Life Expectancy Prediction through Analysis of Immunization and HDI Factors using Machine Learning Regression Algorithms

https://doi.org/10.3991/ijoe.v18i13.33315

A. Lakshmanarao[1(✉)], Srisaila A.[2], Srinivasa Ravi Kiran T.[3], Lalitha G.[4], Vasanth Kumar K.[5]

[1]Aditya Engineering College, Surampalem, India
[2]Department of Information Technology, V.R. Siddhartha Engineering College, Vijayawada, India
[3]P.B. Siddhartha College of Arts & Science, Vijayawada, India
[4]Adarsh College of Engineering, Chebrolu, India
[5]Malla Reddy Engineering College(A), Hyderabad, Telangana
laxman1216@gmail.com

Abstract—One of the most crucial elements in end-of-life judgment is life expectancy. For example, good forecasting aids in determining the course of therapy and planning for the acquisition of wellness services and infrastructure. Physicians, on the other hand, tend to overestimate life expectancy, missing the window of opportunity to begin a plan of care. This study examines the feasibility of estimating life expectancy from a WHO dataset collected from Kaggle using machine learning techniques. Even though much research has been conducted in the past on factors influencing life expectancy, including demographic factors, economic distribution, and death rates. It was observed that the impact of immunizations on the standard of living was not previously considered. In this paper, we analyzed life expectancy based on various features, including immunization features (Polio, Hepatitis B, Diphtheria, etc..), HDI factors (schooling, GDP, etc.) of various countries for 15 years period. We also proposed machine learning algorithms for the prediction of life expectancy. We applied regression algorithms logistic regression, SVM, Decision Tree, and random forest regression and achieved a good r-squared value with the random forest algorithm.

Keywords—life expectancy, kaggle, WHO, machine learning, python

1 Introduction

People are living longer lifetimes. In reality, in recent decades, life expectancy and the highest measured age at death have both increased significantly. Changing human DNA or restricting food consumption to prolong a healthy life has never been realistic, helpful, or ethical. Human life extension is one of the great challenges, and the greatest achievable life duration in humans is still a hot topic of dispute. Many experts argue that human life has an inherent upper limit, albeit they disagree over whether it is 80, 90, or 120 years old. The extending human life span is thought to be roughly 120 years,

with the number of people dying at an ever-increasing age. Traditional analyses or theoretical models haven't been able to explain this difference in a reasonable way yet. The conventional disease strategy, from the perspective of the overall body system, concentrates on the symptoms of diseases rather than the underlying processes of age-related loss. Family doctors need an accurate forecast of life expectancy to decide when to initiate additional care with a patient, and it's also a big factor in end-of-life choices. The ability of a doctor to recognize people in need of hospice care is strongly reliant on his or her expertise with patients. But it is difficult for doctors also to predict the life expectancy. The equitable distribution of health-care services may face an ethical challenge as a result of emerging developments. New techniques to understand the complex biology of aging are needed in ageing research. Despite the fact that health files are widely accessible as electronic data, they are underused in the creation of clinical decision-making.

2 Literature review

A. S. Desuky [1] et al. proposed a methodology for life expectancy after thoracic surgery. They applied J48 and naïve bayes algorithms for life expectancy and achieved best results with J48 algorithm. In [2], authors proposed ML methods for prediction of life expectancy in mouse. Random Forest algorithm given good accuracy with their elected datasets. Kasichainula Vydehi [3] et al. applied various regression techniques for life expectancy prediction and achieved good results. ML regressors and classifier methods for forecasting life expectancy were proposed in [4]. They used approaches like MLR and Random Forest regression to get good results. Alex Zhavoronkov [5] et al. used artificial intelligence techniques for aging research. Reinforcement Learning and GANs applied for finding the aging of human. Parikh RB [6] et al. studied 26,500 oncology patients for prediction of morality. The findings of their experiment are "within 6 months of the index encounter, 1,064 (4.0%) of the 26,500 patients in the study died. The average age of those who died was 67. (94% confidence interval: 66–68) years, with 502 (46.0%) of them being women". V. Bali [7] et al. proposed ML techniques for the analysis and prediction of life expectancy based on several factors influencing human life.

