



PROJECT REPORT :

PREDICTING LIFE

EXPECTANCY USING MACHINE LEARNING

|  |
| --- |
| **PROJECT REPORT ON**  **“PREDICTING LIFE EXPECTANcY USINg MACHINe LEARNiNg”** |

**SUBMITTED BY:**

**Name:** Shevya Solanki

**BRANCH:** Computer Science & Engineering

**PROJECT ID:**  SPS\_PRO\_215

**Category:** Machine Learning

SBID: SB20200035447

**uNder The GUidaNce of:**

**SmartiNterNz rsip-2020**



|  |
| --- |
| **ACKNOwledgemeNt** |

* The internship opportunity I had with **SmartBridge RSIP-2020** was a great chance for learning and professional development. Therefore, I consider myself as a very lucky individual as I was provided with an opportunity to be a part of it. I am also extremely grateful for having a supportive mentors that led me through this internship period and helped me a lot to achieve my goal. I have gained the knowledge and the actual practical application of that in a real world problem, that would definitely going to help in my future career.
* I perceive as this opportunity as a big milestone in my career development & I really consider myself lucky for being the part of this Initiative. I will strive to use gained skills and knowledge in the best possible way, and I will continue to work on their improvement, in order to attain desired career objectives. Hope to continue cooperation with all of you in the future.

Thank You.

Shevya Solanki

|  |
| --- |
| **ABSTRACT** |

* The Global Health Observatory (GHO) data repository under World Health Organization (WHO) keeps track of the health status as well as many other related factors for all countries .The datasets are made available to public for the purpose of health data analysis. The dataset related to life expectancy, health factors for 193 countries has been collected from the same WHO data repository website and its corresponding economic data was collected from United Nation 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 dataset. As the datasets were from WHO, we found no evident errors. The final merged (final dataset) consists of 22 Columns and 2938 rows which meant 20 predicting variables. 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.

|  |
| --- |
| **CONTENT Table** |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **INT 1. INTRODUCTION** | | | | |  | | |
|  |  | | 1.1 Overview | | | |  | | |
|  |  | | 1.2 Purpose | | | |  | | |
| **LITER 2. LITRATURE SURVEY** | | | | |  | | |
|  |  | | 2.1 Existing problem | | | | | | |
|  |  | | 2.2 Proposed solution | | | | | | |
| **3.THEORITICAL ANALYSIS** | | | | |  | | |
|  |  | | 3.1 Block diagram | | | |  | | |
|  |  | | 3.2 Hardware / Software designing | | | | | | |
|  | | | | | | | |
| **4. FLOWCHART** | | | | |  | | |
| **5. RESULT** | | | |  |  | | |
| **6. ADVANTAGES & DISADVANTAGES** | | | | | | | | |
| **7. APPLICATIONS** | | | | | |  | | |
| **8. CONCLUSION** | | | | | |  | | |
| **9. FUTURE SCOPE** | | | | | | **`** | | |
| **10. REFERENCE** | | | | | |  | | |
|  |  |  | | | | |  | | |

|  |
| --- |
| **1. Introduction** |

**1.1 OVERVIEW**

* ***Machine learning*** is a sub-domain of computer science which evolved from the study of pattern recognition in data, and also from the computational learning theory inartificial intelligence. It is the first-class ticket to most interesting careers in data analytics today. As data sources proliferate along with the computing power to process them, going straight to the data is one of the most straightforward ways to quickly gain insights and make predictions. Machine Learning can be thought of as the study of a list of sub-problems, viz: **decision making, clustering, classification, forecasting, deep-learning, inductive logic programming, support vector machines, reinforcement learning, similarity and metric learning, genetic algorithms, sparse dictionary learning,** etc.
* **I**n case of **Life-Expectancy**  , 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 affect 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. In a nutshell, this study will focus on immunization factors, mortality factors, economic factors, social factors and other health related factors as well. Since the observations 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.
* The project tries to create a model based on data provided by the World Health Organization (WHO) to evaluate the life expectancy for different countries in years. The data offers a timeframe from 2000 to 2015. The data originates from here: https://www.kaggle.com/kumarajarshi/life-expectancy-who/data . The output algorithms have been used to test if they can maintain their accuracy in predicting the life expectancy for data they haven’t been trained. Four algorithms have been used:
* Linear Regression  
  Ridge Regression  
  Lasso Regression  
  Elastic Net Regression  
  Decision Tree Regression  
  Random Forest Regression

