ISSN 2085-4552

Identifying Student Competency Patterns in Informatics and Computer Engineering Education at Universitas Sebelas Maret using K-Means Clustering for Academic Guidance

Ratih Friska Dwi Andini[1], Febri Liantoni[2], Aris Budianto[3]
[1,2,3]Informatics Engineering and Computer Education, Sebelas Maret University
[1]ratihfrisska@gmail.com, [2]febri.liantoni@gmail.com, [3]arisbudianto@staff.uns.ac.id

Accepted 25 June 2025
Approved 30 June 2025

***Abstract*— This study evaluates the effectiveness of the K-Means algorithm in clustering student competencies using course score data from students in the Informatics and Computer Engineering Education program at an Indonesian public university. The K-Means algorithm was applied to group students into five distinct competency clusters based on their academic grade patterns. The model's performance was measured using the Silhouette Score, which resulted in a value of 0.3489, indicating a moderate quality of cluster separation. The results suggest potential applications for a student recommendation system for choosing elective courses and as an evaluation tool for the study program. Key limitations of this approach include the algorithm's sensitivity to the initial placement of cluster centers and the dependency on selecting the appropriate number of clusters (k) for optimal results.**

***Index Terms*— Academic performance; Clustering; Data mining; K-Means; Student competence.**

I. INTRODUCTION

Students of the Informatics and Computer Engineering Education study program at Sebelas Maret University will be faced with choosing a course of interest or concentration in the sixth semester. The available areas of concentration include desktop programming, web programming, video game software development, network administrator, and multimedia. Each student has different abilities and competencies in certain fields. However, based on observations, some students do not fully understand their strengths or weaknesses in these areas. This often causes them difficulty in determining the choice of study field, interest, or career path that suits their potential.

At the tertiary level, knowledge of areas of interest, expertise or competence is important in the student learning process. However, low awareness of self-potential often hinders students in determining the right actions to achieve academic and career success. Based on Masturina's research, students who have known their career interests early on find it easier to determine the actions to be taken to develop expertise and skills that are relevant to the needs of the world of work [1].

The abundant academic data of students holds great potential to be explored to help the decision-making process. In this context, data mining technology and machine learning algorithms, such as K-Means, can be utilized to analyze data and identify student competency patterns. The K-Means algorithm is known to be able to group data into certain groups based on similar characteristics. The use of clustering algorithms such as K-Means in education has shown significant results. Previous studies Mohd Talib et al., and Tuyishimire et al., show that this algorithm can identify student performance patterns based on academic data, thus helping educational institutions better understand student needs and potential [2] [3]. This study aims to apply the K-Means algorithm in the process of grouping students based on their academic data patterns. This data-based approach is expected to provide insight to study programs in supporting more appropriate decision making, as well as helping students understand their own competencies to develop their potential optimally.

II. METHODOLOGY

In this study, we have determined the object of research, namely students of the Informatics and Computer Engineering Education study program at a public university in Indonesia. This study uses an exploratory quantitative approach to analyze student competency patterns with the K-Means algorithm.

A. Research Stages

This research was carried out in several stages:

1) Analysis: At the initial stage, research data needs were identified.

2) Design: The process flow for clustering with the K-Means algorithm was designed.
3) Development: A clustering model was created using the Python programming language by leveraging the Scikit-learn library.
4) Determining the Number of Clusters (k): This stage aimed to determine the optimal number of clusters (k) before the final model was implemented. This process involved two methods:
    a. Elbow Method: This method was used to find a potential number of clusters by plotting the explained variance as a function of the number of clusters, then identifying the "elbow" point where adding more clusters no longer yields significant returns
    b. Validation with Silhouette Score: The number of clusters suggested by the Elbow Method was then validated using the Silhouette Score. This metric provides a quantitative measure of clustering quality, where a higher score indicates better-defined clusters
5) Implementation: The results from the formed clustering model were then applied in a decision support system
6) Evaluation: The final stage is to evaluate the quality of the resulting clustering model. This evaluation uses the Silhouette Score metric to measure how well-defined the final clusters are.

