INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH AND ANALYSIS

ISSN(print): 2643-9840, ISSN(online): 2643-9875

Volume 05 Issue 09 September 2022

DOI: 10.47191/ijmra/v5-i9-21, Impact Factor: 6.261

Page No. 2354-2464

Machine Learning Approach to Determine the Drug-Prone Areas in Lhokseumawe City, Indonesia

Novia Hasdyna[1], Sujacka Retno[2]

[1,2] Faculty of Computer and Multimedia, Universitas Islam Kebangsaan Indonesia

ABSTRACT: This research aims to apply Machine Learning algorithms, such as K-means and K-nearest neighbor into a system that can determine drug-prone areas in Lhokseumawe City, Indonesia. The results of the k-means calculation in some clusters of areas that are very prone, prone and not prone to drugs based on drug user data with some criterias, namely the number of users, sub-district, education, occupation and types of drugs. Based on the research that tested by using the k-means algorithm on data on drug users in 2018-2021, the average percentage of clusters in very prone areas is 49.50%, the percentage of clusters in prone areas is 29.26% and the percentage of clusters in non-prone areas is 21.23%. Furthermore, the results of testing with the K-NN algorithm with 98 training data and 20 testing data obtained an average accuracy value of 87%.

Keywords- Machine Learning, K-means, K-Nearest Neighbor, Drug-Prone

I. INTRODUCTION

The development of Machine Learning in technology is so rapid and also with the data that is produced so complex that form in one volume of data so that this data is very difficult to process manually. With the development of technology nowadays many technologies created to process data becomes a result or conclusion that can be obtained from the data itself. One of the most popular technologies today are cloud computing and Geographic Information System (GIS), in addition to analyzing data, various methods and algorithms are needed to obtain an efficient data conclusions [1].

Many algorithms can be used in Machine Learning such as classification, clustering, association, regression and many more. This method is used according to data analytic in need, an efficient and accurate method for determining the scale from high to low is the Clustering method (grouping) [2]. Algorithms that can be used for the clustering process are k-means, k-medoid, fuzzy c-mean, while machine learning algorithms that can be used for data classification are k-nearst neighbor, support vector machine and Naive Bayes algorithm [3][4][5].

The purpose of this study is to analyze and apply machine learning algorithms, which are k-means and k-nearest neighbor into the system based on drug users data in Lhokseumawe City in 2018-2021 obtained from the National Narcotics Agency of Lhokseumawe City. The K-nearest neighbor algorithm will classify drug user data based on the cluster results from the k-means algorithm so that it gets results which areas are very prone, prone and not prone to drugs in Lhokseumawe City, Indonesia. The system can determine drug-prone areas in Lhokseumawe CIty based on criterias such as the number of users, sub-districts, education, occupations and types of drugs.

The urgency of the research is currently the Lhokseumawe city government is finding it difficult to determine clusters of drug-prone areas and circulation based on data that has been recorded at the National Narcotics Agency of Lhokseumawe City. The government still lacks information to take further a action, for handling and preventing drugs with the number of users. Drugs in Aceh recorded in 2019, about 73,000 drug addicts from various types of drugs, spread across all sub-districts and urban districts in Aceh, Indonesia. Lhokseumawe City with territorial areas directly adjacent to the State of Malaysia which is connected by the Malacca Strait becomes a narcotics transit area for Indonesia and is possible to become a drug trafficking area.

Therefore, this research implements an application or system that can map drug-prone areas based on the results of the analysis and application of machine learning models or algorithms. Using machine learning algorithms, such as k-means and k-nearest neighbor, the results of the cluster and data classification of drug-prone areas can make it easier for the public to access the information quickly and also assist the Lhokseumawe City government and related agencies in making policies for handling and drug prevention in the city of Lhokseumawe, Indonesia.

II. STUDI LITERATURE

A. *Machine Learning*

Machine learning is a sub-field of artificial intelligence, a study on how computers are able to learn from data to improve their intelligence [6]. Machine learning is a field of computer science that provides learning capabilities on how computers know things without obvious programming [7]. Machine learning is also used to solve various problems. Algorithms from machine learning are divided into three categories, which is supervised learning, unsupervised learning, and reinforcement learning [8].

