Internship Report

On

**Predicting Life Expectancy Using Machine Learning**

****

**Organization:**

Smartbridge Trading Solutions Pvt Ltd,

A 12/13 , B & B Genesis Building, Sector 16, Noida, Uttar Pradesh 201301

**Report Submitted by:**

Tushar Kumar

Enrollment no.: 41910402717

B.Tech. CSE 3rd year

Amity School of Engineering and Technology

List of Figures

1. Block Diagram

2. Correlation heatmap

3. Level 0 DFD

4. Level 1 DFD

5. User Interface

6. Node-RED flow

Table of Contents

*List of figures*

1. [Introduction](#_Toc42808299) 1
   1. [Overview](#_Toc42808300) 1
   2. [Purpose](#_Toc42808301) 2
2. [Literature Survey](#_Toc42808302) 3
   1. [Existing Problem](#_Toc42808303) 3
   2. [Proposed Solution 3](#_Toc42808304)
3. [Theoretical Analysis 4](#_Toc42808305)
   1. [Block Diagram 4](#_Toc42808306)
   2. [Software Design 4](#_Toc42808307)
4. [Experimental Investigations 5](#_Toc42808308)
   1. [Factors affecting life expectancy 5](#_Toc42808309)
   2. [Correlation among all attributes 6](#_Toc42808310)
   3. [Choosing regression model 7](#_Toc42808311)
5. [Flow Diagram 8](#_Toc42808312)
6. [Result 9](#_Toc42808313)
   1. [User Interface: 9](#_Toc42808314)
7. [Advantages & Disadvantages 10](#_Toc42808315)
   1. [Advantages 10](#_Toc42808316)
   2. [Disadvantages 10](#_Toc42808317)
8. [Applications 12](#_Toc42808318)
9. [Conclusion 13](#_Toc42808319)
10. [Future Scope 13](#_Toc42808320)
11. [Bibliography 14](#_Toc42808321)

[Appendix](#_Toc42808322)

A.[Source Code:](#_Toc42808323)

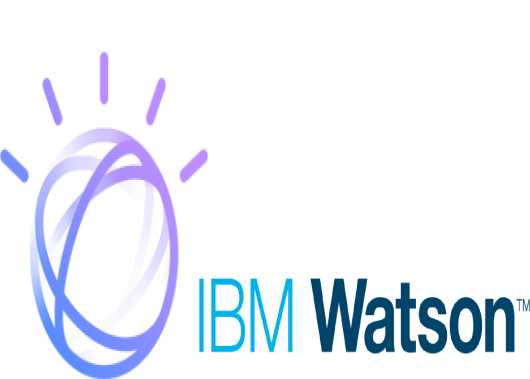
B.[Node-RED flow](#_Toc42808324)

1. Introduction
   1. Overview

This project is focused on predicting “Life Expectancy Rate” of a country for a given set of factors.

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 project 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.

For prediction purpose, the program uses Machine Learning Regression Algorithm. The user interface of the project was developed using Node-Red. The project also uses other IBM services like IBM Cloud, IBM Watson Studio, IBM Machine Learning Service and Cloudant.

* 1. Purpose

Life expectancy is one of the most important factors in end-of-life decision making. Good prognostication for example helps to determine the course of treatment and helps to anticipate the procurement of health care services and facilities, or more broadly: facilitates Advance Care Planning. Advance Care Planning improves the quality of the final phase of life by stimulating doctors to explore the preferences for end-of-life care with their patients, and people close to the patients.

1. Literature Survey
   1. Existing Problem

Although there have been lot of studies undertaken in the past on factors affecting life expectancy considering demographic variables, income composition and mortality rates. It was found that effect of immunization and human development index was not taken into account in the past. Also, some of the past research was done considering multiple linear regression based on data set of one year for all the countries. Hence, this gives motivation to resolve both the factors stated previously by formulating a regression model based on mixed effects model and multiple linear regression while considering data from a period of 2000 to 2015 for all the countries. Important immunization like Hepatitis B, Polio and Diphtheria will also be considered.

