CITIZEN: Jurnal Ilmiah Multidisiplin Indonesia
Vol 5, No. 4, 2025
ISSN: 2807-5994
https://journal.das-institute.com/index.php/citizen-journal

A COMPARATIVE STUDY FOR IMBALANCED DATA TECHNIQUES OF CLASSIFICATION ALGORITHMS

Dede Brahma Arianto[1*], Siti Nurrahmasita[2]
[1]Informatika, Universitas Faletehan, Banten, Indonesia[1*]
[2]Informatika, Universitas Syiah Kuala, Aceh, Indonesia[2]
Corresponding e-mail: dedebrahma@uf.ac.id

DOI: 10.53866/jimi.v5i4.949

Abstract

One of the main challenges in data processing using machine learning is the imbalanced data distribution, where minority classes are often underrepresented, leading to biased predictions in classification algorithms such as K-Nearest Neighbors (KNN), Naive Bayes, and Support Vector Machine (SVM). This study aims to address this issue by applying Random Undersampling (RUS), Synthetic Minority Oversampling Technique (SMOTE), and hybrid approaches such as SMOTETomek. Using the NHANES dataset, this study evaluates the effectiveness of these methods in reducing bias and improving classification performance. The hybrid sampling technique performed the best, increasing sensitivity to minority classes, resulting in more balanced predictions. Models tested using metrics such as accuracy, precision, recall, and F1-score showed that SVM achieved the highest accuracy of 98.8% after hyperparameter tuning. This study also emphasizes the importance of hyperparameter optimization, including parameters such as C and gamma for SVM, k values for KNN, and smoothing factors for Gaussian Naive Bayes, to improve model reliability. These findings emphasize the importance of effective data preprocessing techniques and model optimization in dealing with imbalanced datasets. Implementing these approaches will ensure more accurate data analysis, as well as provide valuable insights for decision-making and policies aimed at improving imbalanced case.

Keywords: *Imbalanced Data, Machine Learning, Classification, RUS, SMOTE, SMOTETomek*

Abstrak

Salah satu tantangan utama pada pengolahan data menggunakan machine learning adalah ketidakseimbangan distribusi data, di mana kelas minoritas sering kali kurang terwakili, yang menyebabkan prediksi yang bias pada algoritma klasifikasi seperti K-Nearest Neighbors (KNN), Naive Bayes, dan Support Vector Machine (SVM). Penelitian ini bertujuan untuk mengatasi permasalahan tersebut dengan menerapkan teknik Random Undersampling (RUS), Synthetic Minority Oversampling Technique (SMOTE), dan pendekatan hibrid seperti SMOTETomek. Dengan menggunakan dataset NHANES, penelitian ini mengevaluasi efektivitas metode-metode tersebut dalam mengurangi bias dan meningkatkan kinerja klasifikasi. Teknik sampling hibrid menunjukkan hasil terbaik, meningkatkan sensitivitas terhadap kelas minoritas, yang menghasilkan prediksi yang lebih seimbang. Model yang diuji menggunakan metrik seperti akurasi, presisi, recall, dan F1-score menunjukkan bahwa SVM mencapai akurasi tertinggi sebesar 99,8% setelah penyesuaian hyperparameter. Penelitian ini juga menekankan pentingnya optimasi hyperparameter, termasuk parameter seperti C dan gamma untuk SVM, nilai k untuk KNN, serta smoothing factors untuk Gaussian Naive Bayes, guna meningkatkan keandalan model. Temuan ini menegaskan pentingnya teknik pra-pemrosesan data yang efektif dan optimasi model dalam menangani dataset yang tidak seimbang. Implementasi pendekatan ini akan memastikan analisis data yang lebih akurat, serta memberikan wawasan berharga untuk pengambilan keputusan kesehatan masyarakat dan kebijakan yang bertujuan meningkatkan kesehatan populasi

Kata Kunci: *Data Tidak Seimbang, Pembelajran Mesin, Klasifikasi, RUS, SMOTE, SMOTETomek*

1. Introduction

In today's big data era, data analysis is becoming increasingly important, especially in the context of evidence-based decision-making. One of the major challenges faced in this analysis is imbalanced data, where the distribution of classes in the dataset is uneven (Tariq et al., 2023). For example, in a dataset designed to evaluate the health status of individuals in the United States, age categories such as Senior and Adult may be unequally distributed. This imbalance can affect the performance of commonly used algorithms. It is difficult to make predictions on an imbalanced dataset because the classifier tends to detect the majority class rather than the minority class. Therefore, the output of the classification will be biased (Chen et al., 2021). Several methods are used to overcome this problem. The resampling method is one of the most effective in solving the problem of imbalanced data. In the resampling method, there are several techniques such as undersampling, oversampling and hybrid sampling (Ghorbani & Ghousi, 2020).

