International Journal of Advances in Data and Information Systems
Vol. 6, No. 2, August 2025, pp. 259~275
ISSN: 2721-3056, DOI: 10.59395/ijadis.v6i2.1381

Customer Transaction Clustering with K-Prototype Algorithm Using Euclidean-Hamming Distance and Elbow Method

Dendy Arizki Kuswardana[1], Dwi Arman Prasetya[2*], Trimono[3], I Gede Susrama Mas Diyasa[4], Wan Suryani Wan Awang[5]
[1,3]Department of Data Science, University of Pembangunan Nasional Veteran Jawa Timur, Indonesia
[2,4]Department Master of Information Technology, University of Pembangunan Nasional Veteran Jawa Timur Indonesia
[5]Faculty of Informatics and Computing, University Sultan Zainal Abidin Besut Campus, Malaysia

Article Info

Article history:

Received May 15, 2025
Revised Jun 06, 2025
Accepted Jun 30, 2025

Keywords:

K-Prototype
Euclidean
Hamming
Elbow
Clustering

ABSTRACT

This study aims to cluster customer transactions in a Japanese food stall using the K-Prototype Algorithm with a combination of Euclidean-Hamming Distance and the Elbow method. Facing intense industry competition, this study seeks to understand customer purchasing behavior to increase loyalty and sales. From 9.721 initial entries, 9.705 cleaned and transformed records were analyzed. K-Prototype was chosen because of its ability to handle numeric features (Total Sales, Product Quantity) and categorical features (Payment Method, Order Type, Day Category and Time Category). The combination of Euclidean-Hamming distances was used for distance measurement. The optimal number of clusters was determined using the Elbow method, with the results recommending three clusters as the most optimal number. A Silhouette score of 0.6191 indicates a Good Structure clustering result, effectively identifying three distinct customer grouping: "Loyal Regulars" (49.5%), "Casual Shoppers" (42.3%), and "Premium Shoppers" (8.2%). Statistical validity was also tested using ANOVA and Chi-Square, the results showed significant differences between the clusters in numerical and categorical variables with a p-value <0.0001. The clusters are statistically valid in both numerical and categorical aspects. These insights provide an understanding of customer characteristics and reveal a strategically valuable cluster for targeted marketing.

Corresponding Author:

Dwi Arman Prasetya
Department Master of Information Technology,
University of Pembangunan Nasional Veteran Jawa Timur
Jl. Rungkut Madya, Gn. Anyar, Kec. Gn. Anyar, Surabaya, Jawa Timur 60294
Email: arman.prasetya.sada@upnjatim.ac.id

1. INTRODUCTION

The culinary industry continues to experience competitive dynamics with an increasing number of new competitors offering a variety of interesting innovations. Competition is not only limited to the uniqueness of the food menu but also extends to creative marketing strategies, services that focus on customer satisfaction, and competitive pricing [1]. Amid changing and evolving consumer needs and preferences, businesses in this sector are required to make various adjustments and innovations to stay relevant [2]. This emphasizes the importance of deeply understanding market characteristics and consumer behavior to create advantages that can support business continuity amid increasingly dynamic industry challenges [3]. The number of culinary

entrepreneurs in Surabaya City continues to increase every year, with data showing growth from 6.32% in 2013 to 6.43% in 2014, then increasing to 6.64% in 2015 [4]. One of the culinary industry segments experiencing rapid growth is Japanese cuisine, with an estimated annual growth of 10% to 15% [5]. This phenomenon reflects the high interest of the people of Surabaya in the taste and uniqueness of Japanese cuisine, so more and more Japanese culinary businesses are popping up in the city of Surabaya [5]. However, the increase in the number of competitors in the market has also triggered increasingly fierce competition. XYZ eatery faces the challenge of maintaining customer loyalty in the midst of this competition by deeply understanding customer purchasing behavior [6]. Relying on the use of traditional, non-data-driven approaches is considered ineffective, requiring innovative strategies to remain competitive and relevant in the market [7].

In facing this challenge, XYZ eatery strives to further strengthen its position in market competition through a data-driven strategy. Through in-depth analysis of customer purchasing behavior and improving services that are more in line with buyer needs [8]. Through customer grouping analysis, it can understand the preferences of each customer grouping, design the right marketing strategy, and offer more relevant products [9]. This helps increase customer loyalty and strengthen competitiveness amid the rapid growth of Japanese cuisine in the city of Surabaya.

This study focuses on grouping customers based on purchasing behavior using the K-Prototype Algorithm, which handles numerical and categorical data simultaneously. The purchase transaction dataset includes numerical data such as total sales and product quantity, as well as categorical data such as order type, payment method, month, and time of purchase. The process begins with data collection and pre-processing, including cleaning, feature engineering, distance calculation, determining clusters, modelling k-prototype, evaluation model, and statistical validation. The distance between customers is measured using Euclidean Distance, which is capable of calculating distances for numerical data, while utilizing Hamming Distance to process categorical data effectively [10], [11]. The Elbow method identifies the optimal number of clusters for effective grouping. Using K-Prototype modeling, the dataset is clustered into three customer groups based on numerical and categorical attributes. An evaluation is performed with the Silhouette Coefficient to assess the quality of the grouping.

Facing difficulties in achieving sales targets every weekend, even though this period should be the peak of sales that allows for the achievement of expected targets. The available customer purchase data has never been used in the analysis of customer purchasing behavior. Understanding customer purchasing behavior is important for designing targeted marketing strategies [12]. This approach is expected to increase customer loyalty and help businesses compete better in the increasingly competitive Japanese culinary market. This study categorizes customers based on purchasing patterns using the K-Prototype Algorithm. The results help design targeted marketing strategies, increase promotion efficiency, and optimize sales and services [13]. By understanding the preferences of each customer group, XYZ eatery can maintain customer loyalty and increase profitability.

In previous research conducted by [8], hostel customers were grouped using the K-Prototypes Algorithm based on score, price, and location to customize promotions and recommendations. However, this study lacks a thorough explanation of the approach employed to determine the optimal number of clusters, as well as the method used to calculate the distance between attributes. It has not been tested on large datasets and without evaluation metrics such as the Silhouette Score. The innovations offered include the use of the Elbow method for cluster optimization, a combination of Euclidean and Hamming Distance for grouping accuracy, and testing on large datasets with evaluation using the Silhouette Score.

