IJMTST is UGC Approved *Journal with Journal Id: 43137*

International Journal for Modern Trends in Science and Technology, 8(S08): 114-119, 2022

ISSN: 2455-3778 online
DOI: https://doi.org/10.46501/IJMTST08S0821
Available online at: https://www.ijmtst.com/vol8si08.html

Life Expectancy Prediction using Machine Learning

Yallamati Prakasarao[1] | Arumalla Nagaraju[2]

[1] CSE, Hyderabad Institute of Technology and Management, Telangana, India.
[2]CSIT, Chalapathi Institute of Engineering and Technology, Guntur, Andhra Pradesh, India.

To Cite this Article
Yallamati Prakasarao and Arumalla Nagaraju. Life Expectancy Prediction using Machine Learning. International Journal for Modern Trends in Science and Technology 2022, 8(S08), pp. 114-119. https://doi.org/10.46501/IJMTST08S0821

Article Info
Received: 26 May 2022; Accepted: 24 June 2022; Published: 28 June 2022.

ABSTRACT

Life span depends on various features like adult mortality, percentage expenditure, alcohol consumption rate. Along with the prognostication of longevity, we also puzzle out how much impact a particular area has with respect to chronic diseases. life expectancy of the people have direct impact on the discussed factors. We study both economical and biological aspects of countries to foresee the expectation of life. To predict life expectancy, we use simple linear regression, KNN, Decision tree and Random Forest algorithms. By comparing these machine learning algorithms, we can understand which among them is more accurate to predict life expectancy.

KEYWORDS: *Machine Learning, Life Expectancy, Regression, Random Forest, Decision Tree.*

1. INTRODUCTION

Human an incredible creation of god. Every creature in the world has a limited life span, to achieve something in the world. We have a limited life span to survive in the current world. To preserve our self from the consequences, even though lot of inventions has been made by human, to prevent from diseases is a major question mark.

Life span prediction has a greater impact in our modern society because of our food habits, different types of diseases and environmental conditions. Investigations about the life span of vertebrates have been made, except the human (HOMO SAPIENS). It is an emerging research area that is gaining interest but involved lot of challenges due to the limited amount of resources (i.e., datasets) available.

In our proposed system the life span of human is predicted by analysis of human. By obtaining the Environmental factors, Food habits, Diseases and Medical history, a lot of investigations will be conducted to predict the sustainability of human. By the machine learning algorithms and data analytics, We can prognosticate and examine the life span of the individual human being and we can use different classification algorithms for this prediction to accomplish higher accuracy.

2. LITERATURE SURVEY

Ayshwaryaa N et al,proposed that Human an incredible creation of god. Every creature in the world has a limited life span, to achieve something in the world.. To preserve our self from the consequences, even though lot of inventions has been made by human, to prevent from diseases is a major question mark. Life span prediction hasa greater impact in our modern society because of our food habits, different types of diseases and environmental conditions.[1].

Linda Mary et al, proposed that the correlation between attributes like diseases, gender, ages and environmental factor are important. In this paper, In order to find or predict the human lifespan with more accuracy we use random forest algorithm.[2].

V.M Shkolnikov et al, proposed that Predicting life span for human being is a vital step. It is an emerging research area that is gaining interest but involved lot of challenges due to the limited number of resources (i.e., datasets) available. By obtaining the Date of birth, Environmental factors, Food habits, Diseases and Medical history, a lot of investigationswill be conducted to predict the sustainability of human.[3].

D.F.Andrews et al, proposed that when there is change in small fraction the data techniques will be resistant. Otherwise, when the efficiency of stastics held high then the techniques will be robust. If the accuracy score is excellent then the result of the predicted one is accurate.[4].

D.M.J Naimark et al, proposed that the expectancy of the life can be grasped to equal to area under a certain region He proposed it is necessary to understand the baseline risk under the control group. By the help of different models wecan predict the life expectancy. [5].

