2nd International Conference on Information Science and Technology Innovation (ICoSTEC)
February, 25 2023. Yogyakarta, Indonesia

Predicting Non-Performing Loan's Risk Level Using K-Means Clustering and K-Nearest Neighbors

Muhammad Mizan Siregar[1], Roslina[2], B. Herawan Hayadi[3]

[1,3]*Magister of Computer Science, Potensi Utama University*
JL. KL. Yos Sudarso Km. 6,5 No. 3-A, Medan

[2]*Departement of Computer and Informatics Technology, Politeknik Negeri Medan*
JL. Almamater No.1, Padang Bulan, Kec. Medan Baru, Medan

[1]mizan.siregar1@gmail.com

[2]roslinanich@gmail.com

[3]b.herawan.hayadi@gmail.com

***Abstract*— In data mining, clustering is an unsupervised learning technique often used to group data by similarity. Clustering, especially the K-means clustering algorithm, is a feasible tool for expanding a dataset label by increasing the cluster's number according to the label's categories. This research extends the credit loan label data set from two categories (non-performing and performing loans) to four risk levels (high risk, medium risk, low risk, and no risk). The combination of three K-nearest neighbor's distance metrics, Euclidean, Manhattan, and Chebyshev distance, with four different K values (K = 3, K = 5, K = 7, and K = 9) produced the best model with accuracy, precision, and recall values of 90%, 90.53571%, and 90%, from the model using the Euclidean distance with K = 9.**

***Keywords*— credit loan, k-means clustering, k-nearest neighbors, risk level**

I. INTRODUCTION

Credit analysis is a principle in the credit risk assessment for small and medium-sized businesses to measure customer creditworthiness in terms of finances [1]. The cooperative, as the lender, will conduct a survey of prospective credit recipients through the 5C analysis (character, capacity, capital, economic conditions, and collateral) to minimize the risk of non-performing loans happening [2].

Data mining can assist cooperatives in analyzing the credit recipients' potential for non-performing loans by comparing the previous credit granting data with the survey data of prospective credit recipients and classifying them in the form of bad credit or non-bad credit classifications [3]. With the selection, exploration, and modeling of previous data, data mining can find knowledge in the form of relationships between one feature and another previously unknown feature [4]. This process, known as knowledge discovery in the database (KDD), generates output such as patterns and relationships between data for further processing using a machine learning algorithm [5].

In the field of non-performing loan prediction using data mining, a study of non-performing loan predicting using 8 data mining algorithms states that we can forecast problematic debtors from their payment history data utilizing machine learning [6]. Another study compares seven machine learning algorithms and shows the four aspects in the dataset, such as clients' historical payment behaviors, business card payments and risks, types of loan products, and customer tenures, that influence the prediction of a non-performing loan [7]. The socio-demographics (age, gender, marital status, and province of residence) play a crucial role in predicting a non-performing loan, as shown in the study of loan recovery rate forecast using linear, non-linear, and rule-based machine learning methods [8].

Clustering is a data mining technique that divides data into several clusters by looking at the level of similarity between data, making it easier to identify the data [9]. Using clustering, especially the K-means clustering algorithm, we can easily divide the data into groups within the scope of unsupervised learning and generate new dataset labels for the subsequent classification process [10]. Recent studies have shown the feasibility of this method, such as combining the K-nearest neighbors (K-NN) and random forest algorithms for the ECG data classification [11], a student's academic performance classification using K-NN and K-means clustering [12], and a clustered K-NN for large data classification [13].

K-NN was chosen as the combination algorithm because it has almost the same data clustering principle as K-means clustering, which is grouping the data groups based on the closest distance using a predetermined number (K) of neighbors [14]. The core of this algorithm is the calculation of the data's distance (distance metric), where the selection of different distance metrics will affect the performance of this algorithm in classifying data [15]. In this research, we implement the euclidean, manhattan, and Chebyshev distances as a comparison to find the best model.

