DOI: https://doi.org/10.34123/jurnalasks.v18i1.870

POLITEKNIK STATISTIKA STIS

Available online at https://jurnal.stis.ac.id/

Jurnal Aplikasi Statistika & Komputasi Statistik
Vol. 18 No. 1, June 2026
e- ISSN: 2615-1367, p-ISSN:2086-4132

JASKS Jurnal Aplikasi Statistika & Komputasi Statistik

Classification of Village Development Status in Bekasi Regency Using Ensemble Learning and SMOTE-Based Class Balancing

Mochamad Ridwan[1*], Erwin Tanur[2]

[1]BPS-Statistics Bekasi Regency, Bekasi, Indonesia, [2]Training and Education Center, Statistics Indonesia, Jakarta, Indonesia
**Corresponding Author: moch.ridwan@bps.go.id*

ARTICLE INFO

Article history:
Received 25 Nov, 2025
Revised 18 June, 2026
Accepted 26 June, 2026
Published 30 June, 2026

Keywords:
Random Forest, SMOTE, Village Potential Statistics (PODES), Village Development Index (IDM), Classification, Machine Learning, Ensemble Learning, Bekasi Regency.

Abstract

Introduction/Main Objectives: This study aims to classify village development status in Bekasi Regency using machine learning based on the 2024 Village Potential Statistics (PODES) and the Village Development Index (IDM). **Background Problems:** Conventional descriptive assessments ignore complex socio-economic relationships, and class imbalance further reduces model predictive performance. **Novelty:** This study integrates PODES data, ensemble learning, and SMOTE to improve classification, providing a reliable, data-driven framework for village profiling and planning. **Research Methods:** Following preprocessing and a 70:30 split, SMOTE was applied to the training data, and four tree-based models (Decision Tree, Bagging, Random Forest, XGBoost) were evaluated using standard classification metrics. **Finding/Results:** The Random Forest model combined with SMOTE achieved the best classification performance, with an accuracy of 0.7778 and consistently high AUC values across all classes. The most influential predictors were the dominant economic sector, number of farmer groups, availability of basic health services, and presence of micro-business units. These findings demonstrate that combining ensemble learning with SMOTE improves village development classification and provides valuable support for evidence-based rural development planning in Bekasi Regency.

1. Introduction

Villages represent the smallest administrative entities that maintain direct interaction with the community. This position makes the village a strategic foundation for delivering public services and facilitating the fulfillment of basic rights at the local level. As both a community and a government institution, the village plays a vital role in the administrative structure of the Republic of Indonesia [1]. Historically, village communities existed long before the establishment of the modern Indonesian state, and the formation of Indonesia itself emerged from these early rural communities that served as the original basis of governance and social organization [2].

To assess the progress of village development, several measurement tools have been introduced, including the Village Development Index (IDM). IDM categorizes villages into five levels: very underdeveloped, underdeveloped, developing, advanced, and independent [3]. This classification is based on three key dimensions (social resilience, economic resilience, and environmental resilience) [4].

Over time, village development in Indonesia has shown a shift from dependency on central government assistance toward increased capacity for self-governance and local resource management. This transformation makes the study of village independence particularly relevant, especially in regions such as Bekasi Regency. As one of Indonesia's major industrial regions and part of the rapidly expanding Jakarta metropolitan area, Bekasi Regency exhibits substantial variation in village development conditions. Based on the 2024 Village Development Index (IDM), the regency consists of 38 developing villages, 59 advanced villages, and 82 independent villages. This heterogeneity reflects the coexistence of industrial, peri-urban, and agricultural village characteristics, making Bekasi Regency an appropriate empirical setting for examining village development status and evaluating the effectiveness of machine learning approaches for village classification.

The classification of village status serves not only as a measure of development achievement but also as a foundation for evidence-based policymaking. Through this classification, local governments can assess the development level of each village, identify gaps, and set priorities for targeted interventions [5]. For the central government, these results are essential for evaluating the effectiveness of rural development policies, including the allocation of Village Funds. At the same time, local governments benefit from more strategic planning, better budgeting decisions, and the identification of model villages, while other stakeholders, such as NGOs, academic institutions, and private entities, gain reliable information for contributing to rural development initiatives [6]. Therefore, the accuracy of the classification method is critical, as it directly influences the quality of policy recommendations produced.

The primary aim of classifying village development status is to provide a comprehensive overview of village conditions based on social, economic, institutional, and environmental indicators. Beyond current mapping, the classification can also be used to project development trends and support monitoring and evaluation of rural development programs implemented by the government. In addition, classification results can assist policymakers in identifying development disparities among villages and prioritizing interventions based on local needs and characteristics.

In quantitative analysis, decision tree algorithms are among the most widely used classification methods [7]. These models divide the dataset into nodes based on the most informative attributes, producing classification rules that are easy to interpret [8]. Common algorithms include ID3, C4.5, and CART (Classification and Regression Tree).

Tree-based models are highly valued because they can handle both numerical and categorical variables, produce intuitive interpretations, and allow easy visualization of the decision-making process [9]. This makes them well suited for analyzing social, economic, and health data, including studies related to village development, as the resulting rules can be easily understood by policymakers at both district and village levels.

However, decision trees also present limitations. They are prone to overfitting, especially when the number of features is large or the data structure is complex. Additionally, decision trees are highly sensitive to variations in the training data, meaning that small changes can result in significantly different tree structures [10]. These limitations reduce the model’s stability and predictive performance when applied to new data.

To address these challenges, ensemble learning methods have been developed. Ensemble learning combines multiple models to produce predictions that are more accurate, stable, and robust than those of a single classifier [11]. The most widely used examples include Random Forest, which combines many decision trees using bagging, and Gradient Boosting, which improves model performance through iterative learning.

In this research, ensemble learning is expected to enhance the accuracy of classifying village development status in Bekasi Regency. With more reliable classification results, the findings can serve as a stronger foundation for designing targeted and data-driven rural development policies. Furthermore, ensemble methods are expected to improve model stability and predictive performance by reducing the limitations commonly associated with single decision-tree classifiers.

The core problem addressed in this study is how to produce an accurate and reliable classification of village development status using multidimensional data that represent social, economic, institutional, and environmental aspects. Although instruments such as IDM are widely available, traditional analytical methods often struggle to capture the complexity of relationships among variables. Moreover, commonly used tree-based models face risks of overfitting and instability, thereby reducing accuracy when applied to new data. This calls for more adaptive and robust approaches, such as ensemble learning, which is expected to provide better classification performance.

Based on this problem, the study aims to identify factors that influence village development levels in Bekasi Regency, develop classification models using tree-based approaches, and compare their performance with ensemble learning methods. The analysis is limited to villages in Bekasi Regency using 2024 PODES data, with predictor variables representing key social, economic, and institutional indicators. The study evaluates model performance based on common classification metrics such as accuracy, precision, recall, and F1-score.

The novelty of this study lies in the application of ensemble learning for classifying village development levels, an approach that remains relatively limited in studies related to village development. While previous studies have primarily focused on descriptive analyses of village development indicators [12], [13], machine learning research has often relied on single-model classifiers such as Decision Tree algorithms [7], [8], [10]. By combining multiple tree-based models within an ensemble learning framework, this research aims to produce more accurate, stable, and reliable predictive models.

2. Material and Methods

2.1. Village Development Index (IDM)

According to Law No. 6 of 2014, a village is a legal community unit with defined territorial boundaries and the authority to manage governmental affairs and local interests based on ancestral rights and customary practices. Both administrative villages and traditional villages possess the autonomy to regulate and administer their development using local resources and values that exist within the community [13].

To provide a more objective measure of village development performance, the Ministry of Villages, Development of Disadvantaged Regions, and Transmigration introduced the Village Development Index (IDM) in 2015. IDM evaluates villages across three core dimensions (social resilience, economic resilience, and ecological or environmental resilience) which together reflect the village's capacity to manage local potential while responding to development challenges [4].

The indicators that form IDM are constructed based on the principles of sustainable development, in which social, economic, and environmental aspects operate in an integrated manner to ensure long-term continuity. Through this framework, village development is expected not only to support equality and social justice but also to strengthen local values, cultural identity, and environmental stewardship through responsible management of natural resources [14].

Furthermore, IDM also captures progress in village independence following the implementation of the Village Law, supported by Village Funds and the presence of village facilitators. By incorporating local characteristics, village typologies, and community social capital, IDM enables government interventions to become more targeted and impactful. Thus, IDM functions not only as a statistical instrument but also as a strategic planning tool for improving village development effectiveness through collaboration between government institutions and local communities [14].

2.2. Decision tree

Decision trees are a learning method that represent a mapping function from input x to output y in the form of a tree structure. The model operates using a recursive partitioning process, which divides the dataset into smaller subsets based on attribute tests (splits) that provide the highest information gain. This process continues until terminal nodes (leaves) are formed, each representing a final decision or prediction. Decision trees are widely used for both classification (categorical outcomes) and regression (numerical outcomes), with their main advantage being high interpretability, as predictions are generated through a sequence of simple and transparent rules [15].

