PROJECT REPORT

PREDICTING LIFE-EXPECTANCY

USING

MACHINE LEARNING

BY:

J.TEJASRI

PROJECT PLATFORM: SMARTBRIDGE

PROJECT ID:SPS_PRO_215

CONTENTS

1.INTRODUCTION

1.1.Overview:

Life expectancy refers to the number of years a person is expected to live. In mathematical terms, life expectancy refers to the expected number of years remaining for an individual at any given age. Based on actuarial science, the life expectancy for a particular person or a population group depends on several individual-level as well as population-level factors such as a persons lifestyle, historical mortality data of country, etc. However, as life expectancy is calculated based on averages, a person may live for many years more or less than expected.

To predict the life expectancy rate of a particular country, we will be using machine learning to draw inferences from the given data set and give a prediction. We will also be creating a UI using Node-RED for making the model accessible to general users.

1.2.Purpose:

1 .Predicting life expectancy plays a vital role in judging the growth and development of the economy. Across countries, high life expectancy is associated with high income per capita. Increase in life expectancy also leads to an increase in the “manpower” of a country.

2. Helps the government bodies take appropriate measures to control the population growth and also direct the utilization of the increase in human resources and skill set acquired by people over many years.

3.This project would also help an individual assess his/her lifestyle choices and alter them accordingly to lead a longer and healthier life. It would make them more aware of their general health and its improvement or deterioration over time.

4.Based on the factors used to calculate life expectancy of an individual and the outcome, health care will be able to fund and provide better services to those with greater need.

5.Insurance sector will be able to provide individualized services to people based on the life expectancy outcomes and factors.

6.To inspect and help properly understand the underlying causes of variations in average Life expectancy of Humans.

7.To concentrate on factors leading to decline in life expectancy and thereby efficiently come up with a solution to negate the cause.

2. LITERATURE SURVEY

2.1 Existing problem:

There have been reports of Declining lifespan of humans all over the world over the past few years. Causes of this declining numbers are due to various new diseases and viruses that get introduced as time goes by. Many have lost lives and have gone far too early before they get a chance to experience the world. Much of this is believed to be our own doing. Humans tend to destroy habitats for their own survival and forgetting that they are actually would have to fight for their own as they continue to do so. It is important that we are able to, in a way, be able to peek into our own future to help prepare us for what is to come.

In the context of life expectancy, we are often devoid of ideas of where to concentrate resources on to increase the average life expectancy of a community. this is because in the modern world there are so many other factors other than medical factors that our lifespan depends on like our Lifestyle, our choice of food, etc. Thus it is necessary to find the right areas to pour resources and spent time working on so that the results may improve in our favor.

2.2 Proposed Solution:

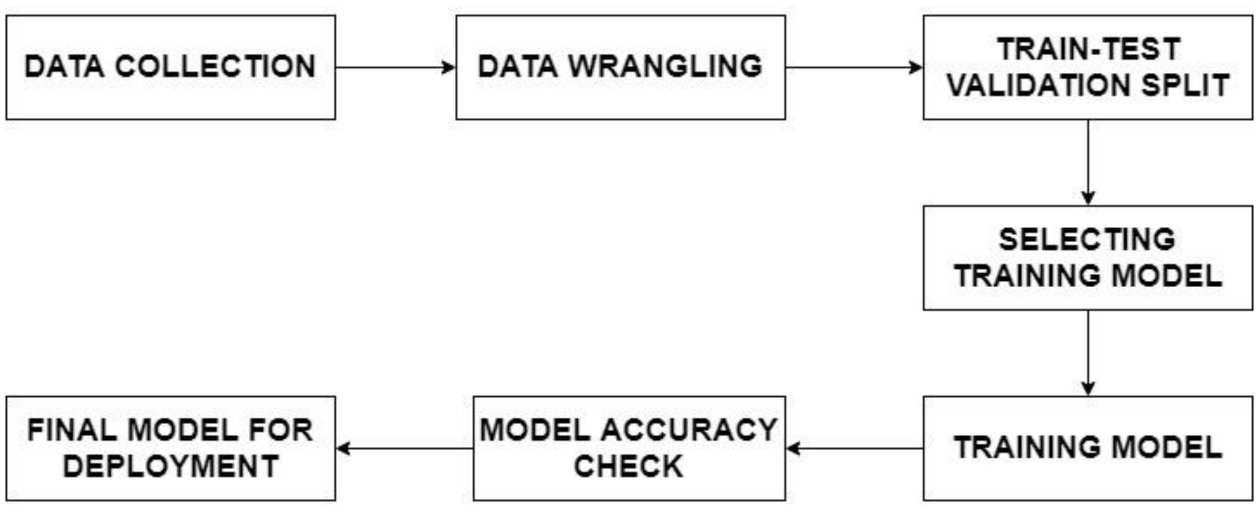
Machine learning, natural language processing, and data mining in general have grown to be increasingly popular methods for processing data within the medical domain. Given examples, machine learning algorithms can be trained to learn which pieces of information are important to execute a task, and which patterns are indicative for producing correct output. Machine learning and language processing techniques have been applied to a broad range of tasks, including medical decision support and decision making, automatic disease detection, automatic diagnostication, identifying the role of genes in the onset of diseases, adverse event detection, identifying interactions between drugs and side-effects of drugs, and phenotyping.

