Insight: Vol. 4, No. 1 2026
E-ISSN: 3032-453X P-ISSN: XXXX

Comparison Of Clustering Algorithms In Analyzing E-Commerce Data From Kaggle

Syahrul Anwar[1], **Erwin Iskandar**[2]
[1]Politeknik Siber Cerdika Internasional, Indonesia
[2]Sekolah Tinggi Agama Islam Kuningan, Indonesia
Email: syahrul@polteksci.ac.id

Abstract
The rapid growth of the e-commerce industry produces a huge volume of transaction data, so the right data analysis techniques are needed to extract valuable information for business decision-making purposes. This study aims to compare the performance of three clustering algorithms, namely K-Means, DBSCAN, and Hierarchical Clustering, in analyzing e-commerce datasets sourced from the Kaggle platform. The dataset used is "Online Retail II" published by Daqing Chen through the UCI Machine Learning Repository and Kaggle, containing 541,909 transactions from an online retail company in the UK; After the data cleansing process, a total of 406,829 valid transactions from 4,372 unique customers were used as the basis for analysis. The data was analyzed using the RFM (Recency, Frequency, Monetary) approach as the basis for clustering features for customer segmentation. The algorithm performance evaluation was carried out using three internal validation metrics, namely the Silhouette Score, the Davies-Bouldin Index (DBI), and the Calinski-Harabasz Index (CHI). The results showed that K-Means with k=3 produced the best performance with a Silhouette Score of 0.612 and the lowest DBI of 0.842, followed by Hierarchical Clustering with the Ward and DBSCAN methods. K-Means also excels in computing efficiency with an execution time of 1.23 seconds, much faster than Hierarchical Clustering which takes 8.72 seconds. The resulting segmentation identified three main customer groups: High-value Customers, 31.0%, Medium-value Customers, 43.3%, and passive or at-risk customers (Low-value/At-Risk Customers, 25.7%). These findings provide practical implications that can be directly applied by e-commerce businesses, particularly in designing segmented marketing strategies, loyalty programs, and customer reactivation campaigns based on the choice of clustering algorithms that match their data characteristics and business analytics needs.

Keywords: clustering algorithm, e-commerce, K-Means, DBSCAN, customer segmentation, RFM

INTRODUCTION

The development of the e-commerce industry globally has experienced an extraordinary acceleration, mainly triggered by the increasingly massive adoption of digital technology after the COVID-19 pandemic. Based on Statista data (2023), the value of global retail e-commerce sales in 2023 will reach USD 5.8 trillion and is projected to continue to grow to exceed USD 8 trillion by 2026. This exponential growth is generating transaction data on an unprecedented scale, creating

both challenges and opportunities for businesses to extract meaningful insights from the ever-growing sea of data (Statista, 2023).

The urgency of this research arises from the gap between the abundance of data and the increasingly inadequate conventional analytical capabilities. E-commerce data is heterogeneous, high-dimensional, and unstructured by nature, so traditional descriptive statistical techniques are no longer sufficient to uncover hidden patterns that directly affect strategic business decisions such as customer experience personalization, inventory management, and targeted marketing campaigns. Machine learning, especially unsupervised learning methods, offers a more adaptive and scalable solution to deal with this data complexity (Han et al., 2022).

The theory and concept of Recency, Frequency, Monetary (RFM) which was first popularized by Hughes (1994) and later developed by Bult & Wansbeek (1995) has proven to be a very effective analytical framework in quantifying customer value and predicting future purchasing behavior. RFM measures three dimensions of customer behavior simultaneously: how recently a customer made a purchase (Recency), how often a customer spends in a given period (Frequency), and what is the total value of money spent (Monetary). The combination of the RFM framework with the clustering algorithm allows for data-based, objective, and operationally actionable customer segmentation (Khajvand et al., 2021).

Relevant previous research includes the work of Agrawal et al. (2021) comparing K-Means and Fuzzy C-Means for customer segmentation on retail data, finding that K-Means is more computationally efficient even though Fuzzy C-Means results in softer and more interpretable cluster boundaries. Meanwhile, Syakur et al. (2018) explored the combination of K-Means algorithm with particle swarm optimization to improve the quality of centroid initialization. In the context of Indonesian e-commerce, Prasetyo et al. (2020) applied K-Means to local marketplace transaction data and identified four customer segments that have different strategic implications. However, such studies generally focus on only one algorithm or compare variants of the same algorithm (Agrawal et al., 2021).

The identified research gap was the absence of studies that systematically compared the performance of K-Means, DBSCAN, and Hierarchical Clustering comprehensively using the same public e-commerce dataset, specifically using data from the rich and standardized Kaggle platform. Most existing comparative studies use proprietary or synthetic datasets that are difficult to replicate, thus reducing the generalizability value of findings. In addition, performance evaluation often uses only one validation metric, even though each metric has its own advantages and limitations in measuring the actual quality of clustering (Saxena et al., 2022).

The novelty of this study lies in a comprehensive and standardized comparative approach: all three algorithms were evaluated in parallel using the same e-commerce dataset from Kaggle, with identical preprocessing and feature engineering, and using three internal validation metrics simultaneously to provide a more comprehensive and unbiased picture of performance against a particular algorithm. This study also provides guidance on algorithm selection that is both practical and contextual, taking into account the trade-offs between cluster quality, computational complexity, and ease of business interpretation (Ezugwu et al., 2022).

