Vol. 8, No. 3, June 2026

Journal of Information Systems and Informatics

Published By Asosiasi Doktor Sistem Informasi Indonesia

Customer Segmentation in an Internet Service Provider: A K-Means Case Study of Telecommunication Company

Merleen Januar[1], Didi Supriyadi[2]

[1]Information Systems Study Program, Telkom University, Purwokerto Campus, Purwokerto Indonesia

[2]Center of Excellence for Sustainability Cities, Village, and Food Security, Research Institute for Intelligent Business and Sustainable Economy, Telkom University, Purwokerto, Central Java, Indonesia

Received:
October 30, 2025
Revised:
May 10, 2026
Accepted:
May 27 2026
Published:
June 22, 2026

Corresponding Author:
Author Name*:
Didi Supriyadi
Email*:
didisupriyadi@telkomuniversity.ac.id

DOI:
10.63158/journalisi.v8i3.1631

Abstract. PT Lintas Jaringan Nusantara, an internet service provider, faces challenges in utilizing customer data, which is mainly used for administrative purposes such as billing and support, limiting deeper analysis. This study applies K-Means clustering under the CRISP-DM framework for customer segmentation-based service-oriented attributes: internet package, price, and NAS location, using 972 customer records. Categorical attributes were transformed using frequency encoding and manual mapping. Model evaluation using the Elbow Method suggested 3 clusters, while the Silhouette Coefficient indicated that 10 clusters were optimal, improving the score from 0.5471 to 0.7704. The resulting clusters show variations in customer characteristics and provide an exploratory overview of grouping patterns. However, the 10 clusters solution should not yet operationally validated, as stakeholder validation involving marketing and customer service teams is still required to assess interpretability, business relevance, and practical applicability. Further validation using additional customer data or alternative datasets is also recommended. Overall, the findings serve as an initial analytical step to support future data-driven decision-making.

Keywords: Customer Segmentation, Internet Service Provider, K-Means Clustering, Silhouette Coefficient, Elbow Method, Telecommunication Customer Analytics

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

1. INTRODUCTION

Telecommunications play a vital role in economic development by providing secure and high-speed infrastructure while enabling digital transformation across sectors, telecommunication standards further support market growth by enabling interoperability, competition, and efficiency in service delivery [1]. The telecommunications industry is highly competitive, with firms competing for customers, market share, and sustainability, intense competition and diverse service alternatives have increased customer switching behavior, although conducted as a case study at PT Lintas Jaringan Nusantara, this study may offer analytical insights relevant to other high-churn service industries [2]. A company's understanding of customer behaviour and preferences is crucial for maintaining competitiveness, internet service providers must identify customer needs to ensure their services are well accepted, while recognizing factors that influence satisfaction is essential for building customer loyalty [3]. PT Lintas Jaringan Nusantara serves as the case study context in this research and has a substantial customer base across Indonesia. The company provides various IT-based services, including IT hardware procurement, Local Area Network (LAN) configuration, system setup, and network infrastructure maintenance. However, based on observations, customer data are primarily utilized for administrative purposes such as billing management and basic service handling. Without customer segmentation, the company risks facing difficulties in identifying customer groups [4].

Customer segmentation divides customers into groups with similar service characteristics, preferences or usage patterns, it supports service improvement and Segmentation, Targeting, and Positioning (STP)-based marketing and in telecommunications, it is important due to differences in digital literacy, usage patterns, and income level [5]. Customer segmentation can be achieved by leveraging clustering algorithms, enabling companies to gain deeper insights into customer heterogeneity, encompassing preferences for services types, usage patterns, and price sensitivity levels [6]. Therefore cluster-based customer segmentation plays an important role in customer relationship management (CRM) by grouping customers based on shared needs characteristics and behaviors helping businesses improve customer relationships deliver better services and gain competitive advantages while maximizing profits [7]. Previous studies have demonstrated that utilizing clustering algorithms provides deeper insights

into the diversity of customer characteristics across various industries including retail, e-commerce, and telecommunications [8], [9], [10]. Despite its limitations, the clustering algorithm is well known for its flexibility, efficiency, and ease of implementation, its wide use across various clustering applications is largely due to its simple implementation process and low computational complexity [11].

Utilizing clustering algorithms, a decision support model was developed to categorize customers based on service-oriented attributes including internet package type, pricing and NAS location. The selection of internet package type, pricing and NAS location is based on the availability of customer data within the company, as detailed demographic and behavioral information is not consistently recorded in the existing customer database. Despite this limitation, these three attributes are considered sufficient for an initial exploratory segmentation, as they represent key service-oriented dimensions that reflect how customers interact with and access the provider's services. Internet package type captures the variation in service offerings consumed by customers, pricing reflects differences in economic engagement levels, and NAS location provides insight into geographical distribution of service usage. Cross-Industry Standard Process for Data Mining (CRISP-DM) is widely recognized as a de facto standard and an industry independent process model for executing data mining projects, and it continues to be extensively adopted in both research and practical applications. As a general framework, it emphasizes the importance of defining clear data mining objectives, particularly within the business understanding phase, building on this foundation, our approach incorporates the formulation of system-specific research questions to refine objectives and support the effective application of the CRISP-DM framework, given its broad acceptance, the integration of our proposed method can be implemented with minimal difficulty [12].

Previous studies have extensively explored customer segmentation, such as the study by Vieri et al. [10] in the telecommunications industry based on customer status whether active subscribers or not, monthly payment history, total payments, subscription duration, and types of services used by customers, as well as study by Adin et al. [13] which conducted segmentation through demographic analysis examining dominant age groups and genders in each segment, income analysis identifying segments contributing the highest revenue to specific services like voice calls, mobile data, and others, product subscription analysis determining top segments subscribing to each product, and

behavioral analysis providing in depth examination of usage indicators in the dataset including average monthly voice calls, SMS, data, and roaming, however, despite the comprehensiveness of these approaches, existing studies tend to overlook the integration of pricing and geographical distribution factors, which may limit the ability of segmentation models to capture variations in customer purchasing power and regional service accessibility, thereby reducing their effectiveness in supporting more targeted and context-aware marketing strategies. Additionally, previous studies by Ramesh & Bhuvaneswari. [14] in the telecommunication network context have predominantly focused on technical aspects such as network performance optimization, signal quality, and service fulfillment based on network parameters, and while such approaches effectively enhance system efficiency and service quality they still exhibit limitations in providing insights into customer characteristics from behavioral and service preferences perspectives, with factors representing economic conditions and customer service needs remaining underutilized in segmentation processes despite their role in reflecting purchasing power, access requirements, and customer profile differences. This study contributes by optimizing the strategic value of customer data for PT Lintas Jaringan Nusantara through the development of more precise and measurable customer segmentation using clustering algorithms in the telecommunications industry, and is conducted as a case study at PT Lintas Jaringan Nusantara.

2. METHODS

This research applies clustering techniques using the K-Means algorithm to identify the characteristics and preferences of each customer group, the clustering process is carried out systematically following the CRISP-DM framework, which includes stages ranging from business understanding, data understanding, data preparation, modeling, to evaluation model as illustrated in figure 1. The framework commences with the business understanding phase focusing on understanding the business context, identifying the main problems to be addressed, and determining how the proposed solution will be implemented; furthermore, In this study, the phase emphasizes the identification of important factors such as business objectives, business requirements, and success criteria, understanding business requirements serves as an essential foundation for directing the model development process to align with the research objectives and ensure that the resulting outcomes are relevant to the company's operational needs—a

process which is then followed by the data understanding phase, this subsequent stage aims to gain a comprehensive understanding of the data used for customer clustering, including data exploration and the identification of data sources, as well as an analysis of data structures and characteristics to obtain deeper insights before finally proceeding to the data preparation stage [15].

