ATLANTIS PRESS

Advances in Engineering Research, volume 189

Ahmad Dahlan International Conference Series on Engineering and Science (ADICS-ES 2019)

Discovering Meaningful Pattern of Undergraduate Students Data using Association Rules Mining

Herman Yuliansyah
Informatics Departments
Ahmad Dahlan University
Yogyakarta, Indonesia
herman.yuliansyah@tif.uad.ac.id

Hafsah
Informatics Departments
Pembangunan Nasional University
Yogyakarta, Indonesia
hafsahotha@upnyk.ac.id

Ika Arfiani
Informatics Departments
Ahmad Dahlan University
Yogyakarta, Indonesia
ika.arfiani@tif.uad.ac.id

Rusydi Umar
Informatics Departments
Ahmad Dahlan University
Yogyakarta, Indonesia
rusydi.umar@mti.uad.ac.id

***Abstract*—Association rules mining is a technique in data mining to discovering a meaningful pattern of data. The main objective of this research is to identify undergraduate students data and to get the profile and insight from the past data. It will have a benefit for improvement in academic activity in the future. This research has two phases. The first phase is preprocessing data, and the second phase is analyzing and measurement data using the Apriori Algorithms. The data preprocessing stage is done by cleaning data from noise and transforming data into the specified parameters. We use four feature/variable data, namely length of study duration, length of thesis duration, and Grade Point Average (GPA), and English proficiency score. The results of this research are variables of English proficiency score, Grade Point Average (GPA), and length of study duration having relations in student data.**

Keywords—data mining, association rules mining, apriori algorithms, frequent itemsets mining, student undergraduate data, knowledge discovery, data patterns

I. INTRODUCTION

This paper proposes an analysis of student data to get identification and to get profile model of student. This article is a continuation of the previous study, and the first author has analyzed the alumni data of a university[1].

Data mining is a technique used to get useful mining data from massive datasets and finding meaningful patterns[2]. Data mining has implemented in broad areas, not only for the education field but also in taurism[3], health[4][5][6], big data[7][8], and so on. Association rules mining (ARM) is a powerful technique in data mining, besides clustering, and classification. Agrawal introduced ARM in 1993[9]. ARM is used to analyze data and to find a pattern from the data. Several fields have implemented the ARM to discover the data, i.e., using an ARM to analyze a virulent form of oxygen[10][11], extract user's TV recommendation for a program recommendation[12], finding time-related in traffic prediction[13][14], network detection from an intruder[15], and frequent traveler in an agro-tourism activity[3], and Early childhood caries (ECC) analysis for risk identification[16].

This article will discuss the implementation of data mining in Education. Students data is complex data. There are many opportunities to mine the data to become new insight for the educator or management. Several researchers have been discussing the implementation of data mining and association rules mining. The evaluation of Student's performance has done by discovering the knowledge based on the internal assessment and semester examination[17]. This evaluation is to identify the number of students who needed individual concern to reduce failed ration and to take appropriate action for the next semester examination. These performances have a relation with the length of study. The investigation of solutions can help the student to manage their performance[18]. Data mining in education also investigates to extract the hidden knowledge by finding relationships among student learning characteristics and behavior[19]. All of the student's performance analysis will affect the improvement of education quality. By improving the quality of education, every university wants to increase the number of students by determining strategies of promotion[20].

Every university has its way to analyze their student data. But most of the university only use a simple procedure to analyze like finding ratio data or percentage data from student data history. Another problem is university only store the data through traditional information system and database or spreadsheet files. However, these data have not utilized by being analyzed to become meaningful information. This unutilized data is a compelling case because data mining can observer various types of data and information repositories. Then analyze the data uses a few method/algorithm to answer and mine the hidden information to be meaningful information.

We are interested in analyzing this student data to identify correlation among the variable of the length of study duration, length of thesis duration, and Grade Point Average (GPA), and English proficiency score. The aim of this analyzes to find a pattern between this variable. The hypothesis in this research is if student's data history is affected by the variable length of study duration, length of thesis duration, Grade Point Average (GPA), and English proficiency score, then association rules mining will discover the meaningful data for university management. The results of this association pattern analysis are expected to be useful for the leaders of the study program or faculty as input material in making decisions.

This article organized as follows: In section 2, the author will present the methodology of the research. In part 3, the result of the study will discuss, and in part 4 is the conclusion of this research.

