PROJECT DOCUMENTATION

PROJECT REPORT ON

Predicting Life Expectancy using Machine Learning

Under

Remote Summer Internship Program 2020 by SmartInternz

Project by :

SARVESH SHASHIDHAR

B.E 3rd Year (ECE) Rashtriya Vidyalaya College of Engineering

Email : sarveshs.ec17@rvce.edu.in

1. INTRODUCTION

a. Overview

The project involved the creation of a Machine Learning Model or Instance to predict the number of years an individual is expected to live based on various factors such as: -

- Country
- Current Year
- Country Economical Status
- The Adult Mortality Rates
- The number of Infant Deaths
- BMI
- Total income and expenditure of the person
- GDP of the nation
- Chance of incurring diseases like Measles, Hepatitis B, HIV/AIDS etc.
- Population of the Nation
- Thinness of the population
- Schooling and Income of the masses

The prediction of the Life Expectancy can be done using basic Regression methods and can provide fairly accurate insights into the average lifespan of the people.

b. Purpose

The purpose of the project is to provide a more clear and consice picture about the average life span of individuals in the world. It has previous historical data inputs that considers all the above mentioned data fields and then helps a person to understand fairly accurately about how long he is expected to live. The model will also tell what all factors contribute greatly to the Life Expectancy and then help the various health organisations take relevant decisions so as to improve the health conditions all across the globe.

2. LITERATURE SURVEY

a. Existing Problem

The existing problem is that the process of data collection is not very reliable and easy. The first challene that will be faced is the collection of the medical records of the various individuals to get the required information for the initial dataset. The second challenge that most probably will occur is that maybe not all data values are present in the dataset and even the recorded values may be wrong and not accurate. One last major problem that exists is that the data set fails to cover other factors such as genetic diseases, other major diseases like Malaria, Cholera etc. and the Age/Gender of the said individual.

b. Proposed Solution

To solve the first problem, there needs to be involvement of the authoritative bodies in support of the studies and data collection to ensure that the correct and required data is provided. The second problem can be solved before the model training in a step called the **Data Pre – processing** where the raw data set is taken and then converted into a clean and more reliable dataset. This step is necessary before engaging in the model training process. There are a few standard steps that are followed in the Data Pre-Processing step: -

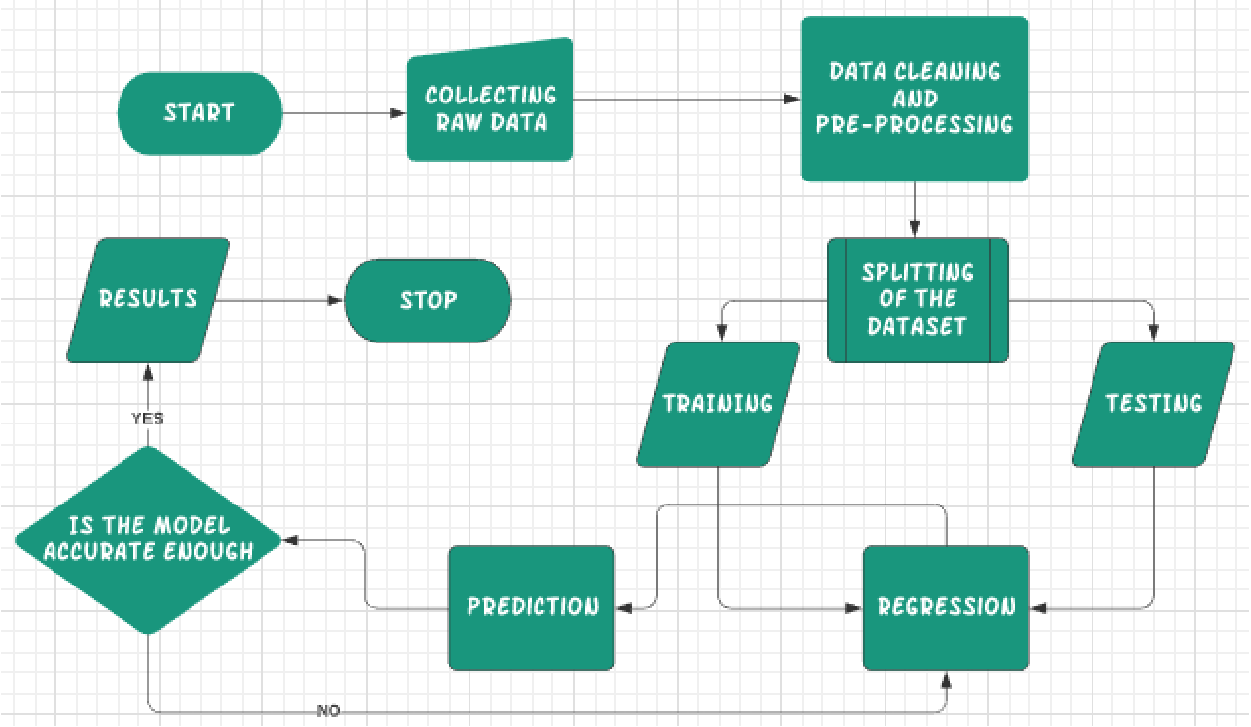
- Detecting and Imputing NULL values in the dataset
- Ensuring the values of all the features are in one value range so that they might be comparable.
- Encoding the string values to numeric values so that they can be analysed.

