JITK (JURNAL ILMU PENGETAHUAN DAN TEKNOLOGI KOMPUTER)
VOL. 11. NO. 2 NOVEMBER 2025
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v11i2.6956

# COMPARATIVE ANALYSIS OF CLASSIFICATION ALGORITHMS IN HANDLING IMBALANCED DATA WITH SMOTE OVERSAMPLING APPROACH

**Agung Nugroho[1*]; Wiyanto[1]; Donny Maulana[1]**

Informatics Engineering[1]
Universitas Pelita Bangsa, Bekasi, Indonesia[1]
www.pelitabangsa.ac.id[1]
agung@pelitabangsa.ac.id*, wiyanto@pelitabangsa.ac.id, donny.maulana@pelitabangsa.ac.id

(*) Corresponding Author
(Responsible for the Quality of Paper Content)



***Abstract***—*Most machine learning algorithms tend to yield optimal results when trained on datasets with balanced class proportions. However, their performance usually declines when applied to data with significant class imbalance. To address this issue, this study utilizes the Synthetic Minority Oversampling Technique (SMOTE) to improve class distribution before model training. Several classification algorithms were employed, including Decision Tree, K-Nearest Neighbors, Logistic Regression, Support Vector Machine, and Random Forest. Experimental results reveal that the Random Forest model produced the highest accuracy (95.70%) and the best F1-score, demonstrating a well-balanced trade-off between precision and recall. In contrast, the Logistic Regression algorithm achieved the highest recall (74.20%), indicating better sensitivity in identifying positive instances despite a lower F1-score. These outcomes highlight the importance of choosing appropriate classification methods based on the specific evaluation goals whether prioritizing accuracy, recall, or overall model balance.*

***Keywords***: *classification, imbalanced data, logistic regression, random forest, SMOTE.*

**Intisari**—*Sebagian besar algoritma klasifikasi menunjukkan kinerja yang baik pada dataset dengan distribusi kelas yang seimbang. Namun, kinerja klasifikasi cenderung menurun ketika menghadapi dataset yang tidak seimbang. Penelitian ini mengatasi permasalahan ketidakseimbangan kelas dengan menerapkan metode SMOTE (Synthetic Minority Oversampling Technique) untuk menyeimbangkan distribusi data. Beberapa algoritma klasifikasi diuji, antara lain Decision Tree, K-Nearest Neighbours, Logistic Regression, Support Vector Machine, dan Random Forest. Berdasarkan hasil pengujian algoritma Random Forest memperoleh akurasi tertinggi sebesar 95,70% serta F1-score tertinggi, yang mencerminkan keseimbangan antara precision dan recall. Sementara itu, algoritma Logistic Regression menghasilkan nilai recall tertinggi sebesar 74,20%, meskipun dengan F1-score yang lebih rendah, yang mengindikasikan kemampuannya dalam mendeteksi kasus positif meskipun keseimbangan prediksi secara keseluruhan menurun. Temuan ini menegaskan pentingnya pemilihan algoritma klasifikasi yang disesuaikan dengan tujuan spesifik, apakah untuk memaksimalkan akurasi, recall, atau keseimbangan prediksi secara keseluruhan.*

***Kata Kunci***: *klasifikasi, data tidak seimbang, regresi logistik, random forest, SMOTE.*

## INTRODUCTION

In many cases, real-time applications produce enormous amounts of data. Classification becomes challenging because to the growing volume of data, its unbounded size, and its imbalance. The most significant issue in data mining is class imbalance[1]. When one of the two classes has more data samples than the other, it is known as the class imbalance problem [2][3].

Data is generally a dataset that has an imbalance in class distribution or referred to as a

dataset with unbalanced classes. Unbalanced data is data that has a significant imbalance in the number of samples between one class and another[4][5][6]. In minority class data, errors often occur in the classification process, therefore optimization is needed on this data.

Many studies have addressed class imbalance with several different approaches to classification algorithms including the Random Forest Classifier [7], Logistic Regression[5][8], Decision Tree Classifier, K-Nearest Neighbours, and Support Vector Machine. Random Forest Classifier provides good performance results in processing a large number of datasets [9], [10]. However, similar research recommends optimizing the Random Forest Classifier to obtain a more optimal classification model [11], [12].

In a number of studies, the problem of class imbalance in data, both unsupervised and supervised, was overcome through the combined application of the K-Nearest Neighbors algorithm and the SMOTE method[13]. This approach is carried out by producing synthetic samples in minority classes through SMOTE to improve data distribution, then using KNN to determine the proximity relationship between samples to improve classification accuracy. The results show that the use of both methods simultaneously can improve the performance of classification models, especially in the introduction of minority classes. The effectiveness of the combination of SMOTE and KNN is shown through an increase in G-mean and F-measure values as indicators of model performance balance [14].

