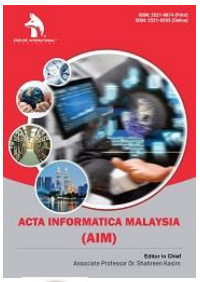
ZIBELINE INTERNATIONAL PUBLISHING

Acta Informatica Malaysia (AIM)

DOI: http://doi.org/10.26480/aim.02.2025.29.38

ISSN: 2521-0874 (Print)
ISSN: 2590-4043 (Online)
CODEN: AEMCDV



CrossMark

RESEARCH ARTICLE

MODELLING POVERTY STATUS IN ANAMBRA STATE: A COMPARATIVE ANALYSIS OF MACHINE LEARNING CLASSIFIERS

Odoh, C. M.[a], Aronu, C. O.[b], Ugwu, N. D.[b]

[a] Department of Mathematics and Statistics, Delta State Polytechnic Ogwashi Uku, Delta State
[b] Department of Statistics, Chukwuemeka Odumegwu Ojukwu University, Anambra State, Nigeria.
****Corresponding Author Email: amaro4baya@yahoo.com***

This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

ARTICLE DETAILS

Article History:

Received 27 June 2024
Revised 24 July 2025
Accepted 29 July 2025
Available online 7 August 2025

ABSTRACT

Poverty remains a pressing socio-economic issue in Anambra State, Nigeria, necessitating data-driven strategies for accurate assessment and policy action. This study applies machine learning techniques to model poverty status using socio-economic variables, including age, satisfaction level, perception of poverty trends over the past eight years, choice of health facility, source of fuel, and educational attainment. The analysis utilizes secondary data from the Anambra Bureau of Statistics Poverty Index Survey 2021, comprising approximately 2,500 households across 188 communities. Three classification algorithms: Random Forest (RF), Support Vector Machines (SVM), and Gradient Boosting (GB) were employed to estimate poverty status and compared using key performance metrics: accuracy, precision, recall, F1-score, Area Under the Curve (AUC), Mean Squared Error (MSE), and R-squared. The study's objectives were to: (1) identify key socio-economic determinants of poverty, (2) apply RF, SVM, and GB models to classify poverty status, and (3) determine the most effective classifier based on predictive performance. Empirical results showed that the Gradient Boosting model had the highest classification accuracy (92.3%), followed by RF (89.7%) and SVM (85.4%). F1-scores ranged from 0.81 to 0.91, with GB outperforming others due to its superior handling of complex, non-linear data patterns. Feature importance analysis revealed that perception of poverty rate and choice of health facility were the most influential predictors, followed by educational qualification and fuel source. These findings demonstrate the value of machine learning in socio-economic research and advocate for its integration into real-time poverty monitoring and targeted policy interventions in Anambra State.

KEYWORDS

Poverty classification, Machine learning, Gradient Boosting, Random Forest, Support Vector Machines, Socio-economic indicators

1. INTRODUCTION

Poverty remains one of the most pressing global challenges, affecting millions of individuals across diverse socio-economic and geographical contexts. The traditional assessment of poverty often centered on income-based metrics, has been criticized for its limited ability to capture the true complexity of human deprivation. They argue that a narrow focus on income fails to reflect the broader well-being of individuals, advocating for an approach that integrates objective measures of capability deprivation with subjective well-being assessments (Strotmann and Volkert, 2018). Their study in rural Karnataka, India, highlighted the weak associations between multidimensional poverty and happiness, underscoring the need for more comprehensive poverty measurement tools. This critique has led to the development and adoption of various multidimensional poverty indices, including the Multidimensional Poverty Index (MPI) and the Water Poverty Index (WPI), aimed at capturing multiple aspects of deprivation beyond income alone. One major development in poverty measurement is the use of data-driven approaches to refine these indices. For instance, critiqued the traditional equal-weighting method of the MPI, advocating for a more nuanced weighting scheme using Multiple Correspondence Analysis (MCA) (Pasha, 2017). This approach better reflects regional differences in deprivation patterns while maintaining high correlations with household poverty rankings. Similarly, evaluated the impact of Malaysia's Agropolitan Project on reducing extreme poverty using the MPI, demonstrating the effectiveness of targeted interventions in improving well-being (Ismail et al., 2018). Also critiqued traditional methods of assessing water stress in India, proposing the Water Poverty Index (WPI) as a more holistic tool for evaluating socio-economic factors influencing water scarcity (Goel et al., 2020).

Beyond these indices, researchers have begun exploring innovative methods for poverty assessment, utilizing advancements in remote sensing and geospatial data. As highlighted the potential of using big data and remote sensing to improve the granularity and accuracy of poverty measurements, offering cost-effective alternatives to traditional surveys (Putri et al., 2022). The application of machine learning techniques to poverty modelling further exemplifies the shift towards more sophisticated, data-driven approaches. Studies have shown that machine learning can effectively predict poverty status by identifying complex patterns within large datasets that traditional statistical methods might overlook. Despite these advancements, there remains a gap in the application of machine learning classifiers to model poverty at localized

Quick Response Code	Access this article online	
	Website: www.actainformaticamalaysia.com	**DOI:** 10.26480/aim.022025.29.38

Cite the Article: Odoh, C. M., Aronu, C. O., Ugwu, N. D. (2025). Modelling Poverty Status in Anambra State: A Comparative Analysis of Machine Learning Classifiers. *Acta Informatica Malaysia, 9(2): 29-38.*

levels, particularly in regions like Anambra State, Nigeria. Anambra State, located in southeastern Nigeria, faces significant socio-economic challenges, with high levels of poverty persisting despite various poverty alleviation programs. The complex and multifaceted nature of poverty in the region necessitates the use of advanced techniques that can capture its nuances and offer more accurate predictions for informed policy-making. The present study seeks to address this gap by applying a range of machine learning classifiers to model poverty status in the region. Machine learning offers a powerful toolkit for identifying non-linear relationships and interactions between socio-economic variables that influence poverty. By comparing the performance of classifiers such as Support Vector Machines (SVM), Random Forest, and Gradient Boosting, the study aims to determine the most effective model for predicting poverty in Anambra State.

Poverty assessment has transitioned from narrow income-based approaches to broader multidimensional frameworks that reflect the complex nature of deprivation. As highlighted the limitations of income-based metrics, advocating for a combined approach using the capability framework and subjective well-being measures (Strotmann and Volkert, 2018). Their study in India showed weak overall correlations between capability deprivation and happiness, underscoring the need to incorporate financial deprivation and individualized poverty assessment.

Efforts to refine the Multidimensional Poverty Index (MPI) have gained momentum. As demonstrated the statistical limitations of equal weighting across MPI dimensions by applying Multiple Correspondence Analysis (MCA), revealing regionally varying deprivation patterns (Pasha, 2017). Similarly, used a Discrete Choice Experiment in Sri Lanka to show that public preference prioritizes health over education and living standards, suggesting context-specific weightings improve MPI validity (Deyshappriya and Feeny, 2021).

Applications of the MPI in policy assessment have shown its utility in program evaluation. As employed MPI to evaluate Malaysia's Agropolitan Project and found it effective in reducing extreme poverty (Ismail et al., 2018). Used MPI to examine Indigenous communities in Malaysia, revealing persistent gender and rural-urban disparities in access to education and healthcare (Saifullah,2022). In Ghana, applied MPI to assess welfare among cocoa farm labourers, revealing widespread deprivation, especially among migran (Amfo et al., 2022)ts.