There is no full proof solution for the third problem. So as to get a more accurate and acceptable system for predicting the life expectancy of the population, one can actually perform a more intensive study and collect more data.

3. THEORETICAL ANALYSIS

a. Block Diagram

The block diagram of the Machine Learning Model is shown below: -



In the model we can see that our process begins with Collecting or Importing the Dataset. Since this is the raw format of the data, we apply data cleaning and pre-processing methods on it to convert the data into a more reliable format. After cleaning the data, we split the data into two parts – Training Data and Testing Data. The training data is used to train the model to notice the relations between the Life Expectancy and the various factors. This training will be based on Regression Algorithms as we need to predict the output data given the input fields. Finally, we would be testing the model on the test dataset to find out what is the accuracy. If the model is sufficiently accurate, then we would tabulate the results else we would make changes to the Regression model to meet the accuracy requirements.

b. Hardware and Software Designing

The Machine Learning Model is not a processor heavy action and therefore, doesn’t need highly powerful and sophisticated systems. The model can be run on Windows 7 or above with a 1.6GHz processor or above. If the system satisfies these conditions, then the Jupyter Notebook and the Node RED application can run smoothly on the system.

When it comes to Software Designing and requirements, then we need some particular softwares as listed below: -

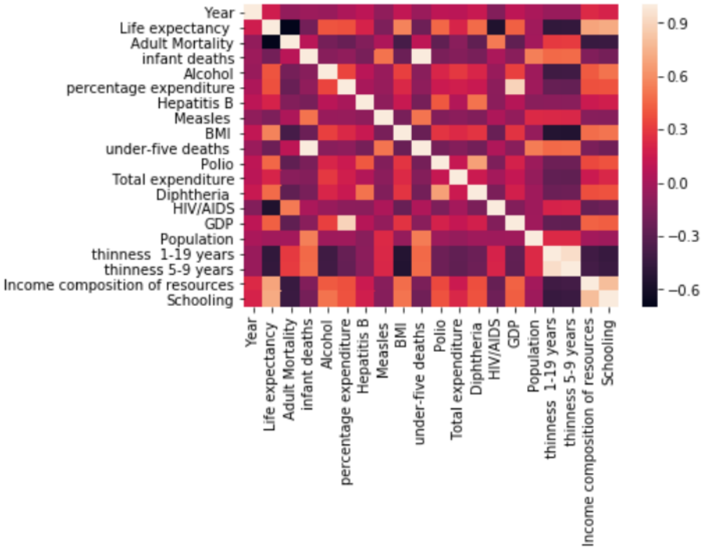
- **IBM Cloud Services –** To store and access the code files and the dataset.
- **IBM Watson Services –** To create a machine learning instance and to access the Jupyter

