

The Impact of K-Means on Association Rules Mining Algorithms Performance


Andre Hasudungan Lubis¹, Rizki Muliono², Nurul Khairina³, Nanda Novita⁴

^{1,2,3}Faculty of Engineering, Universitas Medan Area. Medan, North Sumatera, Indonesia

ABSTRACT

Association Rule Mining (ARM) is one of unsupervised learning approach of machine learning. It acts as a data analysis technique that enables the identification of frequent patterns, correlations, associations, and causal structures within certain datasets. This method widely used in numerous studies and practices to explore knowledges and strengthen decision making. However, dealing a large dataset with high number of transactions may become the shortcoming for the ARM algorithms, such as Apriori, FP-Growth, and Eclat. It leads them to face several challenges, including computational complexity, long mining durations, and memory consumption. Hence, this paper proposes k-means clustering to generates several groups of data to handle the issue, then proceed the ARM algorithms for each individual produced cluster. The study used Elbow method and Silhouette Coefficient as the method to determining optimum number of clusters to be used. The result pointed out that k-means-ARM generates a greater number of rules and provides more contextually relevant and significant correlations. In term of Lift Ratio average score, the k-means-ARM shows the greater value rather than non k-means ARM. The k-means-ARM combination is robust; this approach improves the efficiency and scalability of ARM for large datasets and enhances the interpretability of the discovered association rules.

Keywords— Unsupervised Learning, Apriori, FP-Growth, Eclat, k-means,

 This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Corresponding Author:

Andre Hasudungan Lubis
Faculty of Engineering
Universitas Medan Area
Medan, North Sumatera, Indonesia
Email: andrelubis2201@gmail.com,

Article history:

Received Aug 13, 2024
Revised Aug 18, 2024
Accepted Aug 20, 2024

1. INTRODUCTION

Today is the era of Machine Learning (ML), which is a subset of Artificial Intelligence (AI) by teaching computers to learn from certain data and provides predictions or decisions (Sarker, 2021). This technology is becoming more prominent in daily activities. It has crept into a variety of applications and services, improving user experiences and offers efficiently tasks for them (Tahsien et al., 2020). ML are split into two approaches, supervised and unsupervised learning. In the supervised learning, the algorithm is trained on a labeled dataset that contains both input data and output labels. Meanwhile, the unsupervised learning focuses on coping with unlabeled datasets, which means the algorithm receives only input data with no associated output labels (Alloghani et al., 2020).

As one of the unsupervised learning approach, the Association Rule Mining (ARM) is widely used to enhance the performance of the inventory and business decisions (Fale et al., 2022). ARM is a kind of descriptive model in data mining that illustrates the relationship between frequently occurring items. It assists retailers in identifying the most commonly purchased products by customers (Ghafari & Tjortjis, 2019). The basic purpose of ARM is to find groups of items that frequently appear together in data. These frequently occurring item sets are then used to produce association rules, which express the relationships between items based on co-occurrence patterns (Ibraheem & Hamad, 2023).

ARM has several patterns, including Frequent Pattern, Sequential Pattern, and Structured Pattern (Millham et al., 2021). The Frequent Pattern refers to groups of items that appear together in the dataset with a frequency greater than or equal to a predefined minimum support criterion. On the other hand, Sequential Pattern focuses on temporal relationships in sequential data, such as transaction sequences or event sequences. Then, Structured Patterns are sorts of common patterns or association rules that have certain structural properties or limits (Kaya, 2022). The Frequent Pattern can be classified into three types, including candidate generation, pattern growth, and vertical format (Srinadh, 2022). From each type, there are three algorithms that categorized as the most significant and widely

used for certain purposes, namely Apriori, FP-Growth (Frequent Pattern Growth), and Eclat (Equivalence Class Transformation) (Sharma & Ganpati, 2021).

Apriori is a classical and essential data mining algorithm that widely used in ARM researches. The process is simple, it entails the identification of sets of frequent itemset that exhibit co-occurrence within the dataset, then use them to generate substantial association rules (Fard & Namin, 2020). On the other hand, FP-Growth serves as a substitute for the Apriori, with the objective of enhancing the efficacy of mining frequent patterns. It resulted in a decrease in the necessity for candidate generation and pruning. Furthermore, it is suitable to managing substantial datasets and a significant volume of distinct entities. The FP-Growth algorithm presents a technique for compressing the necessary data for frequent pattern extraction within the vertical FP-tree structure (Aldino et al., 2021; Wicaksono et al., 2020). Moreover, Eclat is an effective algorithm to identify frequent itemset by neglecting explicit generation of candidate itemset. It employs a vertical data format to represent its data. This algorithm offers a distinct advantage over the Apriori, as it facilitates a more efficient process and performance in calculating support from all itemset (Man & Jalil, 2019).

However, these algorithms are vulnerable. They have particular shortcoming which drives to presenting suboptimal results that may not provide meaningful insights for decision-making (Aqra et al., 2019). Apriori is deemed inadequate when handling with large datasets due to its propensity to generate an excessive number of candidates itemset, leading to a combinatorial explosion. Consequently, this results in a significant increase in computational complexity (Fadaei Tehrani et al., 2022). Meanwhile, FP-Growth has a limitation to deals with large datasets with a high number of transactions. Due to its reliance on a vertical format representation of the dataset, it may result inefficiencies and prolonged mining durations. Hence, it may not be the most suitable option for transactional datasets of considerable magnitude due to its memory constraints and possible limitations in processing large-scale data (Kumar & Dubey, 2023). In the same way, Eclat also has difficulties while pruning large datasets with a high number of transactions. The efficiency of Eclat may diminish when dealing with dense datasets wherein a majority of items are present in numerous transactions, leading to a rise in computational complexity, more time consumption and memory needed (Man & Jalil, 2019).