B. Data Source

At this stage, student data of the PTIK Study Program at University X was collected, including course values representing various fields of competence, from semester 1 to semester 5, with a total of 72 students. The data contains attributes such as student identification numbers, student names, and grades for courses that include theoretical and practicum aspects, such as Basic Multimedia, Structured Programming, Web Programming, 2D Animation Engineering, and Computer Network Security.

C. Selection

The selection stage aims to choose the most relevant and unique data for analysis. The first step in this stage is to identify and remove records that appear repeatedly to ensure the uniqueness of each entry.

Additionally, selection involves considering the features or attributes of the data to be used. The selection of relevant features is crucial as it can enhance model performance, reduce computational complexity, and improve interpretability [4][5]. By focusing on appropriate attributes and removing irrelevant or redundant data, simpler and more accurate models can be constructed, which enhances learning accuracy and reduces computation time [6]. After this selection stage, the dataset, containing unique records and relevant features, is ready for the subsequent cleaning process.

D. Cleansing

A procedure was implemented to handle incomplete data. Upon inspection, any student record found with missing grades in several courses was subsequently removed. The deletion method was chosen because the students had either officially withdrawn or had not enrolled in the respective courses, making data imputation irrelevant as there was no underlying academic performance to estimate. This approach is consistent with statistical data preparation practices where the cause of missing data guides the handling strategy.

E. Transformation

The Interquartile Range (IQR) method is a statistical technique used to detect and handle outliers in data analysis. Outliers, which are data points that deviate significantly from others, can distort statistical estimates such as the mean and standard deviation, reducing the reliability of analysis results. In predictive modeling, outliers can affect model stability and accuracy, making their detection and treatment crucial [7].

The IQR method identifies outliers by calculating the first quartile (Q1) and third quartile (Q3) and determining the IQR as the difference between Q3 and Q1. Data points falling below Q1 - 1.5IQR or above Q3 + 1.5IQR are classified as outliers. These outliers can either be removed or normalized, depending on the dataset size and analysis needs. Normalization adjusts extreme values to fit within an acceptable range, ensuring the dataset remains reliable without unnecessary data loss [8].

F. Data Processing

1) Data preparation

Data normalization is a crucial preprocessing step in clustering analysis, as it ensures that features contribute equally to the distance calculations used in clustering algorithms. Normalization can significantly influence the results of clustering by affecting the detection of cluster centers and the overall clustering structure.

Normalization helps in achieving more accurate clustering results by ensuring that no single feature dominates due to its scale. This is particularly important in distance-based clustering methods, where features with larger scales can disproportionately affect the clustering outcome [14] [15].

2) Clustering with K-Means

Apply the K-Means algorithm to form a more homogeneous cluster and determine the number of clusters using the Elbow or Silhouette Score method. The Elbow Method

involves plotting the explained variance as a function of the number of clusters and identifying the "elbow" point where adding more clusters yields diminishing returns in variance explained. This method is straightforward but can sometimes be subjective due to its graphical nature [16] [17] [18] [19]. Some studies suggest that the Elbow Method can be automated to reduce subjectivity, enhancing its effectiveness across various datasets [13].\

Some studies suggest that the Elbow Method can be automated to reduce subjectivity, enhancing its effectiveness across various datasets [18]. The Silhouette Score measures how similar an object is to its own cluster compared to other clusters. A higher silhouette score indicates better-defined clusters [16] [20] [21].

This method is often used to validate the number of clusters suggested by the Elbow Method, providing a quantitative measure of cluster quality[20] [22].