The purpose of this research is to combine the K-NN algorithm with K-means clustering to classify the risk level of the credit assessment. We expand the dataset labels into four risk levels (high risk, medium risk, low risk, and no risk) with K-means clustering and then use the clustering result as new labels for the classification using the K-NN algorithm. We use three different K values (K = 3, K = 5, and K = 7) to analyze the performance values (accuracy, precision, and recall) with a 10-fold cross-validation. Using the combination of the distance metrics, K-NN's K values, and the 10-fold cross-validation, we summarize which combination produces the best model for this classification problem.

II. Methods

A. Dataset

In this study, we use a dataset from previous research about non-performing loan prediction in Mutiara Sejahtera Cooperative [16]. This dataset consists of 60 data, with five categories (full-time employees, membership length, loan amount, loan duration, and loans elsewhere) and one label (non-performing loan). Table 1 shows the 10 dataset samples from each non-performing and performing loan label.

TABLE I
Sample of Dataset

FT	ML	LA	LD	LE	NPL
N	JM	M	M	Y	Y
Y	NM	B	M	Y	Y
N	JM	B	S	Y	Y
Y	NM	B	M	Y	Y
Y	SM	M	M	N	Y
Y	SM	M	S	N	N
N	NM	M	L	N	N
Y	SM	S	L	N	N
Y	SM	M	M	N	N
N	NM	M	L	N	N

Notes: FT = Full-time member (Yes, No), ML = Membership length (New Member, Junior Member, Senior Member), LA = Loan amount (Small, Medium, Big), LD = Loan duration (Short, Medium, Long), LE = Loans elsewhere (Yes, No), NPL = Non-performing loan (Yes, No)

We normalized the data from Table 1 to make the clustering process easier. Table 2 shows the normalization rules, while Table 3 shows the results.

TABLE II
Normalization Rules

Category	Old Value	New Value
FT	No	0
	Yes	1
ML	New Member	0
	Junior Member	1
	Senior Member	2
LA	Small	0
	Medium	1
	Big	2
LD	Short	0
	Medium	1
	Long	2
LE	No	0
	Yes	1

In the Full-Time Member category, we change the "No" value to 0 and the "Yes" value to 1. In the Member Length category, we change the "New Member" value to 0, the "Junior Member" value to 1, and the "Senior Member" value to 2. In the Loan Amount category, we change the "Small" value to 0, the "Medium" value to 1, and the "Big" value to 2. In the Loan Duration category, we change the "Short" value to 0, the "Medium" value to 1, and the "Long" value to 2. In the Loans Elsewhere category, we change the "No" value to 0 and the "Yes" value to 1.

TABLE III
Normalization Results

FT	ML	LA	LD	LE	NPL
0	1	1	1	1	1
1	0	2	1	1	1
0	1	2	0	1	1
1	0	2	1	1	1
1	2	1	1	0	1
1	2	1	0	0	0
0	0	1	2	0	0
1	2	0	2	0	0
1	2	1	2	0	0
0	0	1	2	0	0

B. K-Means Clustering

K-means clustering is an algorithm based on a non-hierarchical clustering method, often used to separate data into two or many groups with similar characteristics [17]. This algorithm uses a random point in initializing the first centroid for each cluster and a simple way to partition the data into clusters [18].

Some studies have shown the performance of this algorithm in the prediction problems, such as mapping the Jamkesda recipient candidates using three different clusters (K = 2, K = 3, and K = 4), with the Davies Bouldin Index values: 0.243, 0.256, 0.275 [19]; predicting scholarship recipient with a combination of simple additive weighting (SAW) method, where the clustering results were ranked using the SAW method [20]; and the optimization of cluster number for the single tuition scholarship prediction, with the most optimum K value of 6 with the silhouette coefficient value of 0.20212705 [21]. These studies are the reference we use in this research, whether in the K-means steps or choosing the K value.

