VISVESVARAYA TECHNOLOGICAL UNIVERSITY

JNANA SANGAMA, BELAGAVI – 590 018, KARNATAKA, INDIA

INTERNSHIP REPORT on project

Predicting Life expectancy of country using machine learning

06/06/2020 – 06/07/2020

SMARTBRIDGE EDUCATIONAL SERVICES Pvt. Ltd

Submitted in partial fulfillment of the requirements for the award of

BACHELOR OF ENGINEERING

IN

COMPUTER SCIENCE & ENGINEERING

Submitted By

Name	USN
VAISHNAVI	4VP17CS084

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

VIVEKANANDA COLLEGE OF ENGINEERING & TECHNOLOGY

[A Unit of Vivekananda Vidyavardhaka Sangha, Puttur (R)]

Affiliated to Visvesvaraya Technological University and Approved by AICTE New Delhi & Govt. of Karnataka

Nehru Nagara, Puttur – 574 203, DK, Karnataka, India

June 2020

Table of Contents

1. **INTRODUCTION** **3**
   a. Overview 3
   b. Purpose 3

2. **LITERATURE SURVEY** **4**
   2.1 Existing problem 4
   2.2 Proposed solution 4

3. **THEORITICAL ANALYSIS** **5**
   a. Block Diagram 5
   b. Hardware /Software designing 6

4. **EXPERIMENTAL INVESTIGATION** **7**
5. **FLOWCHART** **17**
6. **RESULT** **19**
7. **ADVANTAGES & DISADVANTAGES** **20**
8. **APPLICATIONS** **20**
9. **CONCLUSION** **21**
10. **FUTURE SCOPE** **21**
11. **BIBILOGRAPHY** **22**

APPENDIX

A. **Source code**

CHAPTER 1

INTRODUCTION

SMARTBRIDGE is an EdTech organization with a vision to bridge the gap between academia & industry. Our outcome-based experimental learning programs on emerging technologies (Internet of Things, Machine Learning, Data Science, Artificial Intelligence, Robotics) are building skilled entry -level engineers, for the corporate world. SmartBridge is in mission to build technology communities in academia to encourage students towards innovation & entrepreneurship. I did "prediction of life expectancy using machine learning" project.

1.1 Overview

A Machine Learning project "Prediction of Life Expectancy using Machine Learning". This is a typical regression machine learning project using the historical data predicting the future values. Using the various factor that effect the Life will be considered and future life expectancy of a country can be predicted using this model. We are having different years data of a single country. By taking the average of all the data of a country making single column. Then we will train the model.

1.2 Purpose

This model is used to train machine learning models using dataset containing data related to Regional variations, Economic Circumstances, Mental Illnesses, Physical Illnesses, Education 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 healthcare system and some specific disease related deaths that happened in the country are given. By using this data, we are going to build a regression model which predicts the life expectancy of a human in country

CHAPTER 2

LITERATURE SURVEY

2.1 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.We have a data of different countries which effect the life expectancy of humans. We have to predict the life expectancy of a country for a given values. This will help to know how we can increase the life expectancy by keeping effecting values to particular value. We are having data like Country, status like developed country or developing country, GDP, percentage expenditure, schooling, Alcohol , adult morality, hepatitis B, HIV/AIDS, infant deaths, Measles, BMI, under-five deaths, Polio,Total expenditure, Diphtheria, population, thinness 1-19 years, thinness 5-9 years, Income composition of resources etc.

2.2 Proposed Solution

A Machine learning model which is used predict the life expectancy of country. Initially import the required libraries like NumPy, matplotlib, pandas, scikit learn etc. In this, pandas is used to read the dataset, NumPy is used for numerical related works, matplotlib is used to plot the relation graph, heat map etc. Initially you have to find the data which is not used to find the life expectancy, then filter out those data. Take care of missing data. If there is classification in remaining data, then either using label encoder or by using dummies you can classify that data. This process is called data preprocessing. Then start finding which algorithm is very suitable for this model. In my project I found that Linear Regression and Random Forest algorithm is best method to do this project. I got score accuracy of 0.88 for both. Separate the training and testing data for building model. Then import algorithm libraries. Then start building the model. I have done this using IBM Watson to develop a python code. Then to provide a input to predict the life expectancy I have used Node-Red technology from IBM. Then I have connected using credential of deployment.

CHAPTER 3

THEORITICAL ANALYSIS

Machine Learning is a technique required to complete the project. In this project I have used NumPy, Matplotlib, Pandas and scikit learn.

3.1 BLOCK DIAGRAM

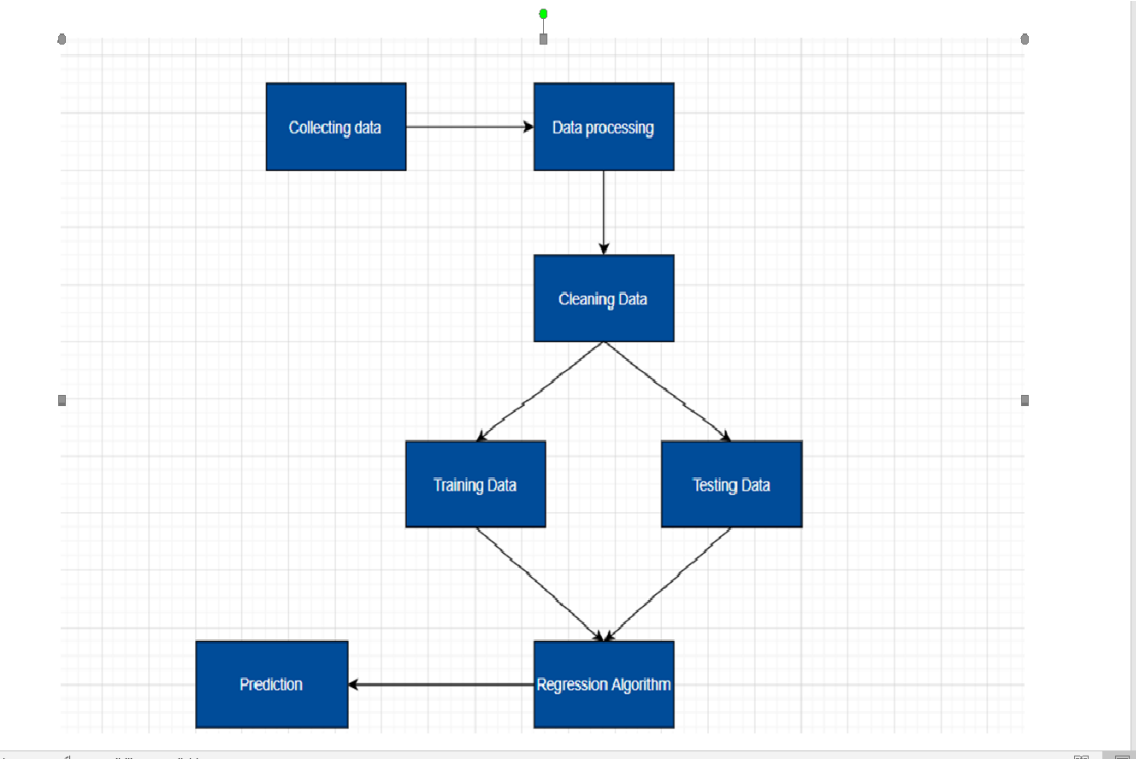


Fig 3.1

3.2 HARDWARE/SOFTWARE DESIGNING

I have designed a User Interface that could help me to calculate the life expectancy by giving input to the various values through that interface. It shows value of Life Expectancy in years.

