International Journal of Statistical Sciences *ISSN 1683-5603*
Vol. 25(1), March, 2025, pp 55-70
DOI: https://doi.org/10.3329/ijss.v25i1.81045

Predicting Life Expectancy using Machine Learning Techniques

Antora Das, Md. Mahfuz Uddin and Md. Rezaul Karim*

Department of Statistics, University of Rajshahi, Rajshahi-6205, Bangladesh
Emails: antora.rustat2222@gmail.com; mahfuz.ru.stat.58@gmail.com
*Correspondence should be addressed to Md. Rezaul Karim
(ORCID: 0000-0001-5461-7709); (Email: mrkarim@ru.ac.bd)

[Received February 20, 2025; Accepted March 15, 2025]

Abstract

Life expectancy is a key measure of a country's overall health, socioeconomic development, and quality of life. The main objective of the study is to identify key factors influencing life expectancy using 'Cleaned-Life-Exp' standardized data from the World Health Organization (WHO) and to compare the performances of life expectancy prediction using various machine learning algorithms. The key influencing features on life expectancy are selected using Boruta and Regularized Random Forest (RRF) algorithms. Eight machine learning models such as Linear Regression (LR), Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Gradient Boosting (GB), XGBoost (XGB), and Neural Network (NN) are evaluated for predictive performance of life expectancy. Evaluation metrics such as the coefficient of determination (R^2), root mean square error (RMSE), and mean absolute error (MAE) are applied to evaluate the performance of the models. Boruta and Regularized Random Forest (RRF) algorithms identified the same 20 significant predictors, including Income Composition of Resources, HIV/AIDS, Adult Mortality, and Schooling, as the most influential features. Among the eight machine learning models evaluated, Random Forest achieves the highest performance (R^2 = 0.969, RMSE = 0.179, MAE = 0.116), highlighting the superiority of ensemble methods. Support Vector Machine (SVM) performs well, while Decision Tree and KNN show moderate performance. Linear Regression and Neural Networks have the lowest predictive performances. This study will help to provide a better predictive framework using machine learning models, which can guide policymakers in improving life expectancy prediction.

Keywords: Life expectancy, Boruta algorithm, Regularized Random Forest algorithm, Linear Regression, Decision Tree, Random Forest, Support Vector Machine, K-Nearest Neighbor, Gradient Boosting, XGBoost, and Neural Network (NN).

AMS Classification: 62P10, 62J05, 68T05.

1. Introduction

Life expectancy is a key measure of a country's overall health, socioeconomic development, and quality of life. It demonstrates the effectiveness of healthcare systems, public policies, and individual health behaviors. Predicting life expectancy is a critical area of research that combines data from various fields, including healthcare, economics, and environmental studies, to estimate

the average number of years an individual or population is expected to live [1]. Several factors, such as genetics, lifestyle choices, healthcare quality, socioeconomic conditions, and environmental factors, influence the prediction of life expectancy [2]. In recent years, advancements in machine learning and statistical modeling have allowed researchers to develop more accurate and nuanced predictions, enabling better public health planning and policy formulation [3]. By analyzing trends and identifying patterns within large datasets, these predictive models can help anticipate changes in life expectancy due to evolving healthcare practices, lifestyle shifts, or environmental conditions, offering valuable insights into the future well-being of populations.

Over the past decades, researchers have increasingly turned to data-driven approaches to understand the factors influencing life expectancy and to predict it accurately. Machine learning (ML) models, with their ability to uncover complex relationships in data, have emerged as powerful tools in this domain [4]. The World Health Organization (WHO) provides comprehensive data that captures various variables impacting life expectancy, including healthcare expenditure, immunization rates, and socioeconomic indices [5]. These variables present an opportunity to delve into the intricate relationships governing life expectancy and to uncover actionable insights through data-driven methodologies. Machine learning has become a robust analytical tool capable of handling complex, non-linear relationships and high-dimensional datasets. By leveraging machine learning techniques, this study aims to predict life expectancy and identify the important factors from a rich dataset of global health indicators [6]. This approach facilitates accurate prediction and highlights key variables that can guide interventions to enhance public health outcomes [7]. This research is motivated by the need to harness advanced computational techniques to address pressing global health challenges. By utilizing various machine learning techniques, the study seeks to advance the understanding of factors affecting life expectancy [8] and provide a robust predictive framework to inform policy-making and resource allocation [9]. Existing studies on life expectancy prediction often neglect feature selection, leading to the inclusion of irrelevant variables, which affects model performance and interpretability. Applying feature selection methods like Boruta or Regularized Random Forest (RRF) can enhance model accuracy, efficiency, and interpretability by identifying key predictors of life expectancy. Most studies focus on a single or limited set of ML algorithms (e.g., linear regression, Random Forest) without evaluating advanced techniques like SVM, XGBoost, or Neural Networks. A comprehensive comparison of multiple ML models can help identify the most suitable approach for life expectancy prediction [10]. The objectives of the study are to identify key factors influencing life expectancy using standardized World Health Organization (WHO) data and to compare the performances of life expectancy prediction using various machine learning algorithms. The workflow diagram of the study process is presented in Figure 1.

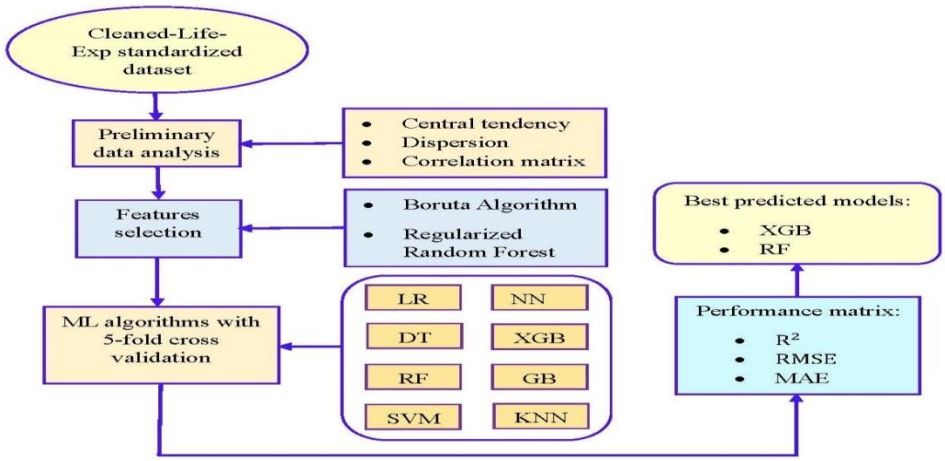

Figure 1: Proposed study process diagram

The rest of the paper is organized as follows: Section 2 explains the materials and methods, including the data source, variables, feature selection techniques, and machine learning models. Section 3 details the results and provides a discussion. Section 4 presents the study's conclusion along with potential future extensions.

