JURNAL EMACS
[Engineering, MAthematics and Computer Science] Vol.8 No.2 May 2026: 145-157
e-ISSN: 2686-2573
DOI:10.21512/emacsjournal.v8i2.14486

Analysis of K-Means Clustering Implementation for New Student Segmentation Based on High School Background and Study Program

Risqy Siwi Pradini[1], Yulianti[2*], Mochammad Anshori[3]
[1-3]Department of Informatics, Faculty of Science and Technology,
ITSK RS dr. Soepraoen Malang
Malang, Jawa Timur, Indonesia 65119
risqy.pradinia@itsk-soepraoen.ac.id; yulianti@itsk-soepraoen.ac.id;
moanshori@itsk-soepraoen.ac.id

*Correspondence: yulianti@itsk-soepraoen.ac.id

Abstract *- Increasing Competition among higher education institutions requires a deeper understanding of prospective students characteristics to develop more effective promotional strategies and student admission policies. The K-Means Clustering Algorithm is used in this study to divide recently admitted students into discrete groups according to their chosen course of study, kind of previous school, and place of origin. The Davies-Bouldin Index (DBI) and the Silhouette Score were used to assess the clustering performance's efficacy, and the Elbow Method was used to determine the ideal number of clusters. The Experimental result indicated that the optimal segmentation consisted of two clusters. The first cluster was primarily composed of students graduating from general senior high schools (SMA), who tended to enroll in highly demanded health-related study programs. Conversely, the second cluster exhibited a more heterogeneous distribution concerning study program preferences and was predominantly constituted of students hailing from Islamic senior high schools (MA) and vocational high schools (SMK). The cluster analysis has been found to give a Silhouette Score of 0.1774 and a Davies–Bouldin Index of 2.1806, meaning that although the separation between clusters is still inadequate, it is good enough to reveal some distinct patterns in the traits of the new students admitted. The results of this research can be used as a useful benchmark for developing marketing and partnership policies with secondary schools and making data-based decisions-making in the new student admission process.*

Keywords: *K-Means Clustering; Student Segmentation; Elbow Method; Silhouette Score; Davies–Bouldin Index; PCA 2D*

I. INTRODUCTION

The admission process of new students is considered to be one of the key processes in the operation of higher education institutions (Elimar et al., 2024). Every year, many universities gather large amounts of data on potential students which include such information as geographical area, type of institution they graduated from, and chosen educational program. Unfortunately, such information is usually only used in a formal way and not taken as a full source of information capable of providing valuable insights for institutional decision-makers (Yunita, 2018).

Even though the K-Means clustering method has been used extensively in education research,

In Press

Copyright © 2026

majority of these studies have focused mainly on analyzing students' performance, learning approaches, or characteristics of students in higher education environments (Izzati et al., 2023). Clustering methods have also been shown to be able to facilitate marketing approaches and segmentation of students in higher education institutions (Andy, 2024), (Poiema & Mantohana, 2025). However, research on student recruitment tends to focus more on individual characteristics like geographical location or school background individually and cannot provide a holistic picture of the characteristics of potential students (Andy, 2024), (Poiema & Mantohana, 2025). Furthermore, admission data is not being widely utilized in higher education institutions as a strategic tool for enhance data-driven decision-making (Andy, 2024).

Therefore, this research aims at developing an integration of cluster analysis with several crucial variables related to new admission students in order to obtain a more comprehensive segmentation process. It is expected that the obtained clusters will provide many useful insights into the process of improving university promotional campaigns, enhancing recruitment strategies, and refining institutional planning grounded in data-driven evidence.

Making decisions based on data has become increasingly significant in view of the rapid advancement of information technology and data analysis methods. Universities will be able to know the features of prospective students and implement their policies flexibly based on empirically obtained facts through the use of data analysis tools (Aziz et al., 2018). Among the most relevant methods to reveal the hidden patterns within new student admission data is clustering.

Clustering is a type of unsupervised learning that attempts to discover clusters or groups of similar objects from a given dataset without using any predetermined class labels. With this technique, it becomes possible to group together the information about students based on their similarities in order to identify some hidden patterns that are not apparent otherwise (Fitri et al., 2023), (Nugroho et al., 2022). Institutions offering higher education may get a better insight into the academic inclinations and demographics of newly admitted students through this information.

Given its efficacy in partitioning large-scale and multidimensional datasets, K-Means Clustering ranks among the most frequently employed clustering methodologies (Solichin & Khairunnisa, 2020). The method makes the clusters compact and well-defined by minimizing the distance of the data points from the centroids of the cluster (Aulia, 2021). In this research, new entrants were classified into three broad categories based on their study program, type of previous school, and the origin of the entrant (whether the city or regency) through K-Means Clustering. These features were selected due to their significant importance in the student admissions process.

