Project Documentation

Project Report on

Predicting Life Expectancy using Machine Learning

Under

Remote Summer Internship Program 2020 by SmartInternz

Project by :

G. Hemant Kumar

B.E. 3rd year (Computer) Datta Meghe College Of Engineering, Airoli, Navi Mumbai

Email: jyothiniranjanburla@gmail.com

1. INTRODUCTION:

- **Overview:**

  Life expectancy is the statistical measure of the average time a human being is expected to live. Life expectancy depends on the 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 given.

- **Purpose:**

  Life expectancy is one of the most important factors in end-of-life decision making. Good prognostication for example helps to determine the course of treatment and helps to anticipate the procurement of health care services and facilities, or more broadly: facilitates Advance Care Planning. Advance Care Planning improves the quality of the final phase of life by stimulating doctors to explore the preferences for end-of-life care with their patients, and people close to the patients. Physicians, however, tend to overestimate life expectancy, and miss the window of opportunity to initiate Advance Care Planning

The project uses a **RandomForestRegressor** algorithm.It isa measure of the relation between the mean value of one variable (e.g. output) and corresponding values of other variables . 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.

2.LITERATURE SURVEY:

- **Existing problem:**

  It has been noted that data collection for predicting the life/health using the machine learning/big data is a big challenge due to considerations relating to privacy and government policy, which will require the collaboration of various health sector bodies. Despite these challenges, Life expectancy can be predicted by proposing a data collection and application approach. As Artificial intelligence and Machine Learning technologies are developing and quickly being implemented, the ease of gathering health data from the public as well as current government agencies such as centralized health servers could be increased

- **Proposed Solution:**

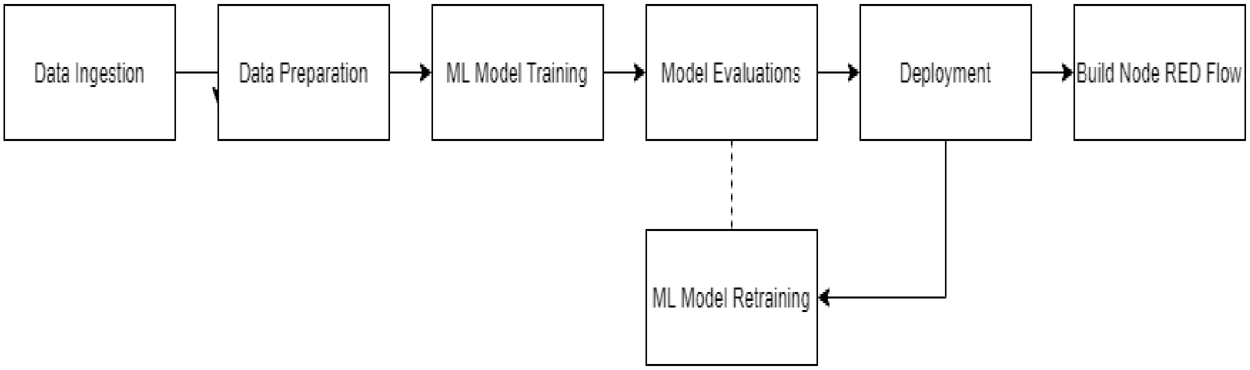
I explored life expectancy and looked for dataset on the following aspects (features) : "Country", "Year", "Status", "Adult Mortality", "infant deaths", "Alcohol", "percentage expenditure", "Hepatitis B", "Measles ", " BMI ", "under-five deaths ", "Polio", "Total expenditure", "Diphtheria ", " HIV/AIDS", "GDP", "Population", " thinness 1-19 years", " thinness 5-9 years", "Income composition of resources", "Schooling"

In summation, the dataset started with 21 unclean variables (including the target) and has been pared down to 12 features to describe the target variable (Life Expectancy). This is very likely only the beginning of the possible things that could be done with this dataset, but nonetheless it serves as a solid foundation for modeling.

After performing data cleaning and data analysis using the statistical tools in python(R) and selected the dependent and independent features and created a machine learning model using **RandomForestRegressor** using that model when we give the inputs(features) the model will give prediction (life expectancy in years) as output. and finally that model is deployed to IBM cloud and madeso that it would be useful for all the people.

3. THEORETICAL ANALYSIS:

- **Block Diagram:**



- **Hardware/Software Requirements:**

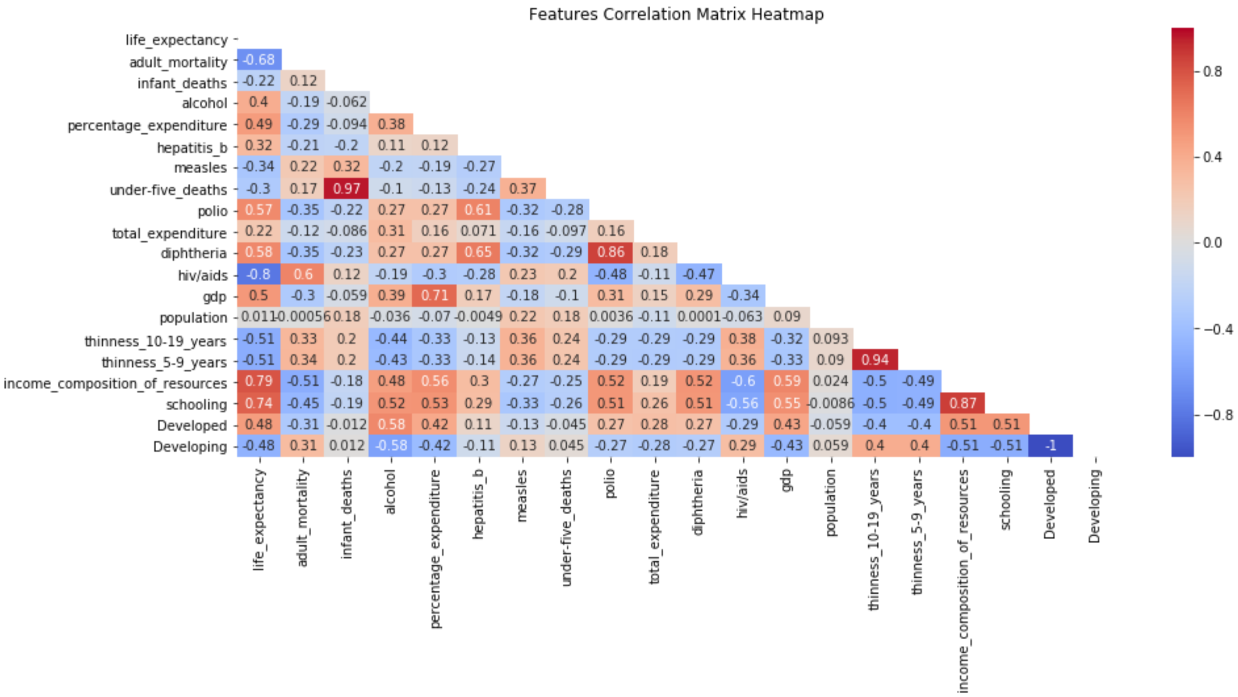
Operating system: Windows XP/Windows7/Windows 10

