Indonesian Journal of Data and Science

Volume 4 Issue 2
ISSN 2715-9930
https://doi.org/10.56705/ijodas.v4i2.71

Research Article

Comparison of Performance of Four Distance Metric Algorithms in K-Nearest Neighbor Method on Diabetes Patient Data

Dewi Ratnasari [1,*]
[1] Universitas Muslim Indonesia, Makassar, Indonesia, dewii.ratna.sari.sudarsono@gmail.com

Correspondence should be addressed to Dewi Ratnasari; dewii.ratna.sari.sudarsono@gmail.com

Received 10 June 2023; Accepted 18 June 2023; Published 31 July 2023

Abstract:

Diabetes is a chronic disease that occurs when the pancreas no longer produces insulin or when the body cannot effectively use the insulin it produces. The aim of this study is to analyze and compare the classification performance on diabetes patient dataset using four distance metric algorithms in the K-Nearest Neighbor (K-NN) method. Based on previous research, the performance values obtained were not sufficiently high, not exceeding 80%. Therefore, some actions are needed with the hope of obtaining new performance values and making comparisons with previous studies. Based on the test results using the confusion matrix, the accuracy level using Euclidean distance measurement obtained the best performance value at k=17 with 10-k fold, with an accuracy of 85.71%, precision of 86.24%, recall of 85.71%, and F-measure of 85.12%. The Manhattan distance measurement obtained the best performance value at k=25 with 10-k fold, with an accuracy of 85.53%, precision of 85.54%, recall of 85.53%, and F-measure of 85.10%. The Minkowski distance measurement obtained the best performance value at k=17 with 10-k fold, with an accuracy of 85.71%, precision of 86.24%, recall of 85.71%, and F-measure of 85.12%. On the other hand, the Hamming distance measurement obtained the best performance value at k=23 with 10-k fold, with an accuracy of 75.32%, precision of 79.27%, recall of 75.32%, and F-measure of 71.45%.

Keywords: K-Nearest Neighbor, Akurasi, Presisi, Recall, F-measure, Diabetes.
Dataset link:

1. Introduction

Diabetes is a chronic disease that occurs when the pancreas no longer produces insulin or when the body cannot effectively use the insulin it produces [1]. Indonesia is one of the top 10 countries with the highest number of diabetes patients in the world. In 1995, Indonesia, which was still classified as a developing country, ranked 7th with 4.5 million diabetes patients. This ranking is predicted to rise to the 5th position by 2025, with an estimated number of patients reaching 12.4 million. In 2021, the International Diabetes Federation (IDF) reported that Indonesia ranked fifth with 19.47 million diabetes patients among a population of 179.72 million. This means that the prevalence of diabetes in Indonesia is 10.6% [2].

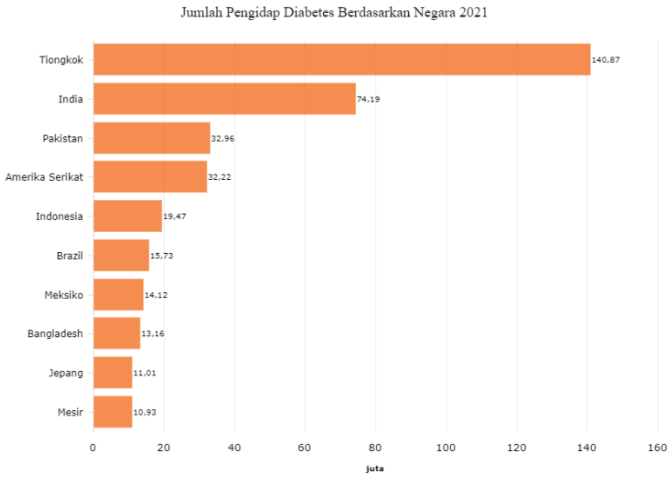

Figure 1. Number of Diabetes Patients by Country in 2021

Machine learning is a part of artificial intelligence. Machine learning is a technique for inferring data with mathematical approaches [3]. Machine learning can improve the performance of search engines, making the accuracy of information searched by users more precise [4]. K-Nearest Neighbor (K-NN) is one of the data mining algorithms widely used in research, particularly in object classification in data. With the principle of finding a group of k from a set of data, K-NN calculates the nearest distance between training data and test data to produce data classification [5].