2. Materials and Methods

2.1 Data Source and Study Variables

The completeness and accuracy of the data are critical to the outcome of this research. The data utilized in this study was sourced from the World Health Organization's (WHO) Global Health Observatory (GHO) data repository, which monitors health indicators and associated factors for nations all over the world. The United Nations was the source of the relevant economic data. This study concentrates on the health-related variables that are most representative and essential to comprehending life expectancy out of the wide range of accessible variables. Data from 193 countries are included in the dataset, which has been painstakingly combined into a single file with 2,938 rows and 22 columns that reflect 20 predictive variables. Economic, social, mortality, and immunization-related factors were the four primary categories into which the variables were divided. By creating a comprehensive dataset with no missing values and using data from 193 countries over numerous years, this study aims to close these gaps. In addition to important vaccinations like Hepatitis B, Polio, and Diphtheria, this dataset includes variables like GDP, education, and health spending. By determining the most important life expectancy predictors, this method enables nations to rank the treatments that have the best chance of enhancing the health and longevity of their citizens [11].

2.2 Statistical Analysis

Some graphical representations (e.g., histogram) and measures of central tendency, dispersion, skewness, and kurtosis are employed to investigate the characteristics of the variables. The relationship between all the variables is measured using Pearson’s correlation coefficient, and the resulting correlation matrix is then displayed as a heatmap using the R corrplot tool.

2.3 Features Selection

Feature selection is crucial in regression and classification to improve model accuracy by excluding irrelevant predictors.

2.3.1 Boruta Algorithm

The Boruta algorithm, based on Random Forest, is a notable method of feature selection [12]. It uses shadow features and random duplicates of the original variables to build decision trees. Feature relevance is assessed by the reduction in model performance caused by shadow features. Z-scores are calculated by dividing the mean accuracy loss by its standard deviation, helping identify essential predictors. The Z-score is regarded as the key metric in the Boruta method. As a result, the set of shadow attribute importance is utilized as a guide to determine how important original attributes are. Next, the highest significance of shadow characteristics is compared with the significance of original features [13]. The step-by-step procedure of the Boruta algorithm is given in [14].

2.3.2 Regularized Random Forest (RRF) Method

The RRF method is an advanced feature selection approach that enhances traditional random forests by incorporating regularization to promote sparse models [15]. Let $\mathrm{gain}(X_j)$ represent the evaluation metric for a given feature X_j. Define F as the set of features previously used for splits in a tree model. The modified measure is defined as:

$$\mathrm{gain}_R(X_j) = \begin{cases} \lambda.\,\mathrm{gain}(X_j) & X_j \notin F \\ \mathrm{gain}(X_i) & X_j \in F \end{cases},$$

where $\lambda \in [0, 1]$ is referred to as a regularization parameter. A lower value of λ imposes a higher penalty on features that are not part of F. The use of $\mathrm{gain}_R(.)$ to determine the splitting feature at each tree node is known as the tree regularization framework. Further details can be found in [15,16].

2.4 Machine Learning Techniques

2.4.1 Linear Regression

Linear Regression is a supervised machine learning technique to predict continuous target variables. It establishes a mathematical relationship between a dependent variable *y* (also known as the response) and one or more independent variables *X* (also known as predictors or explanatory variables) through a linear equation [17].

2.4.2 Decision Tree

A decision tree is a supervised learning method that can be applied for both regression and classification tasks. A decision tree is a supervised learning technique for classification and regression problems. This framework, similar to a flowchart, is used to make decisions or predictions. It includes nodes that indicate decisions or attribute tests, branches that show how these choices turned out, and leaf nodes that provide the results or forecasts. Each internal node signifies a condition applied to an attribute, every branch represents a possible outcome of that condition, and each leaf node corresponds to either a specific class label or a numerical value. The root node stands for the complete dataset, and the primary choice that needs to be created is for inside nodes to show judgments or attribute tests. There are one or more branches on each internal node; branches show the result of a test or choice that leads to a different node, and leaf nodes

stand for the ultimate judgment or forecast. At these nodes, no more splits take place. The step-wise description of the decision tree is given in [18].

2.4.3 Random Forest

Random Forest is an ensemble learning method that constructs multiple decision trees and aggregates their outputs to enhance prediction accuracy and stability. As a supervised machine learning algorithm, it can be applied to both classification and regression problems [19]. It employs bootstrap sampling, where multiple subsets of the original dataset are created through random sampling with replacement. A random subset of features is chosen at each split to reduce the correlation among trees instead of using all available features [20]. Each decision tree is then developed based on a predefined splitting criterion.

Gini impurity, $G = 1 - \sum_{i=1}^{n}(p_i)^2$ for classification and $MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ for regression.

Then, for classification, it takes a majority vote of the predictions from all trees, and for regression, it computes the average of predictions from all trees as $\hat{y}_{RF} = \frac{1}{T}\sum_{t=1}^{T}\hat{y}_t$, where T is the total number of trees and $\hat{y}_t$ is the prediction from the tree t.

A graphical representation of the random forest is available in [21].

