

# **Project Documentation**

Project Report on

**Predicting Life Expectancy using Machine Learning**

Under

**Remote Summer Internship Program 2020 by SmartInternz**

Project by:

**Shubhangi Gupta**

**BTech 3<sup>rd</sup> year (CSE)**

**Vellore Institute of Technology, Bhopal**

**Email: [Shubhangi.gupta2017@vitbhupal.ac.in](mailto:Shubhangi.gupta2017@vitbhupal.ac.in)**

# Index

<b>1. Introduction.....</b>	<b>3</b>
a. Overview.....	3
b. Purpose and Working.....	3
<b>2. Literature Survey.....</b>	<b>4</b>
a. Existing Problem.....	4
b. Proposed Solution.....	5
<b>3. Theoretical Analysis.....</b>	<b>5</b>
a. Block Diagram.....	5
b. Project Requirement	
i)Functional Requirement.....	6
ii) Technical Requirement.....	6
iii) Software Requirement.....	6
<b>4. Flow Chart.....</b>	<b>7</b>
<b>5. Result.....</b>	<b>8</b>
<b>6. Advantages and Disadvantages.....</b>	<b>9</b>
a. Advantages.....	9
b. Disadvantages.....	9
<b>7. Applications.....</b>	<b>9</b>
<b>8. Conclusions.....</b>	<b>10</b>
<b>9. Future Scope.....</b>	<b>10</b>
<b>10. References and Links.....</b>	<b>10</b>

## 1. Introduction:

### 1.1 Overview:

Life expectancy is a statistical measure of the average time a human being is expected to live. This project is about building a model while will consider historical data from a time period of 2000 to 2015 for all the countries. The model trained in factors like Regional variations, Economic Circumstances, Sex Differences, Mental Illnesses, Physical Illnesses, Education, Year of their birth and other demographic factors.

This project will be helpful in predicting the life expectancy of the countrymen so that the preventive measures can be taken accordingly to save them. The project will also prove helpful in predicting the other crucial factors such as the effect and rate of alcohol intake, effects of GDP and etc.

### 1.2 Purpose and working:

The sole purpose of this project is to predict the life expectancy of a person considering the various crucial factors. The project will be helpful in improving the health condition of the country and give insights about some crucial factors such as Alcohol intake, GDP growth, schooling, adult mortality, total and cost expenditure and etc. The project uses a **Random Forest** which is a classification algorithm consisting of many decisions trees. It uses bagging and feature randomness when building each individual tree to try to create an uncorrelated **forest** of trees whose prediction by committee is more accurate than that of any individual tree. The dataset used or the training of the model was downloaded from kaggle.com and Python is used to write the code for machine learning model.

The project is developed using various IBM services mentioned below:-

- 1) **IBM Node-Red:** IBM Node-Red was used to create the UI interface for the machine learning model which will be helpful for the user to input their own values.
- 2) **IBM Watson:** IBM Watson is one of the most popular and useful services provided by the IBM which allows users to create their own ML model along with the help of **IBM Watson Machine Learning**.
- 3) **IBM Auto AI experiment:** It is another unique services provided by IBM which helps us to create the ML model without the use of python and coding.

## 2) Literature survey:

### 2.1) Existing Problem:

Past studies have shown a great deal of research in the estimation of a life expectancy of a Person. After an examination of existing works and techniques in human prediction Life expectancy, I finally concluded that an average can be predicted for individuals who use advanced technologies like Big Data, AI, Machine Learning and devices based on them.

It has been noted that data collection is a big challenge due to considerations relating to privacy and government policy, which will require the collaboration of various health sector bodies. The interworking of a heterogeneous health network is also a statistical analysis challenge. Despite these challenges, Life expectancy can be predicted by proposing a data collection and application approach.