Various object grouping cases can be solved more easily by applying classification techniques. For example, in the medical field, classification applications can be used to classify the level of disease suffered by a patient, making it easier for doctors to provide appropriate therapy solutions. To solve classification problems, various data mining methods can be applied. Accuracy in object classification is crucial; a good classification method is one that produces minimal errors, among which is the K-NN method.

In previous research, diabetes classification tests were conducted using the K-NN method by comparing two distance metric algorithms. The results obtained from the simulation of a 50:50 dataset ratio yielded the best performance values with an accuracy of 76%, precision of 75%, recall of 95%, and F-measure of 84% at k=45 using the Euclidean distance method, and at k=23 with the Manhattan distance method, an accuracy of 75%, precision of 74%, recall of 95%, and F-measure of 84% [6].

Table 1. Comparison Table of Research

No	Peneliti	Metode	Rasio Dataset	*Data training*	Data Testing	Akurasi	Presisi	Recall	F-Measure
1	Andi Maulida Argina	K-NN K = 3	77	69	8	39%	65%	36%	46%
2	Nur Adhim Rosadi	K-NN Menggunakan *Manhattan* K=5	768	384	385	76%	75%	95%	84%
		K-NN Menggunakan *Euclidean* K=5	768	384	385	75%	74%	95%	84%

Based on **Table 1** previous research, the performance values obtained were not sufficiently good, or not above 80%. Therefore, several actions are needed with the hope of achieving new performance values and making comparisons with previous studies.

In this study, we attempted to continue by adding scenarios, comparing four commonly used distance metric algorithms in the K-Nearest Neighbor (K-NN) algorithm, namely Hamming Distance, Euclidean Distance, Minkowski Distance, and Manhattan Distance, and applying cross-validation to the patient database dataset. The dataset used is the Pima Indians Diabetes Database, published by UCI Machine Learning on the Kaggle website.

Based on the above description, the title chosen for this research is "Comparison of Performance of Four Distance Metric Algorithms in K-Nearest Neighbor (K-NN) Method on Diabetes Patient Data.

2. Method

K-Nearest Neighbor (K-NN) is a method for classifying objects based on the closest distance in the training data to the tested object. K-NN can be used to assign new data (test data) to the group of training data with the closest distance, so this method can be used to classify test sound data according to the appropriate sound data group. K-NN will cluster the results of the calculations with the training data that has the most relatives in the specified range value [15].

K-NN method is widely used because it has several advantages, such as producing more accurate and effective data when the training data is large enough. However, this method also has some disadvantages, such as relatively high computational costs due to the need to calculate the distance of the query instance to the entire training sample [16].

The steps in implementing the K-NN method calculation are as follows:

a. Determine the parameter k (number of nearest neighbors).
b. Calculate the squared Euclidean distance between the data to be evaluated and all training data.
c. Sort the formed distances (in ascending order).
d. Select the k nearest distance alternatives.

The workflow of K-NN is illustrated in **Figure 3** [17].

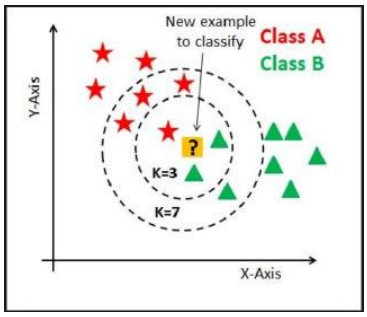

Figure 2. Example of K-NN Workflow

Euclidean Distance

Euclidean distance is a method of calculating the straight-line distance between two different objects. This method can be applied to 1, 2, and 3-dimensional space. The distance calculation in dimensional space can be depicted with the following Equation 1:

$$d = \sqrt{\sum_{i=1}^{n}(x_{1i} - x_{2i})^2 + (y_{1i} - y_{2i})^2 + (z_{1i} - z_{2i})^2 + \ldots} \quad (1)$$

Description:

d : Jarak *Euclidean* (*Euclidean Distance*).

x : sample data

y : test data

i : data variable.

