JURNAL STATISTIKA DAN APLIKASINYA

e-ISSN 2620 – 8369
VOLUME 8 (ISSUE 2) DECEMBER 2024: 215-228
https://doi.org/10.21009/JSA.08208

COMPARISON OF MACHINE LEARNING CLASSIFICATION ALGORITHMS IN GROUPING INCOME DISTRIBUTION INEQUALITIES IN JAVA AND BALI

Qorinul Huda [1*], **Puput Budi Aji**[2]

[1]*Statistician, BPS-Statistics Tulang Bawang Barat Regency*
Jln. Tirta Makmur, Tirta Kencana, Tulang Bawang Tengah, Lampung, 34793, Indonesia

[2]*Statistician, BPS-Statistics Tanjung Jabung Barat Regency*
Jln. Prof. Dr. Soedewi, Pembangis, Kec. Bram Itam, Kabupaten Tanjung Jabung Barat, Jambi, 36514, Indonesia

Corresponding author's e-mail: * **211911012@stis.ac.id**

ABSTRACT

Article History:
Received: 10, December 2024
Revised: 17, December 2024
Accepted: 27, December 2024
Published: 31, December 2024
Available online.

Keywords:
Inequality Income, decision tree, logistic classification, random forest, machine learning.

Inequality is a growing issue in several countries, both developed and developing countries. The level of state income reflected in Gross Domestic Product (GDP) cannot yet describe whether income allocation is equitable or not. High GDP is the goal of a country, but welfare is much more important. Community welfare in a country can be interpreted as how much state income is enjoyed by the community. One benchmark for whether a country's income is equally enjoyed by its people or not is through the Gini index.

As industry 4.0 progresses, economic growth continues to increase. The largest share of Indonesia's GDP is on the islands of Java and Bali. Behind the rapid economic growth on the two islands, there is also inequality in income distribution. This research aims to classify districts and cities on the islands of Java and Bali based on factors that influence inequality using a data mining classification algorithm. The use of machine learning is essential in this research due to its ability to analyze complex and multidimensional data, such as social, economic, and geographical factors that influence inequality. Machine learning algorithms can identify patterns accurately and efficiently, surpassing the limitations of traditional methods like regression, which are often suboptimal for large and unstructured datasets.

This research uses four algorithms, namely Decision Tree, Logistic Classification, Random Forest, and Support Vector Machine (SVM). These four methods will be compared (compared) based on model evaluation, so that they are able to predict testing data for the next period to produce the correct regional classification. This research also accommodates handling of imbalanced data, data imputation, and forecasting using Generalized Regression Neural Network (GRNN). In general, the research findings show that the logistic regression algorithm performs well in classifying income distribution inequality in Java and Bali.

How to cite this article:
Q. Huda, P.B. Aji, "COMPARISON OF MACHINE LEARNING CLASSIFICATION ALGORITHMS IN GROUPING INCOME DISTRIBUTION INEQUALITIES IN JAVA AND BALI", Jurnal Statistika dan Aplikasinya, vol. 8, iss. 2, pp. 215 – 228, December 2024

Journal e-mail: *jsa@unj.ac.id*
Research Article · Open Access

1. INTRODUCTION

Economic development reflects the welfare of the people in the region (Cristea et al., 2021). The main goal in economic development is the highest possible economic growth. Economic growth does not only focus on the national scale, but also on regional scales such as provinces and districts. Development should be viewed as a multidimensional process that encompasses all dimensions including fundamental changes in social structures, attitudes, and national institutions. Development is concerned with acceleration in economic growth, inequality in income distribution, and poverty alleviation (Todaro & Smith, 2020). Undang-undang Nomor 22 Tahun 1999 is the basis for the implementation of regional autonomy in Indonesia, which aims to achieve economic output according to the potential of the region. This regional autonomy makes each region try to increase the scale of its economic output to spur regional income.

Indonesia's development plan for 2020-2024 places income distribution inequality as the main strategic issue that must be resolved (BAPPENAS, 2020). Development inequality is the main problem of economic growth in a region (Islami, 2018). There are several measures used to measure economic inequality, one of which is the Gini index (Karsu & Morton, 2015). Todaro & Smith (2020) explain that the Gini index is a measure of the inequality of income distribution in a region with a value range of 0 (perfect equality) to 1 (perfect inequality).

Inequality in income distribution tends to be large in regions with high economies. According to Figure 1, the largest portion of income contribution is in Java at 58.88 percent, Sumatera at 21.53 percent. While the island of Bali and Nusa Tenggara amounted to 2.92 percent. As a study of the design of the largest economic development, BAPPEDA prepared regional development planning documents for 2023 to 2026 in the Java and Bali regions (BAPPEDA Kulon Progo, 2022).

