Water Level Classification for Detect Flood Disaster Status using KNN and SVM

Jiwa Akbar[1], Muchtar Ali Setyo Yudono[2]
Department of Electrical Engineering, Faculty of Computer Engineering and Design[1], [2]
Nusa Putra University
Sukabumi, Indonesia
jiwa.akbar_te20@nusaputra.ac.id[1], muchtar.alisetyo@nusaputra.ac.id[2]

Abstract*— Flooding occurs when the water's surface elevation exceeds the average level, overflowing river water and creating inundation in low-lying areas. Early warning for potential floods significantly reduces losses, such as human casualties and property damage. In this context, the flood disaster classification system uses water surface elevation data from the Water Resources Agency to predict the likelihood of floods using the K-Nearest Neighbors (KNN) Algorithm. This research aims to classify flood status based on water surface elevation using the K-Nearest Neighbors and Support Vector Machine(SVM) methods in the Ciliwung River. The study results indicate that the SVM algorithm outperforms the KNN algorithm. The SVM algorithm used parameter C ranging from 1 to 10 in the scenarios, and the RBF kernel achieved 100% accuracy. On the other hand, the KNN algorithm achieved 100% accuracy only for K values of 1, 2, 3, 4, and 5 in scenarios where K ranged from 1 to 10..

Keywords—Flood, K-Nearest Neighbors, Support Vector Machine, Ciliwung River, Water Surface Elevation

I. INTRODUCTION

Water is an essential necessity for all individuals and plays a crucial role in human survival [1]. The river serves as a channel to collect and drain water from upstream to downstream sources [2]. A risky situation is when river water is abundant, but without a water filtration and storage system, this can lead to natural disasters, one of which is flooding [3].

Flooding occurs when the Water Level crosses the standard threshold and causes an overflow of river water, which results in inundation in the surrounding area. BNPB noted that flooding disasters were the most frequent from January to April 2021. Floods were recorded 501 times with 267 fatalities [4]. History records the most severe flood disasters occurred in 2002, 2007 and 2013. In 2002, floods submerged 42 sub-districts with 168 sub-districts or approximately 63.4% of the total number of sub-districts in DKI Jakarta. 16,041 hectares or equivalent to 24.25% of the area submerged with a height of 5 meters in DKI Jakarta. A total of 381,266 people and 21 people died as a result of the flood disaster that year .Floods can arise due to various influencing factors, including the low ability of the soil to filter water, the accumulation of garbage along the river flow that blocks the flow of water to the sea and causes water levels in the lower reaches of the river, and the high intensity and duration of prolonged rain [5].

In addition, large population growth also plays a role in the increasing frequency of urban flooding. Floods can cause damage to infrastructure, damage river ecosystems, and threaten the sustainability of development [6].

A study raises the same topic as "Comparison of Machine Learning Algorithms in Water Quality Identification." The results of the study were the accuracy obtained by the Decision Tree method of 60.19%, the Logistic Regression method of 62.80%, the SVM method of 68.59%, and the ANN method of 69.54% [7].

Another study with the same theme approach with the research title "Application of Machine Learning Methods for Flood Disaster Detection." The research used the Systematic Literature Review (SLR) method to solve the problem. From the results of the Systematization of Literature (SLR) analysis conducted, the researchers concluded that implementing this flood detection system has the potential to reduce the risk of loss in the future. The use of machine learning methods can support the classification or detection of whether a flood situation will occur or not[8].

Although the studies mentioned earlier have explored algorithms and machine learning methods for water quality identification and flood disaster detection, there is a lack of thorough evaluation of the K-Nearest Neighbors (KNN) and Support Vector Machine (SVM) methods, especially for water level classification in flood disaster status detection. In addition, existing studies have not provided a comparative analysis of the effectiveness of these methods in real-time flood prediction scenarios. Therefore, further research is needed to assess the accuracy and practical applicability of the KNN and SVM algorithms in this specific context.