I will be building a solution using Machine Learning to predict the life expectancy given a set of inputs. Specifically, linear regression model will be used. We will train the model on a given WHO dataset with real values from the past. It will contain some information like the status of the country, adult mortality rate, infant deaths, alcohol,hepatitis B, measles, BMI, etc. This trained model will then be able to give a prediction for given input data points. Finally, this model will be deployed and made accessible to users using an interactive UI generated using the Node-RED service.

.

3.THEORETICAL ANALYSIS

3.1.Block Diagram:



The above block diagram summarizes the entire process in a simple manner. The blocks include:

1. **DATA COLLECTION:** For training any ML model, the immediate primary need is defining a textbook or syllabus based on which the model can be trained. In the world of ML, this syllabus refers to data. Data can be used to give the model some idea as to how to approach a problem statement presented to it. The data will act as the reference to the model for finding the optimum solution. Since data is the foundation of ML, it is absolutely necessary to ensure that the data collection process is accurate. The output predictions made by the model will only be as accurate as the data it is provided with.
2. **DATA WRANGLING:** The data collected might be huge depending on the its purpose. Often a lot of junk or rather un useful data which does not affect the prediction data might be tangled inside. This is where data scientists come into play. The wrangle the data, studying and understanding each of the factors in it and the relation between each of the factors so as to remove any junk data. This process will be helpful in reducing the time taken for training the model in next stages greatly.
3. **TRAIN TEST VALIDATION SPLIT:** After data is cleaned, it is necessary to consider the accuracy of whatever model we intend to build. An efficient way of finding this is to split the data we collected (after wrangling) into 2 parts; the training set and test set. The training set is dedicated to the training of the model and test set can be used to test the trained model to check accuracy manually since the output of the test set is already known to us (test set was derived from the initial data).often more of the data set is dedicated for the training set.
4. **SELECTING TRAINING MODEL:** This step can be done before or after train test split. In machine Learning there are many models (eg: decision tree,linear regression etc.) which can be used to train the data. Each model can have variations in output from one another. The simplest would be the linear regression model which was one of the earliest models. The models can be seen as different styles of studying

the data. Each style hence can have various results similar to different children who have different ways of learning any subject.

5. **TRAINING MODEL:** After the model is selected, we can proceed to start teaching the machine how to approach problems in the real world.
6. **MODEL ACCURACY CHECK:** After each academic year, every school has exams for the students to test whether they have a strong understanding of what was taught to them. This helps to evaluate each child's progress and find their weak points. Similar exams can be prepared for the machine to check whether they have been able to learn and understand what was taught to them. The test set is used here to test what the machine would predict based on the knowledge it gained. Based on similarity of actual and predicted values, the machine is graded.
7. **FINAL MODEL FOR DEPLOYMENT:** After the model is deemed fit for real world use, the appropriate steps can be taken for deploying the model to the real world.