IBM Cloud Services which includes
Watson Studio
Machine learning
Node RED
Cloud Foundary Service

4. EXPERIMENTAL INVESTIGATIONS:

With Python:

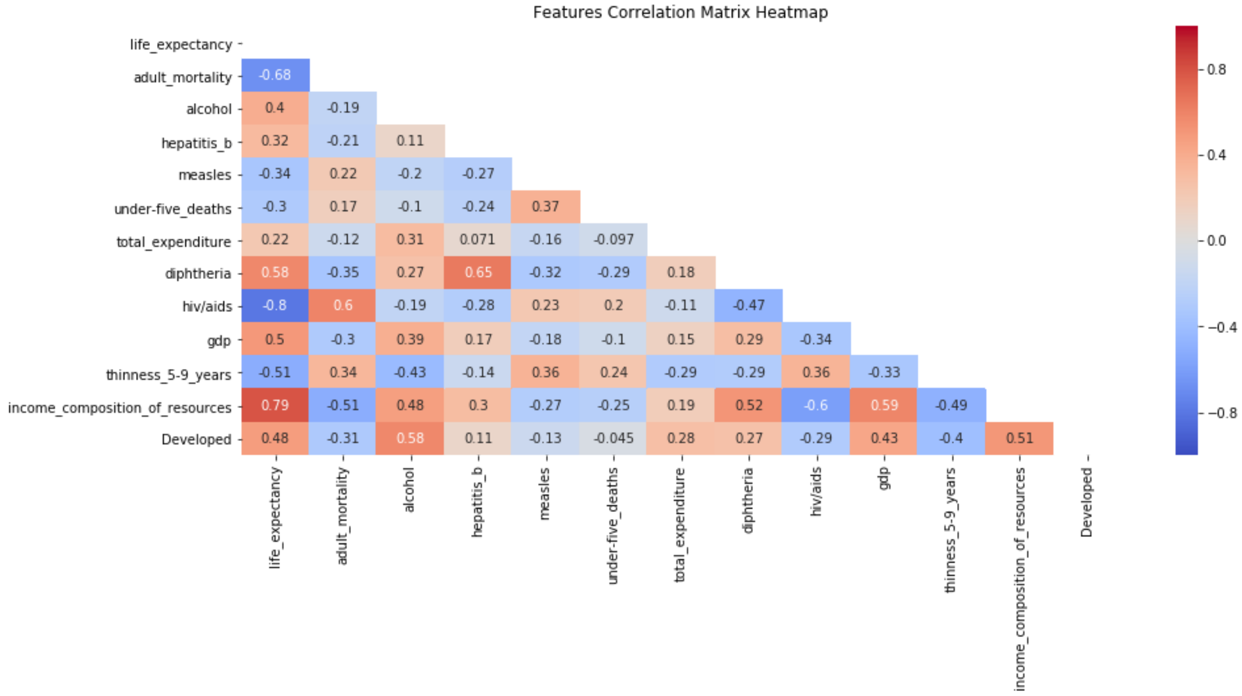
These are some graphs from the refined data analysis, that makes us understand collinearity.



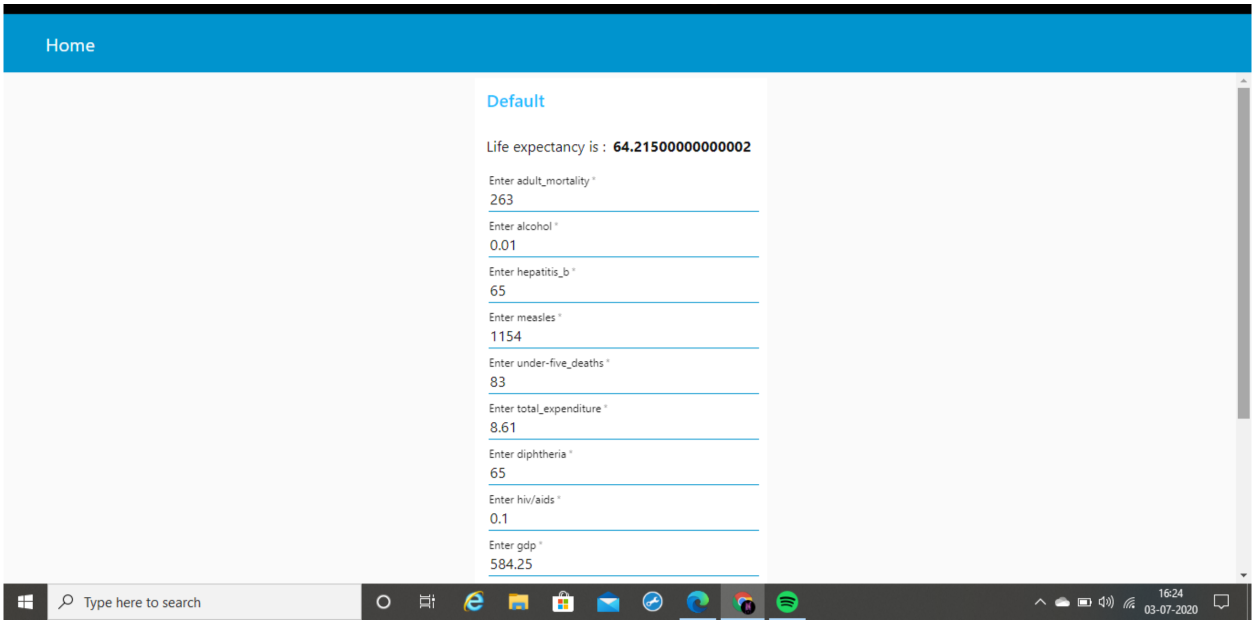
The following are very/extremely highly correlated (correlation > .7 or correlation < -.7):