A.A. Bhosale et al, proposed that expectancy of the life mainly target on predicting models using trends. He proposed life expectancy rely on weight, adult mortality, heart rate, respiration rate for human beings. The inspection provides the standard life expectancy is forecasted by variables that can be easily calculated.[6].

M.K.Z. Sormin et al, proposed to rough calculate the life expectancy of the population across the world so that it will be helpful to the particular country to increase their health of the human beings. The Cyclic Order Weight neural network method is used for the appraise.[7].

K.J.Preacher et al, proposed that slopes, significance and bands of confidence are used to test the steps in multiple linear regression but this trend has been outdated and extended to multi level linear regression. When one dependent value is depended on more independent values then we use multiple linear regression.[8]

3. PROPOSED SYSTEM

The system begins with installation of anaconda software. This process is followed by launching Jupyter notebook which helps to import the certain necessary packages i.e., pandas, NumPy, sklearn etc. After importing all the packages, various machine learning is implemented for identifying an algorithm with high accuracy. In this proposed system, we analyzed the lifespan among human beings based on some of the health and environmental factors. In this work, we also analyze the life expectation of individual people. The lifespan expectancy of each and every human being was analyzed with the help of given data and shown as a result. In our proposed system we are using Random Forest, Decision tree, KNN, Gradient Boosting Algorithms. In our proposed system. Finally, we obtain a better accuracy with the help of random forest algorithms through which better result will be obtained comparatively with other algorithms.

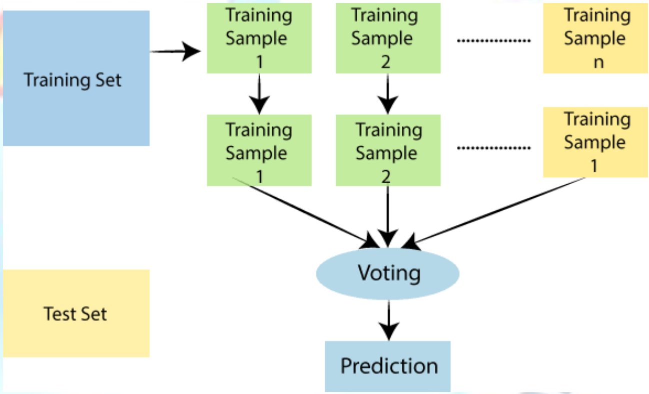


Fig. 1: Random Forest

Random Forest draws many decision trees from our given dataset and it finally combines all the outputs of them into one. Like, If it is used for classification problems , the final result is obtained by taking the majority of the results produced by all the decision trees built by the model. And if it is used for Regression the we take the average value of all the results of the decision tree.

Random forest is more accurate in its prediction than Decision tree because we know that every decision tree have high variance, in random forest we actually combine all the decision trees together so then the final resultant variance is low. Thus in random forest the output depends on many decision trees.

Here in our prediction of life span, when the dataset is given to the random forest regressor , it actually splits the given dataset so that each decision tree gets its unique dataset. And this decision trees then compute their results and finally the average of all the decision trees is taken as our final result.

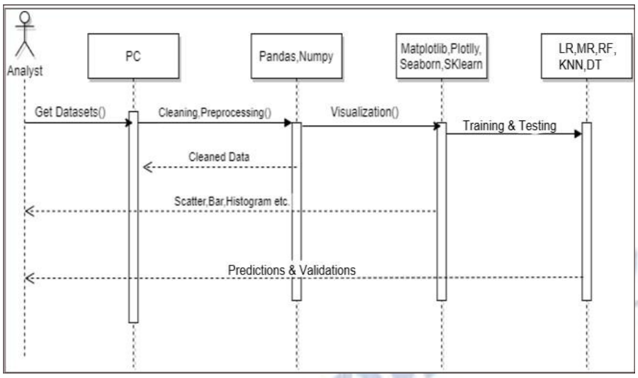


Fig. 2: Sequence Diagram

Fig. 2 represents the sequence flow diagram of the project. Initially, the analyst/user provide the dataset to the PC. By Using libraries like pandas, Numpy we will clean and preprocess the data, the null values will be removed. For the visualisation we use Matplotlib, Plotly, Seaborn, SKlearn, and demonstrate the visualisations by Scatter Bar, Histogram. After visualising the data is splitted to training and testing, training data is used to train the different models and we test the models with the help of testing data and get the finest fulfill model out of all the models.

