Machine Learning Project Report

On

Predicting Life Expectancy using Machine Learning

Made by: - Aditya Borude

Date: 15-06-2020

INDEX

|  |  |  |
| --- | --- | --- |
| **1.** | **Introduction**   * 1. Overview   1.2 Purpose | **1** |
| **2.** | **Literature Survey**  2.1 Existing Problem  2.2 Proposed Solution | **1** |
| **3.** | **Theoretical Analysis**  3.1 Block Diagram   * 1. Software Designing | **2 – 4** |
| **4.** | **Experimental Investigations** | **4** |
| **5.** | **Flowchart** | **5** |
| **6.** | **Results** | **6** |
| **7.** | **Advantages and Disadvantages** | **6** |
| **8.** | **Applications** | **6** |
| **9.** | **Conclusion** | **7** |
| **10.** | **Future Scope** | **7** |
| **11.** | **Bibliography and Appendix** | **7** |

1. Introduction
   1. Overview

Life expectancy refers to the number of years a person is expected to live, based on the statistical average. Life expectancy varies by geographical area, and by era, and various demographic factors.

The project is a web application, developed over a period of 30 days, which is able to predict the average life expectancy of a person in a country based on values of various demographic factors (of that country) calculated over a year.

* 1. Purpose

The purpose of the project is to calculate and output the value of life expectancy, in years, taking the values of various important demographic factors for a country as input for calculation.

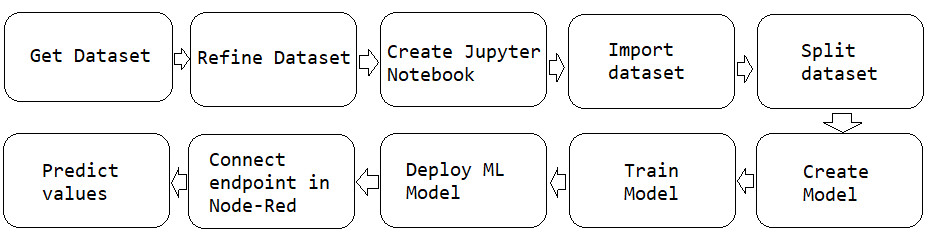
1. Literature Survey
   1. Existing Problem

The life expectancy of a person in a country depends on various factors: economic circumstances, immunizations, illnesses, education, year of birth and other demographic factors. The problem is to devise a way to calculate the life expectancy of people living in a country when various factors such as year, GDP, alcohol intake of people in the country, expenditure on healthcare facilities, and other various demographic factors related to the country are given.

* 1. Proposed Solution

Create a regression machine learning model, trained on a dataset comprising of only those factors which are critical to determining life expectancy. This machine learning model is deployed as a web application to provide user interface for input and output.

1. Theoretical Analysis
   1. Block Diagram



* 1. Software Designing

The online services used in designing this web application are IBM Cloud, IBM Watson Studio and Node-RED. The steps followed to complete the project are as mentioned in the above block diagram.

The dataset is a collection of the value of life expectancy in different countries and the factors (features) required to calculate it. The dataset needs to be refined to remove any discrepancies present in the data such as empty values or incorrect data. The factors which have been used to calculate life expectancy from the dataset are as follows: -

1. Adult Mortality – Adult Mortality Rates of both sexes (probability of dying between 15 and 60 years per 1000 population)
2. Under-five deaths – Number of under-five deaths per 1000 population
3. Infant deaths – Number of Infant Deaths per 1000 population
4. Hepatitis-B – Hepatitis-B (HepB) immunization coverage among 1-year-olds (%)
5. Measles – Measles, number of reported cases per 1000 population
6. Diphtheria – Diphtheria tetanus toxoid and pertussis (DTP3) immunization coverage among 1-year-olds (%)
7. HIV/AIDS – Deaths per 1000 live births HIV/AIDS (0-4 years)
8. Polio – Polio (Pol3) immunization coverage among 1-year-olds (%)
9. Population – Population of the country
10. BMI – Average Body Mass Index of entire population
11. Thinness 5-9 years – Prevalence of thinness among children for Age 5 to 9(%)
12. Thinness 10-19 years – Prevalence of thinness among children and adolescents for Age 10 to 19 (%)
13. Schooling – Number of years of Schooling (years)
14. Alcohol – Alcohol, recorded per capita (15+) consumption (in litres of pure alcohol)
15. GDP – Gross Domestic Product per capita (in USD)
16. Percentage Expenditure – Expenditure on health as a percentage of Gross Domestic Product per capita (%)
17. Total Expenditure – General government expenditure on health as a percentage of total government expenditure (%)
18. Income Composition of Resources – Human Development Index in terms of income composition of resources (index ranging from 0 to 1)

A new project is created in the IBM Watson Studio, in which the Jupyter Notebook is created. A Watson Machine Learning Service is added to the project. The dataset is added in the project. This dataset is stored in the cloud. Jupyter Notebook is used to write different code and text cells to effectively run parts or the whole code while being able to add any additional descriptive text. Use Jupyter Notebook environment to write and test the code. The code written includes the following steps in order – importing required python libraries, loading dataset, exploring and visualising the dataset, splitting dataset into training and test data, scaling the data, building the model and saving the model.

