Comparison of Machine Learning Algorithms for Household's Economic Status Classification

Neneng Nur Sholihah
Department of Informatics
University of Technology Yogyakarta
Yogyakarta, Indonesia

Arief Hermawan
Department of Informatics
University of Technology Yogyakarta
Yogyakarta, Indonesia

ABSTRACT

This research addresses the global commitment to eradicate poverty as outlined in the United Nations' Sustainable Development Goals (SDGs) for 2015-2030. Poverty is a multifaceted issue encompassing income levels, resource availability, education accessibility, hunger, malnutrition, social injustice, and limited access to basic needs. Traditional poverty assessments relying on surveys present challenges in terms of cost, time, and outdatedness. To overcome these challenges, this study leverages machine learning algorithms to classify household economic status. This research compares Random Forest, SVM, Naïve Bayes, and ANN algorithms. The results show that the Random Forest algorithm consistently outperforms others, achieving the highest AUROC values. The classification evaluation results indicate that Random Forest performs the best classification with 93% accuracy. These findings contribute valuable insights for policymakers and development practitioners, enhancing the precision and efficiency of poverty reduction initiatives to align with the UN's goal of a poverty-free world by 2030.

General Terms

Classification, Machine Learning.

Keywords

Comparison, Classification, Household, Economic Status, Machine Learning Algorithms.

1. INTRODUCTION

The United Nation (UN) formulated 17 Sustainable Development Goals (SDGs) for the period 2015-2030, including a commitment to eradicate all forms of poverty worldwide [1]. The issue of poverty has been a concern for several decades which governments, policymakers, and organizations have been striving to eradicate. Typically, poverty is determined by the level of income and the availability of basic resources sufficient to maintain sustainable livelihoods. Moreover, it also involves inaccessibility to education, hunger, malnutrition, social injustice, and limited access to other basic needs [2]. Despite having a job, it cannot guarantee that someone will have a decent life. To address this issue further, the United Nations declares that poverty reduction should be integrated into national policies and tackled from all dimensions, including political, economic, and social aspects, by promoting a people-centric approach to target the most vulnerable groups [3]. Additionally, the UN emphasizes the importance of global cooperation and partnership to achieve sustainable development and eradicate poverty, recognizing that a collective effort involving governments, international organizations, businesses, and civil society is essential for meaningful and lasting progress.

Efforts to reduce poverty have become a crucial mission for all countries, especially for developing nations. In the 18th World Congress, The International Society of Gynecological Endocrinology (ISGE) classified Indonesia as one of the 145 developing countries [4]. The identification as a developing country is often associated with poverty issues, triggering various social, political, and economic challenges in nations like Indonesia [5][6]. The evaluation of poverty and socio-economic development in region is typically conducted through household surveys and annual statistics. However, this method has proven to be expensive, time-consuming, and outdated [7]. It is crucial for policymakers and researchers to monitor poverty to analyze the conditions of the poor and design effective poverty reduction strategies. Traditional poverty measurements rely on survey data, including income, consumption, health, education, and housing [8][9]. Data collection through surveys tends to be time-consuming and involves significant costs. Moreover, the frequency of surveys, typically conducted every 3-5 years, can leave long gaps in data, especially in extremely poor or conflict-affected countries.

The intersection of global development goals and technological advancements has paved the way for innovative approaches to address socioeconomic issues. One such avenue is the application of machine learning algorithms for the classification of household economic status. The advent of data-driven methodologies provides an opportunity to analyze and predict economic conditions at a granular level. This study aims to contribute to the discourse by comparing various machine learning algorithms for the classification of household economic status. The effectiveness of these algorithms will be evaluated based on factors such as accuracy, precision, and recall, with the ultimate goal of identifying models that can aid in targeted poverty alleviation efforts. By harnessing the power of machine learning, this research seeks to provide insights that can inform policymakers and development practitioners in crafting strategies tailored to the unique challenges faced by developing nations. The findings hold the potential to enhance the precision and efficiency of poverty reduction initiatives, bringing us a step closer to realizing the UN's vision of a world free from poverty by 2030.

2. LITERATURE REVIEW

To expand the analysis and establish a robust comparative framework with relevant journals, a review of scientific publications on poverty will be conducted. The aim is to gain a comprehensive understanding of how machine learning algorithms perform in addressing poverty challenges. This research not only seeks to identify trends but also to thoroughly evaluate the strengths and weaknesses of various machine learning algorithms. This comparison is expected to provide a holistic view of the potential application of these algorithms in tackling complex issues like poverty.

In the research [10], discusses the use of machine learning techniques to classify multidimensional poverty in Jordan. Despite facing challenges such as class imbalance in the dataset 1:6 ratio of poor to non-poor households, the authors successfully addressed this issue using techniques like oversampling, undersampling, SMOTE, and class weights. They also overcame the need for large datasets by combining data from five national surveys. The proposed machine learning approach, particularly Light-GBM and Bagged Decision Trees demonstrated superior performance with an f1-score of approximately 80%. The authors recommend using these algorithms to assess and monitor the poverty status of Jordanian households.

