Bigint Computing Journal
Volume 4, Issue 1, 2026, pp. 23-29
ISSN: 3032-5374
DOI: https://doi.org/10.55537/bigint.v4i1.1498
https://journal.aira.or.id/bigint/index

Implementation of the K-Nearest Neighbors Algorithm for Classifying Environmental Security Levels Based on Crime Data

Muhammad Aidil Affan, Alya Winanda

Department of Informatics Engineering, Universitas Medan Area, Medan, Indonesia

A B S T R A C T

This study aims to evaluate the effectiveness and performance of the K-Nearest Neighbors (KNN) algorithm in classifying regional security levels based on crime data. The research adopts a quantitative approach using secondary data, with KNN applied as the classification method and the Confusion Matrix used as the evaluation metric. The dataset consists of crime data from September and October as training data and November data as testing data. The features used include the number of reported crimes, theft cases, and violent crime cases. The results indicate that the KNN algorithm achieves an accuracy of 96.15%. The precision values reach 1.00 for the safe and vulnerable classes, while recall values are 1.00 for the safe and alert classes and 0.80 for the vulnerable class. These findings demonstrate that the KNN algorithm is effective in classifying regional security levels and can support data-driven decision-making based on official crime statistics.

KEYWORD: classification, crime data, environmental security, K-Nearest Neighbor.

* Corresponding author:
Muhammad Aidil Affan and Alya Winanda
Department of Informatics Engineering, Universitas Medan Area, Medan, Indonesia
Email: m.aidilaffan99@gmail.com and alyawnnd@gmail.com
Article History: Received: 2026-01-06; Accepted: 2026-01-25

1. INTRODUCTION

Crime is an act that violates the law and human rights, committed by individuals or groups, and inevitably causes harm to other individuals, communities, and even the state [1]. Criminal activities are often influenced by economic conditions and can result in various forms of loss, including financial, physical, moral, and psychological damage [2]. A high crime rate contributes significantly to public insecurity and undermines citizens' sense of safety within their own country [3]. In addition to this issue, the lack of decision support systems for mapping and managing crime data often hampers the timely dissemination of security-related information to the public [4].

Many criminal offenders evade punishment due to delays in delivering information to law enforcement agencies [5]. Crime data and information are frequently processed manually, resulting in time-consuming procedures and inefficiencies. The absence of systems capable of rapidly and accurately depicting regional security conditions poses challenges for communities, law enforcement authorities, and government institutions. Therefore, there is a need for a system that can accelerate data entry, enable fast and accurate data analysis, and facilitate the timely dissemination of reliable security information, particularly in regions with limited access to crime-related information [6].

To address this issue, it is necessary to apply a machine learning algorithm capable of classifying crime-prone areas based on official crime datasets. One such algorithm is the K-Nearest Neighbors (KNN) algorithm, which is applied in this study using crime data obtained from PUSIKNAS BARESKRIM POLRI. Machine learning is a field that focuses on developing algorithms and statistical models that enable computer systems to solve problems without explicit instructions, instead learning patterns from data [7].

The KNN algorithm is a well-known and widely used method that belongs to the instance-based learning group and is categorized as a lazy learning technique. This method classifies data based on the proximity or similarity between instances [8]. The working principle of KNN involves identifying a group of nearest neighbors from the training data and assigning the class label with the highest similarity to the target data [9]. Applying statistical approaches and machine learning algorithms to classify crime data based on specific characteristics is particularly useful for identifying significant crime patterns [10].

Data mining is a technology that enables the extraction of meaningful information and useful patterns from complex datasets and plays an important role in analyzing and predicting crime to support governmental decision-making processes [11]. The fundamental concept of data mining is to uncover hidden information within databases and is part of the Knowledge Discovery in Databases (KDD) process, which aims to identify valuable patterns and insights from large datasets [12].

