Zero : Jurnal Sains, Matematika, dan Terapan
E-ISSN : 2580-5754; P-ISSN : 2580-569X
Volume 9, Number 1, 2025
DOI: 10.30829/zero.v9i1.24430
Page: 181-190

Integrating Self-Organizing Maps and K-Means in a Multidimensional Approach to Enhance Private University Market Segmentation

[1] **Amalia Nur Alifah**
Department of Data Science, Telkom University, Surabaya, Indonesia

[2] **Wachda Yuniar Rochmah**
Department of Digital Business, Telkom University, Surabaya, Indonesia

[3] **Evellyn Verity Mesak**
Department of Applied Data Analytics, The Australian National University, Canberra, Australia

Article Info

Article history:

Accepted, 28 May 2025

Keywords:

Data-Driven Decision Making;
Educational Marketing;
K-Means Clustering;
School Segmentation;
Self-Organizing Maps (SOM).

ABSTRACT

Educational institutions face challenges in attracting prospective students while maintaining academic quality and resource efficiency. This study applies a hybrid approach that integrates Self-Organizing Maps (SOM) and K-Means to cluster schools based on four attributes, namely the number of accounts, average UTBK scores, geographical distance, and parental income. The analysis's findings produce three distinct clusters. With a high degree of attribute variation, Cluster 2 (279 schools) is a dominant group that suggests the possibility of extensive marketing campaigns. Clusters 1 (45 schools) and 3 (81 schools), on the other hand, are more uniform and call for a more specialized and focused strategy. These results imply that a data-driven approach can help institutions create interventions that are specific to each segment's profile and increase the efficacy of educational marketing strategies. In order to improve segmentation accuracy in the future, this study creates opportunities for investigating new features and dynamic clustering techniques.

Corresponding Author:

Amalia Nur Alifah,
Department of Data Science,
Telkom University, Surabaya, Indonesia
Email: amalialifah@telkomuniversity.ac.id

1. INTRODUCTION

The needs of higher education institutions are increasingly complex and dynamic along with changes in the preferences and expectations of prospective students, which are influenced by the demands of the rapidly developing world of work[1]. To remain relevant, higher education institutions must have the adaptive ability to develop a curriculum that is in accordance with the needs of the times and provide services that are in line with technological advances and globalization[1], [2]. This is not only important to attract prospective students but also to ensure that the graduates produced have adequate competencies to compete in the competitive global job market[2]. In addition, in facing this challenge, a deep understanding of market needs is essential. This understanding allows institutions to develop effective marketing strategies and design educational programs that are relevant to the needs of industry and society as a whole. Thus, institutions can create added value for both students and the world of work[1].

Amidst increasingly fierce competition, the number of higher education institutions in Indonesia has decreased every year. Based on data from the Central Statistics Agency (BPS), the number of universities in 2023 reached 2,966, but decreased to 2,937 in 2024[3]. This decline occurred even though the potential student market

remains large and continues to grow. Every year, almost 3 million students graduate from high school, with around 66% of them continuing on to higher education[4]. In addition, data from the Center for Data and Information Technology (Pusdatin) shows an increase in school participation rates for residents aged sixteen to eighteen years, reaching 74.64% in 2024, with female participation at 72.92% and males slightly lower[5]. This fact reflects a great opportunity for higher education institutions to attract new students. However, the high number of graduates and increasing participation also present challenges in the form of increasingly fierce competition between universities in capturing the attention of prospective students[6].

With the increasingly tight competition in the higher education institution market, institutions need to optimize efficient and targeted marketing strategies to attract prospective students[7]. One approach that has great potential but is still rarely utilized optimally is market segmentation. Segmentation based on historical student data and previous education levels can be a very effective tool to support strategic planning and marketing actions that are in accordance with the characteristics of the target market[1]. By implementing market segmentation, higher education institutions can map schools that have high potential to produce students who are not only large in quantity, but also of high quality in terms of academics and other abilities[1]. This approach allows institutions to focus on more relevant and strategic market segments, thereby increasing marketing efficiency while improving the quality of academic input and output as a whole[8]. Thus, market segmentation is one of the main keys to facing the challenges of competition in the higher education industry.

Several previous studies have used the K-Means Clustering method to segment markets, especially in the field of higher education. Although K-Means is effective in grouping clusters based on the centroid in each cluster, the method has limitations [1]. The position of the centroid in K-Means can change with each iteration, so the clustering results are unstable [9]. In addition, K-Means requires the number of clusters (k) to be determined in advance, making it difficult to determine the exact value of k [10]. K-Means also has several important weaknesses that need to be considered. This method is very sensitive to the presence of outliers and high-dimensional data, so the presence of extreme data or many variables can significantly affect the clustering results [10], [11]. In addition, K-Means is less able to group data that form clusters with non-convex patterns or that have very different sizes, because this algorithm naturally assumes that the clusters are round and uniform [10]. Furthermore, K-Means relies heavily on random centroid initialization so that inappropriate selection of starting points can lead to different clustering results each time the algorithm is run and makes it prone to getting stuck in suboptimal local solutions[10], [12].