Manhattan Distance

Manhattan distance is used to select suitable cases from the case base by calculating the sum of absolute weights of the differences between the current case and other cases. To calculate the weights, the following Equation 2 is used:

$$d = \sum_{i=1}^{n} |xi - yi| \quad (2)$$

Description:

d : Jarak *Manhattan* (*Manhattan Distance*).

x : sample data

y : test data

i : data variable.

n : data dimension

Minkowski Distance

Minkowski distance is a metric in vector space where a norm is defined (normed vector space) and is considered a generalization of Euclidean distance and Manhattan distance.

$$d = \left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{1/p} \quad (3)$$

Description:

d: distance between x and y.

x: cluster center data.

y: data on the attribute.

i: each data.

n: number of data.

xi: data on the i-th cluster center.

yi: data on the i-th data.

p: power.

Hamming Distance

Hamming Distance is a metric used to compare two binary vectors, primarily binary strings or bitstrings.

$$d = \frac{1}{n} \sum_{i=1}^{n} |x_i - y_i| \quad (4)$$

Description:

d: Hamming Distance.

x: sample data.

y: test data.

i: data variable.

n: data dimension.

Analisa Performa

This stage serves to test the performance of a classification method in terms of accuracy, precision, recall, and f-measure.

1) *Accuracy*
   The accuracy Equation 5 is shown below:

$$Accuracy = \frac{tp + tn}{tp + tn + fp + fn} \quad (5)$$

2) *Precision*
   The precision Equation 6 is shown below:

(6)

$$Precision = \frac{tp}{tp + fp}$$

3) *Recall*
The recall Equation 7 is shown below:

$$Recall = \frac{tp}{tp + fn} \quad (7)$$

4) F-measure
The f-measure Equation 8 is shown below:

$$Fmeasure = 2\frac{Precision * Recall}{Precision + Recall} \quad (8)$$

3. Results and Discussion

In the research stage, several processes are described, which consist of several calculations and the implementation of the K-Nearest Neighbor (K-NN) method.

Table 2. Diabetes Dataset

No.	X1	X2	X3	X4	X5	X6	X7	X8	Y
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1
..	…	…	…	…	…	…	…	…	…
763	10	101	76	48	180	32.9	0.171	63	0
764	2	122	70	27	0	36.8	0.340	27	0
765	5	121	72	23	112	26.2	0.245	30	0
766	1	126	60	0	0	30.1	0.349	47	1
767	1	93	70	31	0	30.4	0.315	23	0

Euclidean Distance

Table 3. Cross Validation of Euclidean with K-Fold 10 and K 17

No.	cv-n	*Accuracy*	*Precision*	*Recall*	*F-measure*
1.	cv1	74.03%	73.37%	74.03%	73.49%
2.	cv2	72.73%	72.01%	72.73%	70.59%
3.	cv3	70.13%	68.89%	70.13%	68.88%
4.	cv4	66.23%	66.23%	66.23%	66.23%
5.	cv5	71.43%	70.31%	74.03%	69.48%
6.	cv6	79.22%	79.18%	79.22%	78.18%
7.	cv7	77.92%	77.56%	77.92%	77.00%
8.	cv8	85.71%	86.24%	85.71%	85.12%
9.	cv9	78.95%	78.64%	78.95%	78.73%
10.	cv10	78.95%	78.95%	78.95%	78.95%

Table 4. Cross Validation of Euclidean with K-Fold 5 and K 13

No.	cv-n	*Accuracy*	*Precision*	*Recall*	*F-measure*
1.	cv1	75.97%	75.37%	75.97%	75.06%
2.	cv2	70.78%	70.23%	70.78%	70.43%

3.	cv3	75.97%	75.54%	75.97%	74.67%
4.	cv4	82.35%	82.37%	82.35%	81.66%
5.	cv5	72.55%	72.11%	72.55%	72.28%

