Analyzing the Influence of Various Factors on Life Expectancy

Spoorthi K S
Computer Science & Engineering
PES University
Bangalore, India
spoorthikalkunte@gmail.com

Thrupthi N
Computer Science & Engineering
PES University
Bangalore, India
thrupthi2804@gmail.com

Sahana Rao
Computer Science & Engineering
PES University
Bangalore, India
sahanarao2902@gmail.com

***Abstract*—The life expectancy is a measure which can serve as a comparison between individual countries, and assists in highlighting the important aspects which could affect it, and thus might bring about a vital change in many major avenues like the healthcare, education and socio-economic sectors. The primary objective of this study is to understand the effect of numerous factors including health, existing co-morbidities, educational levels, GDP on the life expectancy and curate a detailed analysis of the same after performing correlation analysis and finding out independent variable relationships, if any, and to use regression methods to model out the scenario, and thus, provide a prevention-betterment report which indicates where countries are falling behind and what they could do in terms of facility improvements and economic development to improve their average life expectancy.**

***Index Terms*—life expectancy, regression models, Random forest**

I. Introduction

Every human in the world would thrive to have higher life expectancy. Thus it becomes important to learn the influence of various factors on life expectancy in order to increase the life expectancy across all countries. Then one can figure out the factors that are leading to the decline in the life expectancy and try to look into those factors for better living and survival. Life expectancy is a common performance measure which is used to compare not only the living conditions of individuals, but also the health facilities and economic factors of countries. Since 1900, when the expected life expectancy was only around 30 years with high global inequality due to industrialization in some counties, the global average life expectancy has more than doubled and is now above 70 years. The inequality of life expectancy is still very large across and within countries. Life expectancy at birth, and at all other ages, has risen rapidly during the last century due to many factors, including reductions in infant mortality, rising living standards, improved lifestyles and better education and higher literacy rates, as well as advances in healthcare and medicine. Demographic and socio-economic factors like education, GNI (Gross National income) per capita, gender play a major role in determining life expectancy. Health factors include the prevalence of various health diseases and immunization provided (if exists), the expenditure of the country on facilities assisting hospitals and care centres, mortality rates, and others, which if managed appropriately can result in healthier, safer conditions and health satisfaction, thus improving overall life expectancy. Thus major factors like state of country,its development, outbreak of epidemic and pandemic diseases, health conditions, life styles cultivated etc are instrumental in influencing the life expectancy of current people. Besides, life expectancy also depends on various other factors like accidents,unexpected natural calamities(earthquakes,tsunami,nuclear disasters).The later ones are not of much importance though, as they are unpredictable incidents which need more structured and broader data to represent and analyze. This study is aimed at investigating various aforementioned factors as determinants of life expectancy, and providing a comparative analysis between different countries, so as to aid policymakers and officials in allocating the countries' resources appropriately, and improve overall standard of living.

II. Related Work/Literature Review

The paper [1] mainly focused on analysing the various factors affecting the life expectancy. They have also compared the differences in important factors affecting longevity between developing and developed countries. They have designed MLR models to study the impact of various factors in order to improve the life expectancy of people.
The more important features were figured out using spearman-rank correlation. The features like HIV/AIDS, income composition of resources, schooling were found to affect the life expectancy significantly. Surprisingly features like consumption of alcohol, population of countries had less effect. They have analysed model without differentiating the "status" (developed or developing). They have also noticed the difference factors influencing life expectancy in developed and developing countries separately. It also proves that these features that affect life expectancy are more significant for developed countries than for developing countries. When not differentiating the status of countries, the model represented was better. The RMSE was 3.818. When considering the status, the model generated by data of developed countries was better, and RMSE is 2.57.
This study conclusively showed that, the factors like schooling,

epidemic and pandemic diseases, income in general are most significant factors affecting people's life expectancy. It throws light on the fact that the factors like schooling, level of medical care have a greater impact on developing countries. On the other hand the factors like thinness 1-19 years, thinness 5-9 years, income composition of resources only influenced that of developed countries.In conclusion, this paper shows that MLR model can be used to show that different features affect life expectancy of different kind of countries.

