Application of Association Rules with Apriori Algorithm to Determine the Pattern of the Relationship between *SBMPTN* Database and Student's Grade Point Average

Ms. Mubeena Begum

Guest Faculty, Department of computer science, Gulbarga University, GUK, Kalaburagi, Karnataka, India

***Abstract*- In a university management, information system is one of the resources that can be utilized to improve competitiveness and provide accurate data to stakeholders. Yogyakarta State University accepts students through a variety of ways, namely *SNMPTN*, *SBMPTN* and self-selection. By utilizing the *SBMPTN* database and students' Grade Point Average (GPA), in this paper will apply data mining technique using Apriori algorithm to determine the pattern of the relationships between *SBMPTN* database and students' GPA. Datawere obtained from the students' *SBMPTN* database of the classes of 2010 and students' GPA.**

Keywords: *SBMPTN* database, GPA, Association Rule, Apriori Algorithm

BACKGROUND

Currently, higher institutions are required to have a competitive edge by utilizing the available resources. Information is an essential requirement that must be provided by an organization including higher institutions/universities, in addition to the resources of facilities, infrastructure and people. Information is the result of processing the data, while the data is the fact that depicts real events. In an organization, a database is used to store important data and it becomes basic information that is then further processed and served as the basis of decision making. Improving the quality of managerial decisions is one way to improve the quality of the organization/higher institution. Educational institutions seek more efficient technologies in order to have a better management, to support decision-making procedures, and to set the right strategy and management plan that is better than the current management.

Data warehouse is a database used by organizations whose large-scale data. Potential and important information in the data warehouse can be analyzed using a technique known as data mining. Data mining helps organizations to find and understand the hidden patterns of data. The information resulting from the application of data mining techniques used to excavate and to predict the potential of an organization. Data mining is a process to find relationships, patterns and new trends by filtering out meaningful very large data stored in the storage, use pattern recognition techniques such as statistical and mathematical techniques (Larose, 2005).

The pattern of relationships in data mining can be a relationship between two or more in one dimension, i.e. the dimension of the product. From here, we can see the relationship between the purchases of a product from other products. In addition, the relationship can also be seen between two or more attributes, and two or more objects (Ponniah, 2001). Data Mining uses a variety of techniques, such as Decision Trees, Neural Networks, Naive Bayes, K-Nearest neighbor and so on. By using a variety of techniques, the patterns of knowledge or information can be found, such as in form of association rules, classification, and clustering. Association rule is one of the main techniques in data mining and as the most common form used in finding the pattern data set (Kantardzic, 2003). Association rule is used to find certain rules that associate the data with other data. Apriori algorithm is one of algorithms which can be used to find the association rule. (Agrawal & Srikant, 1994).

WEKA is a collection of machine learning algorithms for data mining tasks. Thealgorithms can either be applied directly to a dataset or called from your own Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes.

Yogyakarta State University accepts new students approximately 6,000 people eachyear, either through the *SNMPTN*, *SBMPTN*, and self-selection examinations. This large amount of data from incoming students requires some effort to sort it so that it can be processed as useful information. If the data is ignored, then the data will only be meaningless rubbish. By utilizing these *SBMPTN* database and Grade Point Average (GPA), this research will apply data mining technique using Apriori algorithm to determine the pattern of the relationships between *SBMPTN* database and student's GPAat Yogyakarta State University. Data were obtained from the students' *SBMPTN* database of the classes of 2010 and students' GPA. Database *SBMPTN* includes high school district, parent's education level, level of parent's income and student's result forNational Final Examination. From the obtained results, the university can predict the student's GPA.

DATA MINING

Data mining is an interdisciplinary field, the confluence of a set of disciplines, including database systems, statistics, machine learning, visualization, and information science. The term data mining or mining refers to the extraction of knowledge from large amounts of data.(Jan & Kamber, 2006). Data Mining, also known as Knowledge Discovery in Databases (KDD), refers to the extraction of implicit information from the data in the database that were previously unknown, but potentially in knowledge discovery. Steps to find knowledge in data mining is presented in Figure 1. (Zaïane, 1999)