Support Vector Machine algorithm with some oversampling and undersampling techniques is also used in the classification of unbalanced data[15][16]. The use of SVM gives good results, however, similar studies convey a bias in the accuracy value due to the smaller AUC and F-measure values [17][16]. While neglecting or incorrectly classifying minority samples, the bulk of classification algorithms concentrate on classifying majority samples [18]. Rare yet crucial samples in the data calculation process are known as minority samples [19][20].

There are many methods available for unbalanced data classification including algorithmic approaches and data preprocessing approaches[21], [22], [23]. Each of these techniques has advantages and disadvantages. With the data preprocessing approach, several techniques are used in the optimization process, including oversampling the data [24][25]. This research focuses on data preprocessing techniques with an oversampling approach.

Oversampling procedures typically increase the minority class's proportion in the sample beyond its initial proportion. Typically, when it comes to classification modeling, the minority observations are replicated [24][26]. This study's methodology is SMOTE oversampling. The SMOTE or Synthetic Minority Over-Sampling Technique is a technique that uses artificially generated or synthetic data to balance the amount of data from big classes with minor classes [27][28].

This study uses five algorithms. Decision Tree, K-Nearest Neighbors, Logistic Regression, Support Vector Machine, and Random Forest. Because they represent different approaches to classification. Logistic Regression is a linear model, KNN is distance-based, SVM focuses on margins, Decision Tree is interpretable, and Random Forest is an ensemble method.

The SMOTE approach is applied to balance the distribution of data between classes. This method creates new synthetic samples in minority classes without changing the existing data patterns. This technique was chosen because it was able to maintain the characteristics of the original data distribution. Unlike many previous studies that focused mainly on accuracy, this research highlights recall and F1-score as more reliable metrics for imbalanced datasets, especially since accuracy alone can be misleading when one class dominates the data.

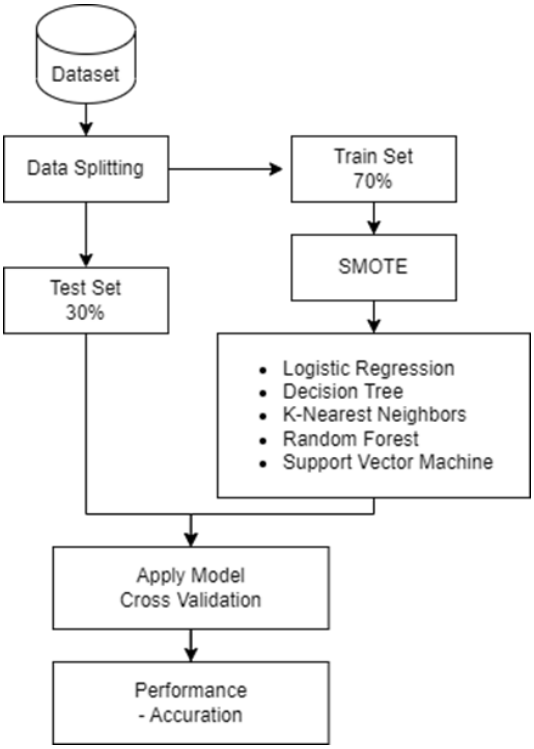
Most previous studies that applied SMOTE with classification algorithms have primarily focused on accuracy as the main evaluation metric [1][3][25]. However, accuracy alone may be misleading in imbalanced datasets, and fewer works have emphasized minority class metrics such as recall and F1-score [5][6]. This study addresses that gap by comparing five algorithms with SMOTE on bankruptcy data, highlighting the trade-off between accuracy and the ability to detect minority cases [13][27][28].

## MATERIALS AND METHODS

The research method used in this study adopts an experimental design that aims to compare classification algorithm models with oversampling techniques using the SMOTE method on several algorithms including Decision Tree Classifier, Logistic Regression, Support Vector Machine, K-Nearest Neighbors, and Random Forest Classifier. To provide a clearer understanding of the research design, Figure 1 shows the process flow scheme used.


Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 11. NO. 2 NOVEMBER 2025
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v11i2.6956




Source: (Research, 2024)

Figure 1. Research scheme

The Taiwan Economic Journal's bankruptcy data from 1999 to 2009 was gathered via Kaggle and used in this analysis [29]. The data set belongs to the category of unbalanced data.

To get more objective model evaluation findings, the dataset was divided into two subsets at the beginning of the study: 70% for training and 30% for testing. Additionally, in order to balance the sample count between the majority and minority classes, the SMOTE approach was used to the training data. Therefore, prior to the training process, the model can learn from more representative data.

