HASIL CEK_60020388_Identification of Toddlers' Nutritional Status using Data Mining Approach

by Sri Winiarti 60020388

Submission date: 08-Mar-2020 02:43AM (UTC+0700)
Submission ID: 1271191109
File name: Paper_22-Identification_of_Toddlers_Nutritional_Status.pdf (521.8K)
Word count: 3618
Character count: 18471

Identification of Toddlers' Nutritional Status using Data Mining Approach

Sri Winiarti, Herman Yuliansyah, Aprial Andi Purnama
Department of Informatics,
Universitas Ahmad Dahlan,
Yogyakarta, Indonesia

***Abstract*—One of the problems in community health center or health clinic is documenting the toddlers' data. The numbers of malnutrition cases in developing country are quite high. If the problem of malnutrition is not resolved, it can disrupt the country's economic development. This study identifies malnutrition status of toddlers based on the context data from community health center (PUSKESMAS) in Jogjakarta, Indonesia. Currently, the patients' data cannot directly map into appropriate groups of toddlers' malnutrition status. Therefore, data mining concept with k-means clustering is used to map the data into several malnutrition status categories. The aim of this study is building software that can be used to assist the Indonesian government in making decisions to take preventive action against malnutrition.**

***Keywords*—Data mining; k-means clustering; malnutrition status of toddler**

I. Introduction

Data mining is a process of extracting large amounts of data to know the data pattern. Some topics in data mining are association rule mining, data clustering and data classification. Association rule mining is data mining techniques for finding associative rules between combinations of items. Several studies apply the association rule mining is to identify the risk factors of early childhood caries [1], to determine the pattern feedback of data alumni tracer study at the university [2] and to visualisation of financial Arabic text [3]. Some studies propose clustering method to solve the problems in their research for example a basic health screening system using Bayesian methods [4], detection of heart disease using decision tree methods [5], and to clasify of Alzheimer Disease using K-Nearest Neighbors (KNN) [6]. Several studies also implement the clustering method to perform automatic color segmentation [7], to perform clustering and analysis of earth-quake epicenter [8], and to decrease the load of computation in high dimensional data [9].

Health and nutritional status of children is one of the measure that reflects the public nutrition situation. Malnutrition is not only a burden to the family, but also a burden for the country. Therefore, Indonesian government through the community health center (PUSKESMAS) has conducted data collection of toddlers' nutritional status by using Excel based application. However, the results cannot show the data grouping of nutritional status automatically. The data that available in PUSKESMAS still not able to determine the nutritional status of toddler, according to the standards set by the Indonesian government. When there is a demand for data related to the community's nutritional status, then the mapping process is done manually. This process becomes not optimal as it will require a long process and can occur duplication of data if thousands of existing data are processed manually.

Previous researches have studied malnutrition in elderly, mothers and toddlers [10]-[12] and Child Care Health Consultation [13]. Malnutrition is the cause and consequence of many geriatric diseases that cause a very significant proportion of state expenditure on health [14]. In [15], author analyzes malnutrition using logistic regression methods and growth charts to reduce the number of children with malnutrition status. This study aims to optimize the data transactions of under five years patients who have malnutrition. The malnutrition patients are grouped according to the nutritional value of children under five years using data mining method with k-means clustering algorithm. Data mining approach is used in this research because data mining are widely used in predicting the various procedures and validity of data. In addition, data mining can improve decision making by finding patterns and trends in complex data [16].

K-means clustering algorithm is also widely implemented in medical science field such as applying k-means clustering to analyze identification of individual characteristics using brainwave signal [17], to identify new candidate drug compounds that have relation with lung cancer drugs [18], to make recommendation of antiarrhythmic drugs [19], and extraction cancer signatures [20]. The other studies are clustering medical data to find direction and effectiveness of the research work [21], enhance cancer subtype prediction [22], color-converted segmentation algorithm for magnetic resonance imaging (MRI) brain images [23] and EEG analysis to detect drowsy driving [24].

