

A Random Forest Approach for Classifying Regional Social Stability in Banten Province Based on Socioeconomic Indicators

Widyawati^{1*}, Ika Ima Nissa², Bagus Setya³, Lisdianto Dwi Kesumahadi⁴, Yuda Pratama Wibawa⁵

^{1,2,3,4,5}Information Systems Study Program, Universitas Pamulang, Banten, Indonesia

E-mail: ^{1*}dosen03354@unpam.ac.id, ²dosen03042@unpam.ac.id, ³dosen03357@unpam.ac.id, ⁴dosen03364@unpam.ac.id, ⁵dosen03045@unpam.ac.id

(Received: 14 May 2026, revised: 6 Jun 2026, accepted: 8 Jun 2026)

Abstract

Social stability is an important indicator in measuring regional welfare and development conditions. This study aims to classify the level of social stability in Banten Province using the Random Forest method based on several social indicators, including poverty, unemployment, Human Development Index (HDI), HDI percentage growth, Gini Ratio, population density, and average years of schooling (RLS). The dataset used in this study was obtained from the regency and city social statistical data in Banten Province from 2020 to 2025. The research process was carried out through several stages, namely data collection, data preprocessing, social stability label determination, splitting training and testing data, Random Forest modeling, model evaluation, feature importance analysis, and result visualization using the Python programming language. The Random Forest algorithm was selected because of its ability to perform classification effectively and reduce overfitting in classification models. The evaluation results showed that the Random Forest model achieved an accuracy value of 100%, with precision, recall, and F1-score values of 1.00. In addition, the feature importance analysis indicated that the unemployment variable had the strongest influence on the classification results, followed by HDI and poverty variables, while the Gini Ratio had the lowest influence. The findings of this study indicate that the Random Forest method can be effectively applied to classify regional social stability in Banten Province based on social indicators and can support data-driven regional policy analysis.

Keywords: Banten Province, Classification, Machine Learning, Random Forest, Social Stability

I. INTRODUCTION

Regional social stability is one of the important indicators in regional development because it is closely related to public welfare, social security, equitable development, and sustainable economic growth. A region can be considered socially stable when its people are able to live in safe, harmonious, productive conditions and possess an adequate level of welfare. Conversely, social instability may trigger various problems such as increasing poverty, unemployment, social inequality, social conflict, and declining quality of life.

In recent years, Indonesian society has faced various social phenomena indicating vulnerabilities in social stability across regions. The increasing prevalence of online gambling, illegal online loans, social conflicts, and economically driven crimes has become a growing concern. These conditions are influenced not only by the rapid development of digital technology but also by socioeconomic pressures such as unemployment, poverty, and income inequality. This phenomenon demonstrates that regional social stability is

affected not only by security factors but also by the overall socioeconomic conditions of society.

One of the major issues receiving public attention in recent years is the increasing use of online loan services and the widespread practice of online gambling. The Financial Services Authority (OJK) reported that the total online loan or fintech peer-to-peer lending debt in Indonesia as of March 2026 reached IDR 101.03 trillion, representing an annual growth of 26.25% [1]. The high level of online borrowing activity indicates economic pressures that drive people to seek instant financing access. On the other hand, online gambling and illegal lending practices have also caused various social impacts, including family conflicts, criminal activities, psychological stress, and declining public productivity. Indirectly, these conditions may affect regional social stability because they are closely related to public welfare, social security, and the quality of socioeconomic life.

Banten Province, as one of the strategic regions in Indonesia, has relatively complex socioeconomic characteristics. As an industrial area and a supporting region for urban development, Banten experiences rapid population

growth and economic activity, yet it also faces various social challenges such as unemployment, economic inequality, and population density. Based on data from the Central Statistics Agency (BPS), several regencies and cities in Banten still exhibit relatively high unemployment and poverty rates. Differences in socioeconomic conditions among regions may influence the level of social stability in each area. Regions with lower welfare levels, higher economic inequality, and uneven human resource quality tend to have greater social vulnerability compared to regions with better socioeconomic conditions.

Differences in development levels, population density, human development index, unemployment rate, and economic inequality result in varying conditions of social stability across regions. Areas with high development levels do not necessarily possess good social stability if significant social inequality or unemployment still exists. Therefore, an analytical approach is needed to objectively classify regional social stability conditions based on socioeconomic indicators.

The advancement of information technology and data science has enabled classification analysis to be conducted using machine learning methods [2], [3], [4]. One of the widely used classification methods is Random Forest [5]. This method is an extension of the Decision Tree algorithm that works by constructing multiple decision trees to improve accuracy and reduce the risk of overfitting [6]. The Random Forest algorithm demonstrates strong capability in handling numerical data, high-dimensional datasets, and generating predictions with relatively high accuracy [7].

In this study, the socioeconomic indicators used include poverty rate, unemployment rate, Human Development Index (HDI), HDI percentage, Gini Ratio, population density, and average years of schooling. These variables were selected because they are considered to have a direct relationship with the social stability of a region. Poverty and unemployment rates may influence public welfare, while HDI and average years of schooling reflect the quality of human resources. In addition, the Gini Ratio describes the level of economic inequality, whereas population density is associated with social pressure within a region.

Previous studies on socioeconomic classification have been conducted using various machine learning methods as shown in Table 1. However, most studies focus on poverty prediction, community welfare classification, or regional development analysis in general. Research specifically addressing the classification of regional social stability based on socioeconomic indicators, particularly in Banten Province, remains relatively limited. Therefore, this study is expected to contribute to the development of regional social stability classification models using the Random Forest algorithm.