**1.2 PURPOSE OF PROJECT**

* **Life Expectancy Prediction** project relies on accuracy of data. The Global Health Observatory (GHO) data repository under World Health Organization (WHO) keeps track of the health status as well as many other related factors for all countries. The data-sets are made available to public for the purpose of health data analysis. The data-set related to life expectancy, health factors for 193 countries has been collected from the same WHO data repository website and its corresponding economic data was collected from United Nation 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. On initial visual inspection of the data showed some missing values. As the data-sets were from WHO, we found no evident errors.. The final merged file (final dataset) consists of 22 Columns and 2938 rows which meant 20 predicting variables. All predicting variables was then divided into several broad categories:​Immunization related factors, Mortality factors, Economical factors and Social factors.

|  |
| --- |
| **2. LITERATURE SURVEY** |

**2.1 EXISTINg PROBLEM**

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. Physicians, however, tend to overestimate life expectancy, and miss the window of opportunity to initiate Advance Care Planning. This research tests the potential of using machine learning and natural language processing techniques for predicting life expectancy from electronic medical records.

Mortality trends in high-income countries between 1900 and 1950 showed a clear age-pattern shift. Mortality at young ages and from infectious conditions was rapidly receding, whereas mortality at older ages and from chronic conditions began to dominate. By the 1960s, major medical improvements in cardiovascular survival led to an increasing prevalence of heart disease at older ages. These developments focused attention on the morbidity as well as the mortality of the increasing older population. By the late 1970s and early 1980s, researchers had devised theoretical frameworks as well as markers of morbidity for assessing healthy aging. We briefly review three of these frameworks—failure of success, compression of morbidity, and dynamic equilibrium—that have guided significant amounts of research on healthy life expectancy in the last decades.

We approached the task of predicting life expectancy as a supervised machine learning task. We trained and tested a long short-term memory recurrent neural network on the medical records of deceased patients. We developed the model with a ten-fold cross-validation procedure, and evaluated its performance on a held-out set of test data. We compared the performance of a model which does not use text features (baseline model) to the performance of a model which uses features extracted from the free texts of the medical records (keyword model), and to doctors’ performance on a similar task as

Prognostication of life expectancy is difficult for humans. Research shows that machine learning and natural language processing techniques offer a feasible and promising approach to predicting life expectancy. The research has potential for real-life applications, such as supporting timely recognition of the right moment to start Advance Care Planning.

* 1. **Proposed solution**
* Four algorithms have been used:

Linear Regression  
Ridge Regression  
Lasso Regression  
Decision Tree Regression  
Random Forest Regression

The data-set aims to answer the following key questions:

1. Does various predicting factors which has been chosen initially really affect the Life expectancy? What are the predicting variables actually affecting the life expectancy?

2. Should a country having a lower life expectancy value (<65) increase its healthcare expenditure in order to improve its average lifespan?

3. How does Infant and Adult mortality rates affect life expectancy?

4. Does Life Expectancy has positive or negative correlation with eating habits, lifestyle, exercise, smoking, drinking alcohol etc.

5. What is the impact of schooling on the lifespan of humans?

6. Does Life Expectancy have positive or negative relationship with drinking alcohol?

7. Do densely populated countries tend to have lower life expectancy?

8. What is the impact of Immunization coverage on life Expectancy?

|  |
| --- |
| **3. THEORETICAL Analysis** |

**3. 1 PROJECT REQUIREMENT :**

* + 1. **FUNCTIONAL REQUIREMENT :**
* Predicting Life Expectancy rate of a country.

**3.1.2 TECHNICAL REQUIREMENTS :**

* Python, IBM Cloud, IBM Watson, Github , Node Red in IBM Watson, Jupyter Notebook Watson.

