

PROJECT NAME: IISPS_INT_2040_Predicting life expectancy of a country using Machine Learning

NAME: Disha Deb

DATE: 22/05/20

PROJECT SCOPE DOCUMENT

INTRODUCTION:

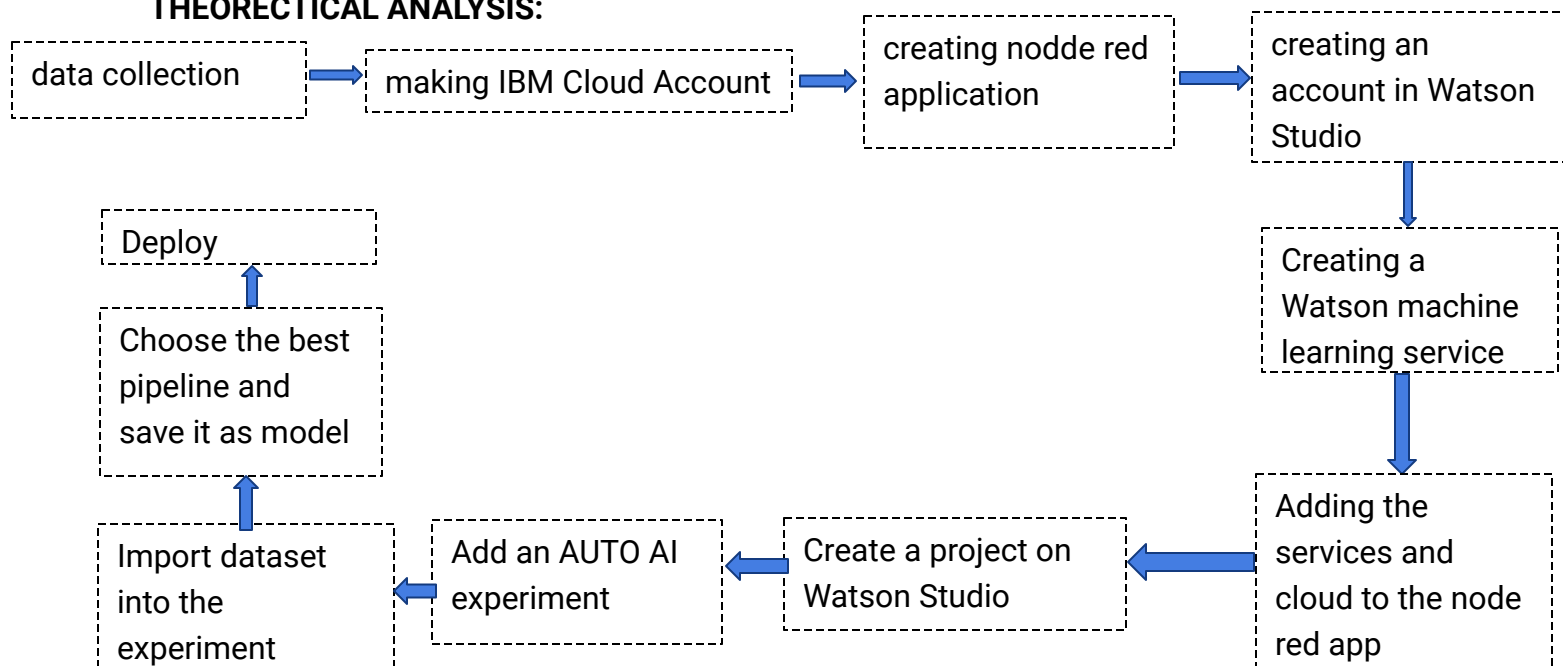
Overview: A data set containing various factors like gender, diseases, education etc. which affect the life expectancy of the residents of a country. A regression model is created to predict the average life expectancy of people in a country.

Purpose: The purpose of the project is to build a regression model to predict the life expectancy of a country based on various factors like economy(GDP, expenditure, income etc.), diseases(Measles, HIV/AIDS etc.). This project will be helpful to social workers who are trying to decide which problem aspect of a country needs to be dealt with first in order to improve the life expectancy.

LITERATURE SURVEY:

Number of surveys have been performed by both Government and independent bodies like Niti Aayog. Surveys have also been conducted by various Universities. Number of life calculators are available on the Internet that make use of the data collected by these surveys to calculate the probable life expectancy of a person. Different surveys have recorded data under different fields depending on the major problems that are prevalent in the area under study like climate, government policies etc. Thus using the data one can study the extent of relation of each factor with each other. Such information is useful for social workers and the Government when deciding policies.

THEORETICAL ANALYSIS:



PROJECT REQUIREMENTS: Life expectancy data set from Kaggle, IBM Cloud, IBM Watson services, Node Red application, GitHub Repository, Zoho writer

EXPERIMENTAL ANALYSIS: Regression Analysis is a form of predictive modeling technique which investigates the relationship between a dependent (target) and independent variable (s) (predictor). This technique is used for forecasting, time series modeling and finding the causal effect relationship between the variables.

An application is created using Node Red provided by IBM Cloud services. Watson services are used to import an AUTO AI experiment into the application. With the help of the AUTO AI a regression model of the life expectancy data set is made. The data set is downloaded from Kaggle. With the help of the Regression model the average life expectancy of the residents of a country is predicted.

Machine Learning uses two sets of data, learning set and testing set. The learning set is used to determine the correlations and predict values while the testing set is used to calculate the error between the actual values and predicted values. What needs to be determined is the percentage of the data set to be used as learning set. Different percentages are tried until a satisfactory efficiency is achieved.

RESULT:

On running the AUTO AI experiment, several pipelines were obtained in the order of accuracy. The best pipeline was saved as both a model as well as a jupyter notebook. The model was then deployed. After deployment, the user can test the model by entering values under the required tab and obtain the predicted value.

By issuing a heatmap from the Jupyter notebook, correlations between the factors was checked. It was observed that life expectancy was closely related to GDP, population and income.

ADVANTAGES AND DISADVANTAGES:

The regression model helps to train the model with the available data in order to predict outcomes for unknown data. The advantage is that the predicted value can be used to prepare for the future. However disadvantage is that it is impossible to attain 0% error in the predicted values. Hence there is no guarantee that the predicted value is correct and dependable.

APPLICATIONS: Social work, making health care policies, identifying areas that require medical aid or economical aid, identifying areas where people are suffering from malnutrition and hunger, etc.

CONCLUSION:

Life expectancy depends on a number of factors like gender, climate, diseases like HIV, physical illnesses, income etc. Using regression model one can determine the factors which affect the life expectancy the most and using the data under those factors one can predict the average life expectancy.

FUTURE SCOPE:

The data set is collected from Kaggle which is maintained by WHO, an international body that seeks to provide health to all the suffering countries. Death by hunger still remains to be a major problem in most of the developing countries. The model will help to decide which country needs an immediate help. Say some aids have been provided to a particular nation and data for the feature values are collected and entered into the regression model. If the predicted value shows an increase in the life expectancy, then that would mean that the efforts are bearing positive outcomes and if not, then new policies and strategies need to be deployed.

BIBLIOGRAPHY:

The entire project has been completed with the help of IBM services. Services like IBM Cloud, IBM Watson Studio and Watson Machine Learning was used. Some guidance was available on IBM website itself as well as from videos uploaded by mentors of SmartBridge. The data set was downloaded as a csv file from Kaggle.com, a site maintained by WHO.