Inform : Jurnal Ilmiah Bidang Teknologi Informasi dan Komunikasi
Vol.9 No.2 July 2024, P-ISSN : 2502-3470, E-ISSN : 2581-0367

Prospective New College Student Dashboard: Insights from K-Means Clustering with Principal Component Analysis

Dyah Putri Rahmawati[1], Sri Hidayati[2], Paramaditya Arismawati [3], Ahmad Wali Satria Bahari Johan [4]

[1,4]*Department of Informatics, School of Computing, Telkom University, Indonesia*

[2]*Department of Information System, School of Industrial Engineering, Telkom University, Indonesia*

[3]*Department of Industrial Engineering, School of Industrial Engineering, Telkom University, Indonesia*

[1]dyahputri@telkomuniversity.ac.id (*)

[2,3,4] [srihidayatisi, paramadityaars, ahmadsatria]@telkomuniversity.ac.id

Received: 2024-06-19; Accepted: 2024-07-11; Published: 2024-07-22

Abstract— Higher education institutions currently compete to attract prospective students, necessitating the implementation of effective and efficient promotion strategies. Universities can create effective promotion strategies by considering the characteristics of prospective students. The clustering method is An approach to understanding prospective students' characteristics. However, clustering analysis with numerous attributes faces the issue of the curse of dimensionality. This research aims to overcome the curse of dimensionality in clustering by applying the K-means clustering method, which is enhanced through dimensionality reduction using Principal Component Analysis (PCA). The enhanced K-means method was applied to clustering data on prospective students at Telkom University Surabaya. The data used in this research pertains to prospective students interested in Telkom University Surabaya (TUS). Attributes include school origin, school province, domicile district/city, type of registration pathway, school type, and choice of study program. This research indicates that utilizing K-means clustering with PCA yielded superior cluster outcomes when evaluated against the Davies-Bouldin Index and Calinski-Harabasz Index, surpassing the performance of ordinary K-means clustering. The cluster analysis also shows that the ideal number of clusters is 3, using three principal components (PCs). The outcomes of the K-means clustering with PCA are incorporated into a dashboard that visually displays comprehensive information about the clusters. This dashboard simplifies examining how potential new students are geographically spread out, alongside how clusters are distributed across various study programs, school types, registration routes, and locations in districts/cities. The analytical data exploration on the dashboard can be utilized to address Business Questions Formulation related to the characteristics of prospective new students based on clustering results at Telkom University Surabaya.

Keywords— Clustering; K-Means; Principle Component Analysis; Dashboard; Prospective New Students.

I. INTRODUCTION

Students constitute a vital asset for higher education institutions, particularly private institutions reliant on tuition fees as a primary source of income. The prevailing trend is that most high school and vocational school graduates prioritize admission to public higher education institutions over private ones. In Indonesia, in 2022, there will be 4,522 higher education institutions, of which 4,140 will be private higher education institutions [1]. The substantial presence of private higher education institutions intensifies competition for attracting high-quality prospective students. The abundance of options in higher education institutions necessitates institutions to devise strategies to captivate the attention of prospective students. Higher education institutions must thoughtfully allocate resources to develop strategies to capture prospective students' interest. They must consider various factors influencing prospective student preferences when formulating promotional strategies for new student admissions. This situation challenges Telkom University Surabaya (TUS) as a private higher education institution to establish itself as a top choice for prospective students. Effective promotional strategies must persuade prospective students that Telkom University Surabaya offers services, opportunities, and educational quality that rival other higher education institutions.

Since its founding in 2018, new student admissions at TUS have steadily increased. However, this growth has not met the targets set by the Telkom Education Foundation. In the 2022 and 2023 admissions cycles, TUS achieved 72.6% and 79% of its target of 1,000 new students, respectively. The shortfall in meeting these targets necessitates a thorough evaluation and strategic response to identify effective solutions. One potential method to increase the number of prospective new students is through targeted promotional activities employing consistent and efficient strategies. Before developing these promotional strategies, institutions must understand the characteristics of their prospective student targets. Data mining techniques are valuable for this purpose, as they analyze data to extract meaningful insights. Clustering, a core function of data mining, helps to identify groups of similar objects. Through clustering analysis, institutions can identify key characteristics and tendencies of prospective students. Clustering analysis can reveal important information, such as high-potential student contributors' geographical distribution and school origins. These popular and underperforming study programs may require special promotions and potential entry pathways. This knowledge will enable TUS to tailor its promotional

DOI : https://doi.org/10.25139/inform.v9i2.8462

efforts more effectively, enhancing its ability to attract and enroll new students.

K-means is a frequently used clustering method known for its good performance. As in a study by Hidayati [2], village grouping data in Surabaya based on poverty indicators was compared using the K-Means, Fuzzy C-Means, Fuzzy Gustafson Kessel, and DBSCAN methods. The results indicated that K-Means and DBSCAN were the best methods based on the Cluster Sum of Squares and Average Silhouette scores. K-Means is an unsupervised learning algorithm that does not require annotated data [3]. This algorithm is highly effective in grouping large amounts of data, providing relatively fast and efficient computing times [4]. However, several clustering methods, including K-Means, encounter issues with the curse of dimensionality when applied to data with many attributes [5]. Common problems include decreased accuracy, poor cluster quality, and extended computing times. Another significant challenge when using high-dimensional data is visually presenting clustering results [6]. One solution to this problem is dimensionality reduction, simplifying the visualization of clustering results. Dimensionality reduction can be performed using Principal Component Analysis (PCA) [7]. PCA reduces high-dimensional data to lower dimensions with minimal information loss [5]. This reduction greatly aids in visualizing clustering results by presenting complex data more clearly and concisely. In previous research [8]–[13], prospective new student data was clustered using the ordinary k-means method. However, there is no Exploratory Data Analysis regarding the characteristics of prospective new students based on the clustering results.