3.2.SOFTWARE DESIGNING:

IBM cloud-

The IBM cloud platform combines platform as a service (PaaS) with infrastructure as a service (IaaS) to provide an integrated experience. The platform scales and supports both small development teams and organizations, and large enterprise businesses. Globally deployed across data centers around the world, the solution you build on IBM Cloud spins up fast and performs reliably in a tested and supported environment you can trust.

The platform is built to support your needs whether it's working only in the public cloud or taking advantage of a multi-cloud deployment model. With our open-source technologies, such as Kubernetes, Red Hat Open Shift, and a full range of compute options, including virtual machines, containers, bare metal, and server less, you have as much control and flexibility as you need to support workloads in your hybrid environment. You can deploy cloud-native apps while also ensuring workload portability.

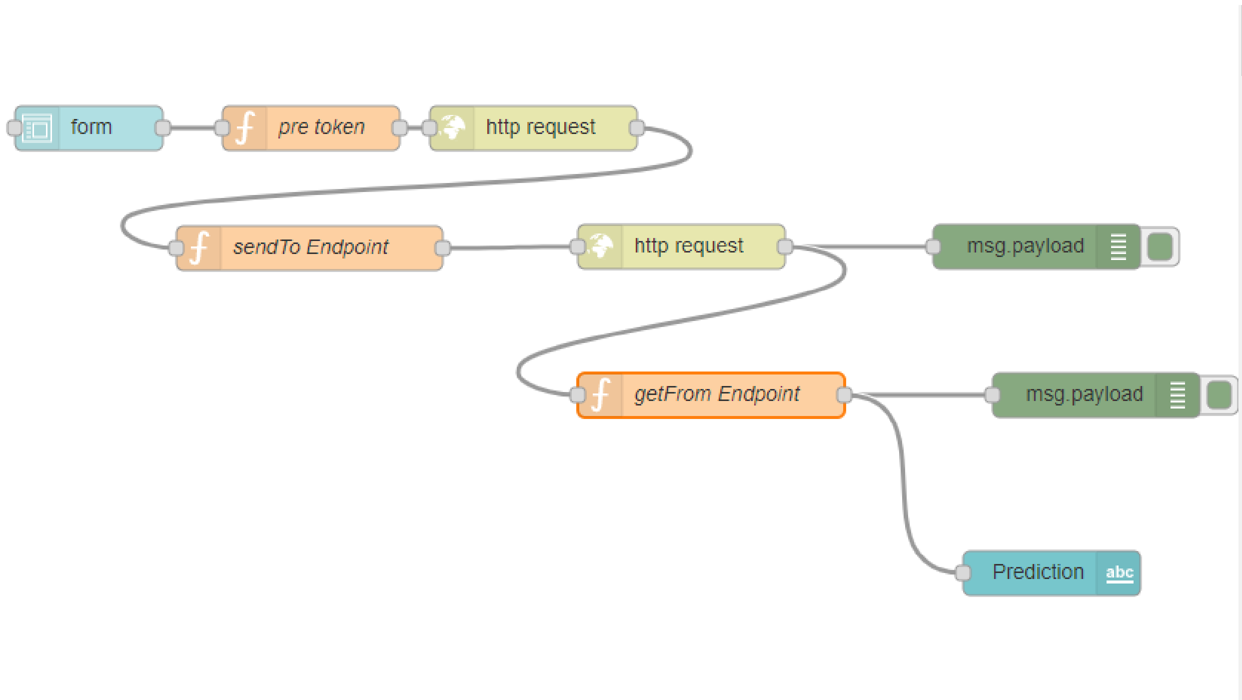
Whether you need to migrate apps to the cloud, modernize your existing apps by using cloud services, ensure data resiliency against regional failure, or leverage new paradigms and deployment topologies to innovate and build your cloud-native apps, the platform's open architecture is built to accommodate your use case.

Node Red:

Node-RED is a programming tool for wiring together hardware devices, APIs and online services in new and interesting ways.It provides a browser-based editor that makes it easy to wire together flows using the wide range of nodes in the palette that can be deployed to its run time in a single-click.

