APRIORI ALGORITHM FOR FINDING RELATIONSHIPS BETWEEN STUDENT SELECTION PATHWAYS SCHOOL DEPARTMENTS WITH STUDENT GRADUATION LEVELS

Fiqih Satria[1], Muhamad Muslihudin[2], Nadya Amalia Nasoetion[3]
[1]Department of Information Systems, UIN Raden Intan Lampung
[2]Department of Information Systems, STMIK Pringsewu Lampung
[3]Department of Islamic Communication and Broadcasting, UIN Raden Intan Lampung
[1,3]Jl. Letkol H. Endro Suratmin, Sukarame, Bandar Lampung, Indonesia
[2]Jl. Wisma Rini No. 09 Pringsewu, Lampung, Indonesia
E-mail: fiqihsatria@radenintan.ac.id[1], muslihudinstmikpsw@gmail.com[2], nadyaamalia@radenintan.ac.id[3]

Abstract

The selection process for New Student Admissions (PMB) of State Islamic Universities in Indonesia, especially at UIN Raden Intan Lampung for undergraduate (S1) programs, is pursued through 3 (three) different selection patterns. Selection of SPAN-PTKIN, UM-PTKIN, and UM independent. The three entry paths have their own character, according to their functions and objectives. With these differences, this study identifies the relationship between student entry pathways, majors/types of previous high-level schools with GPA scores, and the length of the study period of students. Data mining in this study is to uses a priori algorithm. The Apriori algorithm is one of the classic data mining algorithms. The a priori algorithm is used to determine the most dominant factor in predicting student graduation rates. A priori algorithms are used so that computers can learn association rules, looking for patterns of relationships between one or more items in a dataset. The data in this study were taken from student data in SIAKAD, namely the student data of Raden Intan Lampung State Islamic University (UIN), Islamic Community Development Department, Class of 2015, the data used included the type of school of origin, entry route, GPA, and length of time. study period. From the results of the research, it is found that the rules or regulations that graduate students with a study period of 4 years / less and a GPA of 3.51 - 4.00 are students who enter through the Academic Interest (PMA) search path and from their school of high school (SMAN) with Value Support. 14,286 and 60% confidence value. These results can be used by universities in encouraging students from other entry paths and from other schools to graduate on time, such as students who entered through the PMA route and from high school from SMAN with certain efforts.

1. INTRODUCTION

The graduation rate of students in a college or study program is an important element. Especially in relation to the university accreditation process. College accreditation now uses IAPT 3.0 criteria and procedures which have 9 standards. Especially at the 3rd standard regarding students, student graduation is an important point that is included in the accreditation assessment [1]. The study period targeted at the undergraduate program is 4 years or 8 semesters. However, in fact, there are still many students who graduate beyond the targeted study period. In this case, universities or study programs can use student profile data to predict student graduation rates. One of them is by using data mining.

One example of the use of data mining is in terms of increasing credit card facilities. From the existing customer data, patterns will be searched from the data to find information on potential customers and non-

potential customers [2]. Knowing this prediction will be very useful for the company in determining which customers will be given permission to have a credit card. Of course, companies will think twice about giving credit cards to customers who have no potential [2]. In addition to the banking sector, a priori algorithms in the social field can also be used to assess how efficient the communication/media channels are in informing the public about a phenomenon [3]. One of the algorithms that exist in the concept of data mining is the a priori algorithm. The a priori algorithm is one of the data mining algorithms that uses item combination association rules [4].

S L Br Ginting, S A Purba, I D Sumitra in 2017 conducted a research entitled Apriori Algorithm to Show the Correlation of Academic Values with Student Graduation. In this study, the data used is the value of the programming algorithm I course and is associated with the GPA and the length of the

student's study period. Based on application testing and discussion, it can be concluded that there is a relationship between the graduation category and the value of the course based on the association process and data combination. Then based on training and database testing, it can be concluded that the amount of training data (the amount of data in the database) can affect the percentage of match or data mining accuracy. In application testing, it can be seen that if you change the threshold value, various combinations will be produced. A large threshold value is not necessarily the best threshold value, with a high success rate and vice versa. The best threshold value can be influenced by the amount of data and the number of data combinations used [5].

