The Impact of Using the Association Method in Determining the Pattern of COVID-19 Symptoms in the Sumatra Province

Rita Kartika Sari*
Sultan Agung Islamic University Semarang, Central Java, Indonesia.
E-mail: rita.kartika@unissula.ac.id

Ida Nuryana
Institut Dan Bisnis Asia/ STIEASIA, Indonesia.
E-mail: idanuryana1@gmail.com

Meida Rachmawati
Universitas Ngudi Waluyo, Indonesia.
E-mail: meida_r@unw.ac.id

Pandu Adi Cakranegara
Universitas President, Indonesia.
E-mail: pandu.cakranegara@president.ac.id

Danu Eko Agustinova
Universitas Negeri Yogyakarta, Yogyakarta, Indonesia.
E-mail: danu_eko@uny.ac.id

Robbi Rahim*
Sekolah Tinggi Ilmu Manajemen Sukma, Medan, Indonesia.
E-mail: usurobbi85@zoho.com

Received October 08, 2021; Accepted December 25, 2021
ISSN: 1735-188X
DOI: 10.14704/WEB/V19I1/WEB19374

Abstract

COVID-19 (coronavirus disease 2019) is a new type of disease caused by a coronavirus, namely SARS-CoV-2, also known as the Corona virus. The Corona virus outbreak is causing concern all over the world, including in Indonesia. Many people became paranoid as a result of the virus's widespread spread, which was followed by reports of a number of deaths among victims. The spread of the Corona virus in Indonesia has had a negative impact on all aspects of life. Because of the density of settlements and the vastness of Indonesia's territory, this virus became out of control and spread quickly. The lack of socialization in dealing with this virus has an impact on the community's understanding of the importance of following health protocols. The goal of this research is to use data mining techniques to determine the pattern of symptoms caused by the Corona virus. The Association method is used in data mining.

APRIORI is a popular association method that employs a high frequency pattern. This method is expected to be used to determine how likely it is that one case has the first symptom in addition to other symptoms. The parameters Support and Confidence support whether or not the association rules are required. After obtaining the results of the multiplication of Support and Confidence, the rule with the highest multiplication result will be discovered. The rule with the highest result is used as a rule to be applied in the next case, namely "If you have cough symptoms, they will be accompanied by fever symptoms with 90% support and 90% confidence".

Keywords

Corona virus, Data mining, Association Techniques, APRIORI Method, Determination of pattern rules.

Introduction

Early in 2020 we were surprised by the new virus known as the Corona virus (COVID-19) that attacked the human respiratory system which initially attacked China and found to be exact in Wuhan city in November 2019 [1]. The international community has been concerned about the high rate of corona virus spread since the end of last year. Countries around the world have prohibited their citizens from leaving their homes to ensure that the corona virus does not spread [2]. The pandemic of Covid 19, which has spread globally, is going to have an economic and socio-cultural impact and threat to citizens' lives [3]. Several valuable studies on pandemics and data mining techniques [4]–[9] have been published over the past decade [10]. These studies aim to better understand, control and manage pandemics using various data mining methods. These studies are carried out because of the importance of the pandemic control of COVID 19, many studies have been done using data mining methods as it is the most popular and efficient method [11] and has a major impact on the choice of the best techniques for pandemic studies [10].

Various studies have linked to pandemic studies using data mining techniques such as corona virus infections using the Naive Bayes in south Jakarta classification method. Data need to be classified to determine whether corona virus infection is negative or positive. The calculation results show that 55.48% of positive predictions and 44.52% of negative [12]. Furthermore, in Iran, research on the pattern of corona virus infection spread employs the clustering method and a geographic information system (GIS). This is necessary to determine the potential spread of the virus from its origin (the city of Qom) to other parts of Iran. In addition, clustering analysis is used for pattern recognition in order to track the progression of infection. According to the study's findings, Tehran was

the primary site for the spread of the corona virus and was responsible for all patterns of infection spread in Iran [13]. Furthermore, research on the comparison of BLAST, APRIORI, and Decision Tree methods on SARS-CoV-2, SARS-CoV, and MERS-CoV amino acids and codons. This must be done to limit the impact of SARS-CoV-2 (Corona Virus) on humans [11].

