PROJECT ON

"PREDICTING LIFE EXPECTANCY USING MACHINE LEARNING"

Prepared by:

RIYA KALBURGI

1. 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.

1.1. Overview

"Predicting Life Expectancy using Machine Learning" aims, as the name suggests, to predict the lifespan on a human being, based on diverse datasets, 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 various machine learning algorithms such as:
Linear Regression
Logistic Regression
SVM
Clustering
Polynomic Regression
The aim of the project is to find the relationship of the various factors with the lifespan of an individual using the ML Algorithms 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.

1.2. 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.

2. LITERATURE REVIEW

2.1. 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.

2.2. 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 affect 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.

3. 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.

3.1. 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 Node-Red for the User Interface.

3.2. Technical Requirements

1. The dataset must be in csv format.
2. Machine Learning Algorithms must be applied with the help of Python.
3. IBM cloud account.
4. IBM Watson and Node-Red flow.

3.3. Software Requirements

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

4. FLOWCHART

A flowchart is a diagram that depicts a process, system or computer algorithm. They are widely used in multiple fields to document, study, plan, improve and communicate often complex processes in clear, easy-to-understand diagrams. Flowcharts, sometimes spelled as flow charts, use rectangles, ovals, diamonds and potentially numerous other shapes to define the type of step, along with connecting arrows to define flow and sequence.

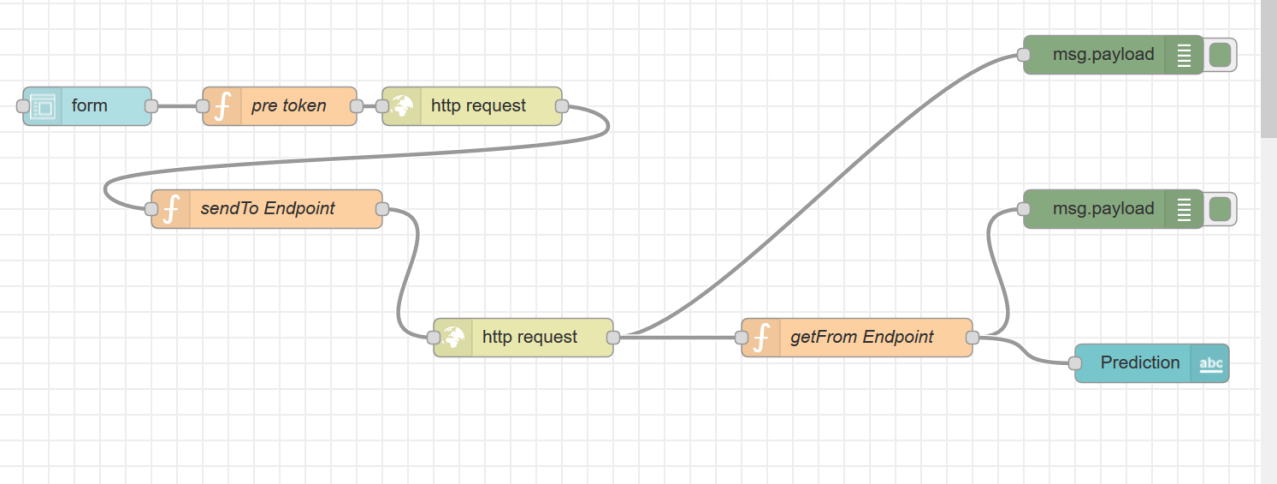

Figure 1: Node-red flow

5. RESULT

The model appears to the user in the form of an interface as shown in the Figure 2. 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 Figure 2.

