PROJECT REPORT On

LIFE EXPECTANCY ANALYSIS & PREDICTION

SUBMITTED BY

MYTHEESH C

ABSTRACT

Life Expectancy is an important metric to assess the health of a nation. This project presents a comparative analysis of life expectancy between developed and developing countries with the help of a Supervised Machine Learning model. The prediction model is trained using three regression models, namely Linear Regression, Decision Tree Regressor and Random Forest Regressor. The selection of model is done on the basis of R 2 score, Mean Squared Error & Mean Absolute Error. Random Forest Regressor is selected for the development of the prediction model for life expectancy. The comparative analysis is done on the basis of HIV/AIDS, Adult Mortality and Expenditure on Healthcare, as they are the important features suggested by the model. The study undertaken suggests that, developed countries have high life expectancy as compared to developing countries. India has high adult mortality as compared to considered developed countries because of the low expenditure on healthcare. The insights from this analysis can be used by Government and Healthcare sectors for the betterment of society.

ACKNOWLEDGEMENT

I am using this opportunity to express my gratitude to everyone who supported me throughout the course of my capstone project. I am thankful for their aspiring guidance, invaluably constructive criticism and friendly advice during the project work. I am sincerely grateful to them for sharing their truthful and illuminating views on a number of issues related to the project.

Further, I have fortunate to have Mr. K. PRASAD as my mentor. He has readily shared his immense knowledge in data science and guides me in a manner that the outcome resulted in enhancing my data skills.

I certify that the work done by me for conceptualizing and completing this project is original and authentic.

Date: July 26, 2022. Name: MYTHEESH C

CERTIFICATE OF COMPLETION

I hereby certify that the project titled “**Life Expectancy Analysis and Prediction**” using ML and PowerBI was undertaken and completed the project on 11th July, 2022.

Mentor: Mr. K. Prasad
Date: 10th July,2022
Place: Karur

TABLE OF CONTENT

1.INTRODUCTION

1.1 OVERVIEW

This project "Life Expectancy Analysis and Prediction" using Machine Learning & PowerBI is a WEB-UI based application that predict the expected average life span of people of a given country based on various features. This project model is built using Flask server. 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 and provides comparative analysis of life expectancy between developed and developing countries. 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.

- Project Requirements: Jupyter Notebook, Power BI, MySQL - Connector.
- Functional Requirements: Flask Server, MySQL Server.
- Technical Requirements: Machine Learning, HTML.
- Software Requirements: Py Charm, MySQL Workbench.

1.2 NEED OF THE STUDY

Predicting a human's life expectancy has been a long-term question to humankind. Past work to generate health-focused forecasts includes that from the UN Population Division, and the Austrian Wittgenstein Centre, which produces life expectancy forecasts with different scenarios to the end of the 21st century. There are so many organizations that are making research in the prediction of life expectancy. Many calculations and research papers have been done and published to create an equation despite it being impractical to simplify these variables into one equation. Currently there are various smart devices and applications such as smartphone apps and wearable devices that provide wellness and fitness tracking. Some apps provide health related data such as sleep monitoring, heart rate measuring, and calorie expenditure collected and processed by the devices and servers in the cloud. However no existing works provide the Personalized Life expectancy. The World Health Organization (WHO) used to produce annual life tables for the countries but after 2011 it said to shift for two-year cycle for the updating of life tables and even still the model is not really updated in every fields. WHO applies standard methods to the analysis of Member State data to ensure comparability of estimates across the countries. This will inevitably result in differences for some Member States with official estimates for quantities such as life expectancy, where a variety of different projection methods and other methods are used.

Life expectancy is the most important factor for 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.

.

1.3 PROPOSED SOLUTION:

Some of the past research was done considering multiple linear regression based on dataset of one year or two years for all the countries. We can resolve it by formulating a regression model while considering data from a period of year 2000 to 2015 for all the countries. Important immunization like Hepatitis B, HIV/AIDS, Polio and Diphtheria will also be considered. We will also focus on Adult Mortality, Alcohol intake, percentage expenditure, Measles, BMI, Death of under 5 years, Schooling, thinness in 1-19 years and 5-9 years and Population related factors as well. Since the dataset is based on different countries, it will be easier for a country to determine the predicting factor which is contributing to value of life expectancy. This will help in predicting the life expectancy (in years) of its population. Foe the solution, first we will examine the dataset provided by World Health Organization (WHO) and find patterns in a dataset and we will attempt to fit various algorithms such as linear Regression and Random Forest Regression Algorithm, Decision Tree Regression Algorithm and see what gives less error and good prediction score. And to pick a good machine learning algorithm. For better usability and input operations we are also going to create a WEB-UI with the help of Flask Server for users.

Front-end: A web page taking the necessary inputs from the user to predict the life expectancy in years.

Back-end: User given input gets processed according to the trained Model and finally gives the desired output of life expectancy.

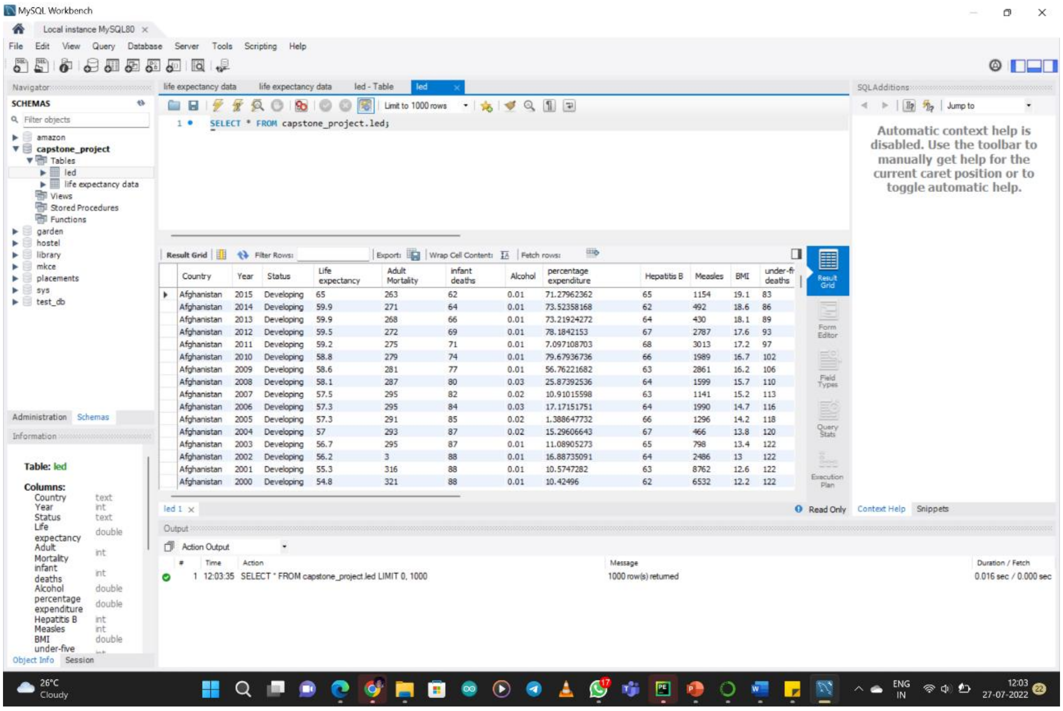
1.4 DATA SOURCES

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 The datasets are made available to public for the purpose of health data analysis. The dataset related to life expectancy, health factors for 193 countries has been collected from the same WHO data repository website and its corresponding economic data was collected from United Nation website. Among all categories of health-related factors only those critical factors were chosen which are more representative. It has been observed that in the past 15 years, there has been a huge development in health sector resulting in improvement of human mortality rates especially in the developing nations in comparison to the past 30 years. Therefore, in this project we have considered data from year 2000-2015 for 193 countries for further analysis. The individual data files have been merged together into a single dataset. On initial visual inspection of the data showed some missing values. As the datasets were from WHO, we found no evident errors. Missing data was handled in R software by using Miss-map command. The result indicated that most of the missing data was for population, Hepatitis B and GDP. The missing data were from less known countries like Vanuatu, Tonga, Togo, Cabo Verde etc. Finding all data for these countries was difficult and hence, it was decided that we exclude these countries from the final model dataset. The final merged file (final dataset) consists of 22 Columns and 2938 rows which meant 20 predicting variables. All predicting variables was then divided into several broad categories:Immunization related factors, Mortality factors, Economical factors and social factors.