A. A. Bhosale [8] et al. energy circle model for life prediction in human lives. They observed that weight has a direct relationship with life expectancy. Heart rate and breathing rate are negatively proportional to life expectancy. In [9], the authors applied various machine learning algorithms for analyzing covid death cases worldwide. Nataliya Boyko [10] et al. investigated ML classifiers and regressors for life expectancy prediction. They applied linear regression with L1 and L2 regularization method and achieved good results. In [11], the authors proposed approach for analysis of personals Life Expectancy through smart devices. Using Monte Carlo simulation from the national community chart, Leng C.H [12] et al. offered a linear regression model to the logit-transformed lifespan ratio between the schizophrenia cohort and the sex-, age-matched referents. In [13], authors used computerized healthcare information to apply a deep learning model with LSTM. They demonstrated that using NLP techniques, they were able to improve the accuracy of their model for forecasting life-expectancy.

The authors achieved better recall and precision. In [14], authors applied ML techniques for cancer detection and life prediction. I. Taylor [15] et al. presented factors effecting quality of life. From their studies, people can improve life expectancy based on human habits.

3 Proposed methodology

Figure 1 shows proposed framework. Initially, we collected Kaggle dataset [16] with multiple features. All the features of the dataset are analyzed for finding the best features which are effecting the life expectancy. Later, we applied four machine learning algorithms for life expectancy prediction. We applied linear regression, random forest, decision tree and SVM for life expectancy prediction. After that, identified the best algorithm and predicted life expectancy.

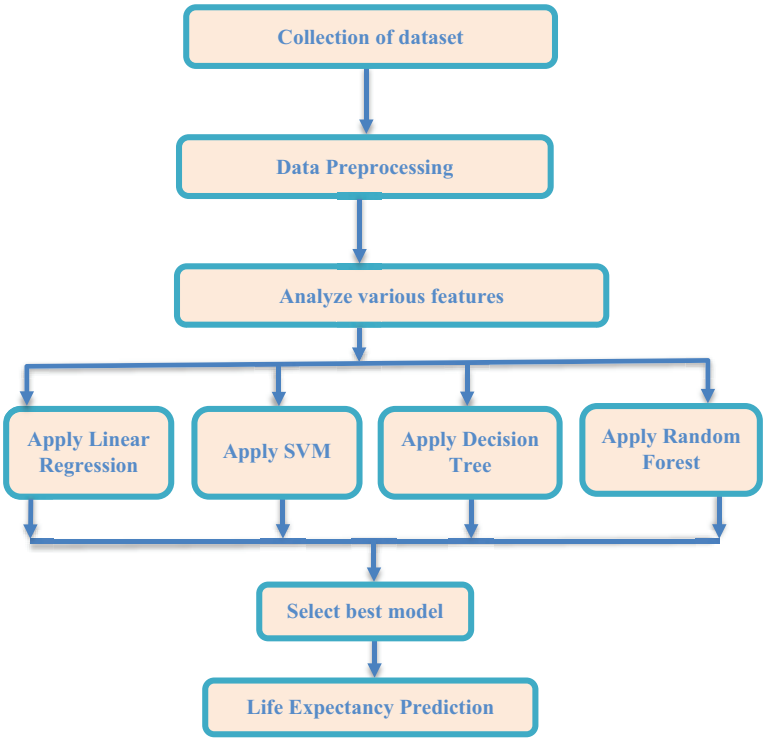

Fig. 1. Proposed method

3.1 Dataset information

The original source of the dataset was WHO (World Health Organization). From the source of the dataset, it is observed that data on life expectancy and health factors for 193 nations was gathered from the WHO data repository website, while economic data