A study compared the performance of Naïve Bayes, Decision Tree, and k-Nearest Neighbors in classifying Malaysia's B40 population, utilizing the 'eKasih' dataset from the National Poverty Data Bank [11]. The research highlighted the significance of data preprocessing, feature selection, engineering, normalization, and sampling methods. To address class imbalance, Synthetic Minority Oversampling Technique (SMOTE) was used. Each classifier underwent tuning with different parameter combinations, including discretization for Naïve Bayes, confidence factor, minimum number of objects for Decision Tree, k-value, and distance function for k-Nearest Neighbors. Feature selection algorithms based on Symmetrical Uncertainty, Correlation, and Information Gain Attributes improved classification accuracy and Kappa statistic by identifying the top eight attributes. The study evaluated classifier performance through 10-Fold Cross-Validation with a Statistical Test to identify significant differences between models. The conclusion emphasized the Decision Tree model's outstanding accuracy compared to other classifiers.

In [12], the researchers conducted an evaluation of poverty determinants using a machine learning approach to predict poverty levels. Their data incorporated information from the Poverty Possibility Index detailing individual data and the Oxford Poverty & Human Development Initiative offering a poverty index across various countries. Through data analysis, the authors sought insights into the relationships among different variables influencing poverty likelihood. Various machine learning models, including linear regression, Decision Tree, Random Forest, Gradient Boosting, and Neural Networks were explored to identify the most suitable model for poverty prediction and classification. The study underscores the key factors impacting poverty based on their significance scores. The findings concluded that the Gradient Boosting Classifier outperformed others in terms of accuracy, reliability, and complexity. Among the factors examined, education level emerged as a crucial determinant of poverty, followed closely by the country of residence.

In [13], a study focused on determining household poverty factors in Pakistan. The researchers utilized logistic regression to investigate both personal characteristics of the household head and overall household traits as determinants. The households were categorized as either poor or non-poor by dividing them into quartiles based on monthly per adult household expenditure, designating the lowest quartile as poor and the other three as non-poor. The significance of coefficients was tested using the Wald test, and results were elucidated through odd ratios. The analysis revealed that an increase in education level correlated with a decreased probability of being poor. Additionally, the presence of remittances played a significant role in reducing the likelihood of poverty. The study also highlighted a higher severity of poverty in rural areas compared to urban areas in Pakistan.

The study [14] investigated factors influencing the state of poverty using Ordinal and Multinomial Logistic Regression models. Poverty was categorized into three states: poverty, near poverty, and above near poverty based on household income percentages relative to the poverty threshold. The threshold, derived from Poland's national median income in 2000, was adjusted annually for inflation. Data spanning 2000 to 2015, collected through biennial household surveys, were analyzed. Two questionnaires in the 2015 Social Diagnosis report gathered information on household composition, living conditions, and individual quality of life. Variables determining the state of poverty included gender, age, education level, residence, household size, family type, socio-economic group, labor-force status, and disability status. The study concluded that the multinomial logit model was more suitable for predicting poverty states due to the ordinal logistic regression model's failure to meet the assumption of parallel lines. Notably, education, residence, labor-force status, and socio-economic group were identified as the most significant factors influencing the state of poverty.

3. METHODOLOGY

3.1 Data Collection

The data used for this research were obtained from the Statistical Service Information System page owned by the Central Statistics Agency of Indonesia. The data utilized are microdata from the National Socio-Economic Survey conducted by Central Statistics Agency for region in the year 2021. In this study, 1059 household data were used, consisting of 667 low-economic status households, 373 medium-status households, and 19 high-status households. The variables in this research consist of 1 response variable (Y) and 36 explanatory variables (X). Detailed information about the classes and variables used in this study can be seen in Table 1.

Table 1. Classes and variables used in the study

Variabel	Description
Class: Economic Status	0: Low 1: Medium 2: High
X1	Urban/Rural
X2	Food Shortage
X3	Homeownership Status
X4	Floor Area of the House (m^2)
X5	Ownership of Another House
X6	Roof Type of the House
X7	Wall Type of the House
X8	Floor Type of the House
X9	Ownership of Toilet Facilities
X10	Toilet Type
X11	Main Source of Drinking Water
X12	Drinking Water Shortage
X13	Source of Bathing/Washing/etc. Water
X14	Source of House Lighting
X15	Cooking Fuel
X16	Ownership of Refrigerator
X17	Ownership of Air Conditioner
X18	Ownership of Water Heater
X19	Ownership of Landline Telephone
X20	Ownership of Computer/Laptop
X21	Ownership of Gold/Jewelry (min 10g)
X22	Ownership of Motorbike
X23	Ownership of Boat

