2021 9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)
Amity University, Noida, India. Sep 3-4, 2021

2021 9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO) | 978-1-6654-1703-7/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICRITO51393.2021.9596123

Life Expectancy: Prediction & Analysis using ML

Dr. Vikram Bali[1*], Dr. Deepti Aggarwal[2*], Sumit Singh[3], Arpit Shukla[4].
[1,2]Department of CS & Engineering, JSSATE, NOIDA, India.
[3,4]Department of Computer Applications, JSSATE, NOIDA, India.
[3]sumit01du@gmail.com

Abstract— Life expectancy (LE) models have vast effects on the social and financial structures of many countries around the world. Many studies have suggested the essential implications of Life expectancy predictions on social aspects and healthcare system management around the globe. These models provide many ways to improve healthcare and advanced care planning mechanism related to society. However, with time, it was observed that many present determinants were not enough to predict the longevity of the generic set of population. Previous models were based upon mortality-based knowledge of the targeted sampling population. With the advancement in forecasting technologies and rigorous work of the past, individuals have proposed this fact that other than mortality rate, there are still many factors needed to be addressed in order to deduce the standard Predicted Life Expectancy Models (PrLE). Due to this, now Life expectancy is being studied with some additional set of interests into educational, health, economic, and social welfare services. In the Analysis, the authors have implemented different machine learning algorithms and have achieved better accuracy based on pertinent features of the dataset.

Keywords— Life Expectancy (LE), Machine Learning (ML), Predicted Life Expectancy (PrLE), Ensemble methods.

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.[1]

Studies have suggested that in early life or the pre-modern era, the average lifespan of human beings was around 30 years in approximately all parts of the world(Fig.1). Since then, industrial enterprise and modernization have valued the rapid increase within the lifespan all around the world. The advancement of technology, better healthcare facilities, and education for all have led to positive changes in the lifestyle of people. Which, in turn, increases the expected average age of a human being. However, there were still many countries with less life expectancy than the rest of the world in the early 1900s. The whole reason for such inequality is the disoriented healthcare facilities in these countries. Developed countries have speedily improved their health care and also the public distribution mechanism. This inequality between developed and developing countries has led to such an improper distribution of life expectancy around the globe.[2]

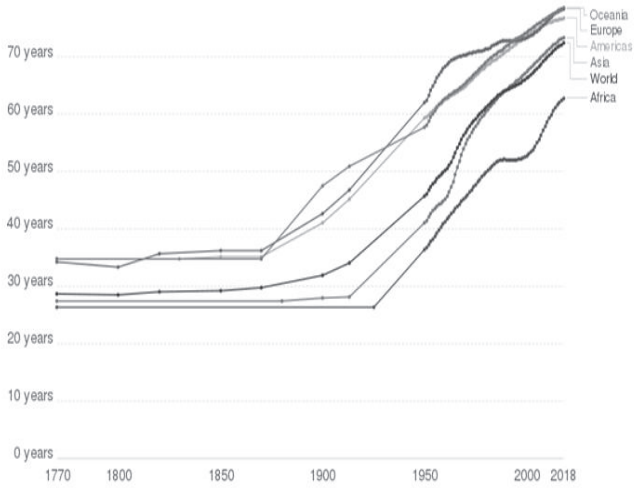

Fig. 1. Shown above is the average life span of different continents around the globe. (Source: UN population division, 2019).

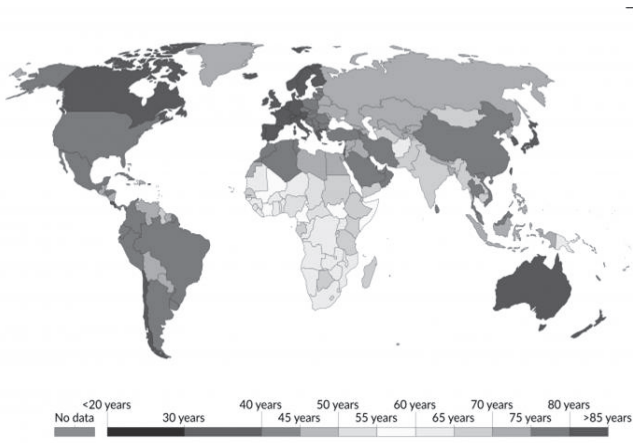

Fig. 2. Avg. no. of years a newborn is expected to remain alive if the mortality rate remains the same throughout the year on the world map. (Source: UN Population Division, 2019).