Data Source Link:https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who

2.DATA UNDERSTANDING AND PREPARATION

Data preparation is the process of cleaning and transforming raw data before building predictive models. Dataset for Life Expectancy prediction is found from various internet sources. Finally, there are nearly 3000 data of different countries is collected and stored in a dataset using MySQL workbench.



```
SELECT * FROM capstone_project.led;
```

Country	Year	Status	Life expectancy	Adult Mortality	infant deaths	Alcohol	percentage expenditure	Hepatitis B	Measles	BMI	under-fi deaths
Afghanistan	2015	Developing	65	263	62	0.01	71.27962362	65	1154	19.1	83
Afghanistan	2014	Developing	59.9	271	64	0.01	73.52358168	62	492	18.6	86
Afghanistan	2013	Developing	59.9	268	66	0.01	73.21924272	64	430	18.1	89
Afghanistan	2012	Developing	59.5	272	69	0.01	78.1842153	67	2787	17.6	93
Afghanistan	2011	Developing	59.2	275	71	0.01	7.097108703	68	3013	17.2	97
Afghanistan	2010	Developing	58.8	279	74	0.01	79.67936736	66	1989	16.7	102
Afghanistan	2009	Developing	58.6	281	77	0.01	56.76221682	63	2861	16.2	106
Afghanistan	2008	Developing	58.1	287	80	0.03	25.87392536	64	1599	15.7	110
Afghanistan	2007	Developing	57.5	295	82	0.02	10.91015598	63	1141	15.2	113
Afghanistan	2006	Developing	57.3	295	84	0.03	17.17151751	64	1990	14.7	116
Afghanistan	2005	Developing	57.3	291	85	0.02	1.388647732	66	1296	14.2	118
Afghanistan	2004	Developing	57	293	87	0.02	15.29606643	67	466	13.8	120
Afghanistan	2003	Developing	56.7	295	87	0.01	11.08905273	65	798	13.4	122
Afghanistan	2002	Developing	56.2	3	88	0.01	16.88735091	64	2486	13	122
Afghanistan	2001	Developing	55.3	316	88	0.01	10.5747282	63	8762	12.6	122
Afghanistan	2000	Developing	54.8	321	88	0.01	10.42496	62	6532	12.2	122

```
#  Time      Action                                          Message                Duration / Fetch
1  12:03:35  SELECT * FROM capstone_project.led LIMIT 0, 1000  1000 row(s) returned   0.016 sec / 0.000 sec
```

Fig 2: MySQL Workbench

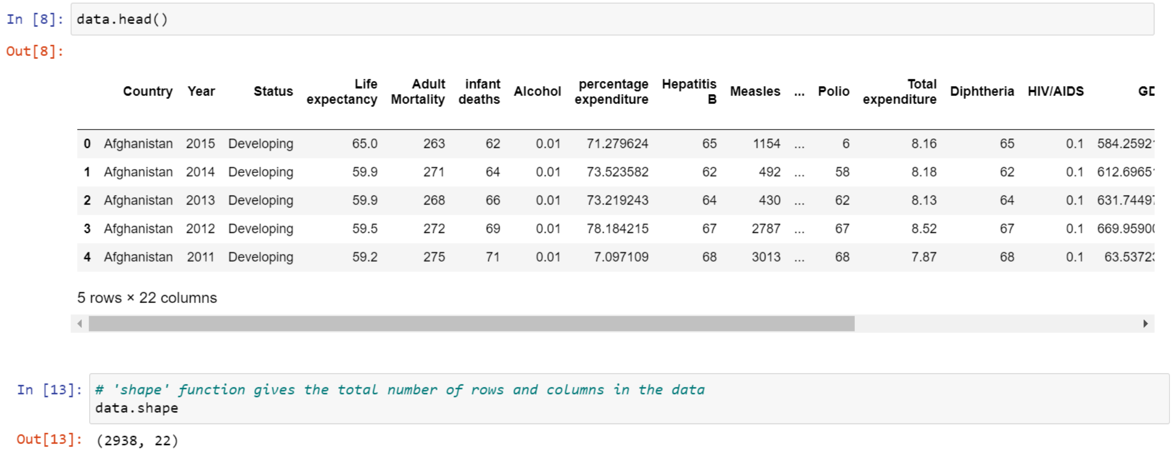
Firstly, we will investigate if there is any null/missing value then fill it with mean value then make every feature of same type i.e. integer or floats then the country, status, year column is changes into numerical values by using label encoder for prediction besides for comparative analysis year and status column is dropped and the dataset is ordered by countries as it will not be used in the analysis. By observing the data, we came to know the data contains 21 columns and 2938 rows with the header row and concluded that there are various Factors affecting Life Expectancy of a country such as:

1. Country’s Adult Mortality
2. Number of Infant Deaths
3. Alcohol consumption
4. Expenditure on health
5. Hepatitis B immunization
6. Measles reported cases
7. Average Body Mass Index of the entire population
8. Number of dead under-five years
9. Polio immunization coverage
10. Government expenditure on health
11. Diphtheria immunization coverage
12. HIV/AIDS cases
13. GDP (Gross Domestic Product of the country)
14. Population of the country
15. Thinness among children for Age 5 to 9
16. Thinness among children for Age 10 to 19
17. Income composition of resources
18. Schooling

To pick a good machine learning algorithm we must need to consider to first check the data and solve the problem within it such as: whether if any value is missing or not? If missing, can we remove that feature or can we put the mean value there? Is all values are of same type? Does object value pay any contribution? If no then can we remove it or if yes then can we change object values into categorical value such as integer? And other problematic factors. We must need to remember, the algorithm must not attempt to infer the function that exactly matches all the data. Being not careful in fitting the data can cause over-fitting, after which the model will answer perfectly for all training examples but will have a very high error for Unseen samples.

2.1 UNDERSTAND THE DATASET

The imported data has been viewed using head() and the number of rows and columns using shape for confirmation.



```
In [8]: data.head()
Out[8]:
```

	Country	Year	Status	Life expectancy	Adult Mortality	infant deaths	Alcohol	percentage expenditure	Hepatitis B	Measles	...	Polio	Total expenditure	Diphtheria	HIV/AIDS	GD
0	Afghanistan	2015	Developing	65.0	263	62	0.01	71.279624	65	1154	...	6	8.16	65	0.1	584.2592
1	Afghanistan	2014	Developing	59.9	271	64	0.01	73.523582	62	492	...	58	8.18	62	0.1	612.6965
2	Afghanistan	2013	Developing	59.9	268	66	0.01	73.219243	64	430	...	62	8.13	64	0.1	631.7449
3	Afghanistan	2012	Developing	59.5	272	69	0.01	78.184215	67	2787	...	67	8.52	67	0.1	669.9590
4	Afghanistan	2011	Developing	59.2	275	71	0.01	7.097109	68	3013	...	68	7.87	68	0.1	63.5372

5 rows × 22 columns

```
In [13]: # 'shape' function gives the total number of rows and columns in the data
         data.shape
Out[13]: (2938, 22)
```

Fig 2.1.1: Understand the Dataset

Data Type:

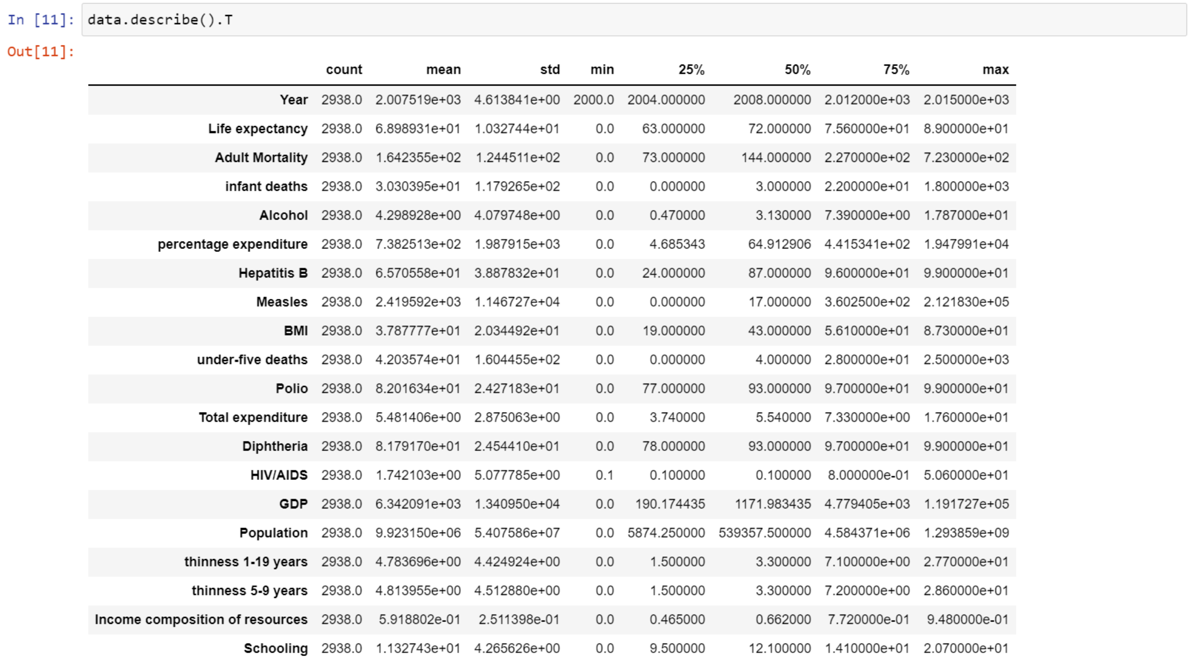
The main data types in Pandas data frames are the object, float and int64. To understand each attribute of our data, it is always good for us to know the data type of each column.

In our dataset, we have numerical and categorical variables. The numeric variables should have data type 'int'/'float' while categorical variables should have data type 'object'.

Summary Statistics:

The below figure illustrates the summary statistics of all the numeric variables namely mean, median (50%), standard deviation, minimum, and maximum values, along with the first and third quantiles.

For example, the average age of a person considered in the study is 34 years, where the minimum age is 21 years and the maximum age is 81 years.



