IMPROVEMENT OF PREDICTION MODEL USING K-NEAREST NEIGHBORS (KNN) AND K-MEANS IN MEDICAL DATA

Wiji Lestari[1]*, Sri Sumarlinda[2], Azizah Binti Rahmat
Universitas Duta Bangsa Surakarta[1], Universitas Duta Bangsa Surakarta[2], Universiti Kuala Lumpur[3]
*Correspondence Email : wiji_lestari@udb.ac.id

ABSTRACT

Improving the performance of a prediction model is very important in its implementation. This study aims to improve the performance of the K-Nearest Neighbors (KNN) classification model with the K-Means clustering algorithm. The dataset used is UCI global data with 300 data and 12 features. The dataset is divided into 200 training data and 100 testing data. The training data is then processed by clustering with K-Means. The cluster centroid from the clustering results will be calculated for its distance from the testing data and produce data classification. The results of the classification process show that the accuracy of the proposed model is 76.45% better when compared to the results of the KNN classification process, for k = 5 the accuracy is 63.37%, k = 10 the accuracy is 64.36% and k = 15 the accuracy is also 64.36%.

KEYWORDS

K-Nearest Neighbors (KNN), K-Means, classification model, clustering, medical dataset

INTRODUCTION

The development of machine learning in terms of performance and application is very fast. Machine learning performance improves with the emergence of new algorithms, improvements from old algorithms and hybrid models or combinations of various algorithms. Data processing is not only structured data but also unstructured data (Chen et al, 2020), (Arowolo et al, 2021). This development requires improvements in machine learning for data handling. In its implementation, of course, it must be adjusted to the character of the data, more complex machine learning, for example deep learning may not be needed for simple data characters. The development of data forms is also rapid, not only text and numbers but also images, sound, video and so on. Data processing with machine learning should be adjusted to the characteristics of the data. In some applications, data even needs to be processed first. Improving model performance is related to improvements or modifications to algorithms and data processing (Arowolo et al, 2021), (Garouani et al., 2023).

The K-Nearest Neighbors (KNN) algorithm is a fairly simple algorithm or often called lazy learning. In its processing, the KNN algorithm only performs the data storage process without a learning process from the data without any model formation process. The stored training data will be used to make decisions in the computational process based on neighbor data. The KNN algorithm is very easy and simple to implement, just by setting one parameter k. The classification model that uses the KNN algorithm, the class

determination decision process can be traced easily. Another advantage of the KNN algorithm is that it works locally, only taking into account a number of k data. The KNN algorithm is suitable for data sets that are grouped locally. The weaknesses of the KNN algorithm include being very sensitive to noise and data outliers. In addition, setting an even number of k can also be a weakness, so it is usually chosen odd. The k parameter functions to regulate the level of generalization of the test data based on its distance from the training data. If k is small, the data generalization is high and vice versa. Improving the performance of the KNN algorithm can be done using various methods such as data processing, improving the computing process, adding algorithms and others (Zhang et.al, 2019), (Musuvathi et al., 2024), (Lestari & Sumarlinda, 2022).

Data collection and processing are important in developing a classification model. These two things greatly affect model performance. Problems related to data collection are usually noise, missing data, outliers and so on. The dataset that will be used in the classification process must have value, validity and veracity. The stages after data collection are usually data preprocessing before data processing (Jadhav & Pellakuri, 2021). Data processing before being processed into the classification model is related to the division of training data and test data (Chen et al., 2020). The division of the data set into test data and training data can use the split validation and m-fold cross validation methods. The m-fold cross validation technique is a development of the split validation method. In the m-fold cross validation method, the dataset used is divided into several parts (folds), for example 3, 5, 10 and so on. In m-fold cross validation, the process can be more comprehensive because the training data and test data vary and alternate in each part (Rizki et al., 2024), (Duan, 2024).

The application of the KNN algorithm as a classification model is very broad, both as a support, main tool and modification of the algorithm. KNN is used to detect unknown radio transmitters based on virtual reference points and RSSD information (Zhang et al., 2019). Development of an energy detection technique model in cognitive radio networks using the KNN algorithm (Musuvathi et al., 2024). The KNN classifier is used for hybrid use to optimize dimensionality reduction techniques for malaria disease data (Arowolo et al., 2021). Combining KNN with K-Means clustering is one effective approach to improving KNN performance, especially when working with large and complex datasets. By utilizing K-Means to reduce the size of the dataset, KNN can run more efficiently without sacrificing too much accuracy (Patel et al, 2024). However, it is important to conduct careful testing and validation to ensure that the results obtained are indeed better than the traditional KNN method. K-Means clustering is an unsupervised learning algorithm that partitions data into clusters based on feature similarity. By applying K-Means clustering before KNN, we can simplify the dataset, reduce noise, and enhance the efficiency of the classification process (Sumarlinda et al., 2022).