Due to certain developments in public healthcare, now emerging countries are also catching up with the other developed countries in terms of life expectancy. In 2019, most of the Central African countries have a low life expectancy of around 52-55 years, whereas, in Japan, recent statistics have shown that life expectancy is around 87 years for women. The lifespan of South Korea was twenty-three

years, a century past. Nevertheless, as of today, the Life Expectancy of India has almost tripled in the last 100 years, and in South Korea, it has almost quadrupled since that time period.

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.[3] 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.[4]

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 (which in turn depends upon the selection of independent variables and their sources of origin).[5] Further, the literature based study will show the different work done by different authors in this field following the different methods used with the help of different flow charts and at last conclusion is discussed.

II. LITERATURE-BASED STUDY WITH LIMITATIONS

After looking into various research work which has taken place by different researchers and authors will derive and describe their method and inference from their research via different data visualization techniques.

Aggarwal D. et al. (2017), in their paper, presented a 5-year predictive tool of patients from a recent one or two hospital visits. In recent years, this is one of the most accurate predictive models, providing results up to 5 years. Their work was based upon the ensemble of different Machine Learning models. The Ensemble technique is a Machine Learning method in which various basic models are combined to achieve the final predictive model. This was the reason behind the deployment of the final model with precise implementation because it was an ensemble with 75 individual models. To attain such a level of precision, they need to be specific about patients and need to have access to each individual's EMR data. It performed far better than the previous models.[6]

Kyle J. Foreman et al. (2018), proposed a paper related to life expectancy estimation in 'Global Health Metrics'. In which their prediction results were mostly centered on the GBD, risk factors, and injuries analysis of 195 countries from 1990 to 2016. Consideration of that number of health-related factors was quite astonishing, and the projection of each attribute against the health scenario was shown. They took 79 health drivers into account for the estimation. They used the ARMA (Auto Regressive Moving Average) model for the estimation of life expectancy based on the health drivers. The demographic and educational factors of each country were not mentioned or taken into account in this particular paper. In the ARMA model, the auto-regression analysis and moving average methods are used and these two methods are applied to time-series datasets(well-behaved).[7]

Beeksma et al. (2019), presented this paper in BMC Medical Informatics. It is one of the recent researches in the concerned subject area. They proposed the idea of using supervised machine learning using recurrent neural networks on deceased patients by using their medical records. They approached the task with supervised machine learning. Then, they trained and tested the data on LSTM (Long Short-Term Memory) recurrent neural networks. The LSTM method is a form of RNN (Recurrent Neural Networks). In the RNN method, the output from the calculation is taken as input within the current phase. LSTM method was formulated to beat the matter of long-term dependencies of RNN. It is mostly used in the time-series data due to the lags between the important unknown events in the time series. Their model was based on non-text and text features of medical records. The first one was the base model (containing non-text features), and another one was the keyword model (containing text features in EMR). The Keyword model proposed a better accuracy of 29% than compared to 20% of the base model. But, some limitations of this model were the data availability and not generalized in predictions.[8]

N. Kerdprasop et al. (2017), in their paper, suggested that there is an association of environmental as well as economical factors to life expectancy. They used the Chi-Square Automatic Interaction Detector (CHAID) method for categorical values and regression on continuous and numerical values. The underlying principle under the CHAID algorithm is a Decision tree. It is a tool used to discover the relationship between the variables. Using this technique, they attempted to relate between economic growth & resources of a country to the life expectancy of its people. They have done the prediction of the life expectancy of people living near the Mekong River. Eventually, their results revealed a strong relationship between GDP growth & environment to the life expectancy of people.

Yang, J. (2016), in their paper, presented a model to make basic forecasting for sex-specific and socio-economic-specific Life Expectancy in the Netherlands using a set of critical assumptions about the future trends. They were using the traditional Li-Lee model for the estimation. This Li-Lee model is an extension of the original Lee-Carter model. The Lee-Carter model was introduced in 1992, and as of today, it is used by many authors to determine future mortality trends. This model is fed with details of age-specific mortality rates and the result is another matrix of forecasted mortality rates.[9]

Creating life expectancy forecasting models may be a tedious job to try and do. There are several challenging factors involved in this process. From identification of indicators, collecting the datasets, to preprocess the available data, and at last final phase of implementation by applying appropriate machine learning algorithms on to the datasets. Testing the accuracy is another matter of discussion, but still, it is one of the involved challenges in the process. Now, the

authors will discuss each of the difficulties briefly below one by one:

1) The first significant challenge in front of us will be to prepare the correct and error-free dataset. The sources of data needed to be legit and trustworthy.

