PROJECT REPORT

INTRODUCTION

Overview

This project is based on Machine Learning in which the goal is to predict the Life Expectancy using historical data. Life Expectancy is a statistical measure of the average time an organism is expected to live, based on the year of its birth, its current age and other demographic factors including gender. The most commonly used measure is life expectancy at birth (LEB). Life Expectancy is depends on various factors like 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 health care system and some specific disease related deaths that happened in the country. This project aims to automate this task and provide life expectancy when values for different factors are given.

Scope

The problem of processing datasets such as electronic medical records(EMR) and their integration with genomics, environmental factors, socioeconomic factor and patient behavior variations have posed a problem for researchers the health industry. Due to rapid innovations in machine learning field such as big data, analytics,visualization, deep learning, health workers now have improved way of processing, and developing meaningful information from huge datasets that have been accumulated over many years .
Big data 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.

Purpose

An important point to bear in mind when interpreting life expectancy estimates is that very few people actually die at the age indicated by life expectancy, even if mortality patterns stay constant.
Most will die much earlier or much later, since the risk of death is not uniform across the lifetime. Life expectancy is the average.
In societies with high infant mortality rates many people die in the first few years of life; but once they survive childhood, people often live much longer. Indeed, this is a common source of confusion in the interpretation of life expectancy figures: It is perfectly possible that a given population has a low life expectancy at birth, and yet has a large proportion of old people.
So life expectancy gives a detail about the health care conditions of a country.

LITERATURE SURVEY

Existing System

In practical terms, estimating life expectancy entails predicting the probability of surviving successive years of life, based on observed age-specific mortality rates. Age-specific mortality rates are usually estimated by counting (or projecting) the number of age-specific deaths in a time interval (e.g. the number of people aged 10-15 who died in the year 2005), and dividing by the total observed (or projected) population alive at a given point within that interval (e.g. the number of people aged 10-15 alive on 1 July 2015).
To ensure that the resulting estimates of the probabilities of death within each age interval are smooth across the lifetime, it is common to use mathematical formulas, to model how the force of mortality changes within and across age intervals. Specifically, it is often assumed that the proportion of people dying in an age interval starting in year and ending in year corresponds to , where is the age-specific mortality rate as measured in the middle of that interval (a term often referred to as the 'central death rate' for the age interval).

Once we have estimates of the fraction of people dying across age intervals, it is simple to calculate a ‘life table’ showing the evolving probabilities of survival and the corresponding life expectancies by age. Period life expectancy figures can be obtained from ‘period life tables’ (i.e. life tables that rely on age-specific mortality rates observed from deaths among individuals of different age groups at a fixed point in time). And similarly, cohort life expectancy figures can be obtained from ‘cohort life tables’ (i.e. life tables that rely on age-specific mortality rates observed from tracking and forecasting the death and survival of a group of people as they become older).

For some countries and for some time intervals, it is only possible to reconstruct life tables from either period or cohort mortality data. As a consequence, in some instances—for example in obtaining historical estimates of life expectancy across world regions—it is necessary to combine period and cohort data. In these cases, the resulting life expectancy estimates cannot be simply classified into the ‘period’ or ‘cohort’ categories.