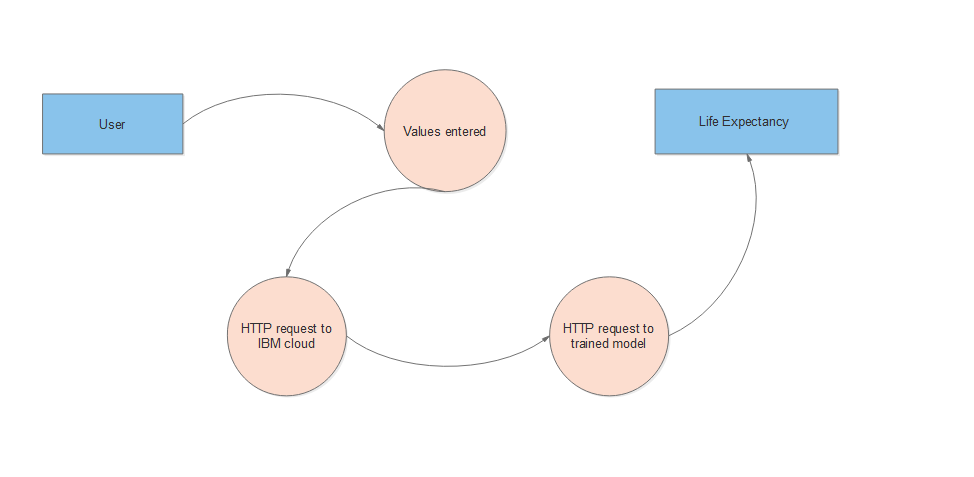
* 1. Proposed Solution

The data-set related to life expectancy, health factors for 193 countries has been collected from WHO data repository website and its corresponding economic data was collected from the United Nations website. Among all categories of health-related factors only those critical factors were chosen which are more representative. It has been observed that in the past 15 years, there has been a huge development in health sector resulting in improvement of human mortality rates especially in the developing nations in comparison to the past 30 years. Therefore, in this project we have considered data from year 2000-2015 for 193 countries for further analysis. The individual data files have been merged together into a single data-set.

The project uses immunization factors, mortality factors, economic factors, social factors and other health related factors to predict life expectancy of a country for a given year using a machine learning model.

Since the observations in this dataset are based on different countries, it will be easier for a country to determine the predicting factor which is contributing to lower value of life expectancy. This will help in suggesting a country, which area should be given importance in order to efficiently improve the life expectancy of its population.

