

# **Project Report**

on

# **PREDICTING LIFE EXPECTANCY USING MACHINE LEARNING**

INTERNSHIP TITLE: Predicting Life Expectancy using Machine Learning - SB27373

PROJECT ID: SPS\_PRO\_215

**REMOTE SUMMER INTERNSHIP PROGRAM by SMARTINTERNZ**

- **Shubham Dua** (shubhamdua02@gmail.com)

---

## **CONTENTS:**

---

1. Introduction
    - a. Overview
    - b. Purpose
  2. Literature Survey
    - a. Existing problem
    - b. Proposed Solution
  3. Theoretical Analysis
    - a. Block Diagram
    - b. Hardware/Software Testing
  4. Experimental Investigations
  5. Flowcharts
  6. Result
  7. Advantages & Disadvantages
  8. Applications
  9. Conclusion
  10. Future Scope
  11. Bibliography
  12. Appendix
    - a. Source Code
-

# 1 INTRODUCTION

## 1.1 Overview

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: Country, Status, infant deaths, GDP, Population, BMI, other 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.

## 1.2 Purpose

Predicting the life expectancy will give the country an idea of the factors which can be improved to increase the lifespan of the people living. For example, by improving the health care facilities and immunization vaccines for infants or by making changes in lifestyle, a person can live a long, healthy and good quality life. This will also benefit the country by increasing manpower that will contribute to the economic growth. We should take full advantage of this new era advanced technology to improve the future by predicting it in the present.

The project uses a **Regression Model** which is a classification algorithm. It is a measure of the relation between the mean value of one variable (the output feature - Life Expectancy) and corresponding values of other variables. The dataset used to train this model was downloaded from kaggle.com (<https://www.kaggle.com/kumarajarshi/life-expectancy-who>), and we used Python to write the code for the Machine Learning model.

## **2 LITERATURE SURVEY**

### **2.1 Existing Problem**

One major problem that we encountered was that collecting data for predicting the average life expectancy is a big challenge due to the considerations related to many privacy and government policies.

A second hurdle that we faced was the features that were included in the dataset with which we predicted the expected lifespan. The factors that we have used to make our predictions were just personal causes and not related to the surrounding, healthcare facilities, demographic, social, regional and economic factors of the country in which the person resides. These country dependent factors can also be an important feature to predict the life expectancy of an individual.

The third issue that we faced is that it is feasible only for individuals using technologies such as mobile health monitoring and tracking devices.

Finally, 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 effect of immunization and human development index was not taken into account in the past. Also, some of the past research has been done considering multiple linear regressions based on the data gathered of one only particular year.

### **2.2 Proposed Solution**

As Artificial Intelligence and Machine Learning technologies are developing and quickly being implemented, the ease of gathering health data from the public as well as current government agencies such as centralized health servers should be increased