We use the K-means clustering algorithm to group the data into clusters based on the K value. This clustering process uses the K-means clustering algorithm steps, as shown below [22]:

1. Determine the K value
   We use two different K values (K = 2 and K = 4) to group the data. To cluster the data into non-performing and performing loan data, we use the first value (K = 2), while the second value (K = 4) clustered the risk level (high risk, medium risk, low risk, and no risk).
2. Determine the Initial centroid
   In determining the initial centroid for the K = 2 clusterings, we select two random data with the non-performing and performing loan labels as the initial centroid. We choose four random data with the C1 and C2 labels as the initial centroid for the K = 4 clustering.
3. Calculate the distance between the data and the centroid
   In this step, we use equation (1) to calculate the distance between the data and each centroid to determine which data enter which cluster.

$$D_{k(i,j)} = \sqrt{\left(X_{1i} - X_{1j}\right)^2 + \ldots + \left(X_{ki} - X_{kj}\right)^2} \qquad (1)$$

4. Group the data based on the closest distance
   From the distance calculation, group each data based on the lowest distance values. The K = 2 clustering will group the data into two clusters (C1 and C2), while the K = 4 will group the data into C1, C2, C3, and C4.
5. Update centroid value
   Calculate a new centroid for each cluster, using equation (2) according to the number of cluster members.

$$y_{new} = \sqrt{\frac{\sum_{i=1}^{n} x_i}{n}} \qquad (2)$$

6. Repeat the process
   Repeat process numbers 3 to 5 until the members of each cluster have not changed.

We first cluster the dataset using the steps above to label each data into NPL (non-performing loan) and PL (performing loan). After evaluating the silhouette coefficient value for each label, we expand the cluster into four risk-level categories (high risk, medium risk, low risk, and no risk) using the K = 4 value. The high and medium risk labels are the results of expanding the NPL label, while the low and no risk are the results of the PL label.

C. K-Nearest Neighbors

K-nearest neighbors is an algorithm based on a non-parametric method, often used as a benchmark in a classification problem because of its overall good performance [23]. This algorithm uses a K value to measure the number of instances with similar values, with a distance metric method as the measuring tool [24].

Some studies have shown the performance of this algorithm in solving classification problems, such as the classification of vocational school's major, with the performance values: accuracy = 84%, precision = 81%, and recall = 84% [25]; classifying the electrical subsidies' recipients for the household, with an accuracy value of 98.07% [26]; and the the Iris flower classification using the K value of 5, with an accuracy value of 96.677% [27]. These studies are the reference we use in this research, both in the classification steps and choosing the K value.

In the classification process, we also use three distance metrics methods, such as the Euclidean, Manhattan, and Chebyshev distance, with the formulas shown in equations (3) to (5) below [28].

$$Euclidean\ (x,y) = \sqrt{\sum_{j=1}^{N} |x-y|^2} \qquad (3)$$

$$Manhattan\ (x,y) = \sum_{j=1}^{N} |x-y| \qquad (4)$$

$$Chebyshev\ (x,y) = log_{\lambda-\infty} \sqrt{\sum_{j=1}^{N} |x-y|^\lambda} \qquad (5)$$

D. Cross-validation Evaluation

Cross-validation is an evaluation method for machine learning that separates the data into K partitions (fold). This method uses an error matrix to evaluate the classification performance, with various indications such as accuracy, precision, and recall [29]. The formulas in equations (6) to (8) show the calculation to evaluate a model's performance using the accuracy, precision, and recall values [30].

$$Accuracy = \frac{True\ Positive + True\ Negative}{Predicted + Actual} \qquad (6)$$

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \qquad (7)$$

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \qquad (8)$$

In this research, we use 10-fold cross-validation to evaluate the models, with the average values from each model as the final performance indicator.

III. Results and Discussions

In the first clustering using the K = 2 value, the result shows that from 60 data, the number of data predicted as C1 (performing loan) totals 23 and 37 for C2 (non-performing loan). From this clustering process, we also obtained an average value of 0.60978787 for the C1 silhouette coefficient and 0.604694649 for the C2. In the beginning, there were 24 PL and 36 NPL data. The first clustering predicted 25 data with PL labels and 24 with NPL labels. This result shows that there is one PL data not clustered correctly. Table IV shows the comparison of the first clustering results compared and the actual data.