Harton Rohul Meisa Tambun and Anofrizen in 2015 conducted a study entitled Design of Data Mining Applications to Display Student Graduation Rate Information Using Apriori Algorithms. What can be taken in this research is that this Data Mining Application can be used to display graduation rate information. The information displayed is in the form of support and confidence values, the relationship between graduation rates and student master data. The higher the value of Confidence and support, the stronger the value of the relationship between attributes. The student master data that is processed by mining includes entry process data, school origin data, student city data, and study program data. The results of this data mining process can be used as a consideration in making further decisions about factors that affect graduation rates, especially factors in graduation data and student master data [6].

This research is different from previous research. This study wanted to find out the relationship between student entry paths, majors/types of previous high schools with GPA scores and the length of student study period. By using this a priori algorithm, it aims to determine the most dominant factor to predict student graduation rates. Researchers hope to be able to assist in finding out information about student entry pathways and majors/types of previous high-level schools that can be used as recommendations in determining student graduation rates and cumulative achievement index (GPA) in the Islamic Community Development study program, Faculty of Da'wah and Communication Studies, Raden State Islamic University. Intan Lampung. In addition, as an effort to assist universities in improving the quality of graduates in the following year.

II. LITERATURE
2.1. Data Mining
Data mining is the process of mining or extracting meaning from such large amounts of data, by extracting the data to find certain patterns and analyzing them to obtain knowledge or information [4]. A method used to extract patterns from data or commonly referred to as Knowledge Discovery in Database (KDD) consists of the following steps [7]:

1. Data cleaning
2. Data integration
3. Data selection
4. Data transformation
5. Data mining
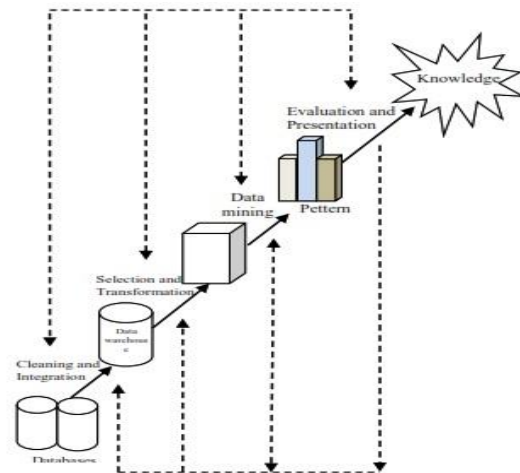6. Pattern evaluation
7. Knowledge presentation

Figure 1. Stages of Knowledge Discovery in Database (KDD) [7]

2.2. Apriori Algorithm

A priori algorithm is a type of association rule in data mining. This association analysis is a data mining technique to find associative rules between a combination of items/attributes [8].

An example of associative rule from buying analysis in a mini market is to know how likely it is that a customer buys a cold drink at the same time as buying a snack. With this knowledge, the mini market owner can arrange the placement of the shelves of the two products close together.

The basic methodology of association analysis is divided into two stages [9]:

a. High frequency pattern analysis
   This stage is looking for a combination of items that meet the minimum requirements of the support value in the database. The support value of an item is obtained by the following formula:

$$\text{Support (A)} = \frac{Jumlah\ transaksi\ mengandung\ A}{Total\ transaksi} \times 100\%$$
.... (1)

$$\text{Support (A} \cap \text{B)}$$
$$\frac{Jumlah\ transaksi\ mengandung\ A\ dan\ B}{Total\ transaksi} \times 100\%$$
.... (2)

b. Associative rule formation
   After all high-frequency patterns are found, then look for the associative rule that meets the minimum requirements for confidence by calculating the confidence of the associative rule A_B

