Efficiency Evaluation of K-Means, K-Medians, and K-Mode Clustering Methods Using SSD

Bounmy Phanthavong[1], Soulith Sengmanotham[2], Sommany Lusavong[3], Phonesouda Souphamith[4]

Department of Computer Science, Faculty of Natural Sciences, National University of Laos, Vientiane Capital, Lao P. D. R

Abstract

This study compares performance of three data clustering algorithms: K-Means, K-Medians, and K-Mode. Using correlation analysis, key variables with the highest interrelationships were identified and then used to determine the optimal number of clusters through the Elbow method. Once the optimal cluster count was established, the clustering was conducted using the three methods, and the efficiency was evaluated based on the Sum of Squares Distance (SSD). A lower SSD indicates a more efficient clustering result. The analysis was performed on test data from 15,602 students across four subjects Mathematics, Physics, Lao Language, and Grography. Physics and Lao Language were found to be the most highly correlated variables. Clustering the data using K-Means, K-Medians, and K-Mode produced SSD values of 0.5535, 0.7476, and 1.4937, respectively. The results demonstrate the K-Means achieved the best performance, delivering the most efficient clustering with the lowest SSD. In conclusion, K-Means outperforms K-Medians and K-Mode, making it the most effective algorithm for clustering the given dataset.

Keywords: K-Means, K-Medians, K-Mode, Sum of Squares Distance (SSD), Elbow method

I. INTRODUCTION

In the rapidly advancing field of data technology, information plays a crucial role in analyzing situations and deriving solutions. This is particularly relevant in decision-making and forecasting events based on data-driven outcomes. Cluster analysis, an unsupervised learning method, is widely used to identify patterns and groupings within data by evaluating the similarity between data points. Among the various clustering methods, K-Means is one of the most commonly used due to its simplicity and effectiveness. However, it comes with limitations, such as the need to predefine the number of clusters and its sensitivity to the initial placement of centroids, which can lead to local optima [1], [2].To address these challenges, researchers have developed advanced techniques. For example, Rena Nainggolan et al. [3] used the Elbow method to optimize the number of clusters in K-Means, improving clustering performance by minimizing the Sum of Squared Error (SSE). Similarly, Zhu and Wang [4] proposed an improved K-Means algorithm, which mitigates some of the inherent limitations of the traditional approach by incorporating a threshold value radius [5]. These enhancements make K-Means more reliable, but they still require careful handling of cluster initialization.Other methods like K-Medians and K-Mode have been explored for specific types of data. K-Medians, which is better suited for datasets with outliers, was discussed by Christopher Whelan et al. [6] in their study of the K-Medians problem, while B. Shathya [7] demonstrated its application in predicting student performance, achieving high accuracy. Despite its robustness to outliers, K-Medians can be computationally intensive, making it less efficient for very large datasets compared to K-Means.K-Mode, on the other hand, is optimized for categorical data. K. Lakshmi et al. [8] introduced an enhancement to K-Mode using a CUCKOO Search Optimization Algorithm to improve clustering efficiency. While K-Mode is ideal for categorical data, it does not handle numeric data as effectively as K-Means or K-Medians.The application of these methods is further highlighted in fields such as education. B. Chaisuwan [9] used K-Means clustering to analyze the communication behavior of digital immigrants, dividing the population into distinct groups based on their social interaction patterns. This study underscores the flexibility of K-Means in real-world applications. Similarly, Sya'iyah et al. [10] and Aggarwal and Sharma [11] applied K-Means to cluster student performance data, reinforcing its value in educational analytics.In conclusion, while K-Means is

highly effective for clustering large datasets due to its simplicity and speed, it has certain limitations that can be mitigated by alternative methods such as K-Medians and K-Mode. Each method has its strengths depending on the type of data and the specific requirements of the analysis. Future research, such as that by Yadav and Sharma [1] and Yuan and Yang [12], continues to explore ways to refine these clustering techniques, making them more adaptable to a wider range of applications.

II. The Purpose Clustering Techniques and Data Technology

In the rapidly evolving field of data technology, clustering techniques play a pivotal role in analyzing vast datasets and uncovering meaningful patterns. Among these, K-Means clustering is widely favored for its simplicity and speed. However, K-Means has certain limitations, including its sensitivity to the initial placement of centroids and the need to predefine the number of clusters, which can lead to suboptimal results. To address these challenges, various enhancements and alternative methods, such as the Elbow method, K-Medians, and K-Mode, have been explored. This study aims to compare the performance of K-Means, K-Medians, and K-Mode algorithms by analyzing their efficiency in clustering student performance data. Using the Elbow method to determine the optimal number of clusters, the Sum of Squared Error (SSE) was minimized to evaluate the effectiveness of each method. Results indicate that K-Means outperforms K-Medians and K-Mode in terms ofclustering efficiency, making it a more suitable option for large datasets. However, alternative methods like K-Medians and K-Mode offer advantages in specific use cases, such as handling outliers or categorical data. This comparison provides insights into selecting the most appropriate clustering method based on data characteristics and analysis goals.