Software Requirements:

IBM cloud services

IBM Watson Studio

IBM Auto AI experiment

IBM Node-Red application

7 Jupyter Notebook

GitHub

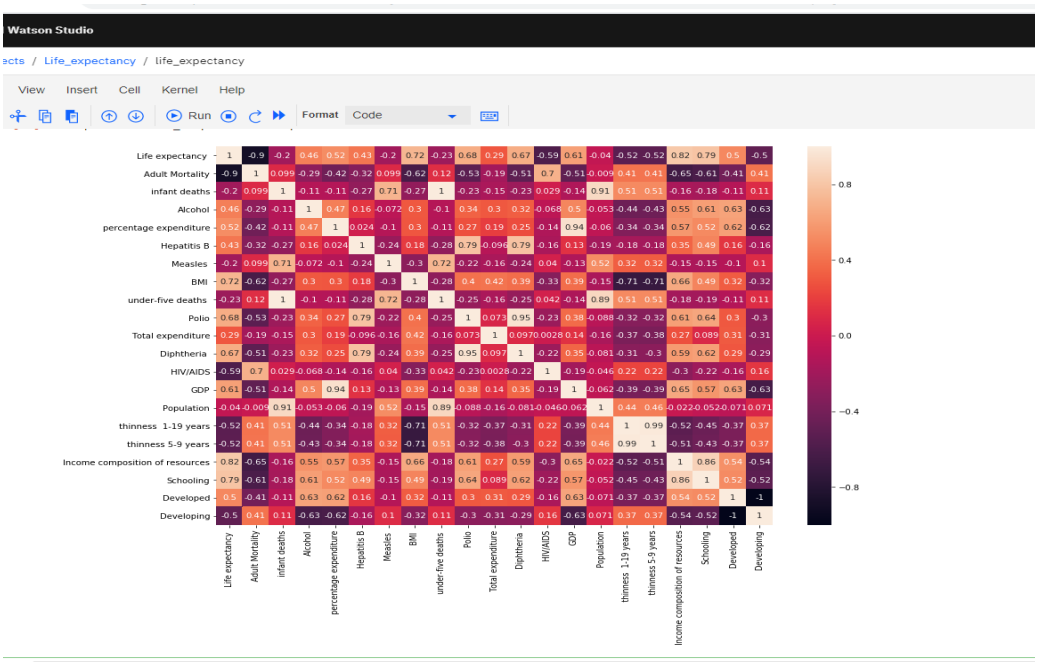
Zoho document writer

CHAPTER 4

EXPERIMENTAL INVESTIGATION

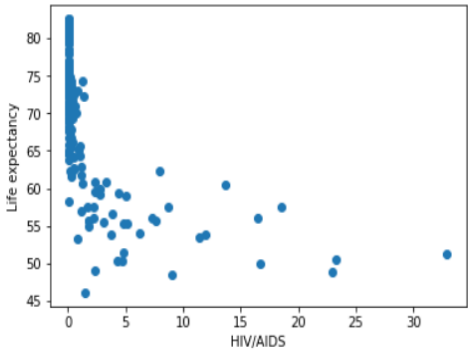
These are some graphs from the refined data analysis, that makes us understand relation between data.

With python

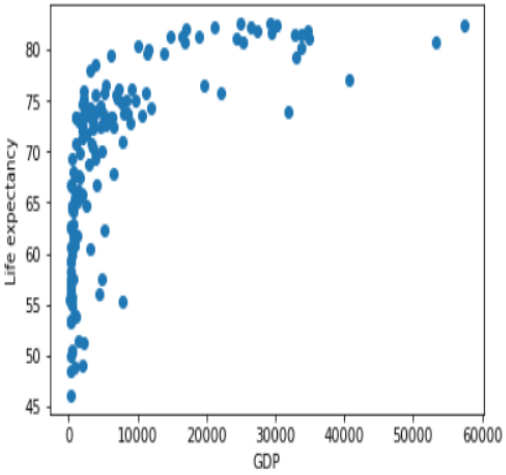


4.1 Relation between different factor and Life Expectancy

```
Out[45]: Text(0, 0.5, 'Life expectancy')
```

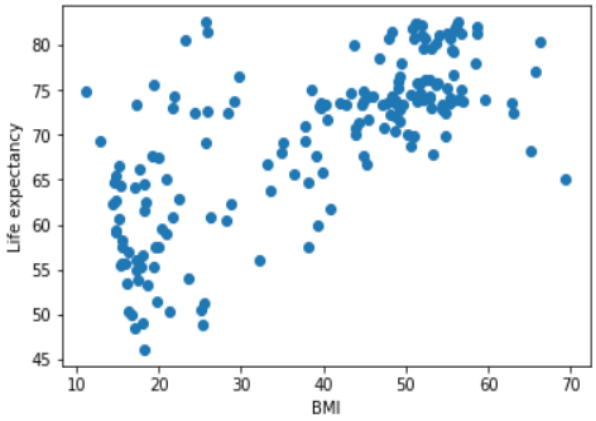



```
Out[47]: Text(0, 0.5, 'Life expectancy')
```



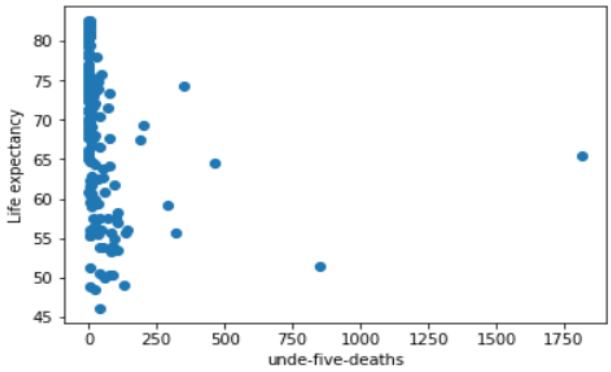
```
plt.pyplot.ylabel('Life expectancy')
```

```
Out[49]: Text(0, 0.5, 'Life expectancy')
```



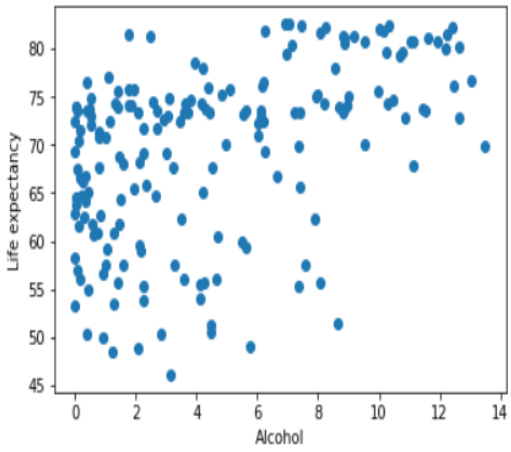

```
plt.pyplot.ylabel('Life expectancy')
```

```
Out[51]: Text(0, 0.5, 'Life expectancy')
```