II. THEORETICAL FOUNDATION

Some previous studies have used artificial intelligence to solve the problem. One related study compared the performance of several artificial intelligence algorithms in rainfall prediction, entitled Comparison of *Data Mining* Methods for Rainfall Prediction Using C4.5, *Naïve Bayes*, and KNN Algorithms. The results of the study showed that of the three algorithms used, the C4.5 algorithm gave the best results in predicting rainfall, with an accuracy rate of 88.03% and an error rate of 11.97% [9].

Another study is titled "Comparison of Data Mining Approaches in Predicting Floods Using Naïve Bayes and KNN

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v13i3.2166, Copyright ©2024
Submitted : Mey 19, 2024, Revised : July 26, 2024, Accepted : August 12, 2024, Published : November 16, 2024

Algorithms." The final results of both algorithms show that the KNN algorithm outperforms in predicting floods with an accuracy of 88.94% and an error rate of about 11.06% [10].

The next study that has the same theme is entitled "Comparison of Support Vector Machine and Random Forest Classification Algorithms on the Status Prediction of the Disaster Mitigation and Preparedness Index (IMKB) of BPS Work Units in Indonesia in 2020". The results indicate that in terms of accuracy, precision, and recall values, the Random Forest classification method shows superior performance than SVM. Specifically, Random Forest recorded 78.22% accuracy, 75.54% precision, and 76% recall, which is significantly higher than SVM [11].

Artificial *Intelligence* (AI) is one branch of computer science that aims to develop computer systems that can perform tasks that usually require human intelligence. AI seeks to make computers able to learn, plan, solve problems, and make decisions in a similar way to humans [12].

Machine learning is a part of artificial intelligence that focuses on developing algorithms and computer models that are able to gain knowledge from data and experience, without the need for direct programming. The main focus of machine learning is to enable computers to identify patterns, forecast, and make decisions based on the information provided [13].

Supervised learning is a machine learning paradigm in which the model is given sample data consisting of previously known pairs of inputs and outputs. For example, a model is given an image along with a label that describes the objects present in the image. The model learns to find patterns in that data and can then make new predictions or classifications when given never-before-seen inputs [14]. Some of the algorithms used in supervised learning include: Linear Regression, K-Nearest Neighbor, Naïve Bayes, Support Vector Machine (SVM), Random Forest, and Neural Network.

In unsupervised learning, machine learning models are supplied with input data without having associated labels. The goal is to find hidden patterns or structures in the data. Models can group data that have similarities based on similar characteristics or reduce the complexity of data by reducing its dimensions [15]. One of the algorithms used in this model is K-Mean Clustering.

A. *K-Nearest Neighbors*

K-Nearest Neighbor (KNN) is a surveillance method that requires training data to classify objects based on closest proximity. The main principle of KNN is to find the shortest distance between the data being evaluated and k nearest neighbors in the exercise data. The K value reflects the number of nearest neighbors used in the training data. K-NN is a classification algorithm guided by the majority of members among k nearest neighbors to determine a new class of objects [16].

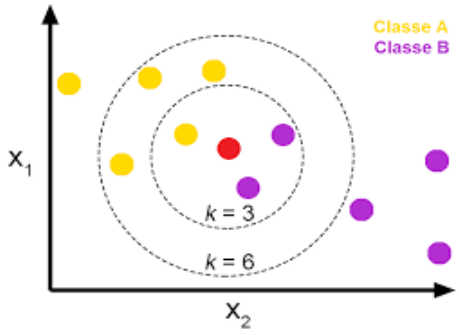

Figure 1. KNN algorithm

B. *Support Vector Machines*

SVM, first introduced by Boser, Guyon, & Vapnik in 1992 when presented at the *Annual Workshop on Computational Learning Theory*, is a *machine learning* method that can be used for classification or prediction. The concept of classification with SVM involves finding the best hyperplane that separates two classes of data. A hyperplane is considered good if it has the largest margin, which is twice the distance between the hyperplane and the support vector. The support vector is the closest point to the hyperplane. SVM is effective in handling high-dimensional data and limited training samples. This method works based on the principle of *Structural Risk Minimization* (SRM), which aims to maximize margins while minimizing expected risks [17].