To overcome the limitations in market segmentation of higher education institutions, this study uses a hybrid approach that combines Self-Organizing Map (SOM) and K-Means. Self-Organizing Map (SOM) is known as a method that is able to map data with more complex patterns, so that it can capture hidden dimensions in the data and provide informative visualizations [13]. After SOM maps the data, K-Means is applied to refine the results by reducing the number of redundant clusters while increasing clustering accuracy [13]. The hybrid SOM-K-Means approach combines the advantages of SOM-based data mapping that is able to capture complex and non-linear data patterns with K-Means that improves clustering accuracy and reduces redundancy in segmentation results. The combination produces both more accurate and consistent segmentations, enabling more targeted marketing strategies.With more accurate segmentation results, higher education institutions can identify prospective student segments that not only have great potential quantitatively, but also have academic quality. These findings are expected to support institutions in designing more targeted and effective marketing strategies, so that they can increase their competitiveness amidst increasingly tight competition.

2. RESEARCH METHOD

This chapter will systematically discuss the stages of the research, starting from initial data exploration to presentation of analysis results. Each stage in this research is interrelated, forming a coherent, directed, and structured workflow. This research uses a combination of the Self-Organizing Map (SOM) and K-Means methods to segment schools with the main goal of improving the quality and quantity of prospective students through the implementation of more effective and efficient marketing strategies. This combination of methods allows for more detailed and accurate processing of complex data.

An overview of the research methodology is explained visually in Figure 1, which presents the workflow starting from data collection, preprocessing, SOM training, K-Means application, to validation and visualization of results. The process begins with the data collection stage which includes recording, scraping, and merging data. The collected data then goes through a preprocessing process, including data cleaning and standardization. After that, SOM is trained using specified parameters, followed by the application of K-Means for further clustering. The final stage includes validation using metrics such as quantization error and topographic error, and interpretation of the results. This flow ensures that each step is well integrated to achieve the research objectives.

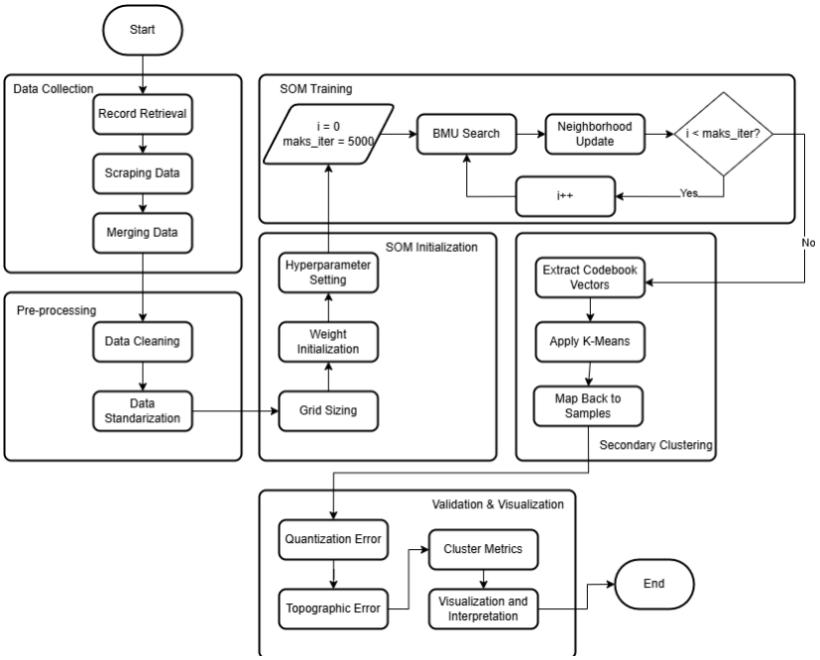

Figure 1. Research Methodology

2.1. Data Understanding and Preprocessing

In the Data Understanding phase, two main steps are conducted: data collection and data description. The process begins with collecting data from various sources, both primary and secondary, which are then integrated into a unified and comprehensive database. Primary data is obtained through administrative processes, including information such as user accounts, school origins, and the average income of parents. This data is calculated based on the number of students from each school to provide a more accurate depiction of the applicant distribution. Meanwhile, secondary data is gathered through web scraping from the official LTMPT (Institute for Higher Education Entrance Tests) website, specifically top-1000-sekolah.ltmpt.ac.id, which contains UTBK (Computer-Based Written Examination) scores. These scores are further extracted to obtain the average values of three key components: scholastic, saintek (science and technology), and soshum (social and humanities). The data from these two sources is then processed and analyzed to deliver deeper and more relevant insights, supporting a systematic and well-directed analysis in the subsequent stages.