Supervised learning is an algorithm that learns based on a set of examples of the desired input and output pairs. This algorithm observes these examples and then produces a model that is able to map the new input to the correct output [9]. Some supervised learning algorithms are k-nearest neighbor, support vector machine, and naive Bayes [10]. While unsupervised learning is a machine learning algorithm whose learning does not use training data, the data given to the system is data that does not have a data label so that the system will study the data based on the features given to it. There is no exact output provided explicitly. One of the most commonly used unsupervised learning algorithms is Clustering or grouping, such as k-means, k-medoid, and fuzzy c-means [11].

B. *K-Means*

K-means is one of the methods in machine learning that can group data or clustering a data into the form of one or more clusters so that data with the same characteristics are grouped into the same cluster and data with different characteristics are grouped into different groups [12]. In detail, K-means algorithm works is as follows:

a. Determine the value of k as the number of clusters to be formed.
b. Determine the random value for the initial cluster center of k centroids, to calculate the distance of each input data to each centroid by using the Euclidean Distance formula as follows:

$$d(xi, \mu j) = \sqrt{\sum (xi - \mu j)^2} \quad (1)$$

Where: xi = data criteria, μj = centroid in js-cluster

c. Group each data based on its nearest to the centroid or find the smallest distance.
d. Update the new centroid value, the new centroid value is obtained from the average of the cluster using the formula:

$$\mu j(t+1) = \frac{1}{Nsj} \sum_{j \in sj} xj \quad (2)$$

Where: μj(t+1) = new centroid in an iteration (t+1), Nsj = Data in Sj-Cluster

e. If the data of each cluster has not stopped, repeat from steps 2 to 5, until the members of each cluster have not changed [13].

C. K-Nearest Neighbor

In machine learning, a method to classify a set of data based on data learning that has been previously classified is K-Nearest Neighbor [14]. K-NN is an algorithm that is included in the supervised learning group, where the results of the new query instance are classified based on the majority of the nearest value to the categories in K-NN. The way the K-Nearest Neighbor algorithm works is to find the closest distance between the data to be evaluated with the closest k-neighbors in a training data. [15].

The following is the sequence process of the K-Nearest Neighbor algorithm:

a. Determine Parameter k number of nearest neighbors.
b. Calculate the distance of each data to the existing data sample, in this research using the Euclidean distance according to equation 3:

$$di = \sqrt{\sum_{i=1}^{p} (x_{2i} - x_{1i})^2} \quad (3)$$

Where:

x1 = Data Sample, x2= Data Testing, I= Data Variable, d= Distance, p= Data Dimension.

c. Sort these data into groups that have a smaller Euclidean distance.
d. Collect Y category (Nearest Neighbor Classification) [16].

III. METHODOLOGY

In this research method, the steps used consist of four stages as follows:

1. Conducting Literature Studies, by collecting relevant primary library/reference sources.
2. Conducting Surveys and Data Collection.
   The data used in this research is data on drug users obtained from the National Narcotics Agency of Lhokseumawe City for 2018-2021 period.
3. Perform data analysis with machine learning algorithms, such as k-means and k-nearest neighbor. At this stage, the first step is the data will be clustered by using k-means algorithm according to equations (1) and (2). After getting the results of the cluster with the k-means algorithm, it will continue the process of calculating and analyzing data using the k-nearest neighbor algorithm according to equation (3) to obtain data classification of drug-prone areas in Lhokseumawe City.
4. Applying the results of data analysis and machine learning algorithms in the form of clustering results with k-means and data classification results with k-nearest neighbors into the system. This stage consists of designing a system and implementing a system that can map drug-prone areas in Lhokseumawe City.

The research steps in more detail can be seen in the research block diagram shown in Figure 1.