To address the issue of unbalanced classes in data, machine learning uses SMOTE, an oversampling technique [30]. In the SMOTE method, synthetic data for a minority class is generated through interpolation between adjacent data points within the feature space, rather than in the original data space. This approach allows for the addition of minority class sample variations without changing the characteristics of the data distribution [27]. This technique adds instances of the minority class by extracting samples of existing minority data using random samples drawn from the k-nearest neighbour values [31][32]. Thus, SMOTE generates new synthetic examples that can expand the decision area of the minority class. The SMOTE method creates synthetic data by increasing the quantity of data in the minor class until it equals the major class [33]. The k-nearest neighbor value of the minor class is used to generate the synthetic data. By creating synthetic data, SMOTE aims to balance the class data. SMOTE was implemented with k=5 nearest neighbors, which is the default parameter widely adopted in previous works [30][31].

The SMOTE technique is effective in correcting data imbalances through the addition of synthetic samples in minority classes. However, the high similarity between synthetic data and original data can increase the likelihood of overfitting, especially in models with complex structures. Therefore, its application requires careful consideration to ensure better generalization. Calculating the distance between data in the minority data is the first step in the SMOTE process. Next, the percentage value of SMOTE is determined, followed by the number of k closest and, finally, the creation of citation data [24]. These stages are described in equation 1.

$$x_{syn} = x_i + (x_{knn} - x_i) \times \delta \quad (1)$$

With $x_{syn}$ is the synthesized data to be generated, $x_i$ is the data to be replicated, $x_{knn}$ is data closest to the data to be replicated and the value of d is a random value between 0 and 1. After the training data is balanced by the SMOTE approach, an evaluation stage is carried out using 10-fold cross-validation to test the performance and stability of the classification model. This technique helps guarantee that the outcomes are not only reliant on a particular selection of data but also demonstrate the model's capacity to identify trends generally throughout the dataset.

Cross-validation is a statistical technique used to measure how well a model or algorithm is able to generalize to new data [34]. This process involves separating the dataset into a training subset and a test subset, each of which is used to build and evaluate the model in turn. There are several cross-validation models, generally a k-fold validation model is used. K-fold validation is used because it can reduce computation time[35]. The k value is the number of iterations used. 10-fold validation is one of the recommended k-fold validation for selecting the best model, because it can provide maximum accuracy estimation [36].

Assessment indicators are very important to evaluate the performance of any classification algorithm. There are many classification assessment indicators including accuracy value, recall value and F1-score value. Accuracy is the percentage of target and non-target samples that are correctly predicted and reflects the ability of

classifying a portion of sample data (test data) to determine all samples (train data)[37]. Accuracy measurement is not influenced by the amount of data but also by the unbalanced data used. The accuracy can be measured by equation 2.

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} x\ 100 \quad (2)$$

In binary classification evaluation, each prediction can be mapped into four possible outcomes. Positive samples that were correctly predicted were recorded as TP, while positive samples that were incorrectly predicted as negative were represented as FN. For the negative class, the correct prediction is recorded as TN, while the incorrect prediction (negative is predicted as positive) falls into the FP category.

Precision measures the consistency of the model in producing correct positive predictions compared to all positive predictions produced. Meanwhile, recall measures the extent to which the model is able to detect all positive instances that actually exist in the dataset. The precission metric is calculated using equation 4 and recall using equation 3.

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

$$Precission = \frac{TP}{TP+FP} \quad (4)$$

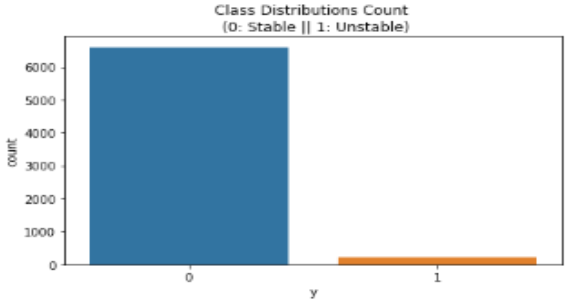
The F1-score measure balances the model's ability to locate all available positive data (recall) and precisely identify positive data (precision) [37]. The F1-score value, which runs from 0 to 1, is obtained by calculating the harmonic mean between the two measurements. A higher F1-score indicates that the model is better able to balance memory and precision. The calculation formula is shown in equation (5).

$$F1 - score = 2\ \frac{precission * recall}{precission + recall} \quad (5)$$

F1-score is frequently employed as the primary metric in classification with an imbalanced data distribution because it may evaluate model performance more fairly than accuracy. Unlike accuracy, which just calculates the percentage of true predictions, the F1-score considers the balance between the majority and minority classes to better indicate the model's ability to detect a little amount of data. F1-score is therefore thought to be more representative when assessing model performance on datasets with class disparity.