Table 1. Data Description

No	Attribute	Type	Description
1	School Origin	Categorical	Originating school
2	Account	Numeric	Registered account
3	PIN	Numeric	Admission PIN used
4	Pass	Numeric	Passed selection
5	Registration	Numeric	Completed registration
6	Income	Numeric	Parental income
7	Scholastic	Numeric	Scholastic score
8	Science and Technology	Numeric	Science & Tech (Saintek) score
9	Social and Humanities	Numeric	Social & Humanities (Soshum) score
10	Distance	Numeric	School-to-university distance
11	Average UTBK Score	Numeric	Average UTBK score

To complement the existing data, the calculation of school location distances was conducted using data scraped from Google Maps. This information was utilized to analyze the tendencies of campus selection based on geographical proximity, providing valuable insights into how location influences student preferences. The entire process aims to support school segmentation analysis based on two primary aspects: the quantity and quality of prospective students. Following data collection, the next step involves data description, which is essential for gaining a deeper understanding of the dataset. This includes identifying data types, structures, column descriptions, data distribution patterns, and correlations between variables. By thoroughly examining these elements, it becomes easier to identify patterns and relationships that are critical for segmentation. Additionally, Table 1 presents a detailed column description to simplify the process of determining which attributes should be further processed

and analyzed in the subsequent stages. This ensures a more systematic and targeted approach to data handling and analysis.

The selection of four main features, namely the number of accounts, the average UTBK score, the geographical distance of prospective students, and the average income of parents is based on the relevance of the strategy to marketing decisions and recruitment of new students. The number of accounts reflects the level of initial interest or engagement of prospective students, while the UTBK score reflects academic readiness. Geographic distance is an important indicator in considering accessibility and study location preferences, while parental income is used as a proxy for socio-economic background that can influence enrollment decisions and the need for subsidy-based interventions or personalized approaches. The selection of these features is also supported by initial discussions with the marketing team and academic institutions, as well as references from previous studies in the field of education segmentation.

Data Preprocessing is a fundamental step in ensuring that the data is clean, reliable, and suitable for analysis[2]. Raw data often contains imperfections, such as inconsistencies, missing values, noise, and outliers that fall outside the expected range[14]. This study utilizes a combination of three primary data sources: data from top-1000-sekolah.ltmpt.ac.id, Google Maps, and administrative admission processes, requiring comprehensive preprocessing to enhance data quality. The preprocessing phase begins with data cleaning, where missing values, especially those arising from mismatched entries between sources are handled using mean or median imputation, depending on the distribution of the data. Inconsistencies in format or entry values are also corrected to ensure dataset integrity. Outliers, given their potential impact on clustering performance, are addressed through a two-stage process: initially via log transformation, which reduces skewness and suppresses the magnitude of extreme values, followed by the removal of extreme outliers using the interquartile range (IQR) rule [15]. Finally, Standardization is conducted to align all variables on the same scale, facilitating effective and unbiased analysis[16]. These preprocessing steps are essential for preparing high-quality data that can support accurate and robust analysis, laying a solid foundation for subsequent modeling and segmentation efforts.

2.2. SOM Implementation

To ensure optimal performance and generalizability of the Self-Organizing Map (SOM), careful consideration was given to the selection of hyperparameters such as the learning rate, neighborhood radius (sigma), and the number of iterations. Improper tuning of these parameters can lead to underfitting, where the model fails to capture important data patterns or overfitting, where the map adapts excessively to noise and loses generalization capability. In this study, the learning rate and sigma values were initialized with commonly used defaults and then refined through empirical testing, gradually decreasing during training to encourage convergence without destabilizing the weight updates.

The standardized data is then mapped into two dimensions through the grid sizing process, namely determining the shape and size of the map. Choosing the right grid size is a crucial step to ensure optimal data representation and support accurate analysis [17]. In this study, the grid sizing was determined at 10 x 10, resulting in a square map. The selection of this size is based on several considerations, including the efficiency of computational complexity, the grid's ability to capture data variations in detail, and adequate resolution and visualization quality [18]. This grid is considered ideal enough to balance between analysis precision and technical performance.

Moreover, the chosen grid size (10×10) represents a trade-off between computational efficiency and the ability to produce meaningful segmentation. A smaller grid risks under-representing the complexity of the data, while a larger grid may result in sparse activation and overfitting. Preliminary experiments with varying grid sizes (ranging from 5×5 to 15×15) were conducted, and the 10×10 configuration demonstrated the most interpretable clustering structure with consistent topology preservation and manageable training time. This choice supports both the granularity required for segmentation and the practicality of implementation.

The next step is weight initialization using the Principal Component Analysis (PCA) method. This method is applied to accelerate convergence by adjusting the initial weights to the main direction of data variance. In addition to accelerating the process, PCA-based initialization also provides initial weights that are more stable, efficient, and representative of the existing data distribution [19]. After weight initialization, parameters such as learning rate, sigma, and neighborhood function are determined, all of which are designed to direct the Self-Organizing Map (SOM) training process to run optimally. Determining these hyperparameters is a key element in ensuring that the SOM model can map the data with high accuracy while producing informative and interpretable visualizations.

$$BMU = \arg\min(||x(t) - w(t)||) \quad (1)$$

The Self-Organizing Map (SOM) training process is carried out using 5000 iterations, where in each iteration the weights are updated based on randomly selected data samples. This approach uses a random method to ensure that the model has the flexibility to capture complex data patterns [18]. In each iteration, the neurons on the map compete to be the best representation of the given input data. The neuron that has the weight with the smallest Euclidean distance to the input vector will be selected as the Best Matching Unit (BMU). This BMU is the focus of the weight update process to optimally approach the input pattern. The BMU can be calculated using Equation