1. Theoretical Analysis
   1. Block Diagram



*Fig. 1:Block Diagram*

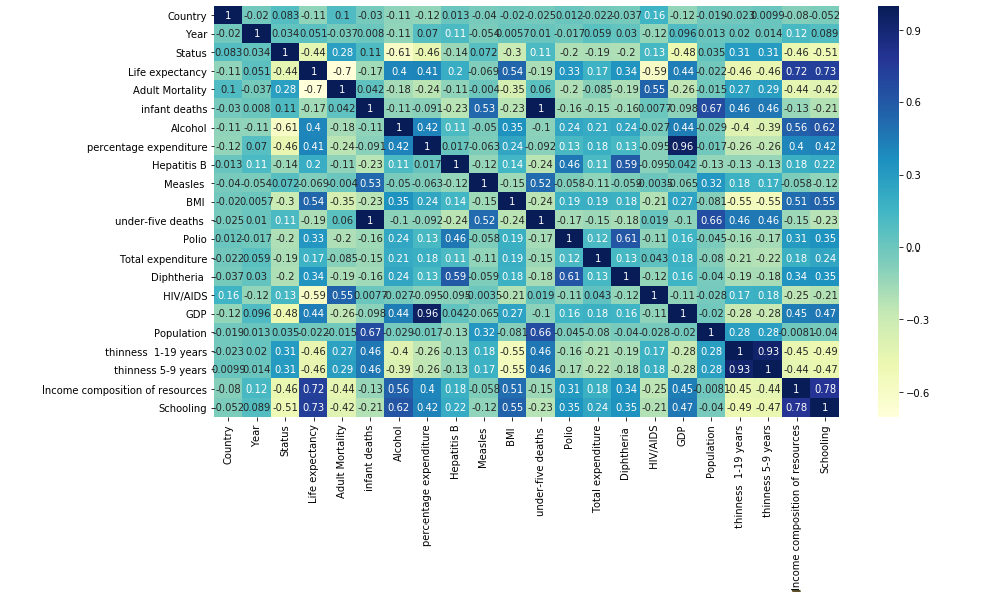
* 1. Software Design

Input is taken from the user using a “Form” element in Node-Red. Then, an HTTP request is made to the IBM cloud that further makes an HTTP request to the deployed model using model’s instance id. After verification of id, the model sends an HTTP response which is finally parsed by the Node-Red application and the result is displayed on the user screen.

1. Experimental Investigations
   1. Factors affecting life expectancy

Following factors are taken into account for predicting the life expectancy of a country.

1. Status: Developed or Developing status of the country.
2. Adult mortality: Adult Mortality Rates of both sexes (probability of dying between 15 and 60 years per 1000 population).
3. Infant deaths: Number of Infant Deaths per 1000 population.
4. Alcohol: Alcohol, recorded per capita (15+) consumption (in litres of pure alcohol).
5. Percentage Expenditure: Expenditure on health as a percentage of Gross Domestic Product per capita (%).
6. Hepatitis B: Hepatitis B (HepB) immunization coverage among 1-year-olds (%).
7. Measles: Measles - number of reported cases per 1000 population.
8. BMI: Average Body Mass Index of entire population.
9. Under-five deaths: Number of under-five deaths per 1000 population.
10. Polio: Polio (Pol3) immunization coverage among 1-year-olds (%).
11. Total expenditure: General government expenditure on health as a percentage of total government expenditure (%).
12. Diphtheria: Diphtheria tetanus toxoid and pertussis (DTP3) immunization coverage among 1-year-olds (%).
13. HIV/AIDS: Deaths per 1 000 live births HIV/AIDS (0-4 years).
14. GDP: Gross Domestic Product per capita (in USD).
15. Population: Population of the country.
16. Thinness 1-19 years: Prevalence of thinness among children and adolescents for Age 10 to 19(% ).
17. Thinness 5-9 years: Prevalence of thinness among children for Age 5 to 9(%).
18. Income composition of resources: Human Development Index in terms of income composition of resources (index ranging from 0 to 1).
19. Schooling: Number of years of schooling.
    1. Correlation among all attributes



*Fig. 2: Correlation heatmap*

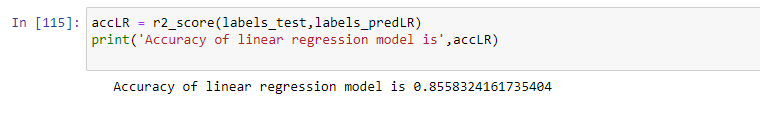
From the heatmap above, it is pretty evident that “Schooling” has the highest level of correlation with “Life expectancy” among all the attributes. “Adult mortality” also has high level of negative correlation with “Life expectancy”.

* 1. Choosing regression model

Two regression models, viz., Linear Regression model and Random Forest Regression model were used to predict life expectancy. The model with maximum accuracy was chosen while deploying the project on IBM cloud.

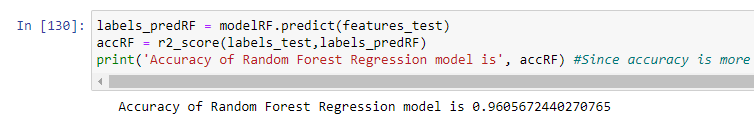
For accuracy of a model, R squared value is calculated. R-squared (R2) is a statistical measure that represents the proportion of the variance for a dependent variable that's explained by an independent variable or variables in a regression model. R-squared explains to what extent the variance of one variable explains the variance of the second variable.