X24	Ownership of Motorized Boat
X25	Ownership of Car
X26	Ownership of Flat-Screen Television (min 30 inches)
X27	Ownership of Land/Property
X28	Main Source of Income
X29	Receives KKS (Prosperous Family Card)
X30	Receives PKH (Family Hope Program)
X31	Receives BPNT (Non-Cash Food Assistance)
X32	Receives social assistance/government subsidy
X33	Number of Household Members
X34	Age of Household Head
X35	Highest Educational Attainment
X36	Employment Status of the Household Head

3.2 Data Preprocessing

In the preprocessing stage, a series of steps are undertaken to prepare the data before proceeding to the classification process. The objective of this pre-processing stage is to ensure the quality and consistency of the data to be utilized in the subsequent classification process [15]. These steps encompass removing duplicates, checking for missing values, employing one-hot encoding, applying label encoding, and addressing data imbalance through the application of the Synthetic Minority Oversampling Technique (SMOTE).

3.3 Feature Selection

Feature selection is a technique utilized to enhance model optimization during the data preprocessing phase. Its operation involves identifying a subset of features from the available set to improve the performance of the implemented model [16]. Additionally, feature selection helps eliminate irrelevant and redundant features in the model, enabling the model to achieve optimal performance. In this study, feature selection will be conducted using Chi-Squared, Correlation, and Information Gain methods. These methods will be employed to identify the most informative attributes from the dataset, contributing to a refined and efficient model for subsequent data analysis.

3.4 SMOTE

SMOTE is a technique employed to balance diverse classes by employing oversampling. The SMOTE method involves duplicating data in the minority class to equalize it with the majority class data [17]. Dataset imbalance can lead to erroneous classification outcomes, where minority class data is frequently misclassified as the majority class [18]. In this study, the low and medium classes significantly outnumber the high class. Hence, SMOTE will be utilized to address this imbalance by generating synthetic data for the high class to attain equilibrium with the other two classes.

In the initial steps of the SMOTE algorithm, computations are conducted by determining the difference between feature vectors in the minority class and the nearest neighbors from the same class. Following this, the calculated difference is multiplied by a randomly generated number ranging from 0 to 1. Subsequently, the result of this computation is added back to the original feature vector, creating a new feature vector [19].

$$X_{new} = X_i + (-X_i) \times \delta \qquad (3.1)$$

Description:
X_i = vector of variables for the minority class
$\hat{X}_i$ = *k-nearest neighbors for* X_i
δ = random value between 0 and 1

3.5 Random Forest

Random Forest is a methodology employing a set of decision trees as its foundational model for classification or regression [20]. This ensemble learning method, comprising decision trees, is designed to provide more precise and stable predictions [21]. In the context of classification with Random Forest, a voting mechanism is utilized to make decisions based on the majority outcomes from the assembled trees [22]. The ensemble learning approach carries various benefits, such as versatility for both classification and regression tasks, the potential to achieve high accuracy, and suitability for handling extensive datasets with multiple dimensions [23]. The Random Forest method introduces randomness in the selection of explanatory variables to mitigate inter-tree correlations. The procedural steps for classification using Random Forest on a training dataset with n observations and p explanatory variables are outlined as follows [24]:

1. Bootstrap Procedure: Drawing a random sample of size n from the training data with replacement.
2. Random Feature Selection Procedure: Building trees without pruning based on bootstrap outcomes until reaching the maximum size. During each splitting process, randomly choosing m explanatory variables, where m < p, and then executing the optimal split.
3. Iterating steps 1-2 k times until acquiring k random trees.
4. Predicting the response of an observation by amalgamating the predictions of k trees. The ultimate prediction is established through a majority vote.

When constructing a tree, each step of separation needs to consider the entropy value to then compute information gain. The variable exhibiting the highest information gain is chosen as the optimal separator or partition. For example, if dataset A is partitioned into various segments (A1, A2, ..., Ak), the formula developed by Claude Shannon in information theory [25] is employed to calculate entropy and information gain.

$$Entropi(A) = -\sum_{i=1}^{k} -p_i \, log_2(p_i) \qquad (3.2)$$

$$IG = Entropi(A) - \sum_{i=1}^{k} \frac{|A_i|}{|A|} \times Entropi(A_i) \qquad (3.3)$$

Description:
A = Data Set
k = Number of partitions in A
p_i = Proportion of A_i relative to A
IG = *Information gain*
$|A_i|$ = Number of observations in partition i
$|A|$ = Total number of observations in A

3.6 Support Vector Machine

Support Vector Machine (SVM) is a classification technique used for analyzing data. The classification process involves both training and testing datasets, where each dataset element contains multiple features and classification attributes. The fundamental principle of SVM is to create a model that predicts classification based on the features of the current test dataset element [26]. The SVM algorithm offers flexibility in choosing various kernels, with linear, polynomial, RBF, and sigmoid kernels generally dominating the options [27]. Linear SVM proves to be effective, especially when dealing with large datasets containing numerous features.