Subsequent research by [14], helped Genta Corp. group products by sales with the K-Means Algorithm, benefitting stock and marketing management. Nevertheless, this study is limited to numerical data, lacking a systematic approach to determining the optimal number of clusters and failing to conduct a comprehensive analysis of customer behavior patterns. An applicable innovation is the use of mixed data clustering with K-Prototypes, optimizing with the Elbow Method, and refining using Euclidean-Hamming distance to improve the accuracy of customer

grouping. Another study was also conducted by [15], applying K-Means for customer segmentation based on rice sales transactions in traditional markets in Tegal City. The results show the effectiveness of clustering in identifying customers and supporting marketing strategies, with C4 SUPER as the most popular rice. However, the limitation is that the dataset is not very varied, covering only three stores. Innovations that can be made include using more diverse data and measuring distance with Euclidean and Hamming distance. The fierce competition in the culinary industry demands a deep understanding of customer behavior to maintain competitiveness. With the increasing number of Japanese culinary businesses in Surabaya, data-driven strategies are important in optimizing marketing and increasing customer loyalty. The traditional, non-data-driven approach is considered less effective, so purchasing pattern analysis is a more appropriate solution.

This study is different from previous studies that focus more on numerical data, this study offers innovation in the use of mixed data types from the food and beverage business sector, measuring distance by combining the rarely used euclidean-hamming distance method, determining the optimal number of clusters using the more systematic elbow method, and using the k-prototype algorithm as a cluster model that can handle both numerical and categorical data. The results of this study are expected to provide a deeper understanding of the characteristics of customers in each cluster, offering insight into their purchasing behavior and preferences amidst increasingly fierce competition in the Japanese culinary industry.

2. RESEARCH METHOD

This study uses a stepwise method to process data before clustering with K-Prototypes. Figure 1. below shows the process flow, starting from data collection to model evaluation, to ensure optimal clustering results. Customer sales transactions during the period January to September 2024, grouping these customers based on the purchasing behavior of each customer, This algorithm, initially introduced by [16], it is designed to efficiently process both numerical and categorical data by utilizing the K-Means cost function for numerical attributes and K-Modes for categorical attributes while incorporating a balancing parameter gamma (γ) to optimize the clustering process for mixed data [17], [18]. The research process, as can be seen in Figure 1, begins with data collecting, data pre-processing, feature engineering, distance calculation, cluster determination, modelling k-prototype, model evaluation model, and statistical validation.

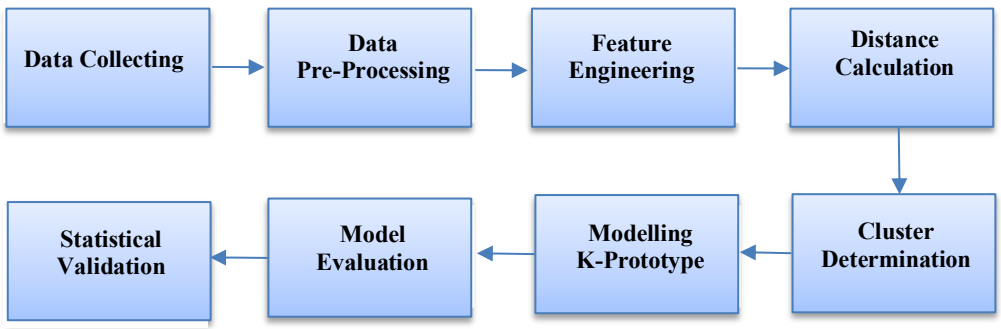

Figure 1. Research Flowchart

Data Collecting

Customer transaction data is obtained from the Point of Sale (POS) system in Excel format. This data includes automatically recorded purchase histories, ensuring completeness of information for further analysis. The data has various features, as shown in Table 1 below.

Table 1. Customer Dataset Features

No.	Features	Description
1.	Order Time	The moment the order was made.
2.	Order Type	The category of the order (e.g., dine-in, takeaway, delivery).
3.	Product	The specific item purchased in the order.
4	Total Sales	The total revenue generated from the order.
5	Payment Method	The method used for payment (e.g., cash, credit card, e-wallet).

Data Pre-Processing

The pre-processing stage is important to do before analysis to prevent bias and ensure accurate and reliable clustering [19]. This will start with retrieving transaction data from the POS system in Excel format, followed by exploring and checking for missing values, duplication, and data types. Transformation is carried out by the Item Counting method in the Product column to calculate the product quantity purchased and save it in the Product Quantity column. In categorical data, Time Extraction and transformation are applied to Order Time to produce the Date, Month, Hour, Day Category, and Time Category columns. After the transformation, Feature Selection, checking for missing values, and data type adjustments are carried out so that it is ready for analysis with K-Prototypes. Table 2. shows the features of the entire dataset that has gone through the pre-processing stage and can be used for feature engineering and modeling using K-Prototype.

Table 2. Customer Dataset Features After Preprocessing

No.	Feature	Description
1.	Order Type	Indicates the type of order, such as dine-in, take-away, or delivery
2.	Total Sales	The total revenue from each transaction in a specific currency.
3	Payment Method	The method of payment used, such as cash, credit card, or e-wallet.
4.	Month	The month when the transaction occurred. generated from the order.
5.	Day Category	Classification of the day, such as a weekday or weekend.
6	Product Quantity	The number of items purchased in a single transaction.
7	Time Category	The time classification of the transaction, such as: afternoon, evening, night

Feature Engineering

After preprocessing, the cleaned data consists of numerical (Total Sales, Product Quantity) and categorical (Order Type, Payment Method, Month, Day Category, Time Category) attributes. Feature engineering is performed separately for each data type. Numerical attributes are processed using Euclidean Distance, while categorical attributes undergo Label Encoding before being measured with Hamming Distance. This approach ensures an accurate representation of both data types in the clustering process. This method ensures precise distance measurement, optimizing data representation before applying the clustering process with the K-Prototypes [20].

Distance Calculation

The prepared data is then used to calculate the level of similarity using euclidean distance for numeric attributes with the following formula [10]:

$$d_{ij} = \sqrt{\sum_{k=1}^{n}(x_{ik} - x_{jk})^2} \quad (1)$$

Using d_{ij}, the similarity distance between the feature vectors of the input image (x_{ik}) (x_{ik}) and the comparison image (x_{jk}) is calculated, where nnn denotes the total number of vector elements. This ensures a thorough and accurate measurement of similarity. This calculation is used to analyze the similarity of data in applications such as pattern recognition and classification. Meanwhile, the hamming distance for categorical attributes is calculated with the following formula [8]:

$$H_1(x, y) = \sum_{i=1}^{n} w_i H(x_i, y_i) \quad (2)$$

In this formula, represents the weight or adjustment cost assigned to the *i*-th component. The *H* function serves as a measure of the hamming distance between the actual value of the i-th

component and its corresponding reference value. The formulation of this function is expressed as follows:

$$H(x,y) = \begin{cases} 0, if\ x_i = y_i \\ 1, if\ x_i \not\equiv y_i \end{cases} \quad (3)$$

Elbow Method

The process of identifying the optimal cluster count is carried out using the elbow method, which evaluates multiple cluster variations and selects the most suitable one based on the calculation results, which are displayed in a graph shaped like an "elbow". In Figure 2., is the calculation stage using the elbow method.

