I. J. Education and Management Engineering, **2025, 4, 12-22**
Published Online on August 8, 2025 by MECS Press (http://www.mecs-press.org/)
DOI: 10.5815/ijeme.2025.04.02

Optimizing Student Performance Prediction via K-Means and k-NN Integration

Muh. Nurtanzis Sutoyo*
Universitas Sembilanbelas November Kolaka, Kolaka, Indonesia
E-mail: mns.usn21@gmail.com
ORCID iD: https://orcid.org/0000-0001-6507-9096
*Corresponding Author

Alders Paliling
Universitas Sembilanbelas November Kolaka, Kolaka, Indonesia
E-mail: alders_p@usn.ac.id
ORCID iD: https://orcid.org/0000-0002-9581-8248

Received: 28 September, 2024; Revised: 05 February, 2025; Accepted: 12 March, 2025; Published: 08 August, 2025

Abstract: This study explores the integration of two methods, namely K-Means and k-NN. K-means is used to identify categories of learning outcome data, while k-NN is used to predict students' learning outcomes into relevant categories. Through the calculation of the Elbow method, it was established that the optimal number of clusters for grouping is three. The learning outcome data, which include Arithmetic and Statistics scores, are processed to produce a mapping that differentiates students into three categories: Adequate, Moderate, and Good. In the 12th iteration, the clustering results using K-Means achieved convergence, with 64 students in the Adequate category (C1), 60 students in the Moderate category (C2), and 59 students in the Good category (C3). This indicates that the students in each group are evenly distributed based on their mathematical and statistical abilities. The prediction results using k-NN for a student with an Arithmetic score of 85 a Statistics score of 75, and a k-value of 61, found that 7 data fell into Category 1 (Adequate), 3 data into Category 2 (Moderate), and dominant 51 data in Category 3 (Good). Thus, the prediction results are placed in Category 3, indicating a 'Good' rating in their academic performance. By using data mining techniques to enhance understanding of student learning outcomes, this study provides a significant contribution to the field of education. It demonstrates substantial progress toward a data-driven learning approach that can be tailored to specific needs and improve student learning outcomes.

Index Terms: Clustering Analysis, K-Means, k-NN, Learning Outcomes Mapping, Elbow Method

1. Introduction

Measurement and prediction of student learning outcomes are crucial components in the education sector to enhance the quality of teaching and learning. The learning outcomes referred to here include not only academic grades but also the competencies and skills that students possess. With the advancement of information technology, particularly in the field of data science, there is significant potential to delve deeper into educational data to gain broader insights into patterns that may not be immediately apparent.

This study selected the K-Means and k-Nearest Neighbors (k-NN) approaches for their complementing strengths in managing educational performance data. K-Means, a prevalent unsupervised learning technique, is especially effective for discerning inherent clusters within numerical datasets, such as students' scores in Arithmetic and Statistics. K-Means facilitates the categorization of students into clusters with analogous learning attributes by splitting the dataset into K clusters according to the nearest mean values. This categorization is crucial for recognizing patterns that can guide specific educational approaches[1,2,3]. Clustering is the process of grouping a number of objects into clusters based on the similarity of their characteristics. The goal is to ensure that objects within the same group have a high level of similarity to one another, and objects in different groups have a low level of similarity [4,5,6,7].

Meanwhile, k-NN is a method used for classification and regression that operates based on instance-based or memory-based learning. This method works by measuring the training data samples that are closest to a new object using a distance function and then making predictions or decisions based on the majority category of the k closest samples [8,9,10,11].

This project aims to amalgamate K-Means and k-Nearest Neighbors (k-NN) methodologies to facilitate a more accurate and pragmatic examination of student learning outcomes within an educational framework. The integration facilitates a bifurcated methodology: K-Means is utilized to categorize students into clusters according to similarities in academic performance, whereas k-NN is applied to forecast future results for individual students based on historical data from their peer groups. This approach seeks to aid educators in recognizing individuals with analogous learning requirements and to enhance the precision and efficacy of early interventions. Data-driven solutions correspond with the tenets of Outcome-Based Education (OBE), which asserts that all educational processes—from curriculum development to assessment—must guarantee that students achieve the anticipated competences and learning objectives by the conclusion of their studies[12,13,14,15].

In many schools across Southeast Sulawesi, assessments of student performance are generally limited to final grades without deeper analysis of learning patterns. This results in a lack of actionable insights for educators to identify students who need specific attention or support. Furthermore, there is an absence of systematic methods to group students with similar academic characteristics or to predict future performance accurately. This gap highlights the need for an integrated, data-driven approach that can both cluster students based on learning outcomes and predict their future academic achievement.

By using K-Means, mapping the learning outcomes can assist in identifying groups of students who have similar learning needs or characteristics. These characteristics can be used as a basis for more focused educational interventions and personalized learning. Furthermore, by using k-NN, this study aims to make accurate predictions about student learning outcomes based on historical data and the characteristics of the grouped students.

It is hoped that this method can provide a significant contribution to the development of more effective learning strategies by understanding and analyzing various components that affect student learning outcomes produced by this data mining technique. Therefore, this research is not only relevant from an academic perspective but also crucial for its application in schools and other educational institutions in the real world. This aligns with the current trend in many educational systems around the world, which encourages the use of big data for more evidence-based and outcome-driven educational reform.

