Journal of Artificial Intelligence and Engineering Applications

Website: https://ioinformatic.org/

15th February 2025. Vol. 4. No. 2; e-ISSN: 2808-4519

Random Forest Algorithm to Improve the Classification Model for Recipients of Direct Cash Assistance in Danalampah Village, Pancalang District, Kuningan Regency

Rio Harsadino[1]*, Rudi Kurniawan[2], Saeful Anwar[3]

[1,2,3]*STMIK IKMI Cirebon*

ryoooharsadino25@gmail.com[1]*, rudi226@gmail.com[2] , saeful.ikmi@gmail.com[3]

Abstract

Data classification plays a crucial role in machine learning, particularly in supporting data-driven decision-making processes. Data imbalance between majority and minority classes often leads to model bias toward the majority class, reducing accuracy in detecting the minority class. In this study, recipients of direct cash assistance (BLT) represent the minority class, whose numbers are significantly fewer than non-recipients. Therefore, specific strategies are required to ensure more accurate and fair classification results. This study aims to analyze the impact of data imbalance on the performance of the BLT classification model, evaluate the model's performance using accuracy, precision, and recall metrics, and explore the effectiveness of data balancing techniques in improving the model's sensitivity to the minority class. The Random Forest algorithm serves as the primary method for building the classification model. This algorithm works by constructing numerous decision trees based on randomly selected data and features, then combining predictions through majority voting. Additionally, data balancing is implemented through manual reduction in the majority class, resulting in a dataset with a more proportional ratio (1:2), comprising 66 records: 22 BLT recipients and 44 non-recipients. The results demonstrate that the Random Forest algorithm applied to the original dataset achieves high accuracy (99.59%), but the recall for BLT recipients is 0%, indicating bias toward the majority class. After data balancing, accuracy slightly decreases to 98.33%, while recall significantly improves to 66.67%, and precision reaches 65.83%. These findings indicate that data balancing successfully enhances the model's sensitivity to the minority class without significantly compromising accuracy. This study concludes that the integration of data balancing techniques and the Random Forest algorithm improves the fairness and representation of classification outcomes. These findings offer practical contributions to the application of social data analysis, particularly in classifying social assistance recipients, to support more targeted decision-making processes.

Keywords: *Data Classification, Machine Learning, Data Imbalance, Random Forest, Data Balancing.*

1. Introduction

The development of information technology has had a significant impact on various aspects, such as technology, business, and education. In the digital era, information technology has become crucial in making decisions more efficiently and effectively. One tool that plays a significant role is the machine learning algorithm, especially Random Forest, which facilitates the analysis of large and complex datasets to recognize new patterns [1]. With the challenge of managing imbalanced data, particularly in the context of social assistance, this study explores the use of the Random Forest algorithm to improve the classification accuracy of recipients of Direct Cash Assistance (BLT) in Danalampah Village. This research is important so that the assistance can be distributed accurately and have the maximum impact on those in need.

The main issue faced is the class imbalance in identifying BLT recipients. This imbalance can cause the model to be more likely to group data into the majority class, leading to bias and weakening the detection of truly deserving BLT recipients. This issue has been discussed in several previous studies, but many have yet to offer effective solutions to the imbalanced data problem [2]. Therefore, this study aims to develop a method that is more responsive to the minority class to enhance the accuracy and fairness of the classification of recipients. Based on previous research, it is known that although the Random Forest algorithm is popular in various applications, handling imbalanced data remains a challenge. [3] showed the success of Random Forest in classification but did not specifically discuss data balancing techniques.

Revealed the benefits of this algorithm in air pollution classification, but did not explore data reduction techniques for better performance. Similarly [4], [5] used Random Forest in classification but did not apply sampling techniques to improve the model's performance on imbalanced data. [6] also used the Random Forest model for its ability to handle outliers and provide more stable results through cross-validation, which helps reduce the risk of overfitting. These findings indicate an opportunity to improve model performance through the integration of manual data reduction and stratified sampling techniques. This study aims to improve the accuracy and sensitivity of the classification model in identifying BLT recipients in an imbalanced class dataset [7]. This imbalance causes bias toward the majority class

(non-BLT recipients), making the model fail to detect BLT recipients effectively. To address this issue, this study develops an approach based on the Random Forest algorithm equipped with data balancing through manual reduction of the majority class. This approach is designed to enhance the model's sensitivity to the minority class without requiring complex methods such as oversampling [5].