Table 1. Comparison of Previous Studies Related to Socioeconomic Classification

Study	Method	Research Focus	Main Findings
Zhou et al. (2022)	Random Forest	Village-level Poverty Identification	Random Forest effectively

Study	Method	Research Focus	Main Findings
[8]			identified poverty conditions based on socioeconomic and geospatial indicators.
Muñetón-Santa & Manrique-Ruiz (2023) [9]	Random Forest, CatBoost, LightGBM	Multidimensional Poverty Prediction	Random Forest achieved competitive performance in predicting multidimensional poverty.
Ahmad et al. (2023) [10]	Machine Learning Classification	Social Aid Beneficiary Selection	Machine learning improved the accuracy of social assistance beneficiary classification.
Zamzuri et al. (2023) [11]	Random Forest, XGBoost	Poverty Forecasting	Ensemble learning methods provided reliable socioeconomic classification performance.
This Study	Random Forest	Regional Social Stability Classification	Random Forest was applied to classify social stability levels in Banten Province based on socioeconomic indicators.

Previous studies have demonstrated the effectiveness of machine learning algorithms, including Random Forest, XGBoost, and other ensemble learning approaches, in addressing socioeconomic classification problems such as poverty prediction and social assistance targeting. However, limited studies have specifically focused on regional social stability classification based on socioeconomic indicators. Therefore, this study extends previous research by applying the Random Forest algorithm to classify regional social stability in Banten Province [8], [9], [10], [11].

Through this research, it is expected that a classification model can be developed to assist local governments and

policymakers in evaluating regional social conditions more quickly and accurately. The classification results may also serve as a consideration in decision-making processes related to equitable development, improving public welfare, and controlling social inequality in Banten Province.

II. RESEARCH METHODOLOGY

A. Problem Identification

Banten Province is one of the regions with relatively high socioeconomic dynamics due to population growth, urbanization, regional development disparities, and differences in social conditions among communities. These conditions may affect the level of regional social stability, which plays an important role in supporting sustainable regional development. However, the analysis of regional social stability is still commonly conducted using conventional approaches, resulting in findings that are not yet optimal in objectively and rapidly describing regional social conditions.

Based on these conditions, this study identifies several problems related to the classification of regional social stability. Research on regional social stability classification based on socioeconomic indicators still requires further development in order to produce more objective and measurable classifications. In addition, the large number of socioeconomic indicators used has complex relationships, requiring classification methods capable of handling multidimensional data effectively. Differences in socioeconomic characteristics among regencies and cities in Banten Province also result in varying levels of regional social stability, thereby requiring an analytical approach capable of describing these conditions more accurately. On the other hand, the classification method used is expected to have high accuracy and be capable of reducing the risk of overfitting during data processing. The classification results are expected to become a source of information that supports local governments in making more targeted social and economic development policies.

B. Research Method

The research method used in this study is a machine learning-based classification method using the Random Forest algorithm. The research was conducted systematically from data collection to classification result analysis.

C. Research Flowchart

As shown in Figure 1, the study was conducted through several systematic stages to ensure that the classification process of regional social stability could be carried out effectively. The research stages began with problem identification and ended with the analysis of classification results. The stages are described as follows.

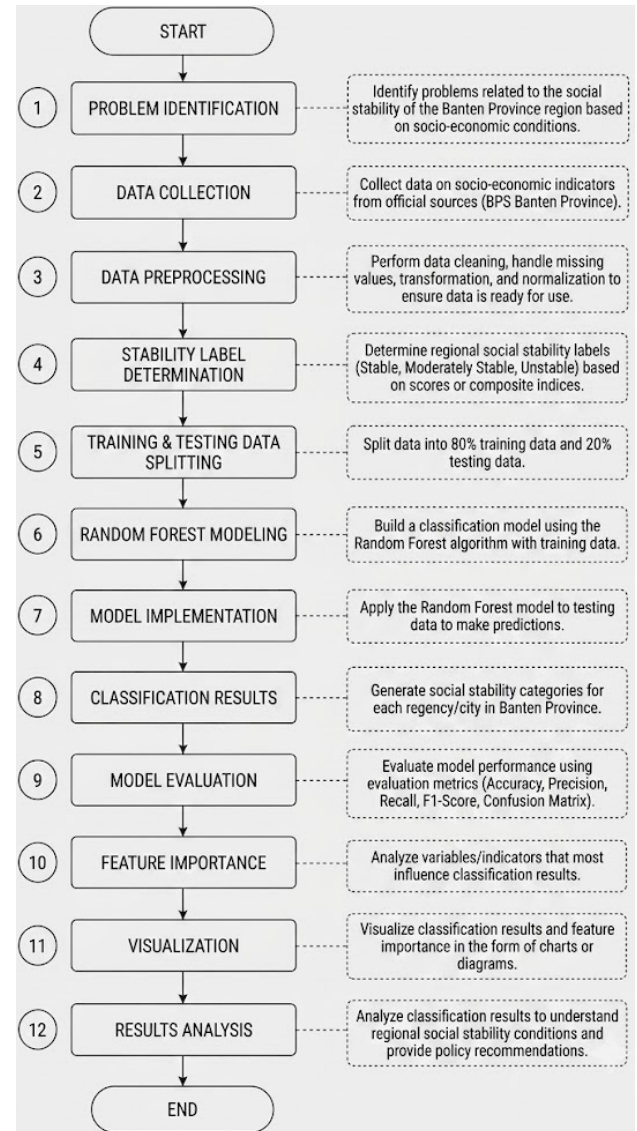


Figure 1. Illustrates The Research Flowchart

a) Problem Identification

The initial stage was carried out by identifying problems related to regional social stability in Banten Province based on the socioeconomic conditions of the community.

b) Data Collection

The data used in this study were collected from official sources, namely the Central Statistics Agency (BPS) of Banten Province. The data consisted of socioeconomic indicators considered to be related to regional social stability conditions. The indicators used included poverty rate [12], [13], unemployment rate [14], Human Development Index (HDI) [15], HDI percentage, Gini Ratio [16], [17], population density [16], and average years of schooling [18]. All of these indicators were used as research variables to support the classification process of regional social stability in Banten Province.