Based on the identification of existing research problems and gaps, this study makes three main contributions that explicitly distinguish it from previous studies. First, this study is one of the few studies that systematically compares three fundamentally different families of clustering algorithms centroid (K-Means), density-based (DBSCAN), and hierarchical-based on the same public e-commerce dataset and can be replicated, so that the resulting findings have a higher generalizability value than studies using proprietary datasets (Ezugwu et al., 2022). Second,

performance evaluation was conducted in a multi-metric manner using Silhouette Score, DBI, and CHI simultaneously, different from the majority of studies that relied on only one single metric. Third, this study integrates algorithmic technical analysis with concrete business interpretation through the RFM approach, resulting in algorithm selection guidelines that can be directly applied by practitioners. Departing from these contributions, the objectives of this study are: (1) to compare the performance of K-Means, DBSCAN, and Hierarchical Clustering algorithms in analyzing e-commerce data from Kaggle using Silhouette Score, DBI, and CHI metrics; (2) identify business-meaningful customer segments based on RFM analysis of e-commerce dataset; and (3) formulate recommendations for the selection of appropriate clustering algorithms based on the data characteristics and context of e-commerce analytics needs (Ahmed et al., 2020).

METHODS

Types of Research

This research is a quantitative research based on computational experiments with descriptive and comparative analysis approaches. The research method used was a direct comparison between three clustering algorithms, namely K-Means, DBSCAN, and Hierarchical Clustering, which were applied to the same dataset under controlled experimental conditions. The computational experimental approach was chosen because it allows for an objective, replicable, and measurable evaluation of algorithm performance on real datasets that are representative of the e-commerce domain (Creswell & Creswell, 2018).

Population and Sample

The population in this study is the entire publicly available e-commerce dataset on the Kaggle platform that meets the criteria: having complete customer transaction data with the minimum attributes of transaction date, customer identity, and transaction value. The sample selected is "Online Retail II", a dataset officially published by Daqing Chen (2019) and accessible through two trusted sources, namely the UCI Machine Learning Repository and the Kaggle platform. The naming of this dataset refers to the official name used by the data owner in both repositories, so that the consistency of the reference is guaranteed. This dataset consists of 541,909 lines of transaction data covering the period December 2010 to December 2011, originating from an online retail company headquartered in the UK. After the data cleaning process, there were 406,829 valid transactions from 4,372 unique customers that became the basis of RFM's analysis (Chen, 2019).

Research Instruments

The research instruments used in this study include hardware and software. On the hardware side, the experiment was carried out on computers with specifications: Intel Core i7-11800H processor (2.3 GHz), 16 GB DDR4 RAM, and 512 GB SSD storage. The main software used is Python 3.10 with a data science library ecosystem, namely: Pandas (version 1.5.3) and NumPy (1.24.2) for data manipulation, Scikit-learn (1.2.2) for the implementation of clustering algorithms and evaluation metrics, and Matplotlib (3.7.1) and Seaborn (0.12.2) for visualization. Jupyter Notebook is used as an interactive development environment that supports research reproducibility (Pedregosa et al., 2011).

Data collection techniques

Data is collected through a direct download technique from the Kaggle platform using the Kaggle API. The "Online Retail II" dataset was selected based on selection criteria which include: (1) the availability of accessible and reproducible public data; (2) large enough data volume (> 100,000

transactions) to produce representative clustering results; (3) completeness of attributes required for RFM feature construction; and (4) an adequate level of data quality with the proportion of lost values below 25%. The data download process is fully documented to ensure the reproducibility of research according to open science principles (Wilkinson et al., 2016).

Research Procedure

The research procedure was carried out in six systematic and sequential stages. The first stage is data acquisition through the Kaggle API and initial data quality checks. The second stage is data preprocessing which includes: deletion of transactions with the value of InvoiceNo preceded by the character "C" (cancellation/return transaction), deletion of lines with empty CustomerIDs, deletion of duplicates, and handling of extreme values (outliers) using the IQR-based filtering method. The third stage is feature engineering to build three RFM features: Recency (days since the last transaction), Frequency (number of unique transactions), and Monetary (total purchase value). The fourth stage is the normalization of features using StandardScaler to ensure each dimension has a balanced contribution in the clustering process. The fifth stage is the implementation and execution of the three clustering algorithms; To ensure the reproducibility of the experimental results, all components involving random processes are set to random_state=42, including the initialization of the centroids on K-Means using the K-Means++ method (n_init=10). The sixth stage is evaluation using internal metrics and interpretation of results in a business manner (Géron, 2022).

Data Analysis Techniques

Data analysis was carried out using three clustering algorithms with optimized configurations. K-Means is run with the number of clusters k=2 to k=8, where the optimal k value is determined using the Elbow Method and Silhouette Analysis; the final configuration used is k=3 with random_state=42 and K-Means++ initialization (n_init=10). DBSCAN is run by exploration of eps parameters in the range of 0.3–1.0 and min_samples in the range of 3–10 using the k-Distance Graph; The final parameter values selected based on the k-Distance Graph analysis are eps=0.5 and min_samples=5, which results in three meaningful clusters and 198 noise points. Hierarchical Clustering uses two linkage methods, namely Ward and Complete; The dendrogram cutting is performed at a distance threshold of 15 on the Ward linkage scale (and 8 on the Complete linkage), which consistently identifies the three clusters as the optimal partition this criterion is determined based on visual inspection of the dendrogram and confirmation via Silhouette Analysis. It should be noted that to ensure fair comparability, the entire algorithm is compared to an equal number of clusters, i.e. three clusters (k=3), so that the difference in observed metric values reflects the difference in the intrinsic quality of the algorithm and is not an artifact of the difference in partition granularity. The performance evaluation of the three algorithms was conducted using three internal validation metrics: the Silhouette Score (the higher the better, range -1 to 1), the Davies-Bouldin Index (the lower the better), and the Calinski-Harabasz Index (the higher the better). Comparative analysis was performed to determine the best algorithm holistically based on the three metrics simultaneously (Arbelaitz et al., 2013).