During tree construction, the selection of the best attribute for splitting is typically based on the information gain metric, which measures the reduction in impurity after the split is performed. The most common impurity measure is entropy, calculated as:

$$Entropy(S) = -\sum_{i=1}^{c} p_i \, log_2(p_i) \tag{1}$$

where S is the dataset, c is the number of classes, and p_i represents the proportion of instances belonging to class i. Once the entropy values are determined, the information gain for each candidate attribute can be computed as follows:

$$Gain(S,A) = Entropy(S) - \sum_{v \epsilon Values(A)} \frac{|S_v|}{[S]} Entropy(S_v) \quad (2)$$

where A is the candidate attribute, $Values\ (A)$ is the set of possible values for attribute A, and Sv is the subset of data in S for which attribute A takes the value v. The attribute with the highest information gain is selected as the splitting node in the decision tree, as it is considered the most effective in separating the data based on class distinction.

2.3. Ensemble learning

Ensemble learning is a machine learning approach that combines multiple models (commonly referred to as base learners or weak learners) to produce predictive performance that surpasses that of any single model on its own. The core idea is that by integrating models with different strengths and weaknesses, the system can form a composite model that is more accurate, robust, and stable overall.

2.3.1. Bagging (Bootstrap Aggregating)

Bagging predictors, or Bootstrap Aggregating, is an ensemble technique introduced by [16] to improve the predictive accuracy of machine learning models. The method generates multiple replicated training datasets using bootstrap sampling, in which data points are drawn randomly with replacement. A separate model is then trained on each replicated dataset, and the final prediction is obtained through aggregation, typically by averaging for regression tasks or majority voting for classification. This approach has proven effective in reducing variance, especially for algorithms that tend to be sensitive to small changes in the input data [16].

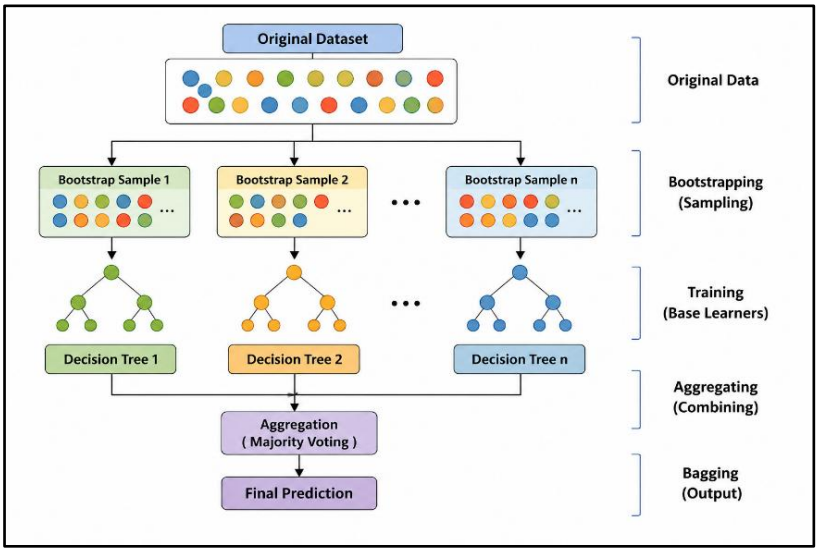

Figure 1. Illustration of the Bagging Process

According to [16], the strength of Bagging lies in its ability to exploit the natural variability in the dataset to enhance model generalization. By training multiple independent models on slightly different versions of the training data and combining their predictions, Bagging can transform an unstable (or weak) predictor into one that is significantly more accurate and robust. This aggregation process minimizes the effect of noise present in any single training sample, making the technique particularly well suited for methods such as decision trees, which are known for their high variance. The illustration of bagging can be seen in Figure 1.

2.3.2. Random Forest

Random Forest is an ensemble-based machine learning algorithm that combines predictions from a large number of decision trees to produce more accurate and stable outputs. Unlike a single decision tree, this algorithm employs the bootstrap aggregating (bagging) technique, in which each tree is trained

on a randomly selected subset of samples and features. This mechanism is specifically designed to reduce correlation among the individual models, thereby effectively minimizing the risk of overfitting and improving predictive validity when applied to new data [17].

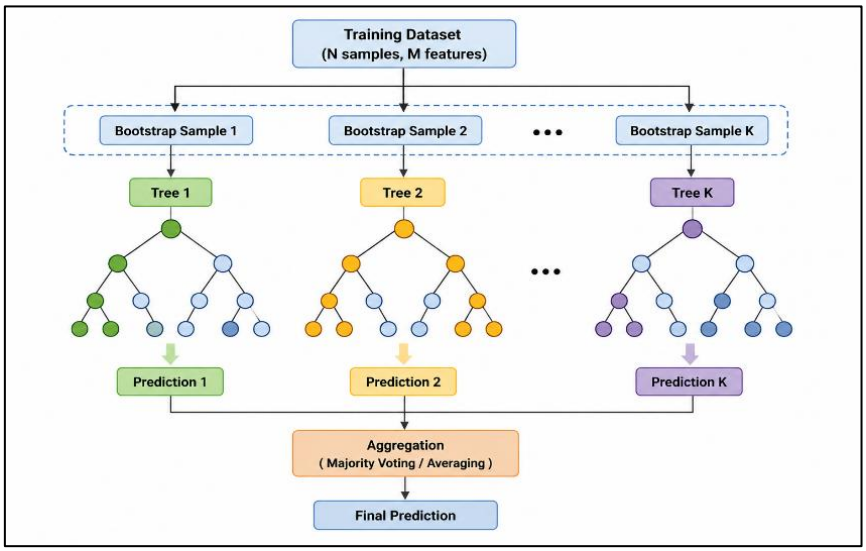

Figure 2. Illustration of the Random Forest Algorithm

In the field of data mining, Random Forest is widely regarded as a powerful predictive method due to its ability to handle both classification and regression tasks using a collective or group-based learning approach. The final prediction is generated by aggregating the outcomes of all decision trees, through majority voting for classification problems or by averaging predictions in regression settings. The illustration of Random Forest can be seen in Figure 2.

2.3.3. Extreme Gradient Boosting (XGBoost)

Extreme Gradient Boosting (XGBoost) is one of the most advanced ensemble machine learning algorithms, developed as an enhanced implementation of the Gradient Boosting (GB) framework. Functionally, XGBoost combines a series of simple predictive models—typically decision trees—to form a strong learner capable of delivering highly effective performance in both regression and classification tasks [18].

The major advancement of XGBoost compared to conventional Gradient Boosting lies in the introduction of features designed to improve model performance and stability. Among the most notable improvements are the inclusion of regularization and tree pruning mechanisms, which help mitigate the risk of overfitting. In addition, XGBoost is well known for its ability to efficiently process large datasets at high speed, supported by optimizations such as block-based structure and multithreaded CPU execution for parallel computation. The illustration of XGBoost can be seen in Figure 3.

2.4. SMOTE

Synthetic Minority Over-sampling Technique (SMOTE) is a data resampling algorithm designed to address bias in imbalanced datasets. SMOTE increases the number of samples in the minority class by generating synthetic observations through linear interpolation between an existing minority sample and its k-nearest neighbors. This process helps clarify the decision boundaries between classes and reduces the dominance of the majority class in model training [19]. Technically, SMOTE computes the difference between the feature vector of a minority instance and its nearest neighbor, multiplies this difference by a random value between 0 and 1, and then adds the result to the original feature vector to create a new synthetic sample. By doing so, SMOTE forces the decision region of the minority class to become more generalized, unlike simple duplication methods that tend to create overly specific decision boundaries and increase the risk of overfitting.

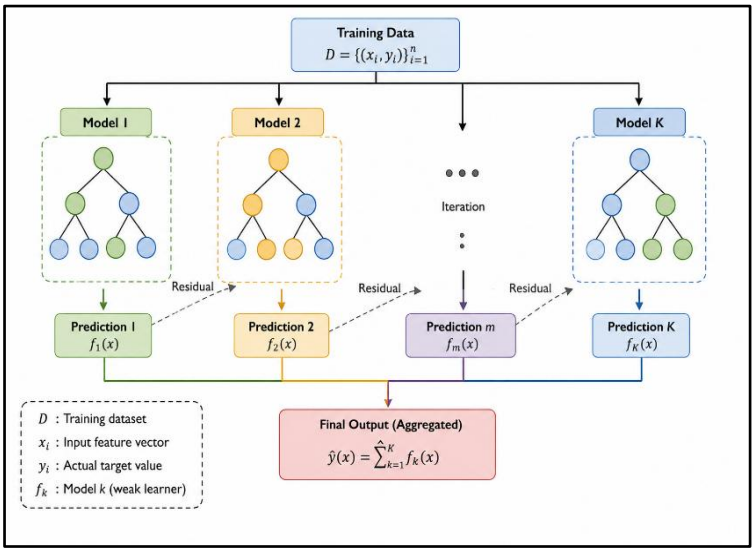

Figure 3. Illustration of the XGBoost Algorithm

2.5. Performance Evaluation

To validate the effectiveness of the algorithms tested, a comprehensive performance evaluation was conducted. The initial assessment is based on the confusion matrix, which maps prediction results into four fundamental components (True Negatives (TN), False Positives (FP), False Negatives (FN), and True Positives (TP)). These components are commonly used to compute standard predictive accuracy. However, relying solely on accuracy as a single metric can be misleading, particularly when dealing with imbalanced datasets or when the cost of misclassification is not uniform. In such situations, a model may appear to perform well simply by prioritizing the majority class, resulting in high accuracy that does not genuinely reflect its ability to detect minority class instances.