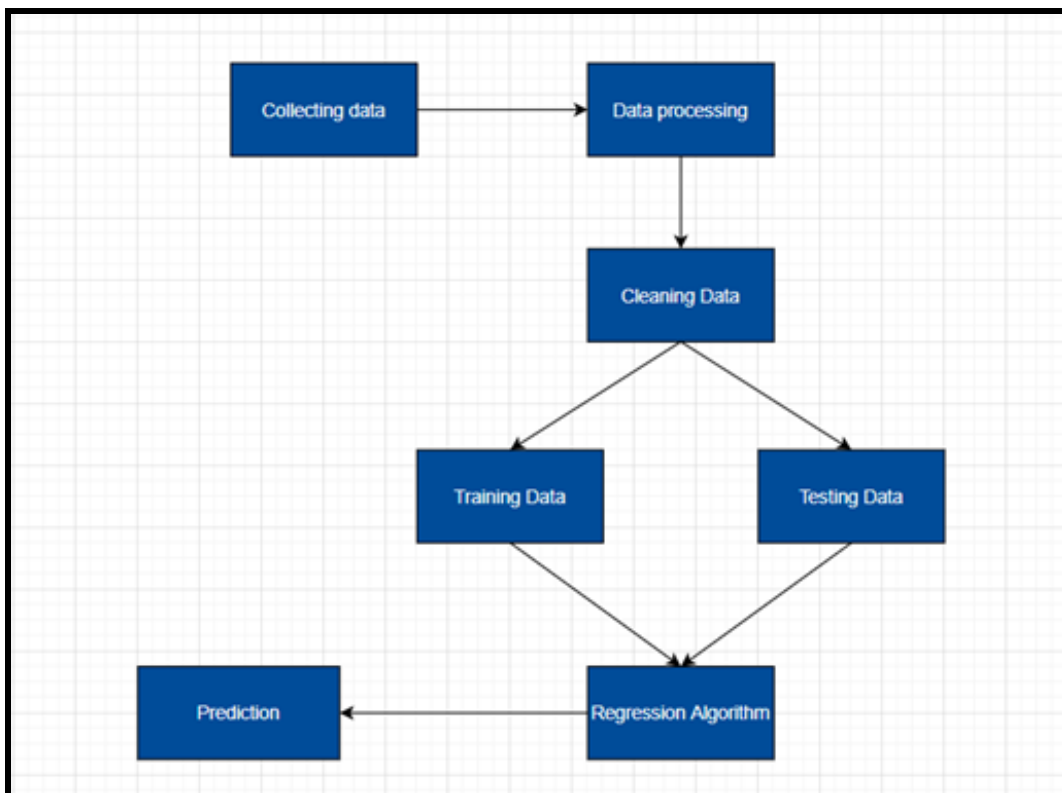
The previous factors were good on a personal level but it is important to know the economic, regional, social and demographic factors like the

population change, birth ratio, foeticide rate, literacy level, levels and dosages of immunizations, history of chronic or genetic diseases, health care facility provided by the country, funds allocated by the government, schemes, cost of medical expenditures (if the cost of basic medical facilities and healthcare is very high, then people will not go to get regular medical checkups done).

For the above problem to be solved, we should have a dataset that consists of various other factors. So the target output variable, that is, expected lifespan of people depends upon variety of factors and not just factors of particular fields. Important immunization like Hepatitis B, Polio and Diphtheria should also be considered.