- Infant Deaths/Under Five Deaths (drop Infant Deaths - Under Five Deaths is more highly correlated to Life Expectancy)
- GDP/Percentage Expenditure (drop Percentage Expenditure - GDP is more higher correlated to Life Expectancy)

- Polio/Diphtheria (drop Polio - Diphtheria is more highly correlated to Life Expectancy)
- Thinness 5-9/Thinness 10-19 (drop Thinness 10-19 as correlations to other variables are slightly higher)
- Income Composition of Resources/Schooling (drop Schooling - Income Composition of Resources is more highly correlated with Life Expectancy)
- Developing/Developed (drop Developing - these two are the same just opposite of one another)



After giving inputs to the deployed machine learning model output is given below:



Default

Life expectancy is : **64.21500000000002**

Enter adult_mortality *
263

Enter alcohol *
0.01

Enter hepatitis_b *
65

Enter measles *
1154

Enter under-five_deaths *
83

Enter total_expenditure *
8.61

Enter diphtheria *
65

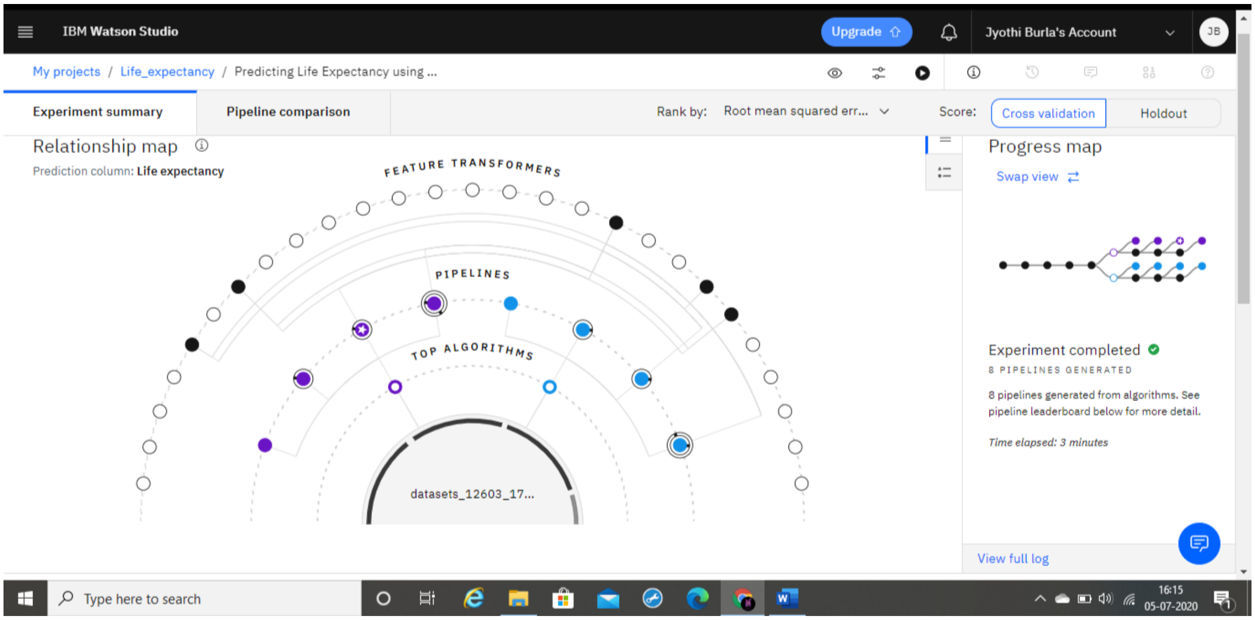
Enter hiv/aids *
0.1

Enter gdp *
584.25

Without Python(AutoAI):

Using the AutoAI graphical tool in Watson Studio will quickly build a model and evaluate its accuracy, all without writing a single line of code. AutoAI guides us, step by step, through building a machine learning model by uploading training data, choosing a machine learning technique and algorithms, and training and evaluating the model.

Relationship Map:



Rank by: Root mean squared err...

Score: Cross validation

Relationship map

Prediction column: **Life expectancy**

Progress map

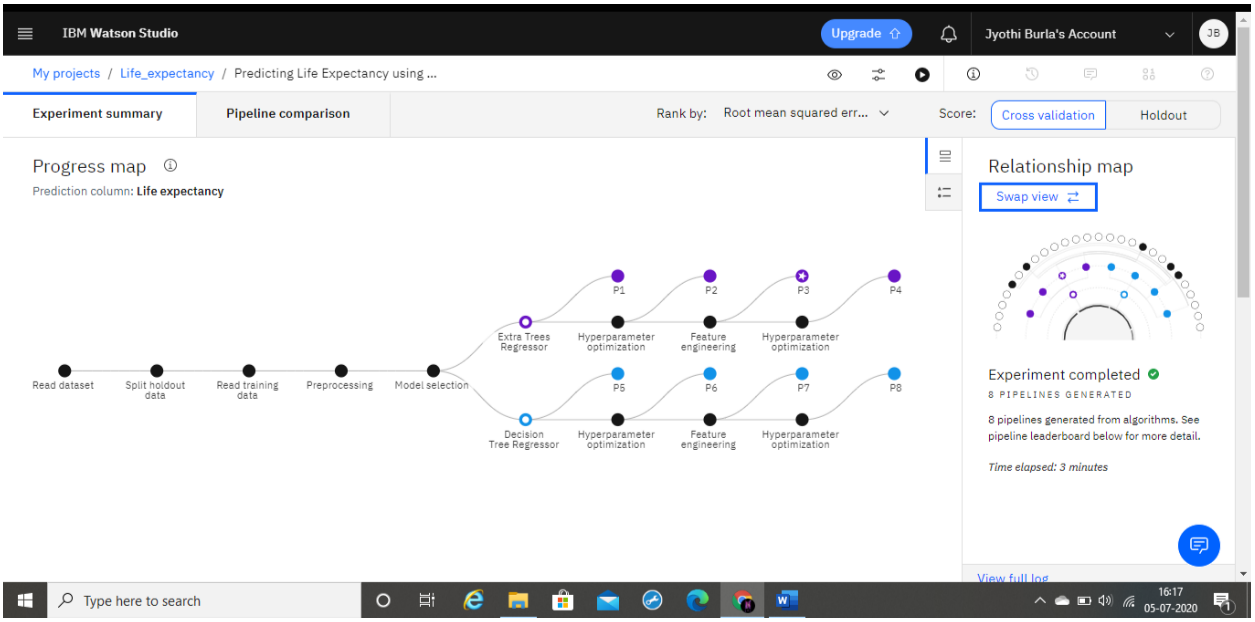
Experiment completed

8 PIPELINES GENERATED

8 pipelines generated from algorithms. See pipeline leaderboard below for more detail.

