Jurnal Ilmu Pendidikan dan Teknologi

https://journalversa.com/s/index.php/jipt Vol. 7, No. 2, April 2025

CLUSTERING ELECTRICITY CUSTOMERS WITH KNN TO DETECT ELECTRICITY MISUSE

Retno Supiyanti[1], Sri Arttini Dwi Prasetyowati[2]

[1,2]Universitas Islam Sultan Agung

Email: oopee.supiyanti@gmail.com[1], arttini@unissula.ac.id[2]

Abstrak: Penelitian ini mengkaji penerapan K-Nearest Neighbors (KNN) untuk mengidentifikasi penyalahgunaan listrik pada pelanggan di Unit Pelaksana Pelayanan Pelanggan Perusahaan Listrik Negara di Kabupaten Demak yang biasa disebut PLN UP3 Demak. Studi ini mengevaluasi efektivitas Teknik Pengambilan Sampel Minoritas Sintetis (SMOTE) dalam mengatasi masalah ketidakseimbangan data. Pola penyalahgunaan listrik dianalisis di seluruh kelompok konsumen, dan dampak SMOTE terhadap akurasi dan sensitivitas prediksi KNN dinilai. Hasil menunjukkan bahwa meskipun SMOTE meningkatkan deteksi instance kelas minoritas, hal ini dapat mengurangi akurasi secara keseluruhan. Sebaliknya, model yang dilatih tanpa SMOTE menunjukkan stabilitas dalam mengidentifikasi pola penggunaan umum, sehingga meminimalkan risiko overfitting. Temuan ini menggarisbawahi trade-off antara mencapai kumpulan data yang seimbang dan menjaga ketepatan klasifikasi dalam deteksi penyalahgunaan listrik.
Kata Kunci: Penyalahgunaan Listrik, KNN, SMOTE, Data Balancing, PLN UP3 Demak, Pengenalan Pola.

Abstract: *This research investigates the application of K-Nearest Neighbors (KNN) for identifying electricity misuse among customers at National Electricity Company Customer Service Implementation Unit in Demak Regency, commonly referred to as PLN UP3 Demak. The study evaluates the effectiveness of the Synthetic Minority Over-sampling Technique (SMOTE) in addressing imbalanced data issues. Patterns of electricity misuse were analyzed across consumer groups, and the impact of SMOTE on KNN's predictive accuracy and sensitivity was assessed. Results indicate that while SMOTE improves the detection of minority class instances, it may reduce overall accuracy. Conversely, models trained without SMOTE demonstrate stability in identifying general usage patterns, minimizing risks of overfitting. The findings underscore the trade-offs between achieving balanced datasets and maintaining classification precision in electricity misuse detection.*
Keywords: *Electricity Misuse, KNN, SMOTE, Data Balancing, PLN UP3 Demak, Pattern Recognition.*

INTRODUCTION

The study investigates the detection of electricity misuse using machine learning techniques at PLN UP3 Demak. It compares the results of the KNN algorithm with and without the use of SMOTE for addressing imbalanced datasets. The goal is to determine the trade-offs

between using SMOTE to balance classes versus maintaining data integrity without synthetic samples [5].

To further evaluate the model, we calculated the Euclidean distance between the test data and the training data. For instance, if we divide the dataset into 10 folds, we can train the model using data from folds 2, 3, 4, and 5. The Euclidean distance formula is employed to measure the distance between test and training samples [6]. For a specific instance, let's assume we calculate the Euclidean distance for the first test data point and the tenth training data point, as follows:

$$\text{Euclidean Distance}=(x1-x2)^2+(y1-y2)^2+\ldots+(n1-n2)^2$$

In this analysis, we find that for k=1k = 1k=1, the model achieves 100% accuracy since all predictions are correct.

Additionally, metrics such as classification reports and confusion matrices are employed to evaluate model performance comprehensively. True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN) are used to calculate overall accuracy, precision, recall, and the F1 score. The Receiver Operating Characteristic (ROC) curve is also generated to analyze the trade-off between true positive rates and false positive rates, with the Area Under the Curve (AUC) serving as a measure of the model's ability to distinguish between positive and negative classes. This methodology provides valuable insights into the model's classification capabilities and potential areas for improvement.

RESEARCH METHOD

Data Collection The dataset used in this study contains electricity consumption data from 115 customers identified as having committed electricity misuse at PLN UP3 Demak, covering the period from February 2023 to April 2024. The dataset includes.

