Life Expectancy Prediction using Machine Learning

Abstract— **Expectancy of life confers to the overall health of a community. Life expectancy is used as one of the most summary indicators for wellness. Based on statistical mediocre the number of years a community life that the researchers aim to pin down the hazards and speculate on future developments. Expectancy of life directly indicates the temporality state of a specific period. The expectancy of life of a community can also be defined as the mortality conditions of a specific year as life expectancy have an abrupt influence on the mortality rate. Based on the death rate expectancy of life is calculated for the living being. This research study attempts to demonstrate the possible impact on life style aspects on life expectancy by machine learning analysis with Python with the goal of increasing alertness regarding the expectancy of life to health and individual life style.**

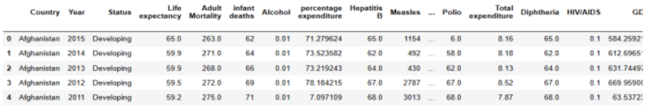
I. Introduction

Life Expectancy is an analytical as well as a statistical measure of the longevity of the population depending upon distinct factors. Over the years, Life expectancy observations are being vastly used in medical, healthcare planning, and pension-related services, by concerned government authorities and private bodies. Advancements in forecasting, predictive analysis techniques, and data- science technologies have now made it possible to develop accurate predictive models. In many countries, it is a matter of political debate about how to decide the retirement age and how to manage the financial issues related to the public matter. Life expectancy predictions provide solutions related to these issues in many developed countries. With the advancement in new systematic, accurate, efficient, and result oriented techniques in the field of Data Science, now predictions of the Life Expectancy of the selected region are becoming more prominent in demand of the government authorities and the private bodies and their policy- making. There have been many vast improvements in the field of data science and analytical techniques, which explains the rise in life expectancy around the world. These significant improvements in the predictive analysis techniques have also led us to more ways so that authors can improve the life expectancy of the distinct population. These improvements were solely dependent upon specific indicators. The extensive research into the prior life expectancy models has suggested us the inclusion of many more indicators than expected, such as; GDP(Gross Domestic Product), healthcare expenditure, family income, educational expenditure, infant mortality rate, adult mortality rate, healthcare plans, and population of the selected region. Recent studies have also revealed the impact of geographical factors, climate conditions on life expectancy. Implicitly, the educational background of people, health plans, economic stability, and the burden of diseases, BMI, and environmental variables also affect the lifestyle of the people. By summarizing all the factors mentioned earlier, the authors have created a distinct set of datasets that have helped us to reach the final destination of prediction in the desired population. However, in the final stages, the selection of the correct and accurate ML algorithms is probably the far most tedious job of this prediction model. Accuracy and reliability factors of final results depend upon the methodology used in the demonstration and as well as the correctness of the dataset

II. Methodology

DATASET DESCRIPTION

This dataset is a combination of data around the globe from numerous nations and aggregated by WHO for an individual country in an individual year. The life expectancy data consists of 193 nations and was collected by the United Nation Website. Considered data from the year 2000-2015 for 193 countries for further analysis. We have taken the life expectancy dataset from Kaggle, and it consists of 22 columns and 2938 rows .The indicators are : country (Nominal), year (Ordinal), status (Nominal), life expectancy (Ratio), adult mortality (Ratio), infant deaths (Ratio) , alcohol (Ratio), percentage expenditure (Ratio), hepatitisb (Ratio), measles (Ratio), bmi (Interval/Ordinal), under-five deaths (Ratio), polio (Ratio), total expenditure (Ratio), diphtheria (Ratio), hiv/aids (Ratio), gdp(Ratio), population(Ratio), thinness_1-19_years (Ratio), thinness_5-9_years (Ratio), income_composition_of_resources (Ratio), schooling (Ratio)



	Country	Year	Status	Life expectancy	Adult Mortality	infant deaths	Alcohol	percentage expenditure	Hepatitis B	Measles	...	Polio	Total expenditure	Diphtheria	HIV/AIDS	GD
0	Afghanistan	2015	Developing	65.0	263.0	62	0.01	71.279624	65.0	1154	...	6.0	8.16	65.0	0.1	584.2592
1	Afghanistan	2014	Developing	59.9	271.0	64	0.01	73.523582	62.0	492	...	58.0	8.18	62.0	0.1	612.6965
2	Afghanistan	2013	Developing	59.9	268.0	66	0.01	73.219243	64.0	430	...	62.0	8.13	64.0	0.1	631.7449
3	Afghanistan	2012	Developing	59.5	272.0	69	0.01	78.184215	67.0	2787	...	67.0	8.52	67.0	0.1	669.9590
4	Afghanistan	2011	Developing	59.2	275.0	71	0.01	7.097109	68.0	3013	...	68.0	7.87	68.0	0.1	63.5372