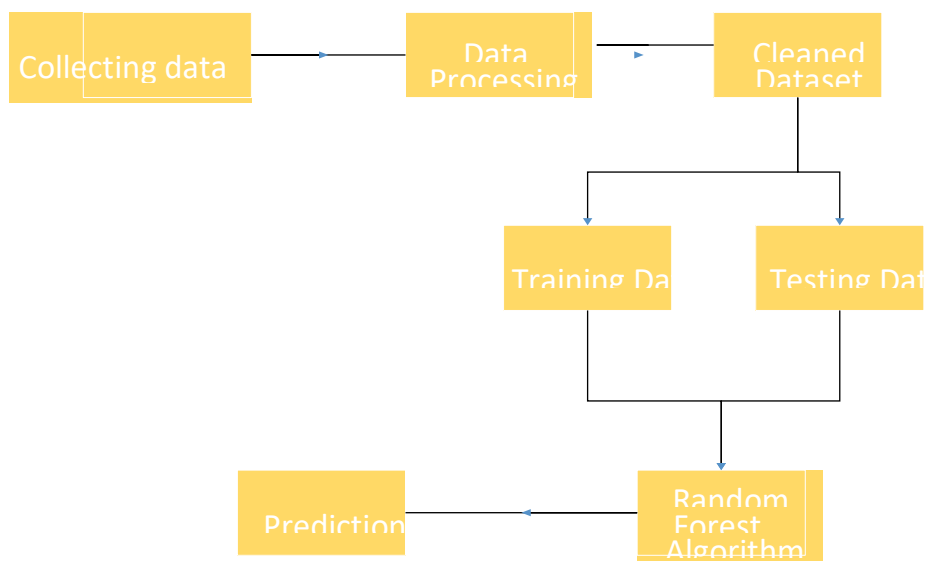
Big data to test the accuracy of PLE prediction and data quality validation Techniques and analysis algorithms with multiple sample groups needs to be developed and tested in real-life situations. As artificial intelligence technology is developing and quickly being implemented, the feasibility of gathering health data from the public as well as current government agencies such as centralized health servers could be increased.

## 2.2) Proposed Solution:

While several studies have been conducted in the past about factors influencing life Expectations, taking into account demographic factors, income distribution and mortality. The effect of the immunization and human development index was found to have not been considered in the past. Some of the past research has also been carried out considering multiple linear regression for all countries, based on a data set of one year. Accordingly, this gives the motivation to resolve the two factors already stated by formulating a Model regression based on model mixed effects and multiple linear regression while considering data for all countries for the period 2000 to 2015. This research will also concentrate on immunization factors, mortality factors, cultural, social and other health-related factors. Since the results in this dataset are focused on different countries, it will be easier for a country to evaluate the predictive factor that leads to lower life expectancy value.

## 3) Theoretical Analysis:

### 3.1) Block Diagram:-



### **3.2) Project Requirements:-**

This project aims primarily to predict life expectancy. The project's basic requirement is the availability of the appropriate dataset which will assist the prediction. So in this project, I have used the standard WHO dataset on Kaggle. The machine learning model is trained based on the provided data, so it could predict an individual's average lifespan in the coming years.

#### **3.2.1) Functional Requirements:**

- Download the dataset of WHO
- Analyze it and clean the dataset
- Create IBM account
- Create the appropriate cloud and node red services
- Train the regression model on different algorithms
- Check for the best one and finalize that algorithm to train our mode
- Build Node red flow for GUI (web app)
- Create scoring end point for integrating our model to node red
- Provide the model with the inputs fields
- The model will return the output as the average predicted lifespan

