JES
Journal of Engineering Sciences

Vol 13, Issue 03, MARCH /2022

ISSN NO:0377-9254

LIFE EXPECTANCY: PREDICTION & ANALYSIS USING ML

V Y R RAMA RAYALU

Vyr.ramarayalu@gmail.com

Amarneni Eashwar sai

bannuchowdary27@gmail.com

SAI HEMANTH BABU SUNKARI

hemanth.sunkari@gmail.com

Bhargav kandala

bkandala@gitam.edu

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.*

1. 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 premodern 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.

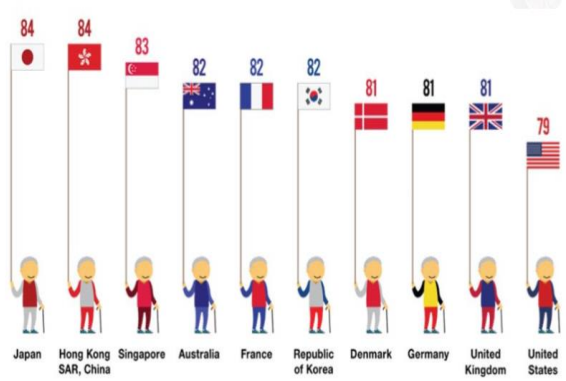

Fig.1: Example figure

However, there were still many countries with less life expectancy than the rest of the world in the early1900s. 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]. 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-threeyears, 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.

2. LITERATURE REVIEW

2.1 Risk prediction in life insurance industry using supervised learning algorithms [1] :

Risk assessment is a crucial element in the life insurance business to classify the applicants. Companies perform underwriting process to make decisions on applications and to price policies accordingly. With the increase in the amount of data and advances in data analytics, the underwriting process can be automated for faster processing of applications. This research aims at providing solutions to enhance risk assessment among life insurance firms using predictive analytics. The real world dataset with over hundred attributes (anonymized) has been used to conduct the analysis. The dimensionality reduction has been performed to choose prominent attributes that can improve the prediction power of the models. The data dimension has been reduced by feature selection techniques and feature extraction namely, Correlation-Based Feature Selection (CFS) and Principal Components Analysis (PCA). Machine learning algorithms, namely Multiple Linear Regression, Artificial Neural Network, REPTree and Random Tree classifiers were implemented on the dataset to predict the risk level of applicants. Findings revealed that REPTree algorithm showed the highest performance with the lowest mean absolute error (MAE) value of 1.5285 and lowest root-mean-squared error (RMSE) value of 2.027 for the CFS method, whereas Multiple Linear Regression showed the best performance for the PCA with the lowest MAE and RMSE values of 1.6396 and 2.0659, respectively, as compared to the other models..

2.2 Impact of Life Expectancy on Economic Growth and Health Care Expenditures in Bangladesh [2]:

Life expectancy is one of the major key indicators of population health and economic development of a country. The main objective of this study was to determine the impact of life expectancy on changes of economic growth and health care expenditure. We also examined trend of life expectancy according to the sex difference. We used multiple regression models to estimate the impact of life expectancy on economic growth and health care expenditure. Elasticity of life expectancy on health care expenditure and economic growth is also estimated. Results show greater life expectancy of females compared with the males over the past 15 years. The higher Gross Domestic Product (GDP) per capita was observed in a longer life expectancy. i.e., one US Dollar (USD) increment in GDP per capita will increase in an average of life expectancy by 33 days. Similarly, increased one unit of per person Health Expenditure Per Capita (HEPC) will increase the life expectancy in an average of 8 days in a year. The higher proportion of total expenditure on health as a percentage of GDP and direct personal expenditure on health by household as a share of private expenditure on health results in also longer life span. We conclude that the increased life expectancy has direct impact on increased per capita real income and higher expenditure on health. This study has some policy implications for Bangladesh, in particular the needs for increased per capita real income and planning for future health and population policies/programs. Therefore, political stability, adequate and suitable social sector policies and government interventions are required to increase life expectancy and economic growth in the country. There is also a need for involvement of health human force in macro and micro policy-makings and critically examine other determinants of health care expenditure..