Time elapsed: 3 minutes

Progress Map:



Rank by: Root mean squared err...

Score: Cross validation

Progress map

Prediction column: **Life expectancy**

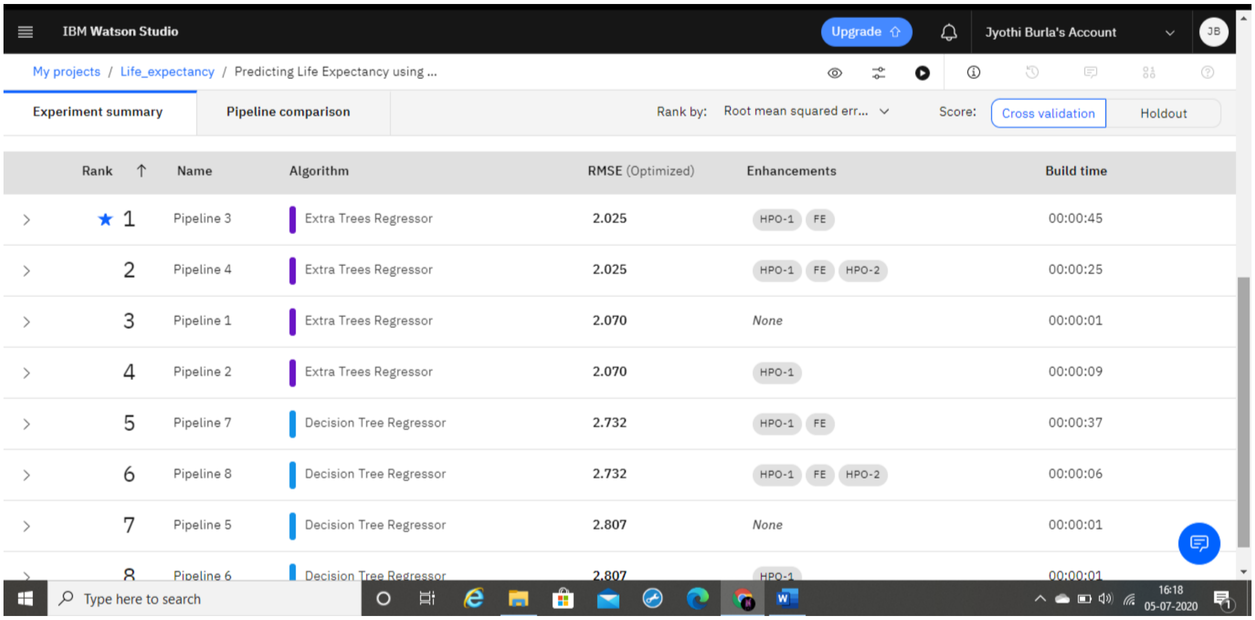
Relationship map

Experiment completed

8 PIPELINES GENERATED

8 pipelines generated from algorithms. See pipeline leaderboard below for more detail.