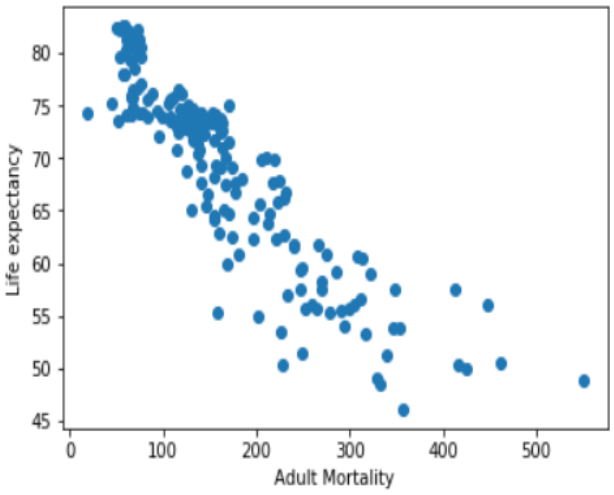


```
plt.pyplot.ylabel('Life expectancy')
```

```
Out[52]: Text(0, 0.5, 'Life expectancy')
```

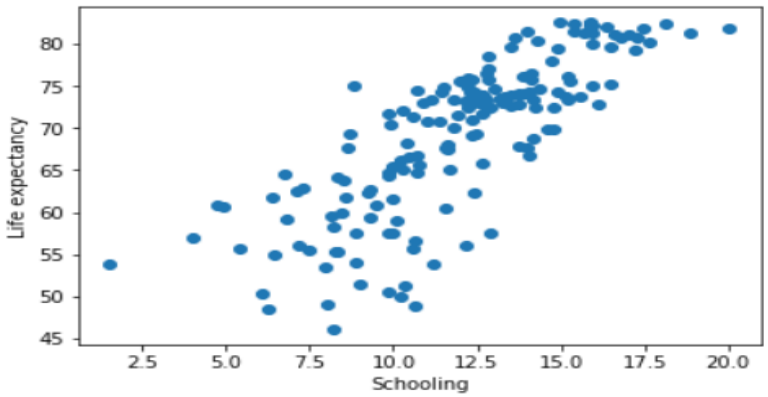



```
Out[53]: Text(0, 0.5, 'Life expectancy')
```

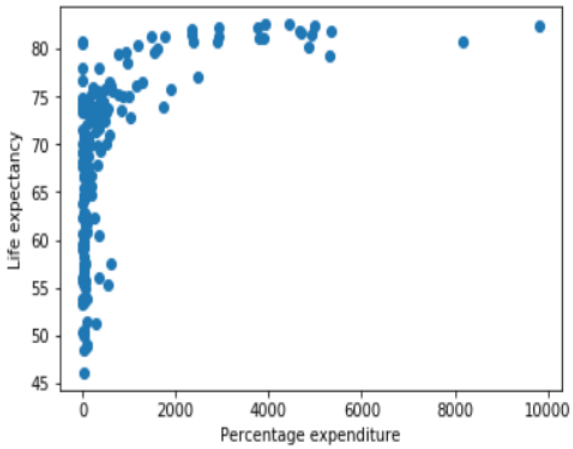


```
plt.pyplot.ylabel('Life expectancy')
```

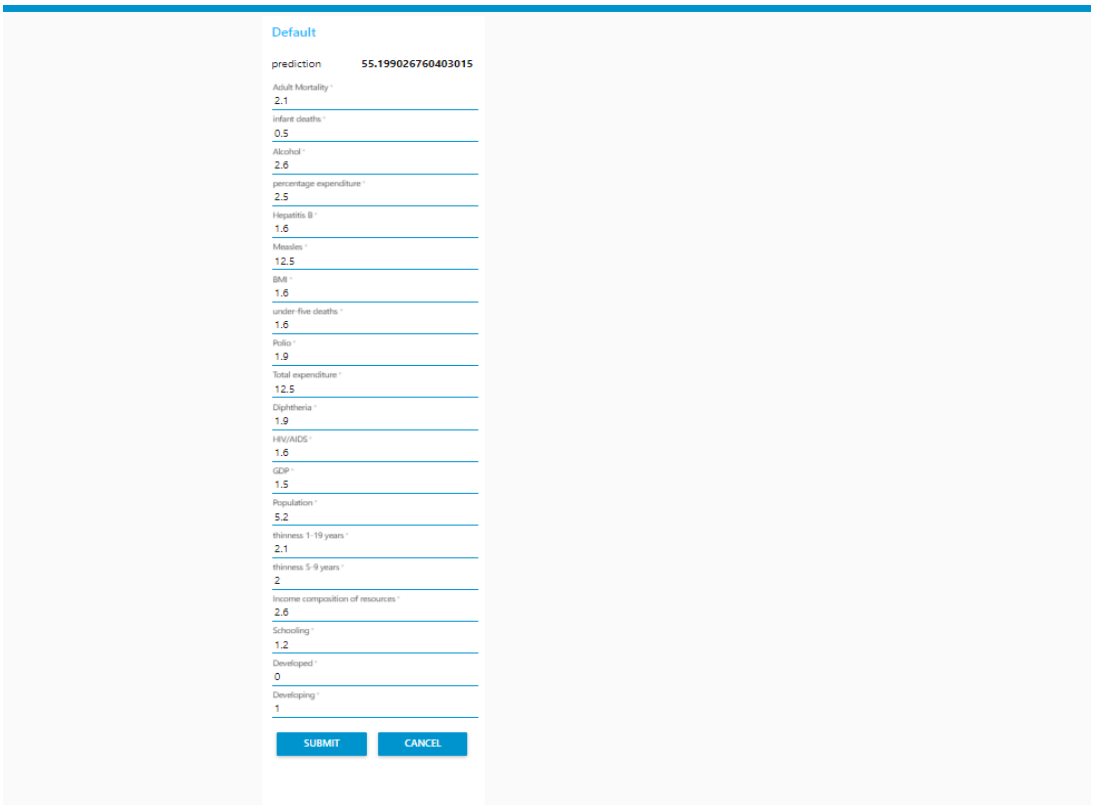
```
[54]: Text(0, 0.5, 'Life expectancy')
```