**Linear Regression Model:**

****

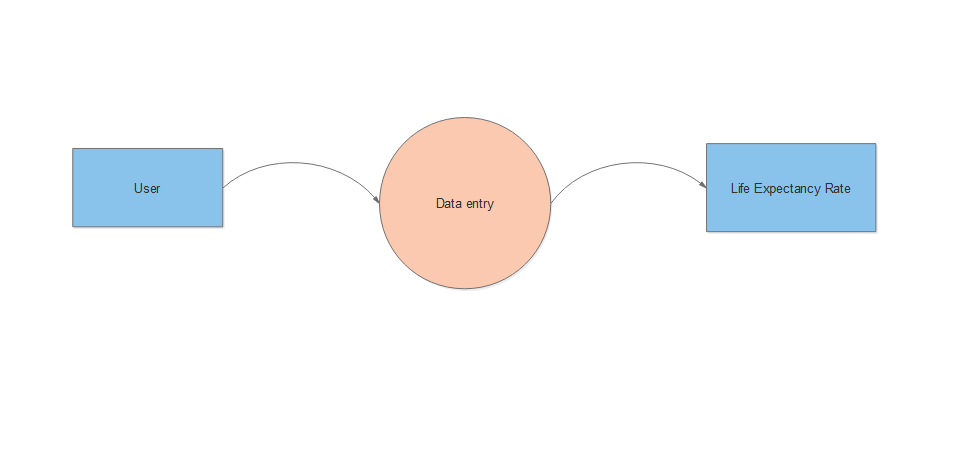
Accuracy for linear regression model comes out to be 0.8558.

**Random Forest Regression Model:**

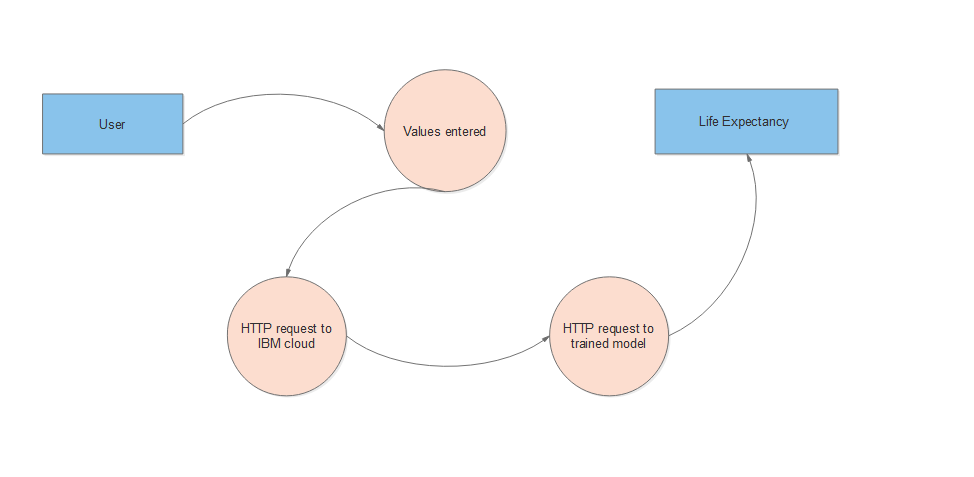
****

Accuracy for random forest regression model comes out to be 0.9606.

