 ISBN: 978-623-92207-1-6

Application of the Apriori Algorithm and FP-Growth to find out the Association Rule between Gender, Education level on wages of SMEs workers in Palembang City

1st Antonius Wahyu Sudrajat
AMIK MDP Palembang
Palembang, Indonesia
wahyu.sudrajat@mdp.ac.id

2nd Idham Cholid
STIE MDP Palembang
Palembang, Indonesia
idham@stie-mdp.ac.id

3rd Ermatita
Sriwijaya University
Palembang, Indonesia
ermatita@ilkom.unsri.ac.id

Abstract—Small, Medium Enterprises (SMEs) are one of the supporting parts of the Indonesian economy by absorbing high labor and production value. One of the factors that greatly influences the development of SMEs in Indonesia is the workforce involved in SMEs business activities. Several factors that influence the workforces are gender, education level, marital status and wages level. The data mining used can help processing data into new knowledge that can be used in decision making. The purpose of this study was to analyze the data of SMEs workers in the Palembang city by comparing the Apriori algorithm and the FP-Growth algorithm to see the association between gender, education level, marital status and wages earned by SMEs's workers in Palembang city. The sample used in this research are 400 SMEs's workers who were randomly selected from 5 sub-districts in Palembang city. The results show that with a confidence level of 0.8, the Apriori Algorithm produces 25 association patterns while the FP-Growth Algorithm has 11 association patterns. In the Apriori Algorithm, it was found that 72% of associations with 1 level confidence while the rest is less than 1. As for the FP-Growth Algorithm, 9% of the association patterns have 1 confidence level and the rest is less than 1. These results indicate that the Apriori algorithm is able to explain more association between gender, education level, marital status and wages of SMEs's workers in Palembang City. The results of this study can be used as a consideration for increasing the capacity of SMEs's workers in Palembang City.

Keywords—*Apriori, FP-Growth, gender, wages, education level, SMEs*

I. INTRODUCTION

Small and Medium Entreprises in a country has many roles for economics growth[1][2]. The Role of SMEs in Indonesia has many aspec in economic, as well as in the GDP, Labour market and many aspect. SMEs workers are one of the factors that support the success of SMEs in the city of Palembang. Related to this, the government needs to understand how the wage rate of SMEs workers in Palembang City so that it can make efforts to improve and increase SMEs workers.

Several studies related to data mining have been carried out for SMEs. Most of the research was conducted to analyze sales transaction data and to discover product marketing patterns [3][4][5]. In this study the authors tried to determine the relationship between gender, education level, marital status and wages of SMEs workers in the city of Palembang by comparing two methods of association rule, namely: the Apriori algorithm and the FP-Growth algorithm. With this algorithm, it can be seen the support and confidence in the association that occurs. The association patterns generated from these two algorithms will then be compared, so that the best pattern can be found which can provide a basis for stakeholder consideration to increase the capacity of SMEs workers in Palembang City.

II. LITERATUR REVIEW

A. Data Mining

Data mining is the process of finding useful information automatically in large data repositories [6]. Data mining is also commonly known by other names such as knowledge discovery (mining) in databases, knowledge extraction and business intelligence and is an important tool for manipulating data for presenting information according to user needs to assist in analyzing the collection of behavioral observations. [3][5].

B. Association Rule

Association rules are one of the main techniques in data mining and perhaps the most common form of pattern finding for unsupervised learning systems. This technique takes all possible patterns of interest in the database, which means that this technique leaves no stone unturned [7]. Two parameter measures in the association analysis are support and confidence.

Support is a measurement to show how much domination an item is from the entire transaction.

$$\text{Support (A)} = \frac{jumlah\ transaksi\ mengandung\ A}{Total\ Transaksi} \quad (1)$$

While confidence is a measurement to show the relationship between two items based on certain conditions.

$$\text{Confidence} = \frac{\sum Transaksi\ mengandung\ A\ dan\ B}{\sum Transaksi\ mengandung\ A} \quad (2)$$

C. *Apriori Algorithm*

Apriori is an algorithm that is widely used in frequent item set searches in the association rule technique. The Apriori algorithm is widely used in transaction data or commonly referred to as market basket. The Apriori Algorithm is the most well-known algorithm for finding high frequency patterns. The high frequency pattern is the pattern of items in the database that have a frequency or support

 ISBN: 978-623-92207-1-6

above a certain threshold called minimum support. The weakness of the Apriori Algorithm is that it has to scan the database every time it performs iterations, so that the time required will increase with the number of iterations. This problem is solved by algorithms such as FP-Growth [3] [4]. In the Apriori algorithm, determining possible candidates is done by paying attention to minimum support and minimum confidence.