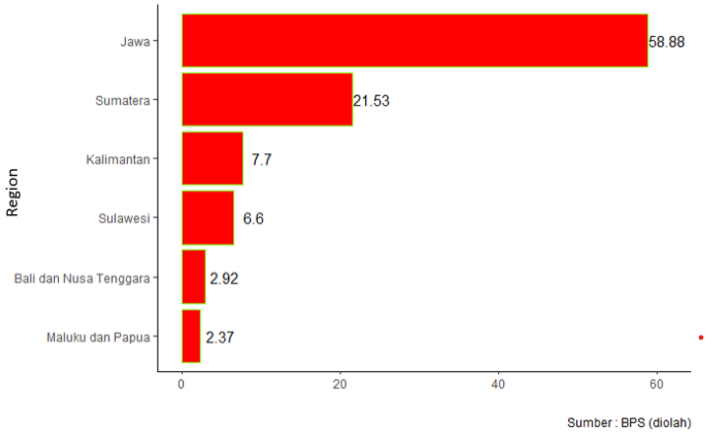

Figure 1. GRDP contribution by region in the third quarter of 2020

With the importance of such planning, accurate classification is needed as an evaluation material and guideline for the implementation of development programs in the area. This can be done using a classification algorithm. This classification will map regions with income distribution inequality. Thus, the best classification algorithm will determine the classification of the region and become a policy suggestion. In data mining algorithms, classification is classified as supervised learning. Supervised is a "control" of the algorithm that greatly influences model building (Purnama, 2019).

Based on the background above, the objectives of this study can be described as follows:

1. Handle the condition of missing data (missing value) and the balance of the classification of the gini ratio of districts and cities in Java Bali in 2020.

2. Classify regencies and cities in Java Bali in 2020 using the Decision Tree, Logistic, Random Forest, and SVM algorithms.
3. Comparing the Decision Tree, Logistic, Random Forest, and SVM algorithms in classifying regencies and cities in Java Bali 2020. So that the best method can be produced for future classification periods.

2. METHODS

Materials dan Data

The data used is secondary data collected from Statistics Indonesia. The dataset consists of 7 variables, namely the Gini index, Gross Regional Domestic Product (GRDP), Average Years of Schooling (AYS), Human Development Index (IPM), Percentage of Poor Population (Po), Life Expectancy Rate (AHH), and Open Unemployment Rate (TPT). Table 1 shows the operational definition of each variable:

Table 1. Operational definition of variables

Variable	Definition
Gini Index	The gini index or gini ratio is an indicator that shows the overall level of expenditure inequality. It ranges from 0 to 1. A gini ratio value closer to 1 indicates higher inequality. This variable is the basis for the classification of inequality into low, medium, and high with the provisions (Todaro & Smith, 2020).
PDRB	PDRB is the aggregate value of all goods and services produced in an area over a period of time (usually one year) that can be calculated through three approaches: the production method, the expenditure method, and the income method (Badan Pusat Statistik, 2024d).
AYS	Average Years of Schooling (AYS) is defined as the number of years spent by the population in formal education. AYS can be used to determine the quality of education of people in an area (Badan Pusat Statistik, 2023).
IPM	IPM is a composite index that measures human development from three basic aspects of a decent standard of living. IPM values range from 0 to 100. The IPM figure provides a comprehensive picture of human development achievements in a region (Badan Pusat Statistik, 2024b).
Po	The percentage of poor people (Po) is the percentage who are below the poverty line. Where Po is calculated from (Badan Pusat Statistik, 2024a): $P_0 = \frac{1}{n}\sum_{i=1}^{q}\left[\frac{z-y_i}{z}\right]^0$ (1) Where z : poverty line, y_i : average monthly per capita expenditure of people below the poverty line, q : number of people below the poverty line, and n : total population.
AHH	Life Expectancy (AHH) is the average estimate of the number of years a person can live since birth (Badan Pusat Statistik, 2012).
TPT	The Open Unemployment Rate (TPT) is the percentage of the number of unemployed people to the total labor force. The labor force is the working-age population (15 years and over) who are employed or have a job but are temporarily unemployed (Badan Pusat Statistik, 2024c).

Research Stages

The first step of research after obtaining data is to perform an initial process so that analysis can be carried out, this process is called preprocessing data (Joshi & Patel, 2021). Data preprocessing is done to see the balance of classification attributes using visualization. Then check for data outliers, imputation of empty data, then repaired. After the data is considered good and complete, comparisons are made with Decision Tree, Logistic, Random Forest, and SVM classification algorithms. Then a performance

analysis is carried out to determine the best classification method. Figure 2 shows the stages of the method used in this research.

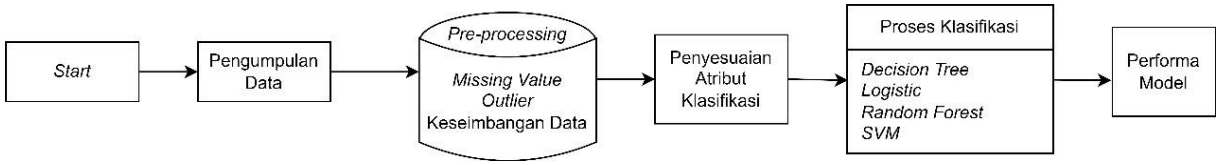

Figure 2. Research stages

Decision Tree

Decision tree is a method that utilizes a tree structure to make decisions (Jijo & Abdulazeez, 2021). Each path in the tree starts from the root node and goes through several data separation stages in each branch. Decision Tree can be used to solve classification and regression problems by grouping data into classes. Decision trees learn through a set of if/else or yes/no questions or other questions that form a hierarchical tree. One of the algorithms is Classification and Regression Tree (CART). The principle of this method is to generate a decision tree from a categorical response variable by sorting all observations into two clusters and sorting again for the next cluster until the minimum number of observations.

