WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Comprehensive Factors Influencing on Life Expectancy Analysis

[1] Athuru Chandana Prasuna, [2] Kolle Harshitha, [3] Kasetty Siva Sai Charan, 4Chiripireddy Srinivasulu ,[5] Thappita Uma Sankar, Dr. M.A Manivasagam[6]

[1] Student, Dept. of Computer Science and Engineering(AI&ML), Siddartha Institute of Science and Technology (SISTK), Puttur, Andhra Pradesh, India. atthurchandana@gmail.com

[2] Student, Dept. of Computer Science and Engineering(AI&ML), Siddartha Institute of Science and Technology (SISTK), Puttur, Andhra Pradesh, India. harshithanani159@gmail.com

[3] Student, Dept. of Computer Science and Engineering (AI&ML), Siddartha Institute of Science and Technology (SISTK), Puttur, Andhra Pradesh, India. itsmecharan19@gmail.com

[4] Student, Dept. of Computer Science and Engineering(AI&ML), Siddartha Institute of Science and Technology (SISTK), Puttur, Andhra Pradesh, India. srinivasuluchiripireddy@gmail.com

[5] Student, Dept. of Computer Science and Engineering (AI&ML), Siddartha Institute of Science and Technology (SISTK), Puttur, Andhra Pradesh, India. umasankarthappita@gmail.com

[6] M.E. Ph.D. Head of the Dept, Dept. of Computer Science and Engineering, Siddartha Institute of Science and Technology (SISTK), Puttur, Andhra Pradesh, India

ABSTRACT

This paper provides an in-depth analysis of life expectancy prediction based on machine learning methods. The main goal of this research is to discover and analyse the influence of socioeconomic, healthcare, environmental, and lifestyle variables on life expectancy in different countries. The dataset applied for this analysis includes factors like GDP, healthcare spending, immunization rates, schooling, BMI, HIV/AIDS prevalence, and nutritional status. These factors are pre-processed and evaluated to eliminate inconsistencies, address missing values, and choose the most powerful features in prediction.

Three machine learning models, Linear Regression, Random Forest, and XGBoost, are used to create predictive models. The performance of each model is measured based on metrics like Mean Squared Error (MSE) and R-squared (R^2). Out of the models considered, the Random Forest model has the best accuracy with an R^2 value of 0.94, showing that it can predict life expectancy accurately based on the chosen factors. Furthermore, feature importance analysis indicates that Schooling, GDP, BMI, and prevalence of HIV/AIDS are important contributors to the variability of life expectancy.

The results from this study reveal insightful information about the most determining factors in influencing life expectancy and propose improvements on how to optimize prediction accuracy. Future research encompasses incorporating more variables, utilizing high-level deep learning methods, and creating stronger models to aid policymakers in making strategic decisions that focus on improving the public's health and life expectancy as a whole.

Keywords: Life Expectancy, Machine Learning, Linear Regression, Random Forest, XGBoost, Health Prediction, Data Analysis, Socioeconomic Factors, Healthcare Expenditure, Feature Importance, Predictive Modelling.

I. INTRODUCTION

Life expectancy is a vital measure of the general health, well-being, and socio-economic progress of a population. [1] The study of life expectancy of a population is important for the evaluation of the degree of economic and social development of a country [1]. Knowledge of the determinants of life expectancy is important for policymakers, healthcare professionals, and researchers seeking to enhance public health levels and improve the quality of life.

Socioeconomic determinants like Gross Domestic Product (GDP), level of education, and spending on healthcare directly influence the access to healthcare facilities, diet, and sanitation.[2] Monumental improvements in life expectancy have been the predominant trend for high income, developed countries over the course of the 20th and 21st centuries[2]. Health-related factors like rates of immunization, burden of disease (like HIV/AIDS), and coverage of major health care interventions shape the determinants of health outcomes of a population. Furthermore, lifestyle indicators like BMI (Body Mass Index), alcohol use, and food consumption can significantly influence the health and lifespan of an individual. [1] The residents of a country with high life standards live longer, on average, and have a small mortality ratio [1].

