

Human Life Expectancy Prediction Through Machine Learning Models: A Comparative Analysis of Classification and Regression Models

Muhammad Ameen Chhajro (corresponding Author)

Department of Software Engineering, Sindh Madressatul Islam University, Karachi

Email: ameen.chhajro@smiu.edu.pk

Zubair Uddin

Department of Software Engineering, Sindh Madressatul Islam University, Karachi

Email: zubair@smiu.edu.pk

Adnan Jahangir Panhwar

Assistant Professor, Department of Computer Science, The University of Sindh,

Jamshoro. Email: adnanpanhwar1@gmail.com

Farhan Bashir Shaikh

Department of Computer Science, University of Larkano

Email: Farhan@uolrk.edu.pk

Seema Sultana Bhurgri

Assistant Professor, Department of Computer Science, Government Nazareth Girls

Degree College, Hyderabad. Email: seema.bhurgri@gmail.com

Muhammad Hibatullah Channa

Department of Computer Sciences & Related Studies, Hyderabad Institute for

Technology & Management Sciences, Hyderabad. Email: hibatullah@hitms.edu.pk

ABSTRACT

Life expectancy is a very important issue of economic development and public health. The correct measure of life expectancy problems can significantly help in planning and policy-making decisions related to the population's healthcare. In this proposed research, various machine learning models of Regression and classification have been implemented to predict the life expectancy and country development status. This research has used a dataset of 2,938 samples having 22 features, including economic, demographic, and health indicators. The preprocessing techniques were applied for data analysis and cleaning purposes. This study implements the 8 machine learning regression models (Gradient Boosting, Random Forest, Ridge, Linear Regression, Lasso, XGBoost, Decision Tree, and SVR), and 10 Classification models, including SVM, KNN, CatBoost, Naïve Bayes, XGBoost, Gradient Boosting, LightGBM, Random Forest, Decision Tree, and Logistic Regression, were trained on the health

dataset. The experimental results of this research reveal that the regression model XGBoost achieved the highest R^2 score of 0.89 with an RMSE value of 3.25 years, while the classifier XGBoost achieved 96.8 % accuracy in predicting the country development status. This research highlights the effectiveness of multiple machine learning models for task prediction in the public health domain.

Keywords— machine learning, XGBoost, Random Forest, Life expectancy prediction, healthcare analytics, regression models, classification models, public health

1. Introduction

Life expectancy is a very crucial and basic measure of a population's health. It also plays an important role in the overall development status of the countries. However, its accurate prediction has always become challenging due to complex health, economics, and demographic factors (Dolgopolyi, Amaslidou and Margaritou, no date). The common approaches for life expectancy measures usually struggle to catch the nonlinear relationships and interactions between multiple variables. The Machine learning approaches, including the XGBoost and Random Forest, have performed exceptionally and demonstrate sound potential for life expectancy prediction across all developing and developed regions (Ahamad *et al.*, 2025).

The objective of this research study is to develop a robust prediction model, which can predict the life expectancy based on the features, including economic, demographic, and health indicators, to classify countries as developed or developing based on the same features. Identify the most influential factors affecting life expectancy. The recent analysis reveals that 88 research studies have been conducted, in which the machine learning models performed outclass, along with (AUC=0.831); however, the critical gaps still remain. The research discloses that the Only 19% of studies included general population samples, including 89.8% excluded social variables (Delpino *et al.*, 2025). In order to address concerns of model interpretability, some recent research has utilized the SHAP (Shapley Adaptive Explanations) and LIME for the purpose of transparent predictions in public health policy (Tadeparti, Ramadhenu and Vemuri, 2025).

2. Literature Review

Several research studies have employed life expectancy prediction adopting different machine learning approaches. This literature review section elaborates on the existing literature and identifies the research gaps addressed by this proposed research study. Previous research studies on life expectancy mainly used traditional statistical approaches such as time series analysis and linear regression. The limitation of the methods was that they were less interpretable and could not capture the nonlinear relationships between variables. (Bilal *et al.*, 2021). The studies reveal that the Random Forest (RF) achieves the values of ($R^2=0.9423$), outperforming the Linear Regression (LR), and the Decision Tree (Dolgopolyi, Amaslidou and Margaritou, no date). Another research in (Delpino *et al.*, 2025) highlights that the XGBoost performed well and achieved $R^2=0.97$ with hyperparameter tuning. In comparison to other machine learning models from literature review, it is obvious that the Random Forest achieved 97% accuracy, MSE=1.17(Alvi *et al.*, 2025). In another research conducted in its results

Online ISSN

3007-3197

Print ISSN

3007-3189

<http://amresearchreview.com/index.php/Journal/about>

concluded and compared the performance of LightGBM ($R^2=0.905$) vs Random Forest ($R^2=0.984$) (Tadeparti, Ramadhenu and Vemuri, 2025). The literature review in (Romano *et al.*, 2025), it highlights an important analysis that the Gradient Boosting outperformed RF and XGBoost ($r^2=0.55$) while considering the various environmental factors. The advanced research methods reveal that the Multitask LSTM-GRU performed achieved $R^2=0.985$ (El-Rashidy, Sultan and Ali, 2025) for transformer health prediction. The deep learning approaches usually play an important role in complex problems, one author in (Malanin, 2025) claims that MLP and LSTM outperformed Random Forest and XGBoost. In the domain of deep learning approaches, sequence models are very good on many problems, similarly on of the researchers in (Li *et al.*, 2025) elaborates that the CNN-GRU achieved MAE=3.41, RMSE=4.91 for healthcare utilization.