2)After identification of legitimacy of sources, dataset need to be categorized the two categories;

a) Structured data: This kind of data is stored in EMR or UN official records as well as on the government's official websites for public reference.

b) Unstructured data: It consists of numerous non-validated information that needs to be converted into measurable and statistically discrete attributes to accommodate the other characteristics of the dataset.

3) Usage of numerous non-standardized abbreviations and improper attributes (with no proper definition in the context of life expectancy) can create ambiguity in the final results. Elimination of redundant information is also an essential process in the initial stages and also a challenging one.

4) The selection of appropriate techniques, machine learning algorithms, and technologies is a big hurdle in order to create the LE prediction model. The next phase will basically determine the consistency and efficiency of the model into the future.

The next phase involves the challenge of validation of the used ML technique and testing of the results. In the next phase, the impact of each indicator on the final results will be tested. It also serves as the refinement phase of the model.

III. METHODOLOGY

It is good to know the objectives and essential theory behind the problem. But to practically showcase, the forecasting is a totally different scenario. So, everyone needs to be aware of the practical aspect of the problem and how it's going to be implemented practically.

There are many kinds of open-source IDEs available in real-world systems that can be used in the coding stages of the work. These data-science-related project works are code into either Python or R language. So, it is needed to choose wisely about the IDEs, in which these codes can be verified and tested after compilation. It is needed to select IDEs over simple text editors for the coding task, because of the debugging and in-build testing features. Spyder IDEs is open-source Anaconda distribution. It is mostly used in the data-science project because of the inclusion of many data-science libraries such as NumPy, SciPy, Matplotlib, and IPython and it can be further extended by adding plugins for numerous other purposes.

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 plotting 2d figures and 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. Seaborn helps to produce some of the most attractive statistical graphics in the data-science domain's problems. Seaborn, ggplot, and shiny are some of the most popular data visualization third-party libraries widely used in both Python and R languages. For small problems or less feature-oriented problems, IDEs are much better options. But for a large scale, where the density of the dataset can become a matter of concern for the researcher. Hence, Specialized Machine learning platforms like the TensorFlow framework were developed by Google Brain Team within Google AI's Organization in 2013. It is widely used in large-scale ML projects and deep neural networks.[10]

The objective is to predict the result of the number of dependent variables in the comparison to number of independent variables. We can use various Machine Learning techniques for solving these problems. Now some of the techniques will be discussed below:

Linear regression: It is one of the simple techniques, in which everyone can predict the number of the outcome of the dependent variable depending upon various features. Multiple regression analysis allows us to develop a mathematical model dependent on numerous features. The stepwise regression method is composed of iteratively adding or removal of dependent features from the set, at the end giving us the best performing model. Bhosale et al. (2010) predicted the life expectancy of humans based upon their heart rate, respiration rate, and blood pressure using the linear mathematical model.

It is one of the simple techniques, which can predict the outcome of the dependent variable depending upon single or multiple independent features.[11]

Ridge regression: Least square method does not differentiate between the important and less-important features in the model. This results in overfitting and multicollinearity in data. The Ridge Regression evades all of the above-stated problems. Ridge regression provides simply sufficient bias to make the estimates fairly dependable approximations to actual population values.

Ridge regression belongs to the class of L2 regularization. L2 regularization adds an L2 penalty, which is equal to the square of the magnitude of coefficients.

Decision tree: The Decision Tree is a quietly used mechanism in classification and continuous-valued prediction problems. Any tree might be trained over the splitting of source into subsets mainly established entirely on the attribute test value. So in this way, it is replicated on each derived subset in the recursive fashion, which is also known as recursive partitioning. This way recursion is done while the subsets at all nodes have a similar value of the target feature, or while splitting it does not gives the value to the predictions. A decision tree classifier may be built without any domain information or parameter settings, making it ideal for exploratory knowledge discovery. Decision trees can also deal with data that has a lot of dimensions. In general, the decision tree classifier is fairly accurate. The traditional inductive approach to learning insights based on classification is the decision tree. The data is easily comprehended, analyzed, and visualized. There are several advantages of working on a decision tree:

- The decision trees execute variable/feature selection implicitly.

- It is capable of dealing with both numerical and categorical data. It can also adjust with problems involving multiple outputs.
- Usually, Decision trees need comparatively less effort from users for data preparation.
- The performance of the tree is unaffected by nonlinear interactions between parameters. Decision trees execute variable screening or feature selection implicitly.

However, many times the decision tree may achieve over-complex trees. Which may result in overfitting. To avoid overfitting mechanisms like boosting and bagging are used. To create unbiased trees, you need to balance the dataset before implementing the decision tree.[12]

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.[13]

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".

