Determination of Nutritional Status Using Classification Method Datamining Using K- Nearst Neighbord (KNN) Algorithm

Muhamad Fatchan[1], Muhamad Taufik Akbar[2], Wahyu Hadikristanto[3] , Andri Firmansyah[4]
Universitas Pelita Bangsa, Cikarang

fatchan@pelitabangsa.ac.id[1], muhamadtaufikakbar18@gmail.com[2],
wahyu.hadikristanto@pelitabangsa.ac.id[3], andrifirmansyah@pelitabangsa.ac.id[4]

Abstract. The Determination of nutritional status aims to determine the nutritional status of children, in health centers in the district of Bojonggambir, the parameters commonly used in determining the nutritional status of children based on body weight according to age (WA/ A), IIn determining the nutritional status of children often miscalculates due to several factors including, the psychological factors of nutritionists due to the large number of cases handled and the limited number of human resources in addition to the number of posyandu managed by the health center in Bojongambir quite a number of thousands of children each year, the purpose of this study is to determine patterns in the process of determining nutritional status and overcome the risk of errors in calculations performed by nutritionists, the method used in this study is the classification of data mining with the k nearst neighbord algorithm, the results of the k nearst neighbord algorithm calculation with the euncledean distance formula against 1001 children's data with value k is determined k = 3, k = 5, k = 7, k = 9, resulting in an average accuracy value of 96.88%, precision 98.59%, recall 95.25% and AUC 0.989 based on these results, it can be concluded the process of determining the nutritional status using an algorithm k nearst neighbord can be applied with very good accuracy results.

Keywords: Nutritional Status, Classification, K-Neart Neigbord, Euclidean Distance

1 Introduction

Nutritional status is a form of the body that results from the pattern of consumption of food and the use of nutrients where nutrients are needed by the body as an energy source as well as regulating body sources, Nutritional status can be determined through laboratory and anthropometric tests, Anthropometric measurements are measurements used to determine the nutritional state of children as a guideline. z-score is an anthropometric index that is used internationally to determine nutritional status and growth, which is expressed as a standard deviation unit (SD) of population z-score is used to calculate anthropometric nutritional status of body weight for age (weight / age), height to age (TB / U), body weight to height [1]

Children's health status is important in a country to prepare a healthy generation, malnutrition is a burden for the family and also the burden of the state. the government through the community health center (PUSKESMAS) provides counseling and recording Toddler nutritional needs, the results of recording can not provide good decisions because they can not process status data nutritional needs automatically, if there is a request for nutritional status data for children under five, then the mapping of the data cannot be done in a manner

INCEESS 2020, July 17-18, Bekasi, Indonesia
Copyright © 2021 EAI
DOI 10.4108/eai.17-7-2020.2302994

fast because the process is still manual so it is less optimal and it is very possible that data duplication occurs.

K-NN is one of the datamining classification algorithms for classifying new objects based on (K). The KNN Classification Method is widely used in machine learning because it is simple to implement [1].efficient and effective use of the KNN algorithm for pattern recognition and KNN classification algorithm is widely used in applications 1) classification and interpretation, 2) Problem solving, 3) Training function, shortcomings of KNN 1) Low Efficiency 2) Depends on good value selection for k [2]. K-NN is a classification algorithm with reliable scalability, to optimize K relies heavily on data [3]. KNN classification is a datamining algorithm that can classify data that is not balanced and one of the datamining algorithms chosen as the top 10 algorithms [4] [5].

2 Methodology

After business objectives and project plans are set, the next steps are to conduct initial data collection, data description, and exploration. This research uses the recapitulation data of the Toddler Weighing Month (BPB) in 2018 and 2019, the data set with 1001 samples, has 9 predictor attributes and 1 class attribute that is used as a target, the target attribute class of good nutrition and malnutrition, with the builder of 511 data including nutrition and 490 data included in malnutrition,

At this stage the identification and construction of answers from the data that have been collected, in this study the data preparation includes all activities to build a data set that will be processed in the modeling process using the k-nearst neighboard algorithm, the data preparation process in this study includes, data selection, data cleaning, data integration and data normalization.

2.1 Preprocessing

In recent years, the data has become larger by the second number examples and number of features. But for very large data this is not practical in machine learning. This causes many problems in machine learning related to predictability and accuracy of predictions At this stage the selected data is data relating to the introduction of patterns of determining nutritional status, among others. z-score data, name data, gender data, age data, weight data, height data, median data, -1SD value data, 1SD value data and nutritional status data, this process includes cleaning and selection data.

2.2 Data Normalization

At the stage of preprocessing data, data transformation is done using the normalization method, where the normalized variable values are in the range 0-1. Normalization of numbers for each of these variables is needed before the process of calculating the distance value by the K-Nearest Neighboard algorithm so that there are no parameters that dominate the calculation of distances between data but still produce.

