INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

RESEARCH ARTICLE

OPEN ACCESS

Received: 26-08-2023

Accepted: 05-01-2024

Published: 02-02-2024

Citation: Sofia G, Hema D (2024) Clustering-based Recommendations for Enhancing Students' Academic Performance by Recognizing Prevalent Assessment Method using Exploratory Data Analysis. Indian Journal of Science and Technology 17(6): 503-513. https://doi.org/10.17485/IJST/v17i6.2175

* **Corresponding author**.

joesofi@gmail.com

Funding: None

Competing Interests: None

Published By Indian Society for Education and Environment (iSee)

ISSN
Print: 0974-6846
Electronic: 0974-5645

Clustering-based Recommendations for Enhancing Students' Academic Performance by Recognizing Prevalent Assessment Method using Exploratory Data Analysis

G Sofia[1]*, D Hema[2]

1 Associate Professor in Computer Science, Lady Doak College, Madurai, Tamil Nadu, India
2 Assistant Professor in Computer Science, Lady Doak College, Madurai, Tamil Nadu, India

Abstract

Objectives: To analyse students' academic performance based on assessment methods and determine the most prevalent one through which students can be categorised for recommending optimal student-centered pedagogies that enhance students' performance. **Methods**: Exploratory Data Analysis identifies the implications of the assessment methods based on the marks obtained by students in Continuous Assessments (CA) and the Cumulative Test (CT). Continuous Assessment (CA) and Cumulative Test (CT) marks of three subjects that come under foundation science, elective, and skill-based course of 100 undergraduate students are collected from a reputed Arts and Science Institution using stratified sampling technique, analyzed, and the recommendations are made based on the statistical observations and cluster analysis. Clustering recognises learning patterns of the students' on the learners' data. The Elbow method determines the number of clusters where the Silhouette score identifies the best suitable clustering technique for the dataset. K-Means Clustering categorises students based on their performance, that helps to give recommendations to improve. **Findings**: Based on Univariate and Bivariate analysis on the dataset, this work identifies Continuous Assessment (CA) as a prevalent evaluation strategy that motivates students to get engaged throughout rather than just before the exam. Based on the Silhouette Score (above .5), K-Means clustering is chosen to discover hidden patterns in the assessment marks depending on the three clusters determined by the Elbow method. It helps to identify the underperformers (46%) and suggest personalised recommendations for improving student's academic performance as per clusters. **Novelty**: This work integrates Statistical Analysis and Clustering Analysis as per the optimal clusters determined by the Elbow method for identifying patterns hidden in assessment marks based on the prevalent assessment types. As a result, it enables more personalised recommendations for recognising the predominant assessment method and boosting academic achievement.

Keywords: Continuous Assessment; Cumulative Test; Statistical Analysis; Exploratory Data Analysis; Univariate; Bivariate; Cluster Analysis; Elbow Method; KMeans Clustering

1 Introduction

In recent years, Educational Data Mining (EDM) has emerged as one of the main terminologies in discovering useful and meaningful information that can be used by decision-makers to make active and knowledgeable data-driven decisions.[1] Educational Data Mining (EDM) focuses on using data mining techniques to extract massive data from the educational context and transform it into knowledge that can improve educational systems and decisions.[2] Student performance modelling is one of the challenging and popular research topics in Educational Data Mining (EDM).[3] The increasing demand for excellence in all areas of life has led to the need to evaluate, monitor and predict the possible Academic Performance (AP) outcome of students at an early stage in their studentship.[4] The evaluation and prediction of student academic performance have always been important parts in EDM. Institutions have enabled the emergence of new strategies aiming at improving students' AP. These strategies include, conducting peer learning and remedial classes for slow learners, multiple admission modes and programmes, adopting student-centered teaching and learning methods and different instructional strategies for outstanding students and so on. Hence, it is necessary to determine students' AP based on different assessment methods to identify students of poor performers at an early stage of their studies and address them with interventions.