Figure 1

Jupyter notebook another similar kind of IDEs vastly used in the same context. Authors will be using one of the tools from the above-mentioned IDEs based on the need and suitability. For Visualization, many python libraries such as Matplotlib (mostly used for plotting2dfiguresand graphs) and seaborn (mainly used for 3-d graphs, heatmaps, and more advanced visualization features). Seaborn is based on the Matplotlib python library. Hence it inherits all the essential features of the latter one.

DATA analysis:

1. Univariate analysis :

Univariate analysis is performed on every parameter to understand the data range of every parameter.

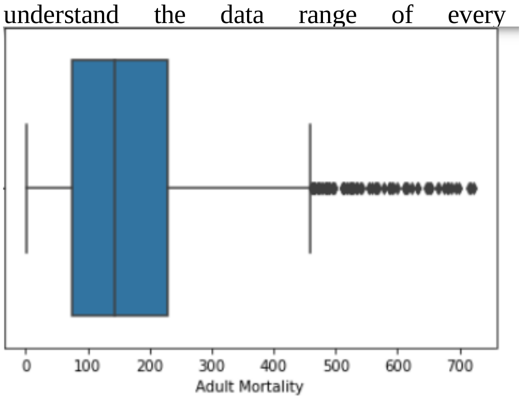


Figure 2

2. Histogram:

Histogram is placed to see how life expectancy in each country is distributed over years from 2001 -2015.

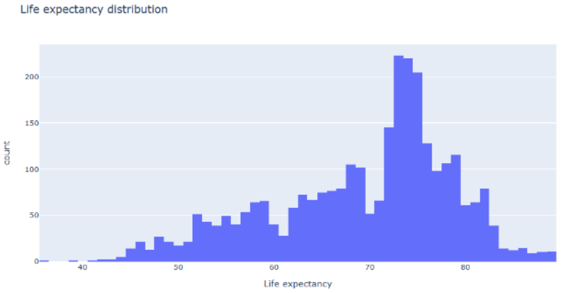
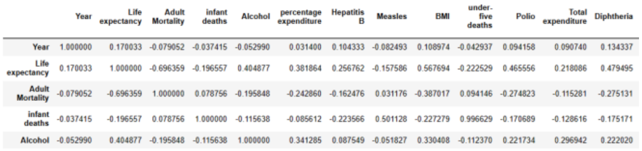


Figure 3

3. Correlation matrix:

Correlation matrix is performed to understand interdependencies of each parameter on life expectancy.



	Year	Life expectancy	Adult Mortality	infant deaths	Alcohol	percentage expenditure	Hepatitis B	Measles	BMI	under-five deaths	Polio	Total expenditure	Diphtheria
Year	1.000000	0.170033	-0.079052	-0.037415	-0.052990	0.031400	0.104333	-0.082493	0.108974	-0.042937	0.094158	0.090740	0.134337
Life expectancy	0.170033	1.000000	-0.696359	-0.196557	0.404877	0.381864	0.256762	-0.157586	0.567694	-0.222529	0.465556	0.218086	0.479495
Adult Mortality	-0.079052	-0.696359	1.000000	0.078756	-0.195848	-0.242860	-0.162476	0.031176	-0.387017	0.094146	-0.274823	-0.115281	-0.275131
infant deaths	-0.037415	-0.196557	0.078756	1.000000	-0.115638	-0.085612	-0.223566	0.501128	-0.227279	0.996629	-0.170689	-0.128616	-0.175171
Alcohol	-0.052990	0.404877	-0.195848	-0.115638	1.000000	0.341285	0.087549	-0.051827	0.330408	-0.112370	0.221734	0.296942	0.222020

Figure 4

4. Heat matrix:

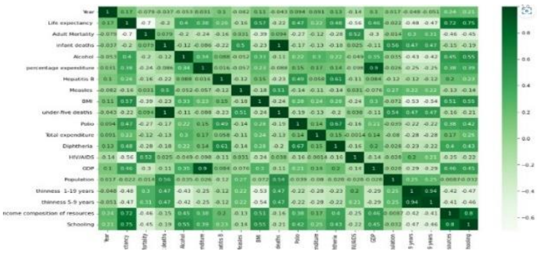


Figure 5

5. Scatter plot:

Scatter plot is performed to understand developed nation and developing nation factor influence like BMI, alcohol, diseases.

