p-ISSN 2338 – 3216
e-ISSN 2528 – 1070

JURNAL STATISTIKA
UNIVERSITAS MUHAMMADIYAH SEMARANG
13(1) 2025: 01-13

COMPARISON OF K-NEAREST NEIGHBOR AND NAÏVE BAYES CLASSIFICATION METHODS FOR STATUS OF TODDLER NUTRITION DATA AT BAQA SAMARINDA SEBERANG COMMUNITY HEALTH CENTER

Muzizah Annabaa' Aulia[1*], Rito Goejantoro[2], Memi Nor Hayati[3]
[1,2,3] Statistics Study Program, Department of Mathematics, Faculty of Mathematics and Natural Sciences (FMIPA), Mulawarman University, Samarinda, Indonesia

***e-mail**: *muzizah09@gmail.com*

Article Info:
Received: September, 24th 2024
Accepted: September, 4th 2025
Available Online: September 5th 2025

Keywords:
Classification; naïve Bayes; k-nearest neighbors; nutritional status of toddlers

Abstract: Classification is a job of assessing data objects to put them into a certain class from a number of available classes. The naïve Bayes method is a statistical classification that can be used to estimate the probability of membership in a class. Meanwhile, the K-Nearest Neighbor (K-NN) method is a supervised method used for classification. The aim of this research is to obtain classification results of the nutritional status of toddlers at the Baqa Samarinda Seberang Community Health Center in 2022 using the naïve Bayes algorithm and the K-NN algorithm. Based on the calculation results for classification of the nutritional status of toddlers at the Baqa Samarinda Seberang Community Health Center using accuracy calculations and confusion matrices, the highest accuracy was obtained using the naïve Bayes method of 82.15% and a Press's Q value of 168 with a training data proportion of 90%: testing data of 10%. Meanwhile, the results of accuracy calculations and the confusion matrix obtained the highest accuracy in the K-NN method of 90.57% at values 3-NN, 5-NN, 7-NN, 9-NN and Press's Q value of 187.65 with a training data proportion of 90% and testing data 10%. From the results of this analysis, it was concluded that the K-NN method worked better than the naïve Bayes method in classifying the nutritional status of toddlers at the Baqa Samarinda Seberang Community Health Center.

1. INTRODUCTION

Data mining is finding patterns and relationships hidden in large data to perform estimation, prediction, association rule, clustering, description, visualization, and classification [1]. There are several classification algorithms including Support Vector Machine (SVM), Decision Tree, K-Nearest Neighbor (K-NN), and naïve Bayes. The K-NN algorithm classification method is one of the data classification methods with a strong level of consistency, by finding a case and calculating the proximity between the new case and the old case based on the weight match [2]. At the same time, naïve Bayes is a statistical classification method used to estimate the probability of members of a group or class. Research conducted by Rahmaulidyah et al (2021) regarding the comparison of naïve Bayes and K-NN classification methods on Value Added Tax Payment Status Data at the Samarinda Ulu Primary Tax Service Office shows the results of classification errors in predicting 19.51% [3]. Then, research by Mustaghfiroh et al (2022) on the classification of

[DOI: 10.14710/JSUNIMUS.13.1.2025.1-13]

COVID-19 patients in Indonesia using the K-NN method with an accuracy of 97.76% [4]. Meanwhile, research conducted by Saeroni et al (2020) regarding the classification of the level of customer fluency in paying premiums using the K-Nearest Neighbor method and Fisher Discriminant Analysis (Case study: PT. Prudential Life Samarinda Customer Data in 2019) shows the results of classification errors in predicting classes using APER of 15% [5].

The results of the Indonesian Nutrition Status Survey (SSGI) of the Ministry of Health show that there are four nutritional problems of toddlers in Indonesia, including stunting, wasting, underweight, and overweight. Stunting is one of the nutrition problems that is of concern to the government because of its high prevalence, reaching 21.6% by 2022. This figure certainly exceeds the threshold set by the World Health Organization (WHO) standard of 20%. Therefore, the purpose of this research is to provide information and scientific insight into classification techniques using naïve Bayes and K-NN in the problem of classifying the nutritional status of toddlers in an area.

