PREDICTING LIFE EXPECTANCY USING MACHINE LEARNING

Composed by

KANAUJ GUHAROY

Project Summary

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. By using supervised machine learning techniques. one can extract a model that will be able to predict the life expectancy of future years. One method of approach is to use LSTM models to achieve this task.

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 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. Following algorithms can be used:

1. Linear Regression
2. Ridge Regression
3. Lasso Regression
4. Elastic Net Regression
5. Linear Regression with Polynomic features
6. Decision Tree Regression
7. Random Forest Regression

A typical Regression Machine Learning project leverages historical data to predict insights into the future. This problem statement is aimed at predicting Life Expectancy rate of a country given various features.

Life expectancy is a statistical measure of the average time a human being is expected to live, Life expectancy depends on various factors: Regional variations, Economic Circumstances, Sex Differences, Mental Illnesses, Physical Illnesses, Education, Year of their birth and other demographic factors. This problem statement provides a way to predict average life expectancy of people living in a country when various factors such as year, GDP, education, alcohol intake of people in the country, expenditure on healthcare system and some specific disease related deaths that happened in the country are given.

Project Requirements

The project will help in determining the Life Expectancy on the basis of the dataset from https://www.kaggle.com/kumarajarshi/life-expectancy-who
The dataset will provide various information like kind of diseases leading to the deaths, etc. Thus, with the help of these information, the Life Expectancy shall be predicted.
The following packages have been imported NumPy, Pandas, Matplotlib, Scipy, Seaborn. Sklearn is the most widely used package for the machine learning process. The fallowing sub packages have been used:

1. train_test_split
2. linear_model
3. model_selection
4. metrics
5. tree
6. ensemble
7. preprocessing

Functional Requirements

The Functional requirements are in the dataset containing:

1. Country
2. Status
3. Life Expectancy
4. Adult Mortality
5. Alcohol
6. percentage expenditure
7. Hepatitis B
8. Measles
9. BMI
10. under-five deaths
11. Polio
12. Total expenditure
13. Diphtheria
14. HIV/AIDS
15. GDP
16. Population
17. thinness 1-19 years

18. thinness 5-9 years
19. Income composition of resources
20. Schooling

Software Requirements

1. Python
2. Excel
3. IBM Cloud
4. IBM Watson
5. IBM Machine Learning
6. IBM Node-Red Service

Project Deliverables

A working model the predicts life expectancy of a Country. It creates the model which gives life expectancy of a country depending on various factors like schooling, GDP, BMI etc.

Project Team

The Predicting Life Expectancy Using Machine LearningProject Team consist of only me, i.e., Kanauj Guharoy.

Project Schedule

The Predicting Life Expectancy Using Machine Learning   Project is scheduled to be delivered by 12th June, 2020.

Algorithms

Machine Learning:

Machine learning is an application of artificial intelligence (AI) that provides systems the ability to learn automatically and improve from experience without being explicitly programmed. Machine learning focuses on the development of computer programs that can access data and use it learn for themselves.

The process of learning begins with observations or data, such as examples, direct experience, or instruction, in order to look for patterns in data and make better decisions in the future based on the examples that one provides. The primary aim is to allow the computers learn automatically without human intervention or assistance and adjust actions accordingly.

The machine learning algorithm is classified into Supervised, unsupervised and reinforced learning. For this project a type of supervised model is used.

Supervised Model

Supervised machine learning algorithms can apply what has been learned in the past to

new data using labeled examples to predict future events. Starting from the analysis of a known training data set, the learning algorithm produces an inferred function to make predictions about the output values. The system can provide targets for any new input after sufficient training. Supervised learning problems can be further grouped into Regression and Classification problems

Classification

A classification problem is when the output variable is a category, such as “red” or “blue” or “disease” and “no disease”. A classification model attempts to draw some conclusion from observed values. Given one or more inputs a classification model will try to predict the value of one or more outcomes.

Regression

A regression problem is when the output variable is a real or continuous value. Many different models can be used, the simplest is the linear regression. It tries to fit data with the best hyper-plane which goes through the points. In this project I am using Elastic Nest Regression as it produces the least error rate among other algorithms available.