Random Forest

Random Forest is an extension of the CART method by applying bootstrap aggregating (bagging) and random feature selection methods. This method is known for its simplicity and effectiveness (Abdulkareem & Abdulazeez, 2021). Breiman introduced the Random Forest algorithm by showing several advantages including being able to produce relatively low errors, having good performance in classification, being able to handle large amounts of training data efficiently, and an effective method for estimating missing value data. Random Forest generates many independent trees (forests) with subsets randomly selected by bootstrapping from training samples and from input variables at each node. Random forest performs classification by adopting an overall approach of various trees through majority occurrence to achieve the final goal (Yoo, C., Han, D., Im, J., & Bechtel, 2019).

Logistic Classification

Logistic regression is a statistical analysis technique to determine the relationship of several variables to a categorical response variable (Roflin et al., 2023). In the case of categorical responses, binary logistic regression is classified, where there are only two classifications in the data. Logistic regression does not model directly the relationship between variables, but is transformed to logit variables with the natural log form of the odds ratio (Fractal, 2003). Logistic regression has the following equation:

$$g(x) = \ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = \ln\left(exp\left(\beta_0 + \sum_{j=1}^{p}\beta_j x_j)\right)\right) = \left(\beta_0 + \sum_{j=1}^{p}\beta_j x_j\right) \quad (2)$$

Support Vector Machine (SVM)

Support Vector Machine was first introduced by Boser, Guyon, Vapnik in 1992 at the Annual Workshop on Computational Learning Theory. Support Vector Machine (SVM) is a technique for making predictions, both in the case of classification and regression that works with the principle of finding the hyperplane with the largest margin. The fact that real-world datasets are rarely linearly separable makes SVM modified by incorporating a kernel function (kernel trick). Basically, the kernel

trick is to map low-dimensional non-linear data and transform it into a higher dimensional space. The goal is to simplify classification by finding the hyperplane. In non-linear SVM, the data x r is first mapped by the function (x) r Φ to a higher dimensional vector space. In this new vector space, the hyperplane that separates the two classes can be constructed. There are three types of non-linear kernels namely polynomial, gaussian, and sigmoid. In SVM, there are two parameters in contributing to the SVM line or n-dimensional hyperplane namely Cost (C) and Gamma (γ).

Generalized Regression Neural Network (GRNN)

GRNN was first proposed by Specht in 1991 which is a variation of radial basis neural networks. GRNNs can be used for regression, prediction, and classification (Martínez et al., 2022). GRNN is composed of four layers: input, pattern neurons, summation neurons, and output. In this research, it is intended for permutation as imputation of blank data in time series data. The advantage of GRNN is that it is consistent when the size of the training data set gets bigger. As the training dataset gets larger, the estimation error approaches 0.

Confusion Matrix

Confusion matrix is a performance measurement of the accuracy of a classification model. The accuracy of a classification model is seen by forming a matrix table called confusion matrix. The table is a two-way table with two binary variables, namely Actual and Predicted by dividing the data set into two groups, namely positive and negative (Zeng, 2019).

Table 2. Confusion matrix

Prediction	**Actual**	
	Positive	**Negative**
Positive	True Positive (TP)	False Positive (FP)
Negative	False Negative (FN)	True Negative (TN)

ROC-AUC Curve

The ROC curve is an analysis method represented in graphical form (Nahm, 2022). The ROC curve is built based on the value obtained in the confusion matrix calculation, namely the False Positive Rate (FPR) with the True Positive Rate (TPR).

$$FPR = \frac{FP}{FP+TN} \text{ dan } TPR = \frac{TP}{TP+FN}$$

If the ROC curve is further away from the baseline line (line crossing from point (0,0)) then the performance of the classification algorithm is better (Gorunescu, 2011).

Cross Validation

Cross validation is a method in data mining techniques that aims to obtain maximum accuracy. This method is also called k-fold cross validation where k times are tried for one model with the same parameters (Santosa, 2007).

The cross-validation stages are as follows:

a. Divide the dataset into k subsets with the same dimensional size.
b. Use each subset for testing data and the rest for training data.
c. With k = 10 as a result of extensive experiments and theoretical proofs, the average accuracy is $\frac{1}{10}\sum_{k=1}^{10} Akurasi_k$

3. RESULTS

Pre-processing

The data used in this study consists of 128 regencies and cities in Java and Bali. Of these, there are 41 empty observation units or 32.02% of the total observation units. This missing value is found in the gini ratio data of districts and cities in Jakarta and Central Java provinces in 2020. Therefore, a data imputation step is needed.

Imputation is carried out in two different treatments for districts/cities in Jakarta and Central Java in 2020. Imputation of the gini ratio of districts and cities in Jakarta uses a constant value technique with the consideration that the city and district areas in Jakarta are not autonomous regions, so that regional finances (APBD) are already included in the provincial government's financial accounts. Meanwhile, the gini ratio is related to the distribution of regional income. So, the imputation of the gini ratio of districts and cities in Jakarta uses the 2020 Jakarta provincial level gini ratio.