Furthermore, the visual clustering results are still difficult to understand. In the K-means clustering study conducted by Susilowati [14], the RFM method was utilized to analyze schools that potentially attract prospective new students specifically. Meanwhile, this study examines a broader range of aspects, including study programs, types of schools, geographical distribution, and registration routes. This comprehensive analysis allows for dimensionality reduction using Principal Component Analysis (PCA).

The results of clustering in this study, which involve grouping data or information, can also be visually represented in graphs, maps, charts, and bars. These visual representations are typically displayed on dashboard pages, a practice known as data visualization [15]. Data visualization is crucial because the human brain processes visual information more efficiently than numerical data [16]. This dashboard facilitates the exploration of analytical data and storytelling characteristics from cluster results, making the information more accessible to understand. As an advanced solution, the dashboard can also assist campus admissions teams in determining further steps for promotional strategies aimed at prospective students.

II. Research Methodology

The dataset utilized in this research pertains to prospective new students at Telkom University Surabaya (TUS) for the years 2021-2023. This data was chosen because public awareness of Telkom University Surabaya has surged since 2021, significantly increasing the number of new student applicants. Additionally, several new study programs have been established since 2021, making data from 2021 to 2023 more representative of the current situation. The attributes extracted from this dataset include School Origin, School Province, Domicile District/City, Registration Type, School Type, and Choice of Study Program. The data undergoes a preprocessing stage to ensure it is ready for analysis. This study employed a combination of the K-Means algorithm and dimensionality reduction using Principal Component Analysis (PCA). The results from the clustering analysis are subsequently used to develop a visualization dashboard. Figure 1 depicts the integration of the K-Means algorithm with the PCA method.

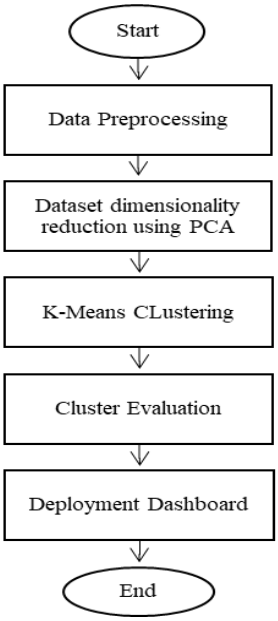

Figure 1. Flowchart of K-Means with Dimension Reduction Using PCA.

A. *Dimensionality reduction using PCA*

Before performing the clustering process, the dataset undergoes dimensionality reduction using PCA. PCA generates a new set of dimensions ranked according to the data variance [17]. This method produces principal components derived from the decomposition of eigenvalues and eigenvectors from the covariance matrix. The stages of the PCA method are illustrated in Figure 2. The steps in PCA) [18]:

1) Calculating the Covariance Matrix: The covariance matrix measures how much two variables change together. If the dataset consists of standardized variables, the covariance matrix is computed as Equation (1), where the Z variable is the standardized data matrix and n is the number of samples.

$$C = \frac{1}{n-1} Z^T Z \quad (1)$$

DOI : https://doi.org/10.25139/inform.v9i2.8462

2) Calculating Eigenvalues and Eigenvectors: The covariance matrix obtained is then used to compute eigenvalues and eigenvectors. Eigenvalues provide information about the variability explained by each principal component. This calculation is based on solving the characteristic as Equation (2), where the λ variable is the eigenvalue and the v variable is the corresponding eigenvector.

$$Cv = \lambda v \tag{2}$$

3) Sorting Eigenvalues: Eigenvalues are sorted in descending order (from largest to smallest). This order indicates the most significant principal components that explain the variability in the data. Larger eigenvalues correspond to components that explain more variance.

4) Forming Principal Components: Principal components are eigenvectors sorted according to the eigenvalues obtained earlier. These eigenvectors form a new basis for the data, where each principal component is a linear combination of the original variables.

5) Forming a New Dataset: A new dataset is formed by multiplying the original standardized data by the selected eigenvectors (principal components).

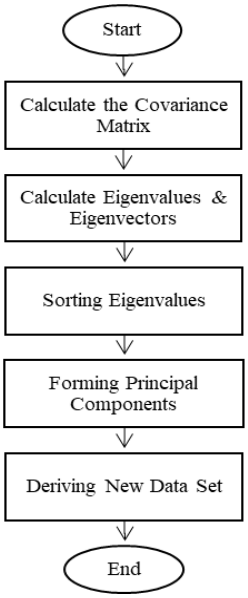

Figure 2. Flowchart of Dimension Reduction with PCA

B. K-Means Clustering

Once the dataset has been dimensionally reduced using PCA, the next step is to apply K-means clustering. The steps of the K-Means algorithm are outlined[19]:

1) Determine the Number of Clusters (k): Decide on the number of clusters *k* to be formed.

2) Initialize Cluster Centers: Randomly select *k* points from the dataset to serve as initial cluster centroids.

3) Calculate Euclidean Distance: For each data point, compute the Euclidean distance to each cluster centroid using the formula in Equation (3). where the d_{pq} variable is the Euclidean distance between objects *p* and *q*, the s variable is number of dimensions in the dataset after PCA, the x_{pr} variable is coordinates of object *p* in dimension *r*, and the x_{qr} variable is coordinates of object *q* in dimension *r*.

$$d_{pq} = \sqrt{\sum_{r=1}^{s} (x_{pr} - x_{qr})^2} \tag{3}$$

4) Assign Data Points to Clusters: Assign each data point to the cluster whose centroid is closest, based on the computed distances.