TABLE IV
First Clustering Result

Label	Actual	Predicted	Silhouette Coefficient
PL (C1)	36	35	0.596017743
NPL (C2)	24	25	0.58548148

We use the clustering result as the dataset's new label and proceed to the next clustering step using the K = 4 value. In the second clustering using the K = 4 value, the result shows that from 35 data previously labeled as C1, we got two expanded clusters, namely C3 with 22 data and C4 with 13 data. Since these clusters come from the previous performing loan labels, we choose C3 as the low-risk (LR) category and C4 as the no-risk (NR) category. From 25 data previously labeled as C2, we got two expanded clusters, namely C1 with 10 data and C2 with 25 data. Since these clusters come from the previous non-performing loan labels, we choose C1 as the high-risk (HR) category and C2 as the medium-risk (MR) category. Table V shows the summary of the second clustering process.

TABLE V
Second Clustering Result

Label		Predicted	Silhouette Coefficient
First Cluster	Second Cluster		
PL (C1)	Low Risk (C3)	22	0.603739727
	No Risk (C4)	13	0.591849615
NPL (C2)	High Risk (C1)	10	0.584526
	Medium Risk (C2)	25	0.576287

From this second clustering result, we use the cluster as the new label for the dataset, with Table VI showing the sample of this result.

TABLE VI
NEW DATASET SAMPLE

FT	ML	LA	LD	LE	RL
0	1	1	1	1	C1
1	0	2	1	1	C1
0	1	2	0	1	C2
1	0	2	1	1	C1
1	2	1	1	0	C3
1	2	1	0	0	C2
0	0	1	2	0	C4
1	2	0	2	0	C3
1	2	1	2	0	C3
0	0	1	2	0	C4

Notes: RL = Risk Level (C1, C2, C3, C4), C1 = High Risk, C2 = Medium Risk, C3 = Low Risk, C4 = No Risk

We build 12 models with a combination of the distance metrics and the K-NN's K value, resulting in the configuration shown in Table VII.

TABLE VII
CONFIGURATION OF MODELS

Model	Distance Metric	K Value
Euclidean3	Euclidean	3
Euclidean5	Euclidean	5
Euclidean7	Euclidean	7
Euclidean9	Euclidean	9
Manhattan3	Manhattan	3
Manhattan5	Manhattan	5
Manhattan7	Manhattan	7
Manhattan9	Manhattan	9
Chebyshev3	Chebyshev	3
Chebyshev5	Chebyshev	5
Chebyshev7	Chebyshev	7
Chebyshev9	Chebyshev	9

The Euclidean3, Euclidean5, Euclidean7, and Euclidean9 models were built with the Euclidean distance metric and differed in the K values. The models' builts with the Manhattan distance metric are Manhattan3, Manhattan5, Manhattan7, and Manhattan9, but with differences in the K values. The Chebyshev3, Chebyshev5, Chebyshev7, and Chebyshev9 models were built using the Chebyshev distance metric but differ in K values. We use these models to process the new dataset, resulting in the prediction result in the form of a confusion matrix, as shown in Table VIII to Table X.

TABLE VIII
CONFUSION MATRIX FOR EUCLIDEAN DISTANCE

Model	Actual	Predicted			
		C1	C2	C3	C4
Euclidean3	C1	8	1	0	1
	C2	1	12	2	0
	C3	0	0	20	2
	C4	0	0	2	11
Euclidean5	C1	7	2	0	1
	C2	1	12	2	0
	C3	0	0	20	2
	C4	0	0	2	11
Euclidean7	C1	8	1	0	1
	C2	0	13	2	0
	C3	0	0	20	2
	C4	0	0	1	12
Euclidean9	C1	8	1	0	1
	C2	1	13	2	0
	C3	0	0	21	1
	C4	0	0	1	12

TABLE IX
CONFUSION MATRIX FOR MANHATTAN DISTANCE

Model	Actual	Predicted			
		C1	C2	C3	C4
Manhattan3	C1	8	1	0	1
	C2	1	12	2	0
	C3	0	0	20	2
	C4	0	0	2	11
Manhattan5	C1	8	1	0	1
	C2	1	12	2	0
	C3	0	0	20	2
	C4	0	0	2	11
Manhattan7	C1	8	1	0	1
	C2	0	13	2	0
	C3	0	0	20	2
	C4	0	0	2	11
Manhattan9	C1	8	1	0	1
	C2	0	13	2	0
	C3	0	0	21	1
	C4	0	0	1	12