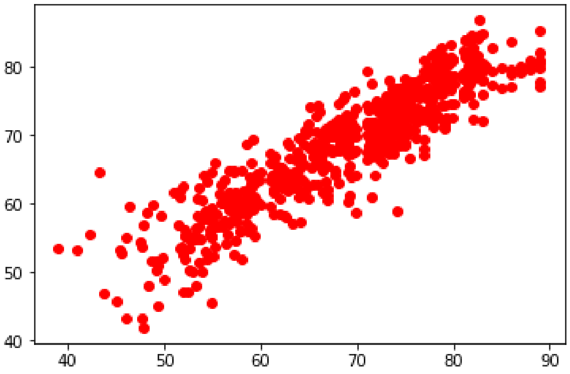
Notebook

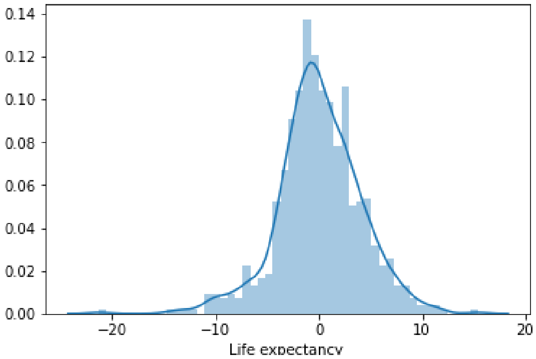
- **Jupyter Notebook –** To write the Python based Regression Codes and to perform the training and testing of the model
- **IBM Node RED Application –** To create an intuitive User Interface so that there is no need to modify the code for testing out various test cases.
- **Zoho Write –** To write and compile the documentation work and reports
- **Git Repository –** To compile and store all the files related to the Machine Learning Model.

4. EXPERIMENTAL INVESTIGATIONS

The process of Experimental Investigation involves creating a fair test for all the variables to get a closer look at how the various variables are related to each other. The Experimental Investigations were performed and the results are displayed below: -



