-linear classification algorithms.

**Node Red Flow**

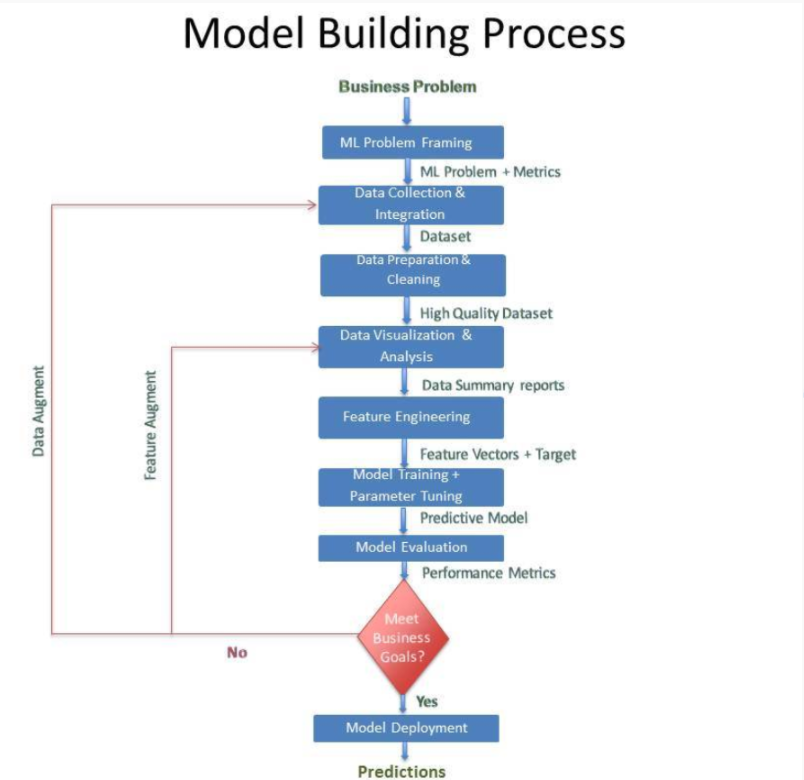
A Node RED starter application was created to implement the front end of the project. In the starter appli- cation, nodes are dragged and dropped to create a flow to integrate the application with the machine learning model.

The following form appears after deployment of the app wherein the user can input values and life expectancy prediction is displayed acccording to the inputs.



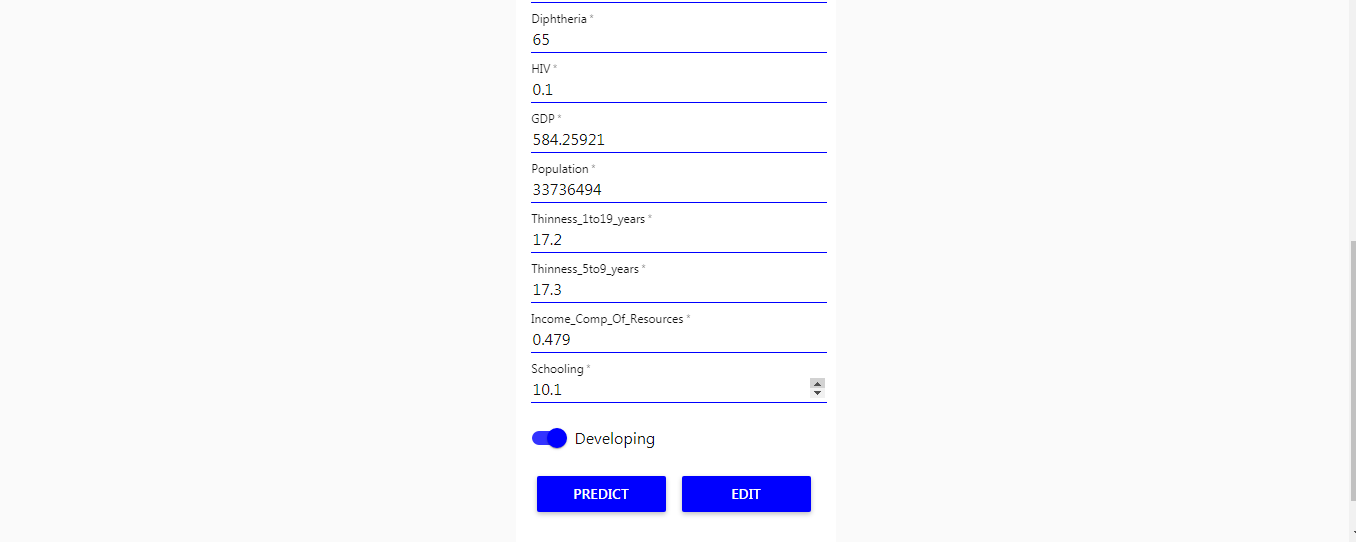
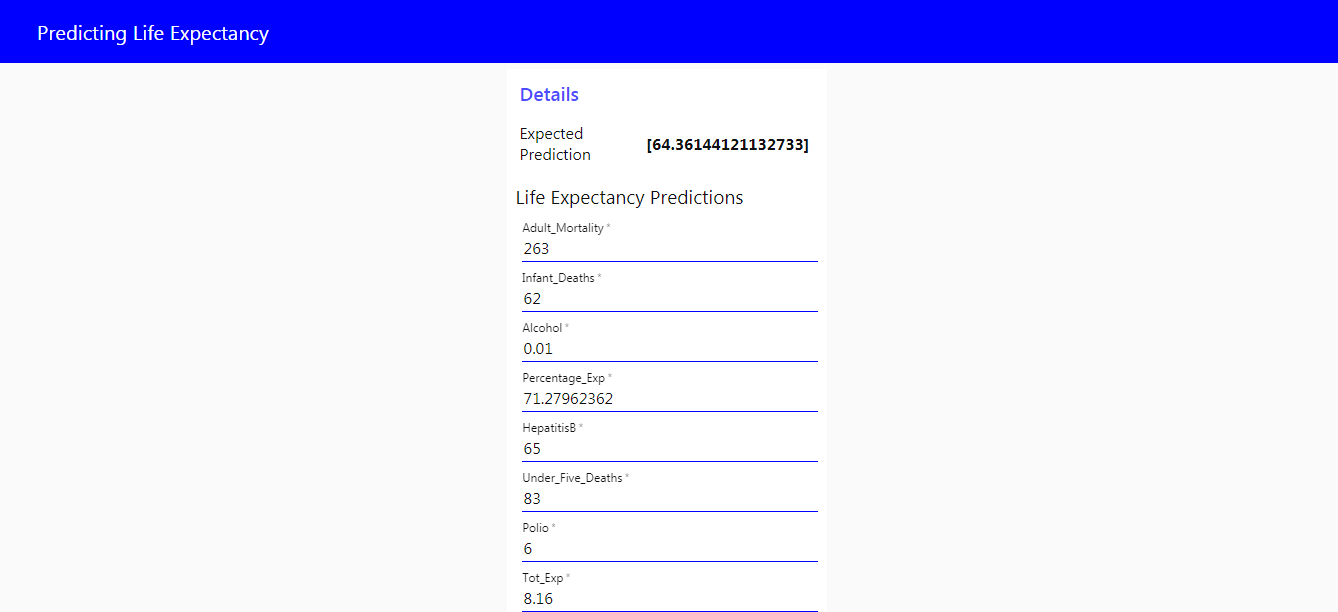
NodeRed flow of the project

