

Application of SMOTE-ENN Method in Data Balancing for Classification of Diabetes Health Indicators with C4.5 Algorithm

Bakti Putra Pamungkas^{[1]*}, Muhammad Jauhar Vikri^[2], Ita Aristia Sa'ida^[3]

Department of Informatics Engineering ^{[1], [2], [3]}

University of Nahdlatul Ulama Sunan Giri

Bojonegoro, Indonesia

baktisensei@gmail.com^[1], vikri@unugiri.ac.id^[2], itaaria@unugiri.ac.id^[3]

Abstract— Data imbalance in health datasets often leads to decreased performance of classification models, especially in detecting minority classes such as diabetics. This study evaluates the effect of the SMOTE-ENN method on improving the performance of the C4.5 algorithm in the classification of diabetes health indicators. The dataset used is the 2021 Diabetes Binary Health Indicators BRFSS from Kaggle, which consists of 236,378 respondent data with unbalanced class distribution: 85.80% non-diabetic and 14.20% diabetic. The SMOTE method was used to add synthetic data to the minority classes, while ENN was applied to remove data considered noise. After balancing, the C4.5 algorithm was used for classification. Evaluation was conducted using accuracy, precision, recall, and F1-score metrics. The results showed that the application of SMOTE-ENN improved accuracy from 79.49% to 80.33% and precision from 29% to 30%. Although the recall value did not increase, this method proved to be able to improve the overall stability of the prediction, especially in terms of the accuracy of the classification of the positive class. The novelty of this research lies in the specific application of the SMOTE-ENN method on large-scale health datasets with the C4.5 algorithm, which has not been widely explored before. Therefore, further exploration of other balancing techniques and algorithms is needed to obtain more optimal classification results on unbalanced data.

Keywords— SMOTE-ENN, Data Imbalance, C4.5, Diabetes, Classification

I. INTRODUCTION

Diabetes is a global health problem with an increasing prevalence. Based on WHO data, more than 422 million people in the world have diabetes, and the disease is responsible for more than 1.5 million deaths each year. The impact is more pronounced in developing countries, where the lack of healthcare facilities is a major challenge in the diagnosis and management of diabetes. In addition, diabetes is also associated with serious complications such as heart disease, kidney damage, neurological disorders, and vascular complications that significantly reduce the quality of life of patients [1].

In health data processing, significant challenges arise from imbalanced data. This imbalance occurs when the amount of data of the minority class (e.g., diabetes cases) is much smaller than the majority class (e.g., non-diabetes). This makes

predictive algorithms tend to be biased towards the majority class, thus reducing the system's ability to detect rare but clinically important critical conditions [2][1].

To overcome this problem, various oversampling techniques have been developed, one of which is SMOTE (Synthetic Minority Over-Sampling Technique). SMOTE generates synthetic data to increase the proportion of minority classes. Research by Rezki et al. (2024) showed that the application of SMOTE can improve the performance of the C5.0, Random Forest, and SVM algorithms in diabetes prediction using the Pima Indian Diabetes dataset. However, they also highlighted the risk of overfitting due to the addition of synthetic data without further cleaning [3][4].

As a more advanced solution, SMOTE-ENN, a combination of SMOTE and Edited Nearest Neighbor, is used to not only augment minority class data but also clean the data from noise. The study by Wang (2022) showed that SMOTE-ENN can improve the accuracy of postoperative complication prediction up to 90% with XGBoost algorithm, emphasizing the importance of the combination of oversampling and data cleaning on medical datasets [4].

Besides SMOTE-ENN, adaptive approaches such as ADASYN (Adaptive Synthetic Sampling) are also being used to balance the data. ADASYN dynamically generates synthetic data based on the classification difficulty of each minority sample. In the study of Marlisa et al. (2024), ADASYN improved accuracy, specificity, and sensitivity in diabetes classification using the K-Nearest Neighbor algorithm, showing that this approach is effective in handling imbalanced data [2].

On the other hand, the C4.5 algorithm is a popular decision tree method due to its ability to handle numerical and categorical data attributes, and classification results that can be interpreted easily. However, the effectiveness of C4.5 in unbalanced datasets remains limited without adequate data balancing techniques [5].