5) Update Cluster Centers: After assigning all data points, recalculate the centroids of the clusters by taking the mean of all data points assigned to each cluster.

6) Iterate Until Convergence: Repeat steps 3 to 5 iteratively until no data points change clusters between iterations, indicating convergence. Convergence is typically determined by assessing whether cluster memberships remain unchanged after an iteration.

C. Cluster Evaluation

In this research, the evaluation of clustering results is crucial to ensure their quality. The cluster evaluation methods employed [20]:

1) Elbow Method: This method assesses the variance decrease within each cluster as the number of clusters increases. A significant bend or "elbow" indicates the optimal number of clusters on a variance plot versus several clusters [21].

2) Silhouette Method: By comparing each data point's fit to its cluster, this approach calculates how well it fits to each other cluster. A higher silhouette score denotes a well-clustered set of data points [22].

3) Davies-Bouldin Index: This method evaluates cluster quality by considering both the distance between clusters and the dispersion within clusters. Lower values of the Davies-Bouldin index indicate better cluster separation and cohesion. [23].

4) The Calinski-Harabasz Index: This index, also known as the Variance Ratio Criterion, measures the total variance ratio between and within clusters. A higher value indicates better clustering because it shows that the clusters are more compact and more separated from each other [24].

Additionally, to compare cluster quality, the research evaluates the performance of K-Means clustering alone and K-Means clustering combined with PCA dimension reduction.

D. Dashboard

The findings from K-Means clustering after PCA dimension reduction are displayed visually through a dashboard. This dashboard, developed on the Tableau Public platform, enhances the ease of interpreting data and aids in

DOI : https://doi.org/10.25139/inform.v9i2.8462

making well-informed decisions. The clustering results are imported into Tableau Public, where a visualization process is carried out by creating various graphs and charts visually illustrating the clustering results [25]. This process involves selecting the appropriate visualization types, adjusting the display to improve readability, and setting interactivity to make it easier for users to explore the data [26].

III. Result and Discussion

The data utilized in this study includes details about prospective new students from 2021 to 2023. This dataset underwent initial preprocessing steps. Preprocessing begins by loading and reading the dataset into a data frame for further analysis. The data is then separated by year, allowing for a more focused and relevant analysis of each period. Next, missing values are checked and addressed by matching and merging data between the 'School Province' and 'Domicile Province' columns, assuming that participants live and attend school in the same province. This method fills in missing values in one column using the values from the other column, thereby increasing the completeness of the data. Duplicate data is removed, and a subset of the data frame is created with selected columns: School Origin, School Province, Domicile District/City, Registration Type, School Type, and Choice of Study Program. Following this consolidation, the data underwent the Label Encoder technique. Label Encoding is a method used in data processing to convert categorical labels into numerical representations by assigning a unique integer to each category. This transformation is crucial for enabling machine learning algorithms to effectively process qualitative data that was previously unsuitable for analysis. An illustrative example of applying Label Encoding to the dataset is depicted in Table I. Following these preprocessing procedures, the dataset was prepared for subsequent Principal Component Analysis (PCA).

Table I
Example of Applying Label Encoding to the Dataset

School Type	Encoded School Type	Study Program	Encoded Study Program
Public	0	Informatics	1
Private	1	Information Systems	4
Public	0	Digital Business	0
Private	0	Information Systems	4

A. Dimensionality reduction using PCA

In this study, Principal Component Analysis (PCA) was conducted using the sklearn.decomposition library, comparing analyses with 2 and 3 principal components (PCs). The PCA results indicated variations in explained variance between the two configurations. With 2 PCs, the explained variance ratios were 0.987826 and 0.012082, resulting in a total explained variance of 0.999908. Meanwhile, using 3 PCs, the explained variance ratios were 0.987826, 0.012082, and 0.000045, yielding a slightly higher total explained variance of 0.999953. Despite the modest increase in explained variance, selecting 3 PCs was preferred due to its significant advantages in visualizing clustering results in three dimensions. 3D visualization allows for more precise and comprehensive data representation, aids in capturing complex patterns and structures, and offers interactive capabilities to better understand cluster distribution and separation. The results of dimension reduction using 3 PCs are detailed in Table II.

Table II
Sample of The Results of Dimension Reduction Using 3 PCs

Prospective student (*i*)	PC-1	PC-2	PC-3
1	186.8548	-177.828	-12.6777
2	395.9267	-122.247	-3.5418
3	-352.972	39.8999	19.3733
4	638.1251	-132.189	18.6049
5	-100.185	230.3305	-3.8406
6	-278.765	-87.7785	-2.8669
7	-273.523	73.334	-2.3855

B. K-Means Clustering

Clustering was performed on two different types of data: the original dataset and the dataset obtained through PCA, which resulted in data sets with 1 PC, 2 PCs, and 3 PCs. K-Means algorithm was employed for clustering, initializing centroids randomly. The number of clusters was determined using the Elbow Method, Silhouette, Davies-Bouldin, and The Calinski-Harabasz Index.

The original dataset and the PCA results were then grouped using the k-means clustering. The initial approach to determine the number of clusters is the elbow method, a technique used to determine the optimal number of clusters in cluster analysis. It aims to identify where adding more clusters does not significantly reduce the objective function value. This study applied K-Means with clusters ranging from 2 to 15. In determining the number of clusters using the elbow method, we examine the decrease in inertia value as an indicator to find the point where the decrease in inertia value is no longer significant.