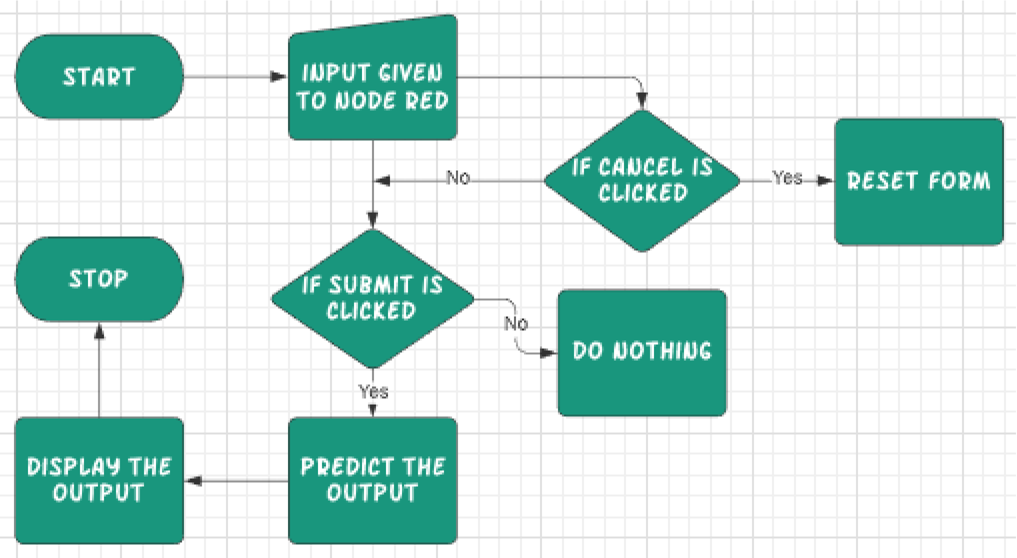




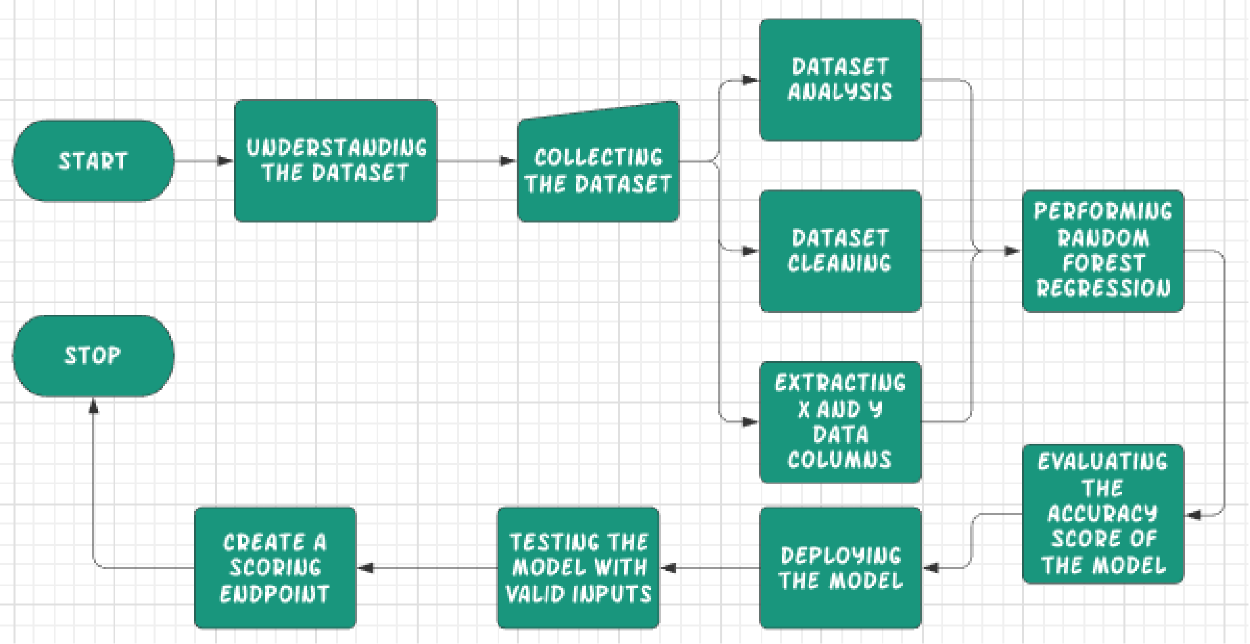
5. FLOW CHART

The Flow chart of the entire model is divided into two parts. One flow chart shows the logic flow when the user tries to use the Node RED UI while the second flow chart shows the back – end logic flow that happens in the Jupyter Notebook.