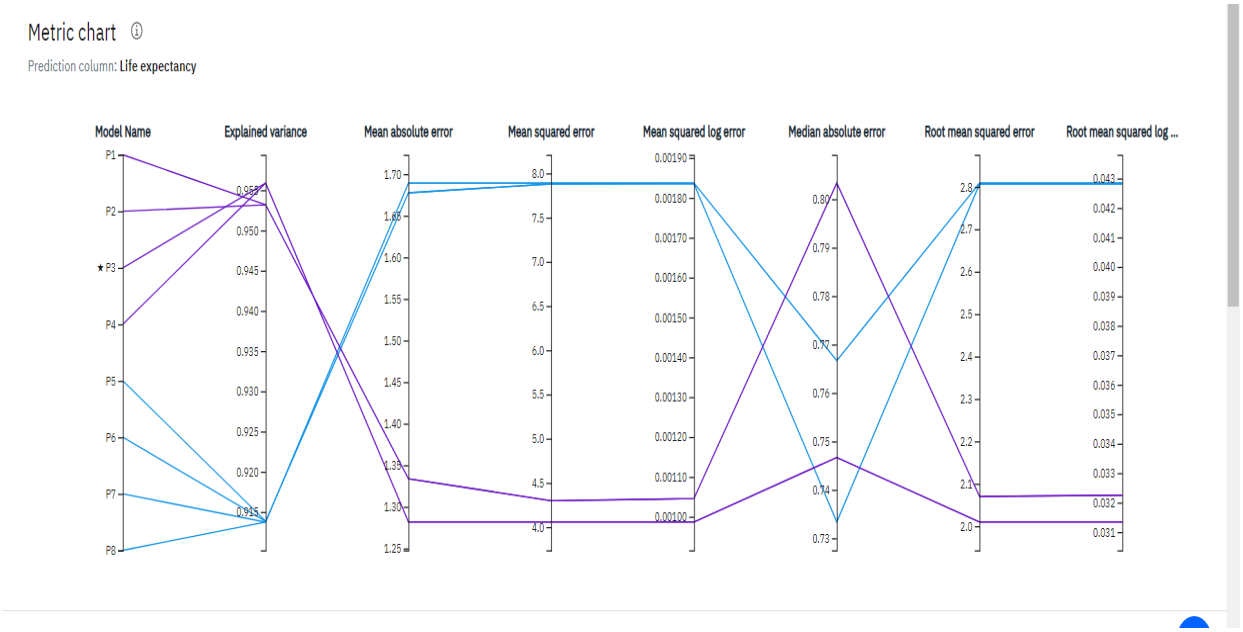

```
Out[56]: Text(0, 0.5, 'Life expectancy')
```

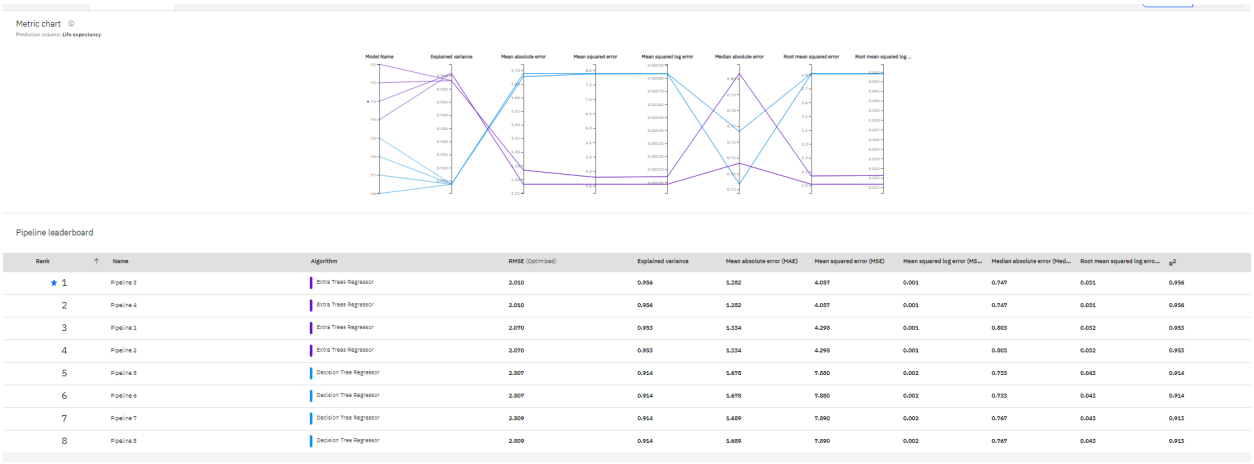


Some random inputs are given as values for a deployed model.

