Journal of Information System and Technology Research

journal homepage: https://journal.aira.or.id/index.php/jistr/

X-Means Clustering for Segmenting Property Customer Payment Behaviors

Edein Fortuna[1], Dwi Arman Prasetya[2,*], Kartika Maulida Hindrayani[3]

[1,2,3]Department of Data Science, Pembangunan National Veteran University of East Java, Indonesia

ARTICLE INFO

Article history:

Received July 11, 2025
Accepted Dec 10, 2025
Available online Jan 31, 2026

Keywords:

Clustering Algorithm
Customer Segmentation
Data Mining
Payment Behavior
X-Means

ABSTRACT

Understanding customer behavior is essential for ensuring the sustainability and competitiveness of property businesses. This study aims to segment customers of PT X based on installment payment patterns using the X-Means clustering algorithm, which automatically determines the optimal number of clusters. From 9,615 transaction records, 386 customer profiles were analyzed using four features: number of transactions, number of late payments, payment differences, and payment status. The analysis produced five customer clusters with a silhouette score of 0.571, reflecting good cluster separation and internal consistency. The results reveal distinct payment behaviors, such as customers who consistently pay on time, those frequently late, and those who have fully completed their payments. These clusters provide practical insights that can support targeted communication, billing, and retention strategies. Furthermore, the study highlights the effectiveness of adaptive clustering techniques in improving segmentation accuracy. The findings contribute to data-driven decision-making in customer management, offering valuable guidance for enhancing operational efficiency and supporting long-term business performance.

Corresponding Author:

Dwi Arman Prasetya,
Department of Data Science, Pembangunan National Veteran University of East Java, Indonesia.
Email: arman.prasetya.sada@upnjatim.ac.id

1. INTRODUCTION

The intense market competition and fluctuations in consumer purchasing power in the property sector have compelled property companies to not only focus on unit sales but also to manage and sustain smooth operational cash flow. Customers fulfillment of their property payment obligations is a key indicator of the continuity of business operations [1]. Irregular or delayed payments by customers can directly impact the company's financial condition, thereby necessitating a more targeted customer management strategy.

In practice, delays in property customer payments not only affect short-term cash flow but also have the potential to disrupt investment plans, debt repayments, and the sustainability of future development projects. Therefore, companies need a system capable of early detection and classification of risks, particularly through the analysis of documented historical payment patterns. A data driven approach not only enhances collection strategies but also contributes to the overall efficiency of company resource management.

Before formulating a customer management strategy, companies should first develop a thorough understanding of their customers. One approach to achieve this is through customer segmentation [2]. Through customer segmentation, businesses can put people into groups according to shared traits or behavioral patterns [3]. In the context of payment behavior, segmentation can assist companies in classifying customers with on-time payments, delayed payments, or other

distinct patterns [4]. Consequently, businesses can design tailored strategies and communication approaches for each customer segment.

In customer data analysis, the success of segmentation largely depends on the selection of appropriate methods and algorithms. It is essential to employ techniques that can accommodate data complexity, especially when customer behavior is highly diverse and not linearly identifiable. Therefore, selecting an adaptive clustering technique capable of accurately capturing data variations is a critical factor in achieving effective segmentation.

This segmentation process can be supported by data mining techniques, which involve extracting valuable information from large datasets [5]. In the business domain, data mining plays a crucial role in uncovering patterns, trends, and hidden relationships within historical data to support decision-making processes [6]. It integrates statistical, mathematical, and computational approaches to efficiently derive insights [7]. One of the primary methods in data mining is clustering, which is used to group data based on similarity without requiring predefined labels or categories.

Clustering is one of the commonly used segmentation analysis techniques in data analysis [8], . It is a *data mining* method categorized under *unsupervised learning*, where data grouping is performed without the use of predefined labels or categories [9]. The fundamental goal of clustering is to find hidden structures in data by grouping objects with similar qualities together and dividing those with different features into distinct clusters.

Clustering is also a part of *machine learning*, specifically within the unsupervised learning approach, which is a branch of *artificial intelligence*. This approach enables systems to learn directly from data by identifying patterns or structures without the aid of labeled information. In practice, clustering is highly valuable for customer segmentation, as it can form groups based on similarities in behavior or characteristics hidden within the data [10]. Therefore, clustering is not only a key technique in *data mining* but also an effective method to support data-driven business strategies [11].

Clustering encompasses various widely used algorithms, such as *K-Means, K-Means*++, DBSCAN, and *Hierarchical Clustering* [12], [13]. Among these, *K-Means* is the most frequently applied due to its simplicity in implementation. However, K-Means demands that the number of clusters (k) be defined in advance, which, if misaligned with the natural structure of the data, might lead to inferior clustering results [14], [15]. This challenge becomes more prominent when dealing with complex and non-uniformly distributed data. Therefore, alternative algorithms such as *X-Means* are utilized. *X-Means* is an extension of *K-Means* that can automatically determine the optimal number of clusters based on internal evaluation metrics, resulting in segmentations that better reflect the underlying data distribution patterns.

X-Means was developed by Pelleg and Moore as an enhancement of the *K-Means* method to address its limitation in determining the number of clusters. *X-Means* is capable of evaluating the data structure and automatically selecting the optimal number of clusters. The algorithm employs the *Bayesian Information Criterion* (BIC) during the clustering process [16], [17]. Through this calculation, the number of clusters does not need to be defined manually but is instead dynamically adjusted according to the data distribution patterns. This provides greater flexibility in clustering, particularly for complex datasets or those lacking a clearly defined cluster structure from the outset [18].

The application of the *X-Means* algorithm in customer segmentation offers advantages in terms of flexibility and effectiveness. This algorithm is capable of handling data without a clearly defined cluster structure, while also minimizing researcher subjectivity in manually determining the number of clusters. Additionally, it provides fast computation in grouping data [19]. Therefore, *X-Means* serves as a more objective and data-driven solution to meet the needs of customer segment analysis in the property industry, which has traditionally relied on conventional and less personalized approaches.

Several studies have analyzed customer payment patterns, such as the research conducted by Nurelasari E., who performed segmentation of customer payment behavior in a multimedia company using the K-Means method and C4.5 for classification. The proposed development in this study is the application of the X-Means method to automatically determine the optimal number of clusters and to generate customer segmentation that aligns with the underlying characteristics of the data [20].

Meanwhile, Sembiring et al. applied the *K-Means* algorithm to cluster vehicle loan arrears data based on three variables: car brand, district, and duration of arrears. The results revealed a dominant pattern of payment delays associated with specific brands and regions [21], [22]. However, the study did not incorporate cluster quality evaluation metrics such as the *Silhouette Score* or the *Davies-Bouldin Index*. This limitation highlights the importance of evaluating the quality of each method to assess how effective, accurate, and representative the applied approach is in producing meaningful segmentation of arrears patterns within the dataset.

Recent studies have also demonstrated the effectiveness of X-Means in discovering meaningful customer or transaction-based segments. Vahidi Farashah et al. applied X-Means clustering to classify customers based on transactional behavior in the telecommunications sector, showing improved interpretability and automatic cluster determination compared to traditional K-Means [23]. Similarly, Yucebas et al. integrated *X-Means* and decision tree models to cluster real-estate properties before conducting price prediction, emphasizing that X-Means effectively captures the heterogeneity in real-estate and financial behavior data [24]. In addition, Tabianan et al. highlighted that clustering methods are valuable tools for identifying patterns in customer purchase behavior and for guiding business strategies through intelligent segmentation [25].