This study aims to improve the performance of the K-Nearest Neighbor (KNN) classification model using K-Means clustering. KNN as a lazy method has weaknesses in the classification process. One of the main weaknesses of KNN is that this algorithm is very sensitive to data size, so its performance tends to decrease on large or high-dimensional datasets. To overcome this problem, one technique that can be used is to combine KNN with K-Means clustering. K-Means is an unsupervised learning algorithm used to divide data into several groups (clusters) based on similar characteristics. By using K-Means, we can reduce the size of the dataset that must be processed by KNN, thereby increasing efficiency and overall performance. In this study, the dataset used was medical data related to cardiovascular.

RESEARCH METHOD

The study used a research and development method approach. This study improved the performance of the KNN algorithm with K-Means clustering. The dataset used is UCI global data, with 12 features, namely age, sex, cp, trestbps, chol, fbs, restecg, thalach, exang, oldpeak, slope and cardio. Research stages as in figure 1 below.

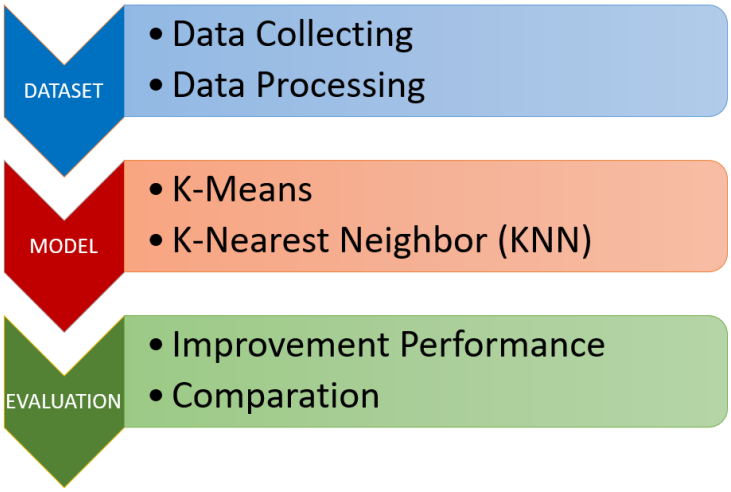

Figure 1. Research Stages

Dataset

The dataset used in this study is UCI global data consisting of 12 features. The dataset consists of 300 data. Before being used, the dataset is processed first, such as removing noise and missing data. Furthermore, the dataset is separated into 2 parts, namely training data and testing data. The training data is 200 and the testing data is 100. Of the 12 features, 1 feature (cardio) as a label and the other features as predictor variables.

Tabel 1. Dataset

id	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	cardio
1	63	1	1	145	233	1	2	150	0	2.3	3	0
2	67	1	4	160	286	0	2	108	1	1.5	2	1
3	67	1	4	120	229	0	2	129	1	2.6	2	1
4	37	1	3	130	250	0	0	187	0	3.5	3	0
5	41	0	2	130	204	0	2	172	0	1.4	1	0
6	56	1	2	120	236	0	0	178	0	0.8	1	0
7	62	0	4	140	268	0	2	160	0	3.6	3	1
8	57	0	4	120	354	0	0	163	1	0.6	1	0
9	63	1	4	130	254	0	2	147	0	1.4	2	1
10	53	1	4	140	203	1	2	155	1	3.1	3	1

Model Development

This classification model improves the performance of KNN with the K-Means clustering algorithm. The detailed model development is as shown in Figure 2 below.