Node-RED provides a browser-based flow editor that makes it easy to wire together flows using the wide range of nodes in the palette. Flows can be then deployed to the run time in a single-click.JavaScript functions can be created within the editor using a rich text editor.A built-in library allows you to save useful functions, templates or flows for re-use.

The light-weight run time is built on Node.js, taking full advantage of its event-driven, non-blocking model. This makes it ideal to run at the edge of the network on low-cost hardware as well as in the cloud. it is easy to extend the range of palette nodes to add new capabilities.The flows created in Node-RED are stored using JSON which can be easily imported and exported for sharing with others.

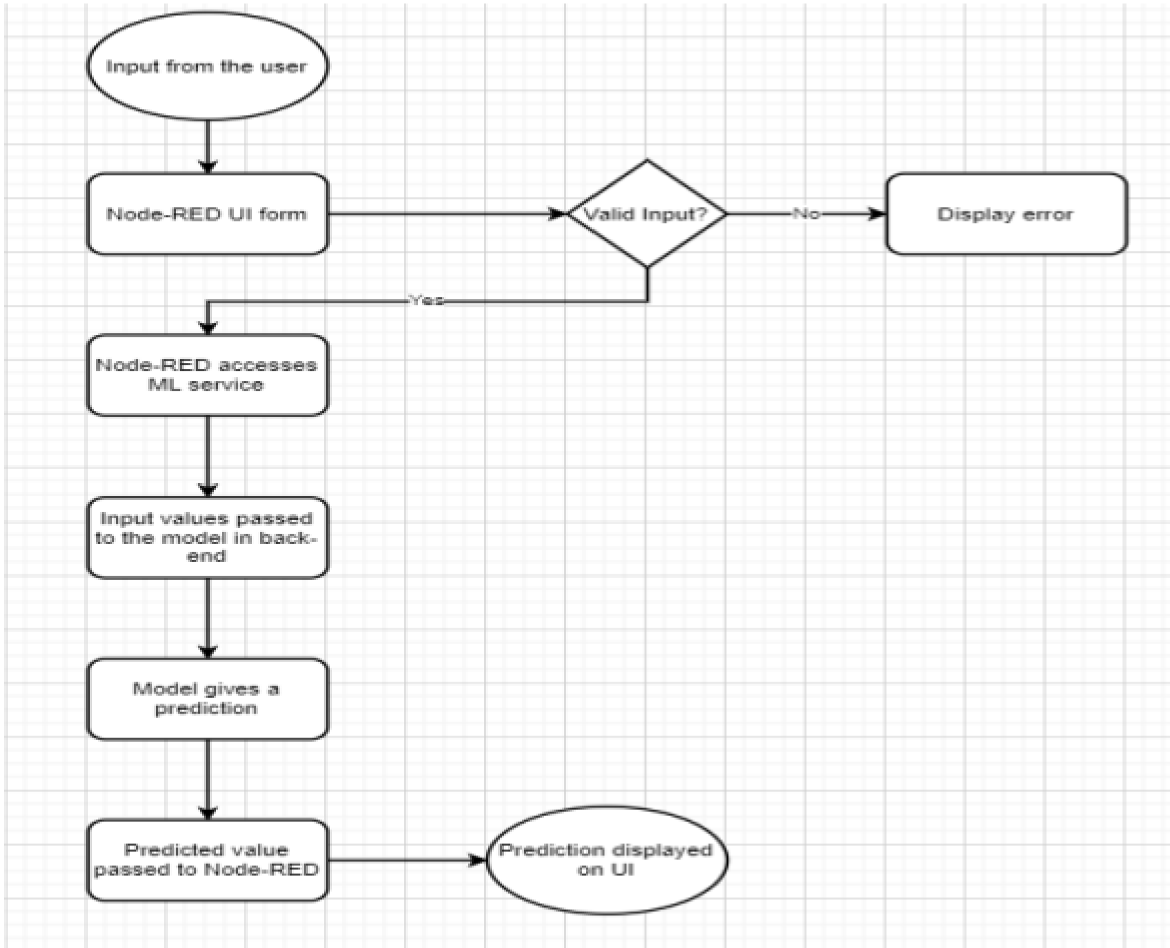


The above figure shows the flow we use in this project. The flow shows the order of the steps taken to get the final output. Each node has different properties. One of the important ones here are the function nodes ‘Pre Token’ and ‘sendTO Endpoint’. These nodes, as the name suggests can be coded to execute a particular function. The code for the entire flow is given in the appendix part of this report at the end. The source code can be imported and can be pasted directly to get the above flow._We created a UI to interact with the user who can input the data he has with him and receive an output through IBM cloud. There is also no issue of creating and hosting a website in the internet as the UI links the cloud directly with the user to transmit and receive data (input or output).

4.EXPERIMENTAL INVESTIGATIONS

The WHO data set for life expectancy contains 22 columns in total. The data set has historical data of life expectancy for the period between 2000 to 2015. There are numerous rows with null values for some of the columns and outliers as well. These null values will be replaced with the mean for the respective columns and the outliers will be adjusted. All the columns except the "Country" and "Status" are of integer type. The country column is not considered into the model training because of its less relation to the "Life Expectancy" column. The "Status" column is changed to integer type such that "Developing" is mapped to 1 and "Developed" is mapped to 0. After these steps, the data is ready for the model to be trained.

5.FLOW CHART



6.RESULT

A machine learning model developed using Linear Regression and IBM Cloud and its services. The model has a high accuracy.

An interactive UI deployed using the Node-RED service for the user to interact with the model in the back-end and get predictions for a given set of input values.

7.ADVANTAGES AND DISADVANTAGES

ADVANTAGES

1. Easily identifies trends and patterns-

Machine Learning can review large volumes of data and discover specific trends and patterns that would not be apparent to humans.

2. No human intervention needed -