The novelty of this research lies in the application of the combination of SMOTE-ENN and the C4.5 algorithm specifically for the classification of diabetes health indicators,

which has not been widely explored in previous studies. This approach is expected to improve classification performance especially in minority classes and make a real contribution to the development of decision support systems in the field of Health [4][5].

Thus, this research not only extends the application of SMOTE-ENN to medical data, but also presents a more optimal alternative to traditional balancing methods in the effort to diagnose and manage diabetes more accurately and efficiently.

II. RESEARCH METHODS

First, this research begins with the collection of datasets downloaded from Kaggle in Excel or CSV format. Next, a pre-processing stage is carried out which includes cleaning, normalization, outlier handling, and data division. After that, to overcome data imbalance, an oversampling technique is applied using the SMOTE-ENN method, which combines minority data synthesis (SMOTE) and data cleaning using Edited Nearest Neighbor (ENN). The balanced data was then divided into training and test data for model training and evaluation purposes. The classification process is performed using the C4.5 algorithm to build a prediction model. Finally, the classification results were evaluated to measure the performance of the model before the study was concluded.

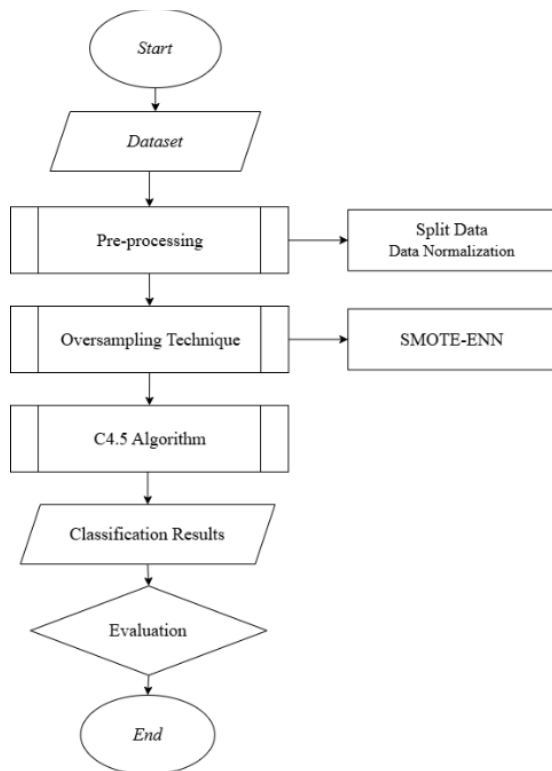


Fig. 1. Research Methods

A. Dataset

The dataset used in this study was obtained from the Kaggle platform due to its relevance to the research objective, which is to build a diabetes prediction model. The data was collected by

accessing the Kaggle website (<https://www.kaggle.com>) and downloading the dataset in CSV format. The dataset used is the Diabetes Binary Health Indicators BRFSS 2021, which consists of 236,378 respondent data with 21 features covering health indicators such as behavior, chronic health conditions, and other risk factors. The target variable in this dataset is `diabetes_binary`, which indicates whether the respondent is indicated to have diabetes or not. The data distribution shows class imbalance, with 85.80% belonging to the non-diabetes class (0.0) and only 14.20% to the diabetes class (1.0). This imbalance is a challenge in the classification process, so data balancing methods such as SMOTE-ENN are applied to improve model performance.

TABLE I. DATASET DIABETES

Diabetes_binary	HighBP	HighChol	...	Education	Income
0.0	0	1.0	...	4.0	5.0
1.0	1	0.0	...	4.0	3.0
1.0	1	1.0	...	4.0	7.0
1.0	0	1.0	...	3.0	4.0
0.0	0	0.0	...	5.0	6.0
0.0	1	0.0	...	4.0	8.0
0.0	1	1.0	...	5.0	3.0
...
0.0	1	0.0	...	4.0	5.0
0.0	0	1.0	...	6.0	10.0
0.0	1	0.0	...	4.0	6.0
0.0	0	1.0	...	6.0	6.0