Meanwhile, the imputation of the gini ratio of districts and cities in Central Java province in 2020 uses a forecasting technique with a Generalized Neural Network that uses historical gini ratio data from 2000 - 2015. With the considerations that have been described in the methodology section and studying past historical patterns. Forecasting gini ratio uses autoregressive lag vector parameters of 1 to 5 and sigma or error evaluation parameters of 0.01.

Table 3. Evaluation of forecasting

RMSE	MAE	MAPE	SMAPE
0.0600	0.0514	18.9060	17.6133

Table 3 is the result of forecasting evaluation with the GRNN method. It can be seen that the RMSE and MAE values are small and close to zero. Thus, the error is said to be minimum and the model is feasible to use.

After imputation, the gini ratio data is classified with the provisions of the rules (Todaro & Smith, 2020). The distribution of the medium gini ratio dominates the regencies and cities in Central Java province in 2020. While the low gini ratio is only 1 district Batang of 0.29 and a high gini ratio of 0.51 in Blora district. From these results, it can be said that Blora district is the only region with the highest gini ratio on the island of Java and Bali in 2020. Thus, adjustments are needed in the next analysis.

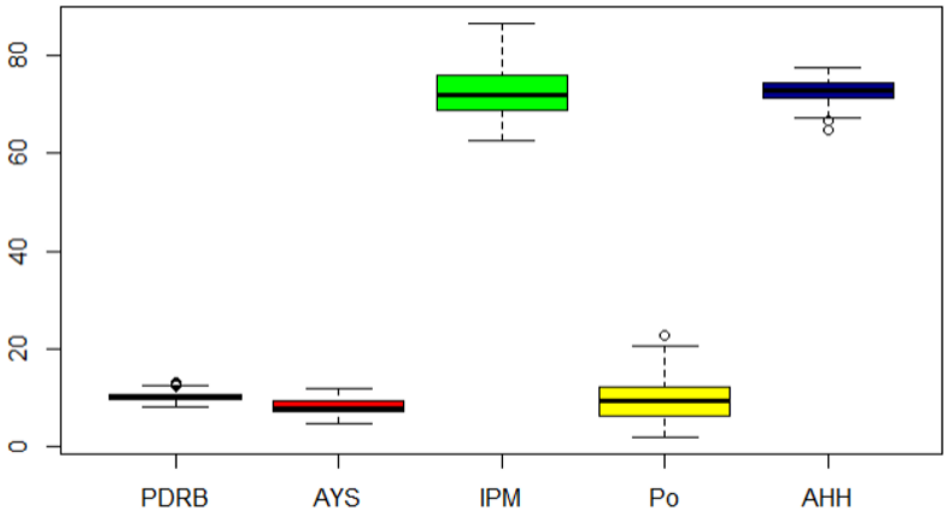

Figure 3. Boxplot of variable data

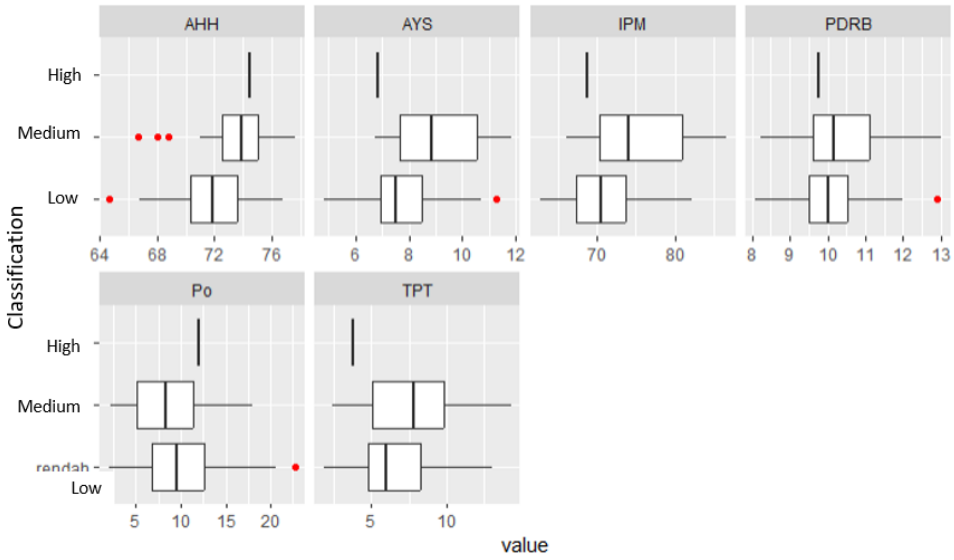

Figure 4. Boxplot of each category of gini ratio

Based on Figure 3 and Figure 4, it can be shown that there are district and city observation units in Java and Bali in 2020 that have outlier status. Based on the literature review, the study of outliers is carried out on each variable, not in each classification as shown in Figure 4. However, Figure 4 shows that there are districts and cities in the medium gini ratio classification that have high AHH values (outliers). The region is an area in the Jakarta province. The outliers in Figure 3 are classified as mild outliers. Mild outliers refer to outlier tolerance (not extreme outliers), thus the researcher decided that the data was free of outliers and no data transformation was performed.