(1)[20] where $x(t)$ is the input vector at the t-th iteration, and $w(t)$ is the weight of the t-th neuron on the SOM grid. By utilizing this process, the SOM is able to adaptively map the data and produce a two-dimensional representation that reflects the data structure intuitively. This iterative process also ensures that the SOM map is well trained for the desired segmentation analysis.

$$w_i(t+1) = w_j(t) + \alpha(t) \cdot hb_j(t) \cdot \left[x(t) - w_j(t)\right] \quad (2)$$

Once the Best Matching Unit (BMU) is found, the next step is to perform a neighborhood update to adjust the weights of the neurons around the BMU. This process uses Equation (2)[21], where $w_i(t)$ represents the weight of the i-th neuron at iteration t, while $x(t)$ is the input vector that triggers the update. The parameter $\alpha(t)$, or learning rate, decreases gradually over time to improve the stability of the model in later iterations. In addition, $hb_j(t)$, the neighborhood function, determines the influence of the BMU on the i-th neuron based on the distance between them. This function uses Equation (3)[17], where d_{ij} is the Euclidean distance between the BMU and neuron j. The parameter $\sigma(t)$, which represents the width of the Gaussian function, also shrinks over time, creating a smaller zone of influence around the BMU[22].

$$hb_j(t) = \exp\left(-\frac{d_{ij}^2}{2\sigma^2}\right) \quad (3)$$

This process is repeated until it reaches a maximum limit of 5000 iterations to ensure that the neuron weights can fully adapt to the input data distribution. The end result of SOM training is neuron weights that accurately represent the structure and patterns in the data. This representation becomes the basis for further segmentation analysis, making SOM a very effective tool in multidimensional data mapping.

2.3. Secondary Clustering with K-Means Clustering

After the Self-Organizing Map (SOM) training process is complete, each neuron in the SOM grid has a final weight vector that represents the distribution of data learned during training. This weight vector serves as a representation of the characteristics of the data associated with the neuron, so that each neuron can be considered as a code or prototypical representation of a set of data. To facilitate the interpretation and analysis of the clustering results obtained from the SOM, a second clustering process is carried out using the K-Means algorithm. At this stage, the weight vectors of the trained neurons are treated as new data to be regrouped. By using K-Means, these vectors are clustered into a number of k predetermined clusters, where the main purpose of this second clustering is to simplify and clarify the segmentation results produced by the SOM, so that the clustering results become more structured and easier to analyze. The number of clusters for K-Means was determined using the Elbow Method, where we evaluated the within-cluster sum of squares (WCSS) for various values of k. The optimal number of clusters was chosen where a significant drop in WCSS was observed, which indicated the most meaningful division of the data. This approach ensures that the selected number of clusters provides a good balance between model complexity and data representation.

$$D_{ij} = \sqrt{\left(X1_i - X1_j\right)^2 + \left(X2_i - X2_j\right)^2 + \cdots + \left(Xn_i - Xn_j\right)^2} \quad (4)$$

The clustering process with K-Means utilizes the calculation of the Euclidean distance between each neuron weight vector and the existing cluster center, as formulated in equation (4) [1], where D_{ij} indicates the distance between the i-th data and the j-th cluster center. After K-Means has finished grouping the neuron weight vectors, each neuron in the SOM grid has a cluster label which is the result of the K-Means grouping. Furthermore, the original data that has previously been mapped to the Best Matching Unit (BMU) neuron in the SOM can be directly given a cluster label based on the cluster of the BMU neuron. This stage is known as mapping back to samples, where each data sample obtains a final label according to the results of the BMU neuron weight vector clustering. Thus, the combination process of SOM and K-Means produces a more robust and structured clustering technique, which is able to group data efficiently based on the representation of neurons in the SOM while simplifying the final clustering results for easier and more informative analysis.

2.4. Visualization and Interpretation

After the clustering process with Self-Organizing Map (SOM) and K-Means is complete, the next step is to evaluate the quality of the mapping and clustering results obtained. One of the main metrics used is Quantization Error (QE), which is a measurement of the average distance between the original input data and the weight of the Best Matching Unit (BMU) neurons that represent the data. QE describes how well the SOM map is able to represent the original data; the smaller the QE value, the more accurate the representation. In addition to QE, the evaluation of the map topology is also carried out using Topographic Error (TE). TE measures the preservation of the structure and geometric relationships of the original data on the SOM map by calculating the average distance between the winning neuron (BMU) and the second best matching neuron for each data. A small TE value indicates that the SOM map has succeeded in effectively maintaining the topology of the data, which is important for maintaining the similarity and relationships between data in the mapping process to lower dimensions.

