A

Project

Report

on

Predicting Life Expectancy using Machine Learning.

Internship under:

The SMARTBRIDGE

NAME: ANKIT SHARMA

EMAIL : **anksh4899@gmail.com**

PROJECT ID : SPS_PRO_215

INTERNSHIP TITLE : Predicting Life Expectancy using Machine Learning - **SB43733**

Category: Machine Learning

Table Of Content

1. INTRODUCTION

OVERVIEW :-

This project “Predicting Life Expectancy using Machine Learning” is an web application that predict the expected average life span of people of a given country based on various features. This project is built using IBM services(Watson studio, Node Red, Watson machine learning).

Life expectancy is a statistical measure of the average time a human being is expected to live, Life expectancy depends on various factors: 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 healthcare system and some specific disease related deaths that happened in the country are given.

Requirements:-

- Project Requirements: IBM Cloud, IBM Watson, Node-RED
- Functional Requirements: IBMcloud
- Technical Requirements: WATSON MachineLearning
- Software Requirements: Python, Watson Studio, Node-Red

PURPOSE:-

The average life Expectancy of a certain country says many things about that particular country. It ultimately helps in predicting the health conditions and the development of the health sector in that particular country. This ultimately helps the nation to find the area which needs attention in an urge to improve its contribution in average lifespan of a human being. The expectancy obviously depends upon the country’s population, GDP, the economy of the country and many more factors. It is not enough to have a long life ,Instead with having a long life one should have a fit life as well

2. LITERATURE SURVEY

EXISTING PROBLEM:- In our regular prediction system, there are many problems exist such as :

- Health relateddisease.
- occupational or social class, area level deprivation, geographical area of residence (urban and rural), housingtenure.
- Race-based inequalities.
- whole concept of life expectancy depends on the interpretation given to "fullhealth".
- Or the factors used to predict the life expectancy of people are based on some associated specific features of particular fields like : morbidity and mortality (smoking, alcohol consumption, overweight and obesity, and physicalactivity).

PROPOSED SOLUTION:-

- For the above problem to get solved we have a dataset consist of various factors. In this system we have taken all the correlated features into consideration. So the target output variable i.e expected life span of the people depends upon variety of factors and not factors of particular fields.

- Important immunization like Hepatitis B, Polio and Diphtheria are also considered.

- The project uses immunization factors, mortality factors, economic factors, social factors and other health related factors to predict life expectancy of a country for a given year using a machine learning model.

- Since the observations in this dataset are based on different countries, it will be easier for a country to determine the predicting factor which is contributing to lower value of life expectancy. This will help in suggesting a country, which area should be given importance in order to efficiently improve the life expectancy of itspopulation.

3.THEOROTICAL ANALYSIS

HARDWARE / SOFTWARE DESIGNING:-

1. Create necessary IBM Cloudservices
2. Create Watson studioproject
3. Configure WatsonStudio
4. Create IBM Machine Learninginstance
5. Create machine learning model in Jupyternotebook
6. Deploy the machine learningmodel
7. Create flow and configurenode
8. Integrate node red with machine learningmodel
9. Deploy and run Node Redapp.

Input is taken from the user using a “Form” element in Node-Red. Then, an HTTP request is made to the IBM cloud that further makes an HTTP request to the deployed model using model’s instance id. After verification of id, the model sends an HTTP response which is finally parsed by the Node-Red application and the result is displayed on the user screen.

4. EXPERIMENTAL INVESTIGATIONS

Following factors are taken into account for predicting the life expectancy of a country.

1. Country
2. Status: Developed or Developing status of thecountry.
3. Year
4. Adult mortality: Adult Mortality Rates of both sexes (probability of dying between 15 and 60 years per 1000population).
5. Infant deaths: Number of Infant Deaths per 1000population.
6. Alcohol: Alcohol, recorded per capita (15+)consumption.
7. Percentage Expenditure: Expenditure on health as a percentage of Gross Domestic Product per capita (%).
8. Hepatitis B: Hepatitis B =immunization coverage among 1-year-olds(%).
9. Measles: Measles - number of reported cases per 1000 population.
10. BMI: Average Body Mass Index of entirepopulation.
11. Under-five deaths: Number of under-five deaths per 1000population.
12. Polio: Polio (Pol3) immunization coverage among 1-year-olds(%).
13. Total expenditure: General government expenditure on health as a percentage of total government expenditure(%).
14. Diphtheria: Diphtheria tetanus toxoid and pertussis (DTP3) immunization coverage among 1-year- olds (%).
15. HIV/AIDS: Deaths per 1 000 live births HIV/AIDS (0-4 years).
16. GDP: Gross Domestic Product per capita (inUSD).
17. Population: Population of thecountry.
18. Thinness 10-19 years: Prevalence of thinness among children and adolescents for Age 10 to 19(%).
19. Thinness 5-9 years: Prevalence of thinness among children for Age 5 to9(%).
20. Income composition of resources: Human Development Index in terms of income composition of resources (index ranging from 0 to 1).