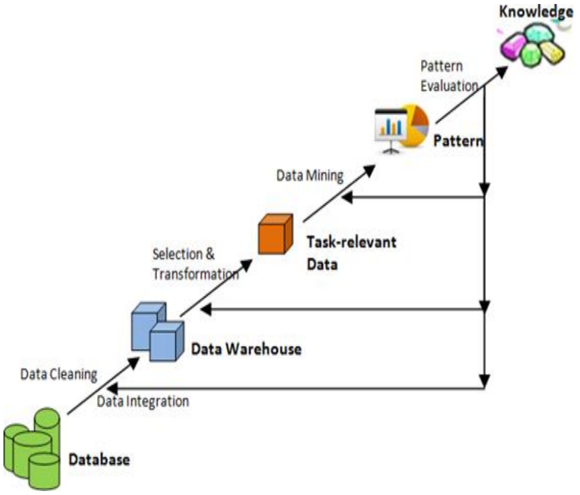


Figure 1 The process of knowledge discovery in data mining

The Knowledge Discovery in Databases process comprises of a few steps leading from raw data collections to some form of new knowledge. The iterative process consists of the following steps:

Data cleaning: also known as data cleansing, it is a phase in which noise data and irrelevant data are removed from the collection.

1. Data integration: at this stage, multiple data sources, often heterogeneous, maybe combined in a common source.
2. Data selection: at this step, the data relevant to the analysis is decided on and retrieved from the data collection.
3. Data transformation: also known as data consolidation, it is a phase in which the selected data is transformed into forms appropriate for the mining procedure.
4. Data mining: it is the crucial step in which clever techniques are applied to extract patterns potentially useful.
5. Pattern evaluation: in this step, strictly interesting patterns representing knowledge are identified based on given measures.
6. Knowledge representation: is the final phase in which the discovered knowledge is visually represented to the user. This essential step uses visualization techniques to help users understand and interpret the data mining results.

ASSOCIATION RULES

Association rules learning is an important topic in data mining which is the discovery of association relationships or correlations among a set of items. While amounts of data are continuously being collected and stored, many industries are becoming interested in learning association rules from their database. The discovery of interesting association relationships among huge amounts of business transaction records can help in many business decision-making processes, such as catalog design, cross-marketing, cross- selling and inventory control. (Gorunescu, 2011). Agrawal et al. (1993) in (Gorunescu, 2011) first developed a framework to measure the association relationship among a set of items. The association rule mining can be defined formally as follows.

$I = \{i_1, i_2, ..., i_m\}$ is a set of items that a store sell such as as milk, bread and coffee. $D = \{t_1, t_2, ..., t_n\}$ is a set of transactions, called a transaction database, where each transaction t has an identifier *tid* and a set of items t-itemset, i.e., t = (tid, t-itemset). For example, a customer's shopping cart going through a checkout is a transaction. X is an itemset if it is a subset of t. For example, a set of items for sale at a store is an itemset. Two measurements have been defined as support and confidence as below. An itemsetX in a transaction database D has a support, denoted as sp(X). This is the ratio of transactions in D containing X,

where X(t) = {t in D|t contains X}.

An itemset X in a transaction database D is called frequent if its support is equal to, or greater than, the threshold minimal support (min_sp) given by users. Therefore support can be recognized as frequencies of the occurring patterns. Two itemsets X and Y in a transaction database D have a confidence, denoted as cf($X \rightarrow Y$). This is the ratio of transactions in D containing X that also contain Y

An association rule is the implication of the form $X \rightarrow Y$, where $X \subset I$, $Y \subset I$, and$X \cap Y = \emptyset$. Each association rule has two quality measurements, support and confidence, defined as: (1) the support of a rule $X \rightarrow Y$ is sp ($X \cup Y$) and (2) the confidence of a rule $X \rightarrow Y$ is cf($X \rightarrow Y$).

Rules that satisfy both a minimum support threshold (min_sp) and a minimum confidence threshold (min_cf), which are defined by users, are called strong or valid.

Mining association rules can be broken down into the following two subproblems:

1. Generating all itemsets that have support greater than, or equal to, user specified minimum support. That is, generating all frequent itemsets.
2. Generating all rules that have minimum confidence in the following simple way:for every frequent itemset X, and any $B \subset X$, let $A = X — B$. If the confidence of a rule $A \rightarrow B$ is greater than, or equal to, the minimum confidence (min_cf), then it can be extracted as a valid rule.