   The confidence value of the A_B rule is obtained from the following formula:

$$\text{Confidence} = P(B|A) =$$

$$\frac{Jumlah\ transaksi\ mengandung\ A\ dan\ B}{Jumlah\ transaksi\ mengandung\ A} \times 100\%$$

…. (3)

III. RESEARCH METHODS

In conducting the analysis, the researcher used the manual method, namely by making several tables in Microsoft Excel to find the support value on Frequent Itemset then continued by using the Tanagra software to find rules with the a priori algorithm. The data used in this study is the student data of the State Islamic University (UIN) Raden Intan Lampung Department of Islamic Community Development Class of 2015 which is stored in the Academic Information System (SIAKAD).

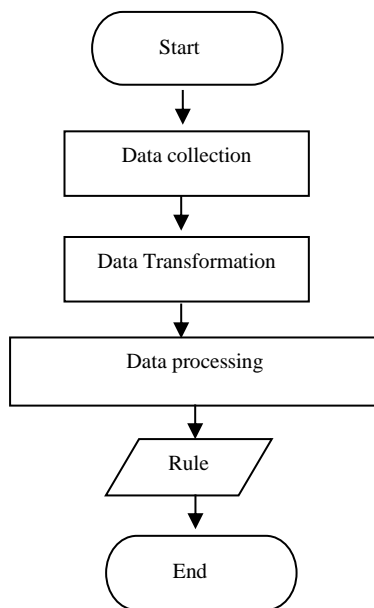The flowchart of the research stages can be seen in Figure 2.

Figure 2. Research Stages

IV. RESULTS
4.1 Data Needs

The data was taken from the personal data of the students of the State Islamic University (UIN) Raden Intan Lampung, Department of Islamic Community Development, Class of 2015 which were stored in the Academic Information System (SIAKAD). Not all attributes in student self-data are used in predictions

using the a priori algorithm, but only Type of School of Origin, Entry Path, GPA, and length of study are used. Determination of timely and untimely graduation based on GPA is divided into 2 categories, namely, study duration of 4 years or less (meaning graduating on time) and study length of more than 4 years (meaning not graduating on time). Type of School Origin consists of 6 categories, namely MAN, MAS, SMAN, SMAS, SMAS PONPES and SMKN. The entry path consists of 4 categories, namely SPAN-PTKIN, UM-PTKIN, UM-Local and Academic Interest Search (PMA).

4.2 Data Transformation

The categorization of graduation data is based on the length of study, namely graduating on time, if the study period is 4 years or less than 4 years and graduating not on time, if the length of study is more than 4 years. From these two categories, categories can be made based on combinations with GPA, as can be seen in Table 1.

Table 1. Graduation Category

Category	Information
X1	Length of study 4 years / less and GPA 3.51 – 4.00
X2	Length of study 4 years / less and GPA 3.10 – 3.50
X3	Length of study 4 years / less and GPA 2.51 – 3.00
Y1	Length of study more than 4 years and GPA 3.51 – 4.00
Y2	Length of study more than 4 years and GPA 3.10 – 3.50
Y3	Length of study more than 4 years and GPA 2.51 – 3.00

4.3 Data processing

Process data processing is to perform calculations using a priori algorithm. The data used is the student data of the State Islamic University (UIN) Raden Intan Lampung, Department of Islamic Community Development Class of 2015, which amounted to 42 data.