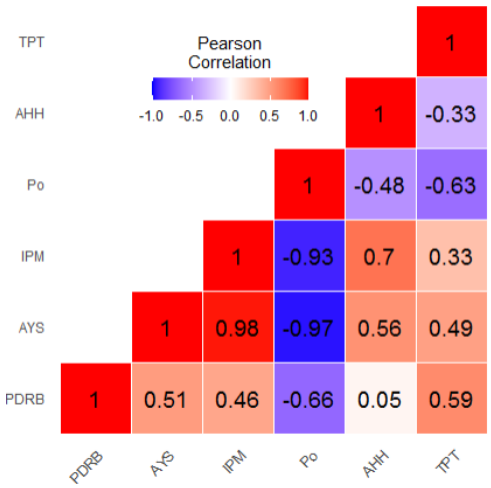

Figure 5. Correlation matrix

Based on Figure 5, there is a symptom of a linear relationship (multicollinearity) between the AYS and HDI variables of 0.93. Other than these two variables, there is no linear relationship. The researcher did not remove variables with multicoll symptoms on the basis of wanting to know the performance of the model against variables that were free of outliers but contained multicoll variables.

The next step is handling unbalanced data. There is one observation with a high gini ratio, namely Blora Regency, which is the result of forecasting and contains forecasting errors and based on Tobler's Law which states that the characteristics of the region are similar to the surrounding areas in an effort to influence each other (Ward & Gleditsch, 2019). So, the first step in balancing is done by including

the classification of the Gini ratio of Blora district in 2020 into the medium class. Thus, the distribution of data classification classes becomes Figure 6.

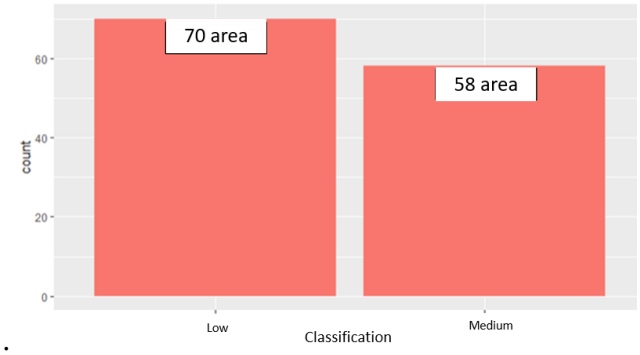

Figure 6. Gini ratio distribution

Figure 6 shows that the imbalanced data ratio is 0.828 and the researcher decided to balance the classification class with the MWMOTE algorithm. This algorithm was chosen with the consideration that it can handle overfitting events by generating synthetic data well and is able to avoid the synthesized data generated into noise data. After Oversampling with the MWMOTE algorithm the dataset class was successfully balanced with a 100% imbalanced ratio.

Classification

The balanced dataset will be randomly separated into training data and testing data. Training data is used for classification modeling and testing data is used for model evaluation. In this study, training data is used with a portion of 80% and testing data with a portion of 20%.

Decision Tree CART

CART decision tree classification modeling in R software has produced a decision tree with the following order of important variables:

Table 4. Important variables for split decision tree

AHH	IPM	AYS	TPT	PO	PDRB
27	21	19	15	10	7

Table 4 shows that the life expectancy variable has the largest contribution in the variation of splitting the model. Table 4 shows that the AHH variable has a large contribution in the form of a score or impurity (entropy) reduction to the model.

Table 5. Confusion matrix of decision tree

Prediction	Actual	
	Low	Medium
Low	9	5
Medium	6	8

Table 5 shows the model evaluation criteria based on the confusion matrix. Where based on Table 5, an accuracy value of 60.71% is generated which is classified as poor classification. In line with these results, the average accuracy value generated by 10-fold validation produces an average with a range of 0.54 to 0.63 after resampling. This also aggravates the classification of poor classification and failure

classification. The model produced a final Cp criterion value of 0.036. Based on output R, the value of ROC is 0,6077 is classified as poor classification.

Logistic Classification

Probability classification modeling is done with a logistic approach. It was found that the variables AYS, HDI, Po, and AHH contributed significantly at the 5% level in classifying the gini ratio of districts and cities in Java Bali in 2020.

Table 6. Confusion matrix of logistic

Prediction	Actual	
	Low	Medium
Low	14	5
Medium	1	8

Table 6 shows the results of evaluating the classification model with a confusion matrix. The number of significant variables contributing to the Gini ratio classification resulted in an accuracy rate of 0.785, which is classified as fair classification. In addition, the False Negative value of 1 is minimal, making this logistic classification algorithm worth using because it minimizes classification errors. Then 10-fold validation is carried out and the resulting resampling average accuracy is 76.93%. Based on output R, the value of ROC is 0,7744 is classified as fair classification.