APRIORI ALGORITHM

Apriori algorithm is an algorithm to find a high frequency pattern. This algorithm controls the development of the *candidate itemset* from the *frequent itemset* with the *support-based pruning* to eliminate *itemset* which do not meet, by specifying *minsup*. The principle of the apriori algorithm is if *itemset* classified as itemset, which has the support of more than a frequent preset, then all subsets also belonged to the frequent itemset, and vice versa. The working system of this algorithm is the algorithm will generate new candidate k-*itemset* of frequent *itemset* in the previous step and calculate the value of the support k-*itemset*. *Itemset* that has a value of support under *minsup* be eliminated. The algorithm stops when no new frequent *itemset* generated.(Wandi, Hendrawan, & Mukhlason, 2012). **A**priori algorithm flow chart is presented in Figure

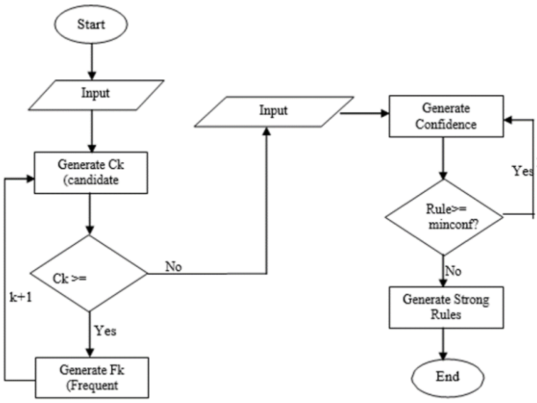


Figure 2. Apriori Algorithm Flow Chart

FINDING ASSOCIATION RULES WITH WEKA

The process of mining in this article are used to determine the pattern of the relationship between the *SBMPTN* database with a cumulative grade point of students. *SBMPTN* database includes:

1. The parent’s education level, consisting of father's last education and mother's last education, which is categorized into 8 categories: Doctorate, Master, Bachelor, Diploma, High School Graduate, Junior High School Graduate, elementary school graduates, no primary school.
2. Parents occupation, consisting of father's and mother's occupation, which are categorized into 8 categories, presented in Table 1.

Table 1. Category of parent's occupation

Occupation	Code
Employee	K1
Teacher	K2
Entrepreneur	K3
Farmers / Fishermen	K4
Lecturer	K5
Labor	K6

Army / Police	K7
No Job	K8

Parent's Income, categorized into 4 categories, which are presented in Table 2.

Table 2. Category of parent's income

Income	Code
Up to IDR 1,000,000	P1
IDR. 1,000,001 - IDR. 5,000,000	P2
IDR. 5,000,001 - IDR. 10,000,000	P3

The average results of the National Final Examination (NFE), categorized into 4 categories, which are presented in Table 3.

Table 3. Categories of the average value of the National Final Exams (NFE)

The average results of the national final exam	Code
Up to 5,5	1
5,5 ≤ average < 7	2
7 ≤ average< 8,5	3
≥ 8,5	4

Student’s high school district consists of 127 districts, including Kab Bangka, Kab Bangka Barat, Kab Banjarnegara, Kab Bantul, Kab Banyuasin, Kab Banyumas, Kab Banyuwangi, Kab Barito Utara, Kab Batang, Kab Belitung, Kab Belitung Timur, Kab Bengkulu Utara, Kab Berau, Kab Blitar, Kab Blora, Kab Boyolali, Kab Brebes, Kab Ciamis, Kab Cilacap, Kab Demak, Kab Flores Timur, Kab Gresik, et al.

Student’s Grade Point Average (GPA), grouped into 4 categories. This grouping is based on the GPA predicate according to academic rules in Yogyakarta State University and presented in Table 4.

Table 4 Category of GPA

GPA	Predicate	Kategori
GPA > 3,5	Cumlaude	A
3<GPA≤3,5	Highly Satisfactory	B
2,5≤GPA≤3	Satisfy	C