[5] As others have argued, one goal of assessment practices is to foster learning. Assessments are used to measure student's performance, assign grades, compare students, meet requirements by the professional body and attain required benchmarks. The evaluation/assessment pattern might be continuous or cumulative, allowing for the continuous improvement of the learning process. All activities undertaken by teachers and students in assessing themselves provide information that can be utilized as feedback to modify the teaching and learning activities. Ultimately it improves the quality of teaching and learning. Continuous Assessment (CA) is a formative assessment through which students will be assessed continually throughout the learning process, and will be graded continuously for a while. It can be evaluated using a variety of components such as MCQs, course-related projects, practical work, assignments, chart presentations, model creation, seminars, open book tests, field trips, and other student-centered activities. Cumulative test (CT) is an assessment in which a student is examined on the overall course content through the descriptive test at the end of the course of study to evaluate learning and long-term retention.

Review of the literature clearly stated that earlier researchers have conducted multiple surveys on EDM studies. Some of them considered a broader scope to outline multiple aspects of educational processes.[2] Although several EDM surveys are available in the literature, only a few specific surveys were on student's performance analysis and prediction. Related work with their findings in the education domain is listed in Table 1.

Recently, some studies have applied cluster analysis for evaluating the students' results and utilised statistical techniques to part their scores regarding students' performance which was not found as efficient. However, the study that combines two techniques, namely, k-means and elbow clustering algorithm to evaluate the student's performance performs better in analyzing and evaluating the progress of the student's performance. Moreover, no previous work focuses solely on the identification of prevalent assessment methods for measuring the student's academic performance addressing them through recommendations based on the clusters to which they belong.

Hence, in this study, the methodology has been implemented to define the diverse fascinating model by applying statistical Analysis and Clustering Analysis along with the Elbow method for determining an optimum number of clusters.

Table 1. Comparative study of related work

Related work	Description	Findings
[1] (2023) Clustering student performance data using k-means algorithms	Students were grouped into four clusters based on their characteristics and performance in school	Found that gender and age of the students play an important role in identifying student performance. Clusters help educators to identify students with the highest risk of failing and underperforming.
[4] (2022) Analysis and Prediction of Students' Academic Performance Based on Educational Data Mining	Clustering-number determination is optimized by using a statistic that has never been used in the algorithm of K-means. Uses the relevant theories of clustering, discrimination and convolution neural network to analyze and predict students' academic performance	Only compulsory course data was analyzed. Elective courses were not included for the analysis
[6] (2022) Data Analysis of Educational Evaluation Using K-Means Clustering Method	K-means technique clusters and analyzes the comprehensive evaluation results. Students are categorized based on their commonality, allowing student managers to provide tailored education management for various types of students	Changes in Student achievement concerning time index after the implementation of targeted education management are reported
[7] (2021) Research on the Application of K-Means Clustering Algorithm in Student Achievement	K-Means algorithm is applied to perform cluster analysis on the final grade data of students majoring in software and information services in a university	Finds subjects with higher importance, and provides a reference for teaching management that can adjust teachers and class hours according to the importance.
[8] (2021) An integrated clustering method for pedagogical performance	Captured underlying rules of association among the students' data attributes using clustering techniques in an interdisciplinary context	14 clusters are made available for potential future examinations
[9] (2020) Clustering Approach for Analyzing the Student's Efficiency and Performance Based on Data	Combined two techniques, namely, k-means clustering algorithm and elbow method to evaluate the student's performance based on Grade and GPA	Students were clustered based on K value

The fundamental goal of statistical analysis in educational data is to uncover patterns that will improve teaching-learning pedagogies and identify new trends in blended, flipped, and face-to-face learning environments in participatory learning. In terms of student performance evaluation, cluster analysis will grasp the features of each category of students more objectively and impartially, create a more reasonable classification of student groups, and provide teachers with personalized policy assistance.[7] The evaluation obtained by cluster analysis is more scientific and reasonable, without subjective factors, and provides decision-making reference for teachers' group teaching and personalized guidance.

Therefore, this work focuses on analyzing students' academic performance in Continuous Assessments and Cumulative Test to determine the optimum assessment method using Statistical Analysis and observe hidden patterns among students' assessment marks, allowing students to be grouped into clusters and make recommendations for improving academic performance using Cluster Analysis based on Elbow method. The outline of this paper is as follows. Following the introduction with review of literature and analysis of related work in Section 1, Section 2 elaborates on the Methodology adopted for Statistical Analysis and Clustering Analysis for the recommendations. Section 3 presents the Results and Discussion by illustrating the recommendations based on the statistical observations and cluster analysis. Finally, Section 4 concludes the work.