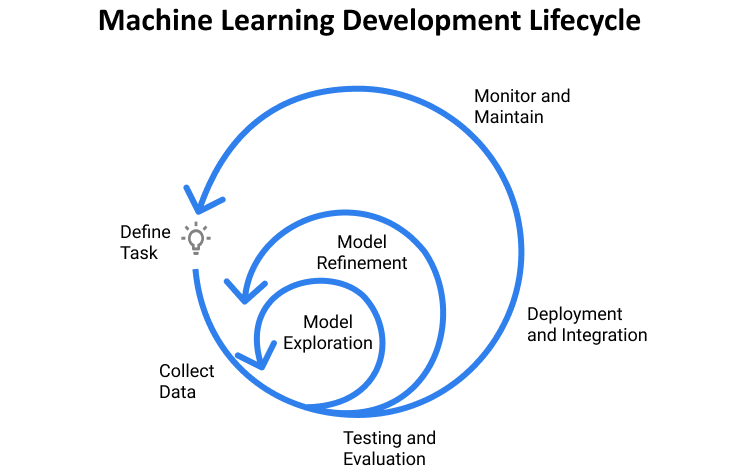
**3.2 HARDWARE  REQUIREMENT :**

*Processor -*i5 7th Gen

*Speed -*2GHz or more

*Hard Disc -*10 Gb or more

*3.*3 HARDWARE SOFTWARE DESIGN



1. [**Planning and project setup**](https://www.jeremyjordan.me/ml-projects-guide/#planning)

a. Define the task and scope out requirements.

b. Determine project feasibility.

c. Setting up project codebase.

**TOOLS**- IBM Watson Studio, Node-Red workflows, IBM cloud.

2. [**Data collection and labeling**](https://www.jeremyjordan.me/ml-projects-guide/#data)

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

a. Define ground truth (create labeling documentation)

b. Build data model

c. Validate quality of data

**TOOLS -**Excel SpreadSheets , CSV files.

3. [**Model exploration**](https://www.jeremyjordan.me/ml-projects-guide/#exploration)

a. Establish baselines for model performance.

b. Overfit simple model to training data.

c. Using various Libraries to explore data.

**TOOLS-**Matplotlib, Sklearn, Seaborn, Keras, Scikit-Learn.

4. [**Model refinement**](https://www.jeremyjordan.me/ml-projects-guide/#refinement)

a. Perform model-specific optimizations ( Hyper parameter tuning)

b. Iteratively debug model as complexity is added

5. [**Testing and evaluation**](https://www.jeremyjordan.me/ml-projects-guide/#testing)

a. Evaluate model on test distribution; understand differences between train and test set distributions (how is “data in the wild” different than what you trained on)

b. Revisit model evaluation metric; ensure that this metric drives

6. [**Model deployment**](https://www.jeremyjordan.me/ml-projects-guide/#deployment)

a. Expose model via a REST API, IBM Watson, Node-Red Kit.

b. Deploy new model to small subset of users to ensure everything goes smoothly, then roll out to all users

