Journal of Information System and Application Development
Volume 3 Number 1 March 2025 page 31-37
P-ISSN: 2988-5698 | E-ISSN: 2988-4721
https://jurnal.unmer.ac.id/index.php/jisad

Evaluation of distance measurement techniques in the k-NN method for toddler nutritional status classification

Rahmatina Hidayati[1*], Anita[1], Sri Lestanti[2]
[1]Program Studi Sistem Informasi, Fakultas Teknologi Informasi, Universitas Merdeka Malang, Jl. Terusan Dieng No. 62-64, Malang, 65146, Indonesia
[2]Program Studi Teknik Informatika, Fakultas Teknologi Informasi, Universitas Islam Balitar, Jl. Imam Bonjol No. 16, Blitar, 66137,Indonesia

E-mail: *rahmatina.hidayati@unmer.ac.id

Abstract. Toddler nutritional status is an essential indicator in assessing public welfare and health. At Rongga Koe Village Health Post, determining nutritional status is still done manually, so it takes a long time and is prone to errors. This study aims to develop a classification system for toddler nutritional status using the K-Nearest Neighbors (k-NN). The data used was 100 samples with five parameters: gender, age, weight, height, and upper arm circumference. The classification process was carried out with variations in the ratio of training and testing data (90:10, 80:20, 70:30, 60:40), as well as the k value and distance calculation method (Euclidean, Manhattan, Chebyshev, Mahalanobis). The results showed that the best combination was obtained at a ratio of 90:10 and a k value = 9 with the Mahalanobis Distance method, which achieved the highest accuracy of 85.7%. This study proves that the K-NN method is effective in helping to classify nutritional status digitally and more efficiently.

Keywords: toddler, classification, k-Nearest Neighbors, nutritional status

Submitted: 11-03-2025 | **Accepted:** 17-04-2025 | **Published:** 21-04-2025

How to Cite:
R. Hidayati, Anita, and S. Lestanti, "Evaluation of distance measurement techniques in the k-NN method for toddler nutritional status classification," *Journal of Information System and Application Development (JISAD)*, vol. 3, no. 1, pp. 31-37, 2025, doi: 10.26905/jisad.v3i1.15393.

INTRODUCTION

Nutritional status is one of the key indicators reflecting societal welfare and is closely associated with public health [1] Toddlers represent a vulnerable age group due to their critical phase of growth and development. Therefore, it is essential to consider factors that may increase the risk of nutritional disorders to prevent malnutrition [2]. Nutritional status can be assessed through two primary methods: laboratory examinations and anthropometric measurements. According to the Decree of the Minister of Health of the Republic of Indonesia No. 1995/MENKES/SK/XII/2010, the Ministry of Health recommends the use of anthropometric parameters, which include age, body weight, body length/height, and Body Mass Index (BMI) [3].

At the Rongga Koe Village Health Post, the nutritional status of toddlers is assessed based on parameters such as sex, age, height, and mid-upper arm circumference. These measurements are recorded in the Healthy Menu Card and documented in the toddler nutritional status monitoring form. Subsequently, the data are compared against reference tables provided in the guideline book to classify nutritional status into three categories: normal, undernutrition, and severe undernutrition [4]. The current method of determining nutritional status is still manual, which requires considerable time. Data mining techniques are necessary to address this issue and optimize data processing and decision-making [5].

DOI: 10.26905/jisad.v3i1.15393

One technique commonly used in data mining is classification [6]. Classification is a data grouping technique that falls under supervised learning, where the goal is to assign class labels to previously unlabeled data [6], [7]. Various classification algorithms have been employed to assess nutritional status, including Artificial Neural Networks (ANN) [8], Support Vector Machine (SVM) [9], Naive Bayes, and k-Nearest Neighbors (k-NN). Among these methods, k-NN gets the highest level of accuracy [10].

This study aims to classify the nutritional status of toddlers at the Rongga Koe Village Health Post using the k-NN method. The analysis includes evaluating different training and testing data ratios [11], calculating distance [12], the optimal number of neighbors (k value) [13].

METHOD

In this study, the k-Nearest Neighbors (k-NN) algorithm is utilized to classify the nutritional status of toddlers by analyzing a combination of anthropometric and demographic data. As a non-parametric and instance-based learning approach, k-NN determines the classification of a new data point by referencing the most frequent class among its k nearest neighbors within the multidimensional feature space. The methodology starts with data preparation and preprocessing, followed by testing multiple train-test data splits (such as 70:30 and 80:20) to assess the model's consistency and predictive capability.

