ATCSJ
Applied Technology and Computing Science Journal

Open Access

p-ISSN 2621-4458 | e-ISSN 2621-4474

Available online to journal2.unusa.ac.id

UNUSA

S4-Accredited – SK No.200/M/KPT/2021

Journal Page is available to https://journal2.unusa.ac.id/index.php/ATCSJ

Comparison of Distance Models on K-Nearest Neighbor Algorithm in Stroke Disease Detection

Iswanto[1], Tulus[2], Poltak Sihombing[3]

[1,2,3]Program Studi Teknik Informatika Sekolah Pasca Sarjana Universitas Sumatera Utara, Indonesia
Jalan Universitas No. 9A Kampus USU, Medan
[1*]iswan1512@gmail.com, [2]tulus@usu.ac.id, [3]poltak@usu.ac.id

Article history:

Received 10 May 2021
Revised 23 June 2021
Accepted 30 July 2021
Available online 31 June 2021

Keywords:

K-Nearest Neighbor
Euclidean
Chebyshev
Manhattan
Minkowski

Abstract

Stroke is a cardiovascular (CVD) disease caused by the failure of brain cells to get oxygen supply to pose a risk of ischemic damage and result in death. This Disease can detect based on the similarity of symptoms experienced by the sufferer so that early steps can be taking with appropriate counseling and treatment. Stroke detecting requires a machine learning method. In this research, the author used one of the supervised learning classification methods, namely K-Nearest Neighbor (K-NN). K-NN is a classification method based on calculating the distance to training data. This research compares the Euclidean, Minkowski, Manhattan, Chebyshev distance models to obtain optimal results. The distance models have been tested using the stroke dataset sourced from the Kaggle repository. Based on the test results, the Chebyshev model has the highest levels of accuracy compared to the other three distance models with an average accuracy value of 95.49%, the highest accuracy of 96.03%, at K = 10. The Euclidean and Minkowski distance models have the same level of accuracy at each K value with an average accuracy value of 95.45%, the highest accuracy of 95.93% at K = 10. Meanwhile, Manhattan has the lowest average compared to the other distance models, which is 95.42% but has the highest accuracy of 96.03% at the value of K = 6.

I. INTRODUCTION

Stroke is cardiovascular disease (CVD) caused by the failure of brain cells to get oxygen supply so that it can pose a risk of ischemic damage and result in death. Stroke can occur because of several factors, including blood pressure, history of atrial fibrillation, cholesterol, diabetes, and so on. During this time handling stroke is done manually, here the patient checks on a specialist in neurological diseases. Then carried out the diagnosis in the patient using ask a question in the form of a complaint perceived by the patient as well as the factors that can trigger a stroke. So it will a conclusion is drawn on the level of risk of stroke in patients. The activity of some kind this can cause problems viz requires no cost and time a little.

[1*] Corresponding author

Based on this explanation, then a risk level diagnostic application is needed stroke so that stroke can be overcome immediately according to the level of risk. The level of risk stroke can be divided into 3, namely low stroke risk level, stroke risk level moderate, and high stroke risk [1].

Most deaths globally each year are caused by a stroke when compared to other causes. In 2016, an estimated 17.9 million people died from a stroke, accounting for 31% of all global deaths. Based on these data, 85% of the deaths were due to strokes and heart attacks. Furthermore, in 2015, the deaths under 70 years of age caused by non-communicable diseases occurred as much as 82%, in middle and lower incomes of countries, ware 37% of deaths caused by stroke [2].

Stroke can be prevented with a healthy lifestyle and leave bad habits like alcohol, tobacco, overeating, and lack of exercise. Those who are at high risk of stroke due to other factors such as hyperlipidemia, diabetes, hypertension, or other diseases need early detection to be handled in a counseling manner and with appropriate drugs. It is hoped that the detection of stroke will be able to help patients who have a high risk of having a stroke. So that prevention and early treatment can be done in patients.

Stroke detecting requires a machine learning method. In this research, the author used one of the supervised learning classification methods, namely K-Nearest Neighbor (K-NN).

K-NN is a method of classifying data based on the nearest distance to training data by some number of K values from the nearest neighbors [3]. In determining new data classes, K-NN uses a majority vote system according to the number of nearest neighbors.