2. LITERATURE REVIEW

2.1. Classification

Classification is a job of assessing data objects to put them into a certain class from a number of available classes. In classification there are two main processes carried out, namely building a model as a prototype to be stored as memory and using the model to carry out recognition or classification or predictions on another data object so that it is known which class the data object belongs to in the model that has been stored [6].

2.2. Naïve *Bayes* Method

A scientist from England, namely Thomas Bayes, stated that naïve Bayes is a classification method using probability and statistics. This method can also predict future opportunities based on previous experience, so it is known as Bayes' Theorem. This theorem is combined with naïve where it is assumed that the conditions between attributes are mutually independent. Naïve Bayes classification is assumed if the presence or absence of certain characteristics of a class has nothing to do with the characteristics of other classes [7].

According to [8], the equation of Bayes' Theorem is as follows:

$$P(C|F) = \frac{P(C)P(F|C)}{P(F)} \quad (1)$$

with:

P(C | F): the probability of C occurring provided that F has occurred
P(C): the probability of C occurring
P(F|C): the probability of F occurring provided that C has occurred
P(F): the probability of F occurring.

Please note that the classification process requires a number of clues (attributes) to determine which class or group is appropriate for the object being analyzed. Therefore Bayes' Theorem can be explained and adapted as follows:

$$P\left(Y \middle| X_1, X_2, \dots, X_p\right) = \frac{P(Y)\ P\left(X_1, X_2, \dots\ X_p \middle| Y\right)}{P(X_1, X_2, \dots\ X_p)} \quad (2)$$

with:

$P\left(Y \middle| X_1, X_2, \dots, X_p\right)$: probability of inclusion of an object with certain variable characteristics in group Y (posterior)

$P\left(X_1, X_2, \dots, X_p \middle| Y\right)$: probability of the appearance of variables in objects that are included in group Y (likelihood)

[DOI: 10.14710/JSUNIMUS.13.1.2025.1-13]

$P(Y)$: probability of group Y appearing before the object enters (prior)
$P(X_1, X_2, \dots, X_p)$: probability of the appearance of variables in objects in general (evidence).

The naïve independence assumption makes the probability conditions simple, so that calculations are easy to do. Next, for the explanation of $P\left(Y\middle|(X_1, X_2, \dots X_p\right)$ which can be simplified to

$$P\left(Y\middle|(X_1, X_2, \dots, X_p\right) = P(Y)\, P(X_1|Y)P(X_2|Y)P(X_3|Y) \dots P\left(X_p\middle|Y\right)$$

$$P\left(Y\middle|(X_1, X_2, \dots, X_p\right) = P(Y)\prod_{k=1}^{p} P(X_k|Y) \quad (3).$$

In Equation (3) is a model from Naïve Bayes' theorem that will be used in the classification process. In general, the naïve Bayes theorem is easy to calculate for observed values of input variables with numeric (non-categorical) types. There is special treatment before processing using naïve Bayes, namely in the following way:

a. Discretize each observation value of a continuous input variable and replace the observation value with a discrete interval value. This approach is carried out by transforming into an ordinal scale.
b. Assumes a particular shape of the probability distribution for continuous observed values and estimates the parameters of the distribution with training data. The Normal Distribution is usually chosen to represent conditional probabilities on continuous observed values in a group $P(Y\,|X_k)$ as follows:

$$P\left(X = x_k\middle|\, Y = y_g\right) = \frac{1}{\sqrt{2\pi}\,\sigma_{gk}} \exp - \left[\frac{(x_k - \mu_{gk})^2}{2\sigma_{gk}^2}\right], g=1,2,..,q \;\&\; k=1,2,\dots,p \quad (4)$$

with:

x_k : value of the k variable
y_g : g group.

2.3. *K-Nearest Neighbor* Method

The K-NN method is a supervised method used for classification (with output variables or dependent variables in the form of categories) [1]. The working principle of this method is to find the shortest distance between the data to be evaluated and the K value of its closest neighbors in the testing data. From the selected K value of the nearest neighbor, the highest class or class voting from the K nearest neighbors will be carried out. If there is a class that has the highest number of neighbors' votes, then a class label will be given as a result of the prediction in the testing data [8].