D. *FP-Growth Algorithm*

The Frequent Pattern (FP) -Growth algorithm provides an alternative way of calculating frequent sets of items by compressing transaction records using a special graphical data structure called the FP-Tree. FP-Tree can be thought of as the transformation of a data set into graphic format [8]. This method reduces the total number of user datasets by producing a compacted database type in conjunction with the FP Tree [9].

E. *Small – Medium Entreprises*

In Indonesia, the definition of SMEs is regulated in the Republic Act Indonesia No. 20 of 2008 concerning SMEs.[10] in the article 1 of the Act, stated that micro-enterprises are productive businesses owned by individuals and / or entities individual business that has the criteria of a micro business as stipulated in the law.[11]

III. METODOLOGY

In this study, the order of the framework carried out is made in stages with steps that are designed so that the research can run well. The work steps in this study are shown in Figure 1.

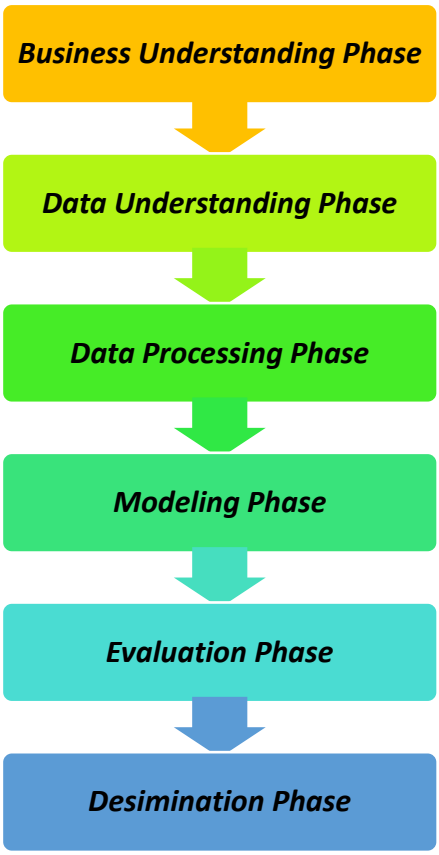

Figure 1. Framework

A. *Business Understanding Phase*

The objective of the study is to determine the association rule between gender, education level and marriage status on the wages of SMEs workers in Palembang city, so that it can be used as a base for consideration in increasing the capacity of SMEs workers in Palembang city.

B. *Data Understanding Phase*

The data source used in this study is primary data, in the form of data from questionnaires distributed to SMEs workers in Palembang City using a margin of error ($\bar{\alpha} = 0.05$).

C. *Data Processing Phase*

Data processing is done by using tabulation of the data in accordance with the variable data collected for later processing using the Rapid Manner soft RapidMinier software.

D. *Modeling Phase*

In this study, there are two algorithms that will be used, namely: the Apriori algorithm and the FP-Growth algorithm in determining the association rule on data SMEs workers and continued by comparing the association rule generated by the two algorithms. The testing process is carried out using the RapidMinier software.

E. *Evaluation Phase*

This stage is carried out to analyze the comparison results and final conclusions. Whether the results of the algorithm can meet the objectives in this study and decisions related to the results of data mining.

F. *Desimination Phase*

The results of this study can be taken into consideration in increasing the capacity of SMEs workers in Palembang City by related agencies, such as the Department of Koperasi and SMEs in Palembang City.

IV. RESULT

A. *Data*

In this study, the data used is data from a survey conducted previously to 600 SMEs workers who are distributed through interview. The data that meet the requirements for data processing are 400 questionnaires spread over 5 districts using proportional numbers.