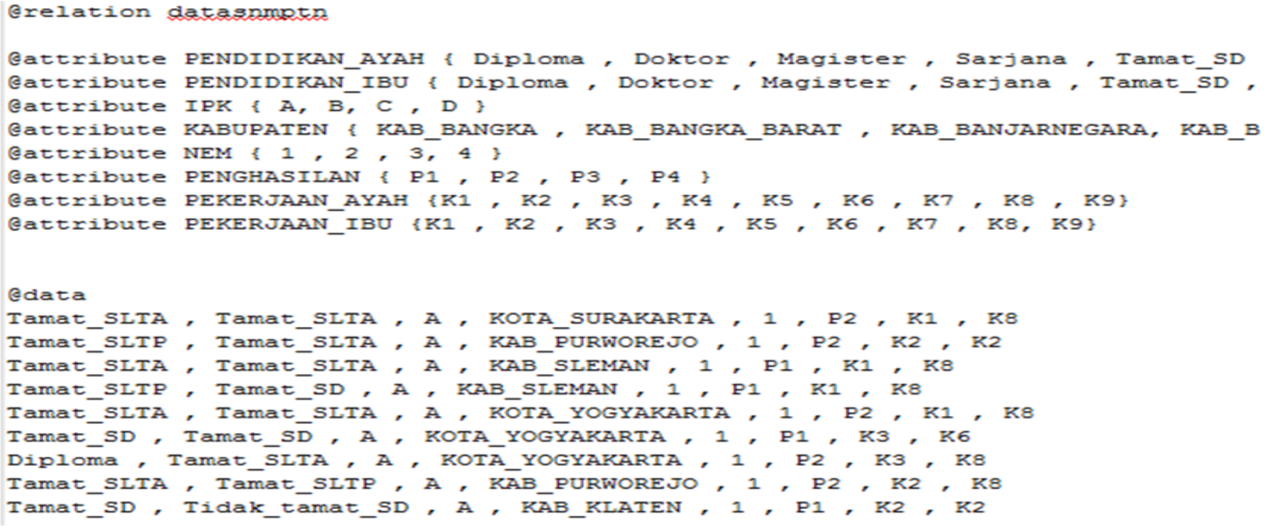
* GPA of less than 2.5 is classified as D.

The amounts of data used were 1500 students’ of 2010 classes that Yogyakarta State University has that were accepted through *SBMPTN*. The GPA calculated from 1st to 3rd semester only. Examples of original data obtained are presented in Table 5 below.

Table 5. Original Data

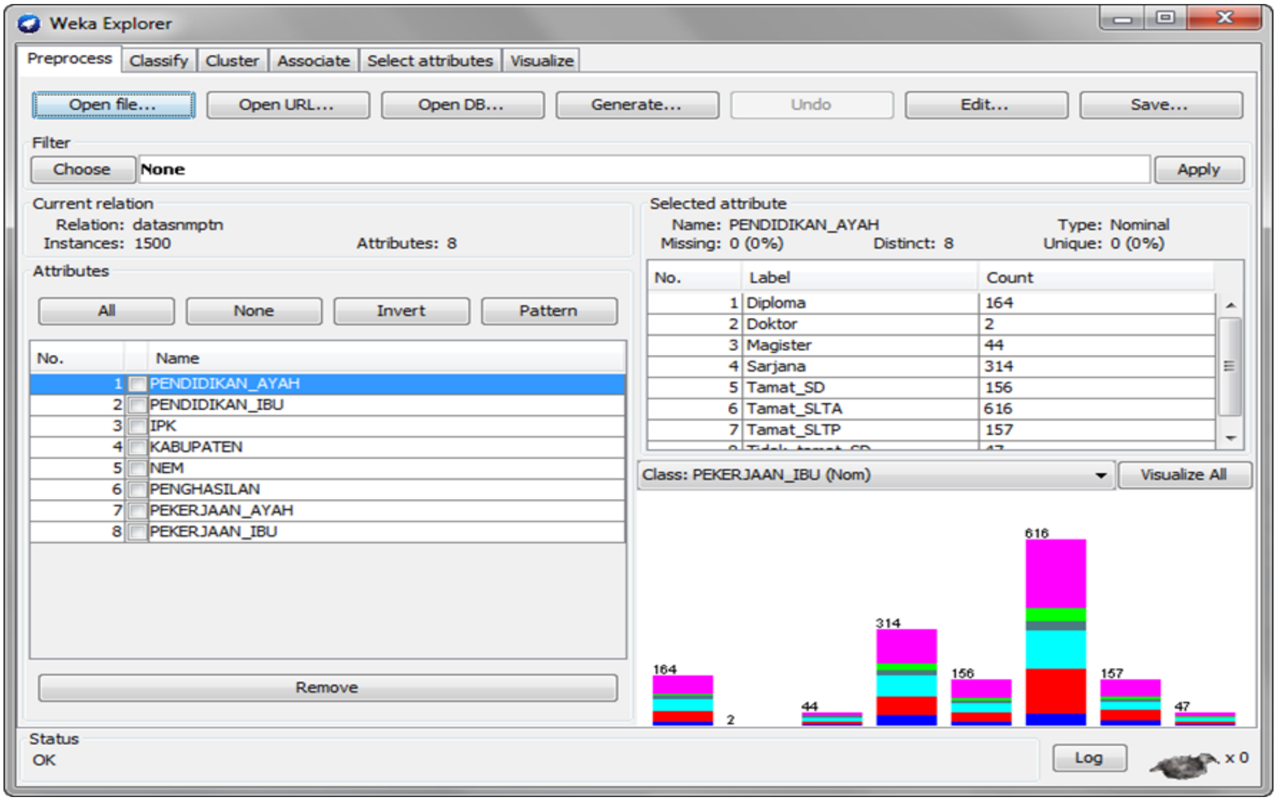
No	Father's Education	Father's Occupation	Mother's Education	Mother's Occupation	Income/month	District of Hg School	AvgNFE	GPA
1	Diploma	Employee	High School Graduate	No Job	IDR. 1,000,001 - IDR. 5,000,000	Kota Surakarta	6.50	3.57
2	Diploma	Teacher	Bachelor	Teacher	IDR. 1,000,001 - IDR. 5,000,000	Kabupaten Purworejo	6.67	3.55
3	Diploma	Employee	High School Graduate	Tidak Bekerja	Up to IDR.1.000.000	Kabupaten Sleman	6.67	2.78
4	Diploma	Employee	High School Graduate	Tidak Bekerja	Up to IDR.1.000.000	Kabupaten Sleman	6.83	3.40
5	Diploma	Employee	Bachelor	Tidak Bekerja	IDR. 1,000,001 - IDR. 5,000,000	Kota Yogyakarta	6.83	3.46