Previous studies have applied various algorithms to crime data analysis, such as clustering crime levels using the K-Means algorithm and predicting crime-prone locations using artificial neural networks [13]. However, studies that specifically classify environmental security levels into categories such as safe, alert, and vulnerable using official data from PUSIKNAS BARESKRIM POLRI remain relatively limited. Based on this gap, this research focuses on implementing the KNN algorithm to classify environmental security levels using regional crime data from the city of Medan, where crime rates continue to increase on a monthly basis. The proposed approach is expected to produce a classification system that is accurate, objective, and easily understood by the public, law enforcement agencies, and government authorities, while also providing a clear representation of regional security conditions.

Although several previous studies have applied data mining and machine learning techniques in crime analysis, most of them focus on regional classification or the prediction of specific criminal incidents. Research that explicitly classifies environmental security levels into the categories of safe, alert, and vulnerable using official crime data from PUSIKNAS BARESKRIM POLRI remains scarce. Therefore, the objective of this study is to address this research gap by applying the K-Nearest Neighbors (KNN) algorithm as a classification method for environmental security levels in the city of Medan. The results are expected to provide structured and easily interpretable information that can support decision-making processes for law enforcement agencies and local governments.

2. METHODOLOGY

This study adopts a quantitative research approach by applying a supervised learning method to classify environmental security levels based on crime data at the sub-district level in the city of Medan. Supervised learning is a machine learning technique in which the learning process is guided by labeled reference values to direct the classification outcomes. The data used in this study are secondary data obtained from the National Criminal Information Center (PUSIKNAS) of Bareskrim Polri. The dataset is provided in Excel file format and contains crime-related attributes for each sub-district.

The overall research procedure is illustrated in Figure 1, which presents the stages of the research process.

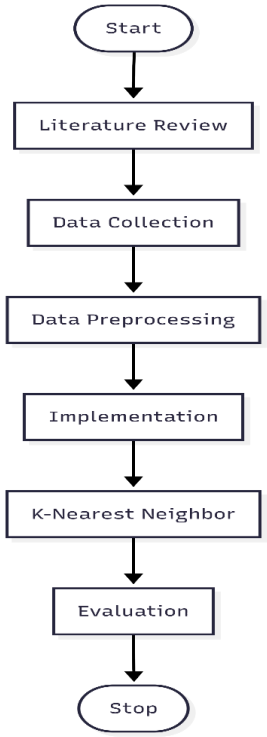

Figure 1. Research stages

Before applying the K-Nearest Neighbors (KNN) algorithm to the dataset, feature scaling is performed as a preprocessing step. Without proper feature scaling, the accuracy of the classification results may decrease significantly [13].

The classification algorithm used in this study is the K-Nearest Neighbors (KNN) algorithm with a value of K=5K = 5K=5, based on the implementation in the developed code. The choice of K=5K = 5K=5 is intended to provide a balance between sensitivity to local data patterns and classification stability, while also reducing the risk of overfitting that may occur when using very small KKK values. The distance between data instances is calculated using the Euclidean Distance metric, which is defined as follows:

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(yi - xi)^2}$$

This study utilizes three datasets: two training datasets consisting of crime data from September and October, and one testing dataset consisting of crime data from November. Prior to applying the KNN algorithm, data preprocessing is conducted to handle missing values in both the training and testing datasets. Missing values in the dataset are replaced with zero to ensure completeness during the classification process.

The threshold-based labeling process for environmental security levels is defined according to the following rules:

1. Vulnerable: total crime value ≥ 50% of the maximum value in the training dataset
2. Alert: total crime value ≥ 25% and < 50%
3. Safe: total crime value < 25%

Model evaluation is performed using leave-one-out cross-validation on the training data to calculate performance metrics, including accuracy, precision, recall, and F1-score. Subsequently, the trained model is applied to the testing dataset (November data) to predict environmental security levels for each sub-district.

The system implementation is developed using the Python programming language with the support of the Pandas, Dataclasses, and Matplotlib libraries. The evaluation metrics calculated in this study include:

1. Accuracy: the percentage of correct predictions
2. Precision, Recall, and F1-score: calculated for each class (Safe, Alert, and Vulnerable)

The main processing stages consist of:

1. Reading and preprocessing crime data from Excel files
2. Automatic labeling of the training data
3. Classification using the KNN algorithm (K=5, Euclidean Distance)
4. Evaluation of model performance
5. Prediction of environmental security levels using the November testing dataset

3. RESULTS AND DISCUSSION

This study produces a classification of environmental security levels based on crime data using the K-Nearest Neighbors (KNN) technique. The training data consist of crime records from September and October, while November data are used as the testing dataset. The variables employed in the classification process include the total number of crimes, the number of fraud cases, and the number of violent crime cases at the sub-district level.

