**PROJECT ON**

***“PREDICTING LIFE EXPECTANCY USING MACHINE LEARNING”***

Prepared by:

HIMANI AGARWAL

**Introduction**

Life expectancy plays an important role when decisions about the final phase of life need to be made. A prediction of Life Expectancy helps to analyze the average life span and thus constitute in making life decisions for the generations to come easier.

**Overview**

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 various years. The data collected is related to a timeframe from 2000 to 2015.

The data originates from here:

<https://www.kaggle.com/kumarajarshi/life-expectancy-who/data>

“Predicting Life Expectancy using Machine Learning”, as the name suggests, aims to predict the lifespan on a human being, based on diverse factors in a demographic region. The life of a human depends on various factors such as Regional variations, Economic Circumstances, Sex Differences, Mental Illnesses, Physical Illnesses, Education, Year of their birth and other demographic factors. The project aims to predict an average life expectancy based on these and several other factors. This project finds the expected solution using machine learning algorithm, Linear Regression. The aim of the project is to find the relationship of the various factors with the lifespan of an individual using the ML Algorithm mentioned above.

The dataset used for the prediction contains data from year 2000 to 2015. It contains more than 2500 entries and around 22 columns with various features like Population, Status, Alcohol, Infant Deaths etc., which aids the prediction of the model.

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.

Purpose:

If life expectancy is longer in a certain country, it says something about the conditions of the place. It says something about the health factors as well as the quality of life. If the conditions in a country and in its economy are good, obviously the life expectancy will be more. But it isn’t enough to have a long life. It must be a healthy life too. A lot of people spend their later years in a miserable condition, in poor health. That’s not acceptable at all. We must strive to ensure that everyone has a healthy life and a life of quality. With today’s new technologies and a positive attitude towards research, it is more possible than ever that a long and healthy life will be possible for more people.

LITERATURE REVIEW

Existing Problem

Few works have been done to provide an individually customized life expectancy prediction. We have reviewed existing works and techniques in the prediction of human LE, and reached a conclusion that it is feasible to predict a PLE for individuals using evolving technologies and devices such as big data, AI, machine learning techniques, and PHDs, wearables and mobile health monitoring devices. We also identiﬁed that the collection of data will be a huge challenge due to the privacy and government policy considerations, which will require collaboration of various bodies in the health industry. The interworking of a heterogeneous health network is also a challenge for data collection. Despite these challenges, a possibility of a PLE prediction by proposing an approach of data collection and application by smartphone, with which users can enter their information to access the cloud server to obtain their own PLE, was shown.

To verify the accuracy of PLE prediction and validation of data quality, big data techniques and analysis algorithms need to be developed and tested in a real-life situation with several sample groups. As artiﬁcial intelligence technology is evolving and being applied rapidly, feasibility may be increasing to collect health data from the public as well as existing health agencies such as centralized health servers.

Proposed Solution

Although there have been lot of studies undertaken in the past on factors affecting life expectancy considering demographic variables, income composition and mortality rates, it was found that effect of immunization and human development index was not taken into account in the past. Also, some of the past research was done considering multiple linear regression based on data set of one year for all the countries. Hence, this gives motivation to resolve both the factors stated previously by formulating a regression model based on mixed effects model and multiple linear regression while considering data from a period of 2000 to 2015 for all the countries. Important immunization like Hepatitis B, Polio and Diphtheria will also be considered. In a nutshell, this study will focus on immunization factors, mortality factors, economic factors, social factors and other health related factors as well. Since the observations this dataset are based on different countries, it will be easier for a country to determine the predicting factor which is contributing to lower value of life expectancy.

The model of “Predicting Life Expectancy using Machine Learning” uses IBM Cloud services, which helps to avoid any storage issues. The UI Presented to the users is a website URL and hence they need not download any application to predict the results, which saves the storage space as that is the need of the hour.

PROJECT REQUIREMENTS

This project fundamentally aims in predicting the life expectancy. The primary requirement of the project is the suitable dataset which will aid the prediction. The machine learning model is trained on the basis of the data provided, such that it could predict the average lifespan of an individual in the coming years.