The biggest advantage of using Machine Learning for predictions is that it learns by itself. It could evolve over time as more data becomes available. The algorithm analyzes the trends in the data and adjusts itself to increase the accuracy.

3. Continuous Improvement-

As ML algorithms gain experience, they keep improving in accuracy and efficiency. This lets them make better decisions. Say you need to make a weather forecast model. As the amount of data you have keeps growing, your algorithms learn to make more accurate predictions faster.

4. Handling multidimensional and multi-variety data-

Machine Learning algorithms are good at handling data that are multidimensional and multi-variety, and they can do this in dynamic or uncertain environments.

DISADVANTAGES

With all those advantages to its powerfulness and popularity, Machine Learning isn't perfect. The following factors serve to limit it:

1. Data Acquisition:

Machine Learning requires massive data sets to train on, and these should be inclusive/unbiased, and of good quality. There can also be times where they must wait for new data to be generated.

2. Time and Resources:

ML needs enough time to let the algorithms learn and develop enough to fulfill their purpose with a considerable amount of accuracy and relevancy. It also needs massive resources to function. This can mean additional requirements of computer power for you.

3. Interpretation of Results:

Another major challenge is the ability to accurately interpret results generated by the algorithms. You must also carefully choose the algorithms for your purpose.

4. High error-susceptibility:

Machine Learning is autonomous but highly susceptible to errors. Suppose you train an algorithm with data sets small enough to not be inclusive. You end up with biased predictions coming from a biased training set. In the case of ML, such blunders can set off a chain of errors that can go undetected for long periods of time. And when they do get noticed, it takes quite some time to recognize the source of the issue, and even longer to correct it.

8.APPLICATIONS

1. Useful in the medical field for studying and isolating causes of deviation in human life expectancy over an observed period of time.
2. Supporting timely recognition of the right moment to start Advance Care Planning.

3.Individuals can predict their own life expectancy by inputting values in the corresponding fields. This could help make people more aware of their general health, and its improvement or deterioration over time. This may motivate them to make healthier lifestyle choices.

4.It could help the government bodies take appropriate measures to control the population growth and also direct the utilization of the increase in human resources and skill set acquired by people over many years. Across countries, high life expectancy is associated with high income per capita. Increase in life expectancy also leads to an increase in the “manpower” of a country. The knowledge asset of a country increases with the number of individuals in a country.