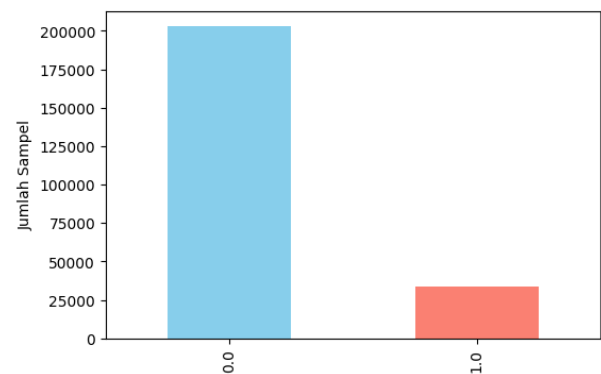


Fig. 2. Comparison of Majority and Minority Class Samples

B. Pre-Processing

The pre-processing stage is an important initial process in data processing before modeling. At this stage, the data that has been collected will be prepared through several steps, such as separating features and labels, dividing data into training and testing data, and normalizing data. This process aims to ensure that the data is in an optimal condition so that it can improve the performance of the model in the classification stage. SMOTE-ENN

1) Separation of Features and Labels

Data separation is done by separating feature attributes (x) and target labels (y) from the dataset [6]. The target label is the diabetes_binary variable, while the other features become attributes used to predict the label [7].

2) Split Data

The dataset is divided into training and testing data using the train_test_split function with a portion of 60% for training and 40% for testing. This separation aims to train the model on training data and evaluate the performance of the model on testing data [8][9].

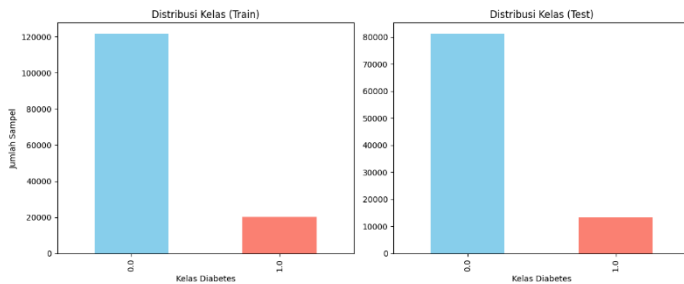


Fig. 3. Split Data

3) Data Normalization

Data features are converted into a certain scale using normalization techniques, such as Min-Max Scaler, so that all attributes are in the same range of values, usually between 0 and 1. This process helps the algorithm work more optimally, especially when the data has a large scale difference between features. Normalization is performed on training data and then applied to testing data to maintain scale consistency [10][11].

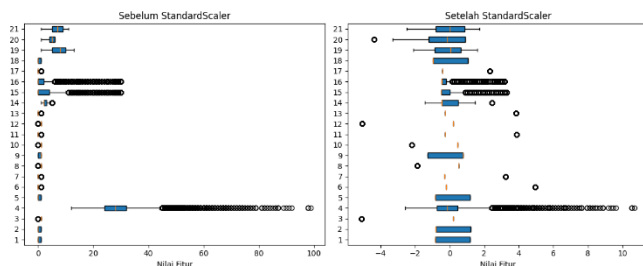


Fig. 4. Data Normalization

amount of data with diabetes. This imbalance can cause the prediction model to be biased towards the majority class and reduce classification performance [4][12].

To overcome this problem, a combination of Synthetic Minority Over-Sampling Technique (SMOTE) and Edited Nearest Neighbors (ENN) method is used. The SMOTE technique works by adding synthetic samples to the minority class, while ENN cleans the data by removing samples that are misclassified or considered noise, thus helping to reduce the risk of overfitting and improve data quality [13][4].

The implementation of SMOTE is done by first identifying the minority class in the dataset, which is the class with a value of 1 in the target variable diabetes_binary. Then, SMOTE is used to generate additional samples based on the nearest neighbor data of the minority class in the training data, so that the distribution between classes becomes more balanced. This step aims to reduce model bias towards the majority class and improve the model's ability to recognize important patterns in both classes. By using the SMOTE-ENN approach, the model is expected to produce more accurate and reliable predictions. The process flow of the SMOTE-ENN method can be seen in the following figure [13].