Steps to Create IBM Cloud services

- WatsonStudio
- Watson Machine Learning
- NodeRed

1. Create **Watson Studio** service instance.

   - Select **Catalog** found at the top right of thepage.
   - Click on **Watson** from the menu on the left, which you can find under **Platform**services.
   - Select WatsonStudio.
   - Enter the **Service name** or keep the default value and make sure to select the **US South** as the **region/location** and your desired **organization**, and**space**.
   - Select **Lite** for the **Plan**, which you can find under **Pricing Plans** and is already selected. Please note you are only allowed one instance of a Lite plan perservice.
   - Click on**Create**.
   - You will be taken to the main page of the service. Click on **GetStarted**.
   - Create a NewProject

2. Create Notebook.

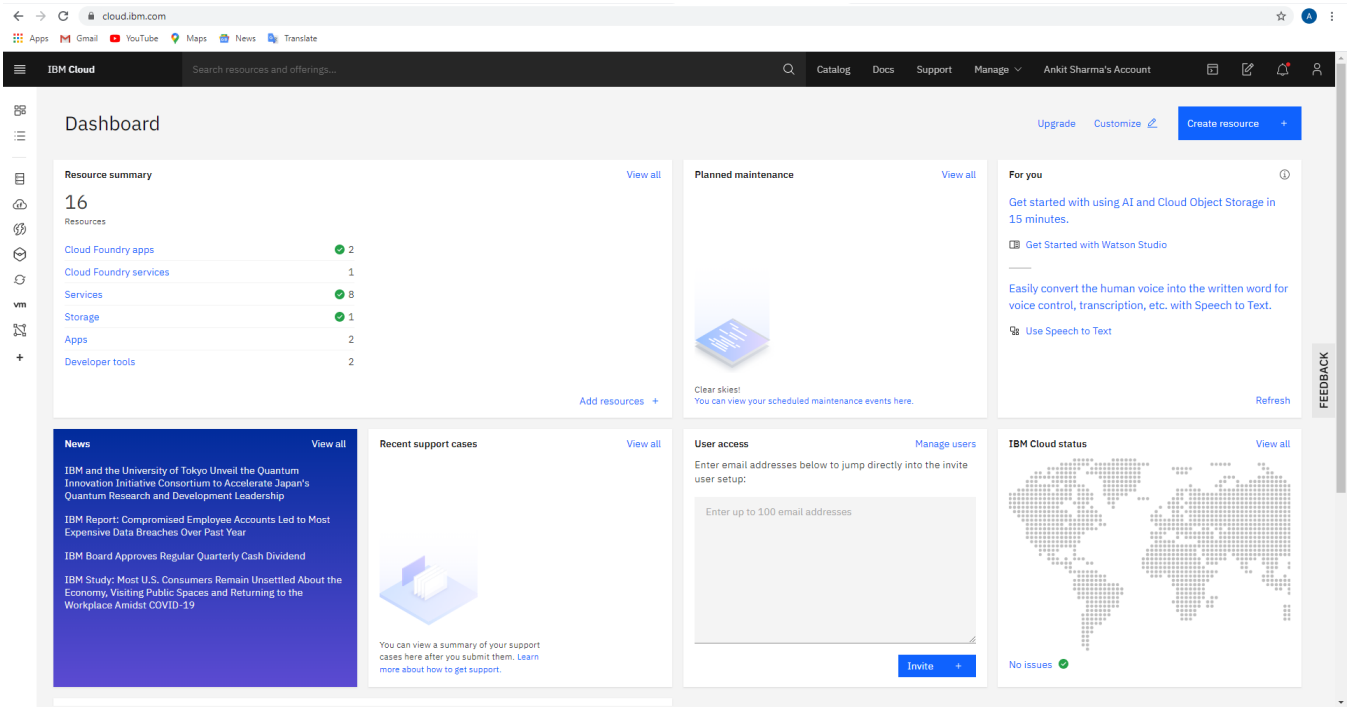
   - Click **Add to project** => **Notebook**
   - And create your Model here.