This novelty lies not only in the application of these techniques but also in the way they are integrated to achieve more specific objectives in education. Here are some aspects of the novelty of this research:

1. K-means (a clustering method) and k-NN (a classification method) have long been used in various fields, but their concurrent application for mapping and predicting learning outcomes is relatively new.
2. Although there are a number of studies using K-Means or k-NN separately, combining the two to map and predict learning outcomes is a novel idea. This helps in selecting groups of students based on similar learning characteristics. By using more dynamic and adaptive methods, their learning outcomes can be predicted.
3. Using this method in the analysis of big data in education, there are new opportunities for more efficient data usage. This research reveals how big data can be utilized to maximize its benefits in an educational context, as it can handle extremely large amounts of data.

2. Related Work

This study refers to the application of the K-Means and k-Nearest Neighbors (k-NN) methods across various disciplines, to understand how both techniques have been successfully implemented in data analysis and machine learning. Previous research has shown that K-Means is effective in grouping large datasets into clusters based on similar attributes, while k-NN has proven to be reliable in making predictions based on the presence and characteristics of the nearest neighbors.

The K-means method has been used for several studies, including Imron et al [16] the K-means algorithm is used to cluster products at Rizki Barokah Store based on demand levels. This store faces issues with expired product stock due to errors in product stock selection. As a result, the store can use sales data to adjust its inventory, reducing purchases of unpopular items and focusing more on popular products. Vulandari et al [17] used the K-Means clustering technique to map crime-prone areas in Central Java based on the crime index, the ratio of police to population, population density, and poverty levels. As a result, regions in Central Java are classified into four categories: safe, moderately vulnerable, vulnerable, and highly vulnerable. Overall, there are 11 safe areas, 4 moderately vulnerable areas, 13 vulnerable areas, and 6 highly vulnerable areas. This helps in identifying areas that require additional supervision, and this data is useful for the general public and police for preventive measures and decision-making.

Nengsih and Zain [18] utilize descriptive modeling with the K-means clustering method for mapping employee attendance. The goal is to address issues that arise at the management level related to mapping and monitoring employee attendance data. The results of this mapping will serve as a crucial reference for future management policies, which will enhance work efficiency and effectiveness following the company's vision and mission. Zhao et al. [19] proposed an extended regularized K-Means clustering method to handle high-dimensional data with correlated features, demonstrating its superiority over traditional K-Means in terms of error rate and variable selection. Although applied to customer segmentation, their approach illustrates the flexibility of K-Means variants in complex prediction tasks involving multidimensional student behavior or learning activity data. This study can assist stakeholders in assessing

and managing the vaccination process, as well as in making public health policy decisions. Iskandarli [20] discusses automatic approaches for analyzing citizen comments in e-government using topic modeling and clustering algorithms. The primary goal is to enhance electronic services and identify the topics of citizen comments. This method addresses the issues of semantic document clustering and vector size by using semantic similarity of words.

Sathe and Adamuthe [21] tested the k-Nearest Neighbors (k-NN) algorithm with cross-validation techniques to predict student success in distance education using data that includes academic grades, attendance, and student participation. The results show that k-NN with k=5 is the most accurate in prediction. This study demonstrates that choosing the right parameters for the k-NN algorithm is crucial for improving prediction accuracy in online education. Hidayati dan Hermawan [22] use the K-Nearest Neighbor (K-NN) algorithm to classify student graduation times using Euclidean and Manhattan distance metrics. This study utilized data from 543 students, with 380 as training data and 163 as test data. There was no significant difference between using Euclidean and Manhattan distances, and the developed model showed that the best accuracy rate was 85.28% at K=7. This study assists universities in planning educational programs to ensure students graduate on time. Ferreira et al [23] used online learning methods to implement the k-Nearest Neighbors (k-NN) prototyping scheme for identifying human activities in embedded devices. According to this study, several substitution strategies can be used to replace training instances when maximum capacity is reached. This will allow k-NN to maintain an equivalent representation of each activity class. Experimental results indicate that the proposed plan enhances the prediction accuracy of k-NN without significantly increasing energy consumption. This suggests that this plan can be effectively used in devices with limited resources.

Liu et al [24] used the k-Nearest Neighbors (k-NN) regression model to predict Amazon EC2 spot instance prices. Accurate predictions can help buyers select offers and avoid overspending, as spot instance prices are highly volatile. The k-NN model is compared with Linear Regression, Support Vector Machine Regression, Random Forest, Multi-layer Perceptron Regression, and gcForest models. The results show that the k-NN model is better at predicting spot instance prices because it is more stable. Priyadarshini et al [25] discusses the detection of fraudulent credit card transactions using soft computing techniques. The primary goal is to enhance predictive accuracy in classifying fraudulent credit card transactions. To address this, researchers have implemented classification models using algorithms such as K-Nearest Neighbor (K-NN), decision trees, random forests, and logistic regression. The proposed models are computationally efficient and can provide more accurate detection of fraudulent transactions.

Fathoni et al [26] analyzed public sentiment towards the spread of COVID-19 in Indonesia using Twitter data. By collecting tweet data in Indonesian, a sentiment analysis was conducted to assist the Indonesian government in taking strategic actions. The data was processed through pre-processing stages, and classification was performed using K-Nearest Neighbor (K-NN), Naïve Bayes, and decision tree methods. The results showed that the K-NN algorithm had the highest accuracy at 95.10%, with the majority of tweets negatively categorizing the spread of COVID-19. Shallal et al [27] discusses a data classification method in the Internet of Things (IoT) environment using the K-Nearest Neighbor (K-NN) machine learning algorithm. This research aims to enhance data security by categorizing data based on their sensitivity levels: low, medium, and high. By differentiating the sensitivity levels of data, computational resources such as encryption and decryption times, as well as battery usage, can be optimized.

This study distinguishes itself from previous ones by integrating K-Means for clustering and k-NN for classification within a single analytical framework. This integration not only enhances the mapping capability to identify groups of students with similar learning characteristics but also strengthens the accuracy of predicting student academic performance by tailoring the k-NN model based on insights gained from the K-Means clustering results.