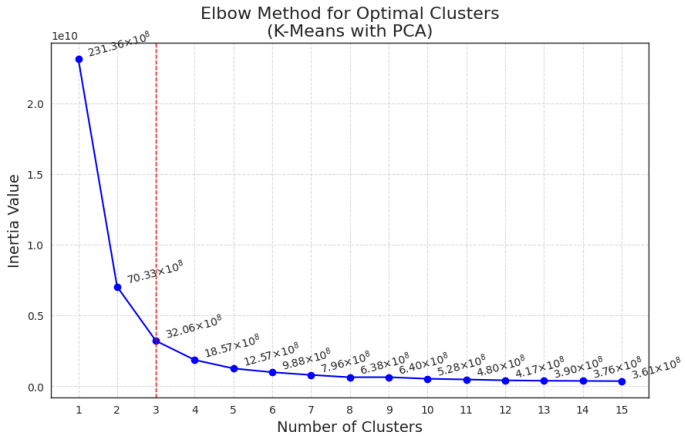

Figure 3. Evaluation Using the Elbow Method for K-Means with PCA

In determining the number of clusters, the elbow method was employed using the decrease in inertia as a reference [27]. This method aims to identify the point where the decrease in inertia slows down significantly, forming an elbow-like bend. Inertia measures the spread of data points within each cluster; lower inertia values indicate better data clustering. When

DOI : https://doi.org/10.25139/inform.v9i2.8462

adding more clusters no longer results in a significant decrease in inertia, this point indicates the optimal number of clusters. From the analysis results, both in the original data and the PCA results, the elbow points were observed to be in n-cluster 3 (see Figure 3 and Figure 4), where the inertia decreases significantly until that point and then stabilizes without much difference, suggesting that further cluster additions do not improve substantially data grouping. However, the analysis does not stop here. The next step involves conducting additional studies to confirm that the optimal number of clusters is 3 using other analytical methods.

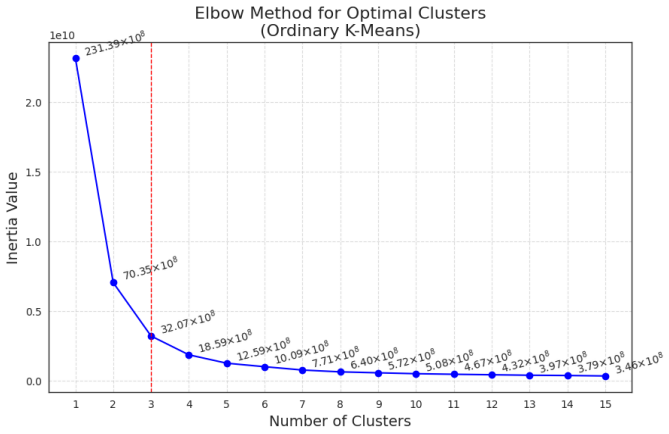

Figure 4. Evaluation Using the Elbow Method for Ordinary K-Means.

Based on the Silhouette Score results in Table III, it can be seen that 2 clusters are slightly better than 3 clusters for all methods used. The highest value is found in K-Means with PCA clustering with 2 PCs in 2 clusters of 0.5711487.

TABLE III
SAMPLE OF THE RESULTS OF DIMENSION REDUCTION USING 3 PCS

Method	Number of Clusters	Silhouette Score	Davies-Bouldin Index	Calinski-Harabasz Index
K-Means with PCA (3 PCs)	2	0.5710683	0.5829994	35028.64047
K-Means with PCA (3 PCs)	3	0.5458587	0.5608999	47383.45213
K-Means with PCA (2 PCs)	2	0.5711487	0.5829098	35033.8662
K-Means with PCA (2 PCs)	3	0.5458581	0.5609001	47383.42874

The silhouette score ranges from -1 to 1, where values closer to 1 indicate that the data points are well placed within their cluster. Conversely, values close to -1 suggest that the data point may be better suited to being in another cluster, and values close to 0 indicate that the data point is near the boundary between two clusters. Thus, a higher Silhouette Score value in the 2-cluster configuration suggests that the data is better grouped in that cluster. However, it is essential to consider the objectives of the analysis in determining the optimal number of clusters. Although higher silhouette values are often obtained by reducing the number of clusters, this may not always meet the desired analysis objectives. In this research, it is hoped that using more than 2 clusters will provide more informative and relevant results. Therefore, even though the Silhouette Score value for 2 clusters is higher, preference is given to using 3 clusters. This choice is further supported by other evaluation metrics, such as the Davies-Bouldin Index and the Calinski-Harabasz Index, which also show good performance for the 3-cluster configuration.

The evaluation results of K-Means clustering with PCA using the Davies-Bouldin Index show that the lowest Davies-Bouldin Index value was obtained from clustering with 3 PCs and 3 clusters (DB = 0.5608999). A lower Davies-Bouldin Index indicates better clustering performance. All results suggest that 3 clusters have a lower Davies-Bouldin Index value compared to two clusters, indicating improved cluster separation.

The Calinski-Harabasz Index evaluation results show that the highest value was obtained from clustering with 3 PCs and 3 clusters (CH = 47383.45213). A higher Calinski-Harabasz Index indicates more well-defined clusters. The 3-cluster configuration once again outperforms the 2-cluster configuration, with the 3-PC setting being slightly superior.

Overall, 3 PCs with 3 clusters can be considered the best clustering solution in K-Means clustering with PCA because it has better Davies-Bouldin Index and Calinski-Harabasz Index values, even though the Silhouette Score is slightly lower. This indicates that the clusters are more separated and compact, which suggests better clustering.

To determine the effect of the dataset dimension reduction stage before the data is clustered using the K-means algorithm, the best results of K-means clustering with PCA are compared with the results of clustering with ordinary K-means. This also proves that PCA can enhance the clustering performance compared to the ordinary k-means used in previous studies. Table IV compares the Silhouette Score, Davies-Bouldin and The Calinski-Harabasz Index.