In [2] the determination of life expectancy has mainly been attributed to, and discussed by weighing demographic variables like total fertility rate (TFR) and adolescent fertility rate, socioeconomic variables like mean year of schooling, and GNI per capita, and health factors like HIV prevalence rate and number of physicians per ten thousand populations in a given year. Data and necessary information were obtained from the WHO, United Nations Development Programs, and World Population Data Sheet. The data obtained from 91 countries was then used for univariate analysis, bivariate analysis and then backward multiple linear regression to obtain the features that affected Life expectancy the most. The study was able to establish that HIV prevalence rate, TFR, mean year of schooling, and GNI per capita were the significant predictors of LE in the low and lower middle income countries, ans showed that life expectancy was inversely proportional to the number of HIV cases, and directly proportional to the inflation rate and GNI.

Since (credits to [3]) , various factors (main ones being economic, social, mortality and immunization) come into play, the use of Lasso penalised linear regression (because of multicollinearity between features makes linear regression model unreliable) , argmin (minimum value of function), Random Forest model (is a classifier containing multiple decision trees, and its categories of output are determined by the mode of those by individual trees-here used to calculate the variable to be considered), XGBoost Model (non-linear model) , SHAP value (to explain XGBoost model) was used to consider the data. The study analysis done in the above research paper is about the prediction of life expectancy in different countries and the analysis of influencing factors to determine the factors that lead to the increase or decrease of life expectancy. According to this study, there are about 24 variables (which include mortality, economic factors, social factors, and immunological factors) from 2000 to 2015 that affect life expectancy for nearly two hundred countries across six continents. The analysis of lasso parameter table in their study, shows that GDP, schooling (positive impact on life expectancy), adult mortality (negative impact), BMI (negative impact on Europe, North America penalised to 0 whereas rest of the continents are positively influenced) and percentage expenditure generally have the greatest impact on life expectancy in six continents. However, this influence of effects varied from country-to-country. Analysis of random forest model shows that feature importance of six continents is different. In the XGboost model, the lowest SHAP values were found to be Haiti in North America and Sierra Leone, Botswana and Eritrea in Africa, with life expectancy below 42.In the interaction analysis, it can be seen that HIV/AIDS is negatively correlated with income composition of resources and positively correlated with adult mortality.

Further according to [4], besides the effect of socio-economic-communicable diseases, it was observed that the non-communicable diseases (NCD) were also leading cause for premature disability and death worldwide. Smoking, hypertension and overweight were identified as the 3 key risk factors that were shared by all NCDs. Between 1989 and 2012, amongst 9,061 community-dwelling individuals (mean age 63.9 years, 60.1 percentage women) in the Dutch population, it was observed from the data that every 9 out of 10 individuals aged 45 years and older developed NCD during their lifetime, with a third of them developing multiple NCDs during follow-up. From the analysis, of the 3 risks is associated with a longer overall life expectancy of about 6 years, and a 2-year compression in lifetime spent with NCDs and most of this increase is due to an extension of disease-free life expectancy. This means that individuals without these risk factors not only live longer than individuals with these risk factors, but also spend less of their lifetime after the onset of symptomatic disease (which is referred to as compression of morbidity).

The paper [5] aimed to identify determinants of life expectancy and designate clusters of Indonesian provinces with similar characteristics, and also provide cooperation strategies to improve life expectancy. They used 2015 published data from the Ministry of Health of Indonesia where all 34 Indonesian provinces were included as analysis units. Four latent variables linked to life expectancy were considered, including health system, socioeconomic factors, demographics and the environment. The theoretical model was tested in Lavaan package in RStudio using the maximum likelihood estimator. To achieve a good-fit model, they simplified latent variables, segregated "health system" into three latent variables: "health insurance", "health workforce" and "healthcare facilities", and statistically insignificant paths were removed based on their P-values less than 0.05. In this study, SEM was used to test for pathways towards life expectancies to other structure in their theoretical model. From all variables, there were six constructs with bilateral correlations towards each other: (1) life expectancy, (2) health workforce, (3) healthcare facilities, (4) environment, (5) mean years of schooling, and (6) expenditure per capita. Magnitude of correlation between six constructs ranged from 0.83 (health workforce and expenditure per capita) to 0.36 (life expectancy and education). Expenditure per capita was found to be strongest among the six constructs. Using k-means clustering, 5 clusters of provinces were generated, with each cluster consisting of provinces within a specific range of life expectancy.