This research contributes to the field of Informatics by presenting a simple approach that can optimize the classification of imbalanced data, a common challenge in data analysis. The findings of this study also have practical benefits for policymakers, social institutions, and government organizations to ensure more accurate and targeted detection of social assistance recipients, particularly when data analysis resources are limited. In this study, the Random Forest algorithm is applied to two dataset scenarios: the original imbalanced dataset and the dataset after data balancing. In the original dataset, the model is tested to evaluate the impact of data imbalance on classification performance, where the majority class dominates and the minority class (BLT recipients) tends to be overlooked. In the balanced dataset, the majority class data is manually reduced to create a more proportional distribution with a 1:2 ratio, thus increasing the representation of the minority class.

Model performance is evaluated using accuracy, precision, and recall metrics. Accuracy is used to measure overall correct predictions, precision measures the accuracy of predicting BLT recipients, and recall measures the model's sensitivity in detecting the minority class. If this study's objectives are achieved, it will contribute to the development of more efficient and fair classification models for imbalanced data. The findings will be useful for practitioners and researchers in applying manual data reduction methods as a practical solution without the need for oversampling or other complex algorithms. The results are also expected to provide further insights into the application of Random Forest in classification with highly imbalanced data, particularly in the context of social assistance distribution, such as BLT.

2. Methods

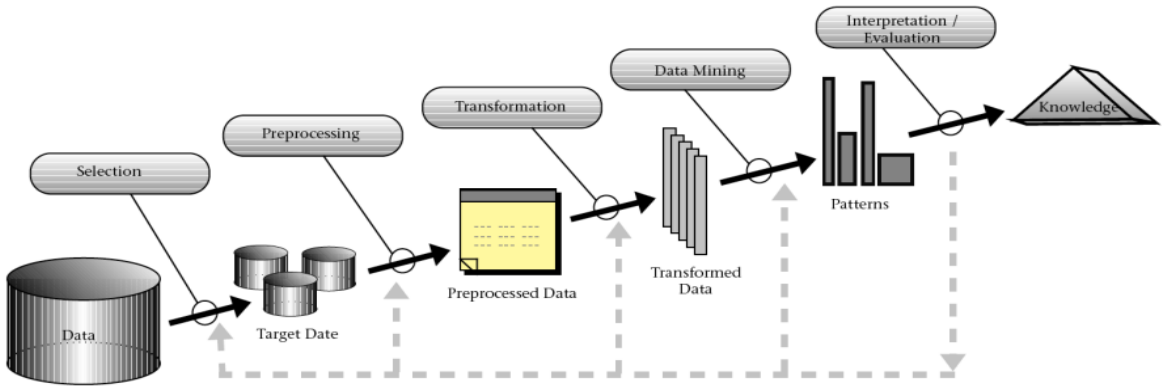

Fig. **1**: Knowledge Discovery In Databases(KDD)

In this research, an experimental approach is used, involving the Random Forest algorithm to improve the performance of classifying Direct Cash Assistance (BLT) recipients. The study follows a systematic process, starting from data collection to model evaluation, with the goal of providing a comprehensive understanding of the data used. This process ensures that the classification model produced delivers accurate and relevant results, while supporting precise decision-making. Each phase is interconnected and follows the principles of Knowledge Discovery in Databases (KDD), which serves as the primary framework for this research. The first step involves data collection, which includes accessing data on both BLT recipients and non-recipients. The data was gathered from documents available at Danalampah Village, containing a total of 272 data entries, including 22 BLT recipients and 250 non-recipients. These records include socio-economic attributes such as National Identification Number (NIK), gender, occupation, and BLT status. The next phase is data selection, where only the relevant attributes for classification such as gender, occupation, BLT distribution mechanism, and recipient status are chosen, while irrelevant attributes such as names and NIK are removed to enhance analysis efficiency.

