UAD Universitas Ahmad Dahlan

Bulletin of Applied Mathematics and Mathematics Education
Volume 5, Number 2, October 2025 pp. 113-120
https://dx.doi.org/10.12928/bamme.v5i2.14613
e-ISSN 2776-1029, p-ISSN 2776-1002

Classification of weather events in Lahat regency using the K-Nearest Neighbor method

Endang Sri Kresnawati, Yulia Resti*, Ning Eliyati, Des Alwine Zayanti, Novi Rustiana Dewi, Irsyadi Yani

Universitas Sriwijaya, Jl. Raya Palembang-Prabumulih Km.32, Inderalaya, Ogan Ilir, 30662, Indonesia

*Corresponding e-mail: yulia_resti@mipa.unsri.ac.id

ARTICLE INFO

Article History
Received 15 October 2025
Revised 23 December 2025
Accepted 23 December 2025

Keywords
Euclidean
KNN
Manhattan
Minkowski
Weather events

How to cite this article:
Kresnawati, E. S., Resti, Y., Eliyati, N., Zayanti, D. A., Dewi, N. R., & Yani, I. (2025). Classification of weather events in Lahat regency using the K-Nearest Neighbor method. *Bulletin of Applied Mathematics and Mathematics Education*, *5*(2), 113-120.

ABSTRACT

Weather event classification in a region is very important for various purposes, such as in the fields of transportation, health, agriculture, and others. Lahat has varying land elevations ranging from 26-106 meters above sea level in the East Merapi sub-district to 341-3032 meters above sea level in the Tanjung Sakti Pumi sub-district. It greatly affects local temperature, rainfall, and atmospheric pressure, which in turn affects the distribution of weather patterns and disasters such as floods. KNN is a prediction method that uses the concept of distance for a number of k nearest observations in determining the similarity between observations. Several metrics can be used for this prediction purpose. This study aims to predict weather events in Lahat Regency using the KNN method with several different distance metrics and then compare them to obtain the performance of the KNN prediction method. The results show that the Euclidean distance metric used in the KNN method has a better performance measurement, followed by the Manhattan and Minkowski metrics. In the Euclidean metric, the accuracy, precision, recall, f1-score, AUC, and MC value are 92.69%, 88.21%, 85.81%, 86.99%, 88.99%, and 76.37%, respectively.

Introduction

The K-Nearest Neighbor (KNN) method classifies new observation points based on their closest distance to a set of k observation points in the modeling data (Ali et al., 2019). The closest distance can be determined using several metrics such as Manhattan (Nayak et al., 2022), Euclidean (Qu et al., 2023), or Minkowski (Nair et al., 2025), (Halder et al., 2024). The distance metric in KNN effectively quantifies the concept of 'similarity' between observations through 'closeness' and directly influences the selection of 'neighbors' that inform class type. Therefore, the distance metric chosen in KNN can significantly impact its performance (Cetin & Buyuklu, 2025).

The KNN method often outperforms other methods, especially when the observation data is numeric (Cubillos et al., 2022). Weather event data generally consists of numeric observation variables, including data in Lahat Regency, South Sumatra Province, Indonesia. Predicting weather events in Lahat is vital since Lahat has a very varied altitude above sea level. It has varying land elevations ranging from 26-106 meters above sea level in the East Merapi sub-district to 341-3032 meters above sea level in the Tanjung Sakti Pumi sub-district (BPS-Statistics Lahat Regency, 2025). This fact greatly affects local temperature, rainfall, and atmospheric pressure, which in turn affects the distribution of weather patterns and disasters such as floods.

This study aims to compare the performance of the KNN method with Manhattan, Euclidean, and Minkowski in predicting weather events in Lahat. We create some k-value for each of metric distances and make a classification of the weather events in Lahat based on the k-value obtained. The k-value can be obtained by examining the smallest error from a set of k experiments (Chandra et al., 2023).