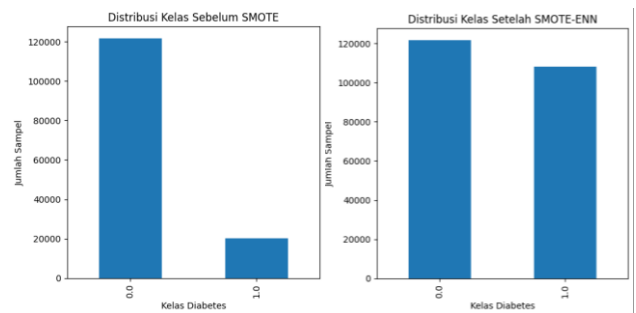


Fig. 5. Before and After SMOTE-ENN Implementations

Figure Explanation:

- Left Panel (Class Distribution Before SMOTE-ENN) Before SMOTE is applied, there is a significant imbalance between class 0 (without diabetes) and class 1 (with diabetes). the number of class 0 samples is much larger than class 1, which may cause bias in the model [14].
- Right panel (class distribution after SMOTE-ENN) After applying enn, the number of minority class samples is slightly reduced. this indicates that enn removes samples that are considered noise or less relevant. the end result is a cleaner and more balanced dataset, which is ready to be used for classification model training [15].

D. C4.5 Algorithm

The C4.5 algorithm is a popular method in data mining used to build decision trees [16]. This algorithm has several advantages, such as being able to handle attributes with continuous and discrete values, overcome attributes with empty

C. SMOTE-ENN

The problem of data imbalance arises when the proportion of the number of majority classes (classes with more samples) and minority classes (classes with fewer samples) is unbalanced. In this Diabetes Health Indicators dataset, imbalance occurs in the target variable diabetes_binary, where the amount of data without diabetes is much greater than the

values (Missing Values), and support the process of pruning decision tree branches to simplify the model [5][17].

The main process of this algorithm involves several steps. First, it calculates the entropy value of the dataset, which is used to measure the level of data uncertainty. the formula used is:

$$Entropy(S) = - \sum_{i=1}^k p_i \log_2 p_i \dots (2.1)$$

Where p_i is the probability of occurrence of class i .

After the entropy value is calculated, the algorithm determines the information gain, which is the reduction of uncertainty after the data is divided based on certain attributes:

$$Gain(S, A) = Entropy(S) - \sum_{v \in values(A)} \frac{|S_v|}{|S|} Entropy(S_v) \dots (2.2)$$

The attribute with the highest gain value is chosen as the root or node of the decision tree [18][19].

However, to avoid bias towards attributes with many categories, C4.5 uses *Gain Ratio*:

$$GainRatio(S, A) = \frac{Gain(S, A)}{SplitInfo(S, A)} \dots (2.3)$$

With

$$SplitInfo(S, A) = \sum_{v \in values(A)} \frac{|S_v|}{|S|} \log^2 \frac{|S_v|}{|S|} \dots (2.4)$$

The attribute with the highest *gain ratio* value will become a node in the decision tree. .

This process continues to repeat until each branch of the tree only contains data with the same class, or no more attributes can be used to further divide the data [20][21]. The algorithm also performs branch pruning to avoid overfitting by removing branches that do not contribute significantly to the accuracy of the model [17][22].

E. Evaluation Metrics

Confusion Matrix is a matrix used in machine learning to evaluate the performance of classification models. This matrix presents the comparison between model predictions and actual values in the form of four elements: *True Positive (Tp)*, *False Positive (Fp)*, *False Negative (Fn)*, and *True Negative (Tn)*. Using these elements, we can calculate various evaluation metrics such as *Accuracy*, *Precision*, *Recall*, and *F1-Score* [23].

Based on the Confusion Matrix results, calculations for several metrics can be done as follows:

1. Accuracy

Accuracy measures how many predictions are correct (both positive and negative) compared to the total amount of data [24].

The equation for calculating accuracy is:

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \times 100\% \dots (3.1)$$

Where:

- TP = *True Positive* (Correct Prediction for Positive Class)
- Tn = *True Negative* (Correct Prediction For Negative Class)
- Fp = *False Positive* (False Prediction For Negative Class)
- Fn = *False Negative* (False Prediction For Positive Class)

2. Precision

Precision measures the accuracy of the positive predictions made by the model, which is how many positive predictions are correct compared to the total positive predictions made [23].