1. Flow Diagram



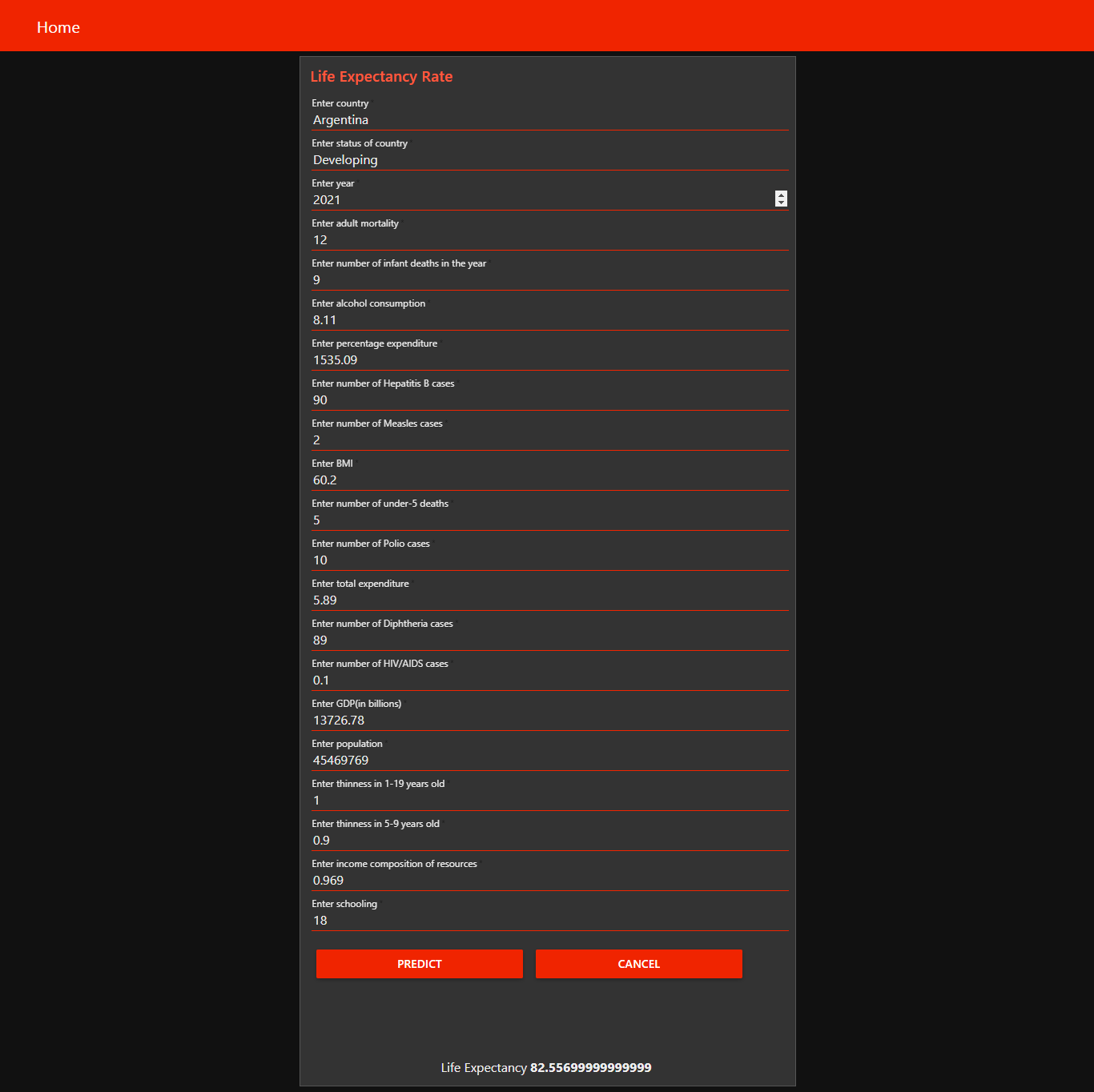
*Fig. 3: Level 0 DFD*



*Fig. 4: Level 1 DFD*

1. Result

**User Interface:**



*Fig. 5: User Interface*

1. Advantages & Disadvantages
   1. Advantages
2. The observations in the dataset used are based on different countries, it will be easier for a country to determine the predicting factor which is contributing to lower value of life expectancy. This will help in suggesting a country, which area should be given importance in order to efficiently improve the life expectancy of its population.
3. Some of the past research was done considering multiple linear regression based on data set of one year for all the countries. But the dataset used for training the model contained data of past 15 years to give a fairly better prediction.
4. The application is easy and simple to use.
5. The machine learning algorithm used in the project is Random Forest regression which is based on the bagging algorithm and uses Ensemble Learning technique. It creates as many trees on the subset of the data and combines the output of all the trees. In this way it reduces overfitting problem in decision trees and also reduces the variance and therefore improves the accuracy.
6. Random Forest algorithm is very stable. Even if a new data point is introduced in the dataset, the overall algorithm is not affected much since the new data may impact one tree, but it is very hard for it to impact all the trees.
7. Random Forest is comparatively less impacted by noise.
   1. Disadvantages
8. As the model is deployed on cloud, so one requires good internet connection to use the application.
9. The model used is Random Forest regression and Random Forest creates a lot of trees (unlike only one tree in case of decision tree) and combines their outputs. By default, it creates 100 trees in Python sklearn library. To do so, this algorithm requires much more computational power and resources.
10. Random Forest require much more time to train as compared to decision trees as it generates a lot of trees (instead of one tree in case of decision tree) and makes decision on the majority of votes.
11. The Node-Red application needs to make HTTP request to IBM cloud and then another HTTP request to the model before providing the prediction. That makes the application a bit slow.
12. Applications
13. The project can be used as a basis to develop personalized health applications.
14. As the model uses a wide range of features for prediction, it will be easier for a country to determine the predicting factor which is contributing to lower value of life expectancy. This will help in suggesting a country, which area should be given importance in order to efficiently improve the life expectancy of its population.
15. The governments can plan and develop their health infrastructures by keeping the most correlated factors in mind.
16. The project can help governments to keep track of their countries’ health status so they can plan for the future accordingly.
17. Conclusion