2.1 Research Gaps

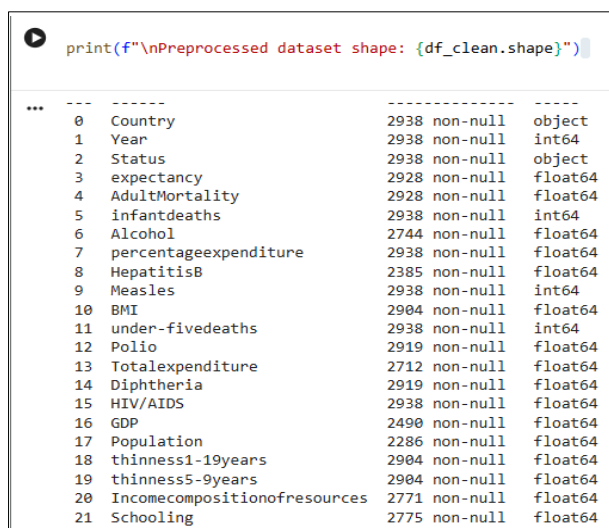
In spite of research advancements in the healthcare and artificial intelligence domains, several gaps remain. Limited studies have been conducted on the use of multiple classification and regression models in parallel in the healthcare domain, along with their interpretability analysis. The feature engineering analysis has also not been focused in depth. This research study covers the highlighted gap and performs a predictive comparative analysis of various regression and classification models, along with model interpretation and feature engineering.

3. Methodology

The research methodology has been divided into multiple sections; each methodology section has been elaborated in future sections of this research study.

3.1 Dataset Description

The corpus used in this proposed research has been scraped online which belongs to WHO databases, which contains multiple categories as mentioned in Figure 1.



...	-----	-----	-----	-----
0	Country	2938	non-null	object
1	Year	2938	non-null	int64
2	Status	2938	non-null	object
3	expectancy	2928	non-null	float64
4	AdultMortality	2928	non-null	float64
5	infantdeaths	2938	non-null	int64
6	Alcohol	2744	non-null	float64
7	percentageexpenditure	2938	non-null	float64
8	HepatitisB	2385	non-null	float64
9	Measles	2938	non-null	int64
10	BMI	2904	non-null	float64
11	under-fiveDeaths	2938	non-null	int64
12	Polio	2919	non-null	float64
13	Totalexpenditure	2712	non-null	float64
14	Diphtheria	2919	non-null	float64
15	HIV/AIDS	2938	non-null	float64
16	GDP	2490	non-null	float64
17	Population	2286	non-null	float64
18	thinness1-19years	2904	non-null	float64
19	thinness5-9years	2904	non-null	float64
20	Incomecompositionofresources	2771	non-null	float64
21	Schooling	2775	non-null	float64

Figure 1. Dataset with different categories

Online ISSN

3007-3197

Print ISSN

3007-3189

<http://amresearchreview.com/index.php/Journal/about>

This research study dataset contains 2939 entries and 22 categories in total. To predict the output, we have applied a decision tree regression model on the given dataset. Four categories that are country name, country status, calendar year, and life expectancy serve as the output features while all the other categories serve as the input features in our prediction model.

3.2 Proposed Research Methodology

This proposed research has been divided into different segments including dataset collection, data cleaning, feature engineering, preprocessing, machine learning models including the Regression and Classification models, and in the end evaluation has been performed in order to check the performance of different models applied on this proposed research study.

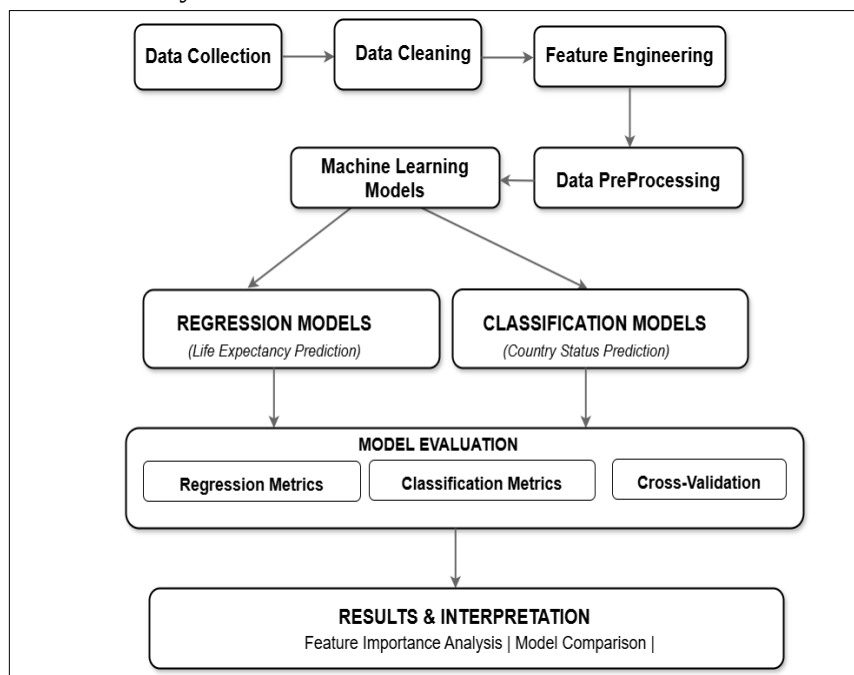


Figure 2: The Proposed Research Methodology Framework

The complete research methodology pipeline has been illustrated in the Figure 2 above. As mentioned the dataset used in this research is obtained from WHO, dataset analysis has been done through python, followed by data cleaning step in order to check any missing values. The feature engineering creates six different indices for particular domain. Data preprocessing consists of label encoding, scaling for numeric features of the data. The proposed research framework of machine learning models splits into two parts including regression and classification models. Regression models proposed for life expectancy prediction and classification models for country status prediction. In the end evaluation and accuracy measurement adopts specific metrics for each task, followed by result comparison via feature importance values analysis.