c. Monitor live data and model prediction distributions

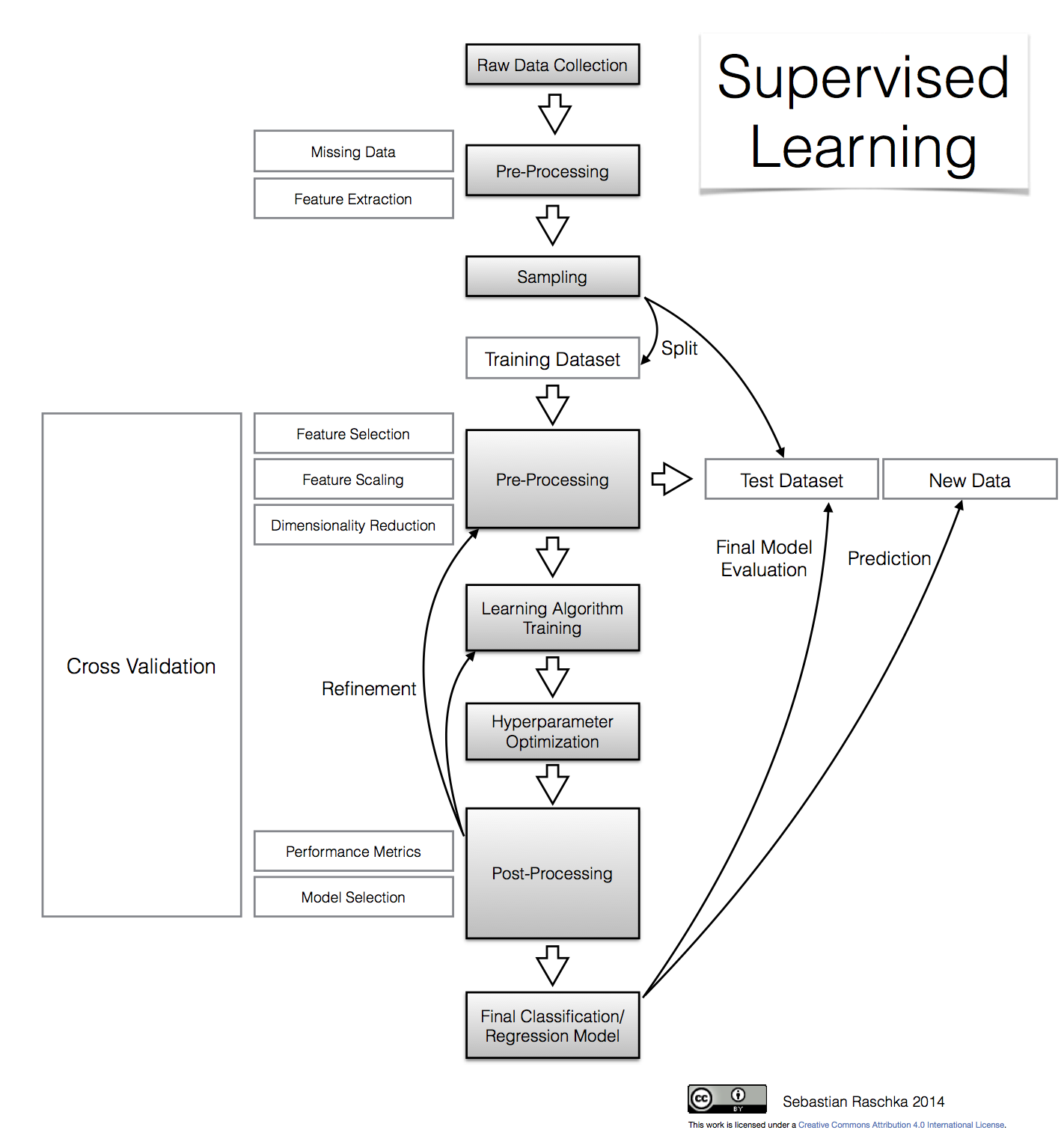
7. [**Ongoing model maintenance**](https://www.jeremyjordan.me/ml-projects-guide/#maintenance)