TABLE X
CONFUSION MATRIX FOR CHEBYSHEV DISTANCE

Model	Actual	Predicted			
		C1	C2	C3	C4
Chebyshev3	C1	4	2	1	3
	C2	0	15	0	0
	C3	0	1	21	0
	C4	0	0	3	10
Chebyshev5	C1	5	1	0	4
	C2	0	12	3	0
	C3	0	1	21	0
	C4	0	0	3	0
Chebyshev7	C1	5	1	1	3
	C2	0	11	4	0
	C3	0	0	22	0
	C4	0	0	3	10
Chebyshev9	C1	5	1	1	3
	C2	0	11	4	0
	C3	0	0	22	0
	C4	0	0	3	10

Using a 10-fold cross-validation and equation (6) to (8), we get the performance value for each model in the form of accuracy, precision, and recall, as shown in Table XI.

TABLE XI
10-FOLD CROSS-VALIDATION EVALUATION

Model	Accuracy (%)	Precision (%)	Recall (%)
Euclidean3	85	85.4711	85
Euclidean5	83.33333	83.59127	83.33333
Euclidean7	88.33333	89.09834	88.33333
Euclidean9	90	90.53571	90
Manhattan3	85	85.4711	85
Manhattan5	85	85.4711	85
Manhattan7	86.66667	87.46032	86.66667
Manhattan9	90	90.53571	90
Chebyshev3	83.33333	84.96667	83.33333
Chebyshev5	80	82.08995	80
Chebyshev7	80	83.13889	80
Chebyshev9	80	83.13889	80

The result from Table XI shows that both Euclidean9 and Manhattan9 model generates the highest accuracy, precision, and recall with 90%, 90.53571%, and 90% values, while the Chebyshev5 model generates the lowest value with 80%, 82.08995%, and 80% respectively. These results show the best model is the model that uses the Euclidean and Manhattan distance with a K value of 9, while the worst model is the model that uses the Chebyshev distance with a K = 5. Next, we analyze the overall performance for each distance metric by taking the average value of their performance in each model, as shown in Table XII.

TABLE XII
OVERALL DISTANCE METRIC PERFORMANCE

Distance Metric	Average Accuracy (%)	Average Precision (%)	Average Recall (%)
Euclidean	86.66667	87.17411	86.66667
Manhattan	86.66667	87.23456	86.66667
Chebyshev	80.83333	83.3336	80.83333

Based on the average distance metric's value shown in Table XII, the Manhattan distance metric yields the best performance, followed by the Euclidean and Chebyshev distance. Both the Euclidean and Manhattan distance yield the same average accuracy and recall, but the Manhattan distance generates a better average precision. Next, we analyze the overall performance for each K value by taking the average value of their performance in each model, as shown in Table XIII.

TABLE XIII
OVERALL K VALUES PERFORMANCE

K Values	Average Accuracy (%)	Average Precision (%)	Average Recall (%)
K = 3	84.44444444	85.30295754	84.44444444
K = 5	82.77777778	83.71743997	82.77777778
K = 7	85	86.56585001	85
K = 9	86.66666667	88.07010582	86.66666667

Analyzing the results from Table XIII, we noticed that K = 9 produces the best average performance, while K = 5 produces the worst average performance. After obtaining the average value of both distance metric and K values, we compared the results with previous studies, as shown in Table XIV.

TABLE XIV
COMPARISON WITH PREVIOUS RESEARCH

	Previous Research	Current Research
Best Model	Distance metric: Euclidean, K value: 9	Distance metric: Euclidean and Manhattan, K value: 9
Worst Model	Distance metric: Chebyshev, K value: 9	Distance metric: Chebyshev, K value: 5
Best Average Distance Metric	Euclidean	Manhattan
Worst Average Distance Metric	Chebyshev	Chebyshev
Best Average K Value	Accuracy: K = 3, Precision: K = 5 and K = 9, Recall: K = 3	Accuracy, Precision and Recall: K = 9
Worst Average K Value	Accuracy: K = 9, Precision: K = 3, Recall: K = 9	Accuracy, Precision and Recall: K = 5

From the results shown in Tables XIV, we found that expanding the labels using K-means clustering algorithm does not affect the best model outcome. Both previous and current research yield the same best model (using Euclidean distance metric with a K value of 9). From both the previous and current research, we found that the Chebyshev is the worst distance metric to use with the dataset; the only difference is the K value used.