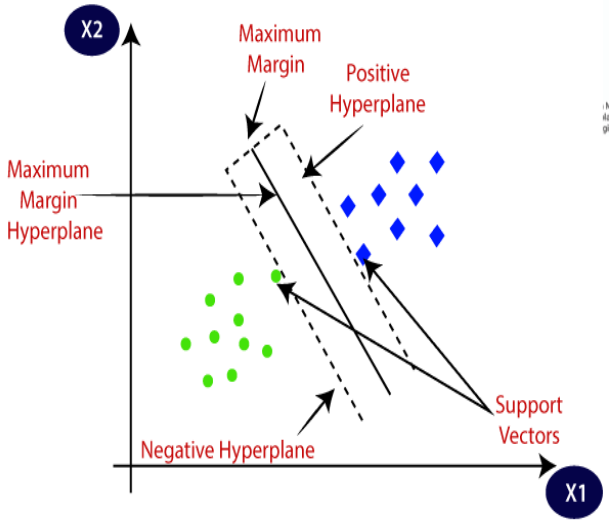

Figure 2. Support Vector Machines.

III. RESEARCH METHOD

The method used in this study is an artificial intelligence method, which uses *the K-nearest neighbors* algorithm and *Support Vector Machines*. Figure 1 is the flow of research conducted from beginning to end. An explanation of each processor will be described in the next discussion.

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v13i3.2166, Copyright ©2024
Submitted : Mey 19, 2024, Revised : July 26, 2024, Accepted : August 12, 2024, Published : November 16, 2024

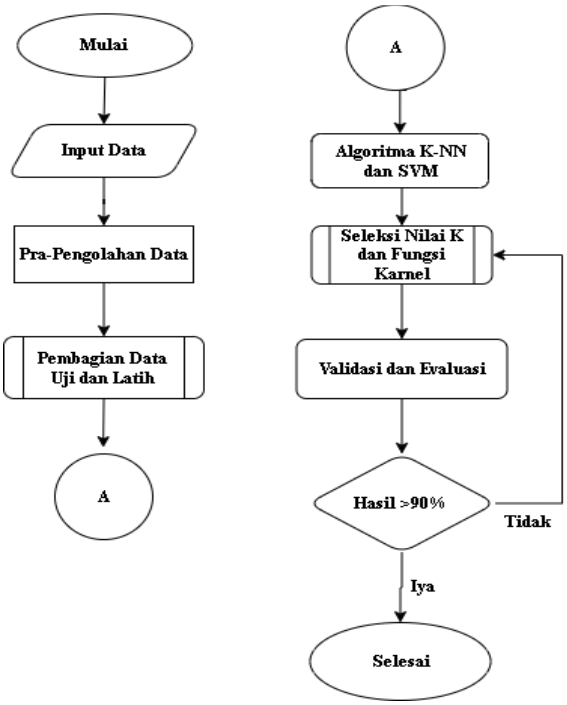

Figure 3. Diagram Flow Research.

A. Input Data

Data input is the process of entering data into the system. The data used has features or attributes that are used to classify classes or targets. Table 1 is an example of the data used .

TABLE I. SAMPLE DATA

Water Level	Class
639	1(Normal)
680	1(Normal)
735	2(Waspada)
845	3(Siaga)
889	3(Siaga)
934	3(Siaga)
931	3(Siaga)
966	4(Awas)
966	4(Awas)
961	4(Awas)
962	4(Awas)

B. Data Pre-Processing

Data pre-processing is the process of processing data before being used as training data and test data. At this stage, the data will be processed if there are anomalies in the data. In this study, data preprocessing was carried out to change the *string* data type on the class label into *integer* data to be entered into the system.

C. Sharing Test Data and Training Data

Sharing test data and training data is a process of sharing data. In this section, what percentage of test data and training data will be used in the study. Training and testing data in this study comprised 55% and 45% of the total 564 data used.