	NOAGENDA	IDPEL	NAMA	GOL	\
0	P2TL/52552/20240104/00007	5.230000e+11	ABDUL MUNIF	P2	
1	P2TL/52554/20240103/00001	5.230000e+11	PARLI	P2	
2	P2TL/52551/20240117/00002	5.230000e+11	MUHAMMAD ZUHRI	P2	
3	P2TL/52551/20240118/00001	5.230000e+11	AHMAT ALI	P2	
4	P2TL/52554/20240116/00003	5.230000e+11	SUPARMIN LASTRI	P2	
..	...	...	...	..	
110	P2TL/52554/20240422/00003	5.230000e+11	KISWAN	P3	
111	P2TL/52552/20240405/00001	5.230000e+11	SANIMAN	P3	
112	P2TL/52551/20240430/00005	5.230000e+11	TK PANGUDI LUHUR	P3	
113	P2TL/52551/20240430/00010	5.230000e+11	KAMJAWI	P3	
114	P2TL/52551/20240417/00005	5.230000e+11	PON PES ROUDLOTUT	P3	

	ALAMAT	DAYA	KWH	Bulan -1	Bulan -2	Bulan -3	...	\
0	DS JAWONG KEMBAN	R1/450	2,479	78.0	65.0	63.0	...	
1	JA MBANGAN BARAT	R1/900	4,957	94.0	79.0	85.0	...	
2	DS NGAWEN WEDUN	R1M/900	4,957	50.0	54.0	49.0	...	
3	DS SABETAN WEDUN	R1M/900	4,957	211.0	205.0	229.0	...	
4	DS PULOREJO KALIS	R1/450	2,479	118.0	116.0	125.0	...	
..	...	...	...	...	...	...	...	
110	MI NUNGAN RT06/04 P	R1/450	2,479	180.0	182.0	200.0	...	
111	DK KARANG KUMPUL	R1/450	2,479	82.0	48.0	36.0	...	
112	DS MERAK DEMPET	S2/450	2,479	10.0	7.0	8.0	...	
113	KA TONSARI	R1M/900	4,957	50.0	36.0	36.0	...	
114	DS KENDUREN RT 3/3	S2/900	4,957	291.0	293.0	276.0	...	

	Bulan -7	Bulan -8	Bulan -9	Bulan -10	Bulan -11	Bulan -12	Bulan -13	\
0	51.0	82.0	63.0	64.0	57.0	64	58	
1	93.0	89.0	96.0	97.0	105.0	96	109	
2	66.0	59.0	58.0	56.0	50.0	51	75	
3	290.0	245.0	339.0	278.0	320.0	84	36	
4	125.0	158.0	132.0	142.0	138.0	127	146	
..	...	...	...	...	...	...	...	
110	210.0	211.0	225.0	217.0	217.0	210	207	
111	34.0	50.0	44.0	77.0	79.0	81	89	
112	5.0	10.0	10.0	1.0	14.0	10	10	
113	36.0	80.0	62.0	50.0	77.0	120	250	
114	260.0	272.0	306.0	319.0	317.0	285	344	

	Bulan -14	Bulan -15	Trend Pattern
0	58	54	[Decreasing, Decreasing, Decreasing, Increasin...
1	106	102	[Decreasing, Increasing, Decreasing, Decreasin...
2	63	73	[Increasing, Decreasing, Increasing, Decreasin...
3	110	553	[Decreasing, Increasing, Decreasing, Decreasin...
4	132	106	[Decreasing, Increasing, Decreasing, Decreasin...
..	...	...	...
110	210	217	[Increasing, Increasing, Decreasing, Decreasin...
111	87	86	[Decreasing, Decreasing, Increasing, Decreasin...
112	12	10	[Decreasing, Increasing, Increasing, Stable, S...
113	300	360	[Decreasing, Stable, Increasing, Decreasing, I...
114	356	327	[Increasing, Decreasing, Decreasing, Decreasin...

[115 rows x 23 columns]

Fig 1. Dataset Study

Data Preprocessing The data preprocessing process encompassed several key steps to ensure the dataset was well-prepared for analysis. First, missing values were addressed through imputation, allowing for the preservation of the dataset's integrity by filling in gaps where data

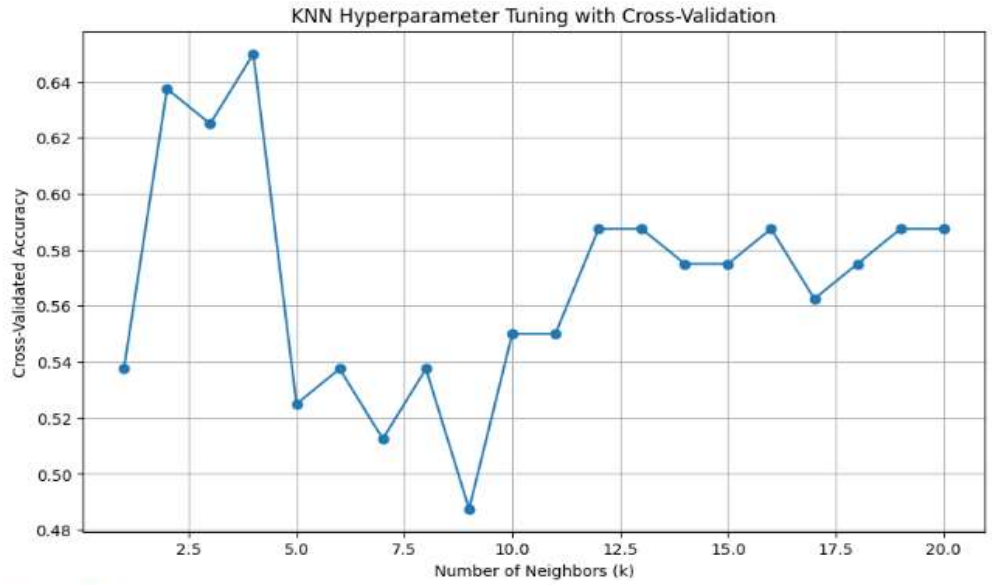
was absent. Next, normalization was conducted using StandardScaler, which standardized the kWh values to create a uniform scale for further analysis. Additionally, categorical encoding was applied to the violation categories, transforming these variables into a format suitable for machine learning algorithms. Finally, the dataset was divided into training and testing sets, with 70% allocated for training and 30% reserved for testing, establishing a clear framework for evaluating model performance.