Data preprocessing follows, where the dataset is cleaned by removing empty, duplicate, or irrelevant data, ensuring that the dataset is of high quality. Categorical attributes are converted to numeric format using encoding, and normalization is applied to align the scales of numerical attributes. The transformation phase addresses the class imbalance by manually reducing the majority class (non-recipients) to achieve a 1:2 class ratio, creating a more representative dataset with 66 entries. This transformation ensures the model can better recognize patterns in the minority class. Furthermore, a polynomial to numerical operator is applied to convert categorical data for compatibility with the Random Forest algorithm. In the data mining phase, the Random Forest algorithm is applied to build classification models using two datasets: the original dataset with 272 entries and the balanced dataset with 66 entries. The data is split into training and testing sets using stratified sampling to maintain the class distribution.

The Random Forest algorithm is applied with default parameters (number of trees: 100, criterion: Information Gain, max depth: 2), and the model is trained on the training data and evaluated using test data. The evaluation phase involves assessing the performance of the Random Forest model based on metrics such as accuracy, precision, and recall. Accuracy measures the overall correctness of the model's predictions, while precision gauges the model's accuracy in predicting BLT recipients, minimizing false positives. Recall measures the model's sensitivity to detecting BLT recipients, minimizing false negatives. The evaluation results allow for comparing the model's performance on both the original and balanced datasets, providing insight into the impact of data balancing on the model's effectiveness. By following

this methodology, the study aims to improve the classification of BLT recipients by addressing data imbalance, ultimately contributing to more accurate and fair decision-making in social assistance distribution.

3. Discussion

The dataset used in this research is the BLT (Bantuan Langsung Tunai) recipient data in Excel format, consisting of 272 rows and 12 attributes. It includes crucial information about the characteristics of BLT recipients and non-recipients in Danalampah Village. The attributes in the dataset cover personal details such as NIK (National Identification Number), name, gender, date of birth, address, job type, BLT eligibility criteria, distribution mechanism, and BLT recipient status. To facilitate data processing and analysis, some attributes were transformed into numeric values for better compatibility with the classification algorithms used in this study.

The dataset was then processed using RapidMiner's "Read Excel" operator, allowing easy data import and conversion into a table format for further analysis. The subsequent step, "Select Attributes," was employed to focus on the most relevant features for classification, such as gender, occupation, and BLT status. This step ensured that the analysis focused on the most significant variables, improving the efficiency and accuracy of the Random Forest classification model. Non-relevant attributes like NIK and name were excluded from the dataset, as they do not directly impact the classification outcome. This process helped streamline the analysis and ensure the model was trained on the most pertinent data.

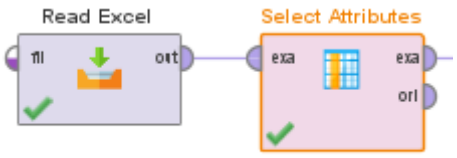

Fig. 1: Data Selection Stages

The data preprocessing phase in this research was completed during the data selection process, ensuring that the attributes used aligned with the requirements of the classification model. Initially, the "Replace Missing Value" operator is commonly employed to address missing data, but it was not necessary in this case since the dataset contained no missing values. The dataset was found to be complete, with no missing values in any of the attributes, ensuring that all entries could be used in the analysis without the risk of bias or errors from missing data. The preprocessing phase also included encoding categorical attributes into numeric formats and normalizing numeric attributes to maintain consistency and prevent bias due to differences in value ranges across attributes. Following preprocessing, the next phase involved data transformation, which aimed to enhance the dataset for more effective analysis. A key part of this transformation was addressing class imbalance between BLT recipients and non-recipients. Manual adjustment of the dataset reduced the majority class (non-recipients) to achieve a 1:2 ratio between recipients and non-recipients.