	Predicted Negative	Predicted Positive
Actual Negative	TN	FP
Actual Positive	FN	TP

Figure 1. Confussion Matrix

Therefore, following the recommendations of Wu T. et al. (2022), this study adopts the Receiver Operating Characteristic (ROC) curve as a more robust primary performance metric. The ROC curve illustrates the trade-off between the true positive rate (%TP) and the false positive rate (%FP), allowing for an evaluation that is independent of decision thresholds and prior probabilities [20]. According to Fitriani R. et al. (2021), the accuracy value is calculated as follows [21] :

$$\text{Acccuration} = \frac{TP + TN}{TP + FN + TN + FP} \times 100\% \quad (3)$$

$$\text{Recall} = \frac{TP}{TP + FN} \times 100\% \quad (4)$$

$$\text{Precision} = \frac{TP}{TP + FP} \times 100\% \quad (5)$$

$$\text{F1} - \text{score} = \frac{2TP}{2TP+FP+FN} \times 100\%$$

As a quantitative indicator, model performance was evaluated using the Area Under the ROC Curve (AUC). Unlike accuracy, AUC provides an unbiased assessment that is not affected by class distribution [22]. An AUC value approaching 1 indicates an ideal classifier with strong ability to distinguish between positive and negative classes, making it the most representative metric for this study. Furthermore, when ROC curves of multiple models intersect, the ROC convex hull analysis can be employed to identify the optimal classifier under specific error-cost constraints.

2.6. Data

This study utilizes the 2024 Village Potential Statistics (PODES) as the primary source of predictor variables. PODES is a nationwide survey conducted periodically by Statistics Indonesia (BPS), covering all villages and urban wards across the country. The dataset provides comprehensive information on demographic, social, economic, infrastructural, institutional, and natural resource conditions within each village. These characteristics make PODES a highly relevant and adequate data source for analyzing variations in development conditions at the village level.

The target variable, on the other hand, is obtained from the 2024 Village Development Index (IDM) published by the Ministry of Village, Development of Disadvantaged Regions, and Transmigration. IDM classifies villages into five levels of development (Independent, Advanced, Developing, Disadvantaged, and Highly Disadvantaged) [23]. However, in the context of Bekasi Regency, none of the villages fall into the Disadvantaged or Highly Disadvantaged categories. Therefore, this study focuses solely on the three relevant categories (Developing, Advanced, and Independent) which serve as the target variable for the classification process. The distribution of IDM categories in Bekasi Regency is illustrated in the figure provided.

Meanwhile, the predictor variables are derived from PODES 2024, as they contain quantitative indicators that objectively describe village conditions and align with the multidimensional assessment framework used in IDM. In addition to comprehensive coverage, PODES is an official dataset produced by the national statistical authority, ensuring high levels of reliability and validity for academic research.

By employing IDM as the dependent variable and PODES as the set of independent variables, this research design is methodologically robust and consistent with the principles of evidence-based policy development. The resulting classification model is expected to support the formulation of more data-driven village development strategies in Bekasi Regency. The complete list of variables used in the analysis is presented in Table 1.

The selection of the 33 predictor variables was guided by theoretical and empirical considerations related to the Village Development Index (IDM). Specifically, the variables were chosen because they represent key aspects of social, economic, and environmental resilience, which constitute the three principal dimensions used in IDM assessment. In addition, only variables that were consistently available in the 2024 PODES dataset, measurable at the village level, and considered relevant to village development conditions were included. Variables that were not directly related to the IDM dimensions or provided redundant information were excluded from the analysis. This theory-driven selection approach was adopted to ensure interpretability and policy relevance while avoiding the inclusion of less informative variables.

Table 1. Research variables

Var	Description	Class	Var	Description	Class
Y	Village Independence Status	1. Developing 2. Advanced 3. Independent	X17	Existence of Featured Products	1. Yes 2. No
X1	Tree Planting on Critical Land, Mangrove Planting, and Similar Activities by the Community	1. Yes 2. No	X18	Existence of Village Cooperatives (KUD)	1. Yes 2. No

Var	Description	Class	Var	Description	Class
X2	Waste/Material Processing or Recycling (Reuse, Recycle) by the Village Community	1. Yes 2. No	X19	Existence of KOPINKRA (Village-level Cooperative Network)	1. Yes 2. No
X3	Promotion of Organic Fertilizer Use in Agricultural Land	1. Yes 2. No	X20	Existence of Savings and Credit Cooperatives (KSP)	1. Yes 2. No
X4	Package A/B/C Education Activities	1. Yes 2. No	X21	Availability of KUR (People's Business Credit) Facilities	1. Yes 2. No
X5	Existence of Community Learning Centers (TBM)	1. Yes 2. No	X22	Availability of KUBE (Group Business Empowerment) Facilities	1. Yes 2. No
X6	Number of Integrated Health Posts (Posyandu)	Numeric	X23	Availability of Banks	1. Yes 2. No
X7	Number of Community Implementers/Leaders	Numeric	X24	Availability of BMT	1. Yes 2. No
X8	Number of Poor Family Certificates (SKTM) Issued by the Village	Numeric	X25	Availability of ATMs	1. Yes 2. No
X9	Number of Persons with Disabilities	Numeric	X26	Availability of Bank Agents	1. Yes 2. No
X10	Number of Family Welfare Movement (PKK) Members	Numeric	X27	Existence of Shop Groups	1. Yes 2. No
X11	Number of Youth Organization (Karang Taruna) Members	Numeric	X28	Existence of Permanent Markets	1. Yes 2. No
X12	Number of Farmer Groups	Numeric	X29	Existence of Semi-Permanent Markets	1. Yes 2. No
X13	Number of Community Groups	Numeric	X30	Number of Minimarkets	1. Yes 2. No
X14	Main Economic Sector of Village Population	Polynomial	X31	Number of Grocery Stores	Numeric
X15	Number of Micro and Small Industries	Numeric	X32	Number of Village-Owned Business Units (Bumdes)	1. Yes 2. No
X16	Number of Industrial Centers	Numeric	X33	Availability of Village Original Revenue (PADes)	1. Yes 2. No

2.7. Workflow

The research process begins with the collection of Potensi Desa (PODES) 2024 data for Bekasi Regency, which contains information on the social, economic, institutional, and environmental characteristics of villages. The collected dataset then undergoes data cleaning and preprocessing procedures, including handling missing values, transforming variables, and encoding categorical attributes to ensure compatibility with machine learning algorithms. In addition, data consistency checks

are performed to improve data quality and ensure the reliability of subsequent analyses.

