

Summer Internship Project Report

Predicting life expectancy using Machine Learning

21/05/2020 - 18/06/2020

Name - Kshitij Kumar

Email - kshitijkumar3@gmail.com

1. Introduction

1.1 Overview

Life expectancy is a statistical measure of the average time a human being is expected to live, Life expectancy depends on various factors: Regional variations, Economic Circumstances, Sex Differences, Mental Illnesses, Physical Illnesses, Education, Year of their birth and other demographic factors. For eg. by

predicting life expectancy we can analyze the aggressiveness of any disease.

A large amount of data that is generated today is unstructured, which requires processing to

generate insights. After preprocessing the data set we will remove the noise from it.

After

removing the noise we will clean the data set.

IBM Watson, machine learning and node red are an integral part of an analysis. The end product will be a web page where you need to give all the required inputs and then submit it.

Afterwards it will predict the life expectancy value based on regression technique.

1.2 Purpose

The purpose of the project is to design a model for predicting Life Expectancy rate of a Country given various features such as year, GDP, education, alcohol intake of people in the

country, Expenditure on health care system and some specific disease related deaths that

happened in the country are given. This project analyses the provided data set and creates a

model to predict life expectancy and machine learning can benefit public health researchers with analyzing thousands of variables to obtain data regarding life expectancy.

We can use demographics of selected regional areas and multiple behavioral health disorders across regions to find correlation between individual behavior indicators and behavioral health outcomes. IBM Watson machine learning and node red are an integral part of analysis. The end product will be a web page where you need to give all the required

inputs and then submit it. Afterwards it will predict the life expectancy value based on regression technique.

2. LITERATURE SURVEY

2.1 Existing problem

Predicting the lifespan of people, or their “Personal Life Expectancy” (PLE) would greatly alter our lives. On one hand, it may have benefits for policy making, and help optimise an individual's health, or the services they receive.

2.2 Proposed Solution

Predicting life expectancy is not a new concept. Experts do this at a population level by classifying people into groups, often based on region or ethnicity.

Also, tools such as deep learning and artificial intelligence can be used to consider complex variables, such as biomedical data, to predict someone's biological age.

Biological age refers to how “old” their body is, rather than when they were born. A 30-year-old who smokes heavily may have a biological age closer to 40.

Calculating a life expectancy reliably would require a sophisticated system that considers a breadth of environmental, geographic, genetic and lifestyle factors – all of which have influence.

With machine learning and artificial intelligence, it's becoming feasible to analyse larger quantities of data. The use of deep learning and cognitive computing, such as with IBM Watson, helps doctors make more accurate diagnoses than using human judgement alone.

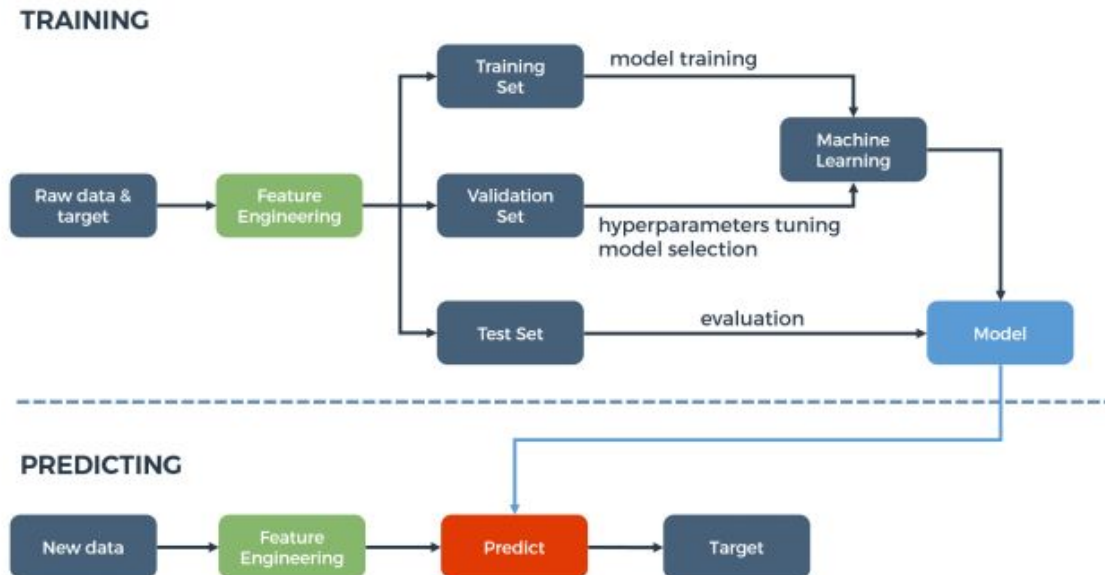
This, coupled with predictive analytics and increasing computational power, means we may soon have systems, or even apps, that can calculate life expectancy.