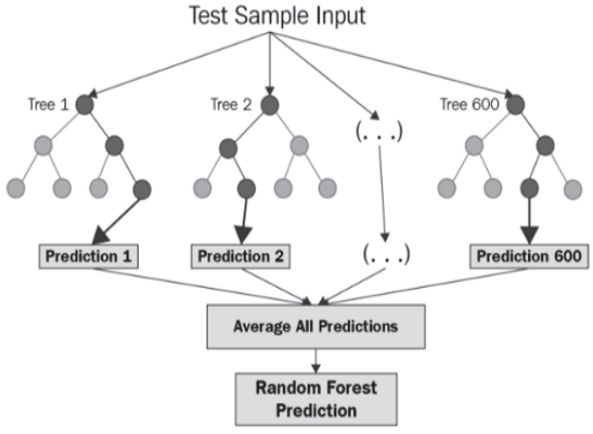

Fig. 3. Flow chart presenting design & description of Random Forest mechanism.

All these factors given below are some of the major parameters involved in the estimation of life expectancy.

TABLE I. LIST OF ATTRIBUTES USED IN LIFE EXPECTANCY PREDICTIVE MODELS.

Attribute	*Brief Description*
Infant mortality rate	It depicts the death of kids beneath a certain age one per a thousand live births.
Alcohol consumption	It represents the average consumption of alcoholic beverages by the net population.
Percentage expenditure	The proportion of total average family expenditure is described by associate item (budget share).
BMI	It is an estimation of body fat in accordance with the weight and height applicable to adult males and females.
GDP	It is additionally referred because of the estimated price of all the products and services created in a very year by a country.
Under-five deaths	It represents the number of total deaths of children after birth till 5 years per 1000 number of births.
Total income composition	The relative share of every income supply or group of sources is expressed as a proportion of the aggregate total income of that cluster or region.
Educational expenditures	It refers to the total sum of expenses done on the educational services and subsidies acquired by each group of people of that region.
Health care expenditures	Whole consumption of health-related services, expenses on the health care plans (including personal care and family healthcare plans).
Population	It simply describes the number of people (active entities) in the region.
Environmental criteria	Major environmental factors, for example, climate change, modernization, and altitude.
Schooling	Total expenditure on educational services by people.
Thinness	Prevalence of thinness among children and adolescents.

IV. RESULT & DISCUSSIONS

In the implementation part initially, authors have created a profile report using the **pandas_profiling** library of python. It shows that a lot of data in GDP and population features for a lot of countries is missing. Imputation is not the best method for handling missing data in this particular dataset, if imputation is used then we are taking information from other countries and putting it in for a different That information would be inaccurate. So, dropping those rows with missing population & GDP is the only option. There might be a bit of loss of data. Other features with missing values, such as BMI, thinness, and hepatitis B, will be filled in with 0s. If those features have no values, 0 is a safe assumption that won't skew the data or cause problems with data evaluation.

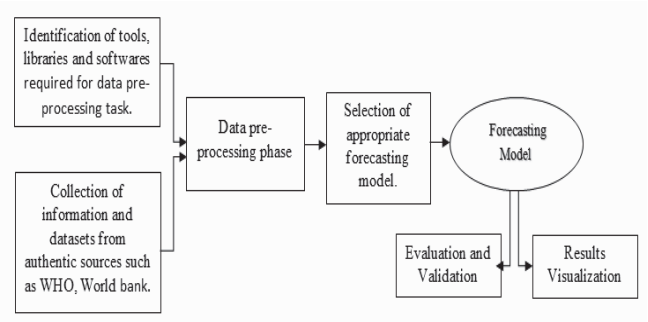

Fig. 4. A flow chart presenting the design & description of each phase of the Basic Forecasting model.

Used models might struggle to handle NaN since it is not a viable option, then 0 have been used in place of NaN. Authors have also dropped country and years because they are not looking to see if a country's life expectancy varies from year to year. The country and year in which the data from the dataset was collected should not be used to forecast life expectancy.

The goal is to forecast life expectancy using health and economic parameters. Having data from several countries across years gives us a greater variety of information.[14]

Then, the baseline model is been created, which is nothing but the mean of life expectancy with its Mean absolute error.

```
Mean Baseline: 68.7 years
Mean absolute error of 8.07
```

Fig. 5. Baseline model result with MAE.

The test set has been divided into three parts: 70% for training, 15% for validation, and 15% for the test set. Next, is to perform necessary data-preprocessing such as naming the feature properly and dropping the country, year, and status feature from the dataset. The next phase in implementation is to create a common function for processing all models at once.

