Predicting Life Expectancy Using Machine Learning

Summer Internship Report

Submitted by:

Name : Nikhil V

Email : nikhilshettigar269@gmail.com

Application ID : SPS_APL_20200005375

Internship Title: Predicting Life Expectancy using Machine Learning - SB47588

Project ID : SPS_PRO_215

1. INTRODUCTION

1.1 Overview

Life expectancy means the estimation of the numbers of years the individual will survive. It further depends on Regional variations, Economic Circumstances, Sex Differences, Mental Illnesses, Physical Illnesses, Education, Year of their birth and other demographic factors It also helps in the treatment of patients which are underlying with certain diseases or illness. For example by predicting life expectancy we can analyse the aggressiveness of any disease. A large amount of data that is generated today is unstructured, which requires processing to generate insights. After pre-processing the dataset we will remove the noise from it. After removing the noise we will clean the data set. Once the data is ready to use we will train our model using this historical data and it is later tested with new data, if the desired accuracy is achieved the model will be deployed for the use. This application is a perfect use case for regression, which determines the relationship between one dependent variable (life expectancy) and a number of independent variables (development indicators).

In this project I am making use of IBM Watson studio and IBM machine learning to build the project. A web framework using Node red will be built so that a person without any knowledge in machine learning can interact with the model.

1.2 Purpose

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. 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. This helps to determine the life expectancy of a region and also to take necessary medical precautions if there are any diseases. This is project is also aimed usage of this project by a normal person by interacting with the model through the webpage where a user needs to enter all the required information and based on which the predicted life expectancy will be displayed.

2. LITERATURE SURVEY

2.1 Existing problem

Some existing techniques for predicting life expectancy have been analysed in which some of them fail to give the good result, reason for this might be following some old techniques or noise in the data There are even some other approaches were smart devices are used which involves human intervention, to overcome this we can use some techniques evolving technologies like Big data, AI, Machine learning. Instead of manually checking the historical data and predicting the insights to the future, a person has to just enter the details in the website and these details are sent to the model and the value will be predicted.

2.2 Proposed solution

There have been a lot studies undertaken to determine the factors which effects the life expectancy of a person, various new factors have been added which plays a major role in identifying the life expectancy.

Although there have been lot of studies undertaken in the past on factors affecting life expectancy considering demographic variables, income composition and mortality rates. It was found that affect of immunization and human development index was not taken into account in the past. Also, some of the past research was done considering multiple linear regression based on data set of one year for all the countries. Hence, this gives motivation to resolve both the factors stated previously by formulating a regression model based on mixed effects model and multiple linear regression while considering data from a period of 2000 to 2015 for all the countries. Important immunization like Hepatitis B, Polio and Diphtheria will also be considered. In a nutshell, this study will focus on immunization factors, mortality factors, economic factors, social factors and other health related factors as well. Since the observations 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 its population. The project relies on accuracy of data. The Global Health Observatory (GHO) data repository under World Health Organization (WHO) keeps track of the health status as well as many other related factors for all countries.

3. THEORITICAL ANALYSIS

3.1 Block diagram

3.2 Hardware / Software designing

This project is aimed at predicting the life expectancy, the primary requirement of this project is a dataset, since the dataset is important to train our model and predict the future based on the historical data. Other requirements are IBM cloud and academic initiative account, IBM Watson studio service, IBM machine learning service and node red service. The UI built can be run any browser.

Requirements:

- Suitable dataset
- IBM cloud account
- IBM academic initiative account
- Watson studio service
- Machine learning service
- Node red service
- Any Browser

