Application of Apriori Algorithm To Determine Product Purchase Patterns In *Online* Stores Using The Kdd Method

Shanti Cahyaningtyas[1*], Asep Arwan Sulaeman[2], Handala Simetris Harahap[3]

[1,2,3]Program Studi Teknik Informatika,Universitas Pelita Bangsa, Indonesia
[*] Corresponding author:
Email: shanti.312210053@mhs.pelitabangsa.ac.id

Abstract.
The growth of e-commerce businesses has led to an increasing volume of sales transaction data stored in Online shops. However, this transaction data is often underutilized, even though it contains valuable product purchasing patterns that can support business decision-making. This study aims to apply the Apriori algorithm to identify product purchasing patterns in an Online shop using the Knowledge Discovery in Databases (KDD) method. The KDD framework is employed to ensure a systematic data analysis process, including data Selection, Preprocessing, data Transformation, application of the Apriori algorithm, and result interpretation. The data used in this study consist of Online retail transaction data obtained from the Kaggle platform, namely the Online Retail dataset, which represents real ecommerce transactions. Data processing is carried out using WEKA software. The analysis focuses on discovering frequent item sets and generating association rules based on minimum support, confidence, and lift ratio values. The results show that the Apriori algorithm can identify products that are frequently purchased together within a single transaction. These purchasing patterns can be utilized as a basis for marketing strategy recommendations, product bundling promotions, more efficient inventory management, and as support for developing product recommendation systems in Online shops.
Keywords: *Data mining and Knowledge.*

I. INTRODUCTION

The The development of information technology and the internet has encouraged the rapid growth of the e-commerce business, both at the global and national levels in Indonesia. Online store platforms provide convenience for consumers to make purchases anytime and from anywhere. This acceleration of digital transactions results in a huge accumulation of transaction data, which is very large, high-dimensional, and complex. Unfortunately, most of these transaction data is often only stored as a digital archive without further processing. In fact, in the raw data pile there are hidden product purchase patterns that if properly identified can provide very strategic business insights [1].

Purchasing patterns that often appear together between products (e.g., the tendency of consumers to buy item B after or at the same time as purchasing item A) are high-value information assets. This information can be used to optimize various marketing strategies, such as bundled product promotions, personalized product recommendation systems, and more efficient warehouse stock arrangements. However, digging up these patterns manually is almost impossible given the massive transaction volume and the huge variety of products. Therefore, data mining techniques are present as the right solution to extract and unearth hidden knowledge from such large transaction data [2].

One of the most popular and effective algorithms in association rule mining is the A Priori Algorithm. This algorithm aims to find frequent itemsets (combinations of products that are often purchased at the same time) and form association rules in the form of logical implications, such as "if you buy product X, then there is a tendency to buy product Y". A priori algorithms have been widely applied in retail and e-commerce research. For example, a previous study on one of the e-commerce in Indonesia managed to identify multi-category purchasing patterns such as "Mother & Baby Needs - Skin Care - Household Care" with a support value of 0.108 and a confidence value of 0.476 [3].

In order for the process of knowledge extraction from raw data to run systematically and produce valid and actionable information, this study adopts the Knowledge Discovery in Databases (KDD) framework. This framework covers crucial stages which include data selection, preprocessing, transformation, application of data mining algorithms, to interpretation of results or evaluation [4]. KDD's approach ensures

that transaction data is not simply processed, but transformed into knowledge that is relevant to the strategic needs of the business.

Although a lot of research has been done on association rules, the research space in the context of local online stores is still very open. Local online stores have unique characteristics—both in terms of product variety, customer shopping behavior, and promotional support patterns—that may differ from global-scale e-commerce. Based on these conditions, this study aims to apply an A Priori Algorithm with the KDD framework on real online store sales transaction data sourced from the Online Retail dataset on the Kaggle platform.

This study is specifically limited to the application of a priori algorithm using WEKA software to produce association rules based on standard parameters, namely minimum support, minimum confidence, and lift ratio. The focus of the research is directed at the analysis of product combinations in a single transaction without involving external attributes such as price, domicile, or payment methods, and does not include sales forecasting.