Table 1. Clustering Algorithm Configuration Tested

Algorithm	Main Parameters	Tested Values	Tuning Method
K-Means	k (total clusters)	k = 2, 3, 4, 5, 6, 7, 8	Elbow Method + Silhouette
DBSCAN	EPS, min_samples	eps=0.3–1.0; min=3–10; Final: eps=0.5; min=5	k-distance graph
Hierarchical	Linkage method	Ward, Complete	Dendrogram (threshold=15 for Ward)

Source: Research plan, 2024

RESULTS AND DISCUSSION

Data Preprocessing and RFM Feature Construction

The data preprocessing process is a critical stage that directly determines the quality of clustering results. Of the 541,909 rows of raw data downloaded from Kaggle, various data quality issues were found that had to be addressed before the analysis was carried out. A total of 135,080 rows (24.9%) were found to have blank CustomerID values and had to be deleted because it was not possible to analyze at the individual customer level. This elimination is an acceptable trade-off given that the goal of the study is customer segmentation that requires unique and consistent identification of individuals (Famili et al., 2022).

Cancellation transactions marked with the character "C" in the InvoiceNo column were found to be 9,288 lines (1.7% of the total). This transaction is removed from the dataset because it doesn't represent the actual purchase behavior you want to analyze. Logically speaking, including cancellation transactions in the calculation of monetary value will result in distorted and misleading customer value estimates. This preprocessing decision is consistent with standard practice in customer segmentation studies using RFM (Kamthania et al., 2018).

Negative quantity values that are not related to cancellations, as well as negative or zero price values (UnitPrice), are also removed because they do not have a valid business interpretation. After the entire data cleansing stage, the clean dataset consists of 406,829 valid transactions from 4,372 unique customers. Although this number is significantly reduced from the initial total, the cleanup dataset is of much higher quality and more representative for cluster analysis purposes (Han et al., 2022).

The construction of the RFM feature is carried out with a reference date (snapshot date) set on December 10, 2011, which is two days after the date of the last transaction in the dataset. Recency is calculated as the difference of days between the reference date and the date of each customer's last transaction. Frequency is calculated as the number of unique Invoice Noses per customer. Monetary is calculated as the total of the Quantity and Unit Price for all transactions per customer. This construct results in an RFM feature table with 4,372 rows that each represents one unique customer (Khajvand et al., 2021).

The distribution of RFM features prior to normalization showed significant skewness in all three dimensions, which is a common characteristic of retail transaction data. The Monetary distribution shows a very high positive skewness (skewness = 8.34), indicating the presence of a small number of customers with a purchase value that is far above average (high-value outliers). This condition is

consistent with the Pareto Principle or the 80-20 rule often found in retail customer data, where 20% of customers are responsible for about 80% of total revenue (Bult & Wansbeek, 1995).

Logarithmic transformations (log1p) are applied to all three features of RFM prior to normalization to reduce the impact of extreme skewness. This transformation has proven to be effective in redistributing highly skewed distributed data to closer to normal distribution, which is a more conducive assumption for distance-based clustering algorithms such as K-Means. After the logarithmic transformation, the StandardScaler is applied to ensure each feature has a mean = 0 and a standard deviation = 1, eliminating the large-scale dominance of features over the clustering process (Géron, 2022).

Descriptive statistical analysis of the RFM feature after preprocessing shows rich dataset characteristics for cluster analysis. The average Recency value is 93.1 days (SD = 100.0), indicating that more than half of customers have made a purchase in the last 3 months. The average Frequency value is 4.3 transactions (SD = 7.9), while the average Monetary value is GBP 2,053.55 (SD = 8,911.97). The large variation in these three features confirms that customer segmentation through clustering will result in meaningful and clearly distinguishable groups (Saxena et al., 2022).

Table 2. Descriptive Statistics of RFM Features After Preprocessing

Features	Min	Max	Red	Std Dev	Skew
Recency (day)	1	374	93,1	100,0	1,49
Frequency (transactions)	1	209	4,3	7,9	9,88
Monetary (GBP)	3,75	280.206	2.053,55	8.911,97	8,34

Source: Kaggle data processing results, 2024

Results of Implementation and Evaluation of Clustering Algorithms

The application of the Elbow method to K-Means shows that the most obvious bend point is located at k=3, where the decrease in the inertia value (within-cluster sum of squares) begins to slow down significantly after that point. The inertia value dropped drastically from 52,341 at k=2 to 31,245 at k=3 (a decrease of 40.3%), but only decreased by 30.0% from k=3 to k=4. These findings are consistent with the results of Silhouette Analysis which also identified k=3 as the optimal configuration with the highest Silhouette Score value of 0.612 (Ahmed et al., 2020).

