Application of Apriori Algorithm in Data Mining to Analyze Malnutrition (Case Study: Secanggang Village)

Indri Sulistianingsih[1*], Wirda Fitriani[2], , Darmeli Nasution[3]

[1,3] Department of Computer System, University of Pembangunan Panca Budi, Indonesia
[2]Department of Computer Engineering, University of Pembangunan Panca Budi, Indonesia

ABSTRACT

Malnutrition is a serious public health issue affecting quality of life in many rural areas. Comprehensive, evidence-based approaches are needed to identify dietary patterns associated with this condition. In this study, we applied the Apriori algorithm for data mining to analyze the relationship between dietary patterns and malnutrition prevalence in Secanggang Village. Using Apriori, we discovered significant associative patterns between foods and malnutrition. Our analysis showed several dietary patterns with sufficiently high support and confidence, indicating potentially strong associations with malnutrition. These patterns may include combinations of less diverse or nutritious foods, which could be risk factors for malnutrition. Our findings provide valuable insights for authorities and health institutions to formulate more effective intervention programs in tackling malnutrition in Secanggang Village. By understanding the relationship between dietary patterns and malnutrition, prevention and treatment efforts can be better targeted to positively impact community health and welfare. Further research and cross-sectoral collaboration are still needed to comprehensively address the complex challenges of malnutrition.

Keywords: Apriori, Data Mining, Malnutrition, Secanggang Village

Corresponding Author:

Indri Sulistianingsih,
Department of Computer System, University of Pembangunan Panca Budi,
St. Gatot Subroto Km. 4.5 Medan Sunggal, Medan 20122, North Sumatera, Indonesia
Email: indie@pancabudi.ac.id

1. INTRODUCTION

Malnutrition, also known as chronic undernutrition, is a condition of inadequate nutrient intake occurring due to insufficient or imbalanced food consumption over an extended period[1]. It can affect all age groups but is more prevalent among infants, young children, and adolescents in their growth and developmental phases. Malnutrition impedes physical and cognitive growth and development, leading to stunted growth, underweight, and cognitive/mental issues in children[2]. It causes susceptibility to diseases, weakened immunity, and poor academic performance[3].

The causes of malnutrition are diverse, including inadequate or imbalanced dietary intake, lack of access to nutritious foods, poor socioeconomic conditions, improper feeding practices, and other health issues[4]. Addressing malnutrition requires various strategies, including improving community nutrition and health, promoting healthy diets, and providing nutritional supplements to those in need. Malnutrition is diagnosed using anthropometric indicators like weight-for-height Z-scores <-3 SD. Malnutrition status in infants can adversely impact physical, mental and cognitive growth[5].

In tackling malnutrition, an effective, evidence-based approach is required to understand dietary patterns and contributing factors. Data mining, specifically the Apriori algorithm, is one viable approach. Apriori is a data mining technique used to identify association rules between items in large datasets[6], [7]. By applying Apriori to daily dietary and malnutrition data in Secanggang Village, we aim to discover significant associative dietary patterns related to malnutrition. This information can provide valuable insights to authorities, health institutions, and local communities in designing more targeted and effective interventions to improve nutritional status and public health in Secanggang Village. Through this research, we hope to make a positive contribution to scientific

knowledge and solving malnutrition at the local level[8]. The application of data mining in analyzing malnutrition can also further the development of health informatics[9], [10].

2. RESEARCH METHODOLOGY

Utilized a quantitative approach with data mining methodologies using the Apriori algorithm. This approach centers on collecting statistical data and objectively, systematically analyzing the data to test hypotheses and uncover patterns between the variables studied.