Theoretically, this research is expected to be able to enrich the research literature on data mining and market basket analysis in the local online store segment. Practically, the research output in the form of association rules is projected to be the basis for data-driven decision making for online store managers, businessmen, and MSMEs in designing bundling promotion strategies, efficient inventory management, and the development of a more accurate product recommendation system to increase customer satisfaction.

II. METHODS

This research is a type of descriptive research with *a data mining approach*. The descriptive approach is used to describe and explain the characteristics of product purchase patterns contained in online store transaction data objectively. Meanwhile, *the data mining* approach is applied to extract and explore hidden *knowledge* in the form of association rules between products purchased by consumers.

In its implementation, this study applies an A priori algorithm to find combinations of products that are often purchased together (*frequent itemsets*) and form association *rules*. In order to ensure that the entire process runs in a structured, systematic, and valid information production, this research flow fully adopts the *Knowledge Discovery in Databases* (KDD) framework. The focus of the analysis was limited to the stages of KDD, the application of the A Priori Algorithm, and the evaluation of association rules based on *the parameters of support*, *confidence*, and *lift ratio.*

The type of data used in this study is secondary data that represents real online store sales transaction data. This transaction data is obtained from the Kaggle platform through *the Online Retail* dataset which has been compiled into a tabular format.

The sample data used in this analysis consists of 100 transaction lines with a total of 10 attributes that include product purchase details. Each transaction line represents a collection of products purchased by customers in a single transaction session (*shopping basket*)

III. RESULT AND DISCUSSION

Research Results

The test is performed to obtain *the frequent itemset* and *association rules*. The resulting association rules were then analyzed using *support*, *confidence*, and *lift ratio values*. This analysis is in accordance with the objectives of the research, which are to apply a priori algorithms, identify product purchase patterns that often appear at the same time, and provide business strategy recommendations such as stock arrangement, bundling promotions, and product recommendations.

The transaction data that has been transformed into ARFF format is then entered into the WEKA software. Based on the results of the *preprocess display*, the data was successfully read with a total of 100 transactions and 10 product attributes. Each product attribute has a face value *yes*, while unpurchased products are treated as *missing value*.

Table 1 Description of Data Used

Yes	Remarks	Quantity
1	Number of transactions	100 transactions
2	The number of products	10 attributes
3	Unused attributes	InvoiceNo
4	Format data	ARFF
5	Analysis tools	WEKA
6	Analysis methods	A priori algorithm

The frequency of product appearances is used to find out how often each product was purchased in 100 transactions. The *support* value in this section is calculated from the number of *yes* values on each product attribute divided by the total transaction

Table 2 Product Pop-Up Frequency

No	Products	Transaction Amount	Support
1	REGENCY_CAKESTAND	44	44%
2	VINTAGE_CLOCK	42	42%
3	CREAM_CUPID_HEARTS	41	41%
4	SET_OF_TEA_CUPS	40	40%
5	WOODEN_PICTURE_FRAME	40	40%
6	WHITE_HANGING_HEART	39	39%
7	RED_RETROSPOT_PLATE	39	39%
8	JUMBO_BAG_RED	35	35%
9	LUNCH_BAG_BLUE	35	35%
10	LUNCH_BAG_PINK	34	34%

A priori algorithm testing is carried out using the parameters that have been determined in Chapter III. The main parameters used are *lowerBoundMinSupport* of 0.02, *minMetric* of 0.6, *metricType* of *Confidence*, and *numRules* of 20 rules, as shown in the table as follows:

Table 3 Parameter Pengujian Apriori

Parameter	Value
lowerBoundMinSupport	0.02
minMetric (Confidence)	0.6
MetricType	*Confidence*
NumRules	20
Delta	0.05
upperBoundMinSupport	1.0
Classindex	-1
Car	*False*

Based on the results of testing using WEKA, the number of *Large itemsets* was obtained as follows:

Table 4Results of the Formation of Large Itemset

No	Itemet Types	Jumlah Large Itemset
1	L(1)	10
2	L(2)	45

3	L(3)	79
4	L(4)	3

After *Large itemsets* are formed, WEKA generates the 20 best association rules in the *Best rules found* section. The resulting association rules are entirely in the form *of a yes* relationship , which *is yes*, so that it is in accordance with the purpose of the research to find the pattern of products that are actually purchased at the same time.