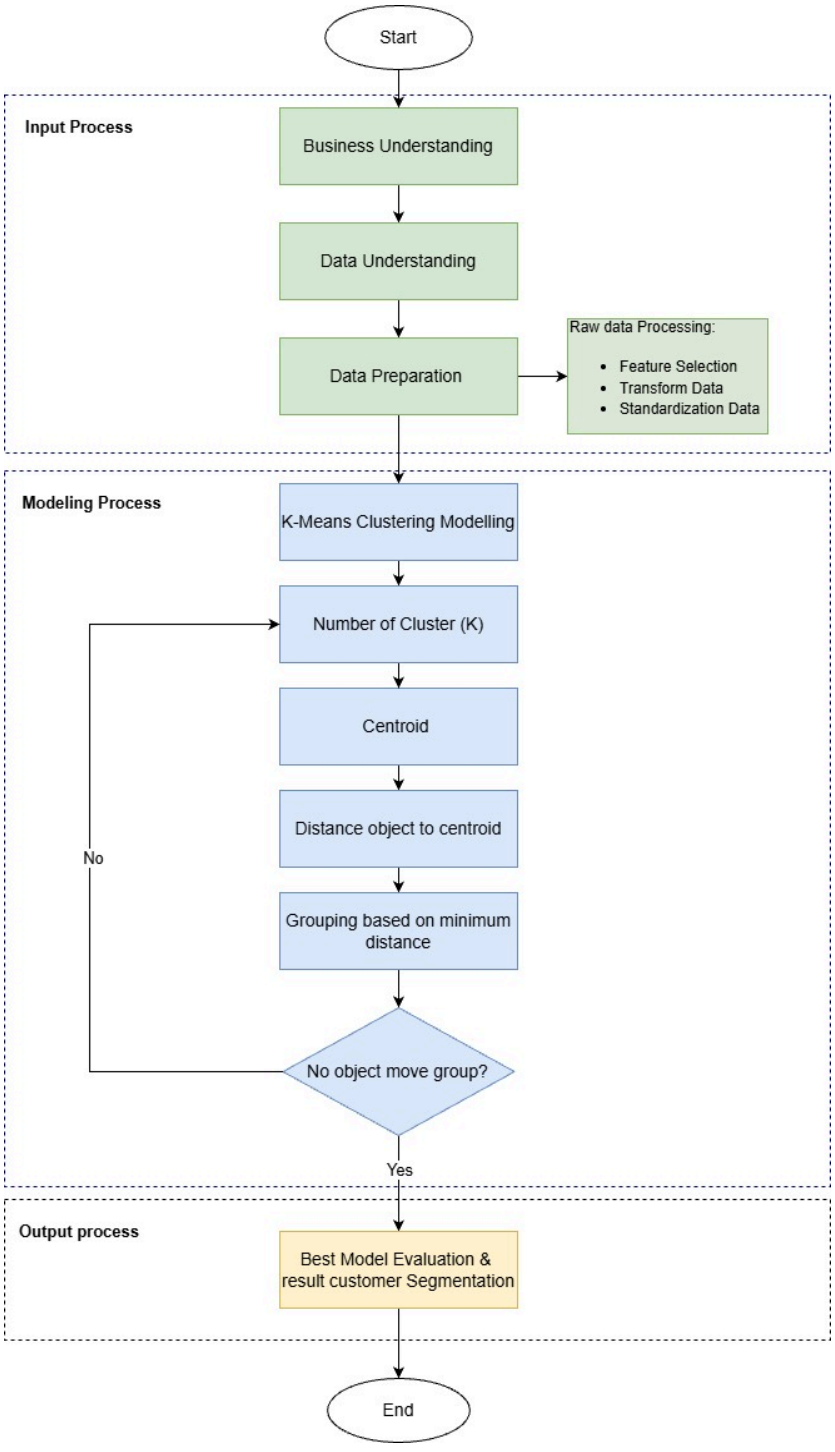

Figure 1. The sequence of stages in the CRISP-DM framework

Based on Figure 1, the CRISP-DM workflow implemented in this study utilizes the K-Means Clustering algorithm and consists of several phases, including business understanding, data understanding, and data preparation; in this stage, the raw dataset, which initially consisted of several attributes, undergoes a feature selection process to obtain the 3 most relevant attributes for customer clustering analysis. Furthermore, an encoding process is applied to transform categorical data into numerical data so that it can be processed by the K-Means algorithm. Afterward, data standardization is performed to ensure that each variable has a balanced scale, preventing the distance calculation process from being dominated by certain features. The next stage involves modeling using the K-Means Clustering algorithm with the implementation of the Elbow Method to determine the optimal number of clusters, followed by centroid calculation and cluster formation, the model is then evaluated using the silhouette coefficient to assess the quality of the resulting clusters, including measuring the separation between groups and determining the most optimal cluster configuration.

2.1. Data Preparation

The dataset for this research was obtained from PT Lintas Jaringan Nusantara, comprising 972 records. The initial raw dataset contains several attributes describing customer service usage and account information, with all personal identifiers such as exact customer names anonymized or removed to ensure data privacy, and details of the customer data attributes are presented in Table 1.

Table 1. Customer Data Attribute

Attribute	Description
CustomerId	Unique ID for each customer
Age	Age of registered customers
Internet Packages	Internet service package used by the customer
Price	Cost of the internet service package used by the customer

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

Attribute	Description
NAS Location	Location of the NAS device serving the customer
Customer Registration	Date and time when the customer first registered or when the services was activated
Due Date	Date and time of the payment deadline or the end of a specific period
SnOnu	Serial number unique alphanumeric code identifying the customer's hardware
isMigration	Customer service migration status
isIsolate	Customer service isolation status
isUnlimited	Customer unlimited service package status
Referral	Customer code obtained through a referral program
Remarks	Customer information details regarding activation status

The initial data distribution presents several customer attributes with varying data types and levels of completeness, presented in Table 2.

Table 2. Distribution Data

Attribute	Data Type	Count	Missing Values
CustomerId	Object	972	0
Age	Float64	971	1
Internet Packages	Object	971	1
Price	Float64	949	23
NAS Location	Object	971	1
Customer Registration	Object	971	1
Due Date	Object	962	10
SnOnu	Object	132	840
isMigration	Float64	0	972

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

Attribute	Data Type	Count	Missing Values
isIsolate	Float64	118	854
isUnlimited	Float64	19	953
Referral	Float64	0	972
Remarks	Object	854	118

Most attributes contain sufficient records for analysis; however, several attributes exhibit missing values that require preprocessing before the clustering process, thereby necessitating a feature selection stage.

2.1.1. Feature Selection

Feature selection, also known as variable selection, represents a widely adopted machine learning technique particularly for managing high-dimensional data [16]. Feature selection in this study is based on business relevance for customer segmentation, where from the original several attributes, only 3 attributes are retained due to their direct contribution to actionable internet service provider decision-making in service optimization, pricing strategy, and network planning. Several attributes were excluded because they do not directly represent key dimensions of customer service usage and are not-consistently recorded in the dataset and low reliability across observation resulting in limited additional contribution to distinguishing in the segmentation process. The selected features internet package type, price, and NAS location are used because they respectively represent service usage, economic value, and network access patterns, which are essential for interpreting customer behavior and supporting business decision-making in an internet service provider context.

After feature selection, these 3 attributes undergo a data quality assessment stage, which includes checking for missing values and identifying potential outliers that may affect clustering performance. Missing values handled using appropriate imputation techniques, where the 'Price' variable is imputed using the median value due to its robustness against extreme values, while the 'Internet Package' and 'NAS Location' attributes are imputed using the mode, as both represent categorical attributes. In addition, outlier detection and treatment are performed on the 'Price' variable using the Interquartile Range (IQR)

method, and identified outlier values were removed to prevent distortion in the distance-based calculations of the K-Means algorithm.