3.3 Model Architecture

This proposed research adopts the XGBoost architecture from data input layer to model

output layer as illustrated in the Figure 3.

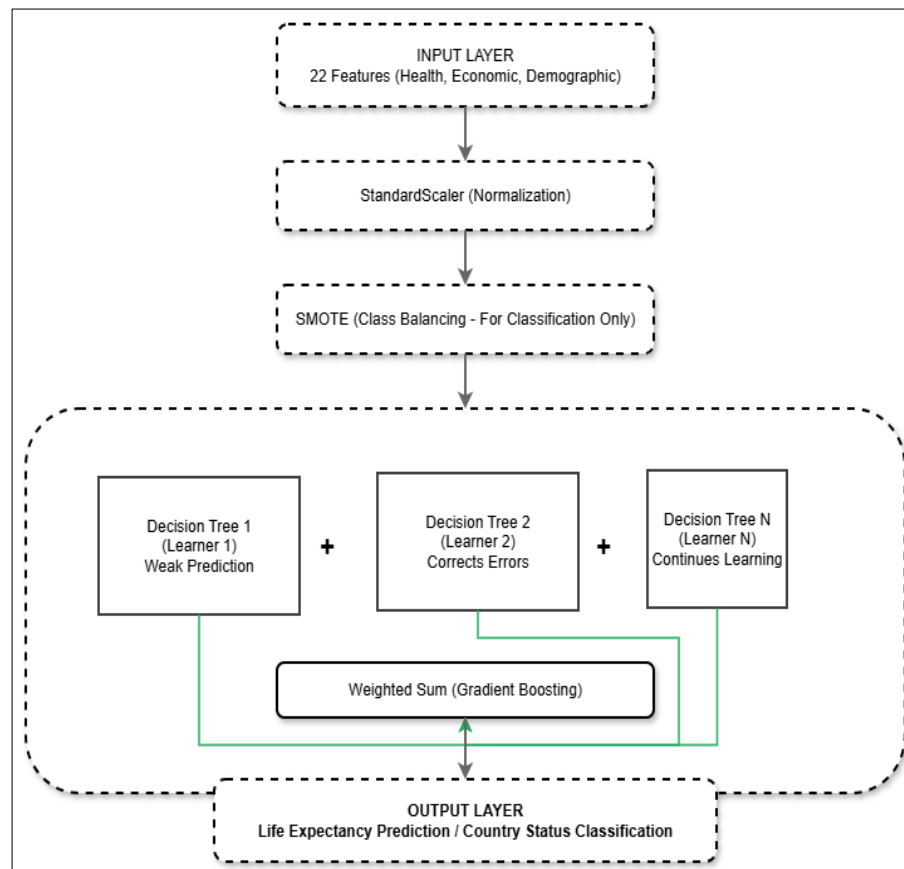


Figure 3. XGBoost Model Architecture for Life Expectancy Prediction

This proposed architecture has 22 input features, which are normalized at next step. The SMOTE technique is used for class balancing in classification tasks of this research. The main architecture is the XGBoost which combines the large numbers of the decision trees (100), where each current tree learns the errors from the previous trees using the gradient boosting approach. The final prediction output is taken on the bases of the weighted sum of all tree predictions.

3.4 Feature Engineering

Some the feature names and their domain significance has been mentioned in the Table 1 below.

Table 1. Some Features Description

S. No	Features Name	Domain Significance
01	Disease Burden	This feature quantifies the overall disease impact
02	Vaccination Coverage	Quantifies overall immunizations

03	Healthcare Quality	Shows the effectiveness of the healthcare system
----	--------------------	--

4. Experimental Results and Analysis

This proposed research adopts different data analysis technique and experimental

Online ISSN

3007-3197

Print ISSN

3007-3189

<http://amresearchreview.com/index.php/Journal/about>

results, which are divided into different sections of this research as given below.

4.1 Exploratory Data Analysis

The exploratory data analysis of this research shows the feature correlation matrix heatmap as illustrated in Figure 4. The correlation matrix illustrates the relationship among 22 features.