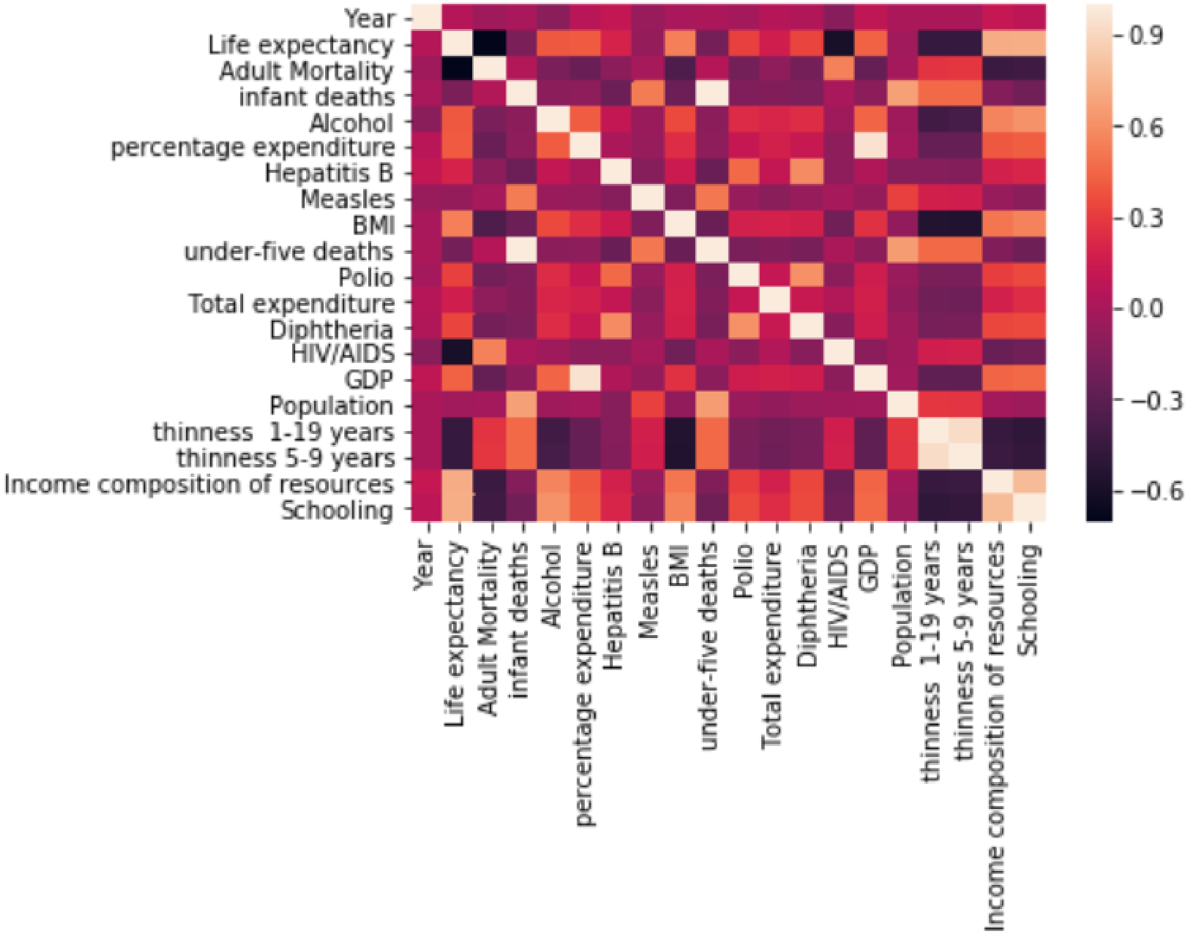
Elastic net linear regression uses the penalties from both the lasso and ridge techniques to regularize regression models. The technique combines both the lasso and ridge regression methods by learning from their shortcomings to improve on the regularization of statistical models.

The elastic net method improves on lasso's limitations, i.e., where lasso takes a few samples for high dimensional data, the elastic net procedure provides the inclusion of "n" number of variables until saturation. In a case where the variables are correlated groups, lasso tends to choose one variable from such groups and ignore the rest entirely.