2.3 Life prediction equation for human beings [3] :

Research on life expectancy focuses on building forecasting models using mortality trends or constructing parameter life expectancy models from sampling populations. Proposed herewith is a model of life expectancy, using empirical analysis, of weight respiration rate, heart rate, & blood pressure for human beings.. The model variants suggest robustly that proxy for technology, education, and healthcare all have a significant and positive effect on average life expectancy. This analysis provides information of use to governments & insurance companies particularly in the developing world, since

average life expectancy is predicted by variables that can be easily measured.

2.4 Prediction Model for Classifying Students Based on Performance using Machine Learning Techniques [4]:

In today's competitive world of educational organizations, the universities and colleges are using various data mining tools and techniques to improve the students' performance. Now a days, when the number of drop out students is increasing every year, if we get to know the probability of a student whether he/she will be able to cope up easily with the course, it is possible to take some preventive actions beforehand. In other words, if we get to know that a student will clear his papers in the course or he will have reappear in papers, a teacher/parent can focus more on such students. The data set of students has been taken from the UCI Machine Learning repository where a sample of 131 students have been provided with twenty-two attributes. The results of six classification algorithms have been compared in order to predict the most appropriate model for classifying whether a student will have a reappear in a course or not.

2.5 XGBoost: A Scalable Tree Boosting System [5]:

Tree boosting is a highly effective and widely used machine learning method. In this paper, we describe a scalable end-to-end tree boosting system called XGBoost, which is used widely by data scientists to achieve state-of-the-art results on many machine learning challenges. We propose a novel sparsity-aware algorithm for sparse data and weighted quantile sketch for approximate tree learning. More importantly, we provide insights on cache access patterns, data compression and sharding to build a scalable tree boosting system. By combining these insights, XGBoost scales beyond billions of examples using far fewer resources than existing systems.

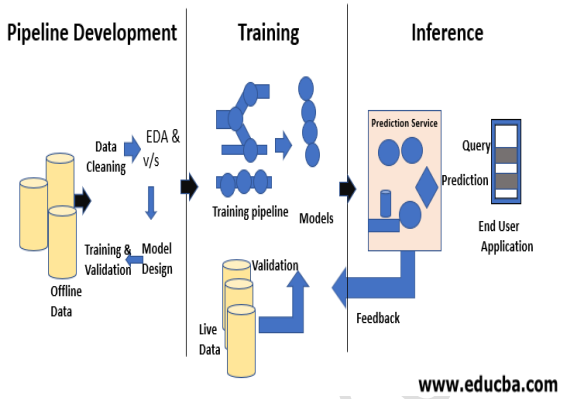

Fig.2: ML model.

2.6 Significance of NonAcademic Parameters for Predicting Student Performance Using Ensemble Learning Techniques [6]:

The academic institutions are focusing more on improving the performance of students using various data mining techniques. Prediction models are designed to predict the performance of students at a very early stage so that preventive measures can be taken beforehand. Various parameters (academic as well as non-academic) are considered to predict the student performance using different classifiers. Normally, academic parameters are given more weightage in predicting the academic performance of a student. This paper compares the two models: one built using academic parameters only and another using both academic and non-academic (demographic) parameters. The primary data set of students has been taken from a technical college in India, which consists of data of 6,807 students containing attributes. Synthetic minority oversampling technique filter is applied to deal with the skewed data set. The models are built using eight classification algorithms that are then compared to find the parameters that help to give the most appropriate model to classify a student based on his performance.

3. IMPLEMENTATION

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 datascience libraries such as NumPy, SciPy, Matplotlib, and IPython and it can be further extended by adding plugins for numerous other purposes.