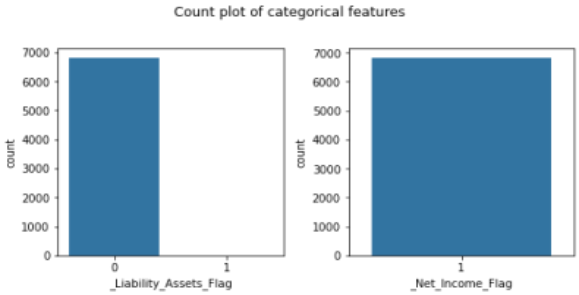
## RESULTS AND DISCUSSION

This study uses the bankruptcy dataset obtained from the Kaggle platform as the main data source. Before the model training process was carried out, the data was processed using the SMOTE method to balance the distribution between classes. After the balancing process is completed, several classification algorithms are applied to analyze the oversampling data and measure the accuracy level of each model. The bankruptcy dataset consists of 96 columns and 6819 data records which have a Financially stable data distribution of 96.77% with 6599 records and financially unstable of 3.23% with 220 records. The data is included in the unbalanced dataset category, so that if classification is carried out, it will produce a low accuracy value due to the dominance of one class, namely 96.77%. The data distribution comparison graph of each class can be seen in Figure 2.




Source: (Research Results, 2024)
Figure 2. Data class distribution

Of the 96 existing features, there is one feature, namely the Net income flag column, which only contains the value 1 across all records. Since there is no variation in this feature, it does not provide any discriminative information for classification and was therefore removed from the dataset before further processing.




Source: (Research Results, 2024)
Figure 3. Categorical features


Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 11. NO. 2 NOVEMBER 2025
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v11i2.6956

Therefore, the "Net income flag" column is removed from the dataset before the next process. The next step is to re-sample the training data using the SMOTE oversampling technique. The data is split into test and train data, with 70% of the data being train and 30% being test, before to resampling. Obtained 4773 train data and 2046 test data with the number of Financially stable data classes of 4619 and the number of Financially unstable data of 154. After re-sampling using SMOTE, the class distribution becomes the same, namely with a total data of 4619 for all data classes. Table 1 presents a comparison of the data before and after the re-sampling process using the SMOTE oversampling technique.

Table 1. Data train with SMOTE

| Label (y) | row data | row data with SMOTE |
|---|---|---|
| 0 | 4619 | 4619 |
| 1 | 154 | 4619 |

Source: (Research Results, 2024)

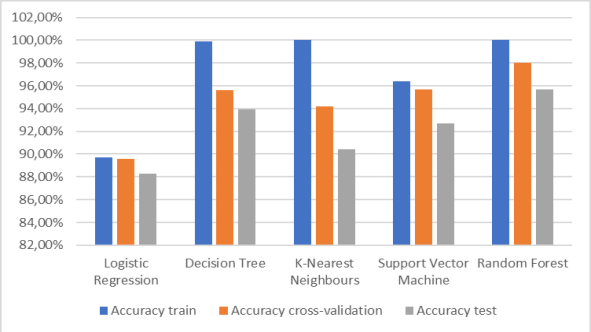
After the re-sampling process, balanced data is obtained which is ready for the next process. The next stage is to train data using several algorithms, namely Logistic Regression, Decision Tree, K-Nearest Neighbours, Support Vector Machine, and Random Forest Classifier with testing using cross validation. Training is carried out with a total of 10 iterations on cross validation to find out the comparative results of each algorithm trained.

Table 2. Data train with SMOTE

| Algorithm | Accuracy train | Accuracy cross-validation | Accuracy test |
|---|---|---|---|
| Decision Tree | 99,90% | 95,60% | 93,90% |
| K-Nearest Neighbors | 100,00% | 94,20% | 90,40% |
| Logistic Regression | 89,70% | 89,60% | 88,30% |
| Support Vector Machine | 96,40% | 95,70% | 92,70% |
| Random Forest | 100,00% | 98,00% | 95,70% |

Source: (Research Results, 2024)

The training process on SMOTE *oversampling* data is carried out on each algorithm that will be compared. The first training was carried out on the Logistic Regression algorithm, the training accuracy value was 89.70%, the cross-validation accuracy value was 89.60%, and the testing accuracy result was 88.30%. Then, Decision Tree showed performance improvements with 99.90% training accuracy, 95.60% cross-validation, and 93.90% testing. The K-Nearest Neighbors (KNN) model gave very high results in training with 100% accuracy, but decreased slightly in validation and testing with scores of 94.20% and 90.40%. Furthermore, the Support Vector Machine (SVM) produces 96.40% training accuracy, 95.70% validation, and 92.70% testing. Meanwhile, Random Forest showed the most optimal performance with 100% training accuracy, 98.00% cross-validation, and 95.70% testing.




