Life Expectancy Prediction

Ishan Kotian (17IT2013)
Burhanuddin Naguthanawala (19IT5001)
Ashish Patil (19IT5008)
Rithvik Chavhan(19IT5012)

Information Technology Department
Ramrao Adik Institute of Technology
Nerul, India

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 data-sets are made available to public for the purpose of health data analysis. 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. 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.

I. INTRODUCTION

Machine learning is a field of computer science which has experienced exponential growth from the past few years. Almost every aspect of life is being changed by big data and machine learning. Health informatics sphere poses a great challenge to this domain. The ultimate aim of applying machine learning is to develop algorithms which can be trained well and make improvements over time. Life expectancy is one of the most important measure in terms of population's health in a country and is used as an indicator by many policy makers and researchers to complement economic measures of prosperity such as GDP etc. Prognosis of life depicts the average age that the members of a particular population group will be when they die. Life expectancy varies with developed and developing countries, ratio of birth to death, mortality rates of different countries and ratio of literate to illiterate population, all affect the survival time in one way or the other. The country's growth, advancements and accessibility of resources all are the factors of affect living rate of population. The life expectancy is calculated as the average survival time which indicates the median age of population where some might live till then, some might live more time span, some might live less but on an average the predicted value is the lifetime of that continent. Prognosis of life is not only instrumental in predicting living rate but also helps in deciding whether there is a tendency of occurrence of disease in a continent. Along with the prediction of life, classification of disease is another aspect of research. Disease prediction is done by considering the economic,

social factors of different countries in the particular continent and then we combine that data to predict it over a continent. The growth of the country affects the occurrence of disease in the country. Development rate of the country is dependent on, GDP, population awareness, illiteracy rate to literacy rate, birth to death ratio, all the factors have a combined effect on the striking of a disease. Therefore, machine learning is the suitable method to predict and classify. Regression, classification predictive algorithms can be used in various ways to achieve the desired output. For prediction of life expectancy, regression algorithms, linear regression and multiple regression are applied, whereas, for classification of diseases occurrence, application of classification algorithms, decision tree, random forest algorithm and k-nearest neighbor algorithm are applied to obtain the desired results.

II. LITERATURE SURVEY

Past studies have revealed a lot of work in the field of predicting life expectancy of a human being. After reviewing existing works and techniques in the prediction of human Life Expectancy, and finally reached a conclusion that it is possible to predict a Average Life Expectancy for individuals using advancing technologies and devices such as big data, AI, machine learning techniques, and PHDs, wear-able and mobile health monitoring devices, IOT. It is noticed that the collection of data is a huge challenge due to the privacy and government policy considerations, which will require collaboration of various bodies in the health industry. The interworking of a heterogeneous health network is also a challenge for data collection. Despite these challenges, a possibility of predicting Life by proposing an approach of data collection and application by smartphone, in which users can enter their information to access the cloud server to obtain their own predicted Lifespan based on the given inputs. To verify the accuracy of PLE prediction and validation of data quality, big data techniques and analysis algorithms need to be developed and tested in a real-life situation with several sample groups. As artificial intelligence technology is evolving and being applied rapidly, feasibility may be increasing to collect health data from the public as well as existing health agencies such as centralized health servers.

III. DATA MINING PROCESS

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.

A. Dataset

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. Missing data was handled in R software by using Miss map command. The

result indicated that most of the missing data was for population, Hepatitis B and GDP. The missing data were from less known countries like Vanuatu, Tonga, Togo, Cabo Verde etc. Finding all data for these countries was difficult and hence, it was decided that we exclude these countries from the final model data-set. 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.