D. KNN and Svm Algorithms

At this stage, the data that has been shared is then integrated into machine learning or KNN and SVM machine algorithms.In the KNN algorithm, there is a method to find the closest distance from a data. One of the methods or equations contained in the KNN algorithm is *Eucledien* Distance, *Manhattan distance* and *Minkowski Distance*[18].

Eucledien Distance is a method for measuring the distance from 2 (two) points in geometric geometric space (covering two-dimensional, three-dimensional, or even more geometric geometric planes). The *Eucledien Distance* formula is as follows.

$$d(x_1, x_2)\sqrt{\sum_{i=0}^{n}(x_{1i} + x_{2i})^2} \quad (2)$$

Manhattan distance is used to calculate the absolute difference between the coordinates of a pair of objects. The *Manhattan distance* formula is as follows.

$$d(x, y)\sum_{i=1}^{n}|x_1 - y_1| + |x_2 - y_2| \quad (3)$$

Minkowski is a metric in a vector space in which a norm is defined (*normed vector space*). Minkowski is also considered a generalization of the eucledien distance and Manhattan distance. To calculate the distance of an object by the minkowski distance method, the p value is either 1 or 2.

$$d(x, y) = (\sum_{i=1}^{n}|x_i - y_i|^p)^{1/p} \quad (4)$$

Information:
d = Distance between x and y
x = Cluster data center
y = Data on attributes
i = Any data
n = Amount of data
x_i = Data on the i-th cluster
y_i = Data pata every i-th record
p = *Power*

In the SVM algorithm, there is amathematical calculation to translate data into higher dimensions. SVM algorithms have karnel features to address more complex data problems. Karnels contained in the SVM algorithm include Linear, Sigmoid, Polynominal, and Radial Basis Funcion (RBF) [19]. The SVM equation is as follows.

$$y(i) * (w * x(i) + b) >= 1 \quad (5)$$

Information:

w = the weight vector
b = biased
x = a feature vector
y = the class label

E. Selection of K Values and Carnel Functions

At this stage, K value selection is one of the processes contained in the KNN algorithm, namely how many K values will be used for data training. The K value is an independent variable which means that the value is free to enter regardless

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v13i3.2166, Copyright ©2024
Submitted : Mey 19, 2024, Revised : July 26, 2024, Accepted : August 12, 2024, Published : November 16, 2024

of the data limit used.

F. Validation and Evaluation

This stage is the process of validating and looking for accuracy from the model that has been created. Good accuracy is above 90% [20]. The process of validation and evaluation of the confusion matrix.

TABLE II. CONFUSION MATRIX 4 CLASS.

	Prediction Class (Recognized)			
Aktual	***Normal(A)***	***Siaga(B)***	***Bahaya(C)***	***Awas(D)***
(A)	t_{pA}	e_{AB}	e_{AC}	e_{AD}
(B)	e_{BA}	t_{pB}	e_{BC}	e_{BD}
(C)	e_{CA}	e_{CB}	t_{pC}	e_{CD}
(D)	e_{DA}	e_{DB}	e_{DC}	t_{pD}

In the case of classification, can use the following equation.

$$\text{Accuracy} = \frac{(TP + TN)}{(TP + FP + FN + TN)} x100\% \quad (6)$$

IV. RESULTS AND DISCUSSION

The data used is data on the water level of the Ciliwung River at the Manggarai floodgate monitoring point, which was taken from the DKI Jakarta open data website in 2020. The amount of data used was 564.

This study's training and testing data consisted of 55% and 45% of the total 564 data used. The scenario carried out in this study for the KNN algorithm is to use K values from the range of 1 to 10, while the SVM algorithm uses C parameters from the range of 1 to 10 and uses the RBF kernel in the training data. Figure 2 shows the accuracy of KNN and SVM.