II. METHODOLOGY

A. *Data Mining Process*

Han [21] stated that the process of finding knowledge in data mining has done through an iterative sequence in the steps shown in Figure 1:

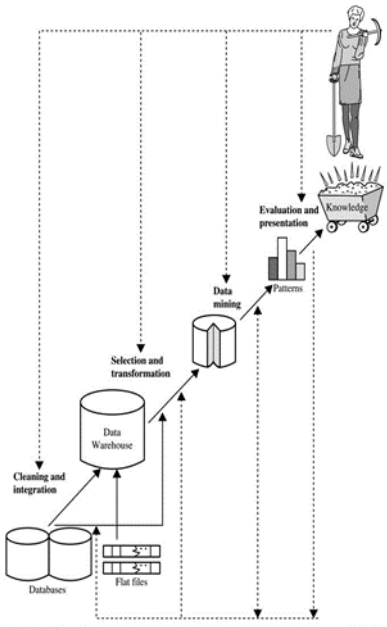

Fig. 1. The sequence of the knowledge discovery process in data mining [21]

Figure 1 shows the knowledge discovery process consists of several steps, including:

1) Data cleaning
   At this stage, data cleaning will be carried out by eliminating data that is noise and inconsistent data.
2) Data integration
   Data integrated if the data source comes from many data sources. To run data cleaning and data integration, this is called the preprocessing stage. The preprocessing stage produces data stored in the data warehouse.
3) Data selection
   At this stage, the variety of relevant data will take for analysis. This data collected from the data warehouse.
4) Data transformation
   At this stage, the data transformed and incorporated it into the appropriate format for the data mining process.
5) Data mining
   A necessary process where intelligent methods are applied to extract data patterns
6) Pattern evaluation
   This stage is carried out to identify interest patterns that represent knowledge based on exciting actions
7) Knowledge presentation
   At this stage will visualize with knowledge representation techniques used to present mining knowledge to users.

B. *Association Rules Mining*

Association rules mining [9] can take from a data set where each example consists of a set of items. The association rule has the form X Y, where X and Y are itemsets, and the interpretation is that if the set X occurs in an example, then the set Y may also occur.

Each association rule is usually associated with two measured statistics from the given data set. The frequency or support of rule X Y denoted fr (X Y), is the number (or alternative relative frequency) of the example in which it occurred. Confidence is a conditional probability which is observed P (X | Y) = fr () / fr (X).

The Apriori Algorithm [22] finds all association rules, between sets X and Y, which exceed the user-defined support and confidence limits. In the association rules mining, unlike most other learning assignments, the result is a set of rules concerning different subsets of feature space.

The association's rules motivated by supermarket basket analysis, but as an independent domain technique, they have found applications in various fields. Mining association rules are part of the frequent itemset field or broader frequent pattern mining.

C. *Data Analysis Process*

The stages that will be carried out to complete this research as in Figure 2:

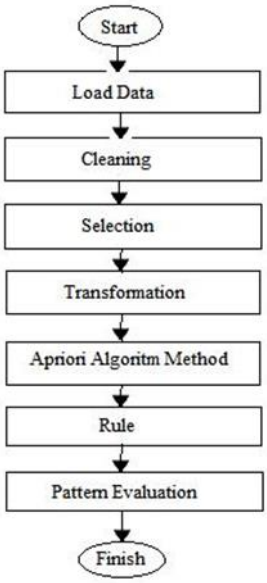

Fig. 2. Data Analysis Process

Figure 2 shows the data analysis process in this research:

1) Load data
   Load data, which is to take the initial data from the Spreadsheet.
2) Initial Data
   Initial data, which displays the initial data after loading the data. The initial data is unprocessed data.
3) Data Cleaning
   Perform cleaning student data, which eliminates inconsistent noise and data or irrelevant data. For example, deleting the same data and not relating to the variables needed.
4) Data Selection
   Selecting student data, which is data that is in the database is often not all used, therefore only the appropriate data to be analyzed will take from the database.

5) Data Transformation
   Transform student data, i.e., data is changed or combined into a format suitable for processing in data mining.
6) Apriori
   Performing an apriori stage, namely the calculation phase of the data by the Apriori algorithm. Until the association rules obtained from student data.
7) Pattern Evaluation
   Pattern evaluation, which determines the rules based on calculations from the a priori algorithm process.

III. RESULTS AND DISCUSSIONS

Table 1 is the data collected from a private university. There are 1437 lines of data that have been successfully obtained from the student's period, namely June 16, 2012, to July 23, 2018.