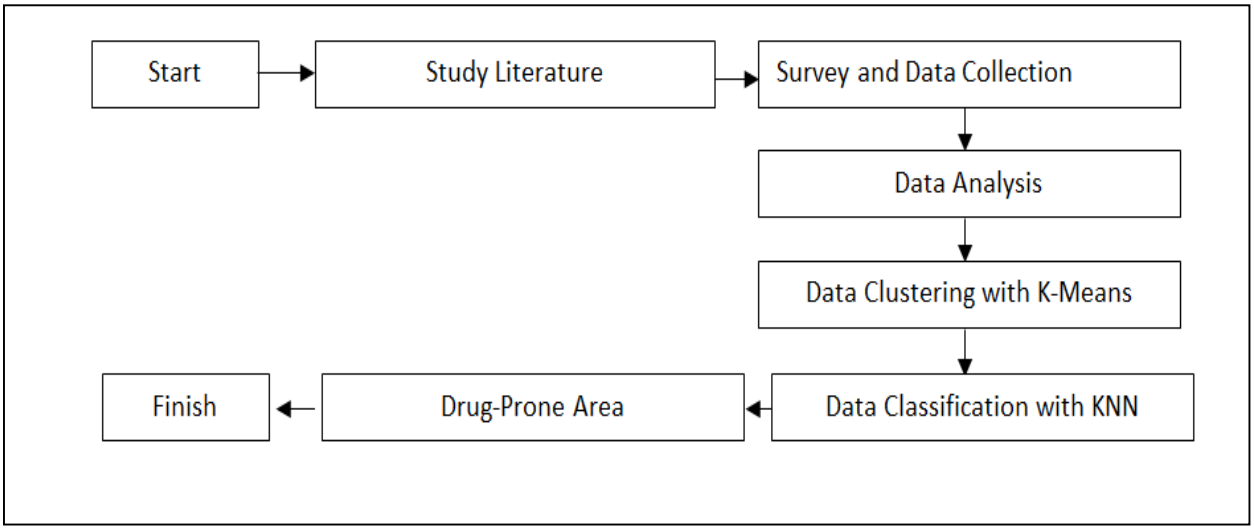


Fig 1. Research Block Diagram

IV. RESULT AND DISCUSSION

A. Data Weighting

Table I. Data Criteria

Criteria Code	Criteria
X1	Number of Users
X2	Sub-district
X3	Occupations
X4	Educations
X5	Type of Drugs

In Table 1, the criteria codes consist of X1, X2, X3, X4 and X5. X1 is the number of users, X2 is the sub-district, X3 is the occupations, X4 is the educations and X5 is the type of drugs.

Table II. Sub-District

Attribute	Value	Weight
Sub-District	Muara Dua	1
	Muara Satu	2
	Banda Sakti	3
	Blang Mangat	4

In Table 2, it consists of sub-district data in Lhokseumawe City, Indonesia, which consists of 4 sub-districts, the sub-districts of Muara Dua, Muara Satu, Banda Sakti and Blang Mangat..

Table III. Occupation Data

Attribute	Value	Weight
Occupation	Self-employed	7
	Civil Servants	8
	Farmer	9
	College Student	10
	Police	11
	Student	12
	Housewife	13
	Laborer	14

In Table 3, it consists of employment data, such as self-employed, civil servants, farmers, college students, police, students, housewives and laborers.

Table IV. Education Data

Attribute	Value	Weight
Educations	Elementary School	15
	Junior High School	16
	Senior High School	17
	College	18

In Table 4, it consists of educational data, such as elementary school, junior high school, senior high school, and college. Elementary school with a weight of 15, junior high school with a weight of 16, high school with a weight of 17 and college with a weight of 18.

Table V. Type of Drugs

Attribute	Value	Weight
Type Of Drugs	Marijuana	19
	Crystals	20
	Ecstasy	21

In Table 5, it consists of data on types of drugs, such as Marijuana, Crystals and Ecstasy. Marijuana weighing 19, Crystals weighing 20 and Ecstasy weighing 21.

B. K-Means Perhitungan Calculation Results

The results of the k-means test are shown in table 6.

Table VI. K-Means Testing Data

Number of Test	Centroid
1	1,2,3
2	4,3,5

Determination of the initial centroid or cluster center as testing data is carried out randomly or randomly. The steps taken are the first to initialize the centroid or the initial cluster center.

Table VII. Centroid Value

No	Village	XI	X2	X3	X4	X5
1	Alue Awe	5	1	7	17	19
2	Batuphat Timur	1	2	7	17	20
3	Batuphat Timur	1	2	7	17	19

The next step is the calculation of the Euclidean Distance Method.