Furthermore, to assess the overall quality of the clustering results, measurements are carried out with several cluster metrics including Silhouette Score, Dunn Index, and Connectivity [21], [22]. Silhouette Score is used to measure the extent to which data is more similar to its own cluster members than to other cluster members, with higher values indicating better and more clearly separated clusters. Dunn Index assesses the ratio between the minimum distance between clusters to the maximum diameter within the cluster, where a large Dunn Index value indicates good and compact cluster separation[23]. Connectivity measures the extent to which locally adjacent data are grouped in the same cluster which a low Connectivity value indicates strong local coherence between cluster members.

These three metrics together provide a comprehensive picture of the quality of data segmentation resulting from the combination of SOM and K-Means[24], [25]. To facilitate interpretation, the clustering results are visualized using the Unified Distance Matrix (U-Matrix) and Data Map, where the U-Matrix displays the boundaries between clusters through color gradations on the SOM map, while the Data Map shows the distribution of data based on K-Means clusters in the SOM grid, making it easier to understand the structure and relationships between clusters visually[4], [17], [26].

3. RESULT AND ANALYSIS

This study aims to segment school clusters by utilizing four main attributes that represent important characteristics of schools. To achieve this goal, a hybrid method that combines Self-Organizing Map (SOM) and K-Means Clustering is applied. SOM plays an important role in mapping school data into an intuitive two-dimensional map, where each neuron in the map represents a particular cluster characteristic. By using SOM, complex and multidimensional data can be projected efficiently so that hidden patterns and structures in the data can be seen more clearly. This method allows the identification of groups of schools with high characteristic similarities, thus facilitating the understanding of different market segments in the world of education.

Furthermore, the results of the neuron representation from SOM are simplified through an advanced clustering process using the K-Means algorithm. K-Means helps group these neurons into a number of more structured and easily interpreted clusters. The combination of these two methods produces cluster segmentation that is not only statistically accurate and relevant, but also easy to understand by decision makers. With more robust and interpretive segmentation results, educational marketing strategies can be formulated more precisely, adjusting to the unique characteristics of each school cluster. This allows for the preparation of more effective promotion and development programs, thereby increasing the efficiency of resource allocation and the results obtained from the educational marketing strategy.

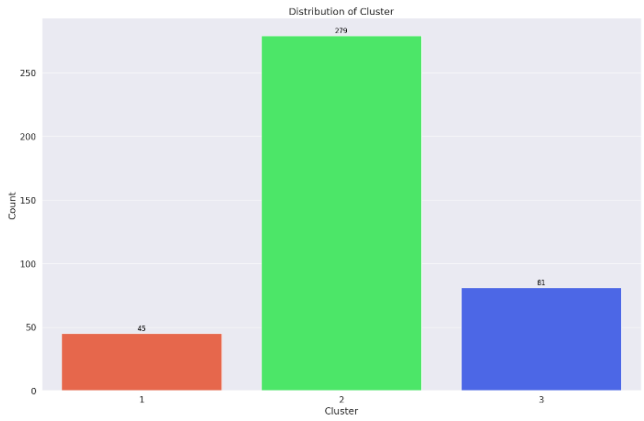

Figure 2. Distribution of Each Cluster

The distribution of schools in each cluster after the clustering process (shown in Figure 2) shows that Cluster 2 is the most dominant group with a total of 279 schools. This indicates that most schools have fairly uniform characteristics and form a strong general pattern in the data. The dominance of Cluster 2 reflects the existence of a majority group that represents conditions or attributes that are often found in many schools. Meanwhile, Cluster 1 and Cluster 3 each consist of a smaller number of schools, indicating that these two clusters contain more specific characteristics and are significantly different from Cluster 2. Although smaller in number, Clusters 1 and 3 are important to note because they describe significant variations in the data, which can represent the needs or characteristics of certain segments. Therefore, an in-depth analysis of each attribute that forms these clusters is essential to understand the differences in characteristics in detail. With this understanding, educational marketing strategies can be designed more precisely, adjusted to the uniqueness and needs of each school segment that has been formed through the clustering process.