Based on the literature review, it is important to continue the research collaboration between data mining and medical science field. The data used in this study refer to nutrition report data from PUSKESMAS Umbulharjo Yogyakarta in 2016. Specification of toddlers' data used in this research is 6 months to 72 months old infants. Parameters that used for the grouping of nutritional status of toddlers namely; height, weight and age.

This research is expects that the PUSKESMAS can access data and data to monitor the nutritional status of children in

This research is supported by Ministry of Research, Technology and Higher Education in the research scheme Higher Education Research Cooperation (Penelitian Kerjasama Antar Perguruan Tinggi/PKAPT) grant number No: 118/SP2H/LT/DRPM/IV/2017 and PEKERTI-058/SP3/LPP-UAD/IV/2017 on 17 April 2017.

every region easily and quickly. This study aims to determine and develop a software that can be used by PUSKESMAS to identify the nutritional status of toddlers using data mining approach to be analyzed in the decision-making process.

II. METHODOLOGY

This research studies the data mapping of malnutrition patients of children under five using data mining approach. The grouping technique uses k-means clustering. The k-means clustering algorithm is the simplest and most common algorithm used to group objects by attribute/feature into k number of clusters, where k is a positive integer and defined by the user. Grouping is done by minimizing the sum of squares distance between the data and the appropriate centroid cluster. The procedure of k-means clustering is shown in Fig. 1 [25].

As shown in Fig. 1, the procedure of k-means clustering can be explained as follows:

- Step 1. Begin by defining k = number of clusters.
- Step 2. Enter each initial partition that classifies the data into the cluster k. It can be done by randomly sampling the data, or systematically as follows: Take the first training data sample k as a single element cluster. Each of the remaining training samples (N-k) collect on the cluster with the nearest centroid. When finished, recompute centroid from the newly acquired cluster.
- Step 3. Perform each sample in a sequence and calculate the distance from the centroid center of each group. If a sample is currently incompatible with the cluster closest to the centroid, replace the sample in this cluster and update the centroid point with the new sample and the sample loss cluster.

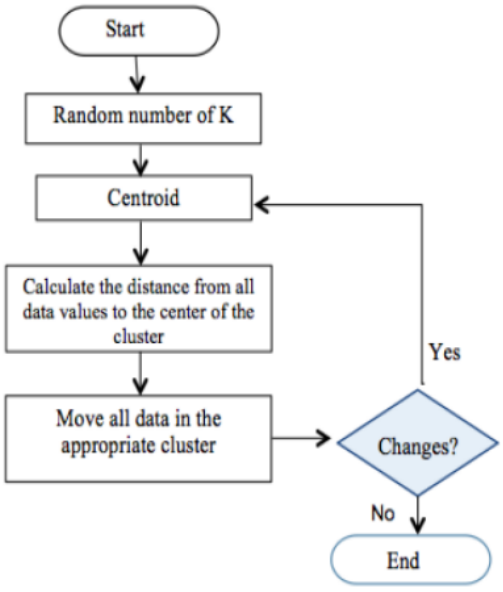


Fig. 1. The procedure of k-means clustering algorithm.

- Step 4. Repeat Step 3 until the target value is reached, i.e. until the training sample matches and there is no new task. If the amount of data is less than the number of clusters, then assign each data as the centroid of the cluster. Each centroid will have a number of clusters. If the amount of data is greater than the number of clusters, for each data, calculate the Distance to all centroids and get the minimum distance. This data is said to belong to a cluster that has a minimum distance value of this data. If you are not sure about the centroid location, you need to be centroid based on your current location. Then set all data to this new centroid. This process is repeated until no data is moved to another cluster again. The k-means algorithm works by using (1).

$$arg_s\ min \sum_{i=1}^{k} \sum_{X_j \ni S_i} || X_j \quad -\mu_i \ ||^2 \qquad (1)$$

Information:

- (X1, X2 ... Xn): the observation results represent a cluster element with a real d dimensional vector.
- n: Number of observations where the observed value to k set (k <= n) S = {S1, S2, ... Sk).
- μi: the mean value of the point at Si.

III. RESULT AND DISCUSSION

Why use data mining concept? Because the concept of data mining can analyze and classify the database so that every organization can make decisions based on this classification and can improve their plan in the future. There are many data mining techniques available where we can detect hidden patterns in the database [26].