2.4.4 Support Vector Machine

The Support Vector Machine (SVM) identifies the optimal hyperplane that provides the greatest separation between data points of various classes in the feature space [22]. Mathematically, the equation for a linear hyperplane is expressed as $w^T x + b = 0$, where w denotes the weight vector (which is perpendicular to the hyperplane), b is the bias term, and x represents the input feature vector. The distance between the data point x_i and the decision boundary is determined using the following formula [23]

$$d_i = \frac{w^T x_i + b}{||w||},$$

where $||w||$ is the weight vector w's Euclidean distance norm. The normal vector w for the linear classifier has the following Euclidean norm:

$$\hat{y} = \begin{cases} 1: & w^T x + b \geq 0 \\ 0: & w^T x + b < 0 \end{cases}$$

We can optimize the hard margin linear SVM classifier as:

$$\underset{w,b}{\text{minimize}} \frac{1}{2}||w||^2 \text{ subject to: } y_i(w^T x_i + b) \geq 1 \text{ for } i = 1, 2, \ldots, m.$$

For the soft margin linear SVM classifier, we can optimize it as:

$$\underset{w,b}{\text{minimize}} \frac{1}{2}||w||^2 + C\sum_{i=1}^{m}\xi_i \text{ subject to: } y_i(w^T x_i + b) \geq 1 - \xi_i \text{ and } \xi_i \geq 0, \text{ for } i = 1, 2, \ldots, m.$$

For dual problems, we can optimize it as:

$$\underset{\alpha}{\text{maximize}} \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j t_i t_j K(x_i, x_j) - \sum_{i=1}^{m}\alpha_i,$$

where $\sum \alpha_i$ denotes the sum of all Lagrange multipliers, $K(x_i, x_j)$ is the kernel function that calculates the degree of similarity between two samples x_i and x_j, and α_i is the Lagrange

multiplier related to the *i*th training sample. The support vectors are the training samples with $i > 0$, and the decision boundary is provided by:

$$w = \sum_{i=1}^{m} \alpha_i t_i K(x_i, x) + b \quad \text{and} \quad t_i(w^T x_i - b) = 1 \Rightarrow b = w^T x_i - t_i .$$

A visual depiction of the SVM can be found in [24].

2.4.5 K-Nearest Neighbor

The K-Nearest Neighbor (KNN) algorithm is a straightforward, non-parametric, supervised, and instance-based machine learning technique for classification and regression. It determines the class or value of a given data point by analyzing the k nearest neighbors within the feature space [25].

2.4.6 Gradient Boosting

Gradient Boosting is a robust supervised machine learning technique applicable to both regression and classification problems. It constructs an ensemble of weak learners, typically decision trees, in a stepwise fashion, with each successive model reducing the errors of the previous one by optimizing a loss function [26]. It includes the following steps [26,27,28]:

i. Assume that X and Y are the input and target, respectively, with N samples. Finding the function f(x) that converts the input characteristics X to the target variables y is our aim. It is the total of trees, or boosted trees. The difference between the actual and anticipated variables is known as the loss function, $L(f) = \sum_{i=1}^{N} L\big(y_i, f(x_i)\big)$.

ii. We want to minimize the loss function L(f) with respect to f as $\hat{f}_0(x) = \operatorname{argmin}_f L(f) = \operatorname{argmin}_f \sum_{i=1}^{N} L\big(y_i, f(x_i)\big)$. If this algorithm is in M stages then to improve the f_m, it adds some new estimator as h_m having $1 \le m \le M$ and $\hat{y}_i = F_{m+1}(x_i) = F_m(x_i) + h_m(x_i)$.

iii. For M stage gradient boosting, the steepest Descent finds $h_m = -\rho_m g_m$ where ρ_m is constant and known as step length and g_m is the gradient of loss function L(f) and $g_{im} = -\left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\right]_{f(x_i)=f_{m-1}(x_i)}$.

iv. The gradient similarity for M trees:

$$f_m(x) = f_{m-1}(x) + \left(\operatorname*{argmin}_{h_m \epsilon H} \left[\sum_{i=1}^{N} L\big(y_i, f_{m-1}(x_i) + h_m(x_i)\big)\right]\right)(x) \text{ and}$$

the current solution will be $f_m = f_{m-1} - \rho_m g_m$. Here the notations such as $F_m(x), h_m(x), g_m(x)$, etc., represent the models, weak learners, and gradients, respectively.

A flowchart of Gradient Boosting can be found in [29].

2.4.7 Extreme Gradient Boosting (XGBoost)

XGBoost is a powerful and scalable machine learning algorithm that enhances the Gradient Boosting framework for improved efficiency [30]. It introduces several enhancements, such as regularization, efficient computation, and advanced optimization techniques, to improve performance and reduce overfitting [31].

2.4.8 Neural Network

Neural Networks (NNs) are supervised machine learning algorithms inspired by the structure and activities of the human brain [32]. They are designed to approximate complex functions and are

used for classification, regression, and deep learning applications. Neural Networks consist of layers of interconnected nodes (neurons), where each neuron processes inputs and produces an output passed to the next layer [33]. Here, the input layer receives the input features $X\left(e.g., x_1, x_2, \ldots, x_p\right)$, hidden layers contain neurons that transform inputs using weights, biases, and activation functions, and finally, the output layer produces the final prediction $\hat{y}$ (classification or regression). The parameters, weights (w) control the influence of each input on the neuron, and bias (b) shift the activation function to allow better fitting of the data. The common activation functions are [34]:

$$\text{sigmoid: } \phi(z) = \frac{1}{1+e^{-z}},$$
$$\text{ReLU: } \phi(z) = \max(0, z),$$
$$\text{Tanh: } \phi(z) = \frac{e^{z} - e^{-z}}{e^{z} + e^{-z}},$$

and

$$\text{softmax: } \phi(z_k) = \frac{e^{z_k}}{\sum_{j=1}^{K} e^{z_j}} \text{ (used for classifiation).}$$

2.5 Model evaluation metrics

The following evaluation matrices are applied here.

2.5.1 Coefficient of Determination (R^2)

The coefficient of determination, or R^2, assesses the accuracy of a model's predictions in relation to the actual data. It shows the proportion of variation in the dependent variable that the independent variables can explain. The values of R^2 range from 0 to 1, with values near 1 suggesting that the model fits well and explains most of the variance. Mathematically, R^2 is calculated as [35]:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

2.5.2 Root Mean Square Error (RMSE)

RMSE quantifies the standard deviation of the residuals (prediction errors), assessing the difference between predicted and actual values. A lower RMSE suggests better performance of the model. The formula for RMSE is as follows [36]:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

2.5.3 Mean Absolute Error (MAE)

MAE means the average of the absolute differences between the predicted and actual values. It assesses the size of prediction errors without accounting for their direction. Smaller MAE values suggest superior model performance. MAE is calculated by the formula [37]:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

where, for R^2, RMSE, and MAE, n is the total number of observations, y_i is the actual value of the target variable for the ith observations, $\hat{y}_i$ is the predictive value of the target variable for the i-the observations and $\bar{y}$ is the mean of the actual values of the target variable.