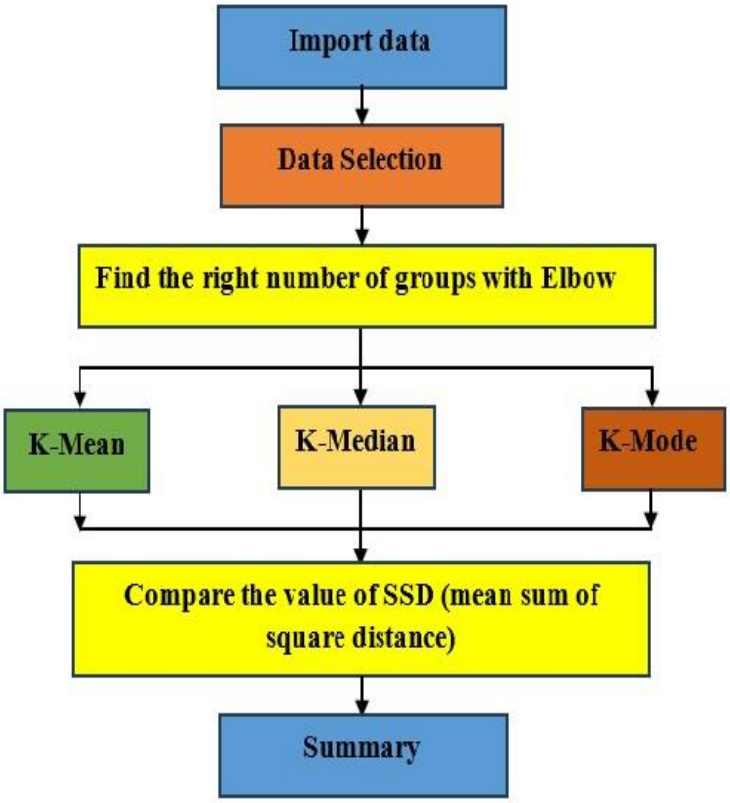

Figure 1 Example Scope of the research

Input Data

The dataset offers valuable insights into the academic performance of students admitted to the National University in the 2014-2015 academic year, specifically focusing on quantitative data to assess and compare results across multiple subjects. Here's a breakdown of the main points and their significance:the scope of data is a dataset is extensive, with 15,602 records, which allows for a robust analysis with statistically

significant findings. Analyzing records from a single academic year also provides a focused snapshot of student performance within that period. The key variables are Student ID (SID) this is a unique identifier for each student, ensuring the ability to track individual records and maintain data integrity. Age by including age, the analysis can explore age-related performance patterns or trends. Age may also serve as a variable in examining maturity or grade level effects on exam results. Subject Scores is Physics Score (PH), Lao Language and Literature Score (LA), Mathematics Score (MA), and Geography and History Score (GO) each represent the student's performance in specific academic areas. Analyzing these scores helps identify areas where students perform well or struggle, which may inform curriculum adjustments or targeted academic support. Total Score, the total score combines individual subject scores, representing the overall academic performance of each student. This metric is essential for broad comparisons, enabling easy identification of top and low-performing students.

The numerical data focus, exclusively on numerical data, the analysis remains streamlined, facilitating various statistical methods such as correlation analysis, clustering, or regression. For instance, the correlation coefficient between scores can reveal how performance in one subject relates to performance in another. Potential Analyses, performance trends by age, subject, or overall performance can be explored, identifying areas for improvement, correlation among subjects for example assessing whether high performance in Mathematics correlates with high performance in Physics could provide insight into interdisciplinary strengths and comparative analysis this the dataset allows for comparisons within subjects or overall performance across different groups, potentially uncovering factors influencing academic success.Overall, this refined dataset serves as a valuable foundation for analyzing student performance across multiple dimensions, potentially aiding in policy decisions, curriculum changes, or targeted educational interventions.

Selection Data

Data selection is a crucial step in data analysis and machine learning, particularly when preparing data for clustering. In clustering, we group data points based on similarities, aiming to discover natural groupings or patterns within the dataset. Effective data selection improves clustering outcomes by ensuring that only relevant variables those that best represent the characteristics we want to analyze are used in the process. The correlation coefficient plays an important role in data selection. It measures the strength and direction of the relationship between two variables, ranging from -1 to 1. A coefficient close to 1 or -1 indicates a strong relationship, while a value close to 0 indicates little or no relationship. When selecting variables, examining correlation coefficients can help in identifying and excluding highly correlated variables to avoid redundancy. This simplification reduces noise, enhances model interpretability, and leads to more efficient clustering. By considering factors like correlation, data selection refines the dataset, making clustering more meaningful and often more accurate.

$$r = \frac{\sum(x_i - \bar{x})\,(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}} \qquad \text{Eq. 1}$$

Note: x_i, y_i represent individual data points,$\bar{x}$,$\bar{y}$ are the mean values of the data, $r$denotes the correlation coefficient.