The utilization of clustering as a means of managing large datasets has proven to be a highly effective approach. This method facilitates the analysis of vast amounts of data by grouping similar items or data points together (Ezugwu et al., 2022). Clustering is classified as an unsupervised learning algorithm, which endeavors to partition data points into clusters. The objective of this algorithm is to group items within each cluster that share common characteristics or are closely related based on a specified similarity measure (Mahdi et al., 2021). Meanwhile, combining clustering and ARM may become effective to handle large datasets. Clustering is used as a preprocessing step to partition the data into manageable subsets, reducing dimensionality and facilitating subsequent analysis. ARM is then applied to each cluster separately to discover association rules within focused subsets of the data (Kaushik et al., 2021). This approach leads to faster ARM with reduced data, making the process more efficient and scalable for large datasets. It improves interpretability by discovering more interpretable and relevant association rules within individual clusters. Combining cluster-level association rules provides a comprehensive view of relationships in the entire dataset (Dol & Jawandhiya, 2023; Telikani et al., 2020).

Clustering algorithms are important in the implementation of ARM. These algorithms help in reducing the dataset size, which is crucial for dealing with large amounts of data (Kanhare et al., 2021). They group similar items together based on their characteristics, such as frequency and price, allowing for more efficient analysis (Moahammed et al., 2021). Clustering also helps in identifying interesting rules that may be missed in a trivial approach, leading to more accurate results (AlZoubi, 2019). Additionally, clustering can help in reducing the number of rules generated, making it easier to understand, interpret, and visualize the discovered knowledge (Mattiev & Kavšek, 2020). It also enables the grouping and pruning of rules, resulting in more compact and accurate classifiers (Zhang et al., 2019). Overall, clustering algorithms play a crucial role in improving the scalability, efficiency, and interpretability of Association Rules Mining.

However, challenges such as choosing the proper clustering algorithm and determining the number of clusters should be carefully evaluated based on the dataset and analysis goals, k-means for instance. The algorithm is one of the most popular clustering algorithm to be employed in numerous researches (Ahmed et al., 2020). The k-means clustering algorithm is widely recognized for its versatility, computational efficiency, and straightforward implementation (Ikotun et al., 2022).

There are several studies conducted that use the combination of k-means and ARM algorithm. A study by Setyorini et al. (Setyorini et al., 2021) implemented the k-means algorithm along with FP-Growth which is providing the optimum rules of various household furniture data transactions. Furthermore, Gayathri and Arunodhay (Gayathri & Arunodhaya, 2021) conducted a research that combined k-means clustering and ARM techniques such as Apriori and Eclat to discover the customer segmentation and personalized marketing. The study revealed that k-means clustering enables the segmentation of customers, while ARM helps identify associated products. These techniques are able to generate combo offer recommendations and execute targeted marketing strategies. This approach proves to be adequate in resolving the marketing challenges faced by the company.

Another related study comes from Dharshinni et al. (Dharshinni et al., 2020), which employed the FP-Growth with k-means algorithm to determine the rules of most frequently menus purchased by customers. The result of the research pointed out that process of grouping menus to obtain menu packages is carried out by the k-means algorithm. Additionally, the FP-Growth algorithm is utilized to identify connections among frequently purchased menus, thereby providing recommendations for menu packages. A study by Dharshinni et al. (Dharshinni et al., 2019) stated that utilization of both the k-means clustering algorithm and the Apriori algorithm in patient data resulted in the generation of more comprehensive insights and expedited computational speed in contrast to solely relying on the Apriori algorithm. Furthermore, Aryanti et al. (Aryanti et al., n.d.) reported a study that involving the application of k-means and Apriori for bundling product selection. The result pointed out that the utilization of k-means and Apriori techniques in the context of product bundling strategy was derived from the outcomes of Market Basket Analysis and sales data of a company. The objective was to generate a set of systematically clustered and measurable product recommendations that cater to the needs of consumers, thereby enhancing the sales and consumer appeal of the company.

The combination of ARM algorithms and k-means also can be employed in different subjects. Enggari and Defit (Enggari & Defit, 2022) observed the association between divorce and internet behavior by using the Apriori algorithm and the k-means algorithm to detect divorce facts and the behavior of internet users. Liu et al. (Liu et al., 2021) utilized the K-Apriori algorithm, which integrates the k-means and the Apriori algorithm, is employed for the purpose of mining frequent item sets and identifying association rules among various factors associated with terrorist attacks. This approach utilizes clustering techniques to effectively analyze and extract meaningful patterns from the data. Moreover, a study by Lisnawita and Devega (Lisnawita & Devega, 2020) implemented the Eclat algorithm combined with k-means to determine books borrowing pattern in university library. Another variant study comes from Laxmi (Laxmi et al., 2020) which focused on analysis of Apriori and k-means algorithms for web mining. The k-means segmentation algorithm is employed to increase efficiency by clustering the initial itemset. The result shown that the combination is help to improve the efficiency of data mining and cloud computing techniques in distributed networks. Yürüsen et al. (Yürüsen et al., 2021) conducted a study which presents the utilization of Apriori algorithm combined with the k-means to analyze the spatio-temporal aspects of Solar PV.