Alternatively, for using AutoAI Modelling, a new AutoAI Experiment should be added to project in Watson Studio. In the experiment, add the dataset and select the required column to be predicted by the model. The AutoAI experiment is able to automatically do all required steps and gives us different models to select from. The selected model can then be saved.

The model saved can be seen in Assets tab of IBM Watson Studio. The model can be deployed from there. The deployment gives the scoring endpoint, which is later integrated with an UI in Node-RED.

Node-RED is a flow-based development tool for visual programming developed originally by IBM for wiring together hardware devices, APIs and online services as part of the Internet of Things. Node-RED provides a web browser-based flow editor, which can be used to create JavaScript functions. In Node-RED, an UI is created in order to get required input values. These values are passed as json payload to the model, connected using the scoring endpoint of its deployment. The predicted value is got back in json format which is then later displayed on the screen.

1. Experimental Investigations

Testing the model on untrained data suggests an amount of error in the predicted value, when compared to the actual life expectancy for the example.

A multiple linear regression algorithm implemented for the machine learning model in Jupyter Notebook gave the following observations: -

**Mean squared error: 15.66**

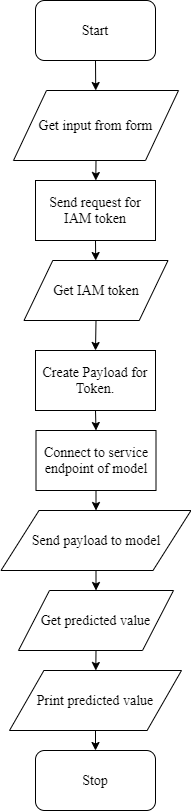
**R2 score: 0.83**

An extra tree regressor algorithm for the machine learning model implemented in the AutoAI experiment gave the following observations: -

**Mean squared error: 4.057**

**R2 score: 0.956**

1. Flowchart



1. Results

Upon using the example of the country ‘India’ and year ‘2006’, the following values were found as output for life expectancy given by the two models,

1. Actual value = 64.8 years
2. Predicted value by AutoAI generated model = 64.94 years

Difference from actual value = +0.14 years

1. Predicted value by Jupyter Notebook model = 68.33 years

Difference from actual value = +3.53 years

1. Advantages and Disadvantages
   1. Advantages
2. The project allows any persons or organisation to be able to predict the value of life expectancy in a country with the help of a few important factor values.
3. Life expectancy is a value very useful for the government, organisations, etc.
4. Knowing the life expectancy in their country will make people conscious about their health and other factors.
   1. Disadvantages
5. Other important factors, like deaths due to natural and man-made disasters, which may affect the average life expectancy in a country are not taken into consideration.
6. The value of life expectancy predicted is the average for a country and not for an individual.
7. The life expectancy of individuals may vary according to the personal conditions of the individual.
8. Applications
9. Life expectancy is one of the factors in measuring the Human Development Index (HDI) of each nation along with adult literacy, education, and standard of living.
10. Life expectancy is used in pricing and underwriting life insurance and insurance products like annuities, as well as in retirement and pension planning.
11. Conclusion

The project is able to predict and give us an approximate value of life expectancy, based on the input values of the factors taken into consideration. This value of life expectancy is an average for a country, and it doesn’t apply to every individual.

1. Future Scope
2. The project can be expanded to consider other factors which can be added to the current dataset and the model be retrained on the new generated dataset.
3. Also, the model could be retrained on a cleaner dataset as collected from different sources; instead of filling empty spaces with the mean values.
4. Bibliography

<https://www.allbusinesstemplates.com/download/?filecode=2KBA4&lang=en&iuid=9f9faa69-9fab-40ee-8457-ea0e5df8c8de>

<https://developer.ibm.com/tutorials/how-to-create-a-node-red-starter-application/>

<https://nodered.org/>

<https://developer.ibm.com/technologies/machine-learning/series/learning-path-machine-learning-for-developers/>

<https://bookdown.org/caoying4work/watsonstudio-workshop/jn.html>

<https://www.kaggle.com/kumarajarshi/life-expectancy-who>

<https://bookdown.org/caoying4work/watsonstudio-workshop/jn.html#deploy-model-as-web-service>

Appendix

<https://github.com/SmartPracticeschool/llSPS-INT-1909-Predicting-Life-Expectancy-using-Machine-Learning>