### 3 THEORETICAL ANALYSIS

#### 3.1 Block Diagrams

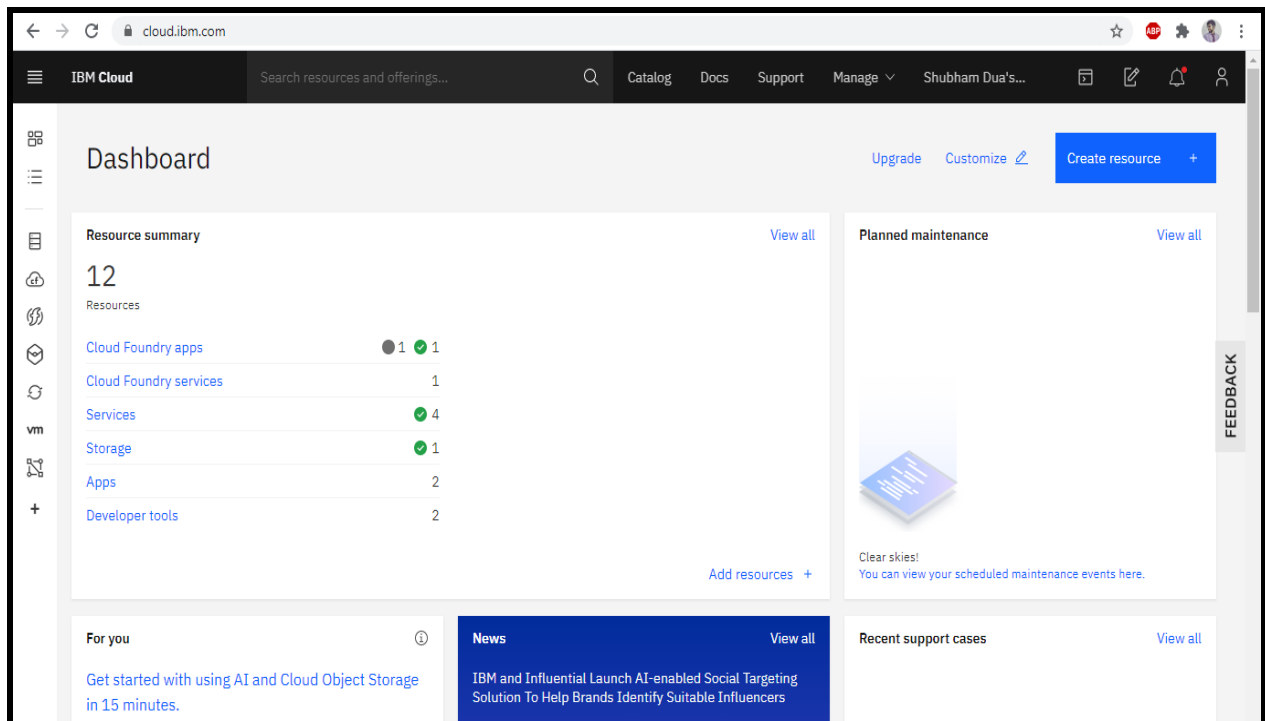


## 3.2 Hardware/Software Designing

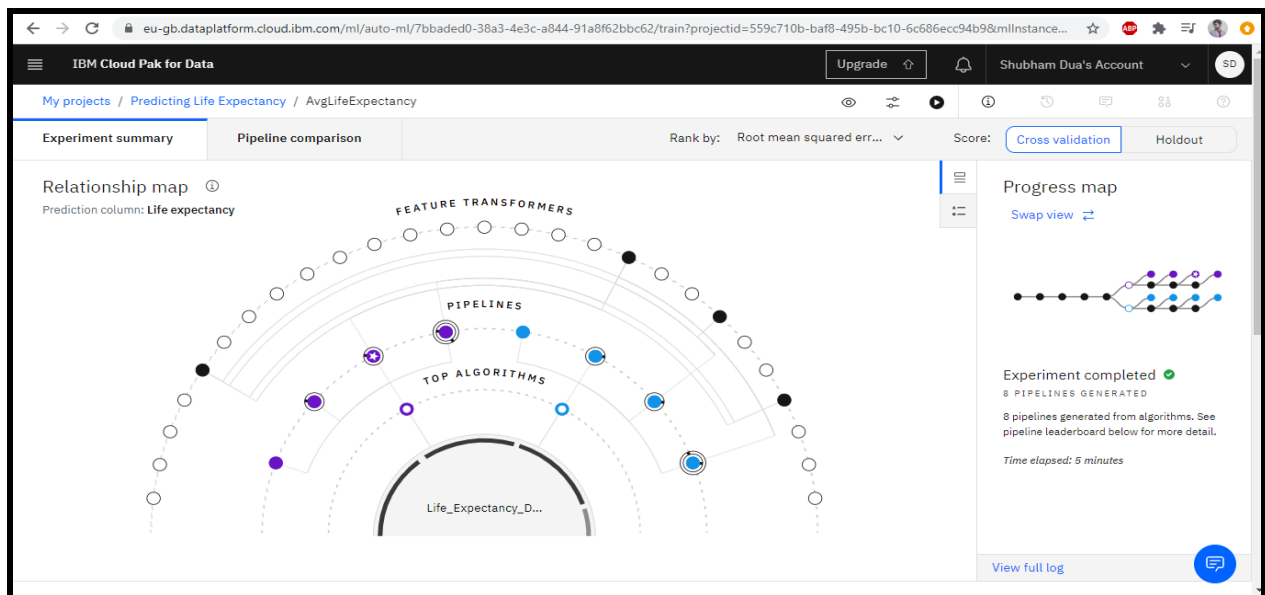
- i. Create necessary IBM Cloud services
- ii. Configure Watson Studio and create Watson studio project
- iii. Create an IBM Machine Learning instance
- iv. Create machine learning model in the Watson notebook
- v. Deploy the machine learning model
- vi. Create a UI using node-red flow and integrate the flow with the Machine Learning model
- vii. Deploy and run the node-red app (it will appear as a "form" to the user)