Elbow Method

The Elbow method is a popular technique for determining the optimal number of clusters, k, for the K-Means clustering algorithm. It is widely used to find the most suitable k by plotting the number of clusters against the total variance or within-cluster sum of squares (WCSS). In the Elbow Method plot, as, k increases, the WCSS typically decreases. However, at a certain point, the rate of decrease sharply diminishes, creating an "elbow" shape on the graph. This "elbow" point is considered the most suitable value for k, where adding more clusters does not significantly improve the model. Is show in figure 2

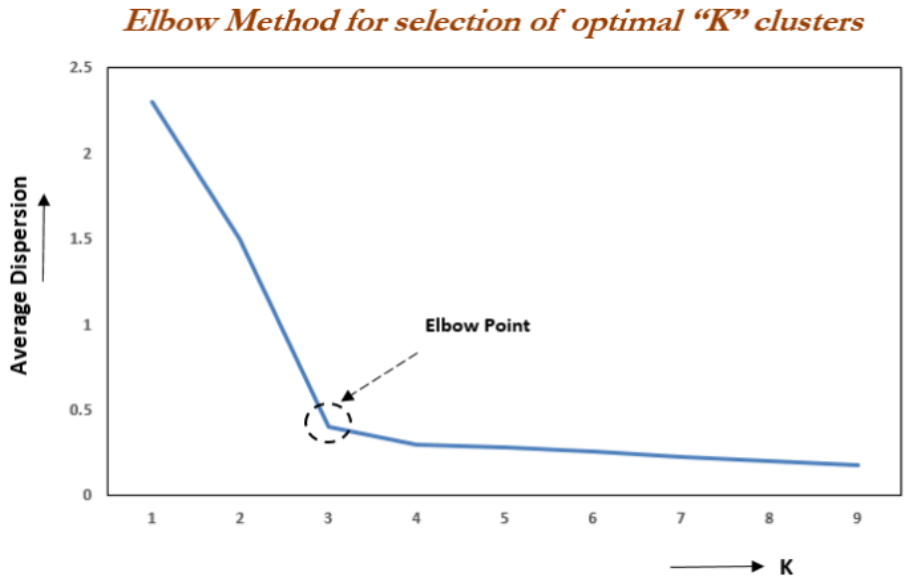

Figure 2 The appropriate value of k for the number of clusters

In clustering, selecting the appropriate number of clusters is critical to obtaining meaningful groupings. Too few clusters may overlook important distinctions in the data, while too many clusters can lead to overfitting. The Elbow Method helps balance simplicity and effectiveness by pinpointing the optimal k value. By helping to select the right k, the Elbow Method improves the clustering model's interpretability and performance, resulting in more meaningful, actionable insights.

K-Means Clustering

The K-Means clustering algorithm groups data into k clusters based on similarity. The process involves the following steps: (1) Input: set the number of clusters, k, and provide the dataset, D. (2) Initialization: calculate initial cluster centroids for k clusters using points from D. (3) Iterative Assignment: for each data point, calculate the distance to each cluster centroid. The distance between points p and q is typically measured by Euclidean distanceand assign each data point to the cluster with the nearest centroid is show in Eq. 2. (4) Centroid update: recalculate the centroid of each cluster by finding the mean of all data points assigned to that cluster. (5) Repeat: repeat the distance calculation and centroid update steps until the centroids stabilize and do not change, indicating convergence.

$$dist(p,q) = \sqrt{\sum_{k=1}^{n}(p_k - q_k)^2} \text{ or } \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \qquad \text{Eq. 2}$$

Additional information on K-Means clustering is distance calculation the Euclidean distance metric helps measure how close each data point is to cluster centroids, ensuring that points within the same cluster are as close to each other as possible. Convergence is reached when points stop switching clusters or when centroid positions stabilize, ensuring that the clusters are well-defined and represent meaningful groupings. Application K-Means clustering is widely used in fields such as customer segmentation, image compression, anomaly detection, and market research, where it helps to reveal natural patterns in data.

K-Mode Clustering

K-Mode clustering is an adaptation of the K-Means clustering algorithm, specifically designed for categorical data. The steps involved in K-Mode clustering are as follows: (1) Initialization: Randomly select the initial number of clusters from the dataset. (2) Assign Clusters: calculate the distance between each data point and the centroid (center point) of each cluster. Then, assign each data point to the cluster with the closest centroid based on the computed distance. (3) Update Centroids Using Mode: for each cluster, update the centroid by finding the mode (most frequent value) for each categorical attribute within the cluster, replacing the previous centroid. (4) Repeat: repeat steps 2 and 3 until the centroids no longer change,

indicating that the clusters have stabilized. Additional information on K-Mode clustering is suitability for categorical data, unlike K-Means, which uses the mean to update centroids (suitable for numerical data), K-Mode relies on the mode, making it ideal for clustering categorical data, such as survey responses, demographic data, and other non-numeric datasets. Distance calculation, in K-Mode clustering, the Hamming distance or matching dissimilarity is commonly used to measure the distance between categorical data points, rather than Euclidean distance, which is unsuitable for categorical variables. Convergence in K-Mode is achieved when cluster centroids stabilize, meaning the mode of each attribute within a cluster does not change. This process minimizes within-cluster variation while preserving meaningful groupings for categorical data. Application K-Mode clustering is often applied in areas where data is primarily categorical, such as customer segmentation based on survey responses, categorizing demographic information, or analyzing categorical attributes in market research.K-Mode clustering offers a reliable solution for handling categorical data, enabling meaningful clustering for datasets that traditional K-Means cannot effectively manage.