2.1.2. Transform Data

Data transformation involves converting data from its original form into a format suitable for analysis requirements [17]. This conversion-based approach involves changing categorical data into numerical data, which can then be used as input for clustering using an available clustering algorithm [18]. In this study, data transformation was conducted using a manual mapping and frequency encoding to convert categorical attributes into numerical representations suitable for clustering analysis. Manual mapping is used for categorical attribute that possess a natural order or meaningful structure, so that their numerical encoding can more accurately represent the relationships among the categories. Frequency encoding, on the other hand, is used for categorical attributes that do not have any natural order. Instead of assigning arbitrary numeric labels, each category is replaced by the proportion or frequency of its occurrence in the dataset. This approach reduces the risk of incorrect interpretation caused by random numeric labeling and helps maintain the meaningful structure of the data for clustering analysis and useful in providing richer variations in distance and assigning appropriate importance to categorical features by reflecting larger separations between categories [17], [19]. However, manual mapping depends on predefined numeric assignments, which can introduce subjectivity and reduce adaptability when used across different datasets and frequency encoding does not reflect the true relationship between categories, as values are based solely on occurrence frequency, which can lead to bias toward dominant categories and may overly influence clustering results. Several techniques were applied:

1) Manual mapping was used for internet package attribute to represent the levels or ordered structure of service capacity based on bandwidth speed. The result of this process are shown in Table 3.

Table 3. Manual Mapping

Internet Package	Converted Data
6 Mbps	0
10 Mbps	1
15 Mbps	2

Internet Package	Converted Data
20 Mbps	3
30 Mbps	4
50 Mbps	5
75 Mbps	6

2) Frequency encoding was used for NAS location attribute to represent the frequency or occurrence rate of each NAS location within the dataset, reflecting the distribution density of customers across different locations. The result of this process are shown in Table 4.

Table 4. Frequency Encoding

NAS Location	Frequency
Mikrotik Mejobo Kudus	391
Mikrotik Panjang Kudus	119
Mikrotik Undaan Kudus	116
Mikrotik Subang Sukamandi	92
Mikrotik Agus Salim	72
Mikrotik Pali Sumsel	58
Mikrotik Pinrang Sulawesi	31
Mikrotik Rokan Hilir	26
Mikrotik Bypass Karawang	18
Mikrotik Kosambi Karawang	14
Mikrotik Cikarang	13

2.1.2. Standardization Data

Standardization using StandarScaler is applied to the service-oriented attributes; 'Internet Package', 'Price', 'NAS Location' to ensure that all features are brought to a comparable scale. This process is essential to prevent attributes with larger numerical ranges, such as price, from disproportionately influencing the distance calculations between data points. By normalizing the scale of all features, each variable contributes equally to the clustering process, thereby improving the

reliability and stability of the K-Means results. Technically, the standardization process is applied to a separate dataset (X_final_scaled), which is specifically used as input for the K-Means algorithm to ensure that distance calculations between features are objective. After K-Means assigns cluster labels to each data point, these labels are mapped back and merged into the original dataset (df_final), which still retains the original numerical values. Consequently, the aggregation step in the profiling stage automatically computes mean values based on the actual price data, resulting in valid cluster interpretations expressed in real monetary units without requiring an inverse transform step and also cluster labels are generated automatically by the K-Means algorithm based on distance calculations on standardized data. However, the interpretation and characterization of each cluster label are carried out manually through the profiling stage.

2.2. Modeling

The K-Means algorithm is widely recognized and commonly utilized owing to its straightforward implementation, effectiveness, and simple underlying concept, requiring the predetermined specification of the number of clusters as a key parameter along with the manual initialization of initial cluster centroids [20]. In the K-Means method the number of clusters K must be predetermined and once obtained through a hierarchical approach [21]. Initial cluster selection can performed using the Elbow method. The Elbow method represents a graphical heuristic commonly employed to identify the optimal number of cluster (k) in algorithms like K-Means, relying on the examination of Within-Cluster Sum of Squares (WCSS) also known as Sum of Squared Errors (SSE) which quantifies the variance of data points relative to their respective cluster centroids as illustrated in the Equation 1 [22].

$$WCSS = \sum_{i=1}^{k} \sum_{x_j \in C_i} \left\| x_j - \mu_i \right\|_2^2 \quad (1)$$

Description:

k : Number of clusters

i : Cluster index

C_i : Set of data points in cluster i

x_j : The j-th data point in cluster C_i

μ_i : Centroid (mean vector) of cluster C_i

$\|\cdot\|_2^2$: Squared Euclidean norm

Lacking any strict rules for determining K though often aligned with subjective user needs, various strategies exist for initial cluster selection such as choosing based on observation count intervals applying hierarchical methods or randomly selecting from the observation set thereby enabling optimal solutions, and at this stage distance measurement proves crucial for assigning observations to cluster by proximity to the nearest centroid with Euclidean distance serving as the primary metric in K-Means [21]. The use of Euclidean distance is deemed appropriate since all categorical attributes have been previously transformed into numerical form prior to clustering as illustrated in the Equation 2 [23].

$$d = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \quad (2)$$

The variable n represents the number of dimensions in Euclidean space, while x_i and y_i refer to points located within that space and d represents the straight-line distance between the two points.

2.3. Evaluation

To evaluate clustering solution quality the silhouette coefficient is employed assuming that an effective clustering solution features compact and well-separated groups [24]. The silhouette method serves as a common unique for assessing cluster quality and ascertaining the optimal number of groups (k) in clustering algorithms, with the coefficient computed for each data point [22]. The silhouette score computation is performed by evaluating the individual silhouette value $s(x_i)$ for each data point x_i. First, the average distance $a(x_i)$ between x_i and all other data points within the same cluster C_I is calculated as illustrated in the Equation 3 [24].

$$a(x_i) = \frac{1}{|C_I| - 1} \sum_{\substack{x_j \in C_i \\ j \neq i}} d(x_i, x_j) \tag{3}$$

Where $|C_I|$ denotes the number of data points in cluster C_I, with $|C_I| > 1$. The value of $a(x_i)$ measures the degree to which the data point x_i is appropriately associated with its cluster. A smaller $a(x_i)$ value indicates that x_i is highly similar to other members within the same cluster, implying that the data point is likely assigned correctly. In contrast, a larger $a(x_i)$ value suggests that x_i is relatively distant from other points in its cluster. The silhouette score calculation also involves determining the minimum average distance between a data point $x_i \in C_i$ and points belonging to other clusters, denoted as $b(x_i)$, which is defined as follows in the Equation 4 [24].

$$b(x_i) = \min_{J \neq i} \left(\frac{1}{|C_J|} \sum_{x_j \in C_J} d(x_i, x_j) \right) \tag{4}$$

A higher $b(x_i)$ value indicates that the data point x_i is substantially different from data points in other clusters, which reflects a desirable clustering result. The Silhouette score for a data point x_i is formulated by considering the need for a low $a(x_i)$ value and a high $b(x_i)$ value, and is defined as follows in the Equation 5 [24].

$$s(x_i) = \frac{b(x_i) - a(x_i)}{max(a(x_i), b(x_i))} \tag{5}$$

It should be noted that the silhouette score satisfies $-1 \leq s(x_i) \leq 1$. A value approaching 1 occurs when $a(x_i)$ is low and $b(x_i)$ is high, indicating that x_i belongs to a compact and well-separated cluster. Conversely, a value near −1 implies that x_i is more similar to data points in other clusters than to those within its own cluster, suggesting that it may have been incorrectly assigned to a cluster. The overall silhouette score for the entire clustering partition C of the dataset X is computed by aggregating all individual silhouette values using a standard averaging approach as follow in Equation 6[24].

$$S(X) = \frac{1}{N} \sum_{i=1}^{N} s(x_i) \tag{6}$$

The silhouette score is not only suitable for internal clustering evaluation but also provides an intuitive objective for clustering, favoring groups that are compact and well-separated from one another.