The equation for calculating precision is:

$$Precision = \frac{TP}{TP + FP} \times 100\% \dots (3.2)$$

Where:

- TP = *True Positive*
- Fp = *False Positive*

3. Recall

Recall measures the model's ability to detect true positive classes. It is the ratio between the number of correct positive predictions and the total number of truly positive data [25].

The equation to calculate recall is:

$$Recall = \frac{TP}{TP + FN} \times 100\% \dots (3.3)$$

Where:

- TP = *True Positive*
- FN = *False Negative*

4. F1-score

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \dots (3.4)$$

III. RESULTS AND DISCUSSION

This study compares the performance of the C4.5 algorithm before and after the application of the SMOTE-ENN method to overcome class imbalance on the Diabetes Binary Health Indicators BRFSS 2021 dataset. Before balancing, the model produces 79.49% accuracy, 29% precision, and 31% recall, which shows that the model is less able to detect diabetics as a minority class.

After the application of SMOTE-ENN, the accuracy of the model increased from 79.49% to 80.33% and the precision increased from 29% to 30%. Although the recall value decreased slightly from 31% to 30%, it shows that the balancing method successfully improved the overall prediction stability, especially in terms of the accuracy of predicting positive classes, although the sensitivity to minority classes can still be improved. Table 2 presents a complete comparison of model performance.

matrix:

TABLE II. MODEL OF PERFORMANCE EVALUATION WITH SMOTE-ENN

Metrics	Without SMOTE-ENN	With SMOTE-ENN
Accuracy	79,49%	80,33%
Precision	29%	30%
Recall	31%	30%
F1-Score	30%	30%

Following the results before balancing the data using the SMOTE-ENN method, the model was re-trained and evaluated. The classification results are shown in the following confusion matrix:

TABLE III. CONFUSION MATRIX C4.5 WITHOUT SMOTE-ENN

	Positive Prediction	Negatif Prediction
Positive Actual	4.174	9.193
Negatif Actual	10.198	70.987

TABLE IV. PERFORMANCE EVALUATION OF MODEL WITHOUT SMOTE-ENN

Metrics	Value
Accuracy	79,49%
Precision	29%
Recall	31%
F1-Score	30%

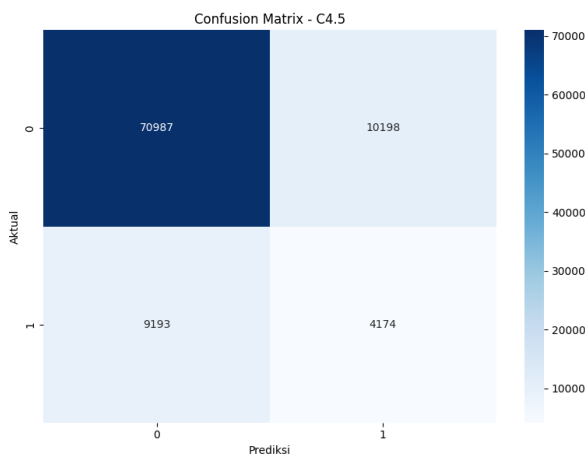


Fig. 6. Confusion Matrix C4.5 without SMOTE-ENN

Following the results after balancing the data using the SMOTE-ENN method, the model was re-trained and evaluated. The classification results are shown in the following confusion

TABLE V. CONFUSION MATRIX C4.5 WITH SMOTE-ENN

	Positive Prediction	Negatif Prediction
Positif Actual	4.070	9.297
Negative Actual	9.298	71.887

TABLE VI. MODEL PERFORMANCE EVALUATION WITH SMOTE-ENN

Metrics	Value
Accuracy	80,33%
Precision	30%
Recall	30%
F1-Score	30%

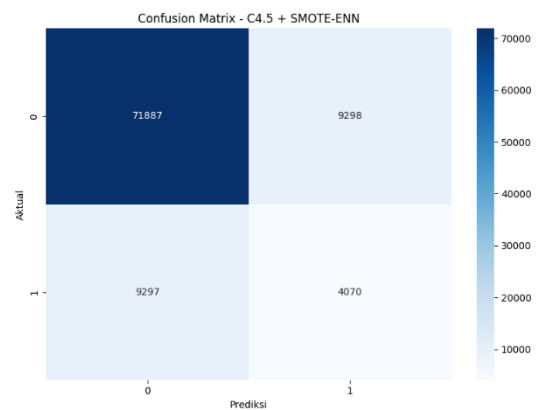


Fig. 7. Confusion Matrix C4.5 with SMOTE-ENN

In general, the improvement in model performance is not statistically significant. However, in practical terms, this balancing process still contributes to the stability of the overall prediction, although it is not optimal in detecting minority classes.