c) Data Preprocessing

The preprocessing stage was conducted to prepare the data before the classification process. At this stage, the data first underwent a missing value checking process to ensure that there were no incomplete or empty data. Furthermore, data cleaning was performed to remove inconsistent data and duplicates that could affect the analysis results. Afterward, data transformation was carried out to ensure that the data format was suitable for processing and classification requirements. The final stage was data normalization, which aimed to standardize the scale among variables so that the classification process could perform more optimally and produce better accuracy.

d) Stability Label Determination

At this stage, the process of determining regional social stability labels was carried out based on specific categories. Initially, the categories consisted of stable, moderately stable, and unstable to describe regional social conditions in general. However, during the research process, the target variable (Y) or classification output was simplified into only two categories, namely stable and unstable, in order to make the classification process more effective and easier to interpret. Label determination was conducted before the modeling process because Random Forest belongs to the supervised learning approach, which requires predefined target labels as references during the model training process [19]. This differs from unsupervised learning methods that do not use target labels in the clustering process. The predetermined labels were subsequently used as classification targets so that the model could learn data patterns and predict regional social stability levels in a more systematic and measurable manner.

e) Training and Testing Data Split

At this stage, the research data were divided into two parts, namely 80% as training data and 20% as testing data [20], [21], [22]. The training data were used to train the Random Forest model so that it could learn patterns and relationships among variables based on the available data. Meanwhile, the testing data were used to evaluate the model performance in classifying previously unseen data. This data splitting process was conducted to ensure that the developed model possessed good predictive capability and could reduce the risk of overfitting during the classification process.

f) Random Forest Modeling

The modeling stage was conducted by developing a classification model using the Random Forest algorithm based on the previously prepared training data. Random Forest is a supervised learning-based classification method that works by constructing several decision trees from the training data and combining the prediction results of each decision tree to determine the final classification result [23]. The basic concept of the Random Forest algorithm can be expressed as Formula 1:

$$f(x) = \text{mode}\{T_1(x), T_2(x), \dots, T_n(x)\} \quad (1)$$

In this equation, $T_1(x), T_2(x), \dots, T_n(x)$ represent a collection of decision trees constructed during the model training process. The final classification result is obtained through a majority voting mechanism, where the class predicted most frequently by all decision trees is selected as the final classification result. This approach enables Random Forest to improve classification accuracy and reduce the risk of overfitting compared to using a single decision tree [24].

g) Model Implementation

After the Random Forest model was successfully developed using the training data, the next stage was implementing the model on the testing data [25]. At this stage, the model was used to predict previously unseen data in order to determine its capability in classifying regional social stability levels. The implementation process was conducted by inputting socioeconomic indicator variables into the trained model, after which the model generated classification outputs in the form of stable or unstable categories. This stage aimed to determine the extent to which the model could accurately recognize data patterns and ensure that the model could be applied objectively and measurably in regional social stability analysis.

h) Classification Results

At the classification result stage, the Random Forest model generated predictions of social stability categories for each region based on the previously processed socioeconomic indicator data. The classification results consisted of grouping regions into stable or unstable categories according to the patterns learned by the model during the training process. This stage provided a more measurable and data-driven overview of social stability conditions in each regency and city in Banten Province. In addition, the classification results could also be used as supporting information for understanding differences in socioeconomic conditions among regions and as a basis for evaluation and regional development policymaking.

i) Model Evaluation

The model evaluation stage was conducted to determine the performance of the Random Forest model in classifying regional social stability. The evaluation used several metrics, namely accuracy, precision, recall, F1-score, and confusion matrix. Accuracy was used to measure the overall prediction correctness of the model, precision to evaluate the accuracy of prediction results, recall to determine the model capability in recognizing relevant data, and F1-score to balance precision and recall values. In addition, the confusion matrix was used to compare prediction results with actual data so that classification errors could be identified more clearly. The accuracy formula used is as Formula 2:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (2)$$

Description:

- TP (True Positive) = positive data correctly predicted as positive

- TN (True Negative) = negative data correctly predicted as negative
- FP (False Positive) = negative data incorrectly predicted as positive
- FN (False Negative) = positive data incorrectly predicted as negative

Through this evaluation stage, the accuracy level and capability of the model in classifying regional social stability can be determined.

j) Feature Importance

The feature importance stage was conducted to determine which variables had the greatest influence on the classification results of regional social stability in the Random Forest model. Through this stage, each socioeconomic indicator was analyzed according to its level of influence in the classification process. The feature importance results could indicate which variables contributed the most to determining stable or unstable categories. Therefore, this stage helped in understanding the factors that most significantly affect regional social stability in a clearer and more measurable manner.

k) Visualization

The visualization stage was conducted to present classification results and feature importance in the form of graphs or diagrams so that the obtained information could be more easily understood and analyzed. Visualization was used to display comparisons of classification results among regions and to show the influence level of each variable on the Random Forest model. Through visualization, the research results could be presented more clearly, informatively, and interpretably.

l) Result Analysis

The result analysis stage represented the final stage of the study, conducted to interpret the classification results of regional social stability in Banten Province. At this stage, the Random Forest prediction results were analyzed to identify social stability conditions in each region based on the socioeconomic indicators used. In addition, analysis was also conducted on the variables that had the greatest influence in the classification process so that the factors affecting regional social stability could be identified. The analysis results were subsequently used as a basis for providing recommendations that could support more targeted social and economic development policymaking.

D. Tools Used

This study used Microsoft Excel for initial data processing and preparation, and Python as the programming language for data classification. Model implementation was carried out using Google Colab as a cloud-based programming platform. In addition, this study utilized several Python libraries, namely Pandas (pandas) for reading, managing, and manipulating research data. The study also used Scikit-learn (sklearn) for machine learning and model evaluation processes, including `train_test_split` for dividing training and testing data, `RandomForestClassifier` for developing the Random Forest

model, `accuracy_score` for calculating accuracy values, `classification_report` for displaying precision, recall, and F1-score values, and `confusion_matrix` for displaying the classification confusion matrix.

