Project Report

PREDICTING LIFE EXPECTANCY USING MACHINE LEARNING

Arpit Shukla
JSSATE, NOIDA
APPLICATION-ID: SPS_APL_20200003844

1.INTRODUCTION

1.1. OVERVIEW

This project is based on Machine Learning in which the goal is to predict the Life Expectancy using historical data. Life Expectancy is a statistical measure of the average time an organism is expected to live, based on the year of its birth, its current age and other demographic factors including gender. The most commonly used measure is life expectancy at birth (LEB). Life Expectancy is depending on various factors like 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 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 health care system and some specific disease related deaths that happened in the country. This project aims to automate this task and provide life expectancy when values for different factors are given.

1.2. SCOPE

This project analyses the provided dataset and creates a model to predict life expectancy. and machine learning can benefit public health researchers with analyzing thousands of variables to obtain data regarding life expectancy.

We can use demographics of selected regional areas and multiple behavioral health disorders across regions to find correlation between individual behavior indicators and behavioral health outcomes.

IBM Watson machine learning and node red are an integral part of analysis. The end product will be a webpage where you need to give all the required inputs and then submit it. Afterwards it will predict the life expectancy value based on regression technique.

1.3. PURPOSE

The purpose of the project is to design a model for predicting Life Expectancy rate of a country given various features 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

2. LITERATURE SURVEY

2.1. EXISTING PROBLEM

The typical regression model that can predict average life expectancy of the country based on some user inputted values such as GDP, BMI, HIV/AIDS, Year, Alcohol intake and etc.

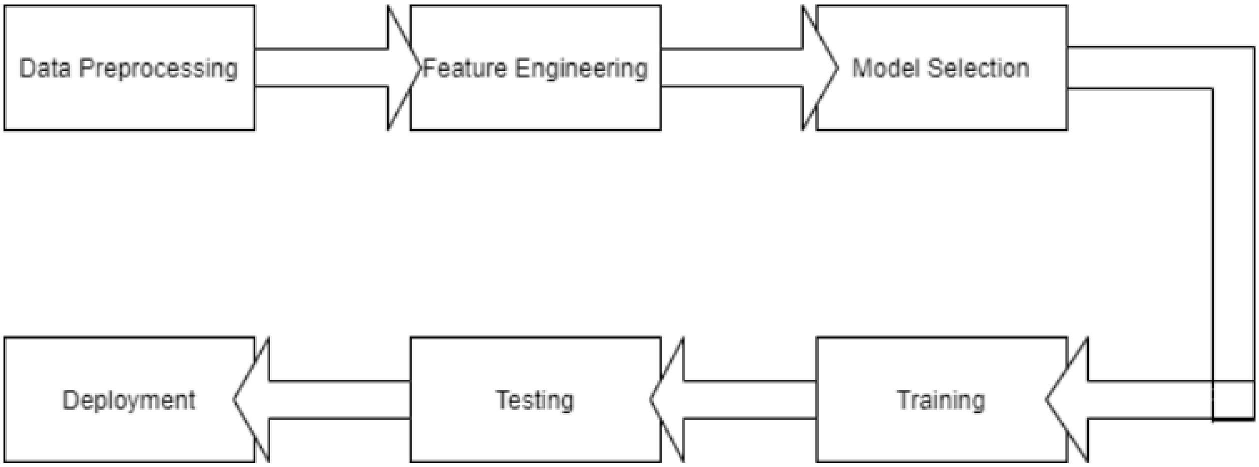
2.2. PROPOSED SYSTEM

Our proposed system makes this whole process of calculating Life Expectancy much easier so any one can calculate the Life Expectancy without any domain knowledge. Our proposed system makes calculation automated and this system has a predicting tool which can predict the Life Expectancy from various attributes value.

So, using machine learning technique we suppose to predict the value of Life Expectancy based on some common attributes like year, GDP, education, alcohol intake of people in the country, expenditure on health care system .We can find this data and get the Life Expectancy value based on their Country and Year.