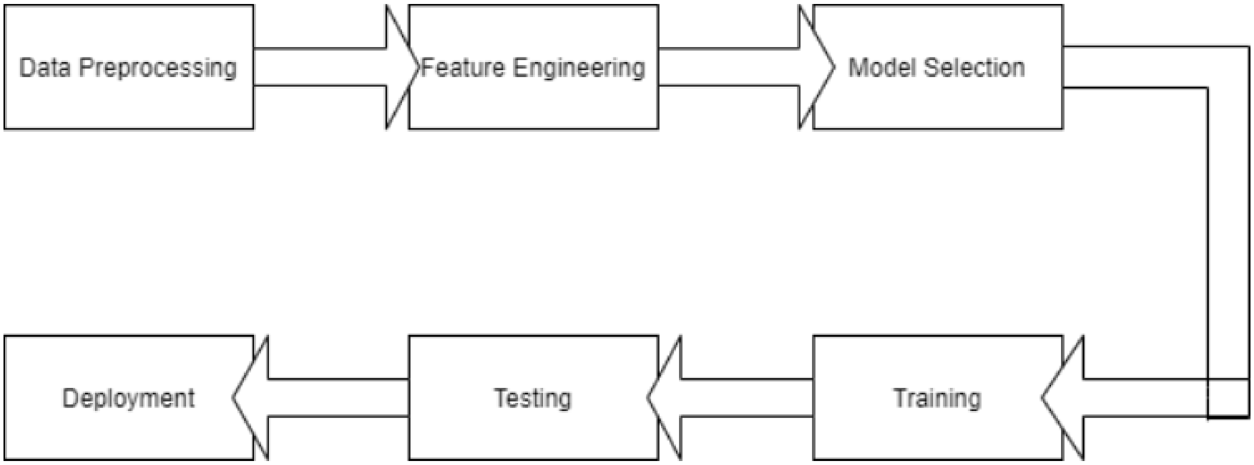
Proposed System

Our proposed system make this whole process of calculating Life Expectancy much easier so any one can calculate the Life Expectancy without any domain knowledge. So, for calculating Life Expectancy we consider some attributes like Regional variations, Economic Circumstances, Sex Differences, Mental Illnesses, Physical Illnesses, Education, Year of their birth and other demographic factors etc. But our proposed system makes this calculation automated and this system has a predicting tool which can predict the Life Expectancy from various attributes value.

So, using machine learning technique we suppose to predict the value of Life Expectancy based on some common attributes like 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 etc. Any one can find this data and get the Life Expectancy value based on the their Country and Year.

THEORITICAL ANALYSIS

Block Diagram



Schedule and strategies

Dataset preparation and preprocessing

Data is the foundation for any machine learning project. The second stage of project implementation is complex and involves data collection, selection, preprocessing, and transformation. Each of these phases can be split into several steps.

Data collection

This is the first step in a machine learning project.We have to find ways and sources of collecting relevant and comprehensive data, interpreting it, and analyzing results with the help of statistical techniques.

The type of data depends on what you want to predict. There is no exact answer to the question “How much data is needed?” because each machine learning problem is unique. In turn, the number of attributes data scientists will use when building a predictive model depends on the attributes' predictive value.

Data visualization

A large amount of information represented in graphic form is easier to understand and analyze. Some companies specify that a data analyst must know how to create slides, diagrams, charts, and templates.
Most of the times visualization helps us in finding correlations and outliers which are not visible when we look at the raw data.

Labeling

Supervised machine learning, entails training a predictive model on historical data with predefined target answers. An algorithm must be shown which target answers or attributes to look for. Mapping these target attributes in a dataset is called labeling.

Data selection

After having collected all information, we choose a subgroup of data to solve the defined problem.

Data preprocessing

The purpose of preprocessing is to convert raw data into a form that fits the required model . Structured and clean data helps in getting more precise results from an applied machine learning model. The technique includes data formatting, cleaning, and sampling.

Data transformation

In this final preprocessing phase, we transform or consolidate data into a form appropriate for machine learning. Data can be transformed through scaling, normalization, attribute decompositions, and attribute aggregations. This phase is also called feature engineering.

Dataset splitting

Any dataset for predictive analysis should be partitioned into three subsets — training, valiidation and test sets

Training set

We create a training set to train a model and define its optimal parameters known as hyperparameters which helps in increasing the accuracy of the model in case of classification or decreasing the loss in case of regression task.

Validation set

The validation set is used to evaluate a given model, but this is for frequent evaluation. We use this data to fine-tune the model hyperparameters. Hence the model occasionally *sees* this data, but never does it "*Learn*" from this. We use the validation set results and update higher level hyperparameters. So the validation set in a way affects a model, but indirectly. A small portion of data is separated from training set and used as validation dataset.

Test set.

The Test dataset provides the gold standard used to evaluate the model. It is only used once a model is completely trained(using the train and validation sets). The test set is generally what is used to evaluate competing models . Many a times the validation set is used as the test set, but it is not good practice. The test set is generally well curated. It contains carefully sampled data that spans the various classes that the model would face, when used in the real world.

Model training

After we have preprocessed the collected data and split it into three subsets,we can proceed with a model training. This process entails “feeding” the algorithm with training data.