was gathered from the United Nations website. Only the most relevant important points were picked from all categories of health-related factors. It has been observed that, in comparison to the previous 30 years, there has been a significant improvement in the health sector, resulting in lower human mortality rates, particularly in developing countries. The dataset has 22 fields and 2938 entries. The features of the dataset are "Country", "Year", "Status", "life expectancy", "Adult Mortality", "infant deaths per 1000 population", "Alcohol", "percentage expenditure", "Hepatitis B", "Measles", "bmi", "under-five deaths", "Polio", "Total expenditure", "diphtheria", "HIV/AIDS", "GDP", "Population", "thinness 1–19 years", "thinness 5–9 years", "Income composition of resources", "Schooling". Among these features "life expectance" is the dependent feature. The remaining features are independent features. Before applying ML algorithms some preliminary analysis done on few features. Figure 2 shows the top 15 countries and bottom 15 countries in immunization percentage. From the figure it is observed that top three countries in polio immunization are srilanka, Hungary and Bolivia.

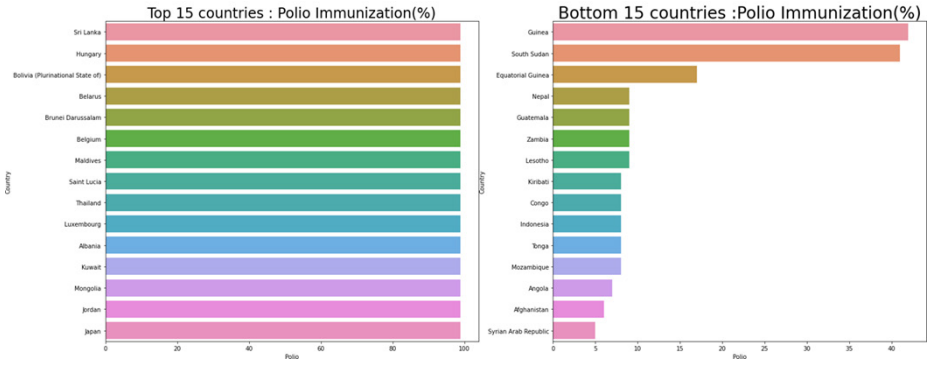

Fig. 2. Top 15 and Bottom 15 countries: polio immunization percentage

Figure 3 shows the top 15 countries with more polio immunization percentage and bottom 15 countries in infant deaths. From the figure it is observed that top three countries in infant deaths (per 1000) are India, Nigeria and Pakistan. The right side of the figure shows the countries with zero infant deaths.

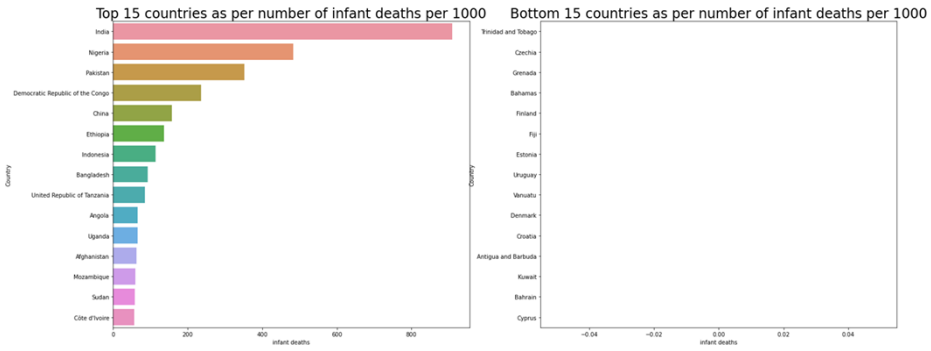

Fig. 3. Top 15 and Bottom 15 countries: infant deaths (per 1000)

Figure 4 shows the top 15 countries and bottom 15 countries with diphtheria cases. The top three countries are Uzbekistan, Bhutan and China.

