EAI Endorsed Transactions
on Scalable Information Systems

Research Article EAI.EU

Life Expectancy Prediction using Recursive Partitioning and Bagging Algorithms

Muhammad Bux Alvi[1,*], Majdah Alvi[1], Yasir Hussain[2], Wajid Rehman[3], Kavita Tabassum[4], Shahnawaz Farhan[5], Noor Fatima[1]

[1]Department of Computer Systems Engineering, The Islamia University of Bahawalpur, Punjab, Pakistan
[2]Numrex, Lahore, Pakistan
[3]UET Lahore, Pakistan
[4]Sindh Agriculture University, Tando Jam, Sindh, Pakistan
[5]Sindh Energy Department, Pakistan

Abstract

Life expectancy is a crucial indicator of the population's health and well-being. Recent research has highlighted the importance of various socioeconomic and health factors in determining the lifespan of individuals. Those factors include Gross Domestic Product (GDP), healthcare expenditure, mortality rates, and education level. This study employs recursive partitioning (decision trees) and bagging (random forest) techniques on the Life Expectancy dataset from the World Health Organization (WHO) to evaluate the effectiveness of predictive models. The dataset was prepared by encoding categorical features, scaling the features, normalizing them, and handling outliers. Mean imputation was used to handle missing values and produce a quality dataset. Optimized models based on recursive partitioning and bagging algorithms achieved performance efficiencies of 92% and 97%, respectively. The bagging algorithm-based model produced a mean squared error of 1.17, a mean absolute error of 2.0, and an R^2-score of 97%. Other key findings included the importance of dataset characteristics—such as HIV/AIDS prevalence, adult mortality, and health resource income—in predicting life expectancy. This research elucidates the impact of feature engineering and data preprocessing strategies on data quality and predictive model precision, offering novel insights for public health policymaking and informing future research directions.

Received on 24 March 2025; accepted on 09 October 2025; published on 13 October 2025

Keywords: Decision Tree, Life Expectancy, Public Health Policy, Random Forest Algorithm, Social Sustainability

doi:10.4108/eetsis.8959

1. Introduction

Life expectancy remains one of the most widely used indicators of population health and social development. Global statistics indicate that life expectancy can vary by more than 30 years between the top- and bottom-ranked countries, representing significant gaps in access to healthcare, socioeconomic conditions, and living environments. Such variation highlights the need for accurate forecasting models to help evidence-based policymaking [1]. A wide range of factors contribute to life expectancy outcomes. Economic indicators, including Gross Domestic Product (GDP) and income distribution, as well as health-related indicators (adult and infant mortality, HIV/AIDS prevalence, vaccination and nutrition levels, and Body Mass Index), education, and investment in healthcare, are all important. Understanding the relative influence of these variables requires methods that can capture both linear and nonlinear relationships [2].

Life expectancy is challenging to predict due to the complex and varied factors that influence it. Many existing models struggle with imputing data and capturing the non-linear relationships between these features. Hence, there is a need for robust machine learning models that can integrate various health and socioeconomic indicators to provide accurate life expectancy forecasts [3].

[*]Corresponding author. Email: mbalvi@iub.edu.pk

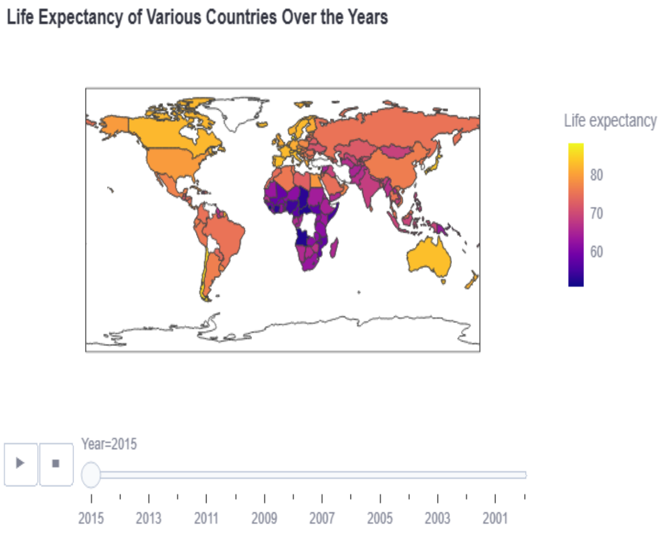

Figure 1. Average life expectancy of countries from 2000—2015 [3]

Figure 1 shows the trends in the average life expectancy of various countries over time. It demonstrates how life expectancy has steadily increased in most countries, a testament to global progress in healthcare and living conditions. It also shows that life expectancy varies significantly from one nation to another, reflecting variations in socioeconomic conditions, healthcare availability, and economic development. This information is critical for tracking global health and development trends. A thorough list of features and their brief explanations for various indicators related to global health and development is provided in Table 1. Dataset features include lifespans (adult and infant mortality), vaccination coverage (against Hepatitis B, Measles, and Polio), and nutritional status, as indicated by BMI, as well as socioeconomic features (GDP and total population count).

Table 1. Dataset Features and their Descriptions

Features	Features description
Country	List of the 179 countries
Year	Year of Observation for health data
Life expectancy	The average life a person is expected to live
Adult Mortality	Number of adult deaths per 1,000 individuals in a population
Infant deaths	Infant deaths per 1,000 live births
Alcohol	Alcohol consumption (liters of pure alcohol per capita) for Aged 15+
Hepatitis-B	Percentage coverage of Hepatitis-B (HepB3) immunization coverage in 1-year-olds
Measles	% of coverage of the Measles containing vaccine first dose (MCV1), immunization among 1-year-olds
BMI	A measurement of nutritional status in adults, defined as a person's weight in kilograms divided by the square of that person's height in meters (kg/m^2)
Under-five deaths	Deaths of children under five per 1,000 live births
Polio	% of coverage of Polio (Pol3) immunization among 1-year-olds
Schooling	Average years that people aged 25+ spent in formal education
HIV/AIDS	Incidents of HIV per 1,000 population (aged 15-49)
Diphtheria	% of coverage of Diphtheria tetanus toxoid and pertussis (DTP3) immunization among 1-year-olds
GDP	GDP per capita in USD
Population	Total population (millions)

This study aims to develop robust machine learning models that predict life expectancy and provide insights for improving it. To achieve the set goal, a comprehensive dataset was acquired and effectively preprocessed to be fit for modeling. The study employed recursive partitioning (Decision Tree) [4] and bagging (Random Forest) [5] techniques on the World Health Organization (WHO) Life Expectancy dataset. By combining systematic preprocessing, such as mean imputation of missing values, normalization, and feature engineering, the proposed models seek to enhance the precision of the predictions while identifying the most significant determinants of life expectancy. The experimental results demonstrated good performance, particularly for the random forest model, which yielded higher accuracy. The developed models were evaluated using the mean squared error (MSE), mean absolute error (MAE), and R^2-scores. Random forest-based model performed better, achieving an R^2-score of 97%, a mean absolute error (MAE) of 2.0, and a mean squared error (MSE) of 1.17.

The key findings underscored the importance of factors such as HIV/AIDS prevalence, adult mortality, and income share of health resources as predictors of life longevity. The research demonstrated the impact of feature engineering and missing data handling strategies on model quality, providing valuable insights for public health policy. The study contributed to the literature by establishing effective models for handling complex datasets and improving the accuracy of life expectancy prediction using machine learning techniques.

The remainder of this paper is organized as follows. Section II provides background information relevant to the study. Section III describes the proposed framework design and implementation details. Section IV discusses the obtained results and interprets the findings. Finally, Section V concludes the paper and highlights directions for future research.

2. Related Work

This section elaborates on the methods, tools, and techniques employed by researchers in machine learning studies on lifespan prediction.

Vikram Bali (2021) presented findings on life expectancy by developing machine learning models, which are among the most relevant models developed in recent times. The authors employed Linear Regression, Ridge Regression, and Decision Tree to predict lifespan and obtained good results [1].

A.A. Bhosale and D. Sundaram in [6] developed a model to predict human life expectancy based on weight, respiration rate, heart rate, and blood pressure. Their study aimed to establish a descriptive assessment of these parameters, highlighting their impact on projected life expectancy. The authors used descriptive analysis to describe various features involved and the life expectancy associated with these features. Additionally, regression analysis was used to determine the relationship between the factors mentioned and life expectancy.

The authors in [7] performed data analysis to measure life expectancy forecasts using regression-based machine learning methods. The study employed cross-sectional data from a WHO repository and Kaggle, which contained data for 193 countries for the years 2000 and 2015, to examine the determinants of life expectancy. They reported Random Forest Regressor to be the model of choice due to its strong predictive capability, resulting in a performance level of 95%.

According to Kerdprasop, economic and environmental factors significantly contributed to life expectancy. The authors analyzed data obtained from the World Bank Database, covering the period 1990 to 2015. They established a CHAID predictive model with a structure similar to the Decision Tree algorithm. They found that economic growth had the highest mean coefficient among all the included variables, demonstrating a strong positive correlation with life expectancy [8].

The authors employed XGBoost and Random Forest Regressor to predict life expectancy for 193 countries using the WHO dataset (from 2000 to 2015). The study described and assessed different health, immunization, socioeconomic, and behavioral factors with the HDI. The results showed that XGBoost outperformed the Random Forest and Artificial Neural Networks models, achieving a mean absolute error (MAE) of 1.554 and a root mean square error (RMSE) of 2.402.

The Random Forest Regressor was found to be the best model, and the study pointed out that the consequences of showing accurate life expectancy could have in terms of decisions being made by public health facilities. The performance of the models was tested using MAE, RMSE, CV-score, and R^2-score of the model obtained is 93.88% [9].