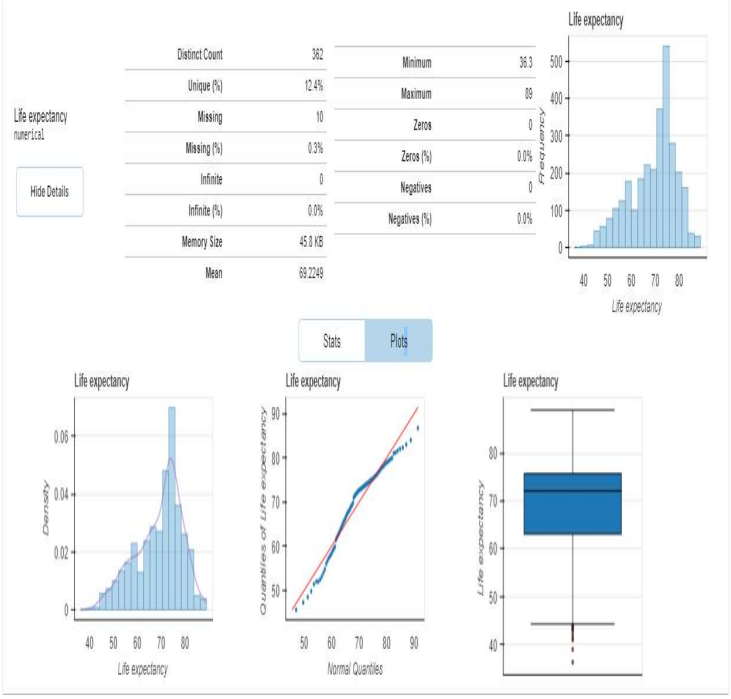
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2938 entries, 0 to 2937
Data columns (total 22 columns):
 #   Column                           Non-Null Count  Dtype
---  ------                           --------------  -----
 0   Country                          2938 non-null   object
 1   Year                             2938 non-null   int64
 2   Status                           2938 non-null   object
 3   Life expectancy                  2928 non-null   float64
 4   Adult Mortality                  2928 non-null   float64
 5   infant deaths                    2938 non-null   int64
 6   Alcohol                          2744 non-null   float64
 7   percentage expenditure           2938 non-null   float64
 8   Hepatitis B                      2385 non-null   float64
 9   Measles                          2938 non-null   int64
 10   BMI                             2904 non-null   float64
 11  under-five deaths                2938 non-null   int64
 12  Polio                            2919 non-null   float64
 13  Total expenditure                2712 non-null   float64
 14  Diphtheria                       2919 non-null   float64
 15   HIV/AIDS                        2938 non-null   float64
 16  GDP                              2490 non-null   float64
