**ML Project**

**Project Scope Document**

**Internship Title:** Predicting Life Expectancy using Machine Learning – SB19684

**Project ID:** SPS\_PRO\_215

**Project Title:** Predicting Life Expectancy using Machine Learning

**By:** Greeshma Maske

1. **Project Description**

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.

1. **Project Requirements**

* Functional Requirements:

To be able to predict the life expectancy accurately using Machine Learning models.

* Hardware Requirements:

Any working laptop/PC with minimum 2.2Ghz processor and at least 8GB of memory with an Internet connection.

* Software Requirements:

Python, IBM Cloud, IBM Watson

For Documentation: Zoho Writer

1. **Project Deliverables**

A machine learning model which should predict the life expectancy of a country given the various factors that may affect it. The factors are: Year, Status, 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.

1. **Work Done**
2. Collect the dataset

A dataset with life expectancy and health factors for 193 countries has been collected from the WHO data repository website and its corresponding economic data was collected from United Nation website. The dataset was obtained from the Kaggle platform.

1. Create necessary IBM Cloud Services

The following services were created on the IBM Cloud platform:

Machine Learning service

Watson Studio service

Node Red Cloudant service

1. Create a new Watson Studio project
2. Create a Jupyter Notebook and import the dataset
3. Build a Machine Learning Model and create endpoints for Node RED Integration

* Import the necessary modules.
* Access the CSV file from your IBM Cloud Object Storage.
* Use info(), describe() and columns to visualize the dataframe.
* The next step is to perform data preprocessing before we can create and train our machine learning model.
* Drop the Country column as won’t be used in our model.
* Identify and handle the missing values using mean imputation.
* Encode the categorical data into binary vectors.
* Perform EDA to  analyze the dataset and uncover its underlying characteristics with visual methods such as graphs and charts.
* To train our ML model, we must first divide our data into features and labels.
* Split the dataset into a training set and a testing set.
* Create the ML Learning model and fit it to the training data set.
* Apply the predict() function to the testing dataset and visualize the predictions.
* Use regression evaluation metrics to evaluate the model.
* Deploy the model using the Watson Machine Learning Client.
* Create endpoints for Node Red integration.

1. Build Node RED flow to integrate ML Services