#def model (x _ train, y _ train, x _ val, y _ val):

In this function, the training and validation have been provided set to the function named model. Based on validation set results MAE and R-2 scores will be calculated to evaluate the individual models.

```
LinearRegression R^2 Score 0.8279750460836258
LinearRegression mean_absolute_error 2.9843251733281515

LinearRegression Coefficient [-1.62239786e-02  7.57469733e-02 -7.79107194e-02  5.42415690e-05
  9.71685991e-04 -7.85154603e-06  3.94840771e-02 -5.56021973e-02
  7.83221800e-03  5.73819624e-02  1.96841990e-02 -5.02079013e-01
  5.16733463e-05 -2.16993157e-09 -1.03715401e-01 -7.46542461e-04
  1.05865569e+01  8.00205603e-01]
LinearRegression Intercept 52.786794904461296

Ridge regression mean_absolute_error 3.046007689629716
Ridge Regression Score R^2 Score 0.8206970478559336

Ridge Regression Coefficient [-2.09742400e+00  3.22130710e+00 -4.40771486e-01  1.89912931e-01
  4.90515150e-02 -1.94233513e-01  8.21989580e-01 -3.23301116e+00
  2.30219885e-01  1.54878916e-01  5.75509122e-01 -2.55138396e+00
  6.58478964e-01  1.29168846e-03 -4.72389104e-01  4.53783671e-02
  2.35023073e+00  2.71450883e+00]
Ridge Regression Intercept 68.67859466493168

Decision tree mean_absolute_error 1.5558441558441558
Decision tree Score R^2 Score 0.9194627223261882

random forest mean_absolute_error 1.2192649350649352
random forest Score R^2 Score 0.9608331126733731
```

Fig. 6. Results of all the models implemented from the model() function.

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. Hence, authors have to evaluate the random forest model on test set results.[15]

```
random forest test mean_absolute_error 1.2769235294117653
random forest test Score R^2 Score 0.9600640226106466
```

Fig. 7. The test set results are based on Random Forest Model.

Permutation importance is a frequently used type of feature importance. It shows the drop in the score if the feature would be replaced with randomly permuted values.[16]

Weight	Feature
0.4274 ± 0.0245	income composition of resources
0.1969 ± 0.0326	adult mortality
0.1513 ± 0.0276	hiv aids
0.0154 ± 0.0049	bmi
0.0090 ± 0.0018	thinness 5 9 years
0.0072 ± 0.0016	alcohol
0.0047 ± 0.0008	under five deaths
0.0045 ± 0.0024	schooling
0.0036 ± 0.0009	gdp
0.0029 ± 0.0017	total expenditure
0.0024 ± 0.0009	infant deaths
0.0018 ± 0.0010	percentage expenditure
0.0017 ± 0.0004	hepatitis b
0.0017 ± 0.0004	polio
0.0014 ± 0.0005	population
0.0014 ± 0.0003	thinness 1 19 years
0.0012 ± 0.0007	measles
0.0011 ± 0.0007	diphtheria

Fig. 8. Permutation feature importance based on Random Forest Model.

These results again show the correctness of previously evaluated feature importance with possible error terms. Features such as population, diphtheria, measles, and thinness are less significant than the other ones and are less effective on life expectancy.

Below are Partial Dependence Plots (PDP) of these important features to find the variation of these features with life expectancy.

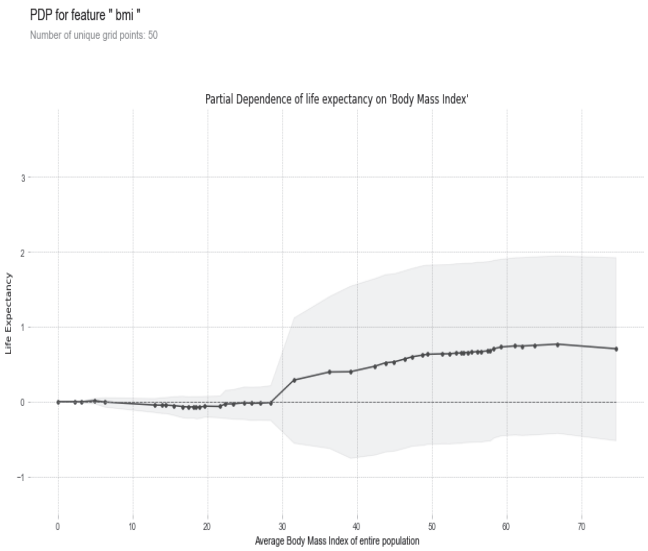

Fig. 9. PDP plot of Life expectancy v/s BMI.