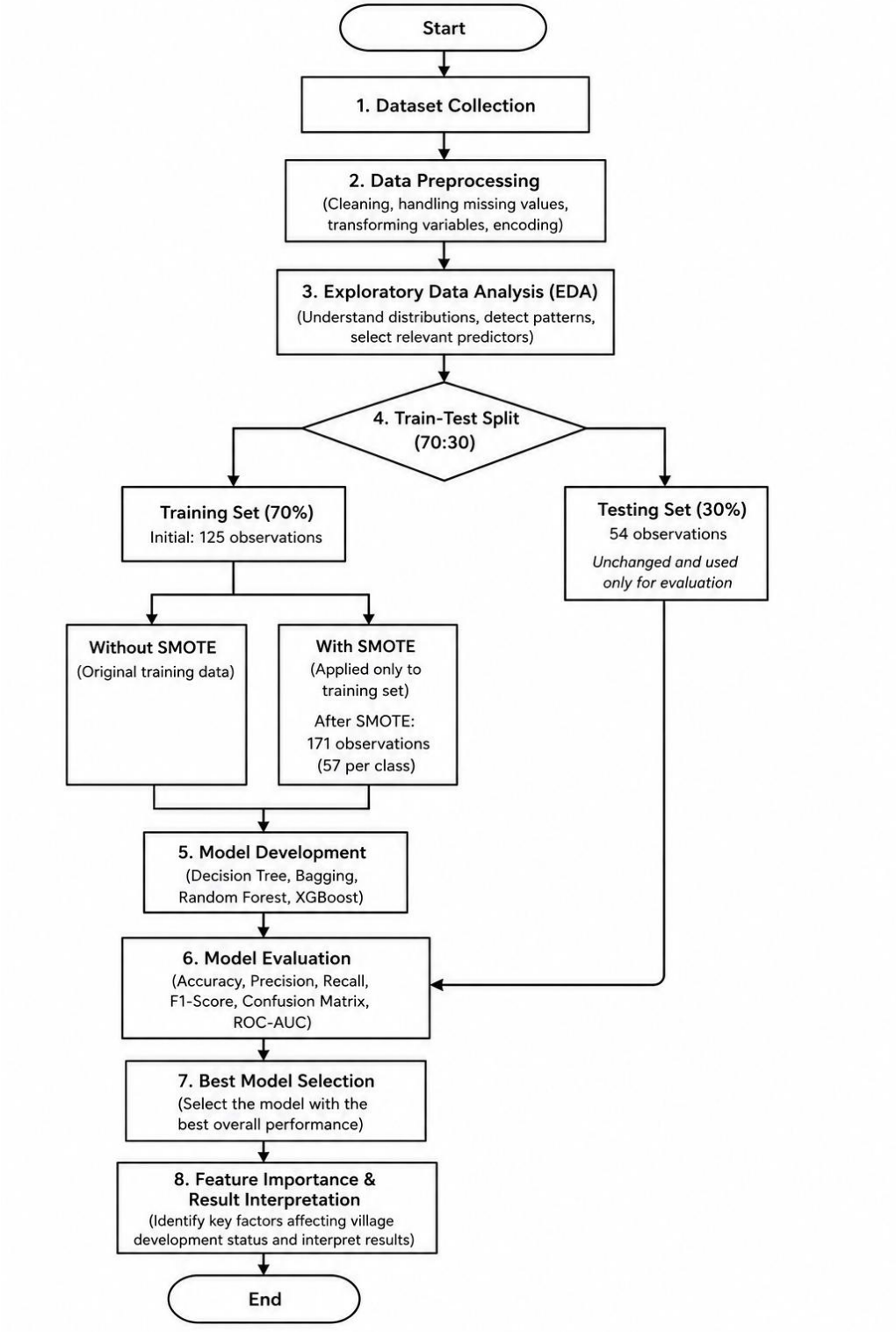

Figure 5. Workflow diagram

The next stage involves Exploratory Data Analysis (EDA) to understand the distribution of variables and identify preliminary patterns within the dataset. Descriptive statistics and visualization techniques are employed to detect anomalies and examine class distribution patterns. The insights obtained from this stage serve as the basis for selecting predictor variables that are conceptually relevant to the Village Development Index (IDM) framework.

The dataset is subsequently divided into training and testing subsets using a 70:30 proportion, where 70% of the observations are allocated for model training and 30% are reserved for model

evaluation. This proportion follows the recommendation of Muraina [25], who emphasizes the importance of an appropriate dataset splitting strategy to reduce modeling bias and improve model generalization capability. Based on this partitioning scheme, the training dataset consisted of 125 observations, while the testing dataset contained 54 observations.

To address the class imbalance problem, the Synthetic Minority Over-sampling Technique (SMOTE) was applied exclusively to the training dataset after the train-test split procedure. This approach was adopted to prevent data leakage and ensure that model evaluation remained unbiased. After applying SMOTE, the number of training observations increased from 125 to 171, resulting in a balanced class distribution of 57 observations for each village development category, while the testing dataset remained unchanged.

Following the data preparation stage, four classification models were developed and compared, namely Decision Tree, Bagging, Random Forest, and XGBoost. The Decision Tree model was used as the baseline classifier, whereas Bagging, Random Forest, and XGBoost represented ensemble learning approaches. All models were trained using the training dataset and subsequently evaluated using the testing dataset. To ensure reproducibility and consistency across experiments, each classification model was implemented using a predefined set of hyperparameters. The hyperparameter settings used in this study are presented in Table 2.

Table 2. Hyperparameter Model

Model	Main Hyperparameters
Decision Tree	max_depth = 10 random_state = 42
Bagging	n_estimators = 200 base_estimator = Decision Tree max_depth = 10 random_state = 42
Random Forest	n_estimators = 200 max_depth = 10 random_state = 42
XGBoost	objective = multi:softprob max_depth = 10 eval_metric = mlogloss random_state = 42
SMOTE	k_neighbors = 5 random_state = 42

The selected hyperparameter values were intended to provide a fair comparison among classification methods while minimizing excessive model complexity. For ensemble-based methods, 200 estimators were employed to improve predictive stability and reduce variance. In addition, SMOTE was implemented using the default value of k_neighbors = 5 to balance the minority classes within the training dataset.

The classification performance of each model was assessed using multiple evaluation metrics, including accuracy, precision, recall, F1-score, confusion matrix, and Area Under the Receiver Operating Characteristic Curve (AUC-ROC). These complementary metrics provide a comprehensive assessment of classification performance across different village development categories. Comparative evaluation results were subsequently used to identify the model that delivered the most reliable classification performance.

The final stage involves interpreting the classification results and deriving substantive conclusions from the analysis. The best-performing model is further examined to identify the most influential variables affecting village development status and to explain the observed classification patterns. The findings are expected to provide evidence-based insights that can support village development planning and policy formulation in Bekasi Regency. The complete research workflow is illustrated in Figure 5.

3. Results and Discussion

3.1. Overview of Data and Class Distribution

This study utilizes the 2024 Potensi Desa (PODES) dataset for Bekasi Regency, comprising 179 villages. The variables used in the analysis include social, economic, institutional, and environmental indicators, as listed in Table 1. These variables were selected to capture key dimensions of village development and to support the classification of village development status based on the Village Development Index (IDM).

The distribution of village development status in Bekasi Regency shows that the majority of villages fall under the Independent category, with a total of 82 villages. This represents the largest proportion among all categories and indicates that most villages in Bekasi Regency possess relatively strong socio-economic and institutional capabilities. In addition, 59 villages are classified as Advanced, reflecting that a considerable number still have potential for further development toward full independence, particularly through the enhancement of basic services, economic opportunities, and community empowerment. Meanwhile, the developing category includes 38 villages, indicating that a smaller share of areas still require more intensive attention in terms of basic infrastructure, social services, and institutional strengthening in order to catch up with the more advanced villages. IDM Distribution in Bekasi Regency can be seen in Figure 6.

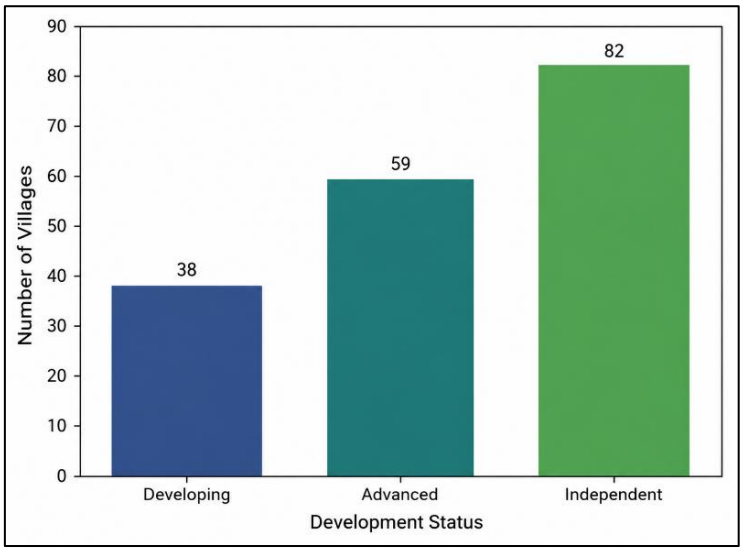

Figure 6. IDM distribution in Bekasi Regency

Overall, this distribution pattern indicates that Bekasi Regency is in a relatively strong position in terms of village development, with a predominance of independent villages reflecting progress across key development sectors. Nevertheless, the presence of advanced and developing villages remains a strategic policy concern to prevent widening disparities and to ensure that development progresses in an equitable and inclusive manner throughout the region.

During the initial exploratory analysis, class distribution revealed an imbalance among the categories (Developing, Advanced, and Independent). The *Independent* category accounted for the largest proportion of observations, while the *Developing* category represented the smallest share.

The correlation heatmap further shows that most numerical variables in the PODES dataset exhibit low to moderate correlations. This indicates that there are no overly strong relationships among variables, suggesting a low risk of multicollinearity and implying that each variable contributes relatively independent information to the village classification model. The highest correlation was observed between X10 and X11 ($r = 0.64$), which reflects a natural association between indicators that tend to develop simultaneously for example, improvements in basic services alongside supporting institutional structures. Moderate correlations were also found between X6–X7 ($r = 0.54$) and X15–X31 ($r = 0.46$), which may represent linkages between basic public service indicators and economic activity within villages. These patterns are reasonable, as more developed villages typically possess more complete public facilities and exhibit more dynamic economic conditions.

Most other variables demonstrated weak correlations (r = 0.00–0.30), including X12, X13, X14, and X16, showing that these indicators capture different dimensions of village conditions. This diversity strengthens the predictive power of machine learning algorithms when the variables are used collectively. Moreover, since no extremely high correlation (r > 0.80) was observed, all variables were retained in the analysis. Overall, the correlation pattern indicates that the numerical variables complement one another and do not exhibit redundancy. Heatmap variabel Numeric can be seen in Figure 7.

3.2. Initial Model Evaluation

The initial evaluation conducted prior to handling class imbalance shows that the performance of the four classification models was still limited due to the unequal distribution of the village development categories. As presented in Table 3, the Decision Tree model achieved an accuracy of 0.6667, reflecting the basic ability of a tree-based classifier to capture relationships among the PODES variables, although it remains sensitive to variance and prone to overfitting. The Bagging model demonstrated an improvement in accuracy to 0.7037, indicating that bootstrap aggregation helped reduce model variance and produce more stable predictions when compared with a single decision tree.