The project makes use of electronic medical records for predicting life expectancy. The potential use of project is not limited to health care in practice, but could also be useful in other clinical applications such as clinical trials. In clinical trials, outcomes often depend on prognostic factors. Automatic processing of medical records would enable quick and systematic stratification of patients based on their prognoses, which could be used to further reduce biases. The project makes a good use of machine learning in predicting life expectancy of a country that can help respective government in making policies that will serve for the benefit of the nation and entire humankind.

1. Future Scope

Some future improvements can be made on current project. They are as follows:

1. As more data comes, that can be fed to the model for more accurate predictions.
2. Currently, the project is just a web application. It can be developed to support other platforms like Android, IOS and Windows Mobile.
3. Other regression models can also be used for prediction and later the best among them should be chosen.
4. User interface can be modified for various countries according to their regional languages.

**11. Bibliography**

[1] IBM Cloud setup [Online]. Available: <https://www.ibm.com/cloud/get-started> .

[2] IBM Developer, “Node-RED starter tutorial” [Online]. Available: <https://developer.ibm.com/tutorials/how-to-create-a-node-red-starter-application/> .

[3] “Node-RED labs on the use of the Watson Developer Cloud services - watson-developer-cloud/node-red-labs.” [Online]. Available: <https://github.com/watson-developer-cloud/node-red-labs> .

[4] “Infuse AI into your applications with Watson AI to make more accurate predictions”. [Online]. Available: <https://www.ibm.com/watson/products-services> .

[5] IBM Watson, “Intro to IBM Watson”, 2018 [Online]. Available: <https://www.youtube.com/watch?v=W3iPbFTAAds&feature=youtu.be> .

[6] “Get an understanding of the principles of machine learning.” [Online]. Available: <https://developer.ibm.com/technologies/machine-learning/series/learning-path-machine-learning-for-developers/> .

[7] IBM Developer, “IBM Watson Machine Learning: Get Started in IBM Cloud”, 2020 [Online]. Available: <https://www.youtube.com/watch?v=NmdjtezQMSM> .

[8] Watson Studio Workshop, “Chapter 4 Build and Deploy models in Jupyter Notebooks” [Online]. Available: <https://bookdown.org/caoying4work/watsonstudio-workshop/jn.html> .

[9] Kumar Rajarshi, “Life Expectancy (WHO) Statistical Analysis on factors influencing Life Expectancy”, 2018. [Online]. Available: <https://www.kaggle.com/kumarajarshi/life-expectancy-who>

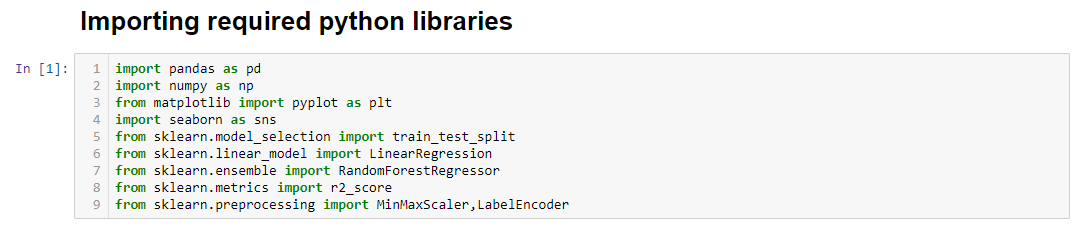
[10] IBM Developer, “IBM Watson: Sign up for Watson Studio and Watson Knowledge Catalog”, 2019. [Online]. Available: <https://www.youtube.com/watch?v=DBRGlAHdj48&list=PLzpeuWUENMK2PYtasCaKK4bZjaYzhW23L> .