B. *Variable*

Table 1 is a sample data of SMEs workers who have normalized the data. In this study using 400 samples of SMEs worker data, with the following format:

1. Gender : Male - Female
2. Education Level (Elementry School – Junior High School – Senior High School - Diploma – Graduated)
3. Marital Status (Married, Not Married, Divorced)
4. Wages (>= Rp 3,2 Juta and < Rp. 3,2 Juta)

ISBN: 978-623-92207-1-6

Table 1 Data sampel 400 SMEs Workers

No	Gender	Education	Marital Status	Wages
1	Perempuan	SMA	Cerai Hidup	500,000
2	Laki-laki	SMA	Menikah	3,000,000
3	Perempuan	SMA	Belum Menikah	2,500,000
4	Perempuan	SMA	Belum Menikah	750,000
5	Laki-laki	SMA	Belum Menikah	1,500,000
6	Perempuan	SMA	Belum Menikah	800,000
7	Laki-laki	SMA	Belum Menikah	1,500,000
8	Laki-laki	SMP	Belum Menikah	1,200,000
9	Perempuan	SMP	Belum Menikah	1,000,000
10	Perempuan	SMK	Belum Menikah	800,000
	…	…	…	…
396	Perempuan	SMA	belum menikah	1,500,000
397	Laki laki	SMK	cerai hidup	2,000,000
398	Laki laki	SMA	Belum Menikah	2,500,000
399	Perempuan	SMA	Belum Menikah	1,500,000
400	Perempuan	SMA	Belum Menikah	1,500,000

Based on data from SMEs workers, the tree design is then made as shown in **Figure 2.**

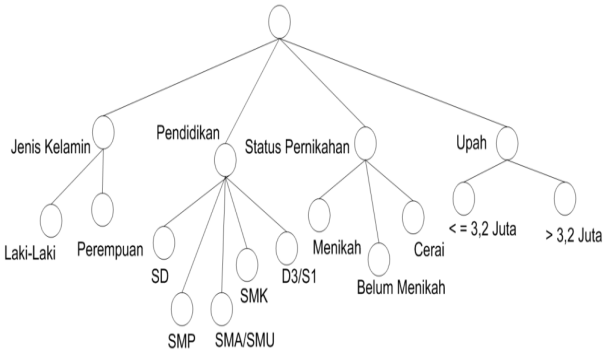

Figure 2. Desain Tree

Furthermore, the data on SMEs workers that have been obtained are then made in tabular data. Then the data is transformed into binary form (1 and 0) in Microsoft Excel.

The support and confidence values included in testing using RapidMiner are:

a. Value of Min. Support = 0,01
b. Value of Min. Confidence = 0,9

C. *Flow of Association Rule Mining Process with RapidMiner*

Testing the Apriori and FP-Growth algorithms using Rapid Miner by using predetermined minimum values of support and confidence. When testing the two algorithms simultaneously, multipy operators are used. The multiply operator is an operator that is used to connect multiple operators so that they can run simultaneously. Figure 3 is the process flow of mining association rule using rapidminer.

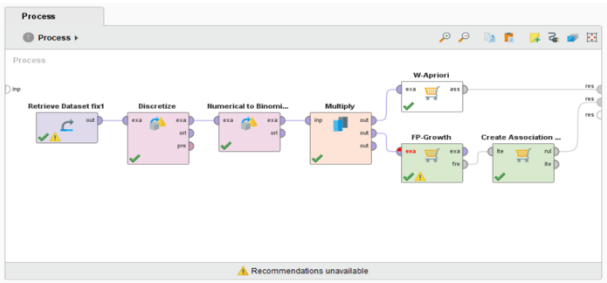

Figure 3. Flow of Association Rule Mining Process

D. *Analysis with the Apriori Algorithm*

The results of the frequent itemset process using the apriori algorithm are shown in Figure 4.

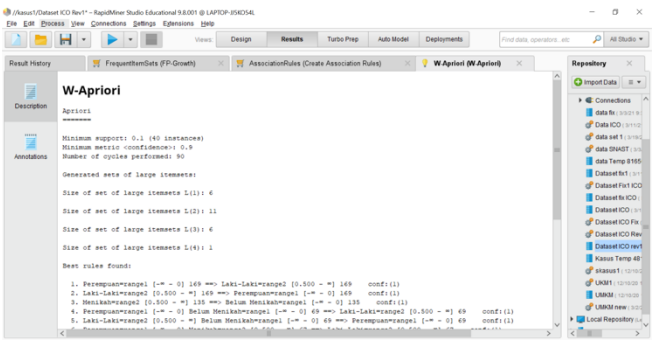

Figure 4. The results of the Rule on Rapid Miner use the Apriori Algorithm

The number of rules generated is as many as 48 rules and has mapped the data of SMEs workers. Figure 5 shows the settings of the parameters in the Apriori Algorithm in Rapid Miner.