At the initial stage, the training data are automatically labeled according to security levels based on the relative crime totals compared to the maximum value in the training dataset. Areas with high crime levels are labeled as Vulnerable, while those with moderate and low crime levels are labeled as Alert and Safe, respectively. This labeling process successfully generates a representative distribution of security classes. The application of the KNN algorithm to the training data shows that most instances are correctly classified according to the assigned labels.

These results indicate that crime patterns in September and October are relatively consistent across sub-districts. When applied to the November testing data, the KNN algorithm classifies environmental security levels for each sub-district based on their proximity to crime patterns observed in previous months. Sub-districts with crime and violence profiles similar to vulnerable areas in the training data tend to be predicted as Vulnerable, while areas with lower crime intensity show a tendency to be classified as Safe.

Overall, these findings suggest that the KNN algorithm is capable of capturing crime patterns effectively. Model performance is evaluated using accuracy, precision, recall, and F1-score metrics on the training dataset. The analysis shows that the model achieves an accuracy of 96.15%, with perfect precision (1.00) for the Safe and Vulnerable classes and perfect recall (1.00) for the Safe and Alert classes. However, despite the high overall

performance, some limitations are observed, particularly a recall value of 0.80 for the Vulnerable class and a precision value of 0.93 for the Alert class. The relatively balanced precision and recall across classes indicate that the model is stable and effective in distinguishing between different security categories.

In general, the results demonstrate that the KNN algorithm is both effective and accurate in classifying environmental security levels based on crime data. By leveraging data similarity, the system can adapt to monthly changes in crime patterns without requiring complex model retraining. Therefore, this approach can serve as a valuable tool for analyzing environmental security conditions.

Table 1. Summary of Crime Data Statistics

Description	Training Data (Sep–Oct)	Testing Data (Nov)
Number of Records	26	13
Number of Sub-districts	13	13
Total Crime Range	4–174	0–113
Fraud Crime Range	4–153	0–95
Violence Crime Range	0–33	0–18
Number of Classes	3 (Safe, Alert, Vulnerable)	3 (Predicted)
K Value	5	

Based on Table 1, the total crime range in the training data spans from 4 to 174 cases, while the testing data range from 0 to 113 cases.

Table 2. Actual vs. Predicted Labels for Training Data (September)

No.	Sub-district	Total Crimes	Fraud Cases	Violence Cases	Actual Label	Predicted Label
1	Percut Sei Tuan	173	149	24	Vulnerable	Vulnerable
2	Sunggal	135	114	21	Vulnerable	Vulnerable
3	Medan Baru	77	72	5	Alert	Alert
4	Medan Area	51	44	7	Alert	Alert
5	Patumbak	49	48	1	Alert	Alert
6	Medan Kota	52	44	8	Alert	Alert
7	Medan Timur	47	43	4	Alert	Alert
8	Deli Tua	52	43	9	Alert	Alert
9	Helvetia	42	32	10	Safe	Safe
10	Pancur Batu	31	21	10	Safe	Safe
11	Medan Tuntungan	34	24	10	Safe	Safe
12	Medan Barat	24	23	1	Safe	Safe
13	Kutalimbaru	13	8	5	Safe	Safe

Table 2 shows that no misclassification occurs in the September dataset, indicating that the KNN model accurately classifies all sub-districts for this period.