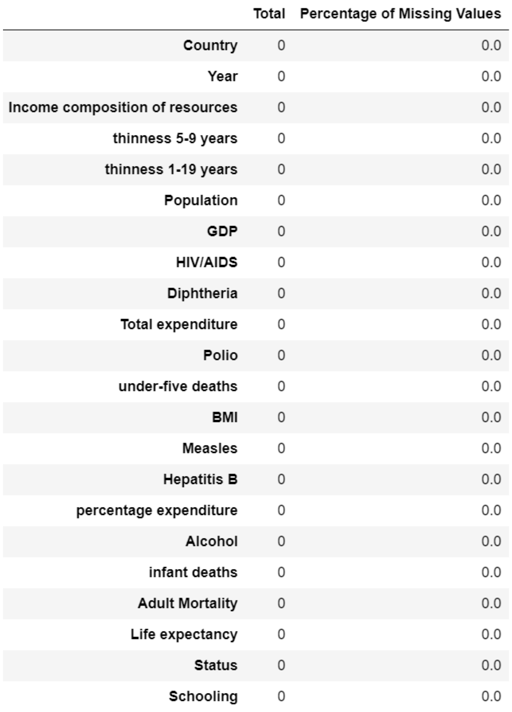
```
In [11]: data.describe().T
Out[11]:
```

	count	mean	std	min	25%	50%	75%	max
Year	2938.0	2.007519e+03	4.613841e+00	2000.0	2004.000000	2008.000000	2.012000e+03	2.015000e+03
Life expectancy	2938.0	6.898931e+01	1.032744e+01	0.0	63.000000	72.000000	7.560000e+01	8.900000e+01
Adult Mortality	2938.0	1.642355e+02	1.244511e+02	0.0	73.000000	144.000000	2.270000e+02	7.230000e+02
infant deaths	2938.0	3.030395e+01	1.179265e+02	0.0	0.000000	3.000000	2.200000e+01	1.800000e+03
Alcohol	2938.0	4.298928e+00	4.079748e+00	0.0	0.470000	3.130000	7.390000e+00	1.787000e+01
percentage expenditure	2938.0	7.382513e+02	1.987915e+03	0.0	4.685343	64.912906	4.415341e+02	1.947991e+04
Hepatitis B	2938.0	6.570558e+01	3.887832e+01	0.0	24.000000	87.000000	9.600000e+01	9.900000e+01
Measles	2938.0	2.419592e+03	1.146727e+04	0.0	0.000000	17.000000	3.602500e+02	2.121830e+05
BMI	2938.0	3.787777e+01	2.034492e+01	0.0	19.000000	43.000000	5.610000e+01	8.730000e+01
under-five deaths	2938.0	4.203574e+01	1.604455e+02	0.0	0.000000	4.000000	2.800000e+01	2.500000e+03
Polio	2938.0	8.201634e+01	2.427183e+01	0.0	77.000000	93.000000	9.700000e+01	9.900000e+01
Total expenditure	2938.0	5.481406e+00	2.875063e+00	0.0	3.740000	5.540000	7.330000e+00	1.760000e+01
Diphtheria	2938.0	8.179170e+01	2.454410e+01	0.0	78.000000	93.000000	9.700000e+01	9.900000e+01
HIV/AIDS	2938.0	1.742103e+00	5.077785e+00	0.1	0.100000	0.100000	8.000000e-01	5.060000e+01
GDP	2938.0	6.342091e+03	1.340950e+04	0.0	190.174435	1171.983435	4.779405e+03	1.191727e+05
Population	2938.0	9.923150e+06	5.407586e+07	0.0	5874.250000	539357.500000	4.584371e+06	1.293859e+09
thinness 1-19 years	2938.0	4.783696e+00	4.424924e+00	0.0	1.500000	3.300000	7.100000e+00	2.770000e+01
thinness 5-9 years	2938.0	4.813955e+00	4.512880e+00	0.0	1.500000	3.300000	7.200000e+00	2.860000e+01
Income composition of resources	2938.0	5.918802e-01	2.511398e-01	0.0	0.465000	0.662000	7.720000e-01	9.480000e-01
Schooling	2938.0	1.132743e+01	4.265626e+00	0.0	9.500000	12.100000	1.410000e+01	2.070000e+01

Fig 2.1.2: Summary Statistics

Missing Values:

Check for the presence of missing values and their percentage for each column. Then choose the right approach to remove them.



	Total	Percentage of Missing Values
Country	0	0.0
Year	0	0.0
Income composition of resources	0	0.0
thinness 5-9 years	0	0.0
thinness 1-19 years	0	0.0
Population	0	0.0
GDP	0	0.0
HIV/AIDS	0	0.0
Diphtheria	0	0.0
Total expenditure	0	0.0
Polio	0	0.0
under-five deaths	0	0.0
BMI	0	0.0
Measles	0	0.0
Hepatitis B	0	0.0
percentage expenditure	0	0.0
Alcohol	0	0.0
infant deaths	0	0.0
Adult Mortality	0	0.0
Life expectancy	0	0.0
Status	0	0.0
Schooling	0	0.0

Fig 2.1.3: Missing Values

Plot the Heatmap – Visualization of Missing Values. From the below figure, we determine

- The horizontal lines in the heatmap correspond to the missing values. But there is no such line. This means there are no missing values.

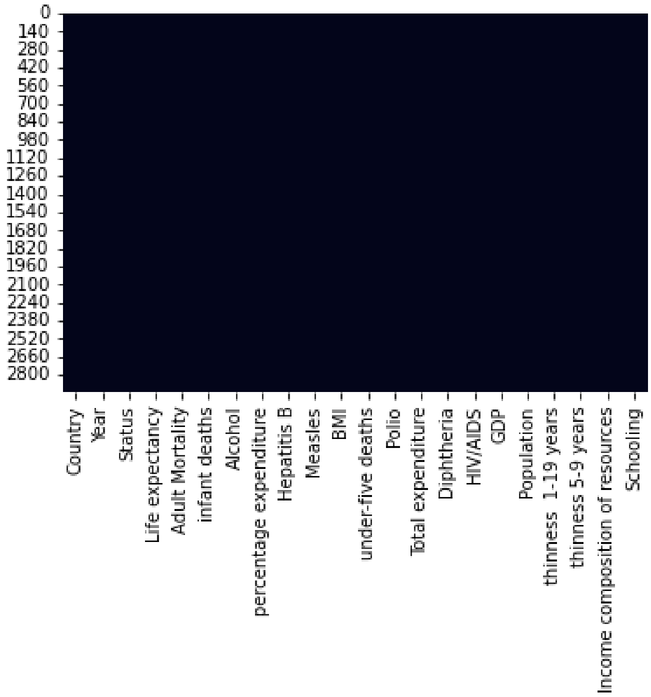


Fig 2.1.4: Heatmap - Missing Values

2.2 DATA PREPROCESSING

By applying correlation function of Panda library, we can see the individual average relation between the features and Life expectancy which also tell us about the factors of the countries on which the life expectancy of its citizen depends. It is observable that Schooling, Income composition of resources, BMI, country’s Status (whether its Developing or Developed), Diphtheria, Polio, Alcohol, Percentage expenditure, Total expenditure, Hepatitis B and Year plays and highly positive correlated with Life expectancy of that country and improving in it may leads to extend the life span of the people of that areas as well as we can also see some feature like Population, Measles, infant deaths, thinness 5-9 years, thinness 1-19 years, HIV/AIDS and Adult Mortality are negatively related to the life expectancy of the people, means in order to increase life span we must need to decrease/improve these features

Below with the help of correlation heat map of the data-set we can a compact understanding of relation between the individual features and how there are inter-related in the scale of 0.5 to 1.0. Where the shades of light represent high correlation while the shades of dark represent poor correlations.