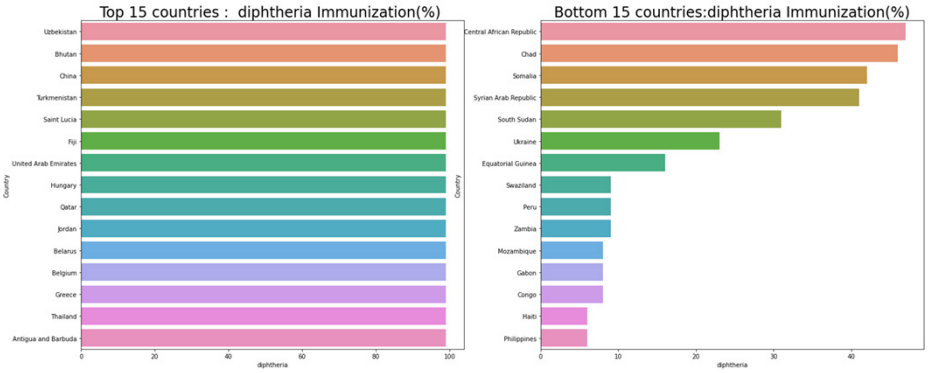

Fig. 4. Top 15 and Bottom 15 countries diphtheria Immunization percentage

Figure 5 shows the top 15 countries and bottom 15 countries with life expectancy. The top three countries are Slovenia, Denmark, cyprus.

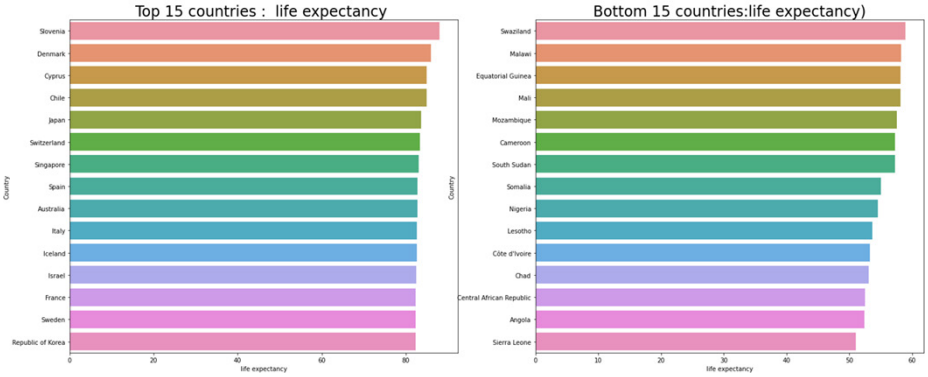

Fig. 5. Top 15 and Bottom 15 countries with life expectancy

Later, we also analyzed which features are directly affecting life expectancy. Figure 6 shows the relation between life expectancy and adult morality and schooling. It is observed that life expectancy is negatively correlated with adult morality. Similarly, Schooling is positively corelated with life expectancy.