The *support* value in Table 5 is calculated from the number of transactions containing all products in *the antecedent* and *consuquent* sections, then divided by the total transactions of 100 transactions. The *confidence* value indicates the level of trust in the rules, while *the lift value* indicates the strength of the relationship between products

Table 5 Results of the WEKA Association Rules

No	Antecedent	Convincing	Antecedent Jjumlah	Number of Rules	support	confidence	lift
1	SET_OF_TEA_CUPS=yes, WOODEN_PICTURE_FRAME=yes,VINTAGE_CLOCK=yes	REGENCY_CAKESTAND=yes	5	5	5%	100 %	2.27
2	WHITE_HANGING_HEART=yes,REGENCY_CAKESTAND=yes, SET_OF_TEA_CUPS=yes	WOODEN_PICTURE_FRAME=yes	6	5	5%	83%	2.08
3	CREAM_CUPID_HEARTS=yes, SET_OF_TEA_CUPS=yes	WOODEN_PICTURE_FRAME=yes	10	8	8%	80%	2.00
4	WHITE_HANGING_HEART=yes,REGENCY_CAKESTAND=yes,WOODEN_PICTURE_FRAME=yes	SET_OF_TEA_CUPS=yes	7	5	5%	71%	1.79
5	CREAM_CUPID_HEARTS=yes, REGENCY_CAKESTAND=yes, SET_OF_TEA_CUPS=yes	WOODEN_PICTURE_FRAME=yes	7	5	5%	71%	1.79
6	CREAM_CUPID_HEARTS=yes, SET_OF_TEA_CUPS=yes	REGENCY_CAKESTAND=yes	10	7	7%	70%	1.59
7	LUNCH_BAG_PINK=yes, VINTAGE_CLOCK=yes	REGENCY_CAKESTAND=yes	15	10	10%	67%	1.52
8	WHITE_HANGING_HEART=yes, RED_RETROSPOT_PLATE=yes	WOODEN_PICTURE_FRAME=yes	12	8	8%	67%	1.52
9	CREAM_CUPID_HEARTS=yes, LUNCH_BAG_PINK=yes	REGENCY_CAKESTAND=yes	12	8	8%	67%	1.52

No	Antecedent	Convincing	Antecedent Jjumlah	Number of Rules	support	confidence	lift
10	REGENCY_CAKESTAND=yes, WOODEN_PICTURE_FRAME=yes	SET_OF_TEA_CUPS=yes	17	11	11%	65%	1.62
11	WHITE_HANGING_HEART=yes, LUNCH_BAG_BLUE=yes	JUMBO_BAG_RED=yes	14	9	9%	64%	1.84
12	SET_OF_TEA_CUPS=yes, VINTAGE_CLOCK=yes	REGENCY_CAKESTAND=yes	14	9	9%	64%	1.46
13	LUNCH_BAG_PINK=yes, LUNCH_BAG_BLUE=yes	VINTAGE_CLOCK=yes	11	7	7%	64%	1.52
14	LUNCH_BAG_BLUE=yes, RED_RETROSPOT_PLATE=yes	VINTAGE_CLOCK=yes	11	7	7%	64%	1.52
15	REGENCY_CAKESTAND=yes, LUNCH_BAG_PINK=yes	VINTAGE_CLOCK=yes	16	10	10%	63%	1.42
16	WHITE_HANGING_HEART=yes, SET_OF_TEA_CUPS=yes, WOODEN_PICTURE_FRAME=yes	REGENCY_CAKESTAND=yes	8	5	5%	63%	1.42
17	CREAM_CUPID_HEARTS=yes, SET_OF_TEA_CUPS=yes, WOODEN_PICTURE_FRAME=yes	REGENCY_CAKESTAND=yes	8	5	5%	63%	1.42
18	CREAM_CUPID_HEARTS=yes, REGENCY_CAKESTAND=yes, WOODEN_PICTURE_FRAME=yes	SET_OF_TEA_CUPS=yes	8	5	5%	63%	1.56
19	REGENCY_CAKESTAND=yes, WOODEN_PICTURE_FRAME=yes, VINTAGE_CLOCK=yes	SET_OF_TEA_CUPS=yes	8	5	5%	63%	1.56
20	WOODEN_PICTURE_FRAME=yes, VINTAGE_CLOCK=yes	WHITE_HANGING_HEART=yes	13	8	8%	62%	1.58

IV. DISCUSSION

The application of the KDD method in this study helps the data processing process to be carried out systematically. At the *selection* stage, the data used was only product attributes because the research focused on product purchase patterns. In *the preprocessing* stage, the data is checked to have an appropriate format

and not contain irrelevant attributes. At the *transformation stage*, the data is converted into ARFF format so that it can be processed using WEKA.