These previous studies motivate to discover the strength of k-means while combined with the three ARM algorithms (i.e. Apriori, FP-Growth, and Eclat). The study aims to scrutinize the impact of k-means on ARM algorithm performance while handling large dataset with a high number of transactions. This article describes on how k-means may able to increase the ARM algorithm by ensure the quality and reliability of the discovered association rules through some validations. The difference between this research and previous research is the usage of three kinds of ARM algorithms at the same time along with the k-means algorithm. The novelty of this research lies in the impact of k-means on ARM algorithms to effectively handle large dataset transactions. This innovative approach combines the strengths of ARM, which is proficient in uncovering intriguing patterns within data, and k-means, which is well-known for its clustering capabilities. By merging these two methodologies, the research aims to improve the efficiency of extracting valuable insights from extensive transaction datasets, thereby facilitating a more comprehensive and nuanced comprehension of complex patterns and associations within the data.

2. RESEARCH METHODS

The study has several stages to fulfill the research objectives. It starts with data collection and ends with conducting the result validation. Fig. 1 shows the research stages.

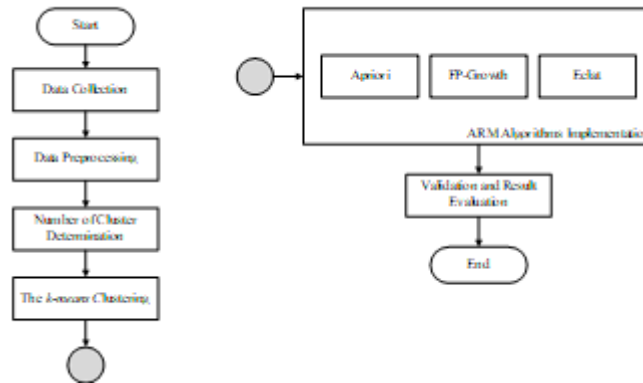


Figure 1. Research Stages

The Study Starts With Collecting Data From <https://www.kaggle.com/datasets/suraj520/car-sales-data>. The Data Is Concerning About Information On Car Sales From A Car Dealership Over The Course Of A Year. The Total Of Data Items Contains As 2,500,000 Rows. However, Due To Time Constraints In Data Load, The Study Limited The Data As 321,054. The Data Is Varied, It Has Several Attributes Including Date, Car Maker, Car Model, Car Year, And Sale Price. In Adherence To Principles Of Transparency And Reproducibility, We Affirm That All Relevant Data Supporting The Findings Of This Research Are Either Explicitly Provided Within The Main Body Of The Paper Or Can Be Found In The Supporting Information Files Accompanying This Manuscript. Table 1 Shows The Sample Of Data Outline.

Table 1. Data Sample

No.	Date	Car Maker	Car Model	Car Year	Sale Price (USD)
1.	8/1/2022	Honda	Civic	2014	10,034
2.	3/15/2023	Nissan	F-150	2016	38,474
...
321,054	12/2/2022	Ford	Silverado	2015	38707

This Stage Encompasses Various Tasks, Including But Not Limited To Eliminating Redundant Data, Rectifying Inaccuracies In The Data, Addressing Instances Of Missing Data, And Other Related Activities. One-Hot Encoding Method Also Employed To Do The Conversion Of Categorical Variables Into A Numerical Format That Used For The ARM Algorithms. It Performed After Data Has Been Clustered. This Stage Is Preliminary, Which It Holds Significant Importance In Data Mining, As The Quality Of The Input Data Directly Influences The Quality Of The Insights And Patterns That Are Discovered By The Algorithms.

The total cluster may influencing the outcome and interpretation of the clustering process, which is fundamental (Sadeghi Moghadam et al., 2021). In this stage, the study proposes certain methods to determining number of clusters to be used such as Elbow method and Silhouette Coefficient. The Elbow method is a simple efficacious approach to determine the optimal number of clusters for a particular

dataset (Jollyta et al., 2023), while Silhouette Coefficient is a metric utilized to facilitates comprehension of the efficacy of the clusters in maintaining proximity among similar data points while simultaneously promoting distance between dissimilar points (Pauletic et al., 2019). k-means supports the processing of transaction data by taking into account the quantity of items presents in each transaction. The k-means algorithm commences by randomly selecting K cluster centers (C_1, C_2, \dots, C_k). Subsequently, the distance between each data point and the cluster centers is calculated by using Equation (1).

$$d_{ik} = \sqrt{\sum_{j=1}^m (x_{ij} - c_{jk})^2} \quad (1)$$

The algorithm then proceeds to assign each data point to the cluster center that exhibits the minimum distance from all the cluster centers. The new cluster center is recalculated using Equation (2), and the distance between each data point and the newly obtained cluster centers is recalculated. Finally, the assignment step is repeated until no data point is reassigned.

$$V_i = \left(\frac{1}{\alpha_i}\right) \sum_{j=1}^{\alpha_i} X_i \quad (2)$$

The integration of k-means clustering with ARM yields the ability to reveal correlations that are unique to distinct segments within a dataset. This facilitates the customization of strategies or decisions based on preferences that are specific to each cluster. The present study employs k-means as a clustering algorithm to facilitate the grouping of large sales transactions into multiple clusters, based on the discernible characteristics or purchasing patterns evident in the data. This approach enables the acquisition of more comprehensive and insightful knowledge regarding the transactional data at hand. Upon acquiring specific clusters, a process of ARM algorithms is conducted for each individual cluster.