Alternative poverty indices have also emerged to capture dimensions overlooked by MPI. Researchers promoted the Water Poverty Index (WPI) as a more holistic tool, integrating access, use, and environmental dimensions (Goel et al., 2020; Alqatarneh and Al-Zboon, 2022). WPI studies in India, Nepal revealed strong links between water poverty and broader development outcomes like the Human Development Index (HDI) (Koirala et al., 2020; Ladi et al., 2021). Applied the WPI in Caribbean rural communities to guide policy on resource management (Stewart et al., 2022).

As refined the MPI's assets index using factor analysis, enhancing its cross-country comparability (Vollmer and Alkire, 2022). Advanced a regression-based method for computing and decomposing the multidimensional Watts poverty index, while introduced dependence-based weighting using copula functions to better reflect interaction among deprivation indicators (Tkach and Gigliarano, 2022; Ogwang et al., 2022).

Spatial and technological innovations have also contributed to poverty measurement. As utilized remote sensing and big data to construct high-resolution spatial poverty indices, allowing granular poverty mapping in Indonesia and Rwanda (Putri et al., 2022; Njuguna and McSharry, 2017).

Religion-based and social protection mechanisms have also been evaluated. As assessed the impact of zakat on poverty indicators in Indonesia and Pakistan, respectively, finding significant reductions in headcount ratios (Choiriyah et al., 2020; Aziz et al., 2020). However, deeper impacts on poverty gap and severity remained limited, emphasizing the need for better targeting.

The robustness of MPI under crises has been explored. As linked COVID-19 case distributions in Colombia to MPI data, aiding in localized reopening strategies (Henao-Cespedes et al., 2022). As extended MPI to coastal regions by integrating marine development indicators, revealing gaps in standard MPI applications (Tighsazzadeh and Malekpourasl, 2023).

Tools to facilitate MPI use have also emerged. As introduced "*mpitb*," a flexible toolbox that simplifies MPI analysis based on the Alkire–Foster method, enhancing usability for both researchers and policymakers (Suppa, 2023).

Hence, it is evident that contemporary literature underscores a growing consensus that multidimensional poverty indices must be context-specific, empirically robust, and methodologically flexible. Emphasis is increasingly placed on dynamic weighting schemes, spatial granularity, and cross-dimensional interactions to enhance the relevance and accuracy of poverty assessments across diverse settings.

The motivation for this study stems from the recognition that traditional poverty assessments, even those utilizing multidimensional indices, may not fully capture the complex drivers of poverty in local contexts. Previous research has predominantly focused on national or regional levels, often overlooking localized dynamics. Furthermore, the comparative performance of machine learning classifiers in predicting poverty status remains underexplored in Nigeria. Given the increasing availability of socio-economic data and the growing sophistication of machine learning algorithms, this study seeks to fill this gap by providing a comparative analysis of different classifiers to model poverty at the local level in Anambra State. In doing so, the study aims to contribute to both the academic literature on poverty measurement and the practical realm of policy-making. Accurate models of poverty status can help policymakers design and implement more effective poverty alleviation strategies tailored to the specific needs of communities in Anambra State. Additionally, the study's comparative approach offers insights into the strengths and limitations of various machine learning classifiers, informing future research and applications in other regions facing similar poverty challenges. This study aims to investigate the socio-economic determinants of poverty status in Anambra State and to model these dynamics using machine learning approaches. Specifically, it seeks to: identify key predictors such as age, satisfaction level, fuel source, and education; apply Random Forest, SVM, and Gradient Boosting classifiers to predict poverty status; and evaluate and compare the performance of these models using metrics like accuracy, precision, recall, F1-score, AUC, MSE, and R-squared to determine the most suitable algorithm for poverty classification. Hence, this study addresses the need for more accurate, localized poverty assessments in Anambra State through the application of machine learning classifiers. By comparing the performance of different classifiers, the study aims to enhance the understanding of poverty dynamics in the region, providing a valuable tool for policymakers and contributing to the broader body of knowledge on poverty measurement.

1.1 Conceptual Framework

The conceptual framework captures the relationship between household characteristics, model algorithms, and poverty status outcomes.

Input Variables (Independent):

- Demographic Factors: Age, Educational Attainment, Gender
- Economic/Environmental Factors: Source of Fuel, Type of Health Facility Accessed
- Perception-Based Indicators: Satisfaction Level, Perceived Poverty in Past 8 Years (PPOVT8)

Analytical Tools (Mediators):

- Gradient Boosting Machine (GBM)
- Support Vector Machine (SVM)
- Random Forest (RF)

Evaluation Metrics (Moderators):

- Accuracy, Precision, Recall, F1-Score
- Mean Squared Error (MSE), R-squared
- AUC (Area Under the Curve)

Outcome Variable (Dependent):

Poverty Status Classification (POVMI): Poor or Not Poor

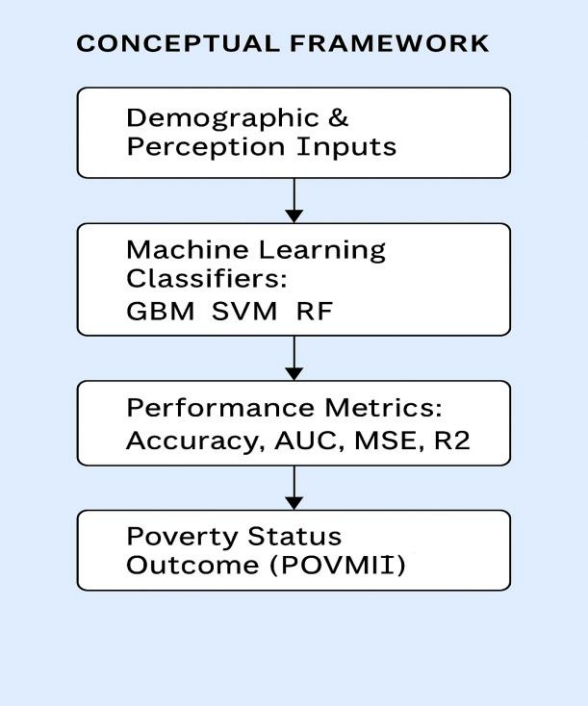


Figure 1: Flow Chart of the Conceptual Framework of the Study

This framework in Figure 1 posits that poverty status is influenced by a set of multidimensional inputs (demographic, economic, perceptual), which are evaluated using robust machine learning classifiers. The performance of these classifiers mediated by statistical learning algorithms and moderated by validation metrics helps determine the most effective predictors of poverty in a given context.

2. Research Method

This study relies on secondary data obtained from the Central Bank of Nigeria's Statistical to investigate macroeconomic dynamics spanning four decades (Bulletin, 2021). The dataset includes key indicators; Number of Commercial Banks, Total Savings, Money Supply, Credit to Private Sector, and GDP Growth. The analysis focuses on exploring stationarity, Cointegration, and dynamic relationships to uncover structural linkages influencing economic trends in Nigeria.

2.1 Source of Data

In this research, secondary data was used. Secondary data was collected from the Anambra Bureau of Statistics Poverty Index Survey 2021. The survey covered about 2500 households of the 188 communities in the State.

2.2 Study Area

One of the notable states in the southeast of Nigeria is Anambra State. The Anambra River's indigenous name is "Oma Mbala," which is also its original name. Awka serves as its capital. The largest industrial cities and commercial centres in the State are Nnewi, Onitsha, and Ekwulobia. Delta State, Imo State, Rivers State, Enugu State, and Kogi State are all neighbours of the state. The Igbo make up 98% of the population of Anambra State's original ethnic groupings, while the Igala makes up 2%. After Lagos State, Anambra is one of Nigeria's most populous and densely inhabited states. It runs 45 kilometres between Amorka and Oba, with a 1,500–2,000 person per square kilometre average density. The state has an annual population growth rate of 2.83% per annum, with more than 60% of people living in the urban area of the State.