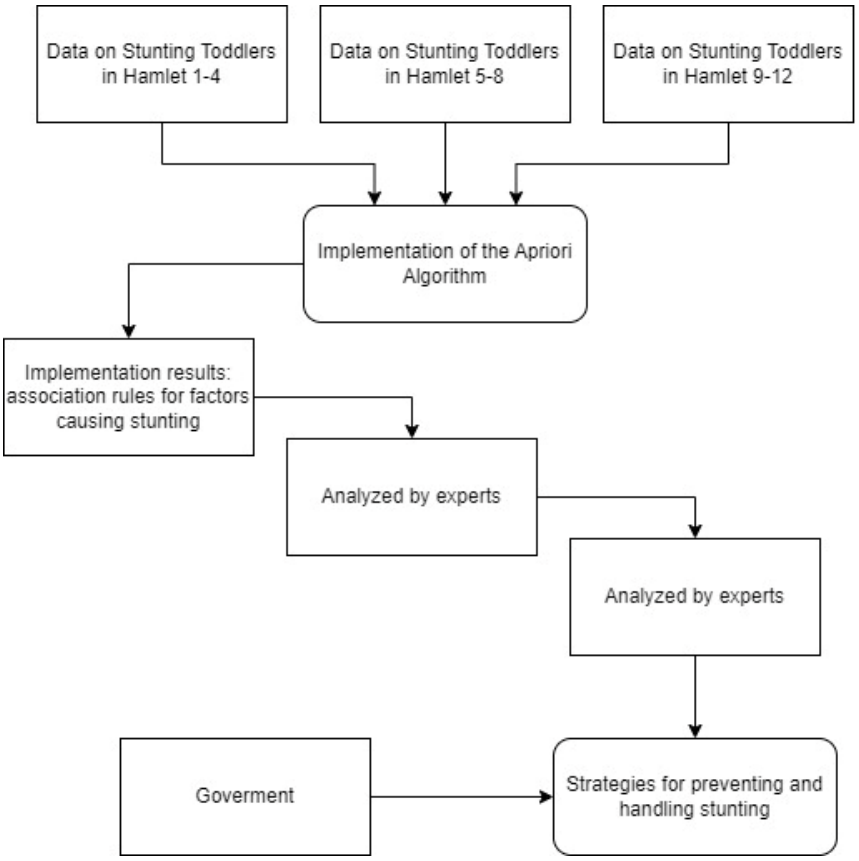

Figure 1. Research Approach

Data mining is an analytical technique for discovering relevant patterns and information in massive amounts of complex data. Data mining using the Apriori algorithm will be used in this study to examine malnutrition data from Secanggang Village. Apriori is well-suited for detecting associative patterns in transactional data[11].

This method was chosen because it enables for the identification of correlations between the numerous factors contributing to malnutrition instances among children in Secanggang Village. We can reveal associative patterns with data mining and Apriori that would be impossible to find with traditional statistical research. Data mining also makes it possible to process enormous datasets quickly and efficiently.

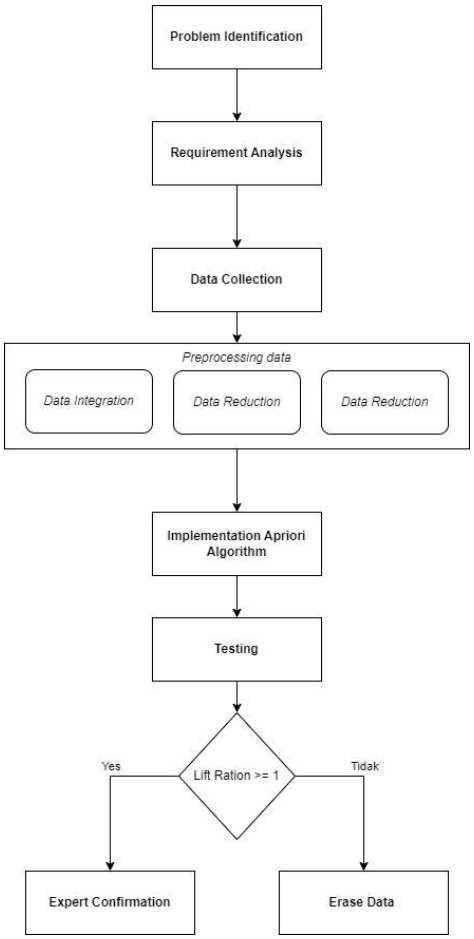

Figure 2. Research Stages

The use of data mining with Apriori is also consistent with our study goal of completely identifying the causes of malnutrition in order to inform well-targeted therapies[12]. We also included expert verification, which may evaluate the Apriori analysis and ensure the quality and value of the research findings.