K-Medians Clustering

K-Medians clustering is a variation of the K-Means algorithm, which uses the median instead of the mean to define cluster centroids. This approach is particularly useful for reducing the influence of outliers. The steps involved in K-Medians clustering are as follows: (1) Initialization: Randomly select an initial number of clusters. (2) Assign Clusters: Calculate the distance between each data point and the current centroid of each cluster, then assign each data point to the cluster with the closest centroid. (3) Sort Data in Each Cluster: For each cluster, sort the data points by distance from the centroid, organizing the points from smallest to largest. (4) Update Centroids Using the Median: Calculate the median of each attribute for data points within each cluster, then use this median value to update the cluster's centroid. (5) Repeat: Repeat Steps 2 through 4 until the centroids stabilize and no longer change, indicating that the clusters have converged. Additional information on K-Medians clustering is robustness to outliers K-Medians is more robust to outliers than K-Means because it uses the median rather than the mean. This makes it particularly suitable for datasets with skewed distributions or extreme values, where outliers might distort cluster centroids if using the mean. Distance calculation while the choice of distance metric can vary, the Manhattan (or L1) distance is commonly used in K-Medians clustering due to its alignment with median calculations. Convergence and stability areachieved when cluster centroids no longer change with each iteration, resulting in stable clusters. Because it is based on median values, K-Medians generally requires more iterations to converge than K-Means, especially with larger datasets. Application K-Medians clustering is applied in scenarios where data contains significant outliers or non-normally distributed data. It is useful in fields such as finance, where transaction data may include extreme values, or in social sciences, where survey responses may have skewed distributions.K-Medians clustering offers a reliable alternative to K-Means for datasets with outliers, producing stable clusters that better represent the core data by focusing on median values for centroids.

Comparing SSD Values for Model Evaluation

To assess clustering quality, the Sum of Squared Distances (SSD) is often used. SSD measures the compactness of clusters by calculating the sum of squared distances between each data point and its assigned cluster centroid. The formula for SSD is:

$$SSD = \frac{1}{N}\left(\sum_{j=1}^{k}\sum_{i=1}^{N}(dist(x_i, c_j))^2\right) \qquad \text{Eq. 3}$$

Note: N is the total number of data points,k is the number of clusters,x_i, is each data point,c_i, is the centroid of thej^{th}cluster,$dist(x_i, c_i)$ represents the distance between a data point and its assigned cluster centroid. Additional information on SSD in clustering evaluation, the SSD metric evaluates howtightly grouped the data points are within each cluster. A lower SSD value indicates that data points are closer to their centroids, reflecting higher cluster compactness and suggesting that clusters are more homogeneous. Role in the Elbow Method SSD id often used in conjunction with the elbow method to determine the optimal number of clusters. By plotting SSD values for a range of cluster counts (k), a significant bend or "elbow" in the plot suggests the most suitable value for k, where additional clusters provide diminishing returns on cluster

compactness.Sensitivity to Outliers SSD can be sensitive to outliers, as outliers increase the average distance from the centroid. Therefore, in data with significant outliers, it might be combined with other metrics or preprocessing techniques to manage the effect of extreme values. Application SSD is widely used across various domains in clustering analysis, including customer segmentation, image analysis, and medical data classification, to verify clustering quality and determine appropriate clustering parameters. The SSD metric is foundational for evaluating and improving clustering results, as it quantifies the internal consistency of cluster by considering data-point-to-centroid distances across all clusters.

III. Experiment and Results

In this experiment, the efficiency of three clustering methods—K-Means, K-Medians, and K-Mode—was evaluated based on their ability to minimize intra-cluster variance using the Sum of Squared Distances (SSD) metric. SSD serves as an indicator of how compactly data points are grouped within each cluster, providing insight into the clustering quality achieved by each method. To prepare the dataset, relevant numerical variables were selected and normalized to mitigate scaling issues, while categorical data was transformed to accommodate the K-Mode clustering process. This preprocessing step was crucial for ensuring that each clustering method had an equal foundation for performance comparison.An essential part of the process involved determining the optimal number of clusters, or k-value, using the elbow method for each clustering technique. This method helps to identify a point where additional clusters provide diminishing improvements to SSD, indicating an efficient balance between cluster compactness and interpretability.In summary, K-Means was identified as the most efficient clustering method in terms of SSD minimization, producing the most compact clusters. While K-Medians demonstrated robustness for data with outliers, K-Mode was less suitable for numerical data due to its categorical focus. These findings provide clear guidance on method selection based on data type and clustering objectives. For purely numerical datasets, K-Means is recommended, while K-Medians can be beneficial in cases with outliers. For categorical data, K-Mode offers an effective clustering solution, despite its higher SSD in mixed datasets.