Referring to the data mining stage in Fig. 1, for the case of nutrition status identification with k-means clustering algorithm, the procedure begins by obtaining patient data from patient's medical record database. Table I shows the patient data with parameters of body height, weight and age of children under five.

TABLE. I. TODDLER DATA BASED ON PUSKESMAS LOCATION

No	Age	Weight	Height
1	42	12.7	91
2	41	12.8	94
3	39	16.8	98
4	33	13.4	94
5	24	10.8	85
6	24	10.3	103
7	24	10.3	103
8	48	16.3	104
9	45	15.7	100
10	44	26.5	104

After the data are loaded, the initial centroids are determined according to 5 groups of toddler's nutritional status, that is Bad with value 0.96, Medium with value 0.73, Good with value 0,73, Over with value 0,355 and Obesity with value 0,04. The data is normalized using the normalization equation.

Normalized value = (initial value - minimum value) / (max value - minimum value) (2)

As shown in Table I, the High Body data have minimum value 85 and maximum value 104, the Weight data have minimum value 10.3 and maximum value 26.5, while the Age data have minimum value 24 and maximum value 48. The data normalization is shown in Table II.

If the obtained data are not consistent, they will change the data centroid through the iteration process. The iteration process will stop if the new ratio value is less than the ratio value in the previous iteration. If the condition has not been achieved, the iteration process will be repeated. The iteration result from the normalization of toddlers' data are shown in Table III.

TABLE. II. NORMALIZATION OF TODDLER DATA

No	Age	Weight	Height	Means
1	0.75	0.148148148	0.315789474	0.404645874
2	0.708333333	0.154320988	0.473684211	0.445446177
3	0.625	0.401234568	0.684210526	0.570148365
4	0.375	0.191358025	0.473684211	0.346680745
5	0	0.030864198	0	0.010288066
6	0	0	0.947368421	0.315789474
7	0	0	0.947368421	0.315789474
8	1	0.37037037	1	0.790123457
9	0.875	0.333333333	0.789473684	0.665935673
10	0.833333333	1	1	0.944444444

TABLE. III. ITERATION RESULT I

No	Status				
	Bad	*Medium*	*Good*	*More*	*Obesity*
1	0.555354126	0.325354126	0.085354126	0.049645874	0.364645874
2	0.514553823	0.284553823	0.044553823	0.090446177	0.405446177
3	0.389851635	0.159851635	0.080148365	0.215148365	0.530148365
4	0.613319255	0.383319255	0.143319255	0.008319255	0.306680745
5	0.949711934	0.719711934	0.479711934	0.344711934	0.029711934
6	0.644210526	0.414210526	0.174210526	0.039210526	0.275789474
7	0.644210526	0.414210526	0.174210526	0.039210526	0.275789474
8	0.169876543	0.060123457	0.300123457	0.435123457	0.750123457
9	0.294064327	0.064064327	0.175935673	0.310935673	0.625935673
10	0.015555556	0.214444444	0.454444444	0.589444444	0.904444444

TABLE. IV. DISTANCE DATA ON THE FIRST ITERATION

Distance Data		
Membership	*Min Distance*	*Min Squared Distance*
More	0.049645874	0.002464713
Good	0.044553823	0.001985043
Good	0.080148365	0.00642376
More	0.008319255	6.921E-05
Obesity	0.029711934	0.000882799
More	0.039210526	0.001537465
More	0.039210526	0.001537465
Mediun	0.060123457	0.00361483
Medium	0.064064327	0.004104238
Bad	0.015555556	0.000241975
	Wcv	0.0228615

TABLE. V. CLUSTER CENTER DISTANCE DATA D

Cluster Center Distance d		
C1	C2	0.23
C1	C3	0.47
C1	C4	0.605
C1	C5	0.92
C2	C3	0.24
C2	C4	0.375
C3	C4	0.135
C3	C5	0.45
C4	C5	0.315
BCV		3.74

TABLE. VI. NEW CLUSTER CENTER DATA ON THE FIRST ITERATION

New Cluster Center				
Bad	*Medium*	*Good*	*More*	*Obesity*
			0.404645874	
		0.445446177		
		0.570148365		
			0.346680745	
				0.010288066
			0.315789474	
			0.315789474	
	0.790123457			
	0.665935673			
0.944444444				
0.944444444	0.728029565	0.507797271	0.345726392	0.010288066

The distance data on the first iteration are presented in Table IV. The center distance data d are presented in Table V and the new cluster center data are presented in Table VI.

