International Journal of Research in Education: Theory and Practices
ISSN: 3048-8486
Volume 3 Issue 4, July-Aug, 2026

An Intelligent XGBoost-Based Machine Learning Framework for Accurate Life Expectancy Prediction Across Developed and Developing Countries

S.Gouthami[1],Yeguri Supriya[2]

[1]Assistant Professor, Department of MCA, Megha Institute of Engineering and Technology for Women, Edulabad (Village), Ghatkesar (Mandal), Medchal District, Telangana – 501301, India.

[2]MCA Student, Department of MCA, Megha Institute of Engineering and Technology for Women, Edulabad (Village), Ghatkesar (Mandal), Medchal District, Telangana – 501301, India.

Abstract. Life expectancy is one of the most important indicators used to evaluate the overall health status, healthcare quality, and socio-economic development of a country. Accurate prediction of life expectancy enables governments and healthcare organizations to formulate effective public health policies and allocate medical resources more efficiently. In recent years, machine learning techniques have demonstrated superior capability in modeling complex healthcare datasets compared with conventional statistical approaches. This study presents a robust life expectancy prediction framework based on the Extreme Gradient Boosting (XGBoost) algorithm for estimating life expectancy in both developed and developing countries. Initially, the collected dataset undergoes comprehensive preprocessing, including missing value imputation, normalization, and categorical encoding to improve data quality. The M5P decision tree algorithm is then employed to identify the most informative attributes influencing life expectancy. Subsequently, the processed data are divided into training and testing subsets, and the XGBoost regression model is optimized using RandomizedSearchCV to determine the most suitable hyperparameter configuration. Model performance is evaluated using Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and the coefficient of determination (R^2). Experimental analysis indicates that the optimized XGBoost model achieves excellent predictive performance with a testing R^2 score of 0.97, MAE of 1.03, and MSE of 2.45, outperforming several conventional machine learning approaches. Feature importance analysis further reveals that HIV/AIDS prevalence, adult mortality, and income composition are among the strongest determinants affecting life expectancy. The proposed framework provides an efficient and reliable decision-support tool for healthcare planners, policymakers, and researchers working toward improving global health outcomes.

keywords: Life Expectancy Prediction, XGBoost, Machine Learning, Feature Selection, M5P Algorithm, RandomizedSearchCV, Healthcare Analytics, Predictive Modeling, Regression Analysis, Public Health.

I. Introduction

Life expectancy represents one of the most widely accepted indicators for assessing the health status and socio-economic development of a population. It reflects the average number of years an individual is expected to live under prevailing mortality conditions and serves as an important benchmark for evaluating national healthcare systems. Governments, healthcare organizations, and international agencies frequently utilize life expectancy statistics to monitor public health progress, identify healthcare inequalities, and formulate long-term development strategies. Because life expectancy is influenced by numerous demographic, environmental, medical, and economic factors, accurately predicting this indicator remains a challenging yet essential task for policymakers and researchers [1], [2].

Several determinants contribute directly or indirectly to life expectancy, including access to healthcare services, disease prevalence, vaccination coverage, nutrition, sanitation, education, income level, environmental conditions, and healthcare expenditure. These factors differ considerably between developed and developing countries, leading to substantial variations in longevity across regions. For example, countries with advanced healthcare infrastructure, better educational systems, and higher economic stability generally exhibit greater life expectancy than regions experiencing limited medical resources and higher disease burdens. Understanding the relationships among these variables enables governments to develop targeted healthcare interventions that improve the overall quality of life [3].

Traditional statistical models have long been employed to estimate life expectancy using demographic and socio-economic variables. Although these approaches provide useful insights, they often struggle to capture the complex nonlinear interactions existing among multiple influencing factors. As healthcare datasets continue to grow in size and complexity, conventional regression techniques become less effective for generating highly accurate predictions. Consequently, machine learning algorithms have emerged as powerful alternatives capable of automatically identifying hidden patterns within multidimensional healthcare data while significantly improving prediction accuracy [4].