a. Understand that changes can affect the system in unexpected ways

b. Periodically retrain model to prevent model staleness

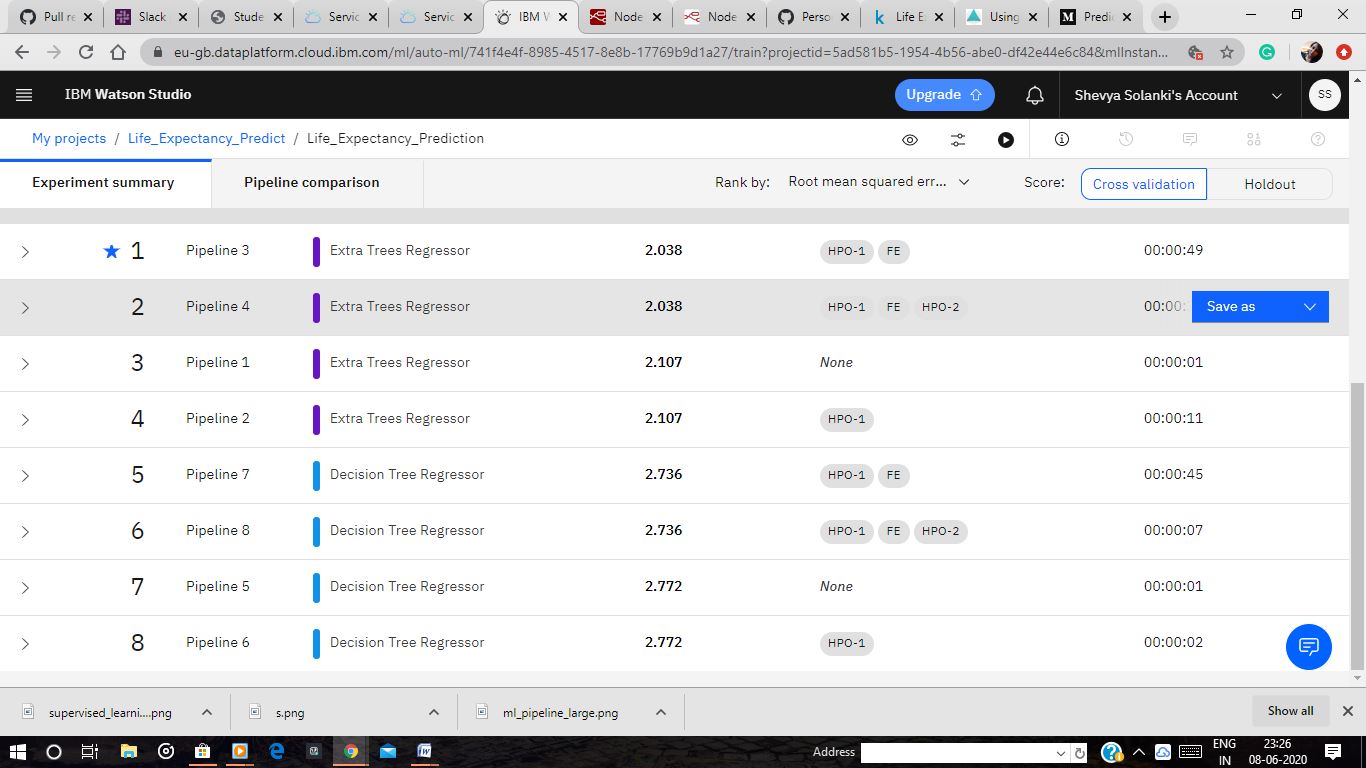
c. If there is a transfer in model ownership, educate the new team