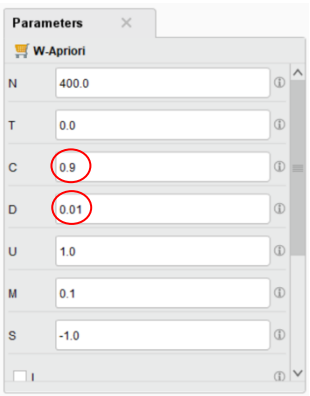

Figure 5. Setting parameter Apriori

E. *Analisis dengan Alforitma FP-Growth*

The number of rules generated is as many as 48 rules and has mapped the data of SMEs workers. Figure 5 shows the settings of the parameters in the Apriori Algorithm in Rapid Miner.

In Figure 6, the result of the frequent item set process using the FP-Growth algorithm is shown. The results of the frequent item set generated by the FP-Growth algorithm are 7 itemset. Figure 7 is the parameter setting of the association rule.

ISBN: 978-623-92207-1-6

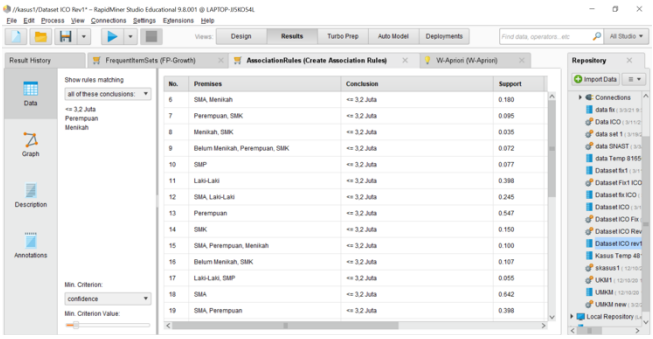

Figure 6. Frequent itemset FP-Growth

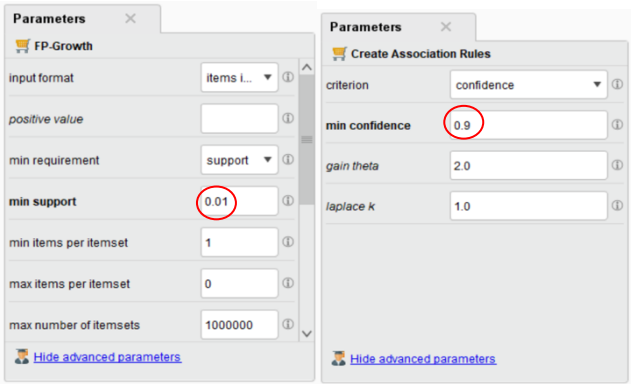

Figure 7. Setting Parameter Association Rule

From testing the FP-Growth algorithm using Rapidminer, 11 rule associations are generated. This is much less than the Apriori algorithm. Figure 8 is the association rule using the FP-Growth algorithm. While figure 9 is a graph of the rules of data for SMEs workers using Rapid Miner.