3) Analysis of clustering results

After the clusters are formed using the K-Means algorithm, the next stage is to thoroughly analyze and interpret the characteristics of each cluster to understand the students' competency patterns. This analysis process will involve several steps:

a. Descriptive Statistical Analysis: For each cluster, a descriptive statistical analysis will be performed. This includes calculating the mean and standard deviation (SD) for the course grades that constitute the clustering features. This step aims to build a quantitative and measurable academic profile for each cluster, thereby avoiding general descriptions such as "stable value" or "highest performance" that lack a statistical basis.

b. Comparative Analysis Between Clusters: The statistical profiles of each cluster will then be systematically compared with one another. A comparison of mean scores between clusters will be used to objectively identify in which competency areas a cluster excels or is weaker compared to the others.

c. Interpretation and Relation to Concentration Areas: Next, these quantitative profiles will be interpreted and linked to the concentration areas available in the study program. This analysis aims to map each cluster profile to the most relevant area of expertise or career interest. For instance, a cluster with high mean scores in 'Web Programming' and 'Multimedia Basics' courses would be interpreted as having a strong inclination towards the web and multimedia development concentrations.

d. Data Visualization: To facilitate understanding and presentation of the findings, the analysis will be supported by data visualizations. Diagrams such as bar charts or radar charts will be used to clearly and visually depict the competency profile of each cluster

III. RESULT AND DISCUSSION

A. Student Data

Student data that will be analyzed by the researcher includes the attributes of course values from 1st semester to 5th semester.

Table 1 Example of a research dataset

NIM	Multimedia dasar	Pemrograman Terstruktur	Komunikasi Data dan Jaringan	Fotografi	Desain Web	Alghoritma dan struktur data	Administrasi jaringan komputer	Desain Grafis Percetakan
1	80,75	90,5	89,45	81,5	88,02	82	92	8
2	85,25	96,5	94,45	81,4	86,73	92	92,5	8
3	84,75	94	92,5	85,8	82,895	87	84	82,1
4	81	85,5	81,4	85,25	82,665	91	91,5	84,0
5	84	92,5	87,5	84	83,955	81,5	90,5	81,7
7	79,75	83	69,45	82,3	81,32	62,5	85,5	8
8	81,75	91	84,45	83,35	76,955	74	79	83,6
9	83,5	92	90	85,2	78,68	81,5	90	83,8
10	84,5	93,5	87,5	81,45	79,055	76	86	83,8
11	83,75	94	94,45	84,3	86,255	91,5	90,5	85,0
12	85	96,5	76,4	83,45	100	90,5	91,5	85,0
13	84,5	95	87,5	84,25	84,195	83,5	89	84,2
14	82,75	60,5	95	80,8	84,245	86,5	74	80,7
15	78,75	92	89,45	82,6	89,05	75,5	89	83,2
16	85	92	100	85,55	87,115	87,5	87	84,3
19	81,75	89	76,4	85,1	80,625	85,5	82,5	83,
20	82,25	96	86,95	82,05	84,745	92	92	84,0
22	85	93,5	88,9	80,95	81,345	81,5	90,5	82,3
23	83,5	96,5	82,5	83	86,23	90,5	85	81,9
24	81,25	93,5	88,9	83,75	87,645	96	88	85,0

B. Data Cleaning

In the data cleaning process, duplicate data and empty data have been cleaned. From the results of the check, there are 12 student data that do not have grades in several courses. This condition occurs because the student has resigned or did not take the course in question in that period.

C. Data Transformation

The process of cleaning data from the oulier using the IQR method found that there were 7 outliner data, then deleted.

```
# Menghitung IQR
Q1 = df[average_columns].quantile(0.25)
Q3 = df[average_columns].quantile(0.75)
IQR = Q3 - Q1

# Menentukan batas bawah dan atas untuk outlier
lower_bound = Q1 - 1.5 * IQR
upper_bound = Q3 + 1.5 * IQR

# Menghapus outlier menggunakan IQR
df_clean_iqr = df[~((df[average_columns] < lower_bound) | (df[average_columns] > upper_bound)).any
```