Manhattan Distance

Table 5. Cross Validation of Manhattan with K-Fold 10 and K 25

No.	cv-n	*Accuracy*	*Precision*	*Recall*	*F-measure*
1.	cv1	75.32%	74.68%	75.32%	74.65%
2.	cv2	72.73%	72.46%	72.73%	69.97%
3.	cv3	70.13%	69.13%	70.13%	69.32%
4.	cv4	61.04%	60.38%	61.04%	60.67%
5.	cv5	72.73%	72.01%	72.73%	70.59%
6.	cv6	83.12%	83.35%	83.12%	82.41%
7.	cv7	77.92%	77.45%	77.92%	77.32%
8.	cv8	81.82%	81.55%	81.82%	81.44%
9.	cv9	75.00%	74.45%	75.00%	74.60%
10.	cv10	85.53%	85.54%	85.53%	85.10%

Table 6. Cross Validation of Manhattan with K-Fold 5 and K 23

No.	cv-n	*Accuracy*	*Precision*	*Recall*	*F-measure*
1.	cv1	77.27%	77.04%	77.27%	76.04%
2.	cv2	68.83%	68.12%	68.83%	68.37%
3.	cv3	74.68%	74.04%	74.68%	73.30%
4.	cv4	83.66%	83.63%	83.66%	83.14%
5.	cv5	77.78%	77.26%	77.78%	77.15%

Minkowski Distance

Table 7. Cross Validation of Minkowski with K-Fold 10 and K 17

No.	cv-n	*Accuracy*	*Precision*	*Recall*	*F-measure*
1.	cv1	74.03%	73.37%	74.03%	73.49%
2.	cv2	72.73%	72.01%	72.73%	70.59%
3.	cv3	70.13%	68.89%	70.13%	68.88%
4.	cv4	66.23%	66.23%	66.23%	66.23%
5.	cv5	71.43%	70.31%	71.43%	69.48%
6.	cv6	79.22%	79.18%	79.22%	78.18%
7.	cv7	77.92%	77.56%	77.92%	77.00%
8.	cv8	85.71%	86.24%	85.71%	85.12%
9.	cv9	78.95%	78.64%	78.95%	78.73%
10.	cv10	78.95%	78.95%	78.95%	78.95%

Table 8. Cross Validation of Minkowski with K-Fold 5 and K 13

No.	cv-n	*Accuracy*	*Precision*	*Recall*	*F-measure*
1.	cv1	75.97%	75.37%	75.97%	75.06%
2.	cv2	70.78%	70.23%	70.78%	70.43%
3.	cv3	75.97%	75.54%	75.97%	74.67%
4.	cv4	82.35%	82.37%	82.35%	81.66%
5.	cv5	72.55%	72.11%	72.55%	72.28%

Hamming Distance

Table 9. Cross Validation of Hamming with K-Fold 10 and K 11

No.	cv-n	*Accuracy*	*Precision*	*Recall*	*F-measure*
1.	cv1	61.04%	51.76%	61.04%	52.83%
2.	cv2	64.94%	42.17%	64.94%	51.13%
3.	cv3	67.53%	67.28%	67.53%	60.02%
4.	cv4	67.53%	69.89%	67.53%	58.52%
5.	cv5	75.32%	79.05%	75.32%	71.45%

6.	cv6	59.74%	40.92%	59.74%	48.57%
7.	cv7	63.64%	57.70%	63.64%	55.97%
8.	cv8	66.23%	66.37%	66.23%	55.99%
9.	cv9	69.74%	79.27%	69.74%	60.57%
10.	cv10	65.79%	60.67%	65.79%	54.44%

Table 10. Cross Validation of Hamming with K-Fold 5 and K 3

No.	cv-n	*Accuracy*	*Precision*	*Recall*	*F-measure*
1.	cv1	65.58%	62.53%	65.58%	61.45%
2.	cv2	64.94%	61.15%	64.94%	59.43%
3.	cv3	60.39%	50.93%	60.39%	52.43%
4.	cv4	65.36%	62.91%	65.36%	63.08%
5.	cv5	69.93%	69.55%	69.93%	65.02%

Analisa Performa

Berikut gambar grafik *cross validation* pengujian *k-fold* dengan menggunakan pengukuran persamaan jarak *Euclidean, Manhattan, Minkowski,* dan *Hamming* serta penerapan hasil performa dalam bentuk *boxplot*.