Functional Requirements

1. The dataset should be preprocessed before applying prediction.

2. The data model must be created on the basis of preprocessed data.

3. The data model must then be converted into a module for further use, after the data is updated.

4. The data should be implemented using IBM Watson which should then be connected to NodeRed for the User Interface.

**Algorithm:**

1. Importing the required libraries.

2. Importing the dataset and storing it as a dataframe.

3. Pre-processing

a. Removing 'NA' values and replacing using interpolation

b. Removing outliers using inter quartile range

4. Visualisations

5. Splitting the dataset into training and test sets

6. Building a model using train set

7. Testing the model using the test set

8. Evaluating the model and finding out the accuracy using RMSE value

9. Deploying the node-red model using scoring endpoint generated for the UI Dashboard.

### **Part 1. Loading packages**

The following packages have been imported NumPy, Pandas, Matplotlib, Scipy, Seaborn. Sklearn is the most widely used package for the machine learning process. The following sub packages have been used:

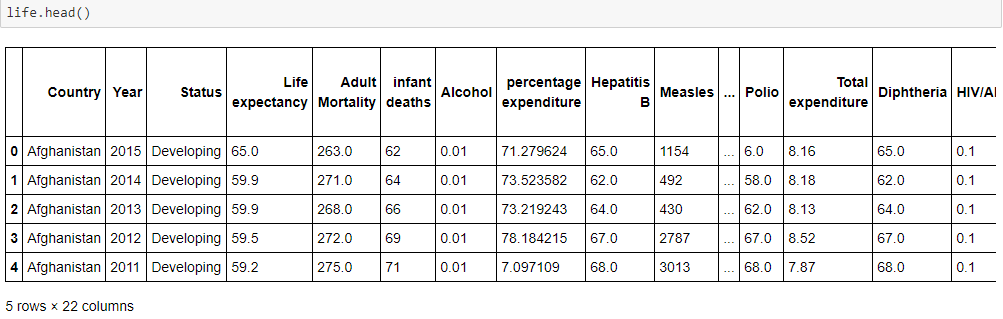
* train\_test\_split
* linear\_model
* model\_selection
* metrics
* tree
* ensemble
* Preprocessing

### **Part 2. Reading the data**

The data is saved as a csv file as LifeExpectancy.csv and it is read and stored in the ‘life’ variable. Below the first 5 rows are shown. The data contains 21 columns and 2938 rows with the header row. The table contains data about:

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

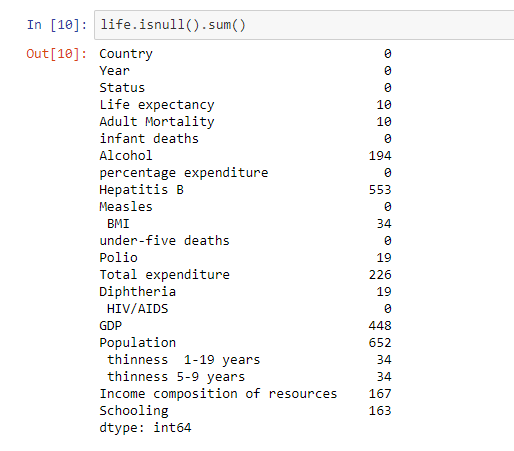
With the exclusion of Country name and Status (either developed or developing) all of the data is numeric. The values are either in years, percentages, millions or dollars in the case of Gross Domestic Product (GDP)



Part 3. Preprocessing

3.a) Removing 'NA' values and replacing using interpolation

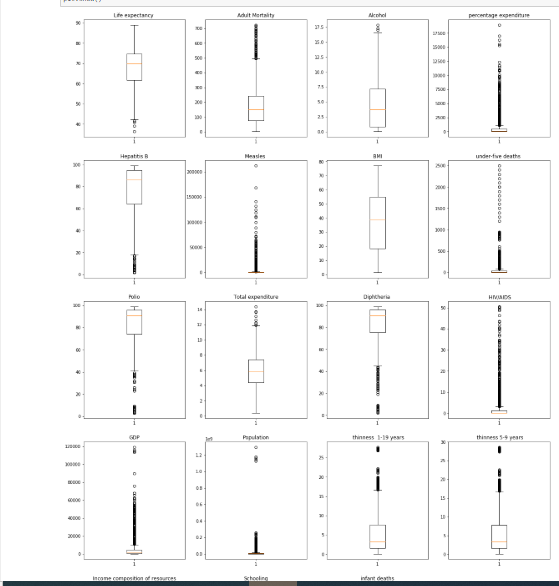
Now, we will be dealing with the ‘NA’ values of our dataset as it can prevent data loss which results in removal of the rows and columns. For this, we go for interpolation method.