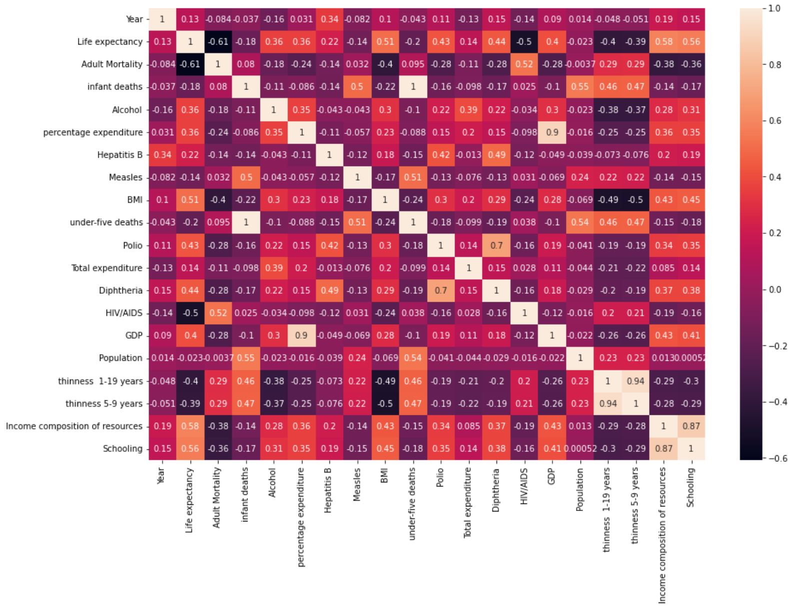


Fig 2.2.1: Correlation Heatmap

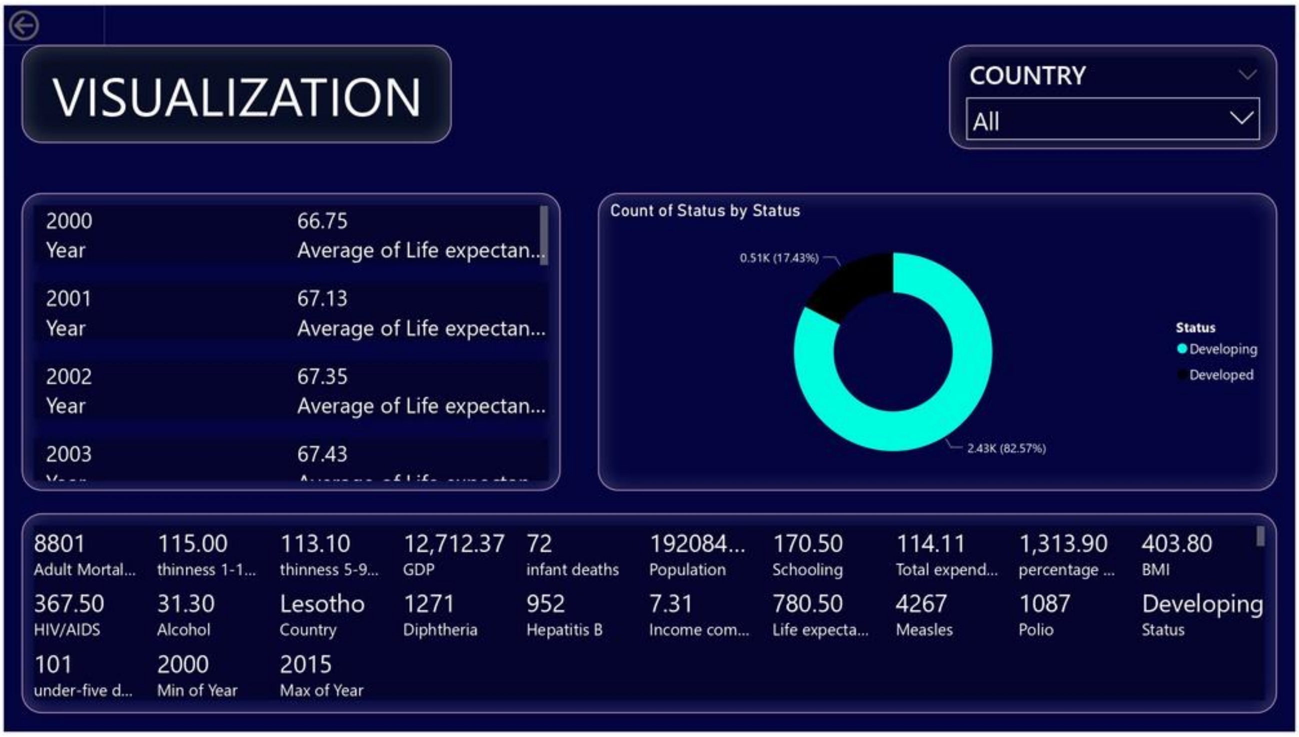
Here we got the round understanding for the factoring on which life expectancy depend and how much government must need to improve. As well as the relation between different features. Now, as we look at data and understand it not a classification problem because the target variable is not categorical (i.e. the output cannot be classified into classes — like it belongs to either Class A or B or something else). Here we have to find the value of life expectancy that is why we will Implement Regression Models for the prediction of life expectancy.

2.3 VISUALIZATION OF DATA

PowerBI Report:

To get our Power BI analytics in a Jupyter notebook with the new PowerBI client Python package. The new package lets you embed Power BI reports in Jupyter notebooks easily. We will be able to export data from visuals in a Power BI report to the Jupyter notebook for in-depth data exploration. We can also filter the report for quick analysis or use bookmarks to apply a saved view.

Fig 2.3.1: PowerBI Report



VISUALIZATION

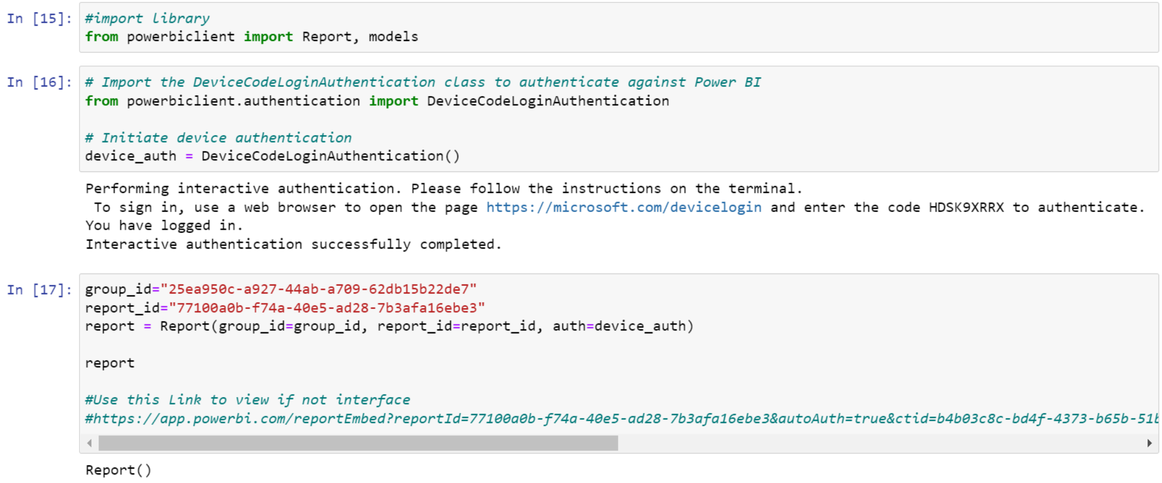
COUNTRY
All

2000 Year — 66.75 Average of Life expectan...
2001 Year — 67.13 Average of Life expectan...
2002 Year — 67.35 Average of Life expectan...
2003 Year — 67.43 Average of Life expectan...

Count of Status by Status

8801 Adult Mortal...
115.00 thinness 1-1...
113.10 thinness 5-9...
12,712.37 GDP
72 infant deaths
192084... Population
170.50 Schooling
114.11 Total expend...
1,313.90 percentage ...
403.80 BMI
367.50 HIV/AIDS
31.30 Alcohol
Lesotho Country
1271 Diphtheria
952 Hepatitis B
7.31 Income com...
780.50 Life expecta...
4267 Measles
1087 Polio
Developing Status
101 under-five d...
2000 Min of Year
2015 Max of Year

The following code is used to integrate PowerBI report to Jupyter Notebook.



```
In [15]: #import library
         from powerbiclient import Report, models

In [16]: # Import the DeviceCodeLoginAuthentication class to authenticate against Power BI
         from powerbiclient.authentication import DeviceCodeLoginAuthentication

         # Initiate device authentication
         device_auth = DeviceCodeLoginAuthentication()

         Performing interactive authentication. Please follow the instructions on the terminal.
          To sign in, use a web browser to open the page https://microsoft.com/devicelogin and enter the code HDSK9XRRX to authenticate.
         You have logged in.
         Interactive authentication successfully completed.

In [17]: group_id="25ea950c-a927-44ab-a709-62db15b22de7"
         report_id="77100a0b-f74a-40e5-ad28-7b3afa16ebe3"
         report = Report(group_id=group_id, report_id=report_id, auth=device_auth)

         report

         #Use this Link to view if not interface
         #https://app.powerbi.com/reportEmbed?reportId=77100a0b-f74a-40e5-ad28-7b3afa16ebe3&autoAuth=true&ctid=b4b03c8c-bd4f-4373-b65b-51

         Report()
```

Fig 2.3.2: PowerBI Report code

Besides, we interfaced PowerBI report for some statistical analysis to following questions by using PowerBI client to display the report in Jupyter notebook.

1. How does Infant and Adult mortality rates affect life expectancy?
2. Do densely populated countries tend to have lower life expectancy?
3. What is the impact of Immunization coverage on life Expectancy?

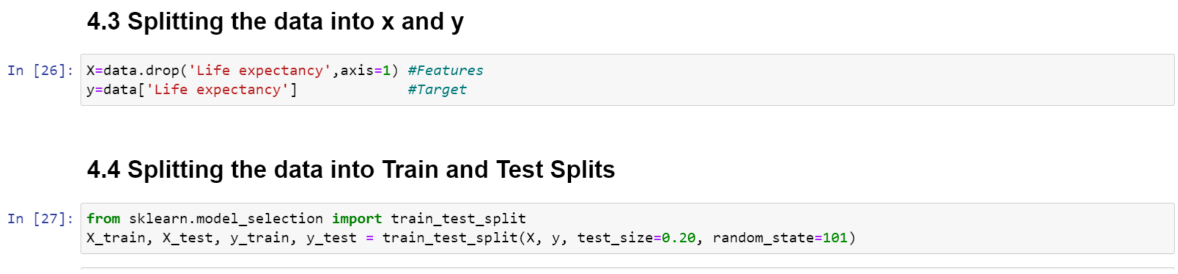
2.4 FEATURE SELECTION

The variables which we already discussed context (various factors) are taken into account of variable here. Some of the variables are,

- Country
- Year
- Status
- Life expectancy

- Adult Mortality
- Percentage expenditure
- Diphtheria
- GDP

In this project "Life expectancy" has been taken as a **target** variable rest of the columns containing various factors have been taken as a **feature** variable for target. Split the data for further analysis.