How far or close a point is from its neighbors can be calculated using the Euclidean distance which is presented as follows [7]:

$$d\left(x_i\,, x^*{}_j\right) = \sqrt{\sum_{k=1}^{p}\left(x_{ik} - x^*{}_{jk}\right)^2} \quad (5)$$

2.4. Data Standardization

If the Euclidean distance is smaller, the more similar the cases or objects are. However, the Euclidean distance is very sensitive to sample size and the magnitude of the variance. If the object under study has very different variants, then the Euclidean distance becomes inaccurate. Therefore, it is necessary to standardize the research variables [9].

$$Z_{rv} = \frac{X_{rv} - \bar{X}_v}{S_v} \quad (6)$$

2.5. Evaluation of the Accuracy of Classification Results

[DOI: 10.14710/JSUNIMUS.13.1.2025.1-13]

The confusion matrix is a table for recording the results of classification work. Table 1 is an example of a confusion matrix that classifies problems in six classes, there are only six classes, namely class 0, class 1, class 2, class 3, class 4 and class 5.

Table 1 Confusion Matrix for Six Class Classification

f_{ab}		Prediction result class (b)					
		Class= 0	**Class=1**	**Class=2**	**Class=3**	**Class=4**	**Class=5**
Original Class (a)	Class = 0	f_{00}	f_{01}	f_{02}	f_{03}	f_{04}	f_{05}
	Class = 1	f_{10}	f_{11}	f_{12}	f_{13}	f_{14}	f_{15}
	Class= 2	f_{20}	f_{21}	f_{22}	f_{23}	f_{24}	f_{25}
	Class = 3	f_{30}	f_{31}	f_{32}	f_{33}	f_{34}	f_{35}
	Class = 4	f_{40}	f_{41}	f_{42}	f_{43}	f_{44}	f_{45}
	Class = 5	f_{50}	f_{51}	f_{52}	f_{53}	f_{54}	f_{55}

$$\text{Accuracy} = \frac{\text{Amount of data correctly predicted}}{\text{Number of predictions made}} \times 100\%$$

$$= \frac{f_{00} + f_{11} + f_{22} + f_{33} + f_{44} + f_{55}}{\begin{matrix} f_{01} + f_{02} + f_{03} + f_{04} + f_{05} + f_{10} + f_{12} + f_{13} + f_{14} + f_{15} + \\ f_{20} + f_{21} + f_{23} + f_{24} + f_{25} + f_{30} + f_{31} + f_{32} + f_{34} + f_{35} + \\ f_{40} + f_{41} + f_{42} + f_{43} + f_{45} + f_{50} + f_{51} + f_{52} + f_{53} + f_{54} \end{matrix}} \times 100\% \quad (7)$$

The stability test was carried out to test whether the allocation of each sample in a group was relatively stable or not as a result of changes in the number of samples studied [10]. The test that can be carried out is by calculating the Press's Q value which is formulated as follows [11]:

Hypothesis :

H_0 : Inaccurate classification

H_1 : Accurate classification

Test statistic:

$$Press's\ Q = \frac{[n_{tes} - (mq)]^2}{n_{tes}(q-1)} \quad (8)$$

Test criteria: reject H_0 if Press's Q > $\chi^2_{\alpha,1}$.

2. METHODOLOGY

3.1 Data and Source Data

The data used is data on 531 toddlers residing in Samarinda Seberang in 2022 taken from the Baqa Samarinda Seberang Health Center. The research variables consist of input variables and target variables which are detailed in Table 2 as follows.