3. Deploy Model as WebService

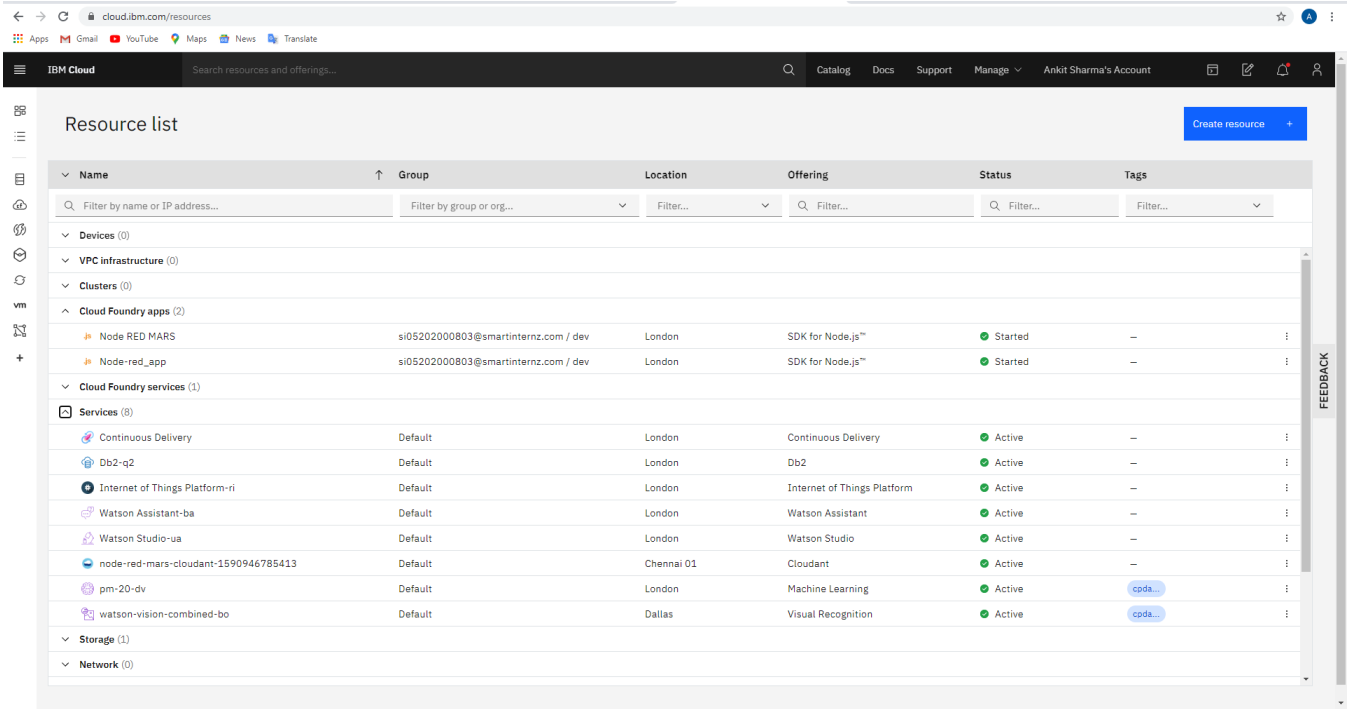
4. Build Node-RED Flow To Integrate MLServices