5.Based on the factors used to calculate life expectancy of an individual and the outcome, health care will be able to fund and provide better services to those with greater need.

6.Insurance sector will be able to provide individualized services to people based on the life expectancy outcomes and factors.

9.CONCLUSION

Prognostication of life expectancy is difficult for humans. Our research shows that machine learning techniques offer a feasible and promising approach to predicting life expectancy. The research has potential for real-life applications, such as supporting timely recognition of the right moment to start Advance Care Planning.

Good prognostication has the potential to contribute significantly to end-of-life decision making, therefore we believe that any increase in prognostic accuracy is worth pursuing. Additionally, human prognostication is costly, time-consuming, requires medical expertise, and is a subjective task. Without compromising prediction accuracy, the model is able to make predictions quickly, automatically and systematically, while it does not depend on human medical expertise.

Life Expectancy prediction can play a vital role in numerous sectors of the industry and machine learning can be used to achieve this with a high level of accuracy.

10.FUTURE SCOPE

The application can be further improved by integrating it with the functionality to provide suggestions and medications to the individuals using the application. This will help predict as well as increase the individual's life expectancy. The Scalability and flexibility of the application can also be improved with advancement in technology and availability of new and improved resources. With the growth in Artificial Intelligence and Computer vision, we can also try to take into account the physical health and appearance of a person for predicting the life expectancy. Mental health can also be considered as a major factor while predicting life expectancy with the help of sentiment analysis systems.

This research should be considered to be exploratory. In order to replicate and extend this research, we can expand the data set substantially, by collecting additional data of both deceased and active patients. This will allow us to zoom in on specific illness trajectories, and to rephrase the task in such a way that it will match clinical settings more closely, for example by aiming to make predictions about patients while they are still active.

We will be able to compare a range of predictive models, alternative patient representations, and (interpretations of) output variables in future work. To provide a better comparison between automatic and human prognostication, we can investigate the prediction accuracy of both the system and general practitioners by presenting them with the same task and test data.

Additionally, we can work towards obtaining insight about the driving forces behind good prognostication. We can explore which information is used by the model, to make the model for automatic prognostication more transparent, and improve our understanding of this complex task.

11.BIBLIOGRAPHY

APPENDIX

A. NODE RED FLOW CODE -

```
[
```

{

"id": "4eea5586.897cec",

"type": "tab",

"label": "Flow 1",

"disabled": false,

"info": ""