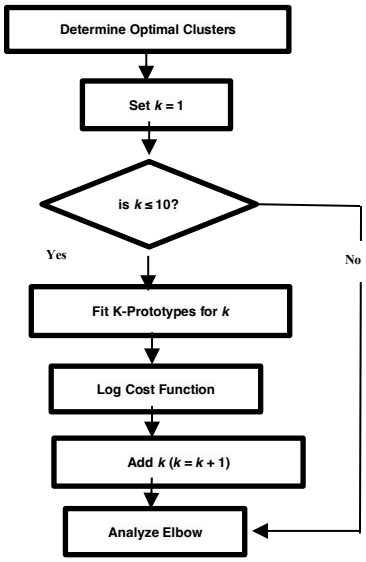

Figure 2. Elbow Method

The best cluster is determined when there is a significant drop or sharp "angle" on the graph between the two cluster values [20]. The Sum of Squared Errors (SSE) measures clustering accuracy by calculating the total variance of data points from their cluster centers. A lower SSE value indicates more compact clusters, signifying a better clustering performance [19]. Through the Sum of Squared Errors (SSE) value using the following formula [21]:

$$SSE = \sum_{K=l}^{k} \sum_{x_l \in S_K}^{n} ||(X_i - C_k|\,|_2^2 \quad (4)$$

The SSE formula determines the overall sum of squared differences between data points X_i and the centroid of its respective cluster C_k. Their respective cluster centers, providing a measure of clustering compactness and accuracy. For each cluster *k,* the Euclidean distance is calculated between X_i in cluster S_K and C_K, then summed. A lower SSE value indicates better cluster separation. Clustering is then performed with the K-Prototype Algorithm, which uses distance measurements from The Euclidean distance is utilized for numerical data, whereas the Hamming distance is employed for categorical data. Using the following formula [16]:

$$d(i,j) = \sum_{k=1}^{P}(x_{ik} - x_{jk})^2 + \gamma \sum_{i=1}^{m} \delta(x_{is}, x_{js}) \quad (5)$$

Euclidean Distance is employed to quantify similarity in numerical data, while Hamming Distance is utilized to assess differences in categorical data [20]. Numerical distance is calculated by $(x_{ik} - x_{jk})^2$ while categorical distance using $\delta(x_{is}, x_{js})$ is zero if the same and one if different.

These two distances are combined with the γ parameter to balance the contribution of each data type.

Modelling K-Prototype

The modeling stage shown in Figure 3., uses the K-Prototypes Algorithm to group customer transactions based on numerical and categorical characteristics. Data distance is measured using Euclidean for numerical and Hamming for categorical attributes. The clustering results are stored in the Cluster column, and their distribution is analyzed to support more effective marketing strategies.

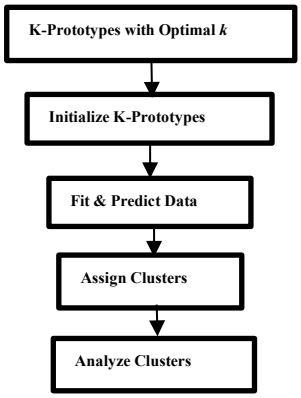

Figure 3. Modelling K-Prototype

K-Prototype is a clustering method that combines K-Means to handle numerical data and K-Modes to process categorical data [22], [23]. This research applies the K-Prototypes Algorithm, which combines distance measurement results with the Euclidean method for numerical data and Hamming for categorical data to measure the proximity between data [17]. The total distance calculation is done by summing the two metrics, where the gamma weighting coefficient (γ) is used to adjust the contribution of each data type. This approach, introduced by [24], enables a more optimal clustering process for mixed data.

This equation calculates the mixture distance by combining the Euclidean distance for numerical data and the Hamming distance for categorical data in the K-Prototypes algorithm [16]:

$$\left(X_j, Z_i\right) = \left(\sum_{l=1}^{m_r}(x_{jl}^r - z_{il}^r)^2 + \gamma_i \sum_{l=l+1}^{m_c} \delta(x_{jl}^c, z_{il}^c\right)^{\frac{1}{2}} \quad (6)$$

The K-Prototypes algorithm combines K-Means for numerical data and K-Modes for categorical data to group mixed data effectively. The process includes [16]:

a. Determine the number of clusters (k).
b. Initialize the cluster center.
c. Calculate distances using Euclidean (numeric) and Hamming (categorical).
d. Group objects based on closest distance.
e. Update the cluster center with the average (numeric) and mode (categorical).
f. Repeat until the cluster is stable or the maximum iteration is reached.

Evaluation Model

Following the clustering process using the K-Prototypes algorithm, the model was evaluated using the Silhouette Score to measure the quality of the formed clusters. Euclidean distance was employed to quantify the similarity between numerical features, while Hamming distance was applied to measure differences in categorical attributes, ensuring an appropriate distance calculation for each data type. These two distances are combined in a dissimilarity matrix as input for the Silhouette Score calculation. The resulting value indicates how well objects in one

cluster are grouped compared to other clusters, where a higher score indicates a more optimal grouping.

The model is evaluated using the Silhouette Coefficient, which assesses clustering quality by comparing the average distance of a data point within its own cluster (a_i) to the nearest cluster (b_i). A higher score indicates well-separated clusters, whereas a lower score indicates poor separation, as follows [25]:

$$SW_i = \frac{b_i - a_i}{\max\{a_i, b_i\}} \quad (7)$$

This value ranges from -1 to 1, where a value near 1 signifies good clustering. A value of 0 indicates that the data is at the boundary between two clusters. In contrast, a negative value indicates that the data has a stronger similarity with another cluster compared to its current cluster. Table 3. shows the categories of model evaluation using the Silhouette Coefficient to indicate the standard of the modeling results [26].

Table 3. Silhouette Score Categories

Group	Score	Standard
1	0,71 – 1,00	Solid Structure
2	0,51 – 0,70	Good Structure
3	0,26 – 0,50	Weak Structure
4	≤ 0,25	Bad Structure

Statistical Validation

In evaluating the significance of differences in characteristics between clusters, this study uses the Analysis of Variance (ANOVA) test for numerical data and the Chi-Square test for categorical data. ANOVA compares the means of several independent groups by calculating the ratio of variation between and within groups to assess the relevance of features to clusters. This ratio is expressed using the following formula [27]:

$$F(\lambda) = \frac{S_B^2}{S_B^2}(\lambda) \quad (8)$$

Meanwhile, Chi-square feature selection is a common approach used to select features by ranking them based on the Chi-square statistic, from the highest to the lowest value. The Chi-Square test is used to identify whether there is a statistically significant relationship between two categorical variables and also to measure the strength of that relationship. This test compares the observed frequencies with the expected frequencies under the assumption of independence between variables. The greater the deviation between these values, the more likely the relationship is significant, using the following formula [27]:

$$X^2 \sum_{i=1}^{k} \frac{(f_0 - f_2)^2}{f_h} \quad (9)$$

In both tests, a p-value < 0.05 is generally used as the significance limit to conclude that the difference or relationship between groups is statistically significant [28]. In this analysis, testing of the null hypothesis (H_0) and alternative (H_1) is used to determine whether there is a significant difference or relationship between variables. Table 4. contains the type of test, type of variable, and the content of each hypothesis used.