It is expected that the resulting segmentation will provide a deeper insight into characteristics of recently admitted students, including the geographical distribution of admitted candidates, the prevalent programs of study preferences in each school category, and student preferences that are difficult to detect via conventional methods of analysis. This information may help colleges formulate more effective promotion efforts, improve collaboration with high schools, and critically assess their admission policy in different programs. At the same time, by encouraging data-based decisions in educational institutions, this study provides practical significance along with theoretical value in the

In Press

domain of machine learning applications in education.

II. METHODS

Figure 1. Research flowchart.

As shown in figure 1 below, the research methodology used to form clusters of new incoming students using the K-Means algorithm is made up of a number of steps. These steps include data gathering, data preprocessing (feature selection, handling missing values, and variable coding), selection of the best number of clusters, carrying out the cluster analysis using the K-Means algorithm, and cluster validation. The following sections provide a comprehensive elucidation of each stage.

1. Data Collection

Data on recently admitted students who have successfully completed the admissions process was first collected for this study. There were 3,116 records of the students involved in the dataset. There were various attributes present in the data which were city/regency, study program, and school previously attended by the students. The data was extracted from the admission database of the university.

2. Data Preprocessing

This step was done in order to prepare the dataset to be used in the clustering algorithm. Three major preprocessing techniques were applied.

a. Feature Selection

The selection of the features used in clustering was done based on the relevance of the feature to the research objective, which in turn is to find the pattern of students' backgrounds. These features are city/regency of origin, study program, and previous school.

For the *previous school* attribute, we first changed the name of each school into school category based on their name prefixes, namely "SMA," "SMK," and "MA." This process helped in reducing data sparsity and made the data more concise in representing their classes, thus improving clustering efficiency (Yaumi et al., 2020).

b. Missing Value Removal

Records with missing or invalid values in important attributes were removed from the dataset. The purpose of this step was to increase the quality of clustering and reduce possible bias due to missing information (Lashiyanti et al., 2023), (Wahyu & Rushenda, 2022).

c. Categorical Encoding

During this phase, categorical features were encoded into numbers for further processing via the K-Means algorithm. Categorical encoding was performed for three attributes: city/regency of origin, study program attended, and type of school. One-hot encoding approach was used for encoding that creates binary vectors and preserves categorical data without implying any ordinal relation (Purnama et al., 2022). This step is essential because the K-Means algorithm computes distances between data points and therefore requires numerical input (Fadilah & Wijayanto, 2023). After encoding of categorical variables, the obtained encoded vectors were transformed into a DataFrame prior to merging with the dataset. This conversion facilitates data exploration and integration.

Before combining the DataFrames, their indices were reset to avoid any issues regarding the indices that may occur due to the joining process. In the end,

In Press

the dataset was formed through the horizontal joining of the original features and the encoded features, thus forming a dataset that can be used for clustering analysis. This way, all the categorical variables were converted into numerical values, enabling the algorithm to calculate the distances between the points better (Fadilah & Wijayanto, 2023).

d. Determination of the Number of Clusters

Elbow method was employed prior to the clustering algorithm to obtain the optimal number of clusters. Elbow method is widely used to find out the optimal number of clusters (K) that can be created in a given dataset. It helps to compute the Within Cluster Sum of Squares (WCSS), which is the variance in the data points within a cluster, across a range of K values (Maori & Evanita, 2023).

The ideal number of clusters was has been selected based on the calculation of the within-cluster sum of squares (WCSS). The data points of each cluster will be more compact if the value of WCSS is smaller. WCSS is calculated using calculate WCSS(1).

$$WCSS = \sum_{i=1}^{k} \sum_{x_j \in C_i} \| x_j - \mu_i \|^2$$

Equation (1)

where (k) denotes the number of clusters, (xj) signifies a data point that is allocated to cluster (Ci), and (μi) represents the centroid of the cluster. A lower WCSS value indicates greater homogeneity within each cluster, resulting in more compact cluster formations.

Then, the values of WCSS are represented by a line graph in order to determine the elbow point. In this case, the specific value represents the best number of clusters since, after that, any further number of clusters will have little impact on reducing the variance within the clusters. In other words, the effectiveness of clustering reduces even though the number of clusters increases (Sulistiyawan et al., 2021).

The Silhouette Score and the Davies–Bouldin Index (DBI), are both measures used for validating the cluster internally that were used after determining the best number of clusters using the Elbow Method. These metrics were utilized to assess the quality of the generated clusters concerning separation (the distinction between various clusters) and cohesion (the compactness of data points within the same cluster).