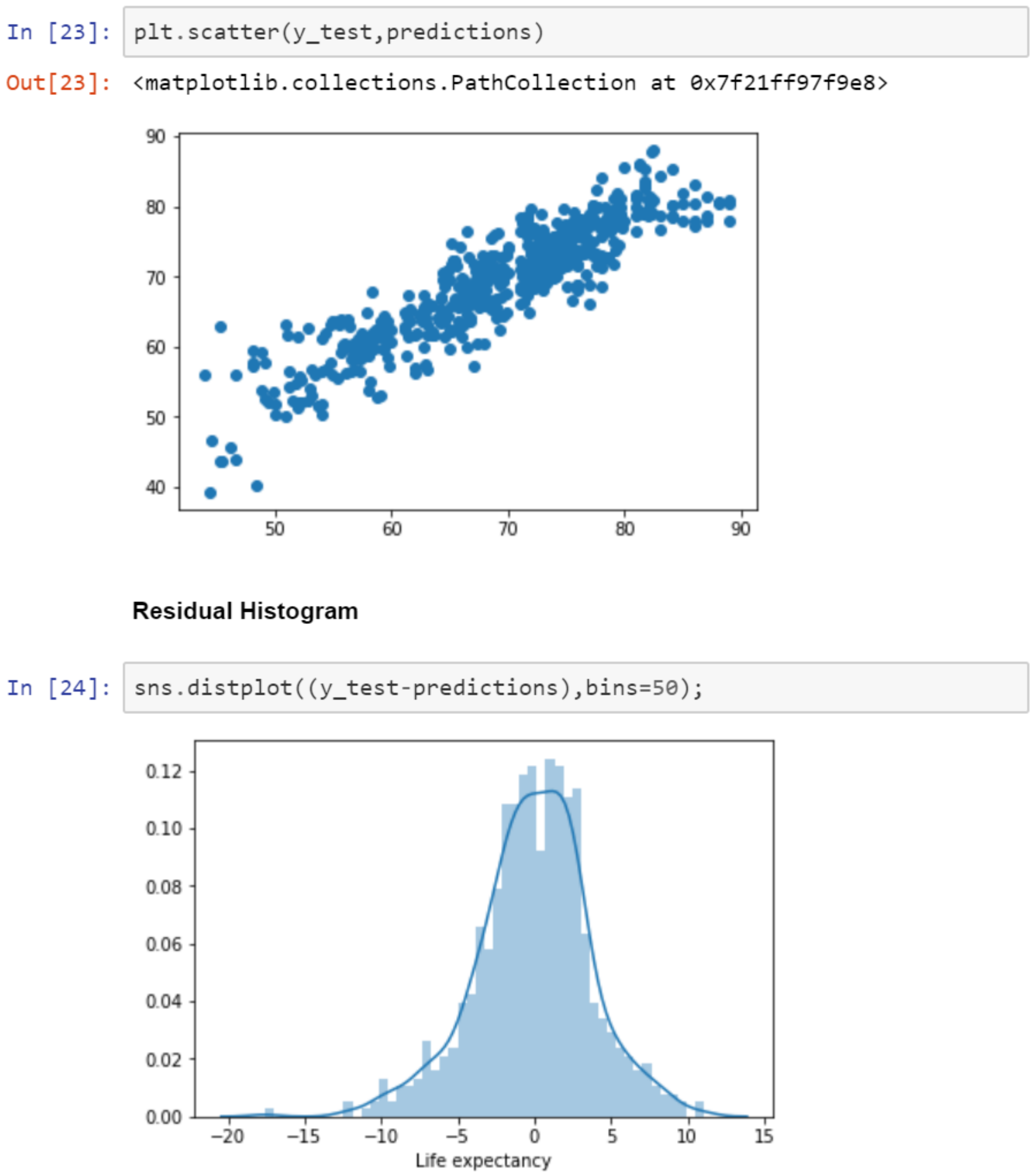
To eliminate the limitations found in lasso, the elastic net includes a quadratic expression in the penalty, which, when used in isolation, becomes ridge regression. The quadratic expression in the penalty elevates the loss function toward being convex. The elastic net draws on the best of both worlds – i.e., lasso and ridge regression.

