# Project Report

**Name:** Vamshi Padala (padalavamshi899@gmail.com)

**Title:** Predicting life Expectancy using Machine Learning

**Category:** Machine Learning

Internship at smartinternz.com@2020

**CONTENTS:**

**APPENDIX**

# 1.INTRODUCTION

### 1.1 PROJECT OVERVIEW:

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. This research tests the potential of using machine learning and natural language processing techniques for predicting life expectancy from electronic medical records.

### 1.2 PURPOSE:

The purpose of this study is to train machine learning models using a dataset containing data from over 800 medical examinations, in order to investigate what information can be learned regarding life expectancy. The models are trained as binary classifiers and four different target features are used. The features used are limited to features understandable to the author, who is not a medical professional. Results point to significant features primarily being smoking and body mass index and secondarily alcohol consumption, physical activity and coffee consumption.

# 2 LITERATURE SURVEY

### 2.1 EXISTING PROBLEM

Life expectancy plays an important role when decisions about the final phase of life need to be made. 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 (ACP) is the process during which patients make decisions about the health care they wish to receive in the future, in case the patient loses the capacity of making decisions or communicating about them . Successful ACP enhances the quality of life and death for palliative patients, by providing timely palliative care and documenting preferences regarding resuscitation and euthanasia, among other things .

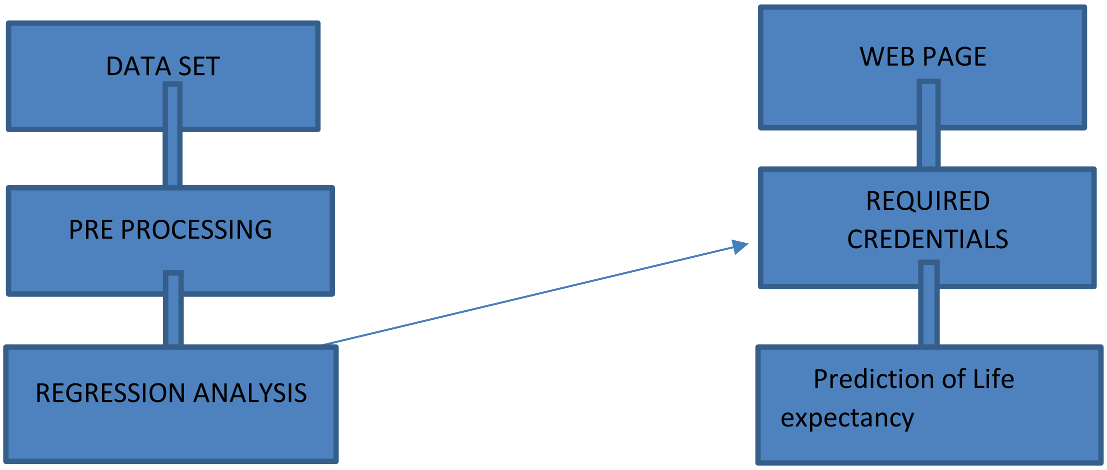
Accurate prognosis of life expectancy is essential for general practitioners (GPs) to decide when to introduce the topic of ACP to the patient, and it is a key determinant in end-of-life decisions. Increasing the accuracy of prognoses has the potential to benefit patients in various ways by enabling more consistent ACP, earlier and better anticipation on palliative needs, and preventing excessive treatment. This study focuses on automatic life expectancy prediction based on medical records.

### 2.2 PROPOSED SOLUTION

The project tries to create a model based on data provided by the World Health Organization (WHO) to evaluate the life expectancy for different countries in years. The data offers a timeframe from 2000 to 2015. The data originates from here: https://www.kaggle.com/kumarajarshi/life-expectancy-who/data The output algorithms have been used to test if they can maintain their accuracy in predicting the life expectancy for data they haven’t been trained.