},

{

"id": "589463fd.46e5bc",

"type": "ui_form",

"z": "4eea5586.897cec",

"name": "",

"label": "",

"group": "71318a64.ec50b4",

"order": 0,

"width": 0,

"height": 0,

"options": [

{

"label": "Year",

"value": "a",

"type": "number",

"required": true,

"rows": null

},

{

"label": "BMI ",

"value": "b",

"type": "number",

"required": true,

"rows": null

},

{

"label": "Adult_Mortality",

"value": "c",

"type": "number",

"required": true,

"rows": null

},

{

"label": "Infant_Deaths",

"value": "d",

"type": "number",

"required": true,

"rows": null

},

{

"label": "Alcohol",

"value": "e",

"type": "number",

"required": true,

"rows": null

},

{

"label": "Percentage_Expenditure",

"value": "f",

"type": "number",

"required": true,

"rows": null

},

{

"label": "HepatitisB",

"value": "g",

"type": "number",

"required": true,

"rows": null

},

{

"label": "Under_Five_Deaths",

"value": "h",

"type": "number",

"required": true,

"rows": null

},

{

"label": "Polio",

"value": "i",

"type": "number",

"required": true,

"rows": null

},

{

"label": "Total_Expenditure",

"value": "j",

"type": "number",

"required": true,

"rows": null

},

{

"label": "Diphtheria",

"value": "k",

"type": "number",

"required": true,

"rows": null

},

{

"label": "HIV",

"value": "l",

"type": "number",

"required": true,

"rows": null

},

{

"label": "GDP",

"value": "m",

"type": "number",

"required": true,

"rows": null

},

{

"label": "Population",

"value": "n",

"type": "number",

"required": true,

"rows": null

},

{

"label": "thinness_1to19_years",

"value": "o",

"type": "number",

"required": true,

"rows": null

},

{

"label": "thinness_5to9_years",

"value": "p",

"type": "number",

"required": true,

"rows": null

},

{

"label": "Income_Comp_Of_Resources",

"value": "q",

"type": "number",

"required": true,

"rows": null

},

{

"label": "Schooling",

"value": "r",

"type": "number",

"required": true,

"rows": null

},

{

"label": "Measles",

"value": "s",

```
"type": "number",
"required": true,
"rows": null
},
{
"label": "Developing(1 or 0)",
"value": "t",
"type": "number",
"required": true,
"rows": null
}
],
"formValue": {
"a": "",
"b": "",
"c": "",
"d": "",
"e": "",
"f": "",
"g": "",
"h": "",
```

"i": "",

"j": "",

"k": "",

"l": "",

"m": "",

"n": "",

"o": "",

"p": "",

"q": "",

"r": "",

"s": "",

"t": ""

},

"payload": "",

"submit": "submit",

"cancel": "cancel",

"topic": "",

"x": 70,

"y": 100,

"wires": [

[

"a3297459.0b8638"

]

]

},

{

"id": "a3297459.0b8638",

"type": "function",

"z": "4eea5586.897cec",

"name": "pre token",

"func": "//make user given values as global
variables\nglobal.set(\"a\",msg.payload.a);\nglobal.set(\"b\",msg.payload.b);\nglobal.s
et(\"c\",msg.payload.c);\nglobal.set(\"d\",msg.payload.d);\nglobal.set(\"e\",msg.payloa
d.e);\nglobal.set(\"f\",msg.payload.f);\nglobal.set(\"g\",msg.payload.g);\nglobal.set(\
"h\",msg.payload.h);\nglobal.set(\"i\",msg.payload.i);\nglobal.set(\"j\",msg.payload.j)
;\nglobal.set(\"k\",msg.payload.k);\nglobal.set(\"l\",msg.payload.l);\nglobal.set(\"m\"
,msg.payload.m);\nglobal.set(\"n\",msg.payload.n);\nglobal.set(\"o\",msg.payload.o);\ng
lobal.set(\"p\",msg.payload.p);\nglobal.set(\"q\",msg.payload.q);\nglobal.set(\"r\",msg
.payload.r);\nglobal.set(\"s\",msg.payload.s);\nglobal.set(\"t\",msg.payload.t);\n\n//f
ollowing are required to receive a token\nvar
apikey=\"W1lNn30RFFcSXiKcgvhS7Yulm0_KrArTky_jVsqztgzX\";\nmsg.headers={\"content-
type\":\"application/x-www-form-
urlencoded\"};\nmsg.payload={\"grant_type\":\"urn:ibm:params:oauth:grant-
type:apikey\",\"apikey\":apikey};\nreturn msg;\n",

"outputs": 1,

"noerr": 0,

"x": 220,

"y": 100,

"wires": [

[

"44b7b843.f7da68"

]

]

},

{

"id": "b82dfa87.68f3a8",

"type": "http request",

"z": "4eea5586.897cec",

"name": "",

"method": "POST",

"ret": "obj",

"paytoqs": false,

"url": "https://eu-gb.ml.cloud.ibm.com/v3/wml_instances/10785a24-3689-4228-8f08-3dd21ff563a5/deployments/dc403feb-b8ec-4017-9139-3fe602151210/online",

"tls": "",

"persist": false,

"proxy": "",

"authType": "",

"x": 470,

"y": 180,

"wires": [

[

"734403c2.36572c",

"65a92058.150b5"

]

]

},

{

"id": "f639a7da.840fa8",

"type": "debug",

"z": "4eea5586.897cec",

"name": "",

"active": true,

"tosidebar": true,

"console": false,

"tostatus": false,

"complete": "payload",

"targetType": "msg",

"x": 750,

"y": 280,

"wires": []

```
},
{
"id": "65a92058.150b5",
"type": "function",
"z": "4eea5586.897cec",
"name": "getFrom Endpoint",
"func": "msg.payload=msg.payload.values[0][0];\nreturn msg;",
"outputs": 1,
"noerr": 0,
"x": 490,
"y": 280,
"wires": [
[
"f639a7da.840fa8",
"79438857.ba6f48"
]
]
},
{
"id": "734403c2.36572c",
"type": "debug",
```