The Value of K and distance models can influence the accuracy of the K-NN algorithms [4]. In this research, the author compared the Euclidean, Minkowski, Manhattan, and Chebyshev distance calculation models on K-NN in determining the closest neighbor distance with values of K = 1 to K = 10 using the stroke dataset sourced from the Kaggle Repository. So that shows the most optimal distance models and value of K in detecting stroke diseases.

II. RELATED WORKS

Previous research has used the K-NN algorithm with the Euclidean distance model to diagnose heart disease. The highest accuracy results obtained were 97.4% at a value of K = 7 [5].

Subsequent research discusses the effect of the K value on the classification of the K-NN algorithm. The results are that the test of K values, namely 1, 8, and 15, resulted are different levels of accuracy [3].

Subsequent research does the effect of distance models on K-NN to classify medical data sets. The distance models used are Euclidean, Cosine, Minkowski, and Chi-square. From the research results, using the distance model affects the results of the accuracy of data classification [6].

Subsequent research also conducted comparisons of distance models on K-NN to classify textual data. The distance model used is Euclidean, Chebyshev, Manhattan, and Minkowski [4][7]. The Euclidean and Minkowski distance models have the same accuracy value for each K with the best accuracy for most K values [8]. The Manhattan distance calculation model is slightly better than the euclidean for small K value [9]. Meanwhile, the worst distance model is Chebyshev based on the accuracy of each K value [4].

Classification of stroke dataset has also researched previously using the Naive Bayes algorithm. Based on this research, known that the average accuracy of the data classification of stroke diseases was 89.65% [1].

III. METHODS

The data used in this research is a stroke dataset sourced from the Kaggle Repository. The data has 5,110 records with two classes, namely "0" for stroke patients and "1" who did not suffer a stroke.

The first step for data classification is pre-processing data to replace the blank data with "0" and character data with numeric so that the distance calculation process can do well.

Furthermore, the dataset divides into two parts, namely training data and test data. In this research, the data divide directly to be 80% for training data and the remaining 20% for test data [10].

The data set has been divided into training data and test data and then classified into the K-NN algorithm by determining the K value first, namely 1 to 10. The calculating distance of the test data to the training data using distance models to see the nearest distance so that the new class the test data can be determined based on the K value of the nearest neighbor.

The data classification process uses a distance model, namely Euclidean, Minkowski, Manhattan, Chebyshev use the following equation [6]:

Euclidean Distance, with the equation :

$$D_{euclidean}(x,y) = \|x-y\|_2 = \sqrt{\sum_{j=1}^{N} |x-y|^2} \quad (1)$$

Minkowski Distance, with the equation :

$$D_{minkowski}(x,y) = \|x-y\|_\lambda = \sqrt[\lambda]{\sum_{j=1}^{N} |x-y|^\lambda} \quad (2)$$

Manhattan Distance with the equation ::

$$D_{manhattan}(x,y) = \|x-y\|_1 = \sum_{j=1}^{N} |x-y| \quad (3)$$

Chebyshev Distance with the equation :**:**

$$D_{chebyshev}(x,y) = \|x-y\|_\lambda = lim_{\lambda-\infty} \sqrt[\lambda]{\sum_{j=1}^{N} |x-y|^\lambda} \quad (4)$$

The description of the above equation is:
D = the result of the calculation of the distance of the x and y data.
N = the number of features in the dataset.
λ = the parameter used for the Minkowski distance.

In detail, the stages carried out in this research in classifying the dataset are described in the following steps [10]:
Step 1. Pre-processing the dataset
Step 2. Divide the dataset into training data by 80% and test data by 20%
Step 3. Determine the value of K
Step 4. Calculating the distance between the test data and the training data using the Euclidean distance model (equation 1), Minkowski (equation 2), Manhattan (equation 3), Chebyshev (equation 4)
Step 5. Sort the data based on the closest distance as much as K
Steo 6. Determine the test data class based on the closest training data class

The final stage in this research is to analyze the performance of the K-NN algorithm based on the accuracy of data classification using the four distance calculation models mentioned above. The level of accuracy using the following equation [4]:

$$Accuracy = \frac{Correct\ prediction}{All\ predictable\ data}\ x\ 100\% \quad (5)$$

Based on the result accuracy of the distance calculation model for each K value, it can be known that the most optimal distance calculation model and the best K value for detecting stroke disease.