At the stage of preprocessing data, data transformation is done using the normalization method, where the normalized variable values are in the range 0-1. Normalization of numbers for each of these variables is needed before the process of calculating the distance value by the K-Nearest Neighboard algorithm so that there are no parameters that dominate the calculation

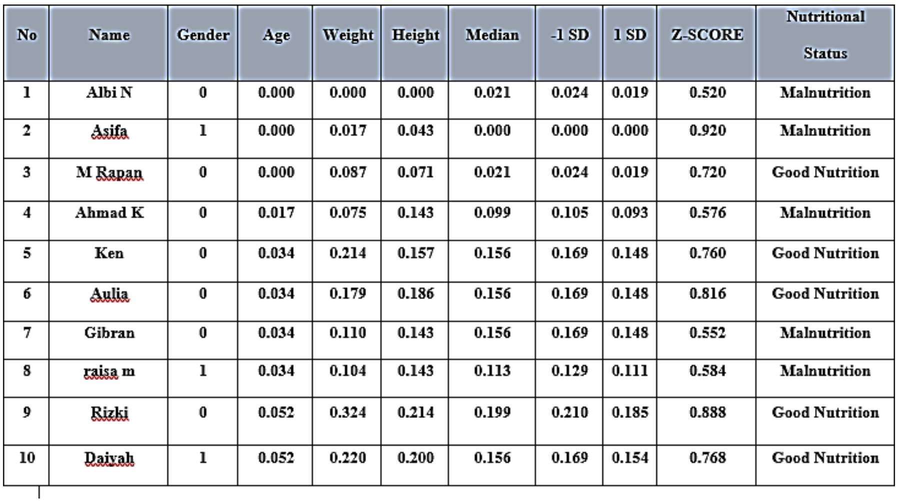
of distances between data but still produce the same analysis. feature selection is a method for reducing the complexity of data mining and machine learning, also part of KDD data preprocessing [6].

2.3 Clasification

Based on the data mining stages, for the K-Nearest Nighbord algorithm, the steps are as follows :

1) Determination of optimization of the K value used does not have standard rules, but in this study the k value used is k = 3, k = 5, k = 7 and k = 9.
2) Calculate the distance of training data and test data using eunclidean distance, in the manual calculation 10 training data samples are used and 1 sample of testing data is taken from the BPB data set which has been preprocessing data.

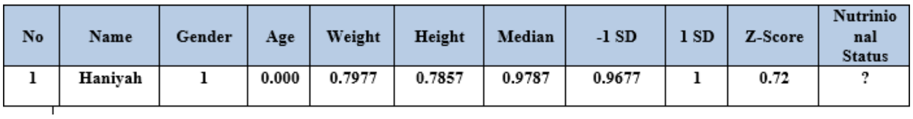
a) Training Data



No	Name	Gender	Age	Weight	Height	Median	-1 SD	1 SD	Z-SCORE	Nutritional Status
1	Albi N	0	0.000	0.000	0.000	0.021	0.024	0.019	0.520	Malnutrition
2	Asifa	1	0.000	0.017	0.043	0.000	0.000	0.000	0.920	Malnutrition
3	M Rapan	0	0.000	0.087	0.071	0.021	0.024	0.019	0.720	Good Nutrition
4	Ahmad K	0	0.017	0.075	0.143	0.099	0.105	0.093	0.576	Malnutrition
5	Ken	0	0.034	0.214	0.157	0.156	0.169	0.148	0.760	Good Nutrition
6	Aulia	0	0.034	0.179	0.186	0.156	0.169	0.148	0.816	Good Nutrition
7	Gibran	0	0.034	0.110	0.143	0.156	0.169	0.148	0.552	Malnutrition
8	raisa m	1	0.034	0.104	0.143	0.113	0.129	0.111	0.584	Malnutrition
9	Rizki	0	0.052	0.324	0.214	0.199	0.210	0.185	0.888	Good Nutrition
10	Daiyah	1	0.052	0.220	0.200	0.156	0.169	0.154	0.768	Good Nutrition

Fig.1. Training Data

b) Testing Data



No	Name	Gender	Age	Weight	Height	Median	-1 SD	1 SD	Z-Score	Nutrinional Status
1	Haniyah	1	0.000	0.7977	0.7857	0.9787	0.9677	1	0.72	?