4. EXPERIMENTAL INVESTIGATIONS

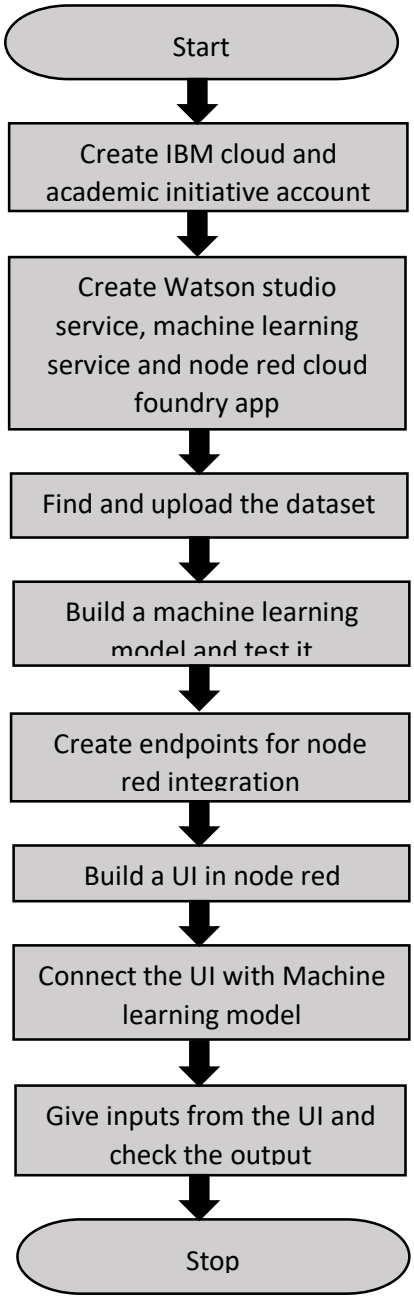
Here we investigate a model for prediction of life expectancy of a country given various factors. A typical Regression Machine Learning project leverages historical data to predict insights into the future. Analysis of various methods of regression gives a insight that random forest regression can achieve a higher a accuracy compared to other methods. Various factors 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.

Various factors which are required in this prediction are:

- Year
- Status (Developing/Developed)
- Adult mortality
- Infant deaths
- Alcohol
- Percentage expenditure
- Hepatitis B
- Measles
- Body mass index (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