## 3 THEORITICAL ANALYSIS

### 3.1 BLOCK DIAGRAM:



## 3.2 Hardware / Software designing

**Hardware:**

- A laptop with at least 4GB RAM
- A 2GB GPU

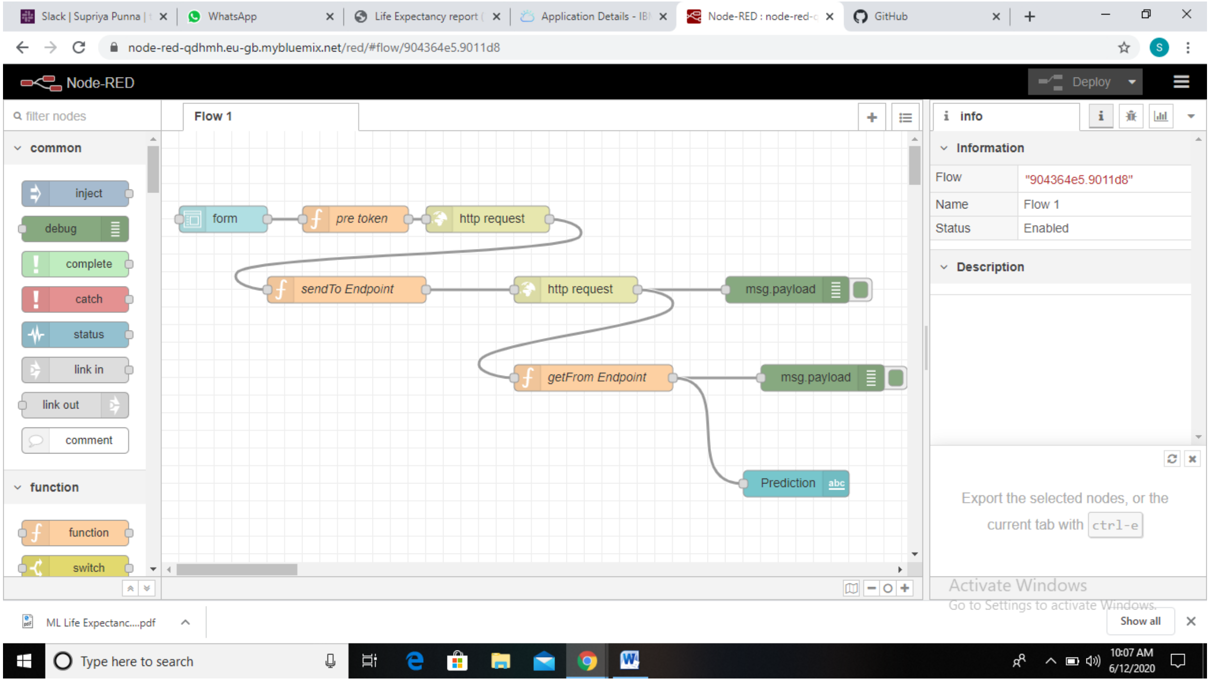
**Software:**

- **IDE**- Spyder, Jupyter
- **Scientific Computation Library** – Pandas
- **Visualization Libraries** – Matplotlib , Seaborn
- **Algorithmic Libraries** – Scikit-Learn , Stats models
- **Dependencies** – Data from internet

# 4 EXPERIMENTAL INVESTIGATIONS

Increasing age is a risk factor for many diseases; therefore developing pharmacological interventions that slow down ageing and consequently postpone the onset of many age-related diseases is highly desirable. In this work we analyse data from the Drug Age database, which contains chemical compounds and their effect on the lifespan of model organisms. Predictive models were built using the machine learning method random forests to predict whether or not a chemical compound will increase Caenorhabditis elegans' lifespan, using as features Gene Ontology (GO) terms annotated for proteins targeted by the compounds and chemical descriptors calculated from each compound's chemical structure. The model with the best predictive accuracy used both biological and chemical features, achieving a prediction accuracy of 80%. The top 20 most important GO terms include those related to mitochondrial processes, to enzymatic and immunological processes, and terms related to metabolic and transport processes. We applied our best model to predict compounds which are more likely to increase C. elegans' lifespan in the DGIdb database, where the effect of the compounds on an organism's lifespan is unknown. The top hit compounds can be broadly divided into four groups: compounds affecting mitochondria, compounds for cancer treatment, anti-inflammatories, and compounds for gonadotropinreleasing hormone therapies.