2 Methodology

The work is accomplished through the following steps. First, the acquired data was organized and presented for the statistical analysis of students' academic performance using Exploratory Data Analysis. The prominent assessment method is recognized based on the interpretations. Following that, the number of clusters applicable to the dataset is found using the Elbow method, after which the appropriate clustering method is applied based on which recommendations are made. The Process flow of this work is given in Figure 1.

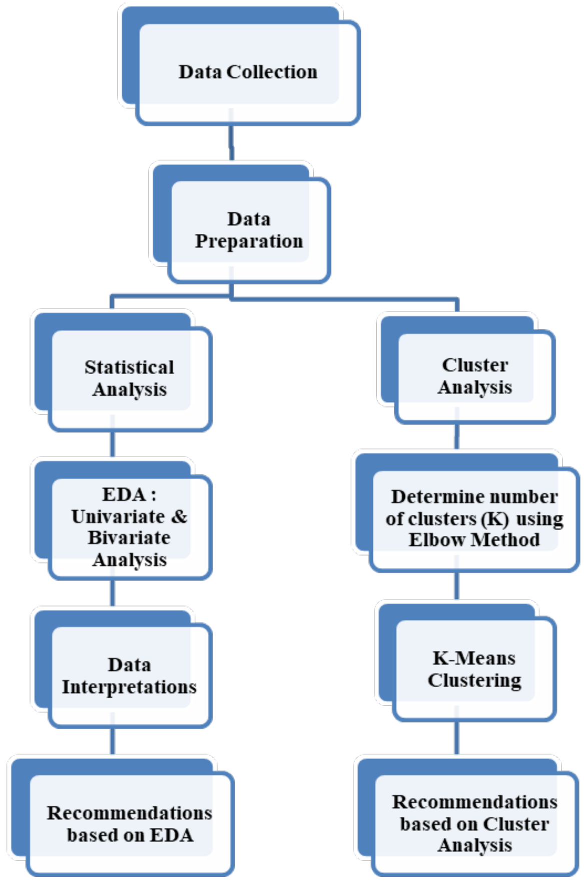

Fig 1. Process Flow chart

2.1 Data Collection

Data was collected from three distinct courses that come under foundation science, elective and skill-based course in a reputed institution. The aggregate of Continuous Assessment (CA) marks and Cumulative Test (CT) marks of 100 first-year Undergraduate College students scored in three different subjects were collected for this study. The performance of the students is measured by the scores of formative assessments conducted as Continuous Assessments or Cumulative Test and the final grade is calculated by combining them with summative assessment results.

2.2 Data Preparation

Data preparation includes Data Organization and Data Presentation. As the collected data must be organized for accurate analysis, inconsistencies have to be identified and removed by data cleaning. The Continuous Assessment (CA) continuously monitors and evaluates students' progress or performance and thus it involves more than one assessment component. As a result, there is a possibility of missing values due to the absence of students for a component. Hence, the data is organized by identifying missing values in the Continuous Assessment scores replacing them with 0.

Data Presentation is an extension of data cleaning in which data is organised for simple analysis. Descriptive statistical techniques are the most effective way of presenting and summarising the fundamental characteristics of numerical data based on its arrangement. The descriptive statistical tools Mean, Median, Interquartile Range, Skewness, and Kurtosis are applied in this study to present the data for further research.

2.3 Statistical Analysis

Statistical analysis has recently gained prominence in several significant areas for analysing quantitative data to uncover patterns and make conclusions. Exploratory Data Analysis (EDA) is an essential phase of Statistical Analysis that involves analysing a dataset to uncover patterns, anomalies (outliers), and form hypotheses based on the data.

2.3.1 Exploratory Data Analysis

EDA can be performed in three ways: Univariate, Bivariate, and Multivariate Analysis. The Univariate and Bivariate tools are used to perform essential descriptive statistical operations in this work.

- **Univariate Analysis**

Univariate analysis describes a specific data set and identifies patterns of data based on a single variable. As a result, it computes summary statistics for CA and CT scores independently, displaying all patterns and shapes of the distribution.