This balancing ensured that the model would not be biased towards the majority class, improving its ability to identify patterns in the minority class. After balancing, the "Polynomial to Numerical" operator was used to convert categorical attributes into numeric forms, making them compatible with the Random Forest classification algorithm. This conversion enabled the model to handle non-linear relationships that may exist between input and output variables. Next, the "Set Role" operator was applied to define the role of each attribute in the dataset. This step organized the data structure and made it suitable for the Random Forest model. The attribute "NO" was assigned as the ID, while "Status Penerima BLT" was designated as the label, representing the target variable to be predicted. This ensured that the dataset was properly structured for the classification model, which was essential for accurate learning and prediction. After completing the transformation steps, the dataset was ready for the next stage of data mining, where the Random Forest algorithm would be used to develop the classification model based on the prepared data.

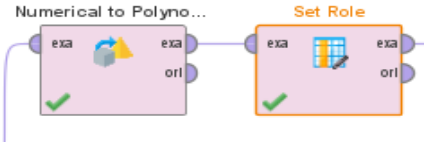

Fig. 2: Transformation Stages

3.1. Data Mining

After the data undergoes the transformation phase, which includes balancing and attribute adjustment processes, the next step is Data Mining. Data Mining aims to explore patterns and relationships within the data using algorithm-based analytical methods. In this study, the Random Forest algorithm was chosen due to its ability to handle complex data, including datasets with imbalanced class distributions. Once the transformation process is completed, the next stage is Data Mining. At this stage, the Random Forest algorithm is applied to build a classification model. This process is carried out using two dataset scenarios: The first scenario uses the original dataset, which consists of 272 data points, as shown in Table 4.8. This dataset includes all initial data without any class ratio adjustments. The original dataset has an imbalanced distribution, with 22 BLT recipients (minority class) and 250 non-BLT recipients (majority class). Testing this dataset aims to evaluate the model's performance under imbalanced data conditions. The second scenario involves inputting a balanced dataset, consisting of 66 data points, as shown in Table.

This dataset is the result of the transformation phase, where the majority class data has been manually reduced to create a more proportional class distribution, with a ratio of 1:2 (22 BLT recipients and 44 non-BLT recipients). The purpose of testing this dataset is to assess how class balancing improves the model's sensitivity to the minority class. The Data Mining phase is conducted using these two dataset scenarios to ensure the classification model's performance. The first scenario uses the original dataset with 272 data points, where the

majority class (non-BLT recipients) dominates over the minority class (BLT recipients). This dataset is used to evaluate how the Random Forest algorithm performs on imbalanced data. The second scenario utilizes the balanced dataset, where the majority class has been manually reduced to achieve a 1:2 class ratio, totaling 66 data points.

This dataset aims to measure the impact of class balancing on the model's sensitivity in identifying the minority class. These two scenarios provide comprehensive insights into the model's performance under different data distribution conditions. The application of the Data Mining algorithm begins with splitting the dataset using the Split Data operator. This operator is used to separate the dataset into two main parts: the training set and the testing set. This division is essential to ensure that the model is trained on specific data and tested on unseen data, enabling objective evaluation of the model's performance. In this study, the division is done with a particular proportion to maintain proportional class representation in both the original and balanced datasets. Below is Figure 4.12, which illustrates the Split Data operator.

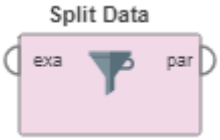

Fig. 3: Operator Split data

In the Split Data operator, there are several parameters that need to be configured to ensure the optimal separation of the dataset. These parameters include the data split proportion, randomization method, and class distribution in both the training and testing data. The following are the parameters used in the Split Data operator, as shown in Table 4.10: Parameters for Split Data.

Table 1: Font Specifications for A4 Papers

No	Attribute name	Target role
1	Sampling Type	Stratified Sample

In the Split Data operator, the Sampling Type parameter is used to determine the method of splitting the dataset into training and testing data. Available options for this parameter include Linear Sample, Shuffled Sample, and Stratified Sample, each with distinct characteristics. The Linear Sample method divides the data sequentially according to the order in the dataset. While simple, this method can lead to bias if the class distribution is imbalanced, as it doesn't account for uneven class distribution. The Shuffled Sample method, on the other hand, shuffles the data before splitting it into training and testing datasets. This helps reduce bias caused by data order but doesn't guarantee balanced class distribution in the training and testing sets. Given the class imbalance in the dataset, especially with the original dataset where the majority class (non-BLT recipients) is much more dominant than the minority class (BLT recipients), Stratified Sample was selected. This method ensures that the class distribution in both the training and testing sets remains proportional to the original dataset.