3. Methodology

To understand and implement effective strategies in mapping and predicting student learning outcomes, this study employs an integrative approach by combining two popular data analysis methods: K-Means for clustering and k-nearest Neighbors (k-NN) for classification. The proposed methodology aims to optimize the strengths of both techniques in clustering and predicting student learning outcomes, thereby enabling more accurate identification and intervention tailored to the learning needs of students. Figure 1 below shows the steps taken in the research.

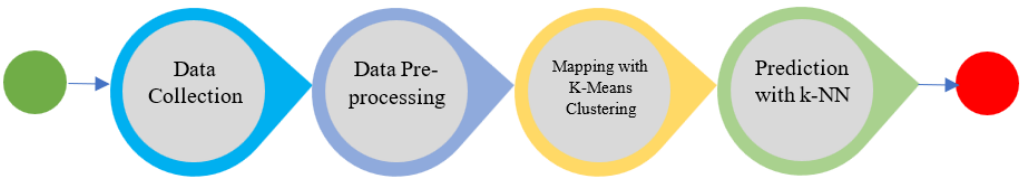

Fig. 1. Research Steps

3.1. Data Collection

The dataset used in this study consists of academic performance records from 183 students, collected from three senior high schools in Southeast Sulawesi, Indonesia, all of which provided informed consent to participate in the data collection process. The data include students' scores in Arithmetic and Statistics, sourced from both exam results and ongoing teacher assessments conducted during a specific academic semester. This dataset was chosen due to its structured and numeric nature, which is suitable for clustering and classification analysis. However, since the data originate from a limited geographical area and involve only schools that voluntarily agreed to participate, there is a possibility of selection bias, and the findings may not be fully generalizable to other regions or educational contexts.

3.2. Data Pre-processing

To ensure the integrity and quality of the data to be used, the pre-processing stage is a crucial step in this research. Incorrect or inconsistent data cleaning is one of the key activities performed during this process. This step is vital for preparing the next stages, allowing the K-Means and k-NN algorithms to be analyzed effectively.

3.3. Mapping with K-Means Clustering

The student's learning outcomes are grouped using K-Means with the following steps [28,29,30,31,32]:

- Determine the ideal number of clusters *K*, using the Elbow method[33]. For the initial cluster centroids, assign values randomly.
- Calculate the distance from each point in the dataset to each centroid. The equation used to compute the distance is the Euclidean distance.

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \quad (1)$$

  Where *d(x, y)* is the distance from data point *x* to the cluster center, *xi* = the *x* data at the *i* observation, *yi* = the *y* coordinate of the center at the *i* observation, and *n* = the number of observations. Each data point belongs to the cluster with the nearest centroid, which is selected based on the minimum distance to the centroid.
- Next, recalculate the centroid position for each cluster after computing the distance of the dataset for each cluster. The new centroid position is calculated based on the average (mean) of all data within the cluster.

$$centroid = \sum \frac{a_i}{n} \quad (2)$$

  Where *ai* = Membership value of each cluster and *n* = Number of cluster members.
- The centroid updates and distance calculations must be repeated until convergence criteria are met. Examples of these criteria could include conditions where cluster assignments no longer change, changes in centroid positions are below a certain threshold, or the maximum number of iterations has been reached.
- Each cluster created by K-Means will be named based on its dominant characteristics.

3.4. Prediction with k-Nearest Neighbors (k-NN)

To ensure that this technique is effective in predicting learning outcomes, the implementation stages of k-Nearest Neighbors (k-NN) in this study require several important steps. Here are the necessary actions [34,35,36,37]:

- In k-NN, determining the ideal number of nearest neighbors *k* is crucial. The value of *k* is selected through a cross-validation process, which is also known as cross-validation, to find a balance between overfitting and underfitting. A *k* value that is too small can cause the model to become sensitive to noise in the data, while a *k* value that is too large can cause the model to ignore important details in the data.
- For each data point, the training set is used to train the k-NN model by identifying the *k*. This is done using the Euclidean distance metric.

4. Result and Discussions

In this section, a comprehensive analysis of the processed and modeled data is presented, integrating the K-Means and k-NN methods. Data preprocessing is carried out on data collected from various schools to ensure the quality of the data sent. Table 1 below shows the preprocessed data amounting to 183.

Table 1. Learning Outcome Data

No	Arithmetic	Statistics
1	90	70
2	50	70
3	55	80
4	70	80
5	65	50
6	60	65
7	75	85
8	55	60
9	55	75
10	80	85
…	…	…
100	75	75
101	60	80
102	80	60
103	65	80
104	80	80
105	80	50
…	…	…
180	50	85
181	60	65
182	60	65
183	80	55

Clustering is performed after the data has been successfully collected. The Elbow Method is used to determine the ideal value of k before beginning the clustering process. To ensure that the selected number of clusters can produce effective and significant data separation while minimizing variation within each cluster, this test is crucial. Figure 2 shows the results of the Elbow test.