Node RED UI Logic Flow

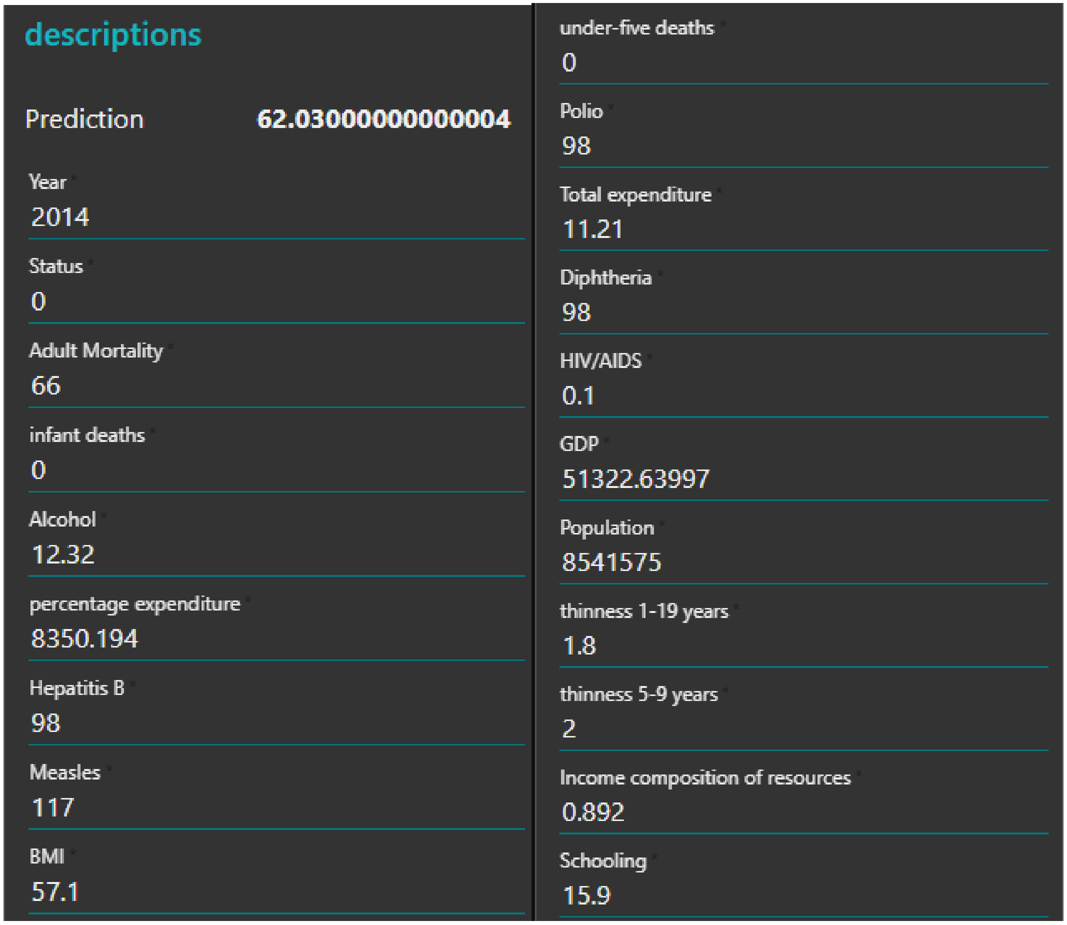


Jupyter Notebook Back End Flow


START
UNDERSTANDING THE DATASET
COLLECTING THE DATASET
DATASET ANALYSIS
DATASET CLEANING
EXTRACTING X AND Y DATA COLUMNS
PERFORMING RANDOM FOREST REGRESSION
EVALUATING THE ACCURACY SCORE OF THE MODEL
DEPLOYING THE MODEL
TESTING THE MODEL WITH VALID INPUTS
CREATE A SCORING ENDPOINT
STOP

6. RESULT

Based on the given data, the model aims at predicting the expected life span of the person. The model analyses previous data and then uses the same to predict the life expectancy of a new data case. The Node RED User Interface output is shown below: -



descriptions

Prediction 62.03000000000004

Year *
2014

Status *
0

Adult Mortality *
66

infant deaths *
0

Alcohol *
12.32

percentage expenditure *
8350.194

Hepatitis B *
98

Measles *
117

BMI *
57.1

under-five deaths *
0

Polio *
98

Total expenditure *
11.21

Diphtheria *
98

HIV/AIDS *
0.1

GDP *
51322.63997

Population *
8541575

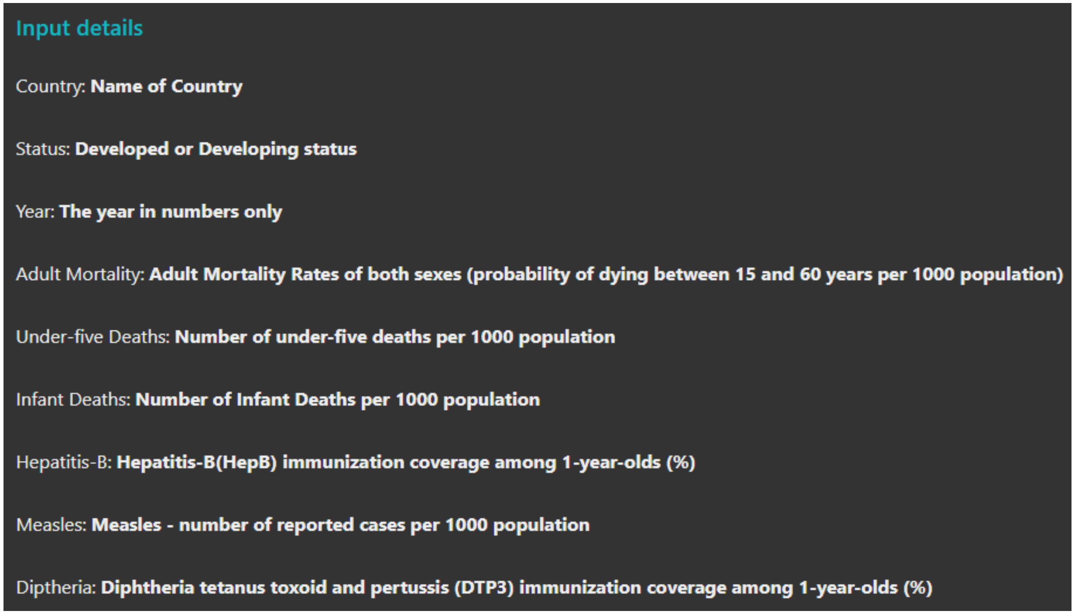
thinness 1-19 years *
1.8

thinness 5-9 years *
2

Income composition of resources *
0.892

Schooling *
15.9

The UI also displays the information about the various fields as shown below: -



Input details

Country: **Name of Country**

Status: **Developed or Developing status**

Year: **The year in numbers only**

Adult Mortality: **Adult Mortality Rates of both sexes (probability of dying between 15 and 60 years per 1000 population)**