Apriori, FP-Growth, and Eclat algorithm are implemented after the clusters has been configured. In this study, the “car make” The potency of these ARM algorithms can be evaluated through the metrics of support and confidence (Kaushik et al., 2021). Support is serving to quantify the frequency with which transactions contain both the antecedent and consequent of a given rule. This metric is instrumental in elucidating the relationship between the two components of the rule (Jain, 2021). It characterized as the proportion of transactions within the database that comprise both X and Y itemset as shown in Equation (3). Confidence, on the other hand, is Confidence refers to the proportion of transactions that comprise both the antecedent and consequent, and is commonly referred to as the rule's confidence. This metric serves as an indicator of the strength of the association between the antecedent and consequent. A higher confidence value signifies a more robust relationship between the two components (Jain, 2021). It described as the percentage of transactions in the database that contain itemset X and also include itemset Y as shown in Equation (4).

$$\text{sup}(X \rightarrow Y) = \frac{n(X \cup Y)}{n} = P(XY) \quad (3)$$

$$\text{conf}(X \rightarrow Y) = \frac{n(X \cup Y)}{n(X)} = \frac{P(XY)}{P(X)} \quad (4)$$

In the context of association rule mining, Lift is a widely utilized metric for measuring support. Lift is defined as the ratio of the expected support to the anticipated support. The lift value indicates the degree to which a rule can accurately predict an outcome compared to a baseline assumption (Babu & Sreedevi, 2023; Jain, 2021). In this study, Lift is used as the means to evaluate the result by using Equation (5). The Lift metric employed to compare the performance of non-k-means ARM algorithms and the original.

$$\text{Lift} = \frac{P(XY)}{P(X).P(Y)} \quad (5)$$

3. RESULT OF STUDY

This section presents the noteworthy outcomes of the extensive research endeavors, elucidating significant findings and perspectives that have arisen from our thorough examination. At first stage, the study undertakes the crucial stage of data preprocessing, wherein it meticulously cleans, transforms, and refines the raw data to ensure its quality and appropriateness for subsequent analysis. Then, the Elbow Method and Silhouette Coefficient are employed as the means to determine the optimum number of k clusters. The implementation of k -means algorithm is conducted as the next stage of the research, clusters are partitioned followed by employing the ARM algorithms to obtain the Support and Confident value for each of them. Lastly, the Lift score is calculated as the consideration for the comparison between k -means-ARM and original ARM.

3.1. Determining the Number of Cluster

The Elbow Method is an essential tool to determining the optimal number of clusters for a given dataset. The Elbow Method is a fundamental technique utilized in cluster analysis to visualize the Sum of Squared Errors (SSE) across varying numbers of clusters (Humaira & Rasyidah, 2020). This approach facilitates the attainment of an appropriate balance between excessively intricate and overly simplistic cluster solutions, thereby ensuring that the clustering outcomes are both meaningful and insightful (Cui, 2020).

$$SSE = \sum_{k=1}^K \sum_{x_i \in S_K} \|x_i - c_k\|_2^2 \quad (6)$$

Equation (6) is utilized to evaluate the Sum of Squared Errors (SSE) for each value of k ranging from 2 to 20. The computed SSE values for each k are presented in Table 2.

Table 2. The Value of SSE from cluster 2 to 20

Cluster	SSE	Cluster	SSE
2	401262.2202	12	54630.75883
3	255338.5115	13	50274.43929
4	160333.9149	14	48771.63053
5	136628.2846	15	44230.44369
6	117272.059	16	40636.73494
7	98481.29966	17	37454.35006
8	83702.88495	18	35526.91627
9	70691.44961	19	33964.85889
10	60261.61001	20	32157.14449
11	60332.62805		

As per the data presented in Table 2, each cluster exhibits a range of SSE values. Notably, the analysis reveals that cluster 11 ($k=1$) displays the highest SSE score, amounting to a total of 60332.62805. The outcome of the Elbow Method is illustrated in Fig. 2.

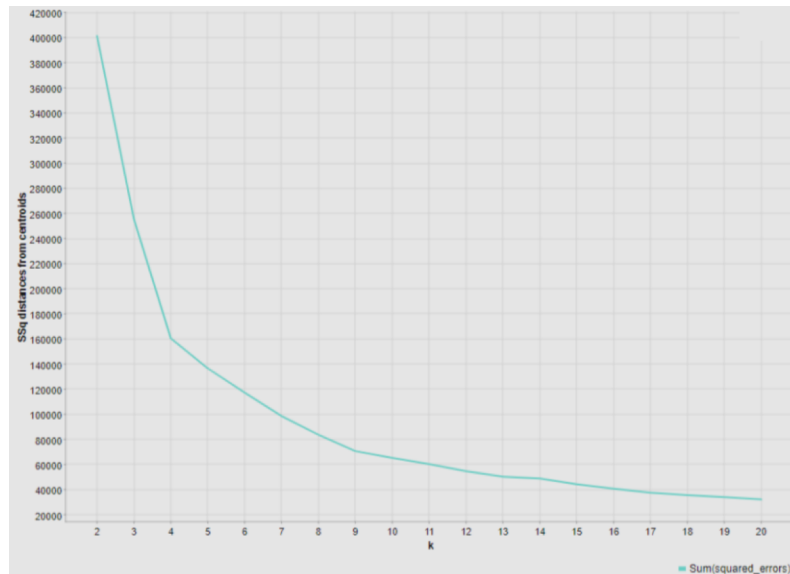


Figure 2. Elbow Method

Fig. 2 depicts the relationship between the Sum of Squared Errors (SSE) and the number of clusters. The results indicate that the SSE value is maximal when the number of clusters is 2, and it experiences a sharp decline when the number of clusters increases to 4 and a continuous decrease until cluster number $k=20$ is attained. Notably, Fig. 2 reveals a distinct elbow at $k= 11$. Hence, it can be inferred that the aggregate number of clusters employed amounts to 11. Besides applying the Elbow Method, we also utilize the Silhouette Coefficient as the second phase of the optimal cluster evaluation. The Silhouette Coefficient is a metric that serves to evaluate the efficacy of clustering outcomes by taking into account the degree of cohesion within clusters and the degree of separation between clusters (Pauletic et al., 2019).