TABLE IV
SAMPLE OF THE RESULTS OF DIMENSION REDUCTION USING 3 PCS

Method	Number of Clusters	Silhouette Score	Davies-Bouldin Index	Calinski-Harabasz Index
Ordinary K-Means	2	0.5708939	0.5833175	35024.08474
Ordinary K-Means	3	0.5455901	0.5611901	47368.71152
K-Means with PCA (3 PCs)	3	0.5458587	0.5608999	47383.45213

The Original K-means method and K-means with PCA demonstrate that utilizing 3 clusters results in slightly improved separation according to evaluations based on the Calinski-Harabasz Index and the Davies-Bouldin Index. Combining three principal components (PCs) with 3 clusters is the optimal clustering solution in K-means clustering with PCA. This conclusion is drawn from its superior Davies-Bouldin Index and Calinski-Harabasz Index values despite a marginally lower Silhouette Score than the results obtained from ordinary K-means analysis. This indicates that the clusters are more distinct and cohesive, which signifies higher-quality clustering. Table V shows a sample of the clustering results. This data will create an analytical data

DOI : https://doi.org/10.25139/inform.v9i2.8462

exploration on the dashboard, enabling visualization and storytelling of the characteristics derived from the clustering results.

TABLE V
SAMPLE DATA WITH CLUSTERING RESULTS USED IN THE EXPLORATION OF ANALYTICAL DATA

School Origin	School Type	School Province	Domicile District/City	Registration type	Study Program	Cluster Result
SMAN 3 TANGERANG	*NEGERI*	*Banten*	*Kota Tangerang Selatan*	*Jalur Unggulan*	*S1 Informatika*	2
PKBM IBNU ALI	*SWASTA*	*Jawa Timur*	*Kab. Sidoarjo*	*Jalur Akademik Rapor*	*S1 Sistem Informasi*	0
SMAN 7 KEDIRI	*NEGERI*	*Jawa Timur*	*Kota Kediri*	*Beasiswa OPES*	*S1 Bisnis Digital*	2
SMAN 1 SUNGAI LILIN	*NEGERI*	*Sumatera Selatan*	*Kab. Musi Banyuasin*	*Beasiswa APERTI BUMN*	*S1 Sistem Informasi*	2
SMAS DHARMAWANGSA	*SWASTA*	*Sumatera Utara*	*Kota Medan*	*Jalur Undangan Prioritas*	*S1 Bisnis Digital*	2
...	...	...	...	...	...	...
SMKN PP KALASEY	*NEGERI*	*Sulawesi Utara*	*Kota Manado*	*Jalur Undangan Prioritas*	*S1 Teknik Telekomunikasi*	1
SMKN PP KALASEY	*NEGERI*	*Sulawesi Utara*	*Kota Manado*	*Beasiswa KIP-K*	*S1 Teknik Telekomunikasi*	1

C. *Dashboard*

The k-Means clustering results using PCA with three principal components (PCs) and producing 3 clusters were then visualized via a dashboard using Tableau Public. This dashboard utilizes Tableau Public to interactively present information related to the formed clusters [28]. Visualizing the resulting clusters allows users to intuitively see patterns and relationships between data. With this dashboard, users can easily explore the characteristics of each cluster. Figure 5 presents the dashboard home page.

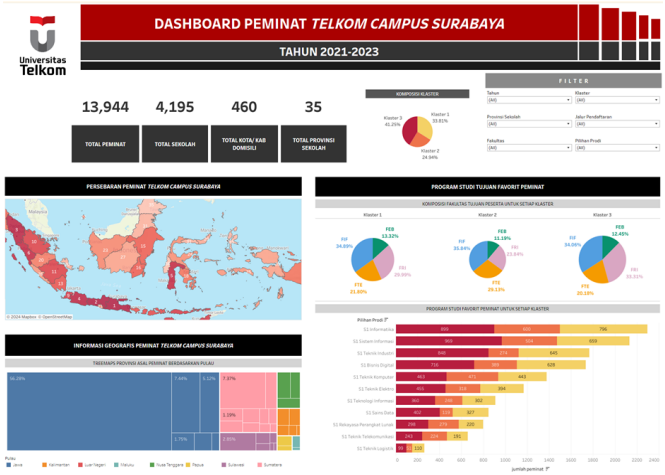

Figure 5. Dashboard Home Page

Figure 6 shows part of the front page of the dashboard. There are filter options that make it easy for users to filter data based on cluster, year, school province, registration type, study program, and faculty. The pie chart shows that Cluster 3 has the largest proportion, making up 41.25% of the total. Cluster 1 is the second largest, 33.81%, and Cluster 2 is the smallest, 24.94%.

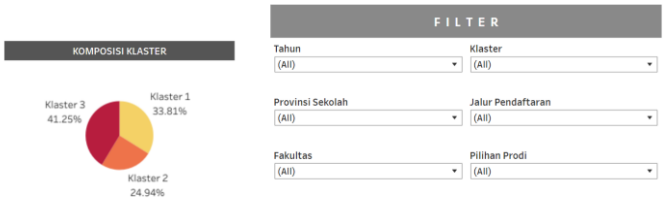

Figure 6. Cluster Composition and Filter Features

Figure 7 contains two visualizations regarding the geographic distribution of prospective students for Telkom Campus Surabaya. The first is Map Visualization. The map shows the distribution of prospective students across Indonesia, with provinces colour-coded based on the number of students. The darker the colour of the province, the greater the number of enthusiasts. A tooltip example is shown for Central Kalimantan, indicating it ranks 27th with 59 prospective students. The second is Treemap Visualization. The treemap displays the provinces of origin for prospective students, categorized by islands. The largest segment is from Java, contributing 56.21% of prospective students. Central Java within Java Island is highlighted with 7,037 prospective students, making up 7.49% of the total.