3. Results and Discussion

3.1 Preliminary Data Analysis

Table 1 summarizes the descriptive statistics of the variables, all standardized for comparability. The mean values are close to zero, and the standard deviations are approximately one, confirming successful normalization. The skewness values indicate that Population (17.880), Infant deaths (9.777), and Measles (9.432) exhibit high positive skewness, while Polio (-2.074) and Diphtheria (-2.048) show left-skewed (negatively skewed) distributions. Kurtosis values highlight heavy-tailed distributions, particularly in Population (378.495) and Infant deaths (115.760). Variables such as Year (-1.215) and BMI (-1.302) show low kurtosis, suggesting flatter distributions. These distributional characteristics imply that extreme values may impact predictive modeling. The presence of heavy tails in variables like Under-five deaths (109.487) and HIV/AIDS (34.805) suggests potential data irregularities. Normality assumptions may not hold for variables with extreme deviations, requiring alternative statistical methods. Identifying skewed and heavy-tailed distributions helps in selecting appropriate machine learning models. Proper handling of non-normal distributions ensures more reliable predictions. The presence of high kurtosis in mortality-related variables suggests potential health disparities. Socioeconomic factors may contribute to the observed variation in Life Expectancy. Accounting for such statistical properties enhances model interpretability and generalization.

Table 1: Summary statistics for all variables in the data set

Variable Name	Min.	Max.	Mean	Med.	SD	Skew.	Kurt.
Year	-1.629	1.622	0.000	0.104	1.000	-0.006	-1.215
Status	-0.459	2.177	0.000	-0.459	1.000	1.716	0.947
Life expectancy	-3.458	2.077	0.0001	0.302	0.999	-0.639	-0.236
Adult Mortality	-1.318	4.492	-0.0004	-0.167	0.998	1.175	1.752
Infant deaths	-0.257	15.010	0.000	-0.232	1.000	9.777	115.760
Alcohol	-1.767	3.274	-0.019	-0.240	0.986	0.613	-0.722
Percentage expenditure	-0.371	9.429	0.000	-0.339	1.000	4.647	26.506
Hepatitis B	-3.189	0.806	-0.107	0.312	1.008	-1.604	1.639
Measles	-0.211	18.296	0.000	-0.209	1.000	9.432	114.582
BMI	-1.862	2.444	-0.012	0.233	1.002	-0.202	-1.302
Under five deaths	-0.262	15.322	0.000	-0.237	1.000	9.485	109.487
Polio	-3.396	0.702	-0.005	0.446	1.000	-2.074	3.687
Total expenditure	-2.229	4.669	-0.003	-0.067	0.967	0.637	1.392
Diphtheria	-3.387	0.703	-0.005	0.450	1.000	-2.048	3.465
HIV / AIDS	-0.323	9.624	0.000	-0.323	1.000	5.391	34.805
GDP	-0.828	7.828	-0.059	-0.393	0.933	3.519	14.997
Population	-0.479	21.002	-0.010	-0.162	0.885	17.880	378.495
Thinness 1-19 years	-1.072	0.547	0.009	-0.326	0.998	1.681	3.878
Thinness 5-9 years	-1.058	5.264	0.008	-0.326	0.998	1.749	4.274
Income composition of resources	-2.976	1.704	-0.009	0.225	1.002	-1.079	1.180
Schooling	-3.571	2.593	-0.004	0.091	1.008	-0.567	0.684

The following Figure 2 shows the histogram of our dependent variable 'Life expectancy'. From the histogram, we observe that the distribution of 'Life expectancy' is slightly negatively skewed and platykurtic as the majority of the data points are concentrated toward the right side of the distribution and is flatter compared to a normal distribution.

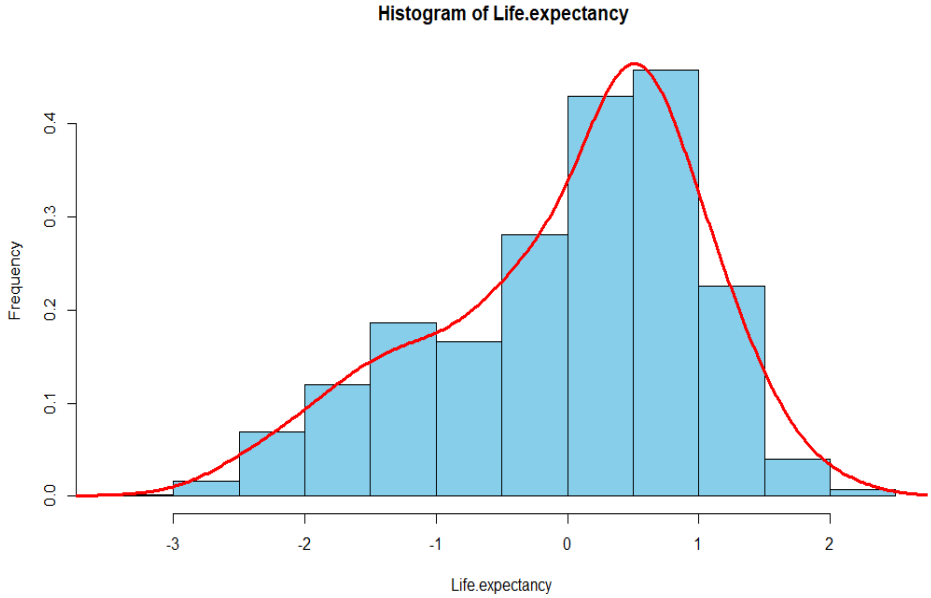

Figure 2: Histogram of the dependent variable Life expectancy (standardized)

3.2 Correlation Measure

The following Figure 3 represents the correlation plot of all the variables in the dataset and also the correlation plot of the dependent variable 'Life expectancy' with the corresponding variable. From the two plots, we observed that the correlation between 'Under-five deaths' and 'Infant deaths' is the highest (0.996), i.e. they are almost perfectly positively correlated. Also, we see that the value of correlation between 'GDP' and 'Percentage expenditure' is 0.899, i.e., they are very strongly positively related. The relationship between 'Year' and 'Status' is the lowest (-0.002), i.e., poorly negatively correlated.