Modeling

During this stage, we train numerous models to see which one of them provides the most accurate predictions. We can use cross validation to find the most suitable hyperparameters. In this stage we observe the loss from our model and introduce new parameters like l1,l2 regularization ,weight decay to avoid overfitting.

Deployment

The wml_credentials (created during watson studio instantiation phase) are used to save the model and create a scoring endpoint for our model which will be used in node red application.
A flow is constructed using different components of nod red like forms, https requests, text fields, functions.
Input is given to the application through a form and the functions are supplied with API keys, Instance IDs and scoring endpoint to connect to the model and create an output. The output is displayed through a text field.

Requirements and Deliverables

Technical Requirements

- A basic knowledge of machine learning algorithms and practices along with mathematics knowledge is required to perform the regression task.
- In-depth knowledge of python and different machine learning libraries like sklearn, pandas, numpy and also visualization libraries like matplotlib, seaborn.
- Knowledge and practice is required to use different IBM Cloud services like watson studio and node red.

Hardware Requirements

- Any device like Smart Phone, Desktop, Laptop and Tablet or Similar kind of device.
- Internet Connectivity

Software Requirements

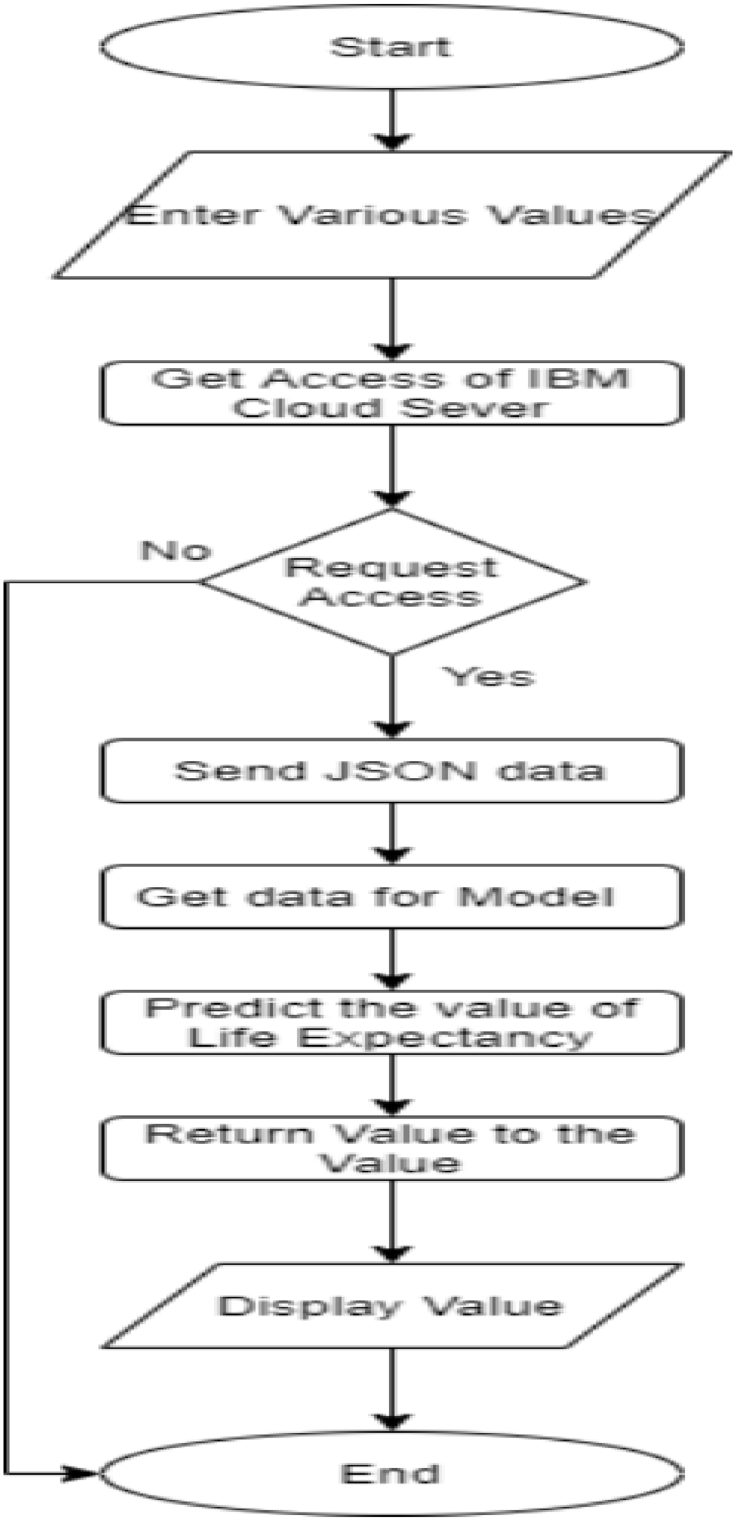
- Any environment which can run python. For example Anaconda distribution, which is excellent for all sorts of data science purposes.
- IBM Cloud account.
- Any compatible web browser to run ibm cloud services.
- Watson studio.
- Node red.