Source: (Research Results, 2024)
Figure 4. Training data comparison graph

The number of iterations in the train cross validation process is 10 times. Table 2 and Figure 4 show a comparison of the training result values of several algorithms compared. From the comparison, it can be seen that the K-Nearest Neighbors and Random Forest algorithms reached 100% for the training accuracy value, while the highest cross validation accuracy value was achieved by the Random Forest algorithm. Furthermore, after the data is trained on each algorithm, it is continued by predicting the test data.

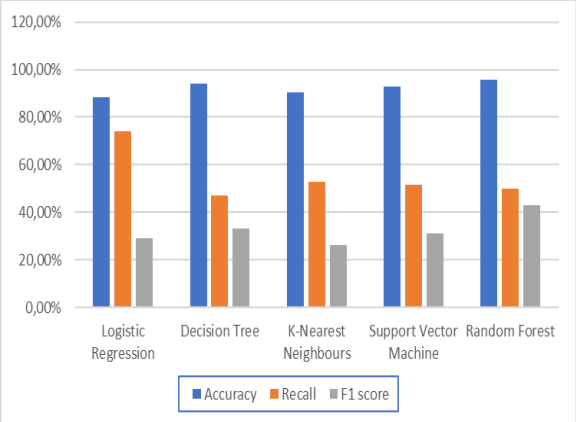
Table 3. Test results

| Algorithm | Accuracy | Precision | Recall | F1 score |
|---|---|---|---|---|
| Logistic Regression | 88.30% | 18.10% | 74.20% | 29.10% |
| Decision Tree | 93.90% | 25.66% | 47.00% | 33.20% |
| K-Nearest Neighbors | 90.40% | 17.40% | 53.00% | 26.20% |
| Support Vector Machine | 92.70% | 22.38% | 51.50% | 31.20% |
| Random Forest | 95.70% | 37.87% | 50.00% | 43.10% |

Source: (Research Results, 2024)

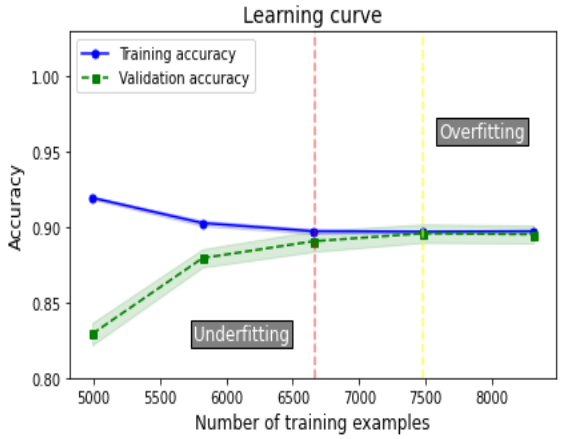
The first test with the Logistic Regression technique yielded an accuracy of 88.30%, recall of 74.20%, and F1-score of 29.10%. The Decision Tree approach performed better in further tests, with an accuracy of 93.90%, a recall of 47.00%, and an F1-score of 33.20%. In contrast, the K-Nearest Neighbors (KNN) algorithm achieved an accuracy of 90.40%, recall of 53.00%, and F1-score of 26.20%. The test results for the Support Vector Machine (SVM) revealed an accuracy of 92.70%, recall of 51.50%, and F1-score of 31.20%. The Random Forest algorithm, which was used in the last test,

produced the highest accuracy of 95.70%, recall 50.00%, and F1-score of 43.10%.




Source: (Research Results, 2024)
Figure 5. Accuracy comparison chart

The number of iterations in the cross-validation testing process is 10 times. Table 3 and Figure 5 show a comparison of the test result values of several algorithms compared. From this comparison, Random Forest has the highest accuracy of 95.70% with a recall value of 50.00%, while the highest recall value is Logistic Regression with a value of 74.20% with an accuracy value of 88.30%. This shows that Logistic Regression is able to find most of the data from the actual data class.




Source: (Research Results, 2024)
Figure 6. Learning curve of Logistic Regression algorithm

Figure 6 shows the performance of the training dataset on the Logistic Regression algorithm. The model experiences overfitting on a fairly small amount of data. This can be seen from the considerable difference in model performance between the training data and the testing data at a small amount of data. After the amount of data reaches around 6500, the performance difference between the training data and the testing data starts to shrink and the two curves start to show a flatter trend. This shows that the model is good enough to predict the new data and does not experience overfitting anymore.

The Random Forest algorithm was proven to provide 95.70% accuracy, higher than the accuracy value achieved by other algorithms in this test, the Random Forest algorithm is able to predict data with a high level of accuracy. However, the recall value of the Random Forest algorithm is only 50.00%, which means that this algorithm is less effective in finding most of the data from the actual data class. Meanwhile, the Logistic Regression algorithm has the highest recall value with a value of 74.20%, which shows that this algorithm is able to find most of the data from the actual data class. Despite having a slightly lower accuracy value compared to the Random Forest algorithm, which is 88.30%, the Logistic Regression algorithm remains a good choice for some types of classification problems.