We will follow the following steps:

- a) Create IBM cloud services
- b) Configure Watson Studio
- c) Create Node-Red Flow to connect all services together
- d) Deploy and run Node-Red app

3. THEORETICAL ANALYSIS

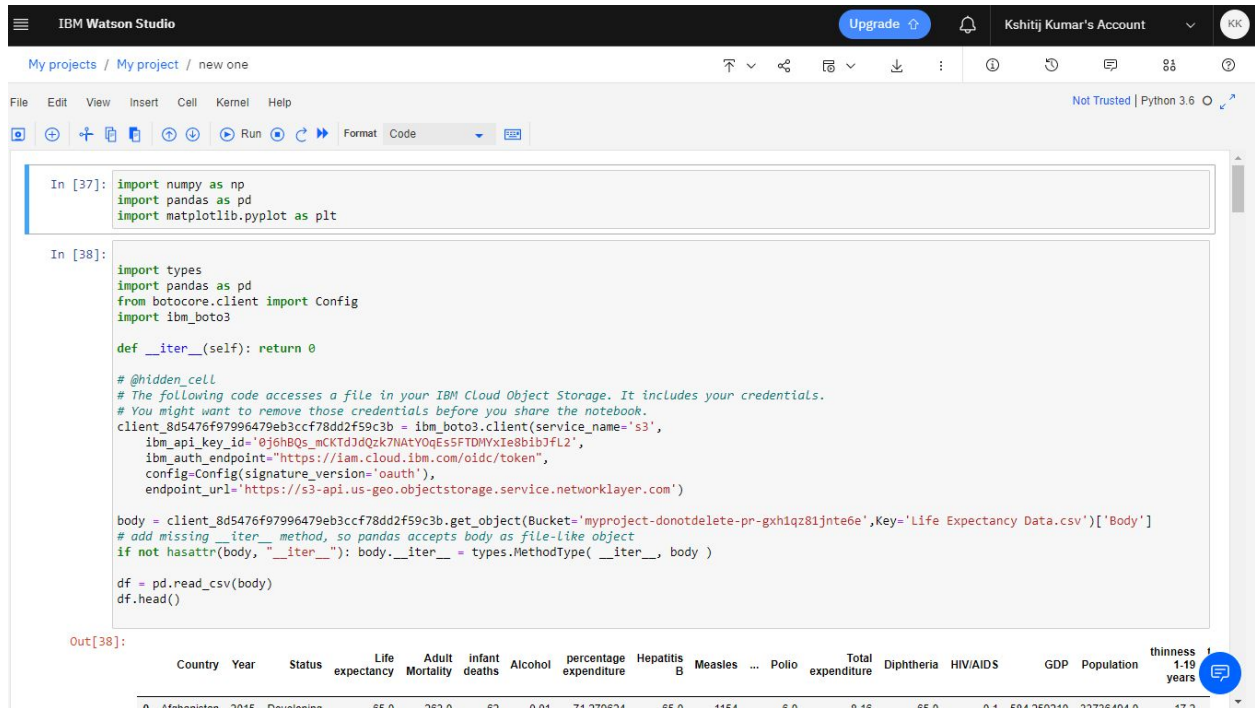
3.1 Block Diagram



3.2 Hardware/Software designing

1. Python notebook containing all the code.
2. A node red application which can input data and outputs a prediction for life expectancy.
3. A json file containing the architecture of node red project.
4. URL of the node red application.