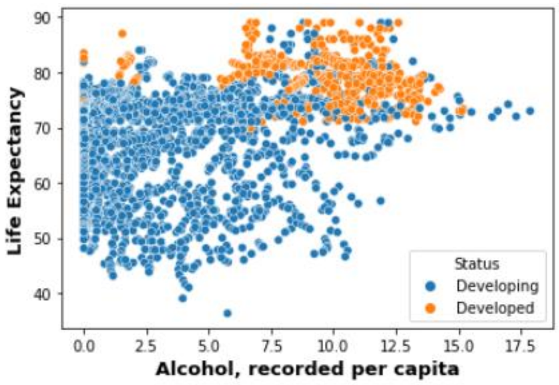


Figure 6

Machine learning algorithms:

We can use various Machine Learning techniques for solving these problems. Now we are using Random Forest Algorithm technique will be discussed below

Random forest: It is a type of supervised machine learning algorithm that combines several algorithms of similar techniques. The random forest can solve regression as well as a classification problem. Random decision forest or random forest is an ensemble method, which consists of a multitude of decision trees for classification and prediction problems. The output is the mean/average of the prediction values of the individual trees. The random forest method provides the necessary correctness required to the decision tree caused by overfitting the training set.

Correlation between the individual models is the key in this method. Because decision trees are extremely responsive and competent to the data with which they are typically taught, any changes in the training set can result in significantly different tree structures as shown in the figure.

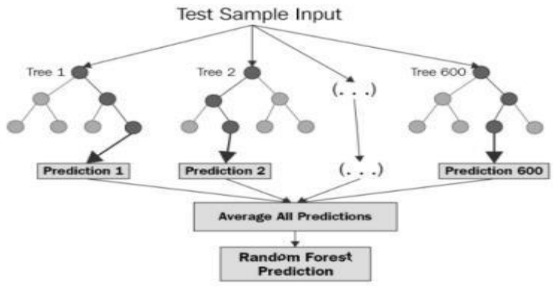


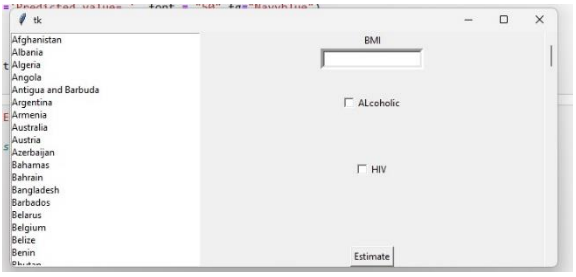
Figure 7

The random forest utilizes this approach by allowing individual trees to randomly sample from the dataset with replacement, ensuing in different trees. This operation is referred to as "bagging".

DATA Cleaning:

There are multiple parameters when analyzed are country specific than individual specific. Eg:GDP, Population, Developing, adult mortality rate, etc., so designing life expectancy model to an individual taking this parameters don’t give enough awareness. Whereas predicting expectancy based on individual health status such as BMI, alcohol status or illness status like HIV gives more awareness which is the main objective of the project.

So the parameters which are country specific are learnt into data frame for prediction by taking median of the values and considering on values from GUI which are specific to individual.



Afghanistan
Albania
Algeria
Angola
Antigua and Barbuda
Argentina
Armenia
Australia
Austria
Azerbaijan
Bahamas
Bahrain
Bangladesh
Barbados
Belarus
Belgium
Belize
Benin

BMI

☐ ALcoholic

☐ HIV

Figure 8

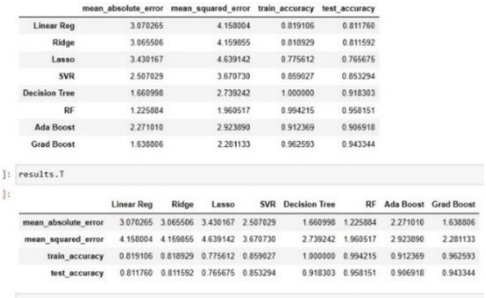
All other Values which are impacting life expectancy are taken from WHO data based on country.