However, it should be noted that the Logistic Regression algorithm's learning curve shows signs of overfitting at a fairly small amount of data. This suggests that the model may not have reached its peak, and it is still possible to optimize it with other methods to get even better accuracy values. Therefore, to improve the performance of the Logistic Regression model, better optimization and parameter adjustment are required. Furthermore, potential overfitting issues are not limited to Logistic Regression. The Decision Tree and K-Nearest Neighbors algorithms achieved extremely high training accuracies (99.90% and 100%, respectively) but considerably lower test accuracies (93.90% and 90.40%). This discrepancy suggests that both models may have memorized the training data rather than generalized effectively, which is a known limitation of high-variance classifiers such as decision trees and instance-based methods like KNN.

The results indicate that high accuracy does not always reflect good performance in imbalanced datasets. Random Forest reached the highest accuracy but had low recall, while Logistic Regression achieved the highest recall despite lower accuracy. These differences stem from the characteristics of each algorithm, such as overfitting in Decision Trees or robustness in ensemble methods. However, since no statistical significance test was conducted, the claim of “highest accuracy” or “highest recall” should be viewed with caution, as the differences between algorithms may not be statistically meaningful.


Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 11. NO. 2 NOVEMBER 2025
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v11i2.6956

## CONCLUSION

With an accuracy of 95.70% and a recall of 50.00%, Random Forest appears as the algorithm with the best performance in the experiments that have been conducted. These results show that the algorithm is superior to other algorithms tested. Meanwhile, the Logistic Regression algorithm obtained the highest recall value of 74.20%, with an accuracy of 88.30%. These results indicate that Logistic Regression is more effective in identifying most of the actual positive cases in the dataset. However, the learning curve analysis shows that the Logistic Regression model tends to experience overfitting when trained on a relatively small amount of data. Despite this, the model has not yet reached its optimal performance, suggesting that further optimization or the use of alternative methods may improve its accuracy and overall performance. Future research can try ensemble methods, cost-sensitive approaches, or deep learning models, and should use multiple datasets with statistical significance testing to make the results more reliable.

## ACKNOWLEDGEMENTS

This research is sponsored by Universitas Pelita Bangsa, so that this research can be carried out properly and hopefully can be useful for the knowledge of science.

## REFERENCE

[1] M. Altalhan, A. Algarni, and M. Turki-Hadj Alouane, "Imbalanced Data Problem in Machine Learning: A Review," *IEEE Access*, vol. 13, no. January, pp. 13686–13699, 2025, doi: 10.1109/ACCESS.2025.3531662.

[2] Y. E. Kurniawati and Y. D. Prabowo, "Model optimisation of class imbalanced learning using ensemble classifier on over-sampling data," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 11, no. 1, p. 276, Mar. 2022, doi: 10.11591/ijai.v11.i1.pp276-283.

[3] K. Ghosh, C. Bellinger, R. Corizzo, P. Branco, B. Krawczyk, and N. Japkowicz, *The class imbalance problem in deep learning*, vol. 113, no. 7. Springer US, 2024. doi: 10.1007/s10994-022-06268-8.

[4] A. Kumar, S. Goel, N. Sinha, and A. Bhardwaj, "A Review on Unbalanced Data Classification," 2022, pp. 197–208. doi: 10.1007/978-981-19-0332-8_14.

[5] Y. Li, N. Adams, and T. Bellotti, "A Relabeling Approach to Handling the Class Imbalance Problem for Logistic Regression," *Journal of Computational and Graphical Statistics*, vol. 31, no. 1, pp. 241–253, 2022, doi: 10.1080/10618600.2021.1978470.

[6] F. Kamalov, F. Thabtah, and H. H. Leung, "Feature Selection in Imbalanced Data," *Annals of Data Science*, vol. 10, no. 6, pp. 1527–1541, Dec. 2023, doi: 10.1007/s40745-021-00366-5.

[7] Yoga Religia, Agung Nugroho, and Wahyu Hadikristanto, "Klasifikasi Analisis Perbandingan Algoritma Optimasi pada Random Forest untuk Klasifikasi Data Bank Marketing," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 5, no. 1, pp. 187–192, Feb. 2021, doi: 10.29207/resti.v5i1.2813.

[8] L. Zhang, T. Geisler, H. Ray, and Y. Xie, "Improving logistic regression on the imbalanced data by a novel penalized log-likelihood function," *J Appl Stat*, vol. 49, no. 13, pp. 3257–3277, 2022, doi: 10.1080/02664763.2021.1939662.