2.1. Population & Sampel

The table below contains population and sample statistics for toddlers in Secanggang Village:

Table 1. Population and Sample Data for Toddlers in Secanggang Village in 2023

No	Category	Number of Toddler Population	Number of Samples
1	Short	59	10
2	Very Short	27	5
3	Not Stunting	281	50
	Total	367	65

a) The data in the table above are population data and a sample of toddlers from Secanggang Village.
b) There are 367 children under the age of five in Secanggang Village.
c) There are 59 toddlers in the "Short" category, 27 in the "Very Short" category, and 281 in the "Not Stunting" category of the overall population.
d) A total of 65 toddlers were sampled, with 10 toddlers falling into the "Short" category, 5 falling into the "Very Short" category, and 50 falling into the "Not Stunting" category.
e) To reflect the population of toddlers in Secanggang Village in the study, sampling was done at random and in a representative manner.

3. RESEARCH RESULTS

3.1. Design System

Stunting data entered into the system will be preprocessed. If the data is ready for use, enter the minimal support and confidence levels. Find the n-itemset first, then the process stage. Carry out the highest frequency pattern development. Iteration is performed up to n-itemset and association rules are formed[13]. The lift ratio will be used to test the results. The end outcome is a pattern of causes that produce stunting.

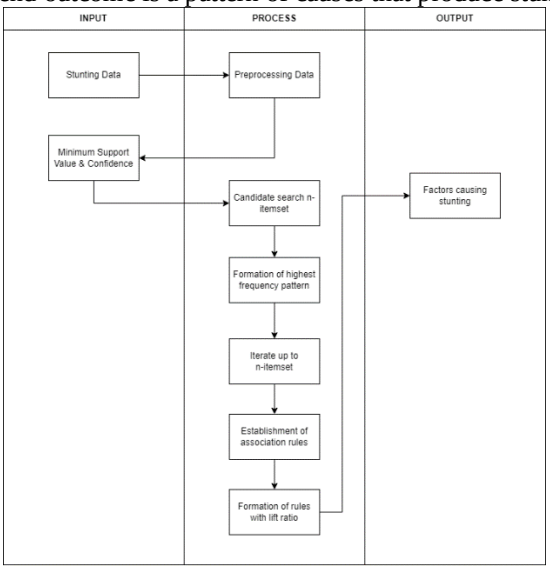

Figure 3. Design System

3.2. Preprocessing

The preprocessing of the stunting dataset from Secanggang Village involved several techniques to transform and prepare the data for effective mining using the Apriori algorithm[14]. Data integration was performed to combine data from multiple excel files into a single dataset. Data reduction techniques like dimensionality reduction were applied to remove unused columns and reduce the dataset size. Normalization and discretization were implemented to transform the data into suitable formats for mining. Variables were grouped into categorical ranges and actual values were replaced with numeric representations.

Table 2. Normalization

Variable	Grouping
Mother's Education	Low (Elementary School) Medium (Junior High) High (High School – College)
Nutrition Status	Bad Less Good
Exclusive breastfeeding (for 6 months)	Yes No
Birth Weight	Low Usual
Birth Length	Low Usual
Gender	Woman Man
Age	0-23 months 24-59 months
Birth Order	1-2 >= 3
Utilization of Integrated Healthcare Center	Never Not Routine Routine

The discretization procedure comes after the normalizing process. Actual toddler and parent data are substituted with predetermined value ranges. The original data is not available to the public. This manual calculation makes use of five toddlers as data. Table 2 shows the clustering of toddler data.

The data is then transformed back to numeric form so that it can be processed using an Apriori method. Toddlers who have or meet a variable are moved to position 1. Toddlers who do not have or meet a variable are renumbered as 0. Table 3 demonstrates this.