$$SI(i) = \frac{b(i)-a(i)}{\max\{a(i)-b(i)\}} \quad (7)$$

To determine the Silhouette Coefficient value for each k number, which ranges from 2 to 20, Equation (7) is employed. The metric values may vary between -1 and 1. Clustering outcomes are deemed suitable if the Silhouette Coefficient values are positive and near 1 (Ullmann et al., 2022). The Silhouette Index calculation results for each k cluster number are presented in Table 3.

Table 3. Silhouette Coefficient Score from cluster 2 to 20

Cluster	Silhouette Coefficient	Cluster	Silhouette Coefficient
2	0.6983	12	0.751
3	0.7523	13	0.6981
4	0.7793	14	0.8178
5	0.7926	15	0.7251
6	0.8165	16	0.6388
7	0.8215	17	0.8039
8	0.7532	18	0.6725
9	0.8288	19	0.7867
10	0.8235	20	0.6771

Cluster	Silhouette Coefficient	Cluster	Silhouette Coefficient
11	0.8476		

According to Table 3, the Silhouette Coefficient attains its highest score at cluster number $k = 11$, with a value of 0.8476. The score is in close proximity to 1 and indicates the presence of well-defined clusters. This finding reinforces the conclusion drawn from the Elbow method, which suggests that the optimal number of clusters is 11 ($k = 11$).

3.2. Clustering

Embarking on the data clustering journey, the second stage delves into the intricate process of grouping similar data points together, exposing concealed patterns and structures that are otherwise obscured within the dataset. In this study, clustering process is held to partition the large-scale dataset into 11 cluster as determined at the first stage before. The used attributes are year and price. Firstly, the k -means algorithm initiates by selecting 11 cluster centers randomly. Then, Equation (1) is employed to calculate the distance between each data point and the cluster centers. Furthermore, Equation (2) is utilized to recalculate the new cluster center. Each data point is assigned to the cluster with the nearest centroid. Table 4 shows the result of data cluster along with total items.

Table 4. Total Data for Each Cluster

Cluster	Total Data
1	29703
2	29047
3	28906
4	29118
5	28730
6	29298
7	29119
8	29533
9	29231
10	29356
11	29012

Based on the data presented in Table 4, it can be observed that Cluster 1 exhibits the highest volume of data in comparison to the other clusters, with a total of 29703. This is followed by Cluster 8 and Cluster 10. Conversely, Cluster 5 displays the lowest amount of data, with a total of 28730. From each of these cluster, all of ARM algorithm will be implemented as the next stage.

3.3. Apriori Algorithm

Apriori is the first algorithm to be test by using k -means data clustering result. However, One-Hot Encoding method is used before the algorithm to transform data into binary form (0 and 1) and resulting total of 12,879 transactions. This form is also used for other ARM algorithms (i.e., FP-Growth and Eclat). Table 5 illustrate the data in One-Hot Encoding form.

Table 5. Data in One-Hot Encoding Form

Transaction ID	Altim a	Corolla	...	F-150	Civic
T00001	1	0	...	0	0
T00002	0	0	...	0	0
...
T12879	0	1	...	0	0

The Apriori process will provide recommendations for the types of products, based on the number of groups generated in the K-Means process. The study involves a comparison of the number of rules generated from varying numbers of clusters, ranging from 2 to 11. The minimum support and minimum confidence values utilized in this experiment are set at 10% and 50%, respectively. Table 6 and Table 7 present the outcomes of the analysis that contrasts the quantity of rules produced from varying numbers of clusters.

Table 6. Comparison of Total Rules from Each Cluster of Apriori

Number of Cluster	Min Support and Min Confidence	Rules	Lift Ratio
1	10-15%	591	45.83%
2	10-15%	558	46.35%
3	10-15%	644	45.49%
4	10-15%	578	46.65%
5	10-15%	693	47.45%
6	10-15%	801	48.03%
7	10-15%	873	48.75%
8	10-15%	868	48.29%
9	10-15%	812	48.14%
10	10-15%	788	47.4%
11	10-15%	792	46.66%

According to the data presented in Table 6, the cluster with the greatest number of rules is cluster 7, with a total of 873 rules and 48.75% or lift ratio score. This is closely followed by clusters 8 and 9. The findings indicate that cluster 7 holds the utmost importance in relation to the quantity of rules generated by the Apriori algorithm, signifying a greater degree of intricacy in the data.

3.4. FP-Growth Algorithm

The next algorithm to be test is FP-Growth. Similar to the Apriori, this algorithm uses data from Table 4 and also employs the One-Hot Encoding method as listed in Table 5. In this study, the minimum support and minimum confidence thresholds employed were established at 10% and 50%. The results of the analysis, which compares the number of rules generated from different cluster quantities, are presented in Table 7.