By using equation Ratio = BCV / WCV, the ratio result is 163.594. When the ratio is compared with the previous ratio, the value of the new ratio is greater than the value of the previous ratio. Therefore, the iteration process is still continued. Fig. 2 and 3 show the interface of the developed software.

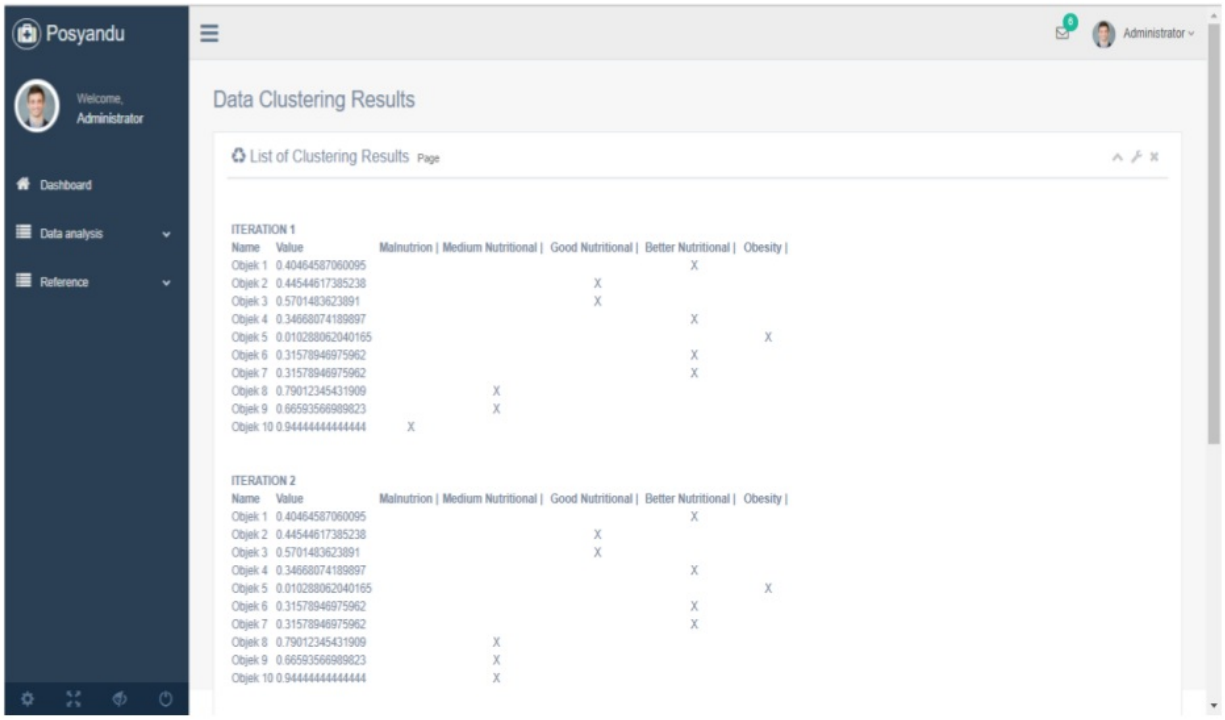
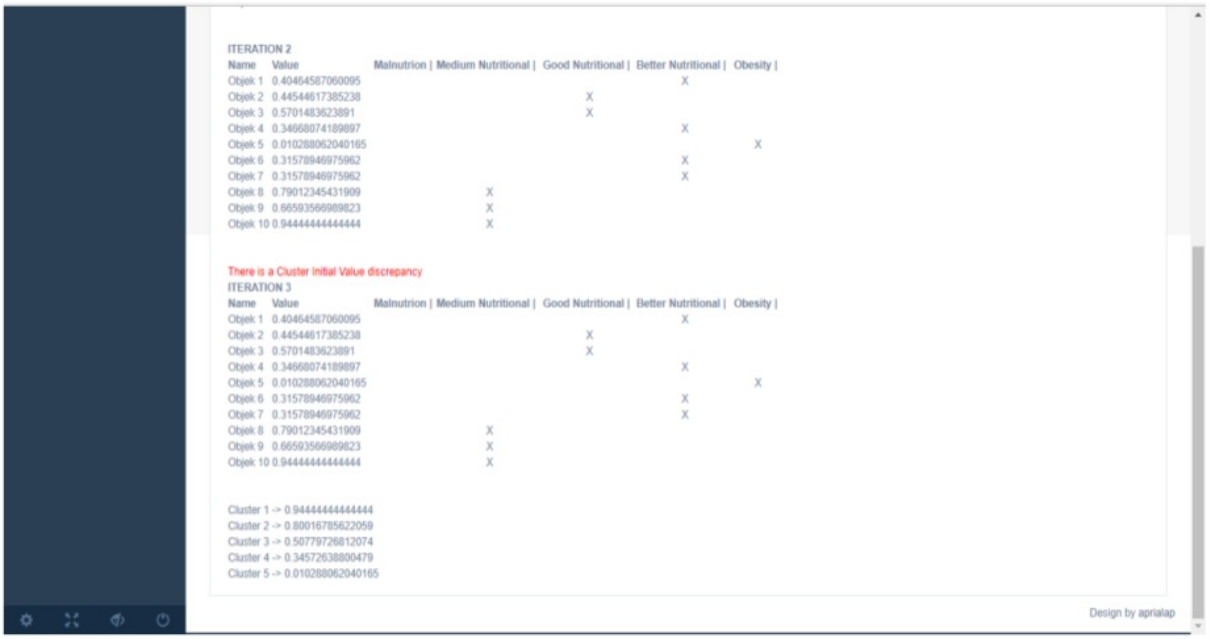