TABLE III. MODEL ACCURACY

K value	Parameter C	Karnel	Accuracy KNN	Accuracy SVM
1	1	rfb	100%	100%
2	2	rfb	100%	100%
3	3	rfb	100%	100%
4	4	rfb	100%	100%
5	5	rfb	100%	100%
6	6	rfb	98%	100%
7	7	rfb	98%	100%
8	8	rfb	98%	100%
9	9	rfb	98%	100%
10	10	rfb	98%	100%
	Average	-	**99%**	**100%**

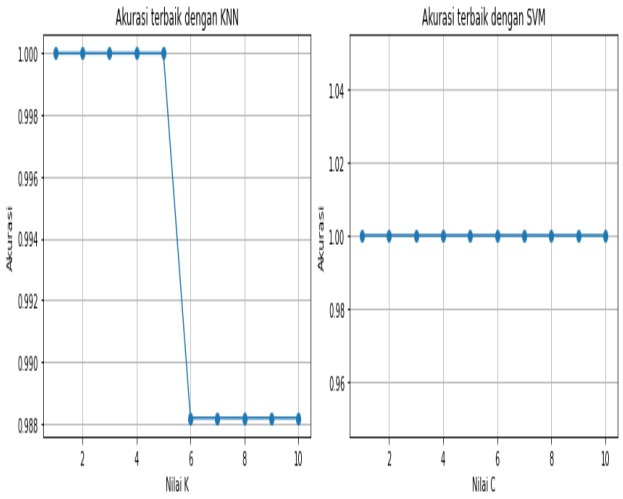

Figure 4. KNN and SVM ccuracy.

Based on Figure 2. It can be seen that in the KNN algorithm, scenarios from K values range 1 to 10 only K= 1, 2, 3, 4, and 5 values have an accuracy value of 100%. While in the SVM algorithm, scenarios from parameter range C 1 to 10 have an accuracy value of 100%. The SVM algorithm has better accuracy influenced by the use of the right carnel and in this study the data used is also not too large. The SVM algorithm has more flexible karnel to get good accuracy values. While in the KNN algorithm, there is a K value as a parameter to determine the accuracy value. The KNN algorithm doesn't look good enough at determining classes, especially on a more complex number of data classes.

Here is the heat map of the confusion matrix to K = 5 values in the KNN algorithm.

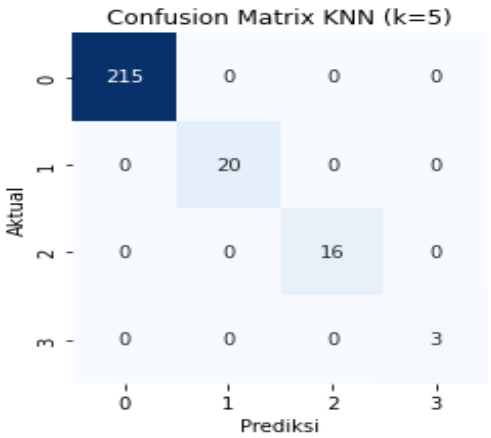

Figure 5. Heat Map Value K = 5.

In Figure 2. there is a heat map of the value K= 5. Figure 2. is a confusion matrix to find the accuracy of the model that has been made. If the calculation is carried out using equation (1), it will produce the following accuracy.

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v13i3.2166, Copyright ©2024
Submitted : Mey 19, 2024, Revised : July 26, 2024, Accepted : August 12, 2024, Published : November 16, 2024

$$\frac{215+20+16+3}{215+20+16+3} x100\% = 100\%$$

Here is the heat map of the confusion matrix for contuh parameter C = 6 in the SVM algorithm.

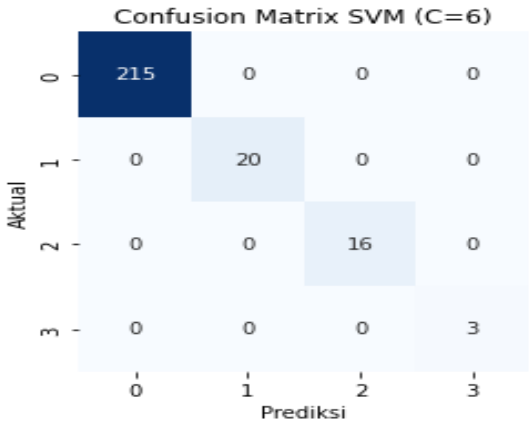

Figure 6. Heat Map Parameter C= 6.