Building on these findings, this research aims to perform segmentation of historical customer payment data from a property company, PT X, to understand the characteristics of customer payment patterns. The findings are intended to serve as a foundation for developing differentiated service, collection, and communication strategies for each customer

segment, while also supporting data-driven decision-making to manage risks and improve operational efficiency. Compared to other clustering techniques, such as *K-Means*, *DBSCAN*, and *Hierarchical Clustering*, the *X-Means* algorithm offers several advantages. Unlike *K-Means,* which requires a predefined number of clusters, *X-Means* automatically determines the optimal number based on internal evaluation criteria, thereby reducing subjectivity in cluster selection. *DBSCAN*, while effective for identifying arbitrarily shaped clusters and noise, often struggles with datasets that vary in density. Meanwhile, *Hierarchical Clustering* provides interpretability through dendrogram structures but is computationally expensive for large-scale datasets. Given these considerations, *X-Means* is more adaptive and efficient for identifying natural groupings within complex behavioral datasets such as property customer payment records.

This research contributes to the development of customer segmentation in the property sector by applying an adaptive clustering method with internal validation to produce more objective and actionable groupings. The research gap addressed in this study lies in the limited application of adaptive clustering methods specifically *X-Means* for analyzing customer payment behavior in the property domain. Most prior studies rely on conventional clustering algorithms without automated cluster determination or comprehensive evaluation metrics. Therefore, the novelty of this study is the implementation of the *X-Means* algorithm combined with Silhouette Score evaluation to obtain representative, data-driven customer segments. Based on the aforementioned research objectives and identified gaps, the next section outlines the methodology used to implement and evaluate the proposed segmentation model.

2. RESEARCH METHOD

The research involves several stages of data processing, beginning with data preparation and continuing through to method implementation. These stages are illustrated in the workflow presented in Figure 1, which outlines the research workflow.

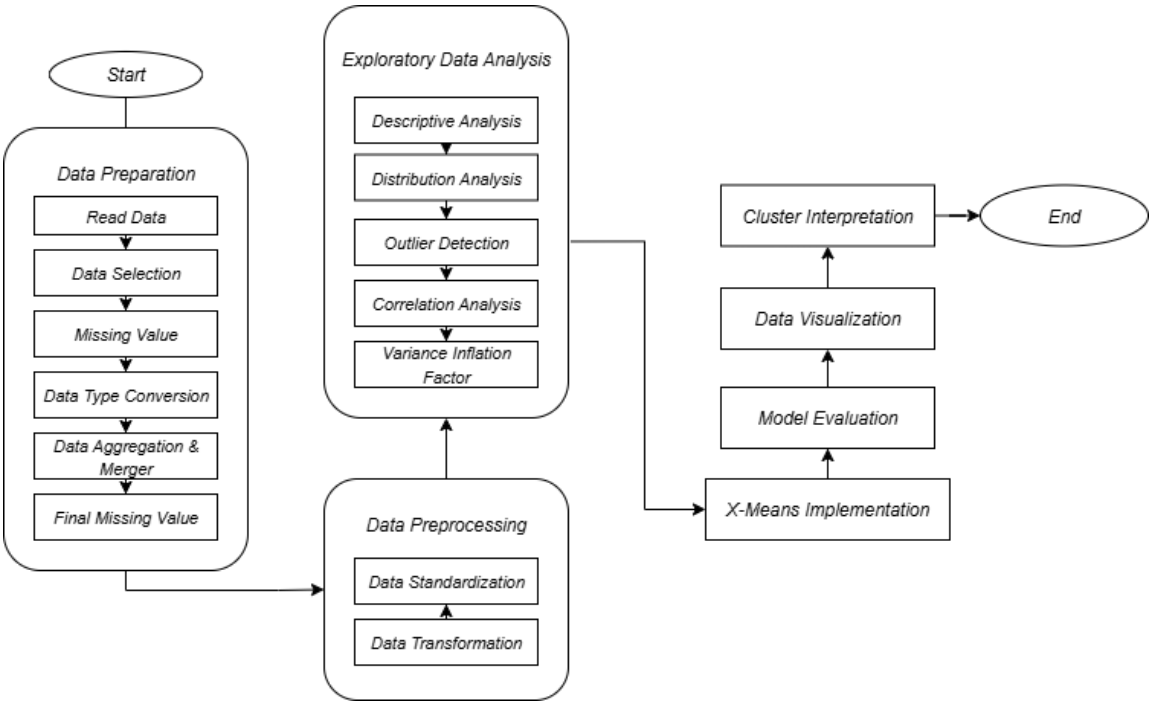

Figure 1. Research Workflow

The research workflow is illustrated in Figure 1. The Research Workflow outlines the sequential stages undertaken in this study, beginning with data preparation, which includes reading data, selecting relevant features, handling missing values, data type conversion, and aggregation. In this stage, four key features are prioritized *payment amount*, *payment frequency*, *delay duration*, and *payment status* as they directly represent the behavioral dimensions of customers in fulfilling their financial obligations. The *payment amount* reflects the customer's purchasing power and financial commitment, while *payment frequency* indicates consistency and reliability in meeting installment schedules. The *delay duration* captures the customer's level of discipline and potential credit risk, and *payment status* summarizes the overall completion or delinquency pattern. Together, these features provide a comprehensive behavioral profile that enables accurate segmentation based on payment tendencies.

This is followed by the data preprocessing stage, involving data standardization and transformation to ensure data consistency and readiness for analysis. The process continues with exploratory data analysis (EDA), where descriptive statistics, distribution patterns, outlier detection, correlation analysis, and multicollinearity detection using the *Variance Inflation Factor* (VIF) are performed. Upon completion of the EDA, the *X-Means* clustering algorithm is implemented to

segment customers based on payment behavior. The resulting clusters are then evaluated using appropriate metrics, visualized for better interpretability, and finally interpreted to derive meaningful insights that can support strategic business decisions.

2.1 Dataset

The data used in this research are secondary data, consisting of installment records for each sold property unit within the period from 2019 to November 4, 2024, comprising a total of 9,615 entries. The primary data, which includes the price of each property unit, was obtained from a property company located in Sidoarjo, East Java.

2.2 Data *Preprocessing*

Before data preprocessing, a data selection process was carried out to determine the features to be merged into a single dataset, followed by data preparation. The resulting dataset consists of 7 features: *Unit Number, Number of Transactions, Total Payment, Price, Difference, Payment Status*, and *Late Amount*, comprising a total of 386 rows. Table 1 presents the description and data type of each feature.

Table 1. Metadata

Feature	Description	Data Type	Rationale
Unit Number	Unique identifier for each property unit	string	Used to represent each transaction entity uniquely; serves as a key identifier for data grouping and aggregation.
Number of Transactions	Total number of payment transactions made	int	Indicates the frequency and consistency of payment behavior; higher frequency often reflects stable commitment and lower default risk
Total Payment	The total amount of payment that has been paid in Rupiah	int	Represents actual financial contribution; relevant to understanding the level of payment completion and customer financial engagement
Price	Total price of the property unit that should have been paid in Rupiah	int	Used as a baseline for comparing total payment and determining discrepancies between paid and expected amounts; crucial for assessing customer affordability
Difference	The difference between the total payment and the price indicates an overpayment or underpayment	int	Captures financial deviation—key to identifying late or incomplete payments, which signal potential financial distress
Payment Status	The final status of the payment (for example, paid or unpaid, represented by a code)	String	Reflects final behavioral outcome in the payment cycle; acts as categorical validation for segmentation results
Late Amount	The number of late payments (in number of occurrences)	Int	Quantifies payment irregularities and delay patterns, serving as a strong indicator of risk and credit discipline

Before the transformation process, the data were first examined to identify the presence of missing values and outliers. Outlier detection was conducted using visualization techniques such as *boxplots*, along with analysis of extreme values in numerical features. Values that significantly deviated from the general distribution were considered for removal, reprocessing, or retention depending on the research objective. The normalization process was then carried out using the *RobustScaler* method, which is known for its resilience to the influence of outliers compared to other standardization techniques. This step is particularly important given that certain features, such as the number of transactions and the difference, exhibit uneven distributions.