IV. RESULTS AND DISCUSSIONS

The The result of this research is showed from the test of the dataset using the methods previously discussed. In this research, the testing of the distance models on K-NN uses the python programming language running through Google Colaboratory.

The testing phase begins with processing the dataset by changing the character data to numeric so that the distance calculation process can do well. Furthermore, the stroke datasets are dividing to be 80% training data and 20% test data. The dividing of a dataset is done randomly by the program.

The K value of the nearest neighbor used in the test is K = 1 to K = 10. Based on the predetermined K value, the stroke dataset is classified using the Euclidean, Minkowski, Manhattan, and Chebyshev distance models. The results of each distance model present in the following table:

Table 1. Comparison of Distance Model Accuracy Results

Value of K	Distance Models			
	Euclidean	**Minkowski**	**Manhattan**	**Chebyshev**
1	94,09%	94,09%	93,08%	93,18%
2	95,52%	95,52%	95,62%	95,93%
3	94,81%	94,81%	94,60%	94,81%
4	95,62%	95,62%	95,72%	95,82%
5	95,21%	95,21%	95,82%	95,62%
6	95,82%	95,82%	96,03%	95,93%
7	95,82%	95,82%	95,82%	95,82%
8	95,82%	95,82%	95,82%	95,93%
9	95,82%	95,82%	95,72%	95,82%
10	95,93%	95,93%	95,93%	96,03%
Average	**95,45%**	**95,45%**	**95,42%**	**95,49%**

The test results based on the accuracy comparison table above show that the Euclidean and Minkowski distance calculation models have the same accuracy value for each K value with an average accuracy of 95.45%.

Those were suitable with the results of previous research [4]. And then the Manhattan distance gets the average accuracy obtained is 95.42%, slightly lower than the result Euclidean and Minkowski, only a difference of 0.03%. The highest average accuracy is distance models of the Chebyshev distance model with an average accuracy value of 95.49%. It is different from previous research [4] states that Chebyshev is a distance calculation model with the lowest accuracy. For more details, shown from the graphic image below:

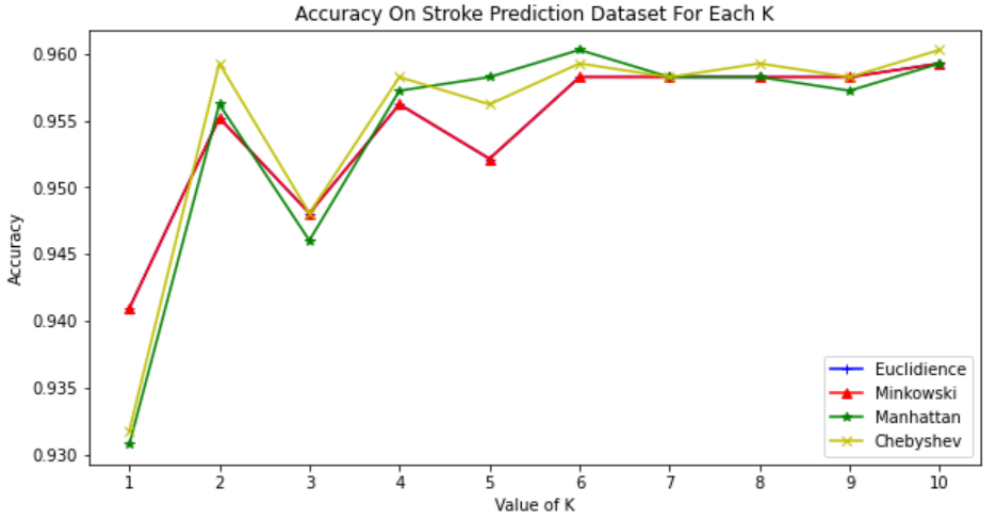

Figure 1. Comparison of Distance Model Accuracy at Each K Value

The accuracy results for each K value in the distance models shown in the following table:

Table 2. Comparison of the Highest Accuracy at K Value

No	Distance	Highest Accuracy	Value of K
1	Euclidean	95,93%	10
2	Minkowski	95,93%	10
3	Manhattan	96,03%	6
4	Chebyshev	96,03%	10

When viewed based on the highest accuracy results at the table above, the use of the Manhattan and Chebyshev distance model has the highest accuracy of 96.03% when K = 6 for Manhattan and K = 10 for Chebyshev. Whereas in the Euclidean and Minkowski distance, the highest accuracy value is 95.93% when the value of K = 10.