When compared to the research of Marlisa et al. (2024) [2], who used ADASYN and K-NN, they obtained a sensitivity (recall) of 71.79% - much higher than the results in this study. Meanwhile, Rezki et al. (2024) [3] showed that using SMOTE on C5.0, Random Forest, and SVM algorithms resulted in an AUC of up to 0.831. However, they also noted the possibility of overfitting due to synthetic data. Similar results occurred in this study, where precision increased slightly, but recall showed no improvement.

Wang (2022) research [4] used SMOTE-ENN and XGBoost, and managed to obtain 90% accuracy with an AUC of 0.90. This corroborates the notion that the success of the balancing method is highly dependent on the algorithm selection.

The novelty of this research lies in the application of the

SMOTE-ENN method with the C4.5 algorithm on large-scale health datasets, which has not been widely explored. The scientific contribution provided is to extend the evidence that balancing techniques such as SMOTE-ENN need to be combined with more adaptive algorithms to achieve optimal performance in minority classes.

The limitations of this research are that no statistical significance test has been conducted on metric changes, no parameter tuning has been done, and only one classification algorithm (C4.5) has been used without comparison. This research is implemented using Python with the SMOTE-ENN approach and evaluation through Confusion Matrix, but further development is still needed to improve classification performance, especially in detecting minority classes.

IV. CONCLUSIONS

Based on the results of the research that has been done, the C4.5 algorithm without handling data imbalance produces an accuracy of 79.49%, but has a low recall for minority classes, which is 31%. After applying the SMOTE-ENN method, the accuracy increased to 80.33% and the precision for the minority class also increased to 30%. However, this method did not provide a significant improvement to the recall of minority classes. This shows that although SMOTE-ENN is able to balance the data distribution, its effectiveness in improving classification performance is highly dependent on the characteristics of the data and the type of algorithm used.