As a result, the use of data mining techniques can provide more information in understanding, controlling, and managing pandemics. This study was carried out based on statistical data from the Ministry of Health of the Republic of Indonesia regarding the spread of the Corona virus on the island of Sumatra on January 22, 2021, with the goal of further analyzing the use of data mining techniques in determining the pattern of Corona virus symptoms by utilizing the association data mining method [14]. This is necessary because even the smallest piece of information can lead to a solution to the Corona virus pandemic. The association method is one of the data mining techniques used to discover interesting relationships between a set of hidden items in a database, which is represented by association rules [15], [16]. The association rule takes the form of a strong relationship that can be quantified using two parameters: support and confidence [17]. The APRIORI method is a type of association rule in which prior knowledge of an itemset with a high frequency of occurrence, known as the frequent itemset, is used. APRIORI employs an iterative approach in which k-itemset is used to investigate (k+1)-itemset. The candidate (k+1)-itemset is generated in this method by combining two itemsets in the domain/size k. Candidate (k+1)-itemsets containing subset frequencies that appear infrequently or below the threshold will be trimmed and will not be used to determine association rules [15]. It is hoped that the research findings will provide information in the form of rules that will be useful in dealing with global pandemic cases, particularly in Indonesia.

Research Methodology

The study relied on symptom data from Corona virus cases in Indonesia on the island of Sumatra obtained from the Ministry of Health of the Republic of Indonesia. The study's data on the spread of the Corona virus is data on the spread on January 22, 2021 (Table 1). This study employs the RapidMiner software to analyze the pattern of Corona virus symptoms based on previously described statistical data. The APRIORI method which is part of the data mining association rules was used to solve the problem.

The APRIORI method is one of the algorithms used in association rules frequently to search objects. In the APRIORI process, an item with a frequent occurrence or also

known as the frequent set is used for previous knowledge. These are general APRIORI methods [18].

Algoritma *Apriori* (T,ε)
1. $L_1 \leftarrow \{Large\ 1-itemsets\}$
2. $K \leftarrow 2$
3. ***While*** $L_{k-1} \neq 0$
3.1. $C_k \leftarrow \{ a \cup \{b\} \mid a \in L_{k-1} \wedge b \notin a \} - \{ c \mid \{ s \mid s \subseteq c \wedge |s| = k-1 \}$ *is not subset* $L_{k-1}\}$
3.2. ***for transactions*** $t \in T$
3.2.1. $C_t \leftarrow \{c \mid c \in C_k \wedge c \subseteq t \}$
3.2.1.1. ***for candidates*** $c \in C_t$
3.2.1.1.1. $count[c] \leftarrow count[c] + 1$
3.2. $Lk \leftarrow \{c \mid c \in C_k \wedge count[c] \geq \varepsilon\}$
3.3. $k \leftarrow k + 1$
4. ***Return*** $\cup_k Lk$

Description: for T is a collection of transaction data from the table/database. "ε" is threshold for support items. Ck is a candidate at stage k. Count[c] is the number of candidates c. "c" is a candidate itemset.

Results and Discussion

The analysis was carried out using the RapidMiner software to determine the pattern of symptoms of the Corona virus by taking data samples on the island of Sumatra on January 22, 2021, which consisted of ten provinces. Table 1 shows a sample of the data used as of January 22, 2021.

Table 1 Data Sample of COVID-19 Symptom

No	Province	Symptoms of COVID-19						
		A	B	C	D	E	F	G
1	Aceh	1	1	1	1	0	0	0
2	North Sumatra	1	1	1	1	0	0	0
3	West Sumatra	1	1	1	0	1	0	0
4	South Sumatra	1	0	1	0	0	1	1
5	Lampung	1	0	1	0	1	1	0
6	Kep. Bangka Belitung	1	1	1	0	1	0	0
7	Jambi	1	0	1	1	0	1	0
8	Riau	1	1	1	0	0	0	1
9	Kep. Riau	1	1	1	1	0	0	1
10	Bengkulu	1	1	0	1	1	0	0

Source: Ministry of Health of the Republic of Indonesia on January 22, 2021

Variable explanation:

A : Cough
B : Have a cold
C : Fever
D : Weakness
E : Dizziness

F : Shortness of breath
G : Sore throat

At this point, the number of general symptoms is seven, and the number of provinces is ten.

The following is the procedure for completing the APRIORI method in determining the pattern of symptoms of the Corona virus in a case study on the Indonesian island of Sumatra:

1) Determine the itemset used for calculations using the APRIORI method, a common symptom of the Corona virus in Sumatra Province.

Table 2 General symptom data of the Corona virus

No	General symptom data of the Corona virus
1	Cough
2	Cold
3	Fever
4	Weak
5	Dizzy
6	Shortness of breath
7	Sore throat

2) Determine the number of each itemset based on each province on Sumatra's island.