In regards to the nearby clusters, the Silhouette Score measures how accurately the particular data point has been placed into its own corresponding cluster. For the calculation of the Silhouette Score, equation (2) is employed. If the value is close to 1, then it means that there is great similarity among the data points belonging to the same cluster (Shutaywi, 2021).

$$S(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

Equation (2)

where $a(i)$ represents the average distance between data point i and all other data points within the same cluster, and $b(i)$ is the average distance between data point i and the closest neighboring cluster. The range of the Silhouette Score is −1 to 1. Superior clustering quality is suggested by elevated values, which show more intra-cluster similarity and greater inter-cluster separation (Shutaywi, 2021).

In addition to the Silhouette Score, the Davies–Bouldin Index (DBI) was used to assess cluster quality. This measure calculates the ratio of the dispersion inside clusters to the distance between cluster centroids. Lower DBI values suggest that the

resultant clusters are more compact and demonstrate enhanced separation from one another. Lower DBI values indicate that the resulting clusters are more compact and exhibit better separation from one another (Ros et al., n.d.).

$$DBI = \frac{1}{K}\sum_{i=1}^{K} \max_{j \neq i}\left(\frac{S_i + S_j}{M_{ij}}\right)$$

Equation (3)

where Mi and Sj represent the within-cluster dispersion of clusters i and j, respectively, and Mij signifies the separation between the centroids of clusters i and j. Higher clustering quality is indicated by a smaller DBI score since it represents more compact clusters that are farther apart from nearby clusters (Ros et al., n.d.).

e. K-Means Clustering Process

The K-Means Clustering methodology was implemented to cluster the data subsequent to its preprocessing and transformation into numerical forms. To attain convergence, the algorithm initially randomizes the initialization of cluster centroids, assigns each data point to the nearest centroid, and subsequently iteratively refines the positions of the centroids (Virgo et al., 2020). This iterative process produces clusters consisting of students with similar characteristics based on the selected feature space.

The data points are allocated to the nearest centroid after the random initialization of the centroids. Centroid locations are updated continuously until convergence is achieved (Virgo et al., 2020). This procedure yields a number of groups containing students with similar attributes in the given feature space.

$$J = \sum_{i=1}^{k} \sum_{x_j \in C_i} \| x_j - \mu_i \|^2$$

Equation (4)

The K-means algorithm uses the proximity of the data points to the cluster centers to group them into a predefined number of clusters. The objective function (4) as provided in equation aims to minimize the objective function (4). The objective function is similar to the Within Cluster Sum of Squares(WCSS) where K-means attempts to minimize the within cluster variance.

After clustering, the centroid of each cluster is calculated using the mean of all the data points that belong to the particular cluster as shown in equation (5). This procedure is repeated iteratively until the centroid positions stabilize or the algorithm reaches convergence.

$$\mu_i = \frac{1}{| C_i |}\sum_{x_j \in C_i} x_j$$

Equation (5)

In Equation (5), x_j refers to a singular data point within cluster i, μ_i denotes the updated centroid of the i-th cluster, and C_i indicates the total number of data points attributed to cluster i. This iteration of clustering continues till the convergence criterion is achieved.

f. Cluster Analysis

Conducting a descriptive analysis of the clustering results is the last step. ach cluster is analyzed in light of the demographic characteristics such as study programs, schools, and regions of origin. To provide a better representation of clusters, Principal Component Analysis (PCA) is performed for a two-dimensional visualization. Findings from the analysis are employed in determining student profiles for evidence-based institutional decisions.

Dimensionality reduction of one-hot encoded data into two principal components through Principal Component Analysis (PCA) is done to facilitate visualization. A covariance matrix is first

In Press

constructed for the one-hot encoded data using Equation (6) and then eigenvalues and eigenvectors are computed using Equation (7). The selected principal components have large eigenvalues as they account for most of the variance in the dataset.

$$C = \frac{1}{n-1} X^T X$$

Equation (6)

In Equation (6), Crepresents the covariance matrix, Xdenotes the normalized data matrix, and $n$represents the total number of observations.

Subsequently, PCA computes the eigenvalues and eigenvectors using Equation (7).

$$Cv = \lambda v$$

Equation (7)

In Equation (7), Cdenotes the covariance matrix, $v$represents the eigenvector, and $\lambda$is the corresponding eigenvalue. The principal components are selected based on the largest eigenvalues because they account for the greatest proportion of variance in the dataset. In this study, the first two principal components were selected to visualize the clustering results of newly admitted students in a two-dimensional space.

III. RESULTS AND DISCUSSION

Based on the methodological approaches delineated in the preceding chapter, the findings and analyses are articulated in this chapter. The study begins with the data collection and preprocessing stages using data on newly admitted students obtained from the university's student admission information system.