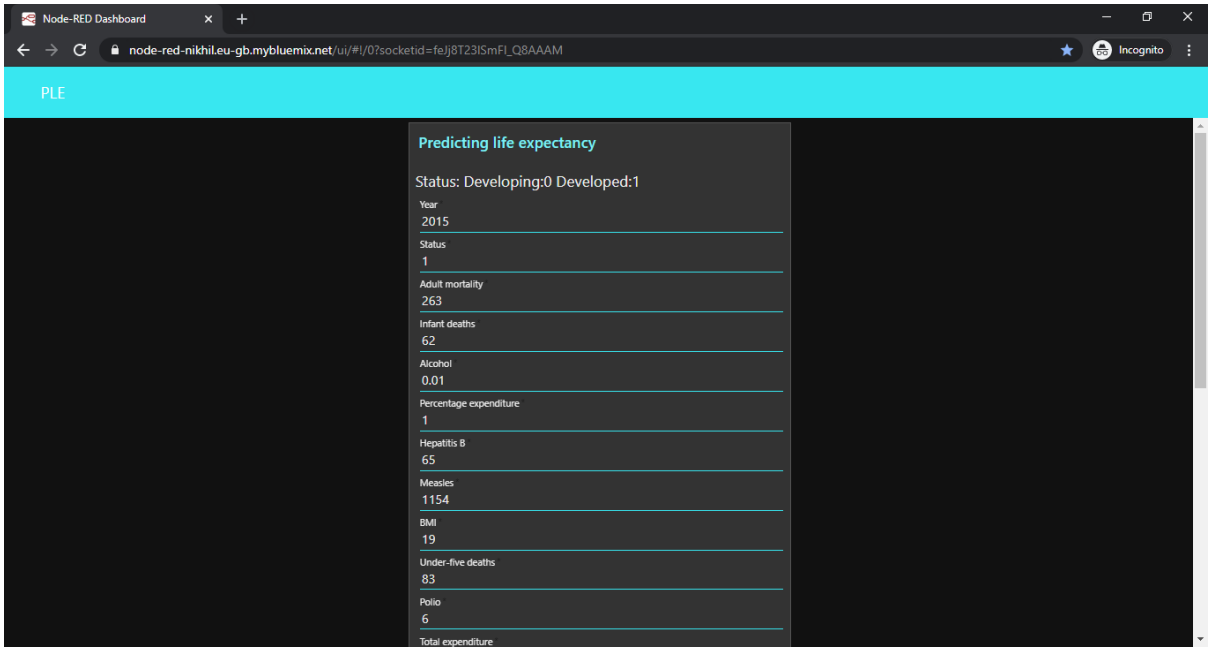
5. FLOWCHART

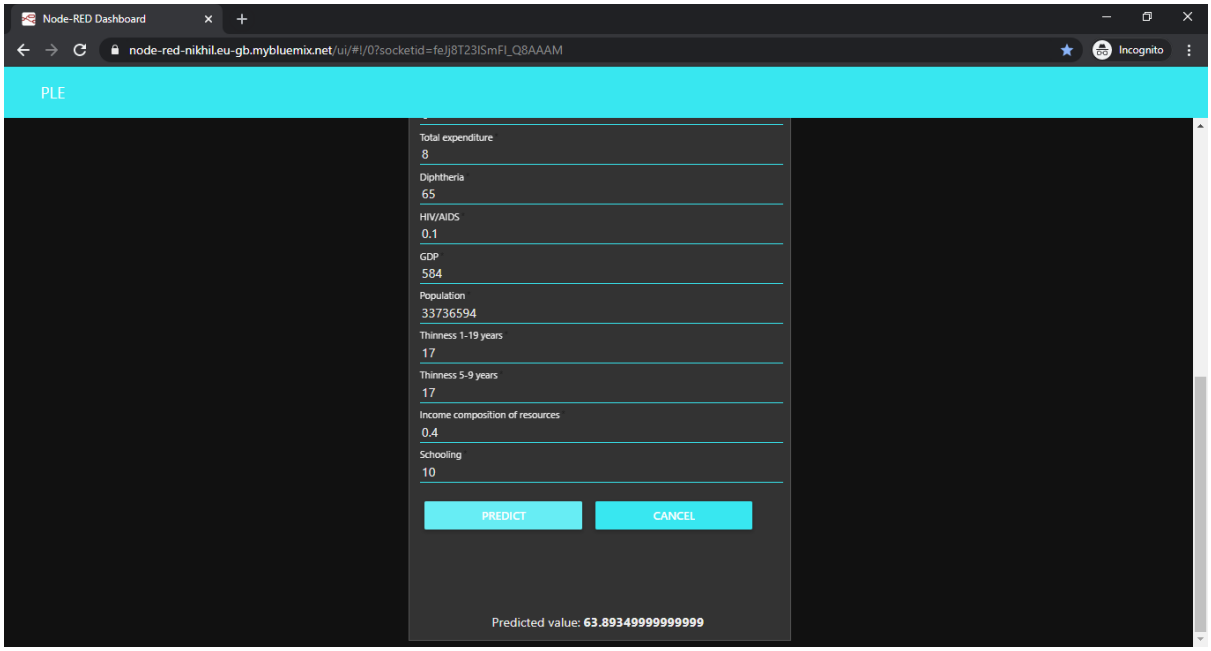
A flowchart is a type of diagram that represents a workflow or process. A flowchart can also be defined as a diagrammatic representation of an algorithm, a step-by-step approach to solving a task. The flowchart shows the steps as boxes of various kinds, and their order by connecting the boxes with arrows. This diagrammatic representation illustrates a solution model to a given problem. Flowcharts are used in analysing, designing, documenting or managing a process or program in various fields.



6. RESULT

After building the model 20% of the whole data was used for testing and achieved an accuracy of 0.954. The UI built using node red makes the task of predicting the life expectancy using different parameters simpler.





7. ADVANTAGES & DISADVANTAGES

Advantages

1. Advantage of the using the Watson studio
2. User friendly
3. Easy to build and deploy
4. Doesn’t require much storage space
5. Strengthens the business systems
6. Easily manage connections amongst partners and customers
7. Data storage can be accessed through different platforms and locations

Disadvantages

1. Lack of security
2. Loss of control on data
3. Depends on network

8. APPLICATIONS

Using this approach it is possible to identify the average life expectancy of a country given various factors, there are many more applications like

- It can be used to monitor health inequalities of a country.
- It can be used to develop statistics for country development process.
- It can be used to analyse the factors for high life expectancy.