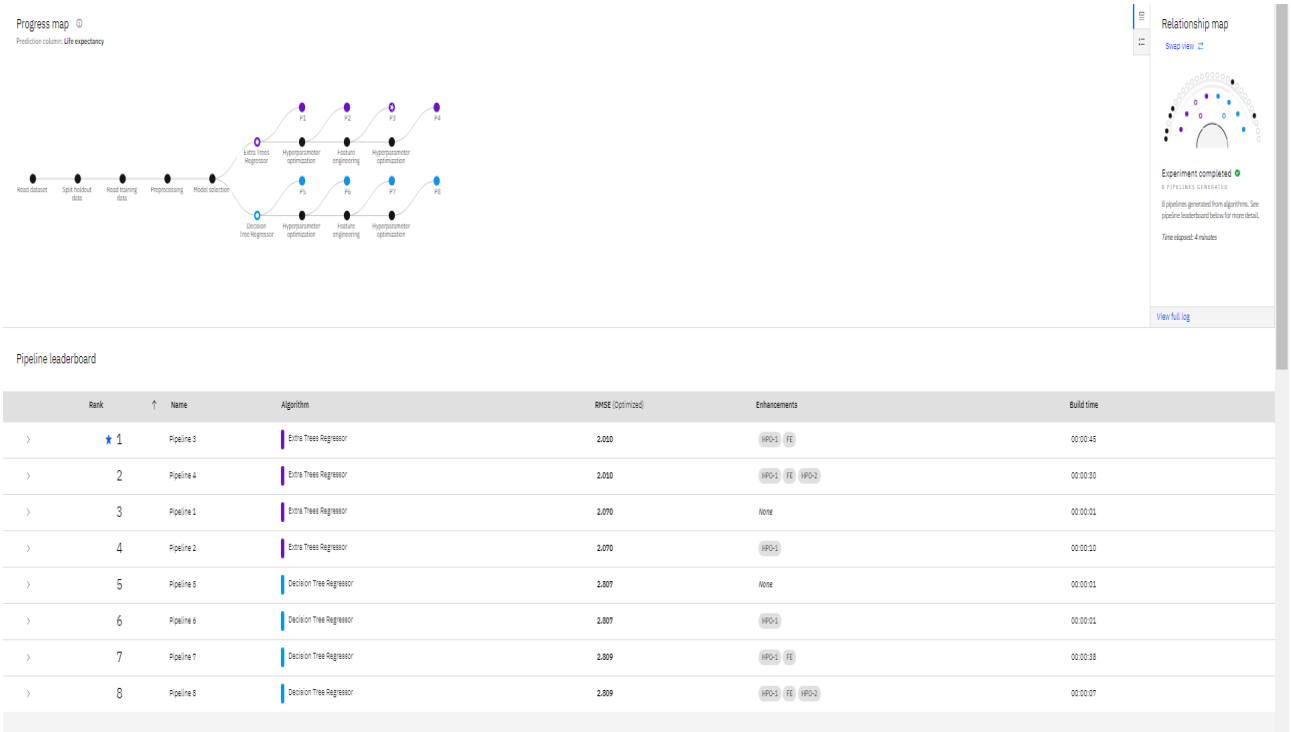


Without Python

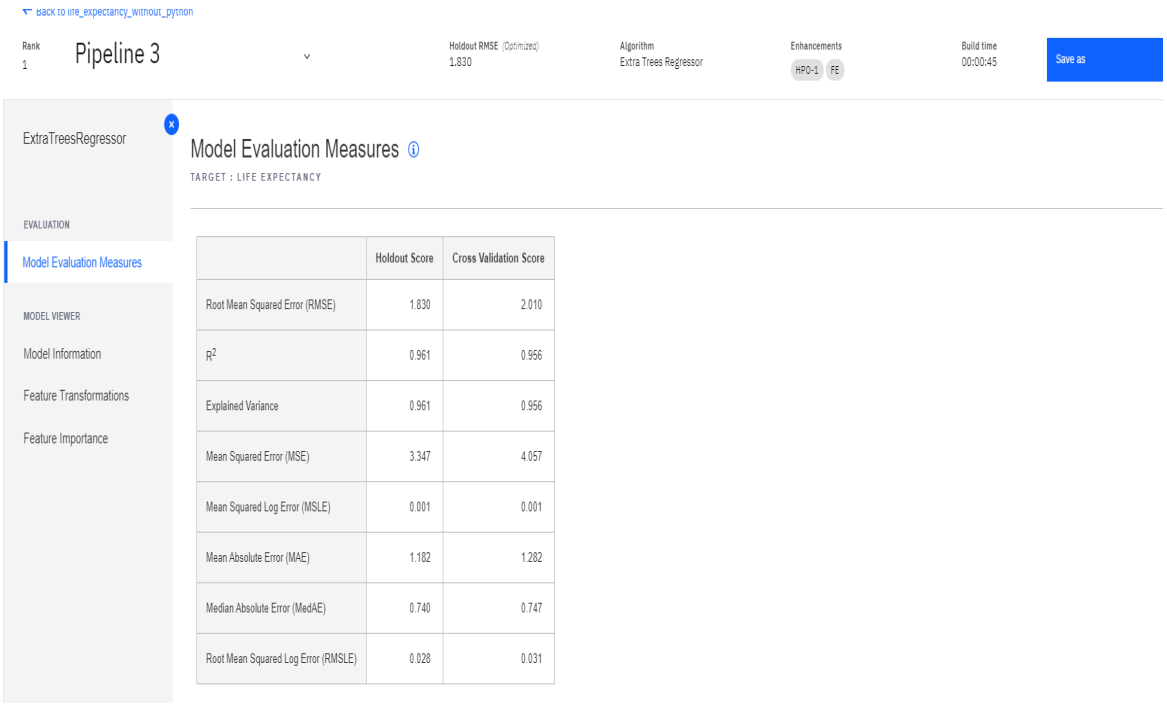




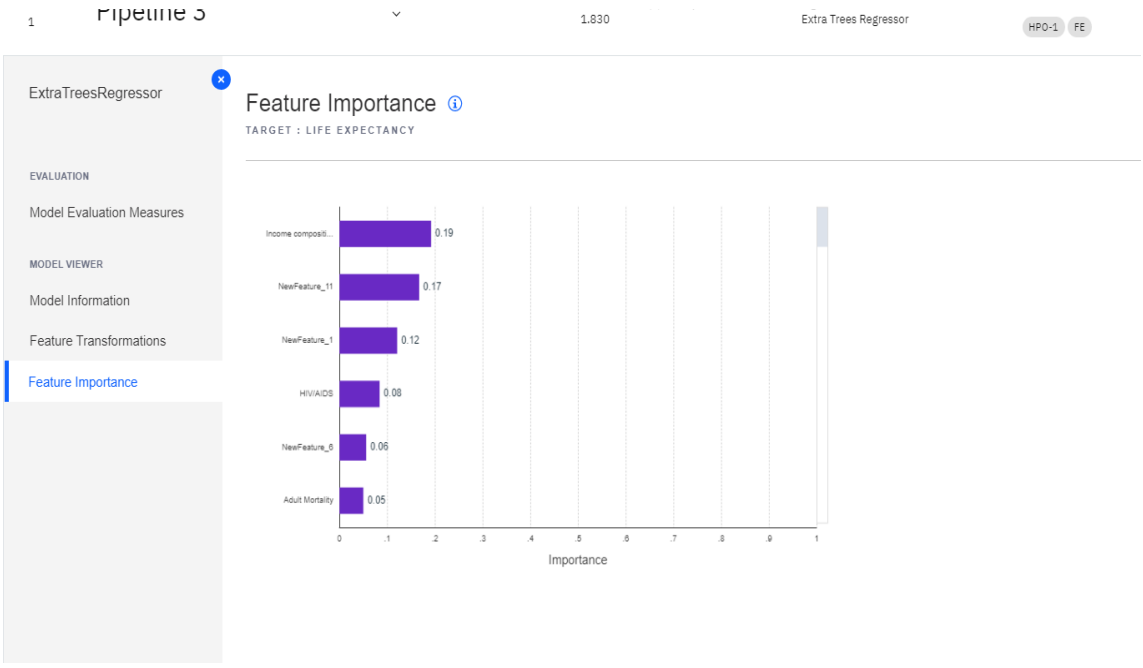
Rank	Name	Algorithm	RMSE (Optimized)	Explained variance	Mean absolute error (MAE)	Mean squared error (MSE)	Mean squared log error (MS...	Median absolute error (Med...	Root mean squared log erro...	R^2
★ 1	Pipeline 3	Extra Trees Regressor	2.010	0.956	1.282	4.097	0.001	0.747	0.031	0.956
2	Pipeline 4	Extra Trees Regressor	2.010	0.956	1.282	4.097	0.001	0.747	0.031	0.956
3	Pipeline 1	Extra Trees Regressor	2.070	0.953	1.334	4.298	0.001	0.803	0.032	0.953
4	Pipeline 2	Extra Trees Regressor	2.070	0.953	1.334	4.298	0.001	0.803	0.032	0.953
5	Pipeline 5	Decision Tree Regressor	2.807	0.914	1.676	7.880	0.002	0.733	0.043	0.914
6	Pipeline 6	Decision Tree Regressor	2.807	0.914	1.678	7.880	0.002	0.733	0.043	0.914
7	Pipeline 7	Decision Tree Regressor	2.809	0.914	1.689	7.890	0.002	0.767	0.043	0.913
8	Pipeline 8	Decision Tree Regressor	2.809	0.914	1.689	7.890	0.002	0.767	0.043	0.913

