Project title : Predicting Life Expectancy using Machine Learning

Name : Shreya Singh

Project Summary :

A typical Regression Machine Learning project leverages historical data to predict insights into the future. 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: Regional variations, Economic Circumstances, Sex Differences, Mental Illnesses, Physical Illnesses, Education, Year of their birth and other demographic factors. This problem statement provides a way to predict average life expectancy of people living in a country when various factors such as year, GDP, education, alcohol intake of people in the country, expenditure on healthcare system and some specific disease related deaths that happened in the country are given.

Project Requirements :

Functional Requirements , Technical Requirements, Software Requirements : Python, IBM Cloud , IBM Watson, Regression model, Visualization Library - Matplotlib, Node Red.
Dataset : 193 countries, 21 features/factors present which affect the Life Expectancy of an individual.

Project Team - Individual

Project Schedule - 1 month

- Project planning and kickoff
- Setting up development environment - Creation of github account
- Creation of IBM cloud account
- Creating a basic Node Red application
- Exploring IBM Watson Studio and usecases
- Studying various algorithms such as classification, regression and clustering to choose best algorithm for predicting the life expectancy
- Building the model using Multivariate Linear Regression
- Comparing the model with the performance of Neural Networks on the same data
- Building the Node Red flow and integrating it with the model

Project Deliverables

The outcome of the project is the prediction of the Life expectancy of an individual based on certain factors such as Schooling, diseases, GDP of the country, population, Alcohol consumption, percentage expenditure etc.
It was found that factors such as the GDP, illnesses such as Diptheria, Alcohol consumption and Schooling, significantly impacted the overall regression output. Other factors that less impacted the outcome were Country, Hepatitis B, Adult Mortality and Status i.e. whether a country was developing or developed. 70 percent data was trained upon and 30 percent kept for testing purposes.
The model successfully calculated the Life expectancy with a mean absolute error as small as 3.4468 and a root mean square error of 5.4594. Below is the scatter plot testing the predicted Life expectancy and the actual Life expectancy as provided in the dataset.