Predicting Life Expectancy using Machine Learning

Functional requirements

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 the dataset.

- Predict life expectancy with maximum accuracy .
- Process the dataset and find dependent and independent variables relationship.

Technical requirements

Technical requirements for Predicting Life Expectancy using Machine Learning as following –

- Use python as base language.
- Use IBM watson studio for data processing .
- Use IBM Node-RED for flow-based development using visual programming

.

Software Requirements

Predicting Life Expectancy using Machine Learning is based on the supervised machine learning using linear regression algorithm. we need following software to develop the project –

- IDE(Integrated development environment Software) for write the python code .
- Database software to stone the dataset of the project .
- Communication software to communicate with the team members and discuss the problem and future planes of the project .

- ## Project Deliverables

Predicting Life Expectancy using Machine Learning deliver the outcome for the given dataset of any county. It will predict the life expectancy based on the

- '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'

- ## Project Team

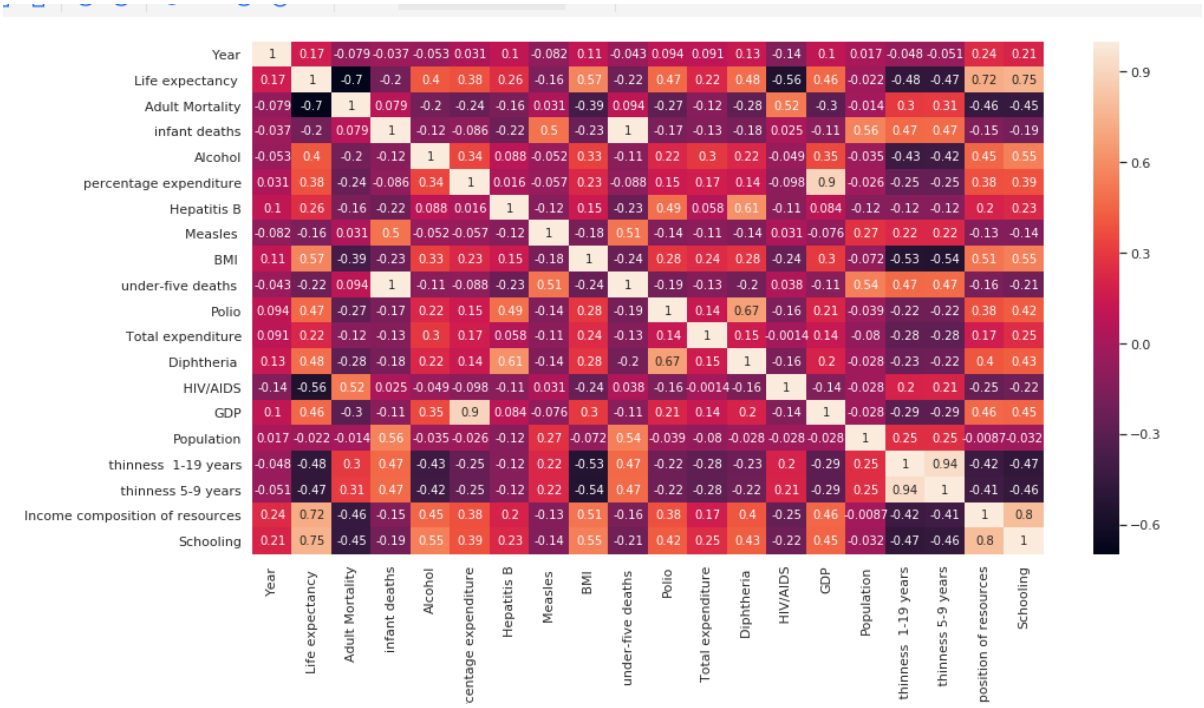
1. TEJAS K N -- Project manager and developer
2. TheSmartBridge -- Project sponser .