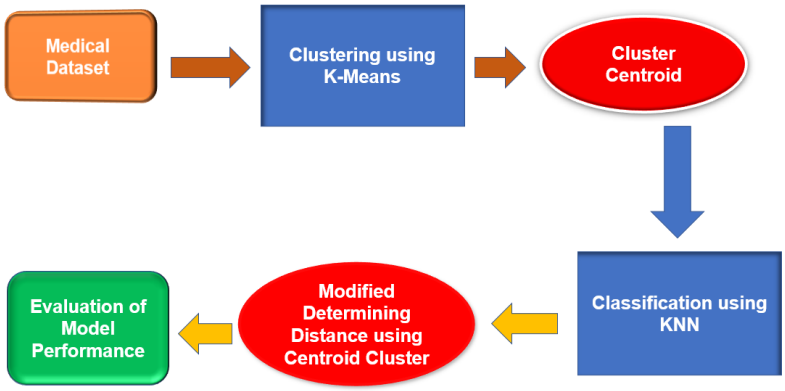

Figure 2. Model Development

Model development begins with input layer with medical dataset, which has been separated into training data and testing data. Furthermore, training data is clustered with K-Means. Clustering results produce cluster centroids. The KNN algorithm uses cluster centroids to determine the distance for determining classification. The classification results of the model will be compared with the KNN classification results with values of k=5, 10 and 15.

K-Means Algorithm

The K-Means clustering algorithm is a clustering method that uses unsupervised learning. The data will be grouped based on their proximity, through the Euclidean distance. K-Means algorithm equations such as formulas (1) and (2) below.

$$d = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^j - c_j \right\|^2 \quad (1)$$

$$c_j^{new} = \frac{1}{n_j} \sum_{i=1}^{n_j} x_{ij} \quad (2)$$

Where:

d = cumulative distance
x = data point
c = cluster centroid
c_j^{new} = new cluster centroid in cluster j
nj = numbers of data in cluster j

K-Nearest Neighbors (KNN)

K-NN algorithm is a kind of supervised learning method for prediction model. The K-Nearest Neighbor (KNN) algorithm is an algorithm with supervised learning and is widely used for prediction and classification. The advantages of the KNN algorithm are high accuracy, intensive on outliers and no assumptions about the data [12], [14], [16-18]. Determining the value of K becomes important. Similarity data with labels used Euclidian distance with the formula [12], [15-16]:

$$d(x, y) = \sqrt{\sum_i^n (x_i - y_i)^2} \quad (3)$$

Where:

d (x, y): distance of data x and y
xi is training data ith
yi is testing data ith

Evaluation

The evaluation of the implementation of this model is by comparing it with the results of the KNN method with k = 5, 10 and 15. The quantity being compared is accuracy. The formula of accuracy is:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (4)$$

Where: TP = true positive; TN = true negative; FP = false positive; FN = false negative.

RESULT AND DISCUSSION

The implementation of the model begins by using the KNN algorithm with variations of k, namely 5, 10 and 15. The dataset used is UCI global data and is separated into training data and testing data. The training data is 200 and the testing data is 100. The implementation uses Rapidminer, as shown in the following figure 3.

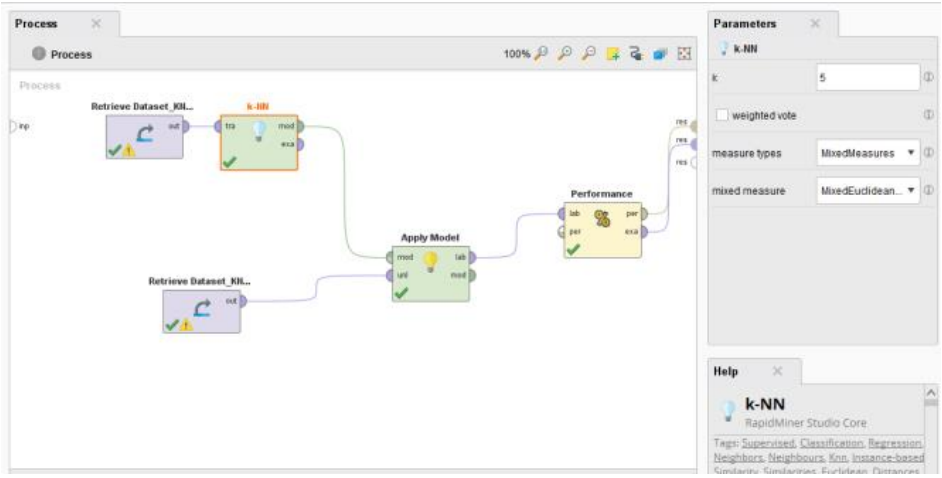

Figure 3. Implementation KNN

The results of implementing dataset classification using KNN are as in table 2 below.