Fig.2. Testing Data

c) Eunclidean Distance Calculation

$$d_1 = \sqrt{(0-0)^2 + (0-0)^2(0.7977-0)^2 + (0.7875-0)^2 + (0.9787-0.021)^2 + (0.9677-0.024)^2 + (1-0.019)^2 + (0.72-0.520)^2}$$

$$= 6.0639$$

$$d_2 = \sqrt{(0-1)^2 + (0-0)^2(0.7977-0.017)^2 + (0.7875-0.043)^2 + (0.9787-0)^2 + (0.9677-0)^2 + (1-0)^2 + (0.72-0.920)^2}$$

$$= 5.0953$$

$$d_3 = \sqrt{(0-0)^2 + (0-0)^2(0.7977-0.087)^2 + (0.7875-0.071)^2 + (0.9787-0.021)^2 + (0.9677-0.024)^2 + (1-0.019)^2 + (0.72-0.720)^2}$$

$$= 5.7860$$

$$d_4 = \sqrt{(0-0)^2 + (0-0.017)^2(0.7977-0.075)^2 + (0.7875-0.143)^2 + (0.9787-0.099)^2 + (0.9677-0.105)^2 + (1-0.093)^2 + (0.72-0.576)^2}$$

$$= 5.2633$$

$$d_5 = \sqrt{(0-0)^2 + (0-0.034)^2(0.7977-0.214)^2 + (0.7875-0.157)^2 + (0.9787-0.156)^2 + (0.9677-0.169)^2 + (1-0.148)^2 + (0.72-0.760)^2}$$

$$= 4.0797$$

$$d_6 = \sqrt{(0-0)^2 + (0-0.034)^2(0.7977-0.179)^2 + (0.7875-0.186)^2 + (0.9787-0.156)^2 + (0.9677-0.169)^2 + (1-0.148)^2 + (0.72-0.816)^2}$$

$$= 4.7239$$

$$d_7 = \sqrt{(0-0)^2 + (0-0.034)^2(0.7977-0.110)^2 + (0.7875-0.143)^2 + (0.9787-0.156)^2 + (0.9677-0.169)^2 + (1-0.148)^2 + (0.72-0.552)^2}$$

$$= 4.8868$$

$$d_8 = \sqrt{0 - 1)^2 + (0 - 0.034)^2(0.7977 - 0.104)^2 + (0.7875 - 0.143)^2 + (0.9787 - 0.113)^2 + (0.9677 - 0.129)^2 + (1 - 0.111)^2 + (0.72 - 0.584)^2}$$

$$= 4.0873$$

$$d_9 = \sqrt{(0 - 0)^2 + (0 - 0.052)^2(0.7977 - 0.324)^2 + (0.7875 - 0.214)^2 + (0.9787 - 0.199)^2 + (0.9677 - 0.210)^2 + (1 - 0.185)^2 + (0.72 - 0.888)^2}$$

$$= 4.3259$$

$$d_{10} = \sqrt{(0 - 1)^2 + (0 - 0.052)^2(0.7977 - 0.220)^2 + (0.7875 - 0.200)^2 + (0.9787 - 0.156)^2 + (0.9677 - 0.169)^2 + (1 - 0.154)^2 + (0.72 - 0.768)^2}$$

$$= 3.6081$$

3 Result and Discussion

After analysing and testing to determine the nutritional status of the Toddler Weighing Month (BPB) recapitulation data, the Bojonggambir sub-district of Tasikmalaya district, the results achieved by researchers were to find out how much the accuracy of the k-nearst neighbor algorithm in classifying the BPB dataset based on test results, obtained accuracy, precision, recall and AUC are shown in table below.

Table 1. Test Result

K Value	Aaccuracy	Precision	Recall	AUC
3	97.50%	98.63%	96.47%	0.984
5	97.10%	98.78%	95.49%	0.990
7	96.80%	98.77%	94.91%	0.989
9	96.10%	98.18%	94.12%	0.994

Based on the table above, the average accuracy of 96.88%, 98.59% precision, 95.25% memory and 95,25% and AUC of 0.989 can be taken, then it can be balanced using the k-near neighbor application to determine nutritional status in.

References

[1] S. Zhang, X. Li, M. Zong, X. Zhu, and D. Cheng, “Learning k for kNN Classification,” *ACM Trans. Intell. Syst. Technol.*, vol. 8, no. 3, 2017, doi: 10.1145/2990508.

[2] M. A. jabbar, B. L. Deekshatulu, and P. Chandra, “Classification of Heart Disease Using K-

Nearest Neighbor and Genetic Algorithm," *Procedia Technol.*, vol. 10, pp. 85–94, 2013, doi: 10.1016/j.protcy.2013.12.340.

[3] S. Yang, H. Jian, Z. Ding, Z. Hongyuan, and C. L. Giles, "IKNN: Informative K- nearest neighbor pattern classification," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 4702 LNAI, pp. 248–264, 2007, doi: 10.1007/978-3-540-74976-9_25.

[4] X. Wu *et al.*, *Top 10 algorithms in data mining*, vol. 14, no. 1. 2008.

[5] S. Zhang, X. Li, M. Zong, X. Zhu, and R. Wang, "Efficient kNN classification with different numbers of nearest neighbors," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 29, no. 5, pp. 1774–1785, 2018, doi: 10.1109/TNNLS.2017.2673241.

[6] S. Zhang, "Cost-sensitive KNN classification," *Neurocomputing*, vol. 391, no. xxxx, pp. 234–242, 2020, doi: 10.1016/j.neucom.2018.11.101.