To create a classification model, a number of training data and test data are needed. The next step is to determine the number of groups or clusters, the 25 existing drug use case data will be grouped into three clusters, namely: Very Prone, Prone, Not Prone. Calculating the distance between the first data and the first centroid is:

$d_{(1.1)} = \sqrt{(5-5)^2 + (1-1)^2 + (7-7)^2 + (17-17)^2 + (19-19)^2}$

= 0

Calculating the distance between the first data and the second centroid is:

$d_{(2.1)} = \sqrt{(5-1)^2 + (1-2)^2 + (7-7)^2 + (17-17)^2 + (19-20)^2}$

= 4,242641

Calculating the distance between the first data and the third centroid is:

$d_{(3.1)} = \sqrt{(5-1)^2 + (1-2)^2 + (7-7)^2 + (17-17)^2 + (19-19)^2}$

= 4,123126

The same calculation is applied to all data. After calculating all data, it will get the closest distance for each data in each cluster as shown in table 8.

Table VIII. Closest Distance in Each Cluster

No.	Village	C1	C2	C3	Closest Distance
1	Alue Awe	0	4,242641	4,123126	C1
2	Batuphat Barat	4,358899	0	9,055385	C2
3	Batuphat Timur	9	9,055385	0	C3
4	Blang Naleung Mameh	6,928203	4,582576	13,22876	C2
5	Blang Panyang	3,464102	1	9,110434	C2
6	Blang Pulo	4	1,732051	10,14889	C2
7	Hagu Barat Laut	6,78233	6,082763	3,316625	C3
8	Hagu selatan	5,567764	3,464102	12,16553	C2
9	Hagu Teungoh	5,830952	6,403124	3,872983	C3
10	Jawa Lhokseumawe	10,34408	8,3666	17,20465	C2
11	Keude Aceh	8	6,244998	15,13275	C2
12	Keude Cunda	6,480741	9,539392	15,06652	C1
13	Meunasah Alue	3,741657	3,316625	6,244998	C2
14	Meunasah Mesjid	1,732051	3,464102	8,717798	C1
15	Meunasah Panggoi	1,414214	3,316625	8,660254	C1
16	Meuria Paloh	4,358899	0	9,055385	C2
17	Mon Geudong	6,403124	4,472136	13,19091	C2
18	Padang Sakti	12,76715	11,04536	20,04994	C2
19	Paloh Punti	7,416198	5,09902	14,07125	C2
20	Pusong Baru	4,582576	1,414214	8,124038	C2
21	Pusong Lhokseumawe	4,582576	1,414214	8,124038	C2
22	Tumpok Teungoh	4,242641	4,582576	5,567764	C1
23	Ujong Blang	4,795832	2,44949	7,211103	C2
24	Uteun Bayi	4,795832	4,242641	5,291503	C2
25	Uteunkot	8,124038	7,681146	16,34013	C2

Next step is Calculating the new Centroid.

After each data distance is calculated for each cluster, the next step is to group the data according to the cluster, the cluster group of a data is calculated from the shortest distance from the data to a cluster, add up all the data contained in the same cluster group and then divide by the number of data, using formula:

$C1_{X1} = \frac{26}{5} = 5.2$

$C1_{X2} = \frac{7}{5} = 1.4$

$C1_{X3} = \frac{141}{5} = 28.2$

$C1_{X4} = \frac{36}{5} = 7.2$

$C1_{X5} = \frac{85}{5} = 17$

$C1_{X6} = \frac{99}{5} = 19.8$

$C2_{X1} = \frac{27}{17} = 1.5882$

$C2_{X2} = \frac{40}{17} = 2.3529$

$C2_{X3} = \frac{531}{17} = 31.2352$

$C2_{X4} = \frac{128}{17} = 7.52941$

$C2_{X5} = \frac{284}{17} = 16.7058$

$C2_{X6} = \frac{340}{17} = 20$

$C3_{X1} = \frac{5}{3} = 1.6666$

$C3_{X2} = \frac{8}{3} = 2.6666$

$C3_{X3} = \frac{66}{3} = 22$

$C3_{X4} = \frac{21}{3} = 7$

$C3_{X5} = \frac{51}{3} = 17$

$C3_{X6} = \frac{59}{3} = 19.6666$

Placement of data in the cluster with the closest distance to the cluster, namely obtaining data close to C1=5, C2=17, C3=3. After the data is grouped according to the cluster, the results of the new centroid can be seen in Table 9.