|  |
| --- |
| **4. FLOWCHART** |

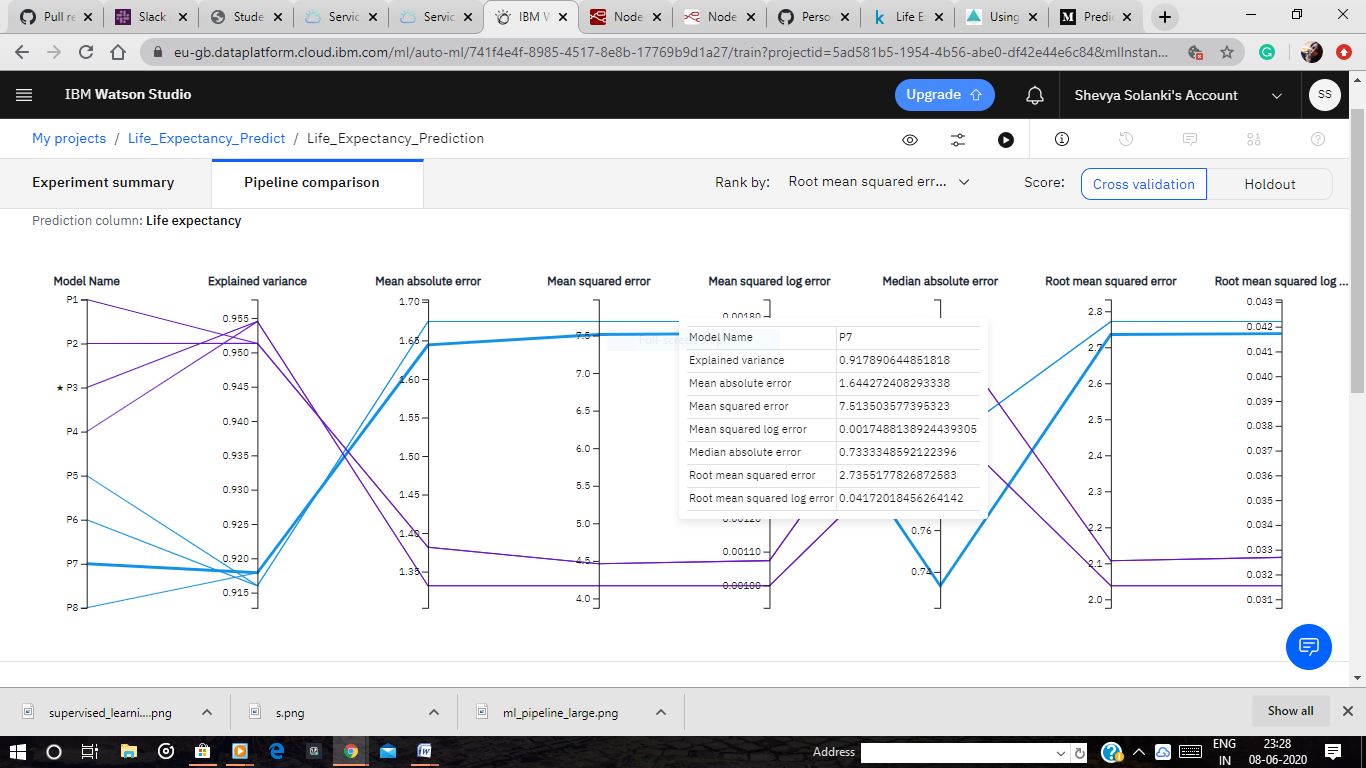
****

|  |
| --- |
| **5. Result** |

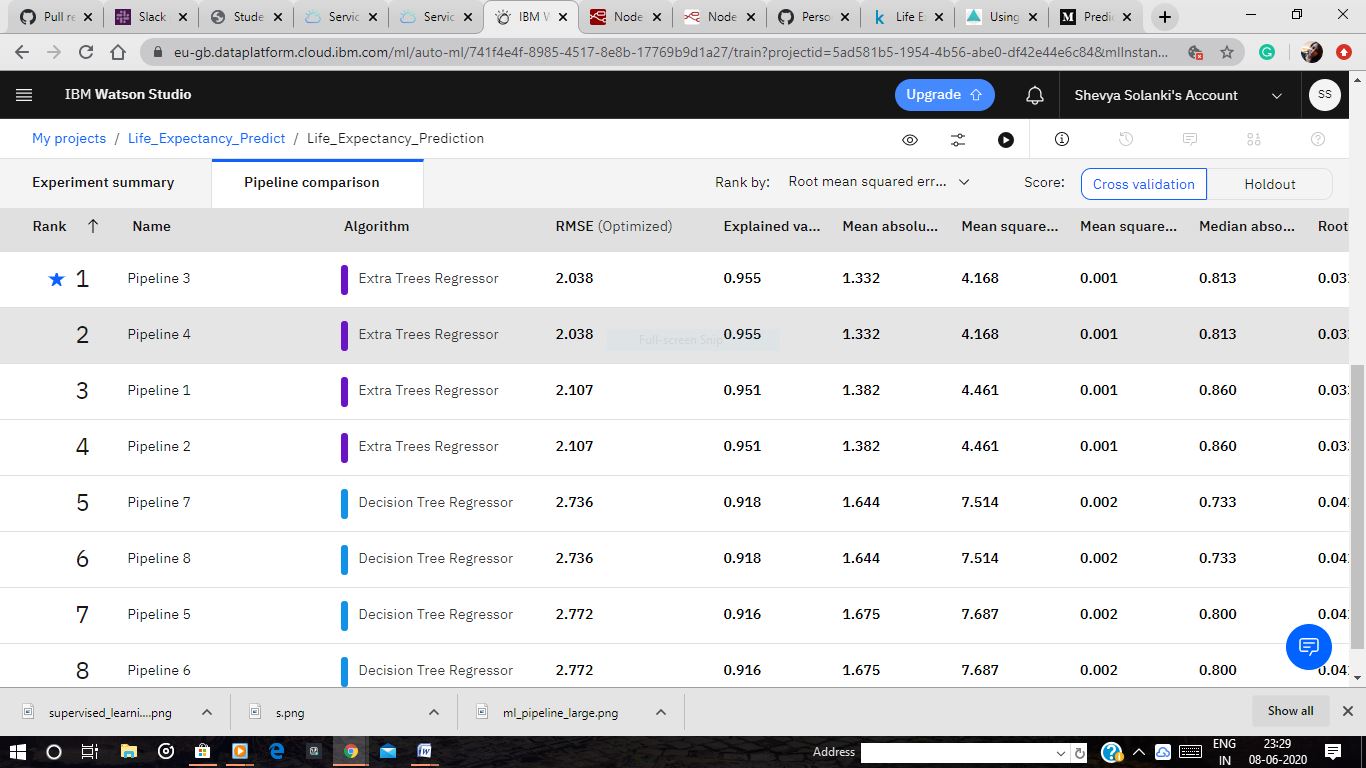
Below given is the Pipelines used on the WHO data set to predict the outcome of the Life- Expectancy data collected. Out of all models Extra Tree Regressor was the pipeline.

****

Below given is the comparison among different Algorithms used to reach out to conclusion.