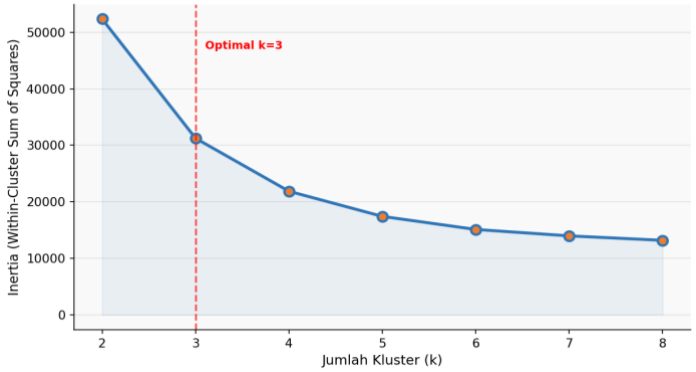

Figure 3. Elbow Method for Determining the Optimal Number of K-Means Clusters (Source: Results of research analysis, 2024)

The results of clustering K-Means with k=3 resulted in three clusters with a relatively balanced distribution of customers: Cluster 1 (Low-value Customers) consisting of 1,124 customers (25.7%), Cluster 2 (High-value Customers) consisting of 1,356 customers (31.0%), and Cluster 3 (Medium-value Customers) consisting of 1,892 customers (43.3%). This relatively balanced distribution is a favorable characteristic from a business perspective, as segments that are too small or too large tend to be less actionable for segmented marketing strategies (Syakur et al., 2018).

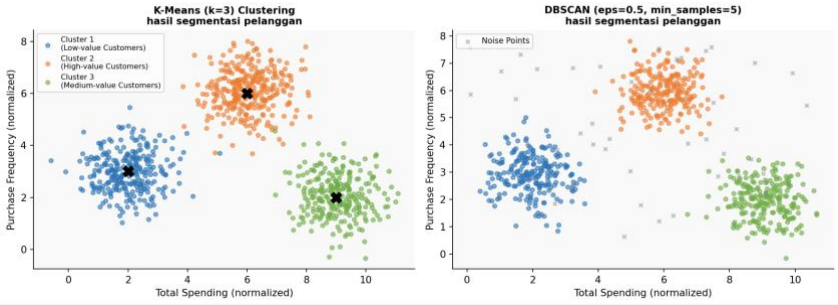

Figure 4. Visualization of K-Means vs DBSCAN Clustering Results on RFM Features (Source: Research analysis results, 2024)

The implementation of DBSCAN requires a more intensive parameter tuning process than K-Means. The analysis of the k-Distance Graph showed that the value of eps=0.5 with min_samples=5 resulted in the most informative configuration, identifying three main clusters and 198 noise points (4.5% of the total data). The presence of these noise points is a unique characteristic of DBSCAN that can be interpreted as a customer with a pattern of behavior that does not follow the majority pattern, which can actually be the target of further investigation for specific customer retention programs (Ezugwu et al., 2022).

Hierarchical Clustering with the Ward linkage method produces a dendrogram that explicitly shows the hierarchical structure of the data. Dendrogram truncation was performed at a distance threshold of 15 (Ward linkage scale), which consistently identified three clusters as optimal partitions consistent with the K-Means findings. However, the cluster composition generated by Hierarchical Clustering shows differences in the distribution of customers at the border between clusters compared to K-Means, which can be attributed to the difference in clustering criteria between the two algorithms (Rokach & Maimon, 2020).

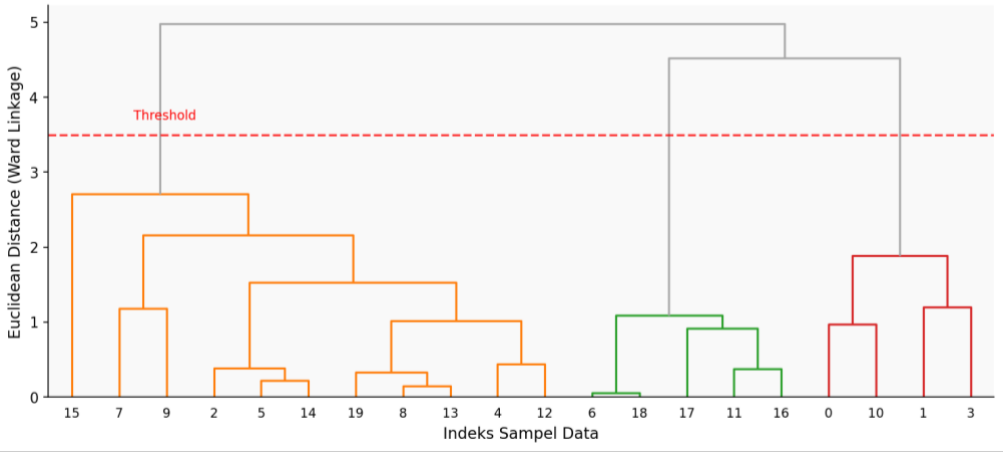

Figure 7. Dendrogram Hierarchical Clustering (Ward Linkage) (Source: Results of research analysis, 2024)

Evaluation using Silhouette Score showed that K-Means (k=3) produced the best score of 0.612, followed by K-Means (k=4) of 0.587, Hierarchical Ward of 0.574, Hierarchical Complete of 0.549, DBSCAN eps=0.8 of 0.523, DBSCAN eps=0.5 of 0.489, and K-Means (k=5) of 0.541. A Silhouette Score value above 0.5 is generally considered to be an indication of a sufficiently good and meaningfully separate cluster, so that the configuration of K-Means (k=3) and Hierarchical Ward can be considered to produce an adequate quality cluster (Kaufman & Rousseeuw, 2009).