4. RESULTS AND ANALYSIS

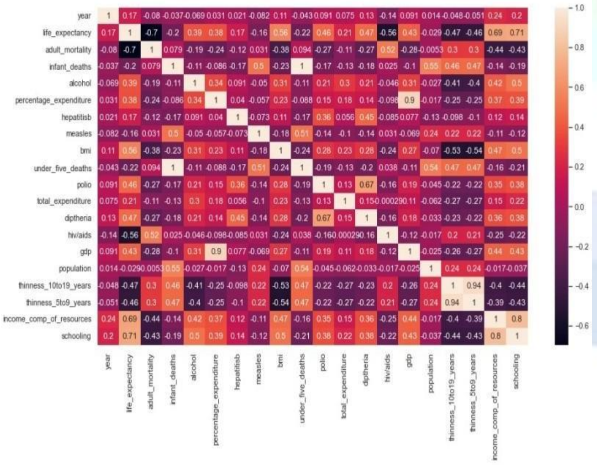
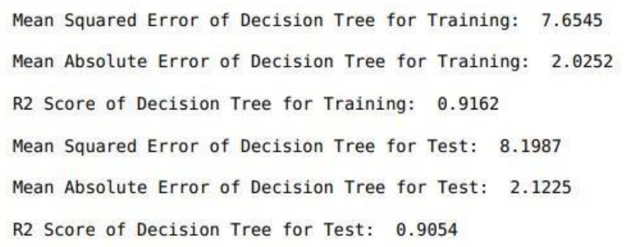


FIG. 3: HEATMAP

Fig. 3 represents the heat map to find out which variables in out dataset has high impact in deciding the life expectancy. So here we have different shades of the same colour where the darker shape implies that it has high impact in predicting the resultant variable then the others. So with heat map we find out all the important variable that have high impact in predicting our final output. So we can see that mortality of the adults, alcohol, percentage cost, hepatitis, measles, bmi, under five deaths, polio, total expenditure, diptheria, hiv, population, schooling have a strong impact in predicting our final resultant variable.



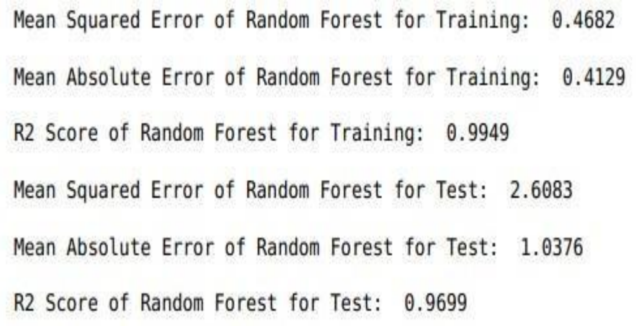
```
Mean Squared Error of Decision Tree for Training:  7.6545

Mean Absolute Error of Decision Tree for Training:  2.0252

R2 Score of Decision Tree for Training:  0.9162

Mean Squared Error of Decision Tree for Test:  8.1987

Mean Absolute Error of Decision Tree for Test:  2.1225

R2 Score of Decision Tree for Test:  0.9054
```

Fig. 4: Evaluation Metrics for Decision tree

If we use decision tree as our model for our prediction purpose, the performance metrics obtained through it are i.e, the mean squared error ,mean absolute error & coefficient of determination value for training dataset is 7.6, 2.0,0.91. Similarly the same values for the test dataset are 8.1,2.1,0.90 .These values can be improved if we use the Random Forest as our model as below.