3. RESULTS AND DISCUSSION

3.1. Customer Segmentation

The optimal number of clusters is determined using the Elbow method, which involves calculating the WCSS (Within-Cluster Sum of Squares), also known as SSE (Sum of Squared Errors), for various numbers of clusters. The WCSS values are visualized in a graph to observe the pattern of their decline. The optimal point is identified at the graph's "elbow" where the rate of decrease begins to flatten, forming a sharp bend resembling an elbow—indicating that adding more clusters no longer yields a significant reduction in WCSS as illustrated in Figure 5.

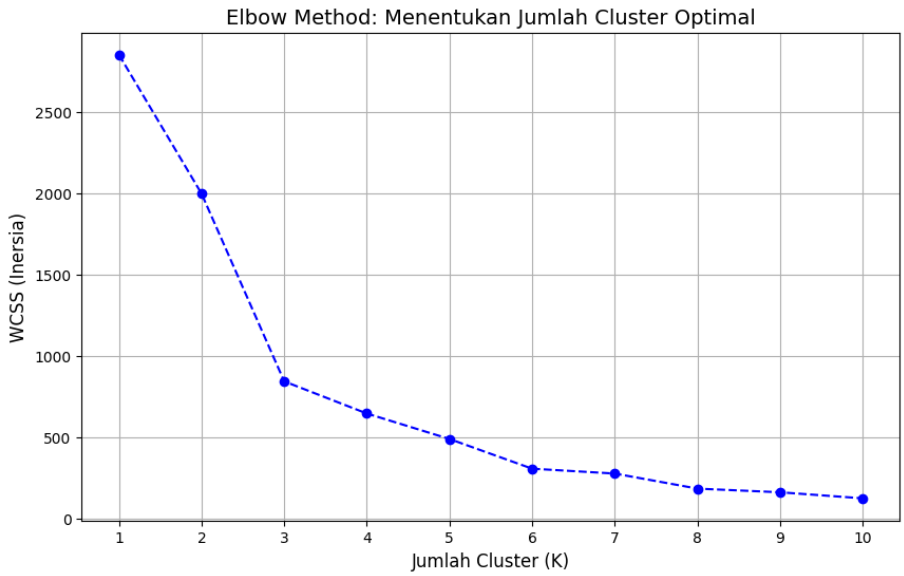

Figure 2. Elbow Method Graph

Based on the Elbow method graph, the Within Cluster Sum of Squares (WCSS) value decreases significantly from cluster 1 to cluster 3. This decline indicates that the addition of clusters within this range effectively improves the homogeneity of data within each group. After the point cluster 3, the rate of decrease in WCSS begins to level off and no longer shows substantial changes compared to the previous clusters. This condition

suggests that adding more than 3 clusters only provides a minor improvement in clustering quality. The WCSS values for clusters 1 through 10 are presented in Table 5.

Table 5. Value WCSS Cluster 1- Cluster 10

Cluster	WCSS value
1	2849.999999999984
2	1996.5148037190813
3	844.1868048688435
4	646.8595891740027
5	490.347097727183
6	308.0524458874026
7	278.1072289282388
8	185.6625820631509
9	163.13947120731447
10	126.02399847830279

Centroid initialization in the K-Means Algorithm modeling with 3 clusters was performed to ensure optimal data clustering process. Accurate determination of central point plays a crucial role in minimizing intra-cluster variance (intra-cluster distance), thereby enabling more representative formation of each cluster's structure and characteristics. Beyond serving as starting points in the iterative clustering process, well-selected centroids facilitate easier interpretation of the resulting segment characteristics, presented in Table 6.

Table 6. Centroid Cluster 3

Cluster	Centroid		
	Price	NAS Location	Internet Package
0	-0.71192302	-0.45660617	-0.72995395
1	-0.06249914	-0.54472571	1.17665612
2	1.10276112	1.47784965	-0.76388032

Centroid points represent the central position of each cluster in the standardized feature space. The centroid values reflect the average characteristics of cluster members based

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

on Price, NAS Location, and Internet Package. In this standardized space, positive values indicate above-average characteristics, while negative values indicate below-average characteristics relative to the dataset mean. The K-Means algorithm assigns data points to clusters based on their proximity to these centroids, making centroid positions the key determinant of cluster membership. Consequently, centroids serve as the primary representation of each cluster and provide a basis for interpreting customer segments and analyzing behavioral patterns across groups Across clusters, differences can be observed by comparing the direction and magnitude of centroid deviations across features. For the Price attribute, cluster 2 exhibits the highest positive deviation from the dataset mean, followed by cluster 1, while cluster 0 shows a negative deviation, indicating a relative ordering of clusters in the Price dimension. A similar comparative pattern is observed in NAS Location attribute, where cluster 2 demonstrates the strongest positive deviation compared to other clusters. In contrast, the Internet Package dimension shows a different pattern, where Cluster 1 has the strongest positive deviation, while cluster 2 and cluster 0 are positioned below the dataset mean.

The visualization results of the 3 clusters can be seen in Figure 6, while a more detailed explanation of the characteristics and interpretation of each cluster is presented in the following discussion.

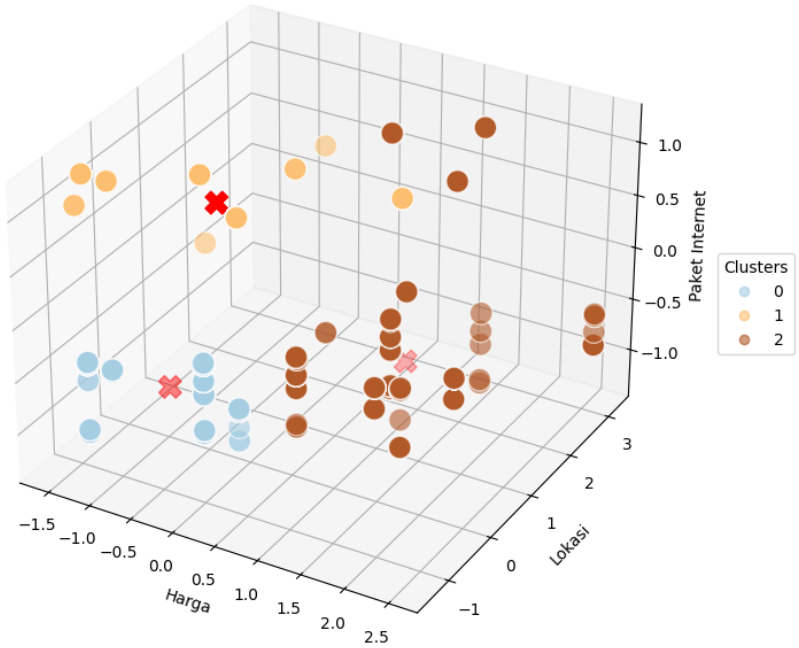

Figure 3. Visualization of Cluster 3

.....

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

The visualization results of the segmentation clustering yield 3 clusters with diverse characteristics, each described as follows are summarized in Table 7.

Table 7. Summary of Cluster 3 Characteristics

Cluster	Characteristics
0	Consist of 341 customers subscribed to a 10 Mbps internet package, with an average price of Rp 141,849 and geographically dominated by the Undaan Kudus region.
1	comprises 368 customers with a 10 Mbps internet package subscription, averaging Rp 165,552 in price, predominantly from the Mejobo Kudus region.
2	Includes 241 customers subscribed to a 50 Mbps package, with an average price of Rp 208,084, distributed across the Pali Sumsel region.