## Node Red Flow

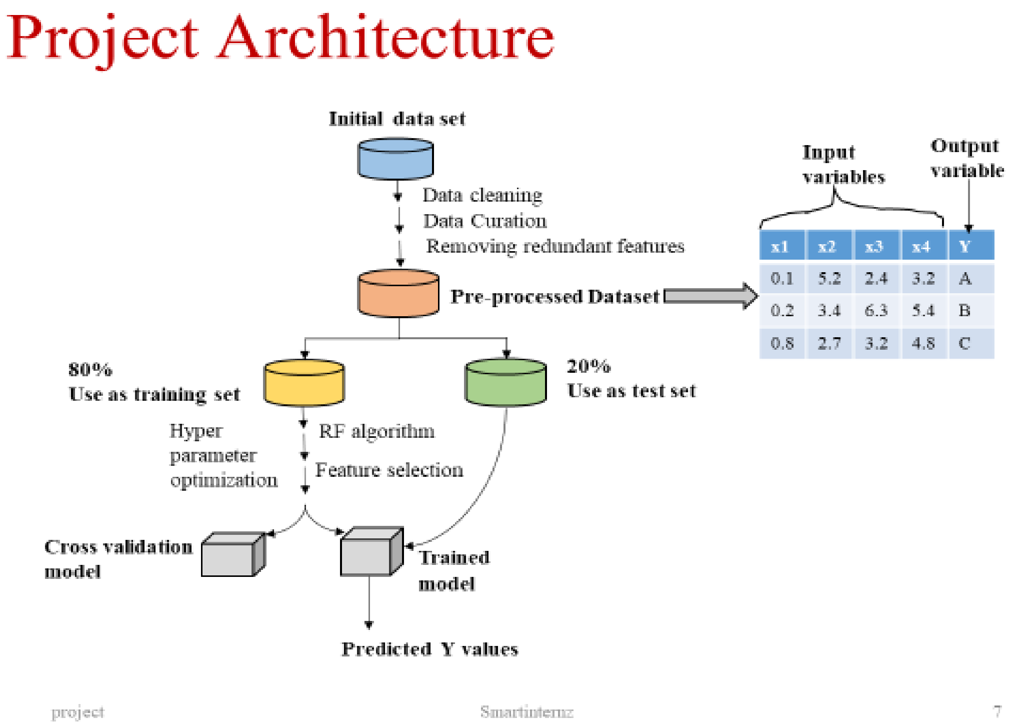


node-red-qdhmh.eu-gb.mybluemix.net/red/#flow/904364e5.9011d8

| Flow | "904364e5.9011d8" |
|---|---|
| Name | Flow 1 |
| Status | Enabled |

Export the selected nodes, or the current tab with ctrl-e

# 5 FLOWCHART



Project Architecture

| x1 | x2 | x3 | x4 | Y |
|---|---|---|---|---|
| 0.1 | 5.2 | 2.4 | 3.2 | A |
| 0.2 | 3.4 | 6.3 | 5.4 | B |
| 0.8 | 2.7 | 3.2 | 4.8 | C |

# 6 RESULT

The user should enter the credentials required for prediction . The average life expectancy prediction will be displayed on the screen.

# 7 ADVANTAGES & DISADVANTAGES

## Advantages of Machine learning

### 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. For instance, for an e-commerce website like Amazon, it serves to understand the browsing behaviours and purchase histories of its users to help cater to the right products, deals, and reminders relevant to them. It uses the results to reveal relevant advertisements to them.

### 2. No human intervention needed (automation)

With ML, you don't need to babysit your project every step of the way. Since it means giving machines the ability to learn, it lets them make predictions and also improve the algorithms on their own. A common example of this is anti-virus softwires; they learn to filter new threats as they are recognized. ML is also good at recognizing spam.

### 3. Continuous Improvement

As ml models  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 multi-dimensional and multi-variety data

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

### 5. Wide Applications

You could be an e-tailer or a healthcare provider and make ML work for you. Where it does apply, it holds the capability to help deliver a much more personal experience to customers while also targeting the right customers.