[DOI: 10.14710/JSUNIMUS.13.1.2025.1-13]

Table 2 Research variable

Research variable	Information	Data Scale
Nutritional Status of Toddlers (Y)	The nutritional status of toddlers is said to be well-nourished if their weight and height according to age are calculated according to the Z score, the value is from -2 SD to +1 SD, poor nutrition if the Z score is less than -3 SD, said to be malnourished if the Z score is - 3 SD to less than -2 SD, said to be at risk of overnutrition if the Z score is more than +1 SD to +2 SD, said to be overnourished if the Z score is more than +2 SD to +3 SD, and said to be obese if the score Z is more than +3 SD	Ordinal
Age (X_1)	Age	Ratio
Weight (X_2)	Toddler weight according to measurements in 2022	Ratio
Height (X_3)	Toddler height according to measurements in 2022	Ratio

3.2 Steps of Analysis

The steps for the data analysis method to classify the nutritional status of toddlers using the naïve Bayes and K-NN methods in this research are as follows:

1. Descriptive statistical analysis
2. Data randomization
3. Data Standardization
5. Classification using the naïve Bayes method
6. Classification using the K-NN method
7. Compare classification results

4. RESULTS AND DISCUSSION

4.1 Descriptive Statistics

This descriptive statistical analysis was carried out to find a general description of the nutritional status data for toddlers at the Baqa Samarinda Seberang Community Health Center in 2022. The characteristics described in the descriptive statistical analysis are the age of the toddler, the height of the toddler, and the weight of the toddler as follows.

[DOI: 10.14710/JSUNIMUS.13.1.2025.1-13]

Table 3 Results of Descriptive Statistical Analysis

Variabel	Minimum	Maximum	Mean	Standard Deviation
(X_1)	8	69	39.34	15.88
(X_2)	2.6	26.3	11.53	3.34
(X_3)	46	116	86.20	12.75

4.2 Data randomization

Data randomization was carried out using R software.

4.3 Data Standardization

Next, a data standardization stage will be carried out because there are quite large differences in unit size between the data variables and this can cause the distance calculations in the classification analysis to be invalid. Then, the results of data standardization will be used in calculating the K-NN method. The results of data standardization can be seen in Table 4 as follows.

Table 4 Data Standardization Results

Sampel	Age (Month)	Weight (kg)	Height (cm)	Nutritional Status of Toddlers
495	-0.65164	-0.42799	-0.32966	Good nutritional
408	1.49108	1.51743	1.475603	Good nutritional
338	1.49108	1.90652	1.711072	Good nutritional
361	1.49108	1.33785	1.475603	Good nutritional
454	0.10462	-0.48785	-0.17268	Good nutritional
⋮	⋮	⋮	⋮	⋮
17	-0.77768	-0.72729	-0.48664	Good nutritional

4.4 Classification Using the naïve Bayes method

The first stage in the classification process uses the naïve Bayes method, namely, calculating the initial probability (prior) for the six groups using training data. The proportion of data used for classification is the proportion of training data 90% : testing data 10%. The results of the prior values for each group can be seen in Table 5 as follows:

Table 5 Prior Probability in Each Group

Nutritional Status of Toddlers	Probability
Poor nutritional	0.00837

[DOI: 10.14710/JSUNIMUS.13.1.2025.1-13]

Malnutrition	0.06485
Good nutritional	0.83054
Risk of overnutrition	0.07322
Excess nutrition	0.01465
Obesity	0.00837

The three input variable values in the testing data are then given the probability values for each independent variable in the six groups. Calculation of the probability value for each independent variable in the six groups assumes that the input variables are normally distributed using Equation (4) as follows:

The probability values for the age variable for the first testing data in each group are as follows:

The probability of the age variable for poor nutritional status of toddlers is calculated as follows:

$$P(X_1 = 44|\ Y_1) = \frac{1}{\sqrt{(2)(3.14)}(20.83267)} \times exp - \left(\frac{(44-30)^2}{2\ (20.83267)^2}\right)$$
$$= 0.015283$$

The probability of the age variable for malnutrition status can be calculated as follows:

$$P(X_1 = 44|\ Y_2) = \frac{1}{\sqrt{(2)(3.14)}(14.17623)} \times exp - \left(\frac{(44-41.96774)^2}{2\ (14.17623)^2}\right)$$
$$= 0.027861.$$