1) Euclidean Distance

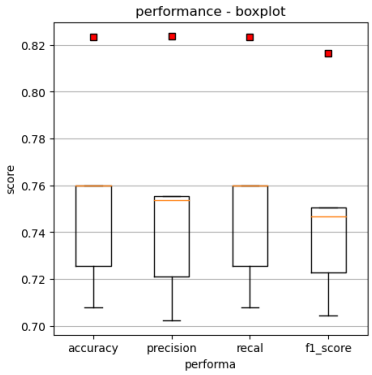

Figure 3. Boxplot of Euclidean Distance K-Fold 5

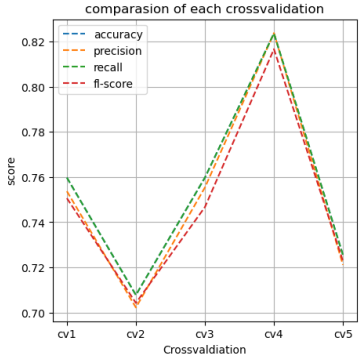

Figure 4. Performance Graph of Euclidean Distance K-Fold 5

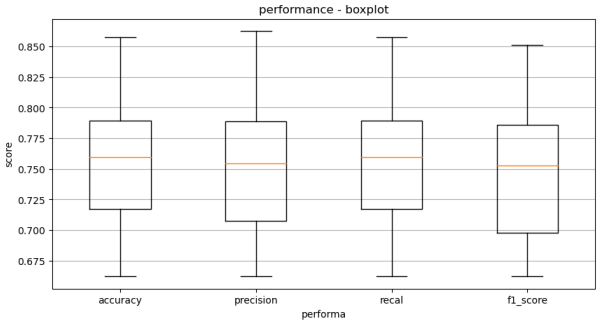

Figure 5. Boxplot of Euclidean Distance K-Fold 10

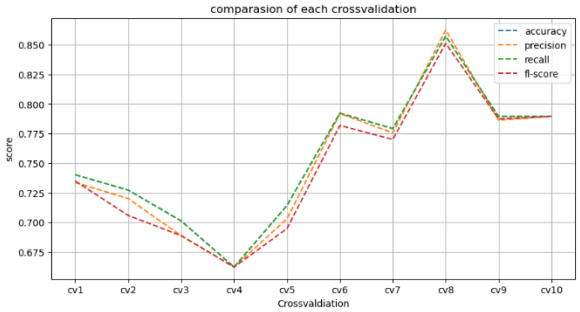

Figure 6. Performance Graph of Euclidean Distance K-Fold 10

2) Manhattan Distance

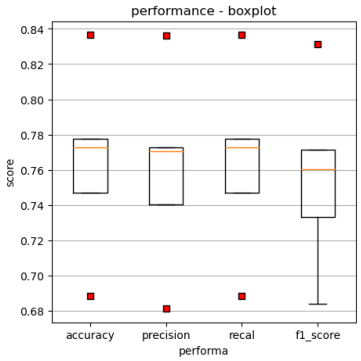

Figure 7. Boxplot of Manhattan Distance K-Fold 5

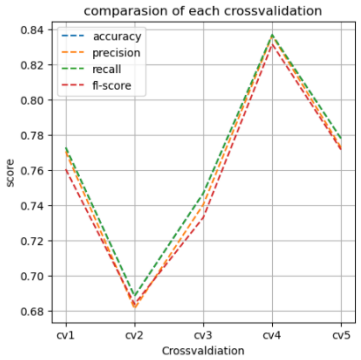

Figure 8. Performance Graph of Manhattan Distance K-Fold 5

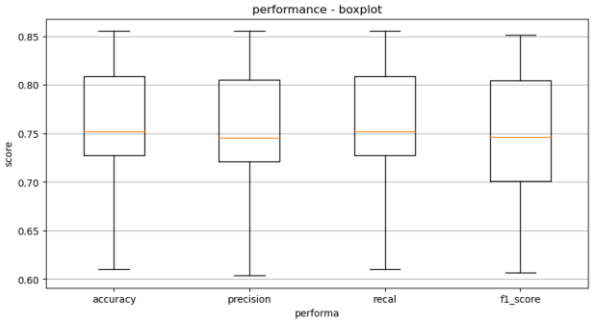

Figure 9. Boxplot of Manhattan Distance K-Fold 10

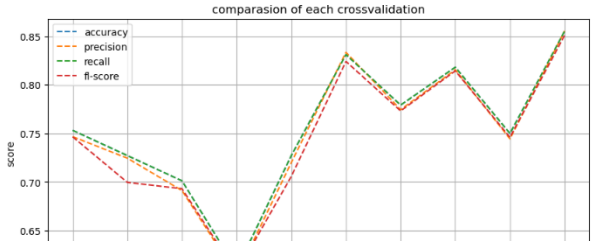

Figure 10. Performance Graph of Manhattan Distance K-Fold 10

3) Minkowski Distance

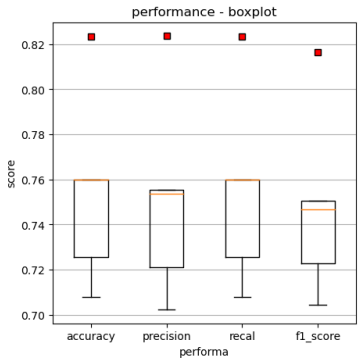

Figure 11. Boxplot of Minkowski Distance K-Fold 5

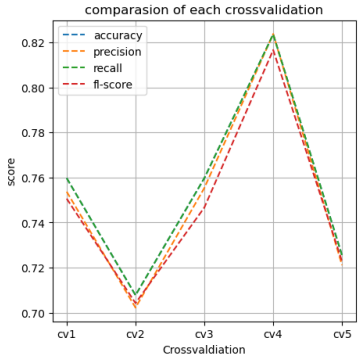

Figure 12. Performance Graph of Minkowski Distance K-Fold 5

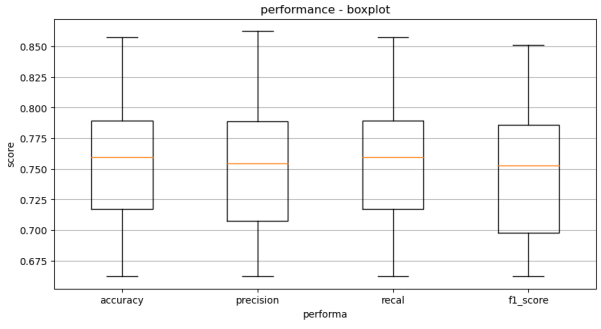

Figure 13. Boxplot of Minkowski Distance K-Fold 10