 17  Population                       2286 non-null   float64
 18   thinness  1-19 years            2904 non-null   float64
 19   thinness 5-9 years              2904 non-null   float64
 20  Income composition of resources  2771 non-null   float64
 21  Schooling                        2775 non-null   float64
dtypes: float64(16), int64(4), object(2)
memory usage: 505.1+ KB
```

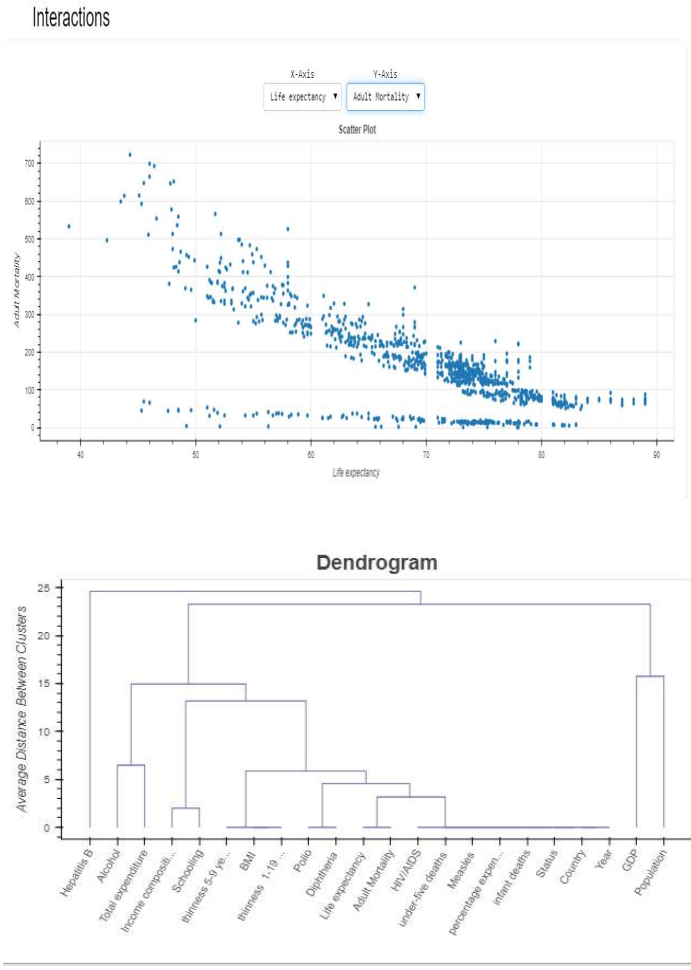
Dataset Statistics

Number of Variables	22
Number of Rows	2938
Missing Cells	2563
Missing Cells (%)	4.0%
Duplicate Rows	0
Duplicate Rows (%)	0.0%
Total Size in Memory	843.6 KB
Average Row Size in Memory	294.0 B

Variable Types

Categorical	2
Numerical	20





It is equally important to plot the data in graphical visualizations in order to understand the data, its characteristics, and its relationships. Henceforth, figures 1 to 4 are constructed as graphical plots of the data based on the summary statistics.

B. Data Mining Implementation

Random Forest Regression:

Random Forest Regression is a supervised learning algorithm that uses ensemble learning method for regression. Ensemble learning method is a technique that combines predictions from multiple machine learning algorithms to make a more accurate prediction than a single model. A Random Forest Regression model is powerful and accurate. It usually performs great on many problems, including features with non-linear relationships. To get a better understanding of the Random Forest algorithm, let's walk through the steps:

