

## **A PROJECT ON “ PREDICTING LIFE EXPECTANCY USING MACHINE LEARNING ”**

**Kick-Off Date:** May 19, 2020

### **Regarding the dataset:**

This dataset is available in the Global Health Observatory (GHO) data repository under World Health Organization (WHO) which keeps track of the health status as well as many other related factors for all countries. The data-sets are made available to public for the purpose of health data analysis. In this dataset, related to life expectancy, health factors for 193 countries have 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 dataset consists of 22 Columns and 2938 rows which meant 20 predicting variables. All predicting variables are divided into several broad categories like Immunization related factors, Mortality factors, Economical factors and Social factors.

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

### **Project description:**

This project gives motivation to take in account the factors like demographic variables, income composition and mortality rates by formulating a regression model based on different regression techniques 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. It is observed that as this dataset is 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. This will help in suggesting a country which area should be given importance in order to efficiently improve the life expectancy of its population.

### Technical Requirements:

1. The dataset must be in csv format.
2. Implementation of Machine Learning Algorithms using Python.
3. IBM cloud account.
4. IBM Watson Studio and Node-Red application.

### Software Requirements:

1. Python IDE
2. Excel
3. IBM Cloud
4. IBM Watson
5. Node-Red for User Interface

### Hardware Requirements:

Processor-i3, 7th generation or higher'  
Speed- 2GHz or more

### Tools used for this analysis include:



- Python libraries particularly Numpy and Pandas for manipulating data structures
- Matplotlib and Seaborn for visualisation
- Scikit-Learn and Statsmodels for regression analysis

### Deliverables:

- The dataset aims to answer the following key questions:
- Does various predicting factors which has been chosen initially really affect the Life

expectancy? What are the predicting variables actually affecting the life expectancy?

- Should a country having a lower life expectancy value increase its healthcare expenditure in order to improve its average lifespan?
- How does Infant and Adult mortality rates affect life expectancy?
- Does Life Expectancy has positive or negative correlation with eating habits, lifestyle, exercise, smoking, drinking alcohol etc.
- What is the impact of schooling on the lifespan of humans?
- Does Life Expectancy have positive or negative relationship with drinking alcohol?
- Do densely populated countries tend to have lower life expectancy?
- What is the impact of Immunization coverage on life Expectancy?
- Finally, a user interface will be implemented where you can get the predicted value of life expectancy given the values of dependent features.

### **Project scope:**

A machine learning model which predicts life expectancy based on the inputs provided by the user, and it will be out of scope for user to modify or to increase the accuracy/predicted output of the ML model.

**Front-end:** A web user interface taking the necessary inputs from the user to implement the designed model.

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

### **Team:**

This project is wholly done by me, **Himani Agarwal**

### **Project Duration:**

1 month