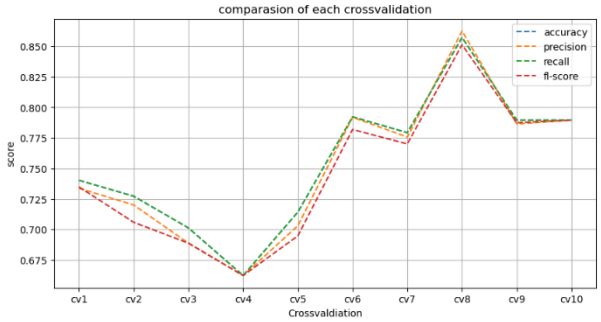

Figure 14. Performance Graph of Minkowski Distance K-Fold 10

4) Hamming Distance

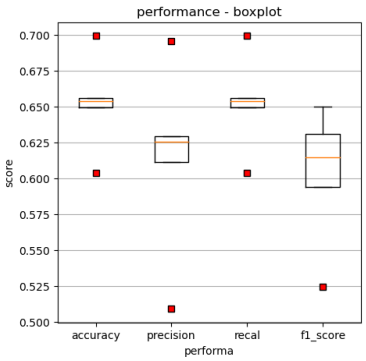

Figure 15. Boxplot of Hamming Distance K-Fold 5

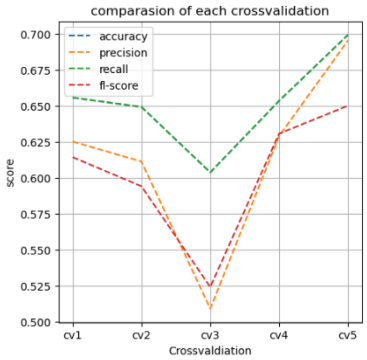

Figure 16. Performance Graph of Hamming Distance K-Fold 5

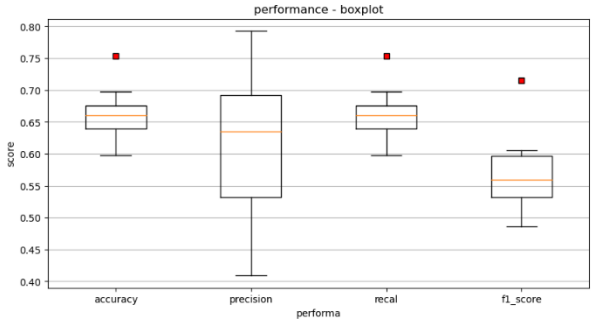

Figure 17. Boxplot of Hamming Distance K-Fold 10

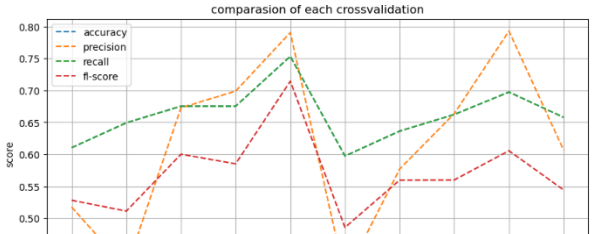

Figure 18. Performance Graph of Hamming Distance K-Fold 10

Decision Making

Table 11. Performance calculation results for each algorithm method

No.	Metode	K-Fold CV	Nilai K	*Accuracy*	*Precision*	*Recall*	*F-measure*
1.	Euclidean Distance	5	13	82.35%	82.37%	82.35%	81.66%
		10	17	85.71%	86.24%	85.71%	85.12%
2.	Manhattan Distance	5	23	83.66%	83.63%	83.66%	83.14%
		10	25	85.53%	85.54%	85.53%	85.10%
3.	Minkowski Distance	5	13	82.35%	82.37%	82.35%	81.66%
		10	17	85.71%	86.24%	85.71%	85.12%
4.	Hamming Distance	5	3	69.93%	69.55%	69.93%	65.02%
		10	11	75.32%	79.27%	75.32%	71.45%

