Predicting Life Expectancy Using Machine Learning

Project Report

By - Pravallika Pulivendula

1. Introduction

1.1. Overview

Life expectancy refers to a prediction of the number of years for which a person will live. This number is determined based on the statistical average, considering many factors, including year and place of birth, race, education level, income, and medical history.

Life expectancy is the single most influential factor that insurance companies use to determine life insurance premiums. Using actuarial tables provided by the Internal Revenue Service, these companies try to minimize the liability risk.

For predicting the Life Expectancy, I have used a Supervised Machine Learning Algorithm (MultiVariate Regression) to train the model with Datasets, and when the user provides input on the Interface designed(developed using Node-RED) it gives the predicted Life Expectancy value.

1.2. Purpose

Purpose of this project is to predict the Life Expectancy of Humans, this helps in retirement planning, charging life insurance policies etc. Because Life Expectancy reflects the overall mortality level of a population.

2. Literature Survey

2.1. Existing Problem

As the technologies are evolving day by day, we humans are able to enjoy longer life expectancies than previously before.

Predicting a human's life expectancy has been a long-term question to humankind, and there have been many attempts to make the prediction accurate and popular since the prevalence of smartphones and apps. However, the effectiveness of those apps is limited due to the constraints of developing a classification of meta-data, such as the complexity and variety of environmental, geographic, genetic, and living factors of humans.

Calculation of life expectancy is a complicated process and requires many variables and circumstances to take into account, there have been several attempts to create an equation despite it being impractical to simplify these variables into one equation.

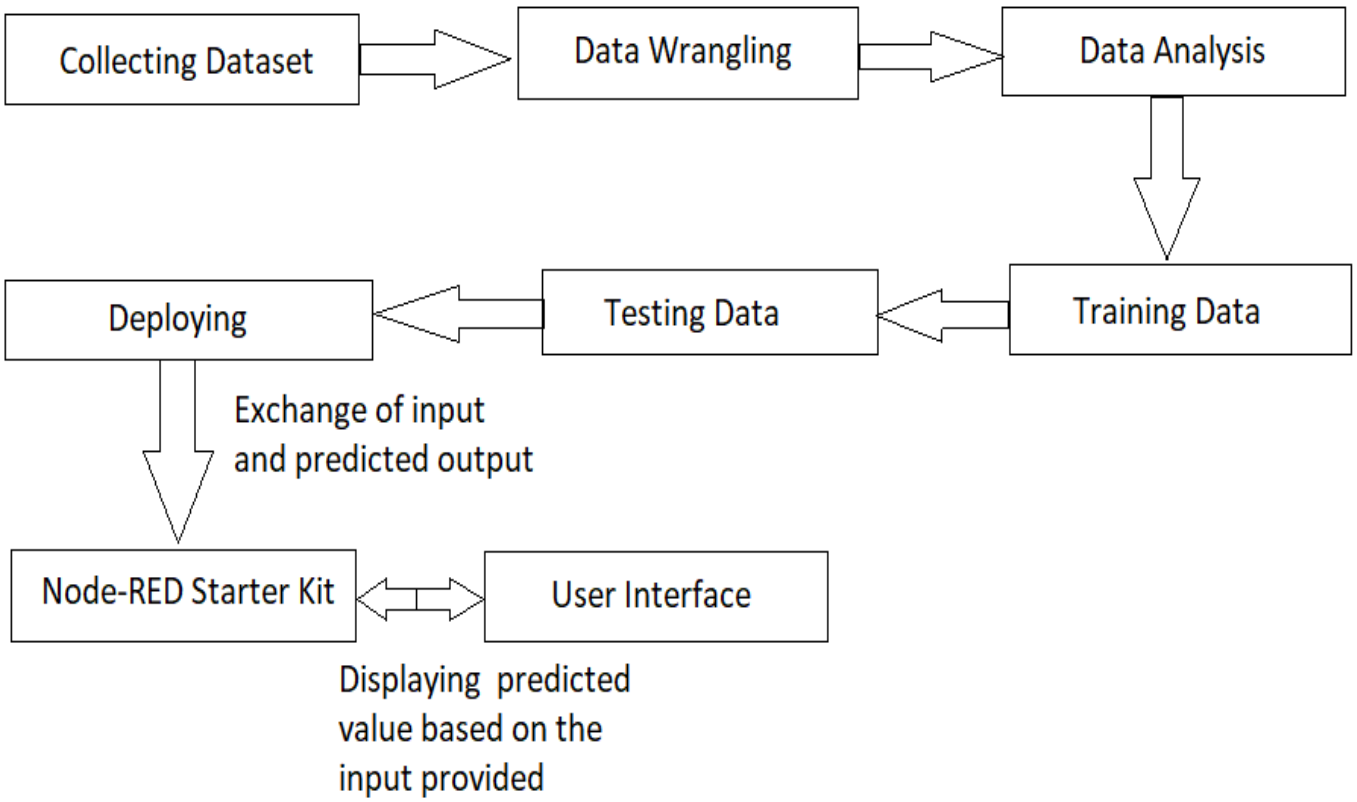
And from this it is very clear that it is feasible to predict the Life Expectancy of individuals using evolving technologies and devices such as big data, AI, machine learning techniques, and PHDs, wearables and mobile health monitoring devices.

2.2. Proposed Solution

The proposed solution involves the creation of the model based on the data provided to evaluate the life expectancy for different countries in years. Different Machine Learning algorithms can be used like Linear Regression, Ridge Regression, Lasso Regression, ElasticNet Regression, Linear Regression with Polynomial features, Decision Tree Regression, Random Forest Regression.