Recent advancements in machine learning have enabled the prediction of life expectancy more accurately by analysing large datasets and determining the most significant factors. [3]The ultimate aim of applying machine learning is to develop algorithms which can be trained well and make improvements over time[3]. Machine learning models can reveal underlying patterns and relationships in complex data that the more conventional statistical approaches

might fail to identify. [4]Many studies have been conducted to estimate life expectancy going back to 1662 with an analysis to create a warning system for the onset, spread and decline of bubonic plague in London[4]. In this paper, we employ diverse machine learning models like Linear Regression, Random Forest, and XGBoost for constructing predictive models for estimating life expectancy.

II. RELATED WORKS

There has been research on predicting life expectancy through the use of machine learning methods and statistical analysis. Abhinaya et al. (2021) performed statistical analysis of factors affecting life expectancy, while Agarwal et al. (2019) applied machine learning methods to predicting disease. Karacan et al. (2020) utilized decision tree algorithms for country-based life expectancy differences.

Bongi Pooja and Archana (2023) verified the performance of machine learning models to predict life expectancy using regression and classification techniques. Pandey and Chhikara (2020) contrasted various regression models to study life expectancy trends. Bali et al. (2021) employed ensemble machine learning models to improve prediction accuracy.

While these studies offer great value to the field, our study contributes to the current literature by including a broader set of factors, such as environmental, economic, and health-related factors. We also compare various machine learning algorithms, such as Random Forest and XGBoost, in order to identify the best method to successfully predict life expectancy.

III. LITERATURE SURVEY

[1] Abhinaya et al. (2021) statistically analyzed life expectancy determinants. In their research, they compared socio-economic and health determinants and placed special emphasis on the impact of government health spending, level of education, and vaccine coverage on enhancing life expectancy. They discovered that economic stability and healthcare facilities are most important for longevity.

[2] Jessica Y. Ho and Arun S. Hendi (2018) examined trends in life expectancy among high-income nations. Their retrospective observational analysis revealed inequalities in life expectancy growth rates, with some nations levelling off because of rising obesity and inequalities in socioeconomic status. Their study emphasized the role of healthcare policies in shaping long-term trends in life expectancy.

[3] Agarwal et al. (2019) used machine learning to forecast life expectancy and disease prognosis. They used supervised learning algorithms in their research to examine health and demographic factors, showing that sophisticated predictive models can be used to improve early healthcare interventions and policy-making.

[4] Karacan et al. (2020) used decision tree methods to examine life expectancy patterns in various nations. In their paper, which was published in the Eastern Mediterranean Health Journal, determinants of life expectancy were set as GDP per capita, healthcare access, and education. The research demonstrated the utility of decision tree models in explaining complex relationships between health outcomes and socio-economic determinants.

[5] Basheer et al. (2019) discussed predictive analytics in different fields, such as predicting application to life expectancy, their research illustrated how machine learning can be applied to structured data and yield actionable insights.

[6] Bhargavi et al. (2024) did a comprehensive analysis of life expectancy using the help of modern data-driven approaches. In their work, they contrasted different regression methods to find strong predictors and thus highlighted the contribution of statistical and machine learning methods to public health studies.

[7] Bongi Pooja and Archana (2023) developed a model based on machine learning algorithms for the prediction of life expectancies. The authors' work incorporated multiple regression and classification algorithms and showed that ensemble approaches provided higher predictive accuracy compared to traditional statistical strategies.

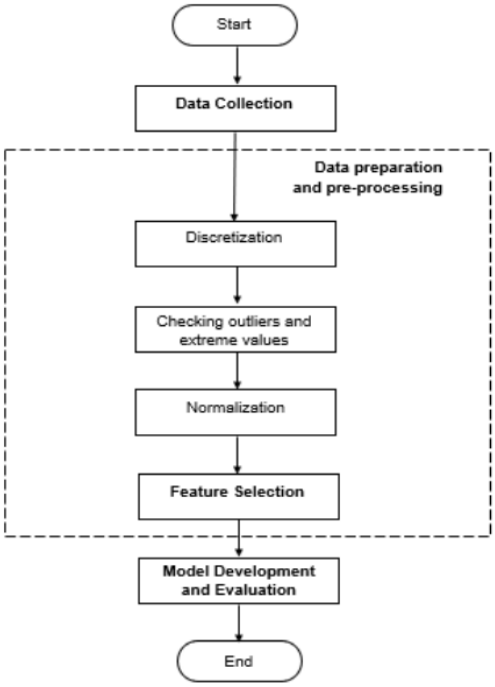
[8] Pandey and Chhikara (2020) examined life expectancy using different multiple regression approaches. The research compared linear regression, decision trees, and ensemble learning algorithms and concluded that non-linear models, particularly gradient boosting, were more accurate in predicting life expectancy trends.