Table 3. Actual vs. Predicted Labels for Training Data (Oktober)

No.	Sub-district	Total Crimes	Fraud Cases	Violence Cases	Actual Label	Predicted Label
1	Percut Sei Tuan	174	153	21	Vulnerable	Vulnerable
2	Sunggal	161	128	33	Vulnerable	Vulnerable
3	Medan Baru	100	84	16	Vulnerable	Alert
4	Medan Area	67	59	8	Alert	Alert
5	Patumbak	58	48	10	Alert	Alert
6	Medan Kota	53	45	8	Alert	Alert
7	Medan Timur	60	50	10	Alert	Alert
8	Deli Tua	50	47	3	Alert	Alert
9	Helvetia	55	44	11	Alert	Alert

No.	Sub-district	Total Crimes	Fraud Cases	Violence Cases	Actual Label	Predicted Label
10	Pancur Batu	40	28	12	Safe	Safe
11	Medan Tuntungan	54	42	12	Alert	Alert
12	Medan Barat	32	31	1	Safe	Safe
13	Kutalimbaru	4	4	0	Safe	Safe

Table 3 indicates that one misclassification occurs in the Medan Baru sub-district, where a vulnerable area is predicted as alert. This suggests a slight overlap in crime characteristics between these two classes.

Table 4. Confusion Matrix Evaluation Results

Class	TP	FP	FN	Precision	Recall	F1-Score
Vulnerable	4	0	1	1.00	0.80	0.89
Alert	13	1	0	0.93	1.00	0.96
Safe	8	0	0	1.00	1.00	1.00
Accuracy						**96.15%**

Table 4 confirms that the model achieves an overall accuracy of 96.15%, with the highest precision observed in the Safe and Vulnerable classes.

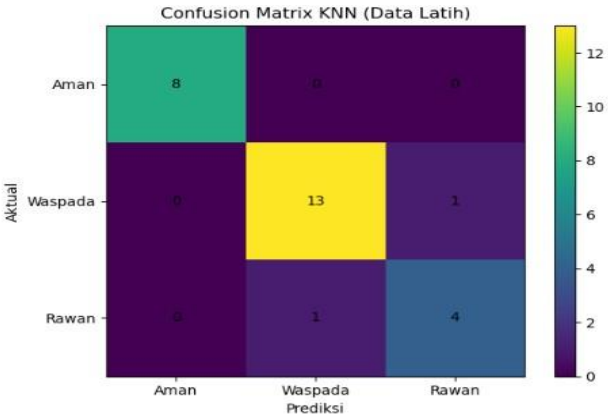

Figure 2. Confusion Matrix

Based on the confusion matrix shown in Figure 2, the KNN algorithm correctly classifies most of the data points. Misclassification is relatively minimal and occurs primarily between the Alert and Vulnerable classes. This indicates that the model has a strong ability to distinguish environmental security levels, although some overlap remains between classes with similar crime characteristics.

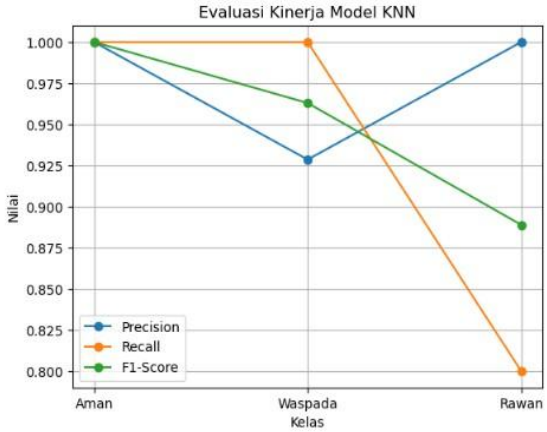

Figure 3. KNN Model Performance Evaluation

The performance evaluation shown in Figure 3 demonstrates that the KNN algorithm achieves very high accuracy (96.15%). The high precision and recall values for the Safe and Vulnerable classes indicate consistent classification performance. However, the slightly lower recall for the Vulnerable class suggests that some high-risk areas may be classified into adjacent categories. These findings are consistent with previous studies reporting that KNN is effective for crime data classification based on data similarity [10], [11].