After the dataset was constructed, data transformation and standardization were performed to prepare it for clustering customer payment patterns. The scale standardization was conducted using the *RobustScaler* method. Following standardization, feature selection was carried out for the clustering process. Based on the exploration of data distribution, outliers, and inter-feature correlations, the most relevant features for representing customer payment behavior were identified as the *Number of Transactions*, the *Late Amount*, the difference between *Price* and *Total Payment*, and *Payment Status*. These four features were used as the input data for the customer segmentation process using clustering techniques.

.

2.3 RobustScaler

RobustScaler is a data *scaling* technique available in the *sklearn. preprocessing* module, designed to address the impact of outliers during the standardization process. Unlike *StandardScaler*, which relies on the mean and standard deviation for transformation, *RobustScaler* utilizes the median and the *interquartile range* (IQR), the difference between the third quartile (Q3) and the first quartile (Q1) [26]. His approach makes *RobustScaler* more robust against the influence of outliers, which can otherwise distort the data distribution and reduce its representativeness.

Standardization using *RobustScaler* is performed using the following formula:

$$x' = \frac{x - median}{IQR} \quad (1)$$

and:

$$IQR = Q3 - Q1 \quad (2)$$

Where x represents the original data value, the *median* is the central value of the distribution, and *IQR* is the difference between the third quartile (Q3) and the first quartile (Q1). The standardized value, denoted as x', indicates how far a data point deviates from the *median* when measured in units of the IQR [27]. Therefore, this method is well-suited for dataset with non-normal distributions or those containing outliers, as it is not significantly influenced by extreme value deviations.

In this study, the use of *RobustScaler* is considered appropriate, as several numerical features, such as *Number of Transactions, Difference,* and *Total Payment*, exhibit wide value ranges and contain outliers. If these data were not transformed using a suitable method, the extreme values could dominate the clustering process and obscure the underlying patterns intended to be revealed. By applying *RobustScaler*, the data are normalized using the *median* and *interquartile range* (IQR), thereby minimizing the influence of extreme values and resulting in more accurate and representative clustering outcomes that better reflect the underlying data structure.

2.4 X-Means Algorithm

The *X-Means algorithm* is an extension of *K-Means* that is considered to offer faster computational performance than *K-Means* [16]. In this research, *X-Means* is employed to cluster customers based on similarities in their payment patterns, using a more flexible approach that adapts to the characteristics of the data. The clustering process begins with a small number of clusters and expands gradually by splitting clusters based on internal evaluations performed by the *X-Means algorithm* to improve clustering quality [28]. Evaluation of cluster splits is conducted using the *Bayesian Information Criterion* (BIC) to ensure optimal results that align with the natural structure of the data. BIC serves as the evaluation basis by calculating the BIC values before and after the split and comparing them to determine whether the division significantly improves the model[29].

BIC combines a measure of model fit to the data (*log-likelihood*) with a penalty for model complexity [30]. Its primary objective is to prevent overly complex models and reduce the risk of overfitting. The general form of the BIC equation is presented as follows:

$$BIC = L - \frac{p}{2} logN \quad (3)$$

Explanation:

a) L is the *log-likelihood* value of the model,
b) p is the number of parameters in the model,
c) N is the number of observations or data points.

A higher BIC value indicates a better model, as it balances model fit and simplicity. In the *X-Means algorithm*, the BIC value is calculated for both pre- and post-cluster split scenarios. If the split results in a higher BIC value, the division is accepted; otherwise, the cluster structure remains unchanged[31]. This approach is repeated until no additional cluster splits increase BIC, resulting in the ideal number of clusters that correspond to the data's inherent structure.

This research employs the *X-Means algorithm* because the customer data being analyzed exhibits uneven distribution, and the optimal number of clusters is not known in advance. *X-Means* offers the advantage of automatically determining the appropriate number of clusters through evaluation using the BIC, allowing the segmentation process to adapt to complex data structures. This is particularly relevant in analyzing customer payment patterns, which often display high variability and are difficult to classify manually. Therefore, the use of *X-Means* is expected to produce more accurate clusters that better reflect real-world conditions.

2.5 Sillhouette Score

The *Silhouette Score* is an evaluation tool intended to objectively examine the quality of clustering results without using labels or ground truth [32]. his metric measures how similar an object is to other objects within the same cluster, compared to objects in different clusters.

The *Silhouette Score* is determined for each data point and ranges from minus one to one [12]. A higher *Silhouette Score*, approaching 1, indicates better clustering quality, as it reflects high similarity among objects within the same cluster and clear separation from those in other clusters. Conversely, a score near 0 suggests that the object lies at the boundary between two clusters, while a negative value indicates that the object is closer to a different cluster than its own, signaling a potential misclassification [33].

The calculation of the *Silhouette Score* is based on two main components:

a) $a(i)$: the average distance between data point i and all other points in the same cluster. (referred to as cohesion),

b) $b(i)$*:* the average distance between data point i and all points in the closest neighboring cluster (referred to as separation).

Using these two values, the *Silhouette Score* $s(i)$ for the-i the data point is calculated using the following formula:

$$s(i) = \frac{b(i)-a(i)}{\max(a(i),b(i))} \quad (4)$$

The average value of all $s(i)$ scores across the dataset serve as a general indicator of clustering quality. *Silhouette Score* is commonly used to evaluate clustering results with varying numbers of clusters to determine the optimal cluster count.

In this research, the *Silhouette Score* is employed as an evaluation tool to assess the quality of the segmentation produced by the clustering algorithm, specifically *X-Means*. This metric is selected for its ability to provide an objective assessment of how well the resulting clusters reflect the natural structure of the data. Additionally, the *Silhouette Score* is useful for comparing different clustering scenarios to identify the most appropriate number of clusters before conducting further interpretation and analysis.

Although the *Silhouette Score* is a widely used evaluation metric, its results should be analyzed in conjunction with visualizations and contextual understanding of the data. In some cases, a high *silhouette* value does not necessarily indicate a segmentation that aligns with business objectives or is easily interpretable in practice. Therefore, a combined approach that integrates quantitative analysis, such as the *Silhouette Score*, with qualitative analysis based on domain knowledge is essential to ensure that the resulting clusters are not only statistically valid but also operationally relevant.

3. RESULTS AND DISCUSSION

The research process was conducted to identify payment patterns formed within each cluster. Subsequently, the model was evaluated, the clusters were visualized, and their characteristics were interpreted accordingly.

3.1 Data Preparation

Data preparation stage began with the aggregation of transaction data based on the *Unit Number*, by grouping all transactions associated with each individual property unit. This process resulted in two new features: *Number of Transactions*, representing the frequency of payments made, and *Total Payment*, indicating the accumulated payment amount per unit.

After the transaction data was summarized, data cleaning was performed on another file containing property price information. In this step, the first two irrelevant rows were removed, and column names were standardized for consistency and ease of use, for instance, renaming column F to *Unit Number* and *Unnamed: 3* to *Price*.

