Predictive Analytics for Future Life Expectancy Using Machine Learning

Mrs. V. Krishna Vijaya [1], Ms. Latchi Naga Mythili[2], Ms. Manam Nandini[3], Ms. Kothuri Keerthi[4], Ms. Manchala Priyanka[5]

[1] *Asst. Professor, Dept. Information Technology, KKR & KSR Institute of Technology and Sciences, Guntur, India*

[2-5]*Student, Dept. Information Technology, KKR & KSR Institute of Technology and Sciences, Guntur, India*

---***---

***Abstract* - Advanced machine learning is used in the project "Predictive Analytics for Future Life Expectancy Using Machine Learning." methods for forecasting forthcoming trends in lifespan. By examining a range of datasets, including historical lifetime records, healthcare data, economic features, demographics, and environmental information, the research develops accurate prediction models. To assure reliability, these models—which were constructed using a variety of machine learning algorithms, including ensemble approaches, decision trees, and regression—go through a thorough training and validation process. This resulting forecasting tool is able to be used to legislators, healthcare professionals, and academics to make informed decisions on public health, including resource allocation. Periodic updates and monitoring enable reaction to evolving trends all while maintaining relevance and effectiveness over time.**

Key Words: Predictive Analytics, Future Life Expectancy, Linear regressor, Decision Tree regressor, Random forest regressor, Xgboost.

1 . INTRODUCTION

This research specifies into the application of advanced technologies like machine learning and data analytics to predict potential life expectancy. Through the examination of a wide range of datasets, including those related to medical care, demography, environmental factors, this research aims to develop accurate and predictive models. Regression modelling and decision trees which are the two machine learning techniques, are used to identify the main factors for impacting changes in trends in lifespan. Choosing relevant features to strengthen the model's resilience and putting the plans in place to deal with missing data efficiently during pre-processing are given a lot of attention. The main objective is to provide policymakers and healthcare professionals with relevant information that comes from in-depth statistical analysis so that they may implement focused interventions of life expectancy.

2. LITERATURE SURVEY

A study on life expectancy by the use of machine learning techniques for prediction was carried out in 2022 by Yallamati Prakasarao and Arumalla Nagaraju. As They proposed a Random Forest algorithm used to Predict Life expectancy by considering factors like Environmental basis, Food Habits, Diseases and Medical histories. As compared to Linear Regression, Decision Tree, KNN the Random Forest model achieved higher accuracy [1].

In 2023, Dr.Renuka Deshpande,Vaishnavi Uttarkar conducted a Study on Life Expectancy using Data Analytics. They have proposed a Random Forest Regressor Algorithm used to Predict life Span by considering the factors like WHO dataset and Demographic factors, Socio economic distribution and vaccination mortality. Compared to Regression models like Random Forest model achieved higher accuracy [2].

In 2023, Brian by 2023, BrianElphas Okango, Aholi Lipesa, Bernard Oguna Omolo, and Evans Otieno Omondi carried out the research on Using a supervised model of machine learning to make predictions lifespan. They suggested an XGBoost Regressor for predicting life expectancy by taking socioeconomic, behavioral, and health-related variables into account qualities. Compared to the Random Forest, Artificial Neural Networks the XGBoost Regressor model achieved higher accuracy [3].

A study on Life Expectancy: Prediction & Analysis using Machine Learning was carried out in 2021 by Dr. Vikram Bali, Dr. Deepti Aggarwal, Sumit Singh, and Arpit Shukla. They have suggested a Random Forest algorithm that takes into the account variables including schooling, young mortality, and HIV/AIDS in order to forecast life expectancy. This Random Forest model is superior for both Decision Tree and Linear Regression models [4].

In the year 2022, A.Lakshmanarao, Srisaila A, Srinivasa Ravi Kiran T, Lalitha G ,Vasanth Kumar K, conducted a study on Life Span prediction through Analysis of immunization and HDI Factors using Machine Learning Regression Algorithms.

They have proposed a Random Forest algorithm which is used to predict the life expectancy by considering factors of WHO dataset, demographic factors, socio economic distribution, Death rates, Immunization and HDI factors. As they have distinguished between the Support Vector Machine, Decision tree, Logistic Regression the Random Forest Regression model has Achieved higher accuracy [5].

In this study Palak Agarwal, Navisha Shetty, Kavita Jhajharia, Gaurav Aggarwal, Neha V Sharma conducted a study on Machine Learning for prognosis of Life spans & Diseases of the people in the year 2019. And they proposed a Random Forest algorithm which is used to predict the Life Span by considering the factors like diseases and Economic Factors. The Random Forest model achieved higher accuracy when compared to the Linear Regression and Decision Tree [6].