#### **3.2.2) Technical Requirements:**

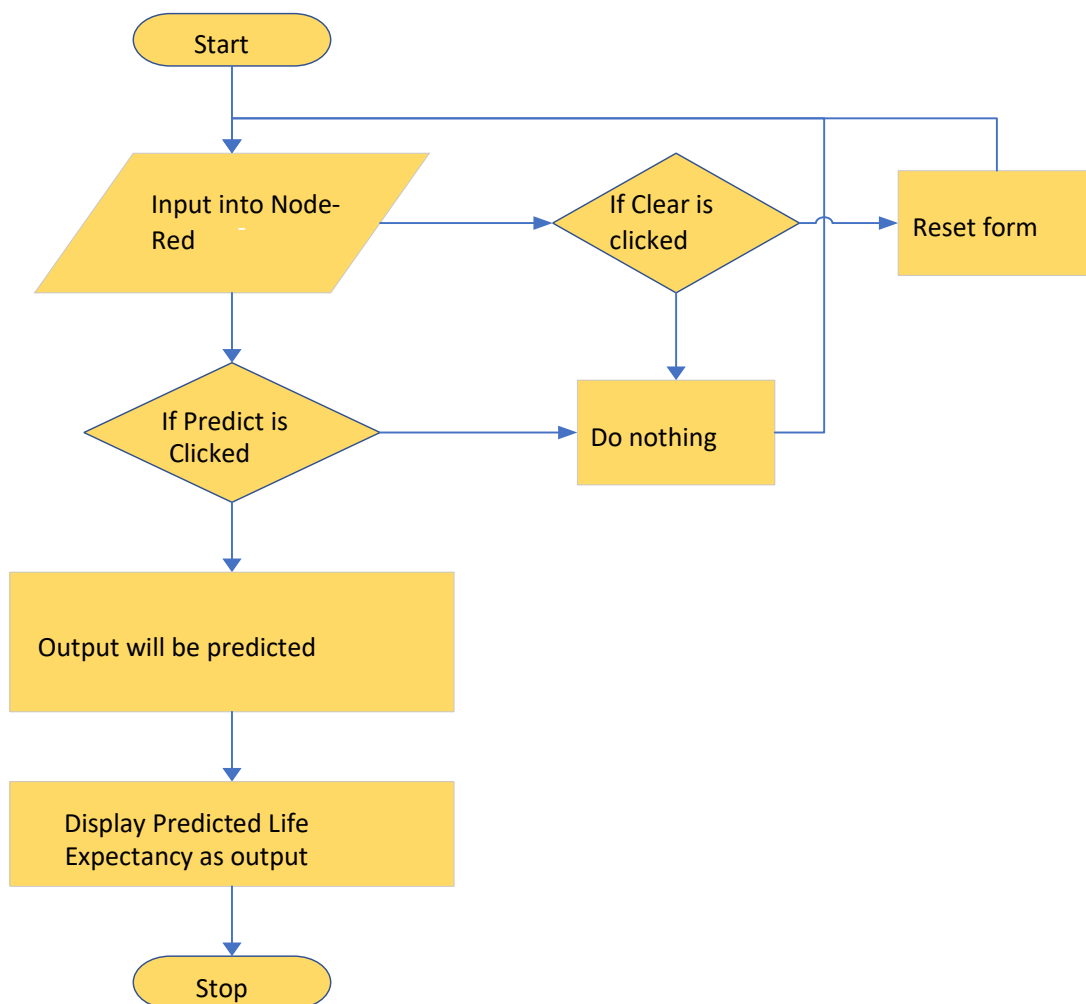
- The GUI must be integrated with the backend trained model.
- The model before training must be given with clean dataset (done by preprocessing)

#### **3.2.3) Software Requirements:**

- IBM cloud services
- IBM Watson services
- IBM Watson Studio
- IBM AutoAI experiment
- IBM Node-Red application

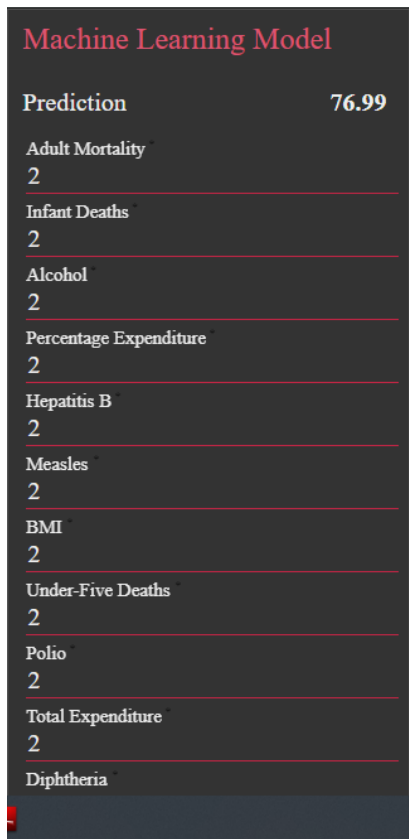
- SmartInternz Project Workspace
- Jupyter Notebook
- Github
- Slack
- Zoho document writer