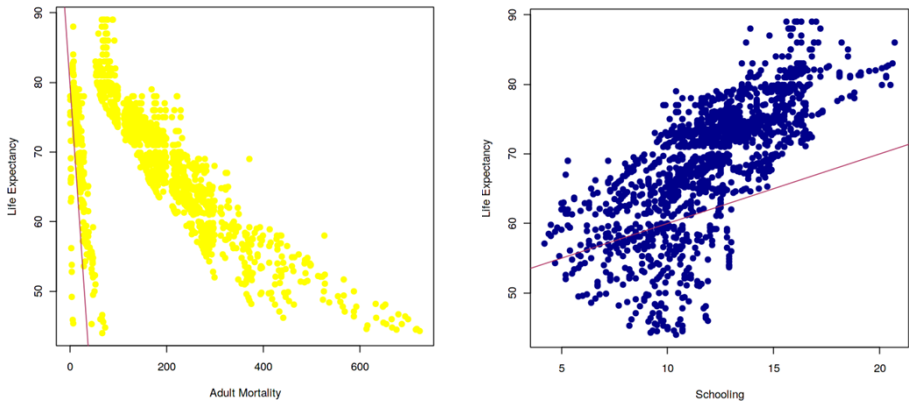

Fig. 6. a) Life Expectancy vs Adult Morality b) Life Expectancy vs Schooling

Figure 7 shows the relation between life expectancy and alcohol and Polio immunization. It is observed that life expectancy is negatively correlated with alcohol. Similarly, Polio immunization is positively corelated with life expectancy.

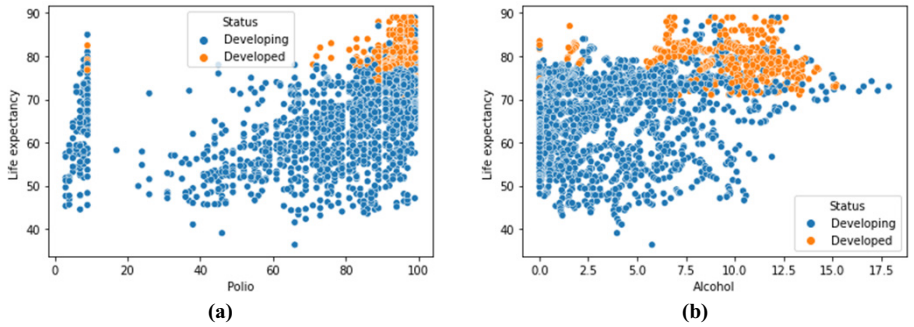

Fig. 7. a) Life Expectancy vs Polio immunization b) Life Expectancy vs Alcohol

4 Experimentation and results

All the implementation are done in google Collaboratory.

4.1 Applying regression analysis

We applied regression analysis technique to identify the features that are not affecting life expectancy. The regression analysis of the experiment was shown in Figure 8.

	coef	std err	t	P>|t|	[0.025	0.975]
const	53.7852	0.528	101.953	0.000	52.751	54.820
x1	-0.0204	0.001	-25.722	0.000	-0.022	-0.019
x2	0.0976	0.008	11.731	0.000	0.081	0.114
x3	0.1401	0.023	5.971	0.000	0.094	0.186
x4	0.0001	8.46e-05	1.640	0.101	-2.72e-05	0.000
x5	-0.0147	0.004	-3.753	0.000	-0.022	-0.007
x6	-1.915e-05	7.69e-06	-2.492	0.013	-3.42e-05	-4.08e-06
x7	0.0489	0.005	10.474	0.000	0.040	0.058
x8	-0.0738	0.006	-11.975	0.000	-0.086	-0.062
x9	0.0284	0.004	6.327	0.000	0.020	0.037
x10	0.1006	0.034	2.977	0.003	0.034	0.167
x11	0.0398	0.005	8.399	0.000	0.030	0.049
x12	-0.4772	0.018	-27.135	0.000	-0.512	-0.443
x13	3.646e-05	1.3e-05	2.797	0.005	1.09e-05	6.2e-05
x14	6.1064	0.637	9.581	0.000	4.857	7.356
x15	0.6737	0.042	16.064	0.000	0.592	0.756

Fig. 8. Regression analysis

It is observed that, expect the feature "percentage expenditure" (which is having p value 0.1), all the remaining features, p values are less than 0.05. So, all these features are good predictors for predicting life expectancy. Later, we also applied regression analysis country wise. All the countries do not have same factors for expecting life expectancy. The features affecting life expectancy is different for different countries. Its difficult to show factors for all countries, so the Table 1 (based on p-values) shows life expectancy influencing factors for some countries ("Afghanistan", "Australia", "Canada", "Egypt", "France", "Ireland", "Japan"). Later, we also checked life expectancy prediction with only these features and achieved good results.