3. Theoretical Analysis

3.1. Block Diagram



3.2. Hardware\Software Designing

3.2.1. Hardware Requirements

Processor : i3 7th generation or higher

Speed : 2GHz or more

Hard Disk Space : 10GB or more

3.2.2. Software Requirements

Zoho Writer

Github

Node RED

Jupyter Notebook

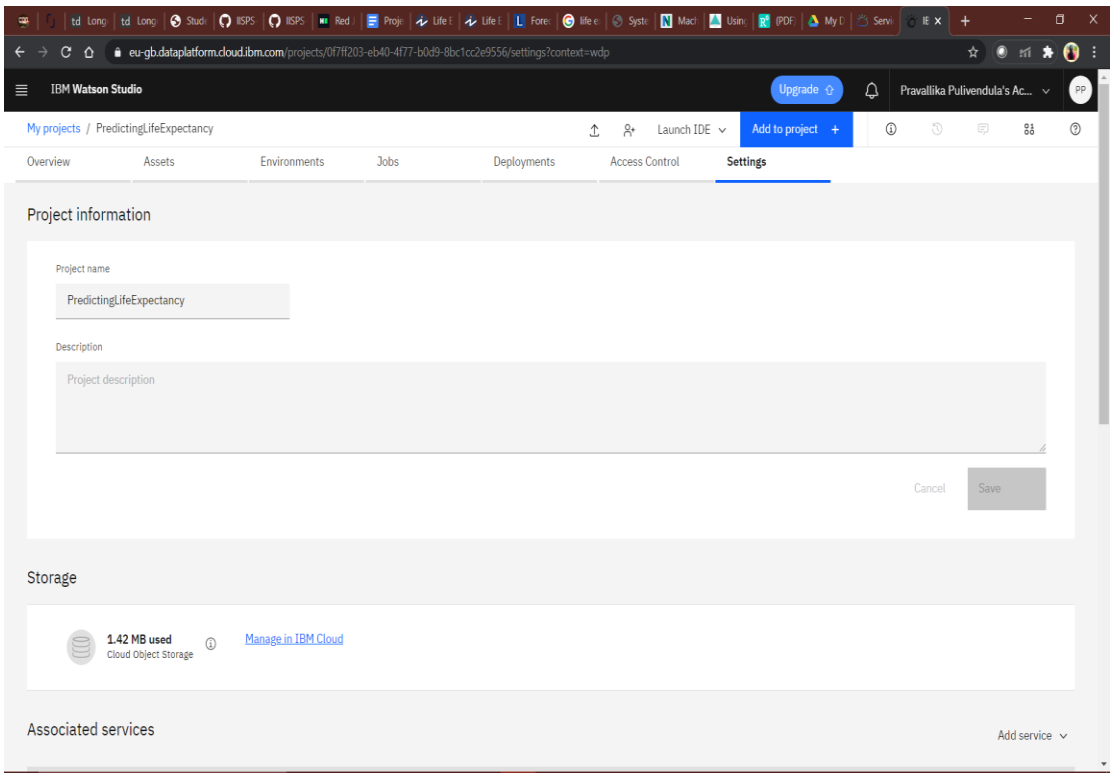
IBM Cloud

IBM Watson Studio

3.2.3. Designing Model

Steps to create Jupyter Notebook

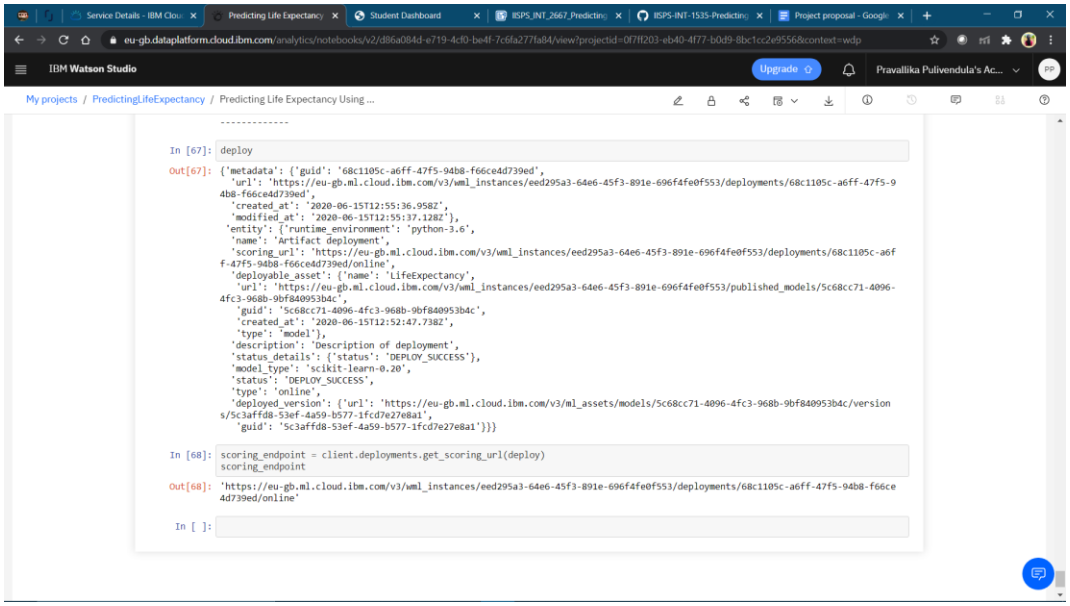
1. After creating IBM Watson Service on IBM Cloud.
2. Goto IBM Watson studio ->Get started -> Create a Project -> Create an empty project ->Name the project -> Click Create -> Add to Project->Notebook