Based on **Table 11**, it can be seen that the performance of accuracy, precision, recall, and f-measure from the Euclidean and Minkowski distance algorithms is the most suitable for use on the diabetes dataset with new performance values at k-fold 10, with k=17, and the accuracy is 85.71%, precision is 86.24%, recall is 85.72%, and f-measure is 85.12%.

4. Conclusion

Based on the testing of performance calculations conducted by the researchers, new performance values are obtained, and this research can be compared with previous research. The comparison of performance values is presented in **Table 12**.

Table 12. Comparison Table of Research

Peneliti	Metode	Nilai K	A	P	R	F
Andi Maulida Argina	*Euclidean*	3	39%	65%	36%	46%
Nur Adhim Rosadi	*Manhattan*	5	76%	75%	95%	84%
	Euclidean	5	75%	74%	95%	84%
Dewi Ratnasari	*Euclidean*	17	85.71%	86.24%	85.71%	85.12%
	Manhattan	25	85.53%	85.54%	85.53%	85.10%
	Minkowski	17	85.71%	86.24%	85.71%	85.12%
	Hamming	11	75.32%	79.27%	75.32%	71.45%

References

[1] G. Mahalisa and N. Arminarahmah, "Diabetes Classification Analysis Using the Euclidean Distance Method Based on the K-Nearest Neighbors Algorithm," *J. Teknol. Komput. dan Sist. Inf.*, vol. 5, no. 3, pp. 178–182, 2022.