Based on the results of the customer clustering distribution, Although Cluster 0 and Cluster 1 are both dominated by customers subscribing to the same 10 Mbps internet package, differences in average pricing and NAS location distribution indicate variations in customer characteristics across the two segments. The disparity in subscription prices is influenced by differences in customer categories and subscription periods. Long-term customers generally remain associated with earlier pricing schemes, whereas newer customers are subject to the company's updated pricing policies. Consequently, customers subscribing to the same internet capacity may incur different subscription costs due to pricing adjustments implemented over time. In addition, variations in NAS location distribution suggest that geographical service allocation and regional operational factors may also contribute to differences in customer grouping between the two clusters. In the cluster 2, this segment represents customers with higher internet bandwidth demands and greater expenditure levels. To determine the quality of the formed clusters, testing was conducted using the silhouette coefficient. This score

reflects the quality of inter-cluster separation based on internal cohesion and external separation metrics, thereby revealing the clarity of the data structure as illustrated in Figure 7.

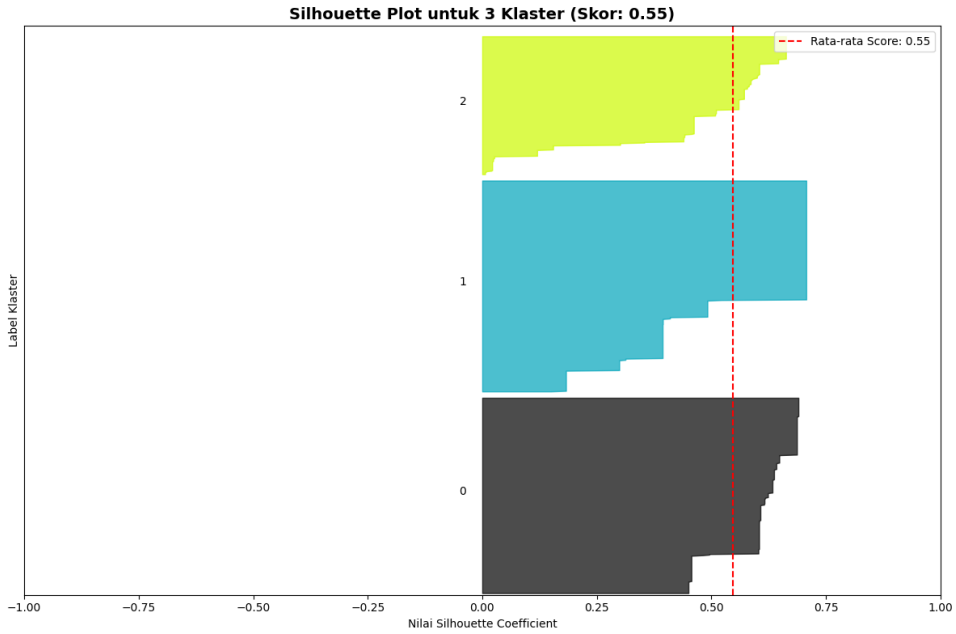

Figure 4. Silhouette Score of 3 Clusters

Cluster validity test results indicate that the red line represents the average silhouette score value. The silhouette coefficient analysis demonstrates that the clustering model with 3 clusters achieved an average silhouette score of 0.5471. Based on the silhouette plot, most data points across all clusters exhibit positive silhouette values, indicating that the majority of customers were assigned to appropriate clusters with relatively small intra-cluster distances and sufficiently large inter-cluster distances. Cluster 1 appears to have the highest concentration of silhouette values in the upper range, suggesting that this cluster has the strongest internal consistency and the clearest separation from other clusters. Meanwhile, Clusters 0 and 2 also show predominantly positive silhouette values, although several observations are located closer to the average silhouette boundary, indicating the presence of customers with characteristics that are relatively similar to those of neighboring clusters. In the evaluation stage, the silhouette score serves as a key metric for assessing the degree of separation between clusters. Therefore, the silhouette method was employed to determine the optimal number of clusters, ensuring that the final clustering configuration achieves the best possible balance between

cohesion within clusters and separation between clusters. Subsequently, the silhouette coefficient for determining the optimal K can be observed in Figure 8.

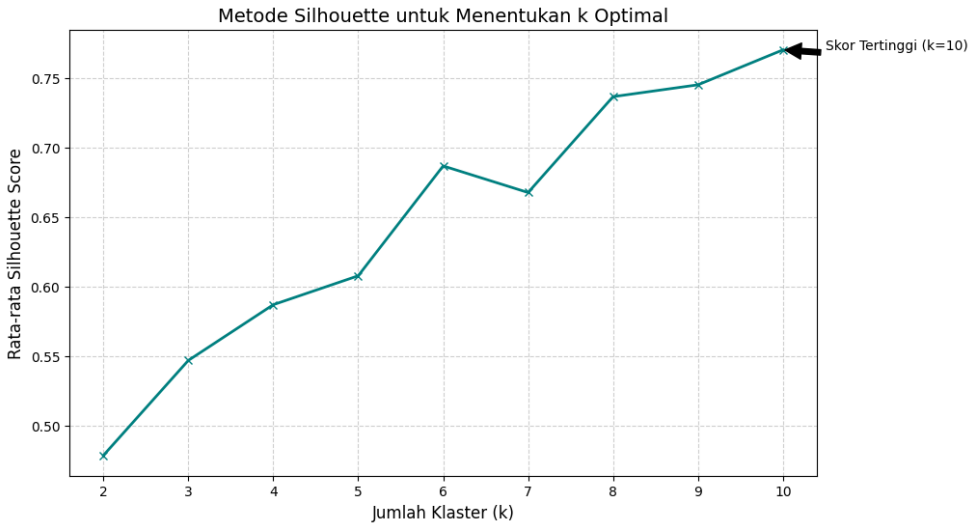

Figure 5. Silhouette Graph for Identifying the Optimal Number of Clusters

The determination of the most optimal number of clusters in the segmentation clustering process can be observed in the figure. The graph visualization on the x-axis displays the tested cluster variations from Cluster 2 to Cluster 10, while the y-axis shows the average silhouette score values generated for each cluster count. Based on the visualization results, the highest average value is achieved at Cluster 10, interpretable as the most optimal cluster number. Silhouette coefficient values approaching 1 indicate high proximity of each observation to its assigned cluster and sufficient distance from other clusters, resulting in low inter-cluster overlap and accurate clustering quality. The centroid points for 10 clusters can be presented in Table 8.

Table 8. Centroid Cluster 10

Cluster	Centroid		
	Price	**NAS Location**	**Internet Package**
0	-1.46032164	-0.73380139	-0.66796217
1	0.17308205	-0.64119531	1.17665612
2	0.56183141	0.9353648	-0.95459068
3	0.26757926	-0.73380139	-0.61920685
4	-1.31561666	-0.75568184	1.17665612

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

Cluster	Centroid		
	Price	NAS Location	Internet Package
5	1.41723576	1.01663507	1.17665612
6	-0.51602463	-0.07738772	-0.83812406
7	1.11296995	1.99336056	-0.96644944
8	1.48404068	0.57902595	-1.04936342
9	2.5251289	3.20468064	-0.78334247