Method

This work using the secondary data since January 1, 2019, until December 31, 2023, from https://www.visualcrossing.com/weather/weather-data-services. The data is data on weather events in Lahat District. The sixteen factors considered are presented in Table 1.

Table 1. Predictor and Target Variables

Variable	Name	Information	Name	Information
Predictor	Max. Temperature (X_1)	25 - 43 °C	Visibility (X_9)	1 - 10 Nm
	Min. Temperature (X_2)	19.3 - 27 °C	UV Index (X_{10})	1.2 – 27.2 Wp
	Ave. Temperature (X_3)	24 – 30.2 °C	Solar energy (X_{11})	61 – 96.9 %
	Max. Feels like (X_4)	25 - 50.9 °C	Humidity (X_{12})	5.4 – 177.8 Km/h
	Min. Feels like (X_5)	19.3 – 30.6 °C	Wind Speed (X_{13})	24.1 – 35.8 °C
	Ave. Feels like (X_6)	17.6 – 25.9 °C	Cloud Layer (X_{14})	0 - 100 Octa
	Dew (X_7)	0 – 359.6 $°C$	Solar Radiation (X_{15})	14.7 – 315.1 Nm
	Wind Direction (X_8)	1.8 – 18.7 Nm	Moon Phase (X_{16})	0 – 0.98 %
Response	Weather Event (Y)	Partially cloudy (13.31%) Overcast (18.73%) Rain (67.96%)		

The research steps are follows:

1) Research data is divided into training data and test data. Training data for modeling is 80% (January 1, 2019 – December 31, 2022) and test data is 20% (January 1, 2023 - December 31, 2023).
2) Determine the shortest distance between training data point A and test data point B using a specific distance metric. The following are the Manhattan (1) (Arora et al., 2021), Euclidean (2) (Debbek et al., 2024), and Minkowski (3) (Lu et al, 2015) distance metrics, respectively:

$$\mathrm{d(A,B)} = \sum_{\mathrm{m=1}}^{\mathrm{p}} |\mathrm{x_{ma}} - \mathrm{x_{mb}}| \tag{1}$$

$$d(A,B) = \sqrt{\sum_{m=1}^{p} (x_{ma} - x_{mb})^2} \tag{2}$$

$$d(A,B) = \left(\sum_{m=1}^{p} |x_{ma} - x_{mb}|^q \right)^{1/q} \tag{3}$$

3) Create a number of k-values for each of the Manhattan, Euclidean, and Minkowski distances and choose the k-value for each distance based on the lowest error. The classification error can be obtained using (4):

$$\text{Error} = 1 - \text{Accuracy} \tag{4}$$

$$\text{Accuracy} = \frac{\sum_{\mathrm{j=1}}^{\mathrm{J}} \frac{\mathrm{TP_j} + \mathrm{TN_j}}{\mathrm{TP_j} + \mathrm{FP_j} + \mathrm{FN_j} + \mathrm{TN}}}{\mathrm{J}} \tag{5}$$

Where class j, j=1,2,3. For j=1 is Partially Cloudy, j=2 is Overcast, and j=3 is Rain. The values in (5) can be obtained using the Confusion Matrix. For the first class, the Confusion Matrix is as in Table 2, the other classes can be obtained in a similar way (Resti et al., 2022).

Table 2. Confusion Matrix

	Actual			
	Class (j)	1	2	3
Classification	1	True Positive (TP_j)	False Negative (FN_j)	False Negative (FN_j)
	2	False Positive (FP_j)	True Negative (TN_j)	False Negative (FN_j)
	3	False Positive (FP_j)	False Positive (FP_j)	True Negative (TN_j)

4) Make predictions using test data based on each distance metric formed in Step (2) with k-values in Step (3).
5) Form a confusion matrix for each prediction based on the selected k-value result as in Table 2.
6) Calculate and compare the metric values for each confusion matrix: accuracy (5), precision (6), recall (7), F1-score (8) (Yani et al., 2025, Resti et al., 2025), AUC (Kresnawati et al., 2024), and Matthew Correlation (MC) (Chicco & Jurman, 2023) for multiclass.