To measure the similarity between instances, suitable distance metrics—typically Euclidean or Manhattan distance—are applied, with the choice depending on data characteristics and scaling. The value of k is optimized through empirical testing and validation procedures, such as cross-validation, to reduce overfitting or underfitting and achieve optimal classification accuracy.

Key input features in this classification include sex, age in months, body weight, height, and mid-upper arm circumference (MUAC)—all of which are essential in evaluating a child's nutritional health. These variables may be used in raw form or converted into standardized nutritional indicators like WAZ (weight-for-age), HAZ (height-for-age), and WHZ (weight-for-height), based on WHO standards.

Data source

The primary data for this study were obtained from the Rongga Koe Village Health Post (Poskesdes Rongga Koe). The dataset consists of 100 data entries and includes five attributes. The data contain categorical and numerical variables, with categorical variables converted into numerical format for analysis.

Handling Missing Value

Missing value or missing data refer to instances where some or all information is unavailable in a data observation [14]. They commonly occur when values in specific assessment attributes are not recorded [15]. If not adequately addressed, missing value can introduce bias into the analysis and reduce the validity of research outcomes. Therefore, appropriate techniques must be applied to handle missing values before model training.

k-Nearest Neighbors (k-NN)

K-NN is a classification method that determines an object's class based on the k value [16]. The parameter k determines the number of nearest neighbors used in decision-making [17], [18]. Determining the nearest neighbor in k-NN depends on the distance calculation method chosen. Some distance calculations that can be used include [19]:

1. Euclidean Distance

   Calculating the distance between 2 points in the feature space

$$d_{euc}(x, y) = \sqrt{\sum_{j=1}^{d}(x_j - y_j)^2} \tag{1}$$

Description:
x = data; y = label (ourput)

2. Manhattan Distance

$$d_{man}(x, y) = \sum_{j=1}^{d} |x_j - y_j| \quad (2)$$

3. Chebyshev Distance

$$d_{che}(x, y) = \max_{1 \le k \le d} |x_j - y_j| \quad (3)$$

4. Mahalanobis Distance

Mahalanobis Distance is closely related to the relationship between variables in a dataset through a covariance matrix [20]. This method measures the distance of a point to the center of the data distribution, considering the variability and correlation between features. In its application, the data center value can be determined using the Mean or Median, depending on the dataset's characteristics. The calculation of Mahalanobis Distance is expressed in matrix and vector form, with the general equation [21]:

If N is odd, then $$med = X_{(\frac{N}{2}+1)} \quad (4)$$

If N is even, then $$med = \frac{x_{\frac{N}{2}} + x_{(\frac{N}{2}+1)}}{2} \quad (5)$$

Description:
X_i = i-th data in a data set, where i =1,2,…,n; N = number of samples in a data set

The computation of Mahalanobis distance involves matrix and vector representation and is expressed through the general formulation presented in Equation (6).

$$D^2 = (x_i - me)C^{-1}(x_i - me)^T \quad (6)$$

Description:
x_i= Variance of each data from 1 to N; me = The central value of a group of variables; C^{-1} = Invers matriks covarian

RESULTS AND DISCUSSION

Results

Table 1 presents a subset of the dataset utilized in this study. Before analysis, the data were preprocessed to identify and address missing values. No missing values were found in the dataset obtained from the health center. Subsequently, categorical variables were encoded into numerical format. For instance, in the sex category, 'M' (male) was encoded as 1, and 'P' (female) as 0.

Table 1. Dataset

No.	Name	Sex	Age (month)	Weight Body	Height	LILA (cm)	Status
1.	Name 1	L	59	16	105	16	Normal nutritional
2.	Name 2	P	59	15	100	15.8	Normal nutritional
3.	Name 3	L	58	17.8	98	17	Normal nutritional
4.	Name 4	L	57	13.3	100	14	Undernutrition
5.	:	:	:	:	:	:	:
6.	:	:	:	:	:	:	:
7.	Name 100	P	48	14	88	16	Normal nutritional

Table 1 presents the dataset, which was subsequently divided into training and testing subsets using proportions: 90:10, 80:20, 70:30, and 60:40. and 60:40. Following this division, the data were classified. This study's classification was performed using the Orange data mining software.