The use of *yes* and *missing values* in the data aims to ensure that the Apriori process does not produce less relevant rules such as product=no == product=no. With this format, the association rules formed focus more on the relationship between products that are actually purchased by customers.

Based on the results of testing using WEKA, the A Priori Algorithm is able to generate product association rules based on online store transaction data. The data used consists of 100 transactions and 10 product attributes. After going through the transformation process, the data was analyzed using *the lowerBoundMinSupport* parameter of 0.02, *minMetric* of 0.6, and *numRules* of 20 rules.

The test results showed that there were *10 Large itemsets* L(1), 45 L(2), 79 L(3), and 3 L(4). The results of the association rules show that all the best rules produced are in the form of *yes* to *yes*, so that it is in accordance with the research objectives to find the pattern of the purchased products at the same time.

The best association rules are *SET_OF_TEA_CUPS = yes*, WOODEN_PICTURE-FRAME=yes, and *VINTAGE_CLOCK=yes* and *REGENCY_CAKESTAND=yes*, with a *confidence* value of 100% and *a lift* of 2.27. These results show a strong relationship between the product combinations. In general, the results of this research can be used as the basis for marketing strategies, bundling promotions, product recommendations and online store stock management.

V. CONCLUSION

Based on the results of the research and discussion on the application of the A Priori Algorithm using the *Knowledge Discovery in Databases* (KDD) framework to analyze product purchase patterns in online stores, several conclusions can be drawn as follows:

1. The KDD framework consisting of *data selection, preprocessing, transformation, data mining, and evaluation/interpretation* stages has been successfully implemented systematically to transform raw transaction data into new knowledge that is structured and strategically valuable for businesses.
2. The *preprocessing* and *transformation stages* are proven to be crucial in improving data quality. Filtering non-essential attributes such as InvoiceNo and converting data formats into *binary value-based Attribute-Relation File Format* (ARFF) (yes for purchased products and *missing value* for non-purchased ones) can make the computational process in the WEKA software more focused and efficient.
3. The application of a priori algorithm with the minimum *support* lower limit (*lowerBoundMinSupport*) parameter of 0.02, *minimum confidence* (*minMetric*) of 0.6, and the rule limit (*numRules*) of 20 rules, succeeded in extracting a number of *frequent itemsets*. This experiment resulted in a variety of itemset combinations, including *Large Itemset* L(1) as many as 10 itemsets, L(2) as many as 45 itemssets, L(3) as many as 79 itemsets, and L(4) as many as 3 itemsets.
4. The analysis of association rules based on *support, confidence*, and *lift ratio* parameters successfully identified the best and strongest buying patterns, namely:
   SET_OF_TEA_CUPS = yes, WOODEN_PICTURE_FRAME = yes, VINTAGE_CLOCK = yes→REGE NCY_CAKESTAND = yes
   This rule has an absolute *confidence* level of 100% and a lift *ratio* of 2.27. A *lift ratio* value greater than 1 (>1) indicates a strong and positively correlated relationship between these products.
5. The practical implications of the association pattern found can be directly used by online store managers as a basis for *data-driven* decision-making, especially in designing more effective marketing strategies such as bundling *products*, optimizing automatic product recommendation features on websites, and setting up warehouse inventory layout and stock management.

REFERENCES

[1] I. D. Hunyadi, N. Constantinescu, and O. A. Țicleanu, “Efficient Discovery of Association Rules in E-Commerce: Comparing Candidate Generation and Pattern Growth Techniques,” Appl. Sci., vol. 15, no. 10, 2025, doi: 10.3390/app15105498.