[2] K. F. Margolang, M. M. Siregar, S. Riyadi, and Z. Situmorang, "Analisa Distance Metric Algoritma K-Nearest Neighbor Pada Klasifikasi Kredit Macet," *J. Inf. Syst. Res.*, vol. 3, no. 2, pp. 118–124, 2022, doi: 10.47065/josh.v3i2.1262.

[3] J. Putra, *Pengenalan Konsep Pembelajaran Mesin dan Deep Learning Edisi 1.3*. Pengenalan Konsep Pembelajaran Mesin dan Deep Learning Edisi 1.3, 2019.

[4] Y. F. Affif Surya Diantika, "Implementasi Machine Learning Pada Aplikasi Penjualan Produk Digital (Studi Pada Grabkios)," no. 15.

[5] R. R. Rahayu and L. Lidiawati, "Implementasi Algoritma K-Nearest Neighbor Untuk Memprediksi Program Studi Bagi Calon Mahasiswa Baru," *Infotek J. Inform. dan Teknol.*, vol. 4, no. 2, pp. 131–141, 2021, doi: 10.29408/jit.v4i2.3546.

[6] N. Rosadi Adhim, "Analisis Performa Metode K-Nearest Neighbor (K-NN) Dalam Klasifikasi Data Pasien Penyakit Diabetes," 2022.

[7] Bustami, "Penerapan Algoritma Naive Bayes," *J. Inform.*, vol. 8, no. 1, pp. 884–898, 2014.

[8] J. Eska, "Penerapan Data Mining Untuk Prekdiksi Penjualan Wallpaper Menggunakan Algoritma C4.5 STMIK Royal Ksiaran," *JURTEKSI (Jurnal Teknol. dan Sist. Informasi)*, vol. 2, pp. 9–13, 2016.

[9] Mardi Y, "Jurnal Edik Informatika Data Mining : Klasifikasi Menggunakan Algoritma C4 . 5 Data Mining Merupakan Bagian Dari Tahapan Proses Knowledge Discovery In Database (Kdd)," *J. Edik Inform.*, p. 215, 2016.

[10] H. Azis, F. Tangguh Admojo, and E. Susanti, "Analisis Perbandingan Performa Metode Klasifikasi pada Dataset Multiclass Citra Busur Panah," *Techno.Com*, vol. 19, no. 3, pp. 286–294, 2020, doi: 10.33633/tc.v19i3.3646.

[11] L. Nurhayati and H. Azis, "Perancangan Sistem Pendukung Keputusan untuk Proses Kenaikan Jabatan Struktural pada Biro Kepegawaian Setda Propinsi Maluku Utara," *Semnasteknomedia Online*, pp. 6–7, 2015.

[12] D. Septiani, "Dan Naive Bayes Untuk Prediksi Penyakit Hepatitis," *J. Pilar Nusa Mandiri*, vol. 13, no. 1, pp. 76–84, 2017.

[13] H. Leidiyana, "Penerapan Algoritma K-Nearest Neighbor Untuk Penentuan Resiko Kredit Kepemilikan Kendaraan Bermotor," *J. Penelit. Ilmu Komputer, Syst. Embed. Log.*, vol. 1, no. 1, pp. 65–76, 2013.

[14] Gavin Hackeling, *Mastering Machine Learning with scikit-learn*. 2014.

[15] M. M. Baharuddin, H. Azis, and T. Hasanuddin, "Analisis Performa Metode K-Nearest Neighbor Untuk Identifikasi Jenis Kaca," *Ilk. J. Ilm.*, vol. 11, no. 3, pp. 269–274, 2019, doi: 10.33096/ilkom.v11i3.489.269-274.

[16] Achmad Ridok, "Klasifikasi Dokumen Berbahasa Indonesia Menggunakan Metode K-NN," *J. Pointer*, vol. 1, p. 44, 2019.

[17] N. L. Suryani, "Pengaruh Lingkungan Kerja Non Fisik Dan Komunikasi Terhadap Kinerja Karyawan Pada PT. Bangkit Maju Bersama Di Jakarta," *JENIUS (Jurnal Ilm. Manaj. Sumber Daya Manusia)*, vol. 2, no. 3, p. 419, 2019, doi: 10.32493/jjsdm.v2i3.3017.