4.3 Splitting the data into x and y

```
In [26]: X=data.drop('Life expectancy',axis=1) #Features
         y=data['Life expectancy']            #Target
```

4.4 Splitting the data into Train and Test Splits

```
In [27]: from sklearn.model_selection import train_test_split
         X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.20, random_state=101)
```

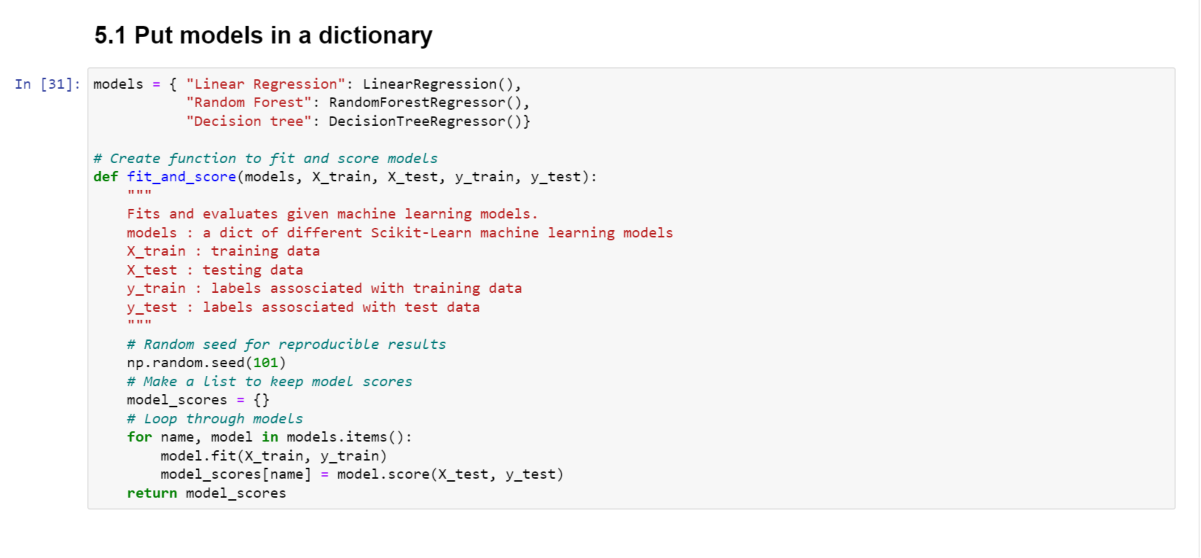
Fig 2.4.1: Feature Selection

3.FITTING MODELS TO DATA

3.1 FINDING THE MOST SUITABLE ALGORITHMM

With the access of Linear Regression, Random Forest [Regressor], Decision Tree [Regressor] are various algorithms, where used to predict the data and are evaluated by R square, Mean Absolute Error and Mean Square Error method.

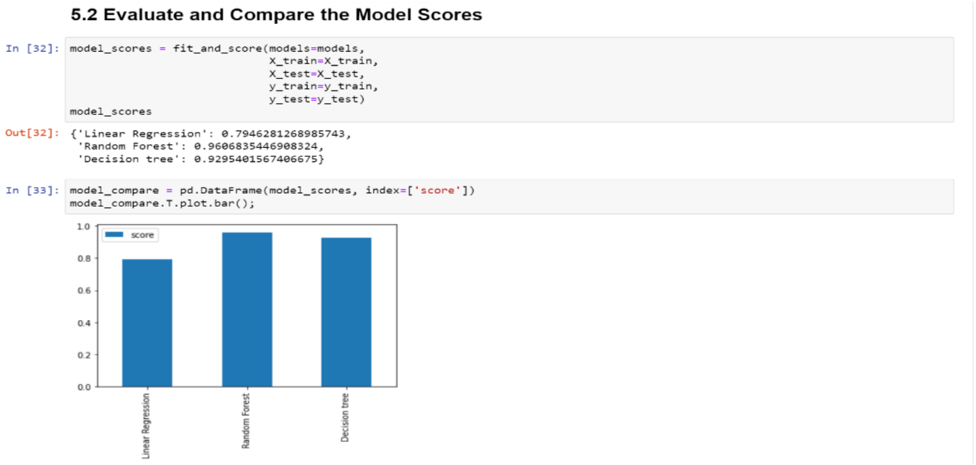
Fig 3.1.1: ML model



5.1 Put models in a dictionary

```
In [31]: models = { "Linear Regression": LinearRegression(),
                    "Random Forest": RandomForestRegressor(),
                    "Decision tree": DecisionTreeRegressor()}

         # Create function to fit and score models
         def fit_and_score(models, X_train, X_test, y_train, y_test):
             """
             Fits and evaluates given machine learning models.
             models : a dict of different Scikit-Learn machine learning models
             X_train : training data
             X_test : testing data
             y_train : labels assosciated with training data
             y_test : labels assosciated with test data
             """
             # Random seed for reproducible results
             np.random.seed(101)
             # Make a list to keep model scores
             model_scores = {}
             # Loop through models
             for name, model in models.items():
                 model.fit(X_train, y_train)
                 model_scores[name] = model.score(X_test, y_test)
             return model_scores
```

Evaluate And Compare the Model:



5.2 Evaluate and Compare the Model Scores

```
In [32]: model_scores = fit_and_score(models=models,
                                      X_train=X_train,
                                      X_test=X_test,
                                      y_train=y_train,
                                      y_test=y_test)
         model_scores

Out[32]: {'Linear Regression': 0.7946281268985743,
          'Random Forest': 0.9606835446908324,
          'Decision tree': 0.9295401567406675}

In [33]: model_compare = pd.DataFrame(model_scores, index=['score'])
         model_compare.T.plot.bar();
```

Fig 3.1.2: Model comparison

Accuracy comparison are made between three model. The below figure shows Random Forest Regressor model has high accuracy score. Apart from those algorithms there are some other resources that are being used here such as Randomized Search CV for hyperparameter tuning [Random Forest regressor]. Besides we find the feature importance of the model and plot feature importance graph.

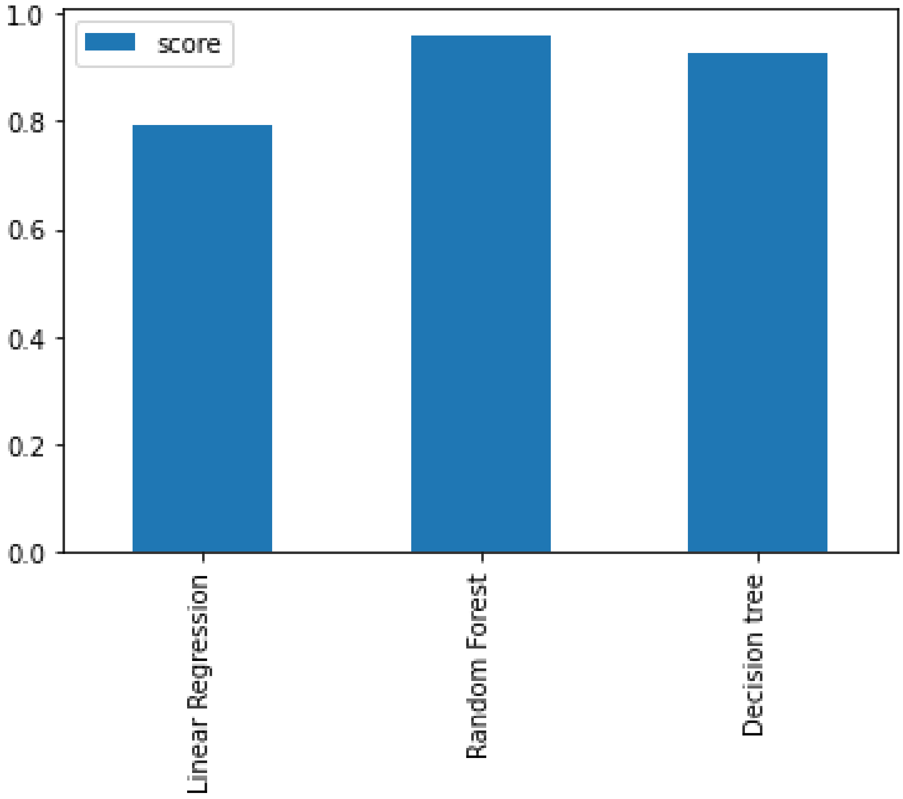


Fig 3.1.3: Model Scores

Model Score:

1. Linear Regression: 0.7946281268985743
2. Random Forest: 0.9606835446908324
3. Decision tree: 0.9295401567406675

Feature Importance:

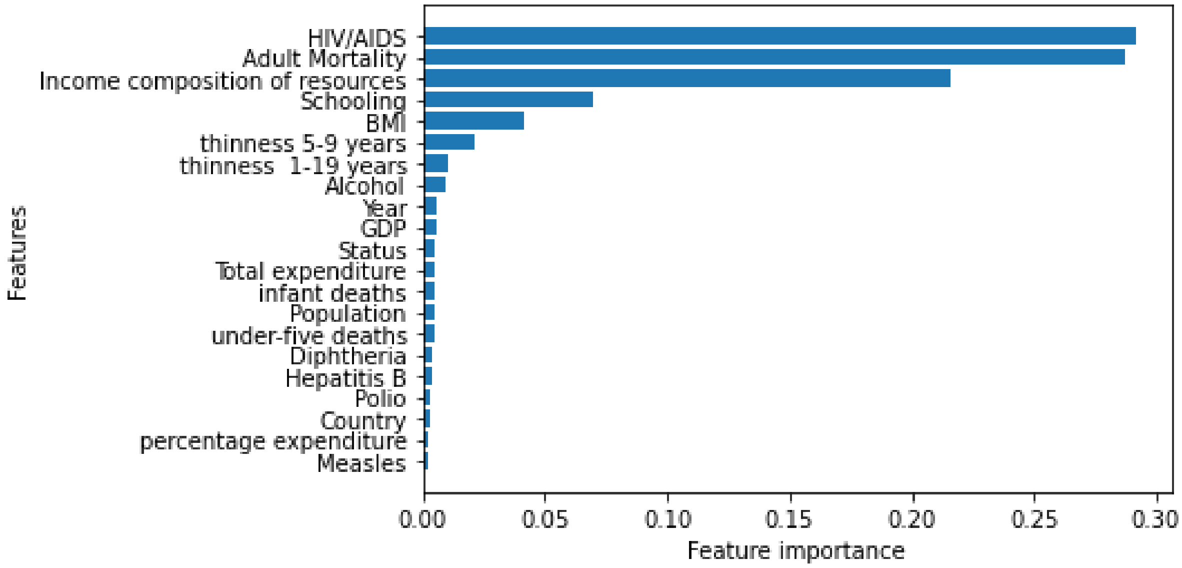


Fig 3.1.4: Feature Importance

From the Feature importance, the comparative analysis of life expectancy between developed and developing countries are made.

The Comparison made are as follows:

1. Comparison of Avg. HIV/AIDS (Deaths per 1000 live births, 0-4 years) among developed & developing countries (2000-2015).