The centroid profiles across the ten clusters reflect the relative positioning of each segment within the standardized feature space, where values represent deviations from the dataset mean. Cluster 0 is consistently below average across all attributes Price −1.40, NAS Location −0.73, Internet Package −0.66, indicating a low-value segment characterized by uniformly negative deviations. Cluster 2 exhibits a mixed structure, with above-average Price 0.56 and NAS Location 0.93 but below-average Internet Package −0.95, suggesting a segment driven by higher pricing and location intensity despite lower package levels. Cluster 3 remains close to the global mean for Price 0.26, yet retains negative deviations in NAS Location −0.73 and Internet Package −0.61, indicating limited differentiation with mild underrepresentation in service-related attributes. Cluster 4 demonstrates a pronounced contrast between strongly negative Price −1.31 and strongly positive Internet Package 1.17, reflecting a structurally heterogeneous configuration between cost and service capacity. Cluster 5 is positioned above the dataset mean across all dimensions, with particularly high values in Price 1.42 and Internet Package 1.17, representing a high-value premium segment. Cluster 6 shows near-neutral positioning in NAS Location −0.07 with moderate negative deviations in Price −0.51 and Internet Package −0.83, indicating mild underperformance relative to the dataset average. Cluster 7 is primarily distinguished by an elevated NAS Location value 1.99, coupled with a positive but comparatively lower Internet Package 0.96, indicating strong spatial concentration effects. Cluster 8 combines high Price 1.48 and moderate NAS Location 0.57 with a pronounced negative Internet Package −1.04, reflecting a cost-intensive but service-limited configuration. Cluster 9 represents the most extreme deviation in the feature space, particularly in NAS Location 3.20 and Price 2.52, while maintaining a below-average Internet Package −0.78, indicating a highly distinctive and structurally isolated segment relative to the overall dataset. Centroid represents the geometric mean of observations

within each cluster and serves as a mathematical summary of cluster location in the feature space. The clustering outcome itself refers to the assignment of individual observations into groups, while centroid is a derived statistical representation of each group, not the grouping result itself and The visualization results of the 10 clusters can be seen in Figure 9, while a more detailed explanation of the characteristics and interpretation of each cluster is presented in the following discussion.

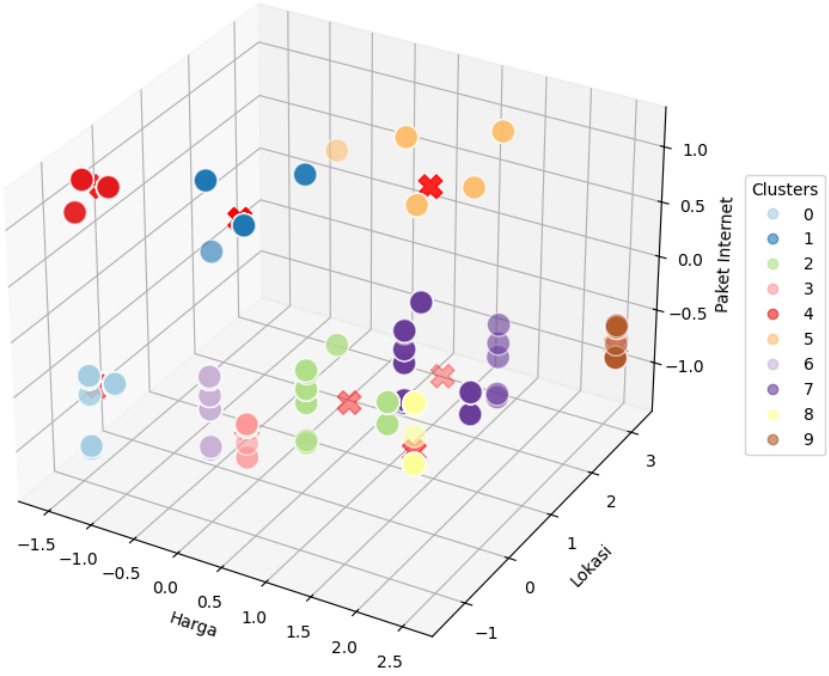

Figure 6. Visualization of Cluster 10

The segmentation clustering visualization produces 10 clusters exhibiting varied characteristics, with each cluster's description summarized in Table 9.

Table 9. Summary of 10 Clusters Characteristics

Cluster	Characteristics
0	Comprises 128 customers subscribed to a 10 Mbps internet package, with an average price of Rp 114,533 geographically distributed in the Panjang Kudus region
1	Consists of 241 customers with a 10 Mbps internet package subscription, averaging Rp 174,151, primarily located in the Mejobo Kudus region.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

Cluster	Characteristics
2	Includes 70 customers subscribed to a 30 Mbps package, with an average price of Rp 188,340, predominantly distributed in the Pali Sumsel region.
3	Contains 69 customers with a 10 Mbps internet package, averaging Rp 177,600, spread across the Undaan Kudus region.
4	Comprises 90 customers subscribed to a 10 Mbps package, with an average price of Rp 119,814, distributed in the Mejobo Kudus region.
5	Consists of 60 customers with a 20 Mbps internet package, averaging Rp 219,562, geographically spread in the Mejobo Kudus region.
6	Includes 144 customers subscribed to a 15 Mbps package, with an average price of Rp 148,999, distributed in the Subang Sukamandi region.
7	Contains 97 customers with a 50 Mbps internet package subscription, averaging Rp 208,456, concentrated in the Pali Sumsel region.
8	Comprises 40 customers subscribed to a 20 Mbps package, with an average price of Rp 222,000, distributed in the Pinrang Sulawesi region.
9	Consists of 11 customers with a 100 Mbps internet package, averaging Rp 259,999, primarily spread in the Agus Salim region.

Clusters 0 and 4 are dominated by 10 Mbps internet packages with relatively lower average prices. However, even though the service package is the same, the prices across clusters differ. Clusters 1 and 3 also consist of 10 Mbps users, but Cluster 3 shows a higher average price. This difference is explained by variations in subscription periods, where some customers are long-term users who still retain older pricing schemes, while newer customers follow updated pricing policies. In addition, some customers are still under promotional pricing periods, which results in lower average package costs even though the internet speed remains the same. Overall, price differences within the same package occur due to changes in company pricing policies over time and the coexistence of multiple pricing schemes within the customer base. clusters 2, cluster 5, cluster 7, cluster 8, and cluster 9 represent customers with higher service levels, ranging from 20 Mbps up to 100 Mbps. In general, higher internet speeds are associated with higher prices. However, there is an interesting variation: for example, within the 20 Mbps package, some regions show higher average prices than others, indicating that location still plays a significant role in pricing differences. Cluster 6 can be considered a mid-level group, as

it consists of customers using the 15 Mbps package. This cluster sits between low-tier users 10 Mbps and mid-to-high-tier users 20 Mbps and above, representing a transitional segment in the overall customer structure. Cluster 9 represents the highest-tier segment, consisting of 100 Mbps users with the highest average price, even though the number of customers is very small. This indicates a niche segment of customers with high bandwidth requirements, likely driven by specific or intensive usage needs. The results of the 10 clusters were subsequently re-evaluated using the silhouette score, as shown in Figure 10.