III. RESULTS AND DISCUSSION

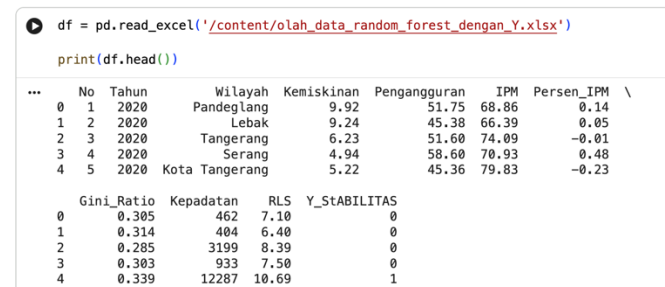
This chapter explains the data processing stages and the implementation of the Random Forest algorithm for classifying regional social stability in Banten Province based on socioeconomic indicators. The research process was carried out systematically, starting from data collection, preprocessing, model development, and evaluation of classification results using the Python programming language.

Data processing was conducted using several Python libraries such as Pandas for data management and Scikit-learn for machine learning processes. The research stages were arranged systematically to ensure that the resulting classification model could provide accurate prediction results and be used as a reference for analyzing regional social stability conditions.

The results presented in this chapter include data processing, Random Forest model implementation, model performance evaluation, feature importance analysis, and visualization of regional social stability classification results in Banten Province. The following sections describe the data processing stages and outputs generated using Python tools and libraries in Google Colab.

A. Data Collection

At this stage, the research data were collected in Microsoft Excel (.xlsx) format and subsequently processed using the Python programming language with the Pandas library (Figure 2). The dataset consisted of several variables related to regional social stability. The implemented code was used to read the dataset and display preliminary data before further processing was conducted.



```
df = pd.read_excel('/content/olah_data_random_forest_dengan_Y.xlsx')
print(df.head())
```

No	Tahun	Wilayah	Kemiskinan	Pengangguran	IPM	Persen_IPM	\
0	1	2020	Pandeglang	9.92	51.75	68.86	0.14
1	2	2020	Lebak	9.24	45.38	66.39	0.05
2	3	2020	Tangerang	6.23	51.60	74.09	-0.01
3	4	2020	Serang	4.94	58.60	70.93	0.48
4	5	2020	Kota Tangerang	5.22	45.36	79.83	-0.23

Figure 2. Dataset Import Process Using Python

Table 2 present the Excel dataset used in this study contained several variables that served as parameters in the classification process of regional social stability using the Random Forest method.

Table 2. Regional Social Stability Research Dataset

No	Year	Region	Poverty (%)	Unemployment (%)	HDI	Percent_HDI	Gini Ratio	Density (pop/km ²)	MYS	Y_STABILITY
1	2020	Pandeglang	9.92	51.75	68.86	0.14		462	7.10	0
2	2020	Lebak	9.24	45.38	66.39	0.05	0.314	404	6.40	0
3	2020	Tangerang	6.23	51.60	74.09	-0.01	0.285	3199	8.39	0
4	2020	Serang	4.94	58.60	70.93	0.48	0.303	933	7.50	0
...	...	...	...	...	...	...	...	...	...	...
45	2025	Tangerang City	519	39.38	82.41	1.08	0.295	11055	11.46	1
46	2025	Cilegon City	3.44	48.80	79.54	0.90	0.329	2832	10.40	1
47	2025	Serang City	5.51	52.24	77.50	0.78	0.329	2800	8.94	0
48	2025	South Tangerang City	2.39	47.75	84.81	0.77	0.375	8508	12.10	1

B. Data Preprocessing

During the preprocessing stage (Figure 3), data inspection and cleaning processes were conducted before the data were used in the Random Forest modeling stage. This stage included checking for missing values, selecting research variables, and adjusting data formats to ensure compatibility with Python processing. Data preprocessing was carried out to ensure that the dataset possessed good quality, thereby improving the classification performance of the model.

```
print(df.info())
print(df.head())
```

<class 'pandas.core.frame.DataFrame'>				
RangeIndex: 48 entries, 0 to 47				
Data columns (total 11 columns):				
#	Column	Non-Null Count	Dtype	
0	No	48 non-null	int64	
1	Tahun	48 non-null	int64	
2	Wilayah	48 non-null	object	
3	Kemiskinan	48 non-null	float64	
4	Pengangguran	48 non-null	float64	
5	IPM	48 non-null	float64	
6	Persen_IPM	48 non-null	float64	
7	Gini_Ratio	48 non-null	float64	
8	Kepadatan	48 non-null	int64	
9	RLS	48 non-null	float64	
10	Y_STABILITAS	48 non-null	int64	

dtypes: float64(6), int64(4), object(1)
memory usage: 4.3+ KB

No	Tahun	Wilayah	Kemiskinan	Pengangguran	IPM	Persen_IPM	Y
0	1	2020	Pandeglang	9.92	51.75	68.86	0.14
1	2	2020	Lebak	9.24	45.38	66.39	0.05
2	3	2020	Tangerang	6.23	51.60	74.09	-0.01
3	4	2020	Serang	4.94	58.60	70.93	0.48
4	5	2020	Kota Tangerang	5.22	45.36	79.83	-0.23

Gini_Ratio	Kepadatan	RLS	Y_STABILITAS
0	0.305	462	7.10
1	0.314	404	6.40
2	0.285	3199	8.39
3	0.303	933	7.50
4	0.339	12287	10.69

Figure 3. Data Preprocessing Using Python

C. Social Stability Label Determination

At this stage, classification labels for social stability were determined as the target variable in the Random Forest modeling process. Social stability labels were divided into two categories, namely stable and unstable, represented in binary values. A value of 1 indicated a stable region, while a value of 0 indicated an unstable region. Label determination was based on the socioeconomic indicator conditions of each research region.