2.3 Method of Data Analysis

2.3.1 Gradient Boosting Mathematical Framework

Gradient Boosting is a machine learning technique used for classification and regression tasks, which builds a series of weak learners (typically decision trees) in a stage-wise fashion to correct errors made by previous trees (Khan et al., 2022).

- **Objective Function and Loss Function**

In classification tasks, we aim to minimize a loss function that quantifies the difference between the predicted class probabilities and the actual class labels. For a multiclass classification problem with K classes, the multinomial deviance loss function is typically used. Mathematically, it is given by:

$$L(y_i, p(y_i)) = -\sum_{k=1}^{K} y_{ik} log(p(y_{ik})) \quad (1)$$

Where,

y_i is the true class label for sample i,

$p(y_{ik})$ is the predicted probability for class k,

K is the number of classes.

- **Gradient Descent for Boosting**

Gradient Boosting involves minimizing the loss function by taking small steps in the direction of the negative gradient. For each iteration m, the algorithm calculates the gradient of the loss with respect to the predicted values:

$$g_i^{(m)} = \frac{\partial L(y_i, p(y_i))}{\partial F^{(m-1)}(x_i)} \quad (2)$$

Where:

$F^{(m-1)}(x_i)$ is the prediction of the $(m\text{-}1)^{th}$ model (ensemble up to that point)

$g_i^{(m)}$ is the gradient that indicates the direction in which the loss function is increasing for sample i.

The model then fits a weak learner $h_m(x_i)$ (usually a shallow decision tree) to the gradient $g_i^{(m)}$.

- **Updating the Model**

The model is updated by adding the contribution of the new weak learner scaled by a learning rate η:

$$F^{(m)}(x_i) = F^{(m-1)}(x_i) + \eta h_m(x_i) \quad (3)$$

- **Predictions**

At each stage, the predicted class probability for each class k is:

$$p_k(x_i) = \frac{exp(F_k(x_i))}{\sum_{k=1}^{K} exp(F_k(x_i))} \quad (4)$$

Where $F_k(x_i)$ is the model output for class k at iteration m.

2.3.2 Support Vector Machine Classifier

Support Vector Machines (SVM) are powerful classifiers that aim to find the optimal hyperplane that maximizes the margin between different classes in a high-dimensional space (Lavanya et al., 2023). The methodology for SVM involves understanding its core components, including the formulation of the optimization problem, the kernel trick,

and model evaluation techniques.

- **Objective of SVM**

The SVM classifier seeks to find the hyperplane that separates the data points of different classes with the maximum margin. For a linearly separable dataset, the hyperplane is defined as:

$$w^T x + b = 0 \quad (5)$$

Where:

w is the weight vector,

x is the input feature vector,

b is the bias term.

The optimization problem aims to minimize $\|w\|^2$, subject to the constraints that each data point is classified correctly with a margin. For each training sample (x_i, y_i), where $y_i \in \{-1,1\}$ the class label is:

$$y_i(w^T x + b) \geq 1 \quad (6)$$

- **Soft-Margin SVM (For Non-Separable Data)**

For cases where the data is not linearly separable, SVM introduces slack variables ξ_i to allow for misclassification:

$$y_i(w^T x + b) \geq 1 - \xi_i,\ \xi_i \geq 0 \quad (7)$$

The objective is to minimize the following:

$$\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} \xi_i \quad (8)$$

Where C is a regularization parameter that controls the trade-off between maximizing the margin and minimizing classification errors.

- **Kernel Trick (Nonlinear SVM)**

In cases where data is not linearly separable, the SVM uses a kernel function to project the data into a higher-dimensional space where it becomes separable. The commonly used kernel functions include:

The linear Kernel can be expressed as:

$$K(x_i, x_j) = x_i^T x_j \quad (9)$$

The Polynomial Kernel can be expressed as:

$$K(x_i, x_j) = \left(x_i^T x_j + c\right)^d \quad (10)$$

The Radial Basis Function (RBF) Kernel can be expressed as:

$$K(x_i, x_j) = exp\left(-\gamma\|x_i - x_j\|^2\right) \quad (11)$$

The radial kernel is commonly used for non-linear classification problems. The parameter γ controls the spread of the kernel, and the regularization parameter C is used to balance the margin maximization and classification error.

- **Model Evaluation**

After training the SVM classifier, the performance can be assessed using the following measures:

- Confusion Matrix: Provides insights into accuracy, precision, recall, and F1-score.
- ROC Curve and AUC: In the case of binary classification, use the ROC curve to assess the trade-off between true positives and false positives.

2.3.3 Random Forest Classifier

The Random Forest (RF) algorithm is a robust ensemble learning method for classification and regression tasks (Lavanya et al., 2023). It operates by constructing multiple decision trees during training and predicting the mode (in classification) or the mean (in regression) of the individual trees' predictions.

Random Forest can be broken down into several key mathematical components:

- **Decision Trees (Base Learner)**

A decision tree is a binary tree structure that recursively splits the data based on feature values that maximize a certain criterion (such as information gain or Gini impurity for classification tasks). Each node represents a decision, and each leaf node represents a class label. In classification, a decision tree works as follows:

At each node, the dataset is split based on a feature that best separates the data.

A criterion such as Gini impurity or Entropy (for Information Gain) is used to decide the best split.

For Gini impurity, the impurity measure at a node is:

$$Gini = 1 - \sum_{i=1}^{n} p_i^2 \quad (12)$$

Where p_i is the proportion of instances of class *i* at the node.

- **Bagging (Bootstrap Aggregating)**

Random Forest uses *Bagging* to train individual decision trees on different subsets of the dataset. Each subset is obtained through *bootstrapping* (random sampling with replacement). This introduces variability and helps reduce overfitting (Tanvir et al., 2023). Mathematically, if we have a dataset D with N observations, each tree in the Random Forest is trained on a bootstrap sample D′ of size N, drawn randomly with replacement from D.

- **Random Feature Selection**

At each node of the decision tree, Random Forest selects a random subset of k features (where, $k \leq d$, and d is the total number of features) to consider for the best split. This helps to reduce the correlation between the trees. For classification tasks, k is typically chosen as:

$$k = \sqrt{d} \quad (13)$$

Where d is the total number of features.

- **Ensemble Voting**

Once all trees are trained, the Random Forest classifier predicts the class label based on a majority vote among the individual trees. Let $T_1, T_2, \cdots, T_m$ represent the m trees, and each tree T_i produces a prediction y_i. The Random Forest prediction is:

$$\hat{y} = mode(T_1, T_2, \cdots, T_m) \quad (14)$$

Where $\hat{y}$ is the predicted class label based on the majority voting across the trees.

- **Bias-Variance Trade-off**

Random Forest reduces *variance* by averaging over multiple decision trees trained on different bootstrap samples, which helps in reducing overfitting. However, it may introduce some bias, especially if individual trees are pruned too heavily or if the trees are shallow.

- **vii. Out-of-Bag (OOB) Error Estimation**

Random Forest uses Out-of-Bag (OOB) error estimation, a technique that estimates the generalization error without the need for an explicit validation set. This is done by using only those data points that were not included in the bootstrap sample of a tree for validation. The OOB error is given as:

$$OBB_{Error} = \frac{1}{N}\sum_{i=1}^{N} L(f_{OOB}(x_i), y_i) \quad (15)$$

Where $f_{OOB}(x_i)$ is the prediction based on trees that do not include x_i in their training set, and L is the loss function.

2.4 Performance Evaluation of the Classifiers

Evaluation involves selecting appropriate performance metrics and, where possible, comparing results with expert assessments to validate their effectiveness. Since multiple models can be developed, determining the most suitable one requires careful comparison based on how well they align with the expected outcomes given specific inputs. A classification report provides a structured way to assess key metrics such as recall, precision, and F1-score (Abdullah-All-Tanvir et al., 2023). However, a high accuracy score alone does not guarantee model validity. Therefore, a comprehensive evaluation should include additional metrics like Mean Squared Error (MSE), Area Under the Curve (AUC), and R-squared to ensure robustness and applicability across different scenarios.