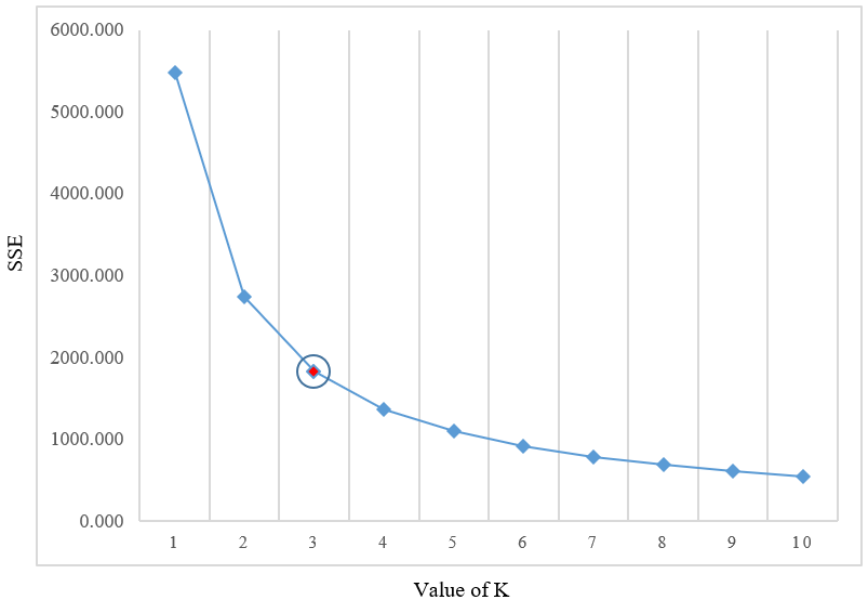

Fig. 2. Elbow Test Results

The ideal value of *k* for the clustering process is 3, as indicated by the results of the Elbow test shown in Figure (Fig. 1). This suggests that dividing the data into three clusters will minimize the variation between clusters and provide a high level of homogeneity within each cluster.

After completing the Elbow test and confirming that the optimal value for *k* is 3, we proceeded to apply the K-Means algorithm for the clustering process. The purpose of this step is to group the data into three identified clusters, thereby enabling further analysis of the characteristics and underlying patterns of each group. Here are the steps for the clustering process:

- Determine the cluster centroids randomly, as follows: C1 = (45, 50); C2 = (70, 80); and C3 = (85, 90).
- Using equation (1), calculate the distance of each data point from the cluster center.
  The distance between the data and the first cluster center (C1)

$$d_{11} = \sqrt{(90-45)^2 + (70-50)^2} = 49.24$$

The distance between the data and the center of the second cluster (C2)

$$d_{12} = \sqrt{(90-70)^2 + (70-80)^2} = 22.36$$

The distance between the data and the center of the third cluster (C3)

$$d_{13} = \sqrt{(90-85)^2 + (70-90)^2} = 20.62$$

Data will become a member of the cluster if they are at the minimum distance from the cluster center. The results of iteration 1 can be seen in the following Table 2.

Table 2. Distance Calculation Results Iteration 1

No	d.C1	d.C2	d.C3	Cluster
1	49.24	22.36	20.62	3
2	20.62	22.36	40.31	1
3	31.62	15.00	31.62	2
4	39.05	0.00	18.03	2
5	20.00	30.41	44.72	1
6	21.21	18.03	35.36	2
7	46.10	7.07	11.18	2
8	14.14	25.00	42.43	1
9	26.93	15.81	33.54	2
10	49.50	11.18	7.07	3
…	…	…	…	…
100	39.05	7.07	18.03	2
101	33.54	10.00	26.93	2
102	36.40	22.36	30.41	2
103	36.06	5.00	22.36	2
104	46.10	10.00	11.18	2
105	35.00	31.62	40.31	2
…	…	…	…	…
180	35.36	20.62	35.36	2
181	21.21	18.03	35.36	2
182	21.21	18.03	35.36	2
183	35.36	26.93	35.36	2

In iteration 1, calculating the distance from each dataset to the cluster centers resulted in 30 members in C1, 128 in C2, and 25 in C3 (Figure 3).

- Using equation (2), find the new cluster centers. The calculation results in new cluster centers at C1 (60.00, 57.50); C2 (71.41, 70.59); and C3 (85.20, 81.80).
- Repeat the second step until the data positions no longer change. In this calculation, the iteration was stopped at the 12th iteration. The following Table 3 shows the position of each data group at iteration 12.

Table 3. Distance Calculation Results Iteration 12

No.	d.C1	d.C2	d.C3	Cluster
1	21.17	31.31	11.87	3
2	27.01	10.98	32.60	2
3	29.75	5.89	26.58	2
4	23.55	11.50	11.65	2
5	10.85	26.55	32.78	1
6	15.85	10.94	25.31	2
7	28.32	18.18	9.35	3

8	18.80	16.47	32.23	2
9	26.02	4.35	26.73	2
10	29.02	22.65	6.87	3
…	…	…	…	…
100	18.34	15.78	7.31	3
101	26.92	4.15	21.59	2
102	7.27	26.15	18.37	1
103	24.79	7.05	16.61	2
104	24.17	21.15	2.28	3
105	9.34	33.20	28.35	1
…	…	…	…	…
180	36.78	12.96	32.23	2
181	15.85	10.94	25.31	2
182	15.85	10.94	25.31	2
183	6.71	29.46	23.35	1

The results of the cluster center calculations for each data point in the 12th iteration indicate that cluster C1 has 64 members (Adequate), C2 has 60 members (Moderate), and C3 has 59 members (Good). Figure 3 below clearly illustrates the data distribution resulting from the clustering.