Recent advances in artificial intelligence have accelerated the adoption of machine learning in medical decision support systems. Algorithms such as Decision Trees, Random Forest, Support Vector Machines, Artificial Neural Networks, Gradient Boosting Machines, and XGBoost have demonstrated remarkable success in solving healthcare prediction problems. Among these techniques, XGBoost has gained considerable pop-

ularity because of its ability to efficiently process structured datasets, prevent overfitting through regularization mechanisms, and produce highly accurate predictions even when handling complex feature interactions. Its scalability and computational efficiency make it particularly suitable for large healthcare datasets containing diverse demographic and clinical variables [5].

One of the major challenges in life expectancy prediction involves selecting the most informative variables from a large collection of available attributes. Including irrelevant or redundant features often increases computational complexity while reducing the generalization capability of predictive models. Therefore, effective feature selection becomes a critical component of developing reliable machine learning systems. Decision tree-based feature selection methods, such as the M5P algorithm, are capable of ranking variables according to their contribution to prediction performance, allowing the model to focus on the most influential health and socio-economic indicators [6].

Another important consideration involves optimizing machine learning models through hyperparameter tuning. Selecting appropriate learning rates, tree depths, subsampling ratios, and the number of estimators greatly influences predictive performance. Automated optimization techniques such as RandomizedSearchCV systematically explore multiple parameter combinations and identify configurations that maximize model accuracy while minimizing prediction error. Such optimization improves model robustness and reduces the risk of overfitting when applied to unseen data [7].

This study proposes an optimized life expectancy prediction framework based on the XGBoost regression algorithm. The methodology begins with comprehensive data preprocessing, including country-wise missing value imputation, normalization, and categorical encoding to improve dataset consistency. The M5P decision tree algorithm is then employed to identify the most significant features affecting life expectancy. Following feature selection, the dataset is partitioned into training and testing subsets, and the XGBoost model is trained using optimized hyperparameters obtained through RandomizedSearchCV. Finally, model performance is evaluated using multiple regression metrics including Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and the coefficient of determination (R^2).

Experimental results demonstrate that the optimized XGBoost framework achieves excellent predictive capability, producing an R^2 score of 0.97 on the testing dataset while maintaining very low prediction errors. Furthermore, feature importance analysis identifies HIV/AIDS prevalence, income composition, and adult mortality as the most influential factors affecting life expectancy. These findings provide valuable insights for healthcare administrators and policymakers seeking evidence-based strategies to improve population health and allocate healthcare resources more effectively. Overall, the proposed framework offers a reliable, scalable, and computationally efficient solution for life expectancy prediction across both developed and developing countries.

II. Survey of Literature

The prediction of life expectancy has become an important research topic in healthcare analytics because it helps governments and healthcare organizations evaluate public health conditions and formulate long-term development policies. The increasing availability of healthcare datasets and advances in artificial intelligence have encouraged researchers to employ machine learning techniques for predicting life expectancy more accurately than traditional statistical models. Various algorithms have been investigated to identify the factors that significantly influence longevity, including socio-economic conditions, disease prevalence, healthcare expenditure, education, nutrition, immunization coverage, and environmental characteristics. Recent studies demonstrate that machine learning models are capable of capturing complex nonlinear relationships among these variables, resulting in improved prediction accuracy and better decision support for healthcare planning.

One of the earliest deep learning approaches for life expectancy prediction was introduced by Beeksma et al., who utilized Long Short-Term Memory (LSTM) Recurrent Neural Networks to analyze electronic medical records. Their framework was designed to capture temporal dependencies present in patient healthcare histories, enabling the model to learn sequential medical patterns that influence survival outcomes. The proposed deep learning model demonstrated that recurrent neural networks can effectively analyze longitudinal healthcare information and outperform several conventional statistical techniques when dealing with time-dependent clinical data. However, the computational complexity of deep neural networks and their dependence on large training datasets limited their practical deployment in resource-constrained healthcare environments.

To investigate differences in longevity across nations, Meshram et al. performed a comparative study on developed and developing countries using several machine learning algorithms, including Linear Regression, Decision Tree Regression, and Random Forest Regression. Their work evaluated prediction performance using statistical measures such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and the coefficient of determination (R^2). The experimental results indicated that Random Forest achieved better prediction performance than conventional regression methods because of its ensemble learning capability. Furthermore, the study identified adult mortality, HIV/AIDS prevalence, and healthcare expenditure as some of the strongest factors affecting life expectancy. Although Random Forest produced reliable predictions, its performance remained limited when modeling highly complex interactions among healthcare variables.