If we use random forest as our model for our prediction purpose , the performance metrics obtained through it are i.e, the mean squared error , mean absolute error & coefficient of determination value for training dataset is 0.46, 0.41,0.99 .Similarly the same values for the test dataset are 2.6,1.03,0.96. Thus random forest is more and much accurate in predicting our final resultant variable i.e, life expectancy We use the cross validation technique to verify that our built models work accurately to the real world data. Here we have used the k-fold cross validation ,from the obtained results we see that the random forest has more cross value score then the other models. Also a model is much accurate if it has the less standard deviation, so here the standard

Random forest as our model for our prediction



```
Mean Squared Error of Random Forest for Training:  0.4682

Mean Absolute Error of Random Forest for Training:  0.4129

R2 Score of Random Forest for Training:  0.9949

Mean Squared Error of Random Forest for Test:  2.6083

Mean Absolute Error of Random Forest for Test:  1.0376

R2 Score of Random Forest for Test:  0.9699
```

Fig. 5: Evaluation Metrics for Random forest

Deviation of random forest is low then others thus it is more accurate in prediction the outputs.

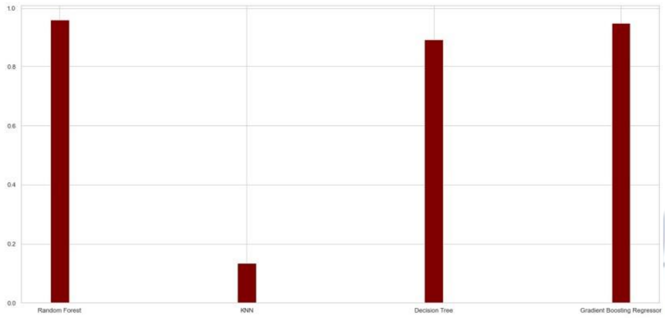


Fig. 6: Comparisons of Algorithms

In the Fig.6 the bar chart depicts the comparisons of accuracies of various machine learning models used in our project. And we observe that the Random Forest is more accurate than other predictive models.

Our next task is to build a Decision Tree by using the bootstrapped data set created in the previous step. Since we're making a Random Forest we will not consider the entire data set that we created, instead we'll only use a random subset of variables at each step.

Let's say we selected Blood Flow and Blocked arteries. Out of these 2 variables, we must now select the variable that best separates the samples. For the sake of this example, let's say that Blocked Arteries is a more significant predictor and thus assign it as the root node.Our next step is to repeat the same process for each of the upcoming branch nodes. Here, we again select two variables at random as candidates for the branch node and then choose a variable that best separates the samples.

Random Forest is a collection of Decision Trees. Each Decision Tree predicts the output class based on the respective predictor variables used in that tree. Finally, the outcome of all the Decision Trees in a Random Forest is recorded and the class with the majority votes is computed as the output class. Thus, we must now create more decision trees by considering a subset of random predictor variables at each step. To do this, go back to step 1, create a new bootstrapped data set and then build a Decision Tree by considering only a subset of variables at each step.Now that we've created a random forest, let's see how it can be used to predict whether a new patient has heart disease or not.The below diagram has the data about the new patient. All we have to do is run this data down the decision trees that we made.The first tree shows that the patient has heart disease, so we keep a track of that in a table as Similarly, we run this data down the other decision trees and keep a track of the class predicted by each tree. After running the data down all the trees in the Random Forest, we check which class got the majority votes. In our case, the class 'Yes' received the most number of votes, hence it's clear that the new patient has heart disease.To conclude, we bootstrapped the data and used the aggregate from all the trees to make a decision, this process is known as Bagging.

Our final step is to evaluate the Random Forest model. Earlier while we created the bootstrapped data set, we left out one entry/sample since we duplicated another sample. In a real-world problem, about 1/3rd of the original data set is not included in the bootstrapped data set.

5. PRACTICAL IMPLEMENTATION OF RANDOM FOREST

In R Even people living under a rock would've heard of a movie called Titanic. But how many of you know that the movie is based on a real event? Kaggle assembled a data set containing data on who survived and who died on the Titanic. To build a Random Forest model that can study the characteristics of an individual who was on the Titanic and predict the likelihood that they would have survived.