Table 2. Initial Data

ID	School Department	Student Selection Pathways	Graduation Category
1	MAN	SPAN	X2
2	MAN	SPAN	Y1
3	MAN	SPAN	Y2
4	MAN	SPAN	Y1
5	SMAS	PMA	Y1
6	SMAN	PMA	X2
7	SMAN	PMA	X2
8	MAN	UMPTKIN	Y2

9	MAN	UMPTKIN	Y2
10	SMAN	UMPTKIN	Y1
11	SMAN	UMPTKIN	Y2
12	SMAS	PMA	X2
13	SMAN	UMPTKIN	Y2
14	SMAN	PMA	Y2
15	SMAN	PMA	Y2
16	SMAN	PMA	X2
17	SMAN	PMA	X2
18	SMKN	PMA	Y2
19	MAS	PMA	X1
20	SMAN	PMA	X1
21	SMAN	PMA	Y2
22	SMAN	PMA	X1
23	SMAN	PMA	Y1
24	SMAN	PMA	Y2
25	SMAN	PMA	Y2
26	SMKN	PMA	X1
27	SMAS PONPES	PMA	Y2
28	SMAN	PMA	X1
29	SMAS	UM-LOKAL	Y2
30	SMAN	UM-LOKAL	Y2
31	SMKN	PMA	Y2
32	SMAS	PMA	Y2
33	SMKN	PMA	X1
34	SMAS	PMA	X1
35	SMAN	PMA	X1
36	SMKN	PMA	X2
37	SMAN	PMA	X2
38	MAS	UM-LOKAL	X2
39	MAN	PMA	X2
40	SMAN	PMA	X1
41	SMAN	PMA	X1
42	SMKN	PMA	X2

From the initial data of 42 students, cleaning data, then the next data is transformed into the data format needed by the a priori algorithm from 42 students, the most data from SMAN schools are 21 people, PMA pathway is 30 people, and graduated with Y2 category as many as 16 people.

4.4 High Frequency Pattern Analysis

The following is the completion stage for analyzing high-frequency patterns based on the data provided in the table above. The process of forming C1 or called 1 itemset with the minimum amount of support used in this study is 15%. The following is the calculation for the formation of 1 itemset:

$$S\ (MAN) = \frac{\sum data\ contains\ MAN}{\sum data} * 100\%$$
$$= \frac{7}{42} * 100\% = 16{,}67\%$$

$$S\ (MAS) = \frac{\sum data\ contains\ MAS}{\sum data} * 100\%$$
$$= \frac{2}{42} * 100\% = 4{,}76\%$$

$$S\ (SMAN) = \frac{\sum data\ contains\ SMAN}{\sum data} * 100\%$$
$$= \frac{21}{42} * 100\% = 50\%$$

$$S\ (SMAS) = \frac{\sum data\ contains\ SMAS}{\sum data} * 100\%$$
$$= \frac{5}{42} * 100\% = 11{,}90\%$$

$$S\ (SMAS\ PONPES) = \frac{\sum data\ contains\ SMAS\ PONPES}{\sum data} * 100\%$$
$$= \frac{1}{42} * 100\% = 2{,}38\%$$

$$S\ (SMKN) = \frac{\sum data\ contains\ SMKN}{\sum data} * 100\%$$
$$= \frac{6}{42} * 100\% = 14{,}29\%$$

$$S\ (PMA) = \frac{\sum data\ contains\ PMA}{\sum data} * 100\%$$
$$= \frac{30}{42} * 100\% = 71{,}43\%$$

$$S\ (SPAN) = \frac{\sum data\ contains\ SPAN}{\sum data} * 100\%$$
$$= \frac{4}{42} * 100\% = 9{,}52\%$$

$$S\ (UM - LOKAL) = \frac{\sum data\ contains\ UM - LOKAL}{\sum data} * 100\%$$
$$= \frac{3}{42} * 100\% = 7{,}14\%$$

$$S\ (UMPTKIN) = \frac{\sum data\ contains\ UMPTKIN}{\sum data} * 100\%$$
$$= \frac{5}{42} * 100\% = 11{,}90\%$$

$$S\ (X1) = \frac{\sum data\ contains\ X1}{\sum data} * 100\%$$
$$= \frac{10}{42} * 100\% = 23{,}81\%$$