"z": "4eea5586.897cec",

"name": "",

"active": true,

"tosidebar": true,

"console": false,

"tostatus": false,

"complete": "payload",

"targetType": "msg",

"x": 710,

"y": 180,

"wires": []

},

{

"id": "166dc001.eb643",

"type": "function",

"z": "4eea5586.897cec",

"name": "sendTo Endpoint",

"func": "//get token and make headers\nvar token=msg.payload.access_token;\nvar instance_id=\"10785a24-3689-4228-8f08-3dd21ff563a5\"\nmsg.headers={'Content-Type': 'application/json',\"Authorization\":\"Bearer \"+token,\"ML-Instance-ID\":instance_id}\n\n//get variables that are set earlier\nvar a = global.get(\"a\");\nvar b = global.get(\"b\");\nvar c = global.get(\"c\");\nvar d = global.get(\"d\");\nvar e = global.get(\"e\");\nvar f = global.get(\"f\");\nvar g = global.get(\"g\");\nvar h = global.get(\"h\");\nvar i = global.get(\"i\");\nvar j = global.get(\"j\");\nvar k = global.get(\"k\");\nvar l = global.get(\"l\");\nvar m = global.get(\"m\");\nvar n = global.get(\"n\");\nvar o = global.get(\"o\");\nvar p =

global.get(\"p\");\nvar q = global.get(\"q\");\nvar r = global.get(\"r\");\nvar s = global.get(\"s\");\nvar t = global.get(\"t\");\n\n\n//send the user values to service endpoint\nmsg.payload = \n{\"fields\":[\"Year\",\"BMI\",\n\"Adult_Mortality\",\n\"Infant_Deaths\",\n\"Alcohol\",\"Percentage_Expenditure\",\"HepatitisB\",\"Under_Five_Deaths\",\"Polio\",\"Total_Expenditure\",\"Diphtheria\",\"HIV\",\"GDP\",\"Population\",\"thinness_1to19_years\",\"thinness_5to9_years\",\"Income_Comp_Of_Resources\",\"Schooling\",\"Measles\",\"Developing\"],\n\"values\":[[a,b,c,d,e,f,g,h,i,j,k,l,m,n,o,p,q,r,s,t]]};\n\nreturn msg;\n",

"outputs": 1,

"noerr": 0,

"x": 210,

"y": 180,

"wires": [

[

"b82dfa87.68f3a8"

]

]

},

{

"id": "44b7b843.f7da68",

"type": "http request",

"z": "4eea5586.897cec",

"name": "",

"method": "POST",

"ret": "obj",

"paytoqs": false,

"url": "https://iam.cloud.ibm.com/identity/token",

"tls": "",

"persist": false,

"proxy": "",

"authType": "basic",

"x": 370,

"y": 100,

"wires": [

[

"166dc001.eb643"

]

]

},

{

"id": "79438857.ba6f48",

"type": "ui_text",

"z": "4eea5586.897cec",

"group": "71318a64.ec50b4",

"order": 1,

"width": 0,

"height": 0,

"name": "",

"label": "Prediction",

"format": "{{msg.payload}}",

"layout": "row-spread",

"x": 720,

"y": 400,

"wires": []

},

{

"id": "71318a64.ec50b4",

"type": "ui_group",

"z": "",

"name": "Machine Learning Model",

"tab": "d7ea07d0.8ba578",

"order": 1,

"disp": true,

"width": "6",

"collapse": false

},

{

"id": "d7ea07d0.8ba578",

"type": "ui_tab",

"z": "",

"name": "Home Page",

"icon": "dashboard",

"disabled": false,

"hidden": false

}

]

Git hub Repo Link for python notebook code-

https://github.com/SmartPracticeschool/llSPS-INT-1541-Predicting-Life-Expectancy-using-Machine-Learning