Undersampling technique is a random sampling method that selects the majority class and adds it to the minority class (Hoyos-Osorio et al., 2021). Random Undersampling (RUS) is one of the methods used in undersampling techniques, the aim of this method is to balance the classes by removing samples from the majority class, so that the number is the same as the minority class (Untoro & Yusuf, 2023). Meanwhile, oversampling is a sampling method by selecting the minority class randomly and duplicating it. Oversampling is another common sampling approach used to address imbalanced class problems (Ahmed et al., 2022). Synthetic Minority Oversampling Technique (SMOTE) is an oversampling method that can synthetically duplicate data so that the problem of different data distributions can be overcome. Hybrid sampling is a combination of oversampling and undersampling techniques. By combining SMOTE and Tomek-Link (SMOTETomek), it is expected that the resulting accuracy performance will be superior to that achieved by using only one of the data balancing techniques (Hairani et al., 2023).

The machine learning classification process with imbalanced data problems will result in poor performance of the classification algorithm used. classification algorithms, such as K-Nearest Neighbors (KNN), Naive Bayes, and Support Vector Machine (SVM). KNN is a method that classifies data based on proximity to other data points. Although KNN is easy to implement and often produces good results, it tends to favor the majority class in situations where one class is much more dominant. This can result in low sensitivity to the minority class, which is often individuals with higher health risks (Syahira & Arianto, 2024). On the other hand, Naive Bayes, which operates on the assumption of independence between features, also faces similar challenges. When applied to imbalanced datasets, these models can produce biased predictions, reducing the accuracy in identifying individuals with more critical health status (Ericha Apriliyani & Salim, 2022). Meanwhile, SVM works by finding a hyperplane that separates classes in the feature space. Although SVM is known for its ability to handle complex data, class imbalance can cause this model to be less sensitive to the minority class. This has the potential to ignore individuals with higher health risks, which is a serious problem in healthcare applications (Indra Buana & Brahma Arianto, 2024).

To overcome this imbalance data problem, various techniques can be applied, such as oversampling, undersampling, and the use of algorithms that are more sensitive to class imbalance. Oversampling can increase the number of examples from the minority class, while undersampling can reduce the number of examples from the majority class. In addition, the use of more appropriate evaluation metrics, such as Confusion Matrix and Classification Report, can provide a more accurate picture of the model's performance in the context of imbalanced data. Considering the challenges faced by KNN, Naive Bayes, and SVM in handling unbalanced data, it is important to apply the right technique so that the classification model can provide more accurate and relevant results.

2. Research Methodology

This research method will consist of several systematic research stages. The scheme of the flow of these stages can be seen in Figure 1.

CITIZEN: Jurnal Ilmiah Multidisiplin Indonesia
Vol 5, No. 4, 2025
ISSN: 2807-5994
https://journal.das-institute.com/index.php/citizen-journal