The authors in [10] employed clustering techniques using features included mental and physical health conditions, disease incidence, and accidents to determine the life expectancy. Three clustering techniques - Density-Based Spatial Clustering of Applications with Noise (DBSCAN), k-means, and fuzzy c-means clustering - were used, and the results were analyzed using the Silhouette score, the Davis Bouldin Index (DBI), and the Calinski-Harabasz index. The authors reported the k-means algorithm to prove better.

A.A. Bhosale and K.K. Sundaram in their work, "Life Prediction Equation for Human Beings," aimed to create a model that predicted the human chance of living long using other measurable factors such as blood pressure, weight, and pulse rate, among others. There was an intention to develop a useful instrument to meet the needs of insurance companies and governments in predicting average life expectancy, given that many characteristics are easily measurable, especially in the developing world. The authors used information collected from a group of people in good health, focusing on factors such as blood pressure, pulse rate, weight, and breathing rate. The paradigm they provided was represented by the empirical equation, which was derived from differential calculus.

$$\text{life} = 0.4467\left(\frac{WT \cdot BP}{HR}\right) + 3.5735 \tag{1}$$

The study identified direct relationships with weight, life expectancy, and blood pressure, revealing inverse relationships with heart and respiratory rates. However, it did not incorporate validation criteria, which greatly hampered the model's stability and practicability due to the omission of genetic predisposition and lifestyle considerations [11].

- **W – Body Weight (kg):** In equation 1, it appears in the numerator, so a higher body weight within healthy limits pushes the predicted life expectancy upward. This suggests that adequate body mass is associated with a better nutritional status.

  numerator,

- **BP – Blood Pressure (mmHg, usually systolic):** In equation 1, its presence in the numerator means that, in this dataset, moderate increases in blood pressure correlated with longer life. The authors were working with healthy subjects, so "higher" here still refers to normal, well-regulated pressure.

- **HR – Heart Rate (beats per minute):** In equation 1, placed in the denominator, heart rate has an inverse effect—when resting heart rate rises, the calculated life expectancy falls. This aligns with clinical evidence that a persistently elevated heart rate can stress the cardiovascular system.

In equation 1, the numerical factors are purely empirical:

- 0.4467 – a scaling coefficient obtained from curve fitting; it adjusts the combined ratio

$$\frac{W \times T \times BP}{HR}$$

So the predicted values match the observed lifespans in the training data.

- 3.5735 – an additive constant that shifts the line vertically, ensuring the formula's baseline prediction aligns with the actual life expectancy when the other terms are near average.

3. Proposed Architecture

Figure 2 presents all the components of the proposed architecture, which are explained in the following subsections. The architecture illustrates the overall approach for creating a data-driven model. It starts with data collection from WHO, a United Nations specialized agency. Data preprocessing is the next step, during which the data is cleaned, missing values are handled, categorical features are transformed, outliers are detected and managed, and the data is normalized. The next step is to perform feature engineering, followed by splitting the dataset. The training set is used to build the predictive models. The predictive models use recursive partitioning (Decision Tree) and bagging (Random Forest) algorithms. Each model is assessed based on suitability indices, followed by model tuning and optimization for increased suitability. The performance of both models is evaluated on the testing set using metrics such as mean squared error (MSE) and mean absolute error (MAE), which measure the difference between the predicted and actual life expectancy values.

3.1. Data Acquisition

The data is acquired from the World Health Organization (WHO) [12], a reliable data source utilized globally. The dataset comprises 22 features and 2,938 data entries. It contains health-related factors, including Polio, Hepatitis-B, Body Mass index (BMI), Measles, and Diphtheria. The economic factors encompass Gross Domestic product (GDP), income composition of resources, health expenditures, and total expenditures. Additionally, significant features in the dataset include schooling, year, country name, and status, as well as infant mortality, adult mortality, and alcohol consumption.

3.2. Data Preprocessing

Data preprocessing is a pertinent step in predictive model building that may include multiple sub-steps. Often, initial data analysis is performed to examine the behavior of the dataset. The data analysis encompassed exploratory data analysis (EDA) to understand the structure of the data. Through EDA, it was identified that the dataset contained missing values, categorical features, outliers, and notable discrepancies among feature values. If unaddressed, these factors can lead to skewed and misleading results.

Missing values handling. Missing values render the dataset incomplete and may be removed if doing so does not compromise the dataset's size or distribution. However, in this case, dropping missing values would result in significant data loss. Therefore, missing values were handled by filling them using techniques such as mean imputation, median imputation, forward-fill, backward-fill, simple imputation, and KNN imputation. These techniques were applied sequentially, and the data distribution was recorded before and after each application. It was found that filling missing values with the mean preserved the original distribution; thus, it was adopted. A similar process was also selected in the previous studies [13, 14]. Table 2 shows the percentage of missing values for dataset features. The dataset features such as Hepatitis-B (18.68%) and Population have missing values of 18.58% and 21.96%, respectively. In contrast, features such as Year, Infant Deaths, and HIV/AIDS have no missing values. It could be concluded that missing value handling is instrumental in assessing data completeness and subsequent investigations. The high percentage of missing values in features like Hepatitis-B and Population may require more sophisticated imputation or data cleaning techniques —using statistical methods or domain-specific knowledge —to fill in the gaps without distorting the overall trends.

Categorical features. The dataset included categorical variables that required transformation into numerical representations to ensure compatibility with machine learning algorithms. Two techniques were considered for processing variables: Label Encoding and One-Hot Encoding. However, Label Encoding was identified as the most suitable approach for this dataset. The dataset contained two categorical variables, namely

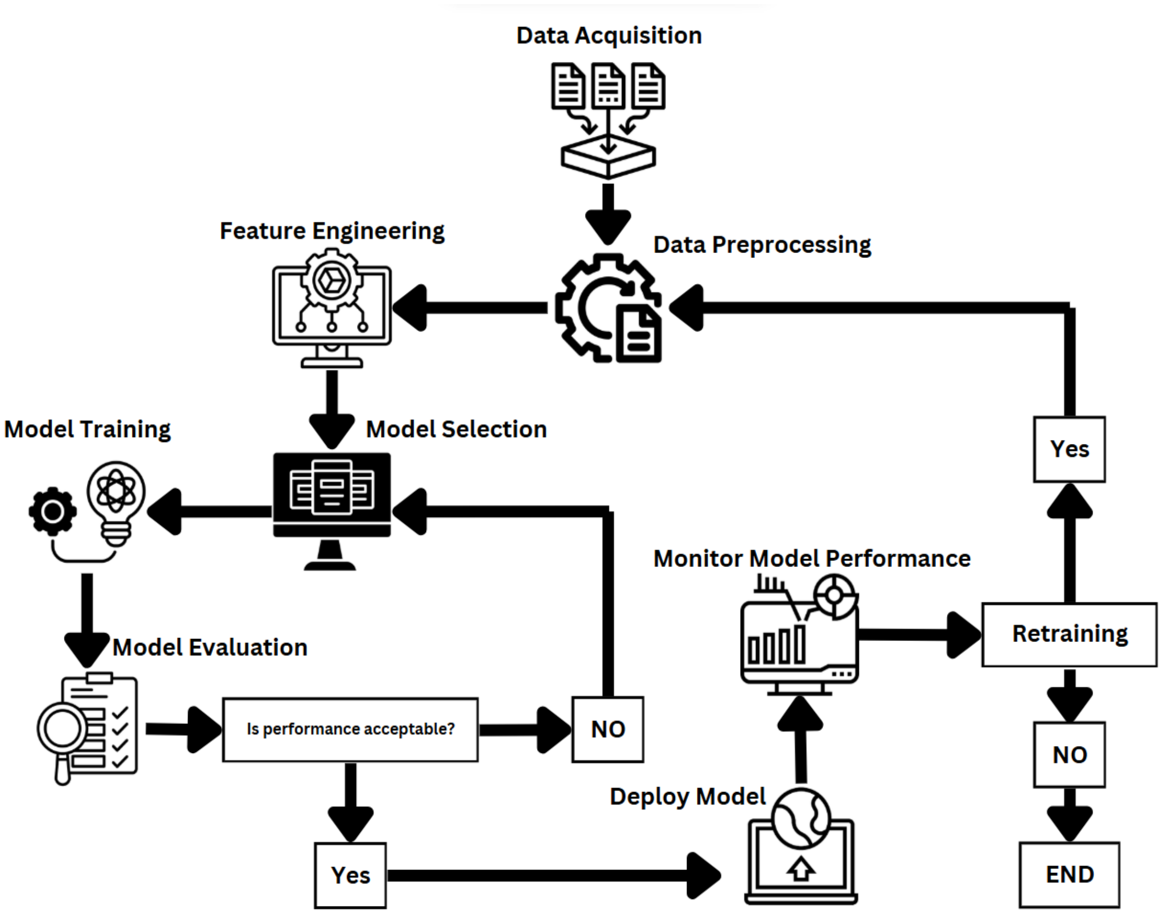

Figure 2. Proposed Life Expectancy Prediction Model Architecture

Country and Status. Both were transformed, ensuring their compatibility with the algorithms while retaining the ability to capture meaningful patterns and relationships [15, 16].

In the context of Decision Tree and Random Forest models, Label Encoding is often preferred over One-Hot Encoding due to its computational efficiency and its compatibility with the way these models process data. Tree-based models inherently handle categorical features effectively by splitting data based on feature values, regardless of whether they are numerically encoded or binary encoded. Label Encoding converts categorical variables into integers, ensuring each variable remains a single column. This is particularly useful for features with high cardinality, such as a "Country" feature containing 190 unique categories, as it avoids the curse of dimensionality that arises with One-Hot Encoding.