$$S\ (X2) = \frac{\sum data\ contains\ X2}{\sum data} * 100\%$$
$$= \frac{11}{42} * 100\% = 26{,}19\%$$

$$S\ (Y1) = \frac{\sum data\ contains\ Y1}{\sum data} * 100\%$$
$$= \frac{5}{42} * 100\% = 11{,}90\%$$

$$S\,(Y2) = \frac{\sum data\ contains\ Y2}{\sum data} * 100\%$$

$$= \frac{16}{42} * 100\% = 38,10\%$$

Tabel 3. Itemset Support Value

Itemset	Support
MAN	16,67%
MAS	4,76%
SMAN	50,00%
SMAS	11,90%
MAS PONPES	2,38%
SMKN	14,29%
PMA	71,43%
SPAN	9,52%
UM-LOKAL	7,14%
UMPTKIN	11,90%
X1	23,81%
X2	26,19%
Y1	11,90%
Y2	38,10%

Next, the process that must be passed is the formation of C2 or called 2 itemset using a combination of itemset which has a minimum amount of support = 15%. Here are the calculations in the formation of C2 or 2 itemset:

$$S\,(X1, PMA) = \frac{\sum data\ contains\ X1, PMA}{\sum data} * 100\%$$

$$= \frac{10}{42} * 100\% = 23,81\%$$

$$S\,(X1, SMAN) = \frac{\sum data\ contains\ X1, SMAN}{\sum data} * 100\%$$

$$= \frac{6}{42} * 100\% = 14,29\%$$

$$S\,(X2, PMA) = \frac{\sum data\ contains\ X2, PMA}{\sum data} * 100\%$$

$$= \frac{9}{42} * 100\% = 21,43\%$$

$$S\,(Y2, SMAN) = \frac{\sum data\ contains\ Y2, SMAN}{\sum data} * 100\%$$

$$= \frac{8}{42} * 100\% = 19,05\%$$

$$S\,(Y2, PMA) = \frac{\sum data\ contains\ Y2, PMA}{\sum data} * 100\%$$

$$= \frac{9}{42} * 100\% = 21,43\%$$

$$S\,(SMAN, PMA) = \frac{\sum data\ contains\ SMAN, PMA}{\sum data} * 100\%$$

$$= \frac{17}{42} * 100\% = 40,48\%$$

Table 4. Support of the combination of 2 itemset

Itemset	Support
X1 dan PMA	23,81%
X1 dan SMAN	14,29%
X2 dan PMA	21,43%
Y2 dan SMAN	19,05%
Y2 dan PMA	21,43%
SMAN dan PMA	40,48%

The specified minimum support is 15%, so the combination of 2 itemsets that do not meet the minimum support will be removed, as shown in the table below:

Table 5. Combination of 2 itemset

Itemset	Support
X1 and PMA	23,81%
X2 and PMA	21,43%
Y2 and SMAN	19,05%
Y2 and PMA	21,43%
SMAN and PMA	40,48%

4.5 Apriori Algorithm Calculation with Tanagra

The Apriori Algorithm in Tanagra can be formed with a predetermined algorithm or steps. This algorithm consists of two algorithms, namely:

1. Algorithm Support

Support determination algorithm can be seen in the algorithm below. The following are the results of the Support from the results of the analysis that the author did.