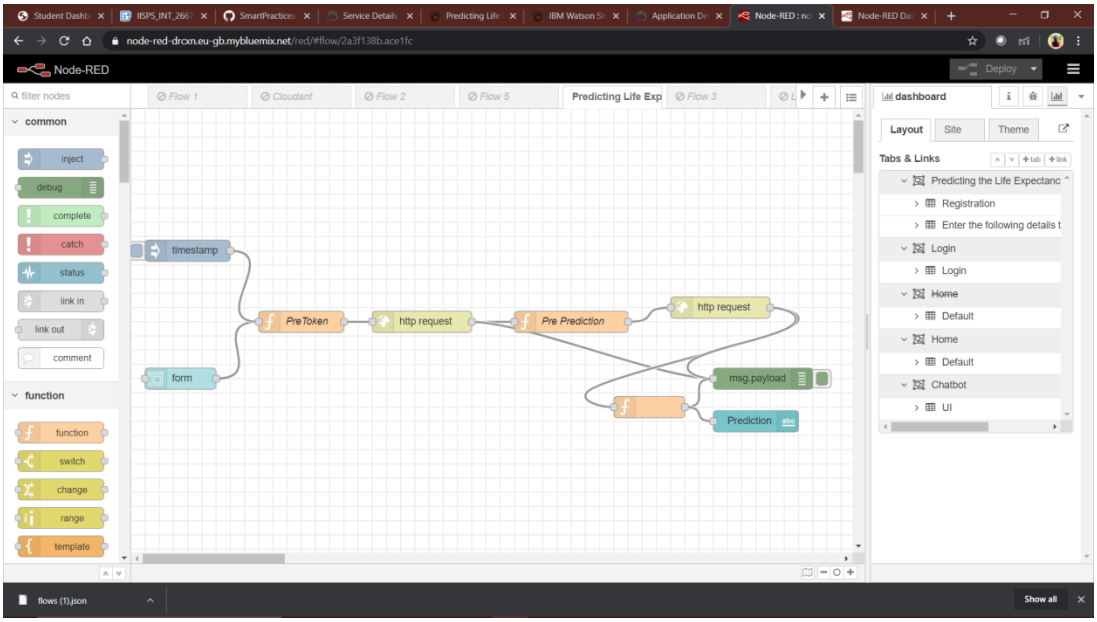
Notebook Link:

https://eu-gb.dataplatform.cloud.ibm.com/analytics/notebooks/v2/d86a084d-e719-4cf0-be4f-7c6fa277fa84/view?access_token=fc7cd3d93788221691d812bf62b4ebf4cb07f005a868a335e7ba3151d9661e2c

3.2.4. Scoring Endpoints



3.2.5. Node-RED Flow



4. Experimental Investigations

```
#Handling missing values
df_data_1 = df_data_1.interpolate(method="linear",limit_direction="forward")
df_data_1.isnull().sum()
```

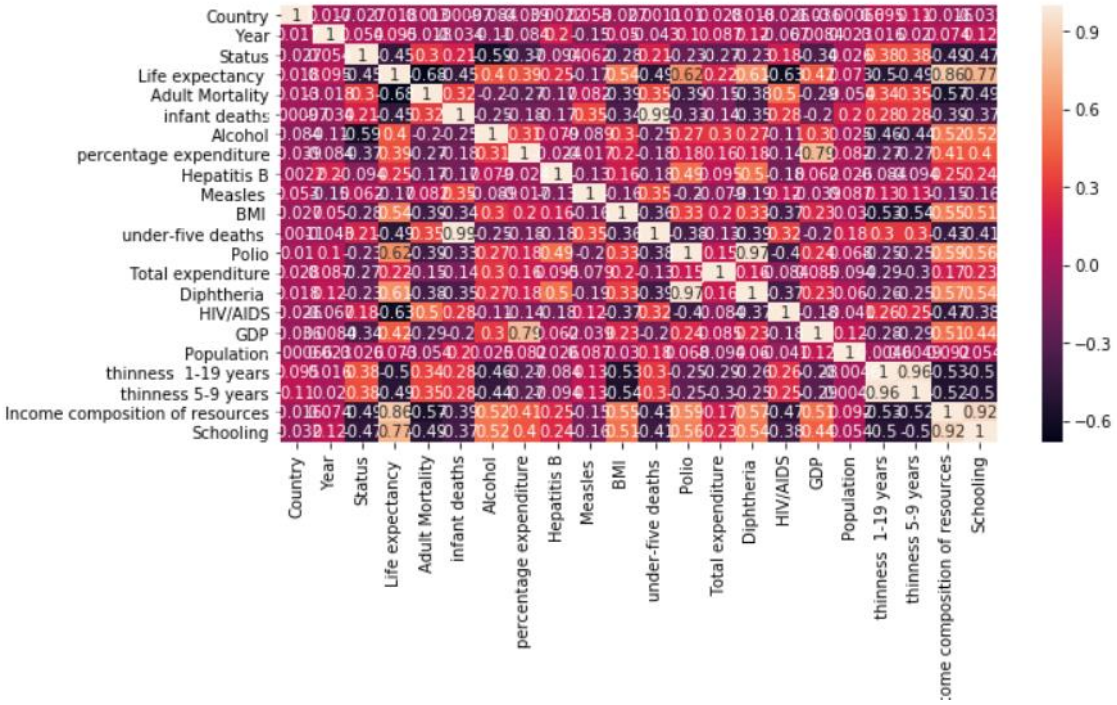
```
Country                          0
Year                             0
Status                           0
Life expectancy                  0
Adult Mortality                  0
infant deaths                    0
Alcohol                          0
percentage expenditure           0
Hepatitis B                      0
Measles                          0
 BMI                             0
under-five deaths                0
Polio                            0
Total expenditure                0
Diphtheria                       0
 HIV/AIDS                        0
GDP                              0
Population                       0
 thinness  1-19 years            0
 thinness 5-9 years              0
Income composition of resources  0
Schooling                        0
dtype: int64
```

```
#Finding and Detecting Outliers using Z-Score
z = np.abs(stats.zscore(df_data_1))
print(z)
threshold = 3
print(np.where(z > 3))
df_data_1 = df_data_1[(z < 3).all(axis=1)]
```