Random Forest

Based on **Error! Reference source not found.**, the random forest model builds 1000 decision trees from the training data as the final result of the model. The line displayed in **Error! Reference source not found.** shows that the more trees formed, the more stable the error and the more accurate the model. The error rate converged after the 500th tree was formed with an Out of Bag (OOB) estimate of error rate of 30.36%. From these iterations, the optimal decision tree is shown in Figure 8.

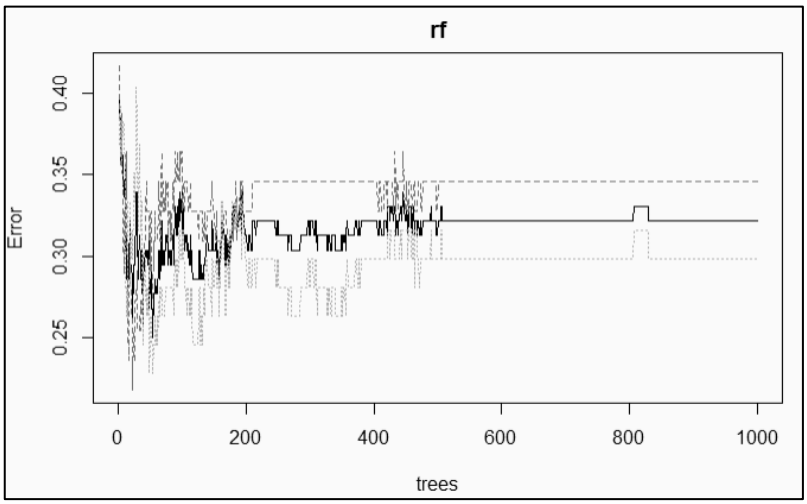

Figure 7. Error patterns based on the number of trees.

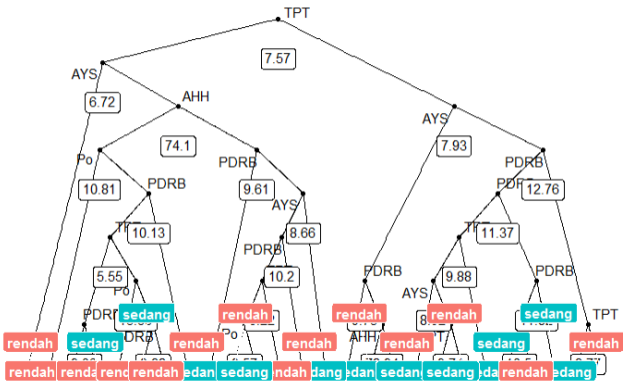

Figure 8. Final decision tree from random forest

Unlike the decision tree image, the decision tree in Figure 8 is generated from the iteration process to converge the error in the random forest. The results of the convergent decision tree in the random forest are fairer in making policies where in each classification the ratio of gini contains all considerations of the research variables.

Table 7. Confusion matrix of random forest

Prediction	Actual	
	Low	Medium
Low	13	6
Medium	2	7

Table 7 is an evaluation of the model with the confusion matrix of the testing data. The accuracy of the random forest model is 71.42%, which is classified as fair classification. The False Negative rate is also relatively small, which means the level of misclassification where a district and city is low inequality but classified as moderate inequality. Then 10 fold validation was conducted which showed an accuracy result of 0.70% which is classified as fair classification. Based on output R, the value of ROC is 0,7026 is classified as fair classification.

SVM

In determining SVM parameters, the parameter tunning process is carried out. After the parameter tunning process, the best value for the cost parameter is 5 and for the gamma parameter is 0.1. With a radial kernel that maximizes accuracy compared to other kernels. Then with a cost value of 5 and a gamma of 0.1 the following output is produced:

```
Call:
svm(formula = Klasifikasi ~ ., data = data.train, kernel = "radial", gamma = 0.1,
    cost = 5)


Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  radial
       cost:  5

Number of Support Vectors:  74

 ( 38 36 )


Number of Classes:  2

Levels:
 rendah sedang
```

Figure 9. Output R SVM algorithm

Figure 9 shows that the minimum number of support vectors is 74. This is interpreted that the number of support vectors or hyperplane separators is 74 with a cost value of 5 and a gamma of 0.1.

Table 8. Confusion matrix of SVM

Prediction	Actual	
	Low	Medium
Low	13	2
Medium	5	8

Based on Table 8, the accuracy value of the SVM model in this study is 75% which is classified as fair classification. Then 10 fold validation is carried out which shows the average resampling accuracy value between 0.73 to 0.75 which is classified as fair classification. Based on output R, the value of ROC is 0,741 is classified as fair classification.