The above PDP plot here shows the variation of these essential features with life expectancy and correlation.

Partial Dependence Plots (PDP) show the dependency between the target variable and a set of independent features, marginalizing over the values of all other features (the 'complement' features). Usually, everyone can interpret the partial dependence as the expected target response as a function of the independent features. Partial PDP plots with more than one feature present a clearer picture and

dependency of these features on outcome i.e., Life expectancy.

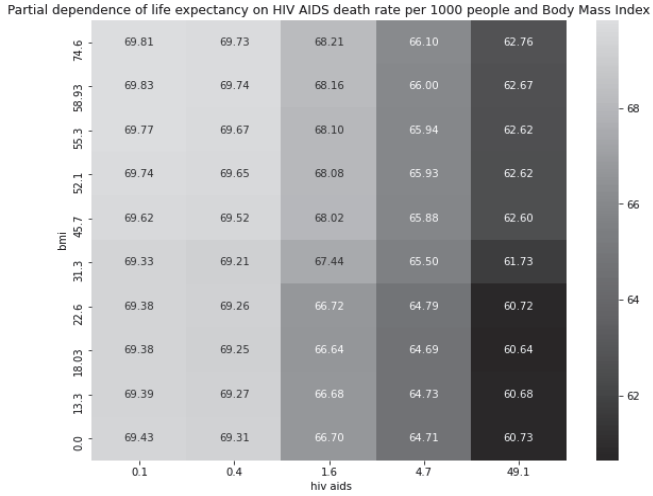

Fig. 10. Partial PDP of life expectancy on HIV-AIDS & BMI.

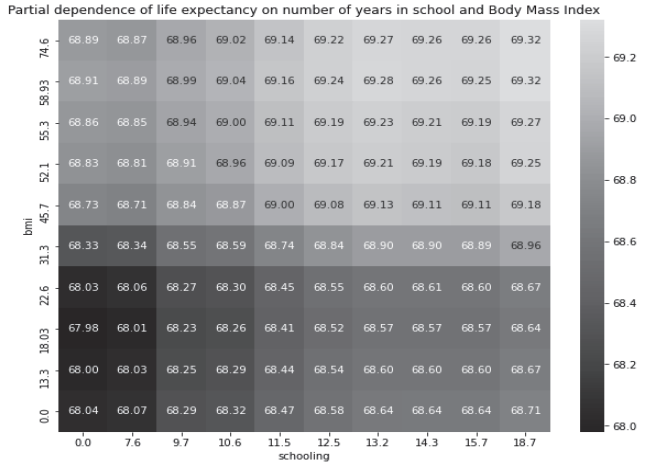

Fig. 11. Partial PDP of life expectancy on Schooling & BMI.

These partial PDP helps us to generalize the behavior and easily understand the common trend between target and dependent features. [17]

V. IMPLICATIONS

Life expectancy results are not just the merely figures derived from the datasets; instead, these are quite influential in solving real-world problems. Most of the demographers use forecasting models to predict the average life span of males and females for specific countries as shown in the table:

TABLE II. DIFFERENCES BETWEEN MALE & FEMALE IN FEW COUNTRIES.

Country	Male	Female	Difference
Russia	64.7	76.3	11.6
Belarus	66.5	78.0	11.5
Lithuania	68.1	79.1	11.0
Bangladesh	60.9	71.1	10.2
Syria	59.9	69.9	10.0
Ukraine	66.3	76.1	9.8
Latvia	69.6	79.2	9.6

These results are quite useful in studying the life structure and living patterns of the population. It helps in drawing appropriate insights from those results. In general, life expectancy predictions are used in the field of research and policy-making decisions. Numerous government policies all around the world are prepared using these results, such as in healthcare services, human resource management decisions, public health wellness, and maintenance. It can also affect the decisions of the policymakers, for example, health expenditure by the concerned authority. Generic Life Expectancy predictions can also help the population to improve their lifestyles and to make efficient, healthier decisions individually. For example, the adverse effects of smoking can endanger human health and may cause some lung-related severe diseases to the individual. Hence, smoking should be avoided to live a healthy long life.