## 4 EXPERIMENTAL INVESTIGATIONS

### IBM DASHBOARD:



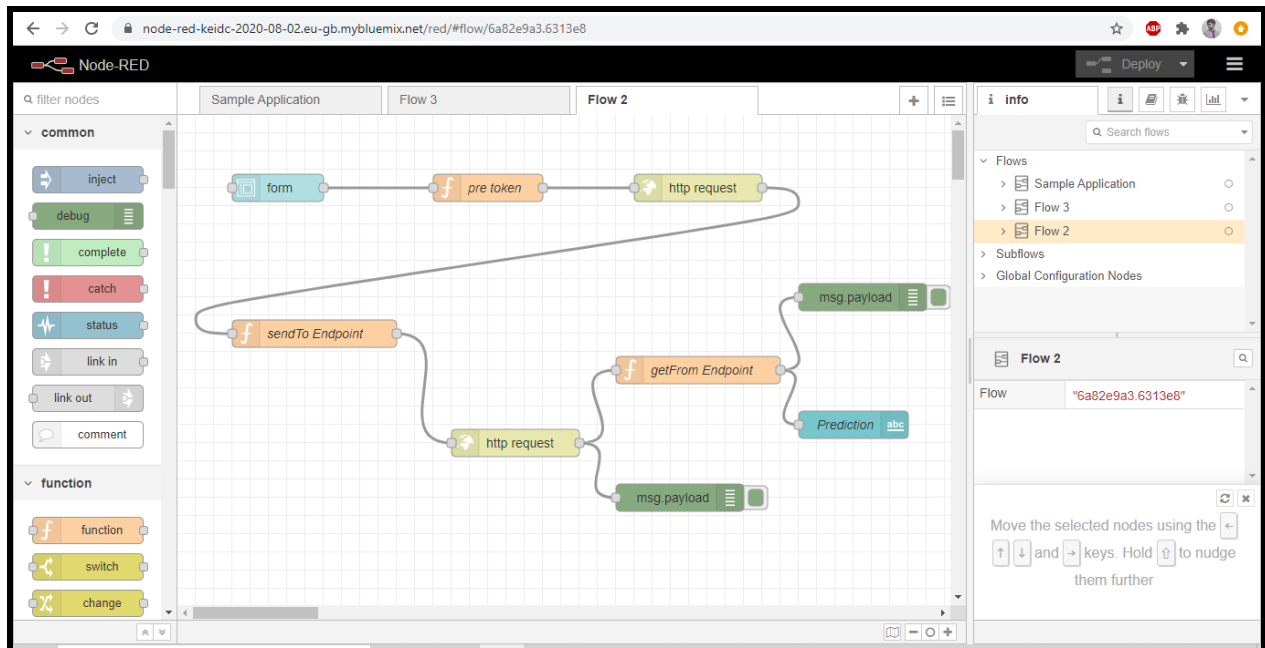
## AUTO AI EXPERIMENT:



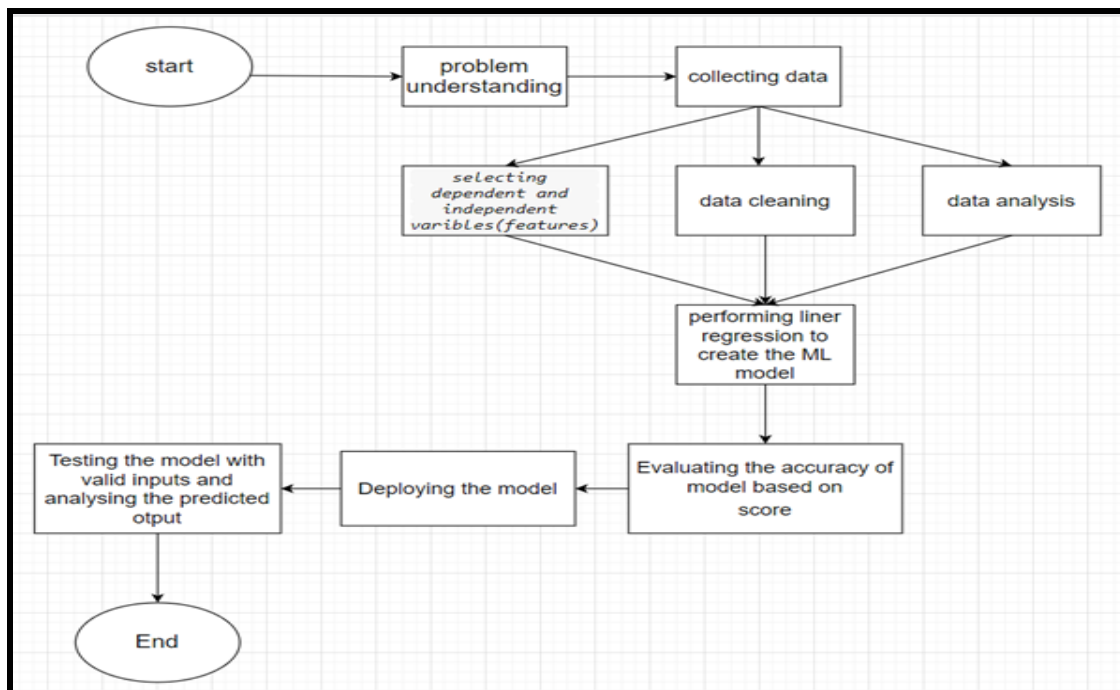
## PIPELINES created by the AUTO AI EXPERIMENT:

Rank	Name	Algorithm	RMSE (Optimized)	Enhancements	Build time
> 1	Pipeline 3	Extra Trees Regressor	2.010	HPO-1 FE	00:00:50
> 2	Pipeline 4	Extra Trees Regressor	2.010	HPO-1 FE HPO-2	00:00:35
> 3	Pipeline 1	Extra Trees Regressor	2.070	None	00:00:01
> 4	Pipeline 2	Extra Trees Regressor	2.070	HPO-1	00:00:11
> 5	Pipeline 7	Decision Tree Regressor	2.732	HPO-1 FE	00:00:38
> 6	Pipeline 8	Decision Tree Regressor	2.732	HPO-1 FE HPO-2	00:00:07
> 7	Pipeline 5	Decision Tree Regressor	2.807	None	00:00:01
> 8	Pipeline 6	Decision Tree Regressor	2.807	HPO-1	00:00:01

## NODE-RED FLOW:



## 5 FLOWCHARTS





## 6 RESULTS

Based on the given data, the auto-AI model or ML model understands the data, and analyzes the factors that affect the results we require, that is, life expectancy. It will predict the output based on the features that we trained. Then when we give any input, based on the trained model, it will validate the features and give the accurate value as predicted output. So the results we get are approximations, they are not 100% accurate, but they are a pretty good indicator of one's life expectancy.