SCREENSHOTS :-

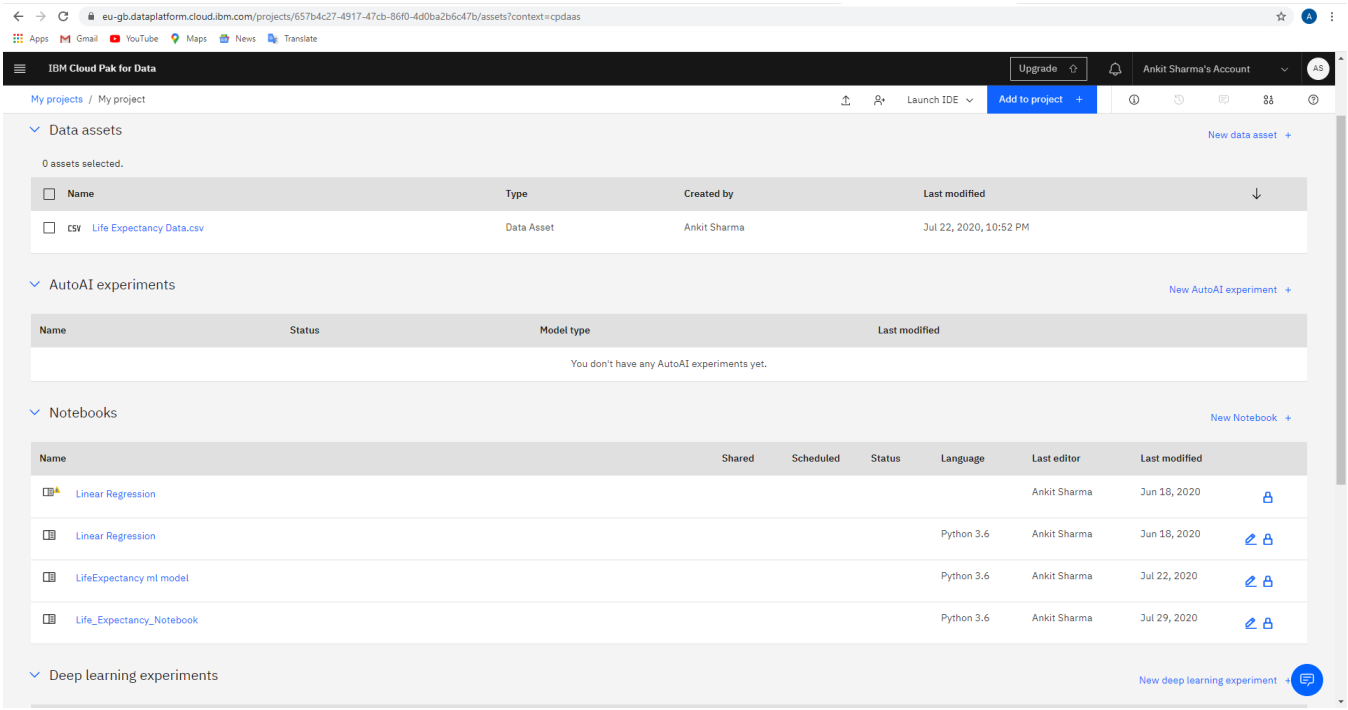
IBM CLOUD DASHBOARD



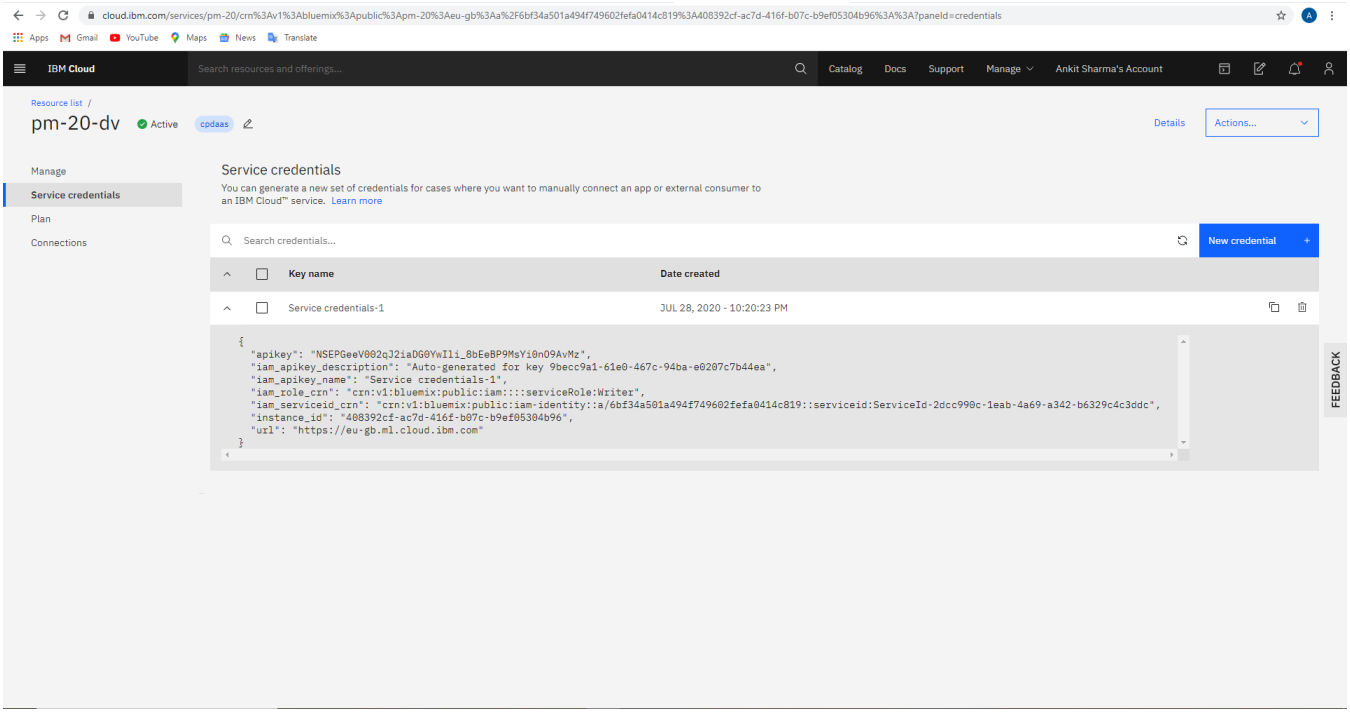
RESOURCES LIST :-



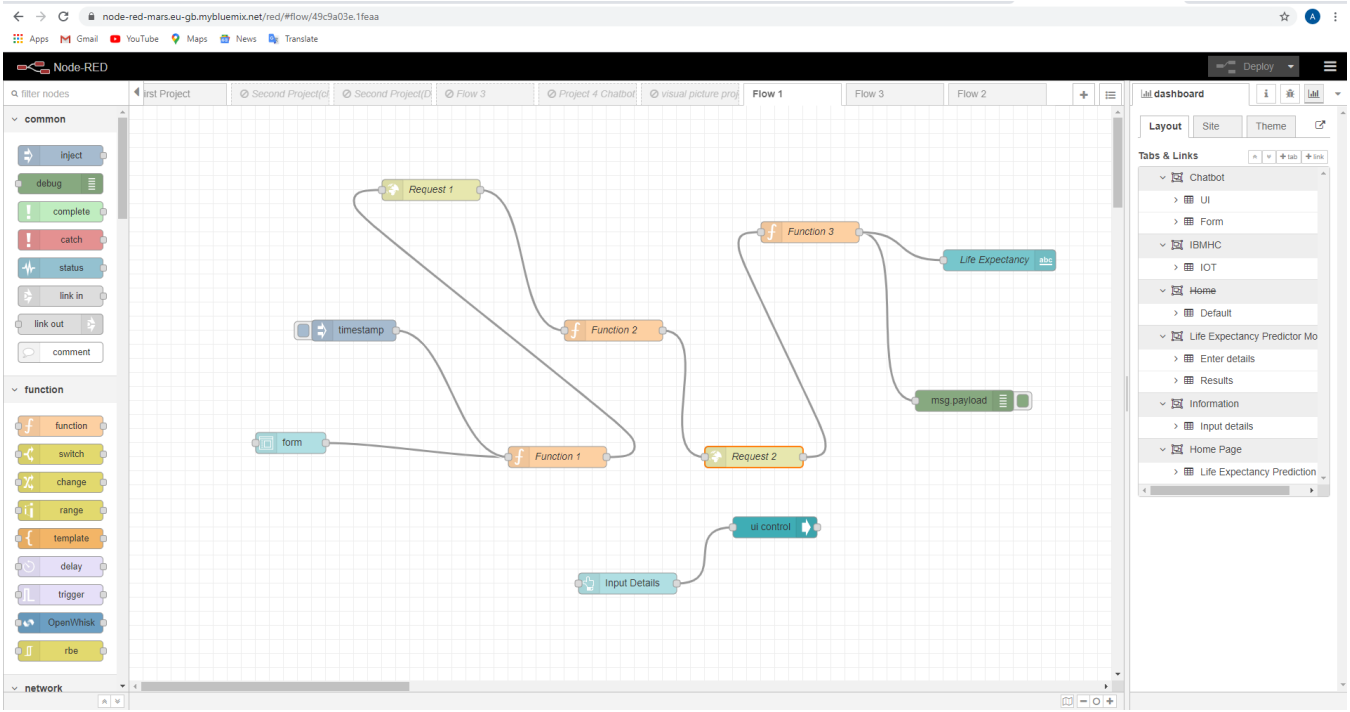
WATSON STUDIO :-



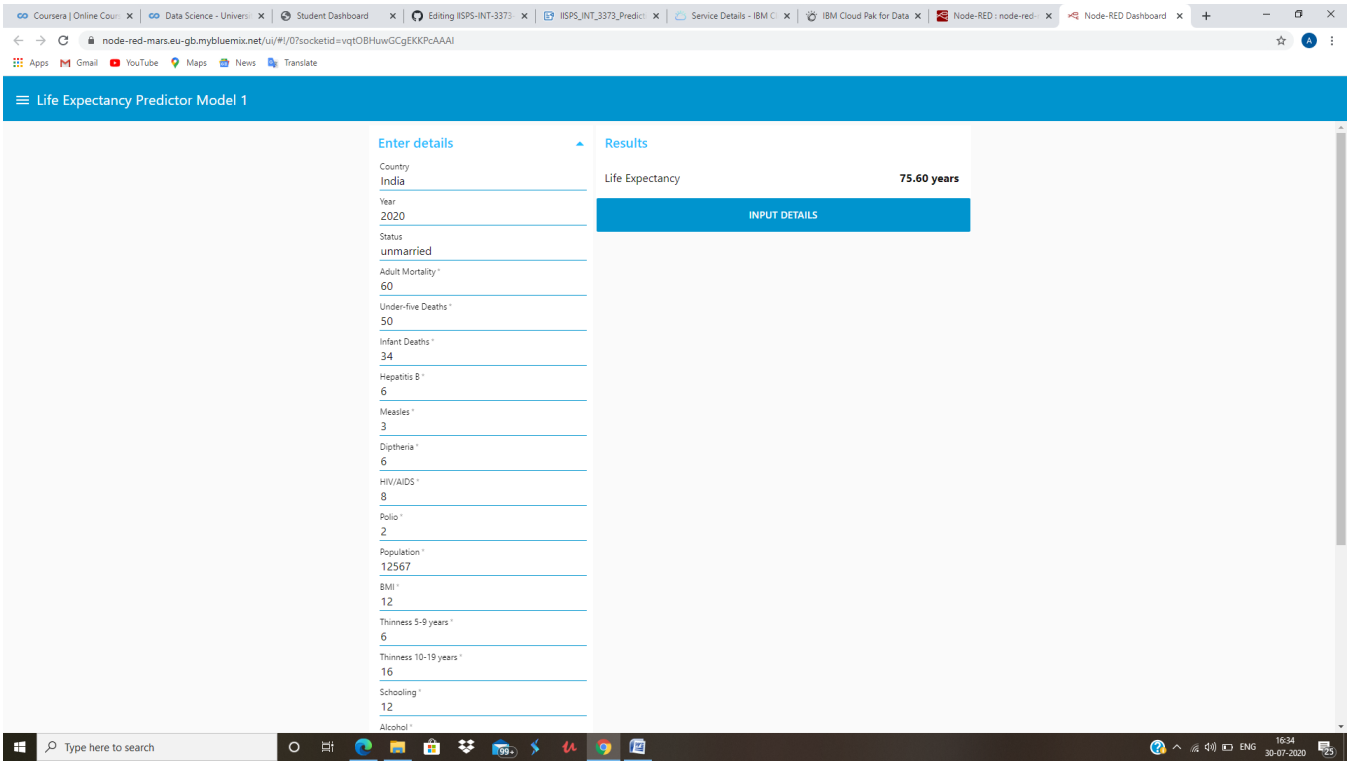
API'S OF MACHINE LEARNING SERVICE :-

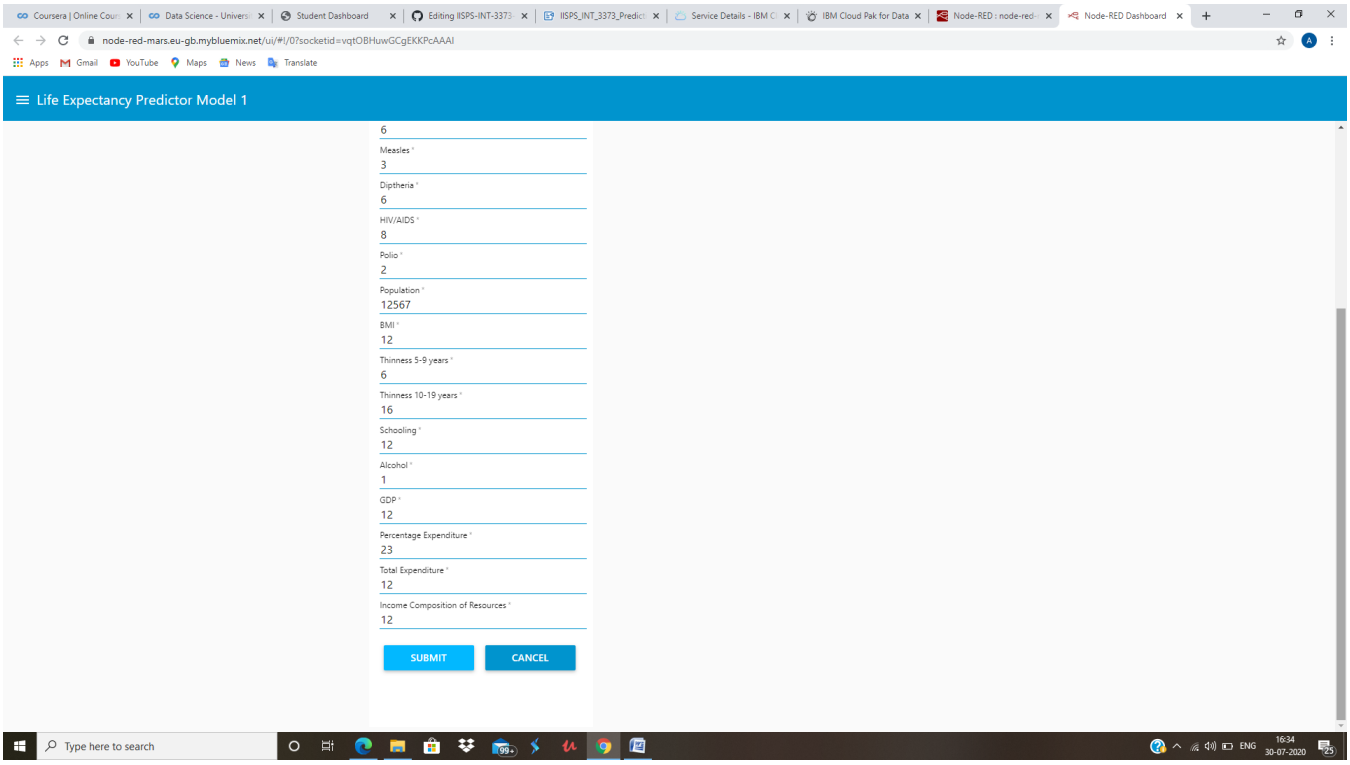


NODERED FLOW :-



OUTPUT :-




Life Expectancy Predictor Model 1
6
Measles *
3
Diptheria *
6
HIV/AIDS *
8
Polio *
2
Population *
12567
BMI *
12
Thinness 5-9 years *
6
Thinness 10-19 years *
16
Schooling *
12
Alcohol *
1
GDP *
12
Percentage Expenditure *
23
Total Expenditure *
12
Income Composition of Resources *
12
SUBMIT
CANCEL

7. ADVANTAGES AND DISADVANTAGES :-

ADVANTAGES :-

- ✓ Since the observations this dataset are based on different countries, it will be easier for acountryto determine the predicting factor which is contributing to lower value of lifeexpectancy.

- ✓ The data-sets are made available to public for the purpose of health dataanalysis.

- ✓ Can be used in any organization to analyze thedata.

- ✓ Random Forest algorithm is very stable. Even if a new data point is introduced in the dataset, the overall algorithm is not affected much since the new data may impact one tree, but it is very hard for it to impact all thetrees.