Life insurance companies are one of the major stakeholders in using these life expectancy predictions. Risk assessment is quite a critical part of the life insurance profession. They classify their applicants based upon certain criteria by using these results. They aim to enhance the process of risk assessment using predictive analytics. Some of the stronghold life insurance organizations still rely on the conventional mortality rate based prediction models. These Firms decide the premium of their plans as per the risk-based on the history of their customers. They categorize their customers on their estimation of survival inferred from their medical history and several other affecting factors.[18]

By taking the example of Bangladesh, the authors will explore a case study implication of life expectancy. In Bangladesh, these results have some policy implications for the country, and it suggests that life expectancy is dependent on economic growth, per person expenditure on health, and individual having a higher education is expected to have a healthier life than the others. Therefore, the study suggested that to have political stability, adequate and satisfactory public health policies, the government needs to involve more often in decision making to increase life expectancy and provide economic stability. Generally, this scenario is real in a sense, for all countries around the world (not just in this particular case). Countries having higher GDP pertain to higher LE than compared to others. Nations, which are spending more wealth on Medical and health care expenditures, are expected to have a higher life expectancy than others. Life expectancy results could be used for statistical comparisons based on illustrative features of the results.[19]

One of the most proficiently used real-time systems using life expectancy analytics is cancer diagnosis based upon socio-economic factors. In this, the patient's EHR is collected and the available data is used for the estimation of lost lifespan for each patient due to different types of cancer. These results are quite helpful in providing the real-world measure of survival of the cancer patient. The results can be estimated at even diagnosis stages of the patients. It also helps in the calculation of future trends of the life loss of a patient. They might also help the policymakers to identify the disease burden and motivate the decision-making efforts to balance socio-demographic inequalities. The consistency criteria have prominently relied upon the variables that affect life expectancy, prediction models. Prior works of literature have shown the introduction of more realistic and significant variables can improve the performance of models.

VI. LIMITATIONS & FUTURE SCOPE

Availability of all health-related data, education, and economic expenditure stats have made possible the proper and error-free estimation of life expectancy models. Earlier life expectancy models were dependent on very few variables and due to the unavailability of advanced data exploratory and validation techniques, the trained model was not so much accurate. Now, in recent studies, various milestones have been achieved in this field. For example, the ensemble of different base models, which includes the results of individual base models, the inclusion of ANN, and RNN techniques in solving such tasks. So, the accuracy and addition of decisive variables into the final life expectancy model will be far most two significant areas of concern in the further research part. The inclusion of newly suggested variables from many kinds of research, such as weather-related trends and effects of natural disasters is still a matter of concern and debate in recent times. However, it will be great a challenge for other authors to access this challenge into the dataset of life expectancy. But, even having such advancements in technology and data science forecasting accompanying such uncertainty into the dataset is still quite not achieved yet.

This paper focuses on presenting the current scenario in this field and tries to propose a generic solution to test the life expectancy models on selected datasets. Validation and accuracy of the trained models are to be verified by the exploitation of numerous model validation techniques and finally, the effect of various indicators will come into the feature. The main aim of future work in this field will be the optimization of results with the inclusion of a few more variables by not affecting the overall performance, complexity, and accuracy of the trained model.

In future authors are planning to explore methods for gaining more insight in the nature of the patterns that are detected by neural networks, as well as making the determinants of a certain prediction transparent.[20]

VII. CONCLUSION

Initially, authors have dropped features such as year, country, and status. The main aim was to analyze the impact of features on the outcome and how it varies. The first task was to find the best-performing model. Among different models, random forest performs best with an MAE of 1.27 and an R^2 score of 96% on the test set. Adult mortality, HIV/AIDS, schooling, and BMI are the most impacting factors on life expectancy among the features. Schooling, Income Composition, and BMI have positively correlated to the outcome. Surprising thing was that some features such as GDP, total expenditure, and infant deaths were not that impactful on the final result. But the initial assumption is proven wrong here about these features.[21]

These results clearly show and prove the importance of health, education, and economic features on Life expectancy. But there is still some room for improvement by including the other features such as environmental and geographical features. The inclusion and dependency of these suggested features on life expectancy is still a matter of debate and a future part of research in this particular domain.

REFERENCES

[1] Noorhannah Boodhun, Manoj Jayabalan, "Risk prediction in life insurance industry using supervised learning algorithms," Complex & Intelligent Systems, vol. 4, no. 2, pp. 145-154, 2018.

[2] Mahumud, R.A., Hossain, G., Hossain, R., Islam, N. and Rawal, L.,;, "Impact of Life Expectancy on Economic Growth and Health Care Expenditures in Bangladesh.," Universal Journal of Public Health, vol. 1, no. 4, pp. 180-186, 2013.

[3] Bhosale, A.A. and Sundaram, K.K., "Life prediction equation for human beings," International Conference on Bioinformatics and Biomedical Technology, vol. IEEE, pp. 266-268, 2019.

[4] Aggarwal, D., Mittal, S., Bali V., "Prediction Model for Classifying Students Based on Performance using Machine Learning Techniques," International Journal of Recent Technology and Engineering, vol.8 p.2S7, 496-503, 2019.