The classification of data is executed by SVM based on the training data provided in the form of features for diverse datasets. The SVM methodology involves finding the optimal hyperplane to separate datasets into classes, emphasizing the importance of maximizing the margin between supporting planes for each class [28]. In the context of kernel functions, the SVM algorithm employs various types, and for non-linear classification tasks, the radial basis function (RBF) is commonly utilized. Kernel functions serve as a technique for approximating multivariable functions by using linear combinations derived from a single univariable function. The mathematical expression for the SVM-linear kernel is presented in equation 3.4, where the coefficient c_i is determined through the resolution of the optimization problem solvable by quadratic programming. The class of the value x_i is detoned as y_i, and $\varphi(x_i)$ represents the transformed data point.

$$f(x) = \sum_{i=1}^{n} c_i y_i \varphi(x_i) \tag{3.4}$$

3.7 Naïve Bayes

The Naïve Bayes algorithm is a straightforward probabilistic method within the classification domain, deriving its probability values through the computation of frequencies and combinations within related sets [29]. Functioning as a probabilistic machine learning model, the Naïve Bayes classifier employs Bayes' theorem for classification purposes. Notably, this algorithm assumes the independence of all attributes [30]. In its operation, Naïve Bayes predicts the probability that a given data sample belongs to a specific class, denoted as the posterior probability P(C|F1...Fn) for a data vector in class C. Thus, equation 3.6 is applied to facilitate this computation.

$$P\,(C|F_1 \dots F_n) = \frac{P\,(C)\,.\;P(F_1 - F_n|C)}{P(F_1 \dots F_n)} \tag{3.5}$$

In equation, the variable C denotes the class, while the variables F1...Fn represent the features essential for conducting the classification. Hence, the probability of a data match with a specific feature in class C (posterior) is obtained by multiplying the probability of class C by the likelihood of the feature in class C. This product is then divided by the overall probability of the feature across all samples (evidence).

3.8 Artificial Neural Network

Artificial Neural Networks (ANN) are a popular machine learning technique inspired by the biological neural network in the human brain [31]. Feedforward neural networks are a common type of ANN that sends the weight values of each artificial neuron as output to the next layer after processing inputs from neurons in the previous layer. One important class of feedforward neural networks is the Multilayer Perceptron (MLP). The backpropagation algorithm is the most widely used training technique for MLP. It adjusts the weights between neurons to minimize errors. This model is good at learning patterns and can easily adapt to new values in the data. However, it may show slow convergence and has the risk of reaching a local optimum [32]. Determining the number of layers, the number of neurons in the hidden layer, and the connections between them is an important problem. These parameters and issues play a crucial role in the performance of artificial neural networks, and the results can vary significantly based on these factors. Different ANN architectures will yield different results for various problems. Nevertheless, it's essential to arrive at an optimal ANN architecture through trial and error.

4. RESULTS AND DISCUSSION

4.1 Dataset

The data utilized in this research comprises 1059 households and 37 variables. After further examination, the data obtained from the Central Statistics Bureau does not contain duplicate data and missing values. Therefore, the process continues with feature selection, one-hot encoding, and label encoding. The distribution of households is divided into 667 for the low-class, 373 for the middle-class, and 19 for the high-class. As shown in Figure 2, there is an imbalance in the data where the number of households with high-class economic status is fewer compared to the other two classes.

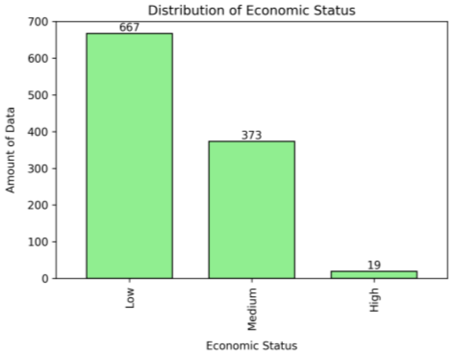

Figure 1. Distribution of Economic Status Classes

This dataset undergoes data splitting with a percentage of 70% for training data and 30% for testing data. The phenomenon of imbalance in class distribution can influence the model's ability to recognize and predict data in the minority class. This can result in classification outcomes biased toward the majority class, causing the minority class to be overlooked or have lower accuracy. Therefore, this study employs the handling of imbalanced data issues using the SMOTE method. Table 2 illustrates the data distribution after addressing the imbalance issue.

Table 2. Data Composition before and after using SMOTE

Class	Training Dataset		Testing Dataset	
	Before SMOTE	After SMOTE	Before SMOTE	After SMOTE
Low	463	463	204	204
Medium	263	463	110	204
High	15	463	4	204
Total	741	1389	318	612

4.2 Feature Selection

Feature selection is a critical step in machine learning, aiming to improve model performance by choosing the most relevant features. An experiment on feature selection algorithms in the dataset has been conducted using the Chi-Squared Attribute, Correlation Attribute, and Information Gain Attribute. The top ten attributes for each of the methods are presented in Table 3.