[9] Bali et al. (2021) investigated machine learning methods for estimating life expectancy from different socio-economic and healthcare variables. Regression models were used in the study to analyse the effect of factors like GDP, health spending, literacy, and disease prevalence on life expectancy. They compared different ML models, with ensemble methods being found to yield better results. The study highlighted the need for feature selection to enhance model accuracy and suggested the inclusion of real-time health data in future prediction models

IV. METHODOLOGY

The dataset utilized in the present study was acquired from trusted sources like the World Health Organization (WHO) and other available public health and economic databases.[5] Data firstly was collected from respective department and the attributes were analysed based on their importance[5].It encompasses a broad set of determinants that affect life expectancy, such as health-related variables such as immunization coverage and disease prevalence, socioeconomic factors such as GDP, education, and government health spending, and environmental determinants such as levels of pollution and availability of clearwater.Lifestyle factors such as alcohol use, BMI, malnutrition indicators were also taken into account. The raw data underwent several preprocessing stages to ensure it was accurate and consistent.

Statistical methods such as mean or mode imputation were used to handle missing values, and in certain instances, they were deleted if they were not significant. Data cleaning was done by standardizing column names and deleting duplicate records. Feature engineering was done by generating new relevant features, e.g., regional averages or health indices.



An Exploratory Data Analysis (EDA) was performed to understand the dataset. Histogram with Kernel Density Estimation (KDE) plots were employed to understand the distribution of life expectancy. Correlation matrices were employed to understand the relationship between different variables, while outlier detection methods were employed to understand their effect on the dataset.

A machine learning model was developed to predict life expectancy using selected factors. Various algorithms were tried, including Linear Regression, Random Forest Regressor, Gradient Boosting Machines and Neural Networks for advanced modelling. Models were compared on Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared (R^2) Score. A model was selected on the basis of the performance metric, interpretability, and generalizability.

V. PROPOSED SYSTEM

Our system uses machine learning to construct a comprehensive and interpretable model to predict life expectancy from a broad set of determinants. In contrast to existing work limited to statistical analysis or univariate regression, our system combines a broad set of socioeconomic, health, environmental, and lifestyle determinants to deliver more holistic insights.

The system follows a systematic procedure:

- Data Preprocessing & Collection – Collection of high-quality data sets from WHO and other sources, imputation for missing values, and normalization.
- Feature Selection and Engineering – Determining the most important features that decide life expectancy and building more informative features.
- Machine Learning Model Construction – Construction and comparison of different regression models, such as Random Forest, XGBoost, and Neural Networks, to identify the top-performing model.
- Interpretability and Insights – Meaningful interpretation of feature importance to provide actionable insights to practitioners and policymakers.
- Web-Based Prediction System – Using the trained model in an easily accessible Flask-based web application, real-time prediction is possible.

By combining a localized model and a list of contributing factors, the system we suggest aims to enhance the accuracy and usability of life expectancy estimates. The information obtained from this model can be utilized to assist government agencies, healthcare institutions, and policymakers in making informed policies to improve public health outcomes.