- **Bivariate Analysis**

This method determines the association between CA and CT to discover the relationship between them. The correlation between CA and CT is calculated and the correlation matrix shows the relationship that exists between them. Independent Samples T-test also known as the two-sample t-test is used to compare the average or mean to determine whether there is a significant difference between the means of the independent groups CA and CT separately. However, to perform the tests, the null and alternate hypotheses are defined:

- **Null Hypothesis**: There is no noteworthy difference between the groups representing CA and CT.
- **Alternate Hypothesis**: There is a noteworthy difference between the groups representing CA and CT.

Determine the p-value for the desired significance level (α = 0.05 or 0.01). If the p-value is smaller than 0.05, then the null hypothesis is rejected, and the alternative hypothesis is accepted. The t-score, also known as the t-value or t-statistic, is a statistical measure used in hypothesis testing and estimating parameters when the sample size is small and the population standard deviation is unknown. The t-score is calculated using the following formula:

$$t = (x - \mu)/(s/sqrt(n)) \quad (1)$$

where,

- x is the sample mean,
- μ is the population mean,
- s is the sample standard deviation,
- n is the sample size.

If the t-score is more significant, this means that there is more difference present between the groups representing CA and CT scores. At the same time, the smaller t-score signifies the similarities between the groups. In hypothesis testing, the t-score is compared to critical values from the t-distribution to determine if the observed difference is statistically significant. If the absolute value of the t-score is larger than the critical value (α = 0.05 or α = 0.01), it suggests that the observed difference is statistically significant, and the null hypothesis is rejected which infers the significant difference between CA and CT.

2.3.2 Data Interpretations and Recommendations

Inferential Statistics entails strategies for forming inferences and drawing conclusions from sample data to infer population opinion. Based on this study, data interpretation generates insights and conclusive results. Based on the evaluated CA and CT scores, interpretations are made, and the optimal assessment technique for improving the effective learning process is recognized. To make recommendations based on interpretations, the findings are presented as charts and reports.

2.4 Cluster Analysis

[10] Machine learning techniques can be utilized to predict the output of the students and identifying the at risk students as early as possible so appropriate actions can be taken to enhance their performance. The student's performance prediction is an essential area as it can help teachers identify students that need additional academic assistance.[11] Clustering is an aspect of machine learning that is of great importance in the area of data mining analysis. Furthermore, clustering involves grouping sets of related data objects into the same group (clusters) considering their unique qualities and similarity.[3] Cluster Analysis (CA) is an important task in data analysis. It aims at revealing implicit structures that were hidden, interesting patterns and relations from datasets, and then adapting the extracted information for the analysis task to ease comparison, interpretation

and relationship assessment.[1] There are three major techniques in the clustering approach: partitioning, hierarchical, and density-based.

Silhouette Score is a metric used for evaluating the performances of clustering methods. The silhouette score for a single data point i is given by the formula:

$$s(i) = \{b(i) - a(i)\}/max\{a(i), b(i)\} \tag{2}$$

where,

- s(i) is the silhouette score for data point i,
- a(i) is the average distance from the i-th data point to other data points in the same cluster,
- b(i) is the smallest average distance from the i-th data point to data points in a different cluster, minimized over clusters.

The silhouette score for the entire dataset is the average of the silhouette score for each data point.[12] Since the Silhouette index does not need a training set to evaluate the clustering performance, it is more relevant to the clustering concept. Hence, this work applied a partitioning based clustering method and evaluated the performance using Silhouette score.

2.4.1 Elbow Method to determine K

The Elbow method determines the optimal number of clusters based on the number of centroids (K) in a dataset for clustering algorithms. It is performed by the interpretation of a line plot with an elbow shape. WCSS is the sum of the squared distance between each point and the centroid in a cluster.

When the WCSS with the k value is plotted in the line, the point at which the bend occurs specifies the optimal number of clusters that can be applied for clustering.

2.4.2 K-Means clustering

As per the various research works done on clustering analysis in education, the K-Means clustering algorithm is one of the predominant methods applied in clustering analysis for the educational environment.[13] K-means, one of the important unsupervised algorithms, is used to identify clusters with the similar characteristics and properties by working with gravity centres. It aims to partition the data into a predefined number of clusters (K), where each data point belongs to the cluster with the nearest mean.