1.Pick at random *k* data points from the training set.

2.Build a decision tree associated to these *k* data points.

3.Choose the number *N* of trees you want to build and repeat steps 1 and 2.

4.For a new data point, make each one of your *N*-tree trees predict the value of *y* for the data point in question and assign the new data point to the average across all of the predicted *y* values.

Furthermore, the following settings were used in the Random Forest Regressor model.

1. Minimum samples leaf: 1,
2. Minimum samples split: 2
3. Number of estimators: 100

Analysis and Summary

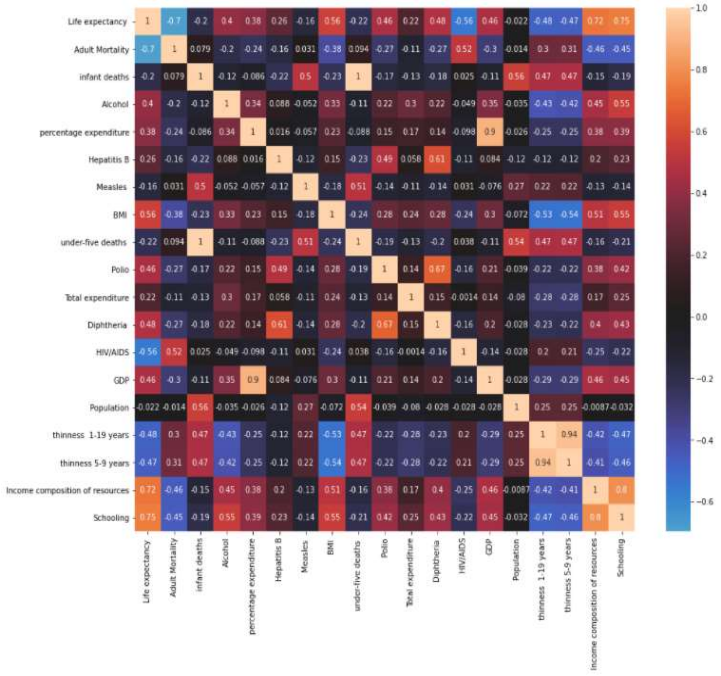
After running Random Forest Regressor algorithm the following results were obtained:

Accuracy: 96.482322%

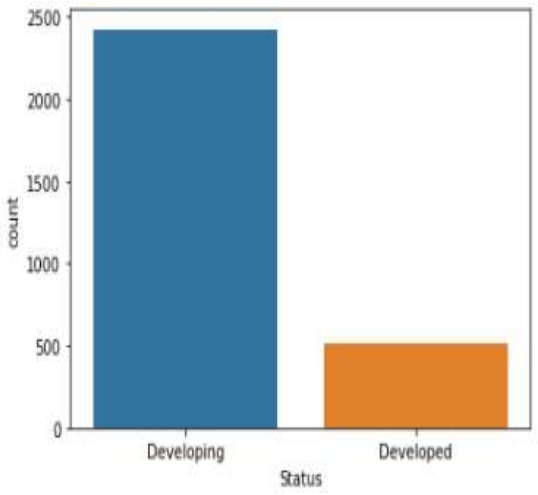
Mean Absolute Error: 17.45376

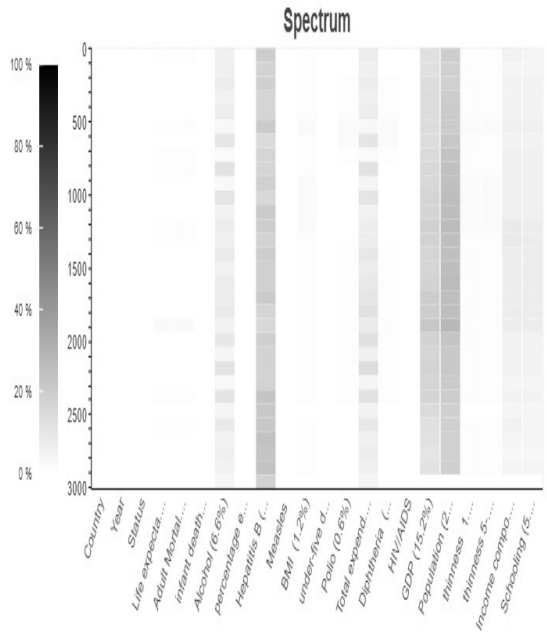
IV. ANALYSIS

Correlation Matrix

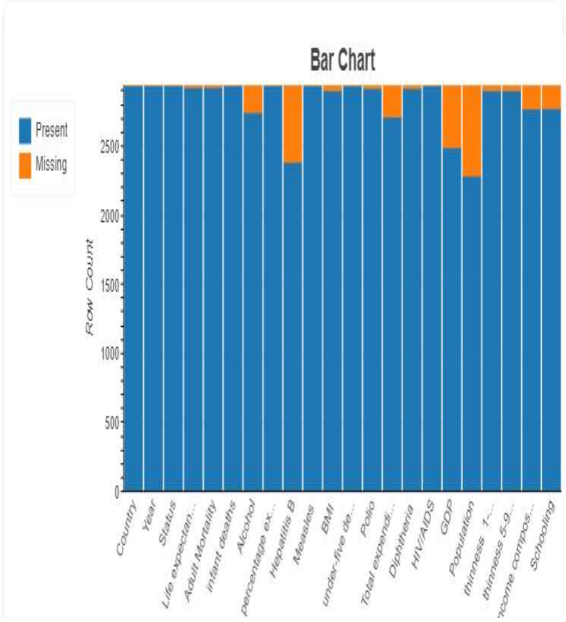


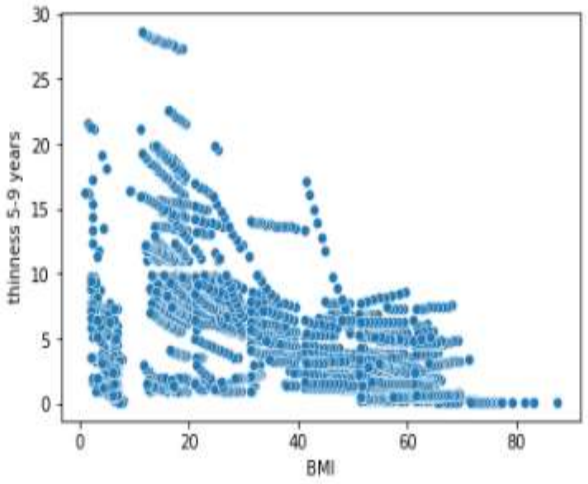
Marks grouped according to various features

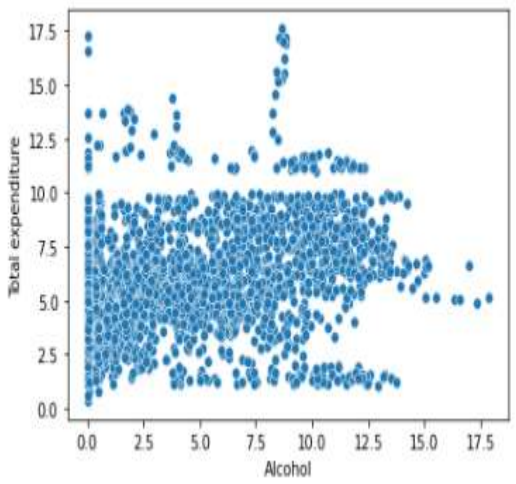


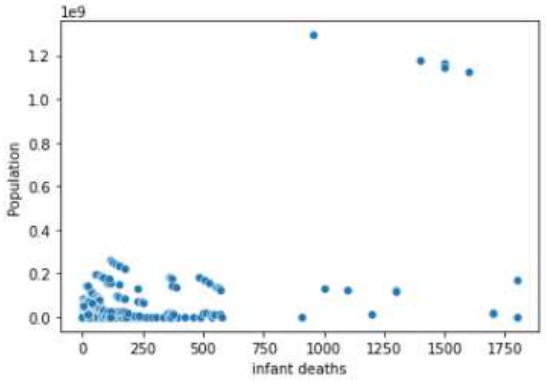


Missing Values









V. RESULT