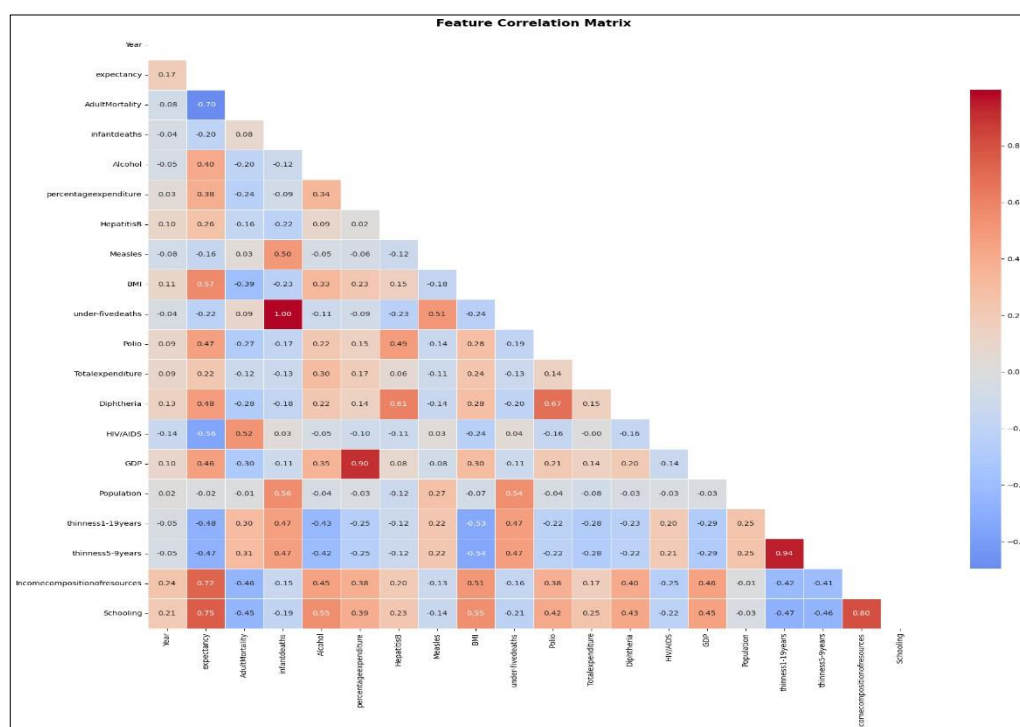


Figure 4. Feature Correlation Matrix

By focusing on the result of correlation Metrix above shows that there is strong positive correlation (0.78) between Income Composition and GDP, which indicate that there is strong association between income distribution and economic development. In opposite to that there is negative correlation between Life Expectancy and Adult Mortality which is (-0.71). The correlation matrix helps identify multicollinearity and guides feature selection decisions.