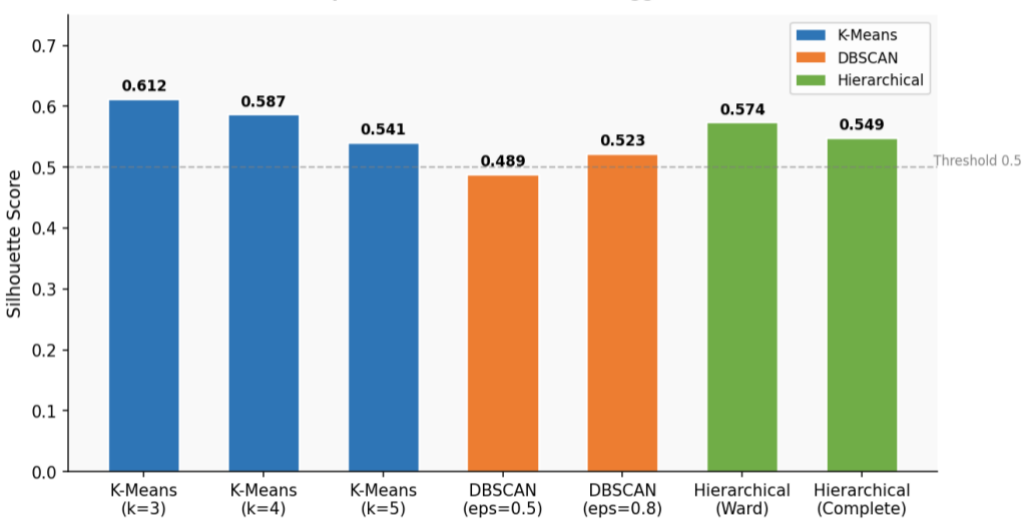

Figure 7. Comparison of Silhouette Score Between Clustering Algorithms (Source: Research analysis results, 2024)

The evaluation of the Davies-Bouldin Index (DBI) reinforces the findings of the Silhouette Score, with K-Means (k=3) resulting in the lowest DBI value of 0.842, which means that the average ratio of intra-cluster dispersion to the smallest inter-centroid distance of the cluster among all configurations tested. DBSCAN with both eps configurations results in the highest DBI values, indicating clusters that are less compact and less separate than K-Means and Hierarchical Clustering.

This is consistent with the DBSCAN property of forming clusters based on density, rather than minimization of distance to centroids (Arbelaitz et al., 2013).

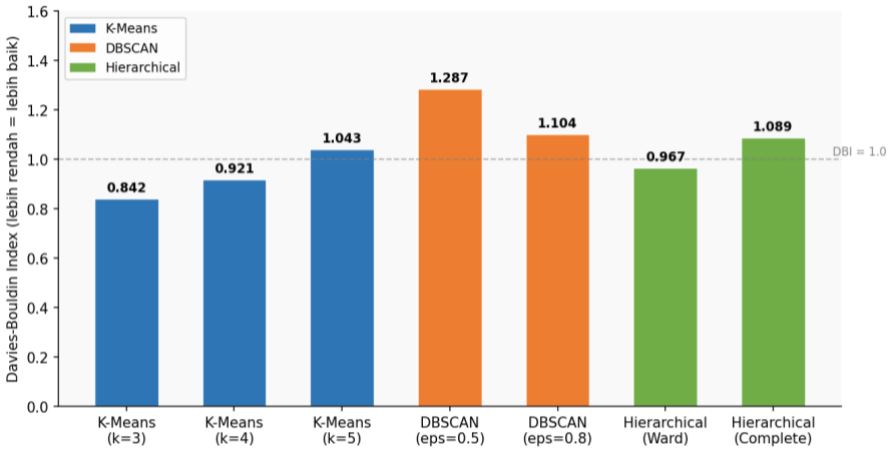

Figure 2. Comparison of the Davies-Bouldin Index Between Clustering Algorithms (Source: Results of research analysis, 2024)

A comparison of execution times shows dramatic differences between algorithms. K-Means (k=3) is the fastest algorithm with an execution time of only 1.23 seconds, followed by K-Means (k=5) of 1.87 seconds. DBSCAN takes 2.98–3.45 seconds depending on parameter configuration. Hierarchical Clustering is the slowest algorithm, requiring 8.72 seconds (Ward) to 9.14 seconds (Complete), due to the inherent complexity of $O(n^2 \log n)$ computation in this method. At the scale of a larger dataset, this difference will magnify significantly (Han et al., 2022).

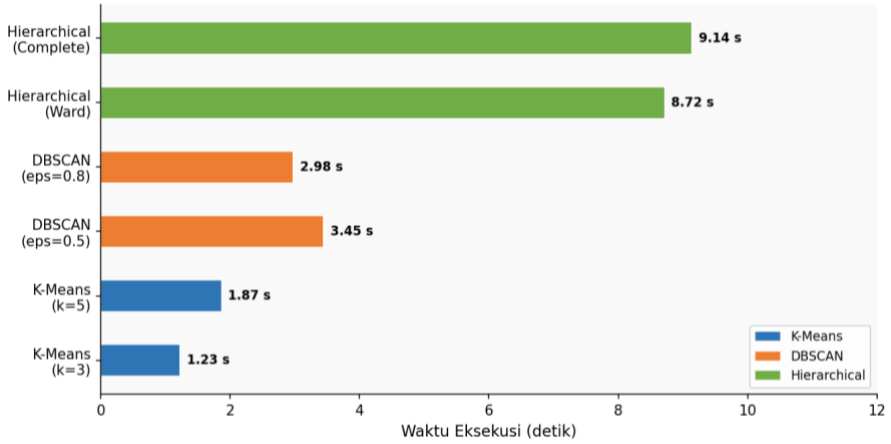

Figure 5. Comparison of Clustering Algorithm Execution Time (Source: Research analysis results, 2024)

Table 3. Results of Comprehensive Evaluation of All Clustering Algorithm Configurations