In the Split Data operator, the Sampling Type parameter is used to determine the method for splitting the dataset into training and testing data. The available options for this parameter include Linear Sample, Shuffled Sample, and Stratified Sample, each with its own characteristics. The Linear Sample method divides the data sequentially according to the order in the dataset. This method is simple and can be used when the order of the data does not affect the analysis results. However, if the class distribution in the dataset is uneven, this method can introduce bias. On the other hand, the Shuffled Sample method randomly shuffles the data before dividing it into training and testing sets.

This method helps reduce potential bias due to the data order, but it does not guarantee balanced class distribution between the training and testing data. For this stage, Stratified Sample was chosen because the dataset being used has class imbalance, particularly in the original dataset where the majority class (non-BLT recipients) is much more dominant than the minority class (BLT recipients). By using Stratified Sample, the class distribution in both the training and testing data remains proportional to the original dataset. This is important to ensure consistent class representation, allowing the model enough opportunity to learn patterns from both classes, both majority and minority.

Additionally, Stratified Sampling improves the validity of the evaluation because the class proportions in the test data are the same as those in the training data, making the model evaluation more representative. This method also reduces model bias toward the majority class, especially when the minority class has significantly fewer data points. By using the Stratified Sample parameter, the data splitting process becomes more accurate and supports a fairer performance analysis in both the original dataset and the balanced dataset scenarios. After determining the Sampling Type parameter with the Stratified Sampling technique to maintain the class distribution proportion in the dataset, the next step is to set the Ratio parameter in the Split Data operator. The ratio used for the data split is 90:10, where 90% of the data is used for model training, and the remaining 10% is used for testing. This ratio ensures that the model has enough training data to learn the patterns in the dataset, especially in the balanced dataset scenario, which has a smaller number of data points. This approach provides the model with more information to enhance its prediction capability. Even though the testing data makes up only 10%, this amount is considered sufficient to provide a reliable performance evaluation, given that the balanced dataset has been designed to represent a more evenly distributed class. Other ratios, such as 80:20 or 70:30, are also commonly used. The 80:20 ratio provides a balance between training and testing, suitable for medium-sized datasets, while the 70:30 ratio is more appropriate for smaller datasets requiring broader evaluation.

However, in this study, the 90:10 ratio was chosen because it optimally benefits both the imbalanced and balanced dataset scenarios, ensuring that the model has a greater chance of learning the patterns without losing the ability to evaluate test data effectively. After the data splitting process using the Split Data operator, the next step in this study is to apply Cross Validation to evaluate the model's performance more accurately and reliably. The Cross Validation technique is used to ensure that the model evaluation results are not dependent solely on a specific data split but reflect the overall generalization capability of the model. In this study, 5-fold Cross Validation is applied, where the dataset is divided into five proportional parts using the Stratified Sample technique. This method is shown in the usage of 5-fold Cross Validation and the application of Stratified Sample in Table 2.

Table 2: Paramater Operator Cross Validation

No	Parameter name	Target role
1	Number of Fold	5
2	Sampling Type	Stratified Type

In this phase, 5-fold Cross Validation and Stratified Sampling techniques were applied to divide the dataset into five balanced parts, ensuring each data point was used for both training and testing. Stratified Sampling was chosen to maintain the original class distribution in each fold, which was important due to significant class imbalance in the dataset. After setting up Cross Validation, the Random Forest operator was used to build the classification model with parameters including 100 trees, Information Gain as the criterion, and a maximum depth of 10. These settings were chosen to balance model complexity and performance, avoiding overfitting while ensuring accurate predictions. Following the training, the Apply Model operator was used to test the model's performance on unseen data, and its effectiveness was evaluated using the Performance (Classification) operator.