The next step involved merging the transaction data with the property price data using the *Unit Number* as the key column. This merge enabled the calculation of a new feature called *Difference*, defined as the difference between the *Price* and the *Total Payment*. This feature was used to derive the *Payment Status*, where a unit was categorized as *Paid in Full* if the difference was less than or equal to 0, and *Not Fully Paid* if the difference exceeded 0.

Subsequently, historical payment dates were analyzed to calculate the number of late payments. By computing the difference in days between the due date (*Payment Date*) and the actual payment date (*Received Date*), transactions exceeding the due date were identified. Transactions with a difference greater than 0 days were classified as late payments. These data were then grouped by *Unit Number* to obtain the total *Number of Late Payments* per unit.

Once all the key features were generated, a merging process was conducted to consolidate all information, including the number of transactions, total payment, property price, difference, payment status, and number of late payments, into a single complete dataset. Finally, to address missing values in the *Number of Late Payments* column, which occurred because not all units had experienced payment delays, the missing values were filled with 0 and converted to an integer data type.

3.2 Data Preprocessing

Before data normalization was performed, the categorical values in the *Payment Status* feature were first converted into a numerical format to ensure compatibility with subsequent analytical processes. This step is essential, as the clustering algorithm used in this research can only process numerical data. The *Payment Status* feature, which originally contained two labels, *Paid in Full* and *Not Fully Paid*, was transformed into *binary* form using *binary encoding*, with *Paid in Full encoded* as 1 and *Not Fully Paid* as 0. This transformation aims to represent the payment status in a numerical format without assigning undue weight to either category.

Once all features were converted into numerical form, data normalization was conducted using the *RobustScaler* method. This method was selected due to its robustness against outliers, in contrast to other standardization techniques such as *Min-Max Scaling* or *StandardScaler*. *RobustScaler* performs transformation based on the median and *interquartile range* (IQR), thereby minimizing the influence of extreme values on the scaling process. This is particularly important because several features, such as *Number of Transactions*, *Difference*, and *Total Payment*, exhibited uneven distributions and contained significant outliers. If not properly addressed, these outliers could distort the clustering process and lead to biased results.

Outlier identification was carried out through *boxplot* visualizations of the numerical features used in the study. The visualizations revealed several extreme values falling outside the *interquartile range* (IQR), represented by data points beyond the whiskers of the *boxplot*. The most prominent outliers were found in the *Number of Transactions* and *Difference* features, where some data points showed values substantially higher than the majority of the dataset. This indicates the presence of customers with exceptionally high transaction activity or large discrepancies between the property *Price* and *Total Payment*, either underpayment or overpayment. A more detailed visualization of these outliers is presented in Figure 2.

These outliers are important to examine, as they can influence the clustering results, particularly if not addressed using appropriate transformation methods. In the context of this study, outliers were not immediately removed but were further analyzed. Outliers considered contextually relevant from a business perspective, such as customers with high transaction intensity or extreme delays, were retained. In contrast, extreme values that appeared inconsistent or potentially indicative of data entry errors were considered for reprocessing or removal

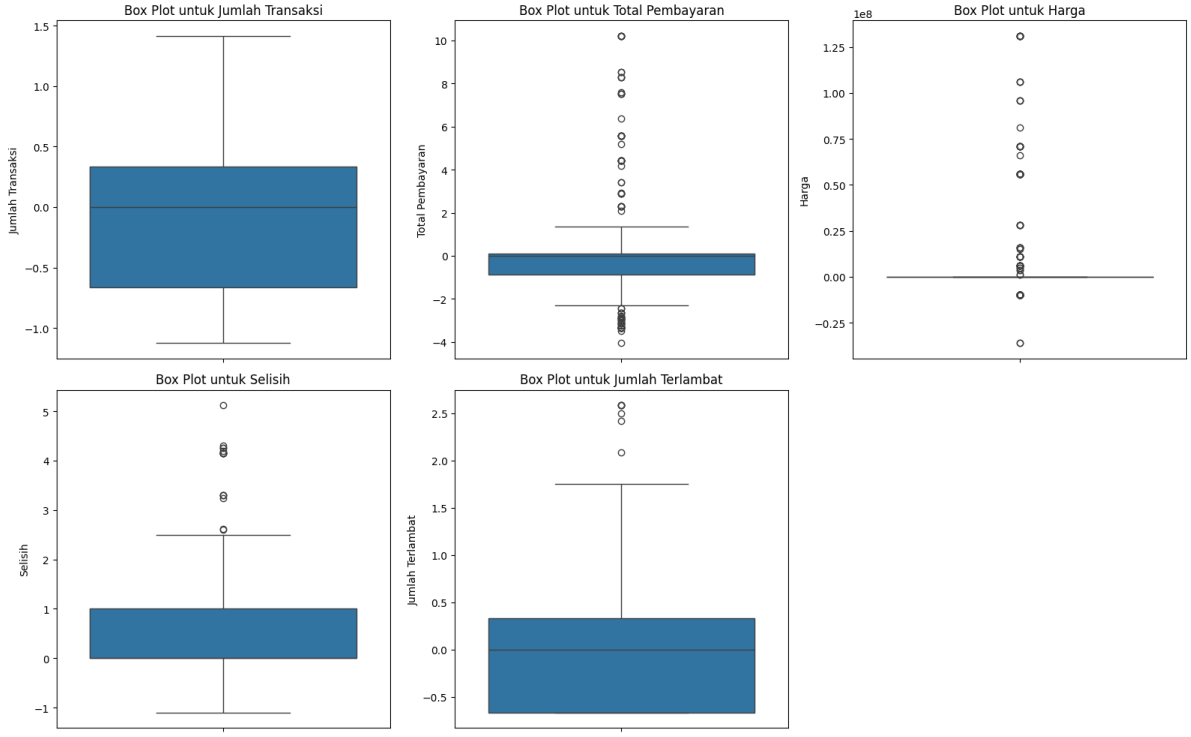

Figure 2. Identification Outlier

Following the outlier handling stage, an analysis of feature relationships was conducted to evaluate the extent to which each feature is interrelated and relevant to mapping customer payment behavior. Understanding feature correlations is essential to avoid information redundancy, which can distort clustering results. Features with very high correlations may lead to multicollinearity, and thus should be considered for elimination or substitution to ensure that the segmentation results remain valid and optimally reflect data variation.

Therefore, a feature correlation analysis was performed to assess how strongly the features are related and how relevant they are to the behavioral mapping of customer payments. Features exhibiting very high correlations pose a risk of multicollinearity and may need to be removed or replaced to maintain the integrity and representativeness of the segmentation. The results of this analysis are visualized in a heatmap, as shown in Figure 3, which provides an initial overview of the strength of associations among features, whether strong, moderate, or weak. This visualization serves as a preliminary basis for determining which features should be retained for cluster formation and which should be eliminated due to information redundancy

Based on Figure 3, it is evident that the *Number of Transactions* is highly positively correlated with *Late Amount*, with a correlation coefficient of 0.89, while *Total Payment* also shows a strong correlation with *Price*, at 0.84. Additionally, the *Difference* feature exhibits a strong negative correlation with *Payment Status*, at −0.69, indicating that a larger difference typically reflects an unpaid status. Other correlations are weaker, such as the correlation between *Number of Transactions* and *Payment Status* (0.35), and between *Total Payment* and *Late Amount* (0.37).

High correlations among certain features suggest the potential presence of overlapping information. When such features are used simultaneously in the clustering process, they may distort the cluster structure due to the lack of independence in their contributions. Therefore, further steps, such as multicollinearity evaluation, are necessary to ensure that only features that are truly relevant and independent are included in the model.