3.1 Data Collection and Preprocessing

The official university entrance (PMB) portal database provided the research data. The initial dataset consisted of 3,116 student records containing attributes such as the city/regency of origin, study program admitted to, and the name of the previous school. Through the transformation of the school name attribute into a newly defined feature termed school level, which categorizes educational institutions into Senior High School (SMA), Vocational High School (SMK), and Islamic Senior High School (MA), feature engineering was implemented to enhance both the pertinence and representativeness of the features.

Next, records containing missing values were removed, resulting in a cleaned dataset of 1,558 records that was ready for further processing. Categorical encoding was then applied using the one-hot encoding technique to the attributes *city/regency of origin*, *study program admitted to*, and *school level*. The transformed features were converted into a numerical DataFrame using the Pandas library and subsequently concatenated with the remaining variables to produce a fully numerical dataset suitable for the clustering process.

Figure 2. Example of the Final Dataset After Data Preprocessing

Figure 2 presents a portion of the final dataset generated after the data preprocessing stage conducted in this study. The original dataset consisted entirely of categorical attributes, including *city/regency of origin*, *admitted study program*, *school name*, and *school category*. These attributes were successfully transformed into a numerical representation using the one-hot encoding technique. This encoding process produced 242

feature columns, where each column represents a unique category derived from the original categorical variables.

In the dataset shown in Figure 2, the first five records are displayed to illustrate the structure of the processed dataset. Each record retains the original categorical information, including the city/regency of origin, admitted study program, and school name, thereby preserving interpretability despite the numerical transformation. The encoded feature columns, displayed on the right side of the table, contain binary values (0 or 1) that indicate the presence or absence of a particular category for each student.

To facilitate the processing of the dataset employing the K-Means clustering methodology, which necessitates numerical input characteristics, this modification is imperative. In addition, the transformation preserves the complete information contained in the original categorical variables. In general, the resultant dataset offers a suitable numerical representation for the clustering endeavor while preserving the contextual information essential for the interpretation of the clustering outcomes.

3.2 Determination of the Optimal Number of Clusters

The optimal number of clusters was determined utilizing the Elbow method by identifying the inflection point on the Within-Cluster Sum of Squares (WCSS) curve concerning the quantity of clusters (k). Subsequently, the resulting clustering solutions were evaluated for cluster compactness and separation through the application of the Davies-Bouldin Index (DBI) and the Silhouette Score. The integration of these evaluative metrics provides a more objective and reliable basis for ascertaining the ideal number of clusters, thereby yielding representative clustering results.

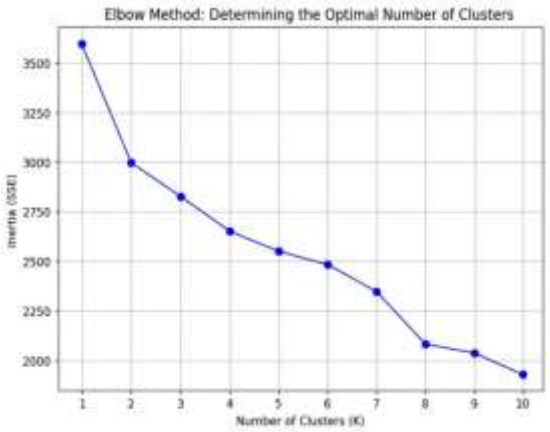

Figure 3. Elbow Method curve showing the most significant elbow point at k=2, indicating the optimal number of clusters.

The findings obtained with the help of the Elbow Method shown in Figure 3 show that there is an evident drop in the Sum of Squared Errors (SSE) as the number of clusters increases from k=1 to k=2. For k≥3, there is a drop in SSE, but this time, its rate is much lower than that for k=2. Typically, the most prominent elbow is observed at k=2 because it is the biggest drop in the curve's slope.

The objective of the Elbow Method is to determine the optimum number of clusters where further increase in the number of clusters will not result in any considerable decrease in the clustering inertia. The graph shows the perfect balance between the complexity of the model and clustering. Though there is a marginal decline in inertia even when $k = 3$, the additional improvement is relatively small and may produce a more complex clustering solution that is less interpretable.

The quality of the clustering solution was evaluated using the Davies-Bouldin Index (DBI) and the Silhouette Score prior to a comprehensive examination of the characteristics of each cluster.

In Press

A Silhouette Score of 0.1774 was obtained from the clustering evaluation. This very low number suggests that the separation between clusters remains narrow, suggesting that some observations share characteristics with neighboring clusters. In other words, there is still some overlap amongst the discovered clusters and the cluster borders are not yet clearly defined.

Furthermore, a Davies–Bouldin Index (DBI) of 2.1806 was obtained from the clustering solution. The clusters have not yet reached ideal compactness and separation, according to this value. A lower DBI generally reflects better clustering quality, whereas the relatively high DBI obtained in this study suggests that the clusters remain relatively similar to one another.