### Disadvantages of Machine Learning

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

**ML** 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. This leads to irrelevant advertisements being displayed to customers. 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

Existing health mobile app can integrate PLE feature as additional services (b) as both functions require the same physiological data to transfer to a monitoring center for PLE calculation.  Applications Along with existing heath applications such as fitness tracking, chronic disease monitoring and real-time patient monitoring, the PLE application can be useful for users to improve their lifestyle and exercise by planning goals on a short and long-term basis. For example, the current PLE outcome of 85 years will be adjusted when the user changes their attributes such as smoking cessation, reducing alcohol consumption, commencing regular exercise, or modifying dietary plans.

## Conclusions and Future Works

Few works have been done to provide an individually customized life expectancy prediction. We have reviewed existing works and techniques in the prediction of human LE, and reached a conclusion that it is feasible to predict a PLE for individuals using evolving technologies and devices such as big data, AI, machine learning techniques, and PHDs, wearables and mobile health monitoring devices. We also identified that the collection of data will be a huge challenge due to the privacy and government policy considerations, which will require collaboration of various bodies in the health industry. The interworking of a heterogeneous health network is also a challenge for data collection. Despite these challenges, we showed a possibility of a PLE prediction by proposing an approach of data collection and application by smartphone, with which users can enter their information to access the cloud server to obtain their own PLE. No attempt has been made to create this novel idea of using smartphone integrating cloud servers for real-time data entry. We investigated obstacles and barriers that can be resolved by future works described below. Previous works have described a five year LE prediction, however it is not oriented as a personalized prediction but rather utilizes a median model-predicted probability of 5-year survival of patients who are either sick or healthy. It is proposed that this can be extended to a lifetime prediction by using big data to generate a generic data, which can be used to create a PLE based on training data as a future solution. Building a generic database will take a considerable amount of time for data collection and analysis, taken from birth to death for various demographic groups to be useful and accurate in representing each attribute classifications. Whilst current applications attempt to show PLE for smartphone users, they are complicated and difficult to collect technical data requested by the questionnaire, as users are unlikely to be able to provide these data themselves. This can be resolved by connecting the app to the central cloud server with the mHealth networks which provide other health related applications. A centralized cloud server plays a key role in collecting, processing, and creating meaningful value using big data, which forms the input of the solution as well as creates generic data against each user's PLE requirements. Service providers shall envisage challenges and hurdles to obtain consent of personal health 'of heterogeneous health networks across developed countries. This will lead to (3) classification of data based on processing big data and each group's traits, which can be used as personalized threshold ranges; (4) When this has been completed in a cloud, it can be connected to a smart device app that can provide questionnaires developed by health specialists and collect answers to customize the user's PLE; (5) Optimization of the generic groups' data is done by developing an algorithm using machine learning for continuously building and optimizing the user's generic data. As the proposed solution requires processing and transmitting health information of users, information security is a key aspect to consider such as privacy as well as ethical requirements recommenced by regulation bodies, such as the Australian national health and medical research (NHMRC). The scope of security and ethical requirements need to be clearly defined and specified for future work as challenges are expected to build a centralized database with incorporation of health networks. For example, North America, Asia, and Europe may have their own unique requirements to satisfy in dealing with health data with different health research guidelines.

## 11 BIBILOGRAPHY

[1] IBM Cloud setup [Online]. Available: https://www.ibm.com/cloud/get-started .

[2] IBM Developer, "Node-RED starter tutorial" [Online]. Available: https://developer.ibm.com/tutorials/how-to-create-a-node-red-starter-application/ .

[3] "Node-RED labs on the use of the Watson Developer Cloud services - watson-developer-cloud/node-red-labs." [Online]. Available: https://github.com/watson-developer-cloud/node-red-labs .

[4] "Infuse AI into your applications with Watson AI to make more accurate predictions". [Online]. Available: https://www.ibm.com/watson/products-services .

[5] IBM Watson, "Intro to IBM Watson", 2018 [Online]. Available: https://www.youtube.com/watch?v=W3iPbFTAAds&feature=youtu.be .

[6] "Get an understanding of the principles of machine learning." [Online]. Available: https://developer.ibm.com/technologies/machine-learning/series/learning-path-machine-learning-for-developers/ .