Table IX. New Centroid Value In Fisrt Iteration

Centroid	X1	X2	X3	X4	X5
C1	5.2	1.4	7.2	17	19.8
C2	1.5882352	2.35294	7.52941	16.70588	20
C3	1.6666666	2.66666	7	17	19.6666

The next step is to compare the result with the previous centroid value, if the value is the same then the iteration process is stopped. However, if the values are not the same, the steps of the data grouping process are repeated. The following are the centroid values obtained in this calculation.

Table X. Second Iteration

Centroid	XI	X2	X3	X4	X5
C1	5	1	7.2	17	19.8
C2	1.533333333	2.4	7.6	16.66667	20
C3	2	2.8	7	17	19.8

In the third iteration the search steps are carried out the same as the steps in the first and second iterations. The iteration is continued until the centroid value is non-repetitive and the same. The new centroid value in the third iteration can be seen in Table 11.

Table XI. Third Iteration

Centroid	XI	X2	X3	X4	X5
C1	5.75	1	7.25	17	19.75
C2	1.5714	2.35714	7.57142	16.6428	20
C3	1.8571	2.571429	7.14285	17	19.8571

In the fourth iteration and so on, the search steps are the same as in the previous search. The new centroid value in the fourth iteration can be seen in table 12.

Table XII. Fourth Iteration

Centroid	XI	X2	X3	X4	X5
C1	5.75	1	7.25	17	19.75
C2	1.666667	2.25	7.66666	16.58333	20
C3	1.666667	2.66666	7.11111	17	19.8888

The new centroid value in the fifth iteration can be seen in table 13.

Table XIII. Fifth Iteration

Centroid	XI	X2	X3	X4	X5
C1	5	1.2	7.2	17	19.8
C2	1.777778	2.33333	7.88888	16.44444	20
C3	1.545455	2.54545	7.09090	17	19.9090

The new centroid value in the sixth iteration can be seen in the table below:

Table XIV. Sixth Iteration

Centroid	XI	X2	X4	X5	X6
C1	4.5	1.33333	7.33333	17	19.8333
C2	1.75	2.375	7.875	16.375	20
C3	1.545455	2.54545	7.09090	17	19.9090

The new centroid value in the seventh iteration can be seen in the table below:

Table XV. Seventh Iteration

Centroid	XI	X2	X3	X4	X5
C1	4.5	1.33333	7.33333	17	19.8333
C2	1.75	2.375	7.875	16.375	20
C3	1.545455	2.54545	7.09090	17	19.9090

In this study the K-means iteration stopped at the seventh iteration with an average value of C1 = 13.2, C2 = 13.81481 and C3 = 12.30303.

C. K-NN Calculation Results

After calculating the results of the cluster with k-means, then an analysis of the k-nn calculation will be carried out on the data of drug users in 2018-2021 in Lhokseumawe City, Indonesia. The following are the results of the calculation of the k-nn distance in table 15.

Table XVI. Calculation results of K1, K2 and K3 values on the First test data

No	Distance	Class	K
2	1,414214	Tinggi	K3
8	1	Rendah	k1
18	1	Sedang	K2

Table XVII. Calculation results of K1, K2 and K3 values on the second test data

No	Distance	Class	K
1	1,732051	Rendah	K2
14	1,414214	Rendah	K1
15	1,732051	Rendah	K3