However, expanding the dataset's labels using K-means clustering affects the average distance metric's performance. Table XIV shows that without the dataset's label expansion, the best average distance metric is the Euclidean distance, while the label's expansion results in the best average distance metric's performance for the Manhattan distance. There is no effect whatsoever with the worst average distance metric's performance, with or without the dataset's label expansion.

The thing that is most affected by the label expansion dataset is the average performance of the K value. Comparing the two studies, we found that the value of K = 9 produces the best average accuracy and recall in the current study but is inversely proportional to the previous. Similarly, on the average precision, the previous research shows that the value of K = 5 produces the best average performance, but the opposite occurs in this research.

IV. CONCLUSIONS

The K-means clustering algorithm is suitable for expanding a dataset labels. We can extend the dataset label by adding the number of clusters using this algorithm, increasing the categories in the dataset label. This research proves that it's feasible to extend the dataset, which initially only contains non-performing and performing loan categories, into risk level (high risk, medium risk, low risk, and no risk) forms. From the comparison with previous research, expanding the dataset labels using the K-means clustering algorithm does not affect the best model result but significantly affects the performance of the average K value. The performance results in this study show the highest accuracy, precision, and recall values at 90%, 90.53571%, and 90% from models that use the Euclidean distance metric and Manhattan distance, with the value of K = 9. The performance results in this study also show the highest average performance for the distance metric yields from the Manhattan distance, with accuracy, precision, and recall values at 86.66667%, 87.23456%, and 86.66667%. The average performance for the K value came from K = 9, with accuracy, precision, and recall values at 86.66666667%, 88.07010582%, and 86.66666667%, respectively.

REFERENCES

[1] N. Yoshino and F. Taghizadeh-Hesary, “A comprehensive method for credit risk assessment of small and medium-sized enterprises based on Asian data,” in *Unlocking SME Finance in Asia*, no. 907, First Edition. | New York : Routledge, 2019. | Series: Routledge studies in development economics: Routledge, 2019, pp. 55–71.

[2] R. M. Dai, S. Suryanto, and S. Novianti, “Analysis of Cooperative Lending Procedure,” *J. Ilmu Keuang. dan Perbank.*, vol. 7, no. 1, pp. 59–70, Jul. 2019, doi: 10.34010/jika.v7i1.1907.

[3] S. A. Rizky, R. Yesputra, and S. Santoso, “Prediction of Smooth Payment of Installments of Prospective Debtors with the K-Nearest Neighbor Method,” *JURTEKSI (Jurnal Teknol. dan Sist. Informasi)*, vol. 7, no. 2, pp. 195–202, Apr. 2021, doi: 10.33330/jurteksi.v7i2.1078.

[4] R. Nofitri and N. Irawati, “Integration of Neive Bayes Method and Rapidminer Software in the Analysis of Trading Company Business Results,” *JURTEKSI (Jurnal Teknol. dan Sist. Informasi)*, vol. 6, no. 1, pp. 35–42, Dec. 2019, doi: 10.33330/jurteksi.v6i1.393.

[5] Y. L. Nainel, E. Buulolo, and I. Lubis, “Application of Data Mining for Drug Sales Estimation Based on the Influence of Brand Image with Expectation Maximization Algorithm (Case Study: PT. Pyridam Farma Tbk),” *JURIKOM (Jurnal Ris. Komputer)*, vol. 7, no. 2, p. 214, Apr. 2020, doi: 10.30865/jurikom.v7i2.2097.

[6] Y. C. Widiyono and S. M. Isa, “Utilization of Data Mining to Predict Non-Performing Loan,” *Adv. Sci. Technol. Eng. Syst. J.*, vol. 5, no. 4, pp. 252–256, 2020, doi: 10.25046/aj050431.