III. Proposed Solution

The problem statement involves measuring the effects of health, socio- economic and demographic factors on global life expectancy, and analysing why and in what arena one country performs better than the other, while keeping in mind the differentiation among the nations in terms of their status as developed or developing nations, over the course of years from 2000-2015.

There are multiple factors which affect the life expectancy of a country, and these dependencies can be explored to analyse how influential or prominent a particular feature is, on different strata of countries.

A. Dataset

We have taken our dataset from kaggle(Life Expectancy(WHO)) which is basically taken from WHO.We found that the percentage expenditure column in this dataset was incorrect.So we got this data from the WorldBank website.This life expectancy dataset consists of enumerable indicators such as diseases, income, country development status, schooling, BMI, alcohol consumption for different countries which helps in analyzing the effect of various socio-economic features on the life expectation. 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 to a country which area should be given importance in order to efficiently improve the life expectancy of its population. 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 data-set. On initial visual inspection of the data showed some missing values. As the data-sets were from WHO, no evident errors were found. 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.

The 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 (Diphtheria, Polio, Hepatitis B, HIV/AIDS), Mortality factors (Measles, Alcohol, BMI, thinness 1-19years, thinness 5-9years, Under 5 deaths, Adult mortality, infant deaths), Economical factors (Percentage expenditure, GDP) and Social factors (Status, population, schooling, income composition of resources, country, year)

B. Data Cleaning and Visualisation

Dealing with missing values:

- The missing values in various columns were identified and were replaced with appropriate imputations.

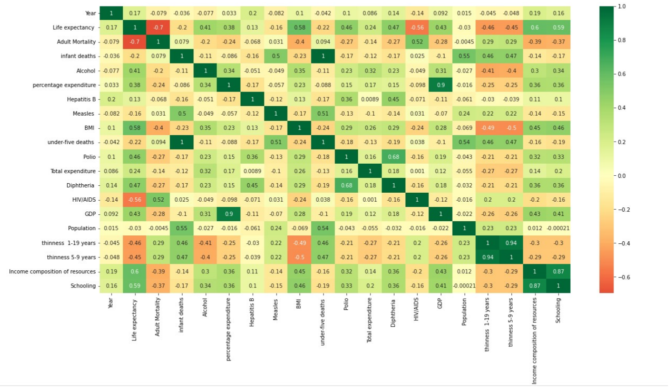

Fig. 1. heat map

- There were few (10) missing values in the life expectancy (dependent variable) column, due to which the corresponding rows were dropped.
- As the dataset contains the LE data ordered from 2015 to 2000 for each country, the missing values were replaced using the method of forward and backward filling country wise.
- For countries for which the entire value set of an attribute was missing, zero imputation was done.
- The outlier plots show that the columns like HIV/AIDS, GDP etc have outliers. We decided not to drop the outliers as their presence might be due to the variation of count in each country based on their population.

Fig. 3. 3 shows the box plot for all features after removing the redundant features.Even though most of the attributes have a higher number of outliers,we came to a conclusion of retaining them owing to the fact that it might have occurred as the population of the countries are huge and removing them might lead to inaccuracies.

After cleaning, the general view of LE can be viewed Violin plot 2 in Fig. 2. Violin plot ,used especially for numerical distribution, helps in comparing and depicting density of the variable (in our case-LE) amongst various countries.From the figure it can be observed that median of life expectancy has slightly increased over the years and is more concentrated within the range of 70-80.

C. Methodology

Firstly we did feature selection of important features influencing life expectancy for building models and drawing conclusions.We used heat maps1 to get rough idea of correlations of various features with LE.