VI. SYSTEM ARCHITECTURE

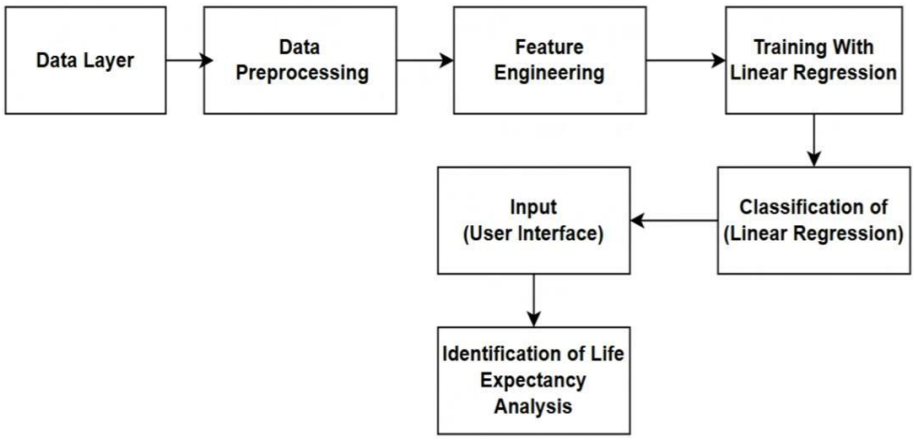


Fig 1. System Architecture

VII. EXPERIMENTAL PROTOTYPE AND ARCHITECTURE

To demonstrate the effectiveness of the proposed system, a prototype was implemented with a web application created using Flask. The prototype is intuitive when it comes to enabling users to input relevant health, economic, and environmental data to make real-time predictions of life expectancy. The backend processes the input data, applies the trained machine learning model, and provides the predicted life expectancy and the contributing factors. The suggested system design adheres to a highly defined three-tier design. The Data Layer consists of a structured database of the pre-processed and cleaned data. Provision is also included for trained machine learning models for prediction within the layer. The Processing Layer consists of backend logic that is executed through the application of Flask. This specific layer is responsible for processing user requests, preprocessing features, and using the selected machine learning model, either Random Forest or XGBoost, to make accurate predictions. The Presentation Layer exists in the form of a simple web interface used to input data and display predictions. The frontend is particularly designed to interact with the backend API, thus ensuring a seamless user experience.

The prototype has several key features designed to optimize both usability and efficiency. The intuitive interface supports easy input of key variables such as GDP, healthcare expenditure, and vaccination coverage. Use of machine learning algorithms ensures that the trained Random Forest and XGBoost models can make precise predictions. Additionally, use of feature importance visualization enables useful insights in terms of the key factors that influence the predicted life expectancy, thus enhancing the interpretability of the results. The system also offers insights relevant to a specific geographical location, allowing it to filter predictions by country or region, thus making it easier for policymakers and researchers to access targeted interventions. Such an architectural configuration offers assurances of scalability, efficiency, and interpretability, hence making it applicable for real-world applications in healthcare and policymaking. The structure of the different components guarantees the robustness of the system and allows for accurate predictions, which can subsequently be used to guide decision-making to optimize public health and optimize life expectancy outcomes..

VIII. CONCLUSION AND FUTURE SCOPE

The study identifies as the most important determinants of life expectancy Schooling, GDP, BMI, and HIV/AIDS prevalence. Results show that compared to conventional models like Linear Regression, ensemble learning models including Random Forest and XGBoost excel. Effective implementation of a Flask-based web application emphasizes even more the practical viability of the suggested system.

Not with standing expected results, the study has some shortcomings including the absence of elements like air quality, healthcare accessibility, and cultural practices. Extensive dataset expansion Variable inclusion help to improve prediction accuracy.

To improve model interpretability, future developments call for including Deep Learning approaches, neural networks, and SHAP explanations. Emphasizing the need of economic, healthcare, and lifestyle elements in life expectancy projections, the Random Forest model shows great accuracy with a R^2 of 0.94. Further developments in real-time data integration and web application usability will help the system to be practically useful for public health planning and policy-making.

XI. RESULT AND DISCUSSION

Data analysis identified a bimodal distribution of life expectancy with two peaks between 70-75 years and a cluster of low life expectancy values.

There were high correlations between life expectancy and variables like GDP, education, and spending on health.

Inverse relationships were found between child mortality and the prevalence of HIV/AIDS and implying that higher levels of these variables significantly lower life expectancy.