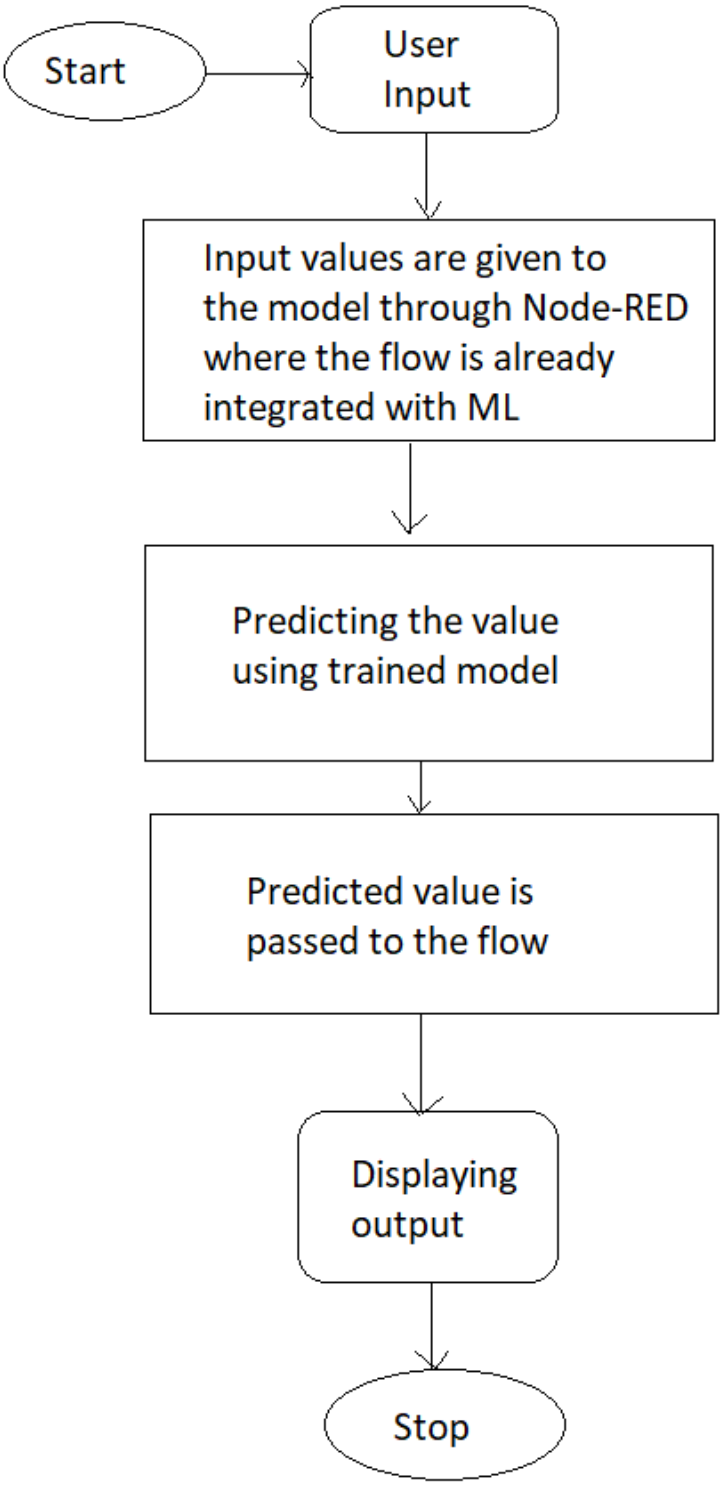
```
[[1.69104231 1.6217623  0.45939851 ... 2.73175989 0.74137957 0.59094031]
 [1.69104231 1.40498625 0.45939851 ... 2.77589328 0.75569475 0.62085876]
 [1.69104231 1.1882102  0.45939851 ... 2.82002667 0.7843251  0.65077721]
 ...
 [1.7231814  1.19632639 0.45939851 ... 0.79891123 0.98950927 0.62085876]
 [1.7231814  1.41310244 0.45939851 ... 0.71064445 0.98950927 0.68069566]
 [1.7231814  1.62987849 0.45939851 ... 1.38569153 0.9561072  0.68069566]]
(array([   0,    4,    5, ..., 2936, 2937, 2937]), array([12, 18, 18, ..., 15,  4, 15]))
```

```
#Data Visualizing
plt.figure(figsize=(10,5))
sb.heatmap(df_data_1.corr(),annot = True)
```