In Figure 3. there is a confusion matrix for the SVM algorithm with a value of C= 6. If equation (1) is used to find the accuracy of the model that has been made, then the calculation is as follows.

$$\frac{215+20+16+3}{215+20+16+3} x100\% = 100\%$$

Based on the results of the analysis, the SVM algorithm has better result accuracy than the KNN algorithm. It is evident from the visualization results contained in Figure 2. High accuracy is greatly influenced by the complexity of the data and the features used.

V. CONCLUSION

Based on the research results, the SVM algorithm has better accuracy than the KNN algorithm. The accuracy generated by the SVM algorithm of all C parameters performed is 100%. As for the KNN algorithm, the best accuracy value is only obtained in the scenario of K values = 1, 2, 3, 4, and 5, which is 100%. This means the SVM algorithm better classifies water levels than the KNN algorithm. The SVM algorithm has a more flexible kernel to get good accuracy values. In the KNN algorithm, a K value is used to determine the accuracy value. The KNN algorithm needs to look better at determining classes, especially on a more complex number of data classes.

REFERENCES

[1] A. Rosyida, R. Nurmasari, and Suprapto, "Analisis Perbandingan Dampak Kejadian Bencana Hidrometeorologi dan Geologi di Indonesia dilihat dari Jumlah Korban dan Kerusakan (Studi: Data Kejadian Bencana Indonesia 2018)," *J. Dialog Penanggulangan Bencana*, vol. 10, no. 1, pp. 12–21, 2019, [Online]. Available: https://perpustakaan.bnpb.go.id/jurnal/index.php/JDPB/article/download/127/97/204

[2] R. Karno and J. Mubarrak, "Analisis spasial (ekologi) pemanfaatan daerah aliran sungai (das) di sungai batang lubuh kecamatan rambah kabupaten rokan hulu," *J. Ilm. Edu Res.*, vol. 7, no. 1, pp. 1–4, 2018.

[3] R. Al Fauzi, "Analisis Tingkat Kerawanan Banjir Kota Bogor Menggunakan Metode Overlay dan Scoring Berbasis Sistem Informasi Geografis," *Geomedia Maj. Ilm. dan Inf. Kegeografian*, vol. 20, no. 2, pp. 96–107, 2022, doi: https://doi.org/10.21831/gm.v20i2.48017.

[4] Q. S. Slat, "Analisis Debit Banjir dan Tinggi Muka Air Sungai Pinateduan di Desa Tatelu Kabupaten Minahasa Utara," *J. Sipil Statik*, vol. 8, no. 3, pp. 403–408, 2020.

[5] Anggraini; Asfilia Nova, "Prediksi Status Banjir Sungai Ciliwung Untuk Deteksi Dini Bencana Banjir Menggunakan Artificial Neural Network Backpropagation.," Universitas Islam Negeri Maulana Malik Ibrahim, 2022.

[6] I. K. A. Sari and D. Sebayang, "Analisis Banjir Dan Tinggi Muka Air Pada Ruas Sungai Ciliwung," *Forum Mek.*, vol. 7, 2018, [Online]. Available: https://doi.org/10.33322/forummekanika.v7i1.85

[7] D. Hartanti and A. Ichsan, "Komparasi Algoritma Machine Learning dalam Identifikasi Kualitas Air," vol. 9, no. 1, pp. 1–6, 2023.

[8] I. M. Faiza, G. Gunawan, and W. Andriani, "Tinjauan Pustaka Sistematis: Penerapan Metode Machine Learning untuk Deteksi Bencana Banjir," *J. Minfo Polgan*, vol. 11, no. 2, pp. 59–63, 2022, doi: 10.33395/jmp.v11i2.11657.