Another significant contribution was presented by Bali et al., who investigated multiple machine learning techniques for predicting life expectancy using demographic, socio-economic, environmental, and healthcare indicators. Their experimental analysis compared Decision Trees, Random Forest, Neural Networks, and regression-based approaches to determine the most suitable predictive model. The researchers observed

that ensemble learning methods consistently outperformed individual learning algorithms due to their ability to reduce variance and improve model stability. Nevertheless, the study highlighted that prediction accuracy could be further improved through effective feature selection and systematic hyperparameter optimization, which were not extensively explored in their framework.

Recognizing the advantages of boosting algorithms, Lipesa proposed an XGBoost-based machine learning framework for life expectancy prediction. XGBoost demonstrated superior predictive capability compared with Artificial Neural Networks and Random Forest models by efficiently handling nonlinear feature interactions and incorporating regularization techniques that minimized overfitting. The study emphasized that selecting suitable hyperparameters significantly enhanced prediction performance and improved model generalization. Despite these improvements, the optimization process relied on relatively limited parameter tuning strategies, leaving opportunities for further enhancement through more comprehensive optimization techniques such as RandomizedSearchCV.

Research has also explored the influence of immunization coverage and socio-economic development on life expectancy. Lakshmanarao et al. developed regression-based machine learning models to examine the relationship between life expectancy, Human Development Index (HDI), and vaccination coverage. Their findings demonstrated that improvements in education, immunization programs, and economic development directly contributed to increased life expectancy. The proposed framework confirmed that combining healthcare and socio-economic indicators improves predictive capability compared with models relying solely on medical variables. However, the regression techniques employed were less effective in modeling nonlinear interactions among multiple attributes.

Several researchers have investigated alternative optimization techniques to improve healthcare prediction models. Guo et al. proposed a Support Vector Machine regression model optimized using Particle Swarm Optimization (PSO) for life expectancy estimation. The optimization algorithm improved parameter selection and enhanced prediction accuracy by searching for globally optimal solutions within the parameter space. Although the optimized SVM demonstrated improved performance over traditional regression models, its computational requirements increased significantly when processing larger healthcare datasets containing numerous variables.

Apart from machine learning algorithms, researchers have also examined demographic and socio-economic determinants affecting life expectancy. Freeman et al. analyzed differences in life expectancy across countries with varying income levels and healthcare infrastructures. Their findings emphasized that healthcare accessibility, education, economic stability, and social equality collectively influence national longevity. Similarly, Girum et al. investigated low- and medium-income countries and reported that infectious diseases, inadequate healthcare services, poor sanitation, and

lower educational attainment substantially reduced life expectancy. These studies provided valuable epidemiological insights but did not employ advanced machine learning techniques capable of generating highly accurate predictive models.

Feature engineering has emerged as another important research direction for improving healthcare prediction systems. Recent studies have demonstrated that selecting only the most informative variables enhances prediction accuracy while reducing computational complexity. Decision tree-based feature selection algorithms, particularly the M5P regression tree, have shown excellent capability in ranking variables according to their predictive significance. Such methods eliminate redundant attributes and improve model interpretability by focusing on highly influential factors. Previous investigations have consistently identified HIV/AIDS prevalence, adult mortality, and income composition as dominant predictors influencing life expectancy across different populations.

Hyperparameter optimization has likewise received considerable attention in modern machine learning research. Techniques such as Grid Search, Bayesian Optimization, Genetic Algorithms, and RandomizedSearchCV have been employed to determine optimal parameter combinations that maximize predictive accuracy. Among these approaches, RandomizedSearchCV offers an efficient balance between computational cost and optimization performance by randomly exploring the parameter search space rather than exhaustively evaluating every possible combination. Consequently, it has become a widely adopted strategy for optimizing ensemble learning algorithms, particularly XGBoost, in healthcare prediction tasks.

Although numerous studies have achieved promising results in life expectancy prediction, several limitations remain. Many existing models employ conventional regression algorithms that struggle to capture highly nonlinear relationships among healthcare indicators. Other approaches rely on limited feature selection strategies or insufficient hyperparameter optimization, reducing prediction accuracy and generalization capability. Furthermore, some deep learning frameworks require extensive computational resources and large-scale training datasets, making them less suitable for practical healthcare applications involving structured tabular data. These limitations highlight the need for more efficient machine learning frameworks capable of combining robust preprocessing, effective feature selection, optimized hyperparameter tuning, and powerful ensemble learning techniques.