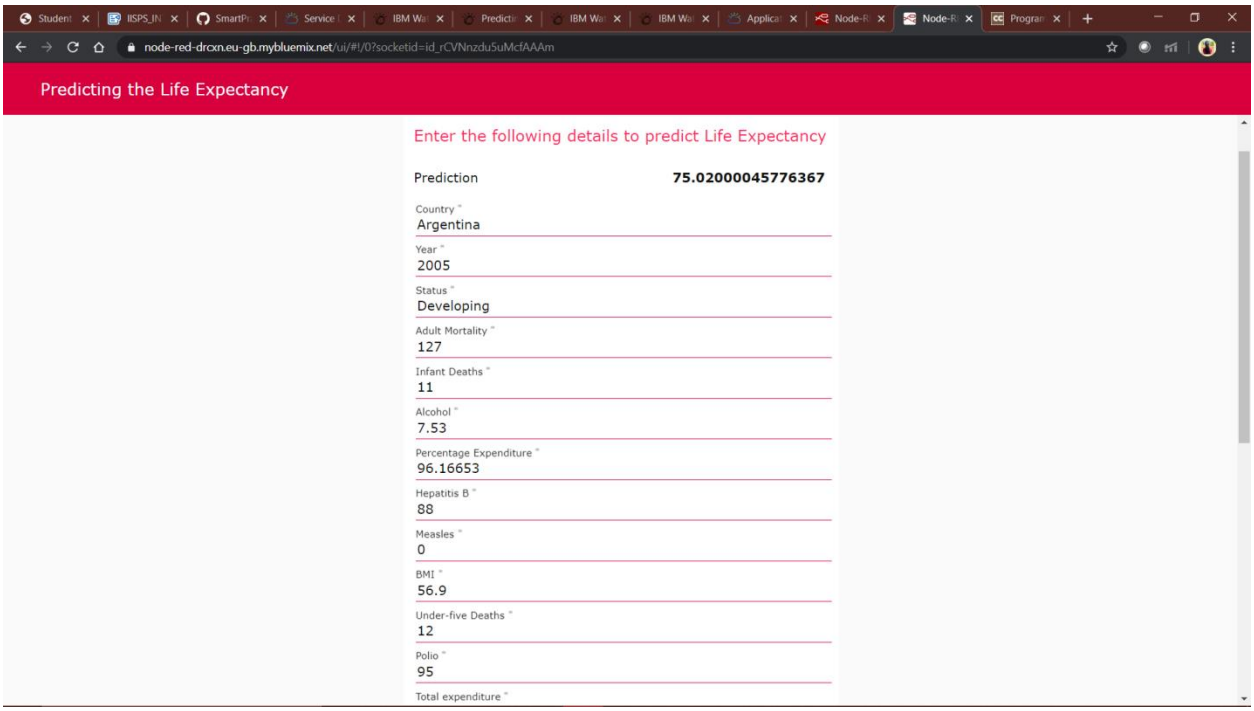
3]: <matplotlib.axes._subplots.AxesSubplot at 0x7f4b7ea70c50>

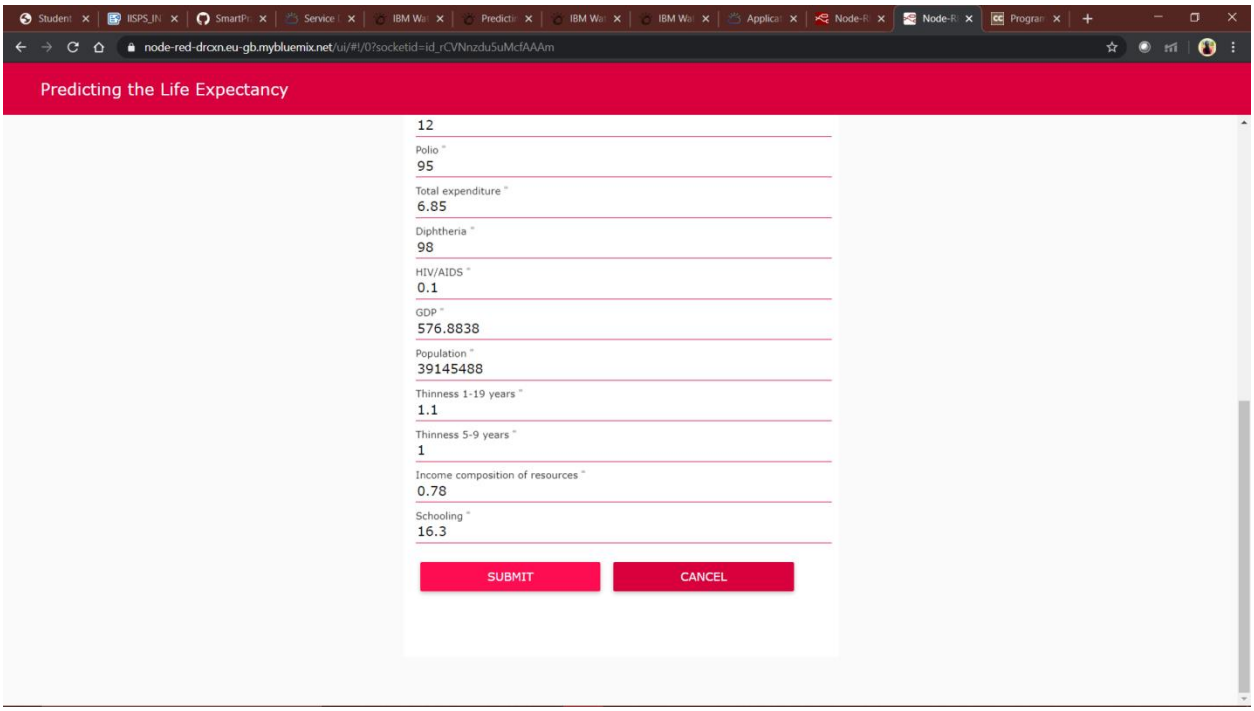


5. Flowchart


Start
User Input
Input values are given to the model through Node-RED where the flow is already integrated with ML
Predicting the value using trained model
Predicted value is passed to the flow
Displaying output
Stop