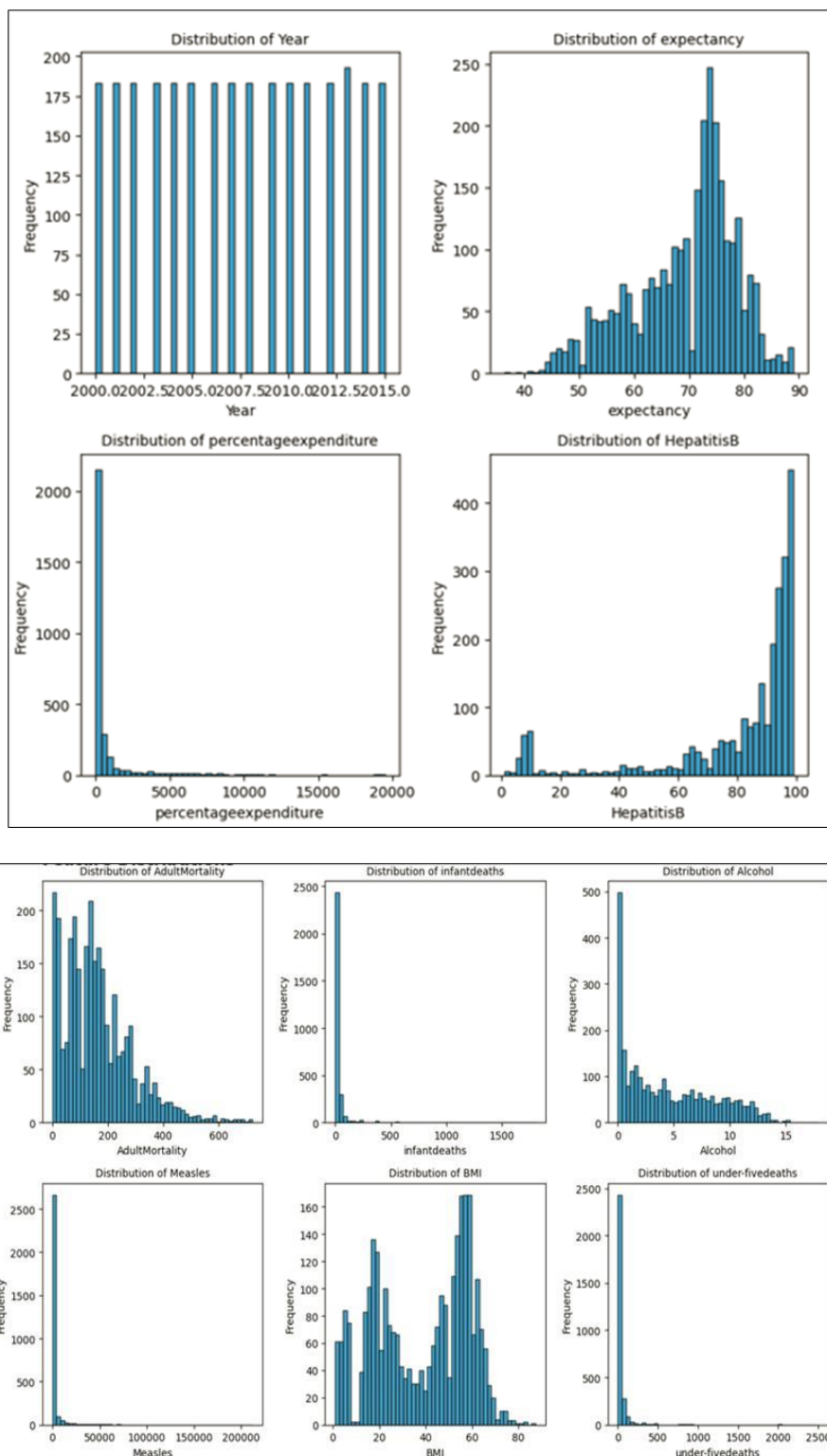
Online ISSN

3007-3197

Print ISSN

3007-3189

<http://amresearchreview.com/index.php/Journal/about>



Online ISSN

3007-3197

Print ISSN

3007-3189

<http://amresearchreview.com/index.php/Journal/about>

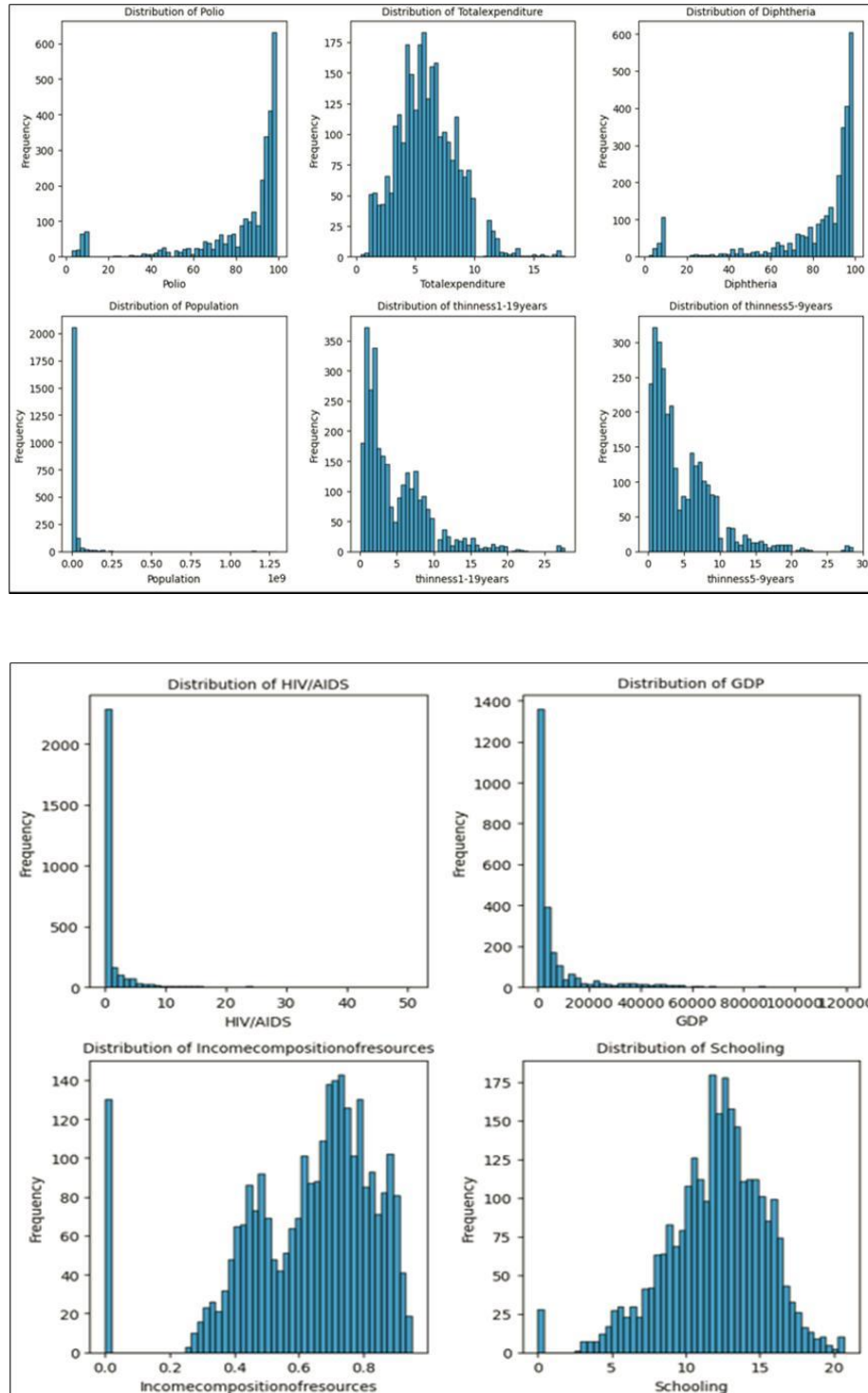


Figure 5. Feature Distributions

The histogram representations in the Figure 5 above show the distribution of each

variable. The key findings coming from the histogram outcomes, show that GDP follows more skewed distribution, which describing the economic disparity. The Adult Mortality reflecting different health patterns describes bimodal distribution. On the other hand, clustering around 10-12 years shows the Schooling years range from 0 to 20.4 years.