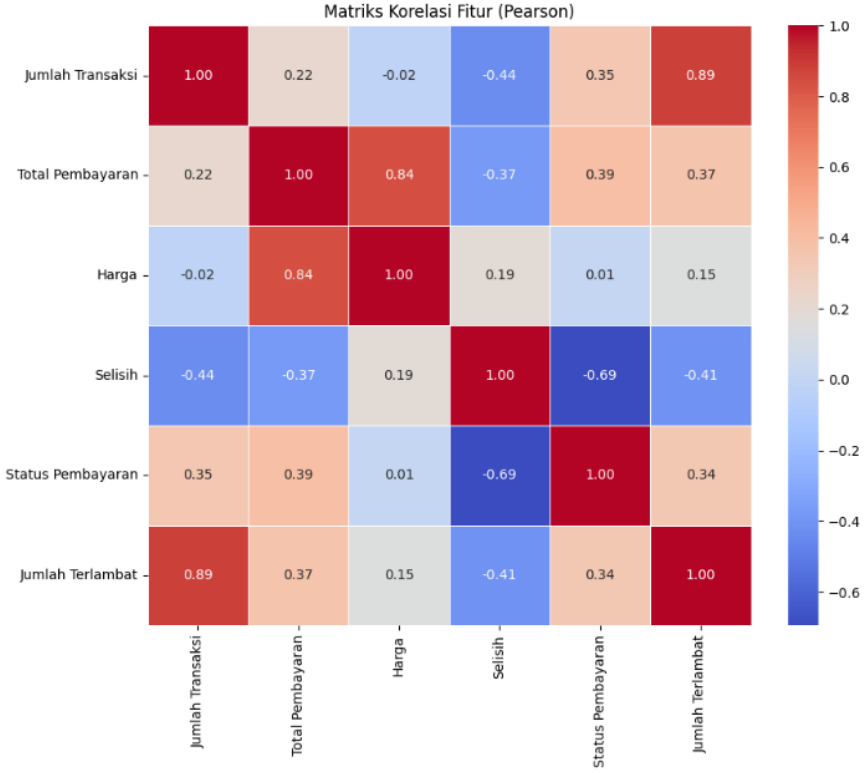

Figure 3. Correlation Pearson

The strongest relationships are observed between *Number of Transactions* and *Late Amount*, *Total Payment* and *Price*, as well as *Difference* and *Payment Status* (in a negative direction). These patterns indicate that a higher number of transactions is associated with a greater likelihood of payment delays; higher property prices tend to result in larger total payments; and a larger difference corresponds to a worse payment status, reflecting an outstanding balance that has yet to be settled.

To reinforce the findings from the correlation analysis and ensure the absence of multicollinearity among features, the *Variance Inflation Factor* (VIF) was calculated. VIF is a method used to measure the extent to which an independent variable can be explained by other variables within a model regression. A high VIF value indicates the presence of information redundancy or multicollinearity. As a general rule, a VIF value exceeding 10 suggests a high degree of multicollinearity, and such features should be considered for elimination from the clustering model to prevent distortion in the segmentation results.

Table 2. VIF

Feature	VIF
Number of Transactions	5,24
Total Payment	829,78
Price	833,87
Difference	342,29

Payment Status	2,02
Late Amount	5,44

Based on the results of the Pearson correlation analysis and VIF calculation, it was found that several features exhibited very high correlations with one another and VIF values that far exceeded the multicollinearity tolerance threshold (>10). For instance, the features *Price* and *Total Payment* had VIF values of 833.87 and 829.78, respectively, indicating a high degree of multicollinearity that could negatively impact the clustering process.

Therefore, the features selected for use in the clustering process, those considered free from correlation issues or high multicollinearity, and capable of effectively representing customer behavior, were narrowed down to four: *Number of Transactions*, *Late Amount*, *Difference*, and *Payment Status*.

3.3 Comparison of X-Means and K-Means Computation Times

This study compares the computational efficiency between the X-Means and K-Means algorithms. This comparison was conducted to see how optimal X-Means is in processing customer payment data compared to the more commonly used baseline method, namely K-Means. Table 3 shows a comparison of the computation between X-Means and K-Means.

Table 3. Comparison of Computation Time

Method	Computational Time (s)
X-Means	0.0243
K-Means	0.1967

Based on Table 3, X-Means was able to complete the clustering process in a shorter time (0.0309 seconds) compared to K-Means (0.1625 seconds). This shows that X-Means works more efficiently on the dataset used. This time difference arises because X-Means has a mechanism that dynamically adjusts the number of clusters through Bayesian Information Criterion (BIC) evaluation. In this way, the process of finding the appropriate number of clusters does not need to be done manually as in K-Means. Thus, the results of this study further strengthen the reasons for choosing X-Means, which has the advantage of better computational speed compared to the comparison method.

3.4 Clustering Results

The *X-Means algorithm* was implemented using the four selected features for clustering. Based on the analysis results, the optimal number of clusters obtained was five. To evaluate the quality of the clustering, the *Silhouette Score* was used, which yielded a value of 0.571. This score indicates that the clustering result is reasonably good, as the clusters are relatively well-separated and exhibit consistent internal structure. It suggests that dividing the data into five clusters effectively reflects distinct patterns or characteristics among the data groups.

The results of this clustering can serve as a foundation for further analysis, such as customer behavior segmentation, identification of potential payment delay risks, or the development of other strategic initiatives relevant to the company's needs. Therefore, the insights gained from the clustering process hold significant practical value in supporting data-driven decision-making.

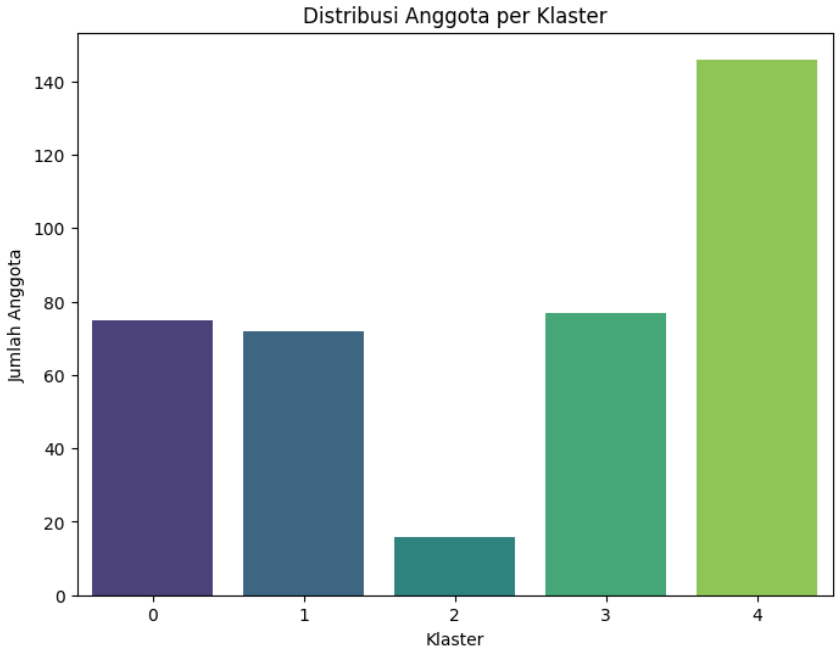

Figure 4. Distribution Cluster

Figure 4 presents a visualization of the distribution of the number of members within each resulting cluster. There are five clusters with unequal membership sizes. Cluster 4 contains the highest number of members, with a total of 145 customers, while Cluster 2 has the fewest, with fewer than 20 members. This disparity reflects the variation in customer behavior characteristics successfully grouped through the clustering process.