One-Hot Encoding, while effective for algorithms that rely on distance-based metrics (e.g., logistic regression or support vector machines), creates a separate binary column for each unique category, significantly increasing the dimensionality of the dataset. This can lead to increased memory usage and computational overhead, as well as a higher risk of overfitting in smaller datasets. In contrast, Label Encoding maps each category to a unique integers, preserving the compactness of the data and ensuring that the models remain computationally efficient.

Importantly, Decision Trees and Random Forests are not sensitive to the ordinal nature of Label Encoded values, as they split data based on thresholds, not on the relative magnitude of the numerical labels. For instance, if "Developing" is encoded as 0 and "Developed" as 1, the model will treat these as distinct categories during splits without assuming any inherent order. Thus, Label Encoding provides a simple, interpretable, and efficient solution for preparing categorical data in tree-based models, making it a natural choice in scenarios where these algorithms are employed [17].

Table 3 illustrates the mapping process of categorical features, such as Country and Status, into their corresponding Label-Encoded values. The Original Features

Table 2. Missing Values Percentage for Each Feature

Feature	Missing Values Percentage
Year	0.00%
Adult Mortality	0.34%
Infant Deaths	0.00%
Alcohol	6.72%
Percentage Expenditure	0.00%
Hepatitis-B	18.68%
Measles	0.00%
BMI	1.15%
Under-five Deaths	0.00%
Polio	0.60%
Total Expenditure	7.87%
Diphtheria	0.60%
HIV/AIDS	0.00%
GDP	14.68%
Population	21.96%
Thinness 1-19 Years	1.15%
Thinness 5-9 Years	1.15%
Income Composition of Resources	5.57%
Schooling	5.49%

column showcases the original categorical data, while the Encoded Features column displays the transformed integer values. This mapping highlights the interpretable nature of Label Encoding, ensuring efficient handling of high-cardinality categorical data without inflating feature space. Such encoding aligns seamlessly with the requirements of Decision Tree and Random Forest models, ensuring both computational efficiency and effective utilization of categorical variables in these algorithms.

Table 3. Mapping of Original and Encoded Features

Original Features		Encoded Features	
Country	Status	Ctry	Stat
Tanzania	Developing	0	0
United States	Developed	1	1
Pakistan	Developing	2	0
Saint Vincent	Developing	3	0
Poland	Developed	4	1

Outlier detection. A variety of methods exist for detecting and handling outliers in datasets. Statistical techniques for outlier detection include Z-score methods and the Interquartile Range (IQR). Visualization techniques such as distplot, boxplot, and scatter plot are also utilized to detect outliers. After detecting the outliers, the next step is to handle them. There are multiple techniques to handle outliers, depending on their type. The outliers may be classified based on whether the outliers affect the dataset distribution or not. One solution is to remove the outliers if they are outside the interquartile range (IQR-based filtering). Another approach for outlier handling is through capping (Winsorization). In the capping technique, values outside the interquartile range are replaced with the nearest boundary value from the IQR. The capping method is used when removing outliers is costly and affects the data. Other techniques included imputation, where outliers were replaced with statistical measures, such as the mean, median, or mode, to preserve data continuity. Transformation methods, such as logarithmic or Box-Cox transformations, can also be employed to reduce the impact of outliers by stabilizing variance and normalizing skewed distributions.

All dataset features, except Country, Year, and Alcohol, contain outliers. The outliers were detected and handled using the Inter Quartile Range (IQR) and Z-score. The outliers were capped with boundary values to minimize their impact on overall analysis [18, 19]. The identification of outliers using the interquartile range (IQR) and box plots is demonstrated in Figure 3. The upper plots show that the extreme outliers in the original dataset significantly affected the overall distribution and statistical summaries. The density plot indicates that the data were skewed due to these outliers, while the box plot clearly shows multiple outliers below the lower limit of the IQR. After applying the capping method, as depicted in the lower plots, the extreme values were replaced with the nearest boundaries of the interquartile range, rather than being removed. This process effectively mitigated the influence of outliers on the data distribution. The resulting density plot exhibits a reduction in skewness, resulting in a more balanced representation of the data. Similarly, the box plot for the capped data confirms the absence of outliers, ensuring that the dataset remained within the defined range.

Data scaling and normalization. The scaling and normalization techniques transform the data. Data scaling and normalization help mitigate the chances for a particular feature to dominate the other features while producing the model results. Scaling adjusted the data range to a specific value, while normalization transformed the data to fit a standard normal distribution. These steps help improving the model performance and avoid biased results [20, 21]. Figure 4 illustrates the scaling process applied to the dataset. Before scaling, Adult Mortality had a much larger range, which could dominate model training and lead to biased outcomes, particularly in algorithms that are sensitive to feature magnitudes. After scaling, both variables were adjusted to a comparable range with a mean of zero and a variance of one, ensuring fair representation and equal

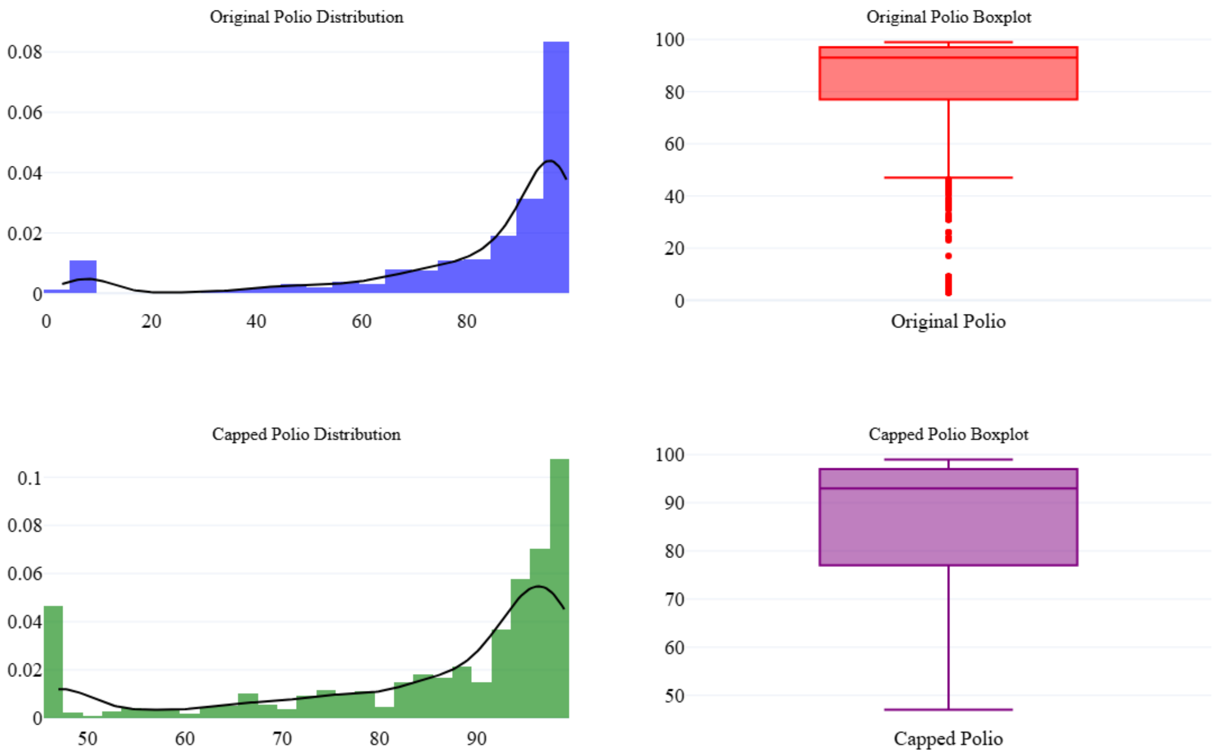

Figure 3. Visualization of the 'Polio' column after capping the outliers using the IQR method.

contribution during analysis. This transformation preserved the relative distribution and patterns of the data while reducing the impact of varying scales. Moreover, scaling facilitated faster and more stable convergence of optimization algorithms by eliminating numerical imbalances.

3.3. Feature Engineering

Feature engineering is a crucial step in machine learning workflows that enhances model accuracy and performance. It involves selecting, transforming, or creating features that capture meaningful patterns from the data. The feature engineering process not only enhances the model's predictive power but also reduces overfitting and computational complexity by eliminating irrelevant or redundant features. By focusing on the most impactful variables, feature engineering enables models to perform more effectively and yield more interpretable results.

Feature Selection. Feature selection plays a pivotal role in reducing data dimensionality, enhancing model interpretability, and improving predictive performance.

Recursive Feature Elimination (RFE) was applied with a Random Forest (RF) estimator to select the top 10 features of the dataset. RFE iteratively removed less significant features based on their importance scores derived from the RF model. Top ten features included Year, Adult Mortality, Alcohol, BMI, Under-Five Deaths, Total expenditure, HIV/AIDS, Thinness 5–9 Years, Income composition of resources, and Schooling. RFE is an effective feature selection method that retains only the most relevant features [22]. Sequential Feature Selection (SFS) is an other feature selection technique to identify the most relevant feature subset. SFS iteratively evaluates combinations of features by adding or removing them based on their contribution to the model's performance. In this study, the backward direction of SFS was applied with a Random Forest (RF) regressor as the base estimator to optimize the selection process. The application of SFS resulted in the selection of 12 key features: Country, Year, Status, Adult Mortality, Infant Deaths, Under-five Deaths, Polio, HIV/AIDS, GDP, Thinness 1-19 Years, Thinness 5-9 Years, and Schooling. SFS maintain a strong model performance and generalizability capability [23].

