International Journal on Applications in Information and Communication Engineering
Volume 8 : Issue 2 : June 2022, pp 8 – 11 www.aetsjournal.com ISSN (Online) : 2394 - 6237

LIFE EXPECTANCY PREDICTION USING MACHINE LEARNING

SIDDHARDHA T

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 peoplehave direct impact on the discussed factors. We study both economical and biological aspects of countries to foresee the expectation of life. To prognosticating life expectancy, we useRandom Forest, K-Nearest Neighbour, Decision Tree, Linear regression algorithms. By comparingthese machine learning algorithms, we can understand which among them is more accurate to predict life expectancy.

Keywords— Random Forest, Regression, Decision Tree, Life Expectancy, Machine Learning.

I. INTRODUCTION

Humanan incredible creation of god. Every creature in the world has a limited life span, to achieve something in the world. Wehave 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.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 thelife 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 examinethe life span of the individual human being and we can use different classification algorithms for this prediction to accomplish higher accuracy.

Siddhardha T , Department of CSE , Madanapalle Institute of Technology & Scienc e, Madanapalle , A.P., INDIA.
(Email : siddhardhasiddu74905@gmail.com)

II. 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 has a 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 accuracywe 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 investigations will 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 stasticsheld 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 thatby the help of different models we can forecast the life expectancy.[5].

A.A. Bhosale et al, proposed thatexpectancy 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. created a way for assessing the population's life expectancy around 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. This pattern has been modernized and prolonged to multi level linear regression in multiple linear regression. We deploy multiple linear regression when one reliant value is dependent on numerous independent ones.[8].

III. 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 analyzedwith the help of given data and shown as a result.We integrate Gradient Boosting, Decision Trees, Random Forests, and K-Nearest Neighbor Algorithms in our suggested system.. 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.

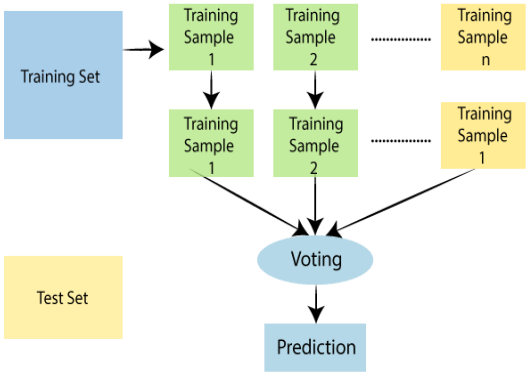

Figure 1: Random Forest

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. As a result, the Random Forest's results are responsible for a huge number of 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.

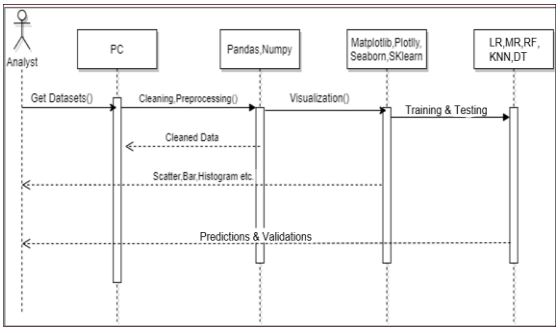

Figure 2:Sequence Diagram

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, Numpywe

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 resolving the data, training and testing data are separated, with training data being leveraged to train the multiple modelsand we test the models with the help of testing data and get the finestfulfill model out of all the models.