Model Training The K-Nearest Neighbors (KNN) algorithm was utilized in this study, employing several specific configurations to enhance its performance. The Euclidean distance metric served as the basis for measuring similarity between data points. To optimize hyperparameters, particularly the value of k, cross-validation was conducted, revealing that the best value was determined to be 4 for the original dataset. The training was carried out under two distinct scenarios: first, the model was trained using the original dataset without any modifications. In the second scenario, the dataset was balanced using the Synthetic Minority Over-sampling Technique (SMOTE), which generated synthetic samples to bolster the representation of the minority class. This approach aimed to improve the model's predictive capabilities and ensure a more balanced dataset for training.

RESULTS AND DISCUSSION

Without SMOTE

Data Preprocessing and K-Nearest Neighbors (KNN) Implementation The dataset was standardized using StandardScaler to ensure uniform feature scaling before splitting into training (70%) and testing (30%) sets. The KNN model was trained with default parameters and the Euclidean distance metric. Cross-validation identified the optimal k value as 4.



Best k value: 4

	precision	recall	f1-score	support
0	0.64	0.82	0.72	22
1	0.43	0.23	0.30	13
accuracy			0.60	35
macro avg	0.54	0.52	0.51	35
weighted avg	0.56	0.60	0.56	35

Fig 2. KNN Evaluation Results Without SMOTE

Figure 2 displays the KNN model's performance with k set to 4. The model effectively identified non-violation cases (class 0) with a precision of 0.64, recall of 0.82, and F1-score of 0.72. However, for violation cases (class 1), it struggled, showing a precision of 0.43, recall of 0.23, and a low F1-score of 0.30. Overall accuracy was 0.60.

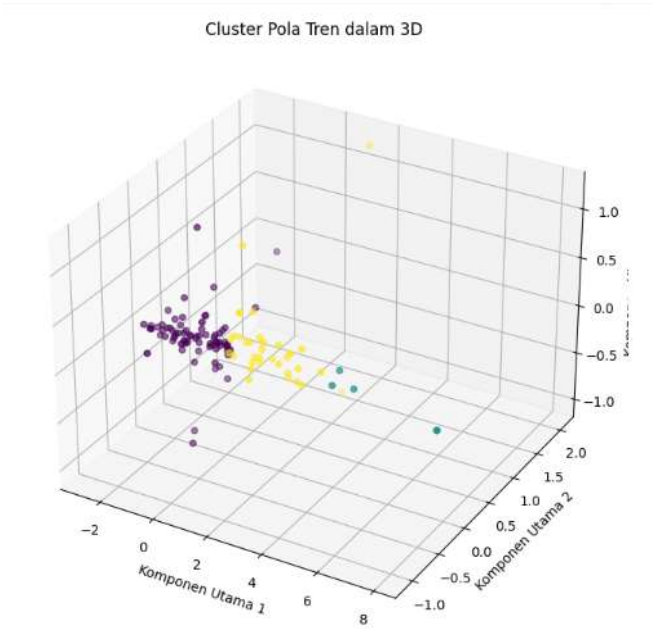


Fig 3. Clustering Results Distribution

Figure 3 illustrates the clustering results, with Cluster 0 ("Increasing Trend") containing the most instances (79), indicating a predominant positive usage trend. Cluster 1 ("Decreasing Trend") had 31 instances, while Cluster 2 ("Seasonal Fluctuation") had only 5, suggesting that most data showed increasing or decreasing trends, with seasonal fluctuations being rare.

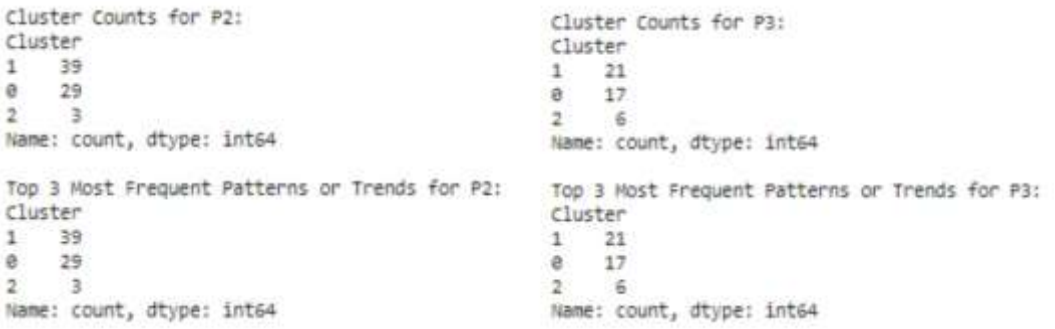


Fig 4. Trends in Datasets P2 and P3

In Figure 4 the largest cluster in both datasets is Cluster 1, representing the increasing trend.

