SmartBridge Educational Services Pvt Ltd.

Project Report

Topic: Predicting Life Expectancy using Machine Learning

Duration: June 6th, 2020 – July 5th, 2020

Guided By -
Mr. Hemant Kumar Gehlot

Made by –
Mr. Mohit Jain

Index:

Project Report

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. A typical Regression Machine Learning project aims at analyzing the historical data in order to predict insights into the future. The problem statement is therefore aimed at predicting life expectancy rate of a country given various features.

1.2 Purpose

The project aims at predicting the average life expectancy rate of people in a country using several features which involve GDP, year, sex, illness, education and other demographic factors. The machine learning approach is used to create a model that could describe how these factors influence the expectancy rate of the individuals in the country. Linear regression algorithm is used to determine the life expectancy rate and its relation with the other factors and predict the future life expectancy rate based on the various factors given. The purpose is to analyse the mathematical relationship between the expectancy rate of the individuals and the features on which the life expectancy rate depends. The algorithm learns from the data given and predicts on the new features given to it based on the insights gained from past data.

2. LITERATURE SURVEY

2.1 Existing Problem

Predicting a human’s life expectancy has been a long-term question to humankind, and there have been many attempts to make the prediction accurate and popular since the prevalence of smartphones and apps. However, the effectiveness of those apps is limited due to the constraints of developing a classification of meta-data, such as the complexity and variety of environmental, geographic, genetic, and living factors of humans. For example, a report showed that people living in a village called Yuzurihara in Japan, also known as “the village of long life”, were ten times more likely to live beyond the age of 85 than anywhere in North America. These people also had similar traits such as smooth skin, flexible joints and thick hair. This implies that geographic and living environments affect the longevity of human life, and the use of statistics can make it possible to forecast a life expectancy of a person who lives in a similar environment village with a similar lifestyle. Several researches have been done on the approaches to predict the life expectancy rate. The problem of processing datasets such as electronic medical records (EMR), and their integration with genomics, environmental factors, socioeconomic factors and patient behaviour variations have posed a problem for researchers in the health industry. The calculation of life expectancy is a complicated process and requires many variables and circumstances to take into account, there have been several attempts to create an equation despite it being impractical to simplify these variables into one equation.

2.2 Proposed Solution

Due to the evolution of data science technologies such as big data virtualization and analytics, data wrangling and with the cloud, health workers now have an improved way of processing and developing meaningful information from huge datasets that have been accumulated over many years. The use of machine learning algorithm and statistical approach can make it quite easier to predict the life expectancy of the individuals. Several variables affecting the life expectancy rate can be recorded and a mathematical relationship can be established to know their influence on the life expectancy of the individuals. The statistical approaches have therefore resolved the problem of processing huge data and analysing them and also the processing power has been reduced due to several approaches of data analytics.

3. THEORETICAL ANALYSIS

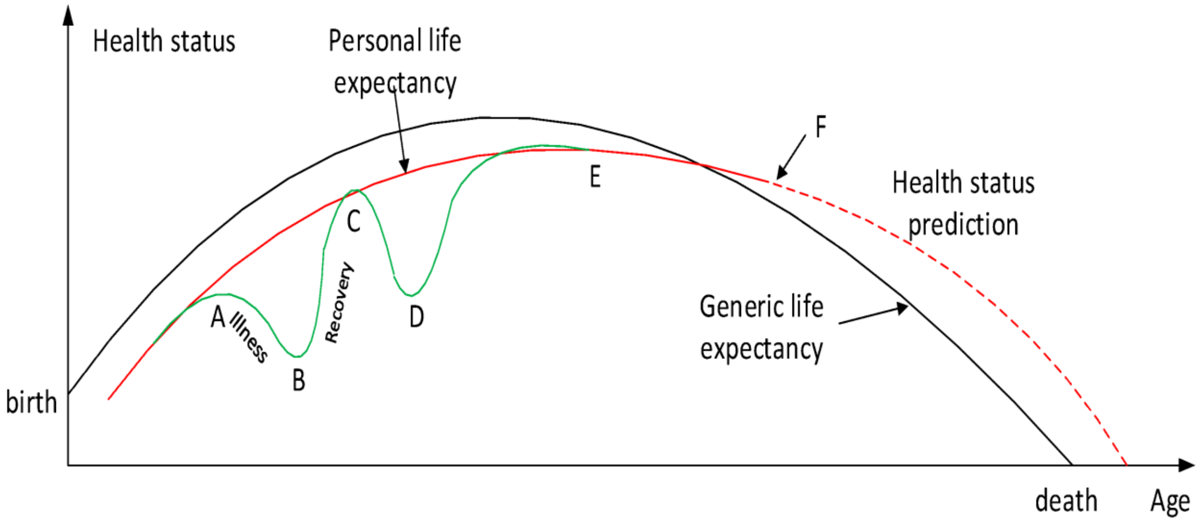
The problem of processing datasets such as electronic medical records (EMR), and their integration with genomics, environmental factors, socioeconomic factors and patient behaviour variations have posed a problem for researchers in the health industry.