The probability of the age variable for good nutritional status can be calculated as follows:

$$P(X_1 = 44|\ Y_3) = \frac{1}{\sqrt{(2)(3.14)}(15.70924)} \times exp - \left(\frac{(44-39.42317)^2}{2(15.70924)^2}\right)$$

$$= 0.02434$$

The probability of the age variable for status at risk of overnutrition can be calculated as follows:

$$P(X_1 = 44|\ Y_4) = \frac{1}{\sqrt{(2)(3.14)}(15.43743)} \times exp - \left(\frac{(44-39.74286)^2}{2(15.43743)^2}\right)$$
$$= 0.02488$$

The probability of the age variable for excess nutrition status can be calculated as follows:

$$P(X_1 = 44|\ Y_5) = \frac{1}{\sqrt{(2)(3.14)}(16.35761)} \times exp - \left(\frac{(44-39.28571)^2}{2(16.35761)^2}\right)$$
$$= 0.02340$$

The probability of the income variable for obesity status can be calculated as follows:

$$P(X_1 = 44|\ Y_6) = \frac{1}{\sqrt{(2)(3.14)}(20.11633)} \times exp - \left(\frac{(44-45)^2}{2(20.11633)^2}\right)$$
$$= 0.01981$$

Next, the calculation of the probability value for each independent variable in the six groups was also carried out for the weight and height variables in each group.

The next stage in carrying out classification using the naïve Bayes method is to calculate the product of the prior probability and posterior probability on the first testing data, with a toddler aged 44 months, weight 10.8 kg and height 85 cm. The calculation of multiplying the prior probability and posterior probability using Equation (5) is as follows:

[DOI: 10.14710/JSUNIMUS.13.1.2025.1-13]

a. First group (toddlers with poor nutritional status)

$$P(Y_1|X_1,X_2,X_3) = P(Y_1) \times P(X_1|Y_1) \times P(X_2|Y_1) \times P(X_3|Y_1)$$
$$= P(Y_1) \times P(X_1 = 44|\ Y_1) \times P(X_2 = 10.8 \mid Y_1) \times P(X_3 = 85 \mid Y_1)$$
$$= (0.00837) \times (0.015283) \times (0.08777) \times (0.02228)$$
$$= 2.50113 \times 10^{-7}.$$

b. Second group (toddlers with malnutrition status)

$$P(Y_2|X_1,X_2,X_3) = P(Y_2) \times P(X_1|Y_2) \times P(X_2|Y_2) \times P(X_3|Y_2)$$
$$= P(Y_2) \times P(X_1 = 44|\ Y_2) \times P(X_2 = 10.8 \mid Y_2) \times P(X_3 = 85 \mid Y_2)$$
$$= (0.06485) \times (0.027861) \times (0.17742) \times (0.04107)$$
$$= 1.31672 \times 10^{-5}$$

c. Third group (toddlers with good nutritional status)

$$P(Y_3|X_1,X_2,X_3) = P(Y_3) \times P(X_1|Y_3) \times P(X_2|Y_3) \times P(X_3|Y_3)$$
$$= P(Y_3) \times P(X_1 = 44|\ Y_3) \times P(X_2 = 10{,}8 \mid Y_3) \times P(X_3 = 85 \mid Y_3)$$
$$= (0.83054) \times (0.02434) \times (0.13464) \times (0.03197)$$
$$= 8.70423 \times 10^{-5}$$

d. Fourth group (toddlers with overnutrition risk status)

$$P(Y_4|X_1,X_2,X_3) = P(Y_4) \times P(X_1|Y_4) \times P(X_2|Y_4) \times P(X_3|Y_4)$$
$$= P(Y_4) \times P(X_1 = 44|\ Y_4) \times P(X_2 = 10.8 \mid Y_4) \times P(X_3 = 85 \mid Y_4)$$
$$= (0.07322) \times (0.02488) \times (0.06602) \times (0.03004)$$
$$= 3.61412 \times 10^{-6}$$

e. Fifth group (toddlers with excess nutrition status)

$$P(Y_5|X_1,X_2,X_3) = P(Y_5) \times P(X_1|Y_5) \times P(X_2|Y_5) \times P(X_3|Y_5)$$
$$= P(Y_5) \times P(X_1 = 44 \mid Y_5) \times P(X_2 = 10.8 \mid Y_5) \times P(X_3 = 85 \mid Y_5)$$
$$= (0.01465) \times (0.02340) \times (0.06026) \times (0.02328)$$
$$= 4.80897 \times 10^{-7}$$