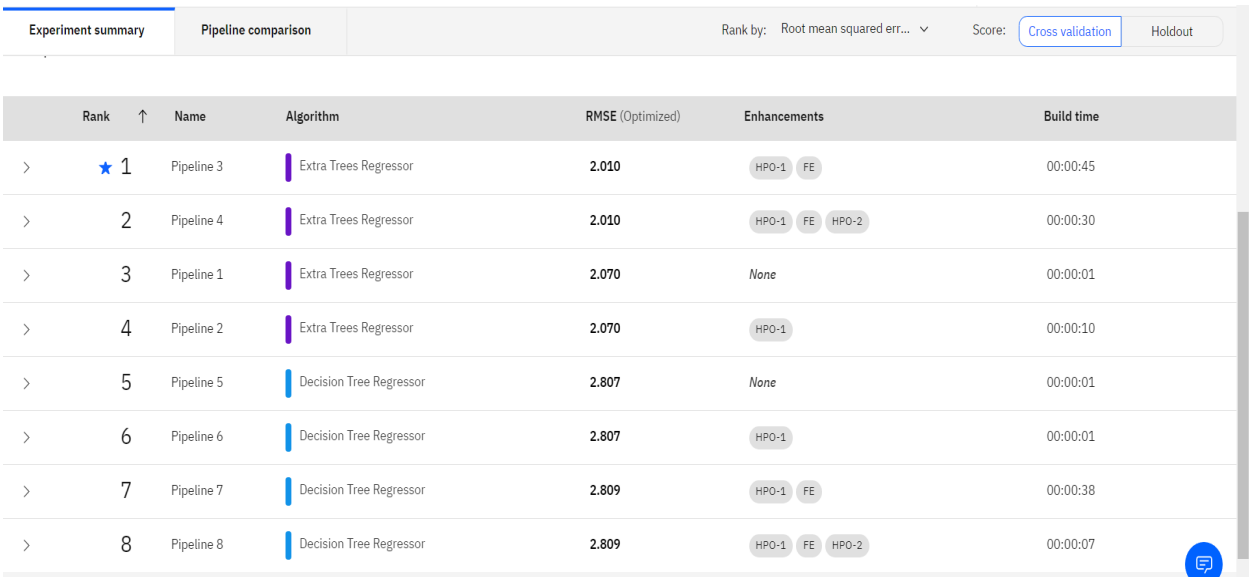


	Rank	Name	Algorithm	RMSE (Optimized)	Enhancements	Build time
›	★ 1	Pipeline 3	Extra Trees Regressor	2.010	HPO-1 FE	00:00:45
›	2	Pipeline 4	Extra Trees Regressor	2.010	HPO-1 FE HPO-2	00:00:30
›	3	Pipeline 1	Extra Trees Regressor	2.070	None	00:00:01
›	4	Pipeline 2	Extra Trees Regressor	2.070	HPO-1	00:00:10
›	5	Pipeline 5	Decision Tree Regressor	2.807	None	00:00:01
›	6	Pipeline 6	Decision Tree Regressor	2.807	HPO-1	00:00:01
›	7	Pipeline 7	Decision Tree Regressor	2.809	HPO-1 FE	00:00:38
›	8	Pipeline 8	Decision Tree Regressor	2.809	HPO-1 FE HPO-2	00:00:07



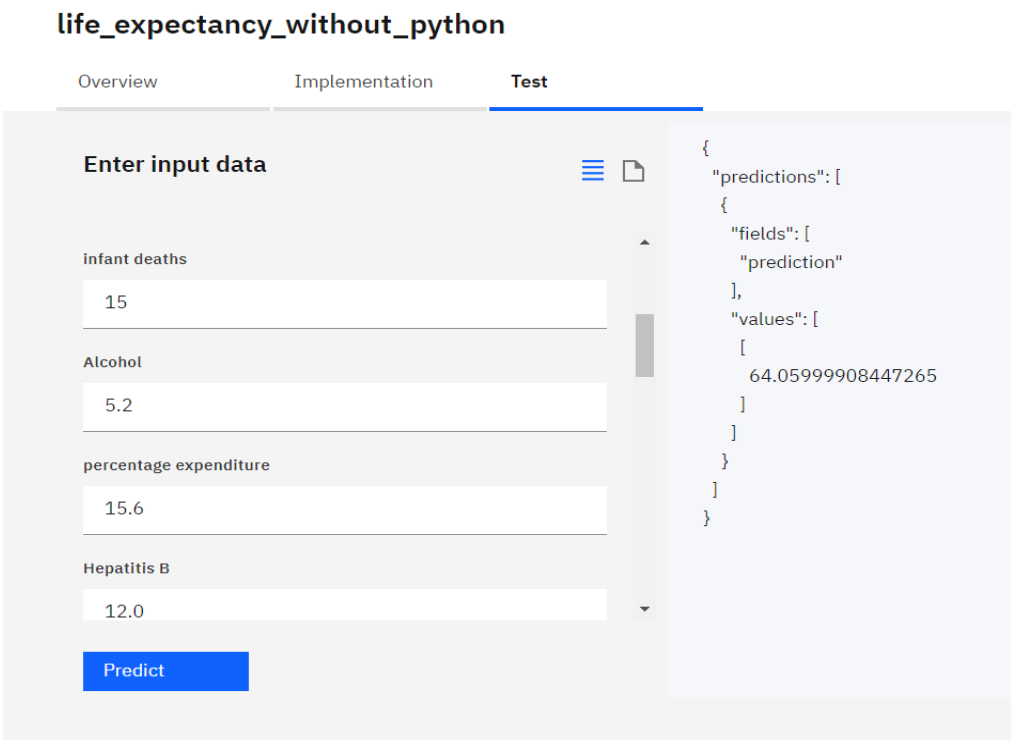
	Holdout Score	Cross Validation Score
Root Mean Squared Error (RMSE)	1.830	2.010
R^2	0.961	0.956
Explained Variance	0.961	0.956
Mean Squared Error (MSE)	3.347	4.057
Mean Squared Log Error (MSLE)	0.001	0.001
Mean Absolute Error (MAE)	1.182	1.282
Median Absolute Error (MedAE)	0.740	0.747
Root Mean Squared Log Error (RMSLE)	0.028	0.031

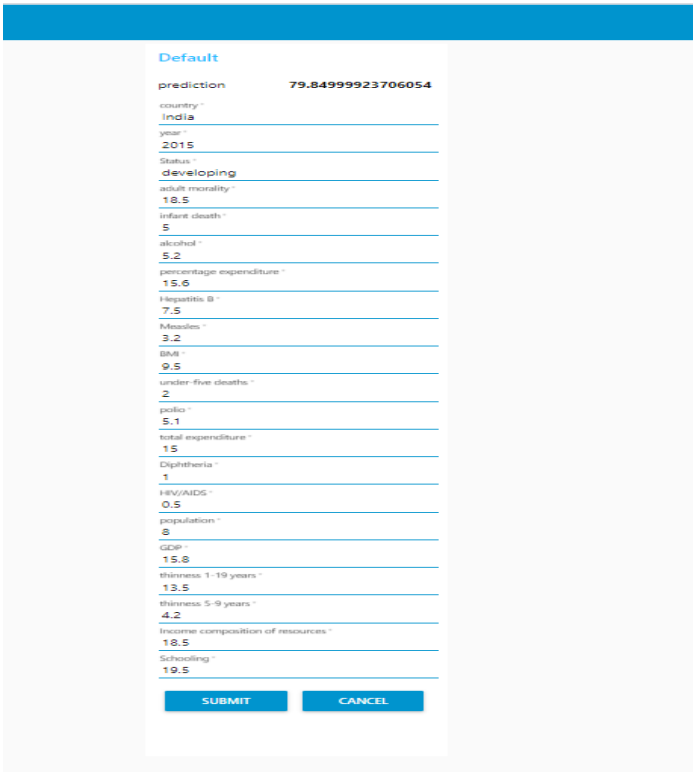




Experiment summary | Pipeline comparison | Rank by: Root mean squared err... | Score: Cross validation | Holdout