Table 4. ANOVA and Chi-Square Test Hypothesis

Types of Statistical Tests	(H_0)	(H_1)
ANOVA	There were no significant mean differences between clusters.	There are significant mean differences between clusters.
Chi-Square	There is no significant relationship (association) between variables and clusters.	There is a significant relationship (association) between variables and clusters.

3. RESULTS AND DISCUSSION

The initial dataset was selected to focus on the required column variables, namely Order Type, Total Sales, Payment Method, Month, Day Category, Product Quantity, and Time Category. This selection ensures that the data is in line with the research objectives and supports the analysis of customer purchasing patterns.

3.1. Prepare the Data Used

Data preprocessing is conducted to prepare the dataset for analysis and processing. Handling missing values, three missing values were found in Products, and 14 in Payment Methods, so the amount of data was reduced from 9.721 to 9.705 rows. Duplication checking shows no duplicate data. In the data transformation stage, several changes were made to facilitate further analysis. The Product Quantity column is processed using the item counting method to calculate the number of items purchased in each Transaction. Meanwhile, the Order Time column is broken down into several new attributes, namely Date, Month, Hour, and Day, in order to extract more detailed transaction time information. The extracted data is then categorized into several groups, namely Order Type (0 = Table, 1 = Non-Table), Payment Method (0 = Cash, 1 = Non-Cash), Day Category (0 = Weekday, 1 = Weekend), and Time Category (0 = 16:00-17:59, 1 = 18:00-19:59, 2 = 20:00-23:00). In addition, unnecessary columns such as Transaction No., Order Time, Date, Product, Day, and Hour are removed so that the analysis focuses more on relevant information.

Re-checking the missing value ensures that no data is lost. Data type adjustments are made by converting Order Type, Payment Method, Month, Day Category, and Time Category from int64 to category. The dataset is ready for the modeling stage, as shown in Table 5.

Table 5. Dataset for Modeling

No.	Order Type	Total Sales	Payment Method	Month	Day Category	Products Quantity	Time Category
1.	0	187000	1	1	0	8	0
2.	1	25000	0	1	0	2	0
3.	0	49000	1	1	0	2	0
4.	0	99000	0	1	0	5	0
5.	0	117000	1	1	0	7	0
...	...	...	...	...	...	...	...
9.720	1	8000	0	9	1	1	2

3.2. Distance Measurement Using Euclidean-Hamming

The measurement of distance is conducted to evaluate the level of dissimilarity between data points, which is essential in the clustering process as it determines how data is grouped based on their similarities and differences. The data is separated by type, where Total Sales and product quantity are categorized as numerical data. In contrast, Order Type, Payment Method, Month, Day Category, and Time Category are categorized as categorical data.

For numerical data, Euclidean Distance method calculates the similarity between two points by finding the square root of the sum of squared differences for each attribute, as shown in Figure 4. part (a) The calculation results show that some data pairs have large distance values, such as 138000.0013043 and 179000.0013687, indicating a significant difference in the numerical scale. Meanwhile, there are smaller distance values, such as 1700.0002941 and 24000, which indicate a higher degree of similarity between data. A lower Euclidean Distance value indicates greater similarity between two data points, while a higher value reflects more significant differences in their numerical attributes.

```
Euclidean Distances (Numerik):
[[     0.         162000.00011111 138000.00013043 ...  98000.00002041
  110000.00004091 179000.00013687]
 [162000.00011111      0.          24000.         ...  64000.000125
   52000.00008654  17000.00002941]
 [138000.00013043  24000.              0.         ...  40000.0002
   28000.00016071  41000.0000122 ]
 ...
 [ 98000.00002041  64000.000125    40000.0002     ...      0.
   12000.00004167  81000.00015432]
 [110000.00004091  52000.00008654  28000.00016071 ...  12000.00004167
       0.          69000.00011594]
 [179000.00013687  17000.00002941  41000.0000122  ...  81000.00015432
   69000.00011594      0.        ]]
```

(a)

```
Hamming Distances (Kategorik):
[[0.  0.4 0.  ... 0.8 0.8 1. ]
 [0.4 0.  0.4 ... 0.8 0.8 0.6]
 [0.  0.4 0.  ... 0.8 0.8 1. ]
 ...
 [0.8 0.8 0.8 ... 0.  0.  0.2]
 [0.8 0.8 0.8 ... 0.  0.  0.2]
 [1.  0.6 1.  ... 0.2 0.2 0. ]]
```

(b)

Figure 4. Distance Measurement
(a) Euclidean Distance, (b) Hamming Distance

Meanwhile, categorical data is measured using Hamming Distance, as shown in Figure 4. part (b) which calculates the number of category differences between attributes. If the attributes are different, the distance is 1; if they are the same, the distance is 0. The calculation results show variations in distance values, such as 0.8, which indicates that 80% of the attributes between two data are different, and a value of 0.6, which indicates a difference of 60%. Conversely, a value of 0 indicates that the two data have identical attributes.

3.3. Cluster Determination Using the Elbow Method

Determine the optimal number of clusters in the process of grouping customer transactions using the K-Prototypes Algorithm to improve the accuracy and effectiveness of data clustering. Euclidean Distance is utilized to assess the similarity between numerical data points, whereas Hamming Distance is employed to assess differences in categorical attributes, and then the two are combined after the category is converted with Label Encoding. The number of clusters is tested from 1 to 10, and the cost function is recorded at each iteration. The Elbow Graph is used to identify the elbow point, which indicates the optimal number of clusters where the reduction in the cost function begins to stabilize. Identifying this point enhances the accuracy of clustering, leading to more meaningful insights into distinct groups.

The Elbow Method is used to determine the optimal number of clusters by identifying the point where the curve exhibits a significant bend, demonstrating a balance between the number of clusters and the variance within each cluster. The clustering process begins by computing distances, where Euclidean Distance is applied to numerical attributes and Hamming Distance is used for categorical attributes to achieve precise data grouping. These calculation results are then combined to ensure more accurate data grouping.