Furthermore, the data were analyzed using WEKA software. The first step was data was converted into .arff format, in this case was stored as data *SNMPTN*.arff as follows.



```
@relation datasnmptn

@attribute PENDIDIKAN_AYAH { Diploma , Doktor , Magister , Sarjana , Tamat_SD
@attribute PENDIDIKAN_IBU { Diploma , Doktor , Magister , Sarjana , Tamat_SD ,
@attribute IPK { A, B, C , D }
@attribute KABUPATEN { KAB_BANGKA , KAB_BANGKA_BARAT , KAB_BANJARNEGARA, KAB_B
@attribute NEM { 1 , 2 , 3, 4 }
@attribute PENGHASILAN { P1 , P2 , P3 , P4 }
@attribute PEKERJAAN_AYAH {K1 , K2 , K3 , K4 , K5 , K6 , K7 , K8 , K9}
@attribute PEKERJAAN_IBU {K1 , K2 , K3 , K4 , K5 , K6 , K7 , K8, K9}


@data
Tamat_SLTA , Tamat_SLTA , A , KOTA_SURAKARTA , 1 , P2 , K1 , K8
Tamat_SLTP , Tamat_SLTA , A , KAB_PURWOREJO , 1 , P2 , K2 , K2
Tamat_SLTA , Tamat_SLTA , A , KAB_SLEMAN , 1 , P1 , K1 , K8
Tamat_SLTP , Tamat_SD , A , KAB_SLEMAN , 1 , P1 , K1 , K8
Tamat_SLTA , Tamat_SLTA , A , KOTA_YOGYAKARTA , 1 , P2 , K1 , K8
Tamat_SD , Tamat_SD , A , KOTA_YOGYAKARTA , 1 , P1 , K3 , K6
Diploma , Tamat_SLTA , A , KOTA_YOGYAKARTA , 1 , P2 , K3 , K8
Tamat_SLTA , Tamat_SLTP , A , KAB_PURWOREJO , 1 , P2 , K2 , K8
Tamat_SD , Tidak_tamat_SD , A , KAB_KLATEN , 1 , P1 , K2 , K2
```

Then, this data *SNMPTN*.arff files was entered to the WEKA workspace and presented in Figure 3 below:



Filter: None

Current relation
Relation: datasnmptn
Instances: 1500
Attributes: 8

Attributes

No.	Name
1	PENDIDIKAN_AYAH
2	PENDIDIKAN_IBU
3	IPK
4	KABUPATEN
5	NEM
6	PENGHASILAN
7	PEKERJAAN_AYAH
8	PEKERJAAN_IBU

Selected attribute
Name: PENDIDIKAN_AYAH
Type: Nominal
Missing: 0 (0%)
Distinct: 8
Unique: 0 (0%)