Table 3 Itemset for each province on the island of Sumatra

No	Province	Symptoms of COVID-19						
		A	B	C	D	E	F	G
1	Aceh	1	1	1	1	0	0	0
2	North Sumatra	1	1	1	1	0	0	0
3	West Sumatra	1	1	1	0	1	0	0
4	South Sumatra	1	0	1	0	0	1	1
5	Lampung	1	0	1	0	1	1	0
6	Kep. Bangka Belitung	1	1	1	0	1	0	0
7	Jambi	1	0	1	1	0	1	0
8	Riau	1	1	1	0	0	0	1
9	Kep. Riau	1	1	1	1	0	0	1
10	Bengkulu	1	1	0	1	1	0	0
Total		10	7	9	5	4	3	3

3) Determine Frequent Itemset (Φ)

In this calculation, = 3 is used for transactions k = 1, then for k = 2 and so on using the constraint ≥ 3

Then it can be known:

F1 = {(Cough), (Cough), (Fever), (Weakness), (Dizziness), (Shortness of Breath), (Sore Throat)
Value of k = 1 ≥ 3

For k = 2 (2 elements), the sets that can be formed are: {(cough and cold); (cough, fever); (cough, weakness); (cough, dizziness); (cough, shortness of breath); (cough, sore throat); (cold, fever), (cold, weakness); (cold, dizziness); (cold, shortness of breath); (cold, sore throat); (fever, weakness); (fever, dizziness); (fever, shortness of breath); (fever, sore throat); (weakness, dizziness); (weakness, shortness of breath); (weakness, sore throat); (dizziness, shortness of breath); (dizziness, sore throat); (shortness of breath, sore throat)}.

Table 4 Examples of sets formed for k= 2 (two elements)

Province	**Symptom**		**Results**
	A: Cough	**B: Cold**	
Aceh	1	1	P
North Sumatra	1	1	P
West Sumatra	1	1	P
South Sumatra	1	0	S
Lampung	1	0	S
Kep. Bangka Belitung	1	1	P
Jambi	1	0	S
Riau	1	1	P
Kep. Riau	1	1	P
Bengkulu	1	1	P
∑			7

From the explanation it can be seen that P is a symptom that appears simultaneously, while S is a symptom that can appear simultaneously or sometimes does not appear.

Based on the Frequent itemset limit, it can be produced F2 ={(cough and cold); (cough, fever); (cough, weakness); (cough, dizziness); (cough, shortness of breath); (cough, sore throat); (cold, fever); (cold, weakness); (cold, dizziness); (fever, dizziness); (fever, shortness of breath); (fever, sore throat)}.

Then a combination of three elements (k= 3) can be done based on F = 2, namely:

{(cough, cold, dizziness); (cough, cold, weakness); (cough, runny nose, shortness of breath); (cough, cold, sore throat); (cough, cold, fever); (cold, fever, weakness); (cold, fever, dizziness); (fever, weakness, dizziness); (fever, weakness, shortness of breath); (fever, weakness, sore throat)}

Table 5 Example of a possible set of 3 elements (k= 3)

Province	**Symptom**			**Results**
	A: Cough	**B: Cold**	**E: Dizzy**	
Aceh	1	1	0	S
North Sumatra	1	1	0	S
West Sumatra	1	1	1	P
South Sumatra	1	0	0	S
Lampung	1	0	1	S
Kep. Bangka Belitung	1	1	1	P
Jambi	1	0	0	S
Riau	1	1	0	S
Kep. Riau	1	1	0	S
Bengkulu	1	1	1	P
∑				4

From the set k= 3, F3 can be produced with a minimum Frequent Itemset (ϕ) namely: (cough, cold, dizziness); (cough, cold, weakness); (cough, cold, fever); (cold, fever, weakness)}.

Next is a combination of four elements (k = 4)

{(cough, cold, dizziness, fever); (cough, cold, dizziness, weakness)}.

Table 6 Examples of possible sets formed from k= 4 (4 elements)

Province	**Symptom**				**Results**
	A: Cough	**B: Cold**	**E: Dizzy**	**C: Fever**	
Aceh	1	1	0	1	S
North Sumatra	1	1	0	1	S
West Sumatra	1	1	1	1	P
South Sumatra	1	0	0	1	S
Lampung	1	0	1	1	S
Kep. Bangka Belitung	1	1	1	1	P
Jambi	1	0	0	1	S
Riau	1	1	0	1	S
Kep. Riau	1	1	0	1	S
Bengkulu	1	1	1	0	S
∑					2

From the combination of k= 4 elements, there is no itemset that meets the frequent itemset (ϕ), then F4 = { }, so that F5, F6, F7 are also empty sets.