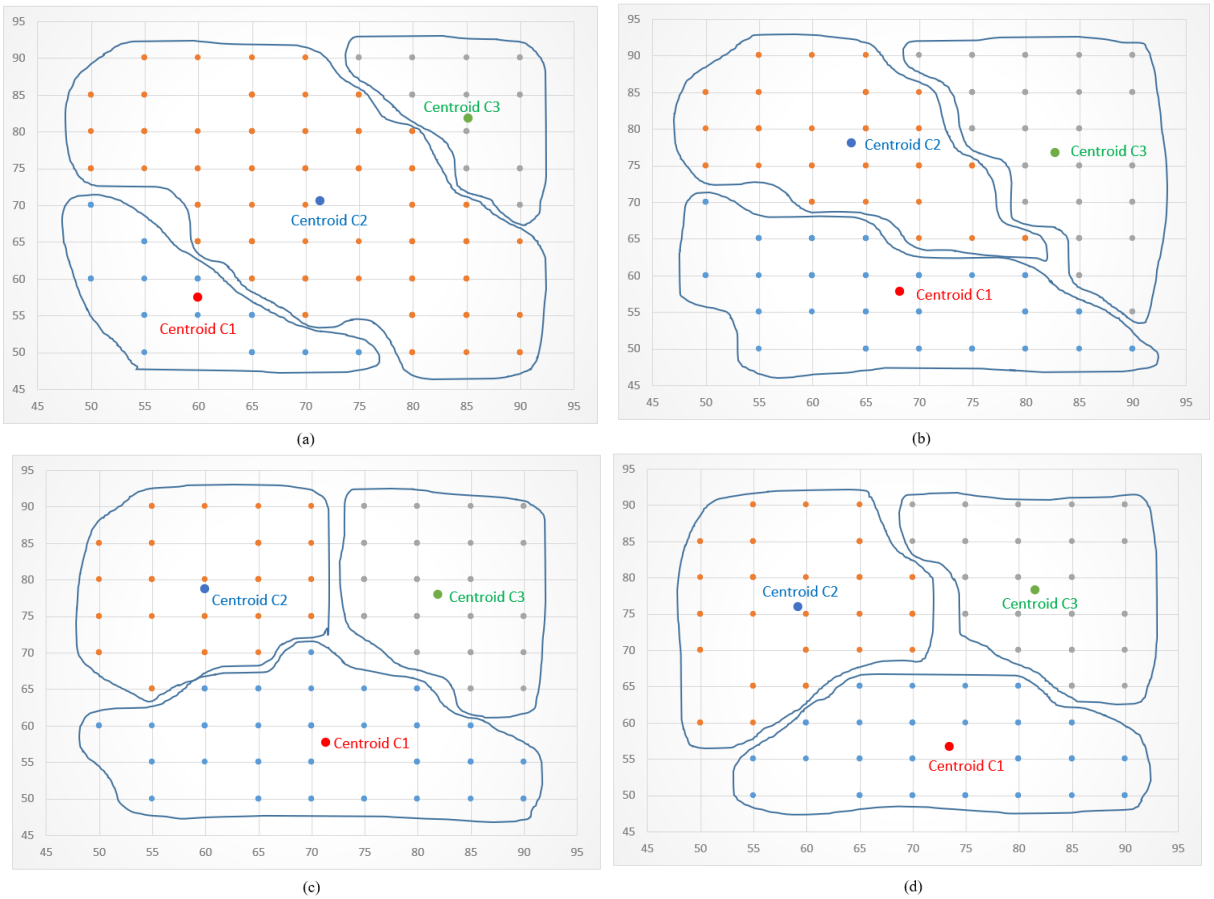

Fig. 3. Distribution of Clustering Results Data, Iteration 1 (a), Iteration 1 (b), Iteration 7 (c), and Iteration 12 (d)

From Figure 3, the changes in the cluster centers and the number of members in each cluster across iterations are evident. During iteration 4, the cluster centers were located at C1 = (68.19, 57.75) with 69 members, C2 = (63.66, 78.04) with 56 members, and C3 = (82.76, 76.72) with 58 members. By iteration 7, the centers had adjusted to C1 = (71.35, 57.32) with 74 members, C2 = (60.00, 78.75) with 52 members, and C3 = (81.83, 77.33) with 57 members. By the 12th iteration, the centers had further evolved to C1 = (73.52, 56.72) with 64 members, C2 = (59.25, 75.92) with 60 members, and C3 = (81.53, 78.31) with 59 members.

After the K-Means clustering is complete, the k-nearest Neighbors (k-NN) method is used to make predictions. It utilizes the identified clusters as a basis for making more accurate predictions about student learning outcomes. For instance, a student scored 85 in mathematics and 75 in statistics. Based on this data and using the results of the previous clustering, which category would be most suitable for this student?

The stages of resolution based on the case are as follows.

- Determine the value of k, let's say k = 61.
- Next, calculate the shortest distance of the data point to be evaluated, x = (85, 75), using Euclidean distance with respect to the clustered data.

$$d_{11} = \sqrt{(85-90)^2 + (75-70)^2} = 7.07$$

The following Table 4 shows the results of the k-NN calculation.

Table 4. k-NN Calculation Results

Data To	Arithmetic	Statistics	Category	d_{ij}
1	90	70	3	7.07
2	50	70	2	35.36
3	55	80	2	30.41
4	70	80	2	15.81
5	65	50	1	32.02
6	60	65	2	26.93
7	75	85	3	14.14
8	55	60	2	33.54
9	55	75	2	30.00
10	80	85	3	11.18
…	…	…	…	…
175	65	75	2	20.00
176	90	75	3	5.00
177	70	60	1	21.21
178	50	60	2	38.08
179	55	65	2	31.62
180	50	85	2	36.40
181	60	65	2	26.93
182	60	65	2	26.93
183	80	55	1	20.62

- After the calculations are completed, sort them based on the nearest distance as shown in Table 5.