From Figure 4, we observe that the correlation coefficient between the dependent variable 'Life expectancy' and 'Schooling' is the highest (0.768), i.e., they are strongly positively related to each other; and with 'Adult Mortality' is -0.697, which indicates the relationship between 'Life expectancy' and 'Adult Mortality' is negatively related. Finally, the value correlation coefficient between the two variables 'Life expectancy' and 'Population' is the lowest (-0.028), i.e., they have a poor negative relationship.

[Here, we abbreviated some variables name with large explanation in short form as A.Mortality = Adult Mortality, I.deaths = Infant deaths, P.expenditure = Percentage expenditure, U5.deaths = Under-five deaths, T.expenditure = Total expenditure, Thinness1 = Thinness 1-19 years, Thinness2 = Thinness 5-9 years, Income = Income composition of resources]

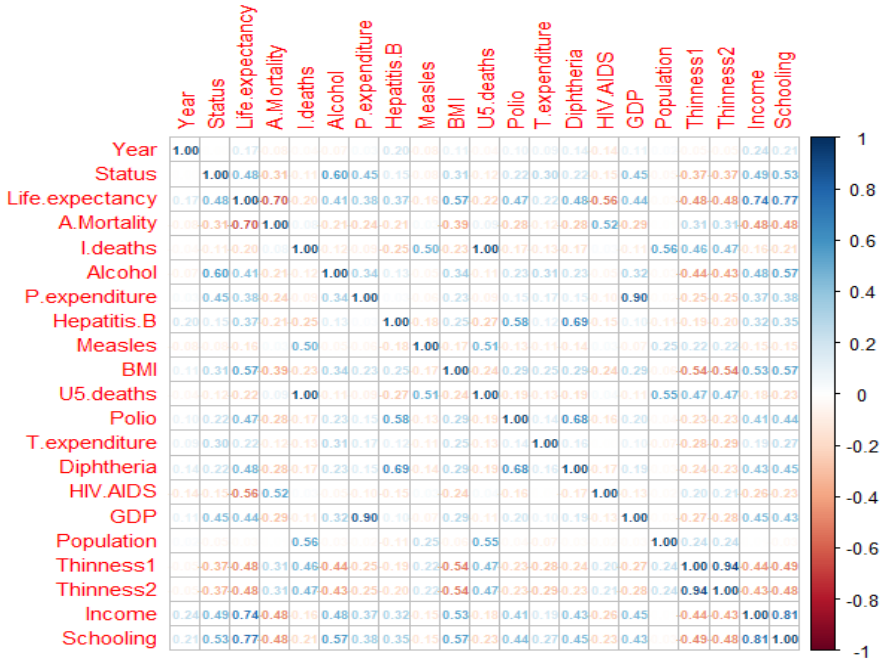

Figure 3: Correlation plot of all variables in the dataset

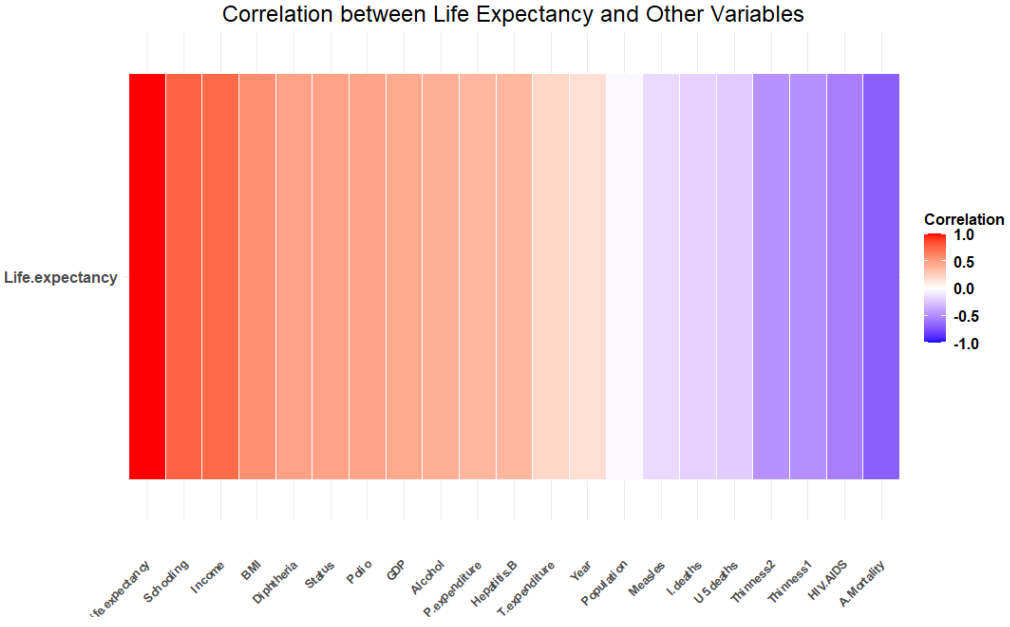

Figure 4: Correlation plot between dependent and corresponding to all other variables

3.3 Features Selection

From the following Figure 5, we see that the Boruta algorithm identified 20 significant features influencing the dependent variable 'Life expectancy'. Variables such as HIV/AIDS, Adult Mortality, Income Composition of Resources, Alcohol, Total expenditure, and Thinness 5-9 years are observed to have high importance scores, indicating a strong influence on 'Life expectancy'. Features like GDP, Population, Polio, and Hepatitis-B show moderate levels of importance. These variables also contribute meaningfully to the model. Variables such as Under-five deaths and Status were assigned lower importance, suggesting minimal relevance for predicting 'Life expectancy'.