Time elapsed: 3 minutes

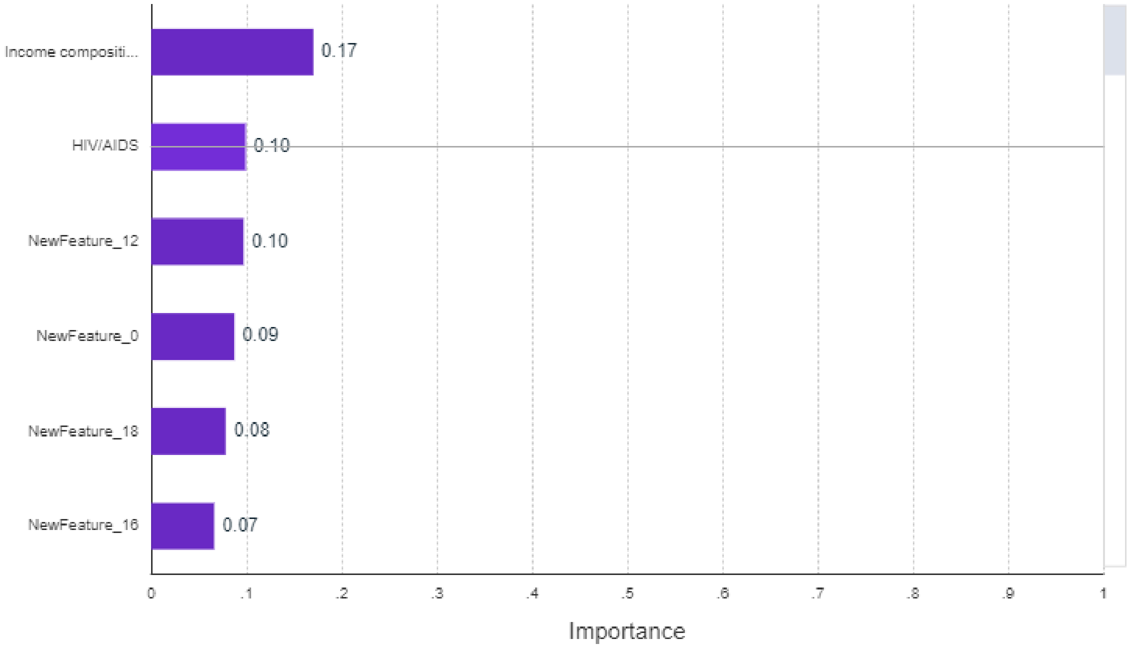


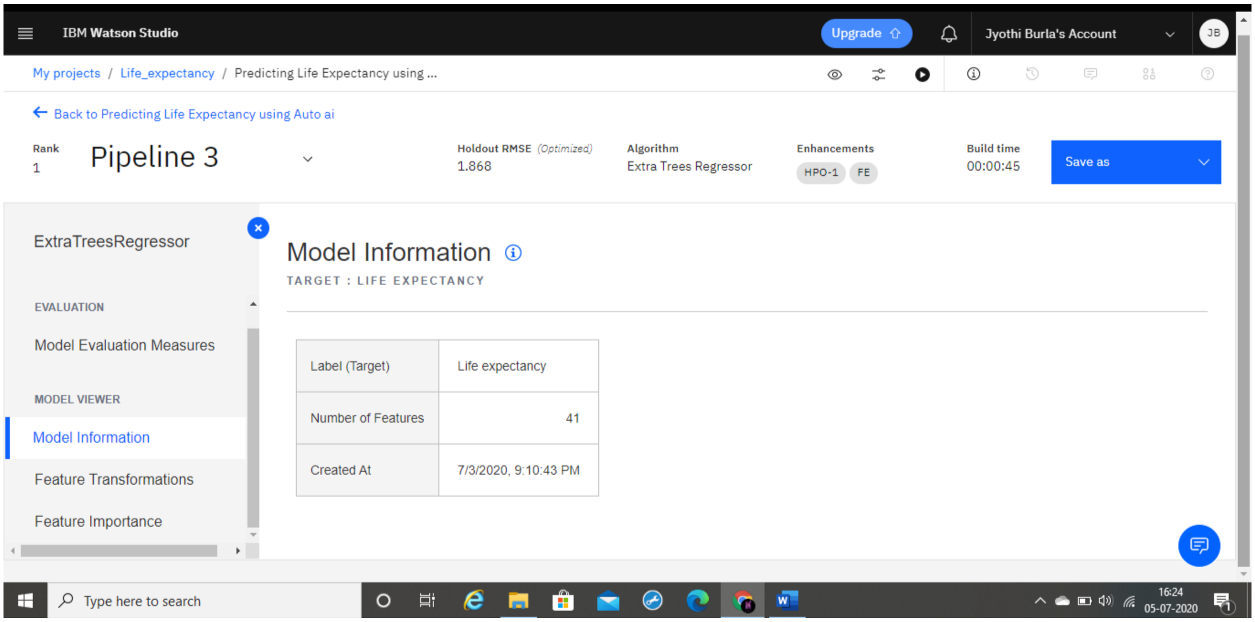
Rank by: Root mean squared err...

Score: Cross validation

Rank	Name	Algorithm	RMSE (Optimized)	Enhancements	Build time
★ 1	Pipeline 3	Extra Trees Regressor	2.025	HPO-1 FE	00:00:45
2	Pipeline 4	Extra Trees Regressor	2.025	HPO-1 FE HPO-2	00:00:25
3	Pipeline 1	Extra Trees Regressor	2.070	*None*	00:00:01
4	Pipeline 2	Extra Trees Regressor	2.070	HPO-1	00:00:09
5	Pipeline 7	Decision Tree Regressor	2.732	HPO-1 FE	00:00:37
6	Pipeline 8	Decision Tree Regressor	2.732	HPO-1 FE HPO-2	00:00:06
7	Pipeline 5	Decision Tree Regressor	2.807	*None*	00:00:01
8	Pipeline 6	Decision Tree Regressor	2.807	HPO-1	00:00:01

Feature Importance:





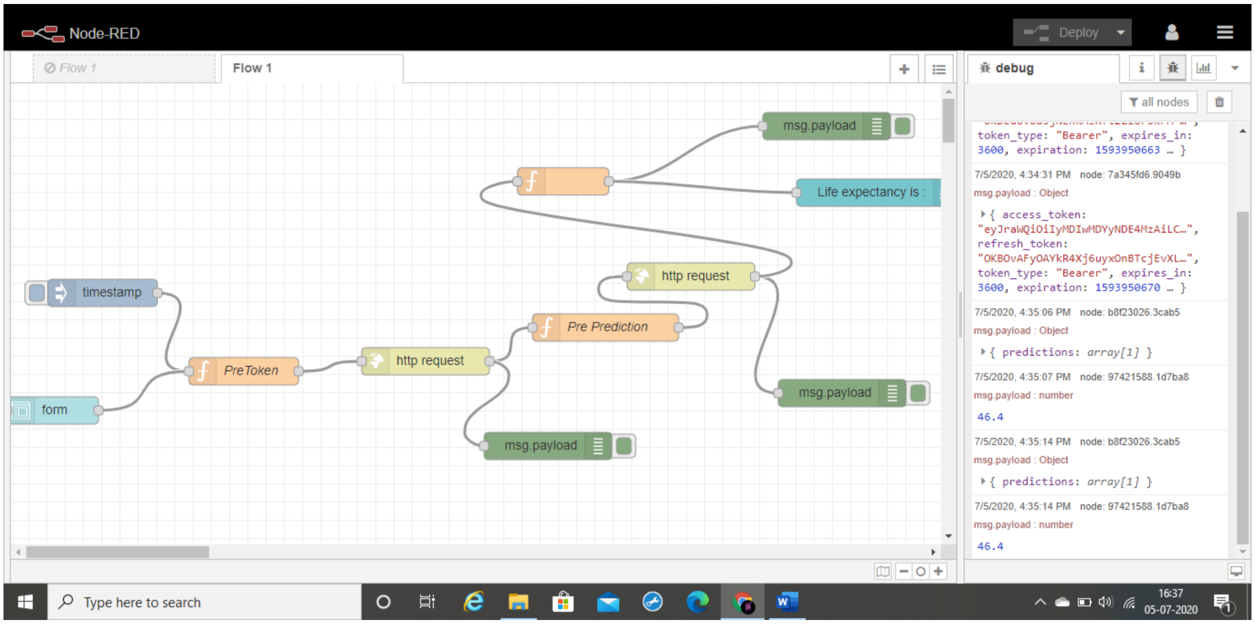
Rank		Holdout RMSE (Optimized)	Algorithm	Enhancements	Build time
1	Pipeline 3	1.868	Extra Trees Regressor	HPO-1 FE	00:00:45