Table 5. Environmental Security Level Classification Results (November)

No.	Sub-district	Total Crimes	Fraud Cases	Violence Cases	Actual Label	Predicted Label
1	Percut Sei Tuan	113.0	95.0	18.0	–	Vulnerable
2	Sunggal	85.0	68.0	17.0	–	Alert
3	Medan Baru	74.0	59.0	15.0	–	Alert
4	Medan Area	38.0	31.0	7.0	–	Safe
5	Patumbak	36.0	30.0	6.0	–	Safe
6	Medan Kota	32.0	25.0	7.0	–	Safe
7	Medan Timur	54.0	49.0	5.0	–	Alert
8	Deli Tua	39.0	34.0	5.0	–	Safe
9	Helvetia	30.0	25.0	5.0	–	Safe
10	Pancur Batu	21.0	14.0	7.0	–	Safe
11	Medan Tuntungan	36.0	31.0	5.0	–	Safe
12	Medan Barat	22.0	17.0	5.0	–	Safe
13	Kutalimbaru	0.0	0.0	0.0	–	Safe

Table 5 shows that the Percut Sei Tuan sub-district is classified as Vulnerable, while most other sub-districts fall into the Safe category. This result indicates that crime intensity is not evenly distributed across regions.

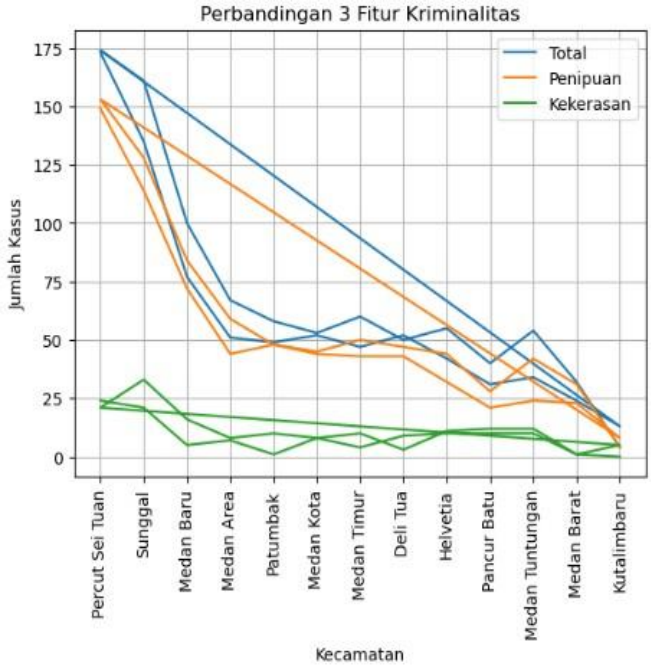

Figure 4. Comparison of Three Crime Features

Based on Figure 4, sub-districts classified as Vulnerable exhibit higher values of total crime, fraud, and violence compared to other regions. In contrast, areas classified as Safe demonstrate relatively low crime intensity across all three features. This difference indicates that these features play a significant role in determining environmental security levels and are key factors in the KNN classification process.

Overall, the classification results show that most regions fall into the Safe or Alert categories, while certain areas, such as Percut Sei Tuan, are classified as Vulnerable. This finding highlights that crime levels vary

significantly between regions and are influenced by specific crime characteristics. Areas classified as vulnerable generally experience higher crime intensity than other regions.

These classification results can serve as decision-support information for law enforcement agencies, where vulnerable areas may be prioritized for security interventions. However, the classification outcomes should be used as a supporting tool rather than the sole basis for policy decisions, considering the potential for misclassification.

4. CONCLUSION

This study successfully demonstrates the capability and performance of the K-Nearest Neighbors (KNN) algorithm in classifying regional environmental security levels based on crime data in the city of Medan. The model achieves an accuracy of 96.15% by leveraging similarity patterns from the training data to predict security levels for November as the testing period. The classification results reflect actual environmental conditions across regions, and the evaluation metrics support the theoretical validity of the KNN algorithm for this task.

Therefore, the KNN algorithm can provide a meaningful representation of environmental security conditions and serve as a decision-support tool for government authorities and law enforcement agencies. Nevertheless, this study is limited by the relatively short time span of the crime data used. As a result, the classification outcomes may not fully capture long-term crime dynamics. Future research should incorporate longer temporal datasets and additional crime indicators to improve robustness and generalizability.