Table 3. Top Ten Attributes for Feature Selection

Feature Selection	Score	Top 10 Rank Features
Chi-Squared Attribute	0.0243767	Floor Area of the House (m^2)
	0.0039458	Floor Type of the House
	0.0015635	Main Source of Drinking Water

	0.0015589	Ownership of Refrigerator
	0.0012869	Ownership of Motorbike
	0.0012186	Ownership of Flat-Screen Television (min 30 inches)
	0.0009386	Ownership of Land/Property
	0.0006928	Receives KKS (Prosperous Family Card)
	0.0006847	Receives PKH (Family Hope Program)
	0.0006446	Receives BPNT (Non-Cash Food Assistance)
Correlation Attribute	0.0230088	Ownership of Air Conditioner
	0.0200383	Ownership of Car
	0.0172138	Ownership of Refrigerator
	0.0161276	Ownership of Computer/Laptop
	0.0142239	Ownership of Flat-Screen Television (min 30 inches)
	0.0089923	Ownership of Gold/Jewelry (min 10g)
	0.0076043	Highest Educational Attainment
	0.0058978	Ownership of Motorbike
	0.0058886	Ownership of Land/Property
	0.0058403	Floor Type of the House
Information Gain Attribute	0.0812799	Floor Area of the House (m2)
	0.0541848	Floor Type of the House
	0.0519056	Main Source of Drinking Water
	0.0426556	Ownership of Refrigerator
	0.0362264	Ownership of Air Conditioner
	0.0283553	Ownership of Computer/Laptop
	0.0245422	Ownership of Motorbike
	0.0225469	Ownership of Car
	0.0186071	Ownership of Flat-Screen Television (min 30 inches)
	0.0080383	Highest Educational Attainment

Variation in attribute priorities exists between Chi-Squared, Correlation, and Information Gain with ownership of refrigerator being crucial in Chi-Squared and Information Gain while ownership of air conditioner takes precedence in Correlation. Consistency is observed as certain attributes including ownership of refrigerator, motorbike, and flat-screen television rank high across multiple methods. Information Gain provides unique insights, emphasizing attributes overlooked by Chi-Squared and Correlation suggesting sensitivity to distinctive information or label distribution. Socio-economic factors represented by attributes like receives KKS, receives PKH, and receives BPNT significantly impact Chi-Squared rankings.

4.3 Model Performance

The selection of the best algorithm for each model is determined through the evaluation of AUROC values, considering its strong ability to distinguish between positive and negative classes. The results of the AUROC calculations are presented in Table 4.

Table 4. AUROC Score Comparison

Algorithms	AUROC Score		
	Model 1	**Model 2**	**Model 3**
Random Forest	0.96248	0.98917	0.99248
SVM	0.87956	0.88243	0.88821
Naïve Bayes	0.80891	0.84241	0.87043
ANN	0.97819	0.98981	0.99541

Model 1 uses chi-squared for feature selection, while Model 2 applies correlation-based feature selection. On the other hand, Model 3 employs Information Gain for feature selection. Model 3 across all algorithms (Random Forest, SVM, Naïve Bayes, and ANN) has the highest AUROC values. This indicates that Model 3 has the best ability to distinguish between positive and negative classes. Therefore, Model 3 obtained through the application of feature selection techniques using Information Gain will be utilized to assess the machine learning algorithm's capability to classify economic status.

Based on the classification evaluation presented in Table 5, several insights can be drawn regarding the classification model performance across the three economic classes. For the Low class, Random Forest stands out with a precision of 0.97 and a recall of 0.90, indicating a strong ability to correctly identify positive cases. SVM achieves a perfect recall (1.0) but with slightly lower precision (0.90). Naïve Bayes shows a high recall (0.99) but a lower precision (0.67). ANN demonstrates a well-balanced performance with precision and recall values of 0.95 and 1.0, respectively. For the Medium class, Random Forest and ANN excel with high precision, recall, and F1-Score values (approximately 0.94 to 0.96). However, SVM exhibits lower precision (0.86), while Naïve Bayes has a lower recall (0.74). For the High class, Random Forest, SVM, and ANN exhibit balanced performance with good precision, recall, and F1-Score values. Nevertheless, Naïve Bayes shows lower precision and recall (0.56) in this class. Overall, Random Forest achieves the highest accuracy at 0.93, followed by Naïve Bayes with an accuracy of 0.93, ANN with an accuracy of 0.81, and SVM with an accuracy of 0.79. This analysis provides a comprehensive overview of the strengths and weaknesses of each model in classifying economic status. The confusion matrix for Random Forest, SVM, Naïve Bayes, and ANN are presented in Figures 2, 3, 4, 5, respectively.