So parameters which are country dependent like Status, AdultMortality, infantdeaths,percentage expenditure, Hepatitis B , Measles, under-fivedeaths, Polio, Total expenditure, Diphtheria, GDP, Population, thinness 1-19 years,thinness 5-9 years, Income composition of resources ,Schooling are taken from Kaggle data set which are relavent to Country selected and individual data like BMI, Alcoholic, HIV are taken from GUI.

This essentially gives dependency of a person/individual personalised life expectancy than the national level life expectancy which in-turn gives an individual factors which can be controlled solely.

III. Results

Random Forest Algorithm The random forest regression variable was fitted into the training data. The results clearly show that Random forest performs the best among given models with an accuracy of approximately 96%. Analysis of features is an important portion of this research. So based upon the random forest, which is performed far superior to the other models, below is the feature importance bar graph. It clearly shows that features such as adult mortality, HIV-AIDS, BMI, schooling, and income composition of resources have a far superior effect on life expectancy than other features. Because random forest performs best and given far better accuracy than other models.



	mean_absolute_error	mean_squared_error	train_accuracy	test_accuracy
Linear Reg	3.070265	4.158004	0.819106	0.811760
Ridge	3.065506	4.159855	0.818929	0.811592
Lasso	3.430167	4.639142	0.775612	0.765675
SVR	2.507029	3.670730	0.859027	0.853294
Decision Tree	1.660998	2.739242	1.000000	0.918303
RF	1.225884	1.960517	0.994215	0.958151
Ada Boost	2.271010	2.923890	0.912369	0.906918
Grad Boost	1.638806	2.281133	0.962593	0.943344

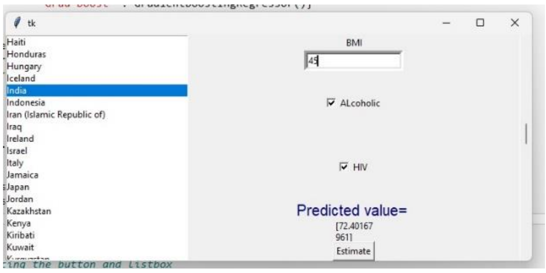
```
]: results.T
]:
```

	Linear Reg	Ridge	Lasso	SVR	Decision Tree	RF	Ada Boost	Grad Boost
mean_absolute_error	3.070265	3.065506	3.430167	2.507029	1.660998	1.225884	2.271010	1.638806
mean_squared_error	4.158004	4.159855	4.639142	3.670730	2.739242	1.960517	2.923890	2.281133
train_accuracy	0.819106	0.818929	0.775612	0.859027	1.000000	0.994215	0.912369	0.962593
test_accuracy	0.811760	0.811592	0.765675	0.853294	0.918303	0.958151	0.906918	0.943344

Figure 9

I. Summary & Conclusion

Here you can see based on BMI, alcohol status how life expectancy is directly impacted. A person having a BMI of 21 and alcoholic has life expectancy as 74.9 while an unhealthy BMI of 45 and alcoholic, HIV has 72.4 years



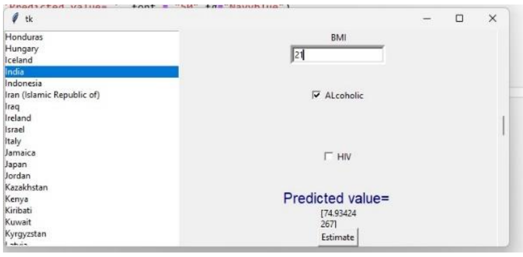
BMI 45

☑ ALcoholic

☑ HIV

Predicted value=
[72.40167
961]

Figure 10



BMI 21

☑ ALcoholic

☐ HIV

Predicted value=
[74.93424
267]

Figure 11

II. References

[1] N. Ali, D. Srivastava, A. Tiwari, A. Pandey, A. K. Pandey and A. Sahu, "Predicting Life Expectancy of Hepatitis B Patients using Machine Learning," 2022 IEEE International Conference on Distributed Computing and Electrical Circuits and Electronics (ICDCECE), 2022, pp. 1-4, doi: 10.1109/ICDCECE53908.2022.9793025.

[2] V. Bali, D. Aggarwal, S. Singh and A. Shukla, "Life Expectancy: Prediction & Analysis using ML," 2021 9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), 2021, pp. 1-8, doi: 10.1109/ICRITO51393.2021.9596123. [3] https://youtu.be/Lvyk94mE6sk