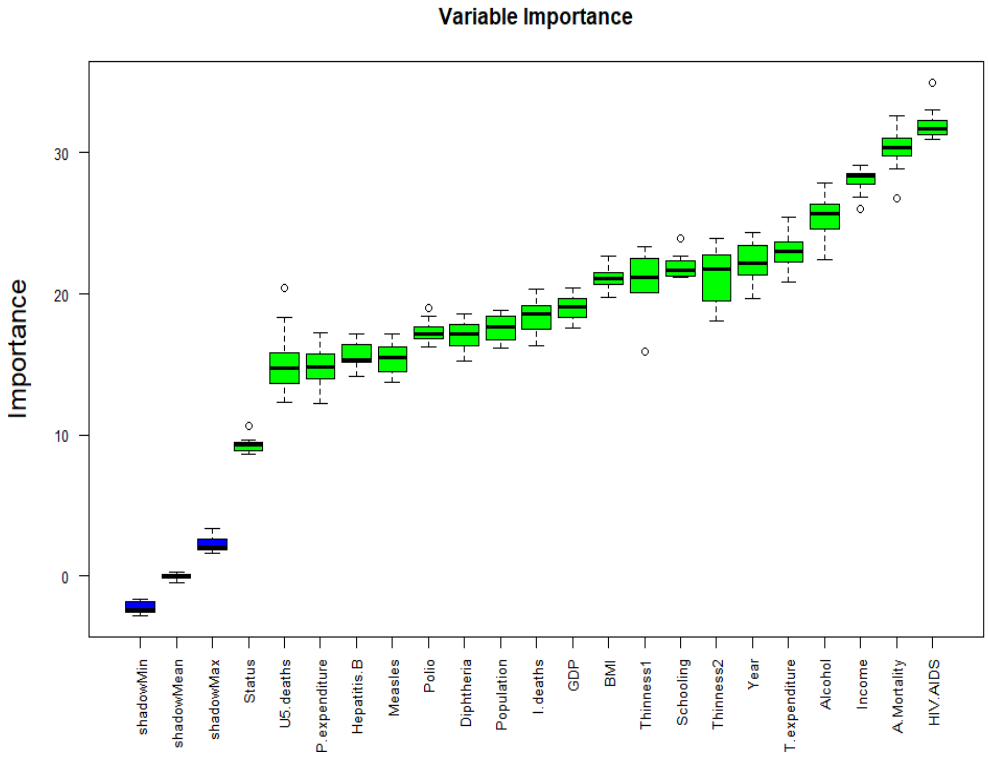

Figure 5: Feature selection by Boruta algorithm

Also, Figure 6 shows that the Regularized Random Forest (RRF) algorithm identified 20 significant features based on their contribution to model accuracy. The most influential predictors include Income Composition of Resources, HIV/AIDS, Adult Mortality, and Schooling, which exhibit the highest mean decrease in accuracy. Other important features like BMI, Thinness indicators, and Under-five deaths were also selected. So, all the 20 features contribute meaningfully more or less to predicting 'Life expectancy'.

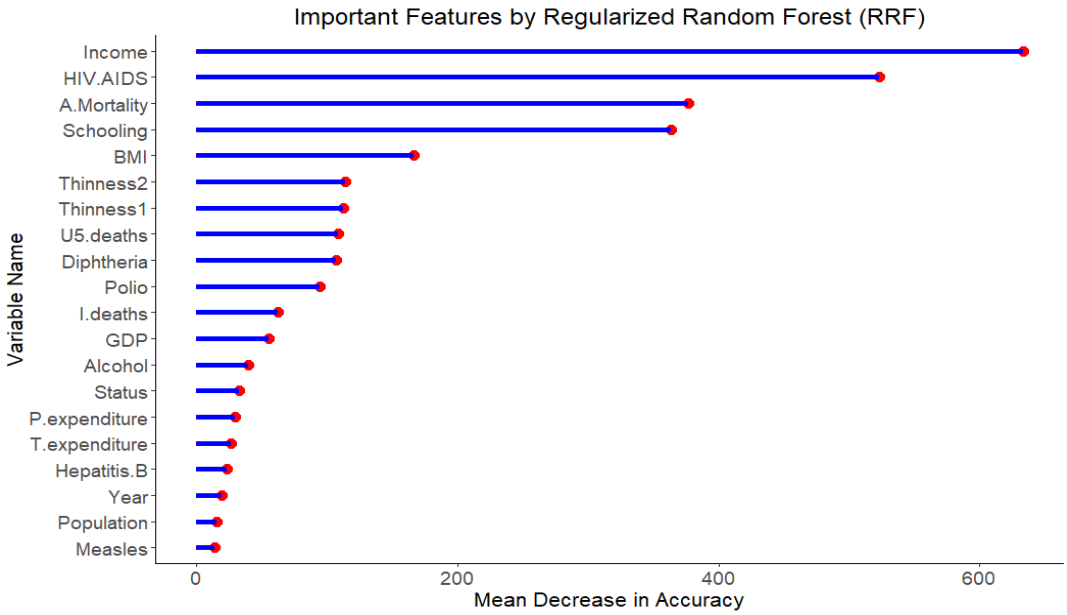

Figure 6: Feature selection by Regularized Random Forest (RRF) algorithm

3.4 Model Performance

At first, the dataset is divided into two parts: 80% as the training set and 20% as the testing set. Then, the models are trained based on the training set, and their performances are assessed based on the testing set. The following Table 2 shows the performance of eight machine learning models. Here, we find that Random Forest performs best with the highest $R^2 = 0.969$ and lowest RMSE $= 0.179$ and MAE $= 0.116$. XGBoost and Gradient Boost show strong performance ($R^2 \approx 0.96$). That is, the proportion of variation of the dependent variable (Life expectancy) is about 97% explained by the independent variables (features) in the model. SVM performs well but is slightly lower than ensemble methods. Decision Tree and KNN have moderate performance. Linear Regression and Neural Networks show the lowest performance in predicting life expectancy. Finally, we conclude that Random Forest, XGBoost, and Gradient Boosting are the most accurate models for predicting Life expectancy, as they demonstrate the highest R^2 values and the lowest errors.

Table 2: Performance of eight machine learning models

Model Name	R^2	RMSE	MAE
Linear Regression	0.842	0.403	0.298
Decision Tree	0.930	0.268	0.183
Random Forest	0.969	0.179	0.116
Support Vector Machine (SVM)	0.946	0.234	0.163
K-nearest neighbor (KNN)	0.926	0.277	0.175
Gradient Boosting	0.959	0.203	0.146
XGBoost	0.961	0.202	0.147
Neural Network	0.907	0.309	0.229

The following Figure 7 represents the bar chart that visually compares the performance of eight machine learning models for predicting Life expectancy using three metrics R^2, RMSE, and MAE. We see that Random Forest is the most accurate model, followed closely by XGBoost and Gradient Boosting. Linear Regression and Neural Networks have the lowest predictive accuracy. That is, the results highlight that ensemble methods are the most effective for predicting Life Expectancy with higher R^2 and lower errors.