9. CONCLUSION

This user interface will be useful for the user to predict life expectancy value of their own country or any other country based on some required details such as GDP, BMI, Year, Alcohol Intake, Total expenditure etc. The advantages of longer life span outweigh its disadvantages. The benefits people and the world can get from a higher life expectancy are irreplaceable and undeniable. It is a truth that life expectancy is a symbol of civilization and better life. Knowing an estimate of how much life we have left pushes us to achieve different things. Higher life expectation is also perceived as greater quality of life and greater income of society.

Our project has automated the entire task of rigorous calculation and removed errors in the existing system and gives the life expectancy to the user. This information can be useful to the society as stated above and this method is also much cheaper than hiring people to do the calculations.

10. FUTURE SCOPE

The result of this approach can be improved by making use of more precise dataset which containing the data with more factors on which the result depends on. We have built this model a static dataset, this should be changed to a real-time data where the down and fall in some values must be automatically updated.

11. BIBILOGRAPHY

https://developer.ibm.com/tutorials/how-to-create-a-node-red-starter-application/
https://www.youtube.com/watch?v=DBRGlAHdj48&list=PLzpeuWUENMK2PYtasC
https://www.kaggle.com/kumarajarshi/life-expectancy-who
https://www.youtube.com/watch?v=NmdjtezQMSM

APPENDIX

A. Source code

#Code for building ML model

```
#Import the data and necessary libraries
import types
from botocore.client import Config
import ibm_boto3
import pandas as pd
def __iter__(self): return 0
body=client_7da32595b948463b99a074a64c68d89d.get_object(Bucket='predictinglifeexpect
ancy-donotdelete-pr-zfjumqstpxwhbb',Key='Life Expectancy Data.csv')['Body']
# add missing __iter__ method, so pandas accepts body as file-like object
if not hasattr(body, "__iter__"): body.__iter__ = types.MethodType( __iter__, body )
life_data = pd.read_csv(body)
life_data.head()
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
#Drop the country column
life_data = life_data.drop('Country', axis = 1)
life_data.columns.values
cloumn_n=['Year', 'Status', 'Life expectancy', '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']
life_data.columns=cloumn_n
cloumn_n2=['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', 'Life expectancy']
life_data=life_data[cloumn_n2]
#data preprocessing
life_data = life_data.dropna(axis = 0, subset = ['Life expectancy'])
life_data.info()
life_data.mean()
life_data = life_data.fillna(life_data.mean())
life_data.info()
life_data.head()
life_data.isnull().sum()
life_data['Status'] = life_data['Status'].map({'Developing' : 0, 'Developed' : 1})
life_data.head()
x=life_data.iloc[:,0:-1].values
y=life_data.iloc[:,-1:].values
from sklearn.model_selection import train_test_split
#Split the data as train and test
xtrain,xtest,ytrain,ytest=train_test_split(x,y,test_size=0.2,random_state=0)
print("Train data size =",xtrain.shape,ytrain.shape,"\nTest data size =",xtest.shape,ytest.shape)
#Random forest regression
from sklearn.ensemble import RandomForestRegressor
regressor=RandomForestRegressor(n_estimators=200,random_state=0)
regressor.fit(xtrain,ytrain)
regressor.score(xtrain, ytrain)
#test the model
pred = regressor.predict(xtest)
pred
regressor.score(xtest, ytest)
plt.scatter(ytest,pred,color="blue")
plt.xlabel('Actual Value')
plt.ylabel('Predicted Value')
plt.show()
pred=np.reshape(pred,(586,1))
sns.set()
sns.distplot((ytest-pred),bins=30,color="blue")
from sklearn import metrics
print('MAE:', metrics.mean_absolute_error(ytest, pred))
print('MSE:', metrics.mean_squared_error(ytest, pred))
print('RMSE:', np.sqrt(metrics.mean_squared_error(ytest, pred)))
from watson_machine_learning_client import WatsonMachineLearningAPIClient
client=WatsonMachineLearningAPIClient(wml_cred)
metadata={
```

```
 client.repository.ModelMetaNames.AUTHOR_NAME:"Nikhil V",
 client.repository.ModelMetaNames.AUTHOR_EMAIL:"SI05202001130@smartinternz.co",
 client.repository.ModelMetaNames.NAME:"Predicting life expectancy"
}
stored_data=client.repository.store_model(regressor,meta_props=metadata)
stored_data
guid=client.repository.get_model_uid(stored_data)
guid
deploy=client.deployments.create(guid)
client.deployments.list()
deploy
scoring_endpoint=client.deployments.get_scoring_url(deploy)
scoring_endpoint
```