Table 5. Closest Distance Sorting Results

Data To	Arithmetic	Statistics	Category	d_{ij}
20	85	75	3	0.00
83	85	75	3	0.00
158	85	75	3	0.00
56	85	80	3	5.00
72	80	75	3	5.00
75	85	70	3	5.00
…	…	…	…	…
62	75	85	3	14.14
58	70	75	2	15.00
60	70	75	2	15.00
77	70	75	2	15.00
87	85	60	1	15.00
125	85	60	1	15.00
134	85	60	1	15.00
139	85	90	3	15.00

- The result of sorting by nearest distance with k = 61, shows 7 data in Category 1, 3 data in Category 2, and 51 data in Category 3. Therefore, the evaluated data falls into Category 3 or 'Good', with a score of 85 for arithmetic and 75 for statistics. Figure 4 displays the number of predictions made by k-NN for each category.

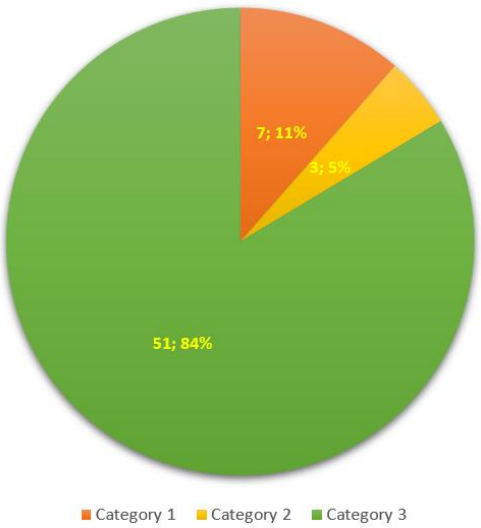

Fig. 4. Statistics of k-NN Calculation Results

The research results indicate that both methods function effectively in education. The student learning outcomes data are divided into three significant clusters based on the analysis results of the K-Means method. Each cluster exhibits different arithmetic and statistical abilities. These clusters are categorized as "Good," "Fair," and "Sufficient," based on the level of proficiency. The "Good" cluster consists of students with high scores who demonstrate consistent numerical strength. Students in the "Fair" group exhibit good abilities, but they lack consistency. Although the "Sufficient" group comprises students with lower scores, they show areas that need improvement. To predict student learning outcome categories based on scores after clustering, the k-NN model is utilized. The model effectively classifies new cases into the appropriate categories, showcasing the reliability of the k-NN model in predicting learning outcomes.

These findings enable educators to implement more tailored learning strategies to meet students' needs. Interventions can be targeted more precisely to individual student needs. Clustering and prediction results can be utilized to adapt curricula and educational resources, ensuring that subject matter is tailored to support students across all levels of ability.

Although the model demonstrates promising results, there are some constraints. For instance, it relies on the quality of input data and requires more data to generalize the outcomes. It is recommended to utilize larger and more diverse datasets, as well as explore additional model parameters that could potentially enhance prediction accuracy.

5. Conclusion

Research utilizing the integration of K-Means and k-NN for mapping and predicting student learning outcomes demonstrates that combining K-Means for clustering and k-NN for classification is an effective methodology for mapping and predicting students' learning outcomes. Initially, K-Means clustering identified three distinct groups with initial cluster centers at C1 = (45, 50), C2 = (70, 80), and C3 = (85, 90). Over twelve iterations, the clusters underwent significant evolution: C1 shifted to (73.52, 56.72) with 64 members classified as 'Adequate', C2 adjusted to (59.25, 75.92) with 60 members classified as 'Moderate', and C3 moved to (81.53, 78.31) with 59 members classified as 'Good'. The continual refinement of clusters enables a better understanding and response to various educational needs. Additionally, the k-NN model, trained using the clustered data, demonstrated strong predictive capabilities, correctly classifying a new instance (85, 75) with k = 61 into Category 3 ('Good').

This research enriches the literature on the application of data analysis in educational practices. It supports data-driven decision-making in education and opens avenues for applying these findings in educational contexts. The integration of these techniques provides educators with powerful tools to anticipate and meet student needs more effectively, allowing for targeted interventions that are both focused and personalized.

However, this study has several limitations that need to be addressed. First, the dataset used in this research is relatively small and sourced from a specific geographic region (Southeast Sulawesi, Indonesia), which may limit the generalizability of the findings to broader populations. Second, the study relies solely on Euclidean distance as a similarity metric without exploring alternatives such as Manhattan or Minkowski distance that could potentially yield different results depending on data characteristics. Additionally, the choice of the k value in k-NN is based on a fixed number (k = 61) and not dynamically optimized through techniques like grid search or advanced hyperparameter tuning.

For future research, it is recommended to expand the dataset to include more diverse student populations across multiple regions to enhance the external validity of the model. Researchers could also explore hybrid distance measures and integrate feature selection or dimensionality reduction techniques to refine model input. Furthermore, comparing this integrated approach with other machine learning methods such as Random Forests, Support Vector Machines, or

Neural Networks could provide a more comprehensive evaluation of performance. Continued validation through cross-validation and real-world educational deployment would further strengthen the model's reliability and practical utility.

References

[1] T. M. Ghazal *et al.*, "Performances of k-means clustering algorithm with different distance metrics," *Intell. Autom. Soft Comput.*, vol. 30, no. 2, pp. 735–742, 2021, doi: 10.32604/iasc.2021.019067.

[2] M. Cui, "Introduction to the K-Means Clustering Algorithm Based on the Elbow Method," *Accounting, Audit. Financ.*, vol. 1, pp. 5–8, 2020, doi: 10.23977/accaf.2020.010102.

[3] A. Jahwar, "Meta-Heuristic Algorithms for K-means Clustering: A Review," *PalArch's J. Archaeol. Egypt/Egyptology*, vol. 17, no. 7, pp. 7–9, 2021, [Online]. Available: https://archives.palarch.nl/index.php/jae/article/view/4630.

