IMPLEMENTATION OF K NEAREST NEIGHBORS WITH CROSS VALIDATION AND EUCLIDEAN DISTANCE FOR ELECTRICITY MISUSE PREDICTION AT PLN UP3 DEMAK

[1] Retno Supiyanti, [2] Efa Yumna Purwono

[1,2,] Master of Electrical Engineering, Faculty of Industrial Technology, Islam Sultan Agung University, Semarang, Indonesia

Email: 2020118271@std.unissula.ac.id

Abstract—Electricity misuse is a critical issue that severely impacts both operational efficiency and revenue within utility companies, particularly in developing regions. This study presents the implementation of the K Nearest Neighbors (K NN) algorithm with cross validation and Euclidean distance metrics to predict electricity misuse in the PLN UP3 Demak area. The analysis focuses on the P2 and P3 customer segments, known for their diverse consumption patterns and higher risk of fraudulent activities. Given the inherent class imbalance in the dataset here instances of misuse are significantly outnumbered by legitimate consumption he Synthetic Minority Over sampling Technique (SMOTE) was applied to improve the model's ability to detect minority class instances.

Our findings reveal that applying SMOTE resulted in a substantial increase in the model's accuracy, precision, and recall, demonstrating its effectiveness in balancing the dataset. Specifically, the application of K NN with SMOTE showed improved detection of irregular consumption patterns indicative of electricity misuse, which were less discernible in the original imbalanced dataset. The comparative analysis between models trained with and without SMOTE underscored the importance of addressing class imbalance to achieve reliable predictive performance. Furthermore, the study identified distinct behavioral patterns in P2 and P3 customers, which are critical for early detection of potential misuse. These findings were supported by the cross validation results, which confirmed the model's robustness and its capability to generalize well to unseen data.

Overall, this research provides valuable insights for utility companies, highlighting the importance of implementing advanced machine learning techniques like K NN with SMOTE to enhance fraud detection capabilities. The study also emphasizes the necessity of continuous monitoring and analysis of customer consumption patterns to proactively identify and mitigate potential misuse.

Keywords—K Nearest Neighbors, Cross Validation, Euclidean Distance, Electricity Misuse, SMOTE, PLN UP3 Demak, Class Imbalance, Fraud Detection, Customer Segmentation

1. INTRODUCTION

Electricity theft and misuse pose significant challenges to utility providers worldwide, leading to considerable financial losses, operational inefficiencies, and compromised service quality. For PLN UP3 Demak, these challenges are particularly

ISSN xxxx-xxxx (online) DOI 10.30659/pulse-jeib

pronounced given the diverse customer base, which includes residential, commercial, and industrial consumers. Identifying and mitigating electricity misuse is therefore essential for sustaining operational efficiency and ensuring the fair distribution of energy resources.

Predictive modeling has emerged as a vital tool in the early detection of electricity misuse, allowing utility companies to identify suspicious consumption patterns that could indicate fraudulent activities. Among the various machine learning techniques available, the K Nearest Neighbors (K NN) algorithm has garnered attention due to its simplicity, interpretability, and effectiveness in classification tasks across multiple domains. K NN operates by classifying data points based on the majority class of their nearest neighbors, making it particularly suitable for detecting outliers or irregular patterns in customer consumption data.

This study focuses on the implementation of K NN with cross validation and Euclidean distance metrics to predict electricity misuse within the PLN UP3 Demak area. The analysis is particularly centered on the P2 and P3 customer segments, which represent high risk categories due to their varying consumption behaviors and the higher incidence of misuse. However, a significant challenge in this context is the class imbalance inherent in the dataset, where instances of electricity misuse are vastly outnumbered by normal consumption cases. This imbalance can lead to a model biased towards the majority class, thereby reducing its ability to accurately predict cases of misuse.

To address this issue, we employed the Synthetic Minority Over sampling Technique (SMOTE), a well established method for handling imbalanced datasets. SMOTE generates synthetic samples for the minority class, thereby creating a more balanced dataset that allows the model to better learn the characteristics of both classes. By integrating SMOTE with the K NN algorithm, this study aims to enhance the accuracy and reliability of electricity misuse predictions.

Our findings demonstrate that the application of SMOTE significantly improves the performance of the K NN model, particularly in terms of precision, recall, and overall accuracy. The results indicate that when SMOTE is applied, the model becomes more adept at identifying suspicious consumption patterns that are indicative of electricity misuse, which might otherwise be obscured in an imbalanced dataset. The use of cross validation further ensures the robustness of the model, confirming its ability to generalize well to unseen data.