4) Determine as (ss-s) antecedent and s as consequent of FK (frequent itemset of size k) which has been obtained in F2 and F3 as many as 17 rules.

a) If symptomatic of cough then symptomatic of runny nose;

b) If symptomatic Cough then symptomatic Fever;

c) If symptomatic Cough then symptomatic Weakness;

d) if symptomatic Cough then symptomatic Dizziness;

e) If symptomatic Cough then symptomatic shortness of breath;

f) if symptomatic Cough then symptomatic Sore Throat;

g) if symptoms of colds then symptoms of fever;

h) if symptoms of colds then symptoms of weakness;

i) if symptoms of colds then symptoms of dizziness;

j) if symptoms of fever then symptoms of weakness;

k) if symptoms of fever then symptoms of dizziness;

l) if symptoms of fever then symptoms of shortness of breath;

m) if symptomatic with fever then symptomatic sore throat;

n) if symptoms of cough and cold then symptoms of dizziness;

o) if symptoms of cough and colds than symptoms of weakness;

p) if symptomatic Cough and runny nose then Fever;

q) if symptoms of colds and fever then weak.

5) After getting 17 new rules, then calculate Support and Confidence on each itemset items.

Table 7 Calculation of Support and Confidence

Rule	***Support***	***Confidence***
if symptomatic of cough then symptomatic of runny nose;	7/10*100%=70 %	7/10*100%=70 %
if symptomatic cough then symptomatic fever;	9/10*100%=90 %	9/10*100%=90 %
if symptomatic cough then symptomatic weakness;	5/10*100%=50 %	5/10*100%=50 %
if symptomatic cough then symptomatic dizziness;	4/10*100%=40 %	4/10*100%=40 %
if symptomatic cough then symptomatic shortness of breath;	3/10*100%=30 %	3/10*100%=30 %
if symptomatic cough then symptomatic sore throat;	3/10*100%=30 %	3/10*100%=30 %
if symptoms of colds then symptoms of fever;	6/10*100%=60 %	6/7*100%=85,7 %
if symptoms of colds then symptoms of weakness;	4/10*100%=40 %	4/7*100%=57%
if symptoms of colds then symptoms of dizziness;	3/10*100%=30 %	3/7*100%=42,8 %
if symptoms of fever then symptoms of weakness;	5/10*100%=50 %	5/9*100%=55,5 %
if symptoms of fever then symptoms of dizziness;	3/10*100%=30 %	3/9*100%=33,3 %
if symptoms of fever then symptoms of shortness of breath;	3/10*100%=30 %	3/9*100%=33,3 %
if symptomatic with fever then symptomatic sore throat;	3/10*100%=30 %	3/9*100%=33,3 %
if symptoms of cough and cold then symptoms of dizziness;	4/10*100%=40 %	4/7*100%=57,1 %
if symptoms of cough and colds than symptoms of weakness;	4/10*100%=40 %	4/7*100%=57,1 %
if symptomatic cough and runny nose then fever;	6/10*100%=60 %	6/7*100%=85,7 %
if symptoms of colds and fever then weak	3/10*100%=30 %	3/6*100%=50%

6) If the Support and Confidence calculations have been carried out, then the itemset selection that reaches the minimum Confidence limit is carried out. In this study, 70% was chosen, so as many as 4 rules were obtained.

Table 8 Rules that have Confidence >70%

if symptomatic cough then symptomatic colds	7/10*100%=70 %	7/10*100%=70%	0,49
if symptomatic cough then symptomatic fever	9/10*100%=90 %	9/10*100%=90%	0,81
if symptomatic of colds then symptomatic of fever	6/10*100%=60 %	6/7*100%=85,7 %	0,514 2
if symptomatic cough and runny nose then fever	6/10*100%=60 %	6/7*100%=85,7 %	0,514 2

After obtaining the results of the multiplication of Support and Confidence, the rule with the highest multiplication result will be discovered. The rule with the highest result is used as a rule to be applied in the next case, namely: "If you have cough symptoms, they will be accompanied by fever symptoms with 90 percent support and 90 percent confidence."

Conclusion

Based on the findings, it was determined that the APRIORI technique, which is part of association data mining, can be used to determine the pattern of Corona virus symptoms. The most common symptom that frequently appears, according to the various rules established, is a cough, which is followed by fever symptoms. The symptoms of the Corona virus on the Indonesian island of Sumatra are quite varied, including cough, fever, runny nose, weakness, dizziness, shortness of breath, and sore throat. Furthermore, this research can be expanded into forecasting by combining artificial neural network methods with genetic algorithms and particle swam optimization to obtain reasonably good forecasting results, or developing association methods by adding datasets to obtain more diverse rule patterns.