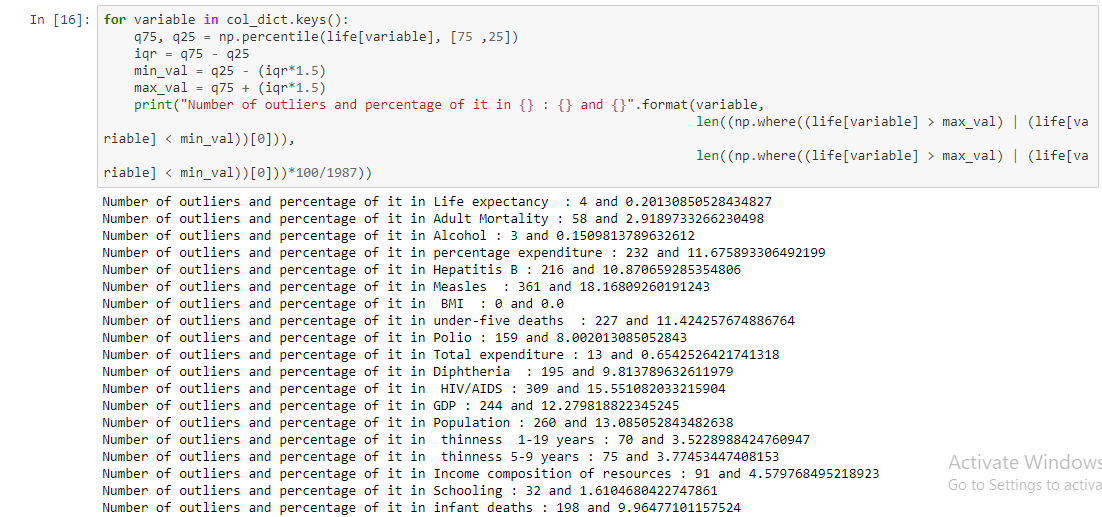


3.b. Removing the outliers

We can get an idea about the outliers in the dataset using the boxplot method.

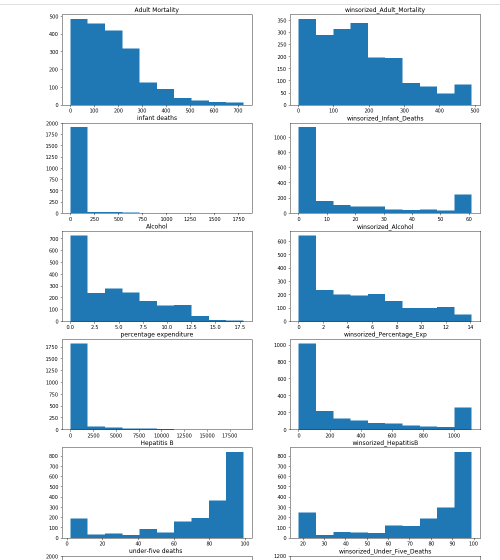


Now, we will be finding the outliers of our dataset using the inter-quartile range method.