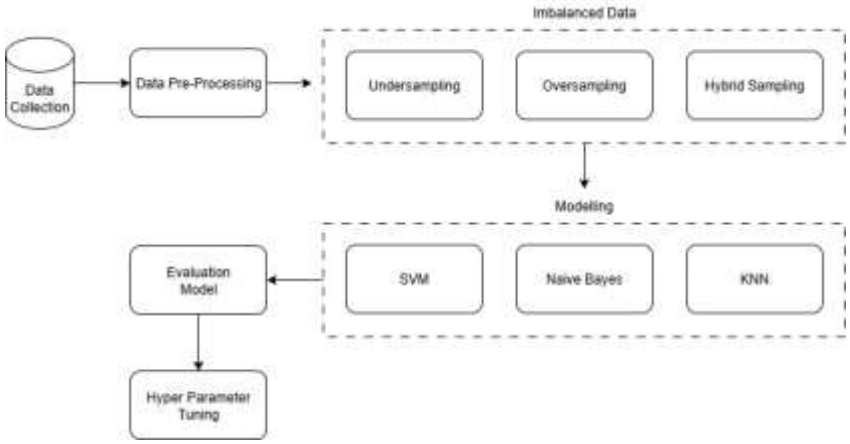

Figure. 1. Research Flowchart

2.1. Data Collection

Data sourced from a open source dataset from National Health and Nutrition Examination Survey (NHANES), this public data has 10 attributes with a total of 2278 records. The following is a table of data that has been collected. Figure 2 shows NHANES data

	A	B	C	D	E	F	G	H	I	J
1	SEQN	age_grou	RIDAGEYR	RIAGENDR	PAQ605	BMXBMI	LBXGLU	DIQ010	LBXGLT	LBXIN
2	73564.0	Adult	61.0	2.0	2.0	35.7	110.0	2.0	150.0	14.91
3	73568.0	Adult	26.0	2.0	2.0	20.3	89.0	2.0	80.0	3.85
4	73576.0	Adult	16.0	1.0	2.0	23.2	89.0	2.0	68.0	6.14
5	73577.0	Adult	32.0	1.0	2.0	28.9	104.0	2.0	84.0	16.15
6	73580.0	Adult	38.0	2.0	1.0	35.9	103.0	2.0	81.0	10.92
7	73581.0	Adult	50.0	1.0	2.0	23.6	110.0	2.0	100.0	6.08
8	73587.0	Adult	14.0	1.0	2.0	38.7	94.0	2.0	202.0	21.11
9	73596.0	Adult	57.0	2.0	2.0	38.3	107.0	2.0	164.0	20.93
10	73607.0	Senior	75.0	1.0	2.0	38.9	89.0	2.0	113.0	17.47
11	73610.0	Adult	43.0	1.0	1.0	28.9	90.0	2.0	95.0	3.24
12	73618.0	Adult	54.0	2.0	2.0	32.7	98.0	2.0	80.0	7.16
13	73619.0	Adult	36.0	2.0	1.0	27.3	85.0	2.0	91.0	9.86
14	73621.0	Senior	80.0	1.0	2.0	24.6	100.0	2.0	97.0	4.33
15	73633.0	Adult	43.0	2.0	2.0	30.5	102.0	2.0	124.0	12.06
16	73639.0	Senior	71.0	1.0	2.0	30.3	133.0	2.0	295.0	22.92
17	73640.0	Senior	67.0	2.0	1.0	22.1	114.0	2.0	150.0	10.09
18	73642.0	Adult	57.0	2.0	2.0	37.8	96.0	2.0	120.0	9.81
19	73656.0	Adult	54.0	1.0	2.0	28.0	89.0	2.0	82.0	8.39
20	73659.0	Senior	70.0	2.0	2.0	46.1	139.0	2.0	154.0	42.67
21	73661.0	Adult	25.0	1.0	2.0	21.0	86.0	2.0	84.0	4.47
22	73663.0	Adult	25.0	1.0	2.0	33.0	108.0	2.0	77.0	8.02

Figure. 2. NHANES Data

2.2. Data Preprocessing

The data obtained needs to be cleaned, a cleaning process was performed to handle missing values, duplication, and anomalies (outliers). Data cleaning is carried out to ensure data quality before modeling. The process includes:

a) Handling missing values
   Handling missing values with imputation methods or removing irrelevant data.
b) Handling outliers
   Identifying and correcting data that has extreme values to avoid distortion in the analysis results.
c) Handling data duplicate
   Removing duplication to prevent bias in the model.
d) Remove columns
   Removing irrelevant or redundant features, namely unique respondent numbers.
e) Pearson correlation heatmap
   Using a heatmap to identify correlations between features to avoid multicollinearity.

