International Journal of Software Engineering and Informatics
Vol. 1, No.1 March 2023 Pp 25-32.
ISSN: xxxx xxxxx,

Market Basket Analysis In Sales Transactions With Apriori Algorithm

Ratnasari[1], Aviv Fitria Yulia[2], Fahlul Rizki[3], Agustinus Eko Setiawan[4]
[1,2,3] Aisyah Pringsewu University, Indonesia

Article Info

Keywords:

data mining,
market basket analysis,
Priori

ABSTRACT

The company's biggest challenge today is how to extract significant data into information that is ready to be used to assist in making decisions and increasing sales. One way to get this knowledge is by doing data mining. In this study, the authors studied vitamin sales transaction data at Viva Health Duku Setro Pharmacy using the association rule method to find association rules from sales transaction data. The algorithm used is the Apriori algorithm. This algorithm is used to find the rules of association using the Rapidminer software tool version 8. The association rules in the A priori algorithm will be used as a reference in determining the goods to be promoted with cross-selling marketing strategies and product bundling marketing techniques. The result of this study is that the a priori algorithm has an accuracy value of 61.17%.

Author Correspondence:

Ratnasari,
Software Engineering,
Aisyah Pringsewu University, Indonesia
Email: ratnasari@aisyahuniversity.ac.id

1 INTRODUCTION

The development of information technology has an influence in many fields, one of which is in the field of product sales. With the large amount of sales transaction data available, companies do not know what the data is for other than stored and is only used when calculating the availability of goods, calculating how many items are sold, how many sales results, and so on. This of course creates a condition called "rich of data *but poor of informations*"[1], where a lot of data is left to mount without knowing that there may actually be information that is very useful for the company. The challenge for companies is how to extract big data into ready-to-use information to assist in making decisions and increasing sales [2].

Viva Healt Pharmacy often conducts sales promotions, but the promotions carried out are not based on consumer patterns or habits, so the expected sales targets are not achieved. In this study, the author conducted a study on vitamin sales transaction data at Viva Health Duku Setro Pharmacy using the market basket analysis method used to find the association rules of sales transactions. The algorithm used is a priori algorithm. Algorhythm will be used to form a frequent itemset so that it can formulate rules of association with the Rapidminer software tool version 8. The rules of association generated by a priori algorithms will be evaluated to find aquarasi. The rules of this association will be used as a reference in determining the goods to be promoted with *cross-selling* marketing strategies and *Product Bundling* marketing techniques.

2 RESEARCH METHODS

2.1.Research Data

The data used is vitamin sales transaction data at Viva Health Duku Setro Pharmacy for 12 weeks in the period September 2017 – November 2017. The tables used as research objects are as follows:

Table 1. Object of Research

No	Table Name	Tuples
1	T_salesHeader	7.863
2	T_salesDetail	21.137
3	M_product	1.379
4	M_customer	2.166
5	M_Pack	24
6	M_subCategory	15
7	M_Category	5

2.2.Research Steps

The research steps used are as follows:

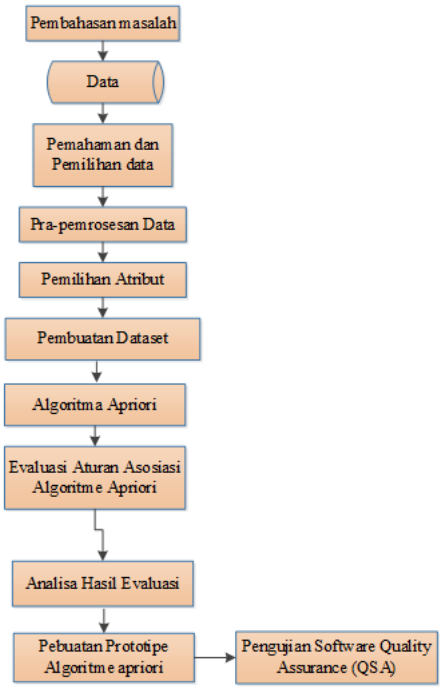

Figure 1. Research Steps

3 RESULTS AND ANALYSIS

3.1.Research Data Structure

The data structure used in this study is as follows:

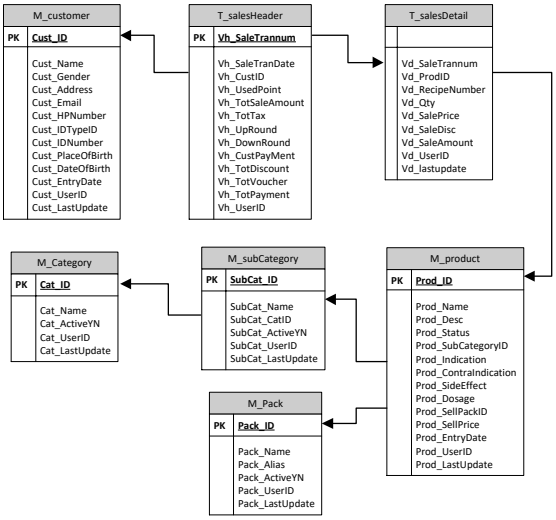

Figure 1. Research Data Structure

3.2.Data Selection

The data that will be used in data mining processing is Vitamin product data, which has the highest 20% occurrence with a minimum support of 2%. The highest 20% product occurrence is taken from the Pareto concept also known as the 80-20 rule which states that for many events, about 80% of the effect is due to 20% of the cause. From the application of the Pareto concept, so that the data used is vitamin sales data as many as 174 records, with detailed sales data as many as 363 records. Of the total data, 141 (81%) were used as training data and 33 (19%) evaluation data.

3.3.Data Pre-processing

Pre-processing techniques or pre-processing used, namely:

1. Data cleaning to remove noise, inconsistent data and incomplete data (missing value).
2. Data integration i.e. combining data from several different tables, this process is done with the following Sql command:

```
Select Vh_SaleTrannum as Trannum, Cast (Vh_SaleTranDate as Date) TranDate, Isnull (Cust_Name,'') Cust_Name,
Prod_ID,Prod_Name, Vd_Qty, Pack_Name, Vd_SaleAmount, cat_ID, SubCat_ID
from store_transaction.. T_salesHeader
Left Join store_transaction.. T_salesDetail on Vh_SaleTrannum = Vd_SaleTrannum
Left Join PR_master.. M_product on Prod_ID = Vd_ProdID
Left Join OP_Master.. M_customer on Cust_ID = Vh_CustID
Left Join PR_Master.. M_Pack on Pack_ID = Prod_SellPackID
Left Join PR_master.. M_subCategory on SubCat_ID = Prod_SubCategoryID
Left Join PR_master.. M_Category on cat_ID = SubCat_CatID
Where Cast (Vh_SaleTranDate as Date) between '2017-09-01' and '2017-11-30'
        and cat_ID not in (4,8,9,32)
        and SubCat_ID = 3
```

3. Data selection is by retrieving relevant data from the database
4. Data transformation is carried out to combine data into a form suitable for the data mining process.

3.4.Attribute Selection

From each item of sales transactions, a large percentage of the number of frequency of occurrence will be calculated against the total number of transactions. The minimum limit value (Treshold) used is 2%, so any item that has the highest percentage of occurrences of 20% and exceeds or equals the minimum limit (Treshold) of 2% will be selected as an attribute.

3.5. Dataset Creation

Here is the structure of the dataset created:

Table 1. Research Dataset Structure

Column Name	Data Type	Information
Trannum	*Char (13)*	Transaction Number
TranDate	*date*	Transaction Date
Cust_Name	*Varchar (100)*	Customer Name
Prod_ID	*Char (8)*	Product ID
Prod_Name	*Varchar (300)*	Product Name
Vd_Qty	*Integer*	Qty
Pack_Name	*Char (30)*	Product Pack Unit
Vd_SaleAmount	*Money*	Total Selling

Data sets with the above structure are made as many as 2 pieces, with the names of the T_Training table and *T_Evaluation*. The T_Training table is filled with vitamin sales data in September to October with sales data of 295 records, while the T_Evaluation table is filled with vitamin sales data in November with sales data of 68 records.