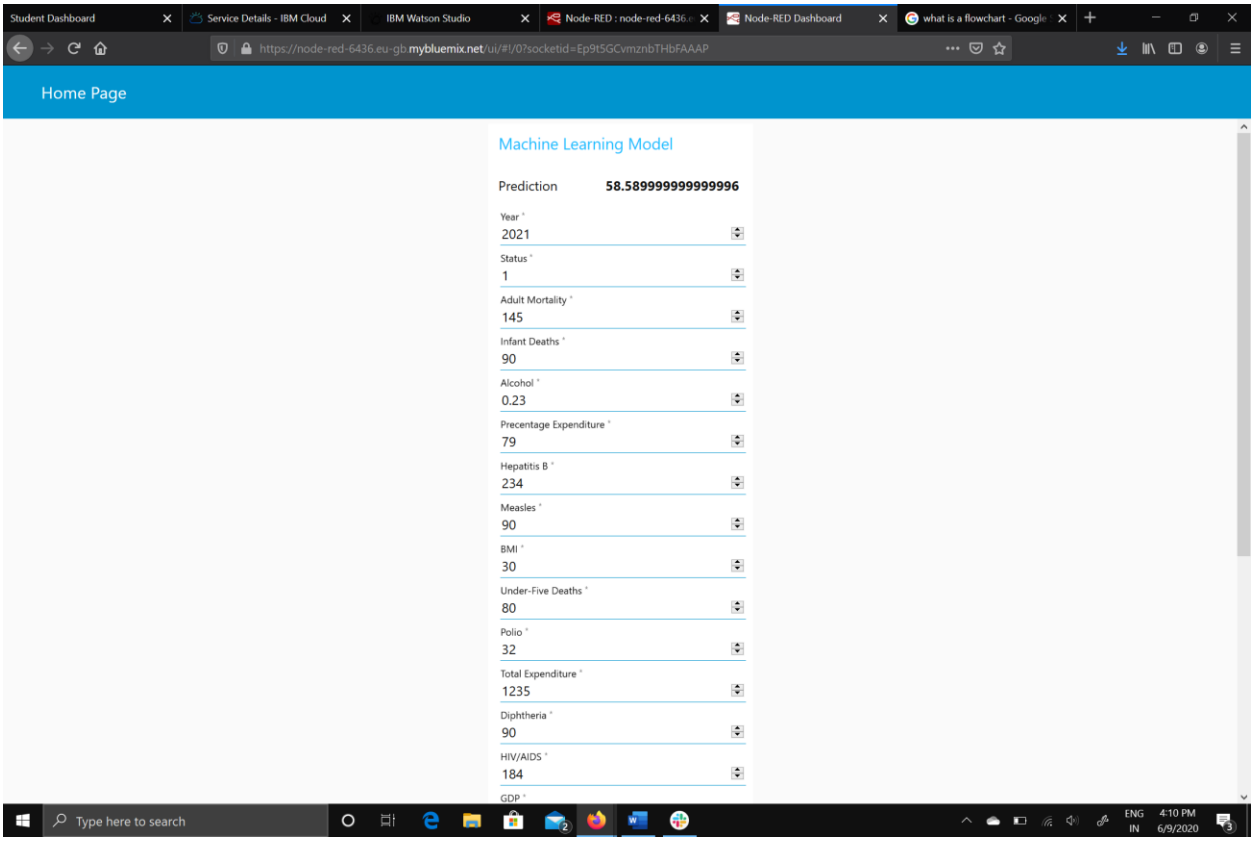

Figure 2: Result

6. ADVANTAGES AND DISADVANTAGES:

6.1. Advantages:

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.

6.2. 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.

7. APPLICATIONS

When will I die?

This question has endured across cultures and civilisations. 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.

8. 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.

9. 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.

APPENDIX

A. Source Code

#Importing the dataset and relevant libraries

```
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_c3f4b847f87649cabe83d0aa92822719 = ibm_boto3.client(service_name='s3',

    ibm_api_key_id='il7E_LE9jfDeeUdI7xzjru5H74pfUrN8MorFw9DQVRl7',

    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_c3f4b847f87649cabe83d0aa92822719.get_object(Bucket='predictinglifeexpectancy-donotdelete-pr-pqe7o5idw5fhmi',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 )

raw_data = pd.read_csv(body)

raw_data.head()

import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

sns.set()
```

```
import statsmodels.api as sm

raw_data.head()

raw_data.info()

raw_data.columns.values

column_names = ['Country', 'Year', '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']

raw_data.columns = column_names

raw_data.info()

data1 = raw_data.copy()

#Data Preprocessing

data1 = data1.dropna(axis = 0, subset = ['Life Expectancy'])

data1.info()

data1.mean()

data1 = data1.fillna(data1.mean())

data1.info()

data1.head()

data1 = data1.drop('Country', axis = 1)

data1.head()

data1['Status'] = data1['Status'].map({'Developing' : 0, 'Developed' : 1})

data1.head(5)

column_2 = ['Year', 'Status', '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', 'Life Expectancy']

data1 = data1[column_2]

data1.head()

data_preprocessed = data1.copy()

data_preprocessed.head()
```

```
data_preprocessed.info()
#Setting the inputs and targets
data_with_targets = data_preprocessed.copy()
targets = data_with_targets.iloc[:,-1]
targets.head()
targets.shape
inputs = data_with_targets.iloc[:,:-1]
inputs.head()
inputs.shape
#Splitting inputs and targets into training and testing dataset to avoid overfitting and underfitting
from sklearn.model_selection import train_test_split
train_test_split(inputs, targets)
x_train, x_test, y_train, y_test = train_test_split(inputs, targets, train_size = 0.8, random_state = 20)
print(x_train.shape, y_train.shape)
print(x_test.shape, y_test.shape)
#Applying the Random Forest Regression for prediction
from sklearn.ensemble import RandomForestRegressor
regressor = RandomForestRegressor(n_estimators = 10, random_state = 0)
regressor.fit(x_train, y_train)
regressor.score(x_train, y_train)
#Predicting the test data
pred = regressor.predict(x_test)
pred
regressor.score(x_test, y_test)
#Plotting the actual v/s predicted graph
plt.scatter(y_test, pred)
plt.xlabel('Actual Value', fontsize = 20)
plt.ylabel('Predicted Value', fontsize = 20)
```

```
plt.show()
sns.distplot((y_test-pred),bins=30)
#Calculating the errors
from sklearn import metrics
print('MAE:', metrics.mean_absolute_error(y_test, pred))
print('MSE:', metrics.mean_squared_error(y_test, pred))
print('RMSE:', np.sqrt(metrics.mean_squared_error(y_test, pred)))
```