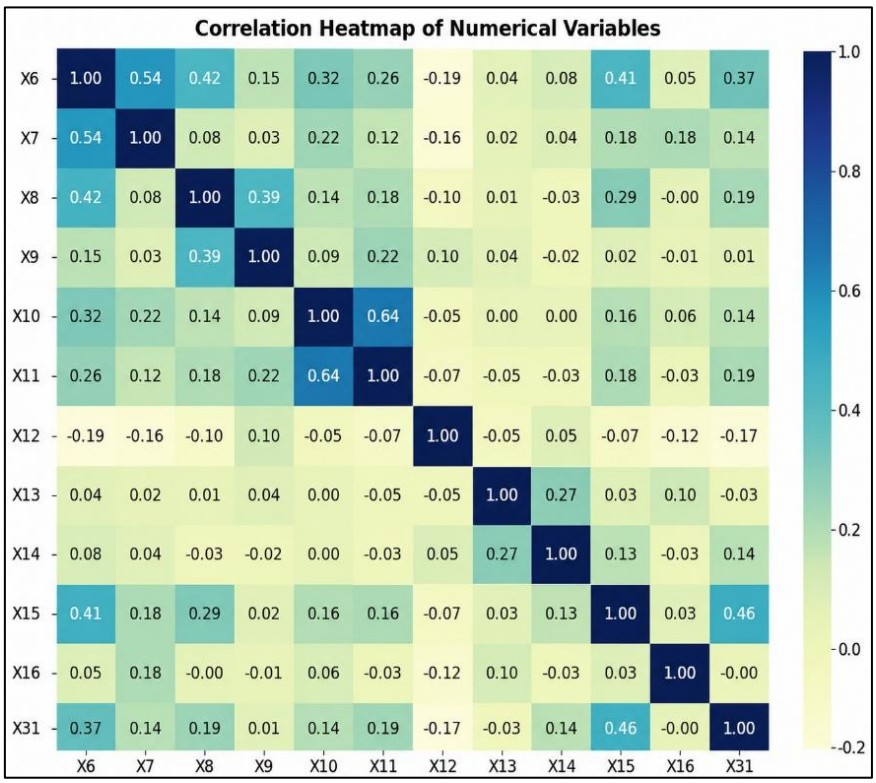

Figure 7. Heatmap of numerical variable

The Random Forest model demonstrated the best performance in the initial stage, achieving an accuracy of 0.7407. This result is consistent with the characteristics of the Random Forest algorithm, which combines hundreds of decision trees to provide more robust predictions, particularly for data with complex and heterogeneous structures. This indicates that the combination of feature randomness and bootstrap sampling is quite effective in enhancing the model's ability to capture inter-variable patterns, even when the data remain imbalanced.

In contrast, the XGBoost model yielded the lowest accuracy, 0.6296, in the initial evaluation. This performance may be attributed to XGBoost's sensitivity to learning rate, max depth parameters, and class distribution imbalance. Since XGBoost optimizes errors iteratively through boosting, class imbalance can cause the model to focus more on the majority class while neglecting minority classes, thereby reducing overall accuracy.

Tabel 3. Initial model accuracy

No	Model	Model Accuracy
1	Decision Tree	0.6667
2	Bagging	0.7037
3	Random Forest	0.7407
4	XGBoost	0.6296

Overall, the evaluation prior to applying SMOTE indicates that class imbalance has a significant impact on model performance, as reflected by the relatively moderate accuracy achieved across all models. These results underscore the importance of implementing imbalance-handling techniques to improve classification accuracy and reduce bias toward the majority class. This finding is consistent with previous studies showing that ensemble methods such as Random Forest generally provide more robust and stable predictive performance than single Decision Tree models [17], [24]. However, even ensemble models may experience performance degradation when trained on imbalanced datasets, making class-balancing techniques such as SMOTE essential for achieving optimal classification performance [19], [20], [25].

3.3. Improving Model Performance Using SMOTE

The application of class balancing techniques using SMOTE improved the performance of most classification models, although the single Decision Tree model experienced a decline in accuracy after oversampling. As shown in Table 4, Bagging, Random Forest, and XGBoost achieved higher accuracy after class balancing, indicating that the initial class imbalance negatively affected predictive performance. These findings suggest that ensemble-based models were better able to benefit from the balanced class distribution generated by SMOTE.

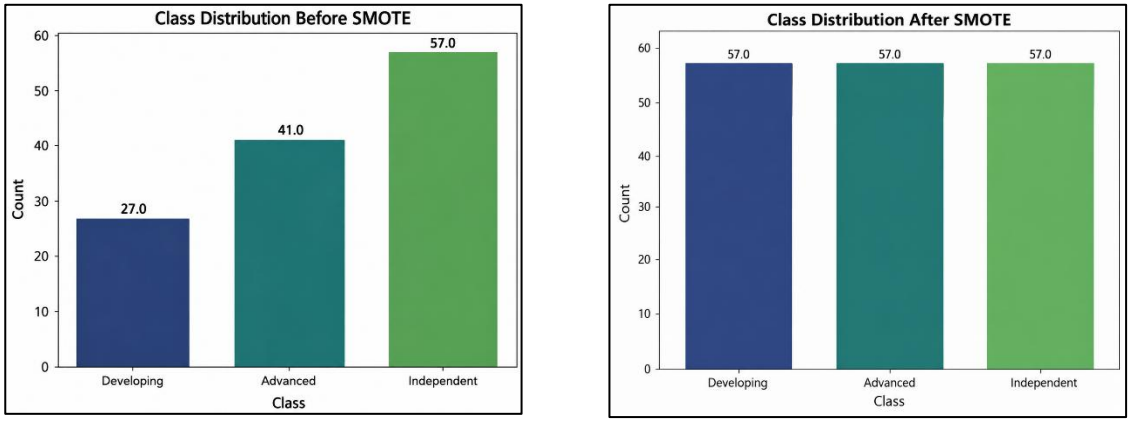

Figure 8. Class distribution of the target variable before and after handling class imbalance

The Decision Tree model experienced a decrease in accuracy from 0.6667 to 0.5741 after applying SMOTE. This decline may be attributed to the sensitivity of single decision trees to changes in data distribution introduced by synthetic samples. Since SMOTE generates new minority-class observations, the resulting data distribution may alter tree-splitting structures and decision boundaries, which can adversely affect the performance of individual tree-based classifiers [19]. In contrast, the Bagging model showed an increase in accuracy from 0.7037 to 0.7407, indicating that ensemble-based approaches are better able to leverage balanced class distributions. This finding is consistent with previous studies suggesting that ensemble methods such as Bagging and Random Forest are generally more robust to data variation and class imbalance because prediction errors are distributed across multiple trees rather than relying on a single classifier [19], [20].

Table 4. Comparison of model accuracy before and after SMOTE

No	Model	Accuracy Before SMOTE	Accuracy After SMOTE
1	Decision Tree	0.6667	0.5741
2	Bagging	0.7037	0.7407
3	**Random Forest**	**0.7407**	**0.7778**
4	XGBoost	0.6296	0.6481

The Random Forest model remained the best-performing model, improving from an initial accuracy of 0.7407 to 0.7778 after applying SMOTE. This indicates that Random Forest is not only robust to data variance but also capable of leveraging a more balanced class distribution to produce more accurate and stable classifications. Improvements were also observed in the XGBoost model, whose accuracy increased from 0.6296 to 0.6481. Although its initial performance was low due to boosting's sensitivity to the majority class, the results after SMOTE demonstrate that XGBoost can function more effectively when class proportion differences are minimized.

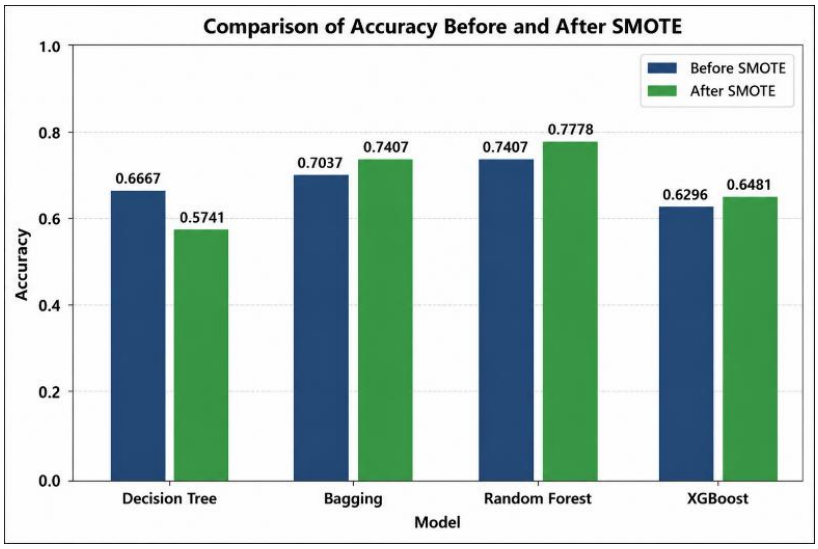

Figure 9. Model accuracy plot

Overall, the application of SMOTE improved the performance of most classification models, particularly Bagging, Random Forest, and XGBoost. Although the Decision Tree model experienced a decline in accuracy, the overall results indicate that class-balancing techniques can substantially enhance predictive performance in imbalanced datasets. These findings highlight that data balancing techniques are a crucial step in classification modeling, particularly when class distributions are uneven [25]. The use of SMOTE significantly contributes to improving classification model accuracy, especially in datasets with imbalanced classes. As reported by Wu et al. (2022), integrating SMOTE into machine learning architectures can achieve higher detection accuracy compared to conventional methods [20]. This improvement occurs because SMOTE enriches minority-class representations, enabling models to learn decision boundaries more effectively [19]. Model Accuracy Plot can be seen in Figure 9.