[5] Chen, Tianqi; Guestrin, Carlos, "XGBoost: A Scalable Tree Boosting System," Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794, 2016.

[6] Aggarwal, D., Mittal, S. and Bali, V., "Significance of Non-Academic Parameters for Predicting Student Performance Using Ensemble Learning Techniques", International Journal of System Dynamics Applications (IJSDA), Vol. 10, Issue 3, Article 3, pp. 38-49, 2020.

[7] Kerdprasop, N. and Foreman, K. J., "Association of economic and environmental factors to life expectancy of people in the Mekong basin," IEEE Conference on Industrial Electronics and Applications (ICIEA), pp. 1984-1989, 2017.

[8] Beekshma., "A neural-network analyzer for mortality forecast," ASTIN Bulletin: The Journal of the IAA, vol. 48, no. 2, pp. 481-508, 2018.

[9] Deb, C., Zhang, F., Yang, J., Lee, S.E. and Shah, K.W., "A review on time series forecasting techniques for building energy consumption," Renewable and Sustainable Energy Reviews, vol. 74, pp. 902-924, 2017.

[10] Sindhwani, N., Verma, S., Bajaj, T., & Anand, R. (2021). Comparative Analysis of Intelligent Driving and Safety Assistance Systems Using YOLO and SSD Model of Deep Learning. International Journal of Information System Modeling and Design (IJISMD), 12(1), 131-146

[11] Aggarwal, D., Bali, V., Agarwal, A., Poswal, K., Gupta, M., Gupta, A. "Sentiment Analysis of Tweets Using Supervised Machine Learning Techniques Based on Term Frequency," Journal of Information Technology Management, vol.13 no.1,pp. 119-141, 2021.

[12] M. R. Hebb, "The Organization of Behaviour," New York, Wiley,, p. 437,1949.

[13] Rosenblatt, F., "The perceptron: a probabilistic model for information storage and organization in the brain," Psychological review, vol. 65, no. 6, p. 386, 1958.

[14] Sormin, M.K.Z., Sihombing, P., Amalia, A., Wanto, A., Hartama, D. and Chan, D.M., "Predictions of World Population Life Expectancy Using Cyclical Order Weight/Bias," Physics: Conference Series (IOP Publishing), vol. 1255, no. 1, p. 012017, 2019.

[15] Nath, B., Dhakre, D.S. and Bhattacharya, D., "Forecasting wheat production in India: An ARIMA modelling approach," Journal of Pharmacognosy and Phytochemistry, vol. 8, no. 1, pp. 2158-2165, 2019.

[16] Bali, V. , Kumar, A. and Gangwar, S., "A Novel Approach for Wind Speed Forecasting Using LSTM-ARIMA Deep Learning Models", International Journal of Agricultural and Environmental Information Systems (IJAEIS), Volume 11, Issue 3, pp. 13-30, ISSN: 1947-3192, EISSN: 1947-3206, 2020.

[17] Box, G.E., Jenkins, G.M., Reinsel, G.C. and Ljung, G.M., Time series analysis: forecasting and control, John Wiley & Sons., 2015.

[18] Foreman, K.J., Marquez, N., Dolgert, A., Fukutaki, K., Fullman, N., McGaughey, M., Pletcher, M.A., Smith, A.E., Tang, K., Yuan, C.W. and Brown, J.C., "Forecasting life expectancy, years of life lost, and all- cause and cause-specific mortality for 250 causes of death: reference and alternative scenarios for 2016–40 for 195 countries and territories," The Lancet, vol. 392, no. 10159, pp. 2052- 2090, 2018.

[19] Mathias, J.S., Agrawal, A., Feinglass , Cooper, A.J., Baker, D.W. and Choudhary, A., "Development of a 5-year life expectancy index in

older adults using predictive mining of electronic health record data," Journal of the American Medical Informatics Association, vol. 20, no. e1, pp. 118- 124, 2013.

[20] Kamalraj, R., Neelakandan, S., Kumar, M. R., Rao, V. C. S., Anand, R. & Singh, H. (2021). INTERPRETABLE FILTER BASED CONVOLUTIONAL NEURAL NETWORK (IF-CNN) FOR GLUCOSE PREDICTION AND CLASSIFICATION USING PD-SS ALGORITHM. Measurement, 109804.

[21] Verberne, S., van den Bosch, A., Das, E., Hendrickx, I. and Groenewoud, S., "Predicting life expectancy with a long short-term memory recurrent neural network using electronic medical records," BMC medical informatics and decision making, vol. 19, no. 1, p. 36, 2019.