2.3. Imbalanced Data

If the dataset has an unbalanced class distribution, special handling is done to balance the number of samples in each class. The techniques used include:

a) Undersampling

Reducing the number of samples from the majority class using the random under sampling method. Random under sampling (RUS) is a technique that reduces the number of examples from the majority class randomly to achieve balance with the minority class. Although this method can reduce imbalance, it also risks losing important information that can affect model performance (Hairani et al., 2023) . The challenge that arises in classification using machine learning is when the dataset used has an unbalanced class distribution. Therefore, handling the problem of class imbalance becomes very crucial. By using the RUS resampling technique, it can improve model performance on minority classes, reduce the risk of overfitting, and improve the efficiency of model training processing (Liu & Tsoumakas, 2020). RUS method it implements selection procedure with the instances of the majority class in the targeted variable are randomly removed until their number matches the instances of the minority class. Figure 3 demonstrates how RUS works (Wongvorachan et al., 2023).

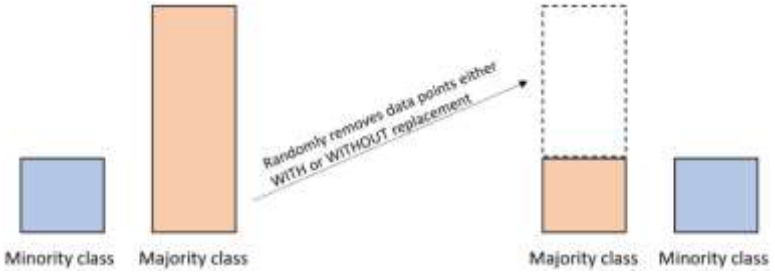

Figure. 3. Random Undersampling Process

b) Oversampling

Increasing the number of samples in the minority class using methods such as SMOTE (Synthetic Minority Over-sampling Technique). SMOTE is an oversampling technique that generates synthetic examples for minority classes by interpolating between existing examples. By increasing the representation of minority classes, SMOTE helps classification models to be more sensitive to underrepresented classes (Elreedy et al., 2024). In practice, the use of SMOTE can result in excessive oversampling and blur the boundaries between classes. SMOTE will determine and select the instances that need to be generated during the oversampling process, so it can overcome the problem of blind oversampling that often occurs in SMOTE and can reduce imbalance in the dataset (Wardoyo et al., 2022). SMOTE operates by randomly duplicating data points of the minority class with replacement until the proportion of the two classes is balanced. Figure 4 demonstrates how SMOTE works (Wongvorachan et al., 2023).

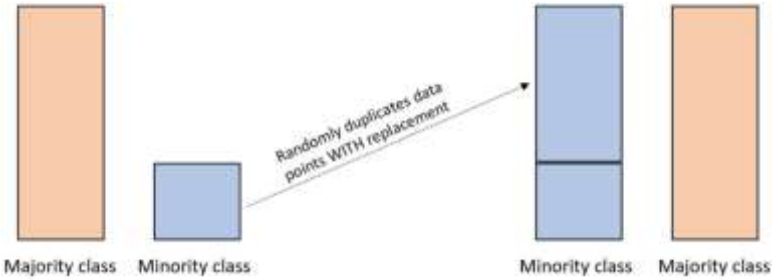

Figure. 4. Random Oversampling Process

c) Hybrid Sampling
Combining oversampling and undersampling techniques such as the SMOTETomek method for more optimal results. SMOTETomek is a combination of SMOTE and Tomek's Link, which aims to improve the quality of the dataset by adding synthetic examples and removing irrelevant examples. This approach aims to reduce bias and improve the accuracy of the classification model (Khleel & Nehéz, 2023). SMOTETomek's distinctive methodology, integrating oversampling and undersampling strategies, seeks to enhance the precision of data cleaning while harnessing the advantages of synthetic data generation. Figure 5 demonstrates how SMOTETomek works (Wongvorachan et al., 2023).