TABLE I. DATASET

Row data	*Length of thesis duration (Month)*	*Length of study duration (Year)*	*Grade Point Average (GPA)*	*English proficiency score*
1.	10.1	3.9	3.8	400
2.	1.6	6.8	2.7	450
3.	2.9	3.9	3.7	440
4.	3.0	3.9	3.6	420
5.	10.1	3.9	3.8	400
...	...	...	...	...
1433.	3.5	6.8	3.5	410
1434.	3.5	6.8	3.4	400
1435.	3.4	4.8	3.0	420
1436.	3.4	4.8	3.1	420
1437.	9.9	6.8	2.5	480

Preprocessing steps such as performing data cleaning, data selection, and data transformation have been carried out in the spreadsheet file until the dataset in Table 1 was successfully prepared. Then after the dataset ready to use, the next step is analyzed in programming. This study uses Python programming to implement apriori algorithms in analyzing the data sets used. Table 2 shows the implementation of the program code used for this analysis. Python programming uses the mlxtend library.

TABLE II. APRIORI ALGORITHMS IN PYTHON PROGRAMMING

Row Code	*English proficiency score*
1.	import csv
2.	import pandas as pd
3.	from mlxtend.preprocessing import TransactionEncoder
4.	from mlxtend.frequent_patterns import apriori
5.	from mlxtend.frequent_patterns import association_rules
6.	dataset = []
7.	tag = ["TT", "TS", "GPA", "EP"]
8.	with open('dataset.csv') as csvDataFile:
9.	csvReader = csv.reader(csvDataFile)
10.	for row in csvReader:
11.	baris = []
12.	z = 0
13.	for i in row:
14.	baris.append(tag[z]+str(i))
15.	z=z+1
16.	dataset.append(baris)
17.	dataset
18.	te = TransactionEncoder()
19.	te_ary = te.fit(dataset).transform(dataset)
20.	df = pd.DataFrame(te_ary, columns=te.columns_)
21.	frequent_itemsets = apriori(df, min_support=0.02, use_colnames=True)
22.	frequent_itemsets
23.	association_rules(frequent_itemsets, metric="confidence", min_threshold=0.11)

Because all data are numbers, before analyzing, we need to mark the data according to each variable. The 7th line code from Table 2 is the tagging process that we have implemented. We use markers like TT for the length of the thesis duration, TS for the length of study, GPA for Grade Point Average, and EP for English proficiency.

The dataset in Table 1 saved in CSV 8, and the code and the lines code is 17 the process of reading and combining the contents of the dataset with the marker. This combination should be done before the implementation of apriori algorithms. It will prevent for confusing the value reading in Apriori algorithms.

The line code 18 to 23 is the implementation of Apriori algorithms. At the beginning of Apriori algorithms implementation, three steps must be prepared to analyze the pattern of relationships between data, namely: data set, the value of minimum support (min_supp), and the value of minimum confidence (min_conf).

In this study, an experiment was conducted to find the optimal min_supp and min_conf to produce an appropriate rule that is easy to learn the results. This experiment was carried out by combining the value of min_supp and min_conf. The value min_supp is the value used as a reference to generate frequent itemset / how often the data appears. Based on Figure 3 it can be seen that for 1% or 0.01 min_supp will produce 140 combinations of data that often appear while for 2% or 0.02 min_ supp will produce 49 combinations of data that frequent and so on.

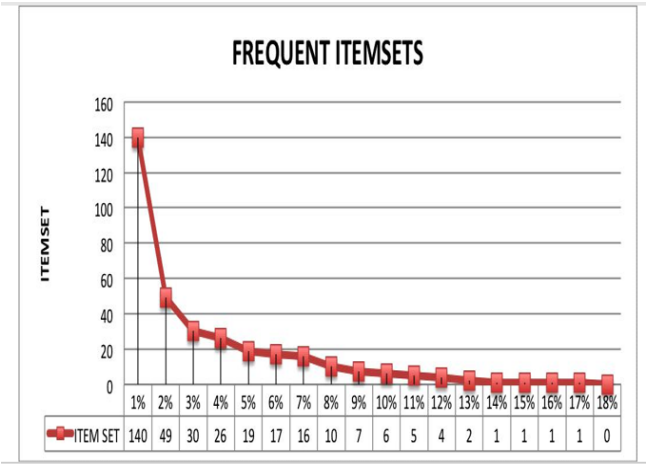

Fig. 3. Correlation of minimal support and number of frequent itemset

Based on the value of min_supp, then the results of the association are determined by determining the value of min_conf. Figure 4 is an experiment to determine the min_conf value of items that appear (frequent itemset). Figure 4 shows the correct amount of min_supp that affects min_conf is a minimum of 1% and 2% support while for a value of min_supp is more than equal to 3% does not produce itemset rules. Figure 4 shows min_sup 1% and 2% still generate the

rules until value of min_conf 28% for 1% min_sup and value of min_conf 25% for 2% min_sup.