Selection Data Results

The results from analyzing correlations among the dataset show that the subjects Physics, Lao Language, and Geography have a correlation of 0.7, which is higher than the correlations among other subjects. It shows as figure 3

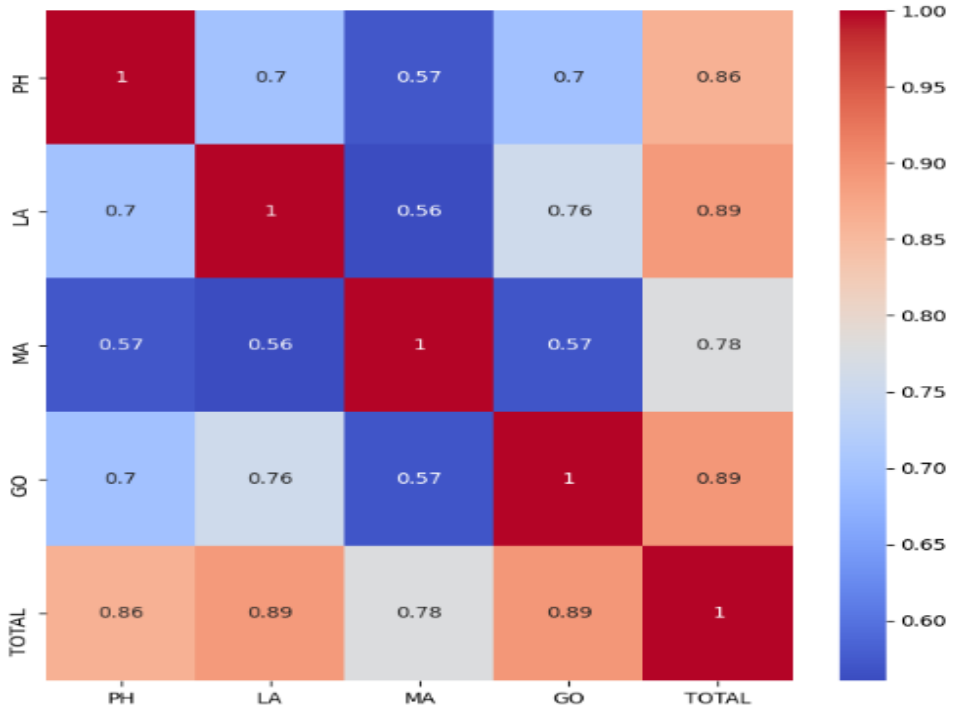

Figure 3Displaying Data Correlation Characteristics with a Correlation Matrix and Heatmap

This correlation of 0.7 indicates a moderately strong positive relationship among these three subjects, suggesting that students who perform well in one of these subjects tend to perform similarly in the others. Here's a closer look: A correlation coefficient of 0.7, which ranges from -1 to +1, shows a significant relationship. In this case, it's positive, meaning scores tend to increase or decrease together. However, it's not a perfect correlation (like 1), so it's moderate, allowing for some variation in performance across these subjects.If these subjects are strongly correlated, it may indicate some overlap in the skills or knowledge required for them, especially if used for clustering or prediction. This could suggest that focusing on one or two of these variables might adequately represent student performance across related subjects, simplifying data without losing significant information.This information can help in selecting the most representative variables for analyses or clustering, aiding in efficient data modeling by reducing redundancy and improving interpretability. Further, these insights can guide educators to investigate shared factors contributing to students' performance across these subjects.

Determining Optimal Data Clusters Using the Elbow Method

Based on the correlation analysis from figure 3, we selected the subjects of Physics and Lao Language for clustering due to their strong correlation, while Geography, sharing a similar correlation pattern with Lao Language, was not included as a separate variable. Using the Elbow method for clustering, we identified that an optimal number of clusters (K) for the data is 3.The Elbow method determines this by plotting the sum of squared distances (SSD) between data points and their respective cluster centroids for varying values of K. As K increases, the SSD decreases as data points are grouped into more clusters. The "elbow" point, where the rate of decrease sharply changes, indicates an optimal number of clusters, balancing data compactness within clusters without overfitting. For this dataset, three clusters best balance complexity and cohesion, aligning well with the observed correlations among variables.