Table 1. Factors affecting life expectancy

Country	Features Having More Influence
Afghanistan	Polio, diphtheria
Australia	bmi
Canada	alcohol, percentage expenditure, GDP
Egypt	adult morality, Total expenditure
France	Polio
Ireland	Alcohol, bmi, GDP
Japan	Polio

4.2 Applying ML regressors

We applied four regression algorithms multiple linear regression, decision tree regressor, SVM and random forest regressor. The dataset contains 2,938 samples. It is partitioned into train and test parts with 2,056 and 882 samples. After applying four algorithms, the results are tabulated as shown in Table 2.

Table 2. Accuracy

Model	R-Squared Value	RMSE
Multiple Linear Regression	0.81	4.051
Support Vector Regression	0.56	7.902
Decision Tree	0.91	2.720
Random Forest	0.96	1.928

From the Figure 9, it is observed that r-squared value is good for both decision tree and random forest algorithms. From the Figure 10, it is observed that RMSE (Root Mean Squared Error) value is also less for both decision tree and random forest algorithms. So, Decision Tree and Random Forest are successfully predicted life expectancy. Random Forest achieves 96% r-squared value, means it is the final model for prediction of life expectancy among four applied algorithms.

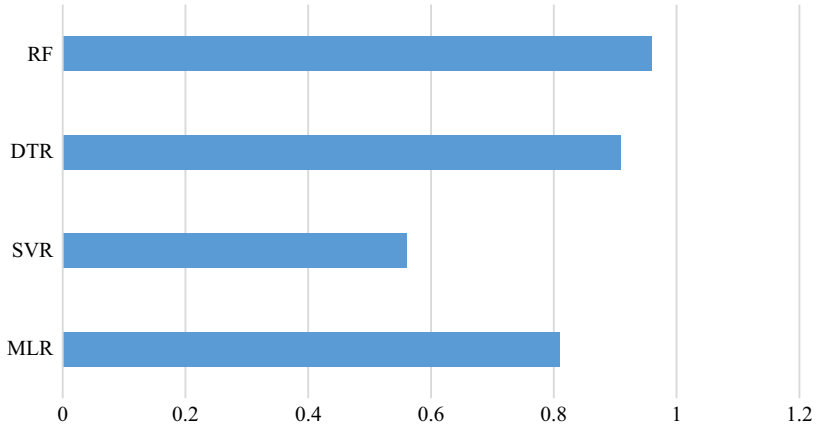

Fig. 9. Comparison of r-squared values of algorithms

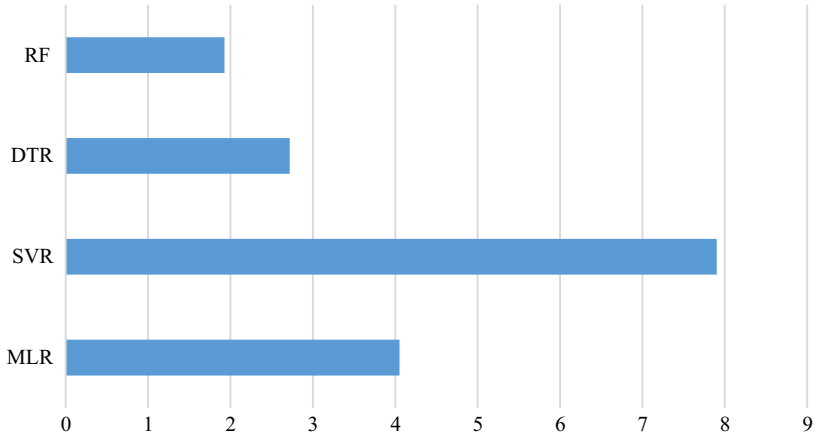

Fig. 10. Comparison of RMSE values of algorithms

Figure 11 shows the predicted values and actual values after applying random forest algorithm for first 150 predictions It is observed from the figure, there are not many deviations from actual and predicted values. So, Random Forest done well with 96% accuracy.