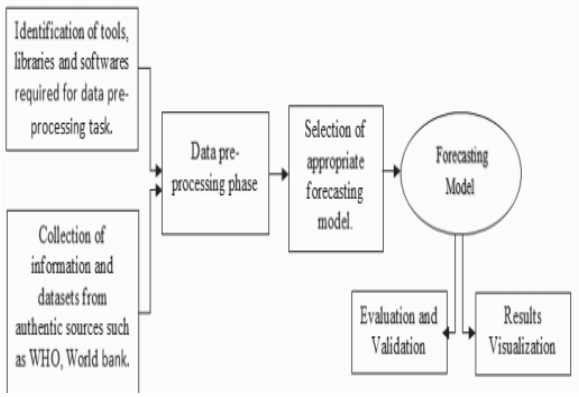

Fig.3: life expectancy prediction using machine learning phases

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.

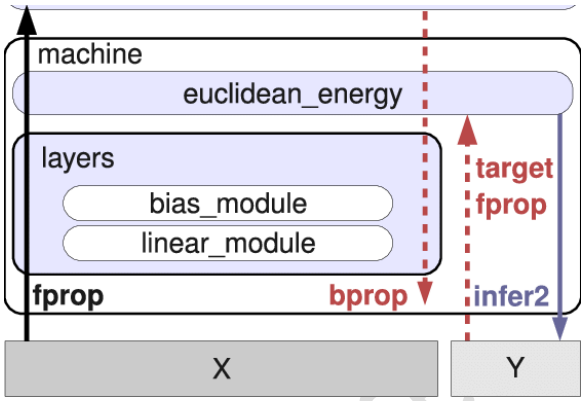

Fig.4: Linear regression model

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.

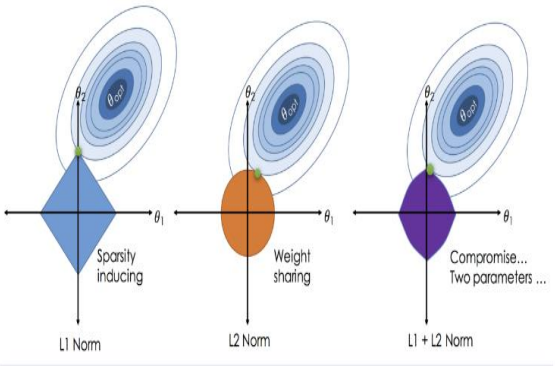

Fig.5: Ridge regression model

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.

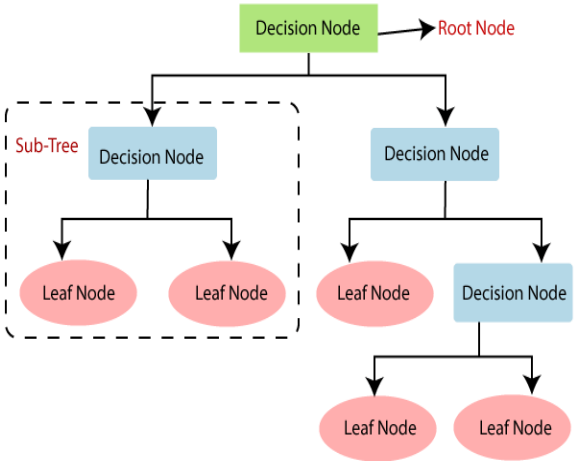

Fig.6: Decision tree model

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.

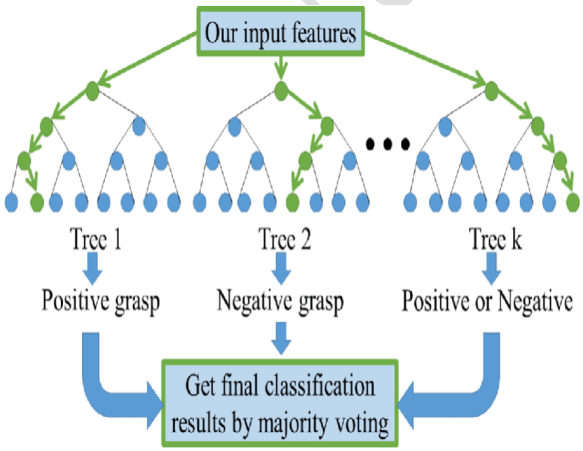

Fig.7: Random forest model

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

Table.1: 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

5. EXPERIMENTAL RESULTS

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.