Additionally, a supplementary visualization was generated to illustrate the spatial distribution of each cluster in a two-dimensional space. This facilitates the observation of cluster structure and proximity, thereby supporting a more comprehensive and intuitive interpretation of the clustering results.

The *t-SNE* visualization in Figure 5 illustrates the distribution of members within each cluster in a more intuitive two-dimensional space. It can be observed that the points representing each cluster are concentrated within specific regions, indicating clear separation between clusters. Different colors for each cluster aid in depicting the structure of the formed segmentation. Red "X"-shaped markers represent the centroids of each cluster.

Based on the visualization, cluster centers such as Cluster 1, Cluster 3, and the majority of Cluster 0 appear to be well-separated, suggesting that the clustering process successfully grouped the data into relatively consistent patterns. However, there are also indications of overlap or proximity between certain clusters, particularly between Cluster 2 and Cluster 4, which may reflect similarities in the characteristics of the data points within those clusters.

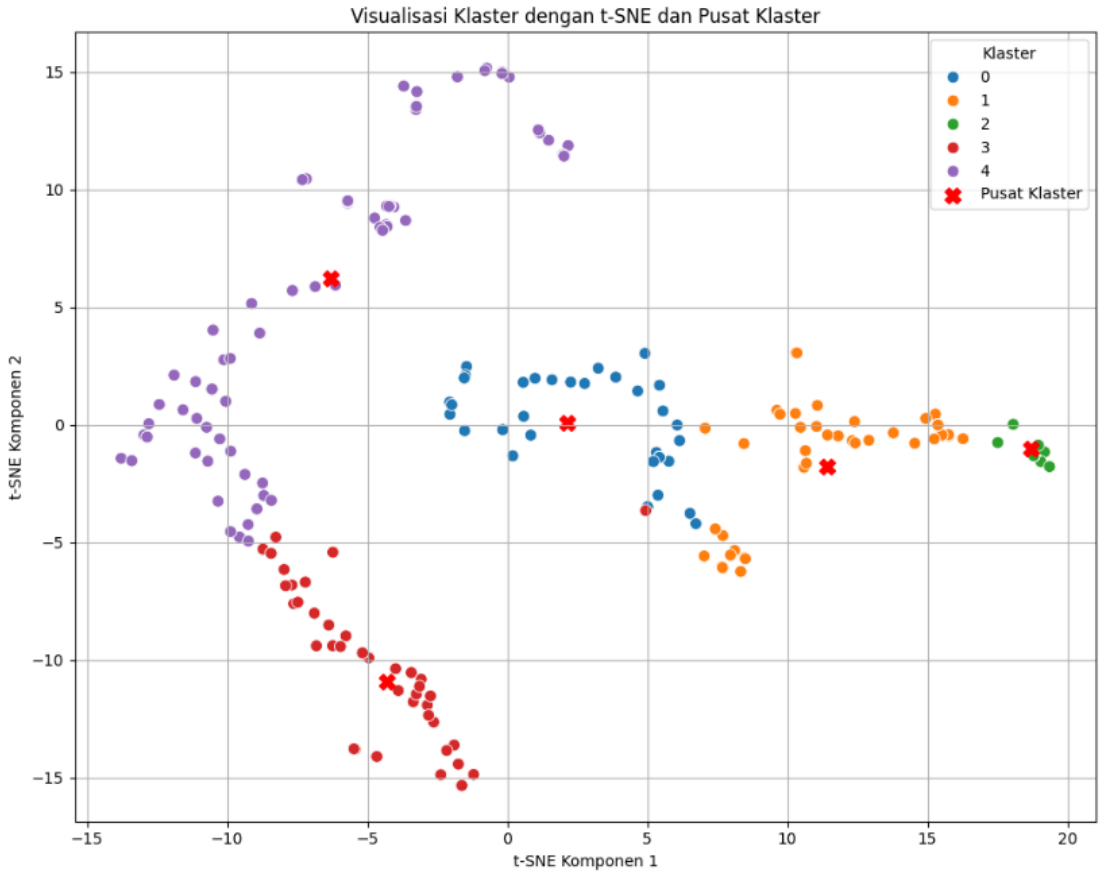

Figure 5. Cluster Center

As an effort to enhance the interpretation of the segmentation results, visualization was carried out using the *t-distributed Stochastic Neighbor Embedding (t-SNE)* method. This technique is useful for projecting high-dimensional data into a two-dimensional space, thereby allowing the distribution patterns between clusters to be more clearly observed. In addition to improving understanding, the *t-SNE* visualization also aids in identifying the relative positions of each cluster. The following table presents the coordinates of each cluster centroid in the two-dimensional space resulting from the *t-SNE* projection.

The *t-SNE* visualization reveals that most clusters appear to be fairly well-separated, particularly Clusters 1 and 3, which form more compact and structured groupings. This indicates notable differences in payment behavior among customers within these clusters. On the other hand, Clusters 2 and 4 appear to have overlapping or closely situated distribution areas, suggesting similarities in customer characteristics such as transaction intensity or patterns of payment delay that are nearly alike.

Table 4. Cluster Center Point Coordinates

Cluster	TSNE-1	TSNE-2
0	2.131000	0.089
1	11.398000	-1.762
2	18.684999	-1.023
3	-4.297000	-9.921
4	-6.317000	6.223

Based on Table 3, each cluster exhibits significantly different centroid coordinates within the two-dimensional space projected by *t-SNE*. For instance, Cluster 1 has a centroid located at (11.39, 1.76), which is relatively distant from that of Cluster 3. This distance supports the visual findings that these two clusters are indeed distinctly separated. It further reinforces the evidence that the *X-Means algorithm* successfully segmented customers into groups that reflect distinct payment behavior characteristics.

The differences in cluster centroid coordinates can serve as a foundation for developing targeted approaches for each customer segment. For example, clusters located in regions with extreme negative coordinate values may be associated with problematic payment behavior and could be prioritized in early intervention or debt collection strategies. Conversely, clusters with stable centroids, well-separated from those associated with risk, may be candidates for loyalty programs or timely payment incentives.

Based on the analysis results, both the moderately high *Silhouette Score* and the *t-SNE* visualization, it is evident that the clusters are formed, providing a strong foundation for the resulting segmentation. These findings are highly valuable for understanding customer behavior, particularly in mapping payment delay risks, identifying transaction patterns, and designing more precise, data-driven business strategies.

3.5 Cluster Interpretation

Table 5. Cluster Average

Cluster	Number of Transactions	Late Amount	Difference	Payment Status
0	28.880000	8.626667	9.168028e+06	0.000000
1	10.611111	1.083333	2.956975e+07	0.00000
2	14.187500	1.125000	6.704086e+07	0.000000
3	13.727273	1.792208	-6.997656e+05	0.987013
4	36.958904	13.972603	2.203878e+04	0.993151

Based on the average values calculated for each cluster as shown in Table 4, it can be observed that each cluster exhibits distinct average values across key features such as *Number of Transactions*, *Late Amount*, *Difference*, and *Payment Status*. These differences confirm that each cluster represents a segment of users with unique and distinguishable behavioral characteristics.

Cluster 4, for instance, stands out with the highest average *Number of Transactions* and *Late Amount* among all clusters, approximately 36.96 transactions and 13.97 delays, respectively. Nevertheless, this cluster also shows a very high average *Payment Status* of 0.993, indicating that the majority of customers in this group ultimately fulfill their payment obligations, despite frequent delays. This suggests a profile of active users who possess the financial capability to pay but exhibit low timeliness or payment discipline.