D. Training and Testing Data Split

At this stage, the dataset was divided into training data and testing data using the `train_test_split` method. Training data

were used to train the Random Forest model, while testing data were used to evaluate the classification performance of the model. In this study, the data were divided using an 80% training data and 20% testing data ratio.

```
X = df[['Kemiskinan',
        'Pengangguran',
        'IPM',
        'Persen_IPM',
        'Gini_Ratio',
        'Kepadatan',
        'RLS']]

y = df['Y_STABILITAS']

X_train, X_test, y_train, y_test = train_test_split(
    X,
    y,
    test_size=0.2,
    random_state=42)

rf = RandomForestClassifier(
    n_estimators=100,
    random_state=42)
```

Figure 4. Training and Testing Data

Figure 4 illustrates the distribution of data into training and testing datasets. The 80:20 split ratio was selected to provide sufficient data for model learning while maintaining an independent dataset for performance evaluation. This approach is commonly used in machine learning studies to assess model generalization capability. By separating unseen testing data from the training process, the evaluation results can better represent the model's predictive performance on new data.

E. Random Forest Modeling

At this stage, the classification model was developed using the Random Forest algorithm. The model was trained using the training data that had been divided in the previous stage.

```
rf.fit(X_train, y_train)

RandomForestClassifier
RandomForestClassifier(random_state=42)

y_pred = rf.predict(X_test)

print(y_pred)

[0 0 0 0 0 1 0 0 1 0]

accuracy = accuracy_score(y_test, y_pred)

print("Akurasi:", accuracy)

... Akurasi: 1.0
```

Figure 5. Random Forest Modeling

Figure 5 presents the implementation of the Random Forest model used in this study. The algorithm constructs multiple decision trees from the training dataset and combines their predictions through a majority voting mechanism. This ensemble learning approach helps improve classification performance and reduces the likelihood of overfitting compared to a single decision tree model. The model was trained using socioeconomic indicators as predictor variables and social stability labels as target classes.

F. Model Implementation

During the implementation stage, predictions were performed using the previously developed Random Forest model (Figure 6). Testing data were used to determine the classification results of social stability levels based on the variables used in this study.

```
<> Python

y_pred = rf.predict(X_test)
```

Figure 6. Model Implementation

G. Classification Results

The prediction results were subsequently displayed to compare the actual data and the classification results generated by the model.

```
<> Python

hasil_prediksi = pd.DataFrame({
    'Aktual': y_test,
    'Prediksi': y_pred
})

print(hasil_prediksi.head())
```

Figure 7. Model Classification Results

Figure 7 displays the classification results generated by the Random Forest model. The output compares the actual class labels with the predicted labels produced by the model. The close agreement between actual and predicted values indicates that the model successfully learned the patterns contained within the socioeconomic indicators. This result demonstrates the effectiveness of Random Forest in distinguishing between stable and unstable regional social conditions.

H. Model Evaluation

At this stage, the Random Forest model was evaluated to determine the classification performance achieved. Model evaluation was conducted using several testing metrics, including accuracy, precision, recall, and F1-score. In addition, a confusion matrix was used to analyze the prediction results against the actual data.

```
print(classification_report(y_test, y_pred))
```

	precision	recall	f1-score	support
0	1.00	1.00	1.00	8
1	1.00	1.00	1.00	2
accuracy			1.00	10
macro avg	1.00	1.00	1.00	10
weighted avg	1.00	1.00	1.00	10

```
print(confusion_matrix(y_test, y_pred))

[[8 0]
 [0 2]]
```

Figure 8. Model Evaluation

Figure 8 presents the evaluation results of the Random Forest model using accuracy, precision, recall, F1-score, and confusion matrix metrics. The model achieved perfect classification performance, with all evaluation metrics reaching 1.00. The confusion matrix indicates that all testing instances were correctly classified without any misclassification errors. Although these results demonstrate excellent predictive capability, they should be interpreted carefully due to the relatively limited dataset size, which may influence model generalization performance.

The evaluation results indicated that the Random Forest model achieved excellent classification performance. Based on the classification report, the precision, recall, and F1-score values for each class reached 1.00. Furthermore, the model achieved an accuracy value of 1.00 or 100%, indicating that all testing data were correctly classified by the model.

The confusion matrix results showed that there were 8 data instances in class 0 and 2 data instances in class 1 that were correctly classified without any prediction errors. Therefore, the Random Forest model was considered capable of classifying regional social stability levels very effectively based on the research variables used in this study.

Although the Random Forest model achieved perfect classification performance, the result should be interpreted carefully due to the relatively small dataset size. The high accuracy may indicate strong separability among the selected socioeconomic indicators. However, there is also a possibility

of overfitting. Therefore, future studies should employ larger datasets and additional validation techniques such as k-fold cross-validation to further evaluate model generalization performance.

I. Feature Importance

At this stage, feature importance analysis was conducted to determine the influence level of each variable on the classification results of regional social stability. Feature importance values were obtained from the Random Forest model based on the contribution of each variable during the decision tree construction process.

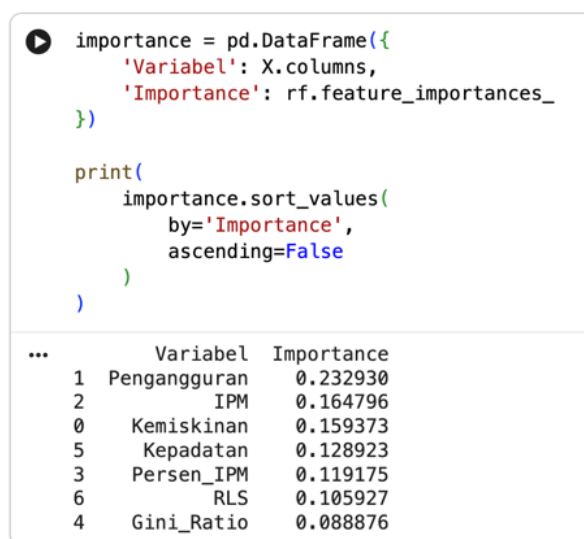


Figure 9. Feature Importance Analysis

Figure 9 illustrates the relative importance of each socioeconomic indicator in determining regional social stability. The unemployment variable achieved the highest importance score, indicating that employment conditions play a dominant role in social stability classification. HDI and poverty also contributed substantially to the model's decision-making process. In contrast, the Gini Ratio showed the lowest contribution, suggesting a relatively weaker influence within the current dataset.