**FLOWCHART**

****

**RESULT**

**The final UI of the project**



**ADVANTAGES & DISADVANTAGES**

## Every coin has two faces, each face has its own property and features. It’s time to uncover the faces of ML. A very powerful tool that holds the potential to revolutionize the way things work.

### Advantages

1.Advantages of using IBM Watson:

* Processes unstructured data
* Fills human limitations
* Acts as a decision support system, doesn’t replace humans
* Improves performance + abilities by giving best available data
* Improve and transform customer service
* Handle enormous quantities of data
* Sustainable Competitive Advantage

2. Easy for user to interact with the model via the UI.

3. User-friendly

4. Easy to build and deploy.

5. Doesn’t require much storage space.

### Disadvantages

1.Disadvantages of using IBM Watson

* Seen as disruptive technology
* Maintenance
* Doesn't process structured data directly
* Increasing rate of data, with limited

2. Not connected to database, hence no record of input.

3. Requires internet connect

**Applications**

**Personalized Life Expectancy**

Individuals can predict their own life expectancy by inputting values in the corresponding fields. This could help make people more aware of their general health, and its improvement or deterioration over time. This may motivate them to make healthier lifestyle choices

**Government**

It could help the government bodies take appropriate measures to control the population growth and also direct the utilization of the increase in human resources and skillset acquired by people over many years. Across countries, high life expectancy is associated with high income per capita. Increase in life expectancy also leads to an increase in the “manpower” of a country. The knowledge asset of a country increases with the number of individuals in a country.

**Health Sector**

on the factors used to calculate life expectancy of an individual and the outcome, health care will be able to fund and provide better services to those with greater need.

**Insurance Companies**

Insurance sector will be able to provide individualized services to people based on the life expectancy outcomes and factors.

**CONCLUSION**

I have developed a model that will predict the life expectancy of a specific demographic region based on the inputs provided. Various factors have a significant impact on the life span such as Adult Mortality, Population, Under 5 Deaths, Thinness 1-5 Years, and Alcohol, HIV, Hepatitis B, GDP, Percentage Expenditure and many more. User can interact with the system via a simple user interface which is in the form of a form with input spaces which the user needs to fill the inputs into.

This was my first internship experience, through this I was able to enhance my Machine Learning skills. Though I am a beginner in programming , with the help of the platform provided , It was an enriching experience as I got to work in a professional like environment. The mentors were very helpful with the webinars they conducted on how to proceed with the project. Our doubts were solved on the slack channel regularly.

I particularly found the IBM cloud experience new and useful in improving my industrial skills. The challenging task in this internship is in creating the Node Red flow as a beginner I have faced difficulties but with the help of the mentors through slack channel platform I succeeded in my project , I found it to be valuable and productive in developing my front end integration skills.

This internship has given me a transparent idea on how to proceed with improving my machine learning skills and I am grateful to Smart Bridge for letting me be a part of the Remote summer internship.

**Finally I can proudly say that my lockdown period was very productive.**

**BIBILOGRAPHY**

**Node-RED Starter Application**

<https://developer.ibm.com/tutorials/how-to-create-a->[node-red-starter-application/](https://developer.ibm.com/tutorials/how-to-create-a-node-red-starter-application/)

**Watson studio cloud**

[https://bookdown.org/caoying4work](https://bookdown.org/caoying4work/watsonstudio-workshop/jn.html)

**Dataset Reference**

[[https://www.kaggle.com/kumaraja rshi/life-expectancy-who](https://www.kaggle.com/kumarajarshi/life-expectancy-who)](https://www.kaggle.com/kumarajarshi/life-expectancy-who)

**IBM cloud services**

<https://www.youtube.com/watch?v=DBRGlAHdj48&amp;list=PLzpeuWUENMK2PYtasCaKK4bZjaYzhW23L>

**Import the data set into the jupyter notebook**

<https://www.youtube.com/watch?v=Jtej3Y6uUng>