Table 5. Comparison of Algorithms Performance

Class	**Random Forest**			**SVM**			**Naïve Bayes**			**ANN**		
	Precision	**Recall**	**F1-Score**	**Precision**	**Recall**	**F1-Score**	**Precision**	**Recall**	**F1-Score**	**Precision**	**Recall**	**F1-Score**
Low	0.97	0.90	0.94	0.90	1.0	0.95	0.67	0.99	0.80	0.95	1.0	0.97
Medium	0.96	0.94	0.95	0.86	0.64	0.74	0.90	0.96	0.93	0.94	0.89	0.92
High	0.86	0.93	0.89	0.69	0.78	0.73	0.89	0.41	0.56	0.89	0.90	0.89
Accuracy	0.93			0.81			0.79			0.93		

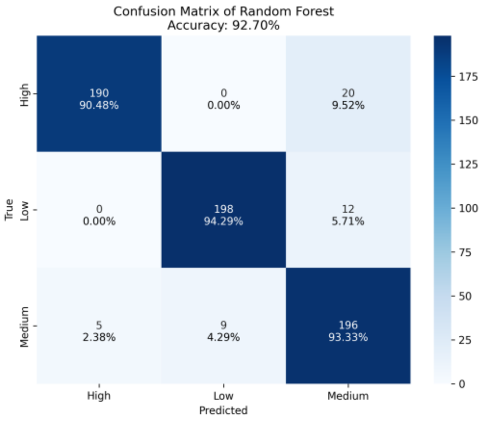

Figure 2. Confusion Matrix of Random Forest

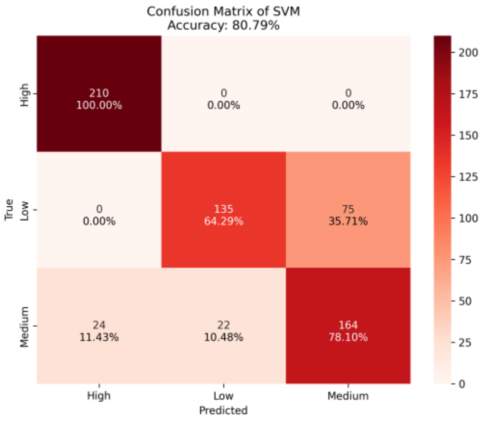

Figure 3. Confusion Matrix of SVM

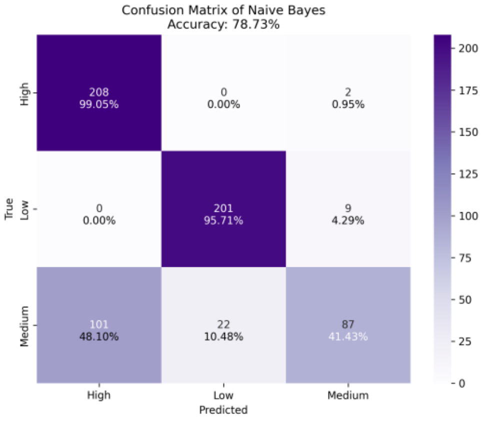

Figure 4. Confusion Matrix of Naïve Bayes

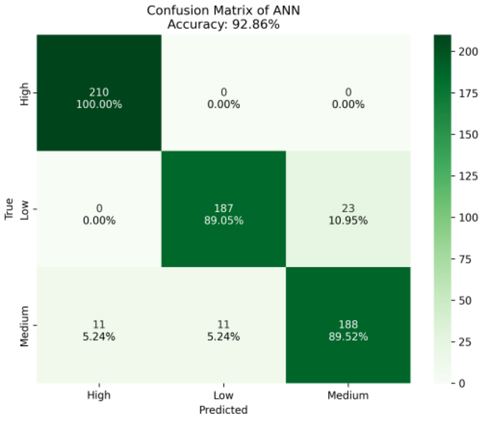

Figure 5. Confusion Matrix of ANN

5. LIMITATIONS AND FUTURE WORK

The main limitation in this study lies in the limited amount of data samples available for nationwide classification. If the utilized dataset inadequately reflects various economic conditions and regional variations, the developed model may face challenges in identifying patterns that are generally applicable at the national level. To address this limitation, further efforts are required in the data collection process for poverty alleviation research. One approach involves gathering additional data that encompasses various aspects of economic conditions on a national scale. Collaboration with governmental agencies, research institutions, or other external data sources can be an effective step in acquiring a more comprehensive dataset that better represents the diversity of economic situations.

Furthermore, in the context of this study, future research considers the incorporation of regional map data. Integrating spatial data into the analysis can provide a richer understanding of economic variations at the national level. Regional maps offer deeper visual insights into the distribution and patterns of the economy across the entire country. Spatial analysis emerges as a robust approach to investigate the interconnection between economic factors and geography. Through this analysis, research can explore how economic conditions vary across different regions, providing a deeper understanding of local factors influencing household economic status.

In this regard, future research also contemplates the utilization of regional map data. The integration of spatial data into the analysis can offer a more comprehensive understanding of economic variations at the national level. Regional maps provide deeper visual insights into the distribution and patterns of the economy across the entire country. Spatial analysis emerges as a robust approach to investigate the interconnection between economic factors and geography. Through this analysis, research can explore how economic conditions vary across different regions, providing a deeper understanding of local factors influencing household economic status. Integrating regional map data can be a significant step in detailing and visualizing the complex relationship between geography and economic status on a national scale.