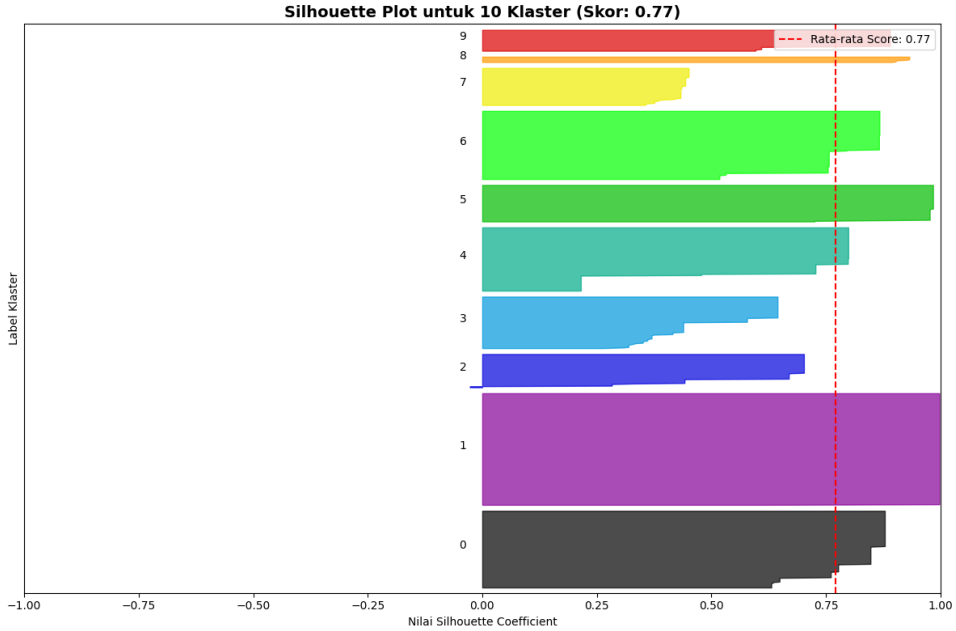

Figure 7. Silhouette Score of 10 Clusters

Based on the validity test results using the Silhouette Coefficient for 10 clusters, the average silhouette score obtained is 0.7704. This indicates that the clustering structure is well-separated and the groups are meaningfully formed. The quality of the result is further supported by most clusters, especially cluster 0, cluster 1, cluster 5, cluster 6, cluster 8, cluster 9 which achieve high silhouette score close to 1.0 and stand clearly above the average line. The data points within these clusters are highly similar to each other, while being distinctly different from points in other clusters. Although there are differences in block thickness that reflect variations in cluster size, such as cluster 1 and cluster 5 being dominant and cluster 2, cluster 3 and cluster 7 representing a smaller group overall distribution remains stable.

The model demonstrates more optimal performance in maintaining a balance between internal cohesion among members within clusters and external separation between different clusters. The comparison of the number of clusters from 3 cluster to 10 cluster has a significant impact on the distribution of the percentage of members in each group, as shown in Table 10 and Table 11.

Table 10. The result percentage in 3 Clusters

Cluster	Percentage
0	35.89%
1	38.74%
2	25.37%

Table 11. The result percentage in 10 Clusters

Cluster	Percentage
0	13.47%
1	25.37%
2	7.37%
3	7.26%
4	9.47%
5	6.32%
6	15.16%
7	10.21%
8	4.21%
9	1.16%

In the 3 Cluster model, Cluster 1 exhibits a strong dominance comprising 38.74% of the total data, indicating that nearly half of the population shares similar characteristics within one large group. Following expansion to 10 Cluster, Cluster 1 remains the largest group but its proportion decreases to 25.37%, demonstrating that increasing the number of clusters successfully splits the large group into more specific subgroups. The comparison of candidate clustering solutions based on WCSS, silhouette score and interpretability is presented in the following Table 12.

Table 12. Comparative Analysis of Clustering Result

Cluster	WCSS Value	Silhouette score	Interpretability Review
3	844.1868048688435	0.5471	Segmentation remains relatively broad, with several customer characteristics still grouped together.
10	126.02399847830279	0.7704	Clusters are more distinct and detailed, although some segments show similar service characteristics.

The table presents a comparison of clustering results using 3 and 10 clusters based on WCSS, Silhouette Score, and interpretability review. The 3 clusters configuration shows a WCSS value of 844.19 with a Silhouette Score of 0.5471, indicating broader customer grouping characteristics. Meanwhile, the 10-cluster configuration produces a lower WCSS value of 126.02 and a Silhouette Score of 0.7704, with customer segments appearing more specific and detailed, although several segments still exhibit similar service characteristics.

3.2. Discussion

The clustering results indicate that the 10 clusters configuration yields a higher silhouette score 0.7704 compared to the 3 clusters solution 0.5471, suggesting improved separation and greater homogeneity within clusters. From an operational perspective, the 10 clusters solution provides a more detailed segmentation structure that allows customer groups to be differentiated based on finer variations in internet package, pricing, and NAS location distribution. Compared to the 3 clusters configuration, which offers a more aggregated view of customer behavior, the 10 clusters may better capture heterogeneity that is relevant for descriptive analysis. However, this increased granularity also introduces higher managerial complexity, as a larger number of clusters requires more effort in interpretation and categorization of customer profiles. There is also a potential risk of over-segmentation, where some clusters may exhibit relatively similar characteristics, thereby limiting the distinctiveness required for clear business decision boundaries. This trade-off between granularity and interpretability should therefore be considered in selecting the most appropriate clustering configuration. To support the interpretability of the clustering results, the identified segments were reviewed in

consultation with domain understanding of marketing and customer service operations teams, particularly with respect to typical usage patterns, NAS location distribution and pricing structures observed in the dataset. This validation remains preliminary in nature, and more structured stakeholder evaluation is suggested as a direction for future work to strengthen the operational applicability of the clustering results. It is also important to acknowledge that the use of frequency encoding and manual mapping for categorical attributes may introduce methodological limitations. Frequency encoding can unintentionally influence distance calculations in K-Means by embedding distribution-based numeric bias into categorical features, while manual mapping may introduce subjectivity in the transformation process. These factors may affect the stability and interpretability of the resulting clusters, particularly for the NAS Location variable, which is a nominal categorical variable without inherent order or magnitude Overall, both clustering configurations provide meaningful exploratory insights. The 3 clusters solution offers a more general overview of customer segmentation, while the 10 clusters solution provides a more detailed but potentially more complex representation of customer heterogeneity. As such, the results should be interpreted as exploratory findings that may require further validation before being used for operational decision-making.

By contrast, Vieri et al. [10] identified an optimal cluster number of 3, achieving a relatively low silhouette score of 0.2514. This difference should not be interpreted solely as inferior clustering performance in that study, but rather as a reflection of differences in data structure, feature composition, and clustering complexity. In Vieri et al. [10], the dataset incorporates five behavioural indicators, including customer status, payment history, total charges, subscription duration, and service type. These attributed are inherently dynamic, heterogeneous, and potentially intercorrelated, which increases dimensional complexity and may obscure natural cluster separations when using distance-based methods such as K-Means. Additionally, behavioural datasets often exhibit higher variance and temporal dependency, making clear partitioning more difficult and leading to lower silhouette scores despite potentially richer behavioural representation. In contrast, the present study uses three service-oriented attributes—Internet Package, Price, and NAS Location—which are more structured and operational in nature. These attributes are less volatile and more directly tied to service configuration rather than evolving customer behaviour. As a result, the feature space is more constrained and tends to produce more compact and well-separated clusters, which contributes to a higher silhouette score.

Similarly, Kumar et al. [25] reported an optimal solution with 3 clusters and a silhouette score of 0.2926 based on a combination of behavioural and demographic attributes. While the higher silhouette score in the present study suggests improved cluster separability, this difference must be interpreted cautiously. Demographic and behavioural attributes typically introduce greater heterogeneity and nonlinear relationships compared to service-level attributes, which can reduce separability under Euclidean distance-based clustering. Moreover, differences in preprocessing strategies—such as encoding schemes, scaling approaches, and feature selection criteria—further limit direct comparability between studies. Therefore, the variation in silhouette scores across studies is influenced not only by the number of clusters but also by fundamental differences in (i) variable types (behavioural vs. service-based), (ii) industry operational focus, (iii) preprocessing pipelines, and (iv) inherent data complexity. Consequently, while the current study demonstrates stronger numerical cluster separation, this should be interpreted as improved structural clarity within a narrower feature space, rather than as absolute superiority in capturing comprehensive customer behaviour patterns. Therefore, the difference in silhouette scores between the two studies is influenced not only by the number of clusters but also by the complexity of attributes, dataset characteristics, and the preprocessing techniques applied. Thus, the higher silhouette score observed in this study primarily reflects better cluster separation, but does not necessarily indicate superiority in capturing customer behavior complexity, as the use of simpler attributes can produce more easily separable clusters that are less informative.