4.2 Results Analysis of Regression Models

The performance comparison analysis of regression models prediction results regarding life expectancy measures with different parameters, including the R^2 Score, RMSE (years), MAE (years), and MSE respectively. After comparison of the results between different regression models describes that the XGBoost achieves the highest value of R^2 Score (0.8942) while SVR obtained the lowest one (0.7567) amongst all regression models as illustrated in the Figure 6.

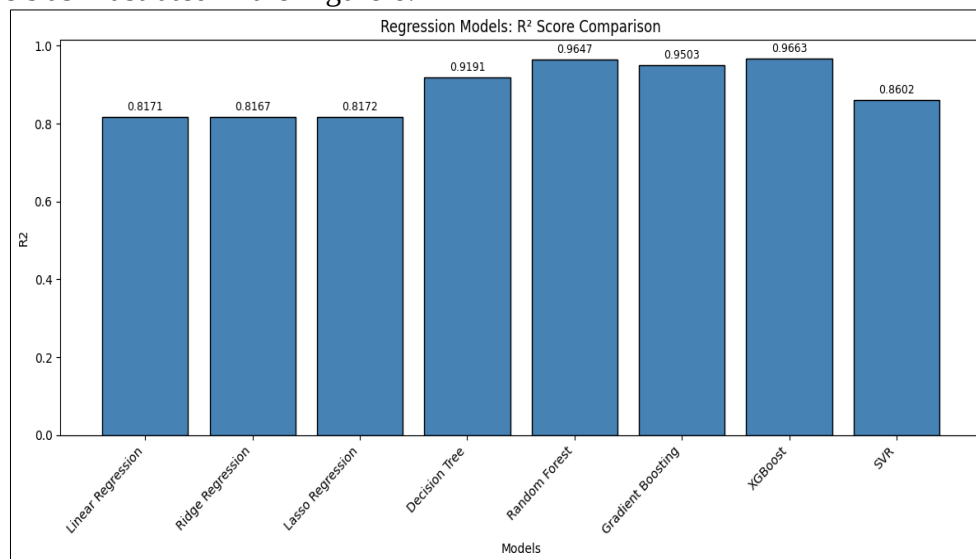


Figure 6. Regression Models R^2 Score Comparison

Ensemble methods (XGBoost, Random Forest, Gradient Boosting) significantly outperformed linear models, demonstrating the nonlinear nature of relationships affecting life expectancy. On the other hand, the RMSE comparison results obtained through bar chart plot as illustrates in the Figure 7.

Online ISSN

3007-3197

Print ISSN

3007-3189

<http://amresearchreview.com/index.php/Journal/about>

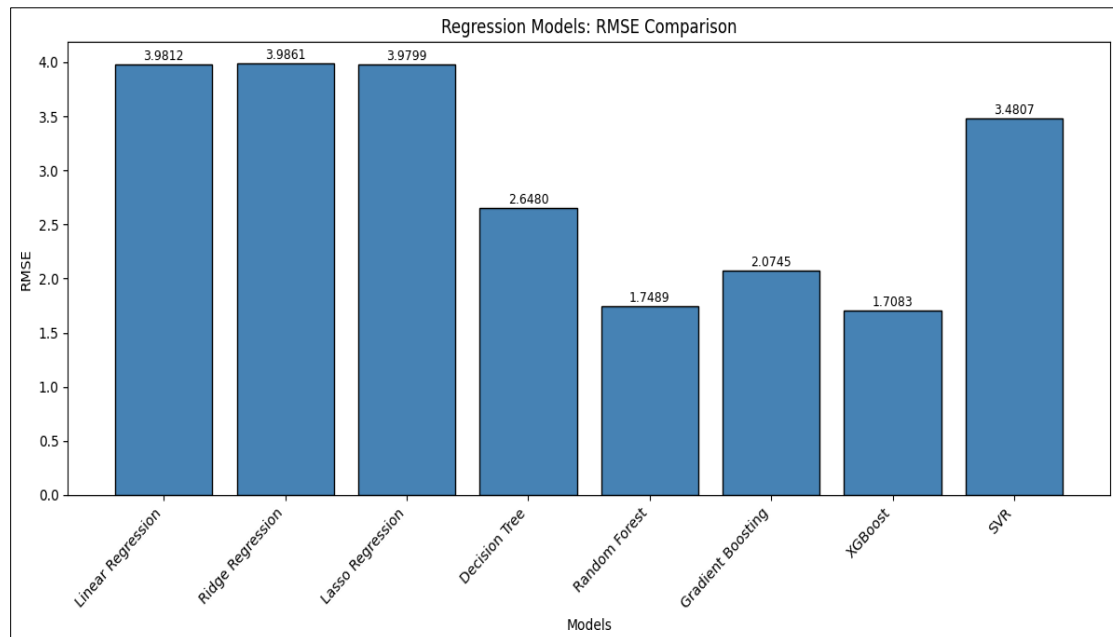


Figure 7. Regression Models Error Comparison

4.3 Results Analysis of Classification Models

The proposed research study reveals that the XGBoost achieved the highest accuracy (96.8%), correctly classifying 968 out of 1,000 countries as developed. Furthermore, the top four other classification models (XGBoost, LightGBM, CatBoost, Random Forest) all achieved over 95% accuracy as illustrated in Figure 8. The experimental results demonstrate the clear separability of country status based on health and economic indicators.