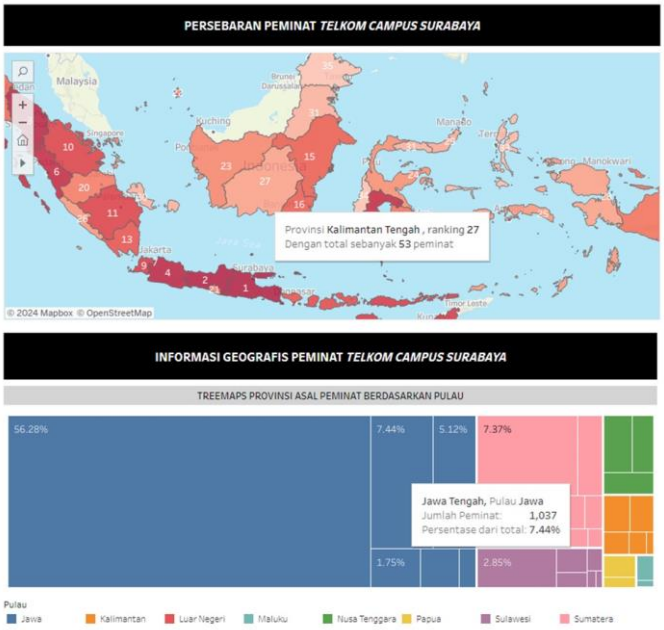

Figure 7. Geographic Distribution of Prospective Students

Figure 8 in the dashboard illustrates the distribution of study programs and faculties across three clusters, where yellow represents cluster 1, orange represents cluster 2, and red represents cluster 3. Insights from the visualization reveal that Cluster 1 is dominated by Logistics Engineering, comprising the highest percentage at 42.31%. Other significant study programs in this cluster include Digital

DOI : https://doi.org/10.25139/inform.v9i2.8462

Business (36.24%) and Informatics (34.68%). Cluster 2 shows the highest percentage of Software Engineering (35.01%), followed by other study programs such as Computer Engineering (34.20%) and Telecommunications Engineering (34.04%). Cluster 3 generally dominates across most study programs, with Industrial Engineering (47.99%), Data Science (47.41%), and Information Systems (45.45%) being the most dominant in this cluster.

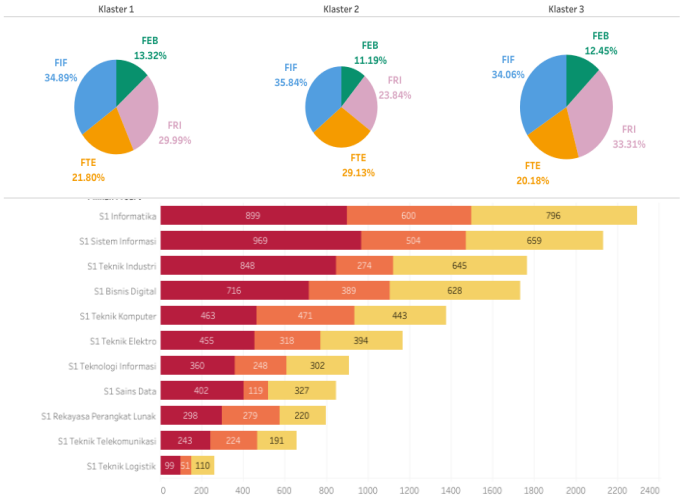

Figure 8. The Distribution of Study Programs

Figure 9 shows the distribution of clusters based on private and public-school types. Regarding school type, clusters public school types dominate 1 and 3, while private schools dominate cluster 2.

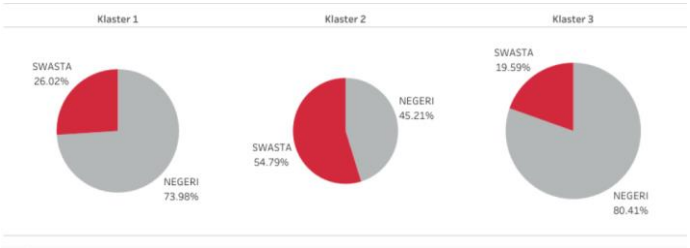

Figure 9. The Clusters Distribution Based on School Types

Based on Figure 10, it is evident that the registration types in Clusters 1 and 3 are predominantly chosen by prospective students applying through scholarship routes. Specifically, Cluster 1 shows more applicants via scholarship pathways (2.50K) than regular admission pathways (2.21K).

Similarly, Cluster 3 prefers scholarship applications (2.92K) over regular applications (2.84K), though the difference is marginal. Conversely, Cluster 2 in Figure 10 is characterized by a dominant preference for regular admission pathways (1.75K) over scholarship applications (1.73K), indicating a nearly balanced but slightly higher inclination towards regular admissions. This dashboard also displays the distribution of prospective new students from various schools and the number of prospective new students per city or district where they live, as presented in Figure 11.