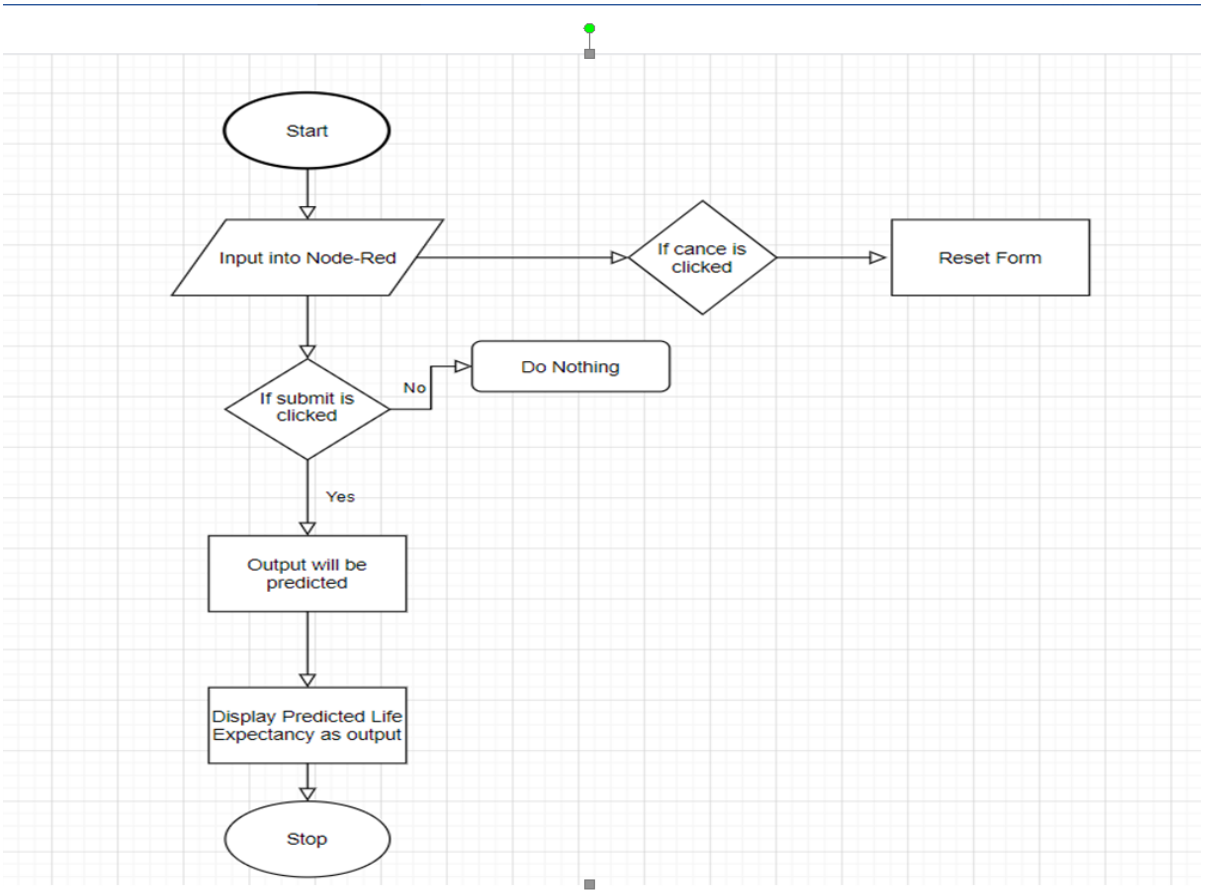
Rank	Name	Algorithm	RMSE (Optimized)	Enhancements	Build time
★ 1	Pipeline 3	Extra Trees Regressor	2.010	HPO-1, FE	00:00:45
2	Pipeline 4	Extra Trees Regressor	2.010	HPO-1, FE, HPO-2	00:00:30
3	Pipeline 1	Extra Trees Regressor	2.070	*None*	00:00:01
4	Pipeline 2	Extra Trees Regressor	2.070	HPO-1	00:00:10
5	Pipeline 5	Decision Tree Regressor	2.807	*None*	00:00:01
6	Pipeline 6	Decision Tree Regressor	2.807	HPO-1	00:00:01
7	Pipeline 7	Decision Tree Regressor	2.809	HPO-1, FE	00:00:38
8	Pipeline 8	Decision Tree Regressor	2.809	HPO-1, FE, HPO-2	00:00:07