[7] IBM Developer, "IBM Watson Machine Learning: Get Started in IBM Cloud", 2020 [Online]. Available: https://www.youtube.com/watch?v=NmdjtezQMSM .

[8] Watson Studio Workshop, “Chapter 4 Build and Deploy models in Jupyter Notebooks” [Online]. Available: https://bookdown.org/caoying4work/watsonstudio-workshop/jn.html .

[9] Kumar Rajarshi, “Life Expectancy (WHO) Statistical Analysis on factors influencing Life Expectancy”, 2018. [Online]. Available: https://www.kaggle.com/kumarajarshi/life-expectancy-who

[10] IBM Developer, “IBM Watson: Sign up for Watson Studio and Watson Knowledge Catalog”, 2019. [Online]. Available: https://www.youtube.com/watch?v=DBRGlAHdj48&list=PLzpeuWUENMK2PYtasCaKK4bZjaYzhW23L .

[11] IBM Developer, “IBM Watson Studio: Create a project”, 2019. [Online]. Available: https://www.youtube.com/watch?v=-CUi8GezG1I&list=PLzpeuWUENMK2PYtasCaKK4bZjaYzhW23L&index=2

[12] IBM Developer, “IBM Watson Studio: Jupyter notebook basics”, 2019 [Online]. Available: https://www.youtube.com/watch?v=Jtej3Y6uUng

# APPENDIX:

## SAMPLE CODES:

### Node red flow (fows (6).json)

```
[{"id":"904364e5.9011d8","type":"tab","label":"Flow 1","disabled":false,"info":""},{"id":"7b657457.6110ac","type":"ui_form","z":"904364e5.9011d8","name":"","label":"","group":"5741eac5.747344","order":0,"width":0,"height":0,"options":[{"label":"Year","value":"a","type":"number","required":true,"rows":null},{"label":"Life expectancy","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":"Hepatitis B","value":"g","type":"number","required":true,"rows":null},{"label":"Measles","value":"h","type":"number","required":true,"rows":null},{"label":"BMI","value":"i","type":"number","required":true,"rows":null},{"label":"under-five deaths","value":"j","type":"number","required":true,"rows":null},{"label":"Polio","value":"k","type":"number","required":true,"rows":null},{"label":"Total expenditure","value":"l","type":"number","required":true,"rows":null},{"label":"Diphtheria","value":"m","type":"number","required":true,"rows":null},{"label":"HIV/AIDS","value":"n",
```