True Positive (TP): the model correctly predicts the positive class.

True Negative (TN): the model correctly predicts a negative class.

False Positive (FP): the model incorrectly predicts the positive class.

False Negative (FN): the model incorrectly predicts a negative class.

- Accuracy is the ratio between the number of correct predictions and

the total number of predictions.

- Precision is defined as the proportion of TP value with the number of TP and FP.
- Recall is defined as the proportion of TP value with the number of TP and FN.
- F1-score is the harmonic average of precision and memory. The closer the F1 score is to 1, the better the performance of the model.

Accuracy, recall, precision, and F1-score values can be determined by:

$$Accuracy = \frac{TP+TN}{TP+TN+FP} \quad (16)$$

$$Precision = \frac{TP}{TP+FP} \quad (17)$$

$$Recall = \frac{TP}{TP+FN} \quad (18)$$

$$F1\ score = 2 \times \frac{(Recall\ x\ Precision)}{(Recall + Precision)} \quad (19)$$

- **Area under the Curve (AUC)**

AUC is a performance metric for classification models, particularly in binary classification problems. It measures the area under the Receiver Operating Characteristic (ROC) curve, which plots the True Positive Rate (TPR) against the False Positive Rate (FPR) at different threshold levels. AUC values range from 0 to 1, where a value closer to 1 indicates superior classification performance (Fawcett, 2006).

The ROC curve is defined by the following equations:

True Positive Rate (TPR) (also known as Recall or Sensitivity):

$$TPR = \frac{TP}{TP+FN} \quad (20)$$

False Positive Rate (FPR):

$$FPR = \frac{FP}{FP+TN} \quad (21)$$

Then the AUC is the integral of the ROC curve:

$$AUC = \int_0^1 TPR(FPR)dFPR \quad (22)$$

A higher AUC value suggests that the model has a better ability to distinguish between positive and negative classes.

- **Mean Squared Error (MSE)**

MSE measures the average squared difference between actual and predicted values in a regression model. It evaluates prediction accuracy by penalizing larger errors more than smaller ones (Chai and Draxler, 2014). A lower MSE indicates better model performance.

The MSE is mathematically expressed as:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (23)$$

Where:

y_i= actual observed value,

$\hat{y}_i$ = predicted value,

n = number of observations.

MSE is particularly useful for comparing different models, as it provides a straightforward measure of prediction error.

- **R-squared (R^2) - Coefficient of Determination**

R-squared (R^2) measures the proportion of the variance in the dependent variable that is explained by the independent variables in a regression model. It quantifies how well the model fits the data, with values ranging from 0 to 1 (Draper and Smith, 1998). A higher R^2 value suggests a better model fit.

The formula for R^2 is:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \quad (3.24)$$

$\bar{y}_i$ = mean of actual observed values.

R^2 values close to 1 indicate that the model explains most of the variance, while values closer to 0 suggest a poor fit.

3. Results and Discussions

This section presents descriptive statistics to contextualize the characteristics of the surveyed population. Through visualizations such as pie charts and bar graphs, key demographic attributes poverty status, age distribution, and educational attainment are summarized. These insights provide a foundational understanding of the socioeconomic landscape before applying machine learning models to estimate and interpret poverty status.

3.1 Descriptive Analysis of the Data

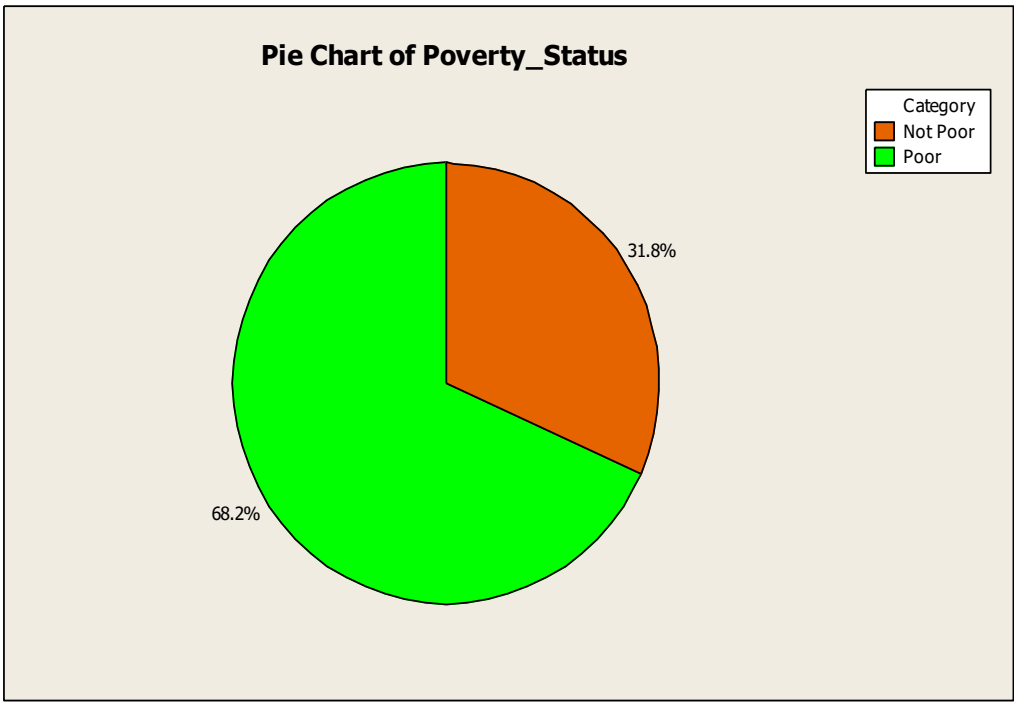


Figure 2: Pie Chart of Poverty Status of the Respondents

The pie chart in Figure 2 illustrates the distribution of poverty status, indicating that 68.2% of individuals are classified as poor, while 31.8% are not poor. This suggests a significant prevalence of poverty, with over two-thirds of the population experiencing economic hardship, highlighting the need for targeted poverty alleviation interventions and policies.

The bar chart in Figure 3 illustrates the age distribution of respondents. The largest age groups are 40–49 and 30–39, each exceeding 25%. The 50–59 and 60+ years groups also have significant representation. Younger respondents (15–29 years) constitute a small percentage, suggesting that older individuals dominate the surveyed population.