We tried to remove the redundant and less important features. The plot shows the correlations of the features with other features throwing light on some of the features that we can drop. Correlation between under-5 deaths and infant deaths is 1 (very high correlation), leading to the conclusion that one of them can be dropped.Life expectancy has barely any correlation to the population, but is highly correlated with adult mortality, BMI, HIV and income composition of resources.So

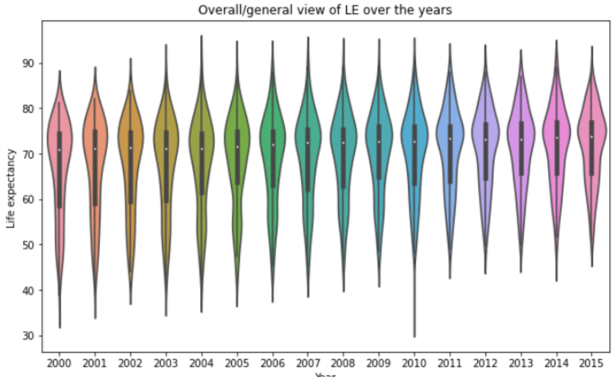

Fig. 2. Violin Plot

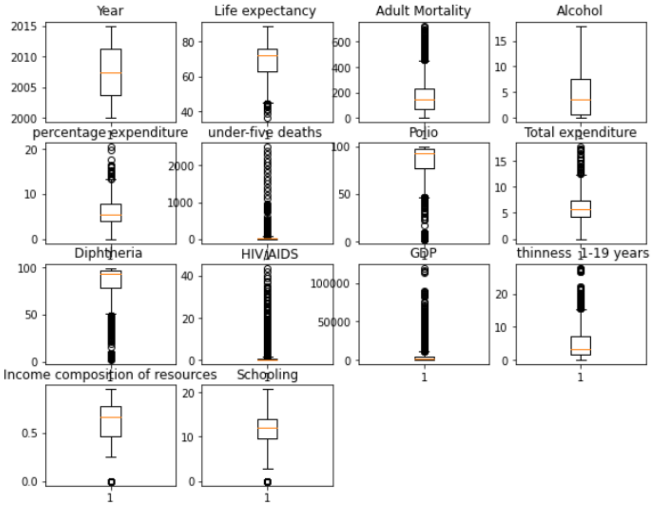

Fig. 3. Outlier analysis

we decided to drop it.Also thinness 1-19 is a superset of thinness 5-9 years, and thus the latter was dropped.

Further we tried to get the most important features using lasso regression feature selection technique. Here we set the alpha value to be 1.The unimportant features were recognized, whose lasso coefficients were zeroes. Other features like year, under-5 deaths, hepatitis B, measles, total expenditure were deemed to be less prominent from this method. This model resulted in an r2 score of 0.7442, and an RMSE of 4.867. From this, we tried to gauge the features like status, country, adult mortality, alcohol consumption, BMI, Polio, Diptheria, HIV/AIDS, GDP, thinness 1-19 years, income composition of resources, schooling which were highly influencing the life expectancy of people. We also attempted Ridge regression which can reduce the standard error by adding some bias in the estimates of the regression, and can benefit a dataset where more features have a similar impact on the response variable. While Ridge performed better than Lasso with an RMSE of 4.582, it gave the same features when a threshold of +0.5 and -0.5 was considered on the model's coefficients. The feature importance can be visualised in the bar plot at Fig.

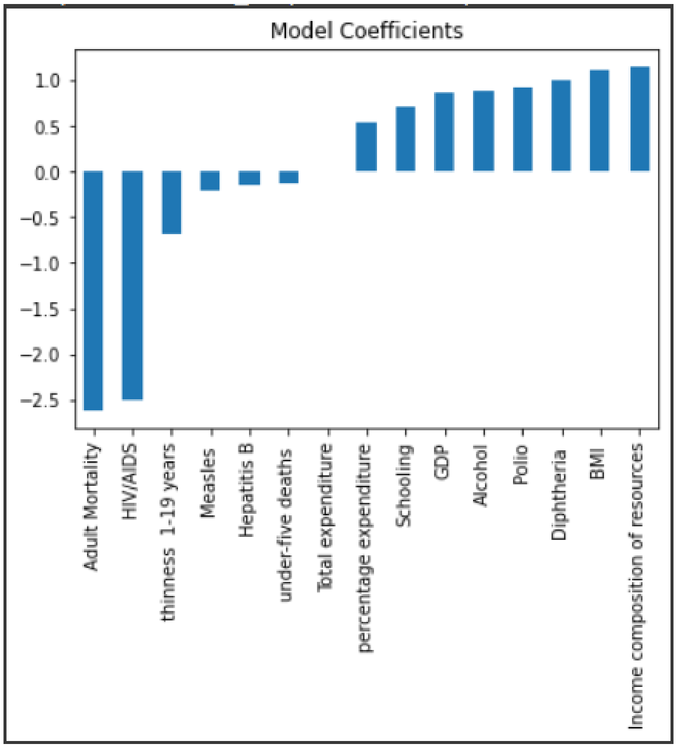

Fig. 4. Plot of coefficients of features (Ridge regression)