Table 1. Indicates That for P2 & P3

Golongan	Cluster	Jumlah Pelanggan	Pola Tren
P2	1	39	Tren Meningkat
	0	29	Tren Menurun
	2	3	Fluktuasi Musiman
P3	1	21	Tren Meningkat
	0	17	Tren Menurun
	2	6	Fluktuasi Musiman

Table 1 indicates that for P2, Cluster 1 (39 instances) is predominant, while Cluster 0 (29 instances) shows a decreasing trend, and Cluster 2 (3 instances) indicates rare seasonal fluctuations. The same pattern is observed in P3, with Cluster 1 leading at 21 instances.

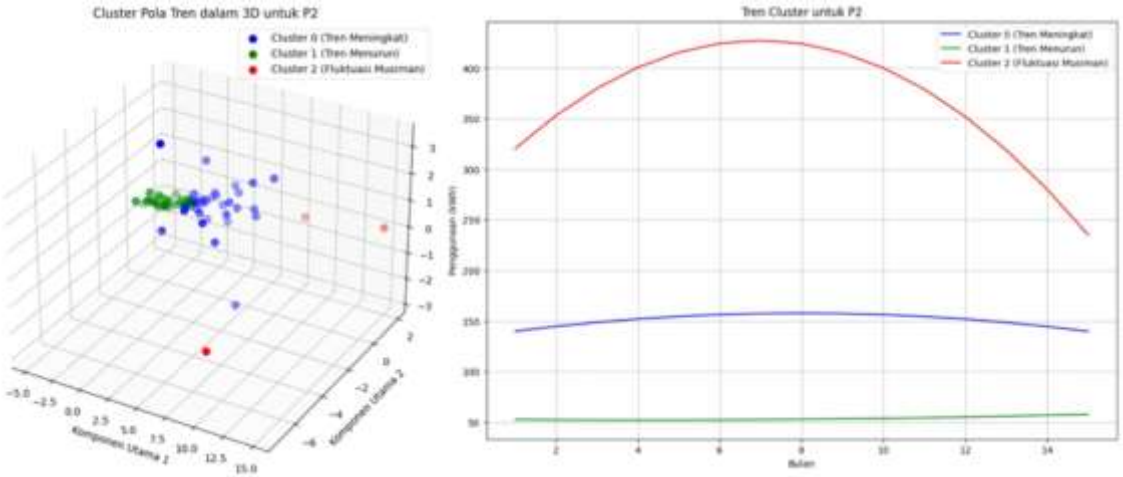


Fig 5. Trend Patterns Without SMOTE

Overall, the increasing trend is dominant, while seasonal fluctuations are infrequent. Figure 5 shows that customers using over 300 kWh monthly experienced a decline, whereas those around 150 kWh remained stable, and those below 100 kWh showed an increase. 39 customers in Cluster 0 of P2 likely engaged in fraud, using between 55 and 58.5 kWh, with an

upward trend. Figure 4.8 indicates that 29 customers in Cluster 1 had usage declining from 325 to 250 kWh. Finally, Figure 4.9 shows that 3 customers in Cluster 2 had usage between 145 and 147 kWh, initially rising before increasing significantly over the first eight months.