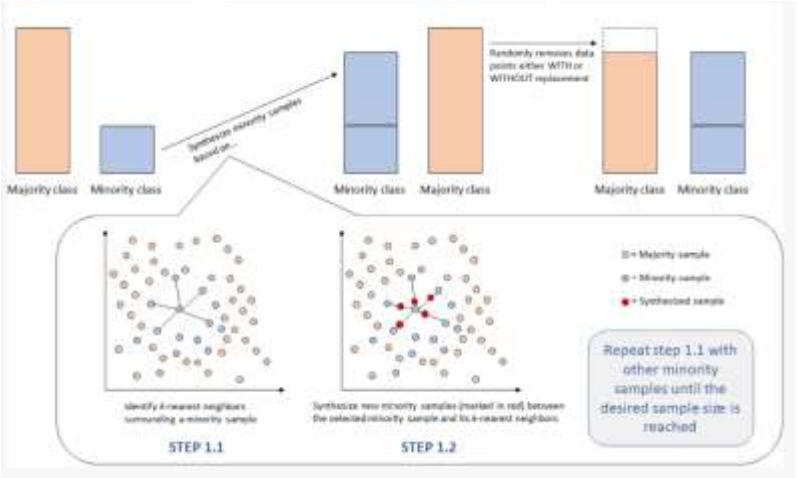

Figure. 5. Hybrid Sampling Process

2.4. Data Splitting

The dataset is divided into several parts to ensure a good model training and evaluation process. The dataset is divided into three main parts, such as training data (80%), validation data (10%) and test data (10%). This proportional division of data ensures a fair and accurate evaluation of the model performance. The model will be trained and evaluated based on the patterns present in this data to ensure optimal classification ability.

2.5. Modelling

In the modeling stage, this study has been balanced using hybrid sampling technique is built using various selected algorithms. The first algorithm used is Support Vector Machine (SVM), SVM An effective algorithm for classification by maximizing the margin between classes using hyperplane. The second is Gaussian Naïve Bayes (GaussianNB), a probabilistic model that is suitable for data with Gaussian distribution assumptions. The third is K-Nearest Neighbors (KNN) Classifier, a method that determines classes based on proximity to the k nearest neighbors.

2.6. Evaluation Model

This stage aims to measure the performance of the model using evaluation matrices such as accuracy, precision, recall, F1-score, and confusion matrix. A good evaluation ensures that the model has good generalization capabilities to new data and meets the research objectives. The evaluation results are used to determine the best model or make further improvements.

2.7. Hyperparameter Tuning

Each model was trained and tested with default parameters and hyperparameters to determine its strengths and weaknesses. To improve the performance of each model, hyperparameter adjustments were made as follows. 1) SVM: The parameters used are C, gamma and kernel. 2) GaussianNB: The parameters used is var_smoothing. 3) KNN: The parameters used are metric, n_neighbors, weight.

3. Result and Discussion

3.1. Data Overview

The NHANES (National Health and Nutrition Examination Survey) dataset is used in this study to predict the age group category of individuals with the labels:

a) Adult (< 65 years) = 0
b) Senior (≥ 65 years) = 1

This dataset has an unbalanced class distribution, where the amount of data for the senior category is much less than the adult category. Figure 6 represent class distribution on target column.

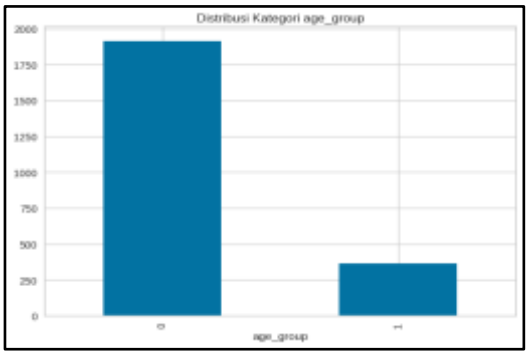

Figure. 6. Class Distribution on Target Column

In Figure 6 above, it can be seen that there is an imbalance case where the data distribution is uneven on the age variable. This uneven distribution can cause the model to tend to be biased towards the majority class. This may cause the model that is built to have low accuracy to errors in making predictions. To handle this problem, it is necessary to have a process stage for handling imbalanced data.

3.2. Handling Imbalanced Data

To handle data imbalance, several sampling techniques are:

a) Undersampling

This technique reduces the amount of data from the majority class (Adult) to balance the amount of data with the minority class (Senior). In this study, the undersampling technique used is RUS. Figure 7 shows the results of the class distribution in the target column using the undersampling technique.