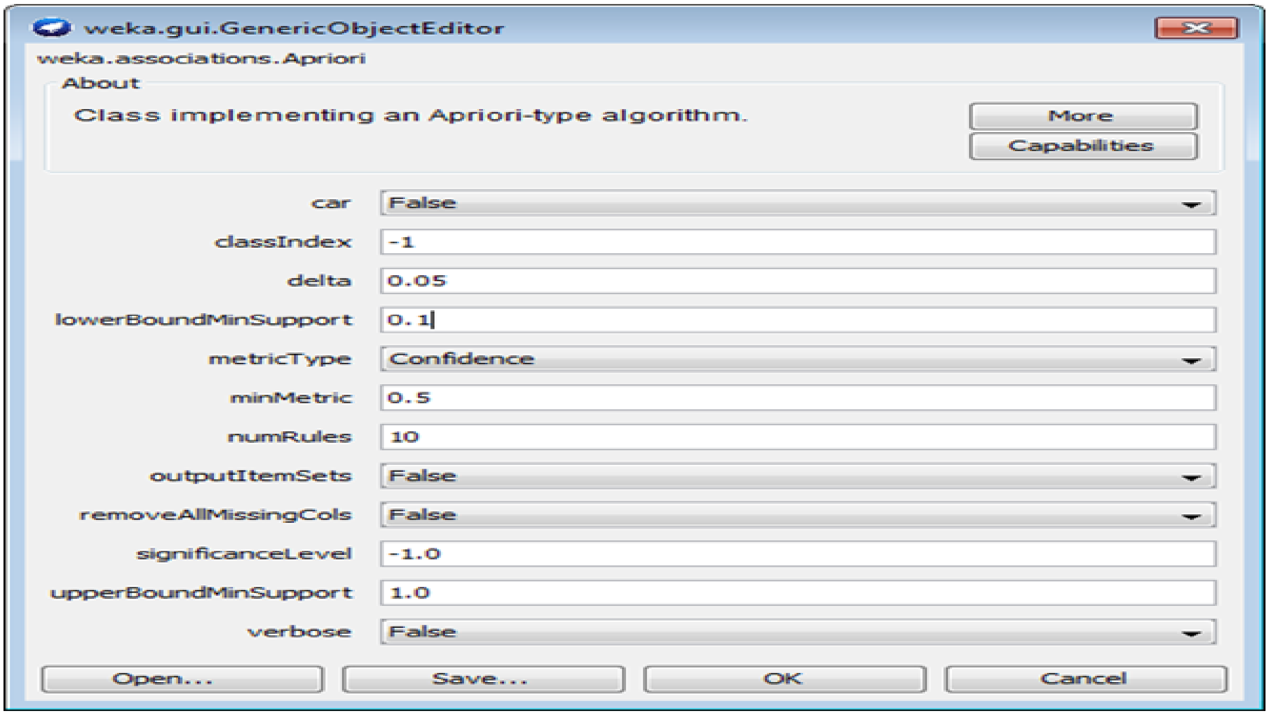
No.	Label	Count
1	Diploma	164
2	Doktor	2
3	Magister	44
4	Sarjana	314
5	Tamat_SD	156
6	Tamat_SLTA	616
7	Tamat_SLTP	157
8	Tidak_tamat_SD	47

Class: PEKERJAAN_IBU (Nom)

Status
OK

Figure 3. The display of data *SNMPTN*.arff in the Weka software

The research used the minimum support of 10%, which means that the itemset whose value is less than 10% (150) will be eliminated. Configuring *minimum support* = 10%, and the *minimum confidence* = 50% on WEKA software is presented in Figure 4.