6. Result


node-red-drcxn.eu-gb.mybluemix.net/ui/#!/0?socketid=id_rCVNnzdu5uMcfAAAm
Predicting the Life Expectancy
Enter the following details to predict Life Expectancy
Prediction
75.02000045776367
Country *
Argentina
Year *
2005
Status *
Developing
Adult Mortality *
127
Infant Deaths *
11
Alcohol *
7.53
Percentage Expenditure *
96.16653
Hepatitis B *
88
Measles *
0
BMI *
56.9
Under-five Deaths *
12
Polio *
95
Total expenditure *


node-red-drcxn.eu-gb.mybluemix.net/ui/#!/0?socketid=id_rCVNnzdu5uMcfAAAm
Predicting the Life Expectancy
12
Polio *
95
Total expenditure *
6.85
Diphtheria *
98
HIV/AIDS *
0.1
GDP *
576.8838
Population *
39145488
Thinness 1-19 years *
1.1
Thinness 5-9 years *
1
Income composition of resources *
0.78
Schooling *
16.3
SUBMIT
CANCEL

7. Advantages and Disadvantages

7.1. Advantages

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

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

When it comes to designing UI, using Node-RED makes the work easier. One doesn't require much knowledge for creating Node-RED flow. Integrating our model with Node-RED flow does our job.

7.2. Disadvantages

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.

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.

One cannot customize UI using Node-RED

8. Applications

Life Insurance

Life expectancy is the single most influential factor that insurance companies use to determine life insurance premiums. Life expectancy is the primary factor in determining an individual's risk factor and the likelihood they will make a claim. Insurance companies consider age, lifestyle choices, family medical history, and several other factors when determining premium rates for individual life insurance policies.

Retirement and Annuity Planning

Life expectancy is critical for retirement planning. Many aging workers arrange their retirement plans' asset allocations based on a prediction of how long they expect to live. Personal, rather than statistical, life expectancy is a primary factor in the character of a retirement plan.

Most retirement plans, including the traditional and Roth, SEP, and SIMPLE IRA plans, also use life expectancy to determine the implementation of required minimum distributions (RMD)

Government

Governments may be able to use predictions to more efficiently allocate limited resources, such as social welfare assistance and health care funding, to individuals and areas of greater need.

9. Conclusion

Life expectancy predictions have the potential to be beneficial to individuals, health service providers and governments.
For instance, they would make people more aware of their general health, and its improvement or deterioration over time. This may motivate them to make healthier lifestyle choices.

As ML algorithms gain experience, they keep improving in accuracy and efficiency. This lets them make better decisions.Machine Learning algorithms are good at handling data that are multidimensional and multi-variety, and they can do this in dynamic or uncertain environments.

10. Future Scope

Technological advances will definitely improve the flexibility and scalability of applications. Machine learning is the key to enabling Artificial Intelligence and the future of healthcare is data-driven.

We can also integrate the application with the advanced information in the future.

11. Bibliography

1. https://www.ibm.com/cloud/get-started
2. https://developer.ibm.com/tutorials/how-to-create-a-node-red-starter-application/
3. https://bookdown.org/caoying4work/watsonstudio-workshop/jn.html
4. https://developer.ibm.com/tutorials/watson-studio-auto-ai/
5. https://youtu.be/x_5kH26xics
6. https://youtu.be/iEadmCNb_hE
7. https://youtu.be/NCSzqhNR-Ik
8. https://youtu.be/O5wqjk_GeJo
9. https://youtu.be/nnVPOFTCmQw