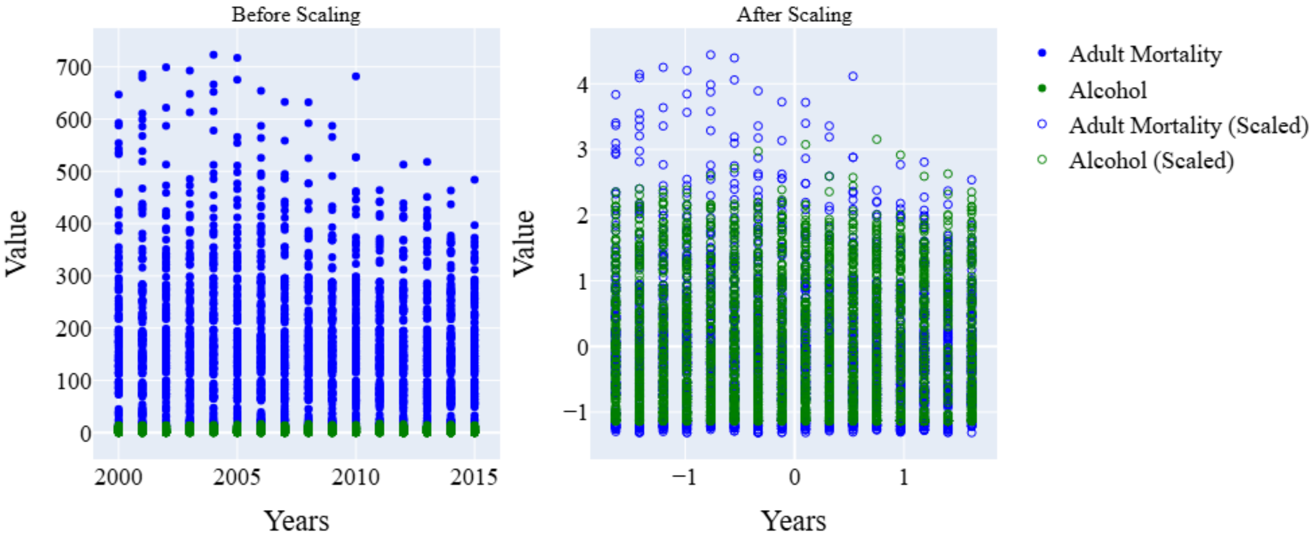

Figure 4. Illustration of scaling process

3.4. Model Building and Optimization

Model building is the process of feeding the training data subset to a machine learning algorithm. The dataset was split into the feature matrix (X) and the target vector (y). The feature matrix included all predictors except Life expectancy, which was used as the target variable. The data set was then split into a training subset (80%) and a testing subset (20%), to ensure a random and reproducible split [24].

The decision tree algorithm works by recursively partitioning the data into smaller subsets based on the values of the input features. The process of building a Decision Tree begins with a root node representing the entire dataset. The algorithm then selects the best feature to split the data based on criteria, such as information gain or Gini impurity. The data is then split into two child nodes based on the selected feature and a splitting criterion, such as a threshold value. This process is recursively applied to each child node until a stopping criterion is met, such as all instances belonging to the same class. The final prediction is made at the leaf node [25, 26].

Mathematical Foundations of Decision Trees. Decision trees are constructed on a strong mathematical foundation that ensures data is well partitioned, enabling accurate prediction. Embedded within this approach are important principles, such as entropy (a measure of impurity) and information gain, which must be used to choose features and split during tree growth.

Entropy (Information Gain) Entropy measures the impurity of a node, with higher entropy indicating higher impurity. The formula for entropy is:

$$H(X) = -\sum(p(x)\cdot\log_2(p(x))) \tag{2}$$

where $H(X)$ is the entropy of node X, $p(x)$ is the probability of class x, and the sum is taken over all classes.
Explanation: In the context of decision trees, entropy quantifies the randomness in the data. A high entropy implies that the data is highly disordered, whereas a low entropy indicates that the data is well-organized or pure.

Proof: Let X be a discrete random variable with possible values $\{x_1, x_2, \dots, x_n\}$ and probabilities $\{p_1, p_2, \dots, p_n\}$. Entropy is defined as:

$$H(X) = -\sum(p(x)\cdot\log_2(p(x)))$$

where the sum is taken over all possible values of X. To prove that entropy is a measure of impurity, it must be shown that:

1. $H(X) \geq 0$ for all X
2. $H(X) = 0$ if and only if X is a constant
3. $H(X) = H(Y)$ if X and Y have the same probability distribution

Proof of 1:

$$[H(X) = -\sum(p(x)\cdot\log_2(p(x))) \geq 0] \tag{3}$$

since $\log_2(p(x)) \leq 0$ for all $p(x) \leq 1$.

Proof of 2: If X is a constant, then $p(x) = 1$ for some x and $p(x) = 0$ for all other x.

$$H(X) = -(1 \cdot \log_2(1)) = 0 \tag{4}$$

Conversely, if $H(X) = 0$, then $p(x) = 0$ for all x, which implies that X is a constant.

Proof of 3: Let X and Y have the same probability distribution. Then:

$$\begin{aligned} H(X) &= -\sum p(x) \cdot \log_2(p(x)) \\ &= -\sum p(y) \cdot \log_2(p(y)) \\ &= H(Y) \end{aligned}$$

Gini Impurity Gini impurity measures the probability of misclassifying an instance. The formula for Gini impurity is:

$$Gini(X) = 1 - \sum (p(x)^2) \tag{5}$$

where $Gini(X)$ is the Gini impurity of node X, and $p(x)$ is the probability of class x.

Explanation: Gini impurity measures the probability of incorrectly classifying a randomly chosen element in the dataset if it were randomly labeled according to the class distribution.

Proof: Let X be a discrete random variable with possible values $\{x_1, x_2, \ldots, x_n\}$ and probabilities $\{p_1, p_2, \ldots, p_n\}$. To prove that Gini impurity is a measure of impurity, it must be shown that:

1. $Gini(X) \geq 0$ for all X
2. $Gini(X) = 0$ if and only if X is a constant
3. $Gini(X) = Gini(Y)$ if X and Y have the same probability distribution

Proof of 1:

$$Gini(X) = 1 - \sum (p(x)^2) \geq 0 \tag{6}$$

since $\sum (p(x)^2) \leq 1$ for all $p(x) \leq 1$.

Proof of 2: If X is a constant, then $p(x) = 1$ for some x and $p(x) = 0$ for all other x.

$$Gini(X) = 1 - (1^2) = 0 \tag{7}$$

Conversely, if $Gini(X) = 0$, then $p(x) = 0$ for all x, which implies that X is a constant.

Proof of 3: Let X and Y have the same probability distribution. Then:

$$Gini(X) = 1 - \sum (p(x)^2) = 1 - \sum (p(y)^2) = Gini(Y)$$

Logarithmic Loss (for Regression Tasks) Logarithmic loss measures the error of the model for regression tasks. The formula for logarithmic loss is:

$$L(y, y') = \frac{(y - y')^2}{2 \cdot \sigma^2} \tag{8}$$

where $L(y, y')$ is the logarithmic loss, y is the true value, y' is the predicted value, and σ is the standard deviation.

Explanation: Logarithmic loss quantifies the accuracy of a regression model by penalizing large errors more than small ones. It is often used as a metric for evaluating regression tasks.

Proof: Let y be the true value and y' be the predicted value. Logarithmic loss is defined as:

$$L(y, y') = \frac{(y - y')^2}{2 \cdot \sigma^2}$$

where σ is the standard deviation. To prove that logarithmic loss is a measure of error, it must be shown that:

1. $L(y, y') \geq 0$ for all y and y'
2. $L(y, y') = 0$ if and only if $y = y'$
3. $L(y, y') = L(y, y'')$ if y' and y'' have the same probability distribution

Proof of 1:

$$L(y, y') = \frac{(y - y')^2}{2 \cdot \sigma^2} \geq 0$$

since $(y - y')^2 \geq 0$ for all y and y'.

Proof of 2: If $y = y'$, then:

$$L(y, y') = \frac{(y - y')^2}{2 \cdot \sigma^2} = 0$$

Conversely, if $L(y, y') = 0$, then $y - y' = 0$, which implies that $y = y'$.

Proof of 3: Let y' and y'' have the same probability distribution. Then:

$$L(y, y') = \frac{(y - y')^2}{2 \cdot \sigma^2} = \frac{(y - y'')^2}{2 \cdot \sigma^2} = L(y, y'')$$

Information Gain Ratio The information gain ratio is a modification of information gain that takes into account the number of splits. The formula for the information gain ratio is:

$$IGR(X, Y) = H(Y) - \sum \left(\left| \frac{X_i}{X} \right| \cdot H(Y_i) \right) \tag{9}$$

where $IGR(X, Y)$ is the information gain ratio, $H(Y)$ is the entropy of the target variable, $|X_i|$ is the number of instances in the i-the split, and $|X|$ is the total number of instances.

Example: Imagine a dataset consisting of students and their corresponding class grade which is either A, B, C, or F. The aim is to find out the grade of a new student considering its attributes and characteristics.