[9] A. S. More and D. P. Rana, "Performance enrichment through parameter tuning of random forest classification for imbalanced data applications," *Mater Today Proc*, vol. 56, pp. 3585–3593, 2022, doi: 10.1016/j.matpr.2021.12.020.

[10] T. E. Tarigan, E. Susanti, M. I. Siami, I. Arfiani, A. A. Jiwa Permana, and I. M. Sunia Raharja, "Performance Metrics of AdaBoost and Random Forest in Multi-Class Eye Disease Identification: An Imbalanced Dataset Approach," *International Journal of Artificial Intelligence in Medical Issues*, vol. 1, no. 2, pp. 84–94, 2023, doi: 10.56705/ijaimi.v1i2.98.

[11] J. Dong and Q. Qian, "A Density-Based Random Forest for Imbalanced Data Classification," *Future Internet*, vol. 14, no. 3, 2022, doi: 10.3390/fi14030090.

[12] A. Newaz, M. S. Mohosheu, M. A. Al Noman, and T. Jabid, "iBRF: Improved Balanced Random Forest Classifier," *Conference of Open Innovation Association, FRUCT*, pp. 501–508, 2024, doi: 10.23919/fruct61870.2024.10516372.

[13] Eva Y Puspaningrum, Yisti Vita Via, Chilyatun Nisa, Hendra Maulana, and Wahyu S.J.Saputra, "Oversampled-Based Approach to Overcome Imbalance Data in the Classification of Apple Leaf Disease with

SMOTE," *Technium: Romanian Journal of Applied Sciences and Technology*, vol. 16, pp. 112–117, Oct. 2023, doi: 10.47577/technium.v16i.9968.

[14] N. M. Djafar and A. Fauzan, "Implementation of K-Nearest Neighbor using the oversampling technique on mixed data for the classification of household welfare status," *Statistics in Transition new series*, vol. 25, no. 1, pp. 109–124, Mar. 2024, doi: 10.59170/stattrans-2024-007.

[15] I. Print, A. F. Pulungan, and D. Selvida, "Kombinasi Metode Sampling pada Pengklasifikasian Data Tidak Seimbang Menggunakan Algoritma Support Vector Machine ( SVM )," *InfoTekJar : Jurnal Nasional Informatika dan Teknologi Jaringan*, vol. 6, no. 2, pp. 276–282, 2022.

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

[17] N. S. Ramadhanti, W. A. Kusuma, and A. Annisa, "Optimasi Data Tidak Seimbang pada Interaksi Drug Target dengan Sampling dan Ensemble Support Vector Machine," *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 7, no. 6, p. 1221, Dec. 2020, doi: 10.25126/jtiik.2020762857.

[18] L. Yuningsih, G. A. Pradipta, D. Hermawan, P. D. W. Ayu, D. P. Hostiadi, and R. R. Huizen, "IRS-BAG-Integrated Radius-SMOTE Algorithm with Bagging Ensemble Learning Model for Imbalanced Data Set Classification," *Emerging Science Journal*, vol. 7, no. 5, pp. 1501–1516, Oct. 2023, doi: 10.28991/ESJ-2023-07-05-04.

[19] D. E, M. Zhang, J. Liu, H. Jiang, and K. Mao, "RE-SMOTE: A Novel Imbalanced Sampling Method Based on SMOTE with Radius Estimation," *Computers, Materials & Continua*, vol. 81, no. 3, pp. 3853–3880, 2024, doi: 10.32604/cmc.2024.057538.

[20] R. Wardoyo, I. M. A. Wirawan, and I. G. A. Pradipta, "Oversampling Approach Using Radius-SMOTE for Imbalance Electroencephalography Datasets," *Emerging Science Journal*, vol. 6, no. 2, pp. 382–398, Mar. 2022, doi: 10.28991/ESJ-2022-06-02-013.

[21] A. U. Reddy, K. T. Devi, B. B. Vamsi, Anushka, and S. Shareefunnisa, "Enhancing Predictive Performance in Binary Classification on Imbalanced Data Using Automated Methodology," in *2024 2nd World Conference on Communication & Computing (WCONF)*, IEEE, Jul. 2024, pp. 1–6. doi: 10.1109/WCONF61366.2024.10692288.

[22] S. Das, "A new technique for classification method with imbalanced training data," *International Journal of Information Technology*, vol. 16, no. 4, pp. 2177–2185, Apr. 2024, doi: 10.1007/s41870-024-01740-1.

[23] A. Damari, Taghfirul Azhima Yoga Siswa, and Wawan Joko Pranoto, "Implementation of the PSO-SMOTE Method on the Naive Bayes Algorithm to Address Class Imbalance in Landslide Disaster Data," *INOVTEK Polbeng - Seri Informatika*, vol. 10, no. 1, pp. 332–343, Jan. 2025, doi: 10.35314/7wcvrb72.