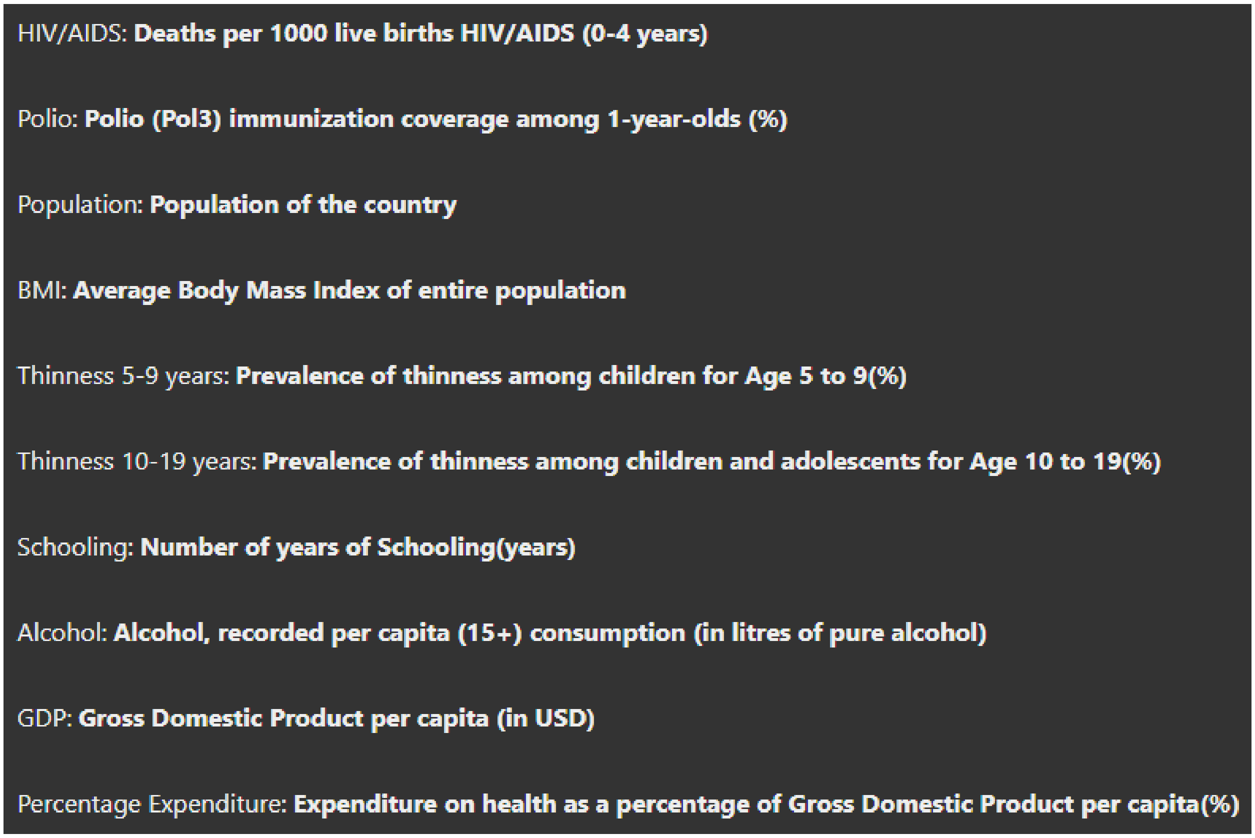
Under-five Deaths: **Number of under-five deaths per 1000 population**

Infant Deaths: **Number of Infant Deaths per 1000 population**

Hepatitis-B: **Hepatitis-B(HepB) immunization coverage among 1-year-olds (%)**

Measles: **Measles - number of reported cases per 1000 population**

Diptheria: **Diphtheria tetanus toxoid and pertussis (DTP3) immunization coverage among 1-year-olds (%)**