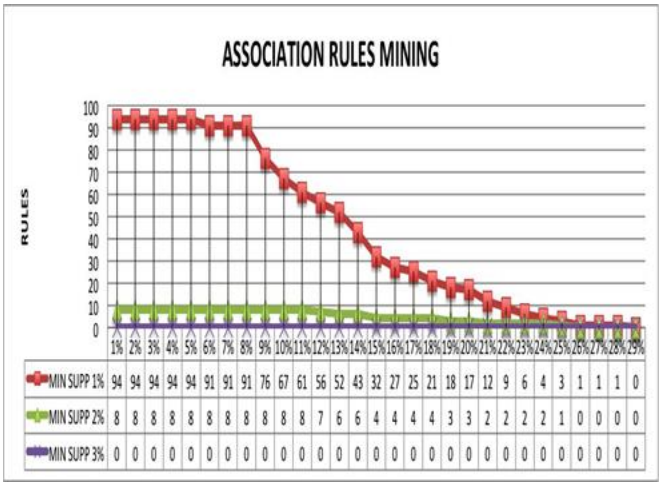

Fig. 4. Experiment result of determination min_supp and min_conf

Figure 4 shows the graphic trend looks constant at min_supp 2% min_conf value between 1% -11%, which results in 8 rules. So that based on this experiment, the min_sup, and min _conf that used in this study are 2% support value and 11% confidence value. Based on the predetermined value of min_supp and min_conf, the rule results that occur shows in Figure 5.

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
0	(EP400)	(GPA3.0)	0.176061	0.116910	0.021573	0.122530	1.048066	0.000989	1.006404
1	(GPA3.0)	(EP400)	0.116910	0.176061	0.021573	0.184524	1.048066	0.000989	1.010377
2	(EP400)	(GPA3.1)	0.176061	0.125261	0.025052	0.142292	1.135968	0.002999	1.019857
3	(GPA3.1)	(EP400)	0.125261	0.176061	0.025052	0.200000	1.135968	0.002999	1.029923
4	(GPA3.2)	(EP400)	0.096729	0.176061	0.025052	0.258993	1.471038	0.008022	1.111917
5	(EP400)	(GPA3.2)	0.176061	0.096729	0.025052	0.142292	1.471038	0.008022	1.053122
6	(TS4.1)	(EP400)	0.082811	0.176061	0.020181	0.243697	1.384163	0.005601	1.089430
7	(EP400)	(TS4.1)	0.176061	0.082811	0.020181	0.114625	1.384163	0.005601	1.035932

Fig. 5. Rules result using python programming

Based on Figure 5, if we create the rules in a sentence, the following result of rules conclusions like:

1) If English proficiency score is 400, then the GradePoint Average (GPA) is 3.0.
2) If the Grade Point Average (GPA) is 3.0, English proficiency score is worth 400.
3) If English proficiency score is worth 400, then the Grade Point Average (GPA) is 3.1.
4) If the Grade Point Average (GPA) is 3.1, then English proficiency score is worth 400.
5) If the Grade Point Average (GPA) is 3.2, then English proficiency score is worth 400.
6) If English proficiency score is 400, then the Achievement Index is 3.2.
7) If the length of study duration is 4.1 years, then English proficiency score is worth 400.
8) If English proficiency score is worth 400, then the Length of Study is 4.1.

Base on the eight rules, we can evaluate the patterns that when English proficiency score is 400 and the Grade Point Average (GPA) is 3.0, 3.1, and 3.2 have two -ways rules. It means that English proficiency and GPA are mutually reinforcing.

The relation between English proficiency and the Length of Study (and vice versa), also gives a contribution to this result, although only two rules that contain these two related variables. The exciting thing is these eight rules; we can not find the correlation related to the length of the thesis duration. This variable disappears in the eight rules. It means that there is a no significant relationship for the length of the thesis duration in this case.

The eight rules have lift value (confidence level) is above 1,048. It means when the lift value more than 1, we can get the conclusion that the rule is a a very strong rule or very confidence rule.

IV. CONCLUSIONS

Based on the result and discussion section, we can get the three-point conclusion of this research are as follows:

1. Variable of English proficiency, Grade Point Average (GPA), and length of study duration is a factor that is strongly related to student data.
2. The optimal value of minimum support and minimum confidence in this study is 2% for min_supp, and 11% for min_conf, then it produces eight rules.
3. The lift values of the eight rules are more than 1,048.

ACKNOWLEDGMENT

Ahmad Dahlan University supports this research in the research scheme Competitive Research Grant (Penelitian Hibah Bersaing/PHB) with grant No: PHB-031/SP3/LPPM-UAD/VI/2018 on 17 June 2018.