The features of the dataset utilized in this investigation might have had an impact on the clustering performance. The region of origin, school type, and approved study program are all categorical variables that make up the student admittance dataset. Post-transformation via one-hot encoding, the resultant high-dimensional feature space may diminish the efficacy of distance-based clustering techniques, such as K-Means, consequently impacting the overall quality of the clustering.

The number of features increased to 242 attributes after the preprocessing stage. The distances between data points may become more uniform in such a very high-dimensional feature space, which would lessen the ability of distance-based clustering algorithms like K-Means to create well-separated clusters.

Although the Silhouette Score and Davies–Bouldin Index (DBI) indicate that the cluster separation is not optimum, the clustering results nonetheless provide significant segmentation based on the prominent traits within each group. Therefore, the clustering model remains suitable for exploratory analysis aimed at identifying patterns in newly admitted students based on their region of origin, school category, and admitted study program.

According to the analysis results, an increase in the number of clusters from k=2 to k=10 has not brought about any considerable changes in clustering quality. The best Silhouette Score achieved was 0.1865, which was for k=10, while the best Davies-Bouldin Index was 2.0561. The results imply that there are no distinct clusters in the underlying data set. Therefore, using $k = 2$based on the Elbow Method remains appropriate because it provides a simpler and more interpretable segmentation for analyzing the characteristics of newly admitted students.

3.3 Clustering Process Using K-Means

The process of clustering is then carried out using the K-Means algorithm, where k is set to 2. The first step involves assigning random values to the centroids of each cluster, followed by assigning each data point to the closest centroid. This continues until convergence is achieved.

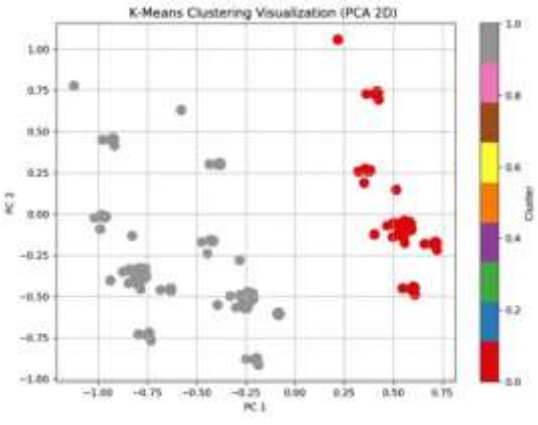

Figure 4. Two-dimensional visualization of the clustering results generated by the K-Means algorithm (k=2) in the PCA feature space.

A two-dimensional depiction of the clustering output produced from the K-means technique is

In Press

shown in Figure 4. In order to ensure that the primary variations inherent in the data are captured in the visualization process, PCA is used to reduce the dimensionality of the data set to two-dimensions.

PCA is an acronym that stands for Principal Component Analysis. The two principal components that account for the highest percentage of variation in the dataset after encoding are represented on the horizontal axis (PC1) and vertical axis (PC2).

The points illustrated in the plot represent newly admitted students that were clustered into two distinct clusters based on the optimal number of clusters (k=2) determined using the Elbow Method. Two distinct clusters are evident from the illustration above and indicate a clear clustering tendency. This graph serves to assist in measuring the extent of separation between the clusters while the interpretation of the clustering results needs a deeper analysis of each cluster's characteristics.

3.4 Cluster Analysis

The clustering analysis suggests that the characteristics of each cluster are different and can be understood from the distribution of the school categories and the study programs being offered by the schools. These help in understanding the clustering process better.

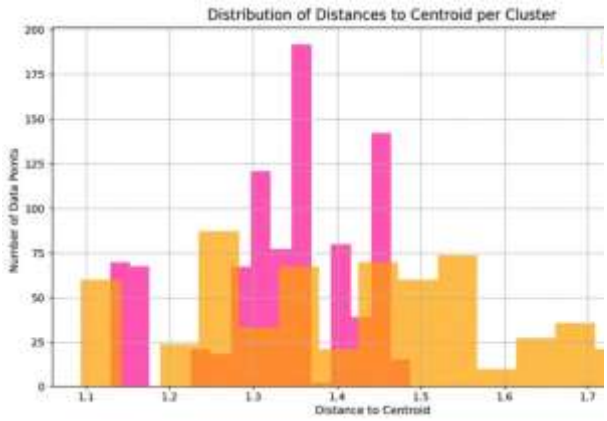

Figure 5. Distribution of distances from data points to the cluster centroids.

The distribution of distances between data points and their respective centroids after applying the K-Means algorithm with k=2 is shown in Figure 5 below. The x-axis shows the Euclidean distance from the assigned centroid while the y-axis shows the number of observations at each distance.