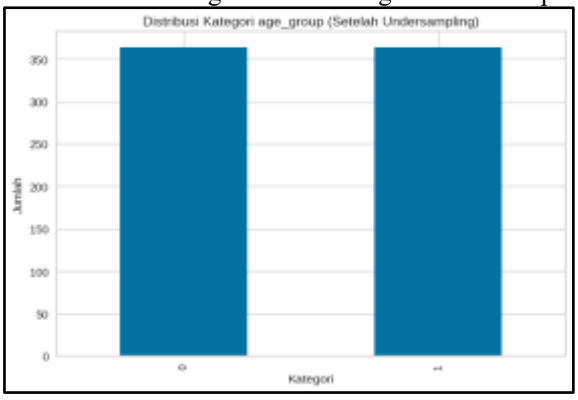

Figure. 7. Class Distribution on Target Column With Undersampling

b) Oversampling

This technique uses a method by increasing the number of minority class data with the SMOTE (Synthetic Minority Over-sampling Technique) technique. SMOTE works by creating synthetic data

among existing samples to balance the amount of data between the two classes. Figure 8 shows the results of the class distribution in the target column using the oversampling technique.

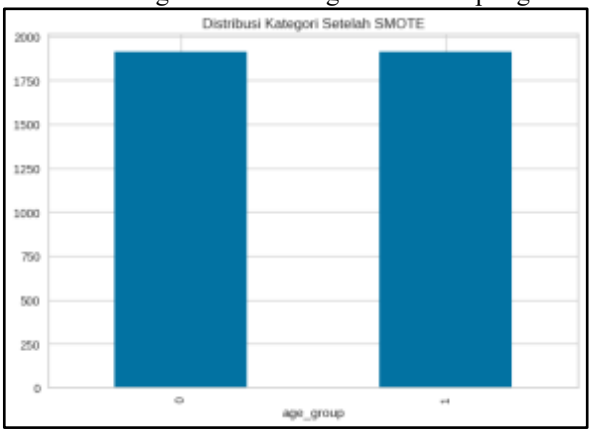

Figure. 8. Class Distribution on Target Column With Oversampling

c) Hybrid Sampling

This technique uses a combination of oversampling and undersampling with the SMOTETomek method to overcome imbalance more effectively. Figure 9 shows the results of the class distribution in the target column using the hybrid sampling technique.

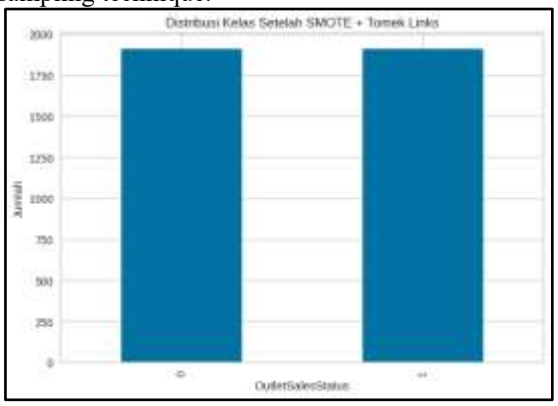

Figure. 9. Class Distribution on Target Column With Hybrid Sampling

After the sampling technique was carried out with the three methods, such undersampling, oversampling and hybrid sampling. The results of the comparison of the three sampling methods above were obtained. The comparison of the results of the application of sampling techniques is shown in Table 1:

Table 1. Comparison of sampling techniques

Sampling Method	**Label**	
	0	**1**
Original	1914	364
Undersampling (RUS)	364	364
Oversampling (SMOTE)	1914	1914
Hybrid Sampling (SMOTETomek)	1811	1811

3.3. Modelling

In the modeling stage, classification was carried out using three machine learning algorithms, such Support Vector Machine (SVM), Gaussian Naive Bayes (GaussianNB) and K-Nearest Neighbors (KNN). Each method was tested with a dataset that had been overcome by the imbalanced case problem using sampling techniques such as Undersampling (RUS), Oversampling (SMOTE), and Hybrid Sampling (SMOTETomek). The purpose of this experiment was to compare the performance of the three models after data balancing. After classification modeling was carried out using the three algorithm models and the imbalanced case method, it was found that the KNN algorithm model with the SMOTETomek method obtained the highest accuracy value of 97.1%. The results of the accuracy comparison on the models are shown in table 2 below.