Table 7. Comparison of Total Rules from Each Cluster of FP-Growth

Number of Cluster	Min Support and Min Confidence	Rules	Lift Ratio
1	10-15%	485	50.34%
2	10-15%	530	52.34%
3	10-15%	692	52.4%
4	10-15%	675	52.66%
5	10-15%	589	53.1%
6	10-15%	743	53.24%
7	10-15%	812	54.43%
8	10-15%	889	55.45%
9	10-15%	864	55.43%
10	10-15%	845	55.28%
11	10-15%	798	54.42%

The data presented in Table 7 reveals that cluster 8 has the highest number of rules, totaling 889, and a lift ratio score of 48.75%. This finding is closely trailed by clusters 9 and 10. The results suggest that cluster 8 is the most significant cluster in terms of the number of rules generated by FP-Growth algorithm, indicating a higher level of complexity in the data. The high lift ratio score further highlights the strong association between the variables in cluster 8.

3.5. Eclat Algorithm

Eclat is the last algorithm to be test which also uses data from Table 4 and utilizes the One-Hot Encoding to convert the data into binary. The minimum support and minimum confidence also established at 10% and 50%, respectively, consistent with other ARM algorithms. The outcomes of the analysis, which involved comparing the number of rules generated from varying cluster quantities, are detailed in Table 8.

Table 8. Comparison of Total Rules from Each Cluster

Number of Cluster	Min Support and Min Confidence	Rules	Lift Ratio
1	10-15%	806	46.2%
2	10-15%	873	47.34%
3	10-15%	903	47.45%
4	10-15%	888	48.68%
5	10-15%	922	49.46%

Number of Cluster	Min Support and Min Confidence	Rules	Lift Ratio
6	10-15%	972	49.58%
7	10-15%	1019	50.88%
8	10-15%	1084	51.11%
9	10-15%	1088	51.02%
10	10-15%	1007	50.62
11	10-15%	1004	50.87

Based on the data presented in Table 8, it can be observed that Cluster 8 exhibits the highest number of rules generated by the Eclat algorithm, amounting to 1084, and a lift ratio score of 51.11%. This finding is closely mirrored by Clusters 7 and 9. The results suggest that Cluster 8 holds greater significance in terms of the complexity of the data, owing to the larger number of rules generated. Furthermore, the elevated lift ratio score underscores the robust association between the variables within Cluster 8.

3.6. Result Comparison between ARM and k-means-ARM

In this study, *k-means* is used to clusters the large datasets into several groups to facilitate the ARM algorithms (Apriori, FP-Growth, Eclat) generating rules from each of them. We also employ those ARM algorithms independently of the *k-means* to handle the dataset. This approach is adopted to ascertain whether any enhancement is observed in the performance of the ARM algorithms due to the incorporation of the *k-means* algorithm. In this case, we use the total of rules generated and the average score of Lift Ratio as the parameter of comparison as shown in Table 9.

Table 9. Comparison between ARM Algorithms and k-means-ARM

Parameter	Non <i>k-means</i>			<i>k-means</i>		
	<i>Apriori</i>	<i>FP-Growth</i>	<i>Eclat</i>	<i>Apriori</i>	<i>FP-Growth</i>	<i>Eclat</i>
Total Rules	126	124	166	7998	7922	10566
Average Lift Ratio	45.29%	44.65%	41.92%	47.19%	53.55%	47.19%

As shown in Table 9, the comparative analysis reveals that the *k-means-ARM* approach generates a greater number of rules in comparison to the non *k-means-ARM*. The utilization of the *k-means-ARM* methodology not only results in a greater quantity of rules but also enhances the interpretability and granularity of the extracted association rules. The *k-means* clustering technique effectively partitions the dataset into distinct clusters, enabling the ARM algorithms to concentrate on uncovering associations within each group of data points separately. This segmentation ensures that the resulting rules are more tailored to the specific characteristics and behaviors exhibited within each cluster. Consequently, the *k-means-ARM* approach not only amplifies the rule count but also enriches the insights derived from association rule mining by offering cluster-specific patterns that may be obscured when using non-clustered data.

Table 9 also shows the discernible enhancements in the average Lift Ratio scores offered by each Association Rule Mining (ARM) algorithm. The Lift Ratio, a crucial metric for evaluating rule significance, measures the degree of association between antecedent and consequent items in a rule, relative to their independent occurrences. A higher Lift Ratio indicates a more impactful association. The data presented in Table 9 highlights the effectiveness of the *k-means*-ARM methodology in extracting significant and contextually pertinent correlations, thereby contributing a valuable dimension to the assessment of algorithmic performance within the dataset's context. Thus, the result pointed out a shed light on the significant potential of *k-means*-ARM as a robust framework for revealing complex patterns within the landscape of the dataset.

5. CONCLUSION

ARM algorithms, as the unsupervised learning approaches, usually used to enhance inventory management and business decisions. Yet, while handling large datasets with a high number of transactions, encounter challenges such as heightened computational complexity, inefficiencies, and prolonged mining durations, which may lead to wasteful memory consumption. To address these challenges, the article proposes combining clustering algorithms, particularly k-means clustering, with ARM. Clustering aids in partitioning data into smaller subsets, thereby simplifying analysis. ARM is subsequently employed to identify association rules within each cluster. The results show that k-means-ARM generates a greater number of rules and provides more contextually relevant and significant correlations. The total rules for each Apriori, FP-Growth, and Eclat are 7998, 7922, 10566 rules. This suggests that k-means-ARM is a robust framework for revealing complex patterns and enhancing decision-making based on association rules. Moreover, the Lift Ratio scores are increased too. The k-means-ARM has greater scores rather than the original. The study highlights the potential of combining clustering with ARM to improve the performance and interpretability of data mining techniques in handling large datasets. The study emphasizes the importance of careful evaluation of clustering algorithms and determination of the optimal number of clusters based on the dataset and analysis goals. However, the study is limited on the validation of clustering process. Therefore, the utilization of Davis Bouldin Index (DBI) or Fowlkes-Mallows Scores can be conducted in the future. Moreover, attributes used are also limited. A high dimensional data is creditable to be employed to explore the robustness of k-means-ARM combination.