Trend Patterns in P2 with SMOTE

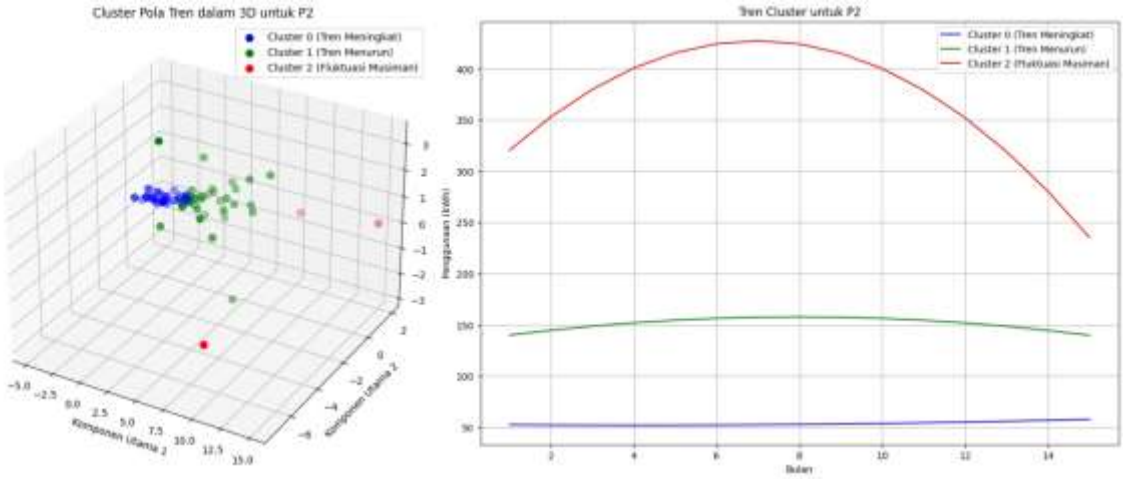


Fig 6. Trends for Cluster P2

Overall, the increasing trend is predominant in both datasets, while seasonal fluctuations are infrequent. **Figure 6** illustrates the trends for Cluster P2, where customers using over 300 kWh monthly experienced a decrease, which is a positive sign despite an initial increase in the first four months. In contrast, customers using around 150 kWh remained stable with no decline, while those using below 100 kWh per month showed an upward trend.

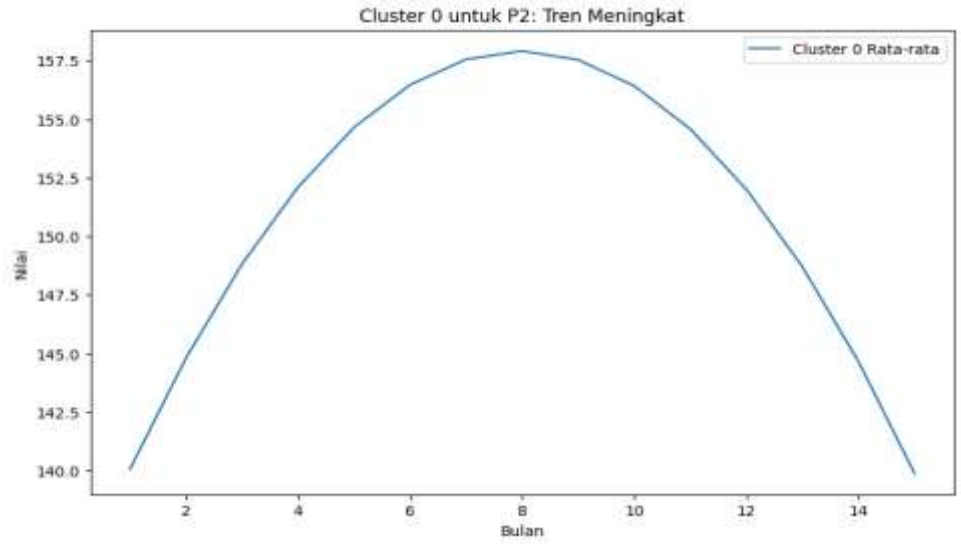


Fig 7. Cluster 0 P2

In **Figure 7**, 36 customers in Cluster 0 of P2 showed signs of potential fraud with an average usage of 140 kWh. This cluster maintained a stable trend, although there was an increase over the initial eight months, followed by a decline in usage.

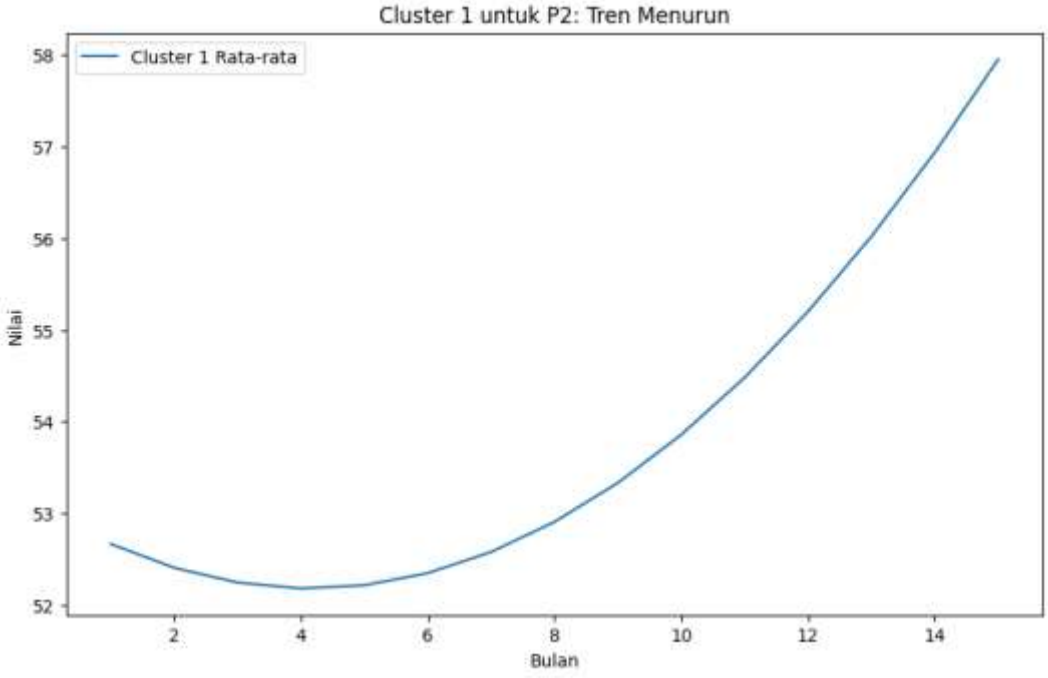


Fig 8. Cluster 1 P2

Figure 8 depicts that 32 customers in Cluster 1 exhibited fraudulent behavior with usage increasing from 53 kWh to 58 kWh, indicating an upward trend. Lastly,