Fig. 2. Iteration process interface 1.



Iteration process interface 2.

IV. SYSTEM TEST

The system is tested using cross validation method by comparing the calculation result of k-means manually and with the result of developed system. Based on the patient data that shown in Table I, the system calculation result is presented in Table VII and the manually calculated result is presented in Table VIII.

TABLE. VII. RESULTS OF CALCULATIONS WITH THE SYSTEM DEVELOPED

Name	Status				
	Bad	*Medium*	*Good*	*More*	*Obesity*
Desta	0.555354126	0.325354126	0.085354126	0.049645874	0.364645874
Aura	0.514553823	0.284553823	0.044553823	0.090446177	0.405446177
Nazwa	0.389851635	0.159851635	0.080148365	0.215148365	0.530148365
Arisa	0.613319255	0.383319255	0.143319255	0.008319255	0.306680745
Evelyn	0.949711934	0.719711934	0.479711934	0.344711934	0.029711934
Amira	0.644210526	0.414210526	0.174210526	0.039210526	0.275789474
Farah	0.644210526	0.414210526	0.174210526	0.039210526	0.275789474
Hasna	0.169876543	0.060123457	0.300123457	0.435123457	0.750123457
Kania	0.294064327	0.064064327	0.175935673	0.310935673	0.625935673
Aira	0.015555556	0.214444444	0.454444444	0.589444444	0.904444444

TABLE. VIII. RESULTS OF CALCULATIONS MANUALLY

Name	Status				
	Bad	*Medium*	*Good*	*More*	*Obesity*
Desta	0.555354126	0.325354126	0.085354126	0.049645874	0.364645874
Aura	0.514553823	0.284553823	0.044553823	0.090446177	0.405446177
Nazwa	0.389851635	0.159851635	0.080148365	0.215148365	0.530148365
Arisa	0.613319255	0.383319255	0.143319255	0.008319255	0.306680745
Evelyn	0.949711934	0.719711934	0.479711934	0.344711934	0.029711934
Amira	0.644210526	0.414210526	0.174210526	0.039210526	0.275789474
Farah	0.644210526	0.414210526	0.174210526	0.039210526	0.275789474
Hasna	0.169876543	0.060123457	0.300123457	0.435123457	0.750123457
Kania	**0.274054321**	**0.054144253**	**0.155935553**	**0.210944673**	**0.525987673**
Aira	0.015555556	0.214444444	0.454444444	0.589444444	0.904444444