3.THEORITICAL ANALYSIS

3.1. BLOCK DAIGRAM



3.2. HARDWARE / SOFTWARE DESIGNING

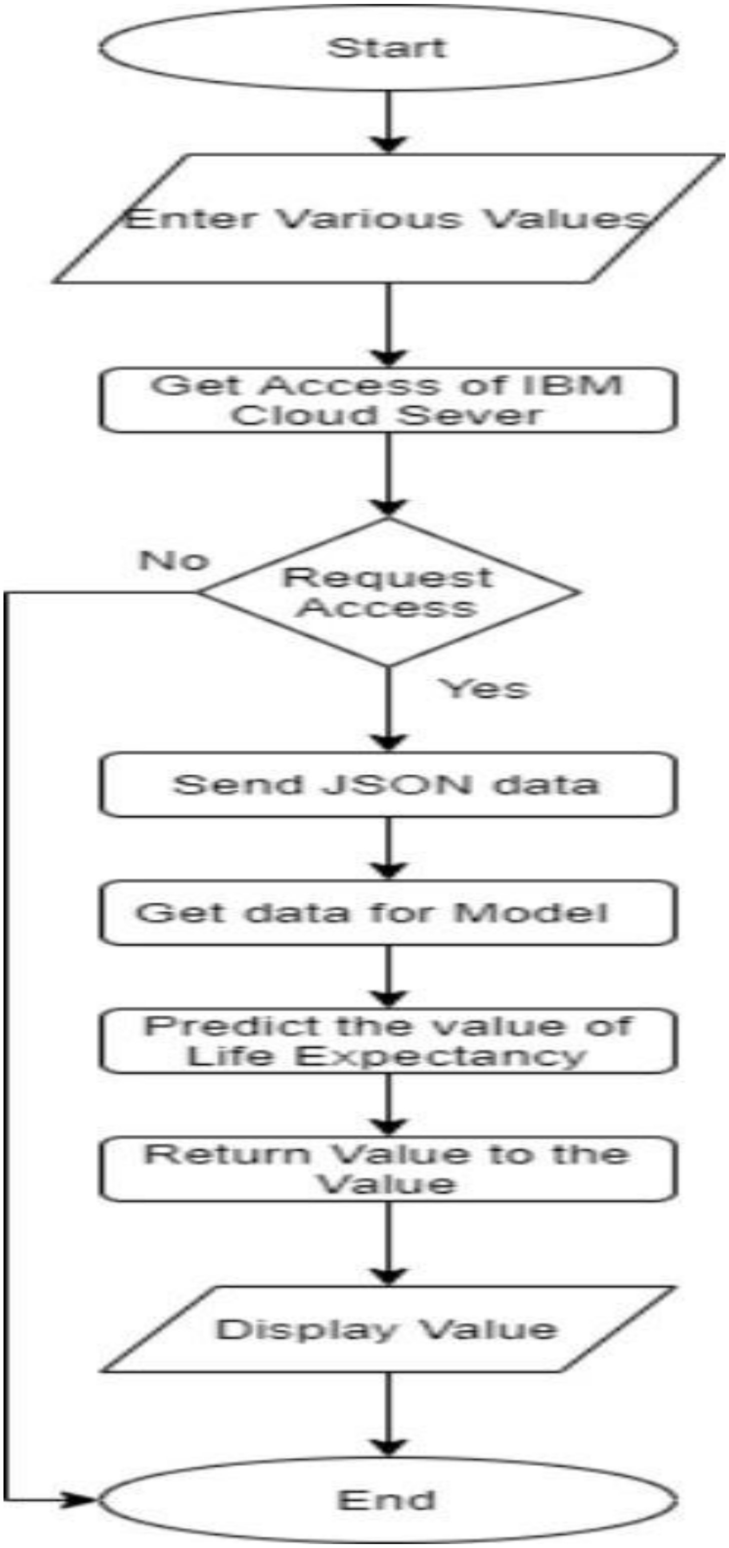
- Project Requirements: Python, IBM Cloud, IBM Watson
- Functional Requirements: IBM cloud
- Technical Requirements: ML, WATSON Studio, Python, Node-Red
- Software Requirements: Watson Studio, Node-Red

3.3. PROJECT DELIVERABLES

- Python notebook containing all the code.
- A node red application which can input data and outputs a prediction for life expectancy.
- A json file containing the architecture of node red project.
- URL of the node red application.