REFERENCES

- Ahmed, M., Seraj, R., & Islam, S. M. S. (2020). The k-means algorithm: A comprehensive survey and performance evaluation. *Electronics*, 9(8), 1295.
- Aldino, A. A., Pratiwi, E. D., Sintaro, S., Putra, A. D., & others. (2021). Comparison of market basket analysis to determine consumer purchasing patterns using fp-growth and apriori algorithm. 2021 International Conference on Computer Science, Information Technology, and Electrical Engineering (ICOMITEE), 29–34.
- Alloghani, M., Al-Jumeily, D., Mustafina, J., Hussain, A., & Aljaaf, A. J. (2020). A systematic review on supervised and unsupervised machine learning algorithms for data science. *Supervised and Unsupervised Learning for Data Science*, 3–21.
- AlZoubi, W. A. (2019). A survey of clustering algorithms in association rules mining. *International Journal of Computer Science & Information Technology (IJCSIT)* Vol, 11.
- Aqra, I., Abdul Ghani, N., Maple, C., Machado, J., & Sohrabi Safa, N. (2019). Incremental algorithm for association rule mining under dynamic threshold. *Applied Sciences*, 9(24), 5398.
- Aryanti, S., Mahdiana, D., & Setiadi, A. (n.d.). Penerapan Metode K-Means Dan Apriori Untuk Pemilihan Produk Bundling. *Journal CERITA: ISSN*, 2461, 1417.
- Babu, M. V., & Sreedevi, M. (2023). A Literature Study on Various Techniques of Association Rule Mining. *TIJER-International Research Journal*, 10(6), 648–654.
- Cui, M. (2020). Introduction to the k-means clustering algorithm based on the elbow method. *Accounting, Auditing and Finance*, 1(1), 5–8.
- Dharshinni, N. P., Azmi, F., Fawwaz, I., Husein, A. M., & Siregar, S. D. (2019). Analysis of accuracy K-means and apriori algorithms for patient data clusters. *Journal of Physics: Conference Series*, 1230(1), 12020.
- Dharshinni, N. P., Bangun, E., Karunia, S., Damayanti, R., Rophe, G., & Pandapotan, R. (2020). Menu Package Recommendation using Combination of K-Means and FP-Growth Algorithms at Bakery Stores: Menu Package Recommendation using Combination of K-Means and FP-Growth Algorithms at Bakery Stores. *Jurnal Mantik*, 4(2), 1272–1277.
- Dol, S. M., & Jawandhiya, P. M. (2023). Classification Technique and its Combination with Clustering and Association