Table 3. Variable

A : Low Mother's Education	L : normal birth body length
B : Medium Mother's Education	M : female
C : High Mother's Education	N : male
D : poor nutritional status	O : Age 0-23 months
E : undernutrition status	P : Age 24-59 months
F : good nutritional status	Q : birth order 1-2
G : receive exclusive breast milk	R : birth order >= 3
H : does not accept exclusive breast milk	S : Never been to a Integrated Healthcare Center
I : low birth weight	T : Not regularly going to Integrated Healthcare Center
J : normal birth weight	U : Routine to Integrated Healthcare Center
K : Low birth body length	

Table 4. Toddler Stunting Data

Toddler	Mother's Education	Status Gizi	Breast Milk Exclusive	Birth Weight	Birth Length	Gender	Age	Birth Order	Utilization of Integrated Healthcare Center
X1	Tall	Good	Yes	Usual	Low	Man	0-23 months	1-2	Not routine
X2	Medium	Less	Yes	Low	Low	Woman	24-59 months	1-2	Never
X3	Low	Bad	Not	Low	Low	Man	24-59 months	>= 3	Not routine
X4	Medium	Less	Yes	Low	Low	Man	0-23 months	>= 3	Routine
X5	Medium	Bad	Yes	Low	Low	Woman	24-59 months	>= 3	Routine

Table 5. Discretization Process

Toddler	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
X1	0	1	0	0	1	0	0	1	1	0	1	0	0	1	1	0	0	1	0	0	1
X2	0	1	0	1	0	0	0	1	1	0	1	0	1	0	0	1	0	1	0	0	1
X3	0	1	0	0	1	0	0	1	1	0	1	0	1	0	0	1	1	0	1	0	0
X4	0	0	1	0	0	1	0	1	0	1	1	0	0	1	1	0	1	0	0	1	0
X5	1	0	0	1	0	0	1	0	1	0	1	0	0	1	0	1	0	1	0	1	0
Sum	1	3	1	2	2	1	1	4	4	1	5	0	2	3	2	3	2	3	1	2	2

3.2. Application of Apriori Algorithm

An Apriori algorithm will be used to implement the preprocessed stunting dataset. Figure 3 depicts the flow of the Apriori algorithm in greater detail. After entering the stunting dataset, the support value calculation method divides the number of each variable by the number of toddlers. The results of computing support values beyond the minimal limit are saved, and two itemsets are created. In the meantime, itemsets with support values that fall below the minimum threshold will be trimmed. The preceding step, namely determining the support value of the two items created, is repeated. Results that satisfy the minimal support value will be trimmed, while those

that do not meet the minimum support value will produce three itemsets. This method is repeated until no more combinations can be generated[15]. This manual computation yields a minimum support value of 0.6.

Table 6 shows the calculation of the support value for a single itemset. Data that meets the minimum support value is highlighted in green, while data that does not match the minimum support value is destroyed and will not be used in subsequent calculations. The formula for calculating the support value is as follows:

Tabel 6. One Itemset Support Calculation

Number	Factors from *Rules*	Sum	*Support*
1.	Low mother's education	1	0,2
2.	**Medium mother's education**	**3**	**0,6**
3.	High mother's education	1	0,2
4.	Poor nutritional status	2	0,4
5.	Intermediate nutritional status	2	0,4
6.	Good nutritional status	1	0,2
7.	Receiving exclusive Mother's Breastfeeding	1	0,2
8.	**Not receiving exclusive Mother's Breastfeeding**	**4**	**0,8**
9.	**Low birth weight**	**4**	**0,8**
10.	Normal birth weight	1	0,2
11.	**Low birth body length**	**5**	**1**
12.	Normal birth body length	0	0
13.	Gender: Female	2	0,4
14.	**Gender: Male**	**3**	**0,6**
15.	Age 0-23 months	2	0,4
16.	**Age 24-59 months**	**3**	**0,6**
17.	Birth order 1-2	2	0,4
18.	**Birth order >= 3**	**3**	**0,6**
19.	Never been to a Integrated Healthcare Center	1	0,2
20.	Not regularly going to Integrated Healthcare Center	2	0,4
21.	Routine to Integrated Healthcare Center	2	0,4

Seven itemsets of factors meet the support value. The support value is then calculated when two itemsets are generated from the factors that meet the support value. Factors that meet the minimum support value are highlighted in bold in Table 7.