#Code for UI in Node red


```
[{"id":"3aa990a6.fecc","type":"tab","label":"Predicting life expectancy","disabled":false,"info":""},{"id":"59db1af2.042a34","type":"function","z":"3aa990a6.fecc","name":"PreToken","func":"global.set(\"ye\",msg.payload.ye)\nglobal.set(\"st\",msg.payload.st)\nglobal.set(\"am\",msg.payload.am)\nglobal.set(\"id\",msg.payload.id)\nglobal.set(\"al\",msg.payload.al)\nglobal.set(\"pe\",msg.payload.pe)\nglobal.set(\"hb\",msg.payload.hb)\nglobal.set(\"me\",msg.payload.me)\nglobal.set(\"bm\",msg.payload.bm)\nglobal.set(\"ud\",msg.payload.ud)\nglobal.set(\"po\",msg.payload.po)\nglobal.set(\"te\",msg.payload.te)\nglobal.set(\"di\",msg.payload.di)\nglobal.set(\"hi\",msg.payload.hi)\nglobal.set(\"gd\",msg.payload.gd)\nglobal.set(\"pp\",msg.payload.pp)\nglobal.set(\"ty\",msg.payload.ty)\nglobal.set(\"tys\",msg.payload.tys)\nglobal.set(\"icr\",msg.payload.icr)\nglobal.set(\"sc\",msg.payload.sc)\nvar apikey=\"Cr58cGUR-w6TjvFJQYqdGV81ccW6apTBwtUX5PsP_5g-\";\nmsg.headers={\"content-type\":\"application/x-www-form-urlencoded\"}\nmsg.payload={\"grant_type\":\"urn:ibm:params:oauth:grant-type:apikey\",\"apikey\":apikey}\nreturn msg;","outputs":1,"noerr":0,"x":240,"y":360,"wires":[["9cd1ed6.757d31"]]},{"id":"9cd1ed6.757d31","type":"http request","z":"3aa990a6.fecc","name":"","method":"POST","ret":"obj","paytoqs":false,"url":"https://iam.cloud.ibm.com/identity/token","tls":"","persist":false,"proxy":"","authType":"","x":450,"y":360,"wires":[["bfa5be64.3c912"]]},{"id":"c275a325.04a87","type":"inject","z":"3aa990a6.fecc","name":"","topic":"","payload":"","payloadType":"date","repeat":"","crontab":"","once":false,"onceDelay":0.1,"x":80,"y":300,"wires":[["59db1af2.042a34"]]},{"id":"4c2e410f.e103e","type":"debug","z":"3aa990a6.fecc","name":"","active":true,"tosidebar":true,"console":false,"tostatus":false,"complete":"payload","targetType":"msg","x":893.0000114440918,"y":107.00000190734863,"wires":[]},{"id":"bfa5be64.3c912","type":"function","z":"3aa990a6.fecc","name":"Pre Prediction","func":"var ye = global.get('ye')\nvar st = global.get('st')\nvar am = global.get('am')\nvar id = global.get('id')\nvar al = global.get('al')\nvar pe =
```

