**Project Report**

**Project by:**

**Pushpak Nandkishor Chaudhari**

(pushpakchaudhari05@gmail.com)

Thakur college of Engineering and Technology

Kandivali (East), Mumbai-400101

**Under:**

Remote summer internship program 2020 by SmartInternz.

**Predicting Life Expectancy using Machine Learning.**

# 1. Introduction

## 1.1 Overview

The overview of the project is to develop a model predicting Life Expectancy by training a model on the Life expectancy dataset. 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 like [Status, Life expectancy ,Adult Mortality, infant deaths, 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].

## 1.2 Purpose

The purpose of the project is to build a scalable, real-time learning model using IBM Watson to predict the life expectancy of a person considering the various factors. The project will be helpful in improving the health condition of the society and give insights about some crucial factors such as [Alcohol intake, GDP growth, schooling, adult mortality, total and cost expenditure and etc]. The project uses a Regression which is a classification algorithm. It is a measure of the relation between the mean value of one variable (e.g. output) and corresponding values of other variables. The dataset used or the training of the model was downloaded from kaggle.com and Python is used to write the code for machine learning model.

# 2. Literature Survey

## 2.1 Existing Problems

The biggest challenge is to collect authentic data from people regarding various features as it requires a systematic collaboration between various departments such as Health, Municipalities, etc. Despite these challenges, Life expectancy can be predicted by proposing a data collection and application approach. 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 could be increased.

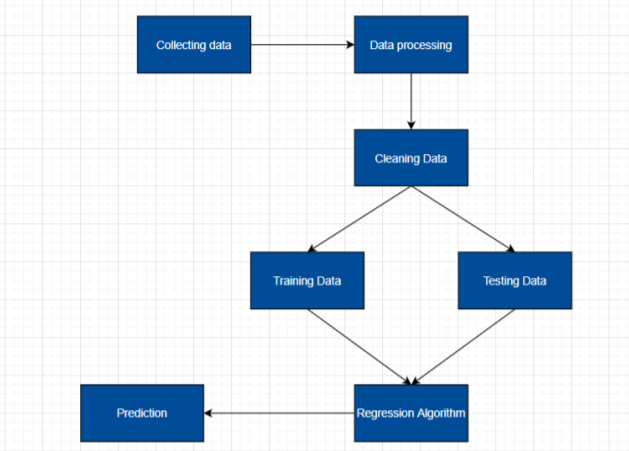
## 

## 2.2 Proposed Solution

The Life Expectancy dataset consisted of the following important features: Birth Rate · Cancer Rate · Dengue Cases · Gross Domestic Product (GDP) · Expenditure · Heart Disease Rate · Population · Area · Population Density . Life Expectancy, measured in number of years. After that, I have performed data cleaning and data analysis using the statistical tools in python and selected the dependent and independent features. Then, I created a machine learning model using regression (a classification algorithm). Using that model when we give the inputs (features) the model will give prediction (life expectancy in years) as output and finally that model is deployed to IBM cloud.