3.4. Best Model Selection

As presented in Table 4, the classification report indicates that the Random Forest model, applied after SMOTE, achieved good and balanced classification performance across all categories of village independence status. It should be noted that the precision, recall, and F1-score values are identical for each class because the confusion matrix exhibits a nearly symmetric misclassification pattern. Under this condition, the numbers of false positives and false negatives become equal for each class, resulting in identical precision and recall values and consequently identical F1-scores.

Table 5. Random Forest model classification report

Class	Precision	Recall	F1-Score	Support
Developing	0.8182	0.8182	0.8182	11
Advanced	0.6667	0.6667	0.6667	18
Independent	0.8400	0.8400	0.8400	25
accuracy			0.7778	54
macro avg	0.7749	0.7749	0.7749	54
weighted avg	0.7778	0.7778	0.7778	54

In the Developing category, the model achieved precision, recall, and F1-score values of 0.8182, indicating that it can accurately identify villages classified as Developing. In the Advanced category, all three metrics reached 0.6667. Although most villages predicted as Advanced were correctly classified, several Advanced villages were misclassified as either Developing or Independent. This outcome may occur because Advanced villages represent a transitional development stage and therefore share characteristics with both neighboring categories. The Independent category demonstrated the strongest classification performance, with precision, recall, and F1-score values of 0.8400. The high recall indicates that the model successfully identified most villages that truly belong to the Independent category. Meanwhile, the less-than-perfect precision suggests that some villages from other classes, particularly the Advanced category, were incorrectly classified as Independent. This finding reflects the socio-economic similarity between Advanced and Independent villages, making the boundary between these two categories less distinct.

Overall, the model achieved an accuracy of 0.7778, with macro-average and weighted-average F1-scores of 0.7749 and 0.7778, respectively. The closeness of these values suggests that the model provides relatively balanced classification performance across all classes without substantial bias toward any particular category. These findings indicate that the integration of SMOTE-based class balancing and the Random Forest ensemble algorithm is effective in improving predictive performance on datasets with imbalanced class distributions, making it a suitable approach for modeling village independence status classification in Bekasi Regency.

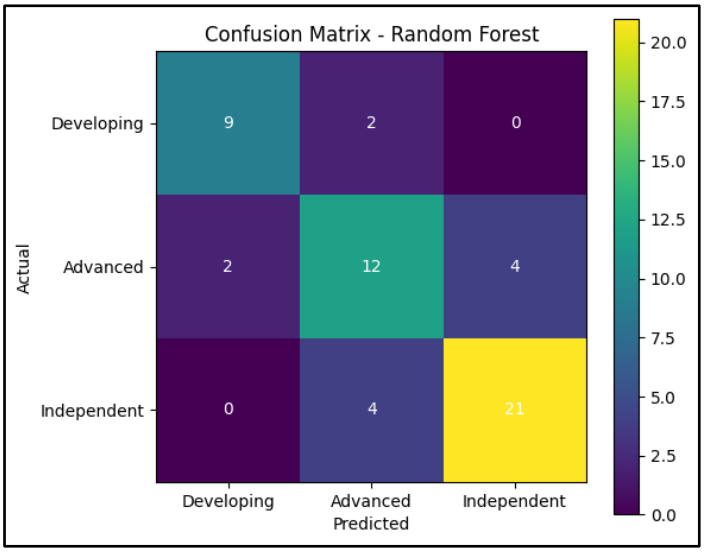

Figure 10. Confusion matrix of the Random Forest Model

The confusion matrix indicates that the Random Forest model (Figure 10), after applying SMOTE, was able to classify the majority of villages into the correct categories. In the Developing category, the model correctly identified 9 out of 11 villages, while 2 villages were misclassified as Advanced, and none were predicted as Independent. For the Advanced category, the model correctly classified 12 out of 18 villages, with 2 misclassified as Developing and 4 misclassified as Independent. This aligns with the relatively lower recall for the Advanced category, reflecting overlapping characteristics with the other two categories.

Meanwhile, in the Independent category, the model correctly identified 21 out of 25 villages, with 4 misclassified as Advanced and none classified as Developing. Overall, these results confirm that most prediction errors occurred between the Advanced and Independent categories, which substantively share

similar socio-economic characteristics, making the boundary between them not always distinct. The multi-class ROC curve illustrates the performance of the Random Forest model in distinguishing among the different village categories (Figure 11). Overall, all three ROC curves lie above the diagonal baseline, indicating that the model has good classification capability across all classes.

In the Developing category, the AUC value is 0.9387, the highest among the three classes. This demonstrates that the model is highly effective in differentiating Developing villages from the other two categories, with a relatively low error rate. The Independent category also shows strong performance, with an AUC of 0.9297. This success is consistent with the confusion matrix results, where most Independent villages were correctly classified. The high AUC indicates that the model can consistently and reliably recognize the characteristic patterns of Independent villages. Meanwhile, the Advanced category has the lowest AUC at 0.8148, though it still falls within the good range. This value confirms that the model's ability to separate Advanced villages from the other two classes is lower compared to the other categories. This aligns with the classification results, which show that some Advanced villages tend to be misclassified as either Developing or Independent, reflecting overlapping characteristics among the categories [26]. Overall, AUC values above 0.80 across all three classes confirm that the Random Forest model exhibits strong and reliable classification performance in predicting village independence levels based on the selected predictor variables.

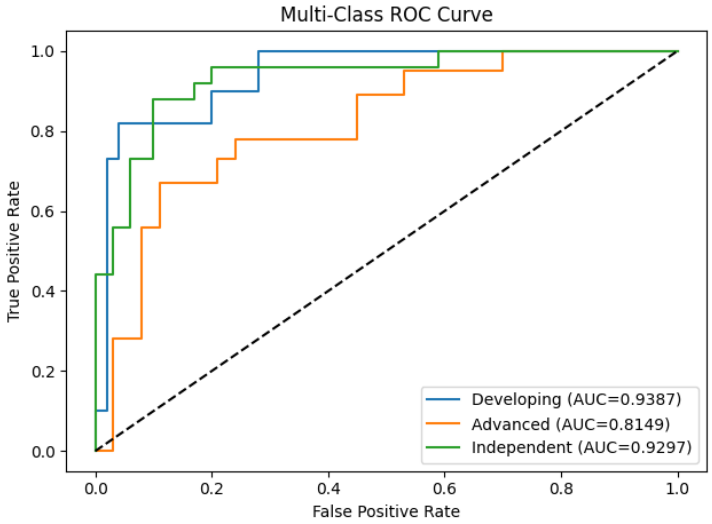

Figure 11. ROC Curve of the best model

3.5. Prediction on Test Data

To provide a clearer picture of model performance on previously unseen data, Table 5 presents a comparison between the actual classification of village independence status and the predictions generated by the selected model for 20 test observations. The table also includes the associated prediction probabilities for each class (Developing, Advanced, and Independent) allowing a more detailed interpretation of the model's level of confidence in each decision.

Overall, the model demonstrates a strong ability to identify the correct class for most observations. Several data points, such as those with indices 15, 117, and 104, were correctly classified as independent, with the highest probability also assigned to that category. This pattern suggests that the model is capable of distinguishing distinct class characteristics when the underlying feature patterns are well represented.

Nevertheless, some misclassifications remain, for example at indices 130 and 14, where the model predicted a different class from the actual label. In many of these instances, the predicted probabilities for the competing classes were relatively close, indicating that the observations may possess overlapping characteristics or fall near the boundary between class definitions. This highlights the inherent complexity of the classification problem and the possibility that some villages share transitional attributes between development categories.

Despite these occasional errors, the predictions on the test dataset reaffirm the model's overall reliability and support the earlier evaluation metrics. The results demonstrate that, after applying

SMOTE to address class imbalance, the model learns meaningful patterns and can generalize well when applied to real or future data.