In Figure 3 is the selection of data itemsets to be processed using Tanagra, not all itemsets are used, only itemsets that have gone through the data cleaning process can be used.
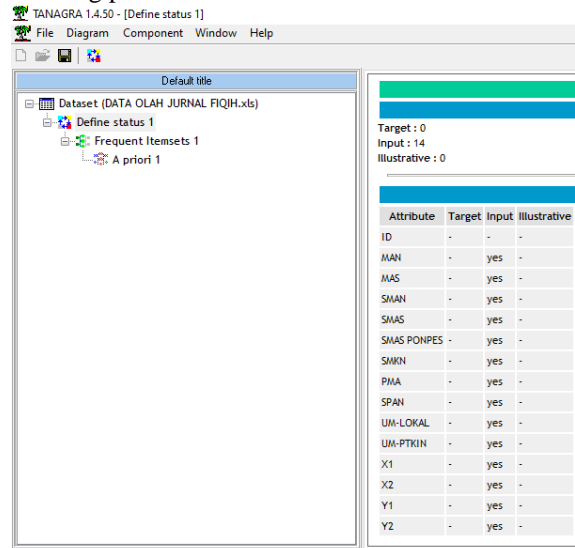

Figure 3. Selection of the data itemset used

Figure 4 shows a combination of 2 itemset used as input. The result is that there are only 5 itemset combinations that have a support value above 15% which will be processed using the a priori algorithm.
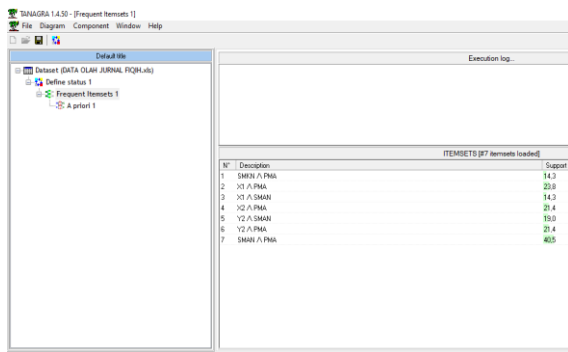
Figure 4. Display Support

2. Algorithm Confidence

Confidence determination algorithm can be seen in the algorithm below which consists of input, output and process. The following is the algorithm of the confident algorithm. If support 50% then min. Confidence is met. Else elimination. From the rule, only itemset combinations that have a confidence value of 50% and above will be chosen to be used as rules. Here are the results of the application using Tanagra:
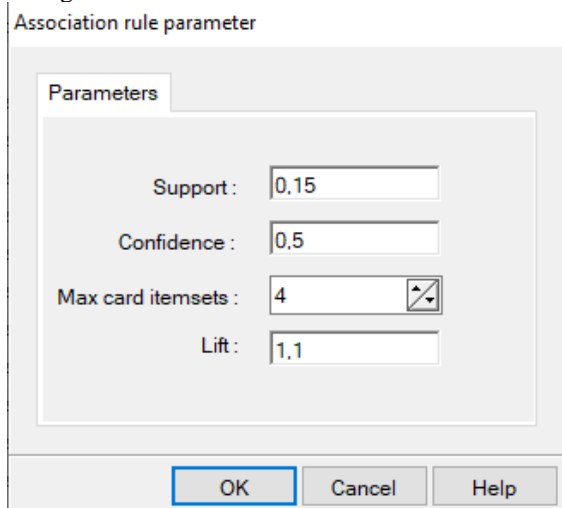

Figure 5. Confident Algorithm

The rules that appear are itemset combinations that have a support value of 15% or more and a confidence value of 50% or more. From processing using Tanagra, 9 rules are generated as shown in the image below.
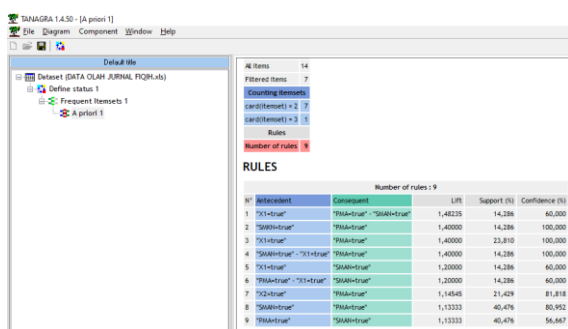

Figure 6. Rules

4.6 Conclusion

From the calculations using the Apriori Algorithm with a Support value of 30% and a Confidence value of 60%, 9 Rules are formed as shown above which can be concluded as follows:

1. If a student passes with the X1 category, then he is a PMA entry student and from SMAN.
2. If the origin is a SMKN, the average entry is through PMA
3. If a student passes the X1 category then he is a PMA entry student.
4. If the origin is high school and graduated with category X1, then the average entry route is through PMA.
5. If a student passes with the X1 category, the origin of the school is from SMAN.
6. If the student's entry is through PMA and passes with the X1 category, the origin of the school is from SMAN.
7. If a student passes the X2 category, he or she is a PMA entryway student.
8. If the origin of the school is from SMAN, the average entrance is through PMA.
9. If the entrance is through PMA, the average school origin is from SMAN

V. CONCLUSION
5.1. Conclusion

Based on the results and discussion in the discussion chapter above, from the data of students at the State Islamic University (UIN) Raden Intan Lampung, Department of Islamic Community Development, Class of 2015 stored in the Academic Information System (SIAKAD), the authors can draw the conclusion that: 1) Application of the Apriori Algorithm in This study is to see the relationship between Student Entry Path and Type of School Origin on the graduation rate of students. 2) It was found that, 1. If a student graduated with category X1 (study length of 4 years / less and GPA 3.51 – 4.00) then he is a student with the entrance to the Academic Interest Search (PMA) and school origin from a State High School (SMAN) with a Support value of 14,286 and a Confidence value of 60%. 3) These results can be used by universities in encouraging students from other schools to enter and graduate on time, just as students who enter through the PMA route and from high school with certain efforts.

5.2. Suggestion

For further research, the authors suggest conducting data mining research using other data mining methods and more complex variables in order to get maximum results. Researchers hope that further research will capture more big data phenomena so that they can be processed into new knowledge that is useful to support the development of other fields of science.

References

[1]   BANPT, "AKREDITASI PERGURUAN TINGGI KRITERIA DAN PROSEDUR IAPT 3.0," *www.banpt.or.id*, 2021. https://www.banpt.or.id/wp-content/uploads/2019/09/Lampiran-02-PerBAN-PT-3-2019-Kriteria-dan-Prosedur-IAPT-3_0.pdf.

[2]   H. D. Honesqi, "KLASIFIKASI DATA MINING UNTUK MENENTUKAN TINGKAT PERSETUJUAN KARTU KREDIT," *J. TEKNOIF*, vol. 5, no. 2, 2017, doi: 10.21063/jtif.2017.v5.2.57-62.

[3]   N. Nasution and F. Satria, "The Influence of COVID-19 News for Religious Activities in Lampung using Apriori Algorithm," *JTKSI*, vol. 4, no. 2, 2021.

[4]   Kusrini, "Proses Data Mining," *Min. Massive Datasets*, 2005.

[5]   S. L. B. Ginting, "Algoritma Apriori untuk Menampilkan Korelasi Nilai Akademik dengan Kelulusan Mahasiswa: Data Mining," *Komputika J. Sist. Komput.*, vol. 6, no. 2, 2019, doi: 10.34010/komputika.v6i2.1706.

[6]   H. R. M. Tambun and A. Anofrizen, "Rancang Bangun Aplikasi Data Mining Untuk Menampilkan Informasi Tingkat Kelulusan Mahasiswa Dengan Algoritma Apriori," *J. Ilm. Rekayasa dan Manaj. Sist. Inf.*, vol. 1, no. 1, 2015.

[7]   J. Han, M. Kamber, and J. Pei, *Data Mining Concept and Techniques, 3rd ed.* 2012.

[8]   D. Nofriansyah, "Konsep Data Mining Vs Sistem Penunjang Keputusan," *Deepublish*. 2015.

[9]   E. T. Kusrini, dan Luthfi, *Algoritma Data mining. Yogyakarta: Andi.* Yogyakarta: Andi, 2009.