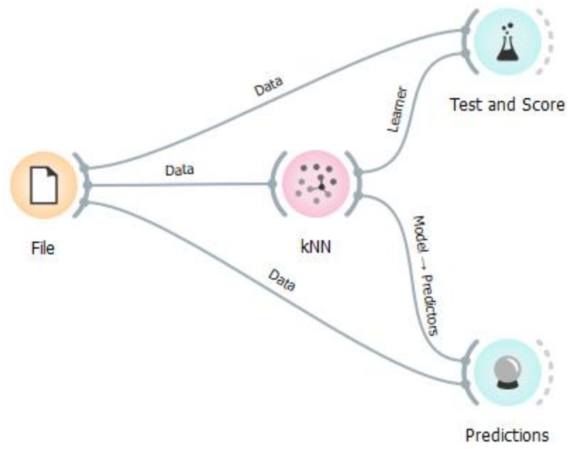

Figure 1. Implementation of Data Analysis

Figure 1 illustrates the k-NN classification process using the Orange data mining software. The classification procedure was repeated multiple times to ensure robustness. In addition to evaluating the division of training and testing data, this study also investigated the impact of distance metrics on classification performance. Table 2 presents the accuracy of the k-NN classifier using Euclidean distance calculations for various values of the number of neighbors (k).

Table 2. Results of k-NN Classification Accuracy Using Euclidean Distance

Data Ratio	k = 3	k = 5	k = 7	k = 9
90:10	81.8%	79.1%	79%	79.2%
80:20	77.5%	75.5%	75.9%	75.9%
70:30	76.7%	75.9%	76.2%	76.1%
60:40	73.9%	73.9%	74.1%	74.8%

Based on Table 2, the highest classification accuracy, 81.8%, was achieved using a data split ratio of 90:10 and a neighborhood value of k = 3. This result indicates that increasing the proportion of training data improves the ability of the k-Nearest Neighbors (kNN) model with Euclidean Distance to learn and recognize data patterns. Furthermore, a tremendous k value tends to reduce classification accuracy.

Table 3. Results of k-NN Classification Accuracy Using Manhattan Distance

Data Ratio	k = 3	k = 5	k = 7	k = 9
90:10	80.7%	81.1%	81.6%	82.7%
80:20	76.6%	77.5%	78.9%	79.6%
70:30	76.9%	77.9%	79.4%	79.7%
60:40	74.8%	76.7%	77.9%	78%

The results presented in Table 3 indicate that the optimal performance is achieved with a 90:10 training-to-testing data ratio and k = 9, yielding an accuracy of 82.7%. In contrast to Euclidean Distance, the Manhattan distance metric exhibits a trend of increasing accuracy as the value of k increases. Suggests that, within the context of the dataset used in this study, the Manhattan algorithm demonstrates greater stability and tolerance to variations in the number of neighbors (k).

Table 4. Results of k-NN Classification Accuracy Using Chebyshev Distance

Data Ratio	k = 3	k = 5	k = 7	k = 9
90:10	79.4%	76.9%	77.7%	76.9%
80:20	76.1%	74.4%	74.1%	74.1%
70:30	75.4%	74.1%	74.1%	74.1%
60:40	72.1%	76.1$	72.2%	73.1%

Table 4 presents the highest accuracy, 79.4%, obtained with a 90:10 training-to-testing data ratio and k = 3. Overall, the accuracy of the Chebyshev algorithm is lower than that of the Euclidean and Manhattan algorithms. Furthermore, the accuracy of the Chebyshev algorithm does not show a significant improvement, even when varying the value of k or the data ratio. Suggests that the Chebyshev algorithm is suboptimal in measuring the proximity between data points in the context of toddler nutritional status classification at the Rongga Koe Village Health Post.

Table 5. Results of k-NN Classification Accuracy Using Mahalanobis Distance

Data Ratio	k = 3	k = 5	k = 7	k = 9
90:10	82.1%	83%	84.4%	85.7%
80:20	79.6%	81.1%	83.2%	84.3%
70:30	79.3%	80.8%	82.7%	82.8%
60:40	78.2%	80.4%	81.9%	82%

Table 5 presents the highest accuracy achieved, 85.7%, obtained with a 90:10 training-to-testing data ratio and k = 9. A clear trend is observed, where an increase in the value of k is directly proportional to an increase in accuracy, similar to the pattern observed with the Manhattan distance metric

Discussion

Table 6. Summary of k-NN Classification Accuracy

Distance Metric	Data Ratio	Highest Accuracy	Best k Value	Accuracy Trend
Euclidean	90:10	81.8%	3	Decreases as k increases
Manhattan	90:10	82.7%	9	Increases with higher k
Chebyshev	90:10	79.4%	3	No consistent trend, generally lower
Mahalanobis	90:10	85.7%	9	Accuracy increases with k