Table 3: Operator Performance (Classification) original dataset scenario

	True cluster 18	True 1	True 2	Class precission
Pred.18	0	0	0	0.00%
Pred.1	0	20	0	100.00%
Pred. 2	1	0	225	99.56%
Class recall	0.00%	100.00%	100.00%	
Accuracy: 99.59% +/- 0.91%				
weighted_mean_recall: 66.67% +/- 0.00%				
weighted_mean_precision: 66.52% +/- 0.33%				

Table 4: Operator Perfrmance(Classification) Balance dataset scenario

	True cluster 18	True 1	True 2	Class precission
Pred.18	0	0	0	0.00%
Pred.1	0	20	0	100.00%
Pred. 2	1	0	40	97.56%
Class recall	0.00%	100.00%	100.00%	
Accuracy: 99.59% +/- 0.91%				
weighted_mean_recall: 66.67% +/- 0.00%				
weighted_mean_precision: 66.52% +/- 0.33%				

The model's performance was assessed using accuracy, weighted mean recall, and weighted mean precision, reflecting the overall accuracy and ability to correctly classify both majority and minority class data. In the original dataset scenario, with 272 data points, the model achieved an accuracy of 99.59%, but it showed a bias toward the majority class, struggling to correctly identify the minority class (BLT recipients). The confusion matrix indicated that the model had a perfect recall for the majority class but failed to predict the BLT status class correctly, leading to a significant imbalance in its performance. In the balanced dataset scenario, with 66 data points, the model achieved an accuracy of 98.33%. Although the balancing helped improve overall performance, the model still exhibited bias toward the majority class and failed to predict the BLT status class correctly. These results demonstrated that while balancing the dataset improved the model's general performance, it did not entirely eliminate the bias against the minority class.

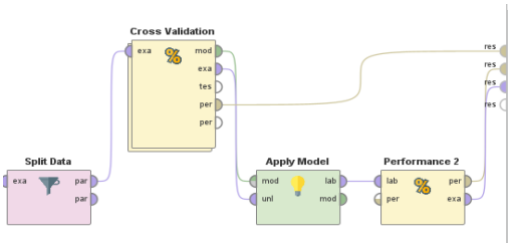

Fig. 2: Data Mining Process Model

3.2. Evaluation

The model evaluation in this study aims to assess the performance of the Random Forest algorithm in classifying Direct Cash Assistance (BLT) recipients and non-recipients in the village of Danalampah. The evaluation process not only measures the overall classification success rate but also evaluates the model's sensitivity to the minority class (BLT recipients), which is often overlooked due to data imbalance. The following is an explanation of each evaluation metric used Accuracy measures how well the model is able to classify all data correctly, both for BLT recipients and non-recipients. It is calculated using the formula:

$$Akurasi = \frac{TP + TN}{TP + TN + FP + FN}$$

In the original dataset (272 data points), the model's accuracy was relatively high (>99%), but this was primarily due to the dominance of the majority class (non-BLT recipients). When the dataset was balanced, the accuracy slightly decreased (~98%), but this drop indicated that the model became fairer in handling both classes. Precision measures how many of the positive predictions (BLT recipients) made by the model were actually correct, compared to all positive predictions made. The precision formula is:

$$Presisi = \frac{TP}{TP + FP}$$

The precision for the BLT recipient class (1) was 95.24%, indicating that the model was good at predicting BLT recipients without much misclassification of the non-recipient data. However, the precision for the BLT recipient status class (18) was 0%, as the model could not recognize this class at all. However, the precision for the BLT recipient status class (18) was 0%, because the model could not recognize this class at all. On the balancing dataset, the precision for the BLT (1) recipient class remained high at 95.24%, indicating an increase in sensitivity without sacrificing prediction accuracy. However, the BLT recipient status class (18) was still not recognized, with a precision of 0%. Recall measures how well the model is able to detect actual BLT recipients. The recall formula is:

$$Recall = \frac{TP}{TP + FN}$$

The recall for the BLT recipient class (1) was 100%, indicating that the model was able to recognize all BLT recipient data. However, the recall for the BLT recipient status class (18) was 0%, as the model did not recognize this class at all. On the balancing result dataset: Recall for the BLT recipient class (1) remained 100%, reflecting the model's increased sensitivity to minority data after balancing. Recall for the BLT recipient status class (18) remained 0%, indicating that balancing the data was not enough to help the model recognize this class.