$$\mathrm{Prec} = \frac{\sum_{\mathrm{j=1}}^{\mathrm{J}} \frac{\mathrm{TP_j}}{\mathrm{TP_j} + \mathrm{FP_j}}}{\mathrm{J}} \tag{6}$$

$$Rec = \frac{\sum_{j=1}^{J} \frac{TP_j}{TP_j + FN_j}}{J} \tag{7}$$

$$\mathrm{F_1Score} = \frac{\mathrm{2Precision\ (Recall\ )}}{\mathrm{(Precision\ + Recall)}} \tag{8}$$

$$AUC = \frac{1}{2}\left(\frac{\sum_{j=1}^{J} \frac{TP_j}{TP_j + FN_j}}{J}\right) + \frac{1}{2}\left(\frac{\sum_{j=1}^{J} \frac{TN_j}{TN_j + FN_j}}{J}\right) \tag{9}$$

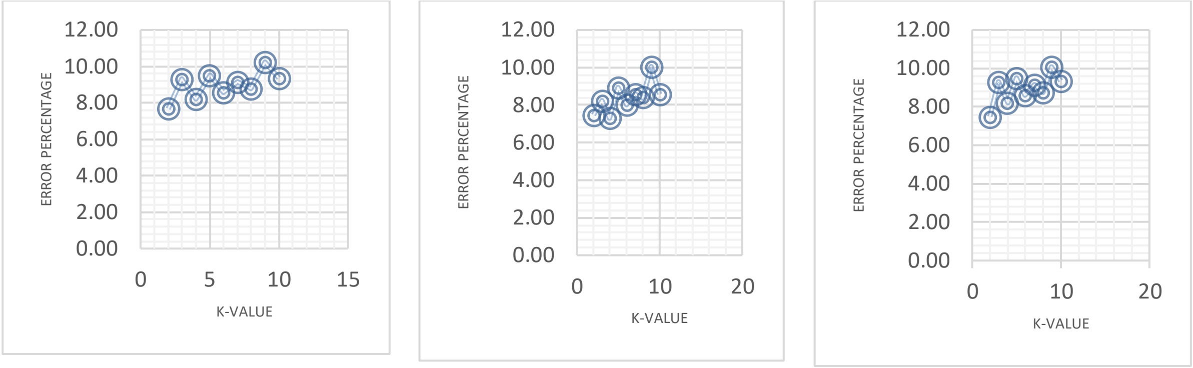
$$\mathrm{MC} = \sqrt{\mathrm{Prec.Rec.Spec.NPV}} - \sqrt{(1-\mathrm{Prec})(1-\mathrm{Rec})(1-\mathrm{Spec})(1-\mathrm{NPV})} \tag{10}$$

$$\mathrm{Spec} = \frac{\sum_{\mathrm{j=1}}^{\mathrm{J}} \frac{\mathrm{TN_j}}{\mathrm{TN_j} + \mathrm{FP_j}}}{\mathrm{J}} \tag{11}$$

$$NPV = \frac{\sum_{j=1}^{J} \frac{TN_j}{TP_j + FP_j}}{J} \tag{12}$$

Results and Discussion

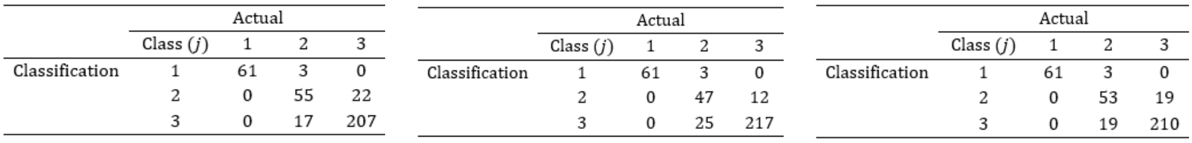
Figure 1 shows that different values of k give different error percentages. The lowest error at each distance metric is then selected for the classification task using KNN.