f. Sixth group (toddlers with obese nutritional status)

$$P(Y_6|X_1,X_2,X_3) = P(Y_6) \times P(X_1|Y_6) \times P(X_2|Y_6) \times P(X_3|Y_6)$$
$$= P(Y_6) \times P(X_1 = 44|\ Y_6) \times P(X_2 = 10.8 \mid Y_6) \times P(X_3 = 85 \mid Y_6)$$
$$= (0.00837) \times (0.01981) \times (0.02602) \times (0.01395)$$
$$= 6.02168 \times 10^{-8}$$

After the posterior values in the six groups are known, the maximum value of the posterior in each group is then determined to determine which objects will be classified into the six groups of toddler nutritional status. Based on the calculation of multiplying the initial probability and the probability of each input variable (posterior) in the six groups, it can be seen that the group that has the largest posterior is toddlers with a good nutritional status group of 8.70423×10^{-5} compared to other toddler nutritional status groups. So it can be concluded that the first testing data, namely a toddler aged 44 months, with a body weight of 10.8 kg and a height of 85 cm, was classified into the third group, namely toddlers with good nutritional status. The calculation of the multiplication of prior probabilities and posterior probabilities for each input variable for each group was carried out for 53 testing data.

After carrying out the classification calculations and getting the classification results using the naïve Bayes method, then evaluate the classification results using the naïve Bayes method using the confusion matrix, accuracy calculations and Press's Q calculations.

The results of the confusion matrix using the naïve Bayes method can be seen in Table 6 as follows:

[DOI: 10.14710/JSUNIMUS.13.1.2025.1-13]

Table 6 Confusion Matrix Classification Results of the Naïve Bayes Method

Initial Classification of Toddler Nutritional Status	Classification Prediction Naïve Bayes Method						Total
	Good nutritional	Poor nutritional	Malnutrition	Excess nutrition	Obesity	Risk of overnutrition	
Good nutritional	43	1*	0	0	0	3*	47
Poor nutritional	0	0	0	0	0	0	0
Malnutrition	1*	0	0	0	0	0	1
Excess nutrition	2*	0	0	0	0	1*	3
Obesity	0	0	0	0	0	0	0
Risk of overnutrition	2*	0	0	0	0	0	2
Total	48	1	0	0	0	4	53

So based on Equation (7), the following accuracy values are obtained:

$$\text{Accuracy} = \frac{43+0+0+0+0+0}{53} \times 100\%$$
$$= \frac{43}{53} \times 100\%$$
$$= 82.15\%$$

Next, to evaluate the accuracy of the prediction (classification) results, the Press's Q value is used. The Press's Q value is calculated using the confusion matrix in Table 6 and Equation (8) to obtain the following results.

Hypothesis testing :

H_0 : Inaccurate classification

H_1 : Accurate classification

Test statistic : Press's Q

Test criteria: reject H_0 if Press's Q > $\chi^2_{\alpha,1}$.

Calculation :

$$Press's\ Q = \frac{[53 - ((44)6)]^2}{53\ (6-1)}$$
$$= 168.001 \approx 168$$

With degrees of freedom worth one and a confidence level of $\alpha = 0.05$, and the value $\chi^2_{(0.05:1)} = 3.841 < Press's\ Q = 168$. So it can be concluded that the classification using the naïve Bayes method rejects H_0 or classification using the naïve Bayes method is accurate.