3.3. Impact of Data Imbalance on Model Performance

The original dataset consists of 22 BLT recipient data (minority class) and 250 non-BLT recipient data (majority class). This significant imbalance causes the model to be biased toward the majority class, as the Random Forest algorithm more frequently identifies patterns from the class with more data. The evaluation results for the original dataset show a very high accuracy (99.59%). However, this accuracy is dominated by the model's ability to recognize the majority class. The model was able to predict the non-BLT class with a precision of 99.56% and recall of 100%, but it failed to recognize the BLT recipient status class (class 18), as shown by a precision and recall of 0% for this class.

This indicates that the model is biased toward the majority class, making it unable to provide fair results for all classes. The data imbalance also means that although the recall for the BLT recipient class (the primary minority class) is high (100%), it is still supported by overall accuracy that does not truly reflect the sensitivity to the minority class. The model overlooks patterns in certain minority data, such as the BLT recipient status class (class 18), which has very few data points. The data imbalance also makes high accuracy not reflective of the fairness of the model's performance. As explained by (Rafrastaraa et al., 2023), in imbalanced datasets, accuracy often masks the model's weaknesses in recognizing the minority class. This finding is consistent with the literature that states metrics like accuracy alone are insufficient for evaluating models on datasets with imbalanced class distributions.

3.4. Accuracy, Precision, and Recall Values in Two Dataset Scenarios

In the original dataset, the accuracy reached 99.59%, but the model was biased toward the majority class. The precision for the BLT recipient class (the primary minority class) was 95.24%, indicating that the model was fairly good at classifying BLT recipients without many misclassifications of non-BLT data. The recall for the BLT recipient class was 100%, meaning the model was able to recognize all BLT recipients without error. However, the model failed to recognize the BLT status class (class 18), with both precision and recall at 0%. In the balanced dataset, the class distribution became more even with 22 BLT recipient data and 44 non-BLT recipient data.

The evaluation results showed a slight decrease in accuracy to 98.51%, but precision (95.24%) and recall (100%) for the BLT recipient class remained the same. However, in this balanced dataset, the performance for the BLT status class remained unchanged. The BLT status class in the balanced dataset was still not recognized, with precision and recall at 0%. These findings support the statement made by [7], which mentions that data balancing can improve the model's sensitivity to the primary minority class, but this technique may not be sufficient to address classes with extremely small amounts of data.

3.5. The Role of Data Balancing in Improving Model Sensitivity

Manual data balancing was performed by reducing the number of majority class data to achieve a more balanced ratio (1:2). The results showed that data balancing helped maintain high performance for the BLT recipient class (the primary minority class) without significantly sacrificing accuracy. However, this balancing was not sufficient to address the bias toward certain classes (BLT status class, class 18). The limitation of manual data balancing is evident from the model's failure to recognize the BLT status class, even though the majority class data had been reduced. This occurred because manual data balancing only focused on redistributing data without adding new relevant information for the minority class.

In this case, the model still failed to understand patterns in classes with very little data. Manual data balancing also risks removing important information from the majority class, which can affect the model's ability to make overall accurate predictions. The accuracy slightly decreased to 98.51%, reflecting the loss of some patterns from the majority class. These results align with findings from [3] who stated that manual data balancing, such as data removal, can eliminate important information from the majority class without adding new information for the minority class. Furthermore, the model's inability to recognize certain classes supports Mualfah's research.

4. Conclusion

This study aims to address the challenge of data imbalance in the classification of BLT (Bantuan Langsung Tunai) recipients using the Random Forest algorithm. Based on the results of the study, the following conclusions can be drawn: The data imbalance in the original dataset caused the model to be biased toward the majority class, with a high accuracy of 99.59% that did not reflect fairness. The model was unable to recognize certain classes, such as BLT status (class 18), which was indicated by precision and recall values of 0%, making the model's performance less representative. In the original dataset, the model showed high accuracy (99.59%) with good sensitivity to the BLT recipient class (precision 95.24%, recall 100%), but it failed to recognize the BLT status class. After data balancing, the accuracy slightly decreased to 98.51%, but sensitivity to the BLT recipient class remained intact, although bias in the BLT status class still persisted. The manual data balancing technique successfully improved sensitivity to the BLT recipient class without significantly sacrificing performance.