global.get('pe')\nvar hb = global.get('hb')\nvar me = global.get('me')\nvar bm = global.get('bm')\nvar ud = global.get('ud')\nvar po = global.get('po')\nvar te = global.get('te')\nvar di = global.get('di')\nvar hi = global.get('hi')\nvar gd = global.get('gd')\nvar pp = global.get('pp')\nvar ty = global.get('ty')\nvar tys = global.get('tys')\nvar icr = global.get('icr')\nvar sc = global.get('sc')\nvar token=msg.payload.access_token\nvar instance_id=\"2d6fc3d0-3802-47c1-b276-b012fc946ce2\"\nmsg.headers={'Content-Type': 'application/json',\"Authorization\":\"Bearer \"+token,\"ML-Instance-ID\":instance_id}\nmsg.payload={\"fields\": [\"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\"], \"values\": [[ye,st,am,id,al,pe,hb,me,bm,ud,po,te,di,hi,gd,pp,ty,tys,icr,sc]]}\nreturn msg;","outputs":1,"noerr":0,"x":640,"y":360,"wires":[["bbf95291.29aed"]]},{"id":"bbf95291.29aed","type":"http request","z":"3aa990a6.fecc","name":"","method":"POST","ret":"obj","paytoqs":false,"url":"https://eu-gb.ml.cloud.ibm.com/v3/wml_instances/2d6fc3d0-3802-47c1-b276-b012fc946ce2/deployments/8d2f6ad7-4d4d-44b1-9a86-9e540342d889/online","tls":"","persist":false,"proxy":"","authType":"","x":850,"y":360,"wires":[["f29af85.e1b2508","2bbf10b8.40116"]]},{"id":"96e47725.da4f28","type":"ui_form","z":"3aa990a6.fecc","name":"","label":"Status: Developing:0 Developed:1","group":"cedcfc24.0e74f","order":1,"width":"0","height":"0","options":[{"label":"Year","value":"ye","type":"number","required":true,"rows":null},{"label":"Status","value":"st","type":"number","required":true,"rows":null},{"label":"Adult mortality","value":"am","type":"number","required":true,"rows":null},{"label":"Infant deaths","value":"id","type":"number","required":true,"rows":null},{"label":"Alcohol","value":"al","type":"number","required":true,"rows":null},{"label":"Percentage expenditure","value":"pe","type":"number","required":true,"rows":null},{"label":"Hepatitis B","value":"hb","type":"number","required":true,"rows":null},{"label":"Measles","value":"me","type":"number","required":true,"rows":null},{"label":"BMI","value":"bm","type":"number","required":true,"rows":null},{"label":"Under-five deaths","value":"ud","type":"number","required":true,"rows":null},{"label":"Polio","value":"po","type":"number","required":true,"rows":null},{"label":"Total expenditure","value":"te","type":"number","required":true,"rows":null},{"label":"Diphtheria","value":"di","type":"number","required":true,"rows":null},{"label":"HIV/AIDS","value":"hi","type":"number","required":true,"rows":null},{"label":"GDP","value":"gd","type":"number","required":true,"rows":null},{"label":"Population","value":"pp","type":"number","required":true,"rows":null},{"label":"Thinness 1-19 years","value":"ty","type":"number","required":true,"rows":null},{"label":"Thinness 5-9 years","value":"tys","type":"number","required":true,"rows":null},{"label":"Income composition of resources","value":"icr","type":"number","required":true,"rows":null},{"label":"Schooling","value":"sc","type":"number","required":true,"rows":null}],"formValue":{"ye":"","st":"","am":"","id":"","al":"","pe":"","hb":"","me":"","bm":"","ud":"","po":"","te":"","di":"","hi":"","gd":

"","pp":"","ty":"","tys":"","icr":"","sc":""},"payload":"","submit":"Predict","cancel":"cancel",
"topic":"","x":180,"y":400,"wires":[["59db1af2.042a34"]]},{"id":"e6eeb798.001a58","type":
"ui_text","z":"3aa990a6.fecc","group":"cedcfc24.0e74f","order":2,"width":0,"height":0,"nam
e":"","label":"Predicted value:","format":"{{msg.payload}}","layout":"row-
center","x":860,"y":180,"wires":[]},{"id":"2bbf10b8.40116","type":"function","z":"3aa990a6
.fecc","name":"","func":"msg.payload=msg.payload.values[0][0]\nreturn
msg;","outputs":1,"noerr":0,"x":604.6000366210938,"y":167.60000610351562,"wires":[["4c
2e410f.e103e","e6eeb798.001a58"]]},{"id":"f29af85.e1b2508","type":"debug","z":"3aa990a
6.fecc","name":"","active":true,"tosidebar":true,"console":false,"tostatus":false,"complete":"f
alse","x":870,"y":420,"wires":[]},{"id":"cedcfc24.0e74f","type":"ui_group","z":"","name":"P
redicting life
expectancy","tab":"d7dec5e9.66baa8","order":2,"disp":true,"width":"9","collapse":false},{"i
d":"d7dec5e9.66baa8","type":"ui_tab","z":"","name":"PLE","icon":"dashboard","disabled":fa
lse,"hidden":false}]