Due to the evolution of data science technologies such as big data visualization and analytics, data wrangling and with the cloud, health workers now have an improved way of processing and developing meaningful information from huge datasets that have been accumulated over many years.

For example, big data and machine learning techniques can benefit public health researchers with analysing thousands of variables to obtain data regarding life expectancy and anxiety disorders. They used the demographics of selected regional areas and multiple behavioural health disorders across regions to find correlations between individual behaviour indicators and behavioural health outcomes.

It may be possible to create a prediction of personal life expectancy, which can be further used to calculate health indexes on a generic level for which the individualized expectancies may be compared against.

This model can also depict the life expectancy predicted by an inference system, which transmits health data over wireless sensor networks. Generic life expectancies may be calculated from multiple sources obtained and analysed by big data. These values can be used to create a personalized graph that most resembles the individual in question, with consideration for personal characteristics such as their age, gender, ethnicity, living environment and current comorbidities or lifestyle habits. There may be an enormous number of variables to consider, of which increasing the number may obviously increase the accuracy of a LE prediction.



This personalized graph can then be compared with other individuals who may be living similar lifestyles and share similar traits to provide an idea of the generic life expectancy.

This concept is described in Figure with the red and black lines superimposed on each other.

During point A, the graph shows that the user was ill and there was a decrease in the overall health status until point C when the user recovered, followed by another case of illness at point C until they recovered finally at point E.

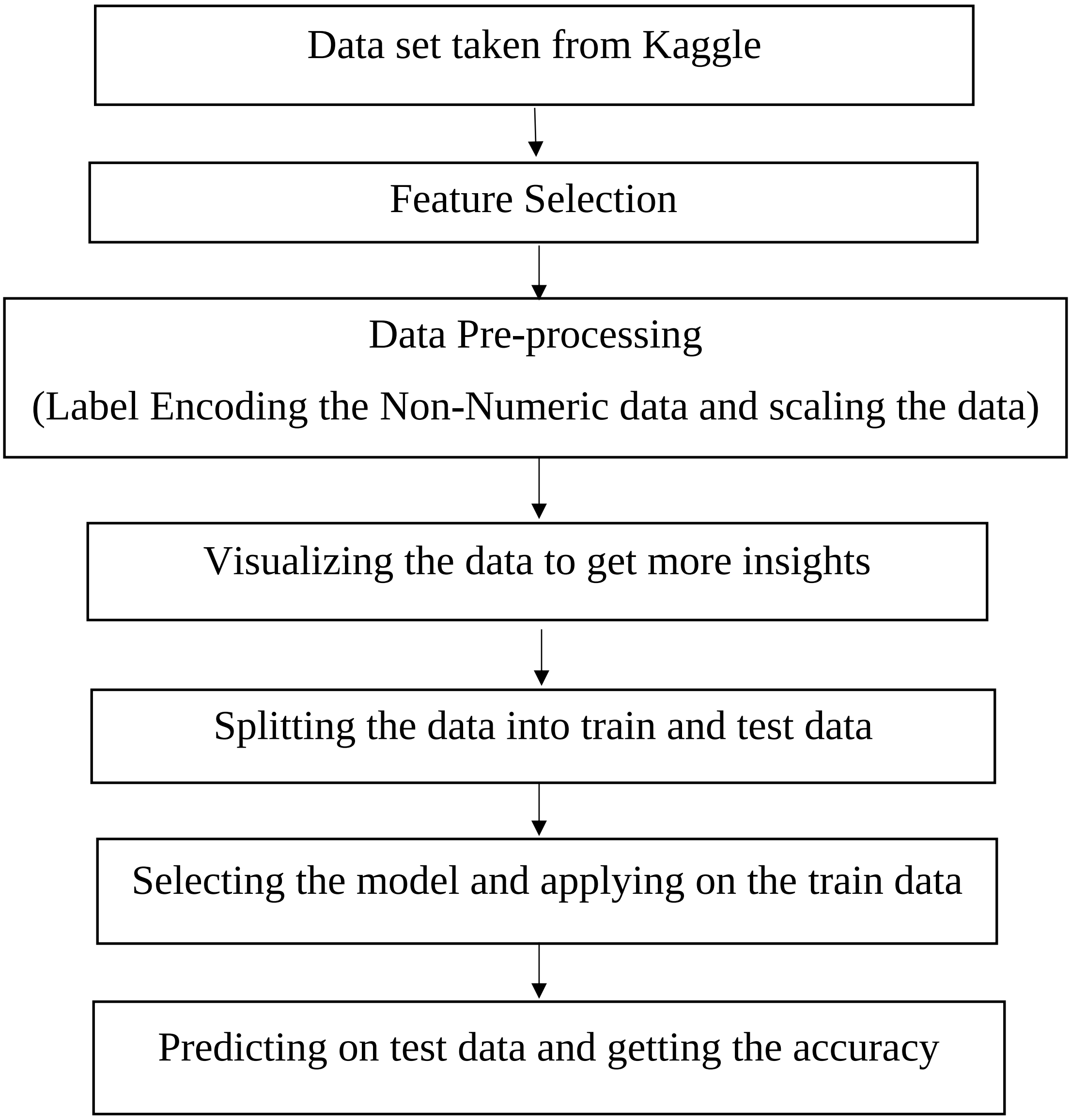
Point F describes the present moment in time, at which point any values following this point would be an inferred prediction of the individual's health status for the future.