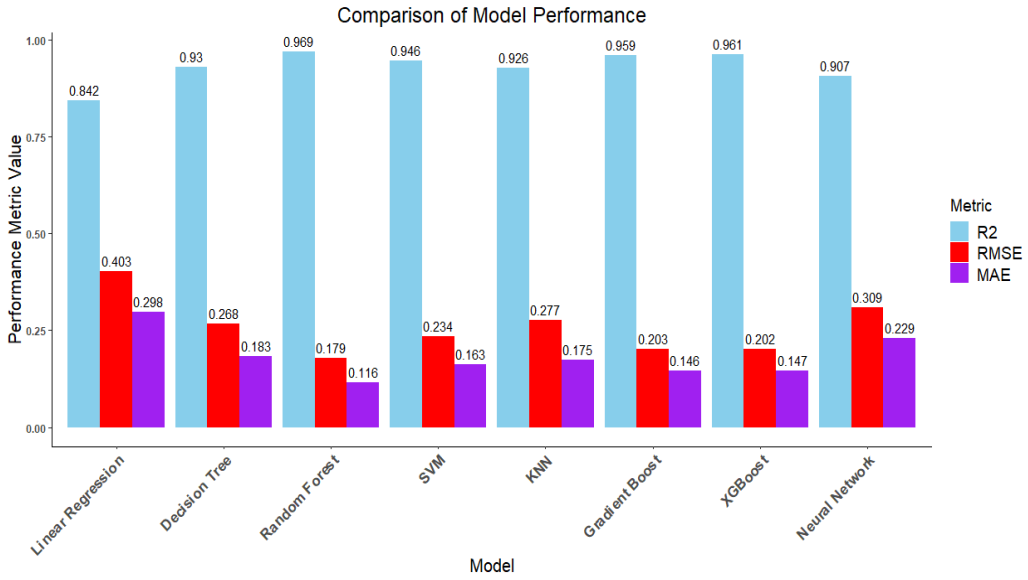

Figure 7: Comparison of performance of eight machine learning models

4. Conclusion and Future Research

This study successfully predicted Life expectancy using machine learning algorithms and identified key influencing factors based on the 'Cleaned-Life-Exp' standardized dataset sourced from the World Health Organization (WHO) databases. Feature selection methods (Boruta and RRF) identified 20 significant predictors, with variables such as HIV/AIDS, Adult Mortality, Income Composition of Resources, and Schooling being the most influential. Among the eight models evaluated, Random Forest achieves the highest performance ($R^2 = 0.969$, RMSE $= 0.179$, MAE $= 0.116$), followed closely by XGBoost and Gradient Boosting ($R^2 \approx 0.96$), highlighting the superiority of ensemble methods. Support Vector Machine (SVM) performs well, while Decision Tree and KNN perform moderately. Linear Regression and Neural Networks have the lowest performance. So, this study provides an accurate predictive framework using machine learning models, which can guide policymakers in improving public health outcomes. Future research can focus on exploring deep learning models, integrating interpretability techniques like SHAP, and developing region-specific or real-time predictive frameworks. Instead of evaluating the machine learning models on a single training-test split, a future extension of the method would be to apply any cross-validation approaches that average the results across multiple folds, leading to a more stable and accurate estimate of the model's performances. Also, further investigations incorporating additional quality of life and environmental variables into the prediction model

would be useful. These advancements will further refine predictions and provide actionable insights to guide global public health strategies and resource allocation.

Acknowledgments: The authors thank the editor and two reviewers for their valuable comments and suggestions in improving the manuscript.

References

[1] T. Choudhury, S. K. Bharti, M. Kumar Gourisaria, J. J. Jena, D. K. Behera, and A. Bandyopadhyay (2024). "Predictive Modeling of Life Expectancy Using Machine Learning Algorithms," 2024 Glob. Conf. Commun. Inf. Technol. GCCIT 2024, 2024, doi: 10.1109/GCCIT63234.2024.10862085.

[2] B. Griffin, V. Loh, and B. Hesketh (2013). "A mental model of factors associated with subjective life expectancy," 2013, doi: 10.1016/j.socscimed.2013.01.026.

[3] O. K. A (2025). "Leveraging Machine Learning for Predictive Models in Healthcare to Enhance Patient Outcome Management Leveraging Machine Learning for Predictive Models in Healthcare to Enhance Patient Outcome Management," no. January, 2025, doi: 10.56726/IRJMETS66198.

[4] B. A. Lipesa, E. Okango, B. O. Omolo, and E. O. Omondi (2023). "An application of a supervised machine learning model for predicting life expectancy," SN Appl. Sci., vol. 5, no. 7, 2023, doi: 10.1007/s42452-023-05404-w.

[5] "Life Expectancy: Data Analysis and Modeling." Accessed: Dec. 12, 2024. [Online]. Available: https://www.kaggle.com/code/topologically/life-expectancy-data-analysis-and-modeling

[6] O. Idrizi and M. Harizaj (2023). "Research Methodology for predicting Life Expectancy using Machine Learning," no. March, pp. 132–134, 2023, [Online]. Available: https://www.enjoyalgorithms.com/blog/life-expectancy-prediction-using-linear-

[7] J. D. Morgenstern et al., (2020). "Predicting population health with machine learning: A scoping review," BMJ Open, vol. 10, no. 10, 2020, doi: 10.1136/bmjopen-2020-037860.

[8] "Analysis on Relevant Factors Affecting Life Expectancy | IEEE Conference Publication | IEEE Xplore." Accessed: Dec. 12, 2024. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9777372

[9] Y. Wang (2021). "The Greatest Factors Affecting Life Expectancy: A Research based on Different Continents and Countries," Proc. - 2021 3rd Int. Conf. Mach. Learn. Big Data Bus. Intell. MLBDBI 2021, pp. 531–541, 2021, doi: 10.1109/MLBDBI54094.2021.00107.

[10] D. Tiwari, B. Nagpal, B. S. Bhati, A. Mishra, and M. Kumar (2023). A systematic review of social network sentiment analysis with comparative study of ensemble-based techniques, vol. 56, no. 11. Springer Netherlands, 2023. doi: 10.1007/s10462-023-10472-w.