Furthermore, the selection of the number of clusters was tested with a variation of 1 to 10 clusters in Figure 5., where the cost function value of each number of clusters was calculated to see the change in the level of variance in the cluster. The Elbow Method graph obtained shows a sharp decrease up to cluster 3, then slopes down afterward. This suggests that the elbow point is at cluster 3, which indicates the optimal number of clusters for data clustering. The selection of this number of clusters aims to maintain a balance between variation in the cluster and model complexity so that it remains optimal in the further analysis process.

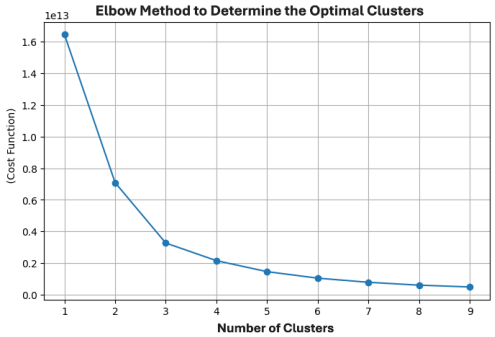

Figure 5. Cluster Determination Using Elbow

3.4. Modelling K-Prototypes Algorithm

The K-Prototypes algorithm clusters numerical and categorical data simultaneously. To ensure stable and interpretable results, the model is initialized using the Huang method, with random_state=42 for reproducibility and max_iter=50 to promote optimal convergence. The elbow method is employed to determine the optimal number of clusters, after which the model is applied to the combined_data dataset by specifying categorical indices and assigning cluster labels.

In this study, the gamma(γ) parameter in K-Prototypes plays an important role in balancing the influence of numerical and categorical features. Sensitivity analysis Figure 6. was performed with various experiments using gamma values ranging from 0.1, 0.3, 0.5, 0.7, 1.0, 1.5, 2.0 with a fixed number of clusters (k=3), evaluated using the silhouette score. Although the numerical score remains constant at 0.6191, the visual trend in Figure 6. shows a decrease in model stability as gamma increases. Therefore, a gamma value of 0.1 was chosen to achieve optimal balance and consistent clustering performance, based on quantitative evaluation and interpretive insights.

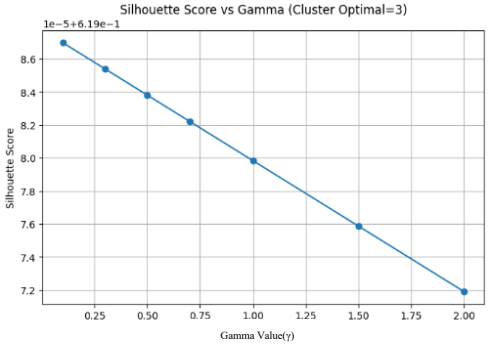

Figure 6. Best Gamma(γ) Parameters

The clustering results are stored in the main dataset as a new column named “Cluster”, and the cluster distribution is displayed for further analysis of customer patterns. Table 6. shows the total of each cluster successfully modeled using the K-Prototype algorithm.

Table 6. Cluster Customer Distribution

Cluster	Total of Each Cluster
0	4.806
1	796
2	4103

3.5. Evaluation Using Silhouette Score

Following the determination of the optimal number of clusters using the Elbow method, the clustering performance was evaluated through the Silhouette Score. This evaluation incorporated Euclidean Distance for numerical attributes and Hamming Distance for categorical attributes to enhance clustering accuracy and reliability. The distance between data is calculated in a dissimilarity matrix, which is used to measure the quality of clustering. The evaluation results show a Silhouette Score of 0.6191, as shown in Table 7., indicating a grouping that falls within the Good Structure level when viewed from Table 3, with a clear separation between clusters and relatively homogeneous data in each cluster.

Table 7. Evaluation Model

Aspect	Silhouette Score
Overall Average	0.6191
Cluster 0	0.6252
Cluster 1	0.4474
Cluster 2	0.6452

As shown in Table 6, Cluster 1 has a silhouette score of 0.4474, which is considered acceptable and reflects a moderate level of separation quality, albeit lower compared to the other clusters. In addition, the analysis of Cluster 1 characteristics shows significant differences compared to the other clusters, particularly in terms of total purchases, number of products, and transaction patterns.

3.6. Statistical Validation Using ANOVA and Chi-Square Tests

ANOVA and chi-square tests were used to ensure that the clusters formed truly represent significant differences between customer groups. This validation analysis was carried out by testing all variables of each type of data, both numerical and categorical by determining the hypothesis decision according to Table.4, to evaluate the consistency and relevance of customer groupings generated by the clustering algorithm.

The results of the ANOVA test on the numerical data in Table 8 show that there are significant differences between clusters. The Total Sales variable has an F-statistic of 19,688.09 and a p-value < 1e-300, indicating that the sales value differs significantly between clusters. Similarly, the Products Quantity variable, with an F-statistic of 5,501.69 and a p-value < 1e-300, suggests that the quantity of products purchased also varies significantly across clusters. These extremely small p-values (approaching zero) imply that the likelihood of these differences occurring by chance is almost negligible. Therefore, the clusters formed can be considered statistically valid in distinguishing numerical characteristics.

Table 8. ANOVA Test

Variables	F-statistic	P-Value	Interpretation
Total Sales	19688.0932	< 1e-300	H_0 is rejected, because p-value < 0.05
Product Quantity	5501.6900	< 1e-300	H_0 is rejected, because p-value < 0.05

Meanwhile, the Chi-Square test on categorical data in Table 9 also shows statistically significant results. The variables Order Type, Payment Method, Month, Day Category, and Time Category each yield p-values far below 0.05 specifically ranging from 5.03×10^{-172} to 2.53×10^{-10} indicating strong associations with the clusters. This means that the distribution of these categorical variables differs meaningfully across clusters. For instance, customer preferences for order type, timing of purchase, and payment method show distinct patterns within each cluster. These findings support that the clusters reflect not only differences in numerical data but also meaningful segmentation based on customer behavior and categorical attributes.

Table 9. Chi-Square Test

Variables	Chi-Square(X^2) Score	P-Value	Interpretation
Time Category	59.1980	$5.03\ x\ 10^{-172}$	P-value < 0.05, H_0 is rejected
Day Category	37.8428	$2.94\ x\ 10^{-12}$	P-value < 0.05, H_0 is rejected
Month	78.9935	$2.53\ x\ 10^{-10}$	P-value < 0.05, H_0 is rejected
Payment Method	53.1041	$6.06\ x\ 10^{-9}$	P-value < 0.05, H_0 is rejected
Order Type	788.8593	$4.28\ x\ 10^{-12}$	P-value < 0.05, H_0 is rejected

3.7. Understanding Numerical Data Visualization

The boxplot shown in Figure 7. shows the distribution of total sales in each cluster. Cluster 0 has a stable distribution with a small range. Cluster 1 shows high variation with many high sales values, indicating large transactions. Meanwhile, Cluster 2 has the lowest and most consistent total sales. These results indicate differences in purchasing patterns, with Cluster 1 having the greatest variation, while Cluster 2 tends to have small transactions.