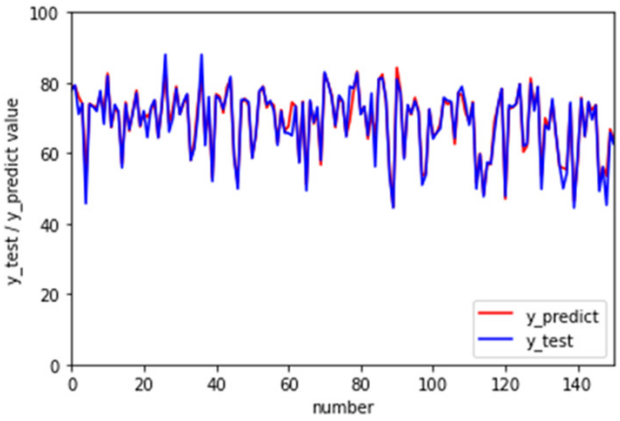

Fig. 11. Actual and predicted values plot for random forest

5 Conclusion

In this paper, we collected a Kaggle dataset (original resource is WHO) with 2,938 samples from 193 countries with several features. The dataset contains immunization features as well as other factors like morality rate, income etc. We analyzed the relation between various features and life expectancy values. We also analyzed country wise features which are having more influence for deciding life expectancy. Later, we applied four machine learning algorithms namely Multiple Linear Regression, Decision Tree Regression, Support Vector Regression and Random Forest for life expectancy prediction. After applying four algorithms, we achieved good results with random forest algorithm.

6 References

[1] Abeer S. Desuky et al. (2016). Improved Prediction of Post-Operative Life Expectancy After Thoracic Surgery. Adv Syst Sci, 16(2); 70–80, http://ijassa.ipu.ru/ojs/ijassa/article/view/351

[2] Michael B. Schultz et al. (2019). Age and Life Expectancy Clocks Based on Machine Learning Analysis of Mouse Frailty. https://doi.org/10.1038/s41467-020-18446-0

[3] Kasichainula Vydehi, Keerthi Manchikanti, T. Satya Kumari, SK Ahmad Shah. (2020). Machine Learning Techniques for Life Expectancy Prediction. International Journal of Advanced Trends in Computer Science and Engineering, 9(4). https://doi.org/10.30534/ijatcse/2020/45942020

[4] Palak Agarwal et al. (2019). Machine Learning for Prognosis of Life Expectancy and Diseases. International Journal of Innovative Technology and Exploring Engineering (IJITEE), 8(10). https://doi.org/10.35940/ijitee.J9156.0881019

[5] Alex Zhavoronkov, Polina Mamoshina, Quentin Vanhaelen, Morten Scheibye-Knudsen, Alexey Moskalev, Alex Aliper. (2019). Artificial intelligence for aging and longevity research: Recent advances and perspectives. Ageing Research Reviews, 49; 49–66, https://doi.org/10.1016/j.arr.2018.11.003

[6] Parikh RB, Manz C, Chivers C, et al. (2019). Machine Learning Approaches to Predict 6-Month Mortality Among Patients With Cancer. JAMA Netw Open. 2(10):e1915997. https://doi.org/10.1001/jamanetworkopen.2019.15997

[7] V. Bali, D. Aggarwal, S. Singh and A. Shukla. (2021). Life Expectancy: Prediction & Analysis using ML. International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), 1–8, https://doi.org/10.1109/ICRITO51393.2021.9596123

[8] A. Bhosale and K. K. Sundaram. (2021). Life Prediction Equation for Human Beings. International Conference on Bioinformatics and Biomedical Technology, 2010, 266–268, https://doi.org/10.1109/ICBBT.2010.5478965