[7] S. I. Serengil, S. Imece, U. G. Tosun, E. B. Buyukbas, and B. Koroglu, “A Comparative Study of Machine Learning Approaches for Non Performing Loan Prediction with Explainability,” *Int. J. Mach. Learn. Comput.*, vol. 12, no. 5, pp. 208–214, 2022, doi: 10.18178/ijmlc.2022.12.5.1102.

[8] A. Bellotti, D. Brigo, P. Gambetti, and F. Vrins, “Forecasting recovery rates on non-performing loans with machine learning,” *Int. J. Forecast.*, vol. 37, no. 1, pp. 428–444, 2021, doi: 10.1016/j.ijforecast.2020.06.009.

[9] M. Iqbal, “Clustering Umrah Pilgrims Data at Auliya Tour Travel Using the K-Means Clustering Method,” *JURTEKSI (Jurnal Teknol. dan Sist. Informasi)*, vol. 5, no. 2, pp. 97–104, Jun. 2019, doi: 10.33330/jurteksi.v5i2.352.

[10] H. Abijono, P. Santoso, and N. L. Anggreini, “Supervised Learning and Unsupervised Learning Algorithms in Data Processing,” *J. Teknol. Terap. G-Tech*, vol. 4, no. 2, pp. 315–318, Apr. 2021, doi: 10.33379/gtech.v4i2.635.

[11] F. Maturo and R. Verde, “Combining unsupervised and supervised learning techniques for enhancing the performance of functional data classifiers,” *Comput. Stat.*, no. 0123456789, Jul. 2022, doi: 10.1007/s00180-022-01259-8.

[12] S. Sugriyono and M. U. Siregar, “Preprocessing kNN algorithm classification using K-means and distance matrix with students’ academic performance dataset,” *J. Teknol. dan Sist. Komput.*, vol. 8, no. 4, 2020, doi: 10.14710/jtsiskom.2020.13874.

[13] R. Ullah, A. H. Khan, and S. M. Emaduddin, “ck-NN: A Clustered k-Nearest Neighbours Approach for Large-Scale Classification,” *ADCAIJ Adv. Distrib. Comput. Artif. Intell. J.*, vol. 8, no. 3, pp. 67–77, 2019, doi: 10.14201/adcaij2019836777.

[14] R. A. Arnomo, W. L. Y. Saptomo, and P. Harsadi, “Implementation of K-Nearest Neighbor Algorithm for Water Quality Identification (Case Study: PDAM Kota Surakarta),” *J. Teknol. Inf. dan Komun.*, vol. 6, no. 1, pp. 1–5, Apr. 2018, doi: 10.30646/tikomsin.v6i1.345.

[15] H. A. Abu Alfeilat *et al.*, “Effects of Distance Measure Choice on K-Nearest Neighbor Classifier Performance: A Review,” *Big Data*, vol. 7, no. 4, pp. 221–248, Dec. 2019, doi: 10.1089/big.2018.0175.

[16] K. F. Margolang, M. M. Siregar, S. Riyadi, and Z. Situmorang, “Distance Metric Analysis of K-Nearest Neighbor Algorithm on Bad Debt Classification,” *J. Inf. Syst. Res.*, vol. 3, no. 2, pp. 118–124, Feb. 2022, doi: 10.47065/josh.v3i2.1262.

[17] M. Rambe, M. Safii, and I. Irawan, “Application Of K-Means Clustering Algorithm On Population Growth In Simalungun Regency,” *Int. J. Basic Appl. Sci.*, vol. 10, no. 2, pp. 61–69, 2021, doi: 10.35335/ijobas.v10i2.55.

[18] D. M. Saputra, D. Saputra, and L. D. Oswari, “Effect of Distance Metrics in Determining K-Value in K-Means Clustering Using Elbow and Silhouette Method,” in *Proceedings of the Sriwijaya International Conference on Information Technology and Its Applications (SICONIAN 2019)*, 2020, vol. 172, no. Siconian 2019, pp. 341–346, doi: 10.2991/aisr.k.200424.051.

[19] M. N. V. Waworuntu and M. Faisal Amin, “Application of K-Means Method for Mapping Jamkesda Recipient Candidates,” *KLIK - Kumpul. J. ILMU Komput.*, vol. 5, no. 2, p. 190, Sep. 2018, doi: 10.20527/klik.v5i2.157.