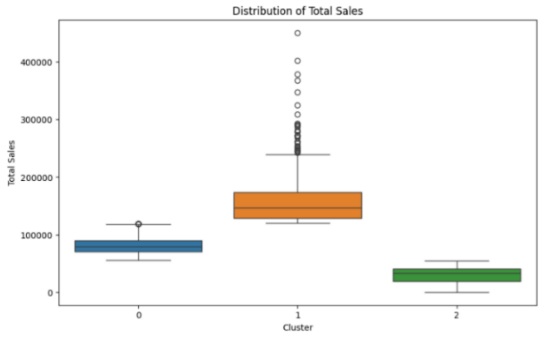

Figure 7. Boxplot of Total Sales

Furthermore, Figure 8. presents the boxplot depicting the distribution of the product quantity purchased within each cluster. Cluster 0 has a fairly stable distribution with a moderate range. Cluster 1 shows greater variation, with several high-volume transactions. Meanwhile, Cluster 2 shows the fewest products purchased, indicating that customers in Cluster 1 tend to buy more products than those in Clusters 0 and 2.

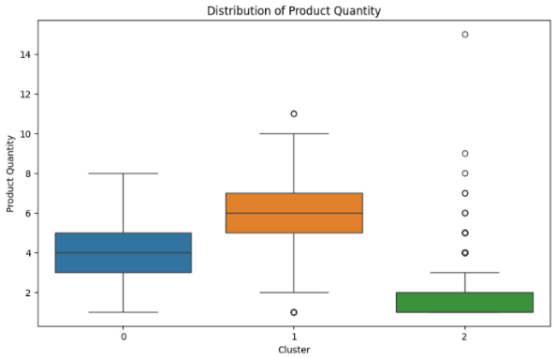

Figure 8. Boxplot of Product Quantity

3.8. Understanding Categorical Data Visualization

Figure 9. shows the distribution of payment methods (Cash = 0, Non-Cash = 1) in each cluster. Cluster 0 has the highest number of transactions using Non-Cash, with 2.852 transactions, compared to Cash, with 1.954 transactions. Cluster 1 has fewer transactions, with 282 transactions using Cash and 514 transactions using Non-Cash. Meanwhile, Cluster 2 has a fairly balanced distribution between Cash, with 1,919 transactions, and Non-Cash, with 2.184 transactions. This data shows that payment method preferences vary in each cluster, with Non-Cash being more dominant in Clusters 0 and 2. In contrast, Cluster 1 has a relatively smaller number of transactions for both payment methods.

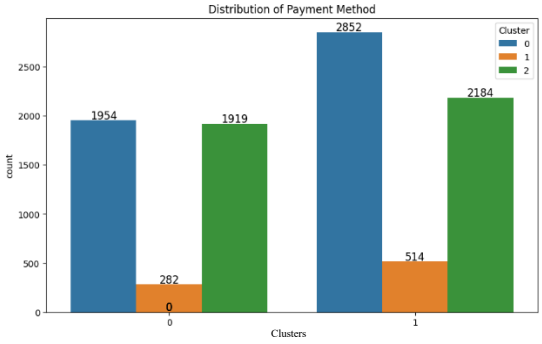

Figure 9. Bar Chart of Payment Method

Figure 10. shows the distribution of customer order types by cluster. Cluster 0 is more dominant with 3.728 Dine-in orders, while Take Away/Delivery is less with 1,078. Cluster 2 has an almost equal distribution between Dine-in with 2.053 and Take Away/Delivery with 2.050. Meanwhile, Cluster 1 has the least number of transactions, with 608 more Dine-in compared to 188 Take Away/Delivery. This data shows that order type preferences vary between clusters, with a higher tendency for Dine-in in some customer groups.

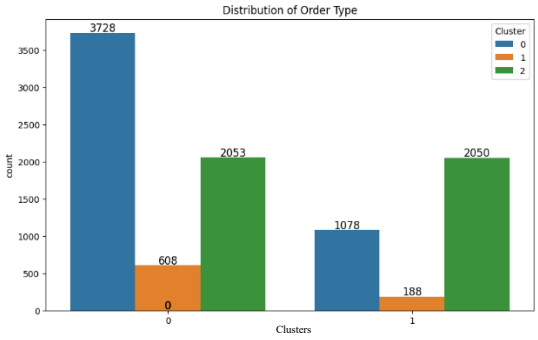

Figure 10. Bar Chart of Order Type

Figure 11. shows the distribution of the number of transactions per month based on customer clusters. Cluster 0 has the highest number of transactions, especially in May at 696 and April at 669, indicating an increase in purchasing activity during that period. Cluster 2 has a relatively stable trend, with a peak in May of 557 and August of 585. Meanwhile, Cluster 1 has far fewer transactions than the other two clusters, with a high of 126 in May. This pattern shows that customer buying patterns vary each month, with significant increases from April to August for some customer groups.

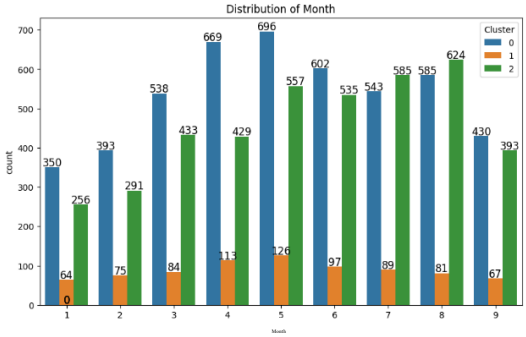

Figure 11. Bar Chart of Month

Figure 12. shows the distribution of purchase time categories based on customer clusters. Cluster 0 has the highest transaction count in the evening time category, specifically between 18:00 and 19:59, with a total of 2,598 transactions, indicating that the majority of customers in this cluster are more active in shopping in the afternoon. Cluster 2 also shows a similar pattern, with the same peak of 2.033 transactions in the 20:00-23:00 (Night) period. Meanwhile, Cluster 1 has far fewer transactions than the other two clusters in all-time categories, with the highest number in the Evening category at 473. In the 16:00-17:59 (Afternoon) category, the number of transactions is lower than in other categories in all clusters. This shows that the afternoon to evening period is the most dominant period for purchasing activities.