References

Zaharah, G. I. Kirilova, and A. Windarti, “The impact of the corona virus outbreak on teaching and learning activities in Indonesia,” *Salam J. Sos. dan Budaya Syar’i*, vol. 7, no. 3, pp. 269–282, 2020.

S. R. Pudjiastuti,, S., and N. Hadi, “The Effect of Corona Virus on the Global Climate,” *Jhss (Journal Humanit. Soc. Stud.*, vol. 4, no. 2, pp. 130–136, 2020.

J. Torales, M. O’Higgins, J. M. Castaldelli-Maia, and A. Ventriglio, “The outbreak of COVID-19 coronavirus and its impact on global mental health,” *Int. J. Soc. Psychiatry*, vol. 66, no. 4, pp. 317–320, 2020.

B. Supriyadi, A. P. Windarto, T. Soemartono, and Mungad, “Classification of natural disaster prone areas in Indonesia using K-means,” *Int. J. Grid Distrib. Comput.*, vol. 11, no. 8, pp. 87–98, 2018.

F. Rahman, I. I. Ridho, M. Muflih, S. Pratama, M. R. Raharjo, and A. P. Windarto, "Application of Data Mining Technique using K-Medoids in the case of Export of Crude Petroleum Materials to the Destination Country," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 835, no. 1, 2020.

A. Waluyo, H. Jatnika, M. R. S. Permatasari, T. Tuslaela, I. Purnamasari, and A. P. Windarto, "Data Mining Optimization uses C4.5 Classification and Particle Swarm Optimization (PSO) in the location selection of Student Boardinghouses," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 874, no. 1, pp. 1–9, 2020.

M. Widyastuti, A. G. Fepdiani Simanjuntak, D. Hartama, A. P. Windarto, and A. Wanto, "Classification Model C.45 on Determining the Quality of Custumer Service in Bank BTN Pematangsiantar Branch," *J. Phys. Conf. Ser.*, vol. 1255, no. 012002, pp. 1–6, 2019.

Z. R. S. Elsi *et al.*, "Utilization of Data Mining Techniques in National Food Security during the Covid-19 Pandemic in Indonesia," *J. Phys. Conf. Ser.*, vol. 1594, no. 1, 2020.

"Training on installing solar water pump for resident of singkar 1 wareng wonosari gunungkidul yogyakarta indonesia | *Jurnal Pengabdian dan Pemberdayaan Masyarakat Indonesia*." https://journal1.ptti.expert/jppmi/article/view/7.

R. Safdari, S. Rezayi, S. Saeedi, M. Tanhapour, and M. Gholamzadeh, "Using data mining techniques to fight and control epidemics: A scoping review," *Health Technol. (Berl).*, pp. 759–771, 2021.

J. E. Huh, S. Han, and T. Yoon, "Data mining of coronavirus: SARS-CoV-2, SARS-CoV and MERS-CoV," *BMC Res. Notes*, vol. 14, no. 1, pp. 1–6, 2021.

W. Miftahul Anwar, Mohammad Iwan Wahyuddin, "Jurnal Mantik Jurnal Mantik," *Mobile-Based Natl. Univ. Online Libr. Appl. Des.*, vol. 3, no. 2, pp. 10–19, 2019.

M. Azarafza, M. Azarafza, and H. Akgün, "Clustering method for spread pattern analysis of corona-virus (COVID-19) infection in Iran," *J. Appl. Sci. Eng. Technol. Educ.*, vol. 3, no. 1, pp. 1–6, 2021.

B. Lei, "Apriori-based spatial pattern mining algorithm for big data," *Proc. - 2020 Int. Conf. Urban Eng. Manag. Sci. ICUEMS 2020*, pp. 310–313, 2020.

N. Distefano and S. Leonardi, "Apriori algorithm for association rules mining in aircraft runway excursions," *Civ. Eng. Archit.*, vol. 8, no. 3, pp. 206–217, 2020.

H. Yu, "Apriori algorithm optimization based on Spark platform under big data," *Microprocess. Microsyst.*, vol. 80, no. November 2020, p. 103528, 2021.

M. Sornalakshmi *et al.*, "An efficient apriori algorithm for frequent pattern mining using mapreduce in healthcare data," *Bull. Electr. Eng. Informatics*, vol. 10, no. 1, pp. 390–403, 2021.

I. Djamaludin and A. Nursikuwagus, "Analysis of Consumer Purchase Patterns in Sales Transactions Using the Apriori Algorithm," *Simetris J. Tek. Mesin, Elektro dan Ilmu Komput.*, vol. 8, no. 2, p. 671, 2017.