4.5 Classification Using the *K-Nearest Neighbor* method

Classification of the nutritional status of toddlers using the K-NN method is carried out by determining the K parameter value first. The K values used in this research are K=1, K=3, K=5, K=7 and K=9. Meanwhile, the proportion of data used for classification is the proportion of training data 90% : testing data 10%. Distance calculations in this study use results from previously standardized data. The calculation of the Euclidean distance between the first testing data and the first training data to the 478th training data is as follows:

[DOI: 10.14710/JSUNIMUS.13.1.2025.1-13]

$$d(x_1, x_1^*) = \sqrt{\begin{matrix}\left((-0.65192) - (0.29380)\right)^2 + \left((-0.42799) - (-0.21848)\right)^2 \\ + \left((-0.32965) - (0.09418)\right)^2\end{matrix}} = 0.99686$$

$$d(x_2, x_1^*) = \sqrt{\begin{matrix}\left((-1.49108) - (0.29380)\right)^2 + \left((1.51743) - (-0.21848)\right)^2 \\ + \left((1.4756) - (0.09418)\right)^2\end{matrix}} = 2.62919$$

$$\vdots$$

$$d(x_{478}, x_1^*) = \sqrt{\begin{matrix}\left((-0.77768) - (0.29380)\right)^2 + \left((-0.39806) - (-0.21848)\right)^2 \\ + \left((-0.64361) - (0.09418)\right)^2\end{matrix}} = 1.21775$$

After calculating the distance using Equation (5), then proceed with ranking the results of the distance calculation to find the smallest to largest sequence. The Euclid distance ranking can be seen in Table 7 as follows.

Table 7 *Ranking of Euclidean Distance Results between First Testing Data and Training Data*

Rank	Data *Training*		First Testing Data	Limit K-NN
	Sample	Classification	d	
1	212	Good nutritional	0.11713	1-NN
2	465	Good nutritional	0.16014	-
3	166	Good nutritional	0.1915	3-NN
4	227	Good nutritional	0.21235	-
5	214	Good nutritional	0.22481	5-NN
6	184	Good nutritional	0.23721	-
7	180	Good nutritional	0.25974	7-NN
8	108	Good nutritional	0.26781	-
9	33	Good nutritional	0.27061	9-NN
⋮	⋮	⋮	⋮	⋮
478	62	Obesity	5.19835	-

After getting the prediction results using the K-NN method, then evaluate the classification results of the K-NN method using the confusion matrix, accuracy calculations and Press's Q calculations.

The results of the confusion matrix using the K-NN method can be seen in Table 8 as follows:

[DOI: 10.14710/JSUNIMUS.13.1.2025.1-13]

Table 8 Confusion Matrix of K-NN Method Classification Results

Initial Classification of Toddler Nutritional Status	Classification Prediction K-NN Method						Total
	Good nutritional	Poor nutritional	Malnutrition	Excess nutrition	Obesity	Risk of overnutrition	
Good nutritional	44	0	2*	0	0	1*	47
Poor nutritional	0	0	0	0	0	0	0
Malnutrition	1*	0	0	0	0	0	1
Excess nutrition	2*	0	0	0	0	1*	3
Obesity	0	0	0	0	0	0	0
Risk of overnutrition	0	0	0	0	0	2	2
Total	47	0	2	0	0	4	53

So based on Equation (7), the following accuracy values are obtained:

$$\text{Accuracy} = \frac{44+0+0+0+0+2}{53} \times 100\%$$
$$= \frac{46}{53} \times 100\%$$
$$= 86.79\%$$

Next, to evaluate the accuracy of the prediction (classification) results, the Press's Q value is used. The Press's Q value is calculated using the confusion matrix in Table 8 and Equation (8) to obtain the following results.

Hypothesis testing :

H_0 : Inaccurate classification

H_1 : Accurate classification

Test statistic : Press's Q

Test criteria: reject H_0 if Press's Q > $\chi^2_{\alpha,1}$.

Calculation :

$$Press's\ Q = \frac{[53 - ((46)6)]^2}{53(6-1)}$$
$$= 187.65$$

With degrees of freedom worth one and a confidence level of $\alpha = 0.05$, and the value $\chi^2_{(0.05:1)} = 3.841 < Press's\ Q = 187.65$. So it can be concluded that the classification using the K-NN method rejects H_0 or the classification using the K-NN method is accurate.