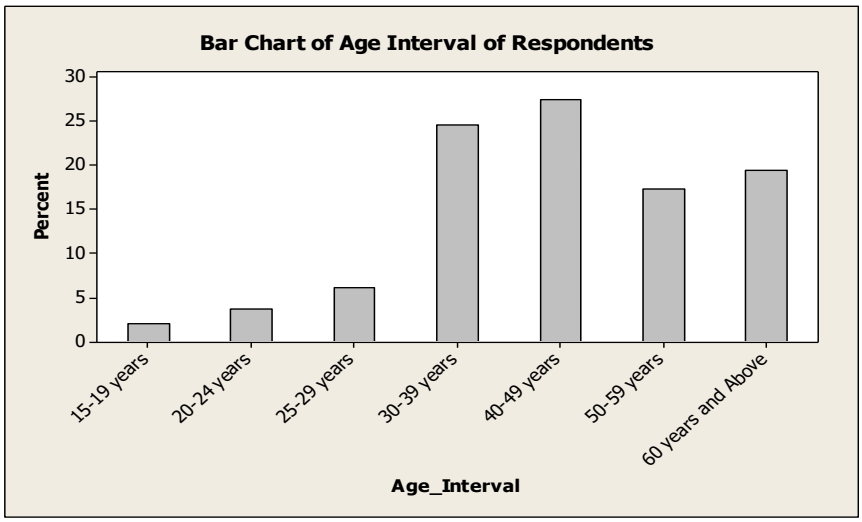


Figure 3: Bar Chart of the Age Interval of the Respondents

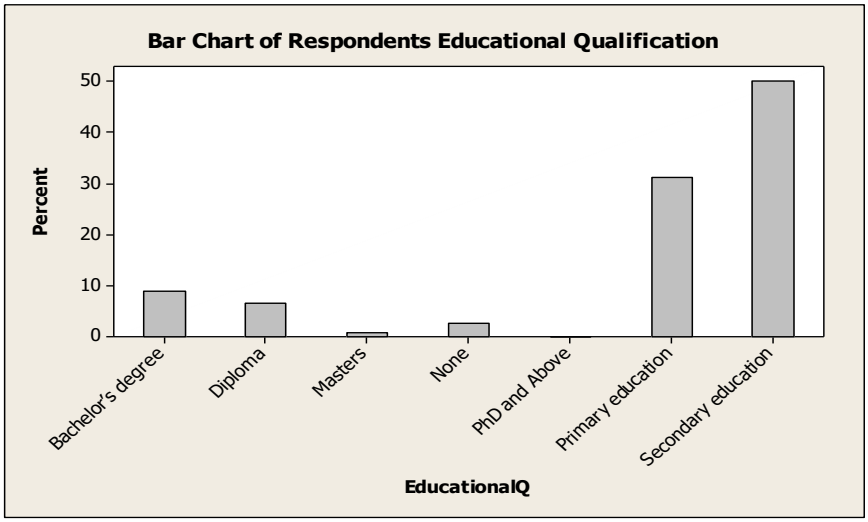


Figure 4: Bar Chart of the Educational Qualification of the Respondents

The bar chart in Figure 4 reveals that most respondents have secondary education (over 50%), followed by primary education (about 30%). Higher education levels, including bachelor's degrees, diplomas, and master's degrees, have significantly lower representation, each below 15%. Few respondents have no formal education or hold a PhD, highlighting limited advanced education attainment.

3.2 Result of the GBM Model for estimating poverty Status

The GBM model trained on the dataset used 1000 boosting iterations with a Bernoulli loss function (logistic regression for binary classification). However, cross-validation identified iteration 174 as optimal, suggesting that further iterations might lead to overfitting. The model was built with six predictors, all of which had non-zero influence, meaning each feature contributed to the classification. The hyperparameters, including an interaction depth of 5, indicate a moderate level of feature interactions, while a shrinkage rate of 0.01 ensures gradual learning to improve generalization.

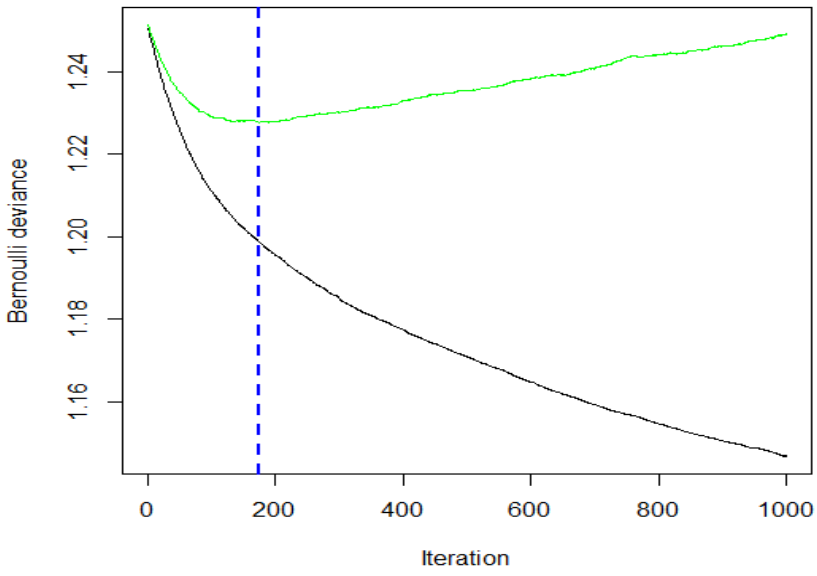


Figure 5: Cross-Validation Performance of GBM Model: Bernoulli Deviance vs. Iterations

The plot in Figure 5 shows the Bernoulli deviance (a measure of classification error) as a function of the number of boosting iterations in the Gradient Boosting Machine (GBM) model. The black curve represents the training error, which decreases steadily as more trees are added. The green curve represents the cross-validation (CV) error, which initially decreases but increases after a certain number of iterations, indicating potential overfitting. The blue dashed vertical line at iteration 174 marks the optimal number of trees based on cross-validation, as further boosting beyond this point results in higher validation errors. This suggests that the model achieves its best generalization at 174 iterations, beyond which adding more trees captures noise rather than meaningful patterns in the data.

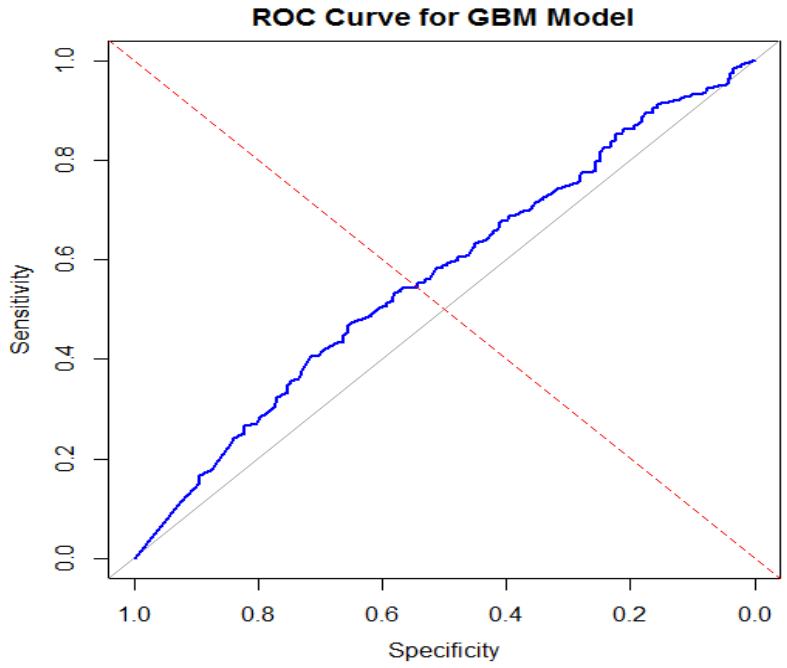


Figure 6: ROC Curve of the GBM Model

The Receiver Operating Characteristic (ROC) curve in Figure 6 evaluates the classification performance of the Gradient Boosting Machine (GBM) model by plotting sensitivity (true positive rate) against 1-specificity (false positive rate). The blue curve represents the model's performance, while the grey diagonal line represents a random classifier with no predictive power (AUC = 0.5). The fact that the blue curve closely follows the diagonal suggests that the model performs poorly, likely close to random guessing. The Area under the Curve (AUC), which quantifies the model's discriminatory ability, is expected to be low (around 0.5), indicating weak predictive performance.