3.6. Apriori algorithm

The a priori algorithm determines the candidates that may appear by paying attention to minimum support and minimum confidence. Support is the visitor value or percentage combination of an item in the database. The support formula is as follows:

$$Support\ (A) = \frac{Jumlah\ Transaksi\ mengandung\ A}{Total\ Transaksi}$$

The Support value of two items is obtained using the formula:

$$Support(A,B) = P(A \cap B)$$

$$Support(A,B) = \frac{\sum Transaksi\ mengandung\ A\ dan\ B}{\sum Transaksi}$$

While confidence is the value of certainty, namely the strong relationship between items in a priori. Confidence can be searched after the pattern of the frequency of appearance of an item is found. The formula for calculating confidence is as follows:

for example found rules A → B then:

$$Confidence\ P(B|A) = \frac{Total\ transaksi\ mengandung\ A\ dan\ B}{Transaksi\ mengandung\ A}$$

The rules of association obtained with a priori algorithms are as follows:

Table 3. A priori algorithm association rules on *training* data

Urut	A		B	Conf
1	03000095 - Vitacimin Orange Tabl 2`S	==>	03000094 - Vitacimin Blueberry Tabl 2`S	0.67
2	03000189 - Fituno Capl 10`S	==>	03000028 - Neurobion Forte Tabl Strp 10`S	0.67
3	03000022 - Xonce Rasa Jeruk Tabl Strp 2`S	==>	03000094 - Vitacimin Blueberry Tabl 2`S	0.56
4	3000094 - Vitacimin Blueberry Tabl 2`S	==>	03000026 - Vitacimin Tabl Strp 2`S	0.45
5	03000022 - Xonce Rasa Jeruk Tabl Strp 2`S	==>	03000026 - Vitacimin Tabl Strp 2`S	0.44
6	03000082 - Theragran-M Tabl Strp 4`S	==>	03000173 - Youvit Multivitamin Mix Berri 7`S	0.44
7	03000026 - Vitacimin Tabl Strp 2`S	==>	03000094 - Vitacimin Blueberry Tabl 2`S	0.38
8	03000173 - Youvit Multivitamin Mix Berri 7`S	==>	02000238 - Imboost Force Capl (3x10`S)	0.35
9	02000238 - Imboost Force Capl (3x10`S)	==>	03000173 - Youvit Multivitamin Mix Berri 7`S	0.32
10	03000094 - Vitacimin Blueberry Tabl 2`S	==>	03000095 - Vitacimin Orange Tabl 2`S	0.3

3.7.Evaluation of a priori Algorithm Association Rules

There are 2 factors that will be evaluated in this study, namely the size of generality (generality) and the measure of reliability (reliability) of the association rules produced. The generality measure is used to determine the degree of occurrence of each item formulated in the association rule to the overall transaction. The sizes used are:

1. Support

   Support is a measure of how often a collection of items in an association occurs together as a percentage of all transactions.

   Formula :

$$s(A \Rightarrow B) = P(A \cap B)$$

   Where:

$$P(A \cap B) = \frac{\sum Transaksi\ mengandung\ A\ dan\ B}{\sum Transaksi}$$

2. Coverage

   Coverage is a measure of how often the collection of each item formulated on the left-hand side (causal factor) occurs as a percentage of all transactions.

   Formula:

$$P(A) = \frac{\sum Transaksi\ mengandung\ A}{\sum Transaksi}$$

The reliability measure is used to determine the level of reliability of the association rules generated in association rules mining. The sizes used are:

1. *Confidence*

   Confidence is a measure of uncertainty or confidence in feasibility associated with any pattern found.

   Formula :

$$a(A \Rightarrow B) = P(B|A$$

   Where:

$$P(B|A) = \frac{\sum Transaksi\ mengandung\ A\ dan\ B}{\sum Transaksi\ mengandung\ A}$$

2. *Added Value*

   Added Value is used to measure how much difference between the confidence value and the chance of a combination of each item on the side of the association rule for all transactions.