REFERENCES

[1] 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, 2016.

[2] O. Jamsheela and R. G, “Frequent Itemset Mining Algorithms : A Literature Survey,” in 2015 IEEE International Advance Computing Conference (IACC), 2015, pp. 1099–1104.

[3] K. Pitchayadejanant and P. Nakpathom, “Data mining approach for arranging and clustering the agro-tourism activities in orchard,” Kasetsart J. Soc. Sci., vol. 39, no. 3, pp. 407–413, 2018.

[4] S. Winiarti, H. Yuliansyah, and A. A. Purnama, “Identification of Toddlers’ Nutritional Status using Data Mining Approach,” Int. J. Adv. Comput. Sci. Appl., vol. 9, no. 1, pp. 164–169, 2018.

[5] I. Riadi, S. Winiarti, and H. Yuliansyah, “Development and Evaluation of Android Based Notification System to Determine Patient ’ s Medicine for Pharmaceutical Clinic,” 2017 4th Int. Conf. Electr. Eng. Comput. Sci. Informatics, no. September, pp. 19–21, 2017.

[6] 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, pp. 1–6.

[7] W. Wang, “Optimization of intelligent data mining technology in big data environment,” J. Adv. Comput. Intell. Intell. Informatics, pp. 129–133, 2019.

[8] V. Gligorijević, N. Malod-Dognin, and N. Pržulj, “Integrative methods for analyzing big data in precision medicine,” Proteomics, vol. 16, no. 5, pp. 741–758, 2016.

[9] R. Agrawal, T. Imieliński, and A. Swami, “Mining association rules between sets of items in large databases,” ACM SIGMOD Rec., vol. 22, no. 2, pp. 207–216, 1993.

[10] Z. A. Othman, N. Ismail, and M. T. Latif, “Association Pattern of NO2 and NMHC towards High Ozone Concentration in Klang,” in 2017 6th International Conference on Electrical Engineering and Informatics (ICEEI), 2017, no. 2, pp. 1–6.

[11] Z. A. Othman, N. Ismail, and M. T. Latif, “Association rules of temperature towards high and low ozone in Putrajaya,” in 2017 6th International Conference on Electrical Engineering and Informatics (ICEEI), 2017, pp. 1–5.

[12] Y. Takama and S. Hattori, “Mining Association Rules from TV Watching Log for TV Program Recommendation,” J. Adv. Comput. Intell. Intell. Informatics, vol. 12, no. 1, pp. 26–31, 2016.

[13] H. Zhou, S. Mabu, W. Wei, K. Shimada, and K. Hirasawa, “Time Related Class Association Rule Mining and Its Application to Traffic Prediction,” IEEJ Trans. Electron. Inf. Syst., vol. 130, no. 2, pp. 289–301, 2010.

[14] X. Li, S. Mabu, H. Zhou, K. Shimada, K. Hirasawa, X. L. Mabu, S. H. Z. Shimada, and K. K Hirasawa, “Genetic Network Programming with Estimation of Distribution Algorithms and its application to association rule mining for traffic prediction,” Evol. Comput., pp. 3457–3462, 2009.

[15] S. Mabu, W. Li, and K. Hirasawa, “A class association rule based classifier using probability density functions for intrusion detection systems,” J. Adv. Comput. Intell. Intell. Informatics, vol. 19, no. 4, pp. 555–566, 2015.

[16] 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.

[17] S. K. Verma, R. S. Thakur, and S. Jaloree, “Fuzzy association rule mining based model to predict students’ performance,” Int. J. Electr. Comput. Eng., vol. 7, no. 4, pp. 2223–2231, 2017.

[18] P. Belsis, I. Chalaris, M. Chalaris, C. Skourlas, and A. Tsolakidis, “The Analysis of the Length of Studies in Higher Education based on Clustering and the Extraction of Association Rules,” Procedia - Soc. Behav. Sci., vol. 147, pp. 567–575, 2014.

[19] S. Hussain, R. Atallah, and A. Kamsin, Cybernetics and Algorithms in Intelligent Systems, vol. 765. Springer International Publishing, 2019.

[20] A. Ikhwan, “a Novelty of Data Mining for Fp-Growth Algorithm,” vol. 9, no. 7, pp. 1660–1669, 2018.

[21] J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques, 3rd ed. USA: Morgan Kaufmann, 2012.

[22] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, and a I. Verkamo, “Fast discovery of association rules,” Adv. Knowl. Discov. data Min., vol. 12, no. 1, pp. 307–328, 1996.