12. Appendix

Source Code

```
import pandas as pd
import numpy as np
from scipy import stats
import seaborn as sb
import matplotlib.pyplot as plt
import math
import types
import pandas as pd
from botocore.client import Config
import ibm_boto3
def __iter__(self): return 0
# @hidden_cell
# The following code accesses a file in your IBM Cloud Object Storage. It includes your credentials.
# You might want to remove those credentials before you share the notebook.
client_2097cdcc80ba4cabb453151a56b106ae = ibm_boto3.client(service_name='s3',
    ibm_api_key_id='g2S5-HZo0If8MtXkeGVnPnf5RLj1BVeCuZIm4tibJvCm',
    ibm_auth_endpoint="https://iam.cloud.ibm.com/oidc/token",
    config=Config(signature_version='oauth'),
    endpoint_url='https://s3.eu-geo.objectstorage.service.networklayer.com')
body = client_2097cdcc80ba4cabb453151a56b106ae.get_object(Bucket='predictinglifeexpectancy-donotdelete-pr-d1io5z70pqobln',Key='datasets_12603_17232_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 )
df_data_1 = pd.read_csv(body)
df_data_1.head()
df_data_1.describe()
df_data_1.columns
df_data_1.shape
df_data_1.dtypes

#Checking for missing values
df_data_1.isnull().sum()

#Handling missing values

df_data_1 =
df_data_1.interpolate(method="linear",limit_direction="forward")
df_data_1.isnull().sum()

#Label Encoding

from sklearn.preprocessing import LabelEncoder

lb_make = LabelEncoder()

df_data_1["Country"] = lb_make.fit_transform(df_data_1["Country"])

df_data_1["Status"] = lb_make.fit_transform(df_data_1["Status"])
df_data_1.head(40)

#Finding and Detecting Outliers using Z-Score

z = np.abs(stats.zscore(df_data_1))

print(z)

threshold = 3

print(np.where(z > 3))
df_data_1 = df_data_1[(z < 3).all(axis=1)]
df_data_1

#Data Visualization

plt.figure(figsize=(20,))
sb.heatmap(df_data_1.corr(),annot = True)

#Dropping country and status columns from the dataset

df_data_1 = df_data_1.drop("Country",axis = 1)
```

```
df_data_1 = df_data_1.drop("Status",axis = 1)
#Dependent and Independent variables
x = df_data_1.iloc[:,df_data_1.columns != 'Life expectancy ']
y = df_data_1.iloc[:,1]
#Splitting Dataset
from sklearn.model_selection import train_test_split
x_train,x_test,y_train,y_test = train_test_split(x,y,test_size =
0.2,random_state = 0)
#Using Linear Regression and training the model
from sklearn.linear_model import LinearRegression
model = LinearRegression()
model.fit(x_train,y_train)
LinearRegression(copy_X = True, fit_intercept = True, n_jobs =
None,normalize = False)
y_prec = model.predict(x_test)
y_prec
y_test
x_test
#Metrics
from sklearn.metrics import mean_absolute_error, mean_squared_error
print("Mean Squared Error: %0.30f"%mean_squared_error(y_test, y_prec))
print("Mean Absolute Error: %0.30f"%mean_absolute_error(y_test,y_prec))
print("Root Mean Squared Error:
%0.30f"%math.sqrt(mean_absolute_error(y_test,y_prec)))
#Deploying the model
from watson_machine_learning_client import
WatsonMachineLearningAPIClient
wml_credentials = {
  "apikey": "XVbCClgW-AyF176G3e2fRNL2SdDQAawomA5Tf0PHH1-d",
  "iam_apikey_description": "Auto-generated for key 56d1a38a-9733-449a-
988b-44746f99aa0a",
  "iam_apikey_name": "Service credentials-1",
  "iam_role_crn": "crn:v1:bluemix:public:iam::::serviceRole:Writer",
  "iam_serviceid_crn": "crn:v1:bluemix:public:iam-
identity::a/0e22e533c8d442798c558d0343fda3b1::serviceid:ServiceId-
a4606563-d004-4fb0-a464-b0c4aee8b544",
  "instance_id": "eed295a3-64e6-45f3-891e-696f4fe0f553",
```

```
  "url": "https://eu-gb.ml.cloud.ibm.com"
}
client = WatsonMachineLearningAPIClient(wml_credentials)
metadata = {
    client.repository.ModelMetaNames.AUTHOR_NAME : "Pravallika",
    client.repository.ModelMetaNames.AUTHOR_EMAIL :
"pravallikap2016@gmail.com",
    client.repository.ModelMetaNames.NAME : "LifeExpectancy"
}
stored_data = client.repository.store_model(model,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
```

Node-RED Flow(Code)


```
[{"id":"2a3f138b.ace1fc","type":"tab","label":"Predicting Life Expectancy
Using Machine
Learning","disabled":false,"info":""},{"id":"95087a18.be8ef8","type":"ui_for
m","z":"2a3f138b.ace1fc","name":"","label":"","group":"4a768920.9ef4a8","
order":3,"width":0,"height":0,"options":[{"label":"Country","value":"country
","type":"text","required":true,"rows":null},{"label":"Year","value":"year","t
ype":"number","required":true,"rows":null},{"label":"Status","value":"status
","type":"text","required":true,"rows":null},{"label":"Adult
Mortality","value":"adult","type":"text","required":true,"rows":null},{"label":
"Infant
Deaths","value":"infant","type":"text","required":true,"rows":null},{"label":"
Alcohol","value":"alcohol","type":"text","required":true,"rows":null},{"label"
:"Percentage
Expenditure","value":"expenditure","type":"text","required":true,"rows":null
},{"label":"Hepatitis
B","value":"hepatitis","type":"text","required":true,"rows":null},{"label":"Me
```

asles
","value":"measles","type":"text","required":true,"rows":null},{"label":" BMI
","value":"bmi","type":"text","required":true,"rows":null},{"label":"Under-
five Deaths
","value":"under","type":"text","required":true,"rows":null},{"label":"Polio","
value":"polio","type":"text","required":true,"rows":null},{"label":"Total
expenditure","value":"total","type":"text","required":true,"rows":null},{"labe
l":"Diphtheria
","value":"diphtheria","type":"text","required":true,"rows":null},{"label":"HI
V/AIDS","value":"hiv","type":"text","required":true,"rows":null},{"label":"G
DP","value":"gdp","type":"text","required":true,"rows":null},{"label":"Popula
tion","value":"population","type":"text","required":true,"rows":null},{"label"
:"Thinness 1-19
years","value":"thin1","type":"text","required":true,"rows":null},{"label":"Th
inness 5-9
years","value":"thin2","type":"text","required":true,"rows":null},{"label":"In
come composition of
resources","value":"income","type":"text","required":true,"rows":null},{"lab
el":"Schooling","value":"schooling","type":"text","required":true,"rows":null}
],"formValue":{"country":"","year":"","status":"","adult":"","infant":"","alcoh
ol":"","expenditure":"","hepatitis":"","measles":"","bmi":"","under":"","polio"
:"","total":"","diphtheria":"","hiv":"","gdp":"","population":"","thin1":"","thin
2":"","income":"","schooling":""},"payload":"","submit":"submit","cancel":"c
ancel","topic":"","x":70,"y":380,"wires":[["9982250a.c486d8"]]},{"id":"998
2250a.c486d8","type":"function","z":"2a3f138b.ace1fc","name":"PreToken",
"func":"global.set(\"country\",msg.payload.country)\nglobal.set(\"year\",ms
g.payload.year)\nglobal.set(\"status\",msg.payload.status)\nglobal.set(\"ad
ult\",msg.payload.adult)\nglobal.set(\"infant\",msg.payload.infant)\nglobal.s
et(\"alcohol\",msg.payload.alcohol)\nglobal.set(\"expenditure\",msg.payload
.expenditure)\nglobal.set(\"hepatitis\",msg.payload.hepatitis)\nglobal.set(\"
measles\",msg.payload.measles)\nglobal.set(\"bmi\",msg.payload.bmi)\nglo
bal.set(\"under\",msg.payload.under)\nglobal.set(\"polio\",msg.payload.poli
o)\nglobal.set(\"total\",msg.payload.total)\nglobal.set(\"diphtheria\",msg.pa
yload.diphtheria)\nglobal.set(\"hiv\",msg.payload.hiv)\nglobal.set(\"gdp\",m
sg.payload.gdp)\nglobal.set(\"population\",msg.payload.population)\nglobal.
set(\"thin1\",msg.payload.thin1)\nglobal.set(\"thin2\",msg.payload.thin2)\n
global.set(\"income\",msg.payload.income)\nglobal.set(\"schooling\",msg.pa