The results indicated that variables with higher importance values had greater influence on the classification results of regional social stability in the Random Forest model. Based on the testing results, the unemployment variable had the highest importance value of 0.232930, making it the most influential variable in the classification process. This was followed by the Human Development Index (HDI) variable with an importance value of 0.164796 and the poverty variable with an importance value of 0.159373. Meanwhile, the Gini Ratio variable had the lowest influence with an importance value of 0.088876. These findings indicate that the unemployment level contributed the most to determining regional social stability in the Random Forest classification model.

The determination of the dependent variable (Y) was conducted using a scoring approach based on socio-economic indicators, namely poverty rate, unemployment rate, Human

Development Index (HDI), Gini Ratio, population density, and mean years of schooling (MYS). Each indicator was classified into positive or negative conditions according to its relationship with regional stability.

Variables that have a positive relationship with regional stability, such as HDI and MYS, were assigned a score of 1 when their values were above the median of the dataset. Conversely, variables that have a negative relationship with regional stability, such as poverty and unemployment rates, were assigned a score of 1 when their values were below the dataset median.

The scores from all indicators were then summed. Regions obtaining a minimum score of three were categorized as stable regions (Y = 1), while regions with scores below three were classified as unstable regions (Y = 0). This approach was employed to generate classification labels, which were subsequently used as the target variable in the Random Forest model.

The feature importance analysis of the Random Forest model revealed that the unemployment rate had the highest importance value, at 0.2329. This finding indicates that unemployment is the most influential variable in classifying regions as stable or unstable. High unemployment levels can reduce community welfare, increase social vulnerability, and intensify economic and social challenges, which may ultimately affect regional stability.

The next most influential variables were the Human Development Index (HDI), with an importance value of 0.1648, and the poverty rate, with an importance value of 0.1594. A higher HDI reflects better human resource quality through improvements in education, health, and living standards. In contrast, a high poverty rate may indicate lower levels of societal welfare and can contribute to various social problems. Therefore, both variables play a significant role in distinguishing between stable and unstable regions.

Meanwhile, the Gini Ratio exhibited the lowest importance value, at 0.0889. This result suggests that income inequality still contributes to the classification process; however, its influence is relatively smaller compared to other socioeconomic variables included in the study. This finding indicates that differences in regional stability among the observed districts and municipalities are more strongly influenced by labor market conditions, human development, and poverty levels than by income inequality.

Overall, the results demonstrate that employment conditions, human development quality, and community welfare are the most dominant factors shaping regional stability characteristics. These findings imply that policies aimed at expanding employment opportunities, improving human resource quality, and reducing poverty may contribute more substantially to fostering stable regional conditions.

J. Result Visualization

The Feature importance visualization was used to illustrate the contribution level of each variable in the classification process of regional social stability within the Random Forest model. The greater the importance value displayed in the

graph, the greater the influence of the variable on the model classification results.

Based on the visualization graph, the unemployment variable showed the highest influence compared to the other variables, followed by HDI and poverty. Meanwhile, the Gini Ratio variable had the lowest influence in the regional social stability classification process.

```
import matplotlib.pyplot as plt

importance = importance.sort_values(
    by='Importance',
    ascending=True
)

plt.figure(figsize=(8,5))
plt.barh(
    importance['Variabel'],
    importance['Importance']
)

plt.xlabel('Importance')
plt.ylabel('Variabel')
plt.title('Feature Importance Random Forest')

plt.show()
```

Figure 10. Result Visualization

A horizontal bar chart was used to make comparisons among variable importance values clearer and easier to analyze.

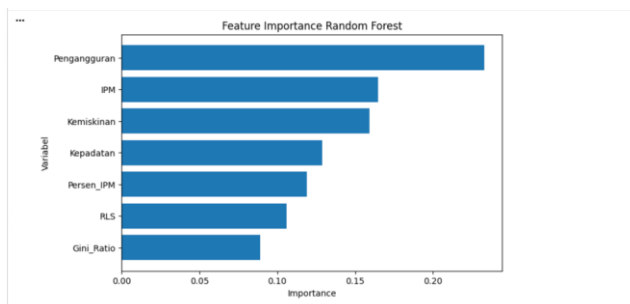


Figure 11. Random Forest Feature Importance

Figures 10 and 11 provide visual representations of feature importance values obtained from the Random Forest model. The graphical visualization facilitates interpretation of the relative contribution of each variable. The dominance of the unemployment variable is clearly visible, highlighting its significant role in differentiating stable and unstable regions. Such visual analysis supports the quantitative findings presented in the feature importance results.

K. Result Analysis

The implementation results of the Random Forest model indicated that the classification process used several social indicator variables, namely poverty rate, unemployment rate, Human Development Index (HDI), HDI growth percentage, Gini Ratio, population density, and average years of schooling

(RLS). These variables were used as input data in predicting regional social stability levels.