Table 6 summarizes classification accuracy for each distance based on variations in data splitting ratios and the number of neighbors (k values). This comparison highlights how different parameter configurations influence the performance of the k-Nearest Neighbors (k-NN) method. This study successfully applied the k-Nearest Neighbors (k-NN) algorithm to classify the nutritional status of toddlers at the Rongga Koe Health Post using the parameters of gender, age, weight, height, and upper arm circumference. The results of testing with various data splitting ratios and distance calculation algorithms revealed that the optimal combination was achieved with a 90:10 training-to-testing ratio and k = 9, using Mahalanobis Distance, which resulted in the highest accuracy of 85.7%. These findings suggest that a larger proportion of training data and a more precise selection of the distance metric and the k value leads to improved classification accuracy.

CONCLUSION AND SUGGESTIONS

This study successfully implemented the k-Nearest Neighbors (k-NN) algorithm to classify the nutritional status of toddlers at the Rongga Koe Village Health Post, utilizing parameters such as gender, age, weight, height, and upper arm circumference. The findings indicate that classification accuracy improves with an increase in the proportion of training data and the selection of an appropriate distance metric and k value. These results highlight the importance of selecting suitable data partitioning techniques and distance calculation to enhance model performance. For future research, it is recommended to utilize a larger and more diverse dataset to improve the generalization of the classification results, making them applicable to a broader region. Additionally, employing more advanced validation techniques, such as k-fold cross-validation, would contribute to greater reliability and stability of the results.

BIBLIOGRAPHY

[1] W. Wahyudi, “Optimasi Klasifikasi Status Gizi Balita Berdasarkan Indeks Antropometri Menggunakan Algoritma C4.5 Adaboost Classification,” *Jurnal Ilmiah Komputerisasi Akuntansi*, vol. 12, no. 2, 2019.

[2] E. Ramon, A. Nazir, N. Novriyanto, Y. Yusra, and L. Oktavia, “Klasifikasi Status Gizi Bayi Posyandu Kecamatan Bangun Purba Menggunakan Algoritma Support Vector Machine (SVM),” *Jurnal Sistem Informasi dan Informatika (Simika)*, vol. 5, no. 2, 2022, doi: 10.47080/simika.v5i2.2185.

[3] Kementrian Kesehatan RI Direktorat Jendral Bina Gizi dan Kesehatan Ibu Tentang Standar Antropometri Penilaian Status Gizi dan Anak, “Keputusan Menteri Kesehatan Republik Indonesia Nomor:1995/MENKES/SK/XII/2010 Tentang Standar Antropometri Penilaian Status Gizi Anak,” 2011.

[4] H. Saleh, M. Faisal, and R. I. Musa, “Klasifikasi Status Gizi Balita Menggunakan Metode K-Nearest Neighbor,” *Simtek : jurnal sistem informasi dan teknik komputer*, vol. 4, no. 2, 2019, doi: 10.51876/simtek.v4i2.60.

[5] R. Shafira, Prajoko, and A. Pambudi, “Penilaian status gizi balita menggunakan metode K-Nearest Neighbor,” *Technologia : Jurnal Ilmiah*, vol. 14, no. 3, 2023, doi: 10.31602/tji.v14i3.11299.

[6] S. Lonang and D. Normawati, “Klasifikasi Status Stunting Pada Balita Menggunakan K-Nearest Neighbor Dengan Feature Selection Backward Elimination,” *JURNAL MEDIA INFORMATIKA BUDIDARMA*, vol. 6, no. 1, 2022, doi: 10.30865/mib.v6i1.3312.

[7] V. Wulandari, W. J. Sari, Z. Alfian, L. Legito, and T. Arifianto, “Implementasi Algoritma Naïve Bayes Classifier dan K-Nearest Neighbor untuk Klasifikasi Penyakit Ginjal Kronik,” *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, vol. 4, no. 2, pp. 710–718, Apr. 2024, doi: 10.57152/malcom.v4i2.1229.

[8] G. P. Insany, I. Yustiana, and S. Rahmawati, “Penerapan KNN dan ANN pada klasifikasi status gizi balita berdasarkan indeks antropometri,” *Jurnal CoSciTech (Computer Science and Information Technology)*, vol. 4, no. 2, 2023, doi: 10.37859/coscitech.v4i2.5079.