- ✓ Some of the past research was done considering multiple linear regression based on data set of one year for all the countries. But the dataset used for training the model contained data of past 15 years to give a fairly betterprediction.
- ✓ Random Forest is comparatively less impacted by noise.

DISADVANTAGES :-

1. Can be only used by the people having the knowledge of dataanalysis.
2. As the model is deployed on cloud, so one requires good internet connection to use theapplication.
3. The model used is Random Forest regression and Random Forest creates a lot of trees (unlike only one tree in case of decision tree) and combines

their outputs. By default, it creates 100 trees in Python sklearn library. To do so, this algorithm requires much more computational power andresources.

4. Random Forest require much more time to train as compared to decision trees as it generates a lot of trees (instead of one tree in case of decision tree) and makes decision on the majority ofvotes.
5. The Node-Red application needs to make HTTP request to IBM cloud and then another HTTP request to the model before providing the prediction. That makes the application a bit slow.

8. APPLICATIONS

a. This will help in suggesting a country which area should be given importance in order to efficiently improve the life expectancy of itspopulation.
b. It will be easier for a country to determine the predicting factor which is contributing to lower value of life expectancy and can be used in various organization to improve the quality ofservice.
c. The project can be used as a basis to develop personalized healthapplications.
d. The governments can plan and develop their health infrastructures by keeping the most correlated factors inmind.
e. The project can help governments to keep track of their country's health status so they can plan for the futureaccordingly

9. CONCLUSION

By doing the above procedure and all we successfully created Life expectancy prediction system using IBM Watson studio, Watson machine learning and Node-RED service. The potential use of project is not limited to health care in practice, but could also be useful in other clinical applications such as clinical trials. The project makes a good use of machine learning in predicting life expectancy of a country that can help respective government in making policies that will serve for the benefit of the nation and entire human kind.