REFERENCES

[1] F. Rahmadayanti and R. Rahayu, "Penerapan metode data mining pada kasus kriminalitas Indonesia," *Jurnal Teknologi Informasi Mura*, vol. 15, no. 1, pp. 52–61, 2023. https://doi.org/10.32767/jti.v15i1.2054

[2] J. A. E. Jurnal *et al.*, "Pengaruh urbanisasi, tingkat kemiskinan, dan ketimpangan pendapatan terhadap kriminalitas di Provinsi Jawa Timur," *Jurnal Aplikasi Ekonomi*, vol. 6, no. 3, 2021. https://doi.org/10.29407/jae.v6i3.16307

[3] B. Solikhin and A. Rifal, "Sistem informasi pengolahan data laporan kasus kriminal pada Subdit Renakta Ditreskrimum Polda Jawa Timur," *DIKE: Jurnal Ilmu Multidisiplin*, vol. 2, no. 1, pp. 17–23, 2024. https://doi.org/10.69688/dike.v2i1.64

[4] R. Puspita and A. Widodo, "Perbandingan metode KNN, decision tree, dan naïve Bayes terhadap analisis sentimen pengguna layanan BPJS," *Informatika*, vol. 5, no. 4, pp. 646–654, 2021. https://doi.org/10.32493/informatika.v5i4.7622

[5] I. Muslim, K. Karo, A. Khosuri, J. Steiven, I. Septory, and D. Pebrian, "Pengukuran jarak pada algoritma k-NN untuk klasifikasi kebakaran hutan dan lahan," *MIB: Media Informatika Budidarma*, vol. 6, pp. 1174–1182, Apr. 2022. https://doi.org/10.30865/mib.v6i2.3967

[6] H. U. Chikodili, P. O. Ogbobe, and M. C. Okoronkwo, "Analysis of crime pattern using data mining techniques," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 12, 2021. https://doi.org/10.14569/ijacsa.2021.0121259

[7] A. P. Setyan, "Prediksi kerawanan lokasi terhadap kasus pencurian kendaraan menggunakan algoritma jaringan syaraf tiruan," Institut Teknologi Sepuluh Nopember, 2021. https://doi.org/10.30591/jpit.v6i3.2627

[8] B. Kommey, D. Opoku, A. Asare-Appiah, G. O. Wiredu, and P. K. Baah, "An ad-hoc crime reporting information management system," *International Journal of Informatics, Information System and Computer Engineering (INJIISCOM)*, vol. 4, no. 2, pp. 136–159, 2023. https://doi.org/10.34010/injiiscom.v4i2.11436

[9] R. S. Nurhalizah and R. Ardianto, "Analisis supervised dan unsupervised learning pada machine learning: Systematic literature review," *JIKI*, vol. 4, no. 1, pp. 61–72, 2024. https://doi.org/10.54082/jiki.168

[10] R. P. R. Pambudi, "Prediction of criminal theft locations at the Binjai Police Station using historical data and the KNN algorithm," *Journal of Artificial Intelligence and Engineering Applications (JAIEA)*, vol. 5, no. 1, pp. 1499–1504, 2025. https://doi.org/10.59934/jaiea.v5i1.1658

[11] A. P. Silalahi, H. G. Simanullang, and M. I. Hutapea, "Supervised learning metode k-nearest neighbor untuk prediksi diabetes pada wanita," *METHOMIKA: Jurnal Manajemen Informatika & Komputerisasi Akuntansi*, vol. 7, no. 1, pp. 144–149, 2023. https://doi.org/10.46880/jmika.vol7no1.pp144-149

[12] P. M. S. Tarigan, J. T. Hardinata, H. Qurniawan, M. Safii, and R. Winanjaya, "Implementasi data mining menggunakan algoritma apriori dalam menentukan persediaan barang," *Jurnal Janitra Informatika dan Sistem Informasi*, vol. 2, no. 1, pp. 9–19, 2022. https://doi.org/10.25008/janitra.v2i1.142

[13] K. Lehmann *et al.*, "Automated classification of crime narratives using machine learning and language models in official statistics," *Stats*, vol. 8, no. 3, p. 68, 2025. https://doi.org/10.3390/stats8030068