Entropy: Let's say the probability distribution of the grades is:

$$P(A) = 0.4$$

$$P(B) = 0.3$$

$$P(C) = 0.2$$

$$P(F) = 0.1$$

The entropy of this distribution is:

$$H(\text{Grades}) = -\begin{pmatrix} 0.4 \cdot \log_2(0.4) \\ +0.3 \cdot \log_2(0.3) \\ +0.2 \cdot \log_2(0.2) \\ +0.1 \cdot \log_2(0.1) \end{pmatrix} = 1.8464$$

Gini Impurity: The Gini impurity of the grades is:

$$\text{Gini(Grades)} = 1 - (0.4^2 + 0.3^2 + 0.2^2 + 0.1^2) = 0.64$$

Logarithmic Loss (for Regression Tasks): Let's say the true grade is A and the predicted grade is B. The standard deviation is $\sigma = 0.5$.

$$L(\text{Grade, Predicted Grade}) = \frac{(A - B)^2}{2 \cdot 0.5^2} = 1.125$$

Decision Tree: A decision tree might split the data based on the student's characteristics, such as their GPA or test scores. The tree might look like this:

- If GPA > 3.5, predict A
- Else if GPA > 3.0, predict B
- Else if GPA > 2.5, predict C
- Else, predict F

The entropy and Gini impurity of the grades at each node of the tree would be calculated based on the probability distribution of the grades at that node.

Assumption of Decision Tree Algorithm.

The assumptions in a decision tree primarily relate to whether the data is categorical or continuous, as well as the nature of the model's output, which can be either linear or non-linear. Here are some key assumptions:

Recursive Partitioning: Decision trees divide data step by step based on feature values. The repeated splitting helps group the data so that the target variable shows similar behavior within each group [27].

Hierarchical Structure: A decision tree can be divided into nodes, which show features; branches, which represent decisions; and leaves, which show the final outcome or prediction. This hierarchical structure clearly illustrates how various features are interconnected and interact with one another.

Greedy Approach: Decision trees are created using greedy methods, meaning that at each node, the best split is chosen based on measures like information gain or impurity around that node. The interesting part is that these locally optimal choices, when combined, form a globally optimal structure for the entire decision tree.

Predictive Accuracy: Decision trees are built so that each split reduces impurity or increases information gain. The algorithm assumes that this process helps find the best feature and split, allowing the data to be divided most effectively into classes or continuous values.

No Feature Interactions: Decision trees split data based only on individual features, without considering how features interact. However, in real-world applications, decision trees can still capture some interactions between features due to their hierarchical structure.

Robustness to Noise: Decision trees make no assumptions about the data; they handle noise and outliers well. However, too much noise or too many outliers can cause overfitting or lead to poorly structured trees.

No Multicollinearity: Decision trees assume that the features are mostly independent of each other. If there is high multicollinearity (strong relationships between features), it can distort the tree's structure and lead to incorrect feature importance scores.

Single Split Decision: For each tree, only one node is split based on a single feature. This binary split is enough to divide the feature space properly, so there's no need to consider mixed feature interactions.

Parameters of Decision Tree.

Decision trees have several parameters that can be adjusted to influence their behavior. Here are some common parameters along with their mathematical definitions where applicable:

Criterion defines the function to measure the quality of a split, that is, the quality function g in section 2. Some of the measures that are frequently used in CART algorithms include Gini impurity and entropy.

Gini Impurity (G): Let n_t be the number of samples at node t, $p(i|t) = \frac{\text{number of samples of class } i \text{ at } t}{n_t}$. Then, the Gini impurity is calculated as:

$$G(t) = 1 - \sum_{i=1}^{c} (p(i|t))^2 \tag{10}$$

where c is the number of classes.

Entropy (H): Entropy is a measure of impurity in the node. It is calculated as:

$$H(t) = -\sum_{i=1}^{c} p(i|t) \log_2(p(i|t)) \tag{11}$$

where $p(i|t)$ is the proportion of samples of class i at node t.

Max Depth (max_depth): This parameter sets the maximum depth of the decision tree, meaning the greatest number of splits it can have. The depth indicates how far the tree extends from the top node to the lowest node.

Min Samples Split (min_samples_split): This defines the minimum number of samples needed to split an internal node and create branches in a decision tree. It determines the number of samples a node must have before it can be divided.

Min Samples Leaf (min_samples_leaf): These are some of the methods used to build the tree: The minimum number of samples allowed in a node is called samples per node. During pruning, if a leaf node has fewer samples than this number, it may be removed or merged with nearby nodes.

Max Features (max_features): By default, it is set to $\sqrt{n}$ for the categorical features and $2\log_2(n)$ for the continuous features where n is the number of instances in the data set. This parameter defines the maximum number of features to consider when finding the best split. It can be specified as an integer, for instance, the number of features, or a decimal value, for instance, the percentage of features.

Splitter: defines the criterion used to determine the preferred split accomplished at the nodes. It often includes "best", by which users select the best split, while "random" is used to select a random split.

Class Weights (class_weight): Some weights related to classes should be used to balance the classes. It can be used to address classification imbalance by giving the minority class decisions larger weights.

Random State (random_state): The basic input used for random number creation; starting figures to count down from. To make the results replicable, different random seeds can be set to a fixed value.

Splitting rules in decision trees control the tree's behavior, and these rules determine the depth and number of splits that can be achieved in a tree, thereby preventing the tree from overfitting. Concepts in this domain refer to specific mathematical definitions and give insight into tuning decision-tree models.

As shown in Table 4, the HIV feature had a 58% importance in the model, followed by Adult Mortality with 16%. The last three features, having the least importance, were Diphtheria, Polio, and Infant Death. The feature importance analysis revealed that HIV had a significantly greater impact on life expectancy compared to other features. In contrast, infant death had the least effect on life expectancy among all features.

Figure 5 illustrates the decision tree for a regression model. It predicts the target variable by iteratively splitting the dataset based on the feature that minimizes the squared error at each step. The root node at the top highlights HIV/AIDS as the most significant factor influencing the target, with a threshold value of -0.215, which divides the dataset into two branches. On the left, where the HIV/AIDS condition is met, further splits are made based on Adult Mortality and Income Composition of Resources, indicating the importance of these features in refining predictions. On the right, where the HIV/AIDS condition is not met, splits are made using Adult Mortality and BMI, demonstrating a different pathway for prediction. Each node displays the squared error, the number of samples, and the average target value for the subgroup. This visualization effectively illustrates the hierarchical decision-making process of the tree and the relative importance of features in the model.

Table 4. Feature Importance Based on Decision Tree

Feature Name	Importance (%)
HIV/AIDS	58.9893
Adult Mortality	16.215
Income composition of resources	15.4081
BMI	2.8836
Schooling	1.2131
Under-five deaths	0.7072
Thinness 1-19 years	0.6957
GDP	0.5579
Year	0.5525
Total expenditure	0.4301
Thinness 5-9 years	0.3827
Percentage expenditure	0.3596
Alcohol	0.3456
Population	0.2987
Hepatitis-B	0.2344
Measles	0.2035
Diphtheria	0.1902
Polio	0.1886
Infant deaths	0.1444

Figure 6 represents the decision boundary visualization for the regression model. It demonstrates how the model predicts the target variable by splitting the feature space into distinct regions. The plot is based on the first two features, Adult Mortality (scaled) and BMI (scaled), while dummy values are used for other features. The color gradient highlights the predicted target values across the feature space, with warmer tones (red) representing higher predictions and cooler tones (blue) representing lower predictions. Each region corresponds to a split created by the decision tree, forming non-linear boundaries that adapt to the data. The scattered points represent the test dataset, with their colors indicating actual target values for comparison. The edges of the regions align with the splits observed in the decision tree (Figure 5), highlighting how the model captures the relationships between these two

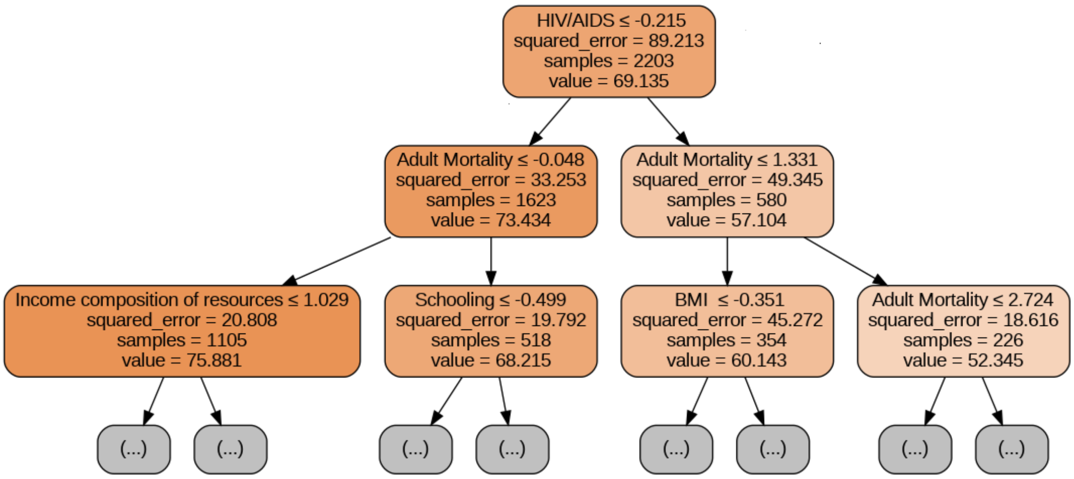

Figure 5. Decision Tree Visualization

features and the target. This visualization effectively shows how the decision tree maps inputs to outputs piecewise.

Random Forest Algorithm.