Project Model

The model was created using Watson studio and Jupyter notebook. The data set contains 22 features. Since regression can only use number values the data set is grouped using countries, to avoid conflict. The null values were filled with their mean and the data was then split into training and testing data. The following heat map was generated from the data



After training the data using Elastic net regression the prediction models were obtained



```
In [23]: plt.scatter(y_test,predictions)

Out[23]: <matplotlib.collections.PathCollection at 0x7f21ff97f9e8>
```

Residual Histogram

```
In [24]: sns.distplot((y_test-predictions),bins=50);
```

Development Phases

1. Collecting the Dataset:

The most important thing for any project is collecting the data as per requirement of the model. Thus, firstly one collect the data from the given source. For the project the dataset was "Life Expectancy". The dataset was provided by the WHO in order for the analysis purpose. One have used this dataset for the prediction purpose.

2. Setting up IBM Cloud Services:

For using the various Cloud services for the project development. One must first create an IBM Cloud account. Once the account is created, one can access various services used for ML projects.

3. Creating a Watson Project:

Once the services required for the project is enabled, one can go with for the creation of the project. Watson Studio allows us to create various project using different tools like Jupyter notebook, Auto AI, R Studio etc.

Configure the Watson studio:

Once one is done with the creation of the Watson project, one can configure the various services associated with it. Also, one can look for the various tools associated with it.

4. Creating Machine Learning Services:

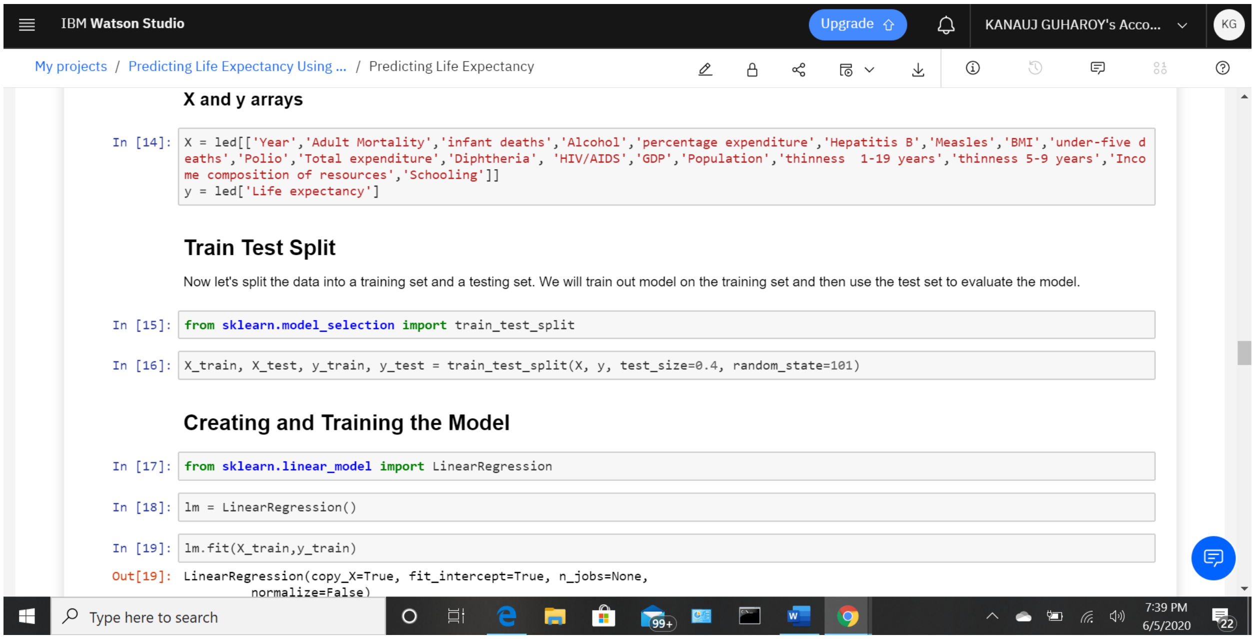
As one is creating the Machine Learning Model for the prediction of the Life Expectancy one must create the Machine learning services in IBM cloud which will help in building up the model.

a) Create Jupyter Notebook and Import Dataset:

Firstly in the project one need to add the Jupyter Notebook(It is the platform for developing the model and actual implementation). Once the Jupyter notebook is created one must import the data. The data set is Inserted to code in pandas data frame.

b) Choosing the appropriate Model for Prediction:

One can use any model for the prediction person and with the help of it one can train and test the dataset. For the project I have been choosing the Random Forest Regression Model for the development purpose.