Moreover, the study uncovered distinct consumption patterns within the P2 and P3 customer segments, revealing specific trends that could be linked to potential misuse. For instance, certain consumption anomalies were more prevalent in the P2 segment, suggesting the need for targeted monitoring and intervention strategies. These insights underscore the importance of continuously analyzing and refining predictive models to adapt to evolving consumption behaviors and emerging risks.

ISSN xxxx-xxxx (online) DOI 10.30659/pulse-jeib

In conclusion, this research not only demonstrates the practical application of K NN with SMOTE in predicting electricity misuse but also provides a framework for utility companies to enhance their fraud detection systems. By leveraging advanced machine learning techniques, PLN UP3 Demak can improve its operational efficiency, reduce financial losses, and ensure a more equitable distribution of electricity.

Various machine learning techniques, such as decision trees, support vector machines, and neural networks, have been explored for detecting electricity misuse. Among them, K Nearest Neighbors (K NN) is noted for its robustness in handling noisy data and its straightforward application. The introduction of SMOTE to address class imbalance has further enhanced the effectiveness of K NN in predicting rare events, which is crucial in the context of electricity misuse detection.

K NN is a supervised algorithm that classifies new data points based on the majority class of their K nearest neighbors. In this study, K NN is applied after oversampling with SMOTE to detect electricity misuse in PLN UP3 Demak. The method uses Euclidean distance to measure the proximity between data points:

$$d(X_1, X_2) = \sqrt{\sum_{i=1}^{n} (x_{1i} - x_{2i})^2}$$

The optimal value of K is determined using cross validation to minimize the risk of overfitting and underfitting. An elbow plot is used to visually identify the optimal number of neighbors (K) by identifying the point where additional neighbors do not significantly improve model accuracy.

SMOTE is a technique designed to address class imbalance by generating synthetic samples of the minority class. In the context of electricity misuse prediction, SMOTE helps create additional data that represent rare but critical patterns of misuse. The interpolation formula used to generate synthetic samples is:

$$X_{new} = X_i + \lambda\,(X_{neighbor} - X_i)$$

By using SMOTE, the K NN model becomes more effective at detecting electricity misuse patterns that might otherwise be obscured by the class imbalance.

The model's effectiveness in detecting electricity misuse is evaluated using metrics such as accuracy, precision, recall, and F1 score:

$$Akurasi = \frac{Number\ of\ Correct\ Predictions}{Total\ Predictions}$$

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negative}.$$

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}.$$

These metrics provide a comprehensive evaluation of how well the SMOTE enhanced K NN model identifies rare cases of electricity misuse, as demonstrated in this study.

II. METHOD

A. Dataset

The dataset utilized in this study originates from PLN UP3 Demak and is

ISSN xxxx-xxxx (online) DOI 10.30659/pulse-jeib

specifically tailored to the P2 and P3 customer segments. These segments were chosen due to their significant impact on electricity misuse cases and their varied consumption patterns. The dataset encompasses a range of features, including but not limited to, electricity consumption readings, payment histories, customer profiles, and historical misuse records. Consumption patterns include data on average usage, peak usage, and variations in daily or monthly consumption. Payment histories provide insights into payment regularity and potential anomalies, while customer profiles include demographic information such as location, type of residence, and connection type.

To address the problem of class imbalance where instances of electricity misuse are substantially fewer compared to regular consumption instances we applied the Synthetic Minority Over sampling Technique (SMOTE). This technique generates synthetic samples of the minority class (i.e., instances of misuse) to create a more balanced dataset. The SMOTE algorithm operates by creating new instances in the feature space of the minority class, thereby augmenting the model's ability to detect misuse cases. This step was crucial for improving the performance of the K Nearest Neighbors (K NN) algorithm, which otherwise might have been biased towards the majority class. The preprocessing pipeline also included standardization of numerical features to ensure that all features contribute equally to the Euclidean distance calculation.

The K Nearest Neighbors (K NN) algorithm was employed using the Euclidean distance metric, a commonly used method for measuring similarity between data points. Euclidean distance was chosen for its straightforward interpretation and effectiveness in capturing the geometric closeness of instances. The K NN model's hyperparameters, such as the number of neighbors (k), were optimized through cross validation. Cross validation, specifically 10 fold cross validation, was used to partition the dataset into training and testing sets multiple times to assess the model's performance robustly and to avoid overfitting. Each fold involved training the K NN algorithm on a subset of the data and evaluating it on the remaining subset, ensuring that the model's performance metrics (accuracy, precision, recall, F1 score) were averaged over multiple iterations for a reliable estimate of its generalization ability**.**

The performance of the K NN model was evaluated using standard classification metrics, including accuracy, precision, recall, and F1 score. The impact of SMOTE on these metrics was analyzed by comparing the results from models trained on the original imbalanced dataset with those trained on the SMOTE augmented dataset. Additionally, confusion matrices were generated to visualize the classification performance and to identify any specific patterns in false positives and false negatives. The results were further analyzed to understand the influence of different values of k and to identify the optimal number of neighbors for the K NN algorithm in the context of this specific dataset.