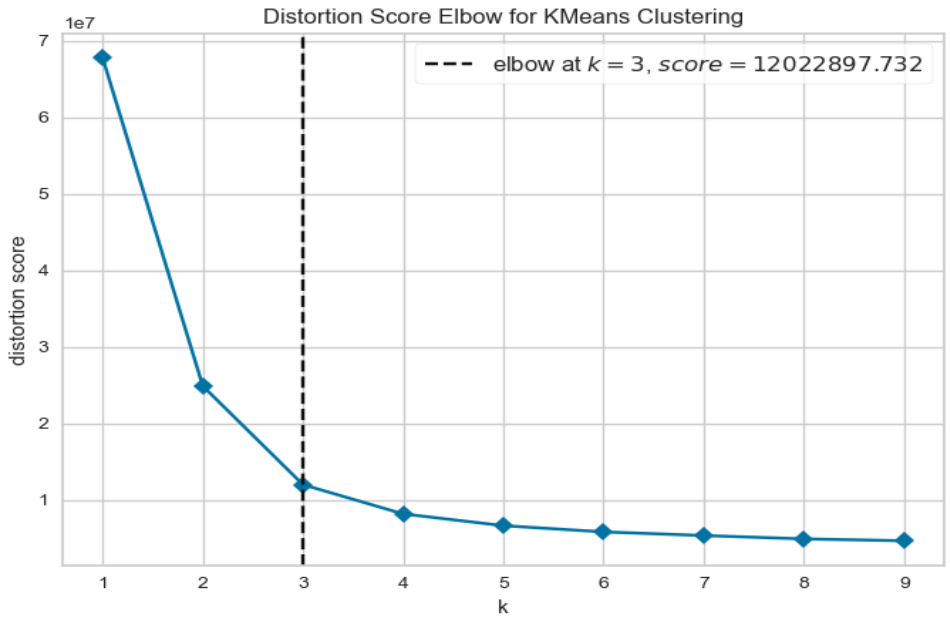

Figure 4 Determining the Optimal Number of Clusters (K) Using the Elbow Method

K-Means Clustering Analysis

Using the K-Means clustering method with k=3, we organized the data into three distinct clusters. For each cluster, a center (or centroid) was computed, representing the average location of points within that cluster. This central point helps to define the characteristics of each group by capturing the typical values of the clustered data. Additionally, the sum of squared distances (SSD) between each data point and its respective cluster center was calculated, resulting in an SSD value of 0.5535. This relatively low SSD suggests that the clusters are well-defined, with minimal variance within each group, affirming that three clusters provide an effective structure for the dataset.

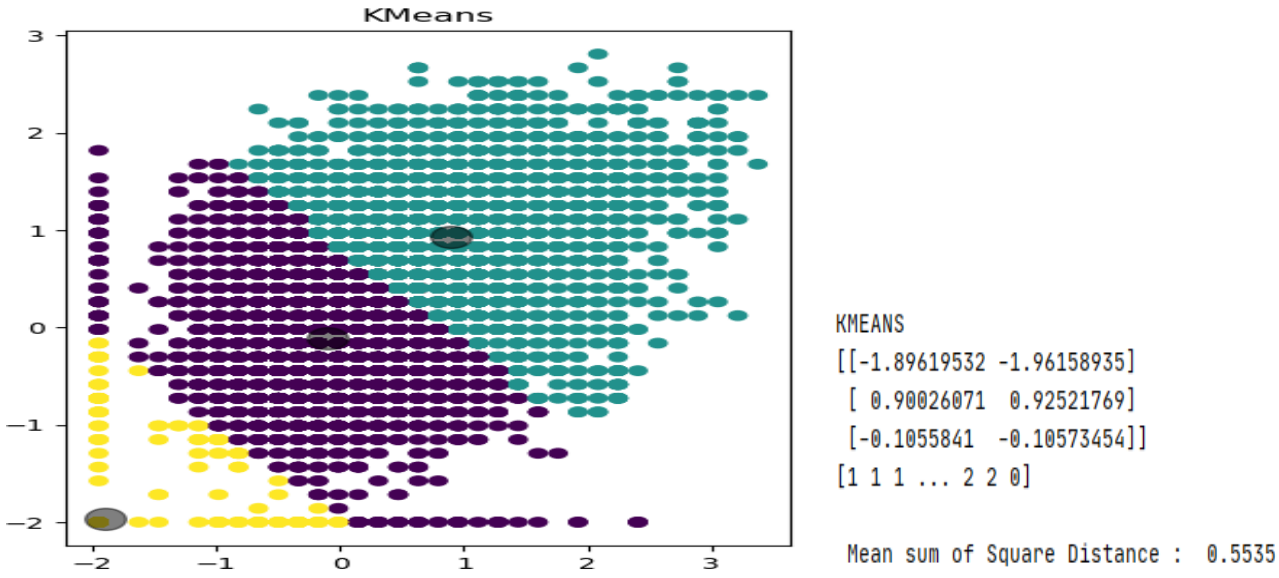

Figure 5 Clusters, Centers, and SSD from K-Means Clustering

K-Mode Clustering Analysis

Based on the experimental results, grouping the data into three clusters (k=3). It was found to be the most appropriate configuration. Each cluster has its own center, and the resulting Sum of Squared Distances (SSD) is 1.4937.In clustering analysis, the selection of the number of clusters (k) is crucial for achieving meaningful groups that accurately represent the data's structure. A lower SSD value indicates that data points are closer to their respective cluster centers, suggesting a more compact and well-defined grouping. The center of each cluster (also called the centroid) is the average position of all points within that cluster, serving as a reference point for defining the cluster's overall position in the data space.Here, an SSD value of 1.4937 reflects the compactness of the clusters and suggests that using three clusters provides a reasonable balance between group coherence and simplicity.