2. Comparison of Avg. Adult Mortality (probability of dying between 15 and 60 years per 1000 population) among developed & developing countries (2000-2015).

3. Comparison of Avg. Life Expectancy among developed & developing countries (2000-2015).

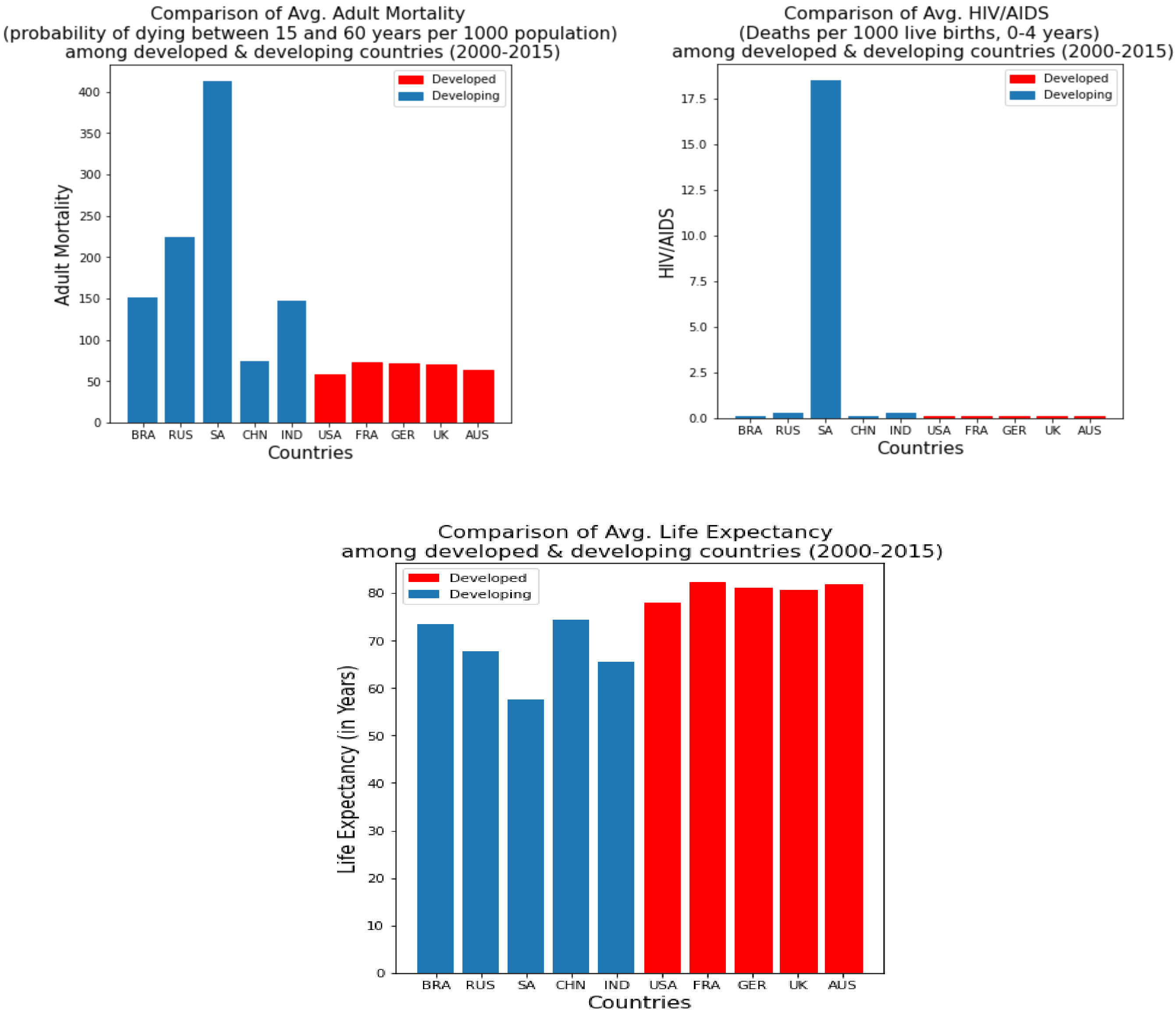


Fig 3.1.4: Comparative Analysis

4.MODEL DEPLOYMENT

4.1 FLASK SERVER

Flask is a web application framework written in Python. It was developed by Armin Ronacher, who led a team of international Python enthusiasts called Poocco. Flask is based on the Werkzeg WSGI toolkit and the Jinja2 template engine. Both are Pocco projects.

To deploy ML model by connected my backend "Life Expectancy Prediction" model with frontend Web pages by using FLASK web python framework. So, we can directly get the query from the user in Webpage, where it sends the query to the Flask server which get the predicted model from ML model and then send the predicted label to webpage.

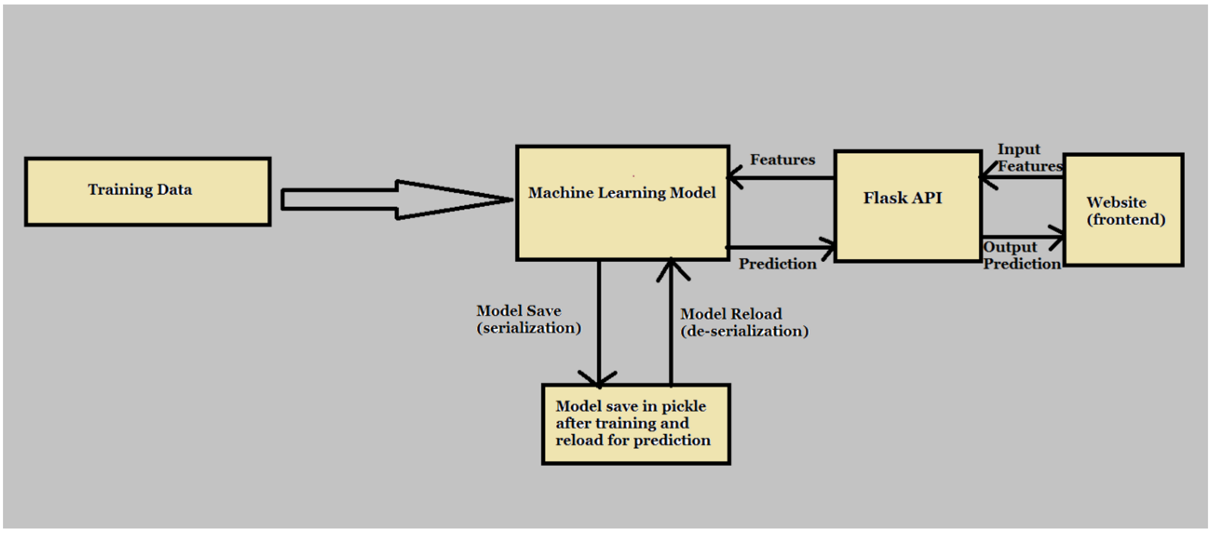
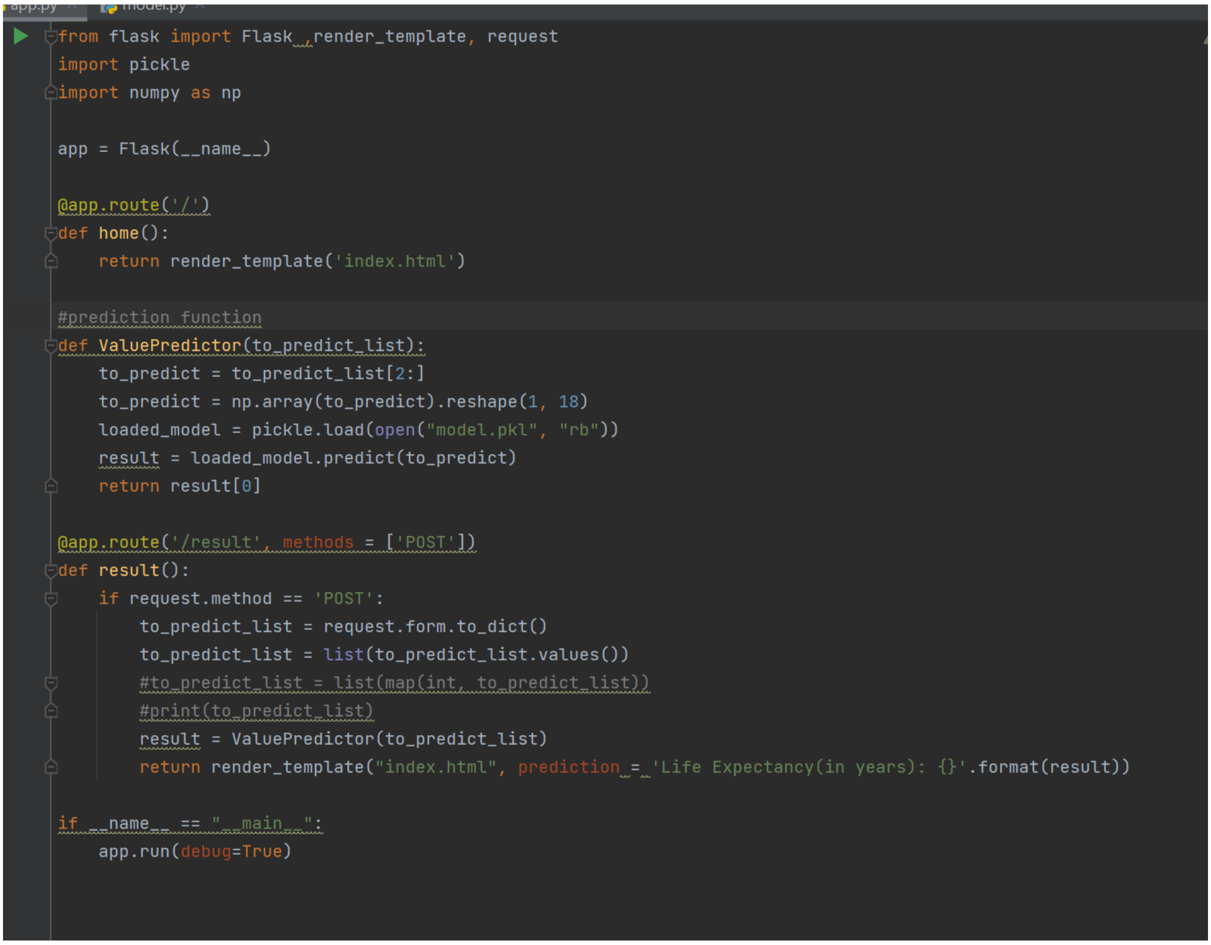