```
from sklearn.ensemble import RandomForestRegressor
regressor = RandomForestRegressor(n_estimators = 100, random_state = 0)
regressor.fit(X_train, y_train)
```

```
RandomForestRegressor(random_state=0)
```

```
# Predicting a new result
y_pred = regressor.predict(X_test)
np.set_printoptions(precision=2)
y_pred = np.array(y_pred)
y_test = np.array(y_test)

print(np.concatenate((y_pred.reshape(len(y_test),1), y_test.reshape(len(y_test),1)),1))
```

```
[[63.27 62.5 ]
 [54.59 53.6 ]
 [82.98 83.3 ]
 ...
 [54.98 55.  ]
 [69.92 69.4 ]
 [74.15 75.  ]]
```

```
print('Random Forest Classifier Accuracy:',(accuracy_score)*100,'%')
```

```
Random Forest Classifier Accuracy: 96.482322951144 %
```

```
from sklearn.metrics import mean_absolute_error, mean_squared_error, r2_score
print(mean_squared_error(y_test,y_pred)**(0.5))
```

```
1.7437636565747747
```

```
from sklearn.model_selection import cross_val_score
accuracies = cross_val_score(regressor,X_train,y_train,cv=10)
accuracies.mean()
```

```
0.9601520673266644
```

VI. CONCLUSION

In this research paper, we have developed a model that will predict the life expectancy of a specific demographic region based on the inputs provided. Various factors have a significant impact on the life span such as Adult Mortality, Population, Under 5 Deaths, Thinness 1-5 Years, Alcohol, HIV, Hepatitis B, GDP, Percentage Expenditure and many more.

VII. FUTURE WORK

As future scope, we can connect the model to the database which can predict the life Expectancy of not only human beings but also of the plants and different animals present on the earth. This will help us analyze the trends in the life span.

A model with country wise bifurcation can

be made, which will help to segregate the data demographically.

ACKNOWLEDGMENT

We would take this opportunity to express our sincere gratitude to Ramrao Adik Institute of Technology, Nerul for giving us the opportunity to explore more on the topic "Student Performance in Exam" as a part of our Data mining mini project..

We are grateful to our respected Principal Sir Dr.Mukesh.D.Patil, for his support and guidance.

We would like to thank Dr. Ashish Jadhav sir, HOD Information Technology for his support.

We would like to profusely thank Mrs. Jyoti Deone ma'am without whom this project would have been a distant reality.

Thanks and appreciation to our classmates for their constant encouragement and support.

REFERENCES

[1] M. I. Jordan and T. M. Mitchell, "Machine learning: Trends, perspectives, and prospects," Science, vol. 349, no. 6245, pp. 255–260, Jul. 2015. 2.

[2] A. L. Beam and I. S. Kohane, "Big Data and Machine Learning in Health Care," JAMA, vol. 319, no. 13, p. 1317, Apr. 2018.

[3] V. M. Shkolnikov, E. M. Andreev, R. Tursun-zade, and D. A. Leon, "Patterns in the relationship between life expectancy and gross domestic product in Russia in 2005– 15: a cross-sectional analysis," Lancet Public Health, vol. 4, no. 4, pp. e181–e188, Apr. 2019.

[4] D. M. J. Naimark, "Life Expectancy Measurements," in International Encyclopedia of Public Health, H. K. (Kris) Heggenhougen, Ed. Oxford: Academic Press, 2008, pp. 83–98.

[5] H. Ouellette-Kuntz, L. Martin, and K. McKenzie, "Chapter Six - A Review of Health Surveillance in Older Adults with Intellectual and Developmental Disabilities," in International Review of Research in Developmental Disabilities, vol. 48, C. Hatton and E. Emerson, Eds. Academic Press, 2015, pp. 151–194.

[6] K. H. Zou, K. Tuncali, and S. G. Silverman, "Correlation and Simple Linear Regression," Radiology, vol. 227, no. 3, pp. 617–628, Jun. 2003.

[7] K. J. Preacher, P. J. Curran, and D. J. Bauer, "Computational Tools for Probing Interactions in Multiple Linear Regression, Multilevel Modeling, and Latent Curve Analysis," J. Educ. Behav. Stat., vol. 31, no. 4, pp. 437–448, Dec. 2006.