Data Set Description:

There are several variables/features in the data set for each person:pclass: passenger class (1st, 2nd, or 3rd)
Sex
age
sibsp: number of Siblings/Spouses Aboard
parch: number of Parents/Children Aboard
fare: how much the passenger paid
embarked: where they got on the boat (C = Cherbourg; Q = Queenstown; S = Southampton)

We'll be running the below code snippets in R by using RStudio, so go ahead and open up RStudio. For this demo, you need to install the caret package and the randomForest package.

```
install.packages("caret", dependencies = TRUE)
install.packages("randomForest")
```

Next step is to load the packages into the working environment.

```
library(caret)
library(randomForest)
```

It's time to load the data, we will use the read.table function to do this. Make sure you mention the path to the files (train.csv and test.csv) train<read.table('C:/Users/zulaikha/Desktop/titanic/train.csv', sep=",", header= TRUE) The above command reads in the file "train.csv", using the delimiter ",", (which shows that the file is a CSV file) including the header row as the column names, and assigns it to the R object train

The main difference between the training set and the test set is that the training set is labeled, but the test set is unlabeled. The train set obviously doesn't have a column called "Survived" because we have to predict that for each person who boarded the titanic. Before we get any further, the most essential factor while building a model is, picking the best features to use in the model. It's never about picking the best algorithm or using the most sophisticated R package. Now, a "feature" is just a variable. o, this brings us to the question, how do we pick the most significant variables to use? The easy way is to use cross-tabs and conditional box plots.

Cross-tabs represent relations between two variables in an understandable manner. In accordance to our problem, we want to know which variables are the best predictors for "Survived".

Let's look at the cross-tabs between "Survived" and each other variable. In R, we use the table function: table(train[,c('Survived', 'Pclass')])

```
 Pclass Survived   1   2   3
                 0  80  97 372
                 1 136  87 119
```

Estimation of the life limit can support current theories on aging which presume the existence of the biological limit for human life. According to our estimates, it is foreseeable that many countries will face growing problems of aging populations, age-related diseases and health care costs. The increase in human longevity will accelerate the population growth rate and probably the constant increaseinthe maximum duration reached or the life expectancy will decrease in the life expectancy in the next half century. Upcoming trends can cause an ethical problem in the fair distribution of health care resources. Aging research requires new approaches to discovering the complex biology of aging. Life expectancy is a statistical measure of the average time a human being is expected to live, Life expectancy depends on various factors: Regional variations, Economic Circumstances, Sex Differences, Mental Illnesses, Physical Illnesses, Education, Year of their birth and other demographic factors. This problem statement provides a way to predict average life expectancy of people living in a country when various factors such as year, GDP, education, alcohol intake of people in the country, expenditure on healthcare system and some specific disease related deaths that happened in the country are given.

6. LIMITATIONS & FUTURE SCOPE

Availability of all health-related data, education, and economic expenditure stats has 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. The determinants of a certain prediction transparent.

7. CONCLUSION AND FUTURE WORK

In this paper a system of the human lifespan can be predicted earlier. By employing data through datasets, the correlation between attributes like diseases, gender, ages and environmental factor are monitored. The Random Forest algorithm is achieved in order to forecast the human lifespan with more precise. The advantage of Random Forest algorithm, gives more flexibility without obtaining the processed data and accurate. Thus, We have analyzed the lifespan among human beings based on some of the health and environmental factors. By prognosticate the human lifespan with dissimilar models Random Forest algorithm gives more precise

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 Furthermore, the future enhancement can be made by using deep learning algorithm which may give better solution.

Conflict of interest statement

Authors declare that they do not have any conflict of interest.

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 BiomedicalTechnology, 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,"ASTINBulletin: The Journal of the IAA, vol. 48, no. 2, pp. 481-508, 2018

[9] 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.

[10] 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

[11] Mathias, J.S., Agrawal, A., Feinglass , Cooper, A.J., Baker, D.W. andChoudhary, 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

[12] 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

[13] 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.