The k-means clustering algorithm is based on the minimization of the sum of squared distances (Euclidean distances) between data points and the centroids of their assigned clusters. The objective function, also known as the ”inertia” or ”within-cluster sum of squares,” is minimized during the optimization process. Here’s the mathematical formula for the k-means objective function:

Given a set of data points X={x1, x2,…, xn} and k cluster centroids C={c1, c2,…, ck}, the objective function is defined as follows :

$$J = \sum_{i=1}^{n} min_{j=1}^{k} \left\|x_i - c_j\right\|^2 \tag{3}$$

where,

- n is the number of data points,
- k is the number of clusters,
- x_i represents the i-th data point,
- c_j represents the j-th cluster centroid.

The algorithm iteratively updates the cluster centroids and assigns data points to clusters until convergence. The optimal number of clusters is usually obtained through several methods, including elbow method, average silhouette method and, gap statistic method.

K-means clustering was applied on the dataset with CA and CT marks for dividing the data points into a number of clusters (k) determined by Elbow method. K-means iteratively minimizes the sum of squared distances between data points and their respective cluster centroids which was visualized using a scatter plot.

2.4.3 Recommendations based on Cluster Analysis

After performing k-means clustering on a dataset, the cluster assignments and cluster centers are used to make various recommendations and gain insights about the data. In this work, clustering of students based on their performance in

CA and CT was performed, and they were represented in different groups. Identification of student subgroups based on their performance in CA and CT potentially aids in planning for personalized strategies to be adopted for enhancing their performance.

3 Results and Discussion

Python is predominantly used in data science to analyze data based on statistical observations and provide recommendations. Hence, this work uses the dataset with attributes representing Continuous Assessment (CA) and Cumulative Test (CT) marks of 100 students from three subjects that are collected using stratified sampling technique.

The data was analysed using Python and the recommendations are made based on the statistical observations and cluster analysis. The following observations are made based on the univariate and bivariate analyses of the dataset.

3.1 Exploratory Data Analysis

3.1.1 Univariate Analysis:

The tabulated values in Table 2 summarises the descriptive statistics based on measures of central tendency and measures of dispersion on the CA and CT scores.

Table 2. Results of Descriptive Statistics

Index	CA	CT
Count	100.0	100.0
Mean	70.97	56.06
Standard Deviation (SD)	9.00	10.76
Min	50.0	27.0
First Quartile (25%)	63.0	50.0
Median	70.0	53.0
Third Quartile (75%)	79.0	62.0
Inter Quartile Range (IQR)	16	12
Max	90.0	83.0

3.1.2 Interpretations based on Descriptive Statistics:

- The lowest mark is obtained in CT, while the highest mark is obtained in CA, indicating that the students were able to perform well and obtain a high score in Continuous Assessments (CA) rather than Cumulative Tests (CT).
- The higher Mean and Median value in CA indicates that the average student's score is higher in CA than in CT.
- The distribution of marks in CA dominates CT in all quartile ranges.
- CA has a lower standard deviation than CT, implying that CA marks are distributed closer to mean values.
- A Lower CA Range(Max-Min) indicates that the dispersion of scores in CA is lower than that of CT.
- The Interquartile range (Q3-Q1) shows that the middle part of data in CA spreads less than CT.

Table 3. Measures of Skewness & Kurtosis

Measures	CA	CT
Coefficient of Skewness	-0.0444	0.5272
Coefficient of Kurtosis	-0.7423	0.6149

The skewness and kurtosis measures are used to determine the shape of the data as given in Table 3.

- The coefficient of kurtosis of CA is less than 3. Hence, the data distribution is platykurtic indicating uniformity in the marks with few extreme outliers.
- In the case of CA, the Coefficient of skewness < 0 and the density plot of CA indicates that it is more negatively skewed i.e. the graph is said to be negatively skewed (left-skewed) with the majority of data values greater than the mean. As illustrated in Figure 2, the majority of the data are clustered on the right side of the graph. However, in the case of CT,

the Coefficient of skewness is > 0 and the density plot of CT reveals that it is more positively skewed ie. the graph is said to be positively skewed (right-skewed) with the majority of data values smaller than the mean. The majority of data are concentrated on the left side of the graph as shown in Figure 2.