Motivated by these observations, the present study develops an optimized life expectancy prediction framework based on the XGBoost algorithm. The proposed methodology integrates comprehensive preprocessing, country-wise missing value imputation, M5P-based feature selection, RandomizedSearchCV hyperparameter optimization, and XGBoost regression into a unified prediction model. Unlike many previous studies, the proposed framework simultaneously improves prediction accuracy, computational efficiency, and model interpretability while preserving excellent generalization performance. Experimental evaluation demonstrates that the optimized XGBoost

model achieves an R^2 score of 0.97, MAE of 1.03, and MSE of 2.45, confirming its superiority over several conventional machine learning algorithms and establishing it as an effective solution for life expectancy prediction in both developed and developing countries.

III. Research Methodology

The proposed research methodology presents a systematic machine learning framework for accurately predicting life expectancy using healthcare, demographic, and socio-economic indicators. The framework is designed to improve prediction accuracy by integrating comprehensive data preprocessing, effective feature selection, optimized model training, and rigorous performance evaluation. Unlike conventional statistical approaches, the proposed methodology employs the Extreme Gradient Boosting (XGBoost) algorithm, which effectively models complex nonlinear relationships among multiple healthcare variables. The complete workflow consists of six major stages: data collection, preprocessing, feature selection, data partitioning, model training with hyperparameter optimization, and performance evaluation. Each stage contributes to improving the reliability, robustness, and generalization capability of the prediction model.

The first stage involves data collection, where the life expectancy dataset is obtained from the Kaggle repository, which compiles information published by the World Health Organization (WHO) and the United Nations. The dataset contains records collected between 2000 and 2015 and includes multiple demographic, healthcare, and socio-economic attributes that influence life expectancy. Among the available variables, life expectancy is considered the target variable, while the remaining attributes serve as predictive features. These variables include adult mortality, infant deaths, immunization rates, healthcare expenditure, HIV/AIDS prevalence, income composition, schooling, and several other health-related indicators. Collecting data from reliable international organizations ensures that the prediction model is trained using accurate and representative information.

The second stage focuses on data preprocessing, which improves data quality before model development. Initially, the dataset is organized according to life expectancy values to facilitate efficient analysis. Missing values present in various attributes are handled using a country-specific mean imputation strategy, where unavailable values are replaced with the average value of the corresponding feature within the same country. This approach preserves regional characteristics while reducing the possibility of introducing bias associated with global averaging. Numerical features are subsequently normalized to ensure consistent value ranges across different variables, preventing attributes with larger numerical scales from dominating the learning process. In addition, categorical variables such as development status are encoded into numerical representations suitable for machine learning algorithms. Correlation analysis is also performed to examine the relationships between individual variables and life expectancy, providing valuable insights into the influence of each predictor before model training.

After preprocessing, feature selection is carried out to identify the most influential variables affecting life expectancy. The proposed framework employs the M5P decision tree regression algorithm implemented through the WEKA environment. The M5P algorithm evaluates every feature according to its contribution to predicting the target variable and ranks the attributes based on their predictive importance. Only the highest-ranked variables are retained for subsequent model development, thereby reducing dataset dimensionality, minimizing computational complexity, and improving prediction accuracy. The feature selection process identifies HIV/AIDS prevalence, income composition of resources, and adult mortality as the most significant predictors of life expectancy. Eliminating redundant and less informative variables also reduces the possibility of overfitting while enhancing model interpretability.

Following feature selection, the dataset is divided into training and testing subsets to evaluate the generalization capability of the machine learning model. A random sampling strategy is adopted to partition the dataset, allocating 80% of the records for model training and the remaining 20% for independent testing. Random partitioning ensures that both subsets represent the overall data distribution while preventing sampling bias. The training dataset enables the model to learn complex relationships among healthcare indicators, whereas the testing dataset evaluates prediction performance on previously unseen observations. This separation provides an unbiased assessment of model effectiveness and reflects its practical applicability to real-world prediction tasks.