In the year 2020 Kasichainula Vydehi, Keerthi Manchikanti, T.Satya Kumari, SK Ahmad Shah, had conducted the study on Life Expectancy Predication using Machine learning techniques they had predicted that the human life span relies upon different features like financial development of the nation , well being developments of the people. In this paper they had proposed a Machine learning model which is used to predict life expectancy through WHO life expectancy data set. Here various Regression and Classification algorithms are been divided into five different ranges in order to predict the life expectancy .From being investigated different models we may conclude that Random forest regression produces the most exact outcomes and had given better rsquared value [7].

In 2021 Abhinaya.V, Dharani.B.C, Vandana.A,Dr.Velvadivu.P, Dr.Sathya.C conducted the study on Statistical analysis on factors influencing life expectancy and they have been analyzed the improvement of the life expectancy of the population had been taken through various factors. The main focus of our study is to determine the predicting factor which is contributing to higher value of life expectancy. The data have been collected from the WHO data repository website and its corresponding economic data was collected from United Nation website for the period 2000 - 2015 for 193 countries. The main factors which are been considered are immunization, mortality, economic, social and other health related factors. Step wise regression and cluster analysis algorithms are been used [8].

In the year 2021,Olgerta Idriz , Miranda Harizaj conducted a study on Research Methodology for predicting Life Expectancy using Machine Learning. They proposed KNN algorithm used to predict the life expectancy by considering factors like mortality rate, alcohol consumption, infant death, covid mortality and other health issues. As compared to Linear Regression the KNN achieved higher Accuracy [9].

In 2023 ,Bongi Pooja1,Ms.U.Archana conducted a study on Life Expectancy Prediction Using Machine Learning. They proposed a Random Forest algorithm used to predict life expectancy by considering factors like demographic, health and Environmental factors. Compared to Decision Tree, KNN the Random Forest model achieved Higher accuracy [10].

In the year 2020, Josep Penuelas, Tamas Krisztin, Michael, Obersteiner, Florian Huber Hannes Winner examined Relationships of the Human intake of N and P, Animal and Vegetable Food, and Alcoholic Beverages with Cancer and Life Span. The Quantity, Quality and all the type of human food have been correlated with human Health, We aimed the shed light on this association by using the integrated data at Country level. They had correlated elemental (nitrogen (N) and phosphorus (P)) Compositions and N:P ratios, they have been used official databases in order to predict the Life Expectancy. By this analysis four generally consistent conclusions had been taken after conducting deep Bayesian analyses of Country-Level data [11].

In 2018C. Cosculluela-Martínez R. Ibar-Alonso G. J. D. Hewings analyzed life Expectancy Index through Age Structure of Population and Environment Evolution The paper discusses the methodology used to compute indexes for life expectancy and other factors. It mentions the use of factorial analysis and VARMA models to calculate the weights of the index. Results show that China, Turkey, and Brazil rank higher in life expectancy index than previously thought. The dimensions of the index include population pyramid, industrial contamination, and citizen contamination. There are no significant differences in life expectancy index between countries, with elder population care being a common factor among top-ranking countries [12].

In the year 2016,Leng C.H et.al proposed a simple linear regression technique with the logit model -transformed survival ratio between the Machine Learning Techniques for Life Expectancy Prediction cohort, gender and age combination referents through simulation from the national life table [13].

Using techniques like MLR and Random Forest regression, Alex Zhavoronkov, Polina Mamoshina, Quentin Vanhaelen, Morten Scheibye-Knudsen, Alexey Moskalev, and Alex Aliper were able to forecast life expectancy in 2019 with good results [14].

Global differences in life expectancy were shown to be persistent in 2020 after Liou, Joe, Kumar, and Subramanian examined 65 years of data from 201 nations. Older age groups showed persistent inequalities, despite convergence at birth. The results highlight the necessity of focused efforts, especially in poor countries, in order to guarantee fair health outcomes. [15].

3. PROPOSED SYSTEM

First the system collects the data based on the demographic variables and understand the links between the data, the system gathers the data and performs pre-processing

(handling missing values, encoding categorical variables, etc.). Following that, methods for feature selection and visualization are applied. A few metrics that are employed in

the training and evaluation of regression models include Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared value. These models include Linear Regression, Decision Tree Regression, Random Forest Regression, and XGBoost Regressor. Serialization techniques are used to save the best-performing model for later deployment and choose it for continued use. With the use of fictitious scenario modeling historical trend analysis, and life expectancy prediction, this technology help users to assess the implications of their choices.