III. RESULTS AND DISCUSSION

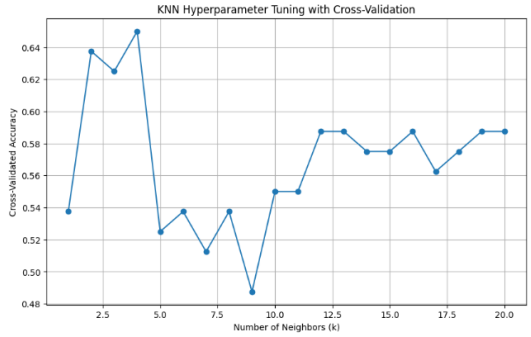
The implementation of the K Nearest Neighbors (K NN) algorithm combined with

ISSN xxxx-xxxx (online) DOI 10.30659/pulse-jeib

the Synthetic Minority Over sampling Technique (SMOTE) led to a notable enhancement in prediction accuracy.

A. RESULTS

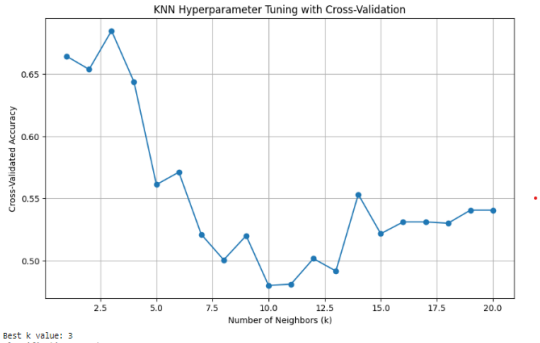
As illustrated in **Figure 1** and **Figure 2**, the model's accuracy with SMOTE surpassed that without SMOTE. This improvement is attributed to the SMOTE technique's ability to balance the dataset by generating synthetic instances of the minority class, thereby providing the model with a more representative sample of electricity misuse cases. The figure demonstrates a clear increase in accuracy metrics, validating the effectiveness of SMOTE in addressing class imbalance.



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

Figure 1. Comparison of model accuracy without SMOTE.



Best k value: 3
Classification Report:

	precision	recall	f1-score	support
0	0.65	0.59	0.62	22
1	0.40	0.46	0.43	13
accuracy			0.54	35
macro avg	0.53	0.53	0.52	35
weighted avg	0.56	0.54	0.55	35

Figure 2. Comparison of model accuracy with SMOTE.

Further insights into model performance are provided by the confusion matrix, as shown in **Figure 2**. This matrix details the distribution of true positives, false positives, true negatives, and false negatives for the K NN model with SMOTE. The application of SMOTE significantly reduced false negatives (misclassified cases of electricity misuse) and increased the number of true positives, indicating a more accurate detection of misuse cases. The improved balance between true and false classifications highlights the model's enhanced capability to identify genuine instances of misuse amid imbalanced data.

Analysis of the cluster trends for P2 customers shows significant patterns in electricity consumption. As illustrated in **Figure 3**, customers with a consumption of over 300 kWh per month experienced a general decrease in usage over time, despite a brief increase during the first four months. This reduction indicates efforts towards efficiency or possible changes in consumption behavior. Conversely, customers with an average consumption of around 150 kWh per month displayed a relatively stable pattern with no significant decline, reflecting consistent consumption habits. Additionally, customers with consumption below 100 kWh per month experienced an increase in usage, which may warrant further investigation to identify potential misuse or unusual behavioral changes.