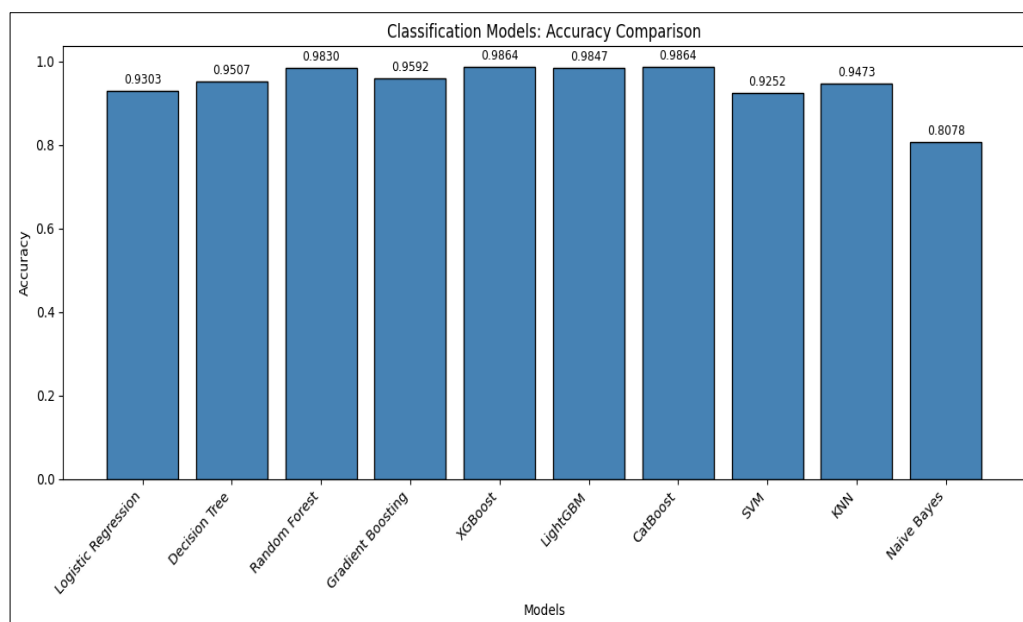


Figure 8. Classification Models Accuracy Comparison

Online ISSN

3007-3197

Print ISSN

3007-3189

<http://amresearchreview.com/index.php/Journal/about>

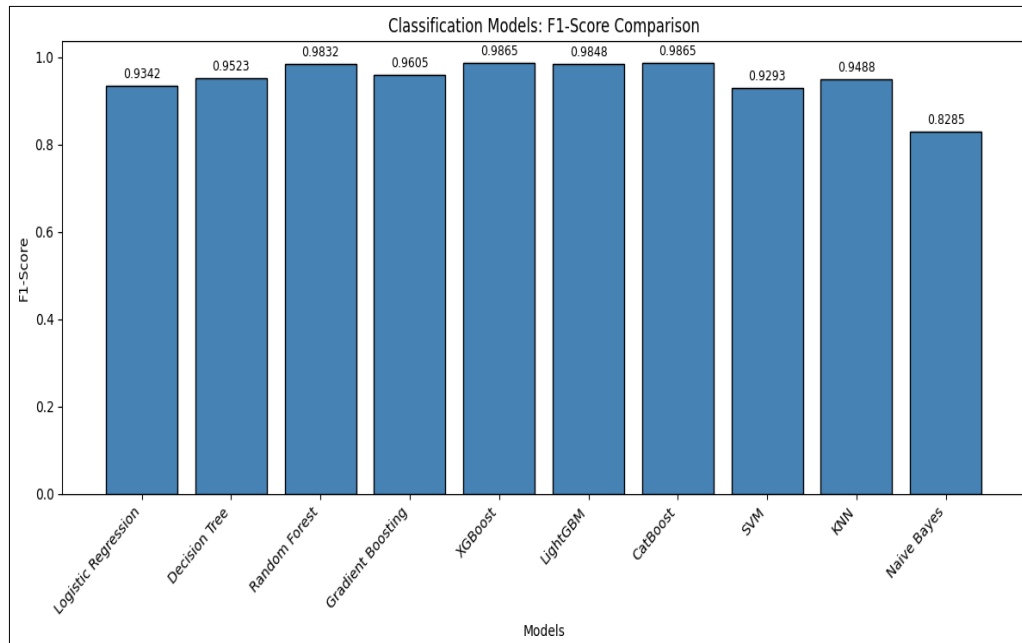


Figure 9. Classification Models F1-Score Comparison

Figure 9 illustrates the F1-scores, which balance precision and recall, and shows that XGBoost achieved a leading position at 0.966. The highest values of consistency in F1-Score across indicate that ensemble methods indicate robust performance without bias toward either class.