Based on the prediction results obtained, the Random Forest model generated classification outputs in the prediction column as shown in Figure 12. A value of 0 represented the unstable category, while a value of 1 represented the stable category. These prediction results were used to identify regional social stability conditions based on the social indicators used in this study.

```
hasil = X_test.copy()
hasil['Prediksi'] = y_pred
print(hasil)
```

	Kemiskinan	Pengangguran	IPM	Persen_IPM	Gini_Ratio	Kepadatan	RLS	\
27	4.85	56.92	72.63	0.89	0.274	1144	7.79	
40	8.51	48.19	71.84	1.35	0.227	483	7.50	
26	6.93	55.60	75.56	0.55	0.286	3272	8.93	
43	4.48	62.94	74.10	1.12	0.252	1169	8.03	
24	9.27	47.63	70.28	0.83	0.243	474	11.84	
37	3.75	39.36	78.83	0.75	0.360	2804	10.39	
12	5.93	52.49	80.08	0.31	0.343	12421	7.51	
19	4.96	59.59	71.99	1.39	0.284	1130	8.92	
4	5.22	45.36	79.83	-0.23	0.339	12287	10.69	
25	8.80	51.61	7.68	0.73	0.271	433	6.60	
Prediksi								
27								0
40								0
26								0
43								0
24								0
37								1
12								0
19								0
4								1
25								0

Figure 12. Testing Data Classification Output

Furthermore, the classification results were exported into an Excel file to simplify the documentation and analysis process of the model prediction results as shown in Figure 13 and 14.

```
hasil.to_excel(
    'hasil_prediksi.xlsx',
    index=False
)
```

Figure 13. Prediction Result Export Implementation Using Python

Python

hasil_prediksi.xlsx

Figure 14. Prediction Result File in Excel Format

During the model implementation stage, the Random Forest classification predictions were displayed based on the testing data used. The prediction outputs showed the classification results of regional social stability levels into stable (1) and unstable (0) categories based on the research variables used. The prediction results were then saved into an Excel file to facilitate further analysis and documentation of the classification results.

Table 3. Testing Data Prediction Results Using Random Forest

No	Year	Region	Poverty (%)	Unemployment (%)	HDI	Percent_HDI	Gini Ratio	Density (pop/km ²)	MYS	Y_STABILITY
1	2020	Pandeglang	0.44	51.75	68.86	0.01	0.212	462	0.29	0
2	2020	Lebak	0.39	1.90	2.78	0.00	0.218	404	0.25	0
3	2020	Tangerang	0.27	2.13	3.09	-0.01	0.198	3199	0.36	0
4	2020	Serang	0.23	2.42	70.93	0.03	0.210	933	0.30	0
...	...	...	...	...	...	...	...	...	...	...
45	2025	Tangerang City	0.22	1.65	3.45	0.05	0.205	11055	0.49	1
46	2025	Cilegon City	0.16	2.01	3.33	0.01	0.228	2832	0.42	1
47	2025	Serang City	0.24	2.18	3.21	0.05	0.228	2800	0.40	0
48	2025	South Tangerang City	0.11	47.75	84.81	0.05	0.260	8508	0.50	1

Based on the testing results using the Random Forest method (Table 3), the model was able to classify regional social stability levels very effectively. This was demonstrated by an accuracy value of 100% and precision, recall, and F1-score values reaching 1.00 for each classification class.

The feature importance results indicated that unemployment was the most influential factor in determining regional social stability, followed by HDI and poverty variables. Meanwhile, the Gini Ratio variable had the lowest influence on the model classification results.

The visualization results also clearly illustrated differences in the contribution levels among variables, thereby helping to understand the factors influencing regional social stability. Therefore, the Random Forest method was considered effective as a classification model for analyzing regional social stability based on the social indicators used in this study.

IV. CONCLUSION

Based on the results of this study, the Random Forest method was successfully implemented to classify regional social stability levels based on variables including poverty, unemployment, Human Development Index (HDI), HDI percentage, Gini Ratio, population density, and average years of schooling (RLS). The model evaluation results showed an accuracy value of 100%, with precision, recall, and F1-score values of 1.00, indicating that the model was able to classify the testing data very effectively.

Based on the feature importance analysis, the unemployment variable was identified as the most influential factor in determining regional social stability, followed by the HDI and poverty variables, while the Gini Ratio variable had the lowest influence on the classification results. The Random Forest method demonstrated advantages in producing high classification accuracy and identifying the influence level of each variable on the classification results.

However, this study still has limitations regarding the amount of dataset used. Therefore, future studies are recommended to use larger datasets, include additional research variables, and compare Random Forest with other classification methods in order to obtain more comprehensive and optimal analysis results.