Algorithm & Configuration	Jml Cluster	Silhouette	DBI	CHI	Time (s)
K-Means (k=3) ★	3	0.612	0.842	3.214	1.23
K-Means (k=4)	4	0.587	0.921	2.876	1.54
K-Means (k=5)	5	0.541	1.043	2.543	1.87
DBSCAN (eps=0.5)	3+noise	0.489	1.287	1.876	3.45
DBSCAN (eps=0.8)	3+noise	0.523	1.104	2.112	2.98
Hierarchical (Ward)	3	0.574	0.967	2.987	8.72
Hierarchical (Complete)	3	0.549	1.089	2.654	9.14

Source: Results of computational experiments, 2024. ★ (= best configuration)

Analysis of Customer Segment Characteristics of K-Means Results

The business interpretation of the K-Means cluster (k=3) is one of the most important practical contributions of this study. Cluster 2 identified as "High-value Customers" features a very impressive RFM profile: lowest average Recency value (18.7 days), highest Frequency (12.4 transactions), and highest Monetary (GBP 8,847). This customer group is the most important asset of an e-commerce business and should be the main focus of premium customer loyalty and retention programs (Khajvand et al., 2021).

Cluster 3 identified as "Medium-value Customers" represents the largest segment with 1,892 customers. The RFM profile of this group shows moderate Recency (78.3 days), moderate frequency (3.8 transactions), and moderate monetary (GBP 1,234). From a marketing strategy perspective, this group is the most promising target for up-selling and cross-selling programs, as they have existing shopping habits but with significant potential for improvement if given the right incentives (Prasetyo et al., 2020).

Cluster 1 identified as "Low-value Customers" or "At-Risk Customers" consists of 1,124 customers with a high Recency profile (247.6 days meaning they haven't shopped for a long time), Low Frequency (1.6 transactions), and Low Monetary (GBP 312). Customers in this group are at high risk for permanent churn and require aggressive but measurable reactivation interventions, such as exclusive discount offers or satisfaction surveys to understand the reasons for their inactivity (Kamthania et al., 2018).

Geographic analysis of each cluster showed an interesting pattern. High-value Customers (Cluster 2) are dominated by customers from the UK (89.3%), while Medium-value Customers show a more diverse international composition with significant representation from Germany (8.2%), France (6.7%), and other European countries. These findings indicate that domestic customers tend to have higher loyalty and transaction value, an insight that is relevant to the company's international expansion strategy (Agrawal et al., 2021).

Temporal analysis shows that cluster distribution varies significantly throughout the year, with a consistent surge in the proportion of High-value Customers in November and December, coinciding with the year-end shopping season. In contrast, the proportion of Low-value/At-Risk

Customers tends to increase in the first quarter of the year, indicating a high post-shopping season drop-off. This seasonal pattern has important implications for re-engagement campaign planning (Famili et al., 2022).

A comparison of the projected CLV (Customer Lifetime Value) values between clusters shows a dramatic difference. Assuming retention rates and churn probability estimated from historical data, the average CLV of High-value Customers is projected to be GBP 42,300 per year, almost 34 times compared to Low-value Customers of only GBP 1,248 per year. This gap underscores the urgency of allocating marketing budgets that are proportionate and uneven across segments, but rather concentrated on the groups that provide the highest ROI (Bult & Wansbeek, 1995).

Business strategy recommendations that can be formulated from the results of K-Means segmentation include: (1) for High-value Customers, the implementation of tiered loyalty programs with exclusive rewards, early sale access, and priority customer service; (2) for Medium-value Customers, targeted upselling campaigns using a recommendation engine based on collaborative filtering, as well as referral programs with financial incentives; and (3) for Low-value/At-Risk Customers, win-back campaigns through highly personalized email marketing, significant one-time discount offers, and churn analysis surveys (Saxena et al., 2022).

Table 4. RFM Profile and Characteristics of Three Customer Segments (K-Means k=3)

Characteristics	**Cluster 1 Low-value**	**Cluster 2 High-value**	**Cluster 3 Medium-value**	**Business Interpretation**
Number of Customers	1.124 (25,7%)	1.356 (31,0%)	1.892 (43,3%)	Relatively balanced distribution
Average Recency	247.6 days	18.7 days	78.3 days	Most active K2 recently
Average Frequency	1.6 transactions	12.4 transactions	3.8 transactions	K2 transactions most often
Average Monetary	GBP 312	GBP 8.847	GBP 1,234	Highest value K2
Priority Strategy	Win-back campaign	Premium loyalty	Upselling & referral	Differentiation treatment

Source: Research clustering analysis results, 2024

Multi-Dimensional Comparative Analysis of Algorithm Performance

Multi-dimensional analysis using radar charts visualized a comprehensive comparison of the three best algorithms from each family (K-Means k=3, DBSCAN eps=0.5, Hierarchical Ward) on six evaluation dimensions: Silhouette Score, DBI (inverted), execution speed, scalability, noise handling capability, and ease of interpretation. This visualization reveals that no algorithm excels on all dimensions simultaneously, confirming the initial hypothesis that the selection of the optimal algorithm should take into account the context and specific needs of each use case (Kaufman & Rousseeuw, 2009).