Figure 1 IQR Implementastion Code

D. K-Means Clustering

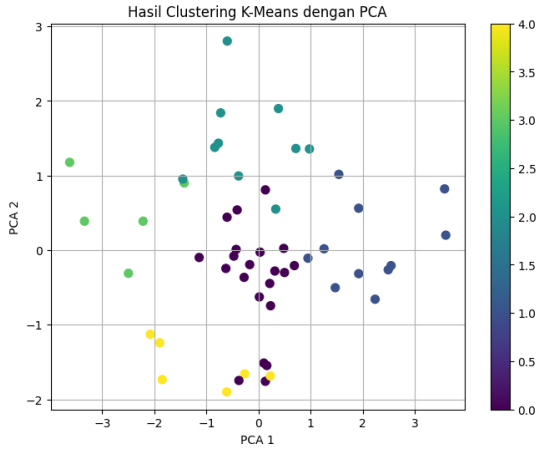

Figure 2 Visualization of Clustering Result with PCA

E. Cluster Interpretation

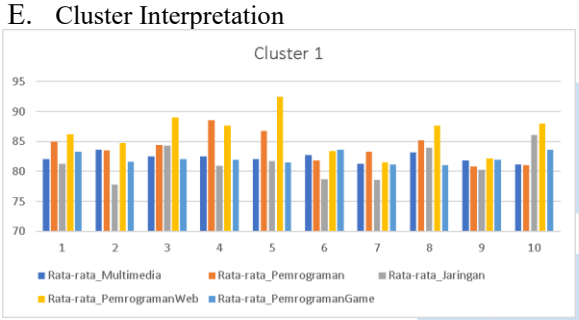

Figure 3 Clustered Column Chart cluster 1

Cluster 1 consists of students with balanced performance. This cluster consists of 10 students with fairly stable performance in all courses. There is no field that is very prominent or very low.

The calculated mean and standard deviation (SD) for each competency area within this cluster are as follows:

Table 2 Mean and Standard Deviation of cluster1

Competency Area	Mean	Standard Deviation (SD)
Web Programming	86.28	3.55
Programming	84.05	2.45
Multimedia	82.30	0.82
Game Programming	82.16	0.96
Networking	81.38	2.58

The qualitative description of "balanced performance" is well-supported, as all mean scores are high (above 81). The provided bar chart also visually confirms this, showing that for each of the 10 students, all five competency scores are consistently proficient.

Although no field shows notably low scores, the statistical analysis reveals that *Web Programming* and *Programming* have the highest average scores, indicating a tendency toward specialization in these areas rather than an evenly distributed skill set.

The concept of stable performance is further supported by the very low standard deviations in *Multimedia* (0.82) and *Game Programming* (0.96), suggesting that students within this cluster not only perform well but also exhibit a high degree of consistency in these specific domains.

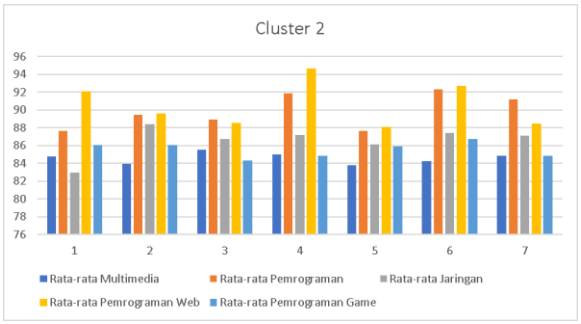

Figure 4 Clustered Column Chart cluster 2

Cluster 2 comprises students with consistently outstanding performance across all courses, particularly in *Programming* and *Web Programming*, where the average scores approach 90. This initial qualitative observation is substantiated by a detailed statistical analysis that quantifies their academic excellence.

Table 3 Mean and Standard Deviation of cluster 2

Competency Area	Mean	Standard Deviation (SD)
Web Programming	90.60	2.53
Programming	89.87	1.92
Networking	86.58	1.68
Game Programming	85.55	0.86
Multimedia	84.59	0.61

The calculated means and standard deviations for each competency area reveal a distinct pattern. *Web Programming* (mean = 90.60, SD = 2.53) and *Programming* (mean = 89.87, SD = 1.92) stand out, supporting the assertion that students in this cluster achieve near-perfect scores in these domains. The consistently high averages across the remaining areas—*Networking* (86.58), *Game Programming* (85.55), and *Multimedia* (84.59)—further confirm that these students excel across the board.