Conversely, Cluster 1 exhibits a markedly different behavior. Users in this cluster recorded the lowest number of transactions and delays, yet have a *Payment Status* of 0. This indicates that although they rarely engage in transactions, none of their payments were completed. A similar pattern is observed in Cluster 2; although users in this cluster performed slightly more transactions than those in Cluster 1, their *Payment Status* remains low, suggesting a relatively high risk of default.

Cluster 0 presents a rather unique condition, with a large and positive *Difference* value and a relatively high frequency of payment delays. However, the *Payment Status* remains at 0, indicating that despite having transactions and a significant outstanding balance (possibly arrears), no payments have been settled. On the other hand, Cluster 3 demonstrates more favorable characteristics, with a *Payment Status* close to 1 and a negative *Difference* value. This may suggest overpayment or refunds, and indicates that users in this group tend to comply with their financial obligations.

When compared holistically, Clusters 1 and 2 appear to pose lower risks, while Clusters 0 and 4 present payment-related challenges despite exhibiting high transaction volumes. Cluster 3 falls in the middle, demonstrating relatively balanced characteristics. These clustering results can be utilized by the company to develop more targeted customer management strategies. For instance, clusters characterized by high transaction frequency but significant payment delays could be targeted with financial education programs or automated reminders.

Meanwhile, customers in clusters with low delays and strong payment status could be offered incentives such as discounts or loyalty programs to reinforce their positive behavior. By analyzing the differences across clusters in this way, the company can gain deeper insights into customer profiles and make more strategic decisions both in terms of risk management and in designing more personalized and relevant product or service offerings. Understanding the profile of each cluster is essential as a foundation for establishing service policies, collection strategies, and resource allocation tailored to the risk level of each customer group.

In general, the interpretation of clusters based on their average values provides a clearer picture that each cluster possesses distinct characteristics or user profiles. These differences reflect variations in user behavior, particularly in terms of transaction frequency, level of payment delays, and final payment status. Such distinctions allow each cluster to be understood as a representation of user segments with differing needs and risk levels.

This information is highly valuable as a foundation for designing more targeted service strategies, and it serves as a reference for developing subsequent policies, such as credit risk management, tailored product offerings, or more personalized communication approaches for each customer segment.

4. CONCLUSION

This research successfully segmented customer data into five clusters with distinct payment behavior patterns using the *X-Means* clustering method. The *X-Means* approach enabled the identification of diverse customer profiles, ranging from on-time payers to those who frequently delay or default on payments. This adaptivity allowed the clustering process to automatically determine the optimal number of clusters without manual intervention, enhancing the accuracy and relevance of the segmentation.

The computational comparison further supports the efficiency of *X-Means*, which completed the clustering process in 0.0243 seconds, significantly faster than *K-Means* at 0.1967 seconds. This improvement demonstrates that *X-Means* can process customer payment data more efficiently by dynamically adjusting the cluster number using *Bayesian Information Criterion* (BIC) evaluation. The *Silhouette Score* of 0.571 also indicates good cluster separation, validated visually through t-SNE visualization with mostly distinct group boundaries.

The practical implications of these findings are significant for property management companies. By leveraging adaptive clustering results, companies can design targeted strategies such as personalized collection methods, service adjustments, and loyalty programs based on customer risk levels. This data-driven segmentation enables better resource allocation, reduces payment delay risks, and enhances operational efficiency in managing diverse customer behaviors.

5. ACKNOWLEDGEMENTS

The author would like to express sincere gratitude to all parties who have provided support in the preparation of this article. Special thanks are extended to my academic supervisors for their invaluable guidance throughout the writing process, as well as to the property company PT X for providing the necessary data and information used in this research. Appreciation is also given to my parents, colleagues, and others who have offered constructive feedback, suggestions, and encouragement. It is hoped that this article will contribute meaningfully to the advancement of knowledge, particularly in the fields of data analysis and customer management.

6. REFERENCES

[1] Sintia Oktapani and Neni Maryani, "Pengaruh Operating Cash Flow, Profitabilitas, dan Leverage terhadap Financial Distress pada Perusahaan Properti dan Real Estate yang Terdaftar di Bursa Efek Indonesia (BEI) Tahun 2018-2022," *El-Mal J. Kaji. Ekon. Bisnis Islam*, vol. 5, no. 4, pp. 3089–3104, 2024, doi: 10.47467/elmal.v5i4.1861.

[2] A. Suyatno, S. Arief, M. Asir, M. A. Anwar, and M. D. Sanusi, "Penerapan Strategi Segmenting dan Targeting dalam meningkatkan Kinerja Pemasaran: Literatur review," *J. Econ. Bussines Account.*, vol. 6, no. 2, pp. 1598–1609, 2023, doi: 10.31539/costing.v6i2.5434.

[3] D. A. Prasetya, A. P. Sari, M. Idhom, and A. Lisanthoni, "Optimizing Clustering Analysis to Identify High-Potential Markets for Indonesian Tuber Exports," *Indones. J. Electron. Electromed. Eng. Med. Informatics*, vol. 7, no. 1, pp. 113–122, 2025, doi: 10.35882/skzqbd57.

[4] B. E. Adiana, I. Soesanti, and A. E. Permanasari, "Analisis Segmentasi Pelanggan Menggunakan Kombinasi Rfm Model Dan Teknik Clustering," *J. Terap. Teknol. Inf.*, vol. 2, no. 1, pp. 23–32, 2018, doi: 10.21460/jutei.2018.21.76.

[5] E. R. Sitorus and I. Nugraha, "Customer Segmentation Analysis with RFM Model (Recency, Frequency, Monetary) and K-Means Clustering: Case Study of Bottled Water Sales at PT XYZ," *JSE (Jurnal Serambi Eng.*, vol. 10, no. 2, pp. 12665–12675, 2025.

[6] M. Risqi Ananda, N. Sandra, E. Fadhila, A. Rahma, and N. Nurbaiti, "Data Mining dalam Perusahaan PT Indofood Lubuk Pakam," *Com. Commun. Inf. Technol. J.*, vol. 2, no. 1, pp. 108–119, 2023, doi: 10.47467/comit.v2i1.124.

[7] R. S. Wahono, *Data Mining Data mining*, vol. 2, no. January 2013. 2023. [Online]. Available: https://www.cambridge.org/core/product/identifier/CBO9781139058452A007/type/book_part

[8] N. Mirantika, T. S. Syamfithriani, and R. Trisudarmo, "Implementasi Algoritma K-Medoids Clustering Untuk Menentukan Segmentasi Pelanggan," *J. Nuansa Inform.*, vol. 17, no. 1, pp. 196–204, 2023.

[9] F. Handayani, "Aplikasi Aplikasi Data Mining Menggunakan Algoritma K-Means Clustering untuk Mengelompokan Mahasiswa Berdasarkan Gaya Belajar," *J. Teknol. dan Inf.*, vol. 12, no. 1, pp. 46–63, 2022, doi: 10.34010/jati.v12i1.6733.

[10] D. Marcelina, A. Kurnia, and T. Terttiaavini, "Analisis Klaster Kinerja Usaha Kecil dan Menengah Menggunakan Algoritma K-Means Clustering," *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 3, no. 2, pp. 293–301, 2023, doi: 10.57152/malcom.v3i2.952.

[11] K. Maulida Hindrayani, P. R. Aji Data Science, T. F. Maulana, and E. Maya, "Business Intelligence For Educational Institution : A Literature Review," *Netw. Secur. Inf. Syst.*, vol. 2, no. 1, pp. 22–25, 2020.