****

Below is the predicted values of the following algorithm :

****

|  |
| --- |
| **6. Advantage & Disadvantage** |

**6.1 AdvaNtages**

* 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. Physicians, however, tend to overestimate life expectancy, and miss the window of opportunity to initiate Advance Care Planning. This research tests the potential of using machine learning and natural language processing techniques for predicting life expectancy from electronic medical records.
* Prognostication of life expectancy is difficult for humans. Research shows that machine learning and natural language processing techniques offer a feasible and promising approach to predicting life expectancy. The research has potential for real-life applications, such as supporting timely recognition of the right moment to start Advance Care Planning.
* Life expectancy plays an important role when decisions about the final phase of life need to be made. 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 (ACP) is the process during which patients make decisions about the health care they wish to receive in the future, in case the patient loses the capacity of making decisions or communicating about them . Successful ACP enhances the quality of life and death for palliative patients, by providing timely palliative care and documenting preferences regarding resuscitation and euthanasia, among other things

**6.2 DIADVANtages**

* The main disadvantage is that NO ONE can predict the future. No one knows when someone will die, who will get cancer or not, who will recover and who won't. A person who appears to have been sickly for many years can surprise everyone by outliving all of his peers. Likewise, a person known to be robustly healthy and in shape can succumb immediately to a stroke, heart attack, or other unexpected calamity.
* Insurance companies and others who make their living from using life expectancy as an indicator know they are basically gambling. They rely on statistics that may or may not, ultimately, turn out to be accurate for an individual. For example, they can say "Hypertension tends to be higher in African Americans," which appears to be a statistically accurate statement, in general. However, that doesn't always translate into accurate risk prediction for James, who is an African American male age 54 with high blood pressure, who turns out to live to 105 because he took his blood pressure medication and started practicing meditation and eating healthy to control his blood .Statistics work in generalities. Humans, however, do not.

|  |
| --- |
| **7. Applications** |

* A life table is a table which shows, for a person at each age, what the probability is that they die before their next birthday. From this starting point, a number of statistics can be derived and thus also included in the table is:
* The probability of surviving any particular year of age
* The remaining life expectancy for people at different ages
* The proportion of the original birth cohort still alive.

Life tables are usually constructed separately for men and for women because of their substantially different mortality rates.

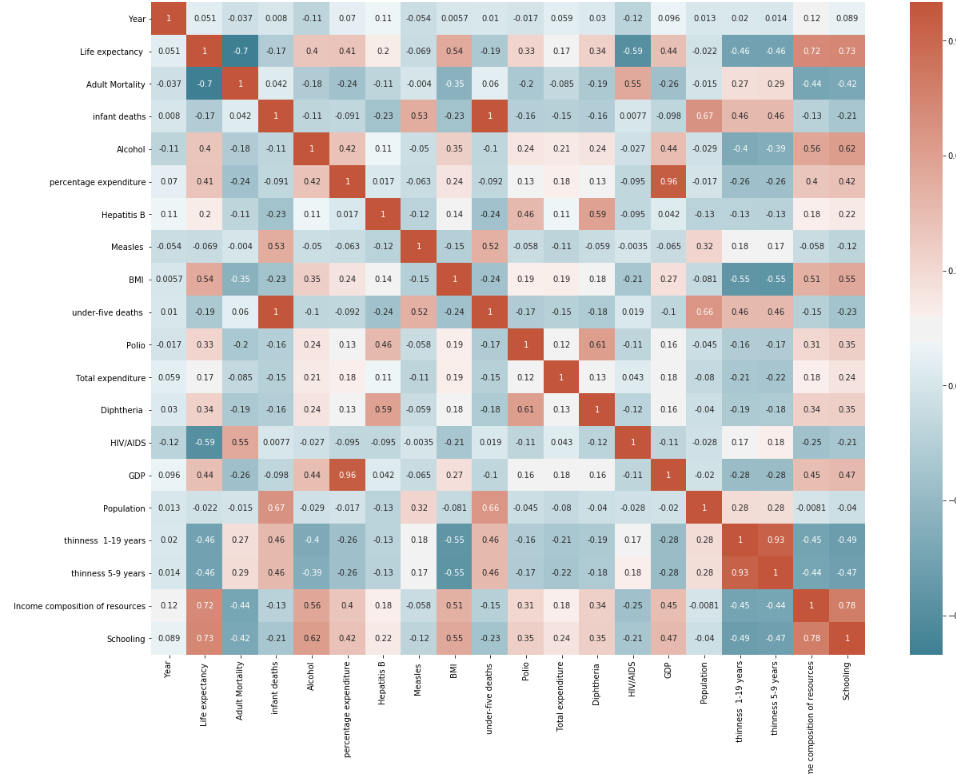
* Life tables are also used in biology. Other characteristics can also be used to distinguish different risk factors for life expectancy, such as smoking-status, occupation, socio-economic class, and others.  More complex analyses for assessing cancer survival, that involves comparisons between two populations or a population in two points in time can also be undertaken.