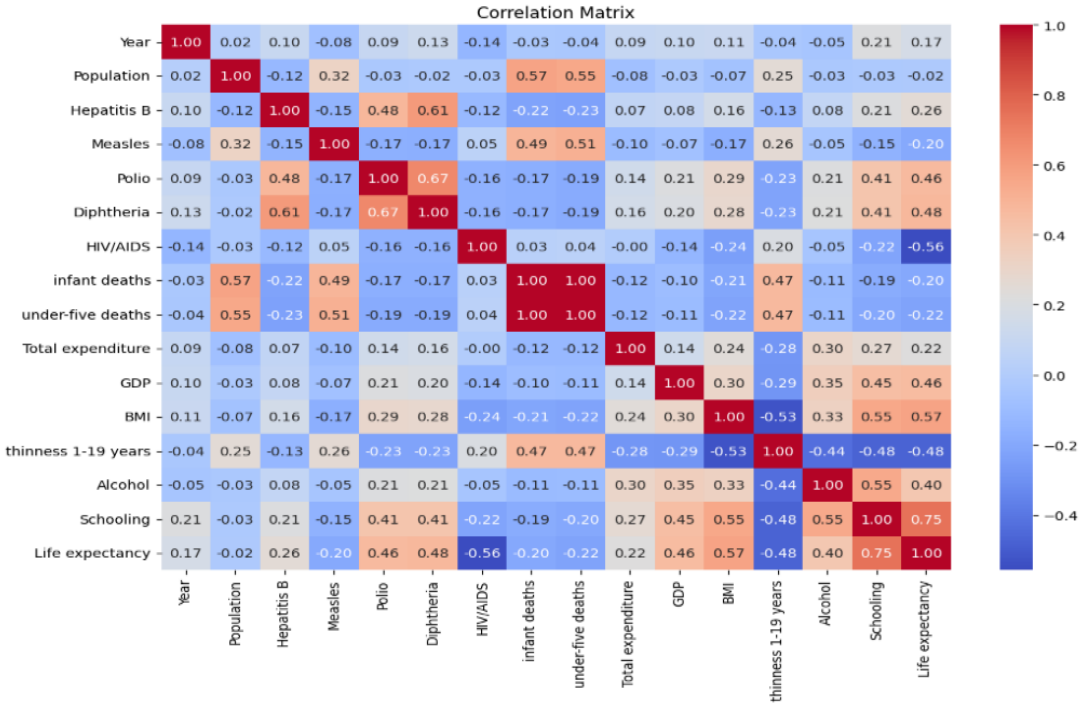


Fig: Heatmap correlations

Out of the models that were tested, the Random Forest Regressor gave the optimal balance between accuracy and interpretability. It yielded:

```
Random Forest Mean Squared Error: 5.54
Random Forest R-squared: 0.94
Random Forest Mean Absolute Error: 1.62

Feature Importances:
                 Feature  Importance
4               HIV/AIDS    0.622452
0              Schooling    0.177894
2                    BMI    0.059347
3   thinness 1-19 years     0.056496
6     Total expenditure     0.027959
5              Diphtheria    0.026790
1                    GDP     0.019833
7           Hepatitis B     0.009229
```

An analysis of feature importance indicated that the primary factors significantly influencing life expectancy are:

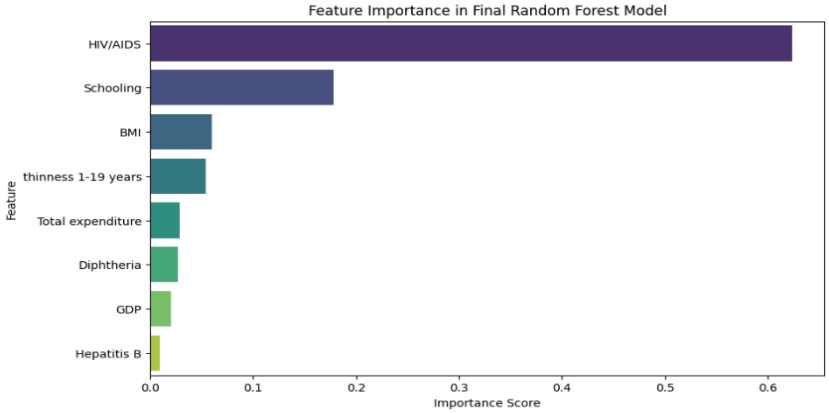
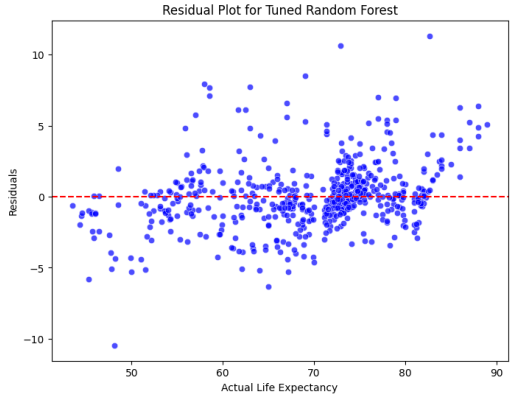


fig: Feature importance

It is associated with greater health consciousness and improved living conditions. GDP i.e More developed countries have improved healthcare and sanitation, Healthcare spending i.e Government health expenditure contributes significantly to life expectancy

Preventable conditions like measles, polio, and HIV/AIDS cause a sharp decline in life expectancy in the affected areas, and therefore vaccination and drug availability become the necessities.



Economically advanced countries with improved education systems are likely to have increased life expectancy, which implies that economic and educational development can directly affect public health outcomes.

Public health investment is needed to improve life expectancy, emphasizing the need for increased financial resources in healthcare in economically disadvantaged regions.

Random Forest is an ensemble method that builds many decision trees and combines their predictions, and it is more interpretable and less prone to overfitting. But it can be computationally expensive for big data.

XGBoost or Extreme Gradient Boosting is a boosting algorithm that constructs trees sequentially, fixing the earlier mistakes. It performs better when used on structured data and is computationally efficient but needs careful hyperparameter tuning to prevent overfitting.

```
Random Forest Cross-Validation R-squared Scores:
[0.77533056 0.84122316 0.81031775 0.8179971  0.7899252  0.88902152
 0.881923   0.69897062 0.78432772 0.88231315]
Average R-squared: 0.82

XGBoost Cross-Validation R-squared Scores:
[0.71936691 0.82467608 0.80000618 0.80583226 0.77779731 0.861431
 0.85650839 0.67119728 0.77888489 0.87867468]
Average R-squared: 0.80
```

X. REFERENCES

1. Abhinaya. V, Dharani. B. C, Vandana. A, Dr. Velvadivu. P, Dr. Sathya. C., “Statistical Analysis On Factors Influencing Life Expectancy”, Coimbatore, India, July 2021.
2. Jessica Y Ho, Arun S Hendi, “Recent trends in life expectancy across high income countries: retrospective observational study”, USA, 15August, 2018
3. Agarwal, P., Shetty, N., Jhajharia, K., Aggarwal, G., & Sharma, N. V. (2019). Machine learning for prognosis of life expectancy and diseases.
4. Karacan, I., Sennaroglu, B., & Vayvay, O. (2020). Analysis of life expectancy across countries using a decision tree. Eastern Mediterranean Health Journal, 26(2), 143–151.
5. Basheer, M. Y. I., Mutalib, S., Hamid, N. H. A., Abdul-Rahman, S., & Malik, A. M. A. (2019). Predictive analytics of university student intake using supervised methods. IAES International Journal of Artificial Intelligence, 8(4), 367–374.
6. K. Bhargavi, J. Rani, K. N. Pavani, D. Sravya “Life expectancy analysis” 4April,2024.
7. Bongi Pooja , Ms .U .Archana “ Life Expectancy Prediction Using Machine Learning” December,2023.
8. Anshu Pandey, Rita Chhikara “Analysis of Life Expectancy using various Regression Techniques” 2020.
9. Dr. Vikram Bali, Dr. Deepti Aggarwal, Sumit Singh , Arpit Shukla “Life Expectancy: Prediction &Analysis using ML” 2021.