#### 4) Flow Chart:



## 5. Result:

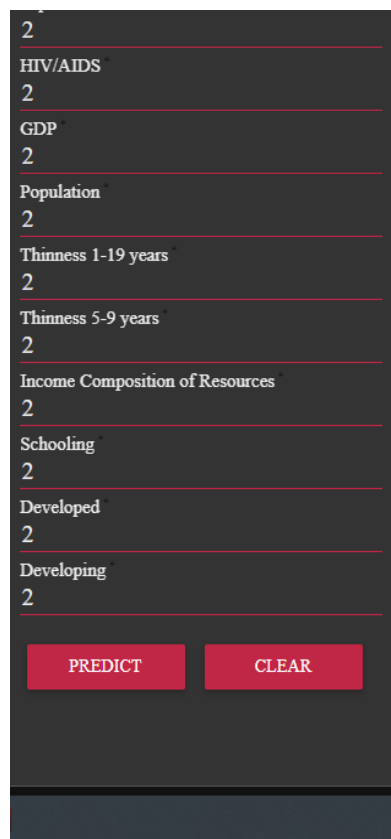
The user friendly Graphical User interface is shown in Figure below This GUI is connected to the trained machine learning model present in the backend (IBM Watson notebook). The user has to fill in the inputs accordingly and click on the “Predict” button present at the end of the form. On clicking the “Predict” button, the user will be displayed the predicted life expectancy at the predict label, based on the inputs provided as shown in Figure .



The screenshot shows a web application titled "Machine Learning Model". It features a list of input fields on the left, each with a numerical value of 2. The input fields are: Adult Mortality, Infant Deaths, Alcohol, Percentage Expenditure, Hepatitis B, Measles, BMI, Under-Five Deaths, Polio, Total Expenditure, and Diphtheria. On the right side, the "Prediction" is displayed as 76.99.

Input	Value
Adult Mortality	2
Infant Deaths	2
Alcohol	2
Percentage Expenditure	2
Hepatitis B	2
Measles	2
BMI	2
Under-Five Deaths	2
Polio	2
Total Expenditure	2
Diphtheria	2

Prediction: 76.99



The screenshot shows the same web application as the previous one, but with the input fields visible. The input fields are: HIV/AIDS, GDP, Population, Thinness 1-19 years, Thinness 5-9 years, Income Composition of Resources, Schooling, Developed, and Developing. Each field contains the value 2. At the bottom, there are two buttons: "PREDICT" and "CLEAR".

Input	Value
HIV/AIDS	2
GDP	2
Population	2
Thinness 1-19 years	2
Thinness 5-9 years	2
Income Composition of Resources	2
Schooling	2
Developed	2
Developing	2

PREDICT CLEAR



## **6. Advantages and disadvantages:**

### **6.1) 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 and abilities by giving best available data
  - Improve and transform customer service
  - Handle enormous quantities of data
  - Sustainable Competitive Advantage
2. Easy for users to interact with the model via the UI.
  3. User-friendly.
  4. Easy to build and deploy.
  5. Doesn't require much storage space.
  6. Data can be analyzed
  7. Factors affecting Life expectancy can be analyzed

### **6.2) 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

## **7. Application:**

Life expectancy is the primary factor in determining an individual's risk factor and the likelihood they will make a claim. This project/idea is useful for Insurance companies as they consider age, lifestyle choices, family medical history, and several other factors when determining premium rates for individual life insurance policies. The principle of

life expectancy suggests that you should purchase a life insurance policy for yourself and your spouse sooner rather than later. Not only will you save money through lower premium costs, but you will also have longer for your policy to accumulate value and become a potentially significant financial resource as you age.

It can be used by researchers to make meaningful research out of it and thus, bring something that will help increase the expectancy considering the impact of a specific factor on the average lifespan of people in a specific country.

## **8. Conclusion:**

Thus, we 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, Alcohol, HIV, Hepatitis B, GDP, Percentage Expenditure and many more. 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.

## **9. 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.

## **10. References and Links:**

- 1) IBM Tutorials - <https://developer.ibm.com/tutorials/https://developer.ibm.com/technologies/machine-learning/series/learning-pathmachine-learning-for-developers/https://scikit-learn.org/stable/modules/preprocessing.html#preprocessing>
- 2) For Dataset - <https://www.kaggle.com/kumarajarshi/life-expectancy-who>

3) Source Code- <https://github.com/SmartPracticeschool/IISPS-INT-1906-Predicting-Life-Expectancy-using-Machine-Learning/upload>

4) Link to access the web-application: <https://node-red-rssvq.eu-gb.mybluemix.net/ui/#!/1?socketid=rtNBfRAQmKGIRJL1AAAN>