3.1 PROBLEM DEFINITION

In particular, vaccination rates and demographic factors like the human development index will be the focus of this data science research, which attempts to thoroughly analyses trends in life expectancy around the world and pinpoint important factors. The goal of the project is to find out how these factors affect life expectancy in different nations by using sophisticated data analytics and machine learning techniques on a dataset from the WHO repository via Kaggle. In order to improve population health outcomes globally, the project intends to identify important predictors and comprehend their relationships in order to support evidence-based public health policies and actions.

3.2 WORKFLOW

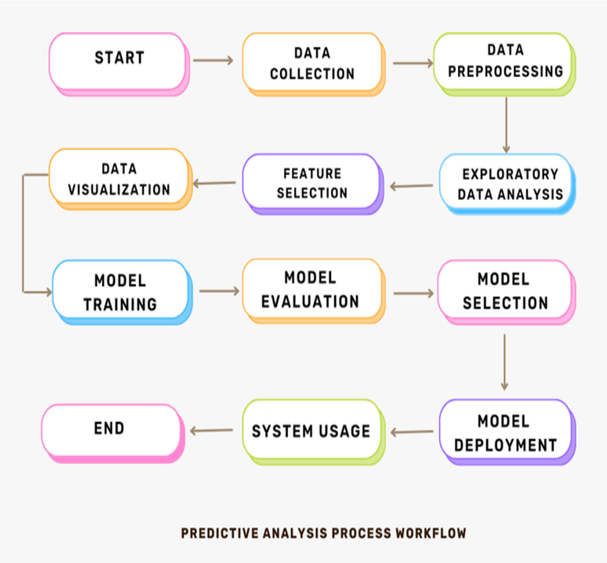


Fig 1: Represents the Workflow of project

3.3 IMPLEMENTATAION METHODS

1. Import and Load data: From a CSV file create a Data Frame containing life expectancy information. Import the libraries that are necessary for both Data processing and Data visualization.
2. Exploratory Information Analysis: In this case we use some of the Primary functions such as Info(), describe(), head(), tail(),to analyze the data in order to determine its brief summary and statistical structure.
3. Data Preparation and the Data Cleaning: Here we should Use the suitable approach to deduce the missing values; and for the numerical characteristics, we use the median; and for categorical features, we use the mode. And also, we use label method for encrypting categorical variable.
4. Features Selection for Visualization: To reduce redundancy and the multicollinearity problems, eliminate significantly related features. Choose specific characteristics and use methodologies like variance threshold and the correlation analysis to visualize the relation among them.
5. The project employs a range of data visualization methods, including bar graph, count plots, and histograms, to shed light on the distribution of the data and the correlations between the variables.
6. Regression model training: XGBoost, Random Forest, Decision Tree, and Linear regression are among the models that will be trained for this project. The dataset is split up into testing, validation, and training sets in order to fully assess model performance.
7. To evaluate how effective the models are, performance measurements such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared are used. And as it compares the model of each with other as we obtained linear regression high as per comparison.
8. Deciding Which Model Is Best: The best performing model is to determine by comparing its R-squared value and root mean square error. This Model is Finally Saved and Loaded by utilizing the {pickle} library, the optimal model is preserved in a file for further utilization. Predictions on fresh data can be made by loading and using the saved model later.
9. Lastly, to show how the trained model is put to use in real-world scenarios, the saved model is used to forecast new input data points.

   With the ability to be modified and expanded for comparable regression tasks in healthcare analytics or adjacent areas, the project offers a full framework for data pre-processing, model training, evaluation, and deployment. It also shows the results to the user.

4. RESULT ANALYSIS

By applying modern forecasting methodologies, our predictive analytics system makes reliable predictions about future changes in life expectancy based on historical data and pertinent socio-economic indices. As soon as the user enters the values and other relevant information, the system gathers and processes the required information. Afterwards, it uses machine learning methods to train an advanced predictive model. Then life expectancy of the population is then accurately predicted for the given horizon using this methodology. By doing this, our platform provides users with insightful knowledge on population health patterns that will help them make well-informed decisions and create policies that will be improve well-being in General.