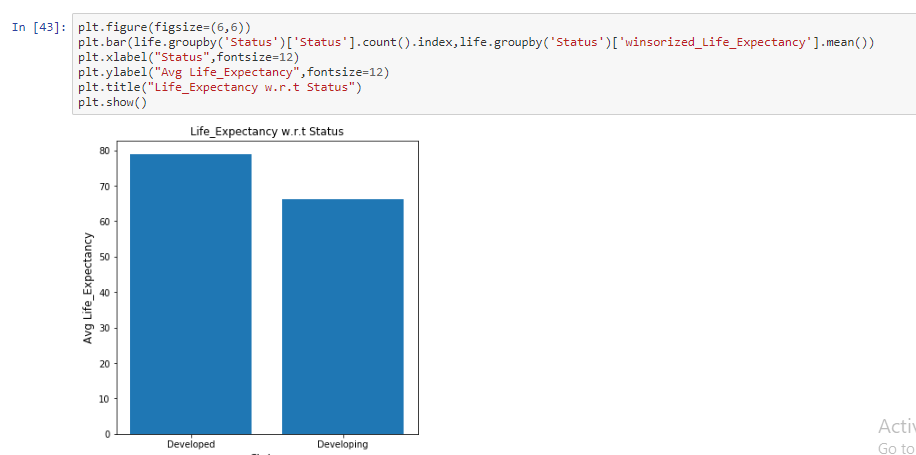


Part 4. Visualisations

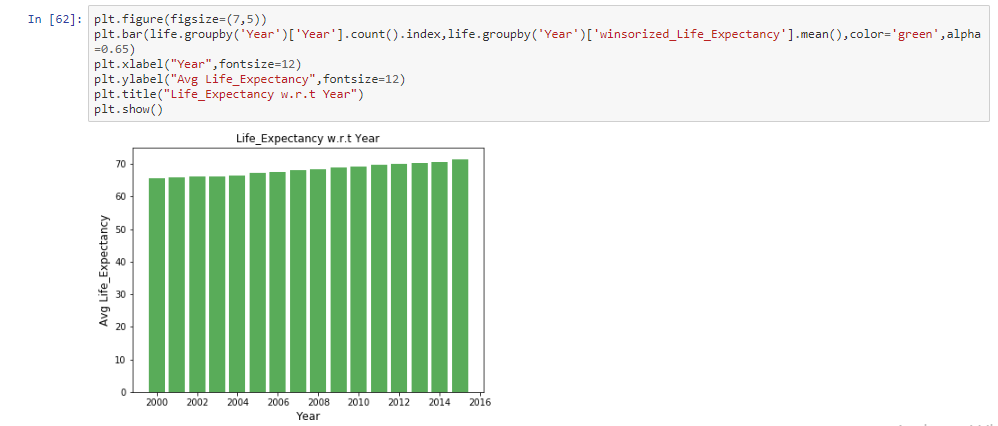
A) Bar plots With outliers and Without Outliers



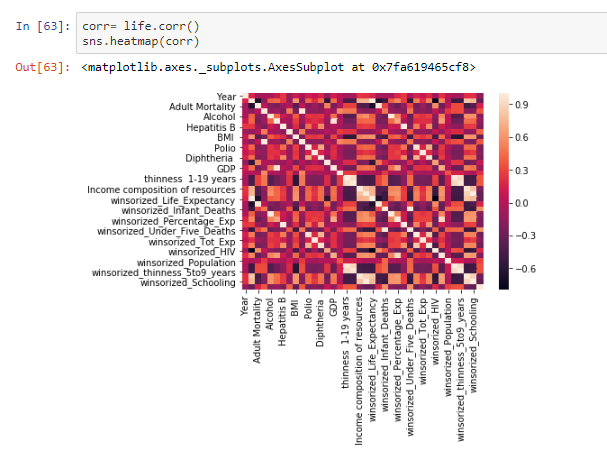
B)Life Expectancy w.r.t Status



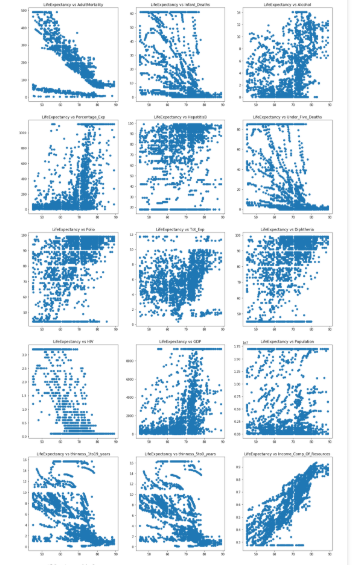
C) Life Expectancy w.r.t year



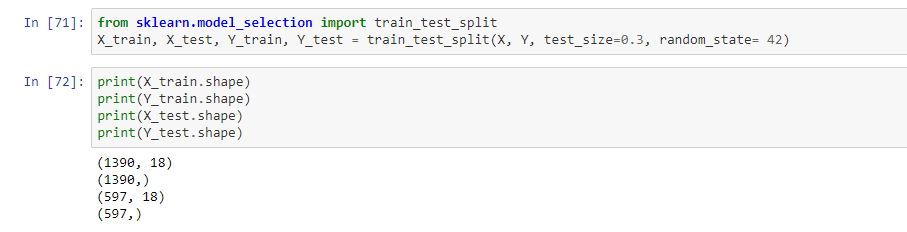
D) Correlation Heatmap tells us which variables are to be considered for building the model



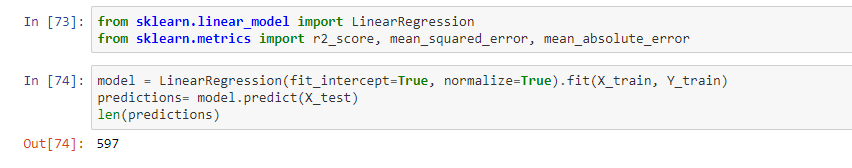
E) Scatter plot of Life Expectancy vs different factors