(a) Manhattan (b) Euclidean (c) Minkowski

Figure 1. Distance metric error rate on weather classification data for Lahat-District

In Manhattan, Euclidean, and Minkowski distance metrics, the k values with the lowest error vary. As shown in Figure 1, the three lowest errors, respectively, for k, are 2, 4, and 2.
The confusion matrix resulting from the classification using KNN for each distance metric based on the selected k-value is given in Figure 2.



		Actual		
	Class (j)	1	2	3
Classification	1	61	3	0
	2	0	55	22
	3	0	17	207

		Actual		
	Class (j)	1	2	3
Classification	1	61	3	0
	2	0	47	12
	3	0	25	217

		Actual		
	Class (j)	1	2	3
Classification	1	61	3	0
	2	0	53	19
	3	0	19	210

(a) Manhattan (b) Euclidean (c) Minkowski

Figure 2. Confusion matrix for the first class

Tabel 3. Performance Metric of KNN Based on Distance-Metric

Distance-Metric	Performance Metric (%)					
	Accuracy	Precision	Recall	F1-score	AUC	MC
Manhattan	92.33	86.38	87.91	87.14	90.44	66.27
Euclidean	92.69	88.21	85.81	86.99	88.99	76.37
Minkowski	92.51	86.88	87.46	87.17	90.14	60.05

Table 3 shows the performance of three different distance metrics using KNN in classifying weather events are satisfactory, more than 85% except MC only. Six performance metrics were measured to obtain information regarding the distance metric that provided the best performance: accuracy, precision, recall, F1-score, AUC, and MC. Euclidean had the three highest scores, followed by Manhattan and Minkowski. The three distance-metrics were accuracy, precision, and MC. Manhattan had the highest recall and AUC, while Minkowski had the highest F1-score.

This fact indicates that Euclidean distance is more relevant in measuring the similarity between predictor variables in Lahat Regency weather data compared to the other two distance metrics. Numerical predictor variables in this study are more suitable for using Euclidean distance, which does not require a transformation process in calculating the distance.

Conclusion

Classifying weather events in Lahat is crucial because Lahat's elevation above sea level varies

10.12928/bamme.v5i2.14613

significantly. This fact has a significant impact on local temperature, rainfall, and atmospheric pressure, all of which impact how weather patterns and natural disasters such as floods are distributed. This paper has classified weather events in Lahat Regency using KNN based on three different distance metrics, namely Manhattan, Euclidean, and Minkowski. The results show that the KNN method performs satisfactorily using the three distance functions, but the best performance is achieved by the Euclidean distance metric with performance metrics of accuracy, precision, recall, F1-score, AUC, and MC of 92.69%, 88.21%, 85.81%, 86.99%, and 76.37%, respectively. Weather classification can also use many other methods such as naive Bayes, logistic regression, random forest, or decision tree, considering that the KNN that has been applied in this study is indeed able to provide satisfactory performance, but this method does not provide an in-depth explanation of the relationship between variables. This research also only focused on Lahat Regency so the model may not be applicable to areas with different weather characteristics. Furthermore, the influence of local topography (mountains, valleys, and elevation variations) was not accounted for in this research dataset.

Acknowledgement

This work was supported by DIPA of Universitas Sriwijaya 2024 Public Service Agency, SP DIPA-023.17.2.677515/2024, on November 24, 2023 and the Rector's Decree 0012/UN9/SK..LP2M. PT/2024, May 20, was issued.

References

Ali, N., Neagu, D., & Trundle, P. (2019). Evaluation of k-nearest neighbor classifier performance for heterogeneous data sets. SN Applied Sciences 1:1559, https://doi.org/10.1007/s42452-019-1356-9

Nayak, S., Bhat, M., Reddy, N.V.S., & Rao, B.A. (2022). Study of distance metrics on k - nearest neighbor algorithm for star categorization. Journal of Physics: Conference Series. 2161s 012004 IOP Publishing doi:10.1088/1742-6596/2161/1/012004

Qu, H., Xu, J., Li, Z., Wei, D., & Wang, F. (2023). Effects of embedded distance measurements interacting with modeling approaches on empirical dynamical model predictions. Ecological Indicators 146 (2023) 109895.