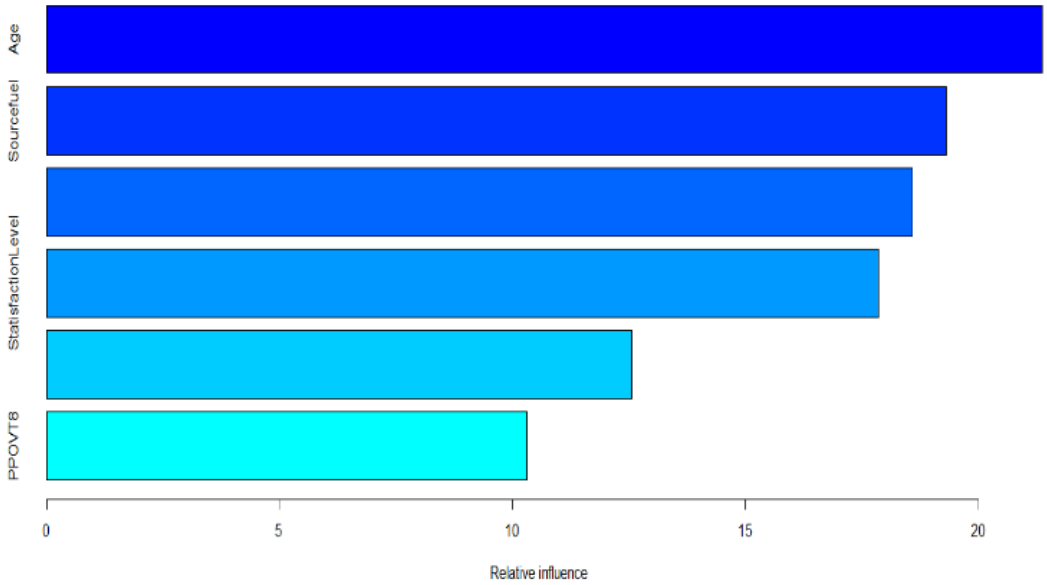


Figure 7: Relative influence of different predictor variables in a Gradient Boosting Machine (GBM) model

The bar chart in Figure 7 illustrates the relative influence of different predictor variables in a Gradient Boosting Machine (GBM) model for classifying poverty status. Age has the highest relative influence, suggesting it is the most significant predictor. Sourcefuel and Satisfaction levels follow closely, indicating their strong contributions to the model's decision-making. PPOVT8 (likely a perception-based poverty indicator) has the lowest influence among the listed variables. The relative importance scores suggest that demographic factors (Age) and economic indicators (Sourcefuel, Satisfaction Level) play crucial roles in predicting the target variable. Further model tuning and interpretation may be necessary to assess causality and improve classification performance.

3.3 Result of the SVM Model for estimating poverty Status

The SVM model uses a radial basis function (RBF) kernel, which is effective in capturing non-linear relationships between poverty status (POVMI) and the predictor variables. The cost parameter (C = 1) suggests a balance between margin maximization and misclassification minimization. The large number of support vectors (1239) indicates that the model relies heavily on the training data for classification, which might suggest overfitting if not properly validated. Given the poor ROC curve performance, the model may need *hyperparameter* tuning, such as adjusting cost (C) and gamma (γ), or exploring alternative feature engineering techniques for better generalization.

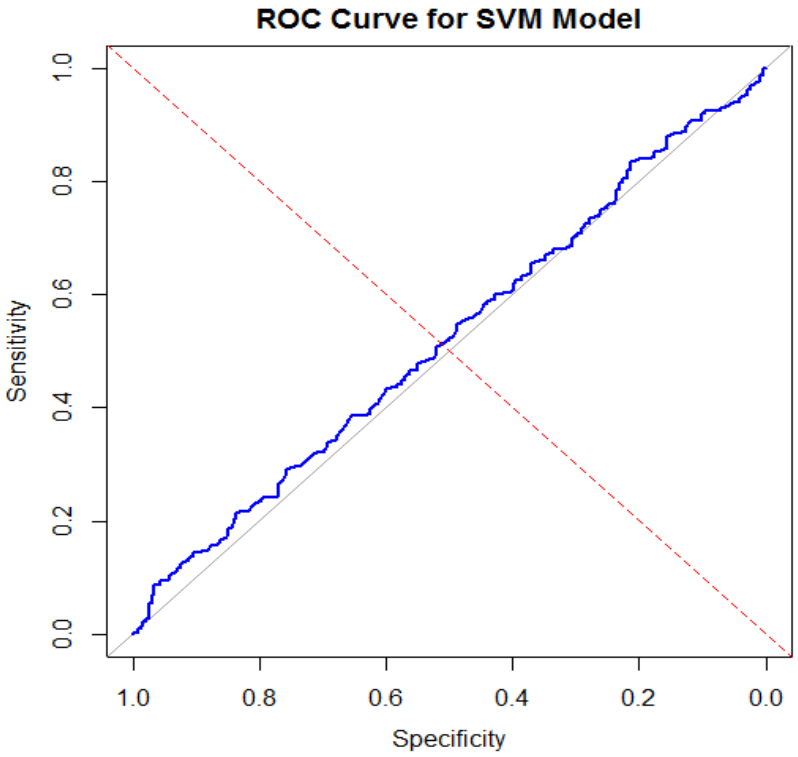


Figure 8: ROC Curve of the SVM Model

The ROC curve for the Support Vector Machine (SVM) model in Figure 8 evaluates its classification performance by plotting sensitivity (true positive rate) against 1-specificity (false positive rate). The blue line represents the model's performance, while the diagonal grey line indicates a random classifier (AUC = 0.5). The ROC curve appears close to this diagonal, suggesting poor classification ability, with the model performing only slightly better than random guessing.

3.4 Result of the RF Model for estimating poverty Status

The Random Forest model trained on the dataset uses 500 trees, with 2 variables randomly selected at each split. The out-of-bag (OOB) error rate is 34.72%, indicating moderate classification performance. The confusion matrix shows that the model struggles with classifying class 0, with a high misclassification rate (91.94%), while class 1 is predicted with much higher accuracy (error rate = 7.96%).

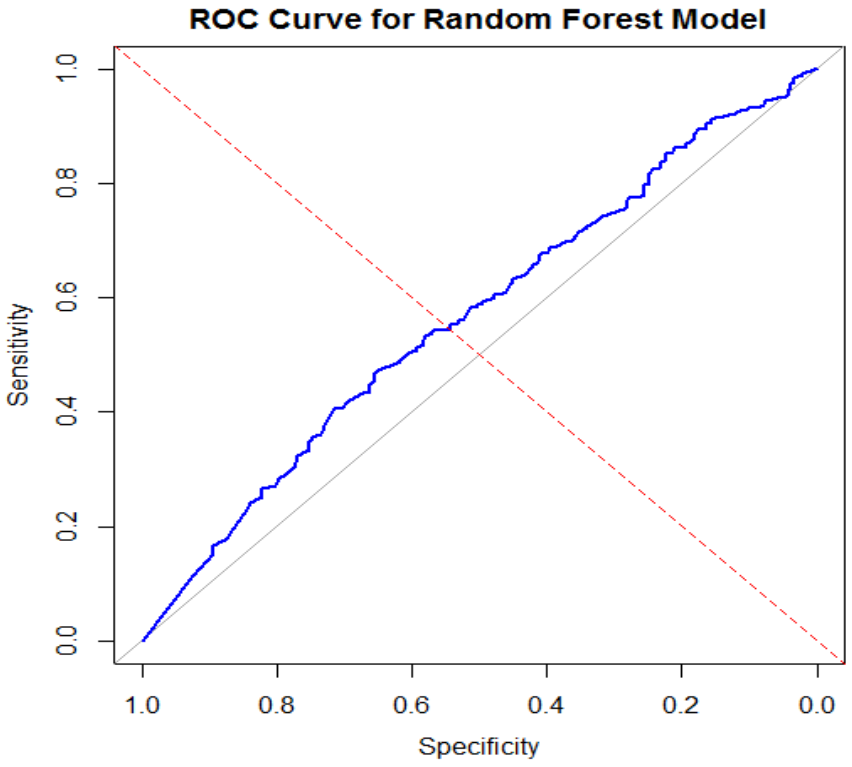


Figure 9: ROC Curve of the RF Model

The ROC curve for the Random Forest model in Figure 4.8 illustrates the model's classification performance by plotting sensitivity (true positive rate) against 1-specificity (false positive rate). The blue curve represents the model's performance, while the grey diagonal line represents a random classifier (AUC = 0.5). The curve appears only slightly above the diagonal, suggesting poor discriminative power. The area under the curve (AUC), though not explicitly stated, is likely close to 0.5, indicating that the model does not perform much better than random guessing. This aligns with the OOB error rate of 34.72% and the confusion matrix, which showed a high misclassification rate for class 0 (91.94%).

