PROJECT REPORT

SMARTINTERNZ INTERNSHIP

PREDICTING LIFE EXPECTANCY USING MACHINE LEARNING

**By:** Greeshma Maske

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

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

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

TABLE OF CONTENTS

Chapter 1: INTRODUCTION #

Section 1.1: Overview #

Section 1.2: Purpose #

Chapter 2: Literature survey #

Section 2.1: Existing Problem #

Section 2.2: Proposed Solution #

Chapter 3: Theoretical Analysis #

Section 3.1: Block Diagram #

Section 3.2: Hardware/Software Designing #

Chapter 4: Experimental Investigations #

CHAPTER 5: Flowchart #

CHAPTER 6: Result #

CHAPTER 7: ADVAntages & disadvantages #

CHAPTER 8: Applications #

CHAPTER 9: Conclusion #

CHAPTER 10: future scope #

CHAPTER 11: bibliography #

APPENDIX

1. Source Code
2. Dataset
3. **INTRODUCTION**
   1. OVERVIEW

A Machine Learning project that uses Python, IBM Cloud and IBM Watson to predict the life expectancy of a country.

* 1. PURPOSE

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.

1. **LITERATURE SURVEY**

2.1 Existing Problem

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. Life expectancy is a measure that is often used to gauge the overall health of a community. Life expectancy at birth measures health status across all age groups. Shifts in life expectancy are often used to describe trends in mortality.

2.2 Proposed Solution

To design a machine learning model using regression techniques that can predict the life expectancy of a country given the various factors that affect the health of its citizens. Using this model, we can assess the extent of effect these factors have on the average life expectancy. This can give us an idea about the changes that need to be brought about in order to improve the health conditions of people around the world.