4. We used these features for building models that predicts the life expectancy.We used models like MLR, XGBoost, Random Forest regressor for predicting the values. Linear regression is used to model a linear relationship on the different features of the data, in order to predict the target variable. XGBoost, or Extreme Gradient Boosting, is a library which optimises gradient boosting, and is highly efficient and flexible. Random Forest algorithm uses ensemble learning methods by creating multiple decision tree framework from the bootstrapped data, averaging the results to output a new result that often leads to strong predictions.

Moving towards country-wise analysis, it was observed countries with least life expectancy and their frequency of occurrence were as follows:Central African Republic(1),Haiti(1),Sierra Leone(13) and Malawi(1).A key point to be noted over here is that all these are developing countries and belong to Africa(Except Malawi).Sierra Leone is identified as the country with least life expectancy in the year of study(2000-2015). The reason can maybe be inferred to economic- health and natural calamities that affected the respective countries during that time. In Africa, AIDS epidemic, malnutrition, curable diseases, and civil strife have taken a tremendous toll on human life.The reason maybe lack of clean water, proper food resources, education about hygiene and an extreme shortage of medical facilities and trained medical personnel as a result it has one of the highest child mortality rates in the world. Haiti observed the drop in its life expectancy(36.3-Least so far in the year range 2000-2015) in the year 2010 because of the fact that in the same year in January,it experienced a 7.0 magnitude earthquake, leaving its capital Port-au-Prince devastated.Political instability,natural disaster lack of infrastructure and planning along with Epidemics and aid mismanagement has contributed towards its downfall.

On the other hand, Austria, Belgium,France,Iceland and Italy were few of the countries which maintained their LE above threshold of 80 and also do note that these were developed countries.Higher LE can be accounted to the facts like raised living standards, improved lifestyle and better education, as well as greater access to quality health services.

In order to account for the growth in each country over 2000-2015, percentage change(rise or fall in LE) was calculated.It was observed that countries with higher LE/developed countries(Belgium,France,Iceland,Italy) had lesser growth(around 0.5) in comparison to countries with lower LE (Eritrea,Zimbabwe,Haiti).This can be interpreted that the countries with higher LE have reached their peak in LE and have been stable in maintaining it across the years. We were delighted to observe that the countries that have relatively lower LE(Haiti,Eritrea,Sierra Leone) are striving towards improving the LE of their country.

In addition to the above mentioned analysis, a study was also conducted while keeping the status of participating countries as a differentiating factor. Since developed and developing countries have vastly different standards of living, with differences in the availability of necessary amenities to all its citizens, it makes total sense that there would be varying factors affecting the life expectancy in the two variations. We obtained the features which were most important and affected the target variable to the highest extent, using SHAP values. SHAP values interpret the impact of having a certain value for a given feature in comparison to the prediction made if that feature took some baseline value. It can be used to determine the influence of each feature on a prediction. First the dataset was separated into two parts: one, consisting sorely of developed countries and the other with developing countries. The dataframe of developed countries of 496 rows, comprising of 31 countries, and the dataframe of developing countries of 2400 rows and 150 countries. Over each of these sets of data, an XGBoost model was trained, and the SHAP values were obtained for the entire training set. Following this, a Random Forest Regression model was trained on only the those features that were deemed to be impactful, since Random Forest performed comparitively well on the global dataset, resulting in lower error values. The resulting model was evaluated using RMSE and r2-score as the evaluation metrics.

IV. Results and Discussions

The project was initiated with the initial milestone of using machine learning models to model and predict the global life expectancy, i.e the life expectancy of all the countries.