10. FUTURE SCOPE

- Look at class within a particular country and see if these same factors are same in determining life expectancy for anindividual.
- Use the Twitter API to incorporate NLP analysis for a country to see how it relates to LifeExpectancy.
- Increase the dataset size with continuing UN and Global Data to incorporate new added features like population, GDP, environmental, and etc in order to test and clarify countrygroupings.

- Mental Health versus LifeExpectancy
- As more data comes, that can be fed to the model for more accuratepredictions.
- Currently, the project is just a web application. It can be developed to support other platforms like Android, IOS and Windows Mobile.
- Other regression models can also be used for prediction and later the best among them should bechosen

11. BIBLIOGRAPHY :-

Appendix :-

Source Code (json file)

[{"id":"49c9a03e.1feaa","type":"tab","label":"Flow 1","disabled":false,"info":""},{"id":"4e08fc99.5e9a24","type":"ui_form","z":"49c9a03e.1feaa","name":"","label":"","group":"cd23403.da5f1c","order":1,"width":0,"height":0,"options":[{"label":"Country","value":"country","type":"text","required":false,"rows":null},{"label":"Year","value":"year","type":"number","required":false,"rows":null},{"label":"Status","value":"stat","type":"text","required":false,"rows":null},{"label":"Adult Mortality","value":"am","type":"number","required":true,"rows":nul

l},{"label":"Under-five
Deaths","value":"ufd","type":"number","required":true,"rows":null},
{"label":"Infant
Deaths","value":"idr","type":"number","required":true,"rows":null},
{"label":"Hepatitis
B","value":"hepb","type":"number","required":true,"rows":null},{"l
abel":"Measles","value":"mesls","type":"number","required":true,"r
ows":null},{"label":"Diptheria","value":"dipt","type":"number","req
uired":true,"rows":null},{"label":"HIV/AIDS","value":"hiv","type":
"number","required":true,"rows":null},{"label":"Polio","value":"poli
o","type":"number","required":true,"rows":null},{"label":"Populatio
n","value":"popn","type":"number","required":true,"rows":null},{"la
bel":"BMI","value":"bmi","type":"number","required":true,"rows":n
ull},{"label":"Thinness 5-9
years","value":"thin5","type":"number","required":true,"rows":null}
,{"label":"Thinness 10-19
years","value":"thin10","type":"number","required":true,"rows":null
},{"label":"Schooling","value":"scl","type":"number","required":tru
e,"rows":null},{"label":"Alcohol","value":"alc","type":"number","re
quired":true,"rows":null},{"label":"GDP","value":"gdp","type":"nu
mber","required":true,"rows":null},{"label":"Percentage
Expenditure","value":"perexp","type":"number","required":true,"ro

ws":null},{"label":"Total Expenditure","value":"totexp","type":"number","required":true,"rows":null},{"label":"Income Composition of Resources","value":"icr","type":"number","required":true,"rows":null}],"formValue":{"country":"","year":"","stat":"","am":"","ufd":"","idr":"","hepb":"","mesls":"","dipt":"","hiv":"","polio":"","popn":"","bmi":"","thin5":"","thin10":"","scl":"","alc":"","gdp":"","perexp":"","totexp":"","icr":""},"payload":"","submit":"submit","cancel":"cancel","topic":"","x":230,"y":480,"wires":[["2564a5db.095a1a"]]},{"id":"64686136.3e70c","type":"debug","z":"49c9a03e.1feaa","name":"","active":true,"tosidebar":true,"console":false,"tostatus":false,"complete":"payload","targetType":"msg","x":1190,"y":420,"wires":[]},{"id":"7297d8b2.7c0888","type":"inject","z":"49c9a03e.1feaa","name":"","topic":"","payload":"","payloadType":"date","repeat":"","crontab":"","once":false,"onceDelay":0.1,"x":320,"y":320,"wires":[["2564a5db.095a1a"]]},{"id":"58bdcc9d.2c3f24","type":"http request","z":"49c9a03e.1feaa","name":"Request 1","method":"POST","ret":"obj","paytoqs":false,"url":"https://iam.bluemix.net/identity/token","tls":"","persist":false,"proxy":"","authType":"","x":430,"y":120,"wires":[["9b72dd63.1645b"]]},{"id":"9b72dd63.1645b","type":"function","z":"49c9a03e.1feaa","name":"Function 2","func":"var country=global.get('country');\nvar