Random Forest, a powerful ensemble algorithm, combines multiple decision trees to create a robust and flexible model, offering high accuracy and the ability to handle complex, non-linear relationships in data [28–31]. Each tree is trained on a random subset of features and data points to introduce diversity and reduce variance. Predictions are made by averaging the predictions of all individual trees in the forest. The Random Forest Regressor was initialized with 100 estimators and trained on the training data. Predictions were made on the test set, and the model's performance was evaluated using the R^2-score. The study also explored the sampling techniques, types of sampling employed, and the general uses of those samples, particularly in the construction of decision trees. **Data Sampling** The dataset has 1,000 record samples. To implement robust model training and introduce randomness, various sampling techniques were employed, which are described below.

Types of Sampling included row samples, column samples, and the combination.

How to Take Samples Sampling can be done in two ways:

- **With Replacement:** The notable thing about the application of this method is that the same sample can be selected more than once.
- **Without Replacement:** The idea here is that each sample is chosen only once, avoiding it from being selected again.

Assumptions of Random Forest.

- **Ensemble Learning**: Random Forest combines multiple decision trees to improve predictive performance and robustness.
- **Independent Trees**: Each decision tree is built independently using a random subset of features, ensuring diversity and reducing correlation between trees.
- **Bootstrap Sampling**: Random Forest uses bootstrap sampling to create the training dataset for each tree, introducing randomness and reducing overfitting.
- **Feature Randomness**: Each split in the decision trees considers only a random subset of available features, introducing additional randomness and diversity.
- **Majority Voting/Averaging**: The final prediction is made by taking the majority vote (classification) or average (regression) across all individual decision trees.
- **Robustness to Overfitting**: The ensemble nature and randomness in the training process make Random Forest robust to overfitting, even with a large number of decision trees.

Parameters of Random Forest.

- **Max Depth**: This parameter controls the maximum depth of each decision tree in the forest. A higher maximum depth allows for more complex trees but can lead to overfitting.
- **Number of Trees**: The number of decision trees to include in the forest. More trees generally lead to better performance, but increase training time.

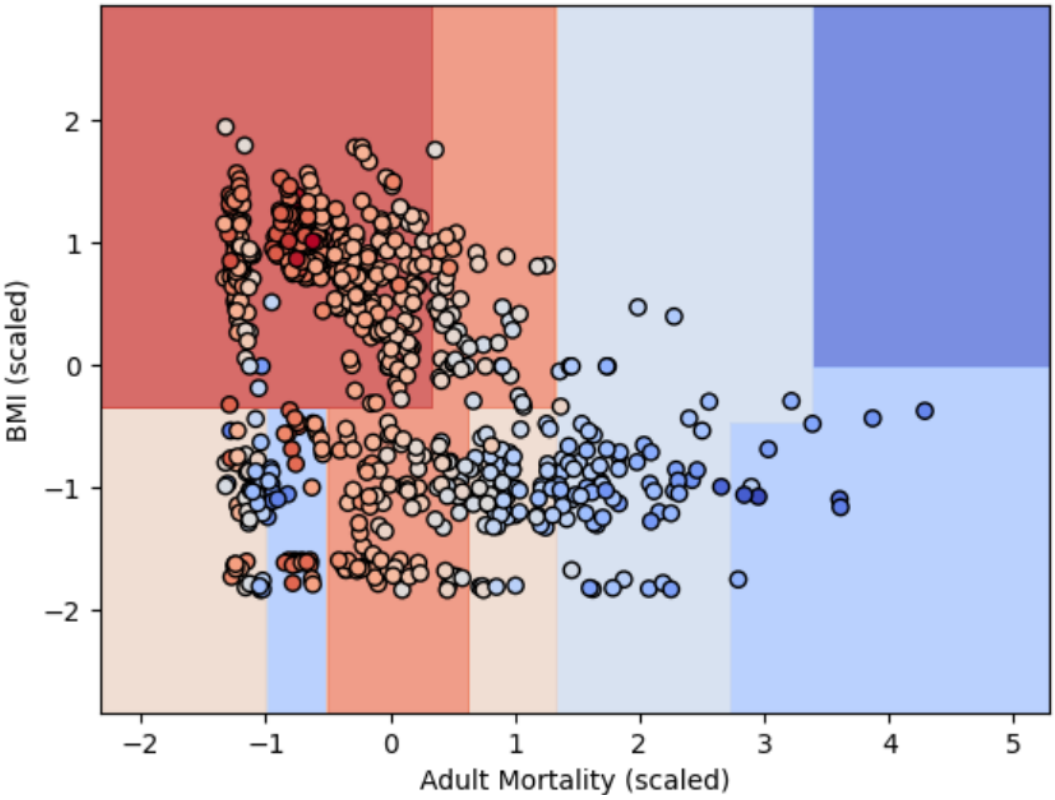

Figure 6. Decision boundary of the regression model using scaled features Adult Mortality and BMI

- **Minimum Samples Per Split**: The minimum number of samples required to split an internal node. Higher values can prevent overfitting but may lead to underfitting.
- **Feature Importance**: Random Forest can calculate the relative importance of each feature in the dataset, helping identify the most relevant predictors [32].

Table 5 shows the importance of features. It was found that the HIV feature had a 59% importance in the model, followed by Adult Mortality with 17% and Income Composition with 12%. The last three features, which had the least importance in the random forest, were Diphtheria, Polio, and Infant Death in that order. This feature importance analysis also showed that HIV had a superior effect on life expectancy compared to other features, while Hepatitis-B had the least effect on life expectancy among all the features.

Table 5. Feature Importance Based on Random Forest

Feature	Importance (%)
HIV/AIDS	59.7174
Adult Mortality	17.2109
Income composition of resources	12.1946
Schooling	1.9719
BMI	1.7492
Under-five deaths	1.1153
Thinness 5-9 years	0.8373
Year	0.7388
Alcohol	0.7188
Total expenditure	0.5555
Thinness 1-19 years	0.4939
Infant deaths	0.4818
Polio	0.4168
Measles	0.3654
GDP	0.3501
Population	0.3092
Diphtheria	0.3033
Percentage expenditure	0.2969
Hepatitis-B	0.1729

3.5. Model Optimization and evaluation

To optimize the model's performance, hyperparameter tuning was conducted using grid search. The hyperparameters tuned included the number of estimators, the maximum depth of the tree, the minimum number of samples required to split a node, and the minimum number of samples required to be at a leaf node. Grid search helped to find the best combination of hyperparameters by exhaustively searching through a specific parameter grid and evaluating the model performance using cross-validation [33]. The performance of the decision tree and random forest models was compared using R^2-score. The random forest model showed superior performance due to its ability to reduce overfitting through ensemble learning. The decision tree model, while simpler and faster to train, tended to overfit the training data, resulting in lower predictive accuracy [34].

4. Result and Discussion

The results of this study support previous findings that the Random Forest (bagging) Regressor outperforms other models based on its superior performance, quantified through evaluation metrics.

4.1. Model Performance and Quantitative Comparison

This section presents a comprehensive comparison of the models employed and provides an in-depth examination of the findings.

1. Random Forest Regressor (Bagging) – Superior Predictive Power
The Random Forest regressor proved to be the most effective for predicting life expectancy due to its ability to identify complex patterns, high accuracy (resulting in reduced prediction errors), and resistance to overfitting. After hyperparameter tweaking, the model achieved the highest R^2 of 0.9716, the lowest MSE of 1.17, and the MAE of 2.00.

2. Decision Tree Regressor (Partitioning) – Competitive performer
The Decision Tree Regressor performed similarly to the bagging model, with an R^2 value of 0.92, an MSE of 1.19, and an MAE of 2.01, after hyperparameter tuning. Nevertheless, the variable importance, ease of interpretability, and the ability to handle non-linear relationships made the Random Forest model outperform the Decision Tree model. The optimal parameters and their values for the Decision Tree model were: criterion: Friedman MSE, max depth: 6, max features: 1.0, and min samples split: 0.25.

3. Model Performance Without Hyperparameter Tuning
The first assessment, conducted without any trial of hyperparameters, indicated that the Random Forest Regressor performed better than the Decision Tree Regressor. The Random Forest model had an MSE of 3.17, MAE of 2.04, and R^2 of 0.94, while the Decision Tree had an MSE of 5.6, MAE of 3.18, and R^2 of 0.80. This goes to illustrate the general superiority of ensemble learning methods, specifically in identifying intricate patterns, as shown in Figure 7

4. Model Performance with Hyperparameter Tuning
The performance of decision tree and random forest-based models was increased after adjusting the hyperparameters. The Random Forest model further optimized its performance, increasing its R^2 from 0.94 to 0.97, and the decision tree model's R^2 increased from 0.81 to 0.92, indicating enhanced accuracy as given in Table 6.

Table 6. The Model Performance after Hyperparameter Tuning

Performance Metrics	**Decision Tree Regression**	**Random Forest Regressor**
Mean squared error	1.19	1.17
Mean absolute error	2.01	2.00
Model score	0.92	0.97

5. RFE combined with RF
RFE combined with an RF estimator was employed to select the ten most important features from the dataset. RFE sequentially deleted features with lower importance in the model using the RF model. The chosen features were year, adult mortality, alcohol, BMI, under-five deaths, total expenditure, HIV / AIDS, thinness (5-9 years), income composition of resources, and schooling. With these features, the model achieved an R^2-score of 0.97162, indicating that the selected features were combined successfully. The result further highlights the capability of RFE to improve the interpretability and accuracy of models when analyzing only relevant predictors.