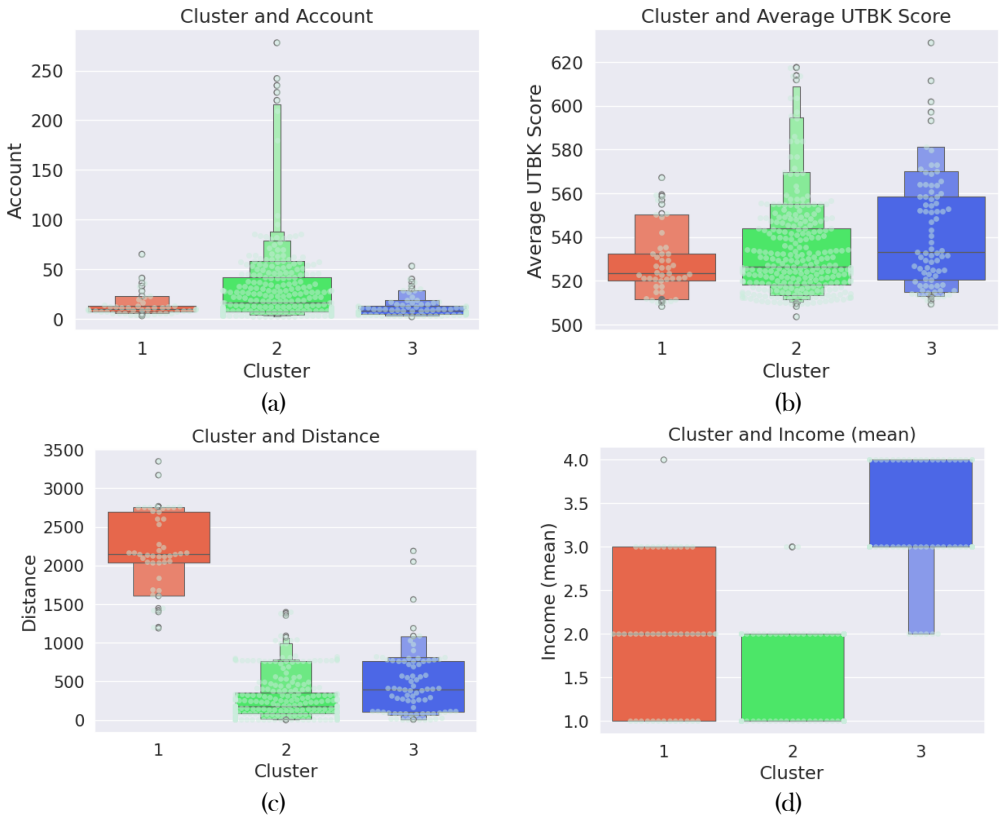

Figure 3. Boxplot Distribution of Each Cluster: (a) Versus Account Attribute; (b) Versus UTBK Score Attribute; (c) Versus Distance Attribute; (d) Versus Income Attribute

The distribution of school accounts in each cluster shows a unique pattern that provides important insights for educational marketing strategies. Based on Figure 3(a), Cluster 2 as the majority cluster has a fairly wide distribution of accounts with significant variations, including the presence of outliers reaching more than 250 accounts. In contrast, Cluster 1 and Cluster 3 have a more stable distribution, with the number of accounts ranging from below 50 to 50 accounts. This difference indicates that Cluster 2 includes schools with quite large scale variations, while Cluster 1 and Cluster 3 are more homogeneous. This information is relevant for developing specific marketing strategies, such as an intensification approach for schools with a large number of accounts in Cluster 2, or a more personalized approach for Cluster 1 and Cluster 3. Thus, institutions can allocate marketing resources more effectively based on the characteristics of each cluster.

Further analysis shows differences in the average UTBK scores in each cluster, as shown in Figure 3(b). Cluster 3 has the highest average UTBK score, followed by Cluster 2 which shows a fairly good score, while Cluster 1 has a lower average score. This provides strategic guidance for educational institutions to direct their promotional focus based on the academic quality of prospective students. Cluster 3 and part of Cluster 2 can be prioritized targets to attract high-quality prospective students, who have the potential to maintain the institution's academic standards. Meanwhile, Cluster 1 requires special attention for academic development, such as providing support for additional educational programs or scholarships to improve the quality of students in this group.

From the aspect of geographical efficiency, the analysis in Figure 3(c) shows the distribution of school locations based on the radius of the prospective student's school distance. Cluster 2 dominates with the majority of schools located within a relatively close radius from the institution. In contrast, Cluster 1 includes schools with locations that are further apart. This condition provides an important consideration in planning the allocation of marketing budgets. By prioritizing promotions in Cluster 2, institutions can reduce travel costs, for example in school visits or sending promotional materials, while maximizing the intensity of interaction with prospective students. In contrast, for Cluster 1, more efficient promotional strategies, such as online promotions or collaboration with local institutions, may be more cost-effective and effective options.

In addition, the distribution of parental income in Figure 3(d) shows that Cluster 3 includes students from relatively strong economic backgrounds compared to other clusters. This information can be used to design inclusive and diverse marketing strategies. Institutions can utilize this data to minimize the risk of dropping out due to financial constraints by providing more affordable scholarship programs or financing schemes for Clusters 1 and 2. On the other hand, Cluster 3, with its high parental income, can be targeted for campaigns for premium education programs or additional services with higher value. By utilizing the clustering results comprehensively, institutions can design strategies that not only increase the number of enrollments but also maintain academic quality, optimize marketing budget efficiency, and ensure more stable financial sustainability for students.