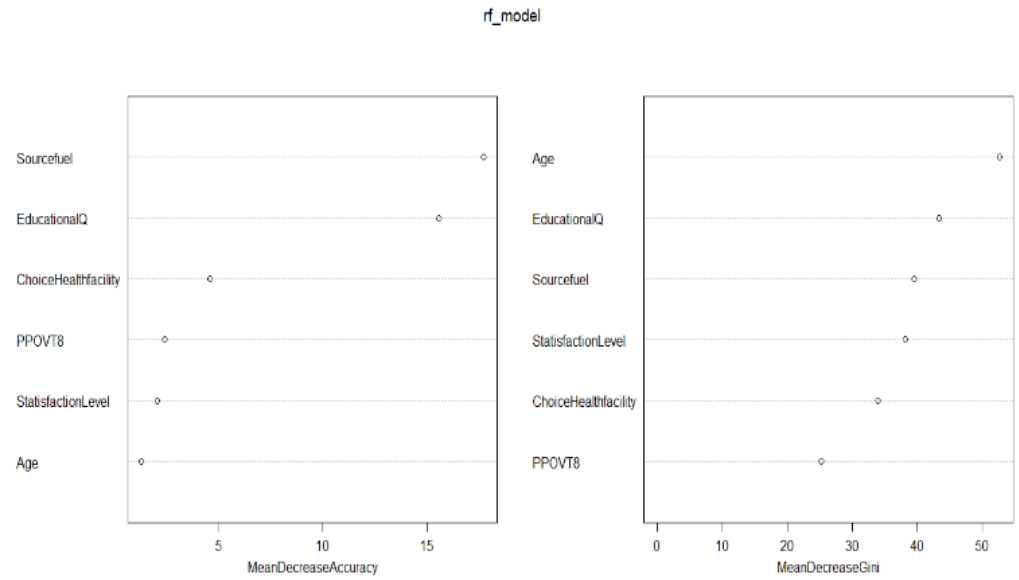


Figure 10: Variable importance plot for the Random Forest model

The variable importance plot for the Random Forest model in Figure 4.9 presents two metrics: Mean Decrease Accuracy (left plot) and Mean Decrease Gini (right plot). The Mean Decrease Accuracy measures the impact of each variable on prediction accuracy, while the Mean Decrease Gini reflects the variable's role in reducing node impurity. Age, EducationalIQ, and SourceFuel are the most influential features, as they exhibit the highest values in both metrics, indicating their strong contribution to model performance. Choice of Health Facility, Satisfaction Level, and PPOVT8 have relatively lower importance, suggesting they contribute less to the model's predictive power. The dominance of Age in both metrics highlights its critical role in determining poverty status (POVMI). At the same time, EducationalIQ's significance suggests that education level plays a key role in economic well-being.

Table 1: Performance Measures of GBM, SVM, and RF Models for Poverty Status Classification

Classifiers	Accuracy	Precision	Recall	F1
GBM	0.6813	0.6813	0.9806	0.8039
SVM	0.3177	0.3177	1.0000	0.4822
RF	0.6688	0.4218	0.1134	0.1788

The performance evaluation of the Gradient Boosting Machine (GBM), Support Vector Machine (SVM), and Random Forest (RF) models for poverty status classification in Table 1 reveals notable differences in predictive effectiveness. GBM outperforms the other classifiers with the highest accuracy (68.13%), precision (68.13%), and F1-score (80.39%), making it the most reliable model. SVM achieves perfect recall (100.00%) but suffers from low accuracy (31.77%) and precision (31.77%), indicating a high false positive rate. RF, while achieving a moderate accuracy (66.88%), struggles with precision (42.18%) and recall (11.34%), resulting in a poor F1-score (17.88%). These results suggest that GBM provides the most balanced and effective classification of poverty status, making it the most suitable model for further analysis and policy recommendations.

Table 2: Performance Comparison of GBM, SVM, and RF Models Based on AUC, MSE, and R-Squared for Poverty Status Classification

Classifiers	AUC	MSE	R-squared
GBM	0.5668	0.2115	0.0256
SVM	0.5225	0.3493	0.0002
RF	0.5668	0.2475	0.0106

The result in Table 2 presents the evaluation metrics of the Gradient Boosting Machine (GBM), Support Vector Machine (SVM), and Random Forest (RF) for poverty status classification. GBM and RF achieve the highest AUC (0.5668), indicating a slightly better discrimination ability than SVM (0.5225), which is close to a random classifier. In terms of Mean Squared Error (MSE), GBM has the lowest value (0.2115), suggesting better prediction accuracy compared to RF (0.2475) and SVM (0.3493). Regarding R-squared, which measures the proportion of variance explained by the model, all classifiers perform poorly, with GBM (0.0256) having the highest value, followed by RF (0.0106) and SVM (0.0002), indicating weak explanatory power. Hence, GBM outperforms the other models in terms of prediction accuracy and variance explanation, while SVM exhibits the weakest performance across all metrics.

3.4 Discussion of Results

The findings reveal a high poverty prevalence in Anambra State, with 68.2% of respondents classified as poor. Age, education level, and fuel source emerged as the most significant predictors across all models, particularly the Gradient Boosting Machine (GBM), which outperformed both Support Vector Machine (SVM) and Random Forest (RF) in accuracy (68.13%), precision, and F1-score. Despite this, all models exhibited low R-squared values, indicating that poverty is shaped by multidimensional and possibly unobserved factors beyond those captured in the dataset. The poor performance of SVM and RF, alongside GBM's moderate predictive strength, highlights the importance of robust feature selection, algorithm tuning, and the incorporation of more diverse indicators.

These findings align with existing literature emphasizing the need for context-specific, multidimensional poverty assessments. Studies support the integration of localized weighting schemes and non-monetary indicators in poverty modelling by (Pasha, 2017; Saifullah, 2022; Vollmer and Alkire, 2022). The relatively low contribution of perception-based variables reinforces concerns raised about the weak correlation between subjective well-being and economic deprivation by (Strotmann and Volkert, 2018). The study underscores the value of machine learning particularly boosting algorithms as a practical tool for micro-targeting poverty interventions, enhancing the effectiveness of policy design through data-driven insights.

4. Conclusion

The findings of this study highlight the challenges and trade-offs in using machine learning models for poverty classification. The Gradient Boosting Machine (GBM) demonstrated the highest accuracy (68.13%) and recall (98.06%), making it the most effective model for identifying poverty status. However, its relatively low precision led to an imbalanced F1-score (80.39%), suggesting that while it identifies most positive cases correctly, it also produces a considerable number of false positives. The Support Vector Machine (SVM) achieved perfect recall (100%), meaning it classified all positive cases correctly, but suffered from poor accuracy (31.77%) and precision (31.77%), leading to a weak F1-score (48.22%). Random Forest (RF), while moderately accurate (66.88%), exhibited the lowest recall (11.34%) and a poor F1-score (17.88%), indicating that it struggled to detect actual poverty cases effectively. These results emphasize the need for careful model selection based on specific policy objectives, whether prioritizing recall, precision, or a balance of both.