However, data balancing was not effective enough to address the bias toward classes with very small amounts of data, such as the BLT status class (class 18), indicating that additional approaches are necessary. In light of these findings, several recommendations are provided for future research and stakeholders involved in this area. These recommendations include efforts to overcome the limitations of the current study and directions for further development in the field of imbalanced data classification and the application of the Random Forest algorithm. The study used a manual data balancing technique to reduce the imbalance between the majority and minority classes. However, this technique has limitations in handling classes with very small amounts of data, such as the BLT status class (class 18). Future research is recommended to use more advanced balancing techniques, such as SMOTE or ADASYN, which can synthetically increase the number of minority class data without reducing the majority class data, thereby enhancing the model's sensitivity to all classes.

The dataset used in this study was limited in size, particularly for certain minority classes such as BLT status. Future research should consider using larger and more representative datasets, either by collecting additional data from other regions or by integrating data from various sources. With a larger and more diverse dataset, the model can learn more relevant patterns, improving its generalization capability and fairness in classification. The Random Forest algorithm used in this study performed well for both the majority and primary minority classes but struggled with extremely small classes. Therefore, future research should explore other algorithms, such as XGBoost, Gradient Boosting, or Neural Networks, which have more adaptive mechanisms for handling data imbalance and can offer more optimal performance, particularly for extreme minority classes.

References

[1] Junaidi Satrio, R. Valicia Anggela, and D. Kariman, "Klasifikasi Metode Data Mining untuk Prediksi Kelulusan Tepat Waktu Mahasiswa dengan Algoritma Naïve Bayes, Random Forest, Support Vector Machine (SVM) dan Artificial Neural Nerwork (ANN)," *J. Appl. Comput. Sci. Technol.*, vol. 5, no. 1, pp. 109–119, 2024, doi: 10.52158/jacost.v5i1.489.

[2] K. G. Putra, "Penentuan Penerima Bantuan Program Keluarga Harapan menggunakan Algoritma Random Forest di Desa Kebonrejo," *Innov. J. Soc. Sci. Res.*, vol. 4, pp. 7242–7257, 2024.

[3] D. Mualfah, W. Fadila, and R. Firdaus, "Teknik SMOTE untuk Mengatasi Imbalance Data pada Deteksi Penyakit Stroke Menggunakan Algoritma Random Forest," *J. CoSciTech (Computer Sci. Inf. Technol.*, vol. 3, no. 2, pp. 107–113, 2022, doi: 10.37859/coscitech.v3i2.3912.

[4] R. Firdaus *et al.*, "Implementasi Algoritma Random Forest Untuk Klasifikasi Pencemaran Udara di Wilayah Jakarta Berdasarkan Jakarta Open Data," *J. FASILKOM*, vol. 14, no. 2, pp. 520–525, 2021.

[5] D. Sudrajat, A. I. Purnamasari, A. R. Dikananda, D. A. Kurnia, and A. Bahtiar, "Klasifikasi Mutu Pembelajaran Hybrid berdasarkan Algoritma C.45, Random Forest dan Naïve Bayes dengan Optimasi Bootsrap Areggating (Bagging) pada masa COVID-19," *JURIKOM (Jurnal Ris. Komputer)*, vol. 9, no. 6, p. 2227, 2022, doi: 10.30865/jurikom.v9i6.5179.

[6] G. Ashari Rakhmat and W. Mutohar, "Prakiraan Hujan menggunakan Metode Random Forest dan Cross Validation," *J. MIND J. | ISSN*, vol. 8, no. 2, pp. 173–187, 2023, [Online]. Available: https://doi.org/10.26760/mindjournal.v8i2.173-187

[7] D. Benaya and B. Prasetiyo, "Implementasi Random Forest dalam Klasifikasi Kanker Paru-Paru," *JOINTER J. Informatics Eng.*, vol. 5, no. 01, pp. 27–31, 2024, doi: 10.53682/jointer.v5i01.331.