The next stage involves model training using the Extreme Gradient Boosting (XGBoost) regression algorithm. XGBoost is selected because of its excellent predictive capability, efficient handling of structured data, built-in regularization mechanisms, and ability to model nonlinear feature interactions. To further enhance prediction accuracy, hyperparameter optimization is performed using RandomizedSearchCV with five-fold cross-validation. Various combinations of learning rate, maximum tree depth, number of estimators, minimum child weight, subsampling ratio, and feature subsampling are evaluated to determine the optimal parameter configuration. The selected hyperparameters allow the model to achieve improved convergence while reducing overfitting and maintaining strong generalization performance across unseen data.

Finally, the trained model undergoes performance evaluation using several regression metrics, including Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and the coefficient of determination (R^2). Additional validation techniques such as learning curves, scatter plots, residual analysis, box plots, violin plots, and SHAP-based feature importance analysis are employed to examine prediction accuracy, model stability, and interpretability. Experimental evaluation demonstrates that the optimized XGBoost model achieves excellent predictive performance, producing a testing MSE of 2.45, MAE of 1.03, and R^2 score of 0.97, confirming its effectiveness for life expectancy prediction. The integration of robust preprocessing, M5P-based feature selection, optimized hyperparameter tuning, and XGBoost regression provides a reliable and computationally efficient framework suitable for healthcare analytics and public health decision-making.

IV. The Current System

Existing life expectancy prediction systems primarily rely on conventional statistical methods and basic machine learning algorithms to estimate the average lifespan of populations using demographic, healthcare, and socio-economic indicators. Earlier studies have employed techniques such as Linear Regression, Decision Tree Regression, Support Vector Machine (SVM), Artificial Neural Networks (ANN), Random Forest, and traditional XGBoost implementations to analyze factors influencing life expectancy. Although these methods have achieved acceptable prediction accuracy, their performance is often affected by limitations in data preprocessing, feature selection, and model optimization. Consequently, many existing models struggle to fully capture the complex nonlinear relationships that exist among healthcare variables, reducing their effectiveness when applied to diverse populations.

In most existing systems, the collected healthcare dataset undergoes only basic preprocessing before model development. Missing values are generally replaced using global statistical measures such as the overall mean or median, which may ignore regional variations and introduce bias into the prediction process. Furthermore, several studies utilize the complete set of available variables without performing an effective feature selection process. The inclusion of redundant and less informative attributes increases computational complexity, prolongs training time, and may reduce the generalization capability of the prediction model. Since healthcare datasets often contain highly correlated variables, inadequate feature selection can also increase the risk of model overfitting.

Several existing models also experience difficulties when handling large healthcare datasets containing numerous demographic and clinical attributes. Traditional regression algorithms generally assume linear relationships among variables and therefore fail to model the complex nonlinear interactions commonly observed in public health data. Similarly, algorithms such as Support Vector Machine require careful parameter selection and become computationally expensive as dataset size increases. Deep learning approaches, although highly powerful, often demand substantial computational resources, extensive training data, and longer training times, making them less practical for many real-world healthcare prediction applications.

V. The Suggested System

The proposed system introduces an intelligent machine learning framework for predicting life expectancy by integrating comprehensive data preprocessing, effective feature selection, optimized hyperparameter tuning, and the Extreme Gradient Boosting (XGBoost) regression algorithm. Unlike conventional prediction methods that rely on basic preprocessing and default model parameters, the proposed framework is designed to improve prediction accuracy, reduce computational complexity, and enhance model

generalization. The system processes healthcare, demographic, and socio-economic information through a structured sequence of operations that enables reliable estimation of life expectancy across both developed and developing countries.

The proposed framework begins with the collection of life expectancy data obtained from internationally recognized healthcare repositories. Before model development, the dataset undergoes a comprehensive preprocessing stage to improve data quality. Missing attribute values are handled using a country-specific mean imputation strategy, ensuring that regional healthcare characteristics are preserved while minimizing prediction bias. Numerical attributes are normalized to maintain consistent feature scales, and categorical variables are transformed into numerical representations suitable for machine learning algorithms. This preprocessing stage significantly improves the consistency and reliability of the dataset before it is supplied to the prediction model.