1. **THEORETICAL ANALYSIS**

3.1 Block Diagram

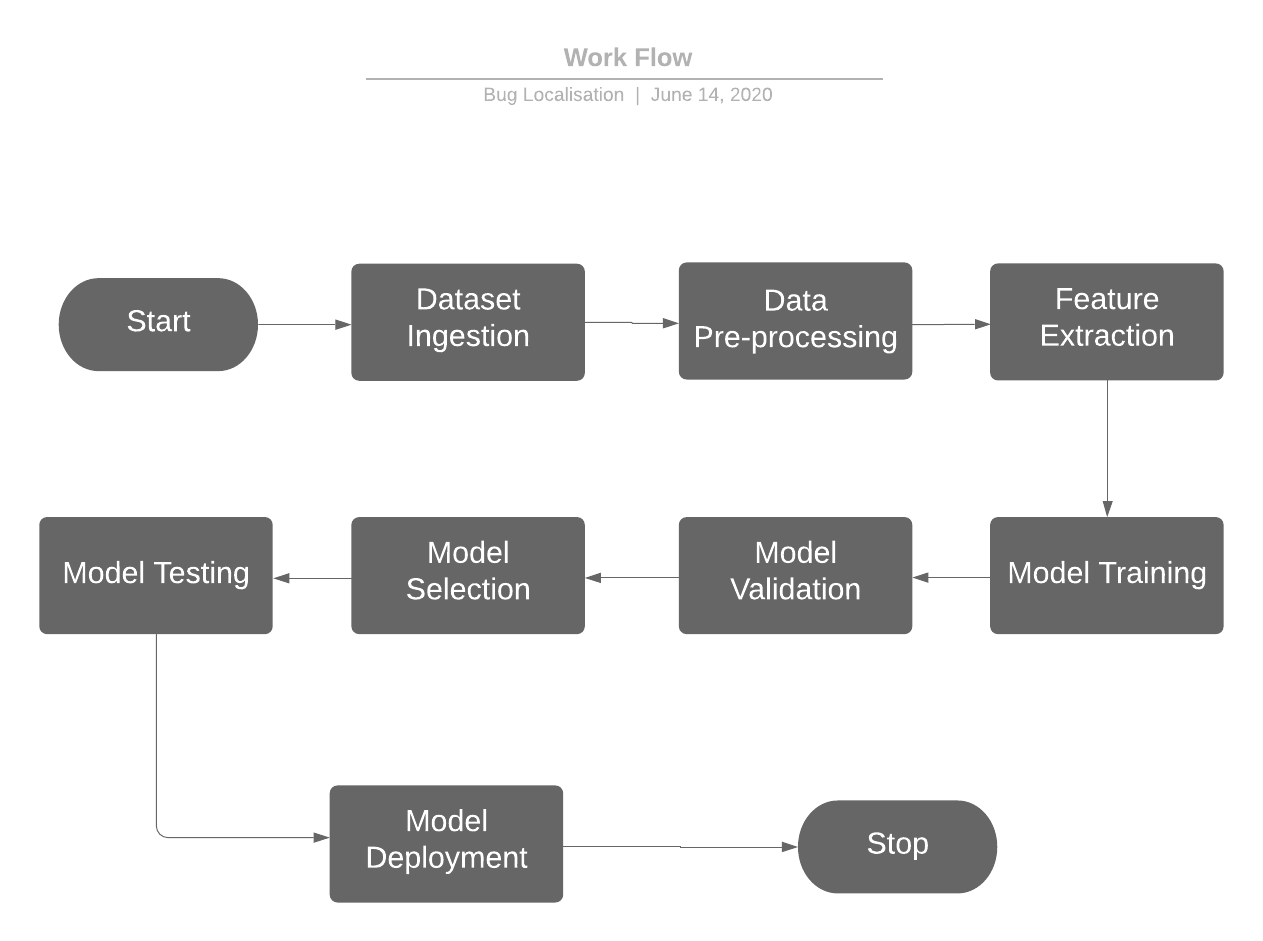


Figure 1: Block Diagram

* 1. Hardware/Software Designing
     1. Hardware Requirements:

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

* + 1. Software Requirements:

Python, IBM Cloud, IBM Watson, GitHub

For Documentation: Zoho Writer

1. **EXPERIMENTAL INVESTIGATIONS**
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 models, 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 models and fit them to the training data set.
* Apply the predict() functions to the testing dataset and visualize the predictions.
* Use regression evaluation metrics to evaluate the models.
* Select the model which gives the best accuracy.
* Deploy the model using the Watson Machine Learning Client.
* Create endpoints for Node Red integration.