4. EXPERIMENTAL INVESTIGATIONS

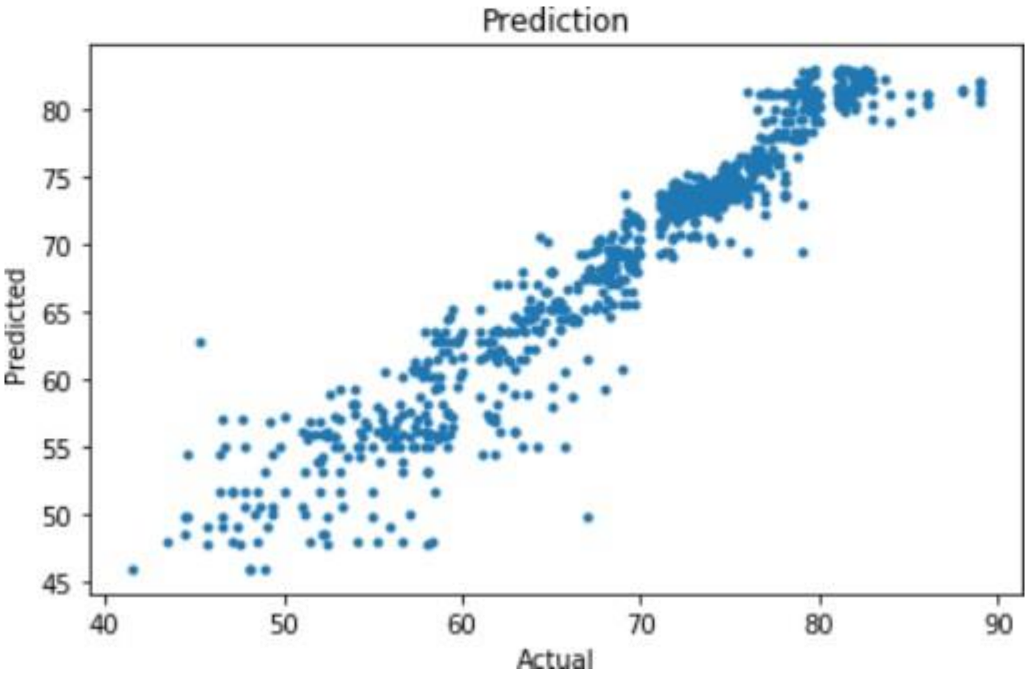
4.1 FLOWCHART



4.2 RESULTS

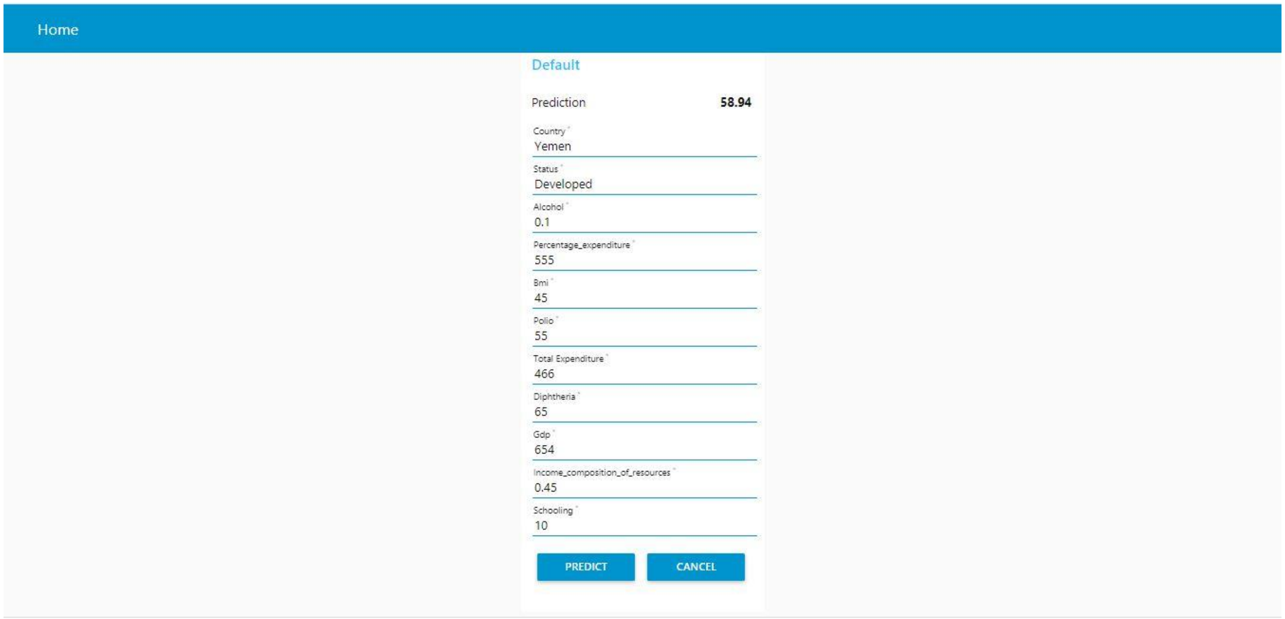
The notebook attached consists of the detailed steps involved in the pipeline. First the data was imported from cloud storage object of IBM Cloud and stored in a data frame. Later it was checked for null values and necessary steps was taken (filling the null values with the mean of the column). Data Visualization was performed on different columns to find the relation between life-expectancy and other variables.
For training phase 20% data was separated for testing purpose. Different pipelines were created for numerical and categorical columns. The model was trained using Random Forest and the M.S.E. came out to be **5.670** and r_2 score of **93.4809**