This interpretation is visually supported by the clustered column chart, where *Programming* and *Web Programming* are consistently represented as the highest-performing areas, often exceeding the 90-point threshold. The chart also highlights the elevated baseline across all competencies.

Moreover, the low standard deviations observed in *Multimedia* (0.61) and *Game Programming* (0.86) point to a high degree of consistency within the cluster, indicating that the outstanding performance is not limited to a few individuals but is uniformly distributed among all students in the group.

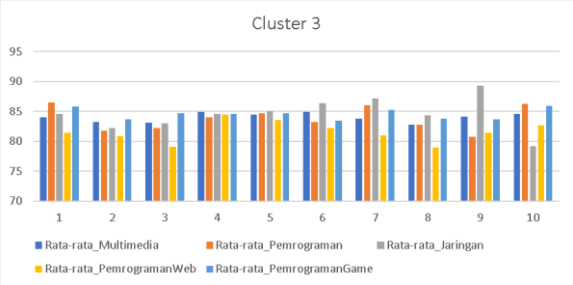

Figure 5 Clustered Column Chart cluster 3

The initial description characterizes Cluster 3 as consisting of 10 students with moderate and balanced performance. It is noted that their overall performance tends to be lower than other clusters and that their scores in the Web Programming course are slightly lower than in other subjects. A detailed statistical analysis provides a quantitative perspective on this profile

Based on the provided data for the 10 records in Cluster 3, the calculated mean and standard deviation (SD) are as follows:

Table 4 Mean and Standard Deviation of cluster 3

Competency Area	Mean	Standard Deviation (SD)
Networking	84.57	2.79
Game Programming	84.55	0.88
Multimedia	83.98	0.69
Programming	83.80	1.89
Web Programming	81.54	1.81

The statistical findings reinforce and elaborate upon the initial qualitative characterization of this cluster.

The average scores across all competency areas fall within a relatively narrow range in the low-to-mid 80s, supporting the classification of this group as exhibiting moderate performance compared to clusters with higher achievement levels. The data also substantiates the observation that *Web Programming* represents the weakest area within the group, with a mean score of 81.54—the lowest among all subjects.

While the performance profile appears balanced, with four out of five means ranging closely between 83.80 and 84.57, subtle variations are present. The highest levels of achievement are observed in *Networking* and *Game Programming*, suggesting these areas as relative strengths.

Notably, the group demonstrates exceptional consistency in *Multimedia* (SD = 0.69) and *Game Programming* (SD = 0.88), indicating a high degree of uniformity in student performance within these subjects. Conversely, the greatest variation is found in *Networking* (SD = 2.79), suggesting a wider disparity in skill levels in this domain.

These patterns are visually reflected in the clustered column chart, which illustrates the close grouping of most scores within the 80–85 range—highlighting the balanced nature of the group's performance. The comparatively lower scores in *Web Programming* are also clearly depicted in the chart, consistent with the statistical summary.

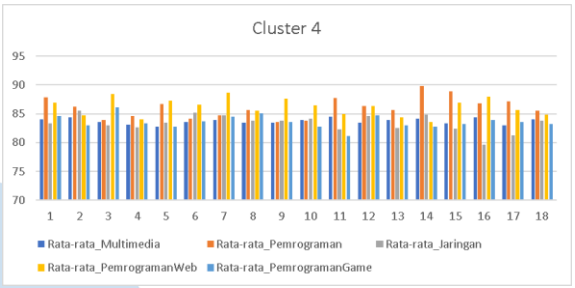

Figure 6 Clustered Column Chart cluster 4

Cluster 4 students with above average performance. This cluster consists of 18 students with slightly higher than average performance, especially in Programming courses. This cluster is the cluster with the largest number of studentsThe initial description characterizes Cluster 4 as the largest group, consisting of 18 students with "above average" performance. A particular strength in Programming courses is also highlighted. A detailed statistical analysis provides quantitative evidence to support and elaborate on this profile.