"type":"number","required":true,"rows":null},{"label":"GDP","value":"o","type":"number","r
equired":true,"rows":null},{"label":"Population","value":"p","type":"number","required":true
,"rows":null},{"label":"thinness_1to19_years","value":"q","type":"number","required":true,"r
ows":null},{"label":"thinness_5to9_years","value":"r","type":"number","required":true,"rows
":null},{"label":"Income_Comp_Of_Resources","value":"s","type":"number","required":true,
"rows":null},{"label":"Schooling","value":"t","type":"number","required":true,"rows":null},{
"label":"Developing(1 or
0)","value":"u","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":"","u":""},"payload":"","submit":"submit","cancel":"cancel","topic":"","x":
70,"y":100,"wires":[["a9ae45ab.595698"]]},{"id":"a9ae45ab.595698","type":"function","z":"
904364e5.9011d8","name":"pre token","func":"//make user given values as global
variables\nglobal.set(\"a\",msg.payload.a);\nglobal.set(\"b\",msg.payload.b);\nglobal.set(\"c\",
msg.payload.c);\nglobal.set(\"d\",msg.payload.d);\nglobal.set(\"e\",msg.payload.e);\nglobal.s
et(\"f\",msg.payload.f);\nglobal.set(\"g\",msg.payload.g);\nglobal.set(\"h\",msg.payload.h);\n
global.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);\nglobal.set(\"p\",msg.payload.p);\nglobal.set(\"
q\",msg.payload.q);\nglobal.set(\"r\",msg.payload.r);\nglobal.set(\"s\",msg.payload.s);\ngloba
l.set(\"t\",msg.payload.t);\nglobal.set(\"u\",msg.payload.u);\n\n//following are required to
receive a token\nvar
apikey=\"NXFuP1HuaBMis_jdlXLASoonqEWaZucTTmk8NDWRQ30S\";\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":[["d7a1e79a.ab7028"]]},{"id":"2323c
5e4.ae16da","type":"http
request","z":"904364e5.9011d8","name":"","method":"POST","ret":"obj","paytoqs":false,"ur
l":"https://eu-gb.ml.cloud.ibm.com/v3/wml_instances/65c1b6c3-5868-4a26-b607-
af6768028f26/deployments/71a45c81-70d5-4558-b63a-
7688e384254a/online","tls":"","persist":false,"proxy":"","authType":"","x":470,"y":180,"wire
s":[["ecc48ad4.476f58","9cc5f34c.5d4c3"]]},{"id":"1a1b2322.13b11d","type":"debug","z":"
904364e5.9011d8","name":"","active":true,"tosidebar":true,"console":false,"tostatus":false,"c
omplete":"payload","targetType":"msg","x":750,"y":280,"wires":[]},{"id":"9cc5f34c.5d4c3",
"type":"function","z":"904364e5.9011d8","name":"getFrom
Endpoint","func":"msg.payload=msg.payload.values[0][0];\nreturn
msg;","outputs":1,"noerr":0,"x":490,"y":280,"wires":[["1a1b2322.13b11d","8eb09b5e.dc59d
8"]]},{"id":"ecc48ad4.476f58","type":"debug","z":"904364e5.9011d8","name":"","active":tr
ue,"tosidebar":true,"console":false,"tostatus":false,"complete":"payload","targetType":"msg",
"x":710,"y":180,"wires":[]},{"id":"f624e90c.552c68","type":"function","z":"904364e5.9011d
8","name":"sendTo Endpoint","func":"//get token and make headers\nvar
token=msg.payload.access_token;\nvar instance_id=\"65c1b6c3-5868-4a26-b607-
af6768028f26\"\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\");\nvar u = global.get(\"u\");\n\n//send the user
values to service endpoint\nmsg.payload = \n{\"fields\":['Year', 'Life expectancy ', 'Adult
Mortality', 'infant deaths',\n       'Alcohol', 'percentage expenditure', 'Hepatitis B', 'Measles ', '
BMI ',\n       'under-five deaths ', 'Polio', 'Total expenditure', 'Diphtheria ',\n       ' HIV/AIDS',
'GDP', 'Population', ' thinness  1-19 years',\n       ' thinness 5-9 years', 'Income composition of
resources', 'Schooling',\n
'Developing'],\n\"values\":[[a,b,c,d,e,f,g,h,i,j,k,l,m,n,o,p,q,r,s,t,u]]};\n\nreturn
msg;\n","outputs":1,"noerr":0,"x":210,"y":180,"wires":[["2323c5e4.ae16da"]]},{"id":"d7a1e7
9a.ab7028","type":"http
request","z":"904364e5.9011d8","name":"","method":"POST","ret":"obj","paytoqs":false,"ur
l":"https://iam.cloud.ibm.com/identity/token","tls":"","persist":false,"proxy":"","authType":"
basic","x":370,"y":100,"wires":[["f624e90c.552c68"]]},{"id":"8eb09b5e.dc59d8","type":"ui_
text","z":"904364e5.9011d8","group":"5741eac5.747344","order":1,"width":0,"height":0,"na
me":"","label":"Prediction","format":"{{msg.payload}}","layout":"row-
spread","x":720,"y":400,"wires":[]},{"id":"5741eac5.747344","type":"ui_group","z":"","nam
e":"Machine Learning
Model","tab":"74a8d81.8975d28","order":1,"disp":true,"width":"6","collapse":false},{"id":"7
4a8d81.8975d28","type":"ui_tab","z":"","name":"Home
Page","icon":"dashboard","disabled":false,"hidden":false}]
```

## Regression.ipynb

https://github.com/SmartPracticeschool/llSPS-INT-1730-Predicting-Life-Expectancy-using-Machine-Learning