Nair, V.G. (2025). Metric-Driven Voronoi Diagrams: A Comprehensive Mathematical Framework. Computation, 13, 212. https://doi.org/10.3390/computation13090212.

Halder, R. K., Uddin, M. N., Uddin, M. A., Aryal, S., & Khraisat, A. (2024). Enhancing K-nearest neighbor algorithm: a comprehensive review and performance analysis of modifcation Journal of Big Data 11:113.

Cetin, A. I. & Buyuklu, A.H. (2025). A new approach to K-nearest neighbors distance metrics on sovereign country credit rating country credit rating. Kuwait Journal of Science 52 (2025) 100324.

Cubillos, M., Wohlk, S., & Wulff, J. N. (2022). A bi-objective k-nearest-neighbors-based imputation method for multilevel data. Expert Systems with Applications 204 (2022) 117298.

BPS-Statistics Lahat Regency. (2025). Lahat Regency in Figures 7 (45), Catalogue: 1102001.1604, ISSN 0215-3971.

Chandra, W., Suprihatin, B., & Resti, Y. (2023). Median-KNN Regressor-SMOTE-Tomek Links for Handling Missing and Imbalanced Data in Air Quality Prediction. Symmetry, 15, 887. https://doi.org/10.3390/sym150408

Resti, Y., Irsan, C., Amini, M., Yani, I., Passarella, R., & Zayanti, D.A. (2022). Performance Improvement of Decision Tree Model using Fuzzy Membership Function for Classification of Corn Plant Diseases and Pests. Science and Technology Indonesia, e-ISSN:2580-4391 p-ISSN:2580-4405 Vol. 7, No. 3, July.

Yani, I., Marwani, Puspitasari, D., & Resti, Y. (2025). The Symmetric Pattern Fuzzy Discretization in Predicting Plastic Type for a Sorting System Using Decision Tree Methods. Science and Technology Indonesia, e-ISSN:2580-4391 p-ISSN:2580-4405, Vol. 10, No. 3, July.

Resti, Y., Yani, I., Puspitasari, D., Thamrin, I., & Saputra, M.A.A. (2025). Ensemble Method of Multiple Decision Trees with Crisp and Fuzzy Discretization for Axial Surface Roughness Prediction. Science and Technology Indonesia e-ISSN:2580-4391 p-ISSN:2580-4405 Vol. 10, No. 3, July.

Kresnawati, S.K., Suprihatin, B., & Resti, Y. (2024). The Combinations of Fuzzy Membership Functions on Discretization in the Decision Tree-ID3 to Predict Degenerative Disease Status. Symmetry 2024, 16, 1560. https://doi.org/10.3390/sym16121560.

Chicco, D. & Jurman, G. (2023). A statistical comparison between Matthew's correlation coefficient (MCC), prevalence threshold, and Fowlkes–Mallow's index. Journal of Biomedical Informatics 144 104426.

Arora, I., Khanduja, N., & Bansal, M. (2021). Effect of Distance Metric and Feature Scaling on KNN Algorithm while Classifying X-rays. The 10th Seminary of Computer Science Research at Feminine, March 08h, 2021, Constantine 2-Abdelhamid Mehri University, Algeria.

Debbek, F.Y., Ehtiba, F.O., & Abdelmula, H.S.B. (2024). Movie Recommendation Engine Based on Cosine Similarity and KNN. Special Issue for IJEIT on Engineering And Information Technology. , Vol.12 ,No. 1, December.

Lu, B., Charlton, M., Brunsdon, C., & Harris, P. (2015). A The Minkowski approach for choosing the distance metric in geographically weighted regression. International journal of geographical information science. Available from: http://dx.doi.org/10.1080/13658816.2015.1087001

This page is intentionally left blank