V. CONCLUSION

This research builds a software that can be used to identify the nutritional status of toddlers using data mining technique, with k-means clustering algorithm. The test is conducted by performing cross validation and gives 90% validation that the system can determine nutritional status of toddler by producing 5 clusters, namely, good nutrition, moderate nutrition, malnutrition, more nutrition and obesity.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for their comments and suggestions that helped to improve the quality and presentation of this paper. This research is supported by Ministry of Research, Technology and Higher Education in the research scheme Higher Education Research Cooperation (Penelitian Kerjasama Antar Perguruan Tinggi/PKAPT) grant No: 118/ SP2H/ LT/ DRPM/ IV/ 2017 and PEKERTI-058/ SP3/ LPP-UAD/ IV/ 2017 on 17 April 2017.

REFERENCES

[1] V. Ivančević, I. Tušek, J. Tušek, M. Knežević, S. Elheshk, and I. Luković, "Using association rule mining to identify risk factors for early childhood caries," Comput. Methods Programs Biomed., vol. 122, no. 2, pp. 175–181, 2015.

[2] H. Yuliansyah and L. Zahrotun, "Designing web-based data mining applications to analyze the association rules tracer study at university using a FOLD-growth method," Int. J. Adv. Comput. Res., vol. 6, no. 27, pp. 215–221, 2016.

[3] H. AL-Rubaiee, R. Qiu, and D. Li, "Visualising Arabic Sentiments and Association Rules in Financial Text," Int. J. Adv. Comput. Sci. Appl., vol. 8, no. 2, pp. 1–7, 2017.

[4] D. Phongphanich, N. Prommuang, and B. Chooprom, "Basic Health Screening by Exploiting Data Mining Techniques," Int. J. Adv. Comput. Sci. Appl., vol. 8, no. 9, pp. 79–85, 2017.

[5] A. Aziz and A. U. Rehman, "Detection of Cardiac Disease using Data Mining Classification Techniques," Int. J. Adv. Comput. Sci. Appl., vol. 8, no. 7, pp. 256–259, 2017.

[6] A. M. Taqi, F. Al-Azzo, and M. Milanova, "Classification of Alzheiemar Disease Based on Normalaized Hu Moment Invariants and Multiclassifiers," Int. J. Adv. Comput. Sci. Appl., vol. 8, no. 11, pp. 10–18, 2017.

[7] A. Prahara, I. T. R. Yanto, and T. Herawan, "Histogram Thresholding for Automatic Color Segmentation Based on k-means Clustering," in Recent Advances on Soft Computing and Data Mining: The Second International Conference on Soft Computing and Data Mining (SCDM-2016), Bandung, Indonesia, August 18-20, 2016 Proceedings, T. Herawan, R. Ghazali, N. M. Nawi, and M. M. Deris, Eds. Cham: Springer International Publishing, 2017, pp. 344–354.

[8] P. Novianti, D. Setyorini, and U. Rafflesia, "K-Means cluster analysis in earthquake epicenter clustering," Int. J. Adv. Intell. Informatics, vol. 3, no. 2, pp. 81–89, 2017.

[9] D. Ismi, S. Panchoo, and M. Murinto, "K-means clustering based filter feature selection on high dimensional data," Int. J. Adv. Intell. Informatics, vol. 2, no. 1, pp. 38–45, 2016.

[10] J. Studnicki, A. R. Hevner, D. J. Berndt, and S. L. Luther, "Comparing alternative methods for composing community peer groups: a data warehouse application.," J. public Heal. Manag. Pract., vol. 7, no. 6, pp. 87–95, 2001.

[11] D. Berndt, A. Hevner, and J. Studnicki, "Data warehouse dissemination strategies for community health assessments, informatik/informatique," J. Swiss Informatics Soc., vol. 1, pp. 27–33, 2001.