[20] R. Sovia, E. P. W. Mandala, and S. Mardhiah, “K-Means Algorithm in Selection of Outstanding Students and SAW Method for Prediction of Outstanding Scholarship Recipients,” *J. Edukasi dan Penelit. Inform.*, vol. 6, no. 2, p. 181, Aug. 2020, doi: 10.26418/jp.v6i2.37759.

[21] K. Fahriya and W. Yustanti, “Optimizing the Number of Clusters on Single Tuition Scholarship on Students’ Socio-Economic Data,” *J. Emerg. Inf. Syst. Bus. Intell.*, vol. 02, no. 02, pp. 73–77, 2021, [Online]. Available: https://ejournal.unesa.ac.id/index.php/JEISBI/article/view/39705.

[22] M. Gunawan, M. Zarlis, and R. Roslina, “Comparative Analysis of Naïve Bayes and K-Nearest Neighbor Algorithms to Predict Student Graduation on Time,” *J. MEDIA Inform. BUDIDARMA*, vol. 5, no. 2, p. 513, Apr. 2021, doi: 10.30865/mib.v5i2.2925.

[23] R. M. F. Lubis, Z. Situmorang, and R. Rosnelly, “Analisis Variation K-Fold Cross Validation On Classification Data Method K-Nearest Neighbor,” *J. Ipteks Terap.*, vol. 15, no. March, pp. 68–73, 2020, [Online]. Available: http://publikasi.lldikti10.id/index.php/jit/article/view/98.

[24] J. P. Pinto, S. Kelur, and J. Shetty, “Iris Flower Species Identification Using Machine Learning Approach,” *2018 4th Int. Conf. Converg. Technol. I2CT 2018*, pp. 1–4, 2018, doi: 10.1109/I2CT42659.2018.9057891.

[25] N. A. Sinaga, R. Ramadani, K. Dalimunthe, M. S. A. A. Lubis, and R. Rosnelly, “Comparison of Decision Tree, KNN, and SVM Methods for Determining Majors in Vocational Schools,” *J. Sist. Komput. dan Inform.*, vol. 3, no. 2, p. 94, Dec. 2021, doi: 10.30865/json.v3i2.3598.

[26] Y. M. Hutahaean and A. W. Wijayanto, “Classification of Households Receiving Electricity Subsidies in Gorontalo Province in 2019 with K-Nearest Neighbor and Support Vector Machine Methods,” *J. Sist. dan Teknol. Inf.*, vol. 10, no. 1, pp. 63–68, 2022, [Online]. Available: https://jurnal.untan.ac.id/index.php/justin/article/view/51210.

[27] T. S. Rao, M. Hema, K. S. Priya, K. V. Krishna, and M. S. Ali, “Iris Flower Classification Using Machine Learning,” *Int. J. All Res. Educ. Sci. Methods*, vol. 9, no. 6, pp. 2082–2090, 2022, [Online]. Available: http://www.ijaresm.com/iris-flower-classification-using-machine-learning.

[28] I. Iswanto, T. Tulus, and P. Sihombing, “Comparison of Distance Models on K-Nearest Neighbor Algorithm in Stroke Disease Detection,” *Appl. Technol. Comput. Sci. J.*, vol. 4, no. 1, pp. 63–68, 2021, doi: 10.33086/atcsj.v4i1.2097.

[29] A. Peryanto, A. Yudhana, and R. Umar, “Image Classification Using Convolutional Neural Network and K Fold Cross Validation,” *J. Appl. Informatics Comput.*, vol. 4, no. 1, pp. 45–51, May 2020, doi: 10.30871/jaic.v4i1.2017.

[30] I. Firmansyah, J. T. Samudra, D. Pardede, and Z. Situmorang, “Comparison Of Random Forest And Logistic Regression In The Classification Of Covid-19 Sufferers Based On Symptoms,” *J. Sci. Soc. Res.*, vol. 5, no. 3, p. 595, Oct. 2022, doi: 10.54314/jssr.v5i3.994.