Table 7. Two Itemset Support Calculation

Number	Factors from *Rules*	Sum	*Support*
1.	**Education of Medium mothers, not receiving exclusive breastfeeding**	**3**	**0,6**
2.	**Medium mother's education, low birth weight**	**3**	**0,6**
3.	**Medium mother's education, low birth length**	**3**	**0,6**
4.	Education of Medium mothers, boys	1	0,2
5.	Medium mother's education, aged 24-59 months	2	0,4
6.	Medium mother's education, birth order >= 3	2	0,4
7.	**Not receiving exclusive breast feeding, low birth weight**	**3**	**0,6**
8.	**Not receiving exclusive breast feeding, low birth length**	**4**	**0,8**
9.	Does not receive exclusive breastfeeding, males	2	0,4
10.	Not receiving exclusive breast feeding, age 24-59 months	2	0,4
11.	Not receiving exclusive breast feeding, birth sequence >= 3	2	0,4
12.	**Low birth weight, low birth length**	**4**	**0,8**
13.	Low birth weight, male	2	0,4
14.	**Low birth weight, age 24-59 months**	**3**	**0,6**
15.	**Low birth weight, birth order >= 3**	**3**	**0,6**
16.	**Low birth length, males**	**3**	**0,6**

17.	**Low birth length, age 24-59 months**	**3**	**0,6**
18.	**Low birth length, birth order >= 3**	**3**	**0,6**
19.	Male, age 24-59 months	1	0,2
20.	Male, birth order >= 3	2	0,4
21.	Age 24-59 months, birth order >= 3	2	0,4

There are 11 possible combinations that satisfy the support value. The components that satisfy the support value are then organized into three itemsets, and the support value is determined. The support value for the six factor combinations that meet the minimum support value is 0.6. Meanwhile, the support values for the other four factor combinations are 0.4 and 0.2. As a result, the following computation does not apply to this combination. Table 8 highlights the components that meet the minimal support value.

Table 8. Three Itemset Support Calculation

Number	*Factors from Rules*	*Sum*	*Support*
1.	**Medium mother's education, not receiving exclusive breastfeeding, low birth weight**	**3**	**0,6**
2.	**Education of Medium mothers, not receiving breast milk exclusive, low birth length**	**3**	**0,6**
3.	**Medium mother's education, low birth weight, low birth length**	**3**	**0,6**
4.	**Does not receive exclusive breastfeeding, low birth weight, low birth length**	**3**	**0,6**
5.	**Low birth weight, low birth length, age 24- 59 months**	**3**	**0,6**
6.	**Low birth weight, low birth length, sequence births >= 3**	**3**	**0,6**
7.	Low birth weight, age 24-59 months, birth order >= 3	2	0,4
8.	Low birth length, male, age 24-59 months	1	0,2
9.	Low birth length, male, birth order >= 3	2	0,4
10.	Low birth length, age 24-59 months, birth order >= 3	2	0,4

The elements that determine the support value are then divided into four itemsets, and the support value is calculated. Table 9 highlights characteristics that fulfill the minimal support value.

Table 9. Four Itemset Support Calculation

Number	*Factors from Rules*	*Sum*	*Support*
1.	**Medium mother's education, not receiving exclusive breast feeding, low birth weight, low birth length**	**3**	**0,6**
2.	**Low birth weight, low birth length, age 24-59 months, birth order >= 3**	**2**	**0,4**

When four itemsets were calculated, only one combination, 0.6, fulfilled the minimal support value. Because the combination could no longer be converted into 5 itemsets, the support computation was suspended. Table 10 contains calculations for factor combinations.