Multiple linear regression was modelled considering the filtered features. We trained the model on 80 percent of the dataset(train data) and validated it on other 20 percent of dataset(test data), and obtained an RMSE of 4.511 and R2 score of 0.7803. We also tried modelling XGBoost with the same set of features, for which we got RMSE of 2.212 and R2 score of 0.947. The Random Forest model gave an RMSE of 1.90 and R2 score of 0.961. From this it is evident that the Random forest model performed best. A scatter plot of the actual vs predicted values is shown in Fig. 5.

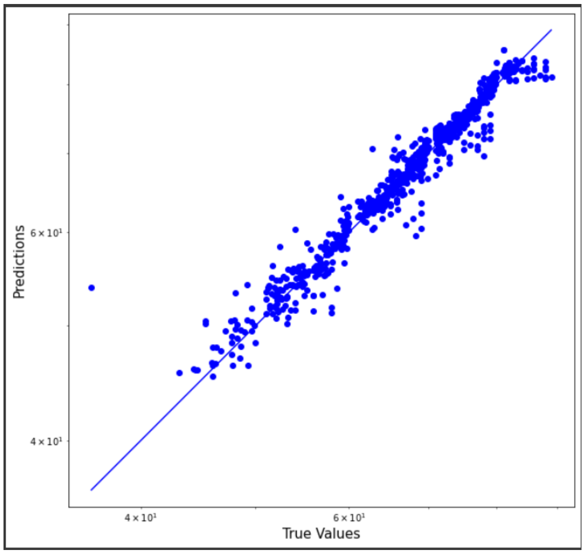

Fig. 5. Plot of true vs predicted target values from the Random Forest model

The status of the nations was also realised to be a highly indicative factor to compare between the development of nations, and would provide more specific results in terms of similar classes of countries. Only those features whose SHAP values crossed a particular threshold were considered. The SHAP values for both developed and developing countries is given in the plots 6 and 7.

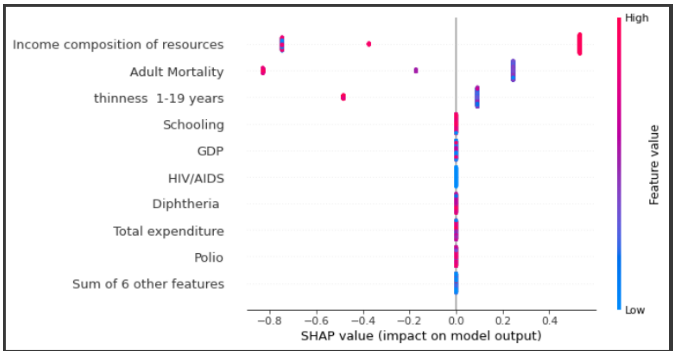

Fig. 6. SHAP values for developed countries

The Random Forest model trained on the dataset of developed countries included the features Income composition of resources, Adult Mortality and thinness 1-19 years, and yielded an RMSE of 2.814, while the model trained on developing countries included Income composition of resources, HIV/AIDS, Schooling, Adult Mortality, BMI, Polio and GDP, and provided an RMSE value of 4.605. The results with

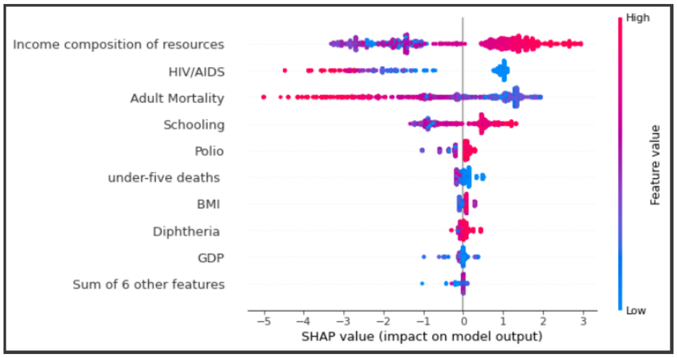

Fig. 7. SHAP values for developing countries

respect to the various Random Forest models have been displayed in the table I.