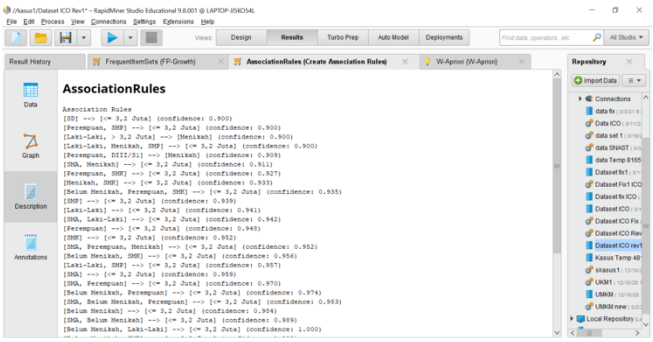

Figure 8. Association Rule using algorithm FP-Growth

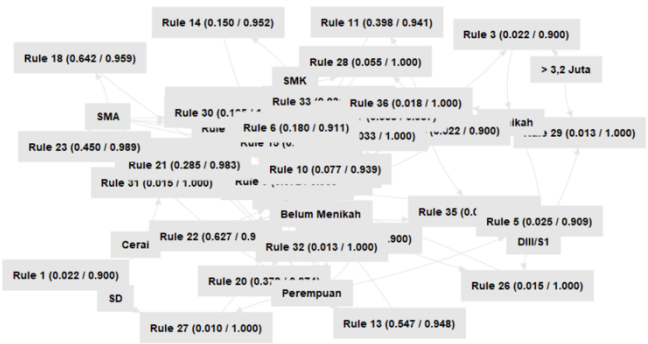

Figure 9. Graphic Rule Data of SMEs Worker using RapidMiner

F. *Comparison Results of Apriori Algorithm and FP-Growth Algorithm*

Here are some comparisons generated by the process of running the Apriori algorithm and the FP-Growth algorithm:

1. The Apriori algorithm produces 48 rules while the FP-Growth algorithm produces 11 rules. This of course is due to differences in how it works, even though the support and confidence values provided are the same.
2. Apriori Algorithm, produces more than FP-Growth, producing confidence of less than 1.
3. There is no known difference in processing time between using the Apriori algorithm and the FP-Growth algorithm, because it uses the same testing software and during the process uses the multiply operator.

Based on the resulting relationship pattern, the Palembang city government can choose one of the relationship patterns, where the level of confidence is equal to 1.

V. CONCLUSION

Based on the analysis and testing that has been done, it can be concluded that several things are related to this research:

1. There are differences in association rule when using the Apriori algorithm method and the FP-Growth algorithm. This is because the Apriori Runtime increases exponentially depending on the number of different items, while FP Growth increases linearly, depending on the number of transactions and items.
2. By using the Apriori method, it produces 48 associates with 22% having a confidance of less than 1, while in the FP-Growth method, 11 associates with 54.5% have a confidence of less than 1.
3. To increase the capacity of workers, especially SMEs, the government needs to pay attention to associates who have confidence as much as 1, either by using Appriori and also FP Growth.Acknowledgment

REFERENCES

[1] D. H. Karadag, "The Role of SMEs and Entrepreneurship on Economic Growth in Emerging Economies within the Post-Crisis Era: an Analysis from Turkey," *J. Small Bus. Entrep. Dev.*, vol. 4, no. 1, 2016, doi: 10.15640/jsbed.v4n1a3.

[2] H. Keskin, C. Sentürk, O. Sungur, and H. M. Kiris, "The Importance of SMEs in Developing Economies," *2nd Int. Symp. Sustain. Dev.*, pp. 183–192, 2010.

[3] Islamiyah, P. L. Ginting, N. Dengen, and M. Taruk, "Comparison of Priori and FP-Growth Algorithms in Determining Association Rules," *ICEEIE 2019 - Int. Conf. Electr. Electron. Inf. Eng. Emerg. Innov. Technol. Sustain. Futur.*, pp. 320–323, 2019, doi: 10.1109/ICEEIE47180.2019.8981438.

[4] D. Wicaksono, M. I. Jambak, and D. M. Saputra, "The Comparison of Apriori Algorithm with Preprocessing and FP-Growth Algorithm for Finding Frequent Data Pattern in Association Rule," vol. 172, no. Siconian 2019, pp. 315–319, 2020, doi: 10.2991/aisr.k.200424.047.

[5] W. P. Nurmayanti, H. M. Sastriana, and A. Rahim, "Market Basket Analysis with Apriori Algorithm and Frequent Pattern Growth (Fp-Growth) on Outdoor Product Sales Data," pp. 132–139.

[6] P.-N. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining (New International Editon)*, no. September. 2013.

[7] M. Liang, *Data Mining: Concepts, Models, Methods, and Algorithms*, vol. 36, no. 5. 2004.

[8] R. Chauhan and H. Kaur, *Predictive Analytics and Data Mining*. 2015.

[9] K. Dharmarajan and M. A. Dorairangaswamy, "Analysis of FP-growth and Apriori algorithms on pattern discovery from weblog data," *2016 IEEE Int. Conf. Adv. Comput. Appl. ICACA 2016*, pp. 170–174, 2016, doi: 10.1109/ICACA.2016.7887945.

[10] *Republic Act Indonesia No. 20 of 2008*. .

[11] T. T. H. Tambunan, *UMKM di Indonesia*. Bogor: Ghalia Indonesia, 2009.