Project Deliverables

- Python notebook containing all the code.
- A node red application which can input data and outputs a prediction for life expectancy.
- A json file containing the architecture of node red project.
- Notebook created using Autoai.
- URL of the node red application.

Experimentation and Results

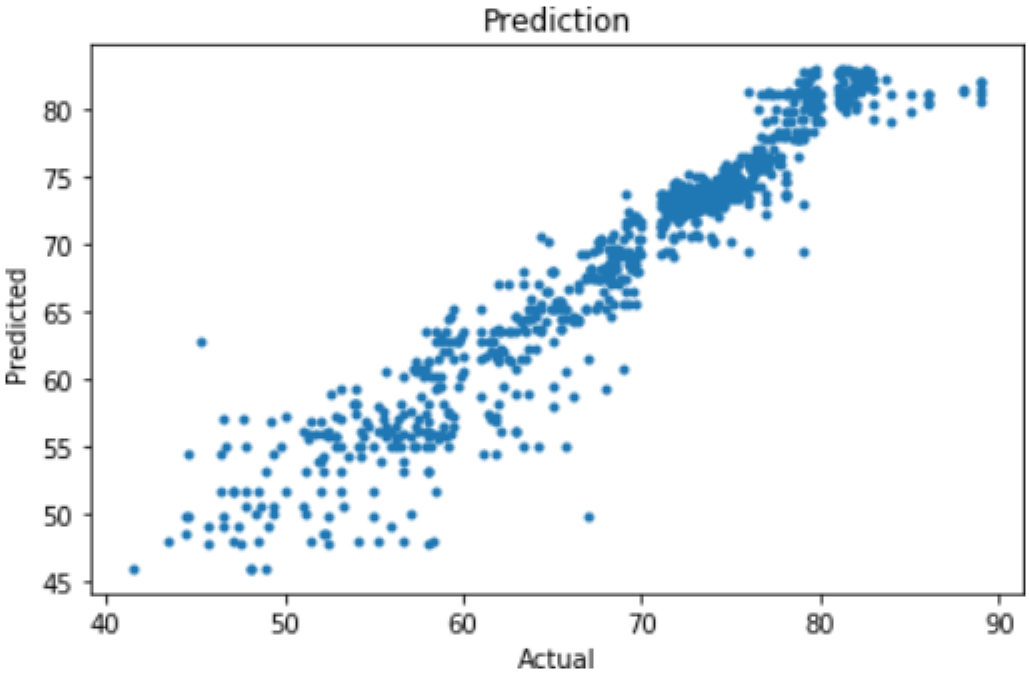
Flowchart



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 dataframe. Later it was checked for null values and necessary steps was taken (filling the null values with the mean of the column). Data Visualization was performed on different columns to find the relation between life-expectancy and other variables.
For training phase 20% data was separated for testing purpose. Different pipelines were created for numerical and categorical columns. The model was trained using ExtraTreesRegressor and the M.S.E. came out to be 2.834 and r_2 score of 97.102 .

Watson studio's machine learning application AUTOAI was also used to make a model and its RMSE came out to be 1.819 and r_2 score of 96.4

The plot between actual and predicted values of Life Expectancy is given below.



The plot is roughly a straight line which indicates a linear relationship between predicted and actual values.

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 most precise life expectancy.

Conclusion

The advantages of longer life span outweigh its disadvantages. The benefits people and the world can get from a higher life expectancy are 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 percieved 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.