In addition to public health domains, life tables are also used by insurance companies and actuary departments.

When used in biology, age specific fertility rates are also included in the calculations. When data have not been available, such as in low income countries, life tables have been modelled using what data are available, usually childhood mortality data.

|  |
| --- |
| **8. Conclusion** |

* The project relies on the accuracy of data. The Global Health Observatory (GHO) data repository under the World Health Organization (WHO) keeps track of the health status as well as many other related factors for all countries the data-sets are made available to the public for the purpose of health data analysis. The data-set related to life expectancy, health factors for 193 countries have been collected from the same 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.
* 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 affect 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. In a nutshell, this study will focus on immunization factors, mortality factors, economic factors, social factors and other health related factors as well. Since the observations 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.
* One thing that I wanted to share was the correlation **Heatmap** because there are some really interesting correlations here:
* There is a strong positive correlation between ‘Schooling’ and ‘Life Expectancy’ of 0.73. This may be because education is more established and prevalent in wealthier countries. This means countries with less corruption, infrastructure, healthcare, welfare, and so forth.
* Similarly to the point above, there is a moderate positive correlation between ‘GDP’ and ‘Life Expectancy’ of 0.44, most likely due to the same reason.
* Surprisingly there’s a moderate positive correlation between ‘Alcohol’ and ‘Life Expectancy’ of 0.40. I’m guessing that this is due to the fact that only wealthier countries can afford alcohol or the consumption of alcohol is more prevalent among wealthier populations.
* There is a very high correlation between thinness of 5-9 year-old and that of 1-19 year-old. Also between population and infant deaths, under 5 deaths, another is between schooling and income composition of resources. On the other hand Life expectancy and Adult Mortality are very highly negatively correlated.



|  |
| --- |
| **9. Future scope** |

* To be concrete, suppose the goal is to forecast female life expectancy at birth in Sweden in 2050. In the year 2000, the expected life span for Swedish women was just over 82 years. What will it be in the middle of the twenty-first century?

Perhaps the future will be characterized by poverty, misery, and a shorter life expectancy. Coming decades could bring nuclear war, massive biochemical terrorism, epidemics more deadly than the AIDS epidemic, catastrophic environmental change, lasting economic depression, or some other disaster or combination of disasters that might cause female life expectancy in Sweden to plummet far below its current level of more than 82 years, perhaps even down to zero. These possibilities are indicated in the figure below. If it does, then my best guess is that the mean value – of the range of possible life expectancies in 2050 – is 70 years, which is close to the current value of female life expectancy in the world as a whole. Discussions among a group of experts and systematic consideration of various scenarios would undoubtedly produce values different from 15% and 70 years, but these values illustrate the approach.

* One possibility is that the pace of age-specific mortality improvement over the next half century will be similar to the pace of improvement over the last 50 or 100 years
* A second possibility is that the pace of life-expectancy increase over the next half century will be similar to the rate of increase over past decades.
* Finally, the third possibility is that mortality improvements will accelerate in the future. Biology and biomedicine may be on the verge of unprecedented breakthroughs in knowledge about specific diseases and about the aging process itself – many knowledgeable scientists are of this opinion.

|  |
| --- |
| **10. Reference** |

The following sources have been used for the project development :

* To collect the data set : <https://www.kaggle.com/kumarajarshi/life-expectancy-who/data>
* <https://cognitiveclass.ai/courses/introduction-to-cloud>
* Introduction to Machine Learning with Python by Andreas C. Muller & Sarah Guido.
* Node-Red Labs of the course.
* Stack overflow.
* <https://bookdown.org/caoying4work/watsonstudio-workshop/jn.html#deploy-model-as-web-service>
* W3Schools. Com
* <https://github.com/>
* <https://bookdown.org/caoying4work/watsonstudio-workshop/jn.html#deploy-model-as-web-service>