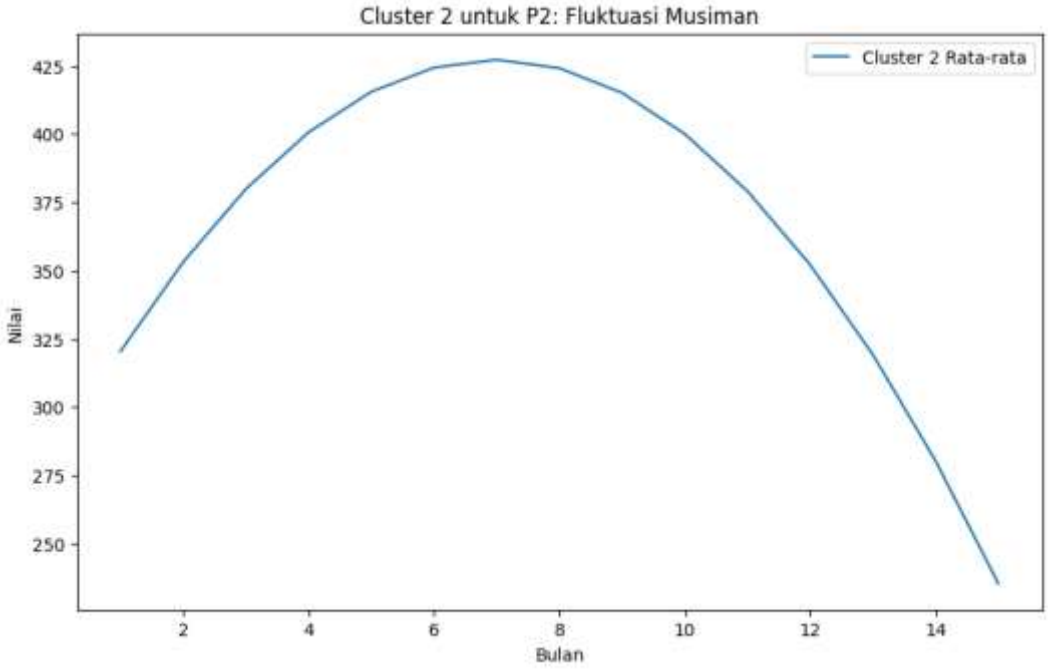


Fig 9. Cluster 2 P2

Figure 9 shows that 3 customers in Cluster 2 had usage ranging from 325 kWh to 240 kWh, initially declining before significantly increasing over the first six months.

Advantages of Using SMOTE: SMOTE (Synthetic Minority Over-sampling Technique) offers significant advantages in data balancing. Firstly, it effectively addresses class imbalance, particularly observed in dataset P2. This technique helps reduce the disparity between majority and minority classes, enabling the KNN model to be trained on data that better represents each class. Consequently, the KNN model becomes more reliable in recognizing patterns from the minority class. While overall accuracy may not show significant improvement, SMOTE enhances model performance, particularly in recall for the minority class (class 1), indicating greater sensitivity in detecting electricity misuse despite a trade-off in precision. Additionally, SMOTE facilitates better pattern identification in clustering analysis, as the more balanced data distribution can reveal previously unnoticed trend patterns.

Disadvantages of Using SMOTE: Despite its advantages, SMOTE also presents some drawbacks. A primary concern is the potential for overfitting in the minority class. SMOTE

generates synthetic samples based on existing data, which may lead the model to become too specific to these synthetic samples and not represent real-world data variations. Furthermore, applying SMOTE does not always improve the overall accuracy of the KNN model. In the analysis conducted, the accuracy of the model with SMOTE was slightly lower than without SMOTE (0.54 with SMOTE vs. 0.60 without SMOTE). This may be due to the artificial imbalance created by synthetic samples, which disrupts the model's ability to accurately classify the majority class. Lastly, using SMOTE adds complexity and processing time, requiring additional computational resources to generate synthetic samples and train the model on a larger dataset, a critical consideration when dealing with very large datasets.

Focus on the Minority Class SMOTE is utilized to enhance model performance on the minority class, albeit at the expense of slight overall accuracy. The lower k value after applying SMOTE indicates that the KNN model requires fewer neighbors to achieve optimal performance on a more balanced dataset. In other words, the model becomes more adaptable in recognizing patterns from the minority class post-SMOTE.