Table 10. Confidence and Lift Calculation

Number	*Factors from Rules*	*Support*	*Confidence*	*Lift*
1.	**Medium mother's education, not receiving exclusive breastfeeding**	**0,6**	**1**	**1,25**
2.	**Medium mother's education, birth weight low**	**0,6**	**1**	**1,25**
3.	Medium mother's education, low birth length	0,6	1	1
4.	Not receives exclusive breast milk, weight low birth	0,6	0,75	0,93
5.	Not receiving exclusive breast feeding, low birth length	0,8	1	1
6.	Low birth weight, low birth length	0,8	1	1
7.	Low birth weight, age 24-59 months	0,6	0,75	1,25
8.	Low birth weight, birth order >= 3	0,6	0,75	1,25
9.	Low birth length, males	0,6	0,6	1
10.	Low birth length, age 24-59 months	0,6	0,6	1
11.	Low birth length, birth order >= 3	0,6	0,6	1

12.	**Medium mother's education, not receiving exclusive breastfeeding, low birth weight**	**0,6**	**1**	**1,25**
13.	Medium mother's education, not receiving exclusive breastfeeding, low birth length	0,6	1	1
14.	Medium mother's education, low birth weight, low birth length	0,6	1	1
15.	Not receiving exclusive breast feeding, weight low birth, low birth length	0,6	1	1
16.	Low birth weight, low birth length, age 24-59 months	0,6	0,75	1,25
17.	Low birth weight, low birth length, birth order >= 3	0,6	0,75	1,25
18.	Medium mother's education, not receiving exclusive breastfeeding, low birth weight, low birth length	0,6	1	1

The following step is to generate association rules by calculating confidence and lift values. The minimum level of confidence is 1.0. Calculating the degree of confidence in a combination of Medium mother's education and not receiving exclusive lactation The method for gaining confidence.

$$Confidence = \frac{3}{3} = 1$$

Combining factors such as medium mother's education, not obtaining exclusive lactation, and first calculating benchmark confidence. Formula for benchmarking confidence.

$$Benchmark\ Confidence = \frac{4}{5} = 0{,}8$$

Calculating the lift, according to the formula.

$$Lift\ Ratio = \frac{1}{0{,}8} = 1{,}25$$

According to the above hand computation, there are three combinations of elements that meet the minimal values of support, confidence, and have a lift value greater than one. Table 11 summarizes the findings.

Table 11. Association Rules of Manual Calculation

Number	*Association Rules*
1.	Mother's education is medium, does not receive breast milk exclusively
2.	Mother's education is medium, low birth weight
3.	Mother's education is medium, does not receive breast milk exclusively, low birth weight

3.3. Expert Confirmation

If association rules are created, the results will be validated by experts using the Delphi technique. The accuracy value will be calculated after experts complete a questionnaire reviewing the findings of the association rules. Nutritionists from Secanggang Village hospitals, community health center nutritionists, and Integrated Healthcare Center (Posyandu) cadres are among the targets. This is due to their collective knowledge of stunting in Secanggang Village. The analysis will be used to compare the output outcomes.

3.4. Testing

Testing is done to determine whether or not the Apriori algorithm implementation software that was produced is working properly. The Python programming language was utilized to construct the software. For the code editor, the author used visual code.

4. CONCLUSION

This study highlights the ability of the Apriori algorithm in data mining to identify important associative patterns between dietary parameters and malnutrition prevalence. The results demonstrate Apriori may successfully discover combinations of foods, nutritional status, and other variables that correspond with established criteria for connections with malnutrition.

By weighting and prioritizing parameters based on user preferences, the system evaluates food patterns to propose the most probable risk factors. This data-driven strategy can assist authorities in devising targeted actions. The detected patterns such as low birth weight, poor nutrition, and short birth length have sufficiently high support and confidence, indicating possible strong associations with malnutrition in children.

While the research is limited to a small dataset, it highlights the capabilities of adopting data mining to unearth relevant insights from health data. The knowledge discovery process can be repeated for bigger volumes

of data across multiple fields. The findings can improve data-informed decision making to solve major public health concerns like malnutrition.