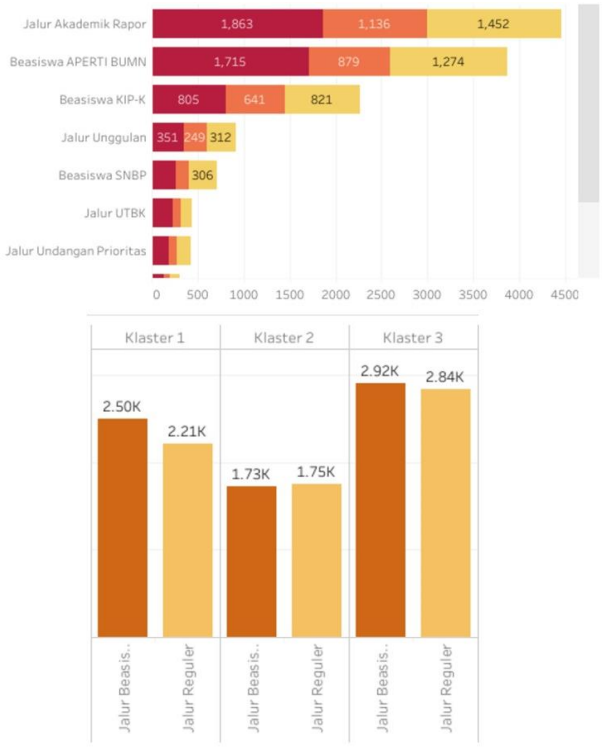

Figure 10. The Clusters Distribution Based on Registration Type

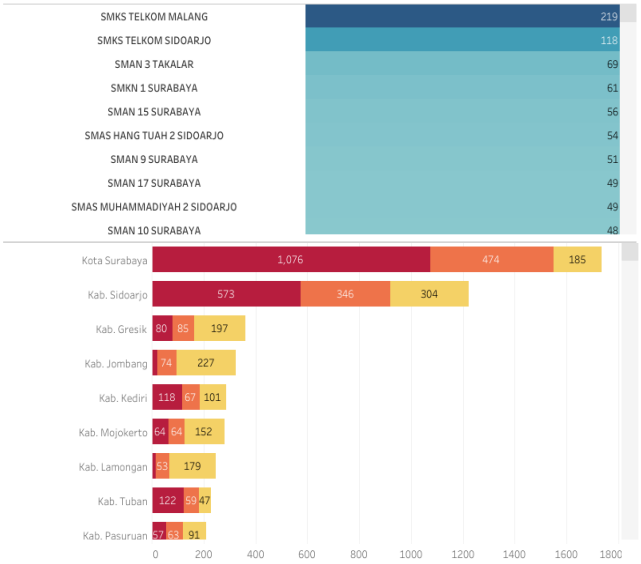

Figure 11. The Distribution of Prospective New Students Based on Schools and Domicile

IV. CONCLUSION

Applying K-means clustering with dimensionality reduction through Principal Component Analysis (PCA) on data from prospective new students at Telkom University Surabaya indicates that the optimal number of clusters is three, utilizing three principal components (PCs). Analysis using the Calinski-Harabasz Index, with a result of 0.5608999, and the Davies-Bouldin Index, with a result of 47383.45213, indicates

DOI : https://doi.org/10.25139/inform.v9i2.8462

that K-means with PCA produces better clusters compared to ordinary K-means.

Subsequently, the clustering results are integrated into a dashboard designed to visually present cluster information. This dashboard is an advanced solution that has not yet been widely developed in other research based on clustering results. It significantly aids in the exploratory data analysis and the storytelling of the characteristics of prospective new students at Telkom University Surabaya. This dashboard facilitates the analysis of prospective new students' geographic distribution and clusters based on study programs, school types, registration types, and district/city domiciles.

The outcomes of this research are anticipated to provide a clearer understanding of the characteristics of data clusters concerning prospective new students at Telkom University Surabaya. Moreover, in the future, this information can provide a basis for creating effective promotional policies and strategies to attract more prospective new students to Telkom University Surabaya.

ACKNOWLEDGEMENTS

This research was supported by funding from LPPM Telkom University. We want to express our sincere gratitude for their financial assistance and support. We also thank Rahmat Sigit Hidayat and Rendika Nurhartanto Suharto for their invaluable contributions and unwavering support throughout this project. Their expertise, dedication, and guidance have been instrumental in completing this work.