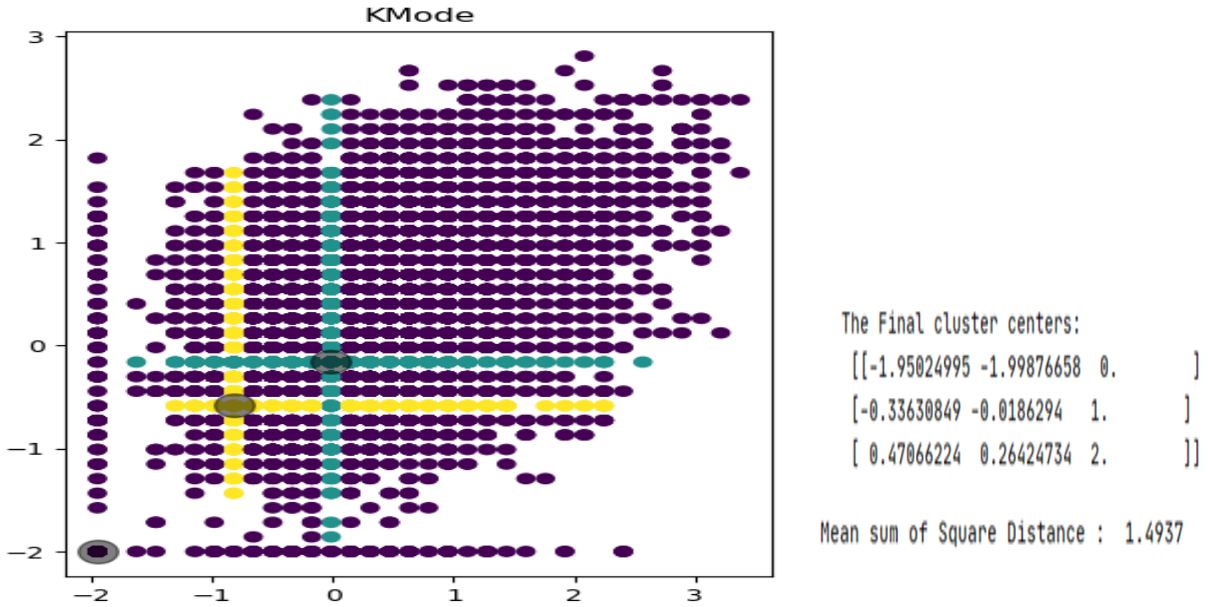

Figure 6Clusters, Centers, and SSD from K-Mode Clustering

K-Medians Clustering Analysis

Based on the experimental results, grouping the data into three clusters (k=3) using K-Medians clustering, as specified in provided the most suitable arrangement. Each cluster has its own center (median), and the resulting Sum of Squared Distances (SSD) is 0.7476.K-Medians clustering is a variation of the traditional k-

means algorithm, where each cluster center is represented by the median rather than the mean. This method is often preferred in scenarios where data contains outliers, as the median is less sensitive to extreme values than the mean. Using medians can result in clusters that better represent the central tendency of data with non-normal distributions or irregular clusters.In this experiment, an SSD value of 0.7476 reflects the clusters' compactness, indicating that data points are closely grouped around their respective medians. This low SSD value suggests a well-defined clustering structure, meaning the data is efficiently grouped into three clusters that capture the underlying data patterns without being skewed by potential outliers.

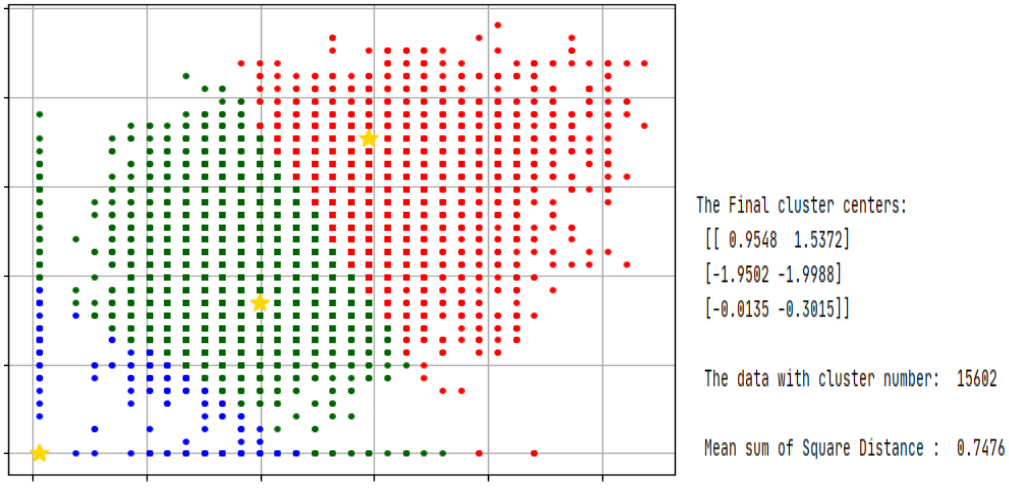

Figure 6 Clusters, Centers, and SSD from K-Medians Clustering

Comparison of Experimental Results Using SSD Values

The results of the data clustering experiments at each stage of this analysis can be compared based on the Sum of Squared Distances (SSD) values, as shown in Table 1.