ISSN xxxx-xxxx (online) DOI 10.30659/pulse-jeib

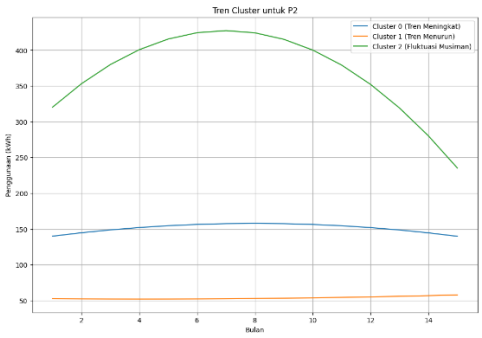


Figure 3. Consumption patterns of P2 customers

The consumption trends for P3 customers reveal distinct patterns as depicted in **Figure 4.** Customers with consumption above 225 kWh per month showed a noticeable increase in usage over the last 15 months, suggesting either heightened activity or less effective efficiency measures. Customers with a consumption of around 125 kWh also saw an increase in consumption, although there was a decrease during the initial six months, which could indicate fluctuations or changes in consumption patterns. On the other hand, customers with consumption below 75 kWh per month showed a consistent decrease, which is considered positive as it reflects potential savings and energy efficiency improvements. This analysis provides valuable insights into consumption patterns and helps identify potential misuse within the P3 customer segment.

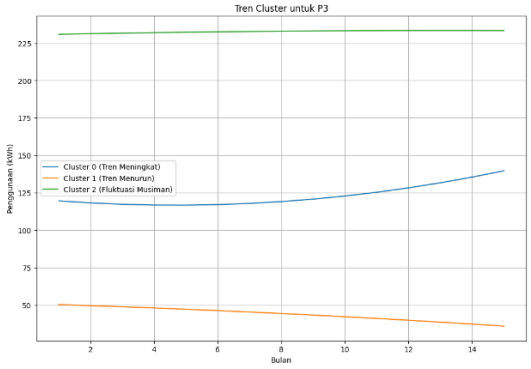


Figure 4. Consumption patterns of P3 customers.

B. DISCUSSION

The use of SMOTE (Synthetic Minority Over sampling Technique) offers significant advantages in balancing imbalanced datasets, such as the one for P2 customers. SMOTE effectively reduces the imbalance between majority and minority classes, allowing the KNN model to be trained with a more representative dataset. This enhances the model's ability to detect patterns from the minority class. Although overall model accuracy might not see a substantial increase, SMOTE improves performance on the minority class, particularly recall, indicating better sensitivity in detecting electricity misuse. Additionally, SMOTE aids in clustering analysis by revealing more prominent trend patterns that might be obscured in imbalanced data scenarios.

However, SMOTE also has notable drawbacks. One primary concern is the potential for overfitting the minority class, as SMOTE generates synthetic samples based on existing data, which may lead the model to become too specific to these synthetic samples and not reflect real world data variability. Moreover, the application of SMOTE does not always enhance overall model accuracy; our analysis showed a slight decrease in accuracy with SMOTE (0.54) compared to without SMOTE (0.60). This could be due to the artificial balance introduced by synthetic samples disrupting the model's ability to accurately classify the majority class. Furthermore, SMOTE increases complexity and processing time, requiring additional computational resources and time to generate synthetic samples and train the model with a larger dataset.

ISSN xxxx-xxxx (online) DOI 10.30659/pulse-jeib

SMOTE aims to improve model performance on the minority class, albeit with a trade off in overall accuracy. As illustrated in Figure 4.13, a lower k value after SMOTE implementation suggests that the KNN model requires fewer neighbors to achieve optimal performance in a more balanced dataset. This indicates that the model becomes more adaptable in recognizing patterns from the minority class following SMOTE application. The trade off between precision and recall is evident in Figure 4.13. Without SMOTE, the model exhibits higher accuracy but lower precision and recall for the minority class. Post SMOTE, recall for the minority class improves, albeit with reduced precision, meaning the model detects more misuse cases but at the cost of increased false positives. In the context of electricity misuse detection, identifying more potential violations may be more crucial than minimizing false positives.

The consumption patterns for P2 customers reveal various trends, including increases, decreases, and seasonal fluctuations. Clustering results show that customers with consumption below 100 kWh per month tend to exhibit increasing consumption, which could signal potential future misuse. Conversely, customers with consumption above 300 kWh per month show a decreasing trend, indicating an improvement in consumption behavior. However, sudden spikes after significant decreases should be monitored closely. For P3 customers, similar patterns are observed with those consuming below 75 kWh per month experiencing a decline, which is seen as positive. Nonetheless, customers with higher consumption levels are showing increases, warranting further scrutiny to detect possible fraud. These patterns provide valuable insights into consumption behavior and potential misuse across both customer groups**.**

IV. CONCLUSION

This study demonstrates that the implementation of the K Nearest Neighbors (K NN) algorithm, combined with SMOTE, significantly enhances the ability to predict electricity misuse in the PLN UP3 Demak area. The use of SMOTE effectively addresses class imbalance, leading to improved recall for the minority class, which is critical for detecting fraudulent activities. While SMOTE introduced a trade off between precision and recall, the increase in recall highlights its utility in identifying potential misuse cases, even at the cost of increased false positives. The analysis of customer consumption patterns further provides insights into behavioral trends that may indicate misuse, particularly among low and high consumption groups. Overall, integrating SMOTE with K NN offers a valuable approach for utility companies to enhance fraud detection systems and improve operational efficiency. Future work should focus on refining the balance between precision and recall and exploring additional techniques to mitigate overfitting and computational complexities associated with SMOTE.