Table 2. Comparison of machine learning algorithm models using the imbalanced case method

Sampling Method	**Model Accuracy**		
	SVM	**GaussianNB**	**KNN**
Original	95.8%	94.7%	95.8%
Undersampling (RUS)	91.8%	95%	95.2%
Oversampling (SMOTE)	96.1%	93.5%	96.9%
Hybrid Sampling (SMOTETomek)	96.6%	95.3%	**97.1%**

3.4. Evaluation Model

Evaluation is done by comparing several performance metrics such as Accuracy, Precision, Recall, and F1-Score. Figure 10 shows that the evaluation results on the SVM algorithm model have an F1-Score value of 96.6%. Then Figure 11 shows that the evaluation results on the naive bayes algorithm model have an F1-Score value of 95.3%. And Figure 12 shows that the evaluation results on the KNN algorithm model have an F1-Score value of 97.1%.

```
=== SVM ===
Accuracy: 0.966
Confusion Matrix:
 [[347  26]
 [  0 392]]
Classification Report:
              precision    recall  f1-score   support

           0      1.000     0.930     0.964       373
           1      0.938     1.000     0.968       392

    accuracy                          0.966       765
   macro avg      0.969     0.965     0.966       765
weighted avg      0.968     0.966     0.966       765
```

Figure. 10. SVM algorithm model evaluation results

```
=== NB ===
Accuracy: 0.953
Confusion Matrix:
 [[350  23]
 [ 13 379]]
Classification Report:
              precision    recall  f1-score   support

           0      0.964     0.938     0.951       373
           1      0.943     0.967     0.955       392

    accuracy                          0.953       765
   macro avg      0.953     0.953     0.953       765
weighted avg      0.953     0.953     0.953       765
```

Figure. 11. Naïve Bayes algorithm model evaluation results

```
=== KNN ===
Accuracy: 0.971
Confusion Matrix:
 [[351  22]
 [  0 392]]
Classification Report:
              precision    recall  f1-score   support

           0      1.000     0.941     0.970       373
           1      0.947     1.000     0.973       392

    accuracy                          0.971       765
   macro avg      0.973     0.971     0.971       765
weighted avg      0.973     0.971     0.971       765
```

Figure. 12. KNN algorithm model evaluation results

3.5. Hyperparameter Tuning

After the algorithm modeling and model evaluation are done. Furthermore, at this stage, the process of changing parameters in each algorithm model is carried out which aims to improve the performance of each model. Each algorithm is given a different parameter value according to the best parameters. Table 3 shows the accuracy results after hyper parameter tuning.

Table 3. Model Evaluation Results With Hyperparameter Tuning

Model	Accuracy	Best Param
KNN	98.2%	'metric': 'manhattan', 'n_neighbors': 7, 'weights': 'distance'
SVM	**98.8%**	'C': 1, 'gamma': 0.001, 'kernel': 'poly'
Naïve Bayes	94.7%	'var_smoothing': 1e-06

In table 3 above, it can be concluded that SVM with tuning is the most powerful model because it has a very good accuracy value with an accuracy of 98.8%, but KNN also offers a good alternative with very high accuracy. Naive Bayes, although simpler, is more suitable for cases with clearer data distributions.

4. Conclusion

This study has shown that data balancing techniques such as oversampling, undersampling, and hybrid sampling can improve the accuracy and performance of classification models on imbalanced datasets. Among the various sampling techniques tested, the use of SMOTETomek gave the best results, addressing the imbalance in a more effective manner. The KNN model showed very good performance when modeling with default parameters. However, after hyperparameter tuning, the SVM model was shown to provide the highest and most stable accuracy compared to KNN and Naive Bayes.

For further research, it is recommended to explore the use of other data balancing techniques, such as ADASYN (Adaptive Synthetic Sampling) or Borderline-SMOTE. In addition, further research can be done to optimize other data processing techniques, such as feature selection or dimensionality reduction, to further improve model performance. The use of ensemble models such as Random Forest or XGBoost can also be tested to see if they can handle data imbalance more effectively

Bibliografi

Ahmed, H. A., Hameed, A., & Bawany, N. Z. (2022). Network intrusion detection using oversampling technique and machine learning algorithms. *PeerJ Computer Science*, *8*, e820. https://doi.org/10.7717/peerj-cs.820