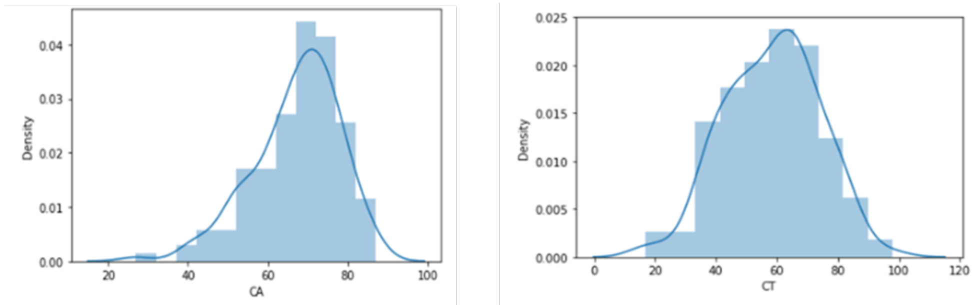

Fig 2. Density Plot for Assessment Marks

The measures of dispersion illustrate that CA has lower dispersion than CT, indicating that CA marks are dispersed closer to the mean and median. It also shows that more students can score higher than average in CA rather than CT. It was also validated by the measure of skewness.

3.1.3 Bivariate Analysis

The relationship between the assessment methods is verified by the measure of the degree of association or correlation between them and independent sample T-test.

3.1.4 Interpretations based on Bivariate Analysis:

- Correlation Matrix in Table 4 shows the correlation between CA & CT. As Pearson's coefficient value is closer to zero, CA & CT are not significantly correlated.

Table 4. Correlation Matrix

	CA	CT
CA	1.00000	0.209402
CT	0.209402	1.00000

- Independent Samples T-test compares the means of two groups CA and CT.

The interpretation of the independent sample T-test is given below.

H0: There is no significant difference between the mean of CA scores and CT scores

HA: There is a significant difference between the mean of CA scores and CT scores.

t value = 10.627963413507434

p value = 3.649380816030137e-21

Since the p-value (probability) is less than (alpha value) 0.05, we reject the null hypothesis. We have sufficient evidence to say that there is a significant difference in the performance of the students in CA and CT. As t value is positive, the mean of CA is greater than the mean of CT that implies CA performance is significantly different and better compared to the performance of students in CT.

3.1.5 Recommendations based on Exploratory Data Analysis

Based on the interpretations concerning univariate and bivariate analysis, it is recommended that Continuous Assessment be a prevalent evaluation method that encourages students to get engaged throughout the course rather than just before the exam. Continuous Assessment (CA) mode of testing is a better way to measure student learning. However, it may be possible that Continuous Assessment is more accurate for measuring overall student performance because students are given different types of tasks to assess their knowledge, such as homework, papers, and quizzes. It is likely to be the most sustained means of assessment in providing relatively prompt feedback that serves to reinforce or correct learned responses. Thus, a climate of continuous improvement of the individual toward personal and professional development is established.

3.2 Cluster Analysis

3.2.1 Elbow Method to determine K

As shown in Figure 3, the optimum value of 'K' determined by the Elbow method, to define the number of clusters for the clustering algorithms is 3 for the dataset used in this work.

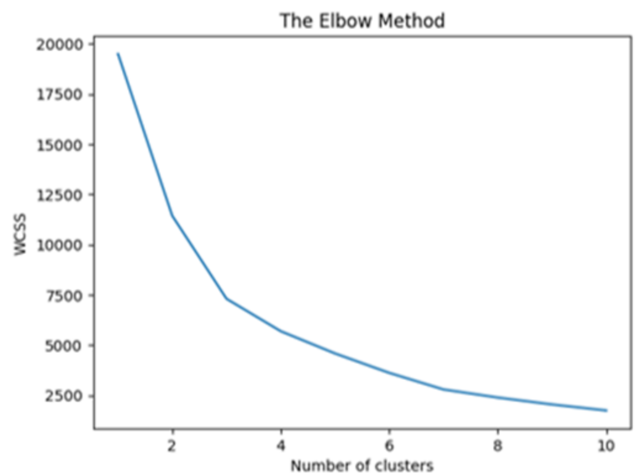

Fig 3. Defining optimum value of K

3.2.2 Applying Clustering Algorithms

Three major clustering algorithms K-Means, BIRCH and DBSCAN are applied on the dataset with the number of clusters as 3. The performance of the algorithms on this dataset is compared using the Silhouette score. Silhouette score can take the values between -1 and +1.

Table 5 clearly states that the K-Means clustering algorithm clusters the data better than the other algorithms. The assessment scores of the students are clustered and examined for hidden patterns to give recommendations.