Jupyter notebook :



The screenshot shows the IBM Watson Studio Jupyter notebook interface. The top bar includes the IBM Watson Studio logo, an 'Upgrade' button, and the user's account name 'Kshitij Kumar's Account'. The notebook is titled 'My projects / My project / new one'. The interface includes a menu bar (File, Edit, View, Insert, Cell, Kernel, Help) and a toolbar with various icons for file operations, running, and formatting. The code area contains two input cells. The first cell (In [37]:) imports numpy, pandas, and matplotlib. The second cell (In [38]:) imports types, pandas, boto3, and Config, and defines a class with a method to read a CSV file from IBM Cloud Object Storage. The output of the second cell (Out[38]:) is a table with 15 columns: Country, Year, Status, Life expectancy, Adult Mortality, Infant deaths, Alcohol, percentage expenditure, Hepatitis B, Measles, Polio, Total expenditure, Diphtheria, HIV/AIDS, GDP, Population, and thinness 1-19 years. The table shows data for Afghanistan in 2015.

```
In [37]: import numpy as np
import pandas as pd
import matplotlib.pyplot as plt

In [38]: import types
import pandas as pd
from boto3.client import Config
import boto3

def __iter__(self): return 0

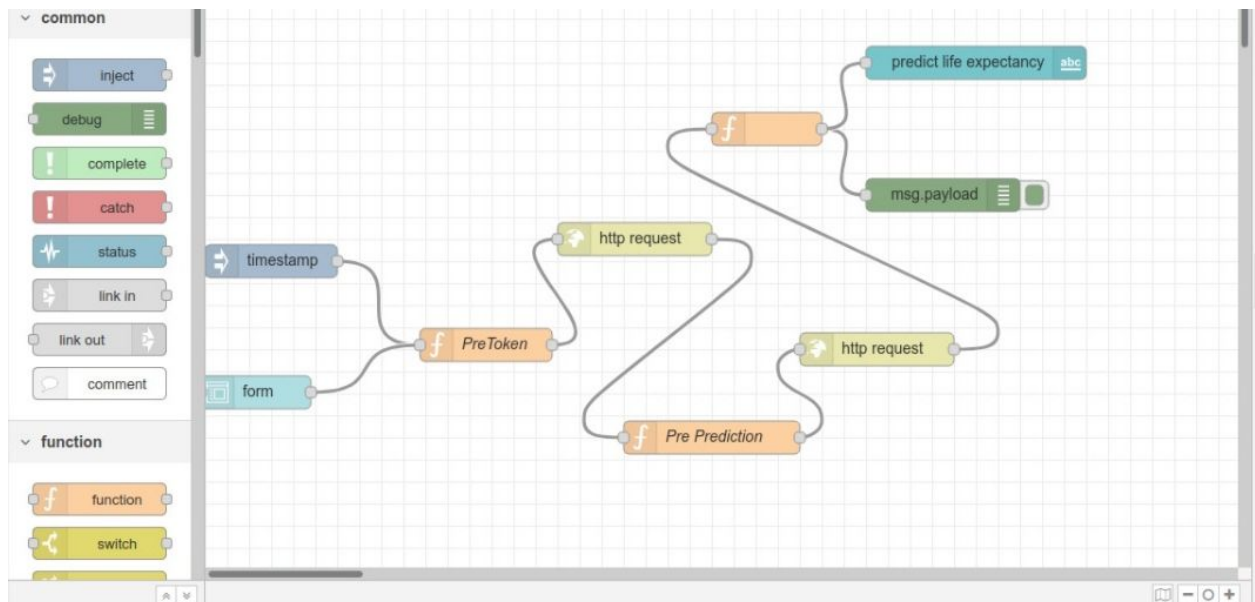
# @hidden_cell
# The following code accesses a file in your IBM Cloud Object Storage. It includes your credentials.
# You might want to remove those credentials before you share the notebook.
client_8d5476f97996479eb3ccf78dd2f59c3b = boto3.client(service_name='s3',
    iam_api_key_id='0j6h8Qc_mCKtdjdQzk7NAtVQqEssFTDMYxIe8b1b3fL2',
    iam_auth_endpoint='https://iam.cloud.ibm.com/oidc/token',
    config=Config(signature_version='oauth'),
    endpoint_url='https://s3-api.us-gio.objectstorage.service.networklayer.com')

body = client_8d5476f97996479eb3ccf78dd2f59c3b.get_object(Bucket='myproject-donotdelete-pr-gxh1qz81jnte6e',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 )

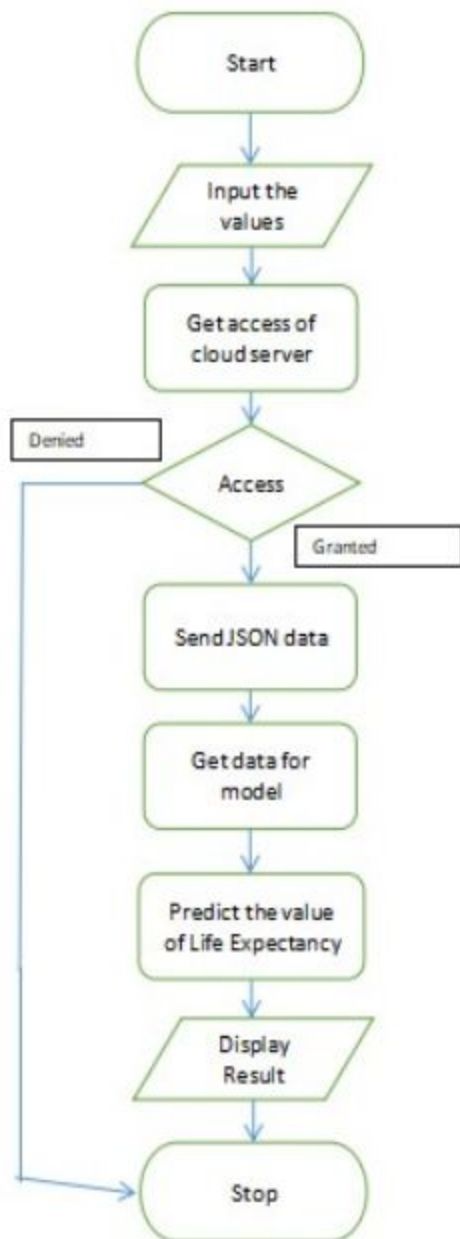
df = pd.read_csv(body)
df.head()
```