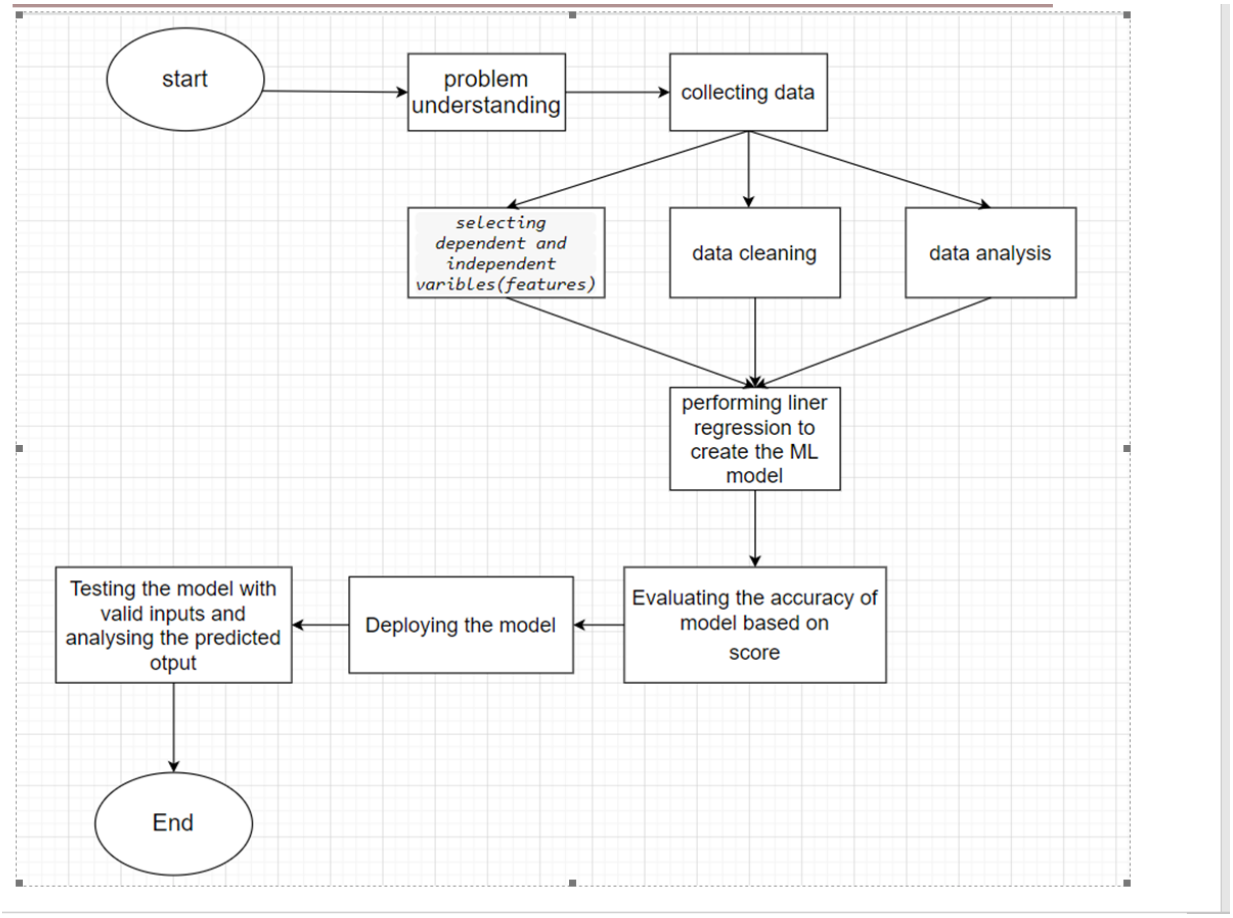



Default
prediction
79.8499923706054
country *
India
year *
2015
Status *
developing
adult morality *
18.5
infant death *
5
alcohol *
5.2
percentage expenditure *
15.6
Hepatitis B *
7.5
Measles *
3.2
BMI *
9.5
under-five deaths *
2
polio *
5.1
total expenditure *
15
Diphtheria *
1
HIV/AIDS *
0.5
population *
8
GDP *
15.8
thinness 1-19 years *
13.5
thinness 5-9 years *
4.2
Income composition of resources *
18.5
Schooling *
19.5
SUBMIT
CANCEL

CHAPTER 5

FLOW CHART

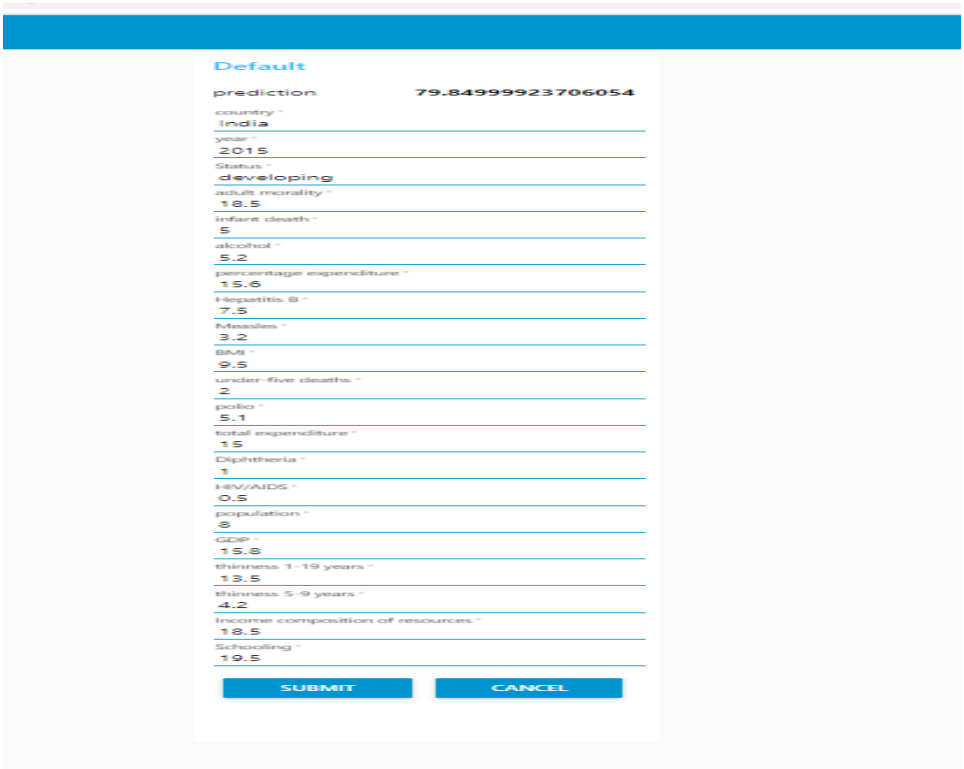



start
problem understanding
collecting data
selecting dependent and independent varibles(features)
data cleaning
data analysis
performing liner regression to create the ML model
Evaluating the accuracy of model based on score
Deploying the model
Testing the model with valid inputs and analysing the predicted otput
End

CHAPTER 6

RESULT

Based on given data ML model understands the data and it will predict the life expectancy. It will consider the factors that are affecting the result and produce the life expectancy. It will predict the data based on how we trained the model, what are the data given as input and it will give the accurate value predicted output. So the results we get are approximation and it will not definitely true, but it works for maximum cases.



CHAPTER 7

1. ADVANTAGES AND DISADVANTAGES

2. **Advantages:**

1.Since we can predict the life span, we can know what factors are influencing the life expectancy on life span in what ways.

2.So, by improving the factors in real life, we can increase the life span.

3. **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

8.APPLICATION

4. 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 suggested that you should purchase a life insurance policy for an individual.
5. This will also help increase the expectancy considering the impact of specific factor on the average life span of people in specific country.

9.CONCLUSION

I had a great experience in various technology. By doing this project I learned how to create a Node-Red flow, How I can use Watson Studio and Machine Learning, how to create a instance for that, How can I integrate Node-Red and Watson Studio etc. I also learned how can I build a machine learning model using Auto-AI. In machine learning how can I process the data, how can I train the model etc. I got knowledge about IBM cloud services. This project made me to learn new technology, which is my hobby too.

10.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.

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.

12.BIBILOGRAPHY

APPENDIX

1. APP/UI WEBPAGE:
   https://node-red-zjaix.eu-gb.mybluemix.net/red/#flow/cabdca81.2d4928

   https://node-red-zjaix.eu-gb.mybluemix.net/red/#flow/c9be6228.7dae3

   https://node-red-zjaix.eu-gb.mybluemix.net/red/#flow/1df14907.d7a2a7
2. Dataset link: :https://www.kaggle.com/kumarajarshi/life-expectancy-who
3. Source code:https://github.com/vaishnavisampigethaya/Life_expectancy

   https://github.com/vaishnavisampigethaya/simple_regression
4. link for demonstration video:
   https://github.com/vaishnavisampigethaya/Life_expectancy