6. CONCLUSION

In conclusion, model 3 utilizing Information Gain for feature selection consistently outperformed models 1 and 2 across all evaluated algorithms (Random Forest, SVM, Naïve Bayes, and ANN) with the highest AUROC values. This highlights model 3's superior ability to distinguish between positive and negative classes, prompting its selection for further evaluation. The classification evaluation results indicate that Random Forest performs the best with a high balance of precision and recall for the Low class, overall high performance for the Medium class, and a good balance between precision and recall for the High class. The best classification model obtained in this study has achieved accuracy in line with the acceptable standards in scientific literature. In this context, all changes such as political, economic, and social influences on the pattern of poverty, whether directly or indirectly, can be addressed because the data used in this study is derived from field surveys conducted over several different years that reflect these changes. Furthermore, since the proposed model is based on this data, it can be concluded that the model is robust enough to cope with any changes that may occur in the near or distant future.

7. REFERENCES

[1] United Nations, “Sustainable Development Goals.” Accessed: Nov. 13, 2023. [Online]. Available: https://www.un.org/sustainabledevelopment/

[2] P. P. Min, Y. W. Gan, S. N. B. Hamzah, T. S. Ong, and M. S. Sayeed, “Poverty prediction using machine learning approach,” Journal of Southwest Jiaotong University, vol. 57, no. 1, 2022.

[3] United Nations, “Poverty – Social Policy and Development Division | DISD.” Accessed: Nov. 14, 2023. [Online]. Available: https://www.un.org/development/desa/dspd/poverty-social-policy-and-development-division.html

[4] The International Society of Gynecological Endocrinology (ISGE), “List of developing countries | ISGE 2018.” Accessed: Jul. 08, 2023. [Online]. Available: https://isge2018.isgesociety.com/registration/list-of-developing-countries/

[5] D. Seftiana, O. D. Arleina, S. Dewi, R. Amalia, and F. Fachrunisah, “Klasifikasi Rumah Tangga Miskin di Kabupaten Jombang dengan Pendekatan Random Forest Cart,” in Pekan Ilmiah Mahasiswa Nasional Program Kreativitas Mahasiswa-Penelitian 2014, Indonesian Ministry of Research, Technology and Higher Education, 2014.

[6] V. Houlden, G.-M. Tsarouchi, and N. Walmsley, “The Impact of Climate Change on the Achievement of the Post 2015 Sustainable Development Goals,” in Climate and Development Knowledge Network, South Africa: Climate and Development Knowledge Network (CDKN), 2015.

[7] S. Hu, Y. Ge, M. Liu, Z. Ren, and X. Zhang, “Village-level poverty identification using machine learning, high-resolution images, and geospatial data,” International Journal of Applied Earth Observation and Geoinformation, vol. 107, p. 102694, 2022, doi: https://doi.org/10.1016/j.jag.2022.102694.

[8] A. Coudouel, J. S. Hentschel, and Q. T. Wodon, “Poverty measurement and analysis,” A Sourcebook for poverty reduction strategies, vol. 1, pp. 27–74, 2002.

[9] S. Carvalho and H. White, Combining the quantitative and qualitative approaches to poverty measurement and analysis: the practice and the potential, vol. 23. World Bank Publications, 1997.

[10] A. Alsharkawi, M. Al-Fetyani, M. Dawas, H. Saadeh, and M. Alyaman, “Poverty classification using machine learning: The case of Jordan,” Sustainability (Switzerland), vol. 13, no. 3, pp. 1–16, Feb. 2021, doi: 10.3390/su13031412.

[11] N. S. Sani, M. A. Rahman, A. A. Bakar, S. Sahran, and H. M. Sarim, “Machine learning approach for Bottom 40 Percent Households (B40) poverty classification,” Int J Adv Sci Eng Inf Technol, vol. 8, no. 4–2, pp. 1698–1705, 2018, doi: 10.18517/ijaseit.8.4-2.6829.

[12] H. Zixi, “Poverty Prediction Through Machine Learning,” in 2021 2nd International Conference on E-Commerce and Internet Technology (ECIT), 2021, pp. 314–324. doi: 10.1109/ECIT52743.2021.00073.

[13] M. T. Majeed and M. N. Malik, “Determinants of Household Poverty: Empirical Evidence from Pakistan,” The Pakistan Development Review, vol. 54, no. 4, pp. 701–717, 2015, [Online]. Available: http://www.jstor.org/stable/43831356

[14] A. Sączewska-Piotrowska, “Determinants of the state of poverty using logistic regression,” Śląski Przegląd Statystyczny, vol. 16, pp. 55–68, Jan. 2018, doi: 10.15611/sps.2018.16.04.

[15] L. B. Adzy, A. Asriyanik, and A. Pambudi, “Algoritma Naïve Bayes untuk Klasifikasi Kelayakan Penerima Bantuan Iuran Jaminan Kesehatan Pemerintah Daerah Kabupaten Sukabumi,” Jurnal Mnemonic, vol. 6, no. 1, pp. 1–10, May 2023, doi: https://doi.org/10.36040/mnemonic.v6i1.5714.