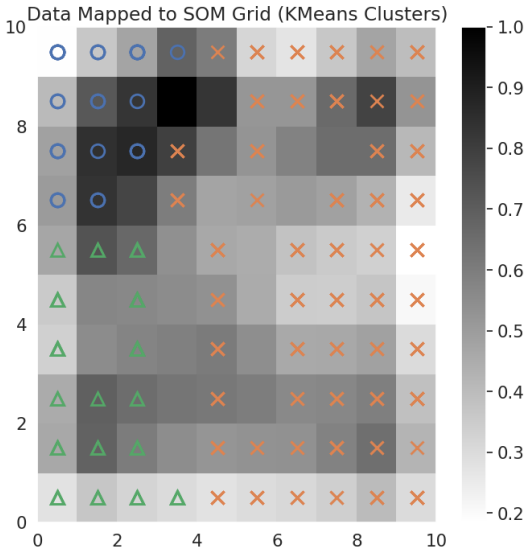

Figure 4. K-Means Cluster Mapping on SOM Grid

As a support for cluster analysis, Figure 4 presents a U-Matrix visualization of Self-Organizing Maps (SOM), which functions to provide an overview of the internal structure of the data. The U-Matrix displays the distance between neurons in the SOM grid, where the color gradation reflects the intensity of the relationship between neurons. Areas with dark colors indicate a greater distance between neurons, indicating a strong cluster separation. Conversely, areas with light colors indicate a smaller distance, reflecting a close relationship between neurons that form a cluster. In this visualization, symbols such as circles, triangles, and crosses are used to represent different clusters. For example, the circle symbol is centered on the upper left, the triangle dominates the lower left, and the cross symbol is more widely distributed throughout the grid, indicating that the cluster has a higher level of variation than other clusters.

This U-Matrix visualization not only strengthens the validity of the clustering results but also provides in-depth insights into school segmentation. By looking at the distribution pattern of symbols on the grid, institutions can understand the unique characteristics of each cluster. For example, clusters with concentrated symbols may indicate groups of schools with homogeneous characteristics, while clusters with widely dispersed symbols reflect higher heterogeneity. This information can be used to develop more targeted marketing strategies, such as offering specific programs that are tailored to the needs and potential of each cluster. In addition, clusters with dark distances between neurons can be identified as areas that require special attention, both in academic development and educational promotion. Thus, the U-Matrix becomes a very useful tool to optimize the effectiveness of marketing and program management of higher education institutions in a strategic and data-driven manner.

In terms of practical application, the segmentation provides actionable insights for optimizing marketing resource allocation. For instance, Cluster 2, characterized by a wide range of socioeconomic and academic attributes, presents opportunities for large-scale outreach strategies such as digital campaigns or regional promotions. In contrast, Clusters 1 and 3 exhibit more uniform profiles, which suggest the need for personalized engagement, such as school-specific visits, scholarship targeting, or academic mentoring programs. Institutions can use these insights to prioritize recruitment efforts, tailor messages to distinct audience segments, and allocate budgets more efficiently based on the strategic value of each cluster. The segmentation framework is adaptable and can be applied by other institutions seeking to enhance the precision and impact of their marketing strategies.

To evaluate the effectiveness of the SOM-K-Means segmentation, we compared it with simpler clustering methods, such as K-Means without the SOM pre-processing step, and traditional random targeting approaches often used in educational marketing. The results show that the SOM-K-Means approach provides more accurate segmentation by capturing complex patterns in the data, which allows for better identification of high-potential student segments. In contrast, random targeting or simpler clustering methods fail to identify meaningful subgroups and often lead to inefficient resource allocation in marketing efforts.

4. CONCLUSION

This study successfully combines the Self-Organizing Maps (SOM) and K-Means methods for school segmentation based on four main attributes, namely the number of school accounts, average UTBK scores, distance of applicants, and average parental income. The segmentation results show that Cluster 2 and Cluster 3 are the main focus because they have great potential in improving the quality and quantity of applicants. In addition, these two clusters also provide opportunities to optimize the efficiency of marketing strategies and better resource management. Based on the distribution analysis, around 89% of schools in the dataset fall into the two priority clusters, making them ideal targets for more targeted and effective promotional strategies. By focusing on these

clusters, educational institutions can achieve optimal results in attracting prospective students who are in line with their academic vision and mission.

The methodological innovation of combining Self-Organizing Maps (SOM) with K-Means clustering significantly improves the segmentation of complex educational data, offering more accurate and insightful results compared to traditional clustering techniques. This hybrid approach enables educational institutions to design targeted and efficient marketing strategies, optimizing resource allocation. However, limitations must be acknowledged, such as variations in data quality across institutions, which could introduce bias, and the assumption of static student preferences, while educational choices and market conditions evolve over time due to socioeconomic shifts, technological influences, or policy changes. To address these challenges, future research should explore adaptive clustering techniques and integrate real-time behavioral monitoring, allowing for continuous adaptation to changing student behaviors and ensuring that segmentation remains relevant in the face of evolving market dynamics.

5. REFERENCES

[1] H. Khusnuliawati and D. R. Putri, "Hybrid clustering based on multi-criteria segmentation for higher education marketing," *TELKOMNIKA (Telecommunication Comput. Electron. Control.*, vol. 19, no. 5, pp. 1498–1506, 2021.

[2] B. L. Ortiz, "Data Preprocessing Techniques for Artificial Learning (AI)/Machine Learning (ML)-Readiness: Systematic Review of Wearable Sensor Data in Cancer Care.," *JMIR Mhealth Uhealth*, 2024.

[3] Agnes Yonathan, "Indonesia Jadi Negara dengan Universitas Terbanyak Kedua di Dunia." https://data.goodstats.id/statistic/indonesia-jadi-negara-dengan-universitas-terbanyak-kedua-di-dunia-Iab4A (accessed May 31, 2025).