Table 2. Result of KNN Implementation

Model	Nilai k	Accurracy
1	5	63.37%
2	10	64.36%
3	15	64.36%

The medical dataset used for classification contains 12 patient attributes. The dataset is potentially subject to noise and class imbalance. KNN may take longer to classify new patients and may be sensitive to outliers in the data. With K-Means: After clustering the data (e.g., into 10 clusters), KNN can be run more efficiently, focusing only on the most relevant clusters. This approach can result in faster diagnosis while maintaining or improving classification accuracy. The next stage of model development is training data clustering with K-Means. The clustering process is selected k = 2, adjusted to the number of labels in the classification process. Implementation of K-means clustering with Raoidminer as in Figure 4 below.

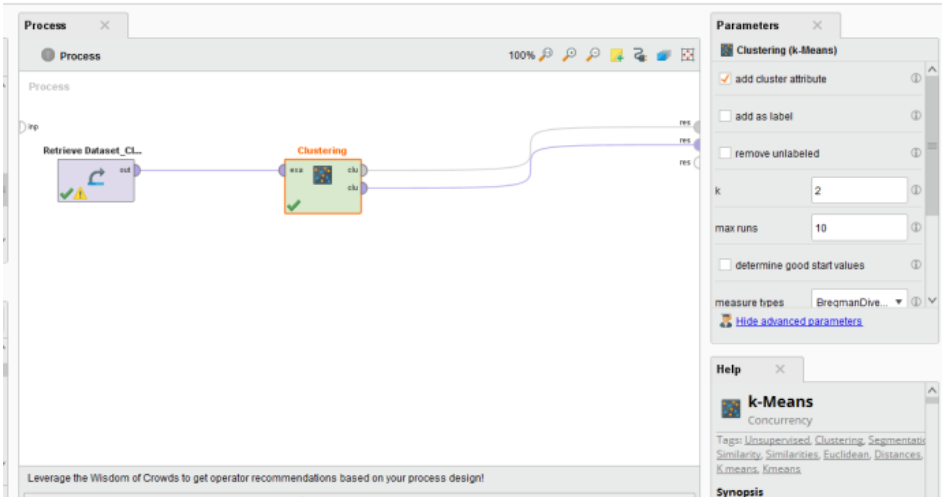

Figure 4. Clustering using K-Means

The clustering results training without label (cardio) and parameter sex (because binominal) data was clustered to produce two clusters. Then the cluster centroid value will be used to calculate the distance on the testing data. The cluster centroid value is as in table 3 below.

Tabel 3. Cluster Centroid

Cluster	Age	Cp	Trestbps	Chol	Fbs	Restecg	Thalach	Exang	Oldpeak	Slope
1	53.068	3.135	128.624	224.481	0.165	1.023	152.045	0.308	1.083	1.632
2	58.045	3.209	139.463	309.045	0.179	1.343	147.716	0.373	1.179	1.612

Next, the distance per data is calculated with the cluster centroid and compared to the closest distance with cluster 1 class 0 and cluster 2 class 1 for the cardio label. The distance calculation is as in the formula below:

$$d = \sqrt{\sum_{i=1}^{10}(x_i - CC_i)^2} \qquad (4)$$

Where: d = distance; x_i = parameters of dataset; CC_i = cluster centroid

The results of the medical data classification process with the proposed model produced an accuracy of 76.45%. This accuracy result is better when compared to the accuracy of the regular KNN algorithm.

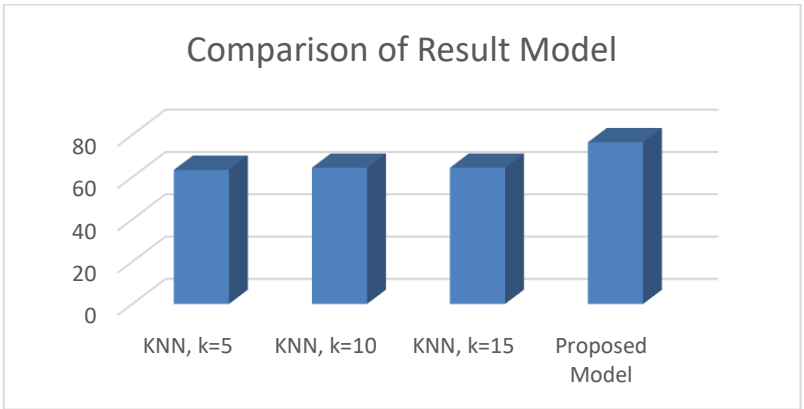

Figure 5. Comparison Model

From Figure 5, it can be seen that the improvement of the KNN algorithm with the K-Means clustering process produces better classification accuracy compared to the regular KNN algorithm.