As shown below the visualization of the data of status and count of developed and developing countries

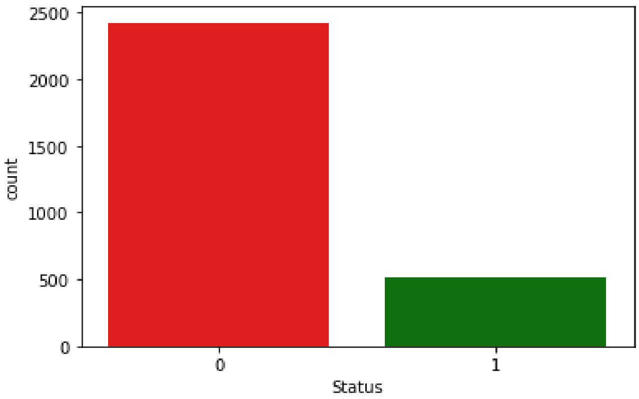


Fig 2: Indicates the population's status and number based on each country's total population.

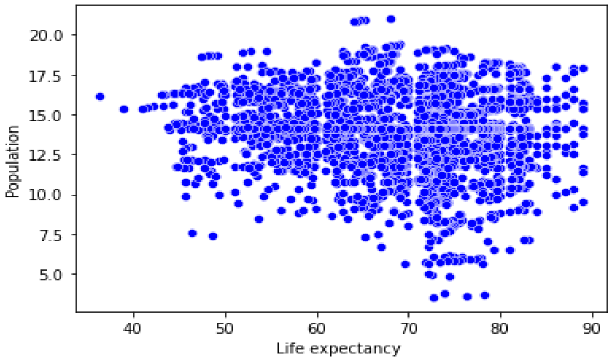
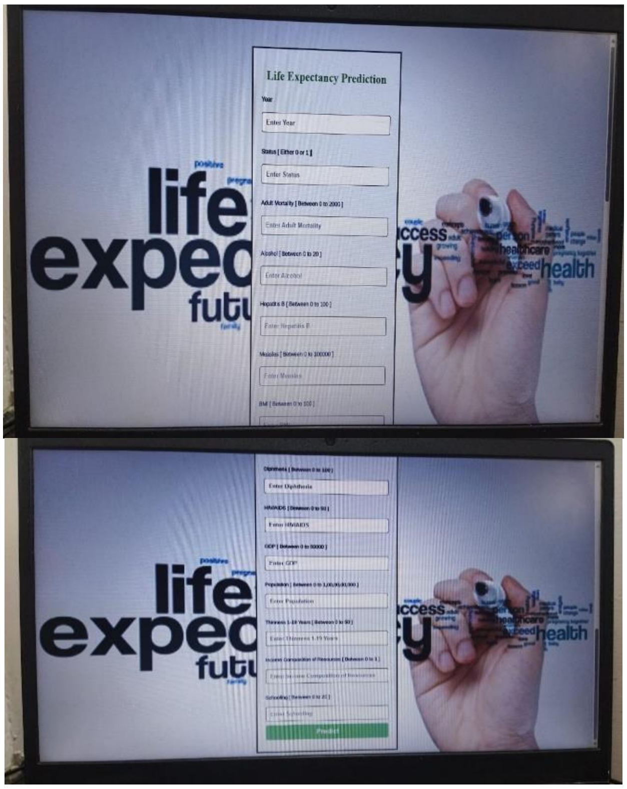


Fig 3: Demonstrates the argument on life expectancy based on a nation's population.

User input form :



Life Expectancy Prediction

Year

Enter Year

Status [Either 0 or 1]

Enter Status

Adult Mortality [Between 0 to 2000]

Enter Adult Mortality

Alcohol [Between 0 to 20]

Enter Alcohol

Hepatitis B [Between 0 to 100]

Enter Hepatitis B

Measles [Between 0 to 100000]

Enter Measles

BMI [Between 0 to 100]

Diphtheria [Between 0 to 100]

Enter Diphtheria

HIV/AIDS [Between 0 to 50]

Enter HIV/AIDS

GDP [Between 0 to 50000]

Enter GDP

Population [Between 0 to 1,00,00,00,000]

Enter Population

Thinness 1-19 Years [Between 0 to 50]

Enter Thinness 1-19 Years

Income Composition of Resources [Between 0 to 1]