Country	Year	Status	Life expectancy	Adult Mortality	Infant deaths	Alcohol	percentage expenditure	Hepatitis B	Measles	Polio	Total expenditure	Diphtheria	HIV/AIDS	GDP	Population	thinness 1-19 years
Afghanistan	2015	Developing	65.0	262.0	67	0.01	71.270674	65.0	116.4	0.0	0.16	65.0	0.1	68.4760210	32736404.0	17.2

Node RED :



4. FLOW CHART



5. RESULT

The notebook attached consists of the detailed steps involved in the pipeline. First the data was imported from cloud storage object of IBM Cloud and stored in a data frame. Later it was checked for null values and necessary steps were taken. Data Visualization in the form of heatmap was performed on different columns to find the relation between life-expectancy and other variables. For the training phase 20% data was separated for the testing phase. Different pipelines were created for numerical and categorical columns.

The screenshot of the web UI has been attached below :

Default

predict life expectancy **81.99799999999996**

Year
2020

Status (1 for developing, 0 for developed)
1

Adult Mortality
1

Infant deaths
1

Alcohol
1

percentage expenditure
1

Hepatitis B
1

Measles
1

BMI
19

under-five deaths
50

Polio
10

Total expenditure
5000000

Diphtheria
12

HIV/AIDS
0.1

GDP
15000

Population
155555555

thinness 1-19 years
1

thinness 5-9 years
1

Income composition of resources
15000

Schooling
1

SUBMIT

CANCEL

6. APPLICATIONS

- a) It can be used to monitor health inequalities of a country.
- b) It can be used to develop statistics for country development process.
- c) It can be used to analyze the factors for high life expectancy.
- d) It is user friendly and can be used by anyone.

7. FUTURE SCOPE

Integrating a data science dashboard which shows different visualizations of Life Expectancy as per the Country and Year. A system to update our model parameters when there is a change in consistency of data like when a sudden epidemic occurs or during a recession, then all the attributes used in our model need to be updated to provide the most precise life expectancy.

8. CONCLUSION

This user interface will be useful for the user to predict life expectancy value of their own country or any other country based on some required details such as GDP, BMI, Year, Alcohol Intake, Total expenditure etc. The advantages of longer life span outweigh its disadvantages. The benefits people and the world can get from a higher life expectancy is irreplaceable and undeniable. It is a truth that life expectancy is a symbol of civilization and better life. Knowing an estimate of how much life we have left pushes us to achieve different things. Higher life expectation is also perceived as greater quality of life and greater income of society.

Our project has automated the entire task of rigorous calculation and removed errors in the existing system and gives the life expectancy to the user. This information can be useful to the society as stated above and this method is also much cheaper than hiring people to do the calculations.

9. ADVANTAGES/DISADVANTAGES

1. The model has been deployed and hence can be used by anyone to predict the result anytime.
2. The model uses Random Forest Classifier which has an R2 score of more than 0.95. Hence, the model is pretty much accurate.

10. BIBLIOGRAPHY

1. Project planning and kickoff :
 - a. <https://www.allbusinesstemplates.com/download/?filecode=2KBA4&lang=en&iuid=9f9faa69-9fab-40ee-8457-ea0e5df8c8de>
 - b. <https://www.youtube.com/watch?v=LOCkV-mENq8&feature=youtu.be>
 - c. <https://github.com/>
 - d. <https://www.zoho.com/writer/help/working-with-text.html>
2. IBM Cloud :
 - a. <https://my15.digitalexperience.ibm.com/b73a5759-c6a6-4033-ab6b-d9d4f9a6d65b/dxsites/151914d1-03d2-48fe-97d9-d21166848e65/>
 - b. <https://developer.ibm.com/tutorials/how-to-create-a-node-red-starter-application/>
 - c. https://www.w3schools.com/howto/howto_make_a_website.asp
3. IBM Watson services :
 - a. <https://www.ibm.com/watson/products-services>
 - b. <https://developer.ibm.com/technologies/machine-learning/series/learning-path-machine-learning-for-developers/>
 - c. <https://bookdown.org/caoying4work/watsonstudio-workshop/jn.html>
4. Dataset:
 - a. [Life Expectancy \(WHO\)](#)

APPENDIX

Notebook:

<https://dataplatform.cloud.ibm.com/analytics/notebooks/v2/84c966c6-0c47-4b72-a8a5-eb26867b02e5?projectid=ccde0cd1-3822-46d6-8470-e74a16f54dae&context=wdp>

Node RED:

<https://myfirstnoderedlondon.mybluemix.net/red/#flow/ce6ad589.a32e48>

Form

link: <https://myfirstnoderedlondon.mybluemix.net/ui/#!/0?socketid=NxJarTmhBSn38S3DA>
[AAC](#)