Formula :

$$Added\ Value\ (A \Rightarrow B) = P(B|A) - P(B)$$

3. Colleration

Correlation is used to determine whether one itemset has a dependency or not.

Correlation (A □ B) is 0 if A and B are independent (have no dependency), and 1 (positive value, >0) or -1 (negative value, <0) if they are dependent. 1 indicates attraction, and -1 indicates repulsion.

Formula :

$$Correlation\ (A \Rightarrow B) = P(A \cap B)/[P(A)P(B)]$$

Where:

P(A) = $\sum Transaksi\ Mengandung\ A$

P(B) = $\sum Transaksi\ Mengandung\ B$

Evaluation of a priori algorithm training data association rules

Table 4. Evaluation of association rules against a priori algorithm training data

Aturan urut	Generalitas		Reabilitas		
	Support	Coverage	Confidence	Added Value	Correlation
1	0,0426	0,0638	0,6667	0,5248	0,0946
2	0,0284	0,0426	0,6667	0,3901	0,1844
3	0,0355	0,0638	0,5556	0,4137	0,0788
4	0,0638	0,1418	0,4500	0,2798	0,0766
5	0,0284	0,0638	0,4444	0,2742	0,0757
6	0,0284	0,0638	0,4444	0,3026	0,0630
7	0,0638	0,1702	0,3750	0,2332	0,0532
8	0,0496	0,1418	0,3500	0,1940	0,0546
9	0,0496	0,1560	0,3182	0,1763	0,0451
10	0,0426	0,1418	0,3000	0,2362	0,0191
Rata-Rata	0,0433	0,1050	0,4571	0,3025	0,0745

3.8. Analysis of Evaluation Results

Analysis of the evaluation results was carried out on a number of association rules with the largest confidence value generated by each algorithm, namely arpiori and FP-Growth, Analysis was carried out by calculating the level of strength and accuracy of each rule implemented against the evaluation dataset on sales transaction data. The strength level of the association rule is determined by the support value which represents the generality aspect and the confidence value which represents the reliability aspect.

Formula :

$$\frac{\sum_{i=1}^{n}(S_i \times C_i)}{n}$$

Where:

n = number of association rules

Si = support value for the i-th association rule

Ci = confidence value for the i-th association rule

Table 5. Analysis of a priori algorithm evaluation results

Urut Aturan Asosiasi	Algoritme Apriori		
	Supp	Conf	Sup x Conf
1	9	0,67	6,03
2	6	0,67	4,02
3	9	0,56	5,04
4	20	0,45	9,00
5	9	0,44	3,96
6	9	0,44	3,96
7	24	0,38	9,12
8	20	0,35	7,00
9	22	0,32	7,04
10	20	0,3	6,00
	Akurasi Algoritme Apriori		**61,17**

From table 7 it can be seen that the accuracy rate of the a priori algorithm is 61.17.

3.9. Application of a priori algorithm prototypes

Apriori algorithm flowchart

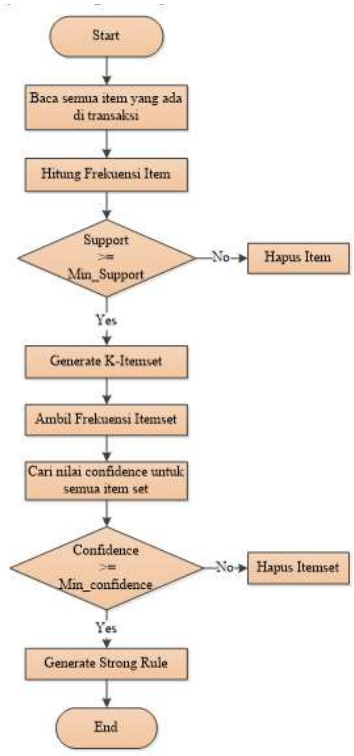

Figure 3. A priori algorithm flowchart

Figure 4. Apriori algorithm prototype GUI

3.10. Testcase Software

Software SQA (*Software Quality Assurance*) testing is done to ensure that the software has a minimum standard of quality. The following are the results of the evaluation of SQA testing:

Table 10. Software Quality Assurance Evaluation Results

No.	Responden	Skor Metrik								Skor
		1	2	3	4	5	6	7	8	
1	Manager Marketing Plan	100	80	60	80	100	80	80	80	82
2	Staff Marketing	80	80	80	100	80	60	60	100	79
3	Staff Marketing	80	100	80	80	80	80	80	80	83
4	Staff Marketing	80	60	80	80	80	100	80	80	80
5	Staff Marketing	80	80	100	80	100	80	80	80	84
6	Staff Marketing	100	80	80	60	80	80	80	100	83
7	Apoteker	80	80	80	100	60	100	100	60	83
8	Apoteker	80	100	80	80	80	80	80	80	83
9	Apoteker	80	80	80	80	80	60	80	80	77
10	Apoteker	100	80	100	60	80	80	80	80	82
									Rata-Rata	82

Based on the survey results in table 10 above, it can be concluded that the Customer Shopping Pattern Analysis Prototype has met quality standards, because the average survey value is 82 and above the optimum value.

3.11. Prototype Verification

The application of the association algorithm in the form of a prototype system to form rules for asosization of vitamin products at Viva Health Duku Setro Pharmacy can be implemented with a priori algorithm prototype accuracy to rapidminer applications of 89.66% with a minimum support used of 2% and minimum confidence of 50%. This percentage is obtained from the following calculation:

$$Akurasi\ Prototipe = \frac{Jumlah\ aturan\ prototipe}{Jumlah\ aturan\ rapidminer}\ x100 = \frac{26}{29}\ x\ 100 = \mathbf{89,66\%}$$

4 CONCLUSION

From the results of this study on Market Basket Analysis on Sales Transactions with a priori Algorithm Case Study of Viva Health Duku Setro Pharmacy, the following conclusions can be drawn:

1. From the research that has been done, the a priori algorithm has an accuracy rate of 61.17% to form the rules for asosization of vitamin products at the viva health Duku Setro pharmacy.
2. The application of the association algorithm in the form of a prototype system to form rules for asosization of vitamin products at Viva Health Duku Setro Pharmacy can be implemented and meet the standards because the average value of the software quality assurance testing survey is above the optimum value of 82, and prototype verification of the rapidminer application is 89.65% with a minimum support used of 2% and a minimum confidence of 50%.

REFERENCE

[1]J. Han and M. Kamber, Data Mining Concept and Tehniques. San Francisco: Morgan Kauffman, 2006.

[2] Maliha Hossain, A H M Sarowar Sattar, Mahit Kumar Paul, Market Basket Analysis Using A priori and FP Growth Algorith, 2019 22nd International Conference on Computer and Information Technology (ICCIT), 18-20 December, 2019

[3]D. T. Larose, Discovering Knowledge in Data: An Introduction to Data Mining. John Willey & Sons, Inc., 2005.

[4]Kusrini, Data Mining Algorithms. Jakarta, 2009.

[5] J. Han and M. Kamber, Data Mining: Concepts and Data Mining: Concepts and Techniques: Chapter 6. Mining Association Rules in Large Databases. Simon Fraser University, 2000.

[6]X. Wu and V. Kumar, The Top Ten Algorithm in Data Mining. New York, USA: CRC Press Taylor & Francis Group, LLC, 2009.

[7] G. E. P. Box, G. M. Jenkins, and G.C.Reinsel, Time Series Analysis: Forecasting and Control. New Jersey: Upper Saddle River, Prentice Hall, 1994.

[8]K. Venkatachari and I. D. Chandrasekaran, "Market Basket Analysis Using Fp Growth and A Priori Algorithm: a Case Study of Mumbai Retail Store," 2016.

[9]I. . M. Putu, Lisna Padma Yanti Tuningrat and A. S. Wiranatha, "No Title," Anal. Sales Forecasting Prod. Soy Sauce Prod. Kecap Manalagi Denpasar Bali, vol. 4 No. 1, ISSN no: 2503-488X, pp. 72–81, 2016.