V. CONCLUSIONS AND RECOMMENDATIONS

From the results of testing and discussion in this research, the conclusion is that the use of distance models can affect the accuracy of the K-NN algorithm in the detection of stroke diseases. The accuracy value tends to increase as the K value increases.

The Chebyshev distance model has the highest accuracy than the other three distance models which an average accuracy value of 95.49% and the highest accuracy of 96.03% at K = 10. The Euclidean and Minkowski distance models have the same level of accuracy at each K value with an average accuracy value of 95.45% and the highest accuracy of 95.93% at K = 10. Meanwhile, Manhattan has the lowest average accuracy comparing to other distance models, which is 95.42%, but it's the highest accuracy of 96.03% at K = 6.

Thus, the most optimal distance model used for the stroke dataset is the Chebyshev distance model with the optimal K value of 10. Furthermore, the Manhattan distance calculation model with the optimal K value is 6.

VI. REFERENCES

[1] D. W. Nugraha, A. Y. E. Dodu, and N. Chandra, "Klasifikasi Penyakit Stroke Menggunakan Metode Naive Bayes Classifier (Studi Kasus Pada Rumah Sakit Umum Daerah Undata Palu)," *semanTIK*, vol. 3, no. 2, 2017.

[2] V. Gerc, I. Masic, N. Salihefendic, and M. Zildzic, "Cardiovascular Diseases (CVDs) in COVID-19 Pandemic Era," *Mater. Socio Medica*, vol. 32, no. 2, 2020, doi: 10.5455/msm.2020.32.158-164.

[3] I. A. Angreni, S. A. Adisasmita, M. I. Ramli, and S. Hamid, "Pengaruh Nilai K Pada Metode K-Nearest Neighbor (KNN) Terhadap Tingkat Akurasi Identifikasi Kerusakan Jalan," *Rekayasa Sipil Mercu Buana*, vol. 7, no. 2, pp. 63–70, 2018.

[4] W. Wahyono, I. N. P. Trisna, S. L. Sariwening, M. Fajar, and D. Wijayanto, "Perbandingan penghitungan jarak pada k-nearest neighbour dalam klasifikasi data tekstual," *J. Teknol. dan Sist. Komput.*, vol. 8, no. 1, pp. 54–58, 2020.

[5] M. Shouman, T. Turner, and R. Stocker, "Applying k-Nearest Neighbour in Diagnosing Heart Disease Patients," *Int. J. Inf. Educ. Technol.*, 2012, doi: 10.7763/ijiet.2012.v2.114.

[6] L. Y. Hu, M. W. Huang, S. W. Ke, and C. F. Tsai, "The distance function effect on k-nearest neighbor classification for medical datasets," *Springerplus*, vol. 5, no. 1, 2016, doi: 10.1186/s40064-016-2941-7.

[7] T. Kirdat and V. V Patil, "Application of Chebyshev distance and Minkowski distance to CBIR using color histogram," *Int. J. Innov. Res. Technol.*, vol. 2, no. 9, pp. 28–31, 2016.

[8] D. Sinwar and R. Kaushik, "Study of Euclidean and Manhattan distance metrics using simple k-means clustering," *Int. J. Res. Appl. Sci. Eng. Technol*, vol. 2, no. 5, pp. 270–274, 2014.

[9] P. Mulak and N. Talhar, "Analysis of distance measures using k-nearest neighbor algorithm on kdd dataset," *Int. J. Sci. Res.*, vol. 4, no. 7, pp. 2101–2104, 2015.
[10] A. M. Argina, "Penerapan Metode Klasifikasi K-Nearest Neigbor pada Dataset Penderita Penyakit Diabetes," *Indones. J. Data Sci.*, vol. 1, no. 2, pp. 29–33, 2020.