[12] T. M. Fahrudin, P. A. Riyantoko, K. M. Hindrayani, and M. H. P. Swari, "Cluster Analysis of Hospital Inpatient Service Efficiency Based on BOR, BTO, TOI, AvLOS Indicators using Agglomerative Hierarchical Clustering," *Telematika*, vol. 18, no. 2, p. 194, 2021, doi: 10.31315/telematika.v18i2.4786.

[13] R. Rahmadhan and M. Wasesa, "Segmentation using Customers Lifetime Value: Hybrid K-means Clustering and Analytic Hierarchy Process," *J. Inf. Syst. Eng. Bus. Intell.*, vol. 8, no. 2, pp. 130–141, 2022, doi: 10.20473/jisebi.8.2.130-141.

[14] K. Maulida Hindrayani, A. Anjani, and A. Lina Nurlaili, "Penerapan Machine Learning pada Penjualan Produk UMKM : Studi Literatur," *Pros. Semin. Nas. Sains Data*, vol. 1, no. 01, pp. 19–23, 2021, doi: 10.33005/senada.v1i01.7.

[15] R. Siagian, P. S. Pahala Sirait, and A. Halima, "E-Commerce Customer Segmentation Using K-Means Algorithm and Length, Recency, Frequency, Monetary Model," *J. Informatics Telecommun. Eng.*, vol. 5, no. 1, pp. 21–30,

2021, doi: 10.31289/jite.v5i1.5182.
[16] S. I. Murpratiwi, I. G. Agung Indrawan, and A. Aranta, "Analisis Pemilihan Cluster Optimal Dalam Segmentasi Pelanggan Toko Retail," *J. Pendidik. Teknol. dan Kejuru.*, vol. 18, no. 2, p. 152, 2021, doi: 10.23887/jptk-undiksha.v18i2.37426.
[17] R. Adhitama, A. Burhanuddin, and A. Febriani, "Penerapan X Means Clustering Pada UMKM Kab Banyumas Yang Mendukung Mega Shifting Consumer Behavior Akibat Covid-19," *J. Informatics, Inf. Syst. Softw. Eng. Appl.*, vol. 4, no. 1, pp. 71–80, 2022, doi: 10.20895/inista.v4i1.429.
[18] S. Renaldi. S, D. A. Prasetya, and A. Muhaimin, "Analisis Klaster Partitioning Around Medoids dengan Gower Distance untuk Rekomendasi Indekos (Studi Kasus: Indekos di Sekitar Kampus UPNVJT)," *G-Tech J. Teknol. Terap.*, vol. 8, no. 3, pp. 2060–2069, 2024, doi: 10.33379/gtech.v8i3.4898.
[19] N. R. Handitia and A. Sofro, "Analisis Klaster Data Pendidikan Kota Surabaya Tahun 2022-2023," *MATHunesa J. Ilm. Mat.*, vol. 13, no. 1, pp. 96–104, 2025, doi: 10.26740/mathunesa.v13n1.p96-104.
[20] E. Nurelasari, "Segmentasi Dan Klasifikasi Perilaku Pembayaran Pelanggan Pada," vol. XXI, no. 1, pp. 69–76, 2019, doi: 10.31294/p.v20i2.
[21] J. B. Sembiring, H. Manurung, and A. Sihombing, "Pengelompokan Data Tunggakan Pembayaran Kredit Mobil Menggunakan Metode Clustering (Studi Kasus: Cv Citra Kencana Mobil)," *J. Manaj. Inform. Jayakarta*, vol. 3, no. July, pp. 275–291, 2023, [Online]. Available: http://journal.stmikjayakarta.ac.id/index.php/JMIJayakarta
[22] H. Ramadhan, M. R. Abdan Kamaludin, M. A. Nasrullah, and D. Rolliawati, "Comparison of Hierarchical, K-Means and DBSCAN Clustering Methods for Credit Card Customer Segmentation Analysis Based on Expenditure Level," *J. Appl. Informatics Comput.*, vol. 7, no. 2, pp. 246–251, 2023, doi: 10.30871/jaic.v7i2.5790.
[23] M. Vahidi Farashah, A. Etebarian, R. Azmi, and R. Ebrahimzadeh Dastjerdi, "An analytics model for TelecoVAS customers' basket clustering using ensemble learning approach," *J. Big Data*, vol. 8, no. 1, pp. 1–24, 2021, doi: 10.1186/s40537-021-00421-1.
[24] S. Yucebas, Ş. Yalpir, L. Genc, and M. Dogan, "Price Prediction and Determination of the Affecting Variables of the Real Estate by Using X-Means Clustering and CART Decision Trees," *JUCS - J. Univers. Comput. Sci.*, vol. 30, pp. 531–560, 2024, doi: 10.3897/jucs.98733.
[25] K. Tabianan, S. Velu, and V. Ravi, "K-Means Clustering Approach for Intelligent Customer Segmentation Using Customer Purchase Behavior Data," *Sustain.*, vol. 14, no. 12, pp. 1–15, 2022, doi: 10.3390/su14127243.
[26] D. A. Prasetya, A. P. Sari, P. A. Riyantoko, and T. M. Fahrudin, "The Effect of Information Quality and Service Quality on User Satisfaction of the Government of Kabupaten Malang," *TIERS Inf. Technol. J.*, vol. 4, no. 1, pp. 32–42, 2023, doi: 10.38043/tiers.v4i1.4328.
[27] P. A. Riyantoko, K. M. Hindrayani, T. M. Fahrudin, and E. M. Safitri, "Southeast Asia Happiness Report in 2020 Using Exploratory Data Analysis," *Netw. Secur. Inf. Syst.*, vol. 2, no. 1, pp. 16–21, 2020.
[28] M. Rizqi Sulistio, N. Suarna, and O. Nurdiawan, "Analisa Penerapan Metode Clustering X-Means Dalam Pengelompokan Penjualan Barang," *J. Teknol. Ilmu Komput.*, vol. 1, no. 2, pp. 37–42, 2023, doi: 10.56854/jtik.v1i2.49.
[29] E. Melik-Gaykazyan *et al.*, "From Fano to Quasi-BIC Resonances in Individual Dielectric Nanoantennas," *Nano Lett.*, vol. 21, no. 4, pp. 1765–1771, Feb. 2021, doi: 10.1021/acs.nanolett.0c04660.
[30] R. C. Tarumingkeng, "Bayesian Information Criterion (BIC)," *Proc. Natl. Acad. Sci.*, vol. 3, no. 1, pp. 1–15, 2015, [Online]. Available: http://dx.doi.org/10.1016/j.bpj.2015.06.056%0Ahttps://academic.oup.com/bioinformatics/article-abstract/34/13/2201/4852827%0Ainternal-pdf://semisupervised-3254828305/semisupervised.ppt%0Ahttp://dx.doi.org/10.1016/j.str.2013.02.005%0Ahttp://dx.doi.org/10.10
[31] R. Kurniawan *et al.*, "Optimizing the Identification of Suitable Congregations for Preachers Using a GMM-PCA-BIC Hybrid Clustering Approach," in *2024 7th International Conference of Computer and Informatics Engineering (IC2IE)*, 2024, pp. 1–6. doi: 10.1109/IC2IE63342.2024.10748223.
[32] T. T. Hmwe, N. Y. T. Thein, and K. M. Cho, "Improving clustering quality using silhouette score," *J. Comput. Appl. Res*, vol. 1, no. 1, pp. 58–62, 2020.
[33] K. R. Shahapure and C. Nicholas, "Cluster quality analysis using silhouette score," *Proc. - 2020 IEEE 7th Int. Conf. Data Sci. Adv. Anal. DSAA 2020*, pp. 747–748, 2020, doi: 10.1109/DSAA49011.2020.00096.