[12] S. Winiarti, S. Kusumadewi, I. Muhimmah, and H. Yuliansyah, "Determining the nutrition of patient based on food packaging product using fuzzy C means algorithm," in 2017 4th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI), 2017, no. September, pp. 1–6.

[13] R. Johnston, B. A. DelConte, L. Ungvary, R. Fiene, and S. S. Aronson, "Child Care Health Consultation Improves Infant and Toddler Care," J. Pediatr. Heal. Care, vol. 31, no. 6, pp. 684–694, Nov. 2017.

[14] S. N. Jang, S. I. Cho, J. Chang, K. Boo, H. G. Shin, H. Lee, and L. F. Berkman, "Employment status and depressive symptoms in Koreans: results from a baseline survey of the Korean Longitudinal Study of Aging.," J. Gerontol. B. Psychol. Sci. Soc. Sci., vol. 64, no. 5, p. 677, 2009.

[15] M. Ohyver, J. V. Moniaga, K. R. Yunidwi, and M. I. Setiawan, "Logistic Regression and Growth Charts to Determine Children Nutritional and Stunting Status: A Review," Procedia Comput. Sci., vol. 116, pp. 232–241, Jan. 2017.

[16] D. Thangamani and P. Sudha, "Identification of Various Deficiencies Using Data Mining Techniques – A Survey," Int. J. Sci. Res., vol. 3, no. 7, pp. 1270–1274, 2014.

[17] A. Azhari and L. Hernandez, "Brainwaves feature classification by applying K-Means clustering using single-sensor EEG," Int. J. Adv. Intell. Informatics, vol. 2, no. 3, pp. 167–173, 2016.

[18] J. Lu, L. Chen, J. Yin, T. Huang, Y. Bi, X. Kong, M. Zheng, and Y. D. Cai, "Identification of new candidate drugs for lung cancer using chemical-chemical interactions, chemical-protein interactions and a K-

means clustering algorithm," J. Biomol. Struct. Dyn., vol. 34, no. 4, pp. 906–917, 2016.

[19] J. Park, M. Kang, J. Hur, and K. Kang, "Recommendations for antiarrhythmic drugs based on latent semantic analysis with K-means clustering," 2016 38th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc., pp. 4423–4426, 2016.

[20] Z. Kakushadze and W. Yu, "*K-means and cluster models for cancer signatures," Biomol. Detect. Quantif., vol. 13, no. July, pp. 7–31, 2017.

[21] S. V and G. H. A, "Appraising Research Direction & Effectiveness of Existing Clustering Algorithm for Medical Data," Int. J. Adv. Comput. Sci. Appl., vol. 8, no. 3, pp. 343–351, 2017.

[22] N. Nidheesh, K. A. Abdul Nazeer, and P. M. Ameer, "An enhanced deterministic K-Means clustering algorithm for cancer subtype prediction from gene expression data," Comput. Biol. Med., vol. 91, pp. 213–221, Dec. 2017.

[23] L.-H. Juang and M.-N. Wu, "MRI brain lesion image detection based on color-converted K-means clustering segmentation," Measurement, vol. 43, no. 7, pp. 941–949, Aug. 2010.

[24] N. Gurudath and H. B. Riley, "Drowsy Driving Detection by EEG Analysis Using Wavelet Transform and K-means Clustering," Procedia Comput. Sci., vol. 34, pp. 400–409, Jan. 2014.

[25] S. Shinde and B. Tidke, "Improved K-means Algorithm for searching Research papers," Int. J. Comput. Sci. Commun. Networks, vol. 4, no. 6, pp. 197–202, 2014.

[26] J. Han, J. Pei, and M. Kamber, Data mining: concepts and techniques. Elsevier, 2011.

HASIL CEK_60020388_Identification of Toddlers' Nutritional Status using Data Mining Approach

ORIGINALITY REPORT

7%	7%	7%	6%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.thesai.org Internet Source	4%
2	Submitted to Universiti Teknologi Malaysia Student Paper	2%
3	ethesis.nitrkl.ac.in Internet Source	2%

Exclude quotes	On	Exclude matches	< 2%
Exclude bibliography	Off		