4. DISCUSSIONS

The above classification model was built on a dataset with a balanced gini ratio classification class, no outliers, and multicollinearity. In addition, there is also an imputation technique with a constant value and forecasting with GRNN resulting in a model performance. The following is a comparison of the performance of each classification algorithm used:

Table 9. Performance comparison of classification algorithms

Algorithm	Confusion Matrix's Accuration	10-Fold Validation Accuration's Mean	AUC	Category
Decision Tree CART	60,71%	0,54% - 0,63%	60,77%	Failure and Poor Classification
Logistic	78,5%	76,9%	77,4%	Fair Classification
Random Forest	71,42%	70%	70,26%	Fair Classification
SVM	75%	73% - 75%	74,1%	Fair Classification

Based on Table 9, it can be concluded that the evaluation method both from confusion matrix accuracy, average accuracy of 10 fold validation, and AUC places logistic regression as the best classification algorithm that can be applied to the conditions of the dataset. Table 9 interprets that the best classification algorithm, namely logistic classification, can classify regencies and cities in Java Bali in 2020 with an accuracy rate of 78.5%. All classification algorithms provide accuracy values that are classified as fair classification. However, the accuracy value is built based on training and testing data which is fixed after a random process. Or in R software with the *set.seed* function. Researchers will try resampling various possible combinations of training data in forming models and testing data in evaluation. The resampling results in the R software producing output as shown in Table 10.

Table 10. Output R resampling model evaluation

Algorithm	Statistics Measurement	Accuracy	Precision	Recall	Sensitivity	Specificity
Decision Tree	Min	0.4000	0.3333	0.2000	0.2000	0.3333
	Median	0.6333	0.6333	0.5000	0.5000	0.6667
	Mean	0.6256	0.6260	0.5733	0.5733	0.6733

Algorithm	Statistics Measurement	Accuracy	Precision	Recall	Sensitivity	Specificity
	Max	0.9000	1.0000	1.0000	1.0000	1.0000
Logistic Regression	Min	0.5455	0.5000	0.3333	0.3333	0.5000
	Median	0.7500	0.7321	0.8167	0.8167	0.8000
	Mean	0.7694	0.7598	0.7667	0.7667	0.7767
	Max	1.0000	1.0000	1.0000	1.0000	1.0000
SVM	Min	0.6429	0.6250	0.5714	0.5714	0.5714
	Median	0.7143	0.7571	0.7143	0.7143	0.7857
	Mean	0.7357	0.7648	0.7000	0.7000	0.7714
	Max	0.9286	1.0000	1.0000	1.0000	1.0000
Random Forest	Min	0.4000	0.3333	0.2000	0.2000	0.6000
	Median	0.6970	0.6667	0.5833	0.5833	0.7333
	Mean	0.7008	0.7224	0.6367	0.6367	0.7667
	Max	0.9167	1.0000	1.0000	1.0000	1.0000

Table 10 shows various measures of model performance from the data resampling process. The possible performance measure values show the distribution of the data. Starting from the minimum value, median, mean, and maximum value. Overall, the best models are logistic regression and Support Vector Machine. Although the analysis in Table 9 uses a subset of resampling, this is enough to illustrate the best algorithm. Table 10 also shows the possible minimum and maximum performance measures of a regression algorithm.

The decision tree algorithm which in Table 9 is placed as poor classification, it turns out that there is a condition where the maximum accuracy value can reach 90%. The same applies to other classification algorithms. Although the accuracy value of the SVM algorithm has a minimum value that is greater than the minimum value of logistic regression. The average performance measure of the logistic classification model is greater than the SVM model. Thus, the best algorithm in classifying the gini ratio of districts and cities in Java and Bali in 2020 is logistic regression. So that in modeling the classification of the level of inequality in income distribution in Java and Bali, the logistic regression method has been produced which has the largest central tendency accuracy value compared to decision tree, SVM, and random forest.

The analysis reveals that accuracy shows varying minimum, median, average, and maximum values for each algorithm. For instance, Decision Tree has a minimum accuracy of 0.4000, indicating that it performs poorly in the worst-case scenario. However, its maximum accuracy reaches 0.9000, demonstrating significant potential when processing data effectively. In contrast, Logistic Regression exhibits more stable performance, with a median accuracy of 0.7500 and a maximum of 1.0000, suggesting it consistently delivers good results.

Precision and recall provide deeper insights into each model's ability to identify positive predictions. Decision Tree shows a maximum precision of 1.0000 but also has lower values in other metrics. On the other hand, both Logistic Regression and SVM demonstrate high precision and recall, indicating they not only predict positive cases accurately but also capture most existing positive instances.

Sensitivity and specificity are also crucial in this evaluation. High sensitivity indicates that the algorithms can detect many positive cases, while specificity shows how well the models avoid errors in predicting negatives. Logistic Regression and SVM perform well in both metrics, while Random Forest exhibits relatively balanced performance.

The use of minimum, median, average, and maximum values provides a comprehensive overview of each algorithm's performance. The maximum value indicates the best potential that can be achieved, while the minimum and median values provide context regarding the variability and stability of the models. Overall, these results suggest that Logistic Regression and SVM are more reliable choices for classifying the level of income distribution inequality in Java and Bali under balanced data conditions, while Decision Tree and Random Forest may require further tuning to achieve optimal performance in this classification study.

5. CONCLUSION

At the pre-processing stage there is missing data in the dependent variable, namely the gini ratio of districts and cities in Jakarta and Central Java provinces in 2020. Handling imputation of the gini ratio in Jakarta uses a constant value, while imputation of the gini ratio in Central Java uses GRNN forecasting. With the condition of a dataset with balanced classes from MWMOTE bootstrapping results, there are no outliers and there is multicollinearity, it has been successfully modeled with decision tree, logistic classification, random forest, and SVM algorithms. With a portion of 80% training data and 20% testing data, it is found that logistic regression is the best classification algorithm based on model performance. When resampling, the SVM algorithm has a higher minimum accuracy value than logistic, but on average the logistic algorithm has greater accuracy than SVM.