Following preprocessing, an efficient feature selection procedure is performed using the M5P decision tree regression algorithm. Instead of utilizing every available attribute, the M5P model evaluates the predictive significance of each variable and identifies only the most influential features associated with life expectancy. The selected attributes include HIV/AIDS prevalence, income composition of resources, and adult mortality, which exhibit the strongest relationship with the target variable. Removing redundant and less informative features reduces model complexity, shortens training time, and improves prediction performance while preventing unnecessary computational overhead.

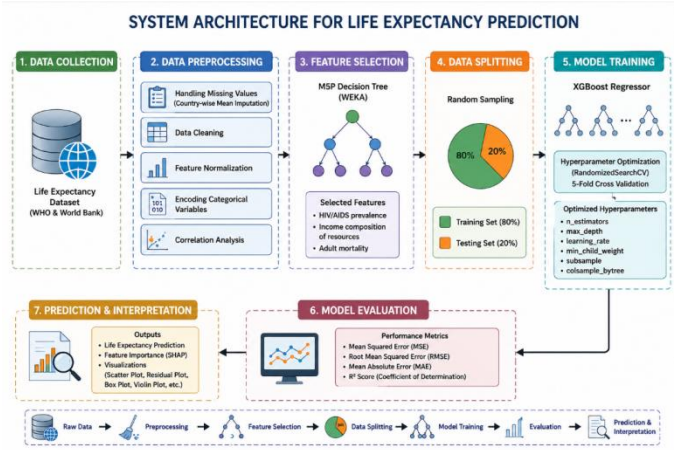
After selecting the optimal features, the dataset is randomly divided into training and testing subsets using an 80:20 ratio. The training dataset enables the machine learning model to learn complex relationships among healthcare indicators, while the testing dataset is reserved for evaluating prediction performance on previously unseen records. This strategy provides an unbiased assessment of the model's generalization capability and ensures reliable validation of the proposed framework.

The core prediction engine of the proposed system is the Extreme Gradient Boosting (XGBoost) regression algorithm. XGBoost is selected because of its exceptional ability to capture nonlinear interactions among multiple healthcare variables while maintaining high computational efficiency. The algorithm constructs an ensemble of gradient-boosted decision trees that iteratively minimize prediction errors through residual learning. Its built-in regularization mechanisms reduce overfitting, allowing the model to produce accurate predictions even when processing complex structured healthcare datasets.

To maximize prediction performance, the proposed framework incorporates RandomizedSearchCV for automatic hyperparameter optimization. Instead of manually selecting parameter values, RandomizedSearchCV evaluates multiple combinations of learning rate, maximum tree depth, number of estimators, minimum child weight, subsampling ratio, and feature sampling parameters using five-fold cross-validation. The

optimal hyperparameter configuration enables the XGBoost model to achieve faster convergence, improved stability, and superior predictive accuracy while maintaining strong generalization across unseen data.

VI. System Architecture



The proposed system architecture presents a structured machine learning pipeline for accurately predicting life expectancy using healthcare, demographic, and socio-economic information. The architecture integrates multiple processing stages, beginning with data acquisition and ending with prediction and performance analysis. Each stage is designed to improve data quality, enhance learning capability, and maximize the predictive performance of the XGBoost regression model. By combining robust preprocessing, efficient feature selection, automated hyperparameter optimization, and comprehensive evaluation, the proposed framework provides a reliable and scalable solution for life expectancy prediction across developed and developing countries.

The first stage of the architecture is data collection, where the life expectancy dataset is obtained from reliable international healthcare sources such as the World Health Organization (WHO) and the World Bank. The dataset contains multiple healthcare, demographic, and socio-economic indicators collected over several years. These variables include adult mortality, infant deaths, immunization coverage, healthcare expenditure, HIV/AIDS prevalence, schooling, income composition of resources, and several other factors that directly or indirectly influence life expectancy. Collecting data from authentic and standardized sources ensures the consistency and credibility of the prediction model.

Following data acquisition, the dataset enters the preprocessing stage, which prepares the raw information for machine learning. Initially, missing values are handled

using a country-wise mean imputation technique to preserve regional healthcare characteristics while reducing data inconsistency. The dataset is then cleaned to remove incomplete and inconsistent records. Numerical attributes are normalized so that all variables contribute equally during model training, while categorical features are converted into numerical values through encoding techniques. Correlation analysis is subsequently performed to examine the relationship between each predictor and life expectancy, providing useful insights into the significance of different healthcare indicators before feature selection.