1. Build Node RED flow to integrate ML Services

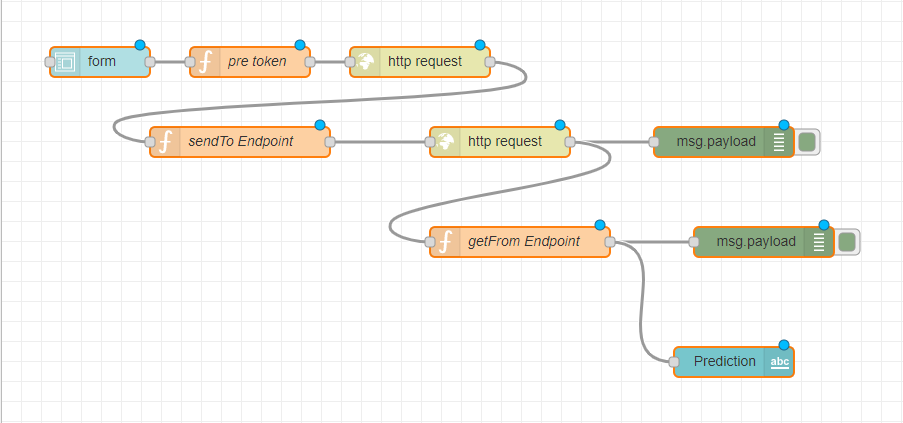


Figure 2: Node RED Flow

1. **FLOWCHART**

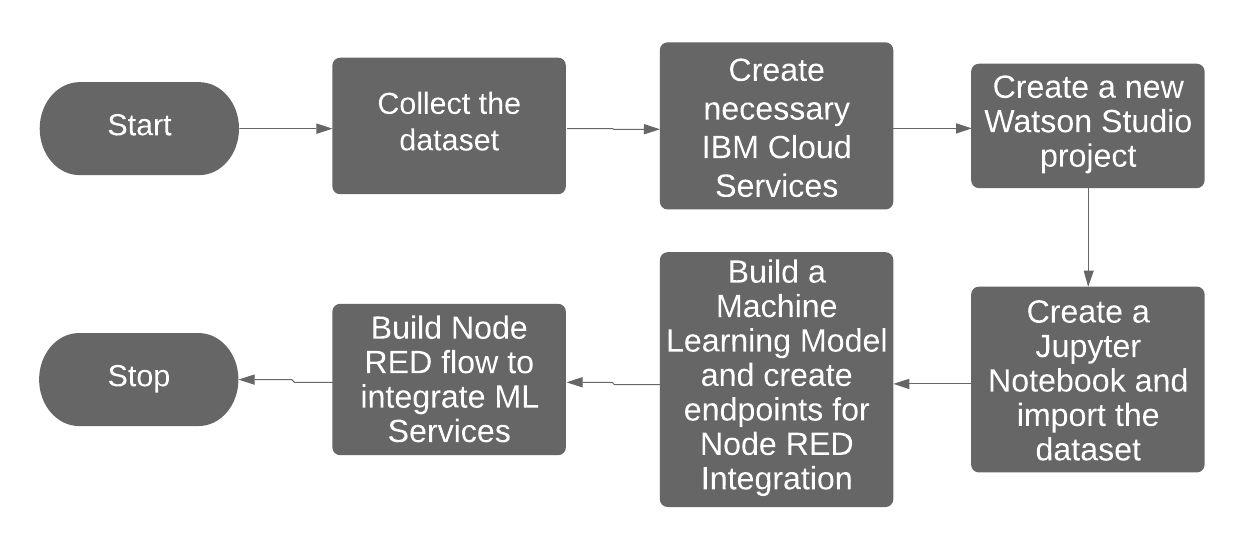


Figure 3: Flowchart

1. **RESULT**

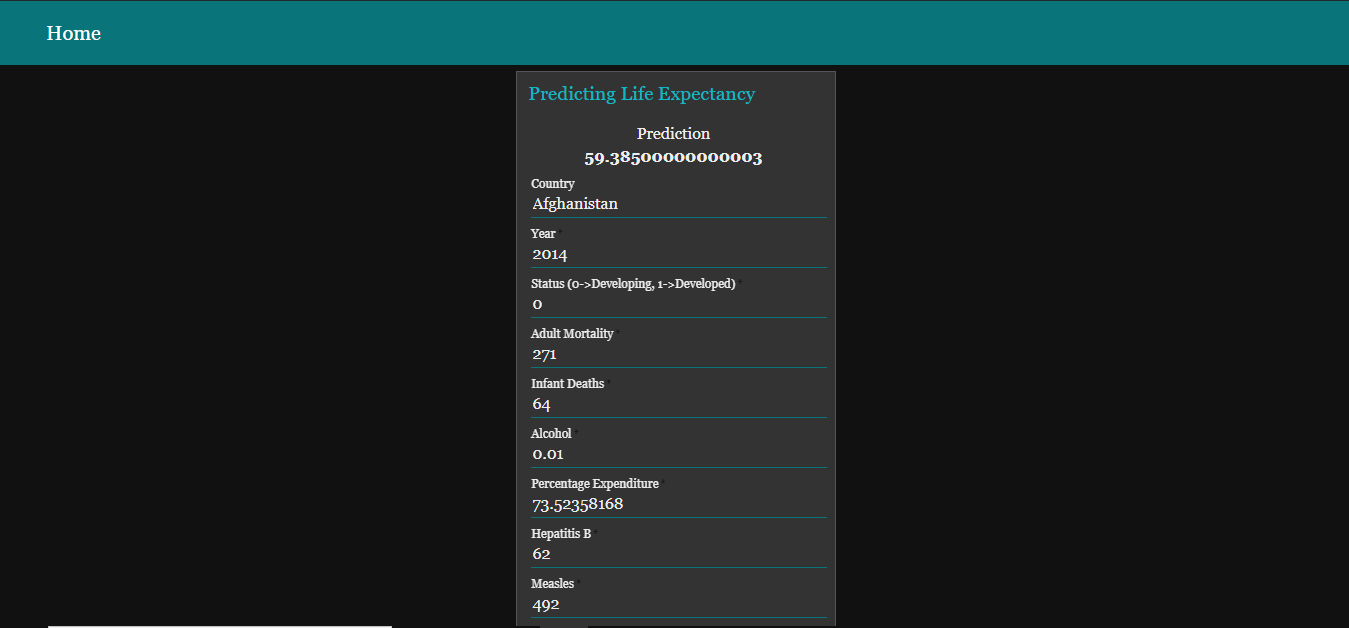


Figure 4: Prediction Page

1. **ADVANTAGES & DISADVANTAGES**
   1. Advantages

* It helps determine which factors matter most, which it can ignore, and how those factors interact with each other.
* It provides a powerful statistical method that allows a business to examine the relationship between two or more variables of interest.
  1. Disadvantages
* It is assumed that the cause and effect relationship between the variables remains unchanged. This assumption may not always hold good and hence estimation of the values of a variable made on the basis of the regression equation may lead to erroneous and misleading results.
* The functional relationship that is established between any two or more variables on the basis of some limited data may not hold good if more and more data are taken into consideration
* It cannot be used in case of qualitative phenomenon.

1. **APPLICATIONS**

* Life expectancy provides a useful measure of average life spans and gives insights into uncertainty about age at death.

* It shows how patterns of change in life expectancy are described by trajectories of mortality improvements over age and time.
* Life expectancy gives an idea of the health issues prevalent in particular country and the healthcare system available to combat these issues.

1. **CONCLUSION**

We implemented a machine learning model using regression analysis to predict the life expectancy of a particular country. We used IBM Cloud, IBM Watson and Node RED to work on this project. The dataset we used related the life expectancy and health factors for 193 countries from the WHO data repository website and its corresponding economic data was collected from United Nation website. The following factors were applied to predict the life expectancy: 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. **FUTURE SCOPE**

In order to extend this work, we can increase the size of the dataset and observe the effect on the results. We can also apply other Machine Learning models in order to obtain higher accuracy.

1. **BIBLIOGRAPHY**

**Appendix**

1. Source Code: <https://github.com/SmartPracticeschool/llSPS-INT-1718-Predicting-Life-Expectancy-using-Machine-Learning>
2. Dataset: <https://www.kaggle.com/kumarajarshi/life-expectancy-who>