[9] C. Algoritma and N. Bayes, "Perbandingan Metode Data Mining untuk Prediksi Curah Hujan dengan," *Penelit. dan Pengabdi. Masy.*, vol. 6, pp. 187–197, 2022, [Online]. Available: https://journal.irpi.or.id/index.php/sentimas Prosiding

[10] A. Naïve, "Comparison of Data Mining Methods for Predition of Floods with Naïve Bayes and KNN Algorithm Perbandingan Metode Data Mining untuk Prediksi Banjir Dengan," pp. 40–48, 2022.

[11] A. A. Nurkhaliza and A. W. Wijayanto, "Perbandingan Algoritma Klasifikasi Support Vector Machine dan Random Forest pada Prediksi Status Indeks Mitigasi dan Kesiapsiagaan Bencana (IMKB) Satuan Kerja BPS di Indonesia Tahun 2020," *Maret*, vol. 7, no. 1, pp. 54–59, 2022, [Online]. Available: http://openjournal.unpam.ac.id/index.php/informatika54

[12] F. Yustiasari Liriwati, "Transformasi Kurikulum; Kecerdasan Buatan untuk Membangun Pendidikan yang Relevan di Masa Depan," *J. IHSAN J. Pendidik. Islam*, vol. 1, no. 2, pp. 62–71, 2023, doi: 10.61104/ihsan.v1i2.61.

[13] I. Darmayanti, P. Subarkah, L. R. Anunggilarso, and J. Suhaman, "Prediksi Potensi Siswa Putus Sekolah Akibat Pandemi Covid-19 Menggunakan Algoritme K-Nearest Neighbor," *JST (Jurnal Sains dan Teknol.*, vol. 10, no. 2, pp. 230–238, 2021, doi: 10.23887/jstundiksha.v10i2.39151.

[14] V. Rani, S. T. Nabi, M. Kumar, A. Mittal, and K. Kumar, "Self-supervised Learning: A Succinct Review," *Arch. Comput. Methods Eng.*, vol. 30, no. 4, pp. 2761–2775, 2023, doi: 10.1007/s11831-023-09884-2.

[15] S. Tibaldi, G. Magnifico, D. Vodola, and E. Ercolessi, "Unsupervised and supervised learning of interacting topological phases from single-particle correlation functions," *SciPost Phys.*, vol. 14, no. 1, pp. 1–18, 2023, doi: 10.21468/SciPostPhys.14.1.005.

[16] P. Putra, A. M. H. Pardede, and S. Syahputra, "Analisis Metode K-Nearest Neighbour (Knn) Dalam Klasifikasi Data Iris Bunga," *J. Tek. Inform. Kaputama*, vol. 6, no. 1, pp. 297–305, 2022.

[17] N. S. H. Pratama, D. T. Afandi, M. Mulyawan, I. Iin, and N. D. Nuris, "Menurunkan Presentase Kredit Macet Nasabah Dengan Menggunakan Algoritma K-Nearest Neighbor," *Inf. Syst. Educ. Prof. J. Inf. Syst.*, vol. 5, no. 2, p. 131, 2021, doi: 10.51211/isbi.v5i2.1537.

[18] A. K. Clustering, "Perbandingan Akurasi Euclidean Distance , Minkowski Distance , dan Manhattan Distance pada Algoritma K- Means Clustering berbasis Chi-Square," no. July, 2019, doi: 10.30591/jpit.v4i1.1253.

[19] K. Kunci, "p-ISSN: 2774-6291 e-ISSN: 2774-6534 Available online at http://cerdika.publikasiindonesia.id/index.php/cerdika/index," vol. 3, no. September, pp. 828–839, 2023.

[20] Sholeh, D. Andayati, and Y. Rachmawati, "Data Mining Model Klasifikasi Menggunakan K-Nearest Neighbor With Normalization For Diabetes Prediction," *TeIka*, vol. 12, no. 1, pp. 77–87, 2022.

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v13i3.2166, Copyright ©2024
Submitted : Mey 19, 2024, Revised : July 26, 2024, Accepted : August 12, 2024, Published : November 16, 2024