The plot between actual and predicted values of Life Expectancy is given below.



The plot is roughly a straight line which indicates a linear relationship between predicted and actual values.

User Interface



Default

Prediction 58.94

Country *
Yemen

Status *
Developed

Alcohol *
0.1

Percentage_expenditure *
555

Bmi *
45

Polio *
55

Total Expenditure *
466

Diphtheria *
65

Gdp *
654

Income_composition_of_resources *
0.45

Schooling *
10

5. ADVANTAGES AND DISADVANTAGES ADVANTAGES

a) Health Inequalities: Life expectancy has been used nationally to monitor health inequalities of a country.

b) Reduced Costs: This is a simple webpage and can be accessed by any citizen of a country to calculate life expectancy of their country and doesnot required any kind of payment neither for designing nor for using.

c) User Friendly Interface: This interface requires no background knowledge of how to use it. It's a simple interface and only ask for required values and predict the output.

DISADVANTAGES:

a) Wrong Prediction: As it depends completely on user, so if user provides some wrong values then it will predict wrong value.

b) Average Prediction: The model predicts average or approximate value with 98.04% accuracy but not accurate value.

6. APPLICATION

a) It can be used to monitor health inequalities of a country.

b) It can be used to develop statistics for country development process.

c) It can be used to analyze the factors for high life expectancy.

d) It is user friendly and can be used by anyone.

7. FUTURE SCOPE

- Integrating a data science dashboard which shows different visualizations of Life Expectancy as per the Country and Year.
- A system to update our model parameters when there is a change in consistency of data like when a sudden epidemic occurs or during a recession, then all the attributes used in our model need to be updated to provide most precise life expectancy.

8.Conclusion

The advantages of longer life span outweigh its disadvantages. The benefits people and the world can get from a higher life expectancy are irreplaceable and undeniable. It is a truth that life expectancy is a symbol of civilization and better life.

Knowing an estimate of how much life we have left pushes us to achieve different things. Higher life expectation is also perceived as greater quality of life and greater income of society.

Our project has automated the entire task of rigorous calculation and removed errors in the existing system and gives the life expectancy to the user. This information can be useful to the society as stated above and this method is also much cheaper than hiring people to do the calculations.

9. BIBLIOGRAPHY

- https://cloud.ibm.com/docs/overview?topic=overview-whatis-platform
- https://developer.ibm.com/tutorials/how-to-create-a-node-red-starter-application
- https://nodered.org/
- https://github.com/watson-developer-cloud/node-red-labs
- https://www.youtube.com/embed/r7E1TJ1HtM0
- https://bookdown.org/caoying4work/watsonstudio-workshop/jn.html
- https://www.kaggle.com/kumarajarshi/life-expectancy-who
- https://www.youtube.com/watch?v=DBRGlAHdj48&list=PLzpeuWUENMK2PYtasCaKK4b
- https://www.youtube.com/watch?v=Jtej3Y6uUng
- https://bookdown.org/caoying4work/watsonstudio-workshop/jn.html#deploy-model-as-webservice
- https://machinelearningmastery.com/columntransformer-for-numerical-and-categorical-data/