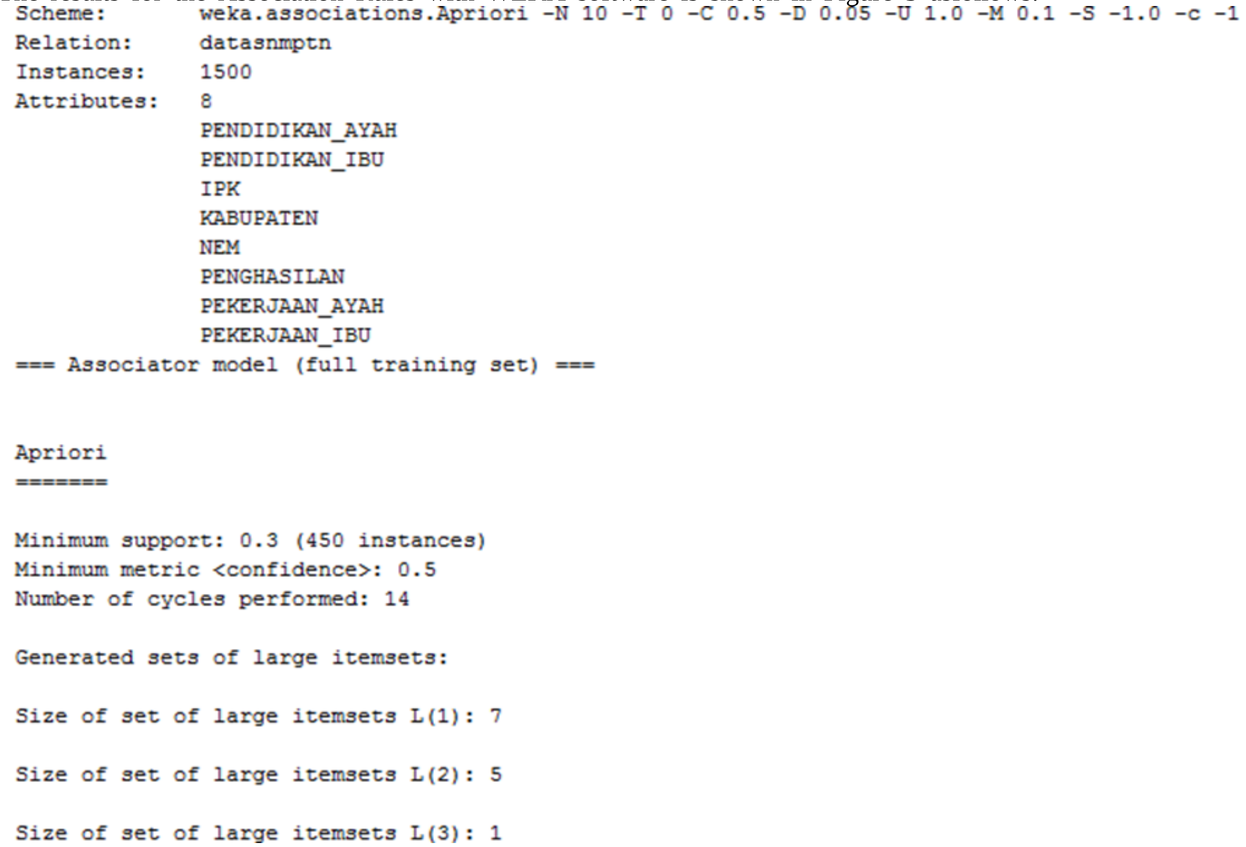
weka.associations.Apriori

About
Class implementing an Apriori-type algorithm.

car	False
classIndex	-1
delta	0.05
lowerBoundMinSupport	0. 1
metricType	Confidence
minMetric	0.5
numRules	10
outputItemSets	False
removeAllMissingCols	False
significanceLevel	-1.0
upperBoundMinSupport	1.0
verbose	False

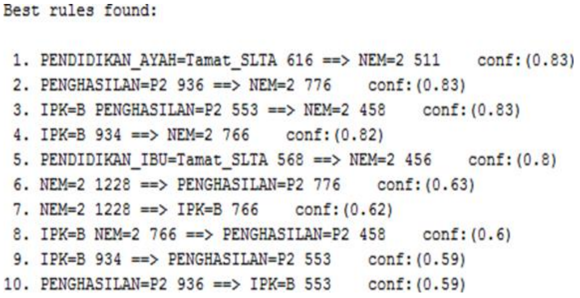
Figure 4 Configuration of Apriori algorithm in WEKA software

The results for the Association Rules with WEKA software is shown in Figure 5 asfollows:



```
Scheme:       weka.associations.Apriori -N 10 -T 0 -C 0.5 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1
Relation:     datasnmptn
Instances:    1500
Attributes:   8
              PENDIDIKAN_AYAH
              PENDIDIKAN_IBU
              IPK
              KABUPATEN
              NEM
              PENGHASILAN
              PEKERJAAN_AYAH
              PEKERJAAN_IBU
=== Associator model (full training set) ===


Apriori
=======

Minimum support: 0.3 (450 instances)
Minimum metric <confidence>: 0.5
Number of cycles performed: 14

Generated sets of large itemsets:

Size of set of large itemsets L(1): 7

Size of set of large itemsets L(2): 5

Size of set of large itemsets L(3): 1
```