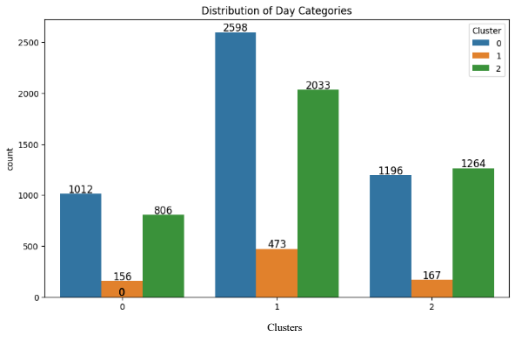

Figure 12. Bar Chart of Time Category

Figure 13. shows the distribution of purchase day categories based on customer clusters. In general, the number of transactions on Weekdays is higher than on Weekends in all clusters. Cluster 0 has the highest number of transactions on Weekdays with 3.046 transactions, while on Weekends it drops to 1,760 transactions. A similar pattern is seen in Cluster 2, where the number of transactions on Weekdays reaches 2.785, but decreases to 1,318 on Weekends. Meanwhile, Cluster 1 has fewer transactions than the other two clusters, with the highest number of transactions on Weekdays being 461 and decreasing to 345 on Weekends. From these results, it can be concluded that the majority of customers make purchases more often on Weekdays than on Weekends.

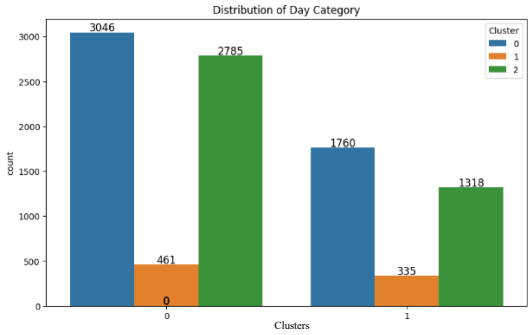

Figure 13. Bar Chart of Day Category

3.9. Customer Distribution in Each Cluster

Figure 14. shows a pie chart of the percentage distribution of all customers across the three clusters. Cluster 0 dominates with 49.5% (4,806 customers), indicating it represents the most prevalent purchasing patterns. Cluster 2 follows with 42.3% (4,103 customers), also reflecting a substantial portion with distinct yet common behaviors. In contrast, Cluster 1 comprises the smallest customer group, accounting for only 8.2% or 796 customers. Despite its smaller size, Cluster 1 should not be dismissed as an outlier or an over-clustered group. The silhouette score of 0.4474, though lower than other clusters, still falls within an acceptable range and indicates reasonable cohesion and separation. More importantly, Cluster 1 exhibits significantly different characteristics from the other clusters recording the highest average total purchase value of Rp157,988 and the greatest product quantity per transactions of 5 to 7 products. These findings highlight that Cluster 1 consists of high-value customers whose behavior, although less common, is highly strategic. Therefore, instead of being excluded, this cluster warrants focused attention through customer retention strategies and exclusive promotional efforts, as it holds strong potential for generating significant business value.

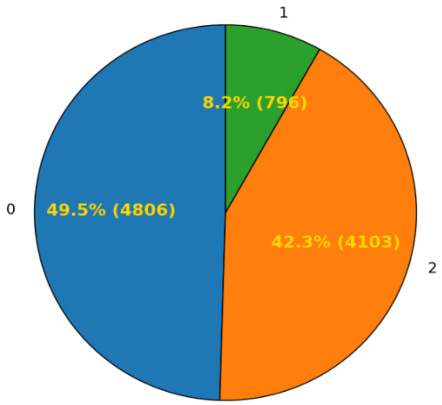

Figure 14. Cluster Percentage Distribution

4. CONCLUSION

Based on the analysis of customer grouping using clustering, it was possible to identify different purchasing patterns. The initial dataset of 9.721 rows was processed through selection, cleaning, and transformation, resulting in 9.705 rows of data ready for analysis. The Euclidean distance dissimilarity measurement is used for numerical variables, while for categorical variables, hamming distance is used, showing significant variations in attributes such as total sales, products quantity, and payment methods. The Elbow Method determines the results of three optimal clusters, with the Silhouette Score showing an evaluation result of 0.6191, indicating that the results of the grouping are within the Good Structure standard. Purchasing behavior in the three clusters can be summarized as follows:

1. Cluster 0 is referred to as "Loyal Regulars"
   Consisting of a stable purchasing pattern, high activity on weekdays, and non-cash preferences, it indicates loyal and regular customers.
2. Cluster 1 is referred to as "Premium Spenders"
   It has a high variation in total sales and product quantity, even though the transactions are few, it shows customers who tend to make large or premium purchases.
3. Cluster 2 is referred to as "Casual Shoppers"
   It has similarities to cluster 0 but with lower sales, indicating customers who are more relaxed and rarely shop in large quantities.

REFERENCES

[1] Z. Hussain, A. Albattat, F. Z. Fakir, and Z. Yi, Eds., *Innovative Trends Shaping Food Marketing and Consumption:* in Advances in Marketing, Customer Relationship Management, and E-Services. IGI Global, 2025. doi: 10.4018/979-8-3693-8542-5.

[2] K. M. Hindrayani and J. Timur, "Business Intelligence For Educational Institution : A Literature Review," vol. 2, no. 1, 2020, doi: https://doi.org/10.33005/ijconsist.v2i1.32.

[3] X. Liu, "The Role of Consumer Behavior in Shaping Market Demand and Economic Trends," *Int. J. Educ. Humanit.*, vol. 15, no. 2, pp. 10–16, Jul. 2024, doi: 10.54097/skmxzd63.

[4] S. Ardian and B. Syairudin, "Development strategy of culinary business employing the Blue Ocean Strategy (BOS)," *IPTEK J. Proc. Ser.*, vol. 0, no. 3, p. 153, Apr. 2018, doi: 10.12962/j23546026.y2018i3.3722.

[5] E. Amalijah and M. Fredy, "Pemetaan Restoran Jepang dan Kuliner Milenial di Surabaya," *J. Sakura Sastra Bhs. Kebud. Dan Pranata Jpn.*, vol. 5, no. 1, p. 169, Feb. 2023, doi: 10.24843/JS.2023.v05.i01.p10.

[6] Muh. R. Ramadhan and N. S. Fadjar, "ANALISIS PENGARUH PENDAPATAN, HARGA, PREFERENSI, DAN GAYA HIDUP TERHADAP PERILAKU KONSUMSI MAKANAN JEPANG (STUDI PADA MAHASISWA FEB UB)," *J. Dev. Econ. Soc. Stud.*, vol. 3, no. 3, pp. 780–789, Jul. 2024, doi: 10.21776/jdess.2024.03.3.09.