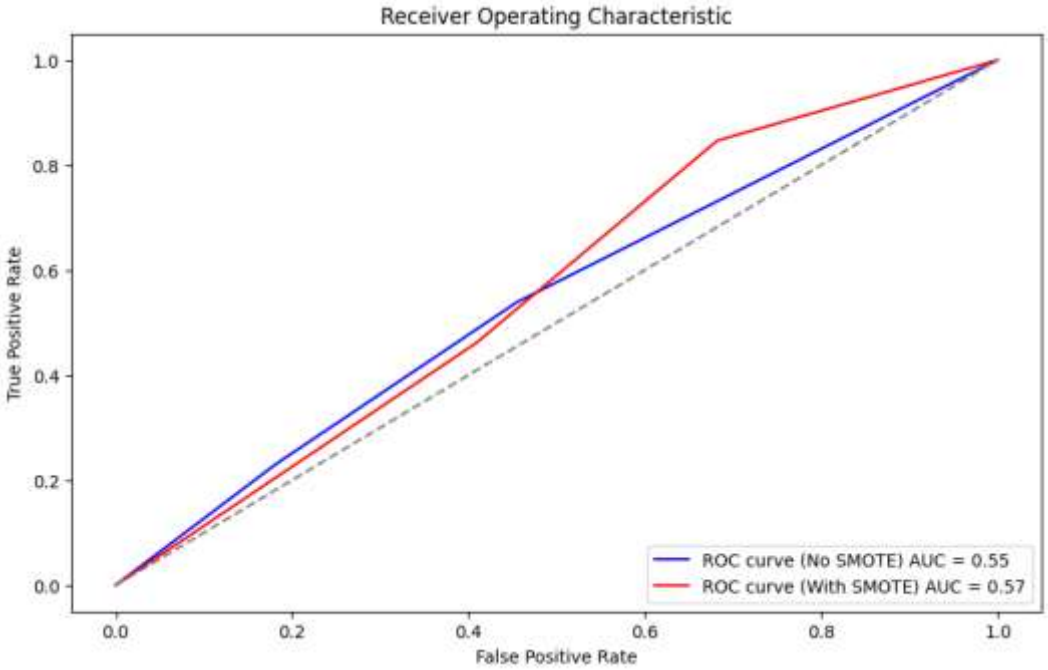


Fig 10. Comparison between methods illustrates that without SMOTE

Trade-off Between Precision and Recall In **Figure 10**, the comparison between methods illustrates that without SMOTE, the model achieved higher accuracy but had low

precision and recall for the minority class. After applying SMOTE, recall for the minority class improved, even though precision decreased. This signifies that the model is more frequently detecting violations, albeit with more false positives. In the context of electricity misuse detection, identifying more violations can be more critical than minimizing false positives.

Relevance in Business Context: Although the overall model accuracy is lower with SMOTE, the primary focus of this research is on detecting patterns of electricity misuse. Therefore, the increased recall for violation cases after applying SMOTE is a significant advantage, as it allows for the identification of more potential offenders, which is more relevant to PLN's operational context.

Discussion

The findings suggest that while SMOTE can enhance recall, it may compromise precision. Overfitting risks arise due to synthetic data, potentially reducing the model's ability to generalize well on unseen data. The trade-offs highlight the need for careful consideration when using data balancing techniques in machine learning applications.

CONCLUSION

This study evaluated the effectiveness of SMOTE (Synthetic Minority Over-sampling Technique) in balancing data within datasets P2 and P3, focusing on the detection of electricity misuse. The findings indicate several key insights:

Effectiveness of SMOTE in Data Balancing: SMOTE proved effective in addressing class imbalance, particularly in enhancing the representation of minority classes. This improvement allowed the KNN model to become more sensitive to patterns in underrepresented classes, even though overall accuracy did not always increase.

Model Performance on Minority Classes: The implementation of SMOTE significantly enhanced the model's performance regarding the minority class, as evidenced by the increased recall rates. This enhancement suggests that SMOTE aids the model in identifying more instances of electricity misuse, which is critical for effective violation detection, despite some compromise in precision.

Customer Trend Patterns: Analysis of customer trends in P2 and P3 revealed significant patterns, such as suspicious increases in electricity consumption among low-consumption

customers and declines among high-consumption customers. These patterns provide crucial insights into potential future misuse of electricity.

Potential Fraud Detection: The increasing trend, particularly among low-consumption customers, highlights the potential for more sophisticated electricity misuse. Continuous monitoring of these patterns is essential, especially following declines followed by significant increases in consumption.

Performance Without SMOTE: Analysis indicated that model performance without SMOTE had advantages in terms of overall accuracy. Without synthetic samples potentially leading to overfitting, the KNN model was able to focus on original data patterns, resulting in more stable and accurate predictions for the majority class. In the context of electricity misuse detection, higher overall accuracy is more reliable, particularly when the primary goal is to identify normal usage patterns and detect suspicious outliers consistently.