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.

	infant deaths	Alcohol	Hepatitis B	Measles	BMI	Polio	Diphtheria	HIV/AIDS	GDP	Life expectancy
0	62	0.010000	65.000000	1154	19.1	6.0	65.0	0.1	584.259210	65.0
1	64	0.010000	62.000000	492	18.6	58.0	62.0	0.1	612.696514	59.9
2	66	0.010000	64.000000	430	18.1	62.0	64.0	0.1	631.744976	59.9
3	69	0.010000	67.000000	2787	17.6	67.0	67.0	0.1	669.959000	59.5
4	71	0.010000	68.000000	3013	17.2	68.0	68.0	0.1	63.537231	59.2
5	74	0.010000	66.000000	1989	16.7	66.0	66.0	0.1	553.328940	58.8
6	77	0.010000	63.000000	2861	16.2	63.0	63.0	0.1	445.893298	58.6
7	80	0.030000	64.000000	1599	15.7	64.0	64.0	0.1	373.361116	58.1
8	82	0.020000	63.000000	1141	15.2	63.0	63.0	0.1	369.835796	57.5
9	84	0.030000	64.000000	1990	14.7	58.0	58.0	0.1	272.563770	57.3
10	85	0.020000	66.000000	1296	14.2	58.0	58.0	0.1	25.294130	57.3
11	87	0.020000	67.000000	466	13.8	5.0	5.0	0.1	219.141353	57.0
12	87	0.010000	65.000000	798	13.4	41.0	41.0	0.1	186.728544	56.7
13	88	0.010000	64.000000	2486	13.0	36.0	36.0	0.1	187.845950	56.2
14	88	0.010000	63.000000	8762	12.6	35.0	33.0	0.1	117.496980	55.3

Fig.8: Represents the sample dataset which contains the variables like Email, address, time on web, time on App, length of membership, gender and salary. Those are the independent variables. And 'yearly amount spent' is the dependent variable.

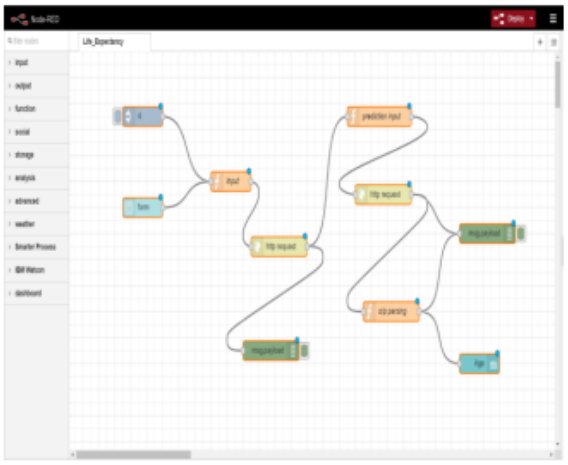

Fig.9: It is the node-red which is used to design the user interface through which an user can enter his/her details whenever required.

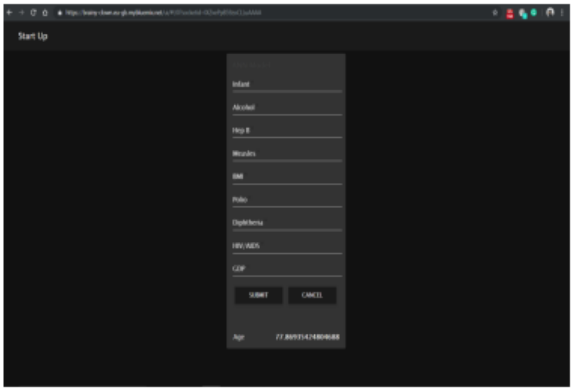

Fig.10: Represent the final user interface

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.

6. 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 R2 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.

7. FUTURE SCOPE

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.

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 NonAcademic 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.