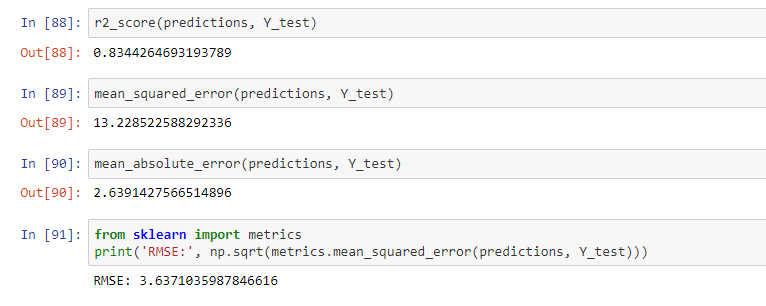
Part 5. Splitting the dataset into train and test datsets



Part 6 and 7. Building and testing the model using training dataset



Part 8. Evaluating the model



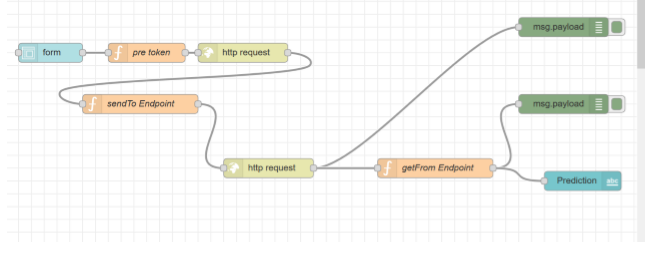
Part 9. Deploying the model

FLOWCHART OF Node-Red Application

Node-RED provides a browser-based flow editor that makes it easy to wire together flows using the wide range of nodes in the palette. Flows can be then deployed to the runtime in a single-click.

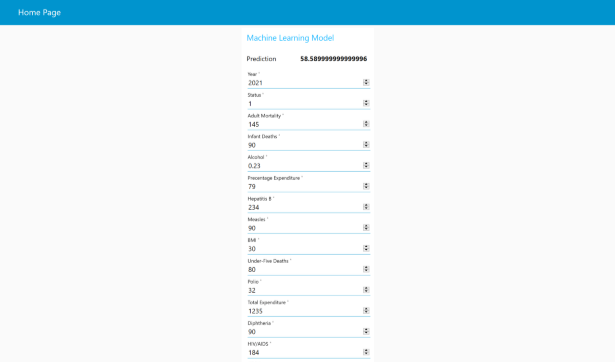
A built-in library allows you to save useful functions, templates or flows for re-use.

The flows created in Node-RED are stored using JSON which can be easily imported and exported for sharing with others.



RESULT

The model appears to the user in the form of an interface as shown in the Figure. The user has to fill in the inputs and click on “Predict” button at the end of the form. On clicking the “Predict” button, the user will be displayed the predicted life expectancy, based on the inputs provided, at the top of the page as shown in the Figure.



Advantages:

Advantages of using IBM Watson:

1. Advantages of using IBM Watson:

• Processes unstructured data

• Fills human limitations

• Acts as a decision support system, doesn’t replace humans

• Improves performance + abilities by giving best available data

• Improve and transform customer service

• Handle enormous quantities of data

• Sustainable Competitive Advantage

2. Easy for user to interact with the model via the UI. 3. User-friendly. 4. Easy to build and deploy. 5. Doesn’t require much storage space.

Disadvantages:

1. Disadvantages of using IBM Watson:

• Only in English (Limits areas of use)

• Seen as disruptive technology

• Maintenance

• Doesn't process structured data directly

• Increasing rate of data, with limited resources

2. Not connected to database, hence no record of input.

3. Requires internet connection.

APPLICATIONS

When will I die? This question has endured across cultures and civilizations. It has given rise to a plethora of religions and spiritual paths over thousands of years, and more recently, some highly amusing apps. This system will be used for people wondering with such questions.

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. The principle of life expectancy suggests that you should purchase a life insurance policy for yourself and your spouse sooner rather than later. Not only will you save money through lower premium costs, but you will also have longer for your policy to accumulate value and become a potentially significant financial resource as you age.

It can be used by researchers to make meaningful researches out of it and thus, bring about something that will help increase the expectancy consider the impact of a specific factor on the average lifespan of people in a specific country.

CONCLUSION

Thus, we have developed a model that will predict the life expectancy of a specific demographic region based on the inputs provided. Various factors have a significant impact on the life span such as Adult Mortality, Population, Under 5 Deaths, Thinness 1-5 Years, Alcohol, HIV, Hepatitis B, GDP, Percentage Expenditure and many more.

User can interact with the system via a simple user interface which is in the form of a form with input spaces which the user needs to fill the inputs into.

FUTURE SCOPE

As future scope, we can connect the model to the database to have the record of predictions. This will help us analyze the trends in the life span.

A model with country wise bifurcation can be made, which will help to segregate the data demographically.