Despite the relatively strong performance of GBM, its tendency to overfit beyond iteration 174 suggests the importance of early stopping and parameter tuning to enhance generalization. The reliance of SVM on 1,239 support vectors indicates possible overfitting, which may require adjusting hyperparameters such as the cost parameter (C) and kernel function to improve its generalizability. Similarly, the RF model's high misclassification rate for class 0 (91.94%) and its poor recall suggests the

need for techniques such as resampling, cost-sensitive learning, or feature selection to address class imbalance. The feature importance analysis of RF further underscores the role of demographic and economic factors, particularly Age, EducationalIQ, and SourceFuel, in poverty classification, reinforcing the necessity of incorporating socioeconomic variables in predictive modelling.

The study highlights Age, Educational Qualification Index (EducationalIQ), and Source of Fuel as key predictors of poverty status in Anambra State, emphasizing the need for education-focused policies and affordable energy access. Policymakers are encouraged to implement vocational training, literacy programs, and renewable energy investments to boost economic well-being. While health facility choice and satisfaction levels showed minimal influence, the findings suggest that poverty reduction efforts should focus on income generation and economic empowerment rather than healthcare access alone. Enhanced data collection methods such as geospatial tracking and integration of big data are recommended to capture poverty's multidimensional nature more effectively. Though Gradient Boosting Machine (GBM) proved the most reliable model, further refinement through ensemble methods and improved feature selection could enhance predictive accuracy.

REFERENCES

Abdullah-All-Tanvir, Iftakhar Ali Khandokar, A.K.M. Muzahidul Islam, Salekul Islam, Swakkhar Shatabda, 2023. A gradient boosting classifier for purchase intention prediction of online shoppers. Heliyon, 9(4): Pp. e15163.

Alkire, S., Kanagaratnam, U., Nogales, R., and Suppa, N., 2022. Revising the Global Multidimensional Poverty Index: Empirical Insights and Robustness. Review of Income and Wealth, 68(S2), Pp. S347–S384.

Alqatarneh, G., and Al-Zboon, K. K., 2022. Water Poverty Index: a Tool for Water Resources Management in Jordan. Water, Air, and Soil Pollution, 233(11).

Amfo, B., Osei Mensah, J., and Aidoo, R., 2022. Migrants and non-migrants' welfare on cocoa farms in Ghana: Multidimensional poverty index approach. International Journal of Social Economics, 49(3), Pp. 389–410

Aziz, Y., Mansor, F., Waqar, S., and Haji Abdullah, L., 2020. The nexus between zakat and poverty reduction, is the effective utilization of zakat necessary for achieving SDGs: A multidimensional poverty index approach. Asian Social Work and Policy Review, 14(3), Pp. 235–247.

Chai, T. and Draxler, R.R., 2014. Root Mean Square Error (RMSE) or Mean Absolute Error (MAE)?—Arguments against Avoiding RMSE in the Literature. Geoscientific Model Development, 7, Pp. 1247-1250. http://dx.doi.org/10.5194/gmd-7-1247-2014

Choiriyah, E. A. N., Kafi, A., Hikmah, I. F., and Indrawan, I. W., 2020. Zakat And Poverty Alleviation In Indonesia: A Panel Analysis At Provincial Level. Journal of Islamic Monetary Economics and Finance, 6(4), Pp. 811–832.

Deyshappriya, N. P. R., and Feeny, S., 2021. Weighting the Dimensions of the Multidimensional Poverty Index: Findings from Sri Lanka. Social Indicators Research, 156(1).

Draper, H.R. and Smith, H., 1998. Applied Regression Analysis. 3rd Edition, John Wiley and Sons Inc., New York, 713. http://dx.doi.org/10.1002/9781118625590

Fawcett, T., 2006. An Introduction to ROC Analysis. Pattern Recognition Letters, 27, Pp. 861-874. https://doi.org/10.1016/j.patrec.2005.10.010

Goel, I., Sharma, S., and Kashiramka, S., 2020. The Water Poverty Index: An application in the Indian context. Natural Resources Forum, 44(3), Pp. 195–218.

Henao-Cespedes, V., Garcés-Gómez, Y. A., Ruggeri, S., and Henao-Cespedes, T. M., 2022. Relationship analysis between the spread of COVID-19 and the multidimensional poverty index in the city of Manizales, Colombia. Egyptian Journal of Remote Sensing and Space Science, 25(1), Pp. 197–204.

Ismail, M. K., Siwar, C., and Ghazali, R., 2018. Gahai agropolitan project in eradicating poverty: Multidimensional poverty index. Planning Malaysia, 16(3), Pp. 97–108.

Khan, S. I., Islam, N., Uddin, J., Islam, S., Nasir, M. K., 2022. Water quality prediction and classification based on principal component regression and gradient boosting classifier approach. Journal of King Saud University – Computer and Information Sciences, 34: Pp. 4773–4781.

Koirala, S., Fang, Y., Dahal, N. M., Zhang, C., Pandey, B., and Shrestha, S., 2020. Application of water poverty index (WPI) in spatial analysis of water stress in Koshi River Basin, Nepal. Sustainability (Switzerland), 12(2).

Ladi, T., Mahmoudpour, A., and Sharifi, A., 2021. Assessing impacts of the water poverty index components on the human development index in Iran. Habitat International, 113.

Lavanya, C., Pooja, S., Abhay, H. K., Abdur, R., Swarna, N., and Vidya, N., 2023. Novel Biomarker Prediction for Lung Cancer Using Random Forest Classifiers. Cancer Informatics, 22: Pp. 1–15.

Njuguna, C., and McSharry, P., 2017. Constructing spatiotemporal poverty indices from big data. Journal of Business Research, 70, Pp. 318–327.

Ogwang, T., 2022. The regression approach to the measurement and decomposition of the multidimensional Watts poverty index. Journal of Economic Inequality, 20(4), Pp. 951–973.

Pasha, A., 2017. Regional Perspectives on the Multidimensional Poverty Index. World Development, 94, Pp. 268–285.

Putri, S. R., Wijayanto, A. W., and Sakti, A. D., 2022. Developing Relative Spatial Poverty Index Using Integrated Remote Sensing and Geospatial Big Data Approach: A Case Study of East Java, Indonesia. ISPRS International Journal of Geo-Information, 11(5).

Saifullah, M. K., 2022. Well-being index of Indigenous people of Peninsular Malaysia: an application of the multidimensional poverty index (MPI). AlterNative, 18(3), Pp. 424–435.

Strotmann, H., and Volkert, J., 2018. Multidimensional Poverty Index and Happiness. Journal of Happiness Studies, 19(1), Pp. 167–189.

Suppa, N., 2023. mpitb: A toolbox for multidimensional poverty indices. Stata Journal, 23(3), Pp. 625–657.

Tanvir, A., Khandokar, I. A., Islam, A.K.M, Islam, S., Shatabda, S., 2023. A gradient boosting classifier for purchase intention prediction of online shoppers. Heliyon, 9 (2023) e15163

Tighsazzadeh, M. N., and Malekpourasl, B., 2023. Assessing Multidimensional Poverty Index in Coastal Regions: Implications for the Makran Region of Iran. Review of Regional Studies, 53(1), Pp. 43–58.

Tkach, K., and Gigliarano, C., 2022. Multidimensional Poverty Index with Dependence-Based Weights. Social Indicators Research, 161(2–3), Pp. 843–872.

Vollmer, F., and Alkire, S., 2022. Consolidating and improving the assets indicator in the global Multidimensional Poverty Index. World Development, 158.