TABLE I
COMPARISON IN TERMS OF RANDOM FOREST MODEL

Model	**Testing RMSE**	**r2-score**
Whole dataset	1.90	0.961
Developed countries only	1.852	0.762
Developing countries only	1.939	0.955

V. CONCLUSION

This study is performed with the objective of analysing the most important factors in terms of demographic, health and socio-economic variables, affecting not only global life expectancy, but also that of specific countries and across classes of developed and developing nations.

Health-economy-Education are considered as the 3 pioneers to determine the status and well-being of a country and this proved to be right once again.Yes, we happened to observe that all the countries with lack of health amenities and economy(like African countries and Haiti) resulted to have an imbalance in maintaining the life expectancy.In countries with average rate of growth in life expectancy like Syrian Arab Republic and Libya, once they faced a dip in LE it proved them to be challenging to recover from that fall and lead to a very slow increase in its LE.On the other hand, countries like Belgium,France had improved healthcare facilities and higher economic status.Thus countries with lower LE can infer from countries having higher LE that proving basic needs to all people of all age groups does have a greater impact on it stability.

Further studies could be done on this data by grouping it in terms of continents, and arriving at the factors currently dominating the standard of living and development of specific continents. The data itself could also be amplified, both in terms of more historical data, and more features, which could present the opportunity to delve deeper into matters like, how only economic, disease-related or disasters categories affect the life expectancy, and how the major events of the decade like the Covid-19 pandemic, the Great Recession of 2008 and the Ebola virus in 2014 affected both the areas of maximum impact and other nations with respect to trade and economic collaborations.

VI. PEER REVIEW

1) We were asked whether we could find any spurious correlations in our data and they asked us to eliminate it if found.We tried figuring out whether our analysis had spurious correlations. We could not find any spurious correlations as such in our analysis.
2) They also suggested us to try time series analysis on our data.We actually tried doing time series analysis.We found that the data had only the trend from 2000-2015.It barely had residuals and seasonality components associated with it.So it was just like linear data.Thus we came to a conclusion of not using time series analysis for our data as it was not much useful.
3) There were around 10 missing values in the target attribute column(Life expectancy).So we told them that we had dropped those corresponding rows from the data as it had missing target values.They recommended us to use those missing values as test data to predict those life expectancy values using the models we have built.Actually,the missing values were the result of not recording the data for some of specific countries.There was no data available for any year for those countries.So we felt it would not be so appropriate to predict the values for those countries' life expectancy values without having any training rows for those countries.
4) We were also asked to model ridge regression along with lasso regression for feature selection.As they suggested we implemented it and found that the important features that we got from this ridge regression technique was almost same as the features we filtered out from lasso regression.

REFERENCES

[1] X. He, J. Hu, C. Liu and Y. Zhang, "Analysis on Relevant Factors Affecting Life Expectancy," 2022 IEEE Asia-Pacific Conference on Image Processing, Electronics and Computers (IPEC), 2022, pp. 569-572, doi: 10.1109/IPEC54454.2022.9777372.

[2] Mondal MN, Shitan M. Impact of Socio-Health Factors on Life Expectancy in the Low and Lower Middle Income Countries. Iran J Public Health. 2013 Dec;42(12):1354-62. PMID: 26060637; PMCID: PMC4441932.

[3] Y. Wang, "The Greatest Factors Affecting Life Expectancy: A Research based on Different Continents and Countries," 2021 3rd International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI), 2021, pp. 531-541, doi: 10.1109/MLBDBI54094.2021.00107.

[4] Licher S, Heshmatollah A, van der Willik KD, Stricker BHC, Ruiter R, de Roos EW, Lahousse L, Koudstaal PJ, Hofman A, Fani L, Brusselle GGO, Bos D, Arshi B, Kavousi M, Leening MJG, Ikram MK, Ikram MA. Lifetime risk and multimorbidity of non-communicable diseases and disease-free life expectancy in the general population: A population-based cohort study. PLoS Med. 2019 Feb 4;16(2):e1002741. doi: 10.1371/journal.pmed.1002741. PMID: 30716101; PMCID: PMC6361416.

[5] Paramita, S.A., Yamazaki, C. Koyama, H. Determinants of life expectancy and clustering of provinces to improve life expectancy: an ecological study in Indonesia. BMC Public Health 20, 351 (2020)