IBM Watson Studio
Upgrade
KANAUJ GUHAROY's Acco...
KG
My projects / Predicting Life Expectancy Using ... / Predicting Life Expectancy

X and y arrays

```
In [14]: X = led[['Year','Adult Mortality','infant deaths','Alcohol','percentage expenditure','Hepatitis B','Measles','BMI','under-five d
         eaths','Polio','Total expenditure','Diphtheria', 'HIV/AIDS','GDP','Population','thinness  1-19 years','thinness 5-9 years','Inco
         me composition of resources','Schooling']]
         y = led['Life expectancy']
```

Train Test Split

Now let's split the data into a training set and a testing set. We will train out model on the training set and then use the test set to evaluate the model.

```
In [15]: from sklearn.model_selection import train_test_split

In [16]: X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.4, random_state=101)
```

Creating and Training the Model

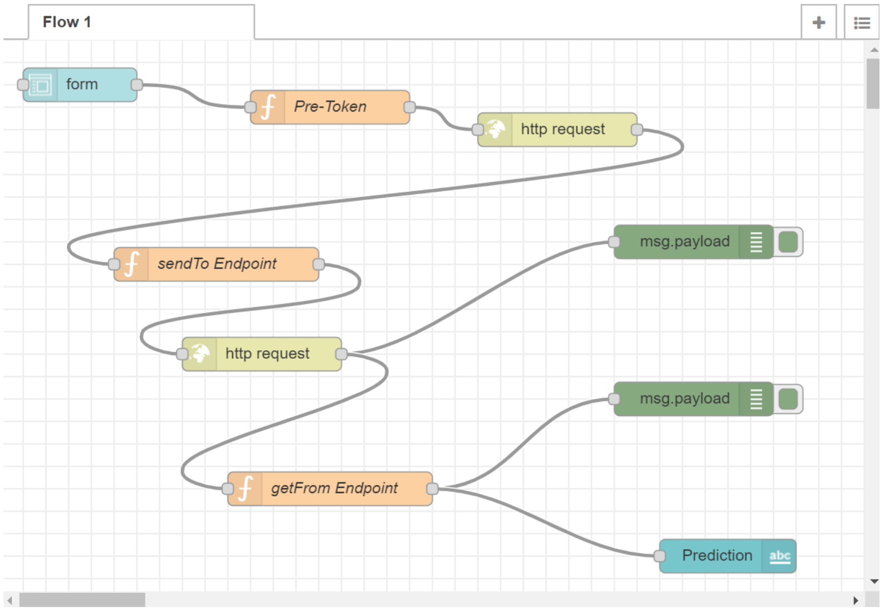
```
In [17]: from sklearn.linear_model import LinearRegression

In [18]: lm = LinearRegression()

In [19]: lm.fit(X_train,y_train)

Out[19]: LinearRegression(copy_X=True, fit_intercept=True, n_jobs=None,
                  normalize=False)
```

c) Deployment of Model:

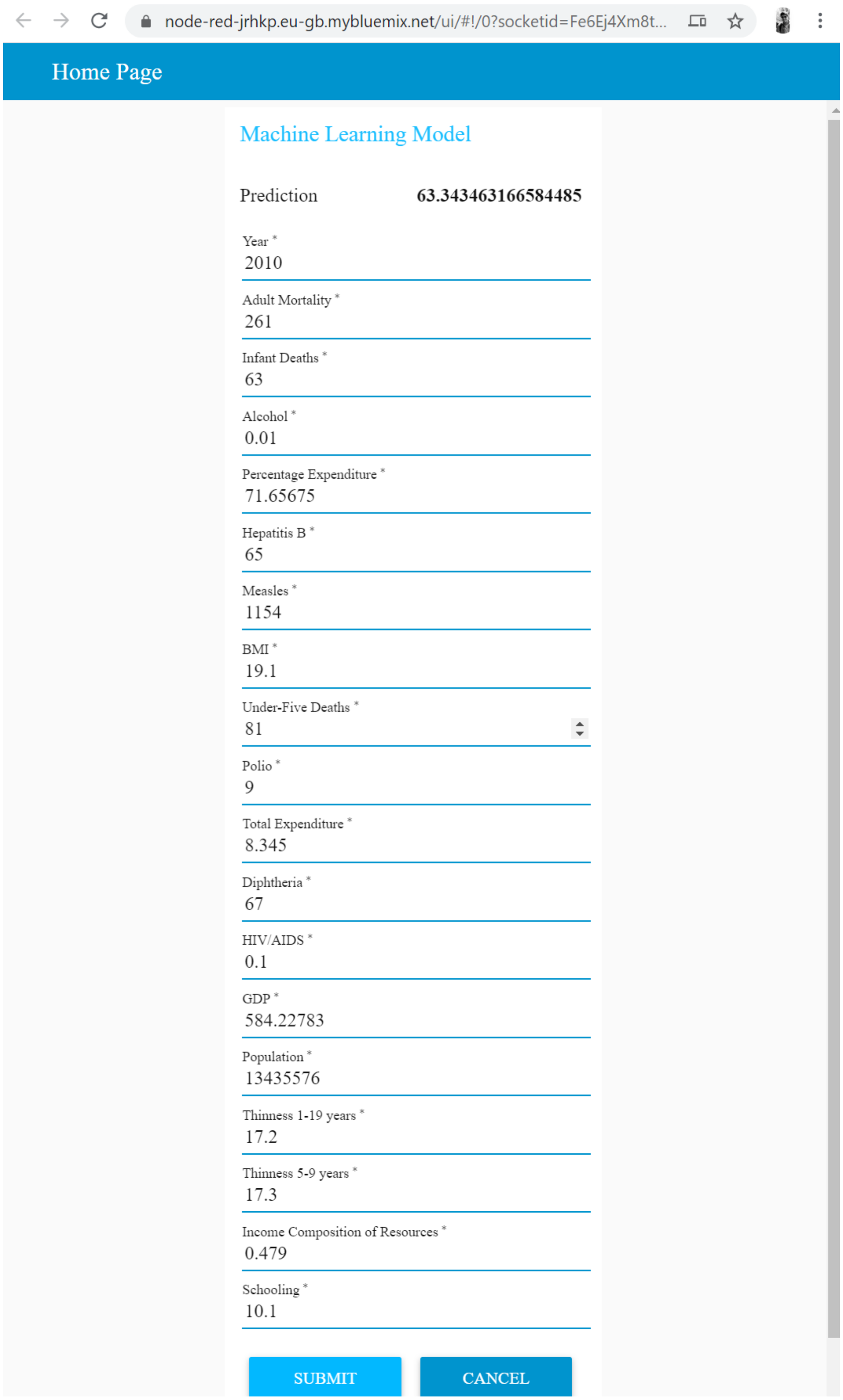
Once we're done with building the model, one must deploy the model. The deployed model will be stored in IBM Cloud Storage.

5. Create a Node-red Flow:

Once the model is deployed, one can create the node red flow to create an API for the model. The API will act as the front end to the model. From here one will get the Life Expectancy prediction data.



Application


node-red-jrhkp.eu-gb.mybluemix.net/ui/#!/0?socketid=Fe6Ej4Xm8t...
Home Page
SUBMIT
CANCEL

Machine Learning Model

Prediction 63.343463166584485

Year *
2010

Adult Mortality *
261

Infant Deaths *
63

Alcohol *
0.01

Percentage Expenditure *
71.65675

Hepatitis B *
65

Measles *
1154

BMI *
19.1

Under-Five Deaths *
81

Polio *
9

Total Expenditure *
8.345

Diphtheria *
67

HIV/AIDS *
0.1

GDP *
584.22783

Population *
13435576

Thinness 1-19 years *
17.2

Thinness 5-9 years *
17.3

Income Composition of Resources *
0.479

Schooling *
10.1

Conclusion

From the project I conclude that the Prediction of Life Expectancy from the given dataset by working on Watson studio as the Back end and Node Red Flow as the Front end is very much possible as shown.