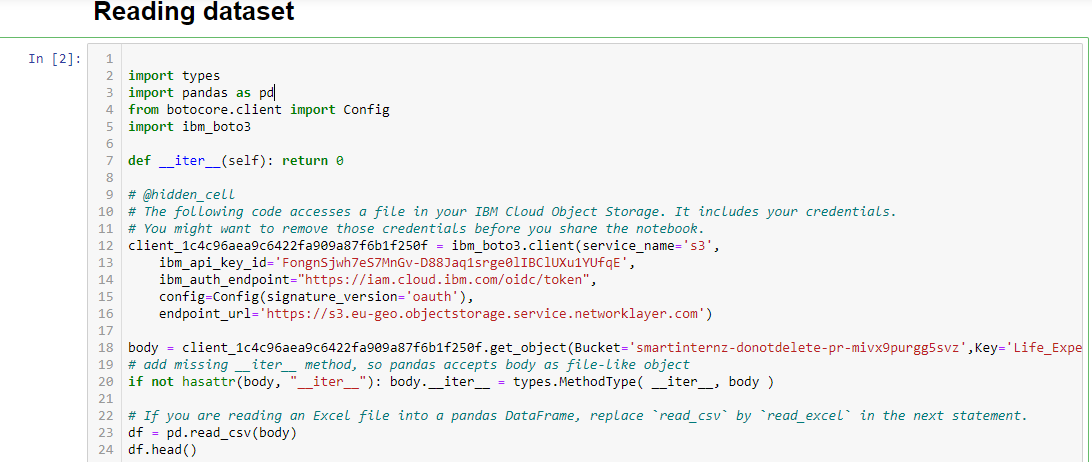
[11] IBM Developer, “IBM Watson Studio: Create a project”, 2019. [Online]. Available: <https://www.youtube.com/watch?v=-CUi8GezG1I&list=PLzpeuWUENMK2PYtasCaKK4bZjaYzhW23L&index=2>

[12] IBM Developer, “IBM Watson Studio: Jupyter notebook basics”, 2019 [Online]. Available: <https://www.youtube.com/watch?v=Jtej3Y6uUng>

Appendix

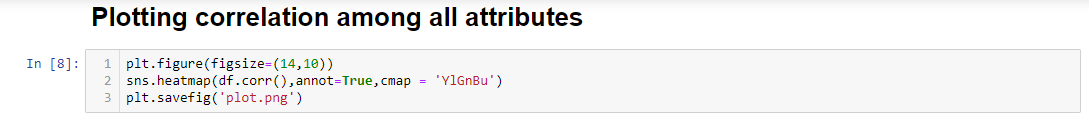
1. Source Code:

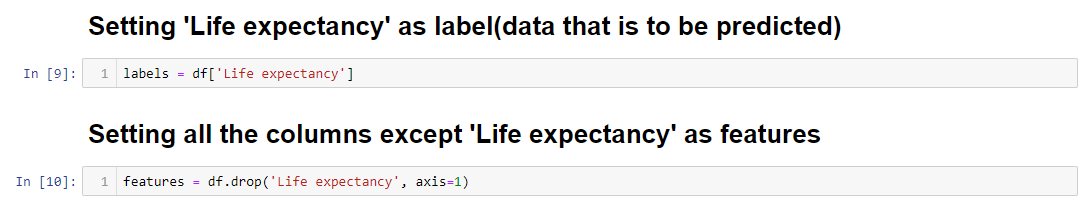


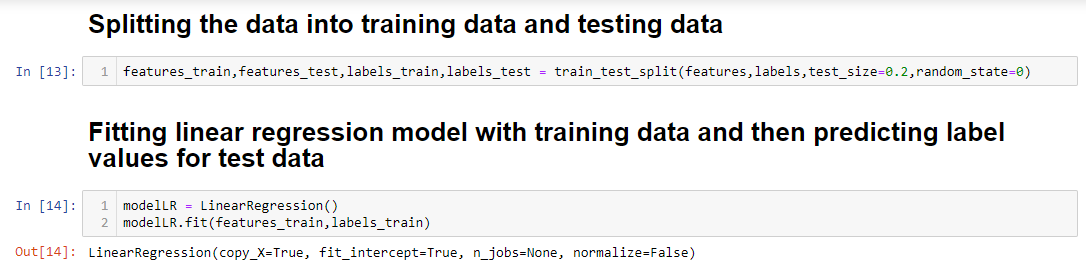


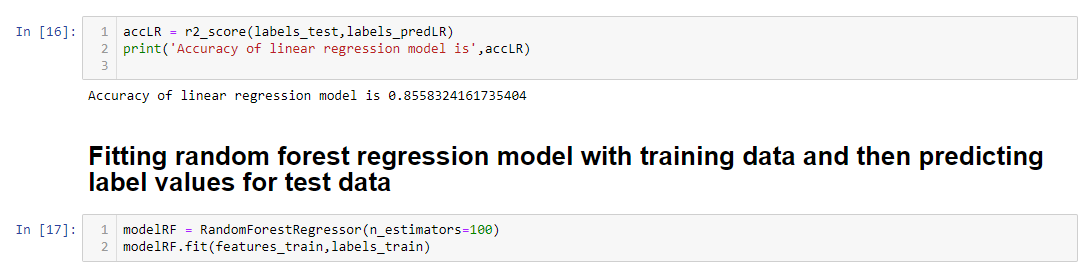


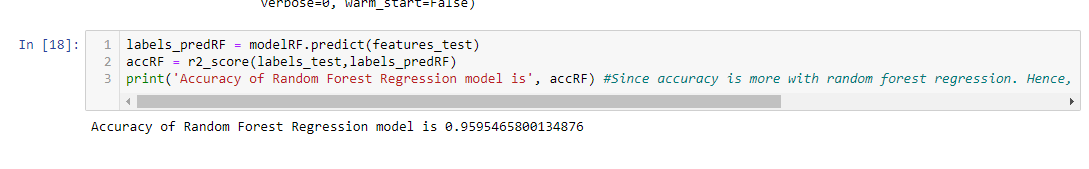


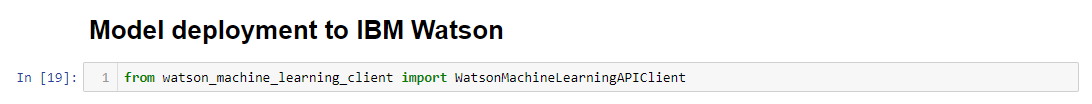








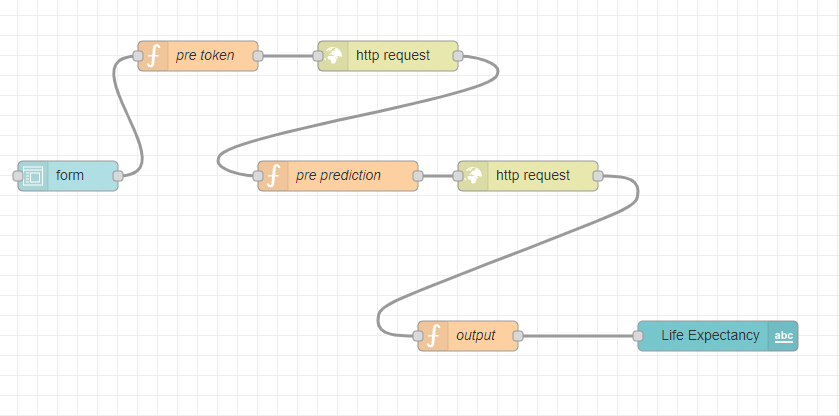








1. Node-RED flow



*Fig. 6: Node-RED flow*