The processed dataset is then supplied to the feature selection module, where the M5P decision tree algorithm identifies the most influential variables affecting life expectancy. Rather than utilizing every available attribute, the M5P model evaluates the predictive importance of each feature and selects only those with the highest contribution to the prediction process. The selected variables, including HIV/AIDS prevalence, income composition of resources, and adult mortality, provide sufficient information for accurate prediction while reducing computational complexity. Eliminating redundant attributes also improves model efficiency and minimizes the possibility of overfitting.

VII. Summary

This research presented an intelligent machine learning framework for predicting life expectancy by integrating comprehensive data preprocessing, M5P-based feature selection, RandomizedSearchCV hyperparameter optimization, and\ capable of handling nonlinear interactions while maintaining high computational efficiency and strong generalization performance. The adoption of country-wise missing value imputation, normalization, and feature selection further enhanced data quality and reduced computational complexity, allowing the prediction model to learn only from the most informative healthcare indicators.

Experimental evaluation demonstrated that the optimized XGBoost model achieved outstanding predictive performance, producing a testing R^2 score of 0.97, Mean Absolute Error (MAE) of 1.03, and Mean Squared Error (MSE) of 2.45, confirming its effectiveness for life expectancy prediction across both developed and developing countries. Feature importance analysis identified HIV/AIDS prevalence, adult mortality, and income composition of resources as the most influential variables affecting life expectancy, providing valuable insights for healthcare professionals and policymakers. These findings indicate that combining effective preprocessing, intelligent feature selection, and automated hyperparameter optimization significantly improves prediction accuracy compared with many conventional machine learning techniques.

Overall, the proposed framework offers a scalable, interpretable, and computationally efficient solution for healthcare analytics and public health decision-making. The integration of optimized XGBoost regression with systematic data preparation enables accurate estimation of life expectancy while providing transparent information about

the factors influencing longevity. The proposed system can assist governments, healthcare organizations, and researchers in developing evidence-based healthcare policies, optimizing resource allocation, identifying high-risk populations, and implementing effective intervention strategies aimed at improving global health outcomes.

.

References

1. M. M. Ahamad, A. M. Khan, M. F. Rahman, and S. M. Islam, "Machine Learning-Based Life Expectancy Prediction in Developed and Developing Regions," IEEE Access, vol. 13, pp. 30524–30546, 2025.
2. M. Tan and Q. Le, "EfficientNetV2: Smaller Models and Faster Training," Proceedings of the 38th International Conference on Machine Learning (ICML), pp. 10096–10106, 2021.
3. T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), pp. 785–794, 2016.
4. L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.
5. J. H. Friedman, "Greedy Function Approximation: A Gradient Boosting Machine," Annals of Statistics, vol. 29, no. 5, pp. 1189–1232, 2001.
6. N. Landwehr, M. Hall, and E. Frank, "Logistic Model Trees," Machine Learning, vol. 59, no. 1–2, pp. 161–205, 2005.
7. M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA Data Mining Software: An Update," ACM SIGKDD Explorations Newsletter, vol. 11, no. 1, pp. 10–18, 2009.
8. G. James, D. Witten, T. Hastie, and R. Tibshirani, An Introduction to Statistical Learning, 2nd ed. New York, NY, USA: Springer, 2021.
9. T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning, 2nd ed. New York, NY, USA: Springer, 2009.
10. J. Dua and C. Graff, "UCI Machine Learning Repository," University of California, Irvine, CA, USA, 2019.
11. World Health Organization, World Health Statistics 2024: Monitoring Health for the Sustainable Development Goals. Geneva, Switzerland: WHO, 2024.
12. The World Bank, World Development Indicators 2024. Washington, DC, USA: World Bank, 2024.
13. F. Pedregosa, G. Varoquaux, A. Gramfort, et al., "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.
14. L. van der Maaten and G. Hinton, "Visualizing Data Using t-SNE," Journal of Machine Learning Research, vol. 9, pp. 2579–2605, 2008.
15. S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," Advances in Neural Information Processing Systems (NeurIPS), vol. 30, pp. 4765–4774, 2017.