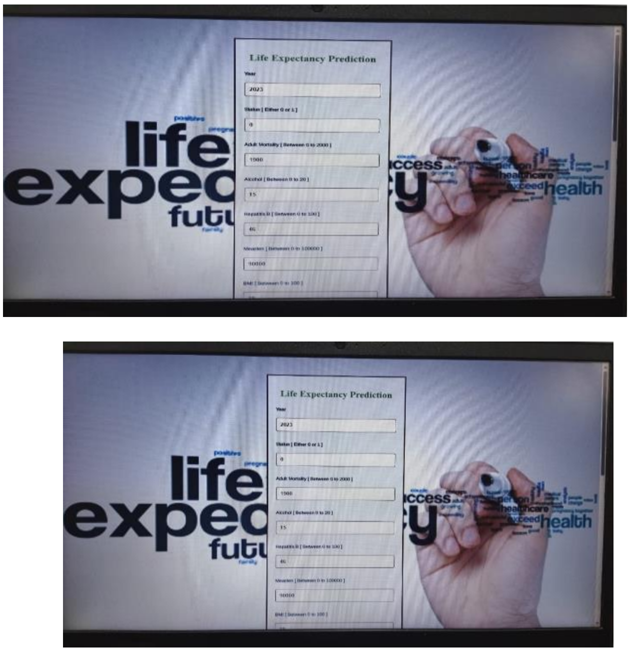
Enter Income Composition of Resources

Schooling [Between 0 to 20]

Enter Schooling

Fig 4: Input form

Taking inputs from user:



Life Expectancy Prediction

Year

2023

Status [Either 0 or 1]

0

Adult Mortality [Between 0 to 2000]

1000

Alcohol [Between 0 to 20]

15

Hepatitis B [Between 0 to 100]

46

Measles [Between 0 to 100000]

10000

BMI [Between 0 to 100]

Life Expectancy Prediction

Year

2023

Status [Either 0 or 1]

0

Adult Mortality [Between 0 to 2000]

1000

Alcohol [Between 0 to 20]

15

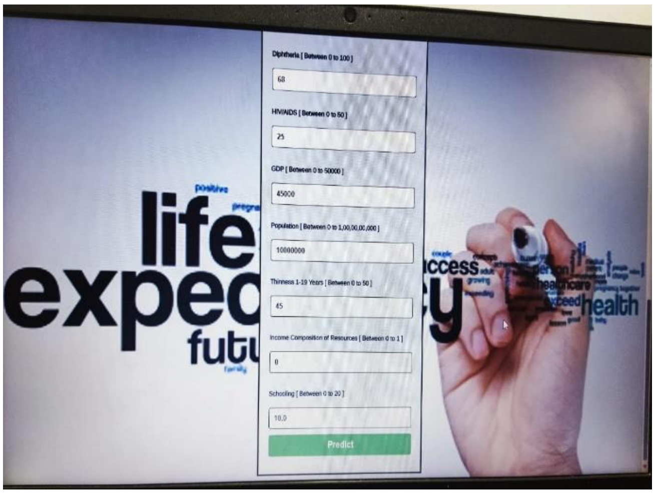
Hepatitis B [Between 0 to 100]

46

Measles [Between 0 to 100000]

10000

BMI [Between 0 to 100]



Diphtheria [Between 0 to 100]
68

HIV/AIDS [Between 0 to 50]
25

GDP [Between 0 to 50000]
45000

Population [Between 0 to 1,00,00,00,000]
10000000

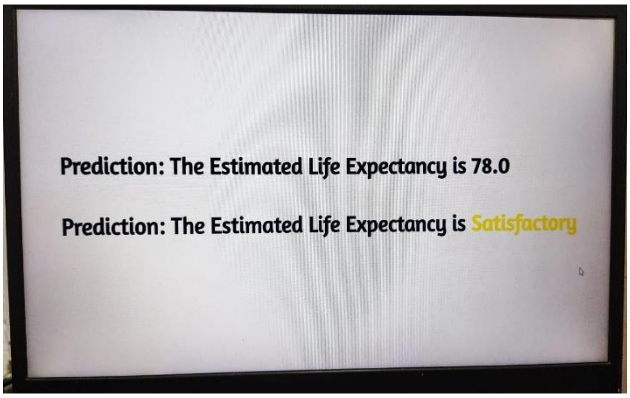
Thinness 1-19 Years [Between 0 to 50]
45

Income Composition of Resources [Between 0 to 1]
0

Schooling [Between 0 to 20]
10.0

Fig 5: Following the user's input of values to estimate lifespan

Final Prediction : -



Prediction: The Estimated Life Expectancy is 78.0

Prediction: The Estimated Life Expectancy is Satisfactory

Fig 4: Outcome of Life Expectancy Prediction

5. CONCLUSION

This including data importation, exploration, cleaning, pre-processing, and feature selection, offers a sophisticated framework for applying predictive analytics of life expectancy estimation. The Random Forest Regressor is proven to be the most successful model after a thorough examination of a model training and testing, confirming its usefulness in Healthcare planning. By integrating a user-friendly backend system, Decision-making becomes easier and interaction is improved with an Emphasis of user accessibility and experience. The project guarantees continuous accuracy and relevance by means of sophisticated visualization Techniques and continuous model refining. In the end by providing a strong basis by incorporating Machine learning Techniques into the development and application of healthcare policies.