Based on the provided data for the 18 records in Cluster 4, the calculated mean and standard deviation (SD) are as follows:

Table 5 Mean and Standard Deviation of cluster4

Competency Area	Mean	Standard Deviation (SD)
Web Programming	86.16	1.57
Programming	86.03	1.84
Multimedia	83.71	0.52
Game Programming	83.50	1.06
Networking	83.44	1.38

The statistical analysis offers a deeper insight into the "above average" profile initially associated with this cluster.

The observed strengths are clearly supported by the data, with *Programming* (mean = 86.03) and *Web Programming* (mean = 86.16) emerging as the highest-

performing areas. These figures confirm the group's strong inclination toward development-oriented competencies.

An important characteristic of this relatively large cluster is its high level of internal consistency. The standard deviation for *Multimedia* is notably low at 0.52, indicating that performance in this subject is nearly identical among all 18 students. Similarly, the low variability in other subjects suggests a generally stable and uniform skill profile across the group.

This interpretation is further supported by the clustered column chart, which illustrates the consistency of student performance. Most bars fall within the 80–89 range, reinforcing the "above average" classification. In addition, the prominence of *Programming* and *Web Programming* scores—often among the highest for each student—visually aligns with the statistical summary.

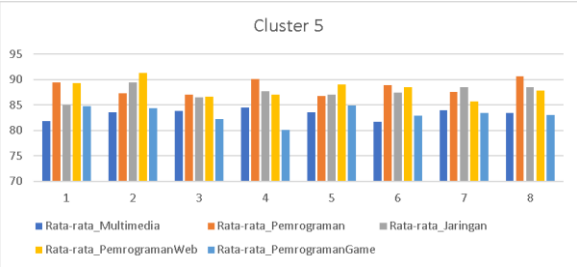

Figure 7 Clustered Column Chart cluster 5

The initial description characterizes Cluster 5 as a group of 8 students with strengths in specific areas. It highlights that this cluster excels in Programming and Networking courses but is slightly lower in Game Programming courses. A descriptive statistical analysis provides quantitative details that confirm and refine this specialized profile.

Based on the provided data for the 8 records in Cluster 5, the calculated mean and standard deviation (SD) are as follows:

Table 6 Mean and Standard Deviation of cluster 5

Competency Area	Mean	Standard Deviation (SD)
Programming	88.45	1.42
Web Programming	88.17	1.77
Networking	87.52	1.35
Multimedia	83.29	0.94
Game Programming	83.20	1.49

The statistical findings offer strong support for the initial qualitative assessment of this cluster.

The group's strengths are clearly reflected in the mean scores for *Programming* (88.45), *Networking* (87.52), and *Web Programming* (88.17), which are among the highest in the cluster. These results confirm the group's solid proficiency in development-related areas, as initially described.

The comparatively lower performance in *Game Programming*, with a mean score of 83.20, substantiates the earlier observation that this area represents a relative weakness within the cluster.

Overall, the standard deviations across subjects are relatively low, particularly in *Multimedia* (SD = 0.94), suggesting that the students in this cluster not only share similar strengths but also exhibit consistent patterns of performance, including their weaker areas.

This interpretation is further supported by the clustered column chart, which illustrates consistently high scores in *Programming*, *Networking*, and *Web Programming* for the majority of the 8 students. The slightly lower bars for *Game Programming* are also clearly visible, aligning with the quantitative analysis.

F. Clustering Evaluation

In this study, accuracy evaluation was conducted to assess the effectiveness of the K-Means and Naïve Bayes algorithms in grouping and classifying student competencies. The clustering results using K-Means were evaluated through the Silhouette Score, which produced a value of 0.34889702766857306, indicating that the clustering results were in the moderate category.