# 3. Theoretical Analysis

## 3.1 Block Diagram



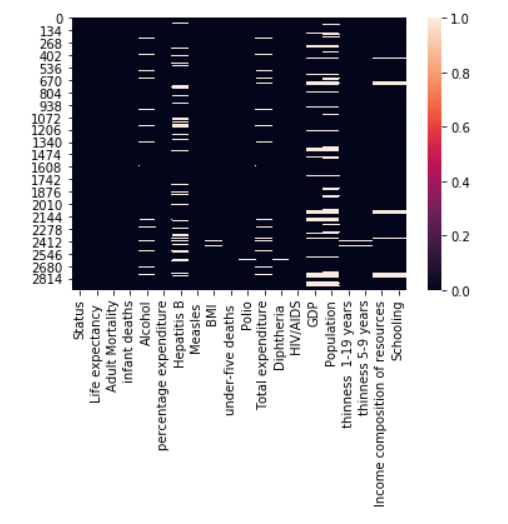
## 3.2 Software Engineering

**Software/Hardware  Requirements:**

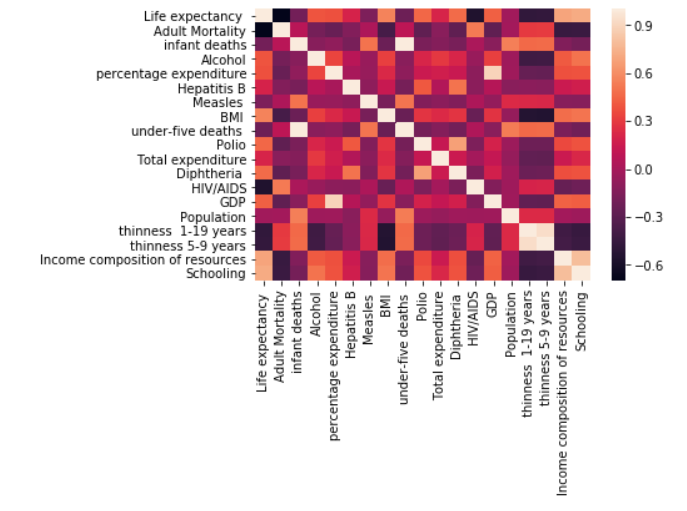
* IBM cloud services
* IBM Watson Studio
* IBM Node-Red application 7
* Jupyter Notebook
* Github
* Zoho document writer

# 4. Experimental Investigation

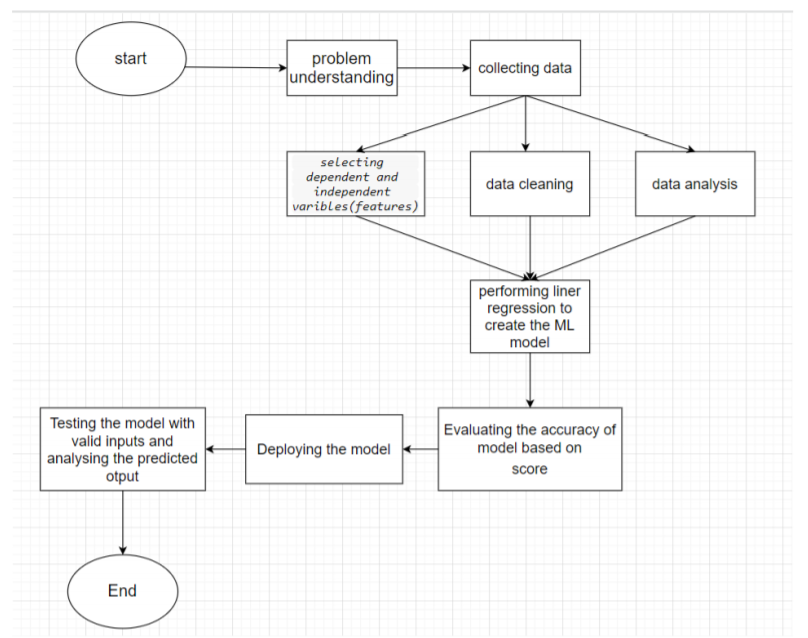
**Plot showing NA values in our dataset:**



**Correlation between the features of the dataset.**



**5. Flowchart**



# 6. Result

* Once the UI and the model are integrated successfully, it asks the user to enter the values for calculating his life expectancy.
* The accuracy achieved with the help of Machine Learning is 99.2%.

# 7. Advantages & Disadvantages

**Advantages:**

* The factors affecting the lifespan of a person get highlighted effectively and can turn out be wake up call for an individual bring about a change in their habits and lifestyle.

**Disadvantages:**

* The dataset misses out on some factors affecting life expectancy due to its limitations.
* Inappropriate pre-processing and data cleaning as well as selecting inappropriate model may lead to errors in the final result.