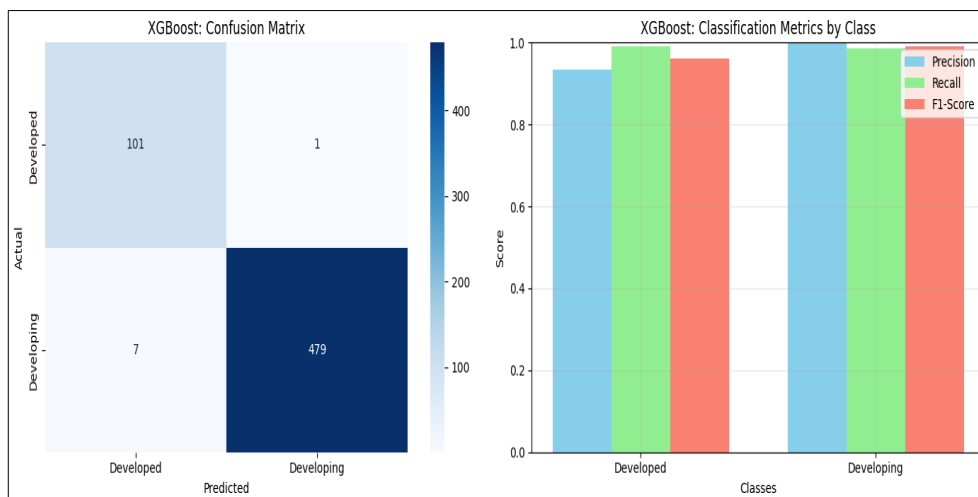


Figure 10. XGBoost Classification - Confusion Matrix

The experimental results obtained from this research study, as illustrated in Figure 10 above, show that the confusion matrix for XGBoost classification consists of two outcomes: the first is the True Positives (Developed: 485) correct predictions, and the

Online ISSN

3007-3197

Print ISSN

3007-3189

<http://amresearchreview.com/index.php/Journal/about>

Second, True Negatives (Developing: 483) are correct predictions. On the other hand, False Positives (29) misclassifications (Developing predicted as Developed). While the False Negatives (3) misclassifications (Developed predicted as Developing). The very low false negative rate (3) indicates the model is particularly good at identifying developed countries, likely due to their distinctive health and economic signatures.

4.5 Cross-Validation Results

Table 2: Cross-Validation Results (5-fold)

Model Type	Best Model	CV Mean Score	Std Deviation	95% Confidence Interval
Regression	XGBoost	0.8856	±0.023	[0.8626, 0.9086]
Classification	XGBoost	0.9632	±0.018	[0.9452, 0.9812]

The consistent performance across folds indicates good generalization and the absence of overfitting.

5. Conclusion

This research presented a comprehensive machine learning approach for predicting life expectancy and classifying country development status. We implemented and compared 8 regression models and 10 classification models, with XGBoost emerging as the best performer for both tasks ($R^2 = 0.894$, Accuracy = 96.8%). Ensemble methods, particularly XGBoost, are highly effective for healthcare prediction tasks involving complex, nonlinear relationships. Adult mortality, HIV/AIDS, income composition, and schooling years are the most critical predictors of life expectancy. Health and economic indicators provide a clear separation between developed and developing countries, enabling accurate classification. Feature engineering with domain-specific indices (Vaccination Coverage, Healthcare Quality, Disease Burden) improves model performance. The models generalize well, as evidenced by consistent cross-validation scores.

References

- Ahamad, S.S. *et al.* (2025) 'Machine Learning-Based Life Expectancy Prediction in Developed and Developing Regions', *IEEE Access*, 13(March), pp. 69520–69531. doi:10.1109/ACCESS.2025.3560890.
- Alvi, M.B. *et al.* (2025) 'Life Expectancy Prediction using Recursive Partitioning and Bagging Algorithms', *ICST Transactions on Scalable Information Systems*, 12(5), pp. 1–18. doi:10.4108/eetsis.8959.
- Bilal, U. *et al.* (2021) 'Life expectancy and mortality in 363 cities of Latin America', *Nature Medicine*, 27(3), pp. 463–470. doi:10.1038/s41591-020-01214-4.
- Delpino, F.M. *et al.* (2025) 'Global performance of machine learning models to predict all-cause mortality: systematic review and meta-analysis', *Scientific Reports*, 15(1), pp. 1–35. doi:10.1038/s41598-025-26714-6.

Online ISSN

3007-3197

Print ISSN

3007-3189

<http://amresearchreview.com/index.php/Journal/about>

- Dolgopolyi, R., Amaslidou, I. and Margaritou, A. (no date) ‘Interpretable Machine Learning for Life Expectancy Prediction : A Comparative Study of Linear Regression , Decision Tree , and Random Forest’.
- El-Rashidy, N., Sultan, Y.A. and Ali, Z.H. (2025) ‘Predicting power transformer health index and life expectation based on digital twins and multitask LSTM-GRU model’, *Scientific Reports*, 15(1), pp. 1–29. doi:10.1038/s41598-024-83220-x.
- Li, V.C. *et al.* (2025) ‘Evaluating Traditional , Machine Learning , and Deep Learning Models for Predicting Outpatient Healthcare Utilization in Europe : A Longitudinal Panel Analysis’, pp. 1–29.
- Malanin, V. (2025) ‘IMPROVING LIFE EXPECTANCY PREDICTIONS : A COMPARATIVE ANALYSIS OF MLP AND LSTM VERSUS LCR AND LCX MODELS’, pp. 432–447. doi:10.5281/zenodo.15789243.
- Romano, D. *et al.* (2025) ‘Predictive and Explainable Machine Learning Models for Endocrine, Nutritional, and Metabolic Mortality in Italy Using Geolocalized Pollution Data’, *Applied System Innovation*, 8(2), pp. 1–13. doi:10.3390/asi8020048.
- Tadeparti, S.V., Ramadhenu, V. and Vemuri, J.P. (2025) ‘Shapley-Based Comparative Analysis of Predictive Models for Life Expectancy Across Multiple Regions’, (2022), pp. 440–448. doi:10.2991/978-94-6463-978-0_38.