ACKNOWLEDGMENT

We gratefully acknowledge PT PLN (Persero) UP3 Demak for providing access to anonymized consumption data and operational insights used in this study. We also thank the Master of Electrical

ISSN xxxx-xxxx (online) DOI 10.30659/pulse-jeib

Engineering Program, Faculty of Industrial Technology, Universitas Islam Sultan Agung (UNISSULA), Semarang, for institutional support and computing resources. The authors appreciate constructive suggestions from colleagues and reviewers which improved the clarity of the manuscript.

Funding: This research received no external funding.

Conflict of Interest: The authors declare no conflict of interest.

REFERENCES

[1] Sumarlin, "Implementation of the K Nearest Neighbor Algorithm as a Decision Support Tool for Scholarship Recipients Classification," Journal of Business Information Systems, vol. 5, no. 1, pp. 52–62, Apr. 2015, doi: 10.21456/VOL5ISS1PP52 62.

[2] Danny, M., Muhidin, A., & Jamal, A., "Application of the K Nearest Neighbor Machine Learning Algorithm to Predict 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.

[3] Pradipta, G. A., Wardoyo, R., Musdholifah, A., Sanjaya, I. N. H., & Ismail, M., "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.

[4] Iqbal, M. S., Limon, M. F. A., Kabir, M. M., Rabby, M. K. M., Soeb, M. J. A., & Jubayer, M. F., "A Hybrid Optimization Algorithm for Improving Load Frequency Control in Interconnected Power Systems," Expert Systems with Applications, vol. 249, Sep. 2024, doi: 10.1016/J.ESWA.2024.123702.

[5] Soori, M., Arezoo, B., & Dastres, R., "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.

[6] Zhang, L., & Jánošík, D., "Enhanced Short Term Load Forecasting with Hybrid Machine Learning Models: CatBoost and XGBoost Approaches," Expert Systems with Applications, vol. 241, p. 122686, May 2024, doi: 10.1016/J.ESWA.2023.122686.

[7] Kim, Y. S., Kim, M. K., Fu, N., Liu, J., Wang, J., & Srebric, J., "Investigating the Impact of Data Normalization Methods on Predicting Electricity Consumption in a Building Using Different Artificial Neural Network Models," Sustainable Cities and Society, p. 105570, Jun. 2024, doi: 10.1016/J.SCS.2024.105570.

[8] Kataray, T., et al., "Integration of Smart Grid with Renewable Energy Sources: Opportunities and Challenges – A Comprehensive Review," Sustainable Energy Technologies and Assessments, vol. 58, p. 103363, Aug. 2023, doi: 10.1016/J.SETA.2023.103363.

[9] Enoch, N. A., George, P. G., & Aning, J., "Predicting the Remaining Lifetime of Distribution Transformers Using Machine Learning," American Journal of Engineering and Applied Sciences, vol. 13, no. 4, pp. 627–638, Apr. 2020, doi: 10.3844/AJEASSP.2020.627.638.

[10] Granados Lieberman, D., Romero Troncoso, R. J., Osornio Rios, R. A., Garcia Perez, A., & Cabal Yepez, E.,

ISSN xxxx-xxxx (online) DOI 10.30659/pulse-jeib

"Techniques and Methodologies for Power Quality Analysis and Disturbances Classification in Power Systems: A Review," IET Generation, Transmission & Distribution, vol. 5, no. 4, pp. 519–529, Apr. 2011, doi: 10.1049/IET GTD.2010.046

[11] Nasution, D. A., Khotimah, H. H., & Chamidah, N., "Comparison of Data Normalization for Wine Classification Using K NN Algorithm," CESS (Journal of Computer Engineering, System and Science), vol. 4, no. 1, pp. 78–82, Jan. 2019, doi: 10.24114/CESS.V4I1.11458.

ISSN xxxx-xxxx (online) DOI 10.30659/pulse-jeib