Combining KNN with K-Means clustering is one effective approach to improve the performance of KNN, especially when working with large and complex datasets. By utilizing K-Means to reduce the size of the dataset, KNN can run more efficiently without sacrificing too much accuracy. However, it is important to conduct careful testing and validation to ensure that the results obtained are indeed better than the traditional KNN method.

CONCLUSION

Improvement of medical dataset classification model using KNN with KNN clustering algorithm can improve model performance. The dataset used is 200 training data and 100 testing data. The training data is clustered to produce cluster centroids used for distance calculation. The model classification results show that the accuracy of the proposed model is better than the KNN results with k values equal to 5, 10 and 15.

REFERENCES

Arowolo, M.O., Adebiyi, M.O., Adebiyi, A.A. and Olugbara, O. (2021). Optimized Hybrid Investigative Based Dimensionality Reduction Methods for Malaria Vector Using KNN Classifier, Journal of Big Data, 8:29, https://doi.org/10.1186/s40537-021-00415-z

Chen, R., Dewi, C., Su-Wen Huang, S. and Caraka, R.E. (2020). Selecting Critical Features for Data Classification Based on Machine Learning Methods, *Journal of Big Data*, 7(52). https://doi.org/10.1186/s40537-020-00327-4

Duan, M. (2024). Innovative Compressive Strength Prediction for Recycled Aggregate/Concrete using K-Nearest Neighbors and Meta-Heuristic Optimization Approaches, *Journal of Engineering and Applied Science*, 71:15, https://doi.org/10.1186/s44147-023-00348-9

Garouani, M., Ahmad, A., Bouneffa, M. and Hamlich, M. (2023). Autoencoder-KNN Meta-Model Based Data Characterization Approach for an Automated Selection of AI Algorithms, *Journal of Big Data*, 10(14), https://doi.org/10.1186/s40537-023-00687-7

Jadhav, A.D and Pellakuri,V. (2021). Highly Accurate and Efficient Two Phase-Intrusion Detection System (TP-IDS) Using Distributed Processing of HADOOP and Machine Learning Techniques, *Journal of Big Data*, 8:131, https://doi.org/10.1186/s40537-021-00521-y

Lestari, L. & Sumarlinda, S. (2022). Implementation of K-Nearest Neighbor (KNN) and Suport Vector Machine (SVM) for Clasification Cardiovascular Disease, *Multiscience* –Vol 2 No 10, January 2022 pp. 30-36.

Musuvathi, A.S.S., Archbald, J.F., Velmurugan, T., Sumathi, D., Devi, S.R. and Preetha, K.S. (2024). Efficient Improvement of Energy Detection Technique in Cognitive Radio Networks Using K-Nearest Neighbour (KNN) Algorithm, *EURASIP Journal on Wireless Communications and Networking*, 2024:10, pp.10-19, https://doi.org/10.1186/s13638-024-02338-8

Patel, P., Balasubramanian, S. and Annavarapu, R.N. (2024). Cross Subject Emotion Identification from Multichannel EEG Sub-Bands using Tsallis Entropy Feature and

KNN Classifier, *Brain Informatics*, 11: 7, https://doi.org/10.1186/s40708-024-00220-3

Rizki, M., Hermawan, A. and Avianto, D. (2024). Optimization of Hyperparameter K in K-Nearest Neighbor Using Particle Swarm Optimization, *JUITA: Jurnal Informatika*, Vol. 12, No. 1.

Sumarlinda, S., Wijiyanto, Lestari, W. (2022). Decision support system for lecturer publication mapping using k-means clustering method, *Journal of Intelligent Decision Support System (IDSS*), Vol. 5, No. 4, December 2022, pp. 140-145.

Zhang, L., Du, T. and Jiang, C. (2019). Detection of an Unknown Radio Transmitter Using an Enhanced K-Nearest Neighbor Algorithm Based on Virtual Reference Point and RSSD Information, *EURASIP Journal on Wireless Communications and Networking.* 2019:71. https://doi.org/10.1186/s13638-019-1383-7