REFERENCES

- [1] H. Khasim, "OJK: Utang Pinjol Masyarakat Indonesia Capai Rp. 101 Triliun," Carapandang. Accessed: May 11, 2026. [Online]. Available: <https://carapandang.com/news/read/ojk-utang-pinjol-masyarakat-indonesia-capai-rp-101-triliun>
- [2] Mita Trianda and Triase Triase, "Implementation of Random Forest Algorithm for Classification of Eligibility For Social Assistance Recipients In Information Systems," *Jurnal E-Komtek (Elektro-Komputer-Teknik)*, vol. 8, no. 2, pp. 275–286, Dec. 2024, doi: 10.37339/e-komtek.v8i2.2197.
- [3] P. Dubath *et al.*, "Random forest automated supervised classification of Hipparcos periodic variable stars," *ArXiv*, Jul. 2011, doi: 10.1111/j.1365-2966.2011.18575.x.
- [4] M. P. Yunizar, L. J. Sinay, and Y. Yudistira, "Performance Analysis Of Random Forest Classification On Unemployment Rate In Maluku Province Based On Data Balancing Method," *Journal of Statistics and Its Applications*, vol. 7, no. 1, Apr. 2025.
- [5] I. Setiawan *et al.*, "Utilizing Random Forest Algorithm for Sentiment Prediction Based on Twitter Data," in *Proceedings of the First Mandalika International Multi-Conference on Science and Engineering 2022, MIMSE 2022 (Informatics and Computer Science)*, Atlantis Press International BV, 2022, pp. 446–456. doi: 10.2991/978-94-6463-084-8_37.
- [6] F. Ahludzikri, R. Herwanto, R. Abdul Aziz, I. Agus, and S. Yusuf Irianto, "Web-Based Research Article Classification Using The Random Forest Algorithm," *Jurnal Ilmu Siber dan Teknologi Digital*, vol. 4, no. 1, pp. 15–31, Nov. 2025, doi: 10.35912/isted.v4i1.5547.
- [7] L. Malem Ginting, M. M. Sigirow, G. Yola Lumbantoruan, and dan Januard Lumbangaol, "Prediksi Indikator Perawatan Menggunakan Random Forest Classification dan Classification and Regression Trees," *Journal of Applied Technology and Informatics*, vol. 1, no. 7, pp. 43–48, Nov. 2021, [Online]. Available: <http://journal-jati.del.ac.id/>
- [8] 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, Mar. 2022, doi: 10.1016/j.jag.2022.102694.
- [9] G. Muñetón-Santa and L. C. Manrique-Ruiz, "Predicting Multidimensional Poverty with Machine Learning Algorithms: An Open Data Source Approach Using Spatial Data," *Soc. Sci.*, vol. 12, no. 5, May 2023, doi: 10.3390/socsci12050296.
- [10] M. Ahmad, H. Prabowo, H. L. H. S. Warnars, and F. L. Gaol, "A Machine Learning Approach for Model Selection of Social Aid Beneficiaries," *Journal of System and Management Sciences*, vol. 13, no. 6, pp. 230–243, 2023, doi: 10.33168/JSMS.2023.0614.
- [11] M. H. A. Zamzuri, N. Sofian, and R. Hassan, "Forecasting of Poverty using the Ensemble Learning Classification Methods," *International Journal on Perceptive and Cognitive Computing*, vol. 9, no. 1, pp. 24–32, Jan. 2023, doi: 10.31436/ijpcc.v9i1.326.
- [12] Badan Pusat Statistik Provinsi Banten, "Jumlah Penduduk Miskin Provinsi Banten (Maret 2025)," BPS Provinsi Banten. Accessed: May 14, 2026. [Online]. Available: <https://banten.bps.go.id/id/statistics-table/2/Mzc4IzI=/jumlah-penduduk-miskin-provinsi-banten--ribu-orang---maret-2025.html>
- [13] Badan Pusat Statistik Provinsi Banten, "Persentase Penduduk Miskin Provinsi Banten, Maret 2025," BPS Provinsi Banten. Accessed: May 14, 2026. [Online]. Available: <https://banten.bps.go.id/id/statistics-table/2/NzgjMg=/persentase-penduduk-miskin-provinsi-banten--maret-2025.html>
- [14] Badan Pusat Statistik Provinsi Banten, "Statistik Sosial Provinsi Banten," BPS Provinsi Banten. Accessed: May 14, 2026. [Online]. Available: <https://banten.bps.go.id/id/statistics-table?subject=519&sortBy=date%2Ctitle&sortOrder=desc%2Casc&utm>
- [15] Badan Pusat Statistik Provinsi Banten, "Kepadatan Penduduk Banten 2026," BPS Provinsi Banten. Accessed: May 14, 2026. [Online]. Available: <https://banten.bps.go.id/id/statistics-table/2/MTA5IzI=/kepadatan-penduduk-banten--2026.html>
- [16] Badan Pusat Statistik Provinsi Banten, "Gini Ratio Provinsi Banten, Maret 2025," BPS Provinsi Banten. Accessed: May 14, 2026. [Online]. Available: <https://banten.bps.go.id/id/statistics-table/2/MzMwIzI=/gini-ratio-provinsi-banten--maret-2025.html>
- [17] Badan Pusat Statistik Provinsi Banten, "Rata-rata Lama Sekolah di Banten 2025," BPS Provinsi Banten. Accessed: May 14, 2026. [Online]. Available: <https://banten.bps.go.id/id/statistics-table/2/NzYjMg=/rata-rata-lama-sekolah-di-banten--2025.html>
- [18] Badan Pusat Statistik Provinsi Banten, "PDRB ADHB TW IV 2025," BPS Provinsi Banten. Accessed: May 14, 2026. [Online]. Available: <https://banten.bps.go.id/id/statistics-table/2/NTAwIzI=/pdrb-adhb--tw-iv-2025.html>
- [19] F. Kruber, J. Wurst, E. S. Morales, S. Chakraborty, and M. Botsch, "Unsupervised and Supervised Learning with the Random Forest Algorithm for Traffic Scenario Clustering and Classification," *ArXiv*, Apr. 2020, doi: 10.1109/IVS.2019.8813994.
- [20] S. Hatmara Sukma, N. Suarna, A. Bahtiar, P. Pramudya Marta, and K. Anam, "Model Prediksi Penerima Bantuan Sosial Berbasis Algoritma Random Forest," *Jurnal Software Engineering and Information System (SEIS)*, vol. 6, no. 1, pp. 45–49, Jan. 2026, doi: <https://doi.org/10.37859/seis.v5i1>.
- [21] M. J. Faisti, R. Hadapiningradja Kusumodestoni, G. Wahyu, and N. Wibowo, "Mental Health Classification Using Naïve Bayes and Random Forest Algorithms," 2025. [Online]. Available: <http://jurnal.polibatam.ac.id/index.php/JAIC>
- [22] H. Hang, X. Liu, and I. Steinwart, "Best-scored Random Forest Classification," *ArXiv*, May 2019, [Online]. Available: <http://arxiv.org/abs/1905.11028>
- [23] R. Genuer, J.-M. Poggi, and C. Tuleau, "Random Forests: some methodological insights," France, Nov. 2008. [Online]. Available: <http://arxiv.org/abs/0811.3619>
- [24] M. A. Tsaqif and W. Maharani, "Comparison of Random Forest and Decision Tree Methods for Emotion Classification based on Social Media Posts," *Building of Informatics, Technology and Science (BITS)*, vol. 6, no. 4, pp. 2240–2248, Mar. 2025, doi: 10.47065/bits.v6i4.6677.
- [25] P. Elisya, A. Sani, A. Hari Wibowo, M. Yamin, and M. Irwan Syahib, "Perbandingan Algoritma C5.0 Dan Random Forest Untuk Klasifikasi Calon Penerima Program Keluarga Harapan (PKH) (Studi Kasus: Desa Bakealu, Kecamatan Wakorumba Selatan, Kabupaten Muna)," *ANIMATOR*, vol. 3, no. 3, pp. 1–9, Dec. 2025.