Table 1 Comparison of Clustering Results Using SSD for K-Means, K-Medians, and K-Mode Methods

	Data Clustering Methods		
	K-Means	***K-Medians***	***K-Mode***
SSD	0.5535	0.7476	1.4937

In clustering analysis, SSD values serve as a metric for evaluating how well the data points fit within their respective clusters. Lower SSD values indicate that data points are closer to their cluster centers, resulting in more compact clusters. By comparing SSD values across different methods or configurations, we can determine which clustering approach best minimizes the distances within clusters, offering insights into the most effective grouping strategy.In this experiment, Table 1 illustrates the SSD values for each clustering method tested, allowing for a direct comparison of their performance. This comparison helps identify the clustering method that most effectively captures the natural structure of the data while minimizing the dispersion of data points around the cluster centers.

IV. Conclusion and Future Works

This paper comparison of clustering methods by incorporating five factors as input data, analyzing their correlations, and selecting the two most strongly correlated factors as input for three clustering methods: K-Means, K-Medians, and K-Mode. The SSD values for these methods were 0.5535, 0.7476, and 1.4937, respectively, showing that K-Means achieved the highest efficiency with the lowest SSD.The findings suggest that K-Means clustering outperformed the other methods in terms of compactness and efficiency when grouping the data. However, the study also highlights that all three methods require predefining the number of clusters, which can be challenging when the data includes complex structures and outliers. To

address these limitations in future analyses, the researchers plan to explore DBSCAN and Fuzzy C-Means methods.

DBSCAN (Density-Based Spatial Clustering of Application with Noise) can form clusters based on data density, making it suitable for data with irregular cluster shapes and outliers, and Fuzzy C-Means allows each data point to belong to multiple clusters with varying degrees of membership, which could provide more flexibility in handling complex data patterns.The use of DBSCAN and Fuzzy C-Means in future studies will help adapt to different data structures and potentially yield improved clustering outcomes for data with unique or complex distributions.

REFERENCES

[1] J. Yadav and M. Sharma, "A Review of K-mean Algorithm," *International Journal of Engineering Trends and Technology*, vol. 4, no. 7, pp. 2972–2976, 2013.

[2] C. Yuan and H. Yang, "Research on K-Value Selection Method of K-Means Clustering Algorithm," *J*, vol. 2, no. 2, pp. 226–235, 2019.

[3] R. Nainggolan, R. Perangin-angin, E. Simarmata, and F. A. Tarigan, "Improved the Performance of the K-Means Cluster Using the Sum of Squared Error (SSE) Optimized by Using the Elbow Method," in *International Conference of SNIKOM*, 2018.

[4] J. Zhu and H. Wang, "An Improved K-means Clustering Algorithm," in *2010 2nd IEEE International Conference on Information Management and Engineering*, 2010, pp. 190–192, doi: 10.1109/ICIME.2010.5478087.

[5] J. Song, F. Li, and R. Li, "Improved K-means Algorithm Based on Threshold Value Radius," *IOP Conference Series: Earth and Environmental Science*, vol. 428, no. 1, 2020, doi: 10.1088/1755-1315/428/1/012001.

[6] C. Whelan, G. Harred, and J. Wang, "Understanding the K-Medians Problem," in *International Conference of Scientific Computing*, 2015.

[7] B. Shathya, "Predicting Students Performance Using K-Medians Clustering," *International Journal of Data Mining Techniques and Applications*, 2015.

[8] K. Lakshmi, N. K. Visalashi, S. Shanthi, and S. Parvathavarthini, "Clustering Categorical Data Using K-Mode Based On CUCKOO Search Optimization Algorithm," *ICTAT Journal on Soft Computing*, 2017.

[9] B. Chaisuwan, "Cluster Analysis of Digital Immigrants Based on Online Communication Behavior and Social Relationship Problem," *NIDA Development Journal*, 2018.

[10] K. Sya'iyah, H. Yuliansyah, and I. Arfiani, "Clustering Student Data Based on K-Means Algorithms," *International Journal of Scientific and Technology Research*, 2019.

[11] D. Aggarwal and D. Sharma, "Application of Clustering for Student Result Analysis," *International Journal of Recent Technology and Engineering*, 2019.

[12] C. Yuan and H. Yang, "Research on K-Value Selection Method of K-Means Clustering Algorithm," *J*, vol. 2, no. 2, pp. 226–235, 2019.