[4] K. K. Sharma and A. Seal, "Clustering analysis using an adaptive fused distance," *Eng. Appl. Artif. Intell.*, vol. 96, no. March, p. 103928, 2020, doi: 10.1016/j.engappai.2020.103928.

[5] J. J. Stankovic, I. Marjanovic, S. Drezgic, and Z. Popovic, "The digital competitiveness of european countries: A multiple-criteria approach," *J. Compet.*, vol. 13, no. 2, pp. 117–134, 2021, doi: 10.7441/JOC.2021.02.07.

[6] J. Pfitzinger and N. Katzke, "A Constrained Hierarchical Risk Parity Algorithm with Cluster-based Capital Allocation," pp. 1–26, 2021.

[7] S. Balovsyak, O. Derevyanchuk, H. Kravchenko, Y. Ushenko, and Z. Hu, "Clustering Students According to their Academic Achievement Using Fuzzy Logic," *Int. J. Mod. Educ. Comput. Sci.*, vol. 15, no. 6, pp. 31–43, 2023.

[8] R. Andrian, M. A. Naufal, B. Hermanto, A. Junaidi, and F. R. Lumbanraja, "K-Nearest Neighbor (k-NN) Classification for Recognition of the Batik Lampung Motifs," *J. Phys. Conf. Ser.*, vol. 1338, no. 1, 2019, doi: 10.1088/1742-6596/1338/1/012061.

[9] S. Chimphlee and W. Chimphlee, "Machine learning to improve the performance of anomaly-based network intrusion detection in big data," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 30, no. 2, pp. 1106–1119, 2023, doi: 10.11591/ijeecs.v30.i2.pp1106-1119.

[10] N. S. B. Mat Said, H. Madzin, S. K. Ali, and N. S. Beng, "Comparison of color-based feature extraction methods in banana leaf diseases classification using SVM and K-NN," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 24, no. 3, pp. 1523–1533, 2021, doi: 10.11591/ijeecs.v24.i3.pp1523-1533.

[11] S. M. H. M. Huzir, N. Z. Mahabob, A. F. M. Amidon, N. Ismail, Z. M. Yusoff, and M. N. Taib, "A preliminary study on the intelligent model of k-nearest neighbor for agarwood oil quality grading," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 27, no. 3, pp. 1358–1365, 2022, doi: 10.11591/ijeecs.v27.i3.pp1358-1365.

[12] H. F. El-Sofany and N. El-Haggar, "The effectiveness of using mobile learning techniques to improve learning outcomes in higher education," *Int. J. Interact. Mob. Technol.*, vol. 14, no. 8, pp. 4–18, 2020, doi: 10.3991/IJIM.V14I08.13125.

[13] H. M. Asim, A. Vaz, A. Ahmed, and S. Sadiq, "A Review on Outcome Based Education and Factors That Impact Student Learning Outcomes in Tertiary Education System," *Int. Educ. Stud.*, vol. 14, no. 2, p. 1, 2021, doi: 10.5539/ies.v14n2p1.

[14] A. Naim, "Applications of E-Learning Tools for Achieving Students Learning Outcomes," *J ournal Pedagog. Invent. Pract.*, vol. 2, no. 1, pp. 125–135, 2021.

[15] M. Usman, I. N. I, S. Utaya, and D. Kuswandi, "The Influence of JIGSAW Learning Model and Discovery Learning on Learning Discipline and Learning Outcomes," *Pegem Egit. ve Ogr. Derg.*, vol. 12, no. 2, pp. 166–178, 2022, doi: 10.47750/pegegog.12.02.17.

[16] M. Imron, U. Hasanah, and B. Humaidi, "Analysis of Data Mining Using K-Means Clustering Algorithm for Product Grouping," *IJIIS Int. J. Informatics Inf. Syst.*, vol. 3, no. 1, pp. 12–22, 2020, doi: 10.47738/ijiis.v3i1.3.

[17] R. T. Vulandari, W. L. Y. Saptomo, and D. W. Aditama, "Application of K-Means Clustering in Mapping of Central Java Crime Area," *Indones. J. Appl. Stat.*, vol. 3, no. 1, p. 38, 2020, doi: 10.13057/ijas.v3i1.40984.

[18] W. Nengsih and M. Mahrus Zain, "Descriptive Modeling Uses K-Means Clustering for Employee Presence Mapping," *Int. J. Inf. Eng. Electron. Bus.*, vol. 12, no. 4, pp. 15–20, 2020.

[19] H. H. Zhao, X. C. Luo, R. Ma, and X. Lu, "An Extended Regularized K-Means Clustering Approach for High-Dimensional Customer Segmentation with Correlated Variables," *IEEE Access*, vol. 9, pp. 48405–48412, 2021, doi: 10.1109/ACCESS.2021.3067499..

[20] G. Y. Iskandarli, "Applying Clustering and Topic Modeling to Automatic Analysis of Citizens' Comments in EGovernment," *Int. J. Inf. Technol. Comput. Sci.*, vol. 12, no. 6, pp. 1–10, 2020.

[21] M. T. Sathe and A. C. Adamuthe, "Comparative study of supervised algorithms for prediction of students' performance," *Int. J. Mod. Educ. Comput. Sci.*, vol. 13, no. 1, pp. 1–21, 2021.

[22] N. Hidayati and A. Hermawan, "K-Nearest Neighbor (K-NN) algorithm with Euclidean and Manhattan in classification of student graduation," *J. Eng. Appl. Technol.*, vol. 2, no. 2, pp. 86–91, 2021, doi: 10.21831/jeatech.v2i2.42777.