REFERENCES

- [1] WHO, "Thermostability of human insulin," *World Heal. Organ.* 2024., vol. 2050, no. 1, pp. 1–7, 2024.
- [2] H. Marlisa, N. Satyahadewi, N. Imro'ah, and N. N. Debaratja, "Application of Adasyn Oversampling Technique on K-Nearest Neighbor Algorithm," *BAREKENG J. Ilmu Mat. dan Terap.*, vol. 18, no. 3, pp. 1829–1838, 2024.
- [3] M. K. Rezki, M. I. Mazdadi, F. Indriani, Muliadi, T. H. Saragih, and V. A. Athavale, "Application of Smote to Address Class Imbalance in Diabetes Disease Categorization Utilizing C5.0, Random Forest, and Support Vector Machine," *J. Electron. Electromed. Eng. Med. Informatics*, vol. 6, no. 4, pp. 343–354, 2024.
- [4] J. Wang, "Prediction of postoperative recovery in patients with acoustic neuroma using machine learning and SMOTE-ENN techniques," *Math. Biosci. Eng.* aimspress.com, 2022.
- [5] R. P. Fadhillah, R. Rahma, and ..., "Klasifikasi Penyakit Diabetes Mellitus Berdasarkan Faktor-Faktor Penyebab Diabetes menggunakan Algoritma C4.5," ... *Penelit. dan ...*, 2022.
- [6] R. Doğan, S. M. Çınar, and E. Akarslan, "A Novel ZIP-Based NILM Method Design Robust to Undervoltage and Overvoltage Conditions," *Arab. J. Sci. Eng.*, 2025.
- [7] U. M. Khaire and R. Dhanalakshmi, "Stability of feature selection algorithm: A review," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 34, no. 4, pp. 1060–1073, 2022.
- [8] Q. H. Nguyen *et al.*, "Influence of data splitting on performance of machine learning models in prediction of shear strength of soil," *Math. Probl. Eng.*, vol. 2021, 2021.
- [9] M. T. Akhir, M. Syarat, G. Memperoleh, G. Sarjana, S. Satu, and T. Informasi, *Perbandingan Kinerja Metode Klasifikasi Naïve Bayes Dan Random Forest Dalam Analisis Sentimen Kasus Narkoba di Indonesia Pada Komentar YouTube SKRIPSI Diajukan oleh : NAILUL ' INAYAH PROGRAM STUDI TEKNOLOGI INFORMASI*. 2023.
- [10] A. Ambarwari, Q. J. Adrian, and Y. Herdiyeni, "Analisis Pengaruh Data Scaling Terhadap Performa Algoritme Machine Learning untuk Identifikasi Tanaman," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 4, no. 1, pp. 117–122, 2020.
- [11] S. Nagibzadeh, *Bilgisayar Bilimleri ve Mühendisli ğ i + tr-2*, no. January 2025. 2024.
- [12] 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.
- [13] M. Seyedtabib and N. Kamyari, "Predicting polypharmacy in half a million adults in the Iranian population: comparison of machine learning algorithms," *BMC medical informatics and decision making*. Springer, 2023.
- [14] H. L. Ngo *et al.*, "The composition of time-series images and using the technique SMOTE ENN for balancing datasets in land use/cover mapping," *Acta Montan. Slovaca*, vol. 27, no. 2, pp. 342–359, 2022.
- [15] M. Lu, L. T. Tay, and J. Mohamad-Saleh, "Landslide susceptibility analysis using random forest model with SMOTE-ENN resampling algorithm," *Geomatics, Nat. Hazards Risk*, vol. 15, no. 1, p. , 2024.
- [16] Dhea Halimah, Muhammad Ridwan Lubis, and Widodo Saputra, "Algoritma C4.5 Untuk Menentukan Klasifikasi Tingkat Pemahaman Mahasiswa Pada Matakuliah Bahasa Pemrograman," *J. Tek. Mesin, Ind. Elektro Dan Inform.*, vol. 1, no. 3, pp. 24–38, 2022.
- [17] P. B. N. Setio, D. R. S. Saputro, and Bowo Winarno, "Klasifikasi Dengan Pohon Keputusan Berbasis Algoritme C4.5," *Prism. Pros. Semin. Nas. Mat.*, vol. 3, pp. 64–71, 2020.
- [18] U. P. Budi and A. Info, "Application Of C4 . 5 Algorithm In Disease Classification," vol. 2, no. 02, pp. 58–62, 2024.
- [19] A. Afifuddin and L. Hakim, "Deteksi Penyakit Diabetes Mellitus Menggunakan Algoritma Decision Tree Model Arsitektur C4.5," *J. Krisnadana*, vol. 3, no. 1, pp. 25–33, 2023.
- [20] L. Y. L. Gaol, M. Safii, and D. Suhendro, "Prediksi Kelulusan Mahasiswa Stikom Tunas Bangsa Prodi Sistem Informasi Dengan Menggunakan Algoritma C4.5," *Brahmana J. Penerapan ...*, 2021.
- [21] F. F. Nugraha, I. Sunandar, and C. Julian, "Penerapan Data Mining Dengan Metode Kalsifikasi Menggunakan Algoritma C4.5," *Teknologi*, vol. 7, no. March, pp. 10–20, 2022.
- [22] M. A. Barata *et al.*, "PERANCANGAN SISTEM ELECTRONIC NOSE BERBASIS," pp. 117–126, 2016.
- [23] M. D. Nguyen *et al.*, "Estimation of recompression coefficient of soil using a hybrid ANFIS-PSO machine learning model," *J. Eng. Res.*, vol. 12, no. September 2023, pp. 358–368, 2024.
- [24] V. R. Prasetyo, M. Mercifia, A. Averina, L. Sunyoto, and B. Budiarjo, "Prediksi Rating Film Pada Website Imdb Menggunakan Metode Neural Network," *Netw. Eng. Res. Oper.*, vol. 7, no. 1, p. 1, 2022.
- [25] S. Sathyanarayanan and B. R. Tantri, "Confusion Matrix-Based Performance Evaluation Metrics," no. November, 2024.