Training Data
Machine Learning Model
Features
Flask API
Input Features
Website (frontend)
Prediction
Output Prediction
Model Save (serialization)
Model Reload (de-serialization)
Model save in pickle after training and reload for prediction

Fig 4.1.1: Model Flowchart

4.2 FLASK API

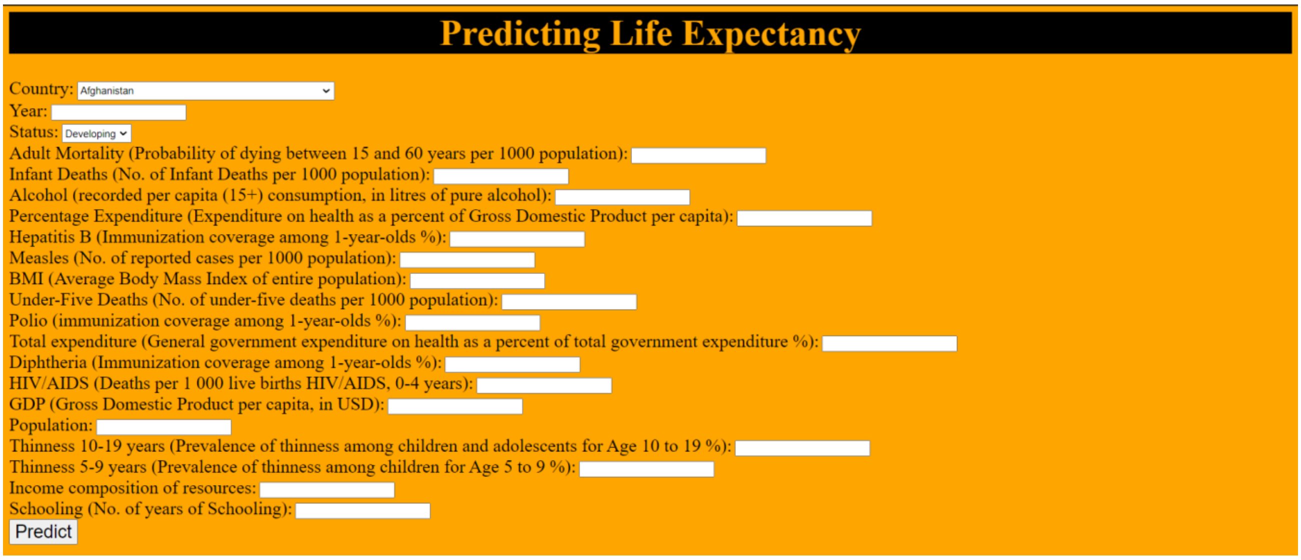


```python
from flask import Flask ,render_template, request
import pickle
import numpy as np

app = Flask(__name__)

@app.route('/')
def home():
    return render_template('index.html')

#prediction function
def ValuePredictor(to_predict_list):
    to_predict = to_predict_list[2:]
    to_predict = np.array(to_predict).reshape(1, 18)
    loaded_model = pickle.load(open("model.pkl", "rb"))
    result = loaded_model.predict(to_predict)
    return result[0]

@app.route('/result', methods = ['POST'])
def result():
    if request.method == 'POST':
        to_predict_list = request.form.to_dict()
        to_predict_list = list(to_predict_list.values())
        #to_predict_list = list(map(int, to_predict_list))
        #print(to_predict_list)
        result = ValuePredictor(to_predict_list)
        return render_template("index.html", prediction = 'Life Expectancy(in years): {}'.format(result))

if __name__ == "__main__":
    app.run(debug=True)
```

Fig 4.1.2: Model App

In the above figure, we will use the flask web framework to handle the POST requests that we will get from the app.py. Here we have imported numpy to create the array of requested data, pickle to load our trained model to predict. we have created the instance of the Flask() and loaded the model into the *model*. we have bounded API with the method predict(). In which predict method gets the data from the HTML passed by the requestor. model.predict() method takes input from the HTML and converts it into 2D numpy array the results are stored into the variable named output. As I mentioned earlier that app.py is going to request the server for the predictions.