Model Information

TARGET : LIFE EXPECTANCY

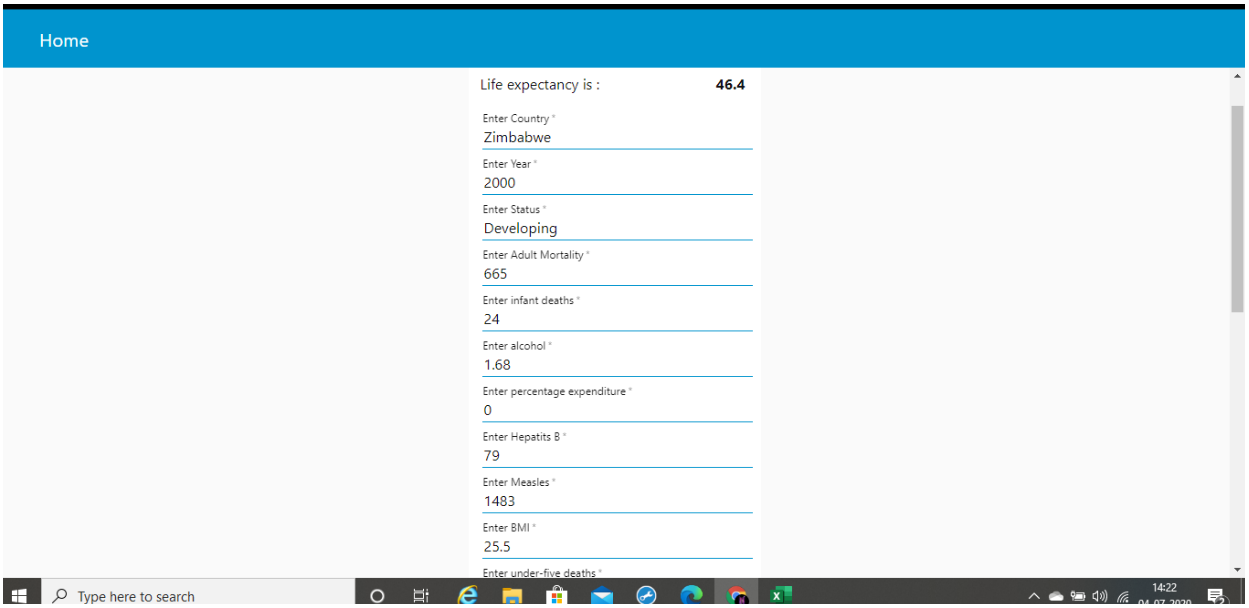
Label (Target)	Life expectancy
Number of Features	41
Created At	7/3/2020, 9:10:43 PM

Node RED Flow:



```
token_type: "Bearer", expires_in:
3600, expiration: 1593950663 … }
7/5/2020, 4:34:31 PM   node: 7a345fd6.9049b
msg.payload : Object
▸{ access_token:
"eyJraWQiOiIyMDIwMDYyNDE4MzAiLC…",
refresh_token:
"OKBOvAFyOAYkR4Xj6uyxOnBTcjEvXL…",
token_type: "Bearer", expires_in:
3600, expiration: 1593950670 … }
7/5/2020, 4:35:06 PM   node: b8f23026.3cab5
msg.payload : Object
▸{ predictions: array[1] }
7/5/2020, 4:35:07 PM   node: 97421588.1d7ba8
msg.payload : number
46.4
7/5/2020, 4:35:14 PM   node: b8f23026.3cab5
msg.payload : Object
▸{ predictions: array[1] }
7/5/2020, 4:35:14 PM   node: 97421588.1d7ba8
msg.payload : number
46.4
```

OUTPUT:



Life expectancy is : **46.4**

Enter Country *
Zimbabwe

Enter Year *
2000

Enter Status *
Developing

Enter Adult Mortality *
665

Enter infant deaths *
24

Enter alcohol *
1.68

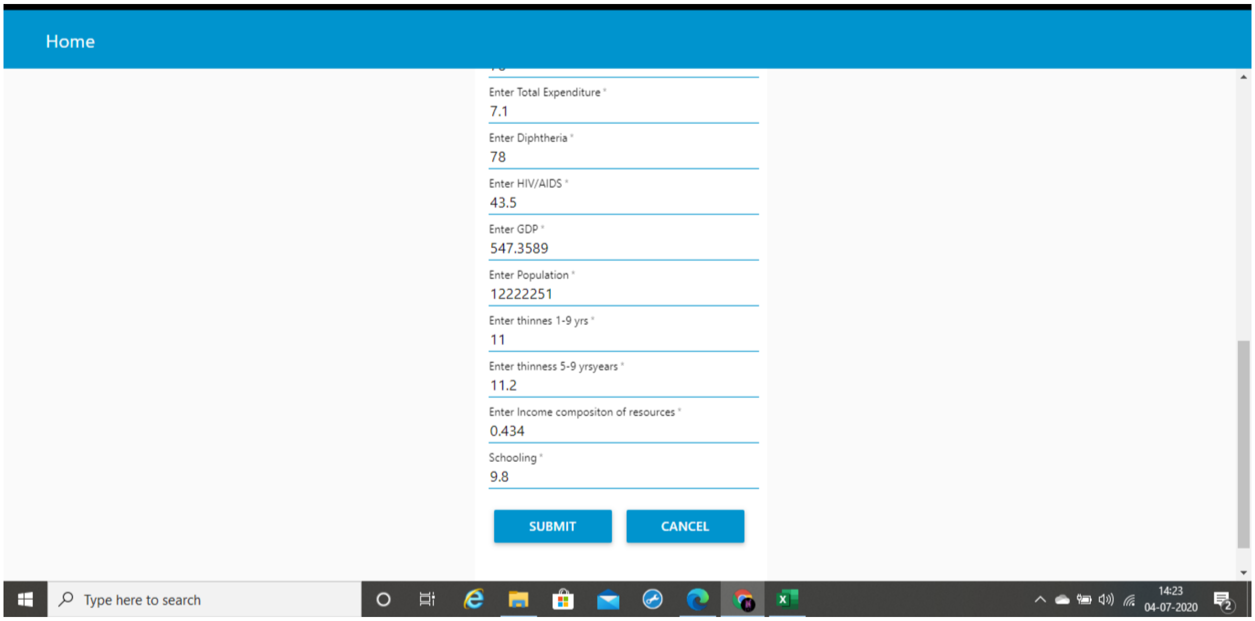
Enter percentage expenditure *
0

Enter Hepatits B *
79

Enter Measles *
1483

Enter BMI *
25.5

Enter under-five deaths *



Enter Total Expenditure *
7.1

Enter Diphtheria *
78

Enter HIV/AIDS *
43.5

Enter GDP *
547.3589

Enter Population *
12222251

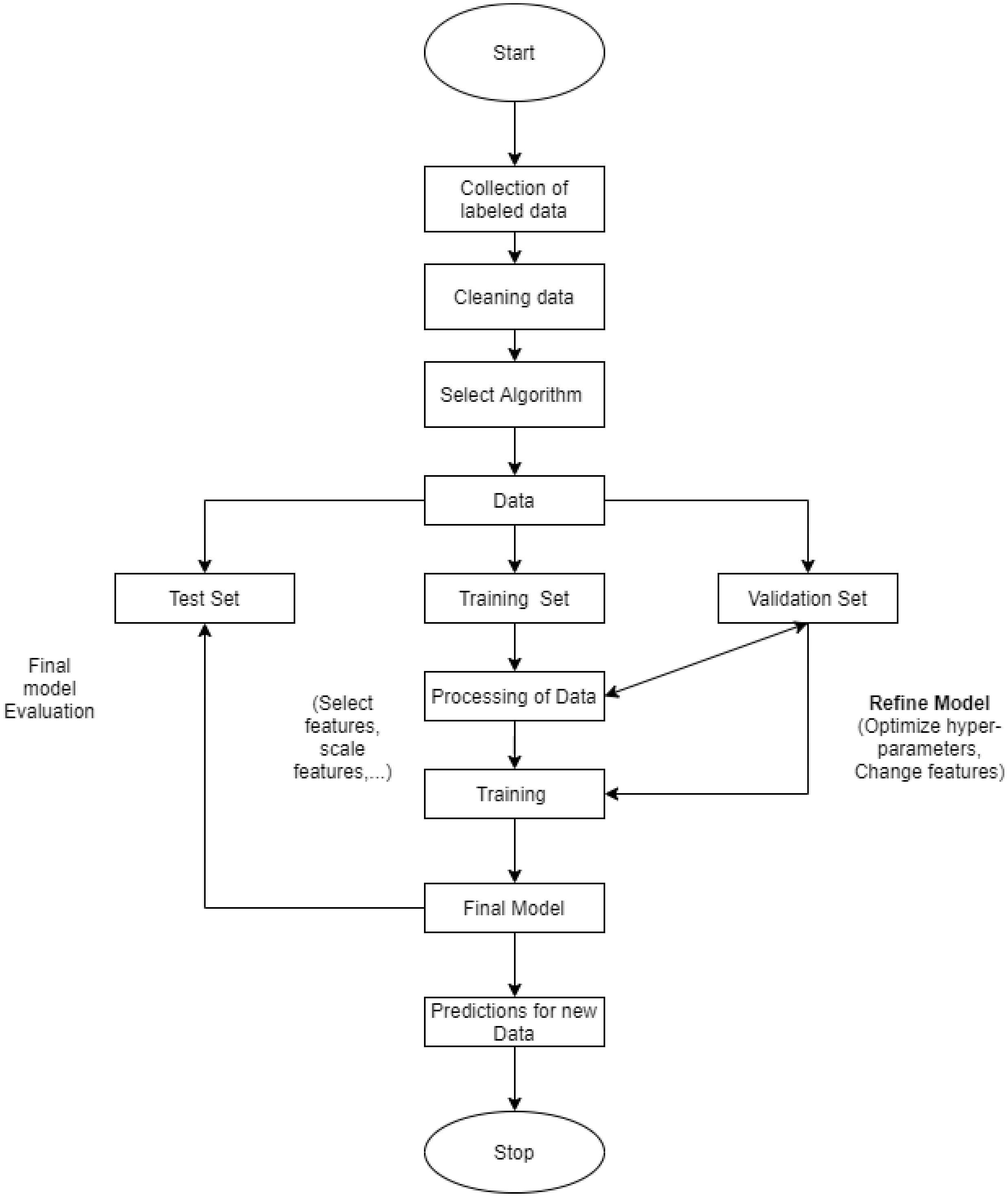
Enter thinnes 1-9 yrs *
11

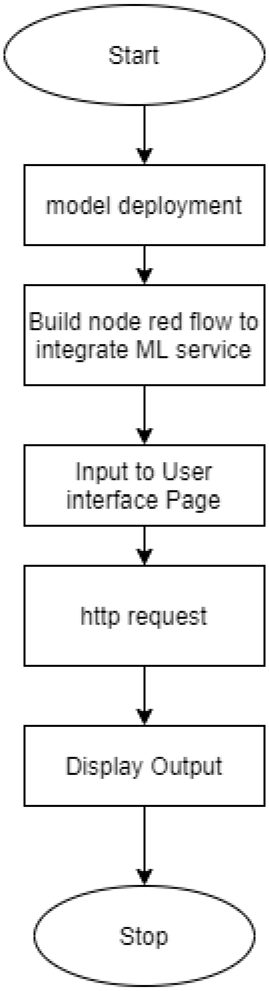
Enter thinness 5-9 yrsyears *
11.2

Enter Income compositon of resources *
0.434

Schooling *
9.8

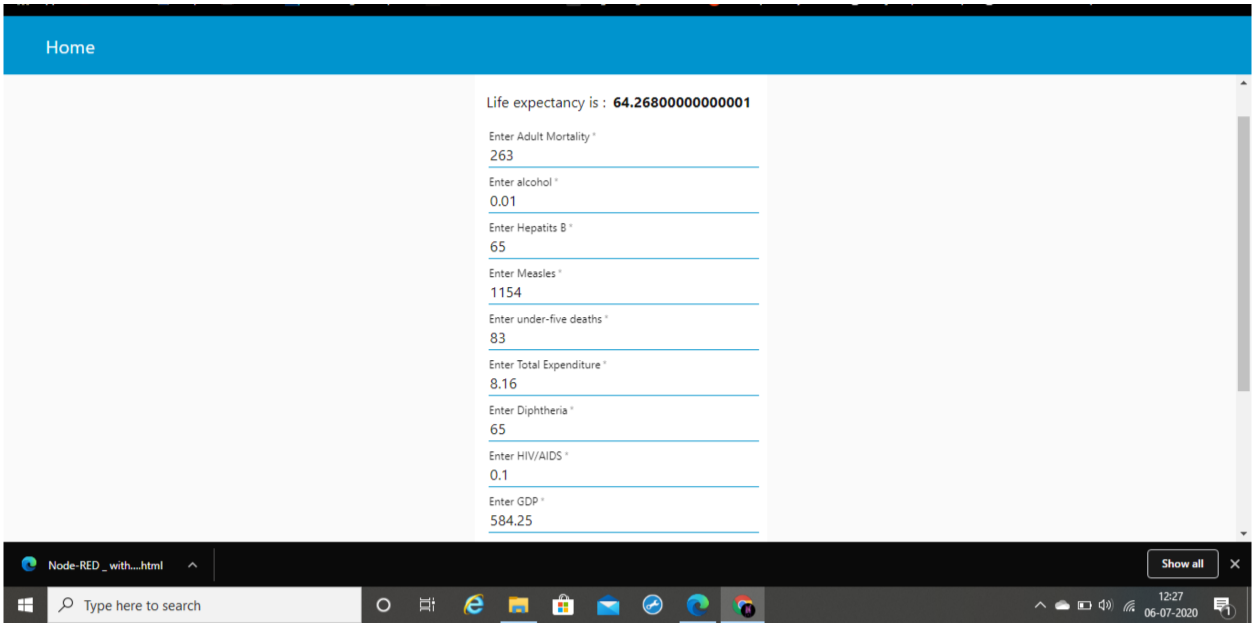
5. FLOW CHART:


Start
Collection of labeled data
Cleaning data
Select Algorithm
Data
Test Set
Training Set
Validation Set
Final model Evaluation
(Select features, scale features,...)
Processing of Data
Refine Model
(Optimize hyper-parameters, Change features)
Training
Final Model
Predictions for new Data
Stop


Start
model deployment
Build node red flow to integrate ML service
Input to User interface Page
http request
Display Output
Stop

6. RESULT:

Based on the given data, the autoAI model or ML model will understands the data, whatever the factors that are affecting the results we require i.e life expectancy. It will predict the output based on the features that we trained . Then based on the given input to the trained model, it will validate the features and give the accurate results with maximum efficiencty as output .



Life expectancy is : **64.26800000000001**

Enter Adult Mortality *
263

Enter alcohol *
0.01

Enter Hepatits B *
65

Enter Measles *
1154

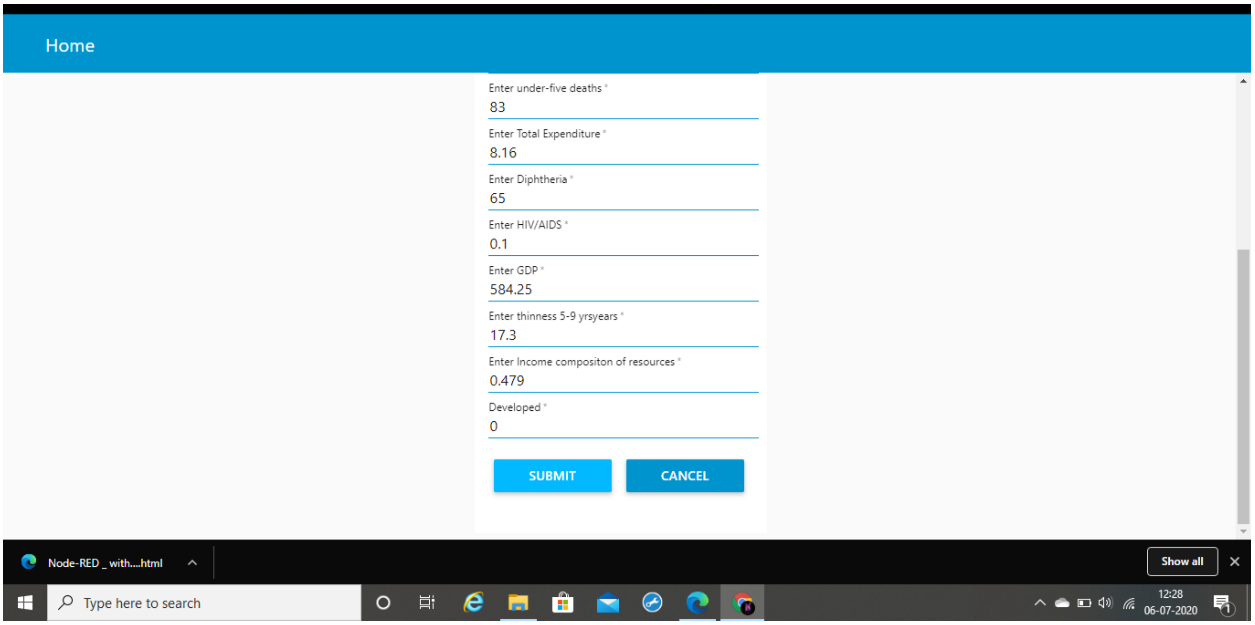
Enter under-five deaths *
83

Enter Total Expenditure *
8.16

Enter Diphtheria *
65

Enter HIV/AIDS *
0.1

Enter GDP *
584.25



Enter under-five deaths *
83

Enter Total Expenditure *
8.16

Enter Diphtheria *
65

Enter HIV/AIDS *
0.1

Enter GDP *
584.25

Enter thinness 5-9 yrsyears *
17.3

Enter Income compositon of resources *
0.479

Developed *
0

7. ADVANTAGES AND DISADVANTAGES:

Advantages:

1. By using this application we will be able to predict the number of years a person can expect to live.
2. It will be helped in the ease of gathering health data from the public as well as current government agencies such as centralized health servers could be increased
3. Good prognostication for example helps to determine the course of treatment and helps to anticipate the procurement of health care services and facilities, or more broadly .

Disadvantages:

1. Anamolies in database can lead to wrong predictions
2. Analysis on the data should be correct in order to get the accurate results
3. Accuracy is not 100%
4. Fake entries in the dataset will give wrong predictions

8. APPLICATIONS:

1.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 an individual.

2. This will also help increase the expectancy considering the impact of a specific factor on the average lifespan of people in a specific country.

9. 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 specific inputs into and then press the “Submit” button in to get the accurate results .

10. FUTURE SCOPE:

For 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

Big data 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.

APPENDIX:

APP/UI Web page using python:
https://node-red-eoxga.eu-gb.mybluemix.net/ui/#!/0?socketid=xP8kRyUlzz1BLLkqAAAK

APP/UI Web page using AutoAI:
https://node-red-eoxga.eu-gb.mybluemix.net/ui/#!/0?socketid=mhs-CFMHAanq5_THAAAO

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

Source Code :
https://github.com/SmartPracticeschool/IISPS-INT-2831-Predicting-Life-Expectancy-using-Machine-Learning/blob/master/Predicting%20Life%20Expectancy%20using%20python.ipynb