[23] P. J. S. Ferreira, J. M. P. Cardoso, and J. Mendes-Moreira, "KNN prototyping schemes for embedded human activity recognition with online learning," *Computers*, vol. 9, no. 4, pp. 1–20, 2020, doi: 10.3390/computers9040096.

[24] W. Liu, P. Wang, Y. Meng, C. Zhao, and Z. Zhang, "Cloud spot instance price prediction using kNN regression," *Human-centric Comput. Inf. Sci.*, vol. 10, no. 1, 2020, doi: 10.1186/s13673-020-00239-5.

[25] A. Priyadarshini, S. Mishra, D. P. Mishra, S. R. Salkuti, and R. Mohanty, "Fraudulent credit card transaction detection using soft computing techniques," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 23, no. 3, pp. 1634–1642, 2021, doi: 10.11591/ijeecs.v23.i3.pp1634-1642.

[26] Fathoni, Erwin, and Abdiansah, "Multilabel sentiment analysis for classification of the spread of COVID-19 in Indonesia using machine learning," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 31, no. 2, pp. 968–978, 2023, doi: 10.11591/ijeecs.v31.i2.pp968-978.

[27] S. Huang, Z. Kang, Z. Xu, and Q. Liu, "Robust deep k-means: An effective and simple method for data clustering," *Pattern Recognit.*, vol. 117, p. 107996, 2021, doi: 10.1016/j.patcog.2021.107996.

[28] P. Govender and V. Sivakumar, *Application of K-Means and Hierarchical Clustering Techniques for Analysis of Air Pollution: A Review (1980–2019)*, vol. 11, no. 1. Turkish National Committee for Air Pollution Research and Control, 2020.
[29] A. Fahim, "Finding the Number of Clusters in Data and Better Initial Centers for K-means Algorithm," *Int. J. Intell. Syst. Appl.*, vol. 12, no. 6, pp. 1–20, 2020.
[30] A. Rafael Braga, D. G. Gomes, B. M. Freitas, and J. A. Cazier, "A cluster-classification method for accurate mining of seasonal honey bee patterns," *Ecol. Inform.*, vol. 59, p. 101107, 2020, doi: 10.1016/j.ecoinf.2020.101107.
[31] C. Shi, B. Wei, S. Wei, W. Wang, H. Liu, and J. Liu, "A quantitative discriminant method of elbow point for the optimal number of clusters in clustering algorithm," *Eurasip J. Wirel. Commun. Netw.*, vol. 2021, no. 1, 2021, doi: 10.1186/s13638-021-01910-w.
[32] F. Liu and Y. Deng, "Determine the Number of Unknown Targets in Open World Based on Elbow Method," *IEEE Trans. Fuzzy Syst.*, vol. 29, no. 5, pp. 986–995, 2021, doi: 10.1109/TFUZZ.2020.2966182.
[33] R. Sammouda and A. El-Zaart, "An Optimized Approach for Prostate Image Segmentation Using K-Means Clustering Algorithm with Elbow Method," *Comput. Intell. Neurosci.*, vol. 2021, 2021, doi: 10.1155/2021/4553832.
[34] Z. K. Maseer, R. Yusof, B. Al-Bander, A. Saif, and Q. K. Kadhim, "Meta-Analysis and Systematic Review for Anomaly Network Intrusion Detection Systems: Detection Methods, Dataset, Validation Methodology, and Challenges," 2023.
[35] J. Salvador-Meneses, Z. Ruiz-Chavez, and J. Garcia-Rodriguez, "Compressed kNN: K-nearest neighbors with data compression," *Entropy*, vol. 21, no. 3, pp. 1–20, 2019, doi: 10.3390/e21030234.
[36] A. Sagar, C. Vega, O. Bouriaud, C. Piedallu, and J. P. Renaud, "Multisource forest inventories: A model-based approach using k-NN to reconcile forest attributes statistics and map products," *ISPRS J. Photogramm. Remote Sens.*, vol. 192, no. August, pp. 175–188, 2022, doi: 10.1016/j.isprsjprs.2022.08.016.
[37] J. Hosseinzadeh, F. Masoodzadeh, and E. Roshandel, "Fault detection and classification in smart grids using augmented K-NN algorithm," *SN Appl. Sci.*, vol. 1, no. 12, 2019, doi: 10.1007/s42452-019-1672-0.

Authors' Profiles

Muh. Nurtanzis Sutoyo, S.Kom.,M.Cs.,IPP is a permanent lecturer in the Information Systems Program at the Faculty of Information Technology Universitas Sembilanbelas November Kolaka, Southeast Sulawesi. They are the eldest of three siblings. They completed their Bachelor's program (S1) at STIMIK Bina Bangsa Kendari (2008), finished their Master's program (S2) at Gadjah Mada University in the Computer Science Program with a concentration in Intelligent Systems (2015), and completed the Professional Engineer Education at Hasanuddin University (2022).

Alders Paliling, S.Kom., MT is a lecturer in computer science at the Universitas Sembilanbelas November Kolaka Indonesia, with expertise in virtual reality, augmented reality, and programming. He completed his undergraduate education at STMIK Dipanegara Makassar, Computer science study program in 2010. He received a master's degree in electrical engineering at Hasanuddin University in 2014. He is also a practitioner in the field of computers.

How to cite this paper: Muh. Nurtanzis Sutoyo, Alders Paliling, "Optimizing Student Performance Prediction via K-Means and k-NN Integration", International Journal of Education and Management Engineering (IJEME), Vol.15, No.4, pp. 12-22, 2025. DOI:10.5815/ijeme.2025.04.02