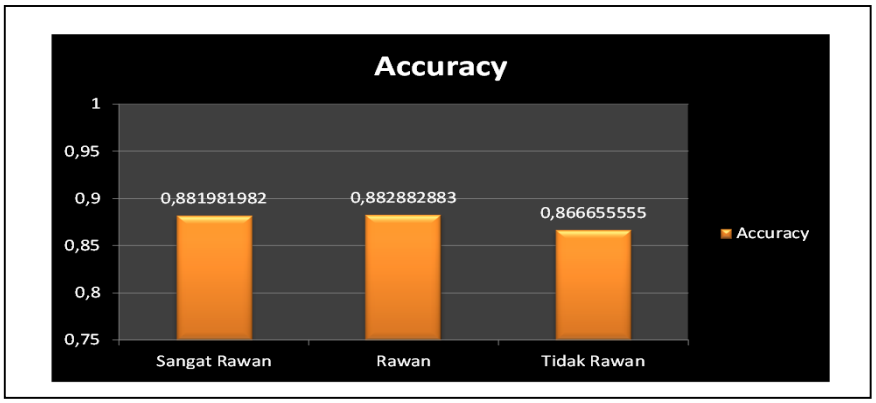
Table XVIII. Calculation results of K1, K2 and K3 values on the third test data

No	Distance	Class	K
4	3,464102	Sedang	K3
26	1	Rendah	K1
98	2,44949	Rendah	K2

Based on the results of the calculation of the values of K1, K2 and K3, the results of the calculation of accuracy are 87% which are shown in table 16.

Table XIX. K-NN Accuracy Calculation Results (ACC)

No.	Class	ACC
1	Sangat Rawan	0,881981982
2	Rawan	0,882882883
3	Tidak Rawan	0,866655555
Average Accuracy		**0,877173473**



D. System implementation

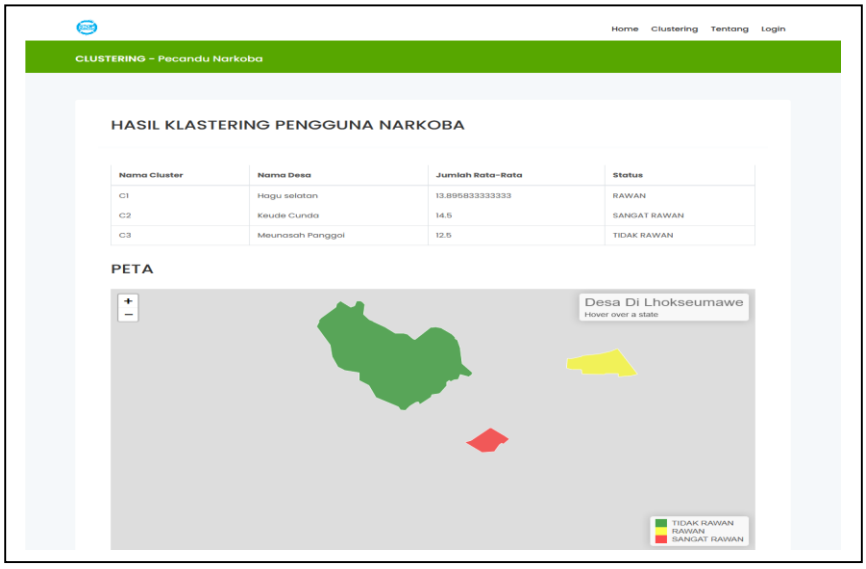


Fig.3. Drug User Data Clustering Page in Lhokseumawe City, Indonesia

Figure 2 contains information on clustering and this page used to determine drug-prone areas based on the selected area cluster. The classification results obtained from the application of machine learning algorithms to determine areas prone to drug users, in 2018 there were 4 villages with very prone status, 8 villages with prone status and 13 villages with non-prone status. In 2019 there were 19 villages with very prone status, 6 villages with prone status and 5 villages with non-prone status. In 2020 there are 24 villages with very prone status, 12 villages with prone status and 3 villages with non-prone status. In 2021 there are 20 villages with very prone status, 12 villages with prone status and 3 villages with non-prone status.

V. CONCLUSIONS

Based on the results of research that has been carried out, the application of machine learning algorithms to determine drug-prone areas in Lhokseumawe City, Indonesia has been successfully implemented. This research shows that the results of the analysis of the calculation of the k-means algorithm obtained that the average percentage of clusters in very prone areas is 49.50%, the percentage of clusters in prone areas is 29.26% and the percentage of clusters in non-prone areas is 21.23%. Furthermore, the results of testing with the K-NN algorithm with 98 training data and 20 testing data obtained an average accuracy value of 87%. Thus, this system can provide information related to drug-prone areas in Lhokseumawe City, Indonesia. So that it can assist the government in taking strategic steps in reducing the level of drug use in Lhokseumawe City, Indonesia.