[11] "Cleaned Countries Life Expectancy Dataset." Accessed: Mar. 11, 2025. [Online]. Available: https://www.kaggle.com/datasets/paperxd/cleaned-life-expectancy-dataset

[12] Z. Jin, J. Shang, Q. Zhu, C. Ling, W. Xie, and B. Qiang (2020). "RFRSF: Employee Turnover Prediction Based on Random Forests and Survival Analysis," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 12343 LNCS, pp. 503–515, 2020, doi: 10.1007/978-3-030-62008-0_35.

[13] H. Ahmadpour, O. Bazrafshan, E. Rafiei-Sardooi, H. Zamani, and T. Panagopoulos (2021). "Gully erosion susceptibility assessment in the kondoran watershed using machine

learning algorithms and the boruta feature selection," Sustain., vol. 13, no. 18, 2021, doi: 10.3390/su131810110.
[14] M. B. Kursa and W. R. Rudnicki (2010). "Feature selection with the boruta package," J. Stat. Softw., vol. 36, no. 11, pp. 1–13, 2010, doi: 10.18637/jss.v036.i11.
[15] E. Izquierdo-Verdiguier and R. Zurita-Milla (2019). "An evaluation of Guided Regularized Random Forest for classification and regression tasks in remote sensing," Int. J. Appl. Earth Obs. Geoinf., vol. 88, no. June 2019, 2020, doi: 10.1016/j.jag.2020.102051.
[16] H. Deng and G. Runger (2012). "Feature selection via regularized trees," Proc. Int. Jt. Conf. Neural Networks, 2012, doi: 10.1109/IJCNN.2012.6252640.
[17] T. Ngo, Hoang (2012). "The Steps to Follow in a Multiple Regression Analysis," SAS Glob. Forum 2012, pp. 1–12, 2012.
[18] "Decision Tree - GeeksforGeeks." Accessed: Dec. 13, 2024. [Online]. Available: https://www.geeksforgeeks.org/decision-tree/
[19] "Random Forest Algorithm." Accessed: Dec. 14, 2024. [Online]. Available: https://www.simplilearn.com/tutorials/machine-learning-tutorial/random-forest-algorithm#working_of_random_forest_algorithm
[20] "Random Forest Algorithm with Machine Learning- Analytics Vidhya." Accessed: Dec. 14, 2024. [Online]. Available: https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/
[21] "Random Forest Algorithm | A Map to Avoid Getting Lost in 'Random Forest.'" Accessed: Dec. 14, 2024. [Online]. Available: https://www.analyticsvidhya.com/blog/2021/05/a-map-to-avoid-getting-lost-in-random-forest/
[22] "Support Vector Machines (SVM) Classifier | by Saba Hesaraki | Medium." Accessed: Dec. 14, 2024. [Online]. Available: https://medium.com/@saba99/support-vector-machines-svm-classifier-f5413109bd10
[23] "Support Vector Machine (SVM) Algorithm - GeeksforGeeks." Accessed: Dec. 14, 2024. [Online]. Available: https://www.geeksforgeeks.org/support-vector-machine-algorithm/
[24] "(a): SVM classifier Courtesy: google | Download Scientific Diagram." Accessed: Dec. 16, 2024. [Online]. Available: https://www.researchgate.net/figure/a-SVM-classifier-Courtesy-google_fig3_370589237
[25] R. K. Halder, M. N. Uddin, M. A. Uddin, S. Aryal, and A. Khraisat (2024). "Enhancing K-nearest neighbor algorithm: a comprehensive review and performance analysis of modifications," J. Big Data, vol. 11, no. 1, 2024, doi: 10.1186/s40537-024-00973-y.
[26] "Gradient Boosting in ML - GeeksforGeeks." Accessed: Dec. 14, 2024. [Online]. Available: https://www.geeksforgeeks.org/ml-gradient-boosting/
[27] R. Sibindi, R. W. Mwangi, and A. G. Waititu (2023). "A boosting ensemble learning based hybrid light gradient boosting machine and extreme gradient boosting model for predicting house prices," Eng. Reports, vol. 5, no. 4, pp. 1–19, 2023, doi: 10.1002/eng2.12599.
[28] A. Natekin and A. Knoll (2013). "Gradient boosting machines, a tutorial," Front. Neurorobot., vol. 7, no. DEC, 2013, doi: 10.3389/fnbot.2013.00021.
[29] "Flow chart of gradient boosting regression model | Download Scientific Diagram." Accessed: Dec. 14, 2024. [Online]. Available: https://www.researchgate.net/figure/Flow-chart-of-gradient-boosting-regression-model_fig5_379187282
[30] "What is the XGBoost algorithm and how does it work?" Accessed: Dec. 16, 2024. [Online]. Available: https://www.analyticsvidhya.com/blog/2018/09/an-end-to-end-guide-to-understand-the-math-behind-xgboost/

[31] B. Mesut, A. Başkor, and N. Buket Aksu (2023). "Role of artificial intelligence in quality profiling and optimization of drug products," A Handb. Artif. Intell. Drug Deliv., pp. 35–54, Jan. 2023, doi: 10.1016/B978-0-323-89925-3.00003-4.

[32] "What is a Neural Network? - GeeksforGeeks." Accessed: Dec. 15, 2024. [Online]. Available: https://www.geeksforgeeks.org/neural-networks-a-beginners-guide/

[33] "Neural network (machine learning) - Wikipedia." Accessed: Dec. 15, 2024. [Online]. Available: https://en.wikipedia.org/wiki/Neural_network_(machine_learning)

[34] "Activation Functions: Sigmoid, Tanh, ReLU, Leaky ReLU, Softmax | by Mukesh Chaudhary | Medium." Accessed: Dec. 15, 2024. [Online]. Available: https://medium.com/@cmukesh8688/activation-functions-sigmoid-tanh-relu-leaky-relu-softmax-50d3778dcea5

[35] "Regression Performance." Accessed: Dec. 16, 2024. [Online]. Available: https://c3.ai/introduction-what-is-machine-learning/regression-performance/

[36] "A Practical Guide to Root Mean Square Error (RMSE) | Aporia." Accessed: Dec. 16, 2024. [Online]. Available: https://www.aporia.com/learn/root-mean-square-error-rmse-the-cornerstone-for-evaluating-regression-models/

[37] "How to Calculate Mean Absolute Error - Shiksha Online." Accessed: Dec. 16, 2024. [Online]. Available: https://www.shiksha.com/online-courses/articles/mean-absolute-error/