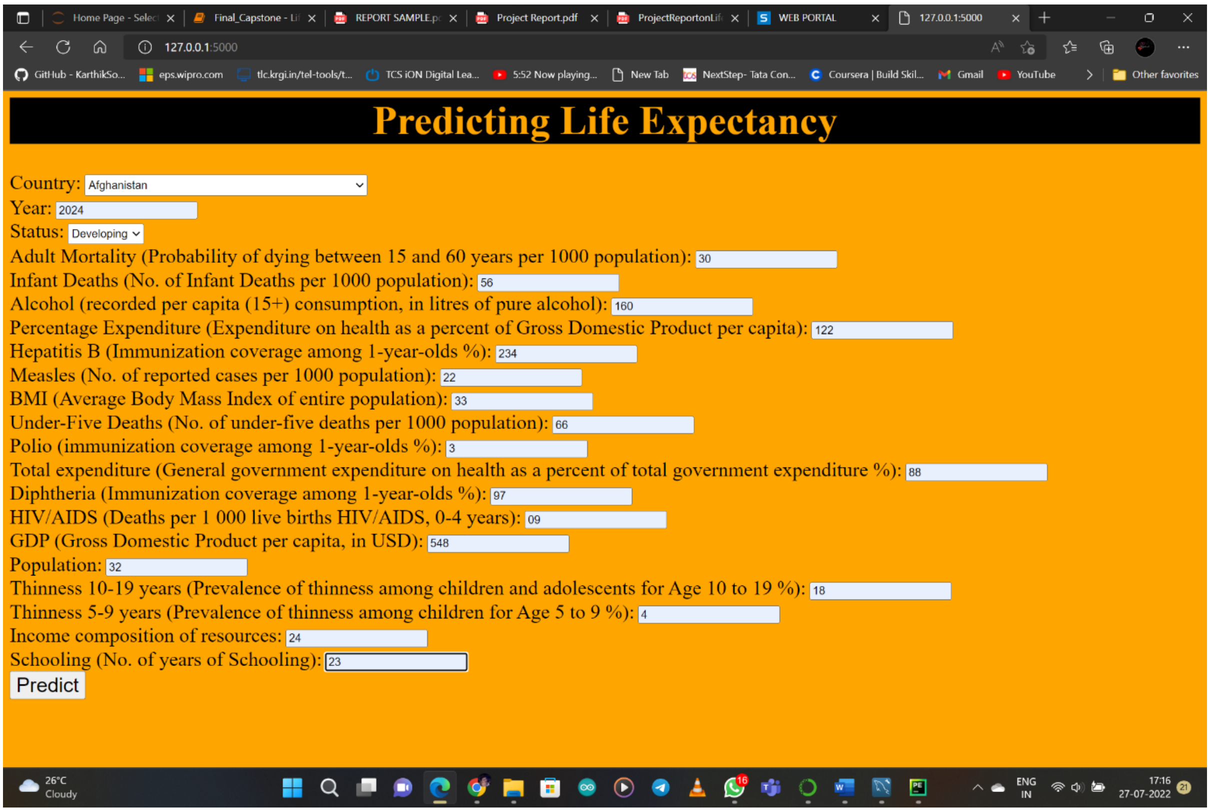
4.3 WEB-UI



Predicting Life Expectancy

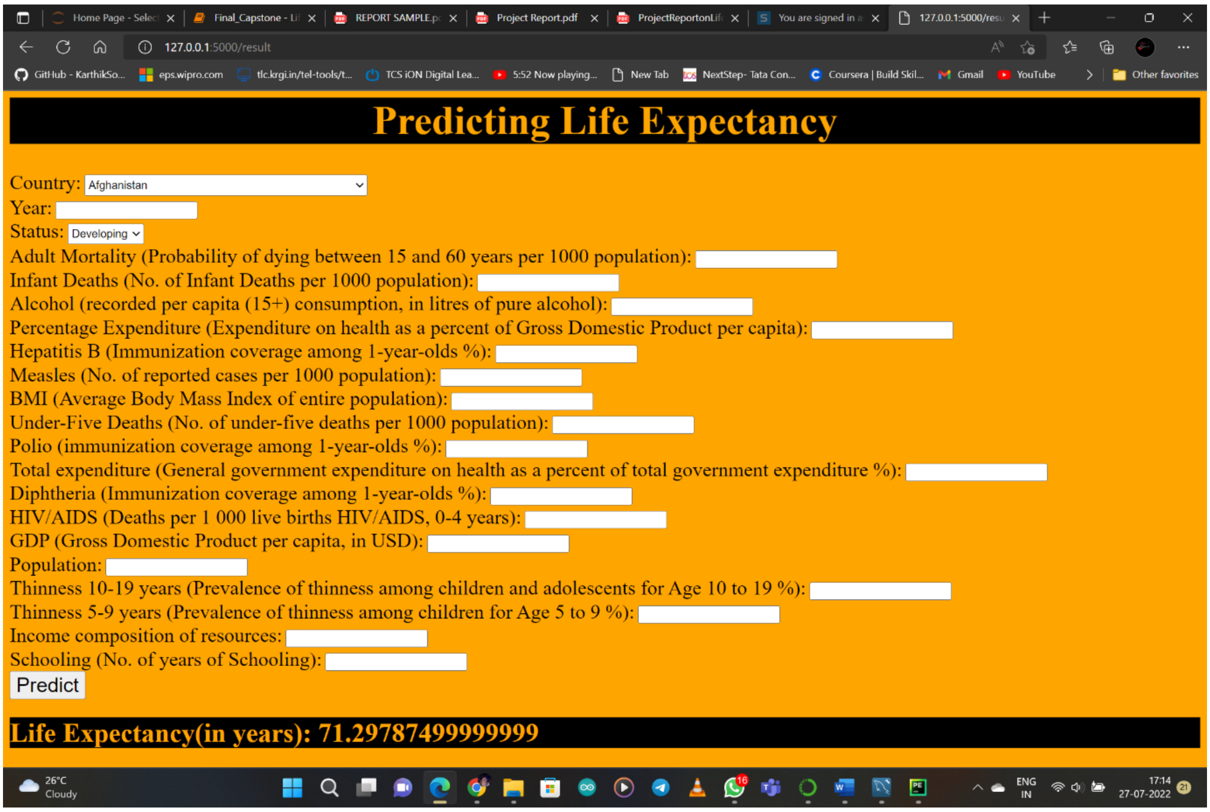
Country: Afghanistan
Year:
Status: Developing
Adult Mortality (Probability of dying between 15 and 60 years per 1000 population):
Infant Deaths (No. of Infant Deaths per 1000 population):
Alcohol (recorded per capita (15+) consumption, in litres of pure alcohol):
Percentage Expenditure (Expenditure on health as a percent of Gross Domestic Product per capita):
Hepatitis B (Immunization coverage among 1-year-olds %):
Measles (No. of reported cases per 1000 population):
BMI (Average Body Mass Index of entire population):
Under-Five Deaths (No. of under-five deaths per 1000 population):
Polio (immunization coverage among 1-year-olds %):
Total expenditure (General government expenditure on health as a percent of total government expenditure %):
Diphtheria (Immunization coverage among 1-year-olds %):
HIV/AIDS (Deaths per 1 000 live births HIV/AIDS, 0-4 years):
GDP (Gross Domestic Product per capita, in USD):
Population:
Thinness 10-19 years (Prevalence of thinness among children and adolescents for Age 10 to 19 %):
Thinness 5-9 years (Prevalence of thinness among children for Age 5 to 9 %):
Income composition of resources:
Schooling (No. of years of Schooling):

Fig 4.1.2: Model Web-UI



Predicting Life Expectancy

Country: Afghanistan
Year: 2024
Status: Developing
Adult Mortality (Probability of dying between 15 and 60 years per 1000 population): 30
Infant Deaths (No. of Infant Deaths per 1000 population): 56
Alcohol (recorded per capita (15+) consumption, in litres of pure alcohol): 160
Percentage Expenditure (Expenditure on health as a percent of Gross Domestic Product per capita): 122
Hepatitis B (Immunization coverage among 1-year-olds %): 234
Measles (No. of reported cases per 1000 population): 22
BMI (Average Body Mass Index of entire population): 33
Under-Five Deaths (No. of under-five deaths per 1000 population): 66
Polio (immunization coverage among 1-year-olds %): 3
Total expenditure (General government expenditure on health as a percent of total government expenditure %): 88
Diphtheria (Immunization coverage among 1-year-olds %): 97
HIV/AIDS (Deaths per 1 000 live births HIV/AIDS, 0-4 years): 09
GDP (Gross Domestic Product per capita, in USD): 548
Population: 32
Thinness 10-19 years (Prevalence of thinness among children and adolescents for Age 10 to 19 %): 18
Thinness 5-9 years (Prevalence of thinness among children for Age 5 to 9 %): 4
Income composition of resources: 24
Schooling (No. of years of Schooling): 23



Predicting Life Expectancy

Country: Afghanistan
Year:
Status: Developing
Adult Mortality (Probability of dying between 15 and 60 years per 1000 population):
Infant Deaths (No. of Infant Deaths per 1000 population):
Alcohol (recorded per capita (15+) consumption, in litres of pure alcohol):
Percentage Expenditure (Expenditure on health as a percent of Gross Domestic Product per capita):
Hepatitis B (Immunization coverage among 1-year-olds %):
Measles (No. of reported cases per 1000 population):
BMI (Average Body Mass Index of entire population):
Under-Five Deaths (No. of under-five deaths per 1000 population):
Polio (immunization coverage among 1-year-olds %):
Total expenditure (General government expenditure on health as a percent of total government expenditure %):
Diphtheria (Immunization coverage among 1-year-olds %):
HIV/AIDS (Deaths per 1 000 live births HIV/AIDS, 0-4 years):
GDP (Gross Domestic Product per capita, in USD):
Population:
Thinness 10-19 years (Prevalence of thinness among children and adolescents for Age 10 to 19 %):
Thinness 5-9 years (Prevalence of thinness among children for Age 5 to 9 %):
Income composition of resources:
Schooling (No. of years of Schooling):

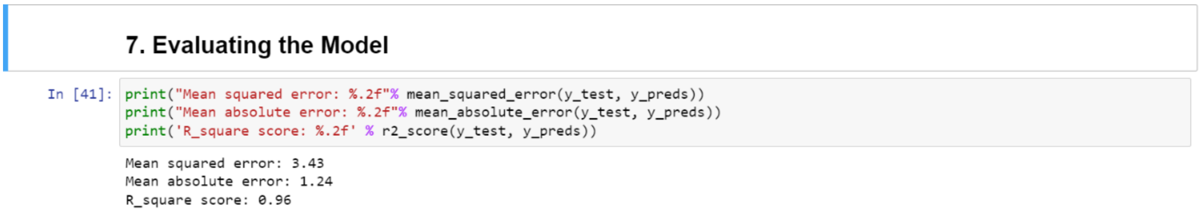
Life Expectancy(in years): 71.29787499999999

Fig 4.1.3: Model Web-UI_1

5.RESULT

The prediction model is trained using three regression models, namely Linear Regression, Decision Tree Regressor and Random Forest Regressor. The selection of model is done on the basis of R 2 score, Mean Squared Error & Mean Absolute Error. Random Forest Regressor is selected for the development of the prediction model for life expectancy and the comparative analysis of life expectancy between developed and developing countries suggests that, developed countries have high life expectancy as compared to developing countries.

By observing various graphs and also the output data we can conclude that some factors influence the life expectancy more than the others. Among such are adult mortality, immunization of the citizens against various diseases, schooling, alcohol consumption etc.



7. Evaluating the Model

```
In [41]: print("Mean squared error: %.2f"% mean_squared_error(y_test, y_preds))
         print("Mean absolute error: %.2f"% mean_absolute_error(y_test, y_preds))
         print('R_square score: %.2f' % r2_score(y_test, y_preds))

         Mean squared error: 3.43
         Mean absolute error: 1.24
         R_square score: 0.96
```

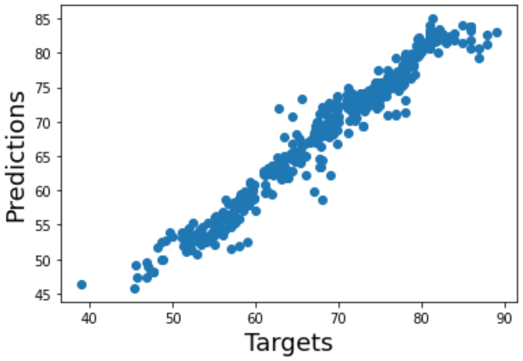


Fig 5.1: Model Evaluation

[Random Forest Regressor Model]

6. CONCLUSION

By doing the above procedure and all we successfully created Life expectancy prediction system using Python, Flask API and PowerBI service. The potential use of project is not limited to health care in practice, but could also be useful in other clinical applications such as clinical trials. The project makes a good use of machine learning in predicting life expectancy of a country that can help respective government in making policies that will serve for the benefit of the nation and entire humankind.

Future Scope:

- Look at class within a particular country and see if these same factors are same in determining life expectancy for an individual.
- Use the Twitter API to incorporate NLP analysis for a country to see how it relates to Life Expectancy.
- Increase the dataset size with continuing UN and Global Data to incorporate new added features like population, GDP, environmental, and etc in order to test and clarify country groupings.
- Mental Health versus Life Expectancy.
- As more data comes, that can be fed to the model for more accurate predictions.
- Currently, the project is just a web application. It can be developed to support other platforms like Android, IOS and Windows Mobile.
- Other regression models can also be used for prediction and later the best among them should be chosen.

7.REFERENCE

- Kumar Rajarshi, “Life Expectancy (WHO) Statistical Analysis on factors influencing Life Expectancy”, 2018. [Online].
  Available: https://www.kaggle.com/kumarajarshi/life-expectancy-who
- Learn how to develop a machine learning model and how to deploy it using Flask [Online]
  Available: https://www.youtube.com/watch?v=MxJnR1DMmsY
- Announcing Power BI in Jupyter notebooks, Embed PowerBI and Jupyter [Online]
  Available: https://powerbi.microsoft.com/fr-ca/blog/announcing-power-bi-in-jupyter-notebooks/
- Deploy ML model using flask [Online]
  Available: : https://github.com/siddiquiamir/ML-MODEL-DEPLOYMENT-USING-FLASK