IV. RESULTS AND ANALYSIS

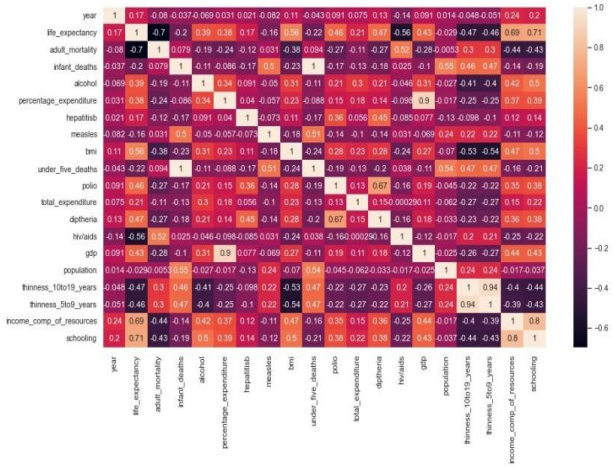

Figure 3:Heatmap

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
```

Figure 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, For the training dataset mean squared error, the mean absolute error and coefficient of determination are 7.6, 2.0, and 0.91. Similarly, the test dataset's values are 8.1,2.1,0.90..These values can be improved if we use the Random Forest as our model as below.

```
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
```

Figure 5: Evaluation Metrics for Random forest

If we use random forest as our model for our prediction purpose , the performance metrics obtained through it are i.e,For the training dataset, the mean squared error, mean absolute error, and coefficient of determination values are 0.46, 0.41, and 0.99, respectively.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.

	CV R2 Mean	Std
Random Forest	0.961226	0.007889
KNN	0.133546	0.056774
Decision Tree	0.892720	0.015107
Gradient Boosting Regressor	0.949733	0.007323

Figure 6:Cross Validation

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.If a model has a low standard deviation, it is far more accurate, so here the standard deviation of random forest is low then others thus it is more accurate in prediction the outputs.

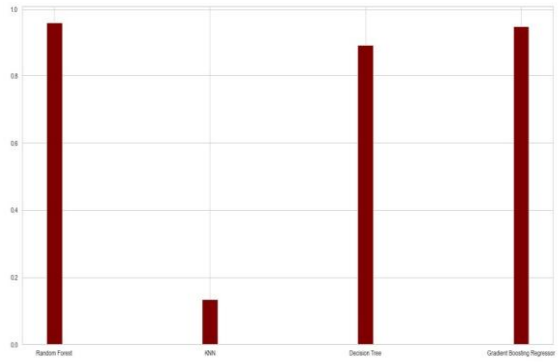

Figure 7:Comparisons of Algorithms

In the Fig.6 the bar chart depicts the comparisons of accuracies of various machine learning models used in our project. The Random Forest model seems to have the highest accuracy of all the forecasting models.

V. 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. Byprognosticate the human lifespan with dissimilar models Random Forest algorithm gives more precise.

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

REFERENCES

[1] Ayshwaryaa N , Kavipriya R , Sobika K , Prof. T. Karthikeyan 1,UG Scholar (Feb 2020) Human Life Span Prediction using Machine Learning, Department of Computer Science and Engineering, Knowledge Institute of Technology, Salem, Tamil nadu, India

[2] Linda Mary, John Ashima Sharma, SiddhantGujarathi (Feb 2019):Detector and Predictor System for lifeaspan using Naives Bayes and Decision Tree Algorithm.

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

[4] D. F. Andrews.(Nov 2018): "A Robust Method for Multiple Linear Regression," Technometrics, vol. 16, no. 4, pp. 523–531.

[5] D.M.J.Naimark. (Aug 2018):Life Expectancy Measurements, in International Encyclopedia of Public Health, H. K. (Kris) Heggenhougen, Ed. Oxford: Academic Press, pp. 83–98.

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

[7] R.SenthamilSelvan '' Performance of FPGA in an Enhanced Level of Watchdog Timer'' on JETIR July 2019, Volume 6, Issue 6, ISSN: 2349-5162.

[8] M.K.Z. Sormin, P. Sihombing, A. Amalia, A. Wanto, D. Hartama and D.M. Chan, "Predictions of World Population Life expectancy using weight/Bias", *Physics: Conference Series (IOP Publishing)*, vol. 1255, no. 1, pp. 012017, 2019.

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