The screenshot shows a web application running in a browser. The address bar displays the URL: `node-red-keidc-2020-08-02.eu-gb.mybluemix.net/ui/#/l/1?socketid=JgaVUI-xhTw1x6MKAAAp`. The page has a green header bar with a hamburger menu icon and the text "Home Page". The main content area is dark-themed and contains a vertical list of input fields for various health and demographic factors. Each field has a label and a numerical value, with a green line indicating the input area. The factors and their values are:

Factor	Value
Year	0
BMI	19.8
Adult_Mortality	0
Infant_Deaths	0
Alcohol	0
Percentage_Expenditure	5.43
HepatitisB	0
Under_Five_Deaths	0
Polio	0
Total_Expenditure	3.45
Diphtheria	0
MMR	0

## 7 ADVANTAGES & DISADVANTAGES

### ADVANTAGES -

- › The life expectancy predictor will give important insights and help people achieve good quality of life in future. The country can plan and improve various healthcare facilities.
- › This model can also help us increase our average lifespan since we now know what all factors positively and negatively affect the life expectancy of a person belonging to a particular country.
- › This project/idea is useful for insurance companies as they consider age, lifestyle choices, family medical history, and several other factors which can be determined with the help of this dataset.

### DISADVANTAGES -

- › First, this dataset and this model is not very easily accessible to all people. Only those people with a stable Internet connection can afford to access this dataset and model.
- › This model can only be used by people who have a good knowledge of how to analyze this data (in order make changes to their own lifestyle).
- › Only those people with access to mobile health monitoring and tracking devices will know how much of their lifestyle has to be changed.
- › The input provided by the user should in the given range to predict good results.
- › A major drawback is that even then, not all results are accurate, and there can easily be errors while analysis and predictions are done using the data.

## 8 APPLICATIONS

- › This will help in suggesting a country which area should be given importance in order to efficiently improve the life expectancy of its population.
- › It will be easier for a country to determine the predicting factor which is contributing to lower value of life expectancy and can be used in various organization to improve the quality of service.
- › The project can be used as a basis to develop personalized health applications.
- › The governments can plan and develop their health infrastructure by keeping the most correlated factors in mind.
- › The project can help governments to keep track of their country's health status so they can plan for the future accordingly.
- › This project/idea is useful for insurance companies as they consider age, lifestyle choices, family medical history, and several other factors which can be determined with the help of this dataset.

## 9 CONCLUSIONS

Thus, we have developed a model that will predict the life expectancy of a person living in a specific region. Various factors like adult mortality, alcohol consumption, population, mortality rate (under the age of 5 years), thinness (through 1-19 years), HIV, Hepatitis B, GDP, literacy level and many others play an important role in the predictions made by this model. The user can interact with the system via a simple UI made by the node-red app.

## 10 FUTURE SCOPE

- › We can connect the model to a database which can predict the life expectancy of not only human beings but also of different species of animals present on the earth.
- › A model with countrywise bifurcation can be made, which will help to segregate the data demographically.
- › Big data and machine learning can benefit public health researchers with analyzing thousands of variables to obtain data regarding life expectancy. We can use demographics of selected regional areas and multiple behavioral health disorders across regions to find correlation between individual behavior indicators and behavioral health outcomes.
- › The accuracy of the model can be increased. This can be done by training more data. Also, the website can be added with many more features to improve the user experience. The user input can be connected to the database for future purposes.

## 11 BIBLIOGRAPHY

1. <https://cloud.ibm.com/>
2. <https://www.kaggle.com/kumarajarshi/life-expectancy-who>
3. <https://bookdown.org/caoying4work/watsonstudio-workshop/jn.html>
4. <https://developer.ibm.com/tutorials/how-to-create-a-node-red-starter-application/>
5. <https://github.com/SmartPracticeschool/IISPS-INT-2459-Predicting-Life-Expectancy-using-Machine-Learning>

## 12 APPENDIX

### A. Source Code

#### Importing required libraries and reading the dataset -

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns
import warnings
warnings.filterwarnings('ignore')
from sklearn import preprocessing, metrics, model_selection, linear_model

# @hidden_cell
# The following code contains the credentials for a file in your IBM Cloud Object Storage.
# You might want to remove those credentials before you share your notebook.
credentials_1 = {
    'IAM_SERVICE_ID': 'iam-ServiceId-57ffe881-6074-44f9-ba71-d4a2ac29d1ff',
    'IBM_API_KEY_ID': 'EA1WL8YcHjISBRFeDtR07CIUBlagQa_Q0YSVNgXcjQ1Z',
    'ENDPOINT': 'https://s3.eu-geo.objectstorage.service.networklayer.com',
    'IBM_AUTH_ENDPOINT': 'https://iam.cloud.ibm.com/oidc/token',
    'BUCKET': 'predictinglifeexpectancyusingmach-donotdelete-pr-wiiomeiwazafry',
    'FILE': 'Life_Expectancy_Data.csv'
}