# 8. Application

* This project can be used by various health care sectors of a nation to look into some particular aspects to enhance the life quality of their citizens.
* Also, the insurance sector can heavily benefit from the project with regards to activities like drawing up insurance plans, setting premium values,etc.

# 9. Conclusion

Thus, we successfully created a model which predicts the life expectancy of an individual with the help of Machine Learning using Python.

# 10. Future Scope

With the help of data integration from various sources, Nations can actually estimate an average life expectancy of its people and work towards the right direction. Also, citizens can gain the information regarding the factors detrimental to an individual’s life and thereby try to change their lifestyles accordingly.

# 11. Bibilography

## Appendix

Node red UI link:

<https://node-red-nwgar.eu-gb.mybluemix.net/ui/#!/0?socketid=jGM5lJ_gJ1Sn_hcyAAAQ>

Project Demonstration link:

<https://drive.google.com/file/d/1S0QoLl9yGG00uau-oHQ3yuByCHeNyYcZ/view>

Github repository link:

<https://github.com/SmartPracticeschool/llSPS-INT-3306-Predicting-Life-Expectancy-using-Machine-Learning>

Dataset:

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

## Source Code

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns

expect=expect.drop(['Year','Country'],axis=1)

sns.heatmap(pd.isnull(expect))

expect.isnull().sum()

expect["Alcohol"]=expect["Alcohol"].fillna(value=expect["Alcohol"].mean())

expect["Income composition of resources"]=expect["Income composition of resources"].fillna(value=expect["Income composition of resources"].mean())

expect[" thinness 1-19 years"]=expect[" thinness 1-19 years"].fillna(value=expect[" thinness 1-19 years"].mean())

expect[" thinness 5-9 years"]=expect[" thinness 5-9 years"].fillna(value=expect[" thinness 5-9 years"].mean())

expect["Schooling"]=expect["Schooling"].fillna(value=expect["Schooling"].mean())

expect["Polio"]=expect["Polio"].fillna(value=expect["Polio"].mean())

expect["Population"]=expect["Population"].fillna(value=expect["Population"].mean())

expect["GDP"]=expect["GDP"].fillna(value=expect["GDP"].mean())

expect["Diphtheria "]=expect["Diphtheria "].fillna(value=expect["Diphtheria "].mean())

expect["Life expectancy "]=expect["Life expectancy "].fillna(value=expect["Life expectancy "].mean())

expect[" BMI "]=expect[" BMI "].fillna(value=expect[" BMI "].mean())

expect["Adult Mortality"]=expect["Adult Mortality"].fillna(value=expect["Adult Mortality"].mean())

expect["Hepatitis B"]=expect["Hepatitis B"].fillna(value=expect["Hepatitis B"].mean())

expect["Total expenditure"]=expect["Total expenditure"].fillna(value=expect["Total expenditure"].mean())

expect.isnull().sum()

correlated\_data=expect.corr()

sns.heatmap(correlated\_data)

y=expect["Expect expectancy "]

X=expect.drop("Expect expectancy ",axis=1)

X["Status"].unique()

X["Status"]=X["Status"].replace(['Developing', 'Developed'],[1,2])

X.shape

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)

X\_train.shape

X\_test.shape

y\_train.shape

y\_test.shape

from sklearn.ensemble import RandomForestRegressor

regr = RandomForestRegressor(n\_estimators=2, random\_state=0)

regr.fit(X, y)

y\_pred = regr.predict(X\_test)

np.set\_printoptions(precision=2)

y\_pred=np.array(y\_pred)

y\_test=np.array(y\_test)

print(np.concatenate((y\_pred.reshape(len(y\_test),1),y\_test.reshape(len(y\_test),1)),1))

accuracy=regr.score(X\_test,y\_test)

print("Accuracy: ",accuracy)

print('Coefficients: \n', regr.coef\_)

print('Mean squared error: %.2f' % mean\_squared\_error(y\_test, y\_pred))

print('Coefficient of determination: %.2f' % r2\_score(y\_test, y\_pred))