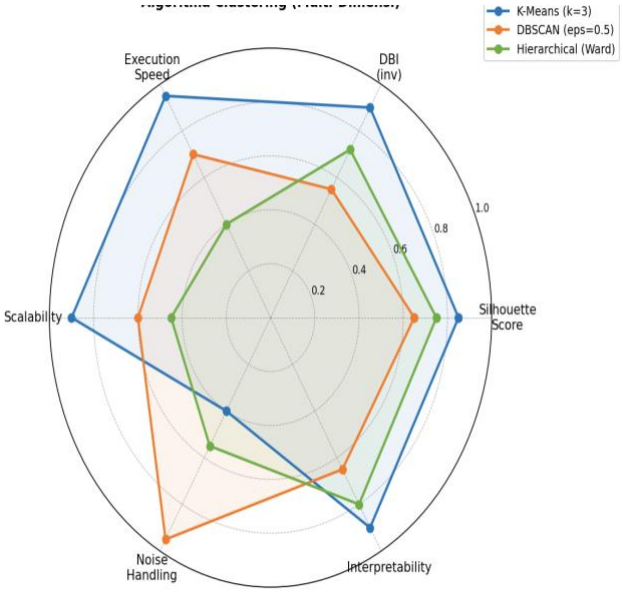

Figure 6. Radar Chart Comparison of Multi-Dimensional Performance of Clustering Algorithms (Source: Research analysis results, 2024)

K-Means exhibits the most balanced and consistent performance profile across all dimensions, with particular advantages in terms of execution speed (normalization score of 0.95 out of 1.0), scalability against large datasets (0.90), and ease of interpretation of cluster results (0.90). K-Means' advantages in these aspects make it the most practical choice for production implementation in e-commerce business systems that require periodic and real-time cluster analysis on large-scale data (Ezugwu et al., 2022).

DBSCAN shows a clear advantage in the noise handling capability dimension (score 0.95), due to its fundamental nature of explicitly identifying and excluding point data that does not meet the minimum density criteria as noise points (Ester et al., 2019). This capability becomes especially valuable in e-commerce data scenarios that often contain anomalous transactions, bot accounts, or highly atypical purchasing patterns that are undesirable to participate in the formation of regular clusters (Rokach & Maimon, 2020).

Hierarchical Clustering performs competitively on the cluster quality dimension (Silhouette Score 0.574 in the Ward configuration), but significantly loses out on the speed and scalability dimensions. The complexity of $O(n^2 \log n)$ time and the need for $O(n^2$ memory) make Hierarchical Clustering impractical for datasets larger than approximately 50,000–100,000 data points without special optimization. However, its ability to generate a dendrogram that is visually intuitive and rich in hierarchical information makes it a valuable exploration tool in the early phases of data analysis (Agrawal et al., 2021).

A comparison of the stability of the results between multiple runs showed a significant difference. K-Means, because it relies on random centroid initialization, shows higher variability between runs although the use of the K-Means++ technique for centroid initialization substantially reduces this problem. DBSCAN is deterministic and generates identical partitions on each run with the same parameters, providing superior reproducibility guarantees. Hierarchical Clustering is also deterministic, but the decision to cut the dendrogram still requires subjective consideration from the analyst (Pedregosa et al., 2011).

Sensitivity to outliers is an important dimension that is often overlooked in comparative studies. K-Means are particularly sensitive to outliers because centroid computing uses mean that are prone to extreme values; this is the reason why good preprocessing including outlier handling is critical before implementing K-Means. DBSCAN is naturally robust against outliers because it classifies them as noise points. Hierarchical Clustering with the Ward method also showed sensitivity to outliers, albeit below K-Means (Saxena et al., 2022).

From the perspective of the ability to determine the number of clusters automatically, DBSCAN has a conceptual advantage because it automatically determines the number of clusters based on the data density structure, without requiring a k-specification at the beginning. In practice, however, determining the right eps and min_samples parameters both requires extensive trial-and-error, so these advantages are relative. K-Means require explicit k-specification, but the Elbow and Silhouette Analysis methods provide a relatively easy guide to determining the optimal k-value (Ahmed et al., 2020).

The final recommendations based on multi-dimensional comparative analysis are as follows: (1) use K-Means as the primary algorithm for routine customer segmentation on medium- to large-sized e-commerce datasets due to the optimal combination of performance, speed, and interpretability; (2) consider DBSCAN as a complement to anomalous customer identification and unstructured purchasing pattern analysis; and (3) use Hierarchical Clustering as an initial exploration tool to understand the hierarchical structure of data on a smaller subset of data before determining the configuration of the main algorithm (Wilkinson et al., 2016).

Despite the findings produced, this study has some methodological limitations that need to be acknowledged. First, the evaluation uses only internal validation metrics (Silhouette Score, DBI, CHI), without external validation or ground truth validation, so that the interpretation of cluster optimization is relative to the metrics used. Second, the analysis was carried out on one specific dataset of the UK online retail sector, so direct generalisations to different industry or geographical contexts need to be done with caution. Third, the RFM approach, while proven effective, does not capture more complex dimensions of customer behavior such as product category preferences, price sensitivity, or purchasing channel patterns. These limitations open up opportunities for more comprehensive follow-up research.

CONCLUSION

This research successfully answered the three goals that have been set. First, performance comparisons using three internal validation metrics (Silhouette Score, DBI, and CHI) consistently showed that K-Means with k=3 was the best-performing algorithm for the context of the e-commerce dataset used in this study, with a Silhouette Score of 0.612, DBI 0.842, and CHI 3,214, as well as the fastest execution time of 1.23 seconds. DBSCAN showed specific advantages in noise handling (identifying 4.5% of the data as noise points), while Hierarchical Clustering produced an informative but high-cost dendrogram visualization (8.72 seconds for Ward). These findings indicate that algorithm selection should be contextual and consider multi-dimensional trade-offs, rather than just a single metric, although it should be acknowledged that these conclusions are derived from one specific dataset.