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_75dedf163fee4ba8ab75cb70eb672d9d = ibm_boto3.client(service_name='s3',
    ibm_api_key_id='EA1WL8YcHjISBRFeDtR07CIUBlagQa_Q0YSVNgXcjQ1Z',
    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_75dedf163fee4ba8ab75cb70eb672d9d.get_object(Bucket='predictinglifeexpectancyusing
mach-donotdelete-pr-wiiomeiwazafry',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 )
```

```
my_data = pd.read_csv(body)
my_data.head()
```

```
len(my_data)
```

```
my_data.info()
```

## **Preprocessing the data -**

```
my_data = my_data.drop('Country', axis=1)
my_data.head()
```

```
my_data.Status.unique()
```

```
my_data.Status = my_data.Status.astype('category')
```

```
encoder = preprocessing.LabelEncoder()
my_data.Status = encoder.fit_transform(my_data.Status)
my_data.Status.unique()
```

```
my_data.isnull().sum()
```

```
my_data = my_data.fillna(0)
my_data.isnull().sum()
```

## Correlation between features -

```
sns.set(rc={'figure.figsize':(20,25)})  
#Create correlation matrix for all variables in the dataframe  
mat = np.triu(my_data.corr())  
sns.heatmap(my_data.corr(), cmap='inferno', linewidths=3, linecolor='black', annot=True,  
square=True , mask=mat)
```

## Train/Test split -

```
X_data = my_data.drop('Life expectancy ', axis=1)  
X_data.head()
```

```
y_data = my_data['Life expectancy ']  
y_data.head()
```

```
scaler = preprocessing.MinMaxScaler()  
X_data = scaler.fit_transform(X_data)
```

```
X_train, X_test, y_train, y_test = model_selection.train_test_split(X_data, y_data, test_size=0.3,  
random_state=101)  
print(X_train.shape)  
print(X_test.shape)  
print(y_train.shape)  
print(y_test.shape)
```

## Creating and training the regression model -

```
model = linear_model.LinearRegression(fit_intercept=True, normalize=True, copy_X=True,  
n_jobs=None)  
model.fit(X_train, y_train)
```

## Making predictions -

```
y_pred = model.predict(X_test)
```

```
sns.set(rc={'figure.figsize':(6,6)})
plt.scatter(y_test, y_pred)

print('Intercept =', model.intercept_)
print('Coefficients = ', model.coef_)

print('Mean Absolute Error: ', metrics.mean_absolute_error(y_test, y_pred))
print('Mean Squared Error: ', metrics.mean_squared_error(y_test, y_pred))
print('Root Mean Squared Error: ', np.sqrt(metrics.mean_squared_error(y_test, y_pred)))
print('R2 score: ', metrics.r2_score(y_test, y_pred))
```

## Deploying the model -

```
!pip install watson-machine-learning-client
client = WatsonMachineLearningAPIClient(wml_credentials)
model_props = {client.repository.ModelMetaNames.AUTHOR_NAME: "Shubham",
               client.repository.ModelMetaNames.AUTHOR_EMAIL: "shubhamdua02@gmail.com",
               client.repository.ModelMetaNames.NAME: "LifeExpectancy_RegressionModel"}

model_artifact = client.repository.store_model(model, meta_props=model_props)
published_model_uid = client.repository.get_model_uid(model_artifact)
published_model_uid

deployment = client.deployments.create(published_model_uid,
name="LifeExpectancy_RegressionModel")

scoring_endpoint = client.deployments.get_scoring_url(deployment)
scoring_endpoint
```