- Rule Mining in Educational Data Mining—A survey. *Engineering Applications of Artificial Intelligence*, 122, 106071.
- Eggari, S., & Defit, S. (2022). Divorce Fact Detection Based on Internet User Behavior Using Hybrid Systems with Combination of Apriori Algorithm and K-Means Method. *Khazanah Informatika: Jurnal Ilmu Komputer Dan Informatika*, 8(1), 8–17.
- Ezugwu, A. E., Ikotun, A. M., Oyelade, O. O., Abualigah, L., Agushaka, J. O., Eke, C. I., & Akinyelu, A. A. (2022). A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects. *Engineering Applications of Artificial Intelligence*, 110, 104743.
- Fadaei Tehrani, A., Sharifi, M., & Rahmani, A. M. (2022). Frequent pattern mining algorithms in fog computing environments: A systematic review. *Concurrency and Computation: Practice and Experience*, 34(24), e7229.
- Fale, P. N., Moundekar, N., RiteshSaudagar, P. K., Rode, M., & Borkar, J. (2022). Review on Optimization of Apriori Algorithm for Finding the Association Rules in Different Business and Other Datasets for Retrieval of Relations Between Different Entities. *International Journal of Scientific Research in Science, Engineering and Technology*, 9(2), 271–276.
- Fard, M. J. S., & Namin, P. A. (2020). Review of apriori based frequent itemset mining solutions on big data. 2020 6th International Conference on Web Research (ICWR), 157–164.
- Gayathri, K., & Arunodhaya, R. (2021). Customer Segmentation and Personalized Marketing Using K-Means and APRIORI Algorithm. *Proceedings of the First International Conference on Combinatorial and Optimization, ICCAP 2021, December 7-8 2021, Chennai, India*.
- Ghafari, S. M., & Tjortjis, C. (2019). A survey on association rules mining using heuristics. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9(4), e1307.
- Humaira, H., & Rasyidah, R. (2020). Determining the appropriate cluster number using elbow method for k-means algorithm. *Proceedings of the 2nd Workshop on Multidisciplinary and Applications (WMA) 2018, 24-25 January 2018, Padang, Indonesia*.
- Ibraheem, H. R., & Hamad, M. M. (2023). A Hybrid Integrated Model for Big Data Applications Based on Association Rules and Fuzzy Logic: A Review. *Iraqi Journal For Computer Science and Mathematics*, 4(2), 171–178.
- Ikotun, A. M., Ezugwu, A. E., Abualigah, L., Abuhaija, B., & Heming, J. (2022). K-means Clustering Algorithms: A Comprehensive Review, Variants Analysis, and Advances in the Era of Big Data. *Information Sciences*.
- Jain, A. (2021). Association Rule Mining in Transactional Data: Challenges and Opportunities. *International Journal of Mechanical Engineering*, 6(3), 4548–4557.
- Jollyta, D., Efendi, S., Zarlis, M., & Mawengkang, H. (2023). Analysis of an optimal cluster approach: a review paper. *Journal of Physics: Conference Series*, 2421(1), 12015.
- Kanhere, S., Sahni, A., Stynes, P., & Pathak, P. (2021). Clustering based approach to enhance association rule mining. 2021 28th Conference of Open Innovations Association (FRUCT), 142–150.
- Kaushik, M., Sharma, R., Peious, S. A., Shahin, M., Yahia, S. Ben, & Draheim, D. (2021). A systematic assessment of numerical association rule mining methods. *SN Computer Science*, 2(5), 348.
- Kaya, M.-F. (2022). Pattern Labelling of Business Communication Data. *Group Decision and Negotiation*, 31(6), 1203–1234.
- Kumar, M., & Dubey, A. K. (2023). An analysis and literature review of algorithms for frequent itemset mining. *International Journal of Advanced Computer Research*, 13(62), 1.
- Laxmi, K. R., Ramya, N., Pallavi, S., & Madhuravani, K. (2020). Study and Analysis of Apriori and K-Means Algorithms for Web Mining. In *Innovations in Electronics and Communication Engineering: Proceedings of the 8th ICIECE 2019* (pp. 693–701). Springer.
- Lisnawita, L., & Devega, M. (2020). Implementation of ECLAT Algorithm Technology: Determining Books Borrowing Pattern in University library. *IOP Conference Series: Earth and Environmental Science*, 469(1), 12036.
- Liu, S., Chen, H., & Yu, Y. (2021). Research on Multi-factors Terrorist Attacks in China Based on K-Apriori Algorithm Research. *Journal of Physics: Conference Series*, 1746(1), 12042.
- Mahdi, M. A., Hosny, K. M., & Elhenawy, I. (2021). Scalable clustering algorithms for big data: A review. *IEEE Access*, 9, 80015–80027.
- Man, M., & Jalil, M. A. (2019). Frequent itemset mining: technique to improve eclat based algorithm. *International Journal of Electrical and Computer Engineering*, 9(6), 5471–5478.
- Mattiev, J., & Kavšek, B. (2020). CMAC: clustering class association rules to form a compact and meaningful associative classifier. *Machine Learning, Optimization, and Data Science: 6th International Conference, LOD 2020, Siena, Italy, July 19--23, 2020, Revised Selected Papers, Part I* 6, 372–384.
- Millham, R., Agbehadji, I. E., & Yang, H. (2021). Pattern mining algorithms. *Bio-Inspired Algorithms for Data Streaming and Visualization, Big Data Management, and Fog Computing*, 67–80.
- Moahmmed, S. A., Alasow, M. A., & El-Alfy, E.-S. M. (2021). Clustering of Association Rules for Big Datasets using Hadoop MapReduce. *International Journal of Advanced Computer Science and Applications*, 12(3).
- Pauletic, I., Prskalo, L. N., & Bakaric, M. B. (2019). An overview of clustering models with an application to document clustering. 2019 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), 1659–1664.
- Sadeghi Moghadam, M. R., Safari, H., & Yousefi, N. (2021). Clustering quality management models and methods: systematic literature review and text-mining analysis approach. *Total Quality Management & Business*

- Excellence, 32(3-4), 241-264.
- Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications and research directions. *SN Computer Science*, 2(3), 160.
- Setyorini, S. G., Sari, E. K., Elita, L. R., & Putri, S. A. (2021). Analisis Keranjang Pasar Menggunakan Algoritma K-Means dan FP-Growth pada PT. Citra Mustika Pandawa: Market Basket Analysis with K-Means and FP-Growth Algorithm as Citra Mustika Pandawa Company. *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, 1(1), 41-46.
- Sharma, A., & Ganpati, A. (2021). Association rule mining algorithms: A Comparative review. *International Research Journal of Engineering and Technology*, 8(11), 848-853.
- Srinadh, V. (2022). Evaluation of Apriori, FP growth and Eclat Association rule mining algorithms. *International Journal of Health Sciences*, II, 7475-7485.
- Tahsien, S. M., Karimipour, H., & Spachos, P. (2020). Machine learning based solutions for security of Internet of Things (IoT): A survey. *Journal of Network and Computer Applications*, 161, 102630.
- Telikani, A., Gandomi, A. H., & Shahbahrami, A. (2020). A survey of evolutionary computation for association rule mining. *Information Sciences*, 524, 318-352.
- Ullmann, T., Hennig, C., & Boulesteix, A.-L. (2022). Validation of cluster analysis results on validation data: A systematic framework. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 12(3), e1444.
- Wicaksono, D., Jambak, M. I., & Saputra, D. M. (2020). The comparison of apriori algorithm with preprocessing and FP-growth algorithm for finding frequent data pattern in association rule. *Sriwijaya International Conference on Information Technology and Its Applications (SICONIAN 2019)*, 315-319.
- Yürüsen, N. Y., Uzunouglu, B., Talayero, A. P., & Estopiñán, A. L. (2021). Apriori and K-Means algorithms of machine learning for spatio-temporal solar generation balancing. *Renewable Energy*, 175, 702-717.
- Zhang, G., Liu, C., & Men, T. (2019). Research on data mining technology based on association rules algorithm. *2019 IEEE 8th Joint International Information Technology and Artificial Intelligence Conference (ITAIC)*, 526-530.