yload.schooling)\nvar apikey=\"XVbCClgW-AyF176G3e2fRNL2SdDQAawomA5Tf0PHH1-d\";\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":300,"wires":[["8f1f2b4b.b18428"]]},{"id":"8f1f2b4b.b18428","type":"http request","z":"2a3f138b.ace1fc","name":"","method":"POST","ret":"obj","paytoqs":false,"url":"https://iam.cloud.ibm.com/identity/token","tls":"","persist":false,"proxy":"","authType":"","x":410,"y":300,"wires":[["c0615549.4677a8","4ff6b5d8.aaef1c"]]},{"id":"2b14be7b.e861e2","type":"inject","z":"2a3f138b.ace1fc","name":"","topic":"","payload":"","payloadType":"date","repeat":"","crontab":"","once":false,"onceDelay":0.1,"x":80,"y":200,"wires":[["9982250a.c486d8"]]},{"id":"4ff6b5d8.aaef1c","type":"debug","z":"2a3f138b.ace1fc","name":"","active":true,"tosidebar":true,"console":false,"tostatus":false,"complete":"payload","targetType":"msg","x":890,"y":380,"wires":[]},{"id":"c0615549.4677a8","type":"function","z":"2a3f138b.ace1fc","name":"Pre Prediction","func":"var country = global.get('country')\nvar year = global.get('year')\nvar status = global.get('status')\nvar adult = global.get('adult')\nvar infant = global.get('infant')\nvar alcohol = global.get('alcohol')\nvar expenditure = global.get('expenditure')\nvar hepatitis = global.get('hepatitis')\nvar measles = global.get('measles')\nvar bmi = global.get('bmi')\nvar under = global.get('under')\nvar polio = global.get('polio')\nvar total = global.get('total')\nvar diphtheria = global.get('diphtheria')\nvar hiv = global.get('hiv')\nvar gdp = global.get('gdp')\nvar population = global.get('population')\nvar thin1 = global.get('thin1')\nvar thin2 = global.get('thin2')\nvar income = global.get('income')\nvar schooling = global.get('schooling')\nvar token=msg.payload.access_token\nvar instance_id=\"eed295a3-64e6-45f3-891e-696f4fe0f553\"\nmsg.headers={'Content-Type': 'application/json',\"Authorization\":\"Bearer \"+token,\"ML-Instance-ID\":instance_id}\nmsg.payload={\"fields\": [\"year\",\"adult\",\"infant\",\"alcohol\",\"expenditure\",\"hepatitis\",\"measles\",\"bmi\",\"under\",\"polio\",\"total\",\"diphtheria\",\"hiv\",\"gdp\",\"population\",\"thin1\",\"thin2\",\"income\",\"schooling\"], \"values\": [[year,adult,infant,alcohol,expenditure,hepatitis,measles,bmi,under,polio,total,diphtheria,hiv,gdp,population,thin1,thin2,income,schooling]]}\nreturn

msg;","outputs":1,"noerr":0,"x":620,"y":300,"wires":[["f8a78aee.67b388"]]
},{"id":"f8a78aee.67b388","type":"http
request","z":"2a3f138b.ace1fc","name":"","method":"POST","ret":"obj","pay
toqs":false,"url":"https://eu-
gb.ml.cloud.ibm.com/v3/wml_instances/eed295a3-64e6-45f3-891e-
696f4fe0f553/deployments/68c1105c-a6ff-47f5-94b8-
f66ce4d739ed/online","tls":"","persist":false,"proxy":"","authType":"","x":83
0,"y":280,"wires":[["7aa4e4dd.83dfdc","4ff6b5d8.aaef1c"]]},{"id":"bc1361d
9.3c08b","type":"ui_text","z":"2a3f138b.ace1fc","group":"4a768920.9ef4a8"
,"order":1,"width":0,"height":0,"name":"","label":"Prediction","format":"{{m
sg.payload}}","layout":"row-
spread","x":880,"y":440,"wires":[]},{"id":"7aa4e4dd.83dfdc","type":"functio
n","z":"2a3f138b.ace1fc","name":"","func":"msg.payload=msg.payload.valu
es[0][0]\nreturn
msg;","outputs":1,"noerr":0,"x":730,"y":420,"wires":[["4ff6b5d8.aaef1c","b
c1361d9.3c08b"]]},{"id":"4a768920.9ef4a8","type":"ui_group","z":"","name
":"Enter the following details to predict Life
Expectancy","tab":"d29659d5.31f418","order":2,"disp":true,"width":"10","co
llapse":false},{"id":"d29659d5.31f418","type":"ui_tab","z":"","name":"DB2"
,"icon":"Login","disabled":false,"hidden":false}]