6. SFS combined with RF)
The SFS method determined 12 features as the most relevant to the prediction outcome. These selected features were country, year, status, adult mortality, infant deaths, under-5 deaths, polio, HIV/AIDS, GDP, the prevalence of thinness (ages 1–19), the prevalence of thinness (ages 5–9), and schooling. The model with these features was trained to have an R^2-score of 0.9623 for the training set and 0.9534 for the test set. This result concluded that SFS fits the best. Figure 8 compares the Decision Tree model performance under different feature selection methods. With all the features, the model produced an R^2 of 0.93 and an MSR of 6.36 on the model. With Recursive Feature Elimination (RFE) using the seven best features, the model achieved an R^2 value of 0.93, while the MSE was found to be 5.72. It was a clear indicator that feature selection is valuable in reducing residual error while maintaining a good level of determination on the model.

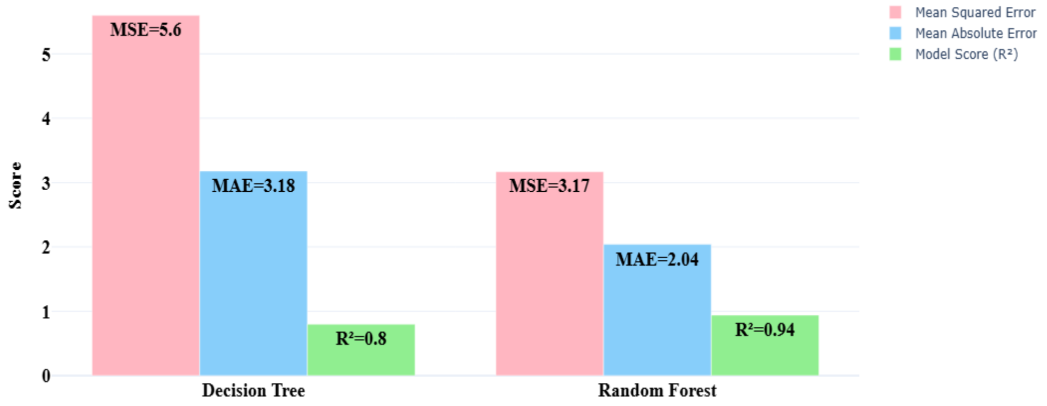

Figure 7. Assessing Model Performance without Hyperparameter Optimization

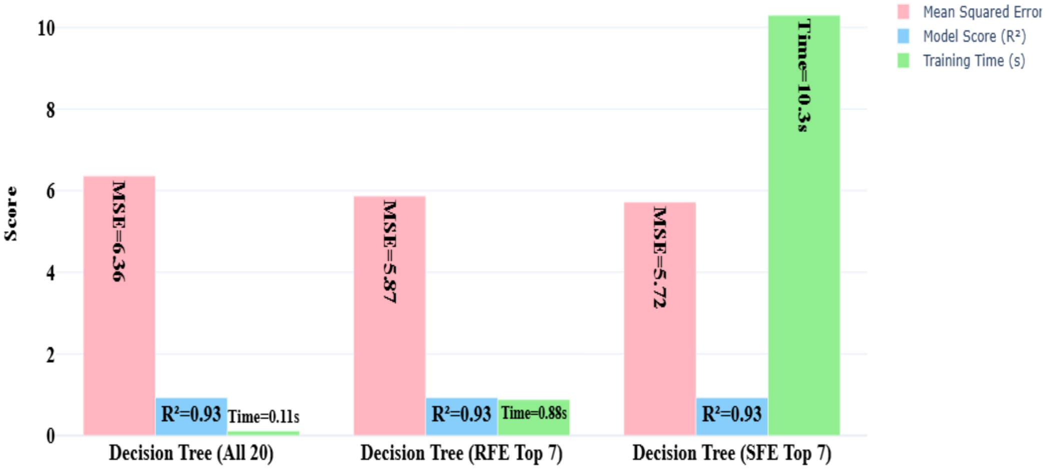

Figure 8. Feature Engineering Model Performance Comparison

Furthermore, Figure 9 presents the performance evaluation of two tree-based regression models. Both the models (the decision tree regressor and the extra trees regressor) used fifteen features. The Decision Tree model yielded an MSE of 11.41, an MAE of 1.79, and a coefficient of determination of 0.86. The Extra Trees Regressor confirmed an improved predictive power compared to the Decision Tree, with a lower MSE value of 11.15, an MAE of 1.76, and a slightly higher R^2 of 0.87. The Extra Tree-based model performed better, a testament to the applicability of ensemble methods in increasing prediction accuracy for life expectancy.

Table 7 presents a quantitative performance comparison of the developed model (current study) and the previous studies. Both the models (repetitive and bagging algorithm-based) demonstrate higher performance and reliability than the prior studies.

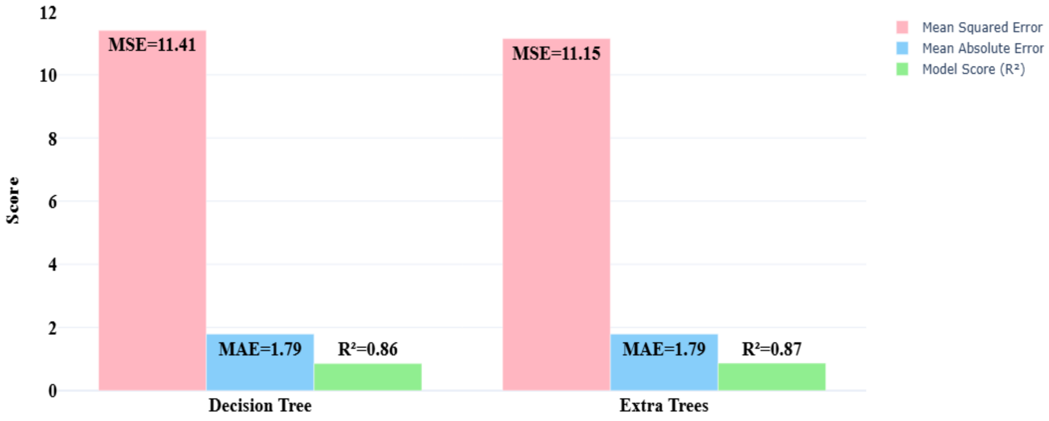

Figure 9. Performance Comparison of Decision Tree and Extra Trees Regressor Using PAC on a 15-Feature Dataset

Table 7. Summary of techniques used, dataset, model efficiency, and results

Ref.	Tech.	Dataset	Efficiency	Limitations
[1]	DT and RF	WHO	DT-R2=0.91 RF-R2=0.96	DT-MSR=1.55 RF-MSR=1.27
[7]	RF	WHO	RT=95	-
[14]	DT and RF	Limited Asian population	10-fold RT=81.42 10-fold RF=88.24	10-fold RMSE=0.27 10-fold RF-RMSE=0.19
[34]	DT and RF	Limited Asian population	20-fold RT=82.04 20-fold RF=87.62	10-fold RMSE=0.26 10-fold RF-RMSE=0.19
[34]	MLR	WHO	Train-R2=80% Test-R2=81%	-
[34]	DT	WHO	Train-R2=98.95% Test-R2=83.15%	-
[34]	DT	WHO	Model score=0.909	MSE=2.843
[34]	RF	WHO	Model score=0.958	MSE=1.930
[35]	Voting Regressor	WHO	Model score=0.947	MSE=4.693
[36]	Voting Regressor	WHO	Train-R2=0.99 Test=0.95%	MSE=4.43 MSE=1.58
[37]	Extra Tree Regression	-	R2=0.9729%	-

4.2. Life Expectancy Data Insight

As shown in Figure 10, the average life expectancy in selected Asian countries (Pakistan, India, China, Bangladesh, and Afghanistan) was analyzed from 2000 to 2015. The data indicated a general upward trend in life expectancy across all countries, with notable differences in the rate of increase.

China consistently maintained the highest average life expectancy throughout the period, peaking significantly around 2012 before stabilizing. Bangladesh experienced a sharp increase in life expectancy around the same time, which then plateaued. India and Pakistan have shown steady, incremental improvements in life expectancy since 2000, demonstrating a substantial upward trajectory, particularly after 2010. This analysis highlights the varying progress of these countries in health and living conditions. The graph shows that in 2005, Pakistan's average life expectancy suddenly dropped, which can be attributed to the 2005 Kashmir earthquake.

As shown in Figure 11, the average life expectancy across different continents from 2000-2015 highlights significant regional disparities. Europe consistently

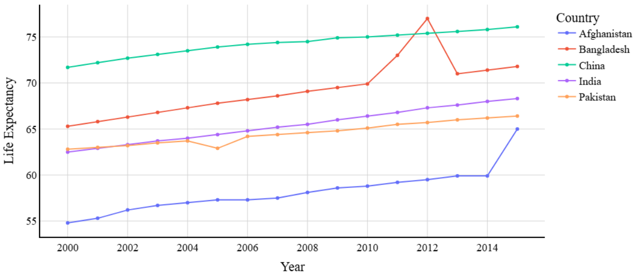

Figure 10. Avg Life Expectancy in Selected Asian Countries (2000–2015)

had the highest life expectancy, reaching approximately 82 years by 2015. Oceania and the Americas followed similar trends, peaking around 78 and 76 years, respectively. Asia showed a steady increase, with life expectancy reaching around 73 years by 2015. In Africa, despite significant improvements, life expectancy remained the lowest, starting below 50 years and rising to approximately 60 years by 2015. These trends reflect ongoing health and development challenges and progress in different regions.