REFERENCES

[1] A. Alpin, "Hubungan Karakteristik Ibu dengan Status Gizi Buruk Balita di Wilayah Kerja Puskesmas Tawanga Kabupaten Konawe," *Nursing Care and Health Technology Journal (NCHAT)*, vol. 1, no. 2, pp. 87–93, 2021.

[2] A. Ayelign and T. Zerfu, "Household, dietary and healthcare factors predicting childhood stunting in Ethiopia," *Heliyon*, vol. 7, no. 4, p. e06733, Apr. 2021, doi: 10.1016/j.heliyon.2021.e06733.

[3] A. Ramdhani, H. Handayani, and A. Setiawan, "Hubungan Pengetahuan Ibu Dengan Kejadian Stunting," in *Prosiding Seminar Nasional LPPM UMP*, 2021, pp. 28–35.

[4] K. Komalasari, E. Supriati, R. Sanjaya, and H. Ifayanti, "Faktor-Faktor Penyebab Kejadian Stunting Pada Balita," *Majalah Kesehatan Indonesia*, vol. 1, no. 2, pp. 51–56, Oct. 2020, doi: 10.47679/makein.202010.

[5] D. P. Lestari, "Upaya pencegahan risiko gizi buruk pada balita: Literature Review," *Jurnal Ilmiah Universitas Batanghari Jambi*, vol. 22, no. 1, pp. 532–536, 2022.

[6] S. Wahyuni, I. Sulistianingsih, Hermansyah, E. Hariyanto, and O. Cindi Veronika Lumbanbatu, "Data Mining Prediksi Minat Customer Penjualan Handphone Dengan Algoritma Apriori," *JURNAL UNITEK*, vol. 14, no. 2, pp. 10–19, Dec. 2021, doi: 10.52072/unitek.v14i2.243.

[7] H. Xie, "Research and case analysis of apriori algorithm based on mining frequent item-sets," *Open J Soc Sci*, vol. 9, no. 04, p. 458, 2021.

[8] S. Supiyandi, M. Zen, C. Rizal, and M. Eka, "Perancangan Sistem Informasi Desa Tomuan Holbung Menggunakan Metode Waterfall," *JURIKOM (Jurnal Riset Komputer)*, vol. 9, no. 2, pp. 274–280, 2022.

[9] B. S. dos Santos, M. T. A. Steiner, A. T. Fenerich, and R. H. P. Lima, "Data mining and machine learning techniques applied to public health problems: A bibliometric analysis from 2009 to 2018," *Comput Ind Eng*, vol. 138, p. 106120, 2019.

[10] M. Sornalakshmi *et al.*, "An efficient apriori algorithm for frequent pattern mining using mapreduce in healthcare data," *Bulletin of Electrical Engineering and Informatics*, vol. 10, no. 1, pp. 390–403, 2021.

[11] C. Wang and X. Zheng, "Application of improved time series Apriori algorithm by frequent itemsets in association rule data mining based on temporal constraint," *Evol Intell*, vol. 13, no. 1, pp. 39–49, 2020.

[12] P.-C. Hsieh *et al.*, "Combination of acupoints in treating patients with chronic obstructive pulmonary disease: an apriori algorithm-based association rule analysis," *Evidence-Based Complementary and Alternative Medicine*, vol. 2020, 2020.

[13] N. Mayasari, "Implementasi Data Mining untuk Memprediksi Itemset Promosi Penjualan Pada CV. Sumber Segar Utama," *Jurnal Teknik dan Informatika*, vol. 6, no. 1, pp. 31–36, 2019.

[14] F. Lv, "Data Preprocessing and Apriori Algorithm Improvement in Medical Data Mining," in *2021 6th International Conference on Communication and Electronics Systems (ICCES)*, IEEE, 2021, pp. 1205–1208.

[15] Y. Zhou, C. Li, L. Ding, P. Sekula, P. E. D. Love, and C. Zhou, "Combining association rules mining with complex networks to monitor coupled risks," *Reliab Eng Syst Saf*, vol. 186, pp. 194–208, Jun. 2019, doi: 10.1016/j.ress.2019.02.013.