[4] W. A. Prastyabudi, A. N. Alifah, and A. Nurdin, "Segmenting the Higher Education Market: An Analysis of Admissions Data Using K-Means Clustering," *Procedia Comput. Sci.*, vol. 234, pp. 96–105, 2024.

[5] P. D. dan T. Informasi, "Statistik Sekolah Menengah Atas (SMA) 2023/2024," 2024.

[6] M. Seyfried, S. Hollenberg, and J. Heße-Husain, "Student selection in higher education–the organisational performance dilemma," *J. High. Educ. Policy Manag.*, vol. 46, no. 6, pp. 671–686, 2024.

[7] Y.-F. Chen and C.-H. Hsiao, "Applying market segmentation theory to student behavior in selecting a school or department.," *New Horizons Educ.*, vol. 57, no. 2, pp. 32–43, 2009.

[8] M. Arpay, "Student Mining Using K-Means Clustering: A Basis for Improving Higher Education Marketing Strategies," *Psychol. Educ. A Multidiscip. J.*, vol. 14, no. 1, p. 1, 2023.

[9] E. Xiao, "Comprehensive K-Means Clustering," *J. Comput. Commun.*, vol. 12, no. 3, pp. 146–159, 2024.

[10] M. Z. Rodriguez *et al.*, "Clustering algorithms: A comparative approach," *PLoS One*, vol. 14, no. 1, p. e0210236, 2019.

[11] Š. Brodinová, P. Filzmoser, T. Ortner, C. Breiteneder, and M. Rohm, "Robust and sparse k-means clustering for high-dimensional data," *Adv. Data Anal. Classif.*, vol. 13, pp. 905–932, 2019.

[12] E. U. Oti, M. O. Olusola, F. C. Eze, and S. U. Enogwe, "Comprehensive review of K-Means clustering algorithms," *Criterion*, vol. 12, pp. 22–23, 2021.

[13] K. K. Jassar and K. S. Dhindsa, "Comparative study and performance analysis of clustering algorithms," *Int J Comput Appl*, vol. 975, p. 8887, 2015.

[14] V. Çetin and O. Yıldız, "A comprehensive review on data preprocessing techniques in data analysis," *Pamukkale Üniversitesi Mühendislik Bilim. Derg.*, vol. 28, no. 2, pp. 299–312, 2022.

[15] C. Fan, M. Chen, X. Wang, J. Wang, and B. Huang, "A review on data preprocessing techniques toward efficient and reliable knowledge discovery from building operational data," *Front. energy Res.*, vol. 9, p. 652801, 2021.

[16] I. Engdahl, "Agreements 'in the wild': Standards and alignment in machine learning benchmark dataset construction," *Big Data Soc.*, vol. 11, no. 2, p. 20539517241242456, 2024.

[17] Y. Miftahuddin and A. R. S. Ridwan, "Application of Self-Organizing Map and K-Means to Cluster Bandwidth Usage Patterns in Campus Environment," *J. Online Inform.*, vol. 10, no. 1, pp. 66–76, 2025.

[18] W. Hua and L. Mo, "Clustering Ensemble Model Based on Self-Organizing Map Network," *Comput. Intell. Neurosci.*, vol. 2020, no. 1, p. 2971565, 2020.

[19] N. Shi and R. Al Kontar, "Personalized pca: Decoupling shared and unique features," *J. Mach. Learn. Res.*, vol. 25, no. 41, pp. 1–82, 2024.

[20] D. L. B. Fortela *et al.*, "Using Self-Organizing Maps to Elucidate Patterns among Variables in Simulated Syngas Combustion," *Clean Technol.*, vol. 2, no. 2, p. 11, 2020.

[21] A. Deshmukh, "Variational Quantum Self-Organizing Map," *arXiv Prepr. arXiv2504.03584*, 2025.

[22] G. T. Breard, "Evaluating self-organizing map quality measures as convergence criteria," 2017.

[23] J. Tian, M. H. Azarian, and M. Pecht, "Anomaly detection using self-organizing maps-based k-nearest neighbor algorithm," in *PHM society European conference*, 2014, vol. 2, no. 1.

[24] R. A. Sary, N. Satyahadewi, and W. Andani, "APPLICATION OF K-MEANS++ WITH DUNN INDEX VALIDATION OF GROUPING WEST KALIMANTAN REGION BASED ON CRIME VULNERABILITY," *BAREKENG J. Ilmu Mat. dan Terap.*, vol. 18, no. 4, pp. 2283–2292, 2024.

[25] A. Pita, F. J. Rodriguez, and J. M. Navarro, "Analysis and Evaluation of Clustering Techniques Applied to Wireless Acoustics Sensor Network Data," *Appl. Sci.*, vol. 12, no. 17, p. 8550, 2022.

[26] S. Myagmarsuren, "Exploring the Use of Silhouette Score in K-Means Clustering for Image Segmentation (Exploring the Use of Silhouette Score in K-Means Clustering for Image Segmentation)," vol. 13, no. 4, 2024, [Online]. Available: http://www.ijert.org