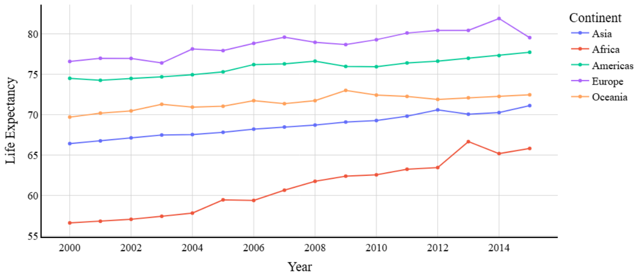

Figure 11. Avg Life Expectancy in Different Continents (2000–2015)

5. Conclusion

This research is a step toward the field of public health predictive modeling. The study developed robust and scalable life expectancy prediction models using recursive partitioning (decision tree) and bagging (random forest) algorithms using the World Health Organization (WHO) dataset. The raw data went through an appropriate data preprocessing sequence and feature engineering process before being fed into the model-building process. Decision tree (repetitive recursion) and random forest (bagging) models were found to predict more accurately. The random forest performed better among all the methods used due to its ability to combine multiple decision trees to achieve better accuracy and avoid overfitting.

In addition to its methodological implications, this study has practical importance for public health. The findings highlight the need for focused health spending in regions with low life expectancy, where issues like low immunization rates of major diseases like poliomyelitis, Hepatitis B, and Diphtheria, along with education and health facilities, are critical.

Furthermore, this study provided actionable recommendations for policymakers, highlighting that the application of big data approaches in critically relevant fields could enhance society's well-being. Health insurance companies can benefit by evaluating the risks of the underlying population more effectively. This study lays the foundation for future work in life expectancy prediction and the ongoing pursuit of global health equality and sustainable development.

5.1. Future Direction

The following research directions have been identified during this research study.

This study advocates for a coordinated global approach to consolidate datasets in light of the current global health situation. The dataset should be available in a single database for global reference for research and policy development. The test datasets should encompass recent trends, cross-regional comparisons, and other emerging health issues to utilize the efficiency of machine learning algorithms. Further research should also examine how additional indicators from socioeconomic, environmental, and health systems can be incorporated to reveal other factors that may influence human longevity. For such a complex and multi-variable dataset, it is possible to supplement the model development through deep learning.

References

[1] Bali, Vikram, et al. "Life Expectancy: Prediction & Analysis using ML." 2021 9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)(ICRITO). IEEE, 2021.

[2] Raja, S. Selvakumar, et al. "HUMAN LIFE EXPECTANCY PREDICTION USING MACHINE LEARNING." Ann. For. Res 66.1 (2023): 4035-4043.

[3] Maps on the Web. Life Expectancy by Country, 2019. URL https://mapsontheweb.zoom-maps.com/post/679623324847964160/life-expectancy-by-country-2019. Accessed: 2024-05-31.

[4] Liu, Lei-Lei, et al. "Dynamic prediction of landslide life expectancy using ensemble system incorporating classical prediction models and machine learning." Geoscience Frontiers 15.2 (2024): 101758.

[5] Gill, Kanwarpartap Singh, et al. ”Predicting Life Expectancy using Machine Learning Approach through Linear Regression and Decision Tree Classification Techniques.” 2023 3rd International Conference on Smart Generation Computing, Communication and Networking (SMART GENCON). IEEE, 2023.

[6] Amit, et al. "Evaluating Models for Better Life Expectancy Prediction." Computational Intelligence and Data Analytics: Proceedings of ICCIDA 2022. Singapore: Springer Nature Singapore, 2022. 389-404.

[7] Deshpande, Renuka, and Vaishnavi Uttarkar. "Life Expectancy using Data Analytics." International Journal for Research in Applied Science & Engineering Technology (IJRASET) 11 (2023): 972-978.

[8] Kerdprasop, Nittaya, and Kittisak Kerdprasop. "Association of economic and environmental factors to life expectancy of people in the Mekong basin." 2017 12th IEEE Conference on Industrial Electronics and Applications (ICIEA). IEEE, 2017.

[9] Ronmi, Akanmode Eyitayo, Rajesh Prasad, and Baku Agyo Raphael. "How can artificial intelligence and data

science algorithms predict life expectancy-An empirical investigation spanning 193 countries." International Journal of Information Management Data Insights 3.1 (2023):100168.

[10] Ho, Dennis Lim Kam, et al. "A Comparative Analysis of Machine Learning Techniques for Exploring Country Clustering Based on Life Expectancy." 2023 International Conference on Networking, Electrical Engineering, Computer Science, and Technology (IConNECT). IEEE, 2023.

[11] A. A. Bhosale and K. K. Sundaram. "Life prediction equation for human beings." In *2010 International Conference on Bioinformatics and Biomedical Technology*, pages 266–268. IEEE, 2010.

[12] Martin Cervantes, Pedro Antonio, Nuria Rueda Lopez, and Salvador Cruz Rambaud. "Life expectancy at birth in Europe: An econometric approach based on Random Forests methodology." Sustainability 12.1 (2020): 413.

[13] World Health Organization. World Health Organization [online], 2024. URL https://www.who.int/. Accessed: 2024-05-31.

[14] Joel, Luke Oluwaseye, Wesley Doorsamy, and Babu Sena Paul. "A review of missing data handling techniques for machine learning." International Journal of Innovative Technology and Interdisciplinary Sciences 5.3 (2022): 971-1005.

[15] Emmanuel, Tlamelo, et al. "A survey on missing data in machine learning." Journal of Big data 8 (2021): 1-37.

[16] Ruiz-Chavez, Zoila, Jaime Salvador-Meneses, and Jose Garcia-Rodriguez. "Machine learning methods based preprocessing to improve categorical data classification." Intelligent Data Engineering and Automated Learning–IDEAL 2018: 19th International Conference, Madrid, Spain, November 21–23, 2018, Proceedings, Part I 19. Springer International Publishing, 2018.

[17] Guedrez, Rabah, et al. "Label encoding algorithm for MPLS segment routing." 2016 IEEE 15th International Symposium on Network Computing and Applications (NCA). IEEE, 2016.

[18] Shah, Deval, Zi Yu Xue, and Tor M. Aamodt. "Label encoding for regression networks." arXiv preprint arXiv:2212.01927 (2022).

[19] Zhang, Kai, and Minxia Luo. "Outlier-robust extreme learning machine for regression problems." Neurocomputing 151 (2015): 1519-1527.

[20] Yang, Jiawei, Susanto Rahardja, and Pasi Fränti. "Outlier detection: how to threshold outlier scores?." Proceedings of the international conference on artificial intelligence, information processing and cloud computing. 2019.

[21] Jo, Jun-Mo. "Effectiveness of normalization preprocessing of big data to the machine learning performance." The Journal of the Korea institute of electronic communication sciences 14.3 (2019): 547-552.

[22] Darst, Burcu F., Kristen C. Malecki, and Corinne D. Engelman. "Using recursive feature elimination in random forest to account for correlated variables in high dimensional data." BMC genetics 19 (2018): 1-6.

[23] Mayer, Joshua, et al. "Sequential feature selection and inference using multi-variate random forests." Bioinformatics 34.8 (2018): 1336-1344.

[24] Alvi, Muhammad Bux, et al. "An effective framework for tweet level sentiment classification using recursive text pre-processing approach." International Journal of Advanced Computer Science and Applications 10.6 (2019).

[25] Faisal, Khulood, et al. "Life expectancy estimation based on machine learning and structured predictors." Proceedings of the 3rd International Conference on Advanced Information Science and System. 2021.

[26] Meshram, Siddhant Sunil. "Comparative analysis of life expectancy between developed and developing countries using machine learning." 2020 IEEE Bombay Section Signature Conference (IBSSC). IEEE, 2020.

[27] Charbuty, Bahzad, and Adnan Abdulazeez. "Classification based on decision tree algorithm for machine learning." Journal of Applied Science and Technology Trends 2.01 (2021): 20-28.

[28] Quinlan, J. Ross. "Induction of decision trees." Machine learning 1 (1986): 81-106.

[29] Loh, Wei-Yin. "Classification and regression trees." Wiley interdisciplinary reviews: data mining and knowledge discovery 1.1 (2011): 14-23.

[30] Mienye, Ibomoiye Domor, and Yanxia Sun. "A survey of ensemble learning: Concepts, algorithms, applications, and prospects." IEEE Access 10 (2022): 99129-99149.

[31] Khan, Azal Ahmad, Omkar Chaudhari, and Rohitash Chandra. "A review of ensemble learning and data augmentation models for class imbalanced problems: combination, implementation and evaluation." Expert Systems with Applications 244 (2024): 122778.

[32] Ali, Peshawa Jamal Muhammad, et al. "Data normalization and standardization: a technical report." Mach Learn Tech Rep 1.1 (2014): 1-6.

[33] Cinaroglu, Songul, and Onur Baser. "Comparative regression performances of machine learning methods optimising hyperparameters: application to health expenditures." International Journal of Bioinformatics Research and Applications 16.4 (2020): 387-407.

[34] Selvaraj, Gayathri, Punithavalli Muthuswamy, and Chaitanya Vasanth Kumar. "Alcohol Expectancy Prediction Using Fuzzy C-Regression Based Structural Brain Imaging." International Journal of Intelligent Engineering & Systems 12.5 (2019).

[35] Pisal, Nurul Shahira, et al. "Prediction of life expectancy for Asian population using machine learning algorithms." Malaysian Journal of Computing 7.2 (2022): 1150-1161.

[36] Pandey, Anshu, and Rita Chhikara. "Analysis of life expectancy using various regression techniques." 2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN). IEEE, 2020.

[37] T. Choudhury, S. K. Bharti, M. Kumar Gourisaria, J. J. Jena, D. Kumar Behera and A. Bandyopadhyay, "Predictive Modeling of Life Expectancy Using Machine Learning Algorithms," 2024 Global Conference on Communications and Information Technologies (GCCIT), BANGALORE, India, 2024, pp. 1-6, doi: 10.1109/GCCIT63234.2024.10862085.