HIV/AIDS: **Deaths per 1000 live births HIV/AIDS (0-4 years)**

Polio: **Polio (Pol3) immunization coverage among 1-year-olds (%)**

Population: **Population of the country**

BMI: **Average Body Mass Index of entire population**

Thinness 5-9 years: **Prevalence of thinness among children for Age 5 to 9(%)**

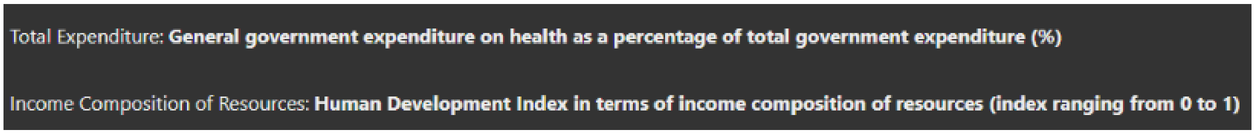
Thinness 10-19 years: **Prevalence of thinness among children and adolescents for Age 10 to 19(%)**

Schooling: **Number of years of Schooling(years)**

Alcohol: **Alcohol, recorded per capita (15+) consumption (in litres of pure alcohol)**

GDP: **Gross Domestic Product per capita (in USD)**

Percentage Expenditure: **Expenditure on health as a percentage of Gross Domestic Product per capita(%)**



Total Expenditure: **General government expenditure on health as a percentage of total government expenditure (%)**

Income Composition of Resources: **Human Development Index in terms of income composition of resources (index ranging from 0 to 1)**

Before creating a Node RED UI application and before deploying the model, we needed to check the accuracy of the said model. This was done in the Jupyter Notebook: -

ACCURACY METRIC	VALUES
Mean Absolute Error	1.3089
Mean Squared Error	4.2364
Root Mean Squared Error	2.0582
R^2 Score	0.9548

This shows that the model is accurate for 95.48% of the test cases.

7. ADVANTAGES AND DISADVANTAGES

a. Advantages

- The factors that affect the life expectancy of the person are clearly analysed
- The model is malleable and can accommodate more factors without major changes in the code
- The high accuracy ensures a fairly correct prediction.

b. Disadvantages

- The model is not 100% accurate and may deliver wrong values sometimes
- The model doesn't include a lot of factors that can determine the life expectancy.
- If the dataset has errors, the model also will have errors and will deliver wrong output.

8. APPLICATIONS

The Machine Learning Model has a lot of real – life applications: -

- The model can be used by people to predict their expected life span and then plan their life accordingly
- The model has various factors that affect life expectancy and the health organisations can check the dependency of the factors on the life expectancy.
- The model is portable and can accommodate many other factors that can help further analysis of life expectancies.

9. CONCLUSION

In conclusion, we can see that a Machine Learning Instance was created and trained with the help of Jupyter Notebook and IBM Cloud Services. The Model was trained with previously collected dataset about the life expectancies of people based on various factors such as country of origin, BMI, Alcoholism rates, Year of Birth, Adult Mortality and many more. The model built was of a regression type and is capable of predicting the life expectancy of new entried based on the training it has performed on the previously known data. The model has been integrated with the simple Node RED started Application to create an interactive User Interface and make the model more intuitive. With an accuracy score of 95-96%, the model is fairly accurate in predicting the Life Expectancy.

10. FUTURE SCOPE

The next step in this project includes the addition of other parameters and data fields that can determine the Life Expectancy of an average person. Some of these fields can be: -

- Obesity
- Standard of Living
- Medical Accessibility
- Hospital Facilities

- Basic Necessities
- Other diseases like Malaria, Cholera etc.
- Population Density
- Pest problems

Once we have reached a saturation in the number of data fields, the model can be modified to perform classification as well about people whether they are healthy or are un-healthy.

This model can also include the data for plants and other organisms to determine the Life Expectancy of the living being in question.

APPENDIX

A. **Source Code: -** https://github.com/SmartPracticeschool/IISPS-INT-3297-Predicting-Life-Expectancy-using-Machine-Learning/blob/master/Life_expectancy_prediction.ipynb
B. **Dataset Link:** - https://www.kaggle.com/kumarajarshi/life-expectancy-who
C. **API/UI Webpage: -** https://node-red-ueohr.mybluemix.net/red/#flow/ceba38bf.7bc2c8
D. **Project Demonstration Video Link: -** https://drive.google.com/file/d/1RPfKiOXTiO8pY41hq5UXzLG7UBIxYFJi/view?usp=sharing