Chen, Y.-R., Leu, J.-S., Huang, S.-A., Wang, J.-T., & Takada, J.-I. (2021). Predicting Default Risk on Peer-to-Peer Lending Imbalanced Datasets. *IEEE Access*, *9*, 73103–73109. https://doi.org/10.1109/ACCESS.2021.3079701

Elreedy, D., Atiya, A. F., & Kamalov, F. (2024). A theoretical distribution analysis of synthetic minority oversampling technique (SMOTE) for imbalanced learning. *Machine Learning*, *113*(7), 4903–4923. https://doi.org/10.1007/s10994-022-06296-4

Ericha Apriliyani, & Salim, Y. (2022). Analisis performa metode klasifikasi Naïve Bayes Classifier pada

Unbalanced Dataset. *Indonesian Journal of Data and Science*, *3*(2), 47–54. https://doi.org/10.56705/ijodas.v3i2.45

Ghorbani, R., & Ghousi, R. (2020). Comparing Different Resampling Methods in Predicting Students' Performance Using Machine Learning Techniques. *IEEE Access*, *8*, 67899–67911. https://doi.org/10.1109/ACCESS.2020.2986809

Hairani, H., Anggrawan, A., & Priyanto, D. (2023). Improvement Performance of the Random Forest Method on Unbalanced Diabetes Data Classification Using Smote-Tomek Link. *JOIV : International Journal on Informatics Visualization*, *7*(1), 258. https://doi.org/10.30630/joiv.7.1.1069

Hoyos-Osorio, J., Alvarez-Meza, A., Daza-Santacoloma, G., Orozco-Gutierrez, A., & Castellanos-Dominguez, G. (2021). Relevant information undersampling to support imbalanced data classification. *Neurocomputing*, *436*, 136–146. https://doi.org/10.1016/j.neucom.2021.01.033

Indra Buana, M., & Brahma Arianto, D. (2024). Analisis Sentimen Ulasan Pengguna Aplikasi ZenPro dengan Implementasi Algoritma Support Vector Machine (SVM). *Adopsi Teknologi dan Sistem Informasi (ATASI)*, *3*(1), 45–52. https://doi.org/10.30872/atasi.v3i1.1092

Khleel, N. A. A., & Nehéz, K. (2023). A novel approach for software defect prediction using CNN and GRU based on SMOTE Tomek method. *Journal of Intelligent Information Systems*, *60*(3), 673–707. https://doi.org/10.1007/s10844-023-00793-1

Liu, B., & Tsoumakas, G. (2020). Dealing with class imbalance in classifier chains via random undersampling. *Knowledge-Based Systems*, *192*, 105292. https://doi.org/10.1016/j.knosys.2019.105292

Syahira, N., & Arianto, D. B. (2024). Prediksi Tingkat Kualitas Udara Dengan Pendekatan Algoritma K-Nearest Neighbor. *Jurnal Ilmiah Informatika Komputer*, *29*(1), 45–59. https://doi.org/10.35760/ik.2024.v29i1.10069

Tariq, M. A., Sargano, A. B., Iftikhar, M. A., & Habib, Z. (2023). Comparing Different Oversampling Methods in Predicting Multi-Class Educational Datasets Using Machine Learning Techniques. *Cybernetics and Information Technologies*, *23*(4), 199–212. https://doi.org/10.2478/cait-2023-0044

Untoro, M. C., & Yusuf, M. A. N. M. (2023). Evaluate of Random Undersampling Method and Majority Weighted Minority Oversampling Technique in Resolve Imabalanced Dataset. *IT Journal Research and Development*, *8*(1), 1–13. https://doi.org/10.25299/itjrd.2023.12412

Wardoyo, R., Wirawan, I. M. A., & Pradipta, I. G. A. (2022). Oversampling Approach Using Radius-SMOTE for Imbalance Electroencephalography Datasets. *Emerging Science Journal*, *6*(2), 382–398. https://doi.org/10.28991/ESJ-2022-06-02-013

Wongvorachan, T., He, S., & Bulut, O. (2023). A Comparison of Undersampling, Oversampling, and SMOTE Methods for Dealing with Imbalanced Classification in Educational Data Mining. *Information*, *14*(1), 54. https://doi.org/10.3390/info14010054