The histogram indicates that the predominant number of observations in both clusters is situated at relatively short distances from their corresponding centroids. This indicates that there is a good amount of compactness and consistency within these clusters, which makes the outcome of this segmentation a valid one. From the diagram, it is clear that the K-Means algorithm has classified the new admissions into groups that have distinct and compact features, aligning with the overarching distribution of the dataset.

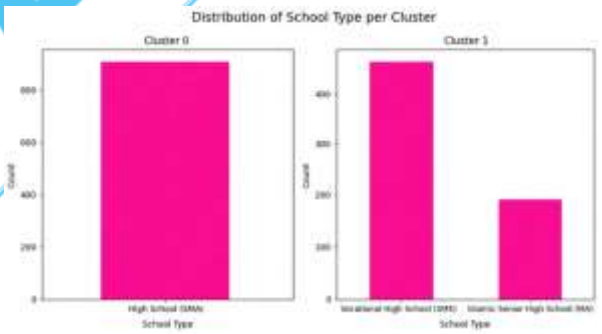

Figure 6. Distribution of school categories of newly admitted students across the identified clusters.

Figure 6 presents the distribution of schools among the students who have been newly admitted in each cluster based on the K-Means algorithm. The students of Cluster 0 are mainly graduates of Senior High Schools (SMA) as they constitute the largest group of schools in this cluster. On the other hand, Cluster 1 contains mainly graduates of Vocational High Schools (SMK), with a minor number of Islamic Senior High School (MA) students.

According to the distribution above, it can be seen that the process of clustering managed to categorize students according to their background

In Press

school. The results above suggest that students having similar educational background will be clustered together. In addition, the data above is very useful for gaining some information about the profiles of applicants and serves as a basis for formulating recruitment strategies.

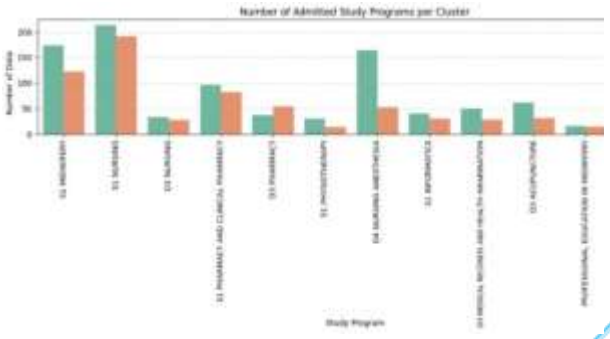

Figure 7. Distribution of newly admitted students by study program across the identified clusters.

Figure 7 demonstrates how the students who were just admitted have been allocated to the available programs of study in both of the clusters formed by using the K-Means method. It has been shown that there are some programs of study such as bachelor's in nursing (S1 Keperawatan) and Bachelors in Midwifery (S1 Kebidanan) in both of the clusters, although with different proportions.

Cluster 0 has a larger number of students studying, for instance, Diploma IV in Nurse Anesthesiology (D4 Keperawatan Anestesiologi), while Cluster 1 has a more equal number of students distributed into various study programs, for example, Diploma III in Pharmacy (D3 Farmasi) and Bachelor of Clinical and Community Pharmacy (S1 Farmasi Klinis dan Komunitas).

These results mean that the above-mentioned clusters differ not only because of the background education of the students but also because of different admission program distribution. Thus, the results obtained allow us to understand better student profiles and, therefore, can help us to plan activities, evaluate educational programs, and recruit students.

The distribution of the students' places of origin by the clusters obtained cannot be provided because of the high number of regions, which will lead to the creation of a very dense graph that will be hard to analyze. This decision was made to maintain the clarity, readability, and overall effectiveness of the presentation of the research findings.

Even though these two groups contain programs of equivalent study, differences in their distributions allow understanding the students' preferences and character better. This will help conduct additional research in relation to the link between the selected program and students' background.

The clustering method conducted using the K-Means algorithm with k=2 showed two different clusters of new admitted students distinguished by different features. The cluster 0 includes mostly students that have finished at the Senior High School (SMA), the largest percentage of whom study such courses as Bachelor of Midwifery (S1 Kebidanan), Bachelor of Nursing (S1 Keperawatan), and Diploma IV in Nurse Anesthesiology (D4 Keperawatan Anestesiologi). In this way, one can suppose that students from SMA usually prefer well-known and popular health-related courses. Moreover, this cluster has good compactness since there is a very short distance between the data points and their centroid.

On the contrary, most of the students in cluster 1 have graduated from Islamic Senior High Schools (MA) and Vocational High Schools (SMK). It is found that the students in this cluster are quite equally distributed among a number of academic majors such as Diploma III in Pharmacy (D3 Farmasi), Bachelor of Clinical and Community Pharmacy (S1 Farmasi Klinis dan Komunitas), and

a number of other academic majors. From the above findings, it can be concluded that the students who have graduated from SMK and MA have different preferences for study programs. In addition, the two-dimensional PCA graph confirms the presence of separation between the two clusters, reinforcing the interpretation that educational background and preferred study program are key distinguishing characteristics of the identified student groups.