[9] S. D. Wahyuni and R. H. Kusumodestoni, “Optimalisasi Algoritma Support Vector Machine (SVM) Dalam Klasifikasi Kejadian Data Stunting,” *Bulletin of Information Technology (BIT)*, vol. 5, no. 2, pp. 56–64, 2024, doi: 10.47065/bit.v5i2.1247.

[10] R. Setiawan and A. Triayudi, “Klasifikasi Status Gizi Balita Menggunakan Naïve Bayes dan K-Nearest Neighbor Berbasis Web,” *JURNAL MEDIA INFORMATIKA BUDIDARMA*, vol. 6, no. 2, 2022, doi: 10.30865/mib.v6i2.3566.

[11] A. Fitri, N. Azizah, and V. P. Ramadhan, “Komparasi Naïve Bayes dan K-NN Dalam Analisis Sentimen di Twitter Terhadap Kemenangan Paslon 02,” *J-INTECH (Journal of Information and Technology)*, vol. 12, no. 2, pp. 228–237, 2024.

[12] A. Setiawan, “Perbandingan Penggunaan Jarak Manhattan, Jarak Euclid, dan Jarak Minkowski dalam Klasifikasi Menggunakan Metode KNN pada Data Iris,” *Jurnal Sains dan Edukasi Sains*, vol. 5, no. 1, 2022, doi: 10.24246/juses.v5i1p28-37.

[13] F. A. A. Putra, A. G. Sulaksono, L. T. Utomo, and A. R. Khamdani, “Klasifikasi Buah dan Sayur Menggunakan Fitur Ekstraksi HOG dan Metode KNN,” *Jurnal Informatika Polinema*, vol. 10, no. 1, 2023, doi: 10.33795/jip.v10i1.1433.

[14] N. Istiana and A. Mustafiril, “Perbandingan Metode Klasifikasi pada Data dengan Imbalance Class dan Missing Value,” *Jurnal Informatika*, vol. 10, no. 2, 2023, doi: 10.31294/inf.v10i2.15540.

[15] A. Fadlil, Herman, and D. Praseptian M, “K Nearest Neighbor Imputation Performance on Missing Value Data Graduate User Satisfaction,” *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 6, no. 4, 2022, doi: 10.29207/resti.v6i4.4173.

[16] A. T. Akbar, N. Yudistira, and A. Ridok, “Identifikasi Gagal Ginjal Kronis dengan Mengimplementasikan Metode Support Vector Machine beserta K-Nearest Neighbour (SVM-KNN),” *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 10, no. 2, 2023, doi: 10.25126/jtiik.20231026059.

[17] H. R. Ngemba *et al.*, “Penerapan Algoritma K-Nearest Neighbor untuk Menentukan Potensi Ekspor Komoditas Pertanian di Provinsi Sulawesi Tengah,” *Jurnal Teknologi dan Manajemen Informatika*, vol. 9, no. 2, 2023, doi: 10.26905/jtmi.v9i2.10235.

[18] R. A. Safitri and R. Hidayati, “Komparasi Metode K-Nearest Neighbor dan Naïve Bayes untuk Mengklasifikasi Resiko Diabetes Di Posbindu Desa Bulupitu,” *SMATIKA JURNAL*, vol. 14, no. 02, pp. 297–303, Dec. 2024, doi: 10.32664/smatika.v14i02.1350.

[19] R. Hidayati, A. Zubair, A. H. Pratama, and L. Indana, “Analisis Silhouette Coefficient pada 6 Perhitungan Jarak K-Means Clustering,” *Techno.Com*, vol. 20, no. 2, 2021, doi: 10.33633/tc.v20i2.4556.

[20] M. N. Kadhim, D. Al-Shammary, A. M. Mahdi, and A. Ibaida, “Feature selection based on Mahalanobis distance for early Parkinson disease classification,” *Computer Methods and Programs in Biomedicine Update*, vol. 7, pp. 1–14, 2025, doi: 10.1016/j.cmpbup.2025.100177.

[21] T. Kantacarini, D. E. Herwindiati, and J. Hendryli, “Kinerja Metode Mahalanobis Distance Yang Dibentuk Dari Dua Ukuran Pusat Dan Dua Dispersi Multivariat (Untuk Ukuran Similaritas Klasifikasi Image),” *Jurnal Ilmu Komputer dan Sistem Informasi*, vol. 9, no. 1, 2021, doi: 10.24912/jiksi.v9i1.12667.