[7] A. V. B. S. Dhivya Devi G. Lakshmi, and S. B. Usman Ak Syed Shujauddin Sameer, "Data-Driven Decision-Making: Leveraging Analytics for Performance Improvement," *J. Inform. Educ. Res.*, vol. 4, no. 3, Aug. 2024, doi: 10.52783/jier.v4i3.1298.
[8] A. S. Girsang, "Clustering Hostels Data for Customer Preferences using K-Prototype Algorithm," *Int. J. Emerg. Trends Eng. Res.*, vol. 8, no. 6, pp. 2650–2653, Jun. 2020, doi: 10.30534/ijeter/2020/70862020.
[9] M. Idhom, A. M. Priananda, A. Raynaldi, R. Nur, S. A. Pamungkas, and A. C. Wardana, "UPAYA REBRANDING SEBAGAI BENTUK KEPEDULIAN TERHADAP UMKM," vol. 2, no. 4, doi: https://doi.org/10.56855/jcos.v2i4.1112.
[10] R. Suwanda, Z. Syahputra, and E. M. Zamzami, "Analysis of Euclidean Distance and Manhattan Distance in the K-Means Algorithm for Variations Number of Centroid K," *J. Phys. Conf. Ser.*, vol. 1566, no. 1, p. 012058, Jun. 2020, doi: 10.1088/1742-6596/1566/1/012058.
[11] J. Tayyebi and A. Deaconu, "Inverse Generalized Maximum Flow Problems," *Mathematics*, vol. 7, no. 10, p. 899, Sep. 2019, doi: 10.3390/math7100899.
[12] E. M. Sipayung, C. Fiarni, and R. Tanudjaya, "DECISION SUPPORT SYSTEM FOR POTENTIAL SALES AREA OF PRODUCT MARKETING USING CLASSIFICATION AND CLUSTERING METHODS," *Proceeding 8 Th Int. Semin. Ind. Eng. Manag.*, pp. 33–39, 2015.
[13] E. Muningsih and S. Kiswati, "SISTEM APLIKASI BERBASIS OPTIMASI METODE ELBOW UNTUK PENENTUAN CLUSTERING PELANGGAN," *Joutica*, vol. 3, no. 1, p. 117, Apr. 2018, doi: 10.30736/jti.v3i1.196.
[14] F. Indriyani and E. Irfiani, "Clustering Data Penjualan pada Toko Perlengkapan Outdoor Menggunakan Metode K-Means," *JUITA J. Inform.*, vol. 7, no. 2, p. 109, Nov. 2019, doi: 10.30595/juita.v7i2.5529.
[15] B. I. Nugroho, A. Rafhina, P. S. Ananda, and G. Gunawan, "Customer segmentation in sales transaction data using k-means clustering algorithm," *J. Intell. Decis. Support Syst. IDSS*, vol. 7, no. 2, pp. 130–136, Jun. 2024, doi: 10.35335/idss.v7i2.236.
[16] Z. Huang, "CLUSTERING LARGE DATA SETS WITH MIXED NUMERIC AND CATEGORICAL VALUES," *Proc. First Pac. Asia Knowl. Discov. Data Min. Conf. Singap. World Sci.*, pp. 21–34, 1997.
[17] S. S. M. Wara, "ANALISIS RESPONS WARGANET TERHADAP DEBAT CALON PRESIDEN 2019 DI TWITTER DENGAN METODE CLUSTERED SUPPORT VECTOR MACHINES," INSTITUT TEKNOLOGI SEPULUH NOPEMBER, 2019. [Online]. Available: https://repository.its.ac.id/64282/1/06211540000101_Undergraduate_Thesis.pdf
[18] D. A. Prasetya, P. T. Nguyen, R. Faizullin, I. Iswanto, and F. Armay, "Resolving the Shortest Path Problem using the Haversine Algorithm," *J. Crit. Rev.*, vol. 7, no. 1, 2020, doi: http://10.22159/jcr.07.01.11.
[19] P. A. Riyantoko, T. M. Fahrudin, D. A. Prasetya, T. Trimono, and T. D. Timur, "Analisis Sentimen Sederhana Menggunakan Algoritma LSTM dan BERT untuk Klasifikasi Data Spam dan Non-Spam," *Pros. Semin. Nas. SAINS DATA*, vol. 2, no. 1, pp. 103–111, Dec. 2022, doi: 10.33005/senada.v2i1.53.
[20] H. Hernández, E. Alberdi, A. Goti, and A. Oyarbide-Zubillaga, "Application of the k-Prototype Clustering Approach for the Definition of Geostatistical Estimation Domains," *Mathematics*, vol. 11, no. 3, p. 740, Feb. 2023, doi: 10.3390/math11030740.
[21] R. Nainggolan, R. Perangin-angin, E. Simarmata, and A. F. Tarigan, "Improved the Performance of the K-Means Cluster Using the Sum of Squared Error (SSE) optimized by using the Elbow Method," *J. Phys. Conf. Ser.*, vol. 1361, no. 1, p. 012015, Nov. 2019, doi: 10.1088/1742-6596/1361/1/012015.
[22] 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, Jul. 2024, doi: 10.33379/gtech.v8i3.4898.

[23] A. R. Adiwidyatma, I. G. S. Mas Diyasa, and T. Trimono, "ANALYSIS OF CLUSTERING METHODS ON THE CAUSAL FACTORS OF DIABETES MELLITUS WITH FUZZY C MEANS METHOD," *J. Lebesgue J. Ilm. Pendidik. Mat. Mat. Dan Stat.*, vol. 5, no. 2, pp. 983–996, Aug. 2024, doi: 10.46306/lb.v5i2.611.

[24] Z. Huang, "Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values," *Data Min. Knowl. Discov.*, 1998.

[25] H. Řezanková, "DIFFERENT APPROACHES TO THE SILHOUETTE COEFFICIENT CALCULATION IN CLUSTER EVALUATION," *21st Int. Sci. Conf. AMSE*, 2018.

[26] 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. Inform.*, vol. 7, no. 1, pp. 113–122, 2025, doi: https://doi.org/10.35882/ijeeemi.v7i1.55.

[27] T. A. Yoga Siswa, "Komparasi Optimasi Chi-Square, CFS, Information Gain dan ANOVA dalam Evaluasi Peningkatan Akurasi Algoritma Klasifikasi Data Performa Akademik Mahasiswa," *Inform. Mulawarman J. Ilm. Ilmu Komput.*, vol. 18, no. 1, p. 62, Feb. 2023, doi: 10.30872/jim.v18i1.11330.

[28] C. Andrade, "The P Value and Statistical Significance: Misunderstandings, Explanations, Challenges, and Alternatives," *Indian J. Psychol. Med.*, vol. 41, no. 3, pp. 210–215, May 2019, doi: 10.4103/IJPSYM.IJPSYM_193_19.