Overall, the results presented here indicate that recently enrolled students can be divided into two major segments depending on the category of school and program choice. The obtained data can be useful for universities in terms of getting acquainted with the profile of the students and help develop specific approaches to their recruitment and services planning tailored to the characteristics of each segment.

IV. CONCLUSION

Based on the analysis in this research, there exist two groups of fresh-admitted students based on the K-Means clustering method. On the contrary, students in Cluster 1 consist mainly of students who graduate from Vocational High Schools (SMK) and Islamic Senior High Schools (MA). This group of students is characterized by having various options of study programs compared to Cluster 0, which consists mostly of students graduating from Senior High Schools (SMA) who choose health studies.

The selection of $k = 2$using the Elbow method was found to be a suitable compromise between complexity and effectiveness of the cluster analysis. Moreover, the importance of the entire preprocessing step such as feature selection, missing value imputation, and one hot encoding should be highlighted in terms of preparation of data for clustering.

The results suggest that it is possible to derive useful information about the attributes of new entrants using clustering and that clustering will be helpful for universities in formulating strategies in recruitment and admissions depending upon the attributes of various student segments.

The values of 0.1774 and 2.1806 were calculated by means of analyzing cluster quality using the Silhouette Score along with the Davies-Bouldin Index (DBI). The obtained results imply that the level of the distance between the clusters is quite low and that there is some level of overlap of the features of the detected student groups. However, even in this case, the results of clustering help to understand the overall tendencies in the behavior of new students due to the combination of the selected features. Therefore, the results of the research may be useful in formulating strategies of recruiting students and building relationships with schools and supporting data-driven decision-making in the university admission process.

There are several limitations associated with this research study. To begin with, the findings cannot be generalized to other institutions where there is diversity in the composition of the student body since the sample used was derived from just one institution. In addition, the process of clustering did not take into account any other variables apart from the three identified, that is, the geographical location of the students, the type of school and the field of study to which they were admitted. This is because using the K-Means clustering algorithm on categorical data using one hot encoding led to a very high dimensional feature space, which may have reduced cluster compactness and separation, as reflected by the relatively low Silhouette Score and relatively high Davies–Bouldin Index.

Future studies should aim at clustering methods that will be suitable for use on categorical data, such

In Press

as K-Modes or K-Prototypes, and a comparison of these techniques' efficiency with that of K-Means. Moreover, an enhanced understanding of student attributes and admission trends can be achieved by using admission data from multiple admission sessions, as well as a wide range of variables, which would enhance the quality and usefulness of the clustering results.

AUTHOR'S CONTRIBUTION

Formal analysis, conceptualization, data curation, methodology, validation, visualization, and writing.

AVAILABILITY DATA AND MATERIALS

The respective organization may provide the dataset used for this paper on request.

REFERENCES

Andy, W. (2024). ScienceDirect Segmenting Segmenting the the Higher Higher Education Education Market : Market : An An Analysis Analysis of of Admissions Admissions Data Data Using Using K-Means K-Means Clustering Clustering. *Procedia Computer Science*, *234*(2023), 96–105. https://doi.org/10.1016/j.procs.2024.02.156

Aulia, S. (2021). Klasterisasi Pola Penjualan Pestisida Menggunakan Metode K-Means Clustering (Studi Kasus Di Toko Juanda Tani Kecamatan Hutabayu Raja). *Djtechno: Jurnal Teknologi Informasi*, *1*(1), 1–5. https://doi.org/10.46576/djtechno.v1i1.964

Aziz, F. N. R. F. J., Setiawan, B. D., & Arwani, I. (2018). Implementasi Algoritma K-Means untuk Klasterisasi Kinerja Akademik Mahasiswa. *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer*, *2*(6), 2243–2251.

Elimar, T., Lestari, A. A., Susyanti, S., Fitri, M. R., & Fatmayanti, E. (2024). Strategi Promosi Penerimaan Mahasiswa Baru di Lingkungan Perguruan Tinggi Keagamaan Islam Negeri. *Leader: Jurnal Manajemen Pendidikan Islam*, *2*(1), 176–185. https://doi.org/10.32939/ljmpi.v2i1.3790

Fadilah, Z. R., & Wijayanto, A. W. (2023). Perbandingan Metode Klasterisasi Data Bertipe Campuran: One-Hot-Encoding, Gower Distance, dan K-Prototype Berdasarkan Akurasi (Studi Kasus: Chronic Kidney Disease Dataset). *Journal of Applied Informatics and Computing*, *7*(1), 57–67. https://doi.org/10.30871/jaic.v7i1.5857