[24] Asniar, N. U. Maulidevi, and K. Surendro, "SMOTE-LOF for noise identification in imbalanced data classification," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 6, pp. 3413–3423, 2022, doi: 10.1016/j.jksuci.2021.01.014.

[25] L. Wang, M. Han, X. Li, N. Zhang, and H. Cheng, "Review of Classification Methods on Unbalanced Data Sets," *IEEE Access*, vol. 9, pp. 64606–64628, 2021, doi: 10.1109/ACCESS.2021.3074243.

[26] D. Ariyadi, T. A. Y. Siswa, and R. Rudiman, "Penerapan Metode PSO-SMOTE Pada Algoritma Random Forest Untuk Mengatasi Class Imbalance Data Bencana Tanah Longsor," *Kesatria : Jurnal Penerapan Sistem Informasi (Komputer dan Manajemen)*, vol. 6, no. 1, pp. 320–329, Jan. 2025, doi: 10.30645/kesatria.v6i1.574.

[27] A. Nugroho and E. Rilvani, "Penerapan Metode Oversampling SMOTE Pada Algoritma Random Forest Untuk Prediksi Kebangkrutan Perusahaan," *Techno.Com*, vol. 22, no. 1, pp. 207–214, Feb. 2023, doi: 10.33633/tc.v22i1.7527.

[28] E. Ileberi, Y. Sun, and Z. Wang, "Performance Evaluation of Machine Learning Methods for Credit Card Fraud Detection Using SMOTE and AdaBoost," *IEEE Access*, vol. 9, pp. 165286–165294, 2021, doi: 10.1109/ACCESS.2021.3134330.

VOL. 11. NO. 2 NOVEMBER 2025
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v11i2.6956

[29] F. Soriano, "Company Bankruptcy Prediction Dataset," Kaggle. [Online]. Available: https://www.kaggle.com/datasets/fedesoriano/company-bankruptcy-prediction

[30] L. Yuningsih, G. A. Pradipta, D. Hermawan, P. D. W. Ayu, D. P. Hostiadi, and R. R. Huizen, "IRS-BAG-Integrated Radius-SMOTE Algorithm with Bagging Ensemble Learning Model for Imbalanced Data Set Classification," *Emerging Science Journal*, vol. 7, no. 5, pp. 1501–1516, Oct. 2023, doi: 10.28991/ESJ-2023-07-05-04.

[31] G. A. Pradipta, R. Wardoyo, A. Musdholifah, I. N. H. Sanjaya, and M. Ismail, "SMOTE for Handling Imbalanced Data Problem : A Review," in *2021 Sixth International Conference on Informatics and Computing (ICIC)*, IEEE, Nov. 2021, pp. 1–8. doi: 10.1109/ICIC54025.2021.9632912.

[32] A. Zhang, H. Yu, Z. Huan, X. Yang, S. Zheng, and S. Gao, "SMOTE-RkNN: A hybrid re-sampling method based on SMOTE and reverse k-nearest neighbors," *Inf Sci (N Y)*, vol. 595, pp. 70–88, 2022, doi: 10.1016/j.ins.2022.02.038.

[33] G. A. Pradipta, R. Wardoyo, A. Musdholifah, and I. N. H. Sanjaya, "Radius-SMOTE: A New Oversampling Technique of Minority Samples Based on Radius Distance for Learning From Imbalanced Data," *IEEE Access*, vol. 9, pp. 74763–74777, 2021, doi: 10.1109/ACCESS.2021.3080316.

[34] A. Nugroho and Y. Religia, "Analisis Optimasi Algoritma Klasifikasi Naive Bayes menggunakan Genetic Algorithm dan Bagging," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 5, no. 3, pp. 504–510, Jun. 2021, doi: 10.29207/resti.v5i3.3067.

[35] A. Wibowo, "10 Fold Cross Validation." Accessed: Dec. 23, 2020. [Online]. Available: https://mti.binus.ac.id/2017/11/24/10-fold-cross-validation

[36] Y. N. FUADAH, I. D. UBAIDULLAH, N. IBRAHIM, F. F. TALININGSING, N. K. SY, and M. A. PRAMUDITHO, "Optimasi Convolutional Neural Network dan K-Fold Cross Validation pada Sistem Klasifikasi Glaukoma," *ELKOMIKA: Jurnal Teknik Energi Elektrik, Teknik Telekomunikasi, & Teknik Elektronika*, vol. 10, no. 3, p. 728, Jul. 2022, doi: 10.26760/elkomika.v10i3.728.

[37] D. Rajput, W. J. Wang, and C. C. Chen, "Evaluation of a decided sample size in machine learning applications," *BMC Bioinformatics*, vol. 24, no. 1, pp. 1–17, 2023, doi: 10.1186/s12859-023-05156-9.