year=global.get('year');\nvar stat=global.get('stat');\nvar am=global.get('am');\nvar ufd=global.get('ufd');\nvar idr=global.get('idr');\nvar hepb=global.get('hepb');\nvar mesls=global.get('mesls');\nvar polio=global.get('polio');\nvar dipt=global.get('dipt');\nvar hiv=global.get('hiv');\nvar popn=global.get('popn');\nvar bmi=global.get('bmi');\nvar thin5=global.get('thin5');\nvar thin10=global.get('thin10');\nvar scl=global.get('scl');\nvar alc=global.get('alc');\nvar gdp=global.get('gdp');\nvar perexp=global.get('perexp');\nvar totexp=global.get('totexp');\nvar icr=global.get('icr');\nvar token=msg.payload.access_token;\nvar instance_id=\"856c1e00-49d9-49ed-93ad-dcfcffe1bf41\"\nmsg.headers = {\n \"Content-type\":'application/json',\n \"Authorization\": 'Bearer'+token,\n \"ML-Instance-ID\": instance_id\n}\nmsg.payload = {\n \"input_data\": [{\"fields\": [\"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\"], \n \"values\": [[country, year, stat, am, idr, alc, perexp, hepb, mesls, bmi, ufd,

polio, totexp, dipt, hiv, gdp, popn, thin10, thin5, icr, scl]]}]\n}\nreturn msg;","outputs":1,"noerr":0,"x":690,"y":320,"wires":[["fbd10670.411b88"]]},{"id":"fbd10670.411b88","type":"http request","z":"49c9a03e.1feaa","name":"Request 2","method":"POST","ret":"obj","paytoqs":false,"url":"https://eu-gb.ml.cloud.ibm.com/v4/deployments/d27025bf-742f-4689-b605-8ee85ff39710/predictions","tls":"","persist":false,"proxy":"","authType":"","x":890,"y":500,"wires":[["93fbed4.2f6b91"]]},{"id":"93fbed4.2f6b91","type":"function","z":"49c9a03e.1feaa","name":"Function 3","func":"msg.payload = msg.payload.predictions[0].values[0][0].toFixed(2)+' years';\nreturn msg;","outputs":1,"noerr":0,"x":970,"y":180,"wires":[["64686136.3e70c","c14c2067.84774"]]},{"id":"c14c2067.84774","type":"ui_text","z":"49c9a03e.1feaa","group":"b4dfeff9.170bf","order":1,"width":0,"height":0,"name":"Life Expectancy","label":"Life Expectancy","format":"{{msg.payload}}","layout":"row-spread","x":1240,"y":220,"wires":[]},{"id":"2564a5db.095a1a","type":"function","z":"49c9a03e.1feaa","name":"Function 1","func":"global.set('country', msg.payload.country);\nglobal.set('year', msg.payload.year);\nglobal.set('status',

msg.payload.stat);\nglobal.set('am',
msg.payload.am);\nglobal.set('ufd',
msg.payload.ufd);\nglobal.set('idr',
msg.payload.idr);\nglobal.set('hepb',
msg.payload.hepb);\nglobal.set('mesls',
msg.payload.mesls);\nglobal.set('polio',
msg.payload.polio);\nglobal.set('dipt',
msg.payload.dipt);\nglobal.set('hiv',
msg.payload.hiv);\nglobal.set('popn',
msg.payload.popn);\nglobal.set('bmi',
msg.payload.bmi);\nglobal.set('thin5',
msg.payload.thin5);\nglobal.set('thin10',
msg.payload.thin10);\nglobal.set('scl',
msg.payload.scl);\nglobal.set('alc',
msg.payload.alc);\nglobal.set('gdp',
msg.payload.gdp);\nglobal.set('perexp',
msg.payload.perexp);\nglobal.set('totexp',
msg.payload.totexp);\nglobal.set('icr', msg.payload.icr);\nvar apikey
=
'LCwNeMqOOHVXqE8D-rV2gdfMZ0lfxB3H0DZjeuHGbj3v';\nms
g.headers={'content-type':'application/x-www-form-urlencoded',
'Accept':'application/json'};\nmsg.payload={'grant_type':'urn:ibm:pa

rams:oauth:grant-type:apikey', 'apikey':apikey};\nreturn msg;","outputs":1,"noerr":0,"x":610,"y":500,"wires":[["58bdcc9d.2c3f24"]]},{"id":"bd7f5e0.f22cfa","type":"ui_button","z":"49c9a03e.1feaa","name":"","group":"b4dfeff9.170bf","order":3,"width":0,"height":0,"passthru":false,"label":"Input Details","tooltip":"","color":"","bgcolor":"","icon":"","payload":"Information","payloadType":"str","topic":"","x":710,"y":680,"wires":[["80f7f35d.08701"]]},{"id":"80f7f35d.08701","type":"ui_ui_control","z":"49c9a03e.1feaa","name":"","events":"change","x":920,"y":600,"wires":[[]]},{"id":"cd23403.da5f1c","type":"ui_group","z":"","name":"Enter details","tab":"b2ef227f.2b617","order":1,"disp":true,"width":"6","collapse":true},{"id":"b4dfeff9.170bf","type":"ui_group","z":"","name":"Results","tab":"b2ef227f.2b617","order":2,"disp":true,"width":"10","collapse":false},{"id":"b2ef227f.2b617","type":"ui_tab","z":"","name":"Life Expectancy Predictor Model 1","icon":"dashboard","order":1,"disabled":false,"hidden":false}]

Nodered flow :-

https://node-red-mars.eu-gb.mybluemix.net/red/#flow/49c9a03e.1feaa

Github link :-

https://github.com/SmartPracticeschool/llSPS-INT-3373-Predicting-Life-Expectancy-using-Machine-Learning

Youtube video link :-

https://youtu.be/8A3zkhWLcX0