6. REFERENCES

[1] 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.

[2] International Journal for Research in Applied Science & Engineering Technology (IJRASET) ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538 Volume 11 Issue IV Apr 2023- Available at www.ijraset.com.

[3] Dr. Renuka Deshpande , Vaishnavi Uttarkar. Life Expectancy using Data Analytics. International Journal for Research in Applied Science & Engineering Technology (IJRASET) ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538 Volume 11 Issue IV Apr 2023- Available at www.ijraset.com

[4] Dr. Vikram Bali, Dr. Deepti Aggarwal, Sumit Singh, and Arpit Shukla. Life Expectancy: Prediction & Analysis using ML. September 2021Conference: 2021 9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO) DOI:10.1109/ICRITO51393.2021.9596123.

[5] A.Lakshmanarao, Srisaila A,Srinivasa Ravi Kiran T,Lalitha G,Vasanth Kumar K. Life Expectancy Prediction through Analysis of Immunization and HDI Factors using Machine Learning Regression Algorithms. iJOE – Vol. 18, No. 13, 2022, https://doi.org/10.3991/ijoe.v18i13.33315.

[6] Palak Agarwal, Navisha Shetty, Kavita Jhajharia, Gaurav Aggarwal, Neha V Sharma, "Machine Learning for Prognosis of Life Expectancy and Diseases", International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075 (Online), Volume-8 Issue-10, August 2019, DOI: 10.35940/ijitee.J9156.0881019 Journal Website: www.ijitee.org

[7] Kasichainula Vydehi, Keerthi Manchikanti, T.Satya Kumari, SK Ahmad Shah, "Machine Learning Techniques for Life Expectancy Prediction", Volume 9, No.4, July – August 2020 International Journal of Advanced Trends in Computer Science and Engineering Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse4594 2020.pdf https://doi.org/10.30534/ijatcse/2020/45942020.

[8] Abhinaya.V, Dharani.B.C, Vandana.A, Dr.Velvadivu.P, Dr.Sathya.C," STATISTICAL ANALYSIS ON FACTORS INFLUENCINGLIFE EXPECTANCY", International Research Journal of Engineering and Technology (IRJET) eISSN: 2395-0056 Volume: 08 Issue: 07 | July 2021 www.irjet.net p-ISSN: 2395-0072 .

[9] Olgerta Idriz,Miranda Harizaj, Research Methodology for predicting Life Expectancy using Machine Learning, purpose

2021. They proposed KNN algorithm used to predict the life expectancy by considering factors.

[10] Bongi Pooja1,Ms.U.Archana ,"life Expectancy Prediction Using Machine Learning". International Journal of Research Publication and Reviews Journal homepage: www.ijrpr.com.

[11] Josep Penuelas, Tamás Krisztin,Michael,Obersteiner , Florian Huber ,Hannes Winner Country-Level Relationships of the Human Intake of N and P, Animal and Vegetable Food, and Alcoholic Beverages with Cancer and Life Expectancy,International Journal of Environmental Research and Public Health, Int. J. Environ. Res. Public Health 2020, 17, 7240; doi:10.3390/ijerph17197240.

[12] Life Expectancy Index: Age Structure of Population and Environment Evolution C. Cosculluela-Martínez1 · R. Ibar-Alonso2 · G. J. D. Hewings3Social Indicators Research. https://doi.org/10.1007/s11205-018-1967-3 .

[13] Leng, C.H, Chou, M.H.; Lin, S.-H; Yang, Y.K.; Wang, J.D, Estimation of life expectancy, loss-of-life expectancy, and lifetime healthcare expenditures for schizophrenia in Taiwan, Schizophr. Res. 2016,171, 97–102, DOI: 10.1016/j.schres.2016.01.03.

[14] 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.

[15] Lathan Liou, William Joe, Abhishek Kumar, S.V. Subramanian, "Inequalities in life expectancy: An analysis of 201 countries, 1950–2015",in Elsevier Social Science & Medicine ,Volume 253, May 2020, 112964, https://doi.org/10.1016/j.socscimed.2020.112964.