[2] B. H. Situmorang, A. Isra, D. Paragya, and D. A. A. Adhieputra, "Apriori Algorithm Application for Consumer Purchase Patterns Analysis," ***Komputasi J. Ilm. Ilmu Komput. dan Mat***., vol. 21, no. 1, pp. 15–20, 2024, doi: 10.33751/komputasi.v21i1.9260.

[3] N. Okviannas, S. Widianto, and R. Komaladewi, "Marketing Strategy Analysis Using Apriori Association Method To Increase Sales In E-Commerce Companies," vol. 6, no. 3, pp. 2313–2322, 2025.

[4] Y. A. Singgalen, "Utilizing Knowledge Discovery in Databases (KDD) for Hotel Guest Feedback Analysis," ***J. Comput. Syst. Informatics***, vol. 6, no. 1, pp. 117–132, 2024, doi: 10.47065/josyc.v6i1.6094.

[5] M. Trifena, N. Heryana, and T. Ridwan, "Penerapan Algoritma Apriori Pada Transaksi Penjualan Air Minum Untuk Meningkatkan Strategi Bisnis (Studi Kasus: PT Sila Tirta Gemilang)," ***J. Rekayasa Inf. Swadharma***, vol. 4, no. 2, pp. 37–46, 2024.

[6] H. A. Maulana and A. N. Rohman, "Apriori-Based Association Rule Mining Approach for Developing a Product Recommendation System in an Agricultural E-Marketplace," ***J. Sisfokom (Sistem Inf. dan Komputer)***, vol. 14, no. 4, pp. 566–572, 2025, doi: 10.32736/sisfokom.v14i4.2486.

[7] I. Utnasari, "Analisis Data Pola Pembelian Konsumen Dengan Algoritma Apriori Pada Transaksi Penjualan Minimarket D Mart," ***J. Sist. Inf. Dan Inform***., vol. 2, no. 1, pp. 1–7, 2024, doi: 10.47233/jiska.v2i1.1254.

[8] R. Helfianur and Z. K. A. Baizal, "E-Commerce Recommender System on the Shopee Platform Using Apriori Algorithm," IND. J. Comput., vol. 7, no. 2, pp. 53–64, 2022, doi: 10.34818/indojc.2022.7.2.650.

[9] N. Oktaviani, "Implementasi Algoritma Apriori Untuk Analisis Pola Pembelian Konsumen Pada Toko Serba," ***JATI (Jurnal Mhs. Tek. Inform***., vol. 8, no. 3, pp. 3706–3711, 2024, doi: 10.36040/jati.v8i3.9624.

[10] S. D. Putri and S. Sitohang, "Analisis Pola Pembelian Konsumen Menggunakan Algoritma Apriori," Comput. Sci. Ind. Eng., vol. 9, no. 7, pp. 1504–1513, 2023, doi: 10.33884/***comasiejournal***.v9i7.7889.

[11] A. F. Achmad, Abdul Rahim, and Naufal Azmi Verdikha, "Implementasi Data Mining Algoritma Apriori Pada Data Transaksi Pik Store," ***J. Inform. Polinema***, vol. 11, no. 2, pp. 203–112, 2025, [Online]. Available: https://jurnal.polinema.ac.id/index.php/jip/article/view/6860

[12] P. Khoirunisa, M. Martanto, A. R. Dikananda, and D. Rohman, "Penerapan Fp-Growth Untuk Analisis Pola Pembelian Produk Skincare," ***J. Inform. Teknol. dan Sains***, vol. 7, no. 1, pp. 166–174, 2025, doi: 10.51401/jinteks.v7i1.5213.

[13] D. A. N. F. Yang, M. Kemajuan, D. A. Febriyanti, A. Amelia, and A. T. Kamila, "Dela Ayu Febriyanti," vol. 4, no. 3, pp. 448–459, 2025.

[14] M. Trifena, K. Hamidah, Y. Umaidah, and A. Voutama, "Implementasi Algoritma Apriori untuk Menentukan Paket Bundel dalam Penjualan Toko Swalayan XYZ," vol. 09, no. 02, pp. 187–197, 2023.

[15] I. Irawan and S. Harlina, ***"Jurnal JTIK (Jurnal Teknologi Informasi dan Komunikasi) Implementasi Algoritma Apriori Pada Aplikasi Penjualan Buah***," vol. 9, no. March, pp. 234–243, 2025.5.