Comparison of Model Performance with and without SMOTE

Metric	Without SMOTE	With SMOTE
Accuracy	0.60	0.54
Precision (Class 1)	0.43	0.40
Recall (Class 1)	0.23	0.46
F1-Score (Class 1)	0.30	0.42

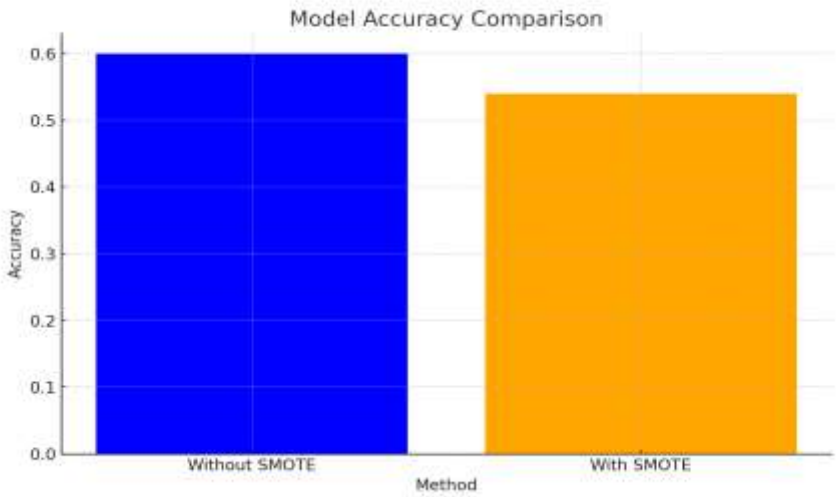


Fig 11. Model Accuracy Comparison

REFERENCES

PLN, "Peraturan Direksi PT. PLN (Persero) Nomor.0028.P/Dir/2023 Tentang Penertiban Pemakaian Tenaga Listrik."

P. P. Pekerjaan Kepada Perusahaan Lainl Anggaran Dasar PLN *et al.*, "Peraturan Direksi : Nomor : 088-Z.P/DIR/2016 Tentang Penertiban Pemakaian Tenaga Listrik (P2TL)."

S. Sumarlin, "Implementasi Algoritma K-Nearest Neighbor Sebagai Pendukung Keputusan Klasifikasi Penerima Beasiswa PPA dan BBM," *Jurnal Sistem Informasi Bisnis*, vol. 5, no. 1, pp. 52–62, Apr. 2015, doi: 10.21456/VOL5ISS1PP52-62.

M. Danny, A. Muhidin, and A. Jamal, "Application of the K-Nearest Neighbor Machine Learning Algorithm to Preduct Sales of Best-Selling Products," *Brilliance: Research of Artificial Intelligence*, vol. 4, no. 1, pp. 255–264, Jun. 2024, doi: 10.47709/brilliance.v4i1.4063.

G. A. Pradipta, R. Wardoyo, A. Musdholifah, I. N. H. Sanjaya, and M. Ismail, "SMOTE for Handling Imbalanced Data Problem : A Review," *2021 6th International Conference on Informatics and Computing, ICIC 2021*, 2021, doi: 10.1109/ICIC54025.2021.9632912.

D. J. Won, I. Y. Chung, J. M. Kim, S. Il Moon, J. C. Seo, and J. W. Choe, "Development of power quality monitoring system with central processing scheme," *Proceedings of the IEEE Power Engineering Society Transmission and Distribution Conference*, vol. 2, no. SUMMER, pp. 915–919, 2002, doi: 10.1109/PESS.2002.1043496.

T. Zhang, "Solving large scale linear prediction problems using stochastic gradient descent algorithms," *Proceedings, Twenty-First International Conference on Machine Learning, ICML 2004*, pp. 919–926, 2004, doi: 10.1145/1015330.1015332.

R. N. Singarimbun, "Adaptive Moment Estimation Untuk Meminimalkan Kuadrat Error pada Algoritma Backpropagation," 2019, Accessed: Aug. 06, 2024. [Online]. Available: https://repositori.usu.ac.id/handle/123456789/20255

M. S. Iqbal, M. F. A. Limon, M. M. Kabir, M. K. M. Rabby, M. J. A. Soeb, and M. F. Jubayer, "A hybrid optimization algorithm for improving load frequency control in interconnected power systems," *Expert Syst Appl*, vol. 249, Sep. 2024, doi: 10.1016/J.ESWA.2024.123702.

M. Soori, B. Arezoo, and R. Dastres, "Optimization of energy consumption in industrial robots, a review," *Cognitive Robotics*, vol. 3, pp. 142–157, Jan. 2023, doi: 10.1016/J.COGR.2023.05.003.

D. Granados-Lieberman, R. J. Romero-Troncoso, R. A. Osornio-Rios, A. Garcia-Perez, and E. Cabal-Yepez, "Techniques and methodologies for power quality analysis and disturbances classification in power systems: A review," *IET Generation, Transmission and Distribution*, vol. 5, no. 4, pp. 519–529, Apr. 2011, doi: 10.1049/IET-GTD.2010.0466.