A moderate Silhouette Score likely reflects a combination of factors related to the data's nature and the methodology used. A primary reason could be the data's inherent structure, where the groups naturally overlap rather than being distinct and well-separated. Additionally, the chosen clustering algorithm, such as K-Means, may be ill-suited if the data's true clusters are non-spherical or have varying densities. Finally, the result may also be due to a suboptimal choice for the number of clusters (k), as this parameter heavily influences the clustering outcome.

```
# Menghitung Silhouette Score
labels=kmeans.labels_
silhouette_avg = silhouette_score(X, labels)
print(f"Silhouette Score: {silhouette_avg}")

Silhouette Score: 0.34889702766857306
```

Figure 8 Silhoutte Score

The silhouette score is used to evaluate the quality of data grouping in a cluster. If the silhouette score is close to +1, the data is considered to be in the correct or well-identified cluster. Conversely, if the score is close to 0, the data is likely to be on the border between two clusters, so there is potential for misplacement of the cluster. Meanwhile, a silhouette score close to -1 indicates that the data is in the wrong cluster, because it is closer to another cluster than the cluster where it is currently located [15].

The clustering results obtained show that the data has a fairly clear separation between one cluster and another, although there are still some data points that are close to the boundaries between clusters. This shows that although the clustering has been done well, there is still room to improve the quality of the separation between clusters to make it more optimal.

Given the moderate performance and the inherent limitations of the primary algorithm, it is prudent to discuss alternative methods that could be explored in future work. For instance, Hierarchical Clustering offers the advantage of not requiring the number of clusters to be pre-specified, instead building a tree-like structure of nested clusters (a dendrogram) that can be insightful for understanding relationships at different levels of granularity, which is particularly useful if the data contains meaningful sub-groups. Alternatively, DBSCAN (Density-Based Spatial Clustering of Applications with Noise) can identify arbitrarily shaped clusters and is robust to outliers, which it automatically flags as noise. This makes DBSCAN a strong candidate if the underlying data patterns are not spherical or if the dataset is known to contain anomalous data points.

IV. CONCLUSIONS

This study successfully implemented the K-Means algorithm to identify five distinct student competency patterns within the Informatics and Computer Engineering Education program at Universitas Sebelas Maret. The analysis revealed diverse student profiles, including a group of 'Elite High-Achievers' excelling in all areas, a specialized cohort of 'Development and Networking Specialists', and a large, stable group of 'Consistent Moderates'. These findings demonstrate that K-Means is effective for uncovering meaningful, data-driven structures in student academic records without requiring initial labels. The practical implications of these findings are significant. The identified cluster profiles can form the foundation for a data-driven academic guidance system to help students make more informed decisions when choosing their area of concentration. Furthermore, these insights provide the study program with a valuable tool for curriculum evaluation and for better understanding the competency landscape of its student body.

However, the study acknowledges key limitations. The clustering quality, as measured by the Silhouette Score, was moderate at 0.3489. This result suggests that while the clusters are distinct, they are not perfectly separated, likely due to the overlapping nature of student skills and the inherent tendency of K-Means to prefer spherical clusters, which may not fully represent the complex distribution of academic competencies. The algorithm's effectiveness is also highly dependent on the selection of an optimal number of clusters (k) and its sensitivity to initial cluster centers.

Based on these limitations, future research should explore the application of alternative clustering algorithms. Methods such as Hierarchical Clustering or DBSCAN could provide different perspectives, potentially identifying non-spherical patterns or nested sub-groups within the data. Further work could also involve enriching the dataset with non-academic features to create more holistic student profiles and developing the proposed recommendation system into a fully functional tool for academic guidance.

REFERENCES

[1] D. Masturina, "Pengaruh Kompetensi diri terhadap perencanaan karir," *Psikoborneo*, vol. 6, no. 2, pp. 198–205, 2018, doi: https://doi.org/10.30872/psikoborneo.v6i2.4558.

[2] N. I. Mohd Talib, N. A. Abd Majid, and S. Sahran, "Identification of Student Behavioral Patterns in Higher Education Using K-Means Clustering and Support Vector Machine," *Applied Sciences (Switzerland)*, vol. 13, no. 5, Mar. 2023, doi: 10.3390/app13053267.