Fitri, E. M., Suryono, R. R., & Wantoro, A. (2023). Klasterisasi Data Penjualan Berdasarkan Wilayah Menggunakan Metode K-Means Pada Pt Xyz. *Jurnal Komputasi*, *11*(2), 157–168. https://doi.org/10.23960/komputasi.v11i2.12582

Izzati, N., Talib, M., Aini, N., Majid, A., & Sahran, S. (2023). *applied sciences Identification of Student Behavioral Patterns in Higher Education Using K-Means Clustering and Support Vector Machine*. https://doi.org/https://doi.org/10.3390/app13053267

Lashiyanti, A. R., Munthe, I. R., & ... (2023). Optimisasi Klasterisasi Nilai Ujian Nasional dengan Pendekatan Algoritma K-Means, Elbow, dan Silhouette. *Jurnal Ilmu Komputer ...*, *6*, 14–20. http://ejournal.sisfokomtek.org/index.php/jikom/article/view/1550%0Ahttp://ejournal.sisfokomtek.org/index.php/jikom/article/download/1550/1026

Maori, N. A., & Evanita, E. (2023). Metode Elbow dalam Optimasi Jumlah Cluster pada K-Means Clustering. *Simetris: Jurnal Teknik Mesin, Elektro Dan Ilmu Komputer*, *14*(2), 277–288. https://doi.org/10.24176/simet.v14i2.9630

Nugroho, M. R., Hendrawan, I. E., & Purwantoro, P. P. (2022). Penerapan Algoritma K-Means Untuk Klasterisasi Data Obat Pada Rumah Sakit ASRI. *Nuansa Informatika*, *16*(1), 125–133. https://doi.org/10.25134/nuansa.v16i1.5294

Poiema, Z., & Mantohana, C. (2025). *Segmentation of Promotion Target Areas Using the K-Means Clustering Algorithm at Universities Segmentasi Wilayah Target Promosi Menggunakan Algortima K-Means Clustering pada Perguruan Tinggi*. *5*(October), 1160–1171. https://doi.org/https://doi.org/10.57152/malcom.v5i4.2125

Purnama, C., Witanti, W., & Nurul Sabrina, P. (2022). Klasterisasi Penjualan Pakaian untuk Meningkatkan Strategi Penjualan Barang Menggunakan K-Means. *Journal of Information Technology*, *4*(1), 35–38.

https://doi.org/10.47292/joint.v4i1.79

Ros, F., Riad, R., & Guillaume, S. (n.d.). *PDBI : a Partitioning Davies-Bouldin Index for Clustering Evaluation*. https://doi.org/https://doi.org/10.1016/j.neucom.2023.01.043

Shutaywi, M. (2021). *Silhouette Analysis for Performance Evaluation in Machine*. 1–17. https://doi.org/https://doi.org/10.3390/e23060759

Solichin, A., & Khairunnisa, K. (2020). Klasterisasi Persebaran Virus Corona (Covid-19) Di DKI Jakarta Menggunakan Metode K-Means. *Fountain of Informatics Journal*, *5*(2), 52. https://doi.org/10.21111/fij.v5i2.4905

Sulistiyawan, E., Hapsery, A., & Arifahanum, L. J. A. (2021). PERBANDINGAN METODE OPTIMASI UNTUK PENGELOMPOKAN PROVINSI BERDASARKAN SEKTOR PERIKANAN DI INDONESIA (Studi Kasus Dinas Kelautan dan Perikanan Indonesia). *Jurnal Gaussian*, *10*(1), 76–84. https://doi.org/10.14710/j.gauss.v10i1.30936

Virgo, I., Defit, S., & Yuhandri, Y. (2020). Klasterisasi Tingkat Kehadiran Dosen Menggunakan Algoritma K-Means Clustering. *Jurnal Sistim Informasi Dan Teknologi*, *2*, 23–28. https://doi.org/10.37034/jsisfotek.v2i1.17

Wahyu, A., & Rushenda. (2022). Klasterisasi Dampak Bencana Gempa Bumi. *JEPIN (Jurnal Edukasi Dan Penelitian Informatika)*, *8*(1), 175–179. https://doi.org/10.26418/jp.v8i1

Yaumi, A. S., Zulfiqkar, Z., & Nugroho, A. (2020). Klasterisasi Karakter Konsumen Terhadap Kecenderungan Pemilihan Produk Menggunakan K-Means. *JOINTECS (Journal of Information Technology and Computer Science)*, *5*(3), 195. https://doi.org/10.31328/jointecs.v5i3.1523

Yunita, F. (2018). Penerapan Data Mining Menggunkan Algoritma K-Means Clustring Pada Penerimaan Mahasiswa Baru. *Sistemasi*, *7*(3), 238. https://doi.org/10.32520/stmsi.v7i3.388