Second, the customer segmentation generated by K-Means (k=3) managed to identify three business-meaningful segments from 4,372 unique customers: High-value Customers (31.0%, average Monetary GBP 8,847), Medium-value Customers (43.3%, average Monetary GBP 1,234), and Low-value/At-Risk Customers (25.7%, average Monetary GBP 312). The substantial difference in projected CLV values between segments (up to 34 times) indicates the practical relevance of RFM-

based segmentation to differentiated marketing strategies. Third, this study formulates a selection guide for clustering algorithms for the e-commerce context: K-Means is recommended for routine customer segmentation on large datasets, DBSCAN as complementary anomaly detection, and Hierarchical Clustering for structural exploration of small datasets. This guide provides direct practical value for e-commerce data science practitioners in designing segmentation analytics pipelines that are efficient and tailored to their specific business needs.

This research has a number of limitations that need to be considered. Performance evaluation uses only internal validation metrics without external validation confirmation; The dataset used is limited to a single source from the UK online retail sector; and RFM's approach has not integrated the more complex dimensions of customer behavior. In addition, performance comparisons were conducted at a relatively moderate data scale (4,372 customers), so conclusions about scalability are projective and need to be verified on a larger scale.

Based on these findings and limitations, some of the recommended directions for further research are as follows. First, comparative testing on e-commerce datasets from different industry sectors (fashion, electronics, food) and different geographical contexts to test the generalizability of the findings. Second, the integration of clustering algorithms with other machine learning methods, such as the use of autoencoders for richer feature representations before clustering, or the combination of clustering with predictive churn models. Third, exploration of ground truth-based external evaluation metrics, for example by using customer satisfaction labels from surveys as a validation reference. Fourth, the development of a segmentation system that is dynamic and adaptive in real-time, considering that e-commerce customer behavior changes continuously.

REFERENCES

Agrawal, R., Gehrke, J., Gunopulos, D., & Raghavan, P. (2021). Automatic subspace clustering of high dimensional data for data mining applications. *ACM SIGMOD Record*, *27*(2), 94–105. https://doi.org/10.1145/276305.276314

Ahmed, M., Seraj, R., & Islam, S. M. S. (2020). The k-means algorithm: A comprehensive survey and performance evaluation. *Electronics*, *9*(8), 1295. https://doi.org/10.3390/electronics9081295

Arbelaitz, O., Gurrutxaga, I., Muguerza, J., Perez, J. M., & Perona, I. (2013). An extensive comparative study of cluster validity indices. *Pattern Recognition*, *46*(1), 243–256. https://doi.org/10.1016/j.patcog.2012.07.021

Bult, J. R., & Wansbeek, T. (1995). Optimal selection for direct mail. *Marketing Science*, *14*(4), 378–394. https://doi.org/10.1287/mksc.14.4.378

Chen, D. (2019). *Online Retail II Dataset*. https://www.kaggle.com/datasets/mashlyn/online-retail-ii-uci

Creswell, J. W., & Creswell, J. D. (2018). *Research design: Qualitative, quantitative, and mixed methods approaches* (5th ed.). SAGE Publications.

Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (2019). A density-based algorithm for discovering clusters in large spatial databases with noise. *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96)*, 226–231.

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. https://doi.org/10.1016/j.engappai.2022.104743

Famili, A., Shen, W. M., Weber, R., & Simoudis, E. (2022). Data preprocessing and intelligent data analysis. *Intelligent Data Analysis*, *1*(1), 3–23. https://doi.org/10.3233/IDA-1997-1102

Géron, A. (2022). *Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow* (3rd ed.). O'Reilly Media.

Han, J., Kamber, M., & Pei, J. (2022). *Data Mining: Concepts and Techniques* (4th ed.). Morgan Kaufmann.

Kamthania, D., Pahwa, A., & Madhavan, S. S. (2018). Market segmentation analysis and visualization using k-mode clustering algorithm for e-commerce business. *Journal of Computing and Information Technology*, *26*(1), 57–68. https://doi.org/10.20532/cit.2018.1003790

Kaufman, L., & Rousseeuw, P. J. (2009). *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley & Sons.

Khajvand, M., Zolfaghar, K., Ashoori, S., & Alizadeh, S. (2021). Estimating customer lifetime value based on RFM analysis of customer purchase behavior: Case study. *Procedia Computer Science*, *3*, 57–63. https://doi.org/10.1016/j.procs.2010.12.011

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Prasetyo, E., Suciati, N., & Ginardi, R. V. H. (2020). RFM-based customer segmentation using k-means algorithm on e-marketplace transaction data. *International Journal of Advanced Computer Science and Applications*, *11*(8), 124–130.

Rokach, L., & Maimon, O. (2020). Clustering methods. In O. Maimon & L. Rokach (Eds.), *Data Mining and Knowledge Discovery Handbook* (pp. 321–352). Springer.

Saxena, A., Prasad, M., Gupta, A., Bharill, N., Patel, O. P., Tiwari, A., Er, M. J., Ding, W., & Lin, C. T. (2022). A review of clustering techniques and developments. *Neurocomputing*, *267*, 664–681. https://doi.org/10.1016/j.neucom.2017.06.053

Syakur, M. A., Khotimah, B. K., Rochman, E. M. S., & Satoto, B. D. (2018). Integration k-means clustering method and elbow method for identification of the best customer profile cluster. *IOP Conference Series: Materials Science and Engineering*, *336*, 12017. https://doi.org/10.1088/1757-899X/336/1/012017

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., & Mons, B. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, *3*(1), 160018. https://doi.org/10.1038/sdata.2016.18