Table 6. Comparison of actual and predicted classifications on 20 test observations

Index	Actual Class	Predicted Class	Probability Developing	Probability Advanced	Probability Independent
130	Advanced	Developing	0.7362	0.2403	0.0235
15	Independent	Independent	0.0200	0.1850	0.7950
117	Independent	Independent	0.0204	0.3521	0.6275
14	Advanced	Independent	0.0717	0.3038	0.6246
152	Developing	Developing	0.7025	0.2608	0.0367
165	Advanced	Advanced	0.1610	0.6490	0.1900
169	Developing	Developing	0.7035	0.2459	0.0506
104	Independent	Independent	0.0870	0.3072	0.6058
167	Advanced	Advanced	0.1938	0.6112	0.1950
13	Advanced	Advanced	0.1550	0.4888	0.3563
53	Advanced	Advanced	0.3716	0.5482	0.0802
144	Developing	Developing	0.5041	0.4649	0.0310
79	Advanced	Advanced	0.0920	0.7260	0.1820
175	Developing	Developing	0.9000	0.0800	0.0200
115	Independent	Independent	0.0589	0.3935	0.5476
106	Advanced	Advanced	0.0666	0.4690	0.4644
2	Advanced	Independent	0.3150	0.3200	0.3650
36	Developing	Developing	0.6324	0.3674	0.0002
9	Advanced	Independent	0.0354	0.4371	0.5275

3.6. Analysis of Dominant Factors (Feature Importance)

As can be seen in Figure 12, the feature importance analysis indicates that the most influential variable in determining village independence levels is X14 (Main Economic Sector of Village Population). The high importance of this variable suggests that the dominant economic structure of a village community plays a central role in shaping whether a village is Developing, Advanced, or Independent. Villages with productive economic bases, such as trade, industry, or services, tend to have stronger economic capabilities compared to those still reliant on traditional sectors. The next significant contributor is X12 (Number of Farmer Groups), highlighting that the presence of agricultural institutions remains an important foundation for strengthening village economies. A higher number of farmer groups suggests a more structured agricultural production system, better access to training, and stronger networks for community-based development.

Additionally, X6 (Number of Integrated Health Posts/Posyandu) also plays a major role in shaping village independence status. This indicates that the availability of basic health services, particularly for mothers and children, not only influences public health quality but also affects productivity and the economic competitiveness of the village. Meanwhile, the high importance of X31 (Number of Grocery Stores/Warung Kelontong) underscores that local economic circulation through small businesses is a strong indicator of community economic activity. A large number of grocery stores reflects increased purchasing power and dynamic trade activity.

Several other variables, such as X8 (Number of Poor Families/SKTM), X27 (Number of Shop Groups), and X15 (Number of Micro and Small Industries), also contribute meaningfully to the model's decision-making. These variables essentially reflect the dynamics of social welfare and local economic activity. In terms of social institutions, X7 (Number of Community Leaders/Executors) is significant in representing the community's capacity to manage village empowerment programs. Meanwhile, variables such as X9 (Number of Persons with Disabilities) and X30 (Number of Minimarkets) provide

supplementary information regarding social service challenges and the modernization of village commerce.

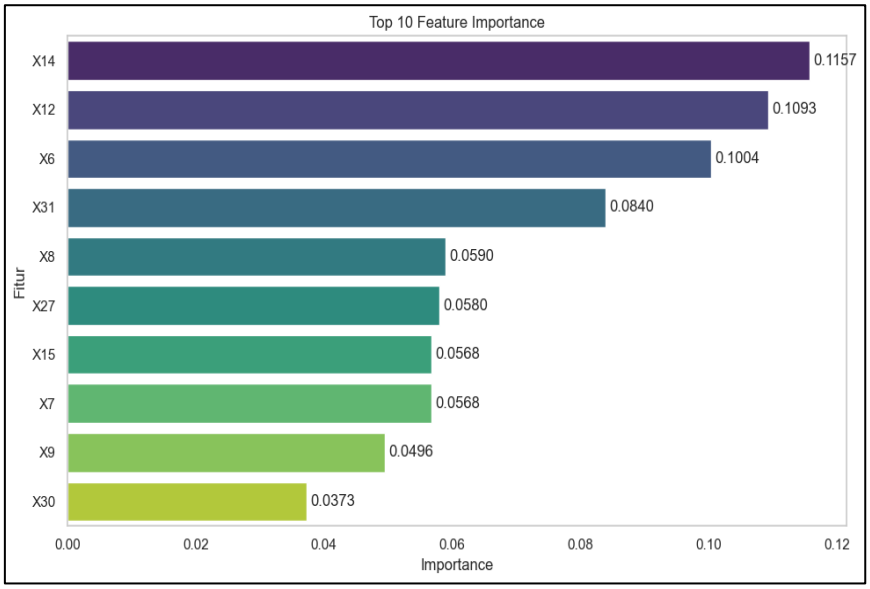

Figure 12. Feature importance best model

Overall, this pattern confirms that village independence is influenced not only by economic factors but also by the presence of social infrastructure, institutions, and robust basic service support. The stronger the local economic structure and community empowerment capacity, the higher the likelihood that a village can sustainably progress toward Independent status.

3.7. Substantive Interpretation of Model Results

The study results indicate that the Random Forest model, after applying SMOTE, was able to deliver strong classification performance in mapping village independence levels in Bekasi Regency. The model achieved an accuracy of 0.7778 with a macro-average F1-score of 0.7749, suggesting that it is not only accurate but also relatively balanced in recognizing all three class categories (Developing, Advanced, and Independent). These findings indicate that the application of data balancing techniques can enhance the algorithm's effectiveness in handling imbalanced class distributions.

Substantively, the model's performance reflects patterns consistent with the realities of village development in Bekasi Regency. In the Independent category, the model achieved its best performance, with precision, recall, and F1-score values of 0.8400. This suggests that Independent villages are easier to distinguish and exhibit more consistent variable patterns compared to the other two categories. Independent villages generally have strong local economic structures, well-developed institutional capacity, and adequate public facilities, making their distinguishing patterns more easily captured by the model.

In contrast, the Advanced category showed the lowest performance, with precision, recall, and F1-score values of 0.6667. This aligns with the confusion matrix results, where some Advanced villages were misclassified as Developing or Independent. This phenomenon reflects that Advanced villages are in a transitional phase and share socio-economic characteristics with both other categories, making the boundaries between classes less distinct. Developing villages exhibited better performance, with an F1-score of 0.8182, indicating that their characteristics can still be consistently recognized by the model.

The findings of this study are consistent with previous research emphasizing the advantages of ensemble learning over single-tree classification approaches. Studies by Joses et al. [11] and Breiman [24] reported that ensemble-based methods generally provide higher predictive stability and better generalization performance because prediction errors are distributed across multiple learners. Furthermore, the observed improvement after applying SMOTE supports the findings of Wu et al. [20] and Pulungan et al. [25], who demonstrated that class-balancing techniques can substantially improve classification effectiveness on imbalanced datasets. The accuracy achieved in this study (0.7778) therefore indicates that the proposed Random Forest–SMOTE framework provides competitive

performance for village development status classification while maintaining balanced predictive capability across all classes.

Further examination of the AUC values in the ROC curve reinforces these findings. The Developing category achieved the highest AUC of 0.9387, followed by Independent at 0.9297, while Advanced had the lowest AUC at 0.8148. Nevertheless, all three AUC values exceed 0.80, indicating that the model possesses strong discriminative ability across all classes. Overall, the model is capable of providing a substantive mapping of village independence status that is consistent with the development patterns observed in Bekasi Regency.

3.8. Strengths & Limitations

This study has several methodological and empirical strengths worth noting. First, the use of an ensemble learning approach, particularly Random Forest, has been shown to provide better classification performance than a single Decision Tree model. This finding is consistent with previous studies demonstrating that ensemble methods reduce model variance, improve predictive stability, and generally achieve better generalization performance by aggregating multiple learners [11], [24]. Second, the application of the SMOTE class-balancing technique effectively improved the model's ability to recognize minority classes, resulting in more equitable classification outcomes that are less biased toward dominant classes. Similar findings have been reported by Chawla et al. [19], Wu et al. [20], and Pulungan et al. [25], who showed that oversampling techniques can substantially improve classification performance in imbalanced datasets. Third, the model leveraged official PODES and IDM data, which are highly valid, comprehensive, and objectively reflect village-level empirical conditions. This strengthens the relevance of the study findings in the context of evidence-based policy and development planning.

However, several limitations should also be acknowledged. First, the model was evaluated using data from a single year (cross-sectional), which limits its ability to capture the longitudinal dynamics of village independence. Time-series analysis could provide deeper insights into village development trajectories over time. Second, the predictor variables were primarily derived from PODES indicators, which are largely quantitative and administrative in nature. Consequently, dimensions such as social capital, community participation, and institutional quality that are not directly captured in numerical form may not be fully represented in the model. Third, the classification results still reveal overlap between the Advanced and Independent categories, primarily due to the similarity of characteristics between villages in these two development stages. This suggests that additional discriminative variables or alternative modeling approaches may be required to improve class separation.

Moreover, this study selected Random Forest as the best-performing model based on empirical evaluation results. Nevertheless, future research could explore other boosting-based algorithms optimized through hyperparameter tuning, which may yield higher predictive performance. The implementation of Explainable Artificial Intelligence (XAI) approaches, such as SHAP or LIME, may also provide deeper insights into the contribution of individual variables to model predictions. Overall, this study provides a robust foundation for applying machine learning to village independence mapping while highlighting opportunities for future methodological improvements, richer data integration, and the incorporation of spatial and temporal dimensions into village development analysis.