[16] O. Somantri, W. E. Nugroho, and A. R. Supriyono, “Penerapan Feature Selection pada Algoritma Decision Tree untuk Menentukan Pola Rekomendasi Dini Konseling,” Jurnal Sistem Komputer dan Informatika (JSON), vol. 4, no. 2, pp. 272–279, Dec. 2022, doi: http://dx.doi.org/10.30865/json.v4i2.5267.

[17] M. Mustaqim, B. Warsito, and B. Surarso, “Combination of synthetic minority oversampling technique (Smote) and backpropagation neural network to handle imbalanced class in predicting the use of contraceptive implants,” Register: Jurnal Ilmiah Teknologi Sistem Informasi, vol. 5, no. 2, pp. 116–127, Jul. 2019, doi: 10.26594/register.v5i2.1705.

[18] L. Qadrini, H. Hikmah, and M. Megasari, “Oversampling, Undersampling, SMOTE SVM dan Random Forest pada Klasifikasi Penerima Bidikmisi Sejawa Timur Tahun 2017,” Journal of Computer System and Informatics (JoSYC), vol. 3, no. 4, pp. 386–391, Sep. 2022, doi: https://doi.org/10.47065/josyc.v3i4.2154.

[19] M. L. Suliztia, “Penerapan Analisis Random Forest pada Prototype Sistem Prediksi Harga Kamera Bekas Menggunakan Flask,” Skripsi, Universitas Islam Indonesia, Yogyakarta, 2020.

[20] A. Primajaya and B. N. Sari, “Random Forest Algorithm for Prediction of Precipitation,” Indonesian Journal of Artificial Intelligence and Data Mining, vol. 1, no. 1, pp. 27–31, 2018, doi: http://dx.doi.org/10.24014/ijaidm.v1i1.4903.

[21] D. Ismafillah, T. Rohana, and Y. Cahyana, “Analisis Algoritma Pohon Keputusan untuk Memprediksi Penyakit Diabetes Menggunakan Oversampling SMOTE,” INFOTECH: Jurnal Informatika & Teknologi, vol. 4, no. 1, pp. 27–36, Jun. 2023, doi: https://doi.org/10.37373/infotech.v4i1.452.

[22] M. W. B. Santoso, R. C. Wihandika, and M. A. Rahman, “Ekstraksi Ciri untuk Klasifikasi Jenis Kelamin berbasis Citra Wajah menggunakan Metode Compass Local Binary Patterns,” Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer, vol. 3, no. 11, pp. 10556–10563, Jan. 2019.

[23] F. Hamami and A. Dahlan, “KLASIFIKASI CUACA PROVINSI DKI JAKARTA MENGGUNAKAN ALGORITMA RANDOM FOREST DENGAN TEKNIK OVERSAMPLING,” 2022.

[24] J. Widiastuti, “Klasifikasi Pembiayaan Warung Mikro Menggunakan Metode Random Forest dengan Teknik Sampling Kelas Imbalanced (Studi Kasus: Data Nasabah Pembiayaan Warung Mikro Bank Syariah Mandiri KC Jambi),” Tugas Akhir, Universitas Islam Indonesia, Yogyakarta, 2018.

[25] R. G. Gallager, “Claude E. Shannon: a retrospective on his life, work, and impact,” IEEE Trans Inf Theory, vol. 47, no. 7, pp. 2681–2695, 2001, doi: 10.1109/18.959253.

[26] M. Awad, R. Khanna, M. Awad, and R. Khanna, “Support vector machines for classification,” Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers, pp. 39–66, 2015.

[27] M. Achirul Nanda, K. Boro Seminar, D. Nandika, and A. Maddu, “A comparison study of kernel functions in the support vector machine and its application for termite detection,” Information, vol. 9, no. 1, p. 5, 2018.

[28] M. J. Al_Dujaili, H. T. H. Salim ALRikabi, and I. R. Niama ALRubeei, “Gender Recognition of Human from Face Images Using Multi-Class Support Vector Machine (SVM) Classifiers.,” International Journal of Interactive Mobile Technologies, vol. 17, no. 8, 2023.

[29] S. Chen, G. I. Webb, L. Liu, and X. Ma, “A novel selective naïve Bayes algorithm,” Knowl Based Syst, vol. 192, p. 105361, 2020.

[30] D. Berrar, “Bayes’ theorem and naive Bayes classifier,” Encyclopedia of bioinformatics and computational biology: ABC of bioinformatics, vol. 403, p. 412, 2018.

[31] J. J. Hopfield, “Artificial neural networks,” IEEE Circuits and Devices Magazine, vol. 4, no. 5, pp. 3–10, 1988, doi: 10.1109/101.8118.

[32] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, “Learning representations by back-propagating errors,” Nature, vol. 323, no. 6088, pp. 533–536, 1986, doi: 10.1038/323533a0.