[9] A. Lakshmanarao, M. Raja Babu, and T. Srinivasa Ravi Kiran. (2021). An Efficient Covid19 Epidemic Analysis and Prediction Model Using Machine Learning Algorithms. International Journal of Online and Biomedical Engineering, 17(11); 176–184, https://doi.org/10.3991/ijoe.v17i11.25209

[10] Nataliya Boyko and Olena Moroz. (2020). Comparative Analysis of Regression Regularization Methods for Life Expectancy Prediction. 3rd International Workshop on Modern Machine Learning Technologies and Data Science, June 5, 2021, Ukraine,CEUR-WS.org/vol-2917/paper27.pdf

[11] James Jin Kang and Sasan Adibi. (2018). Systematic Predictive Analysis of Personalized Life Expectancy Using Smart Devices. Technologies, 6(74), www.mdpi.com/journal/technologies; https://doi.org/10.3390/technologies6030074

[12] C.H. Leng et al. (2016). Estimation of life Expectancy, Loss-of-Life Expectancy, and Lifetime Healthcare Expenditures for Schizophrenia in Taiwan. Schizophr. Res. 171; 97–102, https://doi.org/10.1016/j.schres.2016.01.033

[13] M. Beeksmal, Suzan Verberne, Antal van den Bosch, Enny Das, Iris Hendrickx and Stef Groenewoud. (2019). Predicting Life Expectancy with a Long Short-Term Memory Recurrent Neural Network Using Electronic Medical Records. BMC Medical Informatics and Decision Making, https://doi.org/10.1186/s12911-019-0775-2

[14] F. Zahedi and M. Karimi Moridani. (2022). Classification of Breast Cancer Tumors Using Mammography Images Processing Based on Machine Learning: Breast Cancer Tumors Using Mammography Images. International Journal of Online and Biomedical Engineering, 18(05); 31–42, https://doi.org/10.3991/ijoe.v18i05.29197

[15] I. Taylor, I. Silva, S. Barreto, C. Soares, and J. Mendes. (2020). Raynaud's Phenomenon Impact on Quotidian Quality of Life. International Journal of Online and Biomedical Engineering, 16(09); 88–104, https://doi.org/10.3991/ijoe.v16i09.13993

[16] https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who

7 Authors

A. Lakshmanarao is currently working as Associate Professor in Aditya Engineering College, Surampalem. He completed his B.Tech in CSIT and M.Tech in Software Engineering. He is pursuing Ph.D. in Andhra University, Vishakapatnam. His areas of

interest are Machine Learning, Cyber Security, Deep Learning. He is a life member of Computer Society of India (CSI).

Dr. A. Srisaila, Assistant Professor, IT Dept, V.R. Siddhartha Engineering College completed M.Tech (CSE) from Bapatla Engineering College and Ph.D in CSE from Acharya Nagarjuna University. Her research areas are Software Reliability Engineering, Graphical Passwords, Human Computer Interaction and Big Data Analytics.

T. Srinivasa Ravi Kiran currently working as Assistant Professor & HOD in Department of Computer Science, PB Siddhartha College of Arts & Science, Vijayawada. He completed his Ph.D. in Acharya Nagarjuna University. His areas of interest are machine learning, databases, cyber security. He has published research papers in various conferences and journals.

G. Lalitha is currently working as Associate Professor in Adarsh College of Engineering, Chebrolu. She completed her B.Tech in CSE and M.Tech in CSE. Her areas of interest are Machine Learning, Machine learning and artificial intelligence.

Dr. K. Vasanth Kumar is currently working as Associate Professor in Malla Reddy Engineering College, Autonomous, Hyderabad. He completed his M.Tech (CSE) from JNTU Kakinada and Ph.D. from Sri Satya Sai University of Technology and Medical Sciences, Sehore, MP. His areas of interest are Software Engineering, Cloud Computing, Artificial Intelligence, Cyber Security, Data Structures.

Article submitted 2022-06-14. Resubmitted 2022-08-15. Final acceptance 2022-08-18. Final version published as submitted by the authors.