4. Conclusion

This study demonstrates that the integration of Village Potential Statistics (PODES) 2024 and the Village Development Index (IDM) through a Random Forest classification model combined with the Synthetic Minority Over-sampling Technique (SMOTE) provides a reliable approach for mapping village independence status in Bekasi Regency. The findings confirm that machine learning methods can support a more objective, consistent, and data-driven assessment of village development by effectively utilizing multidimensional village-level indicators.

he analysis reveals that village independence is closely associated with factors related to local economic structure, agricultural institutions, access to basic health services, and community economic activities. These findings suggest that village development is not determined solely by economic performance, but also by the availability of supporting institutions and public services that strengthen community capacity and resilience. Consequently, policies aimed at accelerating village development should prioritize strengthening local economic sectors, expanding access to essential services, and enhancing community-based institutions.

From a practical perspective, the proposed classification framework can support the Bekasi Regency Government in identifying villages that require targeted interventions, monitoring development progress, and allocating resources more efficiently based on empirical evidence. In particular, villages classified as Advanced may require focused policy attention because they represent a transitional stage with characteristics overlapping those of Developing and Independent villages. Targeted programs designed to strengthen economic opportunities, institutional capacity, and public service provision may help accelerate their transition toward higher levels of village independence.

Furthermore, the study highlights the potential of predictive analytics as a complementary tool for evidence-based rural development planning. Beyond describing current village conditions, machine learning models can assist policymakers in identifying development patterns and anticipating future needs, thereby improving the effectiveness of planning and evaluation processes.

Nevertheless, this study is limited by its reliance on cross-sectional data from a single year and by the availability of predominantly quantitative indicators. Future research is therefore encouraged to incorporate longitudinal datasets, additional social and institutional variables, and advanced analytical approaches, including spatial analysis and explainable artificial intelligence (XAI). Such developments would provide deeper insights into village development dynamics and further strengthen the contribution of data-driven methods to rural policy formulation and decision-making.

Ethics approval

Not required.

Acknowledgments

The authors would like to express their sincere gratitude to Mr. Krido Saptono, Head of the BPS-Statistics Bekasi Regency, for his encouragement and institutional support throughout this research. The authors also thank the BPS- Statistics Bekasi Regency for providing access to the 2024 Village Potential Statistics (PODES) data used in this study. Furthermore, the authors sincerely appreciate the anonymous reviewers and the Editor of the Jurnal Aplikasi Statistika & Komputasi Statistik for their valuable comments and constructive suggestions, which have significantly improved the quality of this manuscript.

Competing interests

The author declares that there are no conflicts of interest related to this study.

Funding

This research received no external funding.

Underlying data

The data supporting the findings of this study are available from the corresponding author upon reasonable request.

Credit Authorship

Mochamad Ridwan: Conceptualization, Methodology, Data Curation, Software, Writing (Original Draft, Writing, Review and Editing), Visualization. **Erwin Tanur**: Supervision, Validation, Writing – Review & Editing.

References

[1] B. Rahmasari, "Paradigma Pembangunan Desa Dalam Pengelolaan Keuangan Desa Berdasarkan Undang-Undang Nomor 6 Tahun 2014 Tentang Desa," *Volksgeist: Jurnal Ilmu Hukum dan Konstitusi*, vol. 3, no. 2, 2020, doi: 10.24090/volksgeist.v3i2.4001.

[2] S. Agung, *Pemerintahan asli masyarakat adat: sebuah studi kepemimpinan adat di Lembah Timur Ciamis, Jawa Barat*. Deepublish, 2020.

[3] P. Hendrarso, P. Handoko, M. Faiz Ali Ramdhani, N. Andayani, and R. Tania, "Kajian Pengentasan Desa Tertinggal Melalui Pendekatan Indeks Desa Membangun," *Transparansi : Jurnal Ilmiah Ilmu Administrasi*, vol. 4, no. 1, 2021, doi: 10.31334/transparansi.v4i1.1607.

[4] S. Sriningsih, E. Astuti, B. Ismiwati, and F. Ekonomi, "Implementasi PERMENDESAPDTTRANS NO. 2 Tahun 2016 Terkait Status Desa di Desa Sukarara Lombok Tengah," *Jurnal Kompetitif: Media Informasi Ekonomi Pembangunan, Manajemen dan Akuntansi*, vol. 6, no. 1, 2020.

[5] A. M. Gai, A. Witjaksono, and R. R. Maulida, "Perencanaan dan Pengembangan Desa," 2020, *Dream Litera Buana*.

[6] A. Amka, M. Anshar Nur, and J. Jamalluddin, "Kebijakan dan Strategi Pembangunan Daerah untuk Masa Depan," *CV BRAVO PRESS Indonesia*, (n.d.).

[7] A. H. Nasrullah, "Implementasi algoritma Decision Tree untuk klasifikasi produk laris," *Jurnal Ilmiah Ilmu Komputer Fakultas Ilmu Komputer Universitas Al Asyariah Mandar*, vol. 7, no. 2, pp. 45–51, 2021.

[8] H. Kurniawan, "Deteksi Twitter Bot menggunakan Klasifikasi Decision Tree," *Jurnal Sustainable: Jurnal Hasil Penelitian dan Industri Terapan*, vol. 9, no. 1, pp. 31–37, 2020.

[9] I. Setiawan, R. F. A. Cahyani, and I. Sadida, "Exploring complex decision trees: Unveiling data patterns and optimal predictive power," *Journal of Innovation And Future Technology (IFTECH)*, vol. 5, no. 2, pp. 112–123, 2023.

[10] R. N. Ramadhon, A. Ogi, A. P. Agung, R. Putra, S. S. Febrihartina, and U. Firdaus, "Implementasi Algoritma Decision Tree untuk Klasifikasi Pelanggan Aktif atau Tidak Aktif pada Data Bank," *Karimah Tauhid*, vol. 3, no. 2, pp. 1860–1874, 2024.

[11] S. Joses, D. Yulvida, and S. Rochimah, "Pendekatan metode ensemble learning untuk prakiraan cuaca menggunakan soft voting classifier," *Journal of Applied Computer Science and Technology*, vol. 5, no. 1, pp. 72–80, 2024.

[12] T. H. Handoko, *Manajemen personalia dan sumberdaya manusia*. Bpfe, 2016.

[13] D. S. Lindawaty, "Pembangunan desa pasca Undang-Undang No. 6 Tahun 2014 tentang desa [Village development post Law No. 6 of 2014 on villages]," *Jurnal Politica Dinamika Masalah Politik Dalam Negeri Dan Hubungan Internasional*, vol. 14, no. 1, pp. 1–21, 2023.

[14] Direktorat Jenderal Pembangunan Desa dan Perdesaan | KDPDTT, "Tentang Indeks Desa Membangun.", (n.d.).

[15] H. Blockeel, L. Devos, B. Frénay, G. Nanfack, and S. Nijssen, "Decision trees: from efficient prediction to responsible AI," *Front. Artif. Intell.*, vol. 6, p. 1124553, 2023.

[16] L. Breiman, "Bagging predictors," *Mach. Learn.*, vol. 24, no. 2, pp. 123–140, 1996.

[17] H. A. Salman, A. Kalakech, and A. Steiti, "Random Forest Algorithm Overview," *Babylonian Journal of Machine Learning*, vol. 2024, pp. 69–79, Jun. 2024, doi: 10.58496/BJML/2024/007.

[18] M. Niazkar *et al.*, "Applications of XGBoost in water resources engineering: A systematic literature review (Dec 2018–May 2023)," *Environmental Modelling & Software*, vol. 174, p. 105971, Mar. 2024, doi: 10.1016/j.envsoft.2024.105971.

[19] N. V Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[20] T. Wu, H. Fan, H. Zhu, C. You, H. Zhou, and X. Huang, "Intrusion detection system combined enhanced random forest with SMOTE algorithm," *EURASIP J. Adv. Signal Process.*, vol. 2022, no. 1, p. 39, 2022.

[21] R. D. Fitriani, H. Yasin, and T. Tarno, "Penanganan klasifikasi kelas data tidak seimbang dengan random oversampling pada naive bayes (Studi kasus: Status peserta KB IUD di Kabupaten Kendal)," *Jurnal Gaussian*, vol. 10, no. 1, pp. 11–20, 2021.

[22] Ş. K. Çorbacıoğlu and G. Aksel, "Receiver operating characteristic curve analysis in diagnostic accuracy studies: A guide to interpreting the area under the curve value," *Turk. J. Emerg. Med.*, vol. 23, no. 4, pp. 195–198, 2023.

[23] D. S. Lindawaty, "Pembangunan desa pasca Undang-Undang No. 6 Tahun 2014 tentang desa [Village development post Law No. 6 of 2014 on villages]," *Jurnal Politica Dinamika Masalah Politik Dalam Negeri Dan Hubungan Internasional*, vol. 14, no. 1, pp. 1–21, 2023.
[24] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.
[25] M. P. Pulungan, A. Purnomo, and A. Kurniasih, "Penerapan SMOTE untuk Mengatasi Imbalance Class dalam Klasifikasi Kepribadian MBTI Menggunakan Naive Bayes Classifier," *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 11, no. 5, pp. 1033–1042, 2024.
[26] D. Chicco and G. Jurman, "The Matthews correlation coefficient (MCC) should replace the ROC AUC as the standard metric for assessing binary classification," *BioData Min.*, vol. 16, no. 1, p. 4, 2023.