Figure 5. Rule search with WEKA



```
Best rules found:

 1. PENDIDIKAN_AYAH=Tamat_SLTA 616 ==> NEM=2 511    conf:(0.83)
 2. PENGHASILAN=P2 936 ==> NEM=2 776    conf:(0.83)
 3. IPK=B PENGHASILAN=P2 553 ==> NEM=2 458    conf:(0.83)
 4. IPK=B 934 ==> NEM=2 766    conf:(0.82)
 5. PENDIDIKAN_IBU=Tamat_SLTA 568 ==> NEM=2 456    conf:(0.8)
 6. NEM=2 1228 ==> PENGHASILAN=P2 776    conf:(0.63)
 7. NEM=2 1228 ==> IPK=B 766    conf:(0.62)
 8. IPK=B NEM=2 766 ==> PENGHASILAN=P2 458    conf:(0.6)
 9. IPK=B 934 ==> PENGHASILAN=P2 553    conf:(0.59)
10. PENGHASILAN=P2 936 ==> IPK=B 553    conf:(0.59)
```

From the analysis, there are10 best rules obtained, which are:

FATHER'S EDUCATION=High School Graduate 616 ==> NFE=2 511 conf:(0.83)2. INCOME=P2 936 ==> NFE=2 776 conf:(0.83)

3. GPA=B INCOME=P2 553 ==> NFE=2 458 conf:(0.83)

4. GPA=B 934 ==> NFE=2 766 conf:(0.82)

5. MOTHER'S EDUCATION= High School Graduate 568 ==> NFE=2 456 conf:(0.8)6. NFE=2 1228 ==> INCOME=P2 776 conf:(0.63)

7. NFE=2 1228 ==> GPA=B 766 conf:(0.62)

8. GPA=B NEM=2 766 ==> INCOME=P2 458 conf:(0.6)

9. GPA=B 934 ==> INCOME=P2 553 conf:(0.59)

10. INCOME=P2 936 ==> GPA=B 553 conf:(0.59)

Base on the rule obtained, it shows that the highest confidence value having a relationship with the GPA is 62% on the average value of NFE type 2 (5.5 ≤ average score < 7), obtaining a GPA of type B (3 < GPA ≤ 3,5). It means the possibility of having GPA of type B when the average value of type 2 comes up is 62%.

Furthermore, we can see the relationships pattern between income variable of P2 type (IDR. 1,000,001 - IDR. 5,000,000) and a GPA of B type (3 <GPA≤3,5) that has a confidence value of 59%. Based on this confidence value, it shows that there is no strong correlation pattern between variables in the *SBMPTN* database with student’s GPA.

CONCLUSIONS AND SUGGESTIONS

Conclusions

Based on the association rules derived, there are some conclusions that can be drawn as follows:

By using Apriori algorithm with WEKA software, we can some rules which are a collection of frequent *itemset* that has a relationships pattern with the GPA with the highest confidence value of 62%.

The highest confidence value of 62% is reached on the relationship pattern ofthe average NFE of 2 type (5.5 ≤ average score <7) and the GPA of B category (3 <GPA≤3,5).

There is no pattern of relationship between the GPA of A, C and D categories and *SBMPTN* database.

There is no strong relationships pattern between variables in the *SBMPTN* database with student's GPA.

SUGGESTION

In further analysis, we can use a larger data, for example by using the data of students who enter the university using a way other than *SBMPTN* or using other classes for the objects, so that the rules generated are more diverse and more useful for decision- making in higher institutions. The greater of data are used, the more useful the information is produced.

REFERENCE

[1] Agrawal, R., & Srikant, R. (1994). FastAlgorithms For Mining Association Rules. In Proc.1994 . *Proc. International Conference Very Large DataBases (VLDB).*

[2] Gorunescu, F. (2011). *Data Mining Concepts, Models and Techniques.* Verlag Berlin:Springer.

[3] Jan, J., & Kamber, M. (2006). *Data Mining: Concepts and Techniques, Second Edition.*

[4] San Francisco: Morgan Kaufmann Publishers.

[5] Kantardzic, M. (2003). *Data Mining: Concepts, Models, Methods, and Algorithms.* New Jersey: John Wiley & Sons.

[6] Larose, D. T. (2005). *Discovering Knowledge in Data: An Introduction to Data mining,* New Jersey: John Willey & Sons. Inc.

[7] Ponniah, P. (2001). *Datawarehouse Fundamentals: A comprehensive Guide for IT Professional.* New York: John Willey & Sons. Inc.

[8] Zhu, Z., & Wang, J.-Y. (2007). Book Recommendation Service by Improved Association Rule Mining Algorithm. *Sixth International Conference on Machine Learning and Cybernetics.* Hongkong