4.7 Comparing classification results

Comparing the classification results. After obtaining the results of measuring the level of accuracy using the K-NN method and the Naïve Bayes method, then the accuracy values are compared and the largest value is taken to be used as a better method for classifying toddler nutritional status data at the Baqa Samarinda Seberang Community Health Center in 2022 Comparison of the accuracy values of the two algorithms can be seen in Table 9 as follows

[DOI: 10.14710/JSUNIMUS.13.1.2025.1-13]

Tabel 9 Comparison of Accuracy Levels and Press's Q

Method	Accuracy	*Press's Q*
naïve Bayes	82.15%	168
K-Nearest Neighbor	90.57%	187.65

5. CONCLUSION

Based on the results of data analysis and discussion, it is concluded that the results of measuring the accuracy of the classification of nutritional status of toddlers at the Baqa Samarinda Seberang Health Center in 2022 using the naïve Bayes method with a proportion of 90% training data and 10% testing data obtained the accuracy of the classification in predicting the classification of 82.15% and the Press's Q value of 168. Meanwhile, the K-NN method obtained the accuracy of classification in predicting the classification of 90.57% and the Press's Q value of 187.65. This shows that the K-Neirest Neighbor method works better than the naïve Bayes method in classifying the nutritional status of toddlers at the Baqa Samarinda Seberang Health Center in 2022 seen from the higher accuracy and Press'S Q values.

REFERENCES

[1] Han, J., Kamber, M., & Pei, J. (2012). *Data Mining: Concept and Techniques, Third Edition*. Waltham: Morgan Kaufmann Publishers.

[2] Prasetyo, E. (2012) . *Data Mining: Konsep dan Aplikasi Menggunakan Matlab*. Yogyakarta: Andi Offset.

[3] Rahmaulidyah, F.N., Hayati, M.N., & Goejantoro, R. (2021). Perbandingan Metode Klasifikasi *naïve* Bayes dan *K-Neirest Neighbor* Pada Data Status Pembayaran Pajak Pertambahan Nilai di Kantor Pelayanan Pajak Pratama Samarinda Ulu. *Jurnal Eksponensial, 12(2*), 161-164.

[4] Mustaghfiroh, L., Ariani, M. H., & Bijanto. (2022) . Klasifikasi Pasien Covid-19 di Indonesia menggunakan Metode *K-Nearest Neighbor*. *Jurnal AMRI*, *1*(1), 16-21.

[5] Saeroni, A., Hayati, M.N., & Goejantoro, R. (2020). Klasifikasi Tingkat Kelancaran Nasabah Dalam Membayar Premi dengan Menggunakan Metode *K-Nearest Neighbor* dan Analisis Diskriminan Fisher (Studi kasus: Data Nasabah PT. Prudential Life Samarinda Tahun 2019). *Jurnal Statistika Universitas Muhammadiyah Semarang, 8(2)*, 88-94.

[6] Bustami. (2014). Penerapan Algoritma Naïve Bayes untuk Mengklasifikasikan Data Nasabah Asuransi. *Jurnal Penelitian Teknik Informatika*, *8*(1), 129-132.

[7] Tan, P., Steinbach, M., & Kumar, V. (2006). *Introduction to Data Mining*. Boston: Pearson Education.

[DOI: 10.14710/JSUNIMUS.13.1.2025.1-13]

[8] Prasetyo, E. (2014) . *Data Mining: Mengolah Data Menjadi Informasi Menggunakan Matlab*. Yogyakarta: Andi Offset.

[9] Simamora, B. (2005). *Analisis Multivariat Pemasaran*. Jakarta: PT. Gramedia Pustaka Utama

[10] Hair, J., Black, W., Babin, B., Anderson, R. & Tatham, R. (2006) . *Multivariate Data Analysis 5th Edition* . New Jersey: Pearson Prentice Hall.

[11] Rofiq, A., Wuryandari, T., & Rahmawati, R. (2016). Perbandingan Analisis Diskriminan Fisher dan *naïve* Bayes Untuk Klasifikasi Risiko Kredit (Studi Kasus Debitur di Koperasi Jateng Amanah Mandiri Cabang Sukorejo Kendal). *Jurnal Gaussian, 5(1),* 1-10.

[DOI: 10.14710/JSUNIMUS.13.1.2025.1-13]