7. REFERENCES

Abdulkareem, N. M., & Abdulazeez, A. M. (2021). Machine learning classification based on Radom Forest algorithm: a review. *International Journal of Science and Business*, *5*(2), 128–142. https://doi.org/10.5281/zenodo.4471118

Badan Pusat Statistik. (2012). *Kematian Bayi dan Angka Harapan Hidup Penduduk Indonesia Hasil Sensus Penduduk 2010*.

Badan Pusat Statistik. (2023). Statistik Pendidikan 2023. *Badan Pusat Statistik*, *12*, i–242. https://www.bps.go.id/id/publication/2022/11/25/a80bdf8c85bc28a4e6566661/statistik-pendidikan-2022.html

Badan Pusat Statistik. (2024a). Data dan Informasi Kemiskinan Provinsi Jawa Tengah 2019-2023. *Statistik Kerawanan Sosial*, *16*.

Badan Pusat Statistik. (2024b). *Indeks Pembangunan Manusia 2023*. *18*, 1–282.

Badan Pusat Statistik. (2024c). *Keadaan Angkatan Kerja di Indonesia Februari 2024*.

Badan Pusat Statistik. (2024d). *Tinjauan PDRB Kabupaten/Kota Se-Jawa Tengah Menurut Pengeluaran*. *4*.

BAPPEDA Kulon Progo. (2022). *Konsolidasi Perencanaan Pembangunan Jawa-Bali Tahun 2022*. BAPPEDA Kulon Progo. https://bappeda.kulonprogokab.go.id/detil/828/konsolidasi-perencanaan-pembangunan-jawa-bali-tahun-2022

BAPPENAS. (2020). *Rencana Pembangunan Jangka Menengah Nasional 2020 - 2024*.

Cristea, L. A., Vodă, A. D., & Mihai, D. U. (2021). ECONOMIC GROWTH AND DEVELOPMENT, PROMOTERS OF NATIONAL WELL-BEING. *Revisita Economica*, *73*, 86.

Fractal. (2003). Comparative Analysis of Classification Techniques. *A Fractal White Paper*.

Gorunescu, F. (2011). *Data Mining Concepts, Models and Techniques*. Springer.

Islami, F. S. (2018). FAKTOR-FAKTOR MEMPENGARUHI KETIMPANGAN WILAYAH DI PROVINSI JAWA TIMUR, INDONESIA. *Media Ekonomi Dan Manajemen*, *33(1)*.

Jijo, B. T., & Abdulazeez, A. M. (2021). Classification Based on Decision Tree Algorithm for Machine Learning. *Journal of Applied Science and Technology Trends*, *2*(01), 20–28. https://doi.org/10.38094/jastt20165

Joshi, A. P., & Patel, B. V. (2021). Data Preprocessing: The Techniques for Preparing Clean and Quality Data for Data Analytics Process. *Oriental Journal of Computer Science and Technology*, *13*(0203), 78–81. https://doi.org/10.13005/ojcst13.0203.03

Karsu, Ö., & Morton, A. (2015). Inequity averse optimization in operational research. *European Journal of Operational Research*, *245*(2), 343–359. https://doi.org/10.1016/j.ejor.2015.02.035

Martínez, F., Charte, F., Frías, M. P., & Martínez-Rodríguez, A. M. (2022). Strategies for time series forecasting with generalized regression neural networks. *Neurocomputing*, *491*, 509–521. https://doi.org/10.1016/j.neucom.2021.12.028

Nahm, F. S. (2022). ROC Curve: overview and practical use for clinicians. *Korean Journal of Anesthesiology*, *75*(1), 25–36.

Purnama, B. (2019). *Pengantar Machine Learning*. Informatika.

Roflin, E., Riana, F., Munarsih, E., Pariyana, & Liberty, I. A. (2023). *Regresi Logistik Biner dan Multinomial*. PT Nasya Expanding Management.

Santosa, B. (2007). *Data Mining Teknik Pemanfaatan Data untuk Keperluan Bisnis*. Graha Ilmu.

Todaro, M. P., & Smith, S. C. (2020). Economic Development. Thirteenth Edition. In *Pearson* (Issue 13th Edition). Pearson. https://www.mkm.ee/en/objectives-activities/economic-development

Ward, M. D., & Gleditsch, K. S. (2019). *Spatial Regression Model* (2nd Editio). SAGE Publications, Inc.

Yoo, C., Han, D., Im, J., & Bechtel, B. (2019). Comparison Between Convolutional Neural Networks and Random Forest for Local Climate Zone Classification in Mega Urban areas using Landsat Images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 155–170.

Zeng, G. (2019). On the confusion matrix in credit scoring and its analytical properties. *Communications in Statistics - Theory and Methods*, *49*(9), 2080–2093. https://doi.org/10.1080/03610926.2019.1568485