[3] E. Tuyishimire, W. Mabuto, P. Gatabazi, and S. Bayisingize, "Detecting Learning Patterns in Tertiary Education Using K-Means Clustering," *Information (Switzerland)*, vol. 13, no. 2, Feb. 2022, doi: 10.3390/info13020094.

[4] M. Rong, D. Gong, and X.-Z. Gao, "Feature Selection and Its Use in Big Data: Challenges, Methods, and Trends," *IEEE Access*, vol. 7, pp. 19709–19725, 2019, doi: 10.1109/ACCESS.2019.2894366.

[5] J. Li *et al.*, "Feature Selection," *ACM Computing Surveys (CSUR)*, vol. 50, pp. 1–45, 2016, doi: 10.1145/3136625.

[6] J. Cai, J. Luo, S. Wang, and S. Yang, "Feature selection in machine learning: A new perspective," *Neurocomputing*, vol. 300, pp. 70–79, 2018, doi: 10.1016/j.neucom.2017.11.077.

[7] T. Nyitrai and M. Virág, "The effects of handling outliers on the performance of bankruptcy prediction models," *Socioecon Plann Sci*, p., 2019, doi: 10.1016/J.SEPS.2018.08.004.

[8] H. Aguinis, R. Gottfredson, and H. Joo, "Best-Practice Recommendations for Defining, Identifying, and Handling Outliers," *Organ Res Methods*, vol. 16, pp. 270–301, 2013, doi: 10.1177/1094428112470848.

[9] M. Suárez-Álvarez, D. Pham, M. Prostov, and Y. Prostov, "Statistical approach to normalization of feature vectors and clustering of mixed datasets," *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 468, pp. 2630–2651, 2012, doi: 10.1098/rspa.2011.0704.

[10] V. Starovoitov and Y. Golub, "Data normalization in machine learning," *Informatics*, p., 2021, doi: 10.37661/1816-0301-2021-18-3-83-96.

[11] H. Humaira and R. Rasyidah, "Determining The Appropiate Cluster Number Using Elbow Method for K-Means Algorithm," *Proceedings of the Proceedings of the 2nd Workshop on Multidisciplinary and Applications (WMA) 2018, 24-25 January 2018, Padang, Indonesia*, p., 2020, doi: 10.4108/eai.24-1-2018.2292388.

[12] D. Saputra, D. Saputra, and L. Oswari, "Effect of Distance Metrics in Determining K-Value in K-Means Clustering Using Elbow and Silhouette Method," *Proceedings of the Sriwijaya International Conference on Information Technology and Its Applications (SICONIAN 2019)*, p., 2020, doi: 10.2991/aisr.k.200424.051

[13] A. Onumanyi, D. N. Molokomme, S. Isaac, and A. Abu-Mahfouz, "AutoElbow: An Automatic Elbow Detection Method for Estimating the Number of Clusters in a Dataset," *Applied Sciences*, p., 2022, doi: 10.3390/app12157515.

[14] M. Syakur, B. Khotimah, E. Rochman, and B. D. Satoto, "Integration K-Means Clustering Method and Elbow Method For Identification of The Best Customer Profile Cluster," *IOP Conf Ser Mater Sci Eng*, vol. 336, p., 2018, doi:

10.1088/1757-899X/336/1/012017

[15] K. Shahapure and C. Nicholas, "Cluster Quality Analysis Using Silhouette Score," *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, pp. 747–748, 2020, doi: 10.1109/DSAA49011.2020.00096.

[16] C. Yuan and H. Yang, "Research on K-Value Selection Method of K-Means Clustering Algorithm," *J (Basel)*, p., 2019, doi: 10.3390/J2020016.

[17] N. Sagala and A. Gunawan, "Discovering the Optimal Number of Crime Cluster Using Elbow, Silhouette, Gap Statistics, and NbClust Methods," *ComTech: Computer, Mathematics and Engineering Applications*, p., 2022, doi: 10.21512/comtech.v13i1.7270.