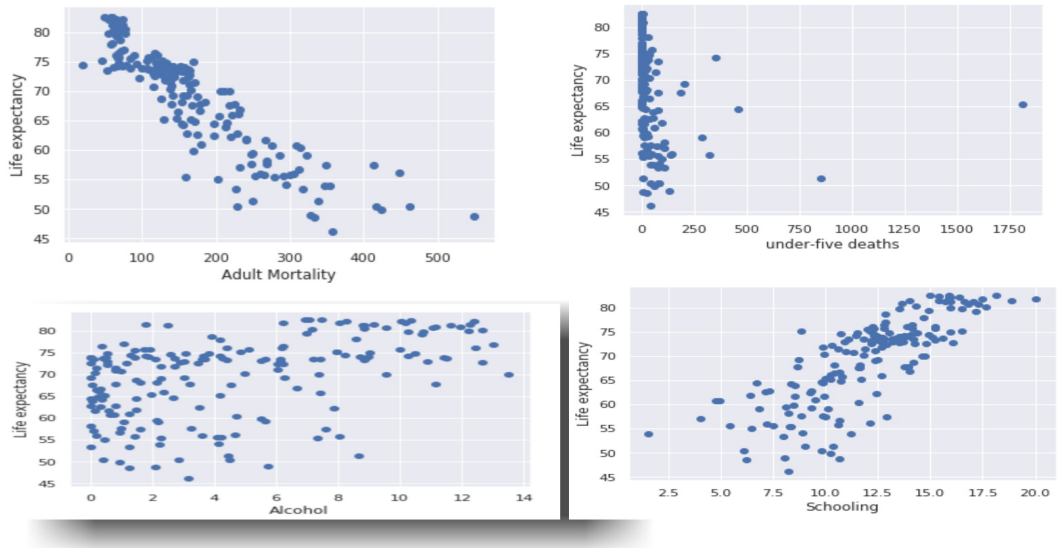
- ## Project Schedule

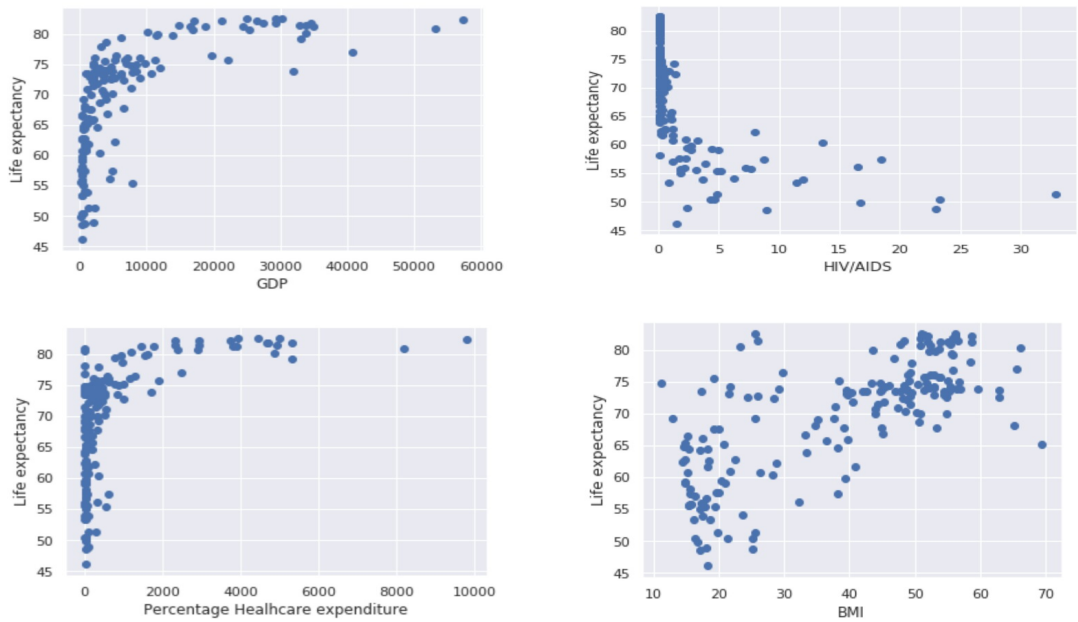
Predicting Life Expectancy using Machine Learning project will be develop completely with maximum accuracy with in given dead line which is 30 days (15 may to 15 june).

- # Exploratory Data Analysis

✓ In this step the **EDA** is done to select the feature variables that will be used later for machine learning.this is done by plotting scatter plot of life expectency v/s different features such as mortality rate,disease,consumption of alchohol etc..







- Using the scatter plot we plot the Life Expectancy against some other varyables to see if there is any correlation between them. There seem to be a positive correlation between The Percentage of Healthcare Expenditure, Schooling, GDP and BMI and Life Expectancy, while there is a negative one between Adult Mortality, AIDS and Life Expectancy,there does not seem to have any correlation between Alcohol, under 5 years – old deaths and Life Expectancy.

• Preprocessing the data

- Data preprocessing is a data mining technique that involves transforming raw data into understandable format.Real world data is often inconsistent , incomplete ,lacking in certain behaviours or trends and likely to contain errors.Data Preprocessing prepares raw data for futher processing.

- The raw data is not suitable for us to start building a model so some preprocessing will be done. First the Status of the country is turned into numberical with the get_dummies function, so we get 2 new columns. The original column is being dropped. Second the data is being grouped by the country and we find the mean values during the 2000 – 2015 year period. Then the Life expectancy column is removed to form the life_labels variable or the output, and the rest is stored as the life_features variable. Now we

consider that we have some null values in the table, the isnull function has been used to find the with the boolean True. Below that the number of null values are displayed in each separate column. It is mostly situated in the Population and GDP columns. Now the missing values are filled with the mean of its respective column. This will create some distortions, but the other option in removing parts of the table will shrink the data so it will be avoided here because the number of rows is not that high. The final shape for the life_features is 193 rows to 20 columns. Finally considering the large differences in the values of the columns, there will be some scaling with the MinMaxScaler function. Now we will split the data into a training part of 70% and a testing of 30%. Cross validation will be initialized with the creation of 5 fold split

- ✓ The following shows the null data in dataset:--
  Adult Mortality -10,Alcohol-2,Hepatitis B-9,BMI-4,Total expenditure-2,GDP-30,Population-48,
  thinness 1-19 years-4,thinness 5-9 years-4,Income composition of resources-17,Schooling-13

- ✓ After filling null values with mean value in the dataset it can be used for train test split data.

• Model selection

- ✓ In this step tried linear regression ,Random forest regression,Decision tree regression we compare to choose best algorithm.We will use both the properties of the Ridge regression and the Lasso and eventually the ElasticNet to see if the score can be improved.

- ✓ After comparing all the algorithms we can conclude the Lasso and the Elastic Net Regression offer which are the same:

  Best Parameters: {‘alpha’: 0, ‘max_iter’: 10}
  R square on the test data of 92%
  MAE of 1.83
  MSE of 6.05

- ## Node red Integration

✓ **Node-RED** is a programming tool for wiring together hardware devices, APIs and online services in new and interesting ways. It provides a browser-based editor that makes it easy to wire together flows using the wide range of **nodes** in the palette that can be deployed to its runtime in a single-click.

✓ Here Node red is used for machine learning service integration.A scoring endpoint is created using WatsonMachineLearningClient API and used.