REFERENCES

[1] Kemendikbudristek, *Statistik Pendidikan Tinggi 2022*. Jakarta, 2022.
[2] S. Hidayati, A. T. Darmaliana, and R. Riski, "Comparison of K-Means, Fuzzy C-Means, Fuzzy Gustafson Kessel, and DBSCAN for Village Grouping in Surabaya Based on Poverty Indicators," *J. Pendidik. Mat.*, vol. 5, no. 2, p. 185, 2022.
[3] S. Sartikha, M. Maria, F. W. Sari, and N. Jannah, "Analisis Profil Mahasiswa Politeknik Negeri Batam dengan Teknik Data Mining Asosiasi dan Clustering," *J. Integr.*, vol. 8, no. 1, pp. 16–21, 2016.
[4] A. Tahta, S. Budi, and B. A. Ridho, "Analisa Perbandingan Metode Hierarchical Clustering, K-Means dan Gabungan Keduanya dalam Cluster Data," *Stud. kasus Probl. Kerja Prakt. Jur. Tek. Ind. ITS)". Inst. Teknol. Surabaya*, 2012.
[5] D. Hediyati and I. M. Suartana, "Penerapan Principal Component Analysis (PCA) Untuk Reduksi Dimensi Pada Proses Clustering Data Produksi Pertanian Di Kabupaten Bojonegoro," *JIEET (Journal Inf. Eng. Educ. Technol.*, vol. 5, no. 2, pp. 49–54, 2021.
[6] E. J. Nam, Y. Han, K. Mueller, A. Zelenyuk, and D. Imre, "Clustersculptor: A visual analytics tool for high-dimensional data," in *2007 IEEE Symposium on Visual Analytics Science and Technology*, 2007, pp. 75–82.
[7] S. Mulyaningsih and J. Heikal, "K-Means Clustering Using Principal Component Analysis (PCA) Indonesia Multi-Finance Industry Performance Before and During Covid-19," *APMBA (Asia Pacific Manag. Bus. Appl.*, vol. 11, no. 2, pp. 131–142, 2022.
[8] 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.
[9] N. A. Rahmalinda and A. Jananto, "Penerapan Metode K-Means Clustering Dalam Menentukan Strategi Promosi Berdasarkan Data Penerimaan Mahasiswa Baru," *J. Tekno Kompak*, vol. 16, no. 2, pp. 163–175, 2022.
[10] M. Farozi, "Metode K-Means Clustering Dalam Merancang Strategi Promosi Penerimaan Mahasiswa Baru Pada STIE Serelo Lahat," *J. Ilm. Inform. Glob.*, vol. 12, no. 2, 2021.
[11] B. Harahap and A. Rambe, "Implementasi K-Means Clustering Terhadap Mahasiswa yang Menerima Beasiswa Yayasan Pendidikan Battuta di Universitas Battuta Tahun 2020/2021 Studi Kasus Prodi Informatika," *Informatika*, vol. 9, no. 3, pp. 90–97, 2021.
[12] R. Budiman, "Penerapan Data Mining Untuk Menentukan Lokasi Promosi Penerimaan Mahasiswa Baru Pada Universitas Banten Jaya (Metode K-Means Clustering)," *ProTekInfo (Pengembangan Ris. dan Obs. Tek. Inform.*, vol. 6, pp. 6–14, 2019.
[13] M. R. Alhapizi, M. Nasir, and I. Effendy, "Penerapan Data Mining Menggunakan Algoritma K-Means Clustering Untuk Menentukan Strategi Promosi Mahasiswa Baru Universitas Bina Darma Palembang," *J. Softw. Eng. Ampera*, vol. 1, no. 1, pp. 1–14, 2020.
[14] D. Susilowati, H. Hairani, I. P. Lestari, K. Marzuki, and L. Z. A. Mardedi, "Segmentasi Lokasi Promosi Penerimaan Mahasiswa Baru Menggunakan Metode RFM dan K-Means Clustering," *MATRIK J. Manajemen, Tek. Inform. dan Rekayasa Komput.*, vol. 21, no. 2, pp. 275–282, 2022.
[15] D. Y. Liliana, I. Ermis, A. R. Zain, and N. A. Azza, "K-Means Clustering untuk Visualisasi Informasi Pemanfaatan Aplikasi Deteksi Dini Depresi," in *Seminar Nasional Inovasi Vokasi*, 2022, vol. 1, pp. 116–123.
[16] C. Ware, *Information visualization: perception for design*. Morgan Kaufmann, 2019.
[17] B. M. S. Hasan and A. M. Abdulazeez, "A review of principal component analysis algorithm for dimensionality reduction," *J. Soft Comput. Data Min.*, vol. 2, no. 1, pp. 20–30, 2021.
[18] C. Ding and X. He, "K-means clustering via principal component analysis," in *Proceedings of the twenty-first international conference on Machine learning*, 2004, p. 29.
[19] K. P. Sinaga and M.-S. Yang, "Unsupervised K-means clustering algorithm," *IEEE access*, vol. 8, pp. 80716–80727, 2020.
[20] S. Dewi and M. A. I. Pakereng, "Implementasi Principal Component Analysis pada K-Means untuk Klasterisasi Tingkat Pendidikan Penduduk Kabupaten Semarang," *JIPI (Jurnal Ilm. Penelit. dan Pembelajaran Inform.*, vol. 8, no. 4, pp. 1186–1195, 2023.
[21] M. A. Syakur, B. K. Khotimah, E. M. S. Rochman, and B. D. Satoto, "Integration k-means clustering method and elbow method for identification of the best customer profile cluster," in *IOP conference series: materials science and engineering*, 2018, vol. 336, p. 12017.
[22] S. P. Lima and M. D. Cruz, "A genetic algorithm using Calinski-Harabasz index for automatic clustering problem," *Rev. Bras. Comput. Apl.*, vol. 12, no. 3, pp. 97–106, 2020.
[23] Y. A. Wijaya, D. A. Kurniady, E. Setyanto, W. S. Tarihoran, D. Rusmana, and R. Rahim, "Davies bouldin index algorithm for optimizing clustering case studies mapping school facilities," *TEM J*, vol. 10, no. 3, pp. 1099–1103, 2021.
[24] X. Wang and Y. Xu, "An improved index for clustering validation based on Silhouette index and Calinski-Harabasz index," in *IOP Conference Series: Materials Science and Engineering*, 2019, vol. 569, no. 5, p. 52024.
[25] S. Batt, T. Grealis, O. Harmon, and P. Tomolonis, "Learning Tableau: A data visualization tool," *J. Econ. Educ.*, vol. 51, no. 3–4, pp. 317–328, 2020.
[26] R. Akbar and M. Octaviany, "Perancangan visualisasi dashboard dan clustering dengan menerapkan business intelligence pada dinas DPMPTSP kabupaten Dharmasraya," *JEPIN (Jurnal Edukasi dan Penelit. Inform.*, vol. 7, no. 3, pp. 340–350, 2021.
[27] A. Karna and K. Gibert, "Automatic identification of the number of clusters in hierarchical clustering," *Neural Comput. Appl.*, vol. 34, no. 1, pp. 119–134, 2022.
[28] N. L. R. A. Nur Laita Rizki Amalia, A. A. S. Ahmad Afif Supianto, N. Y. S. Nanang Yudi Setiawan, V. Z. Vicky Zilvan, A. R. Y. Asri Rizki Yuliani, and A. R. Ade Ramdan, "Student Academic Mark Clustering Analysis and Usability Scoring on Dashboard Development Using K-Means Algorithm and System Usability Scale," *J. Ilmu Komput. Dan Inf.*, vol. 14, no. 2, pp. 137–143, 2021.

DOI : https://doi.org/10.25139/inform.v9i2.8462