ACKNOWLEDGMENT

The authors would like to thank the Ministry of Education and Culture, Indonesia, which has funded this research.

REFERENCES

1) Mete, M. O., & Yomralioglu, T. 2021 Implementation of serverless cloud GIS platform for land valuation. International Journal of Digital Earth, 14(7), 836-850.
2) Ağbulut, Ü., Gürel, A. E., & Biçen, Y. 2021 Prediction of daily global solar radiation using different machine learning algorithms: Evaluation and comparison. Renewable and Sustainable Energy Reviews, 135, 110114.
3) Dinata, R. K., Retno, S., & Hasdyna, N. 2021 Minimization of the Number of Iterations in K-Medoids Clustering with Purity Algorithm. Rev. d'Intelligence Artif., 35(3), 193-199.
4) Kang, M., Shin, S., Zhang, G., Jung, J., & Kim, Y. T. 2021 Mental Stress Classification Based on a Support Vector Machine and Naive Bayes Using Electrocardiogram Signals. Sensors, 21(23), 7916.
5) Dinata, R. K., Akbar, H., & Hasdyna, N. 2020 Algoritma K-Nearest Neighbor dengan Euclidean Distance dan Manhattan Distance untuk Klasifikasi Transportasi Bus. ILKOM Jurnal Ilmiah, 12(2), 104-111.
6) Hasdyna, N., Sianipar, B., & Zamzami, E. M. 2020 Improving the performance of K-nearest neighbor algorithm by reducing the attributes of dataset using gain ratio. In Journal of Physics: Conference Series (Vol. 1566, No. 1, p. 012090). IOP Publishing.

7) Luan, H., & Tsai, C. C. 2021 A review of using machine learning approaches for precision education. Educational Technology & Society, 24(1), 250-266.

8) Truong, G., Le, H., Suter, D., Zhang, E., & Gilani, S. Z. 2021 Unsupervised learning for robust fitting: a reinforcement learning approach. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 10348-10357).
9) Chakraborty, A., Alam, M., Dey, V., Chattopadhyay, A., & Mukhopadhyay, D. 2021 A survey on adversarial attacks and defences. CAAI Transactions on Intelligence Technology, 6(1), 25-45.
10) Nayeem, M. J., Rana, S., Alam, F., & Rahman, M. A. 2021 Prediction of Hepatitis Disease Using K-Nearest Neighbors, Naive Bayes, Support Vector Machine, Multi-Layer Perceptron and Random Forest. In 2021 International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD) (pp. 280-284). IEEE.
11) Vijaya, J., & Syed, H. (2021, April). A Performance Study of Probabilistic Possibilistic Fuzzy C-Means Clustering Algorithm. In International Conference on Advances in Computing and Data Sciences (pp. 431-442). Springer, Cham.
12) Dinata, R. K., Novriando, H., Hasdyna, N., & Retno, S. (2020). Reduksi atribut menggunakan information gain untuk optimasi cluster algoritma k-means. J. Edukasi dan Penelit. Inform, 6(1), 48-53.
13) Retno, S. 2019 Peningkatan Akurasi Algoritma K-Means dengan Clustering Purity Sebagai Titik Pusat Cluster Awal (Centroid). Repositori USU
14) Hasdyna, N., Nababan, E.B ., Efendi, S. 2019 Dimension Reduction in Datasets Using Information Gain To Enhance K-NN Performance. International Journal of Trend In Research and Development 6 (3), 379-383
15) Dinata, R. K., Fajriana, F., Zulfa, Z., & Hasdyna, N. 2020 Klasifikasi Sekolah Menengah Pertama/Sederajat Wilayah Bireuen Menggunakan Algoritma K-Nearest Neighbors Berbasis Web. CESS (Journal of Computer Engineering, System and Science), 5(1), 33-37.
16) Romadhon, M. R., & Kurniawan, F. 2021 A comparison of naive Bayes methods, logistic regression and KNN for predicting healing of Covid-19 patients in Indonesia. In 2021 3rd east Indonesia conference on computer and information technology (eiconcit) (pp. 41-44). IEEE.