4. CONCLUSION

The application of the K-Means clustering algorithm for customer segmentation demonstrates its effectiveness in organizing customer profiles into distinct and interpretable groups using service-oriented attributes, Internet package, price, and NAS Location. The analytical process was conducted within the CRISP-DM framework, including data preprocessing steps such as selection feature, transform data using manual mapping and frequency encoding, feature standardization, and the treatment of missing values and outliers to ensure data quality. Model selection was guided by the Elbow method, which initially suggested a 3 clusters configuration, while further evaluation using the silhouette coefficient indicated that a 10 clusters solution achieved superior internal validity with a higher silhouette score 0.7704 compared to the 3 clusters

model 0.5471. this result suggest that the 10 clusters configuration provides stronger intra-cluster cohesion and clearer inter-cluster separation at a finer level of segmentation granularity. However, despite its superior internal validation performance, the 10 clusters solution has not yet been validated from business or managerial perspective, and its practical applicability in supporting marketing strategies, service optimization, or customer management decisions remains subject to further evaluation. It is also important to note that the study is limited by the relatively small number of features used, which may restrict the richness of customer representation in the segmentation process. Future research is recommended to incorporate stakeholder-based validation to assess the managerial and operational relevance of the identified segments, as well as to expand the feature set with more comprehensive customer attributes to enhance segmentation robustness and business interpretability.

ACKNOWLEDGMENT

The authors would like to express their sincere gratitude to PT Lintas Jaringan Nusantara for their valuable support in facilitating the data collection process required for this research. The authors also extend their appreciation to Telkom University for the facilitation and support provided throughout the course of this study.

REFERENCES

[1] G. E. Corazza et al., "A glimpse on the futures of telecommunication networks: From market, technology and regulation trends to strategic foresight," *J. Open Innov. Technol. Mark. Complex.,* vol. 12, no. 1, Mar. 2026, doi: 10.1016/j.joitmc.2026.100735.

[2] H. Ribeiro, B. Barbosa, A. C. Moreira, and R. G. Rodrigues, "Determinants of churn in telecommunication services: a systematic literature review," *Manag. Rev. Q.,* vol. 74, no. 3, pp. 1327–1364, Sep. 2024, doi: 10.1007/s11301-023-00335-7.

[3] A. H. Abushar, "Factors Affecting Brand Switching Behaviour in the Palestinian Telecommunications Industry in the Gaza Strip," in *Stud. Syst. Decis. Control*, vol. 516, 2024, pp. 291–302. doi: 10.1007/978-3-031-49544-1_26.

[4] J. S. Harini, A. Anusuya, P. Kanimozhi, and T. Ananthkumar, "Churn Prediction and Factor Identification in Telecommunication Industry Using Deep Learning," in *Proc. Int. Conf. Emerg. Technol. Eng. Appl. (ICETEA)*, 2025. doi: 10.1109/ICETEA64585.2025.11099737.

[5] S. Sayuti, B. Berman, D. Sirya, and J. Heikal, "Clustering Customer's Internet Subscriptions in Apartment Using K-Means Clustering Algorithm," 2025, doi: 10.37481/jmeb.v5i3.1514

[6] E. Eslami, N. Razi, M. Lonbani, and J. Rezazadeh, "Unveiling IoT Customer Behaviour: Segmentation and Insights for Enhanced IoT-CRM Strategies: A Real Case Study," *Sensors*, vol. 24, no. 4, Feb. 2024, doi: 10.3390/s24041050.

[7] C. Rungruang et al., "RFM model customer segmentation based on hierarchical approach using FCA," *Expert Syst. Appl.*, vol. 237, Mar. 2024, doi: 10.1016/j.eswa.2023.121449.

[8] J. M. John, O. Shobayo, and B. Ogunleye, "An Exploration of Clustering Algorithms for Customer Segmentation in the UK Retail Market," *Analytics*, vol. 2, no. 4, pp. 809–823, Dec. 2023, doi: 10.3390/analytics2040042.

[9] K. Tabianan et al., "K-Means Clustering Approach for Intelligent Customer Segmentation Using Customer Purchase Behavior Data," *Sustainability,* vol. 14, no. 12, Jun. 2022, doi: 10.3390/su14127243.

[10] J. V. Kristian, T. A. Munandar, and D. B. Srisulistiowati, "Exclusive Clustering Technique for Customer Segmentation in National Telecommunications Companies," 2023, doi: 10.58776/ijitcsa.v1i1.19

[11] A. M. Ikotun et al., "K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data," *Inf. Sci.*, vol. 622, pp. 178–210, Apr. 2023, doi: 10.1016/j.ins.2022.11.139.

[12] R. H. Bemthuis et al., "A CRISP-DM-based methodology for assessing agent-based simulation models using process mining," *J. Simul.,* 2025, doi: 10.1080/17477778.2025.2508245.

[13] E. H. Sharaf Addin et al., "Customer Mobile Behavioral Segmentation and Analysis in Telecom Using Machine Learning," *Appl. Artif. Intell.,* vol. 36, no. 1, 2022, doi: 10.1080/08839514.2021.2009223.

[14] P. Ramesh and P. T. V. Bhuvaneswari, "Machine learning driven clustering for silhouetting 5G network throughput," *Sci. Rep.*, vol. 16, no. 1, Dec. 2026, doi: 10.1038/s41598-026-45902-6.

[15] M. Mustaqim et al., "Analisis Data Pelanggan dengan Algoritma K-Means untuk Peningkatan Penjualan Layanan ICONET DIBANGKA BELITUNG," *J. Akad. Ekon. Manaj.*, vol. 2, pp. 50–60, Dec. 2025, doi: 10.61722/jaem.v2i4.7106

[16] H. Huang et al., "Feature Selection for Unsupervised Machine Learning," in *Proc. IEEE 8th Int. Conf. Smart Cloud (SmartCloud)*, 2023, pp. 164–169. doi: 10.1109/SmartCloud58862.2023.00036.

[17] M. F. Fachri and L. Zahrotun, "Centroid Optimization of K-Means Using Ant Colony Optimization for Culinary MSME Clustering," *J. Inf. Syst. Inform.*, vol. 8, no. 1, pp. 860–888, Mar. 2026, doi: 10.63158/journalisi.v8i1.1443.

[18] L. Bai and J. Liang, "A categorical data clustering framework on graph representation," *Pattern Recognit.*, vol. 128, Aug. 2022, doi: 10.1016/j.patcog.2022.108694.

[19] D. Choi et al., "Deep Clustering for Mixed-type Data with Frequency Encoding and Doubly Weighted Cross Entropy Loss," in *Proc. ITC-CSCC*, 2022, pp. 141–144. doi: 10.1109/ITC-CSCC55581.2022.9894964.

[20] A. M. Ikotun et al., "Benchmarking validity indices for evolutionary K-means clustering performance," *Sci. Rep.*, vol. 15, no. 1, Dec. 2025, doi: 10.1038/s41598-025-08473-6.

[21] N. Huda Ahsina et al., "Analisis segmentasi pelanggan bank berdasarkan pengambilan kredit dengan menggunakan metode K-means clustering," 2022, doi: 10.33197/jitter.vol8.iss3.2022.883.

[22] A. Y. Raya-Tapia et al., "Machine Learning and Clustering for a Sustainable Future: Applications in Engineering and Environmental Science," in *Stud. Comput. Intell.*, vol. 1233, 2025, pp. 1–351. doi: 10.1007/978-3-032-03876-0.

[23] A. S. Nyamawe et al., "Practical Machine Learning; A Beginner's Guide with Ethical Insights," 2025. doi: 10.1201/9781003486817-1.

[24] G. Vardakas, I. Papakostas, and A. Likas, "Deep Clustering Using the Soft Silhouette Score: Towards Compact and Well-Separated Clusters," arXiv preprint, Feb. 2024, doi: 10.1007/s10994-026-07026-w.

[25] S. Kumar et al., "Customer segmentation in e-commerce: K-means vs hierarchical clustering," *Telkomnika*, vol. 23, no. 1, pp. 119–128, Feb. 2025, doi: 10.12928/TELKOMNIKA.v23i1.26384.