This prediction would be an inference made of physiological data analysed by the cloud computation. Whilst the red line shows a trend line of the history of the user's health status and a rough estimate of his or her future health, the graph may also show fine trends as depicted by the green line. This would be used to inform short term information about whether the user's health is improving or declining on a more detailed level.

4. EXPERIMENTAL INVESTIGATION

Several factors affecting the life expectancy rate were taken into consideration which involved GDP, country, status of the country, adult mortality and so on. Research papers were studied in order to know the top most factors that contribute most towards life expectancy prediction. The dataset was taken from Kaggle which consisted of 22 variables and 2939 rows. The existing works were analysed and studied to gain more insights about the problem statement and to improve the solutions to the given problem.

5. FLOW CHART


Data set taken from Kaggle
Feature Selection
Data Pre-processing
(Label Encoding the Non-Numeric data and scaling the data)
Visualizing the data to get more insights
Splitting the data into train and test data
Selecting the model and applying on the train data
Predicting on test data and getting the accuracy

6. RESULT

- Successfully deployed the Machine Learning Model to Predict Life Expectancy **using Python** and created the UI using Node-RED application
  - ➢ R2 score = 0.818
  - ➢ Mean Squared Error = 15.90
- Successfully deployed the Machine Learning Model to Predict Life Expectancy **using Auto-AI of IBM Cloud** and created the UI using Node-RED application.
  - ➢ R2 score = 0.95
  - ➢ Root Mean Squared Error = 1.98

7. Advantages & Disadvantages

1. Comparison to human performance

To put the reported results in perspective, we provided a comparison of the model’s performance to human performance as described by. To make a truly valid comparison, our study design should include judgments about life expectancy from GPs about the actual patients that the medical records used for this research correspond to. Making this comparison was however impossible within the scope of this research, and with the use of this dataset.

2. Data limitations

One of the main challenges we faced during this project was the amount of available data. Our dataset’s size was a fair amount of data according to clinical standards, but is not considered to be a lot of data for training machine learning model.

3. Transparency

When it comes to incorrect predictions, both the baseline and the keyword model tend to make overly pessimistic predictions. It would be interesting to investigate why the models have a tendency toward overly pessimistic predictions, despite being trained with and tested on balanced data.

8. APPLICATIONS

Probing into sanitation:

To dig deeper into how sanitation specifically, I evaluated access to sanitation against statistics on child mortality rates.

Not surprisingly, there’s a strong negative correlation between sanitation and mortality: As sanitation improves, mortality rates decrease for neonates, infants, and children under 5.

According to the World Health Organization, diarrheal disease, the leading cause of death for children under five, is spread by poor sanitation conditions.

9. CONCLUSION

The life expectancy rate of the individuals was predicted by applying multiple linear regression. It was concluded that the factors like adult mortality, GDP, BMI, under five deaths and thinness 1-19 years were the factors that most affected the life expectancy rate and therefore showed a strong correlation to the life expectancy rate. The developed country tends to have high life expectancy due to better standards of living and also better health care facilities. Adult mortality showed the strongest negative correlation to the life expectancy rate. The model predicted the life expectancy quite accurately and the results obtained were acceptable.

10. FUTURE SCOPE

One possibility is that the pace of age-specific mortality improvement over the next half century will be similar to the pace of improvement over the last 50 or 100 years. A second possibility is that the pace of life-expectancy increase over the next half century will be similar to the rate of increase over past decades. Finally, the third possibility is that mortality improvements will accelerate in the future. Biology and biomedicine may be on the verge of unprecedented breakthroughs in knowledge about specific diseases and about the aging process itself – many knowledgeable scientists are of this opinion. Specifically, instead of increasing by 2.5 years per decade, life expectancy may increase by 3, then 4, and then 5 years per decade over the next three decades and perhaps by 6, 8, or even 10 years per decade in the 2030s and 2040s.

11. BIBLIOGRAPHY

- https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-019-0775-2#Sec24
- https://link.springer.com/chapter/10.1007/978-3-030-05075-7_6
- https://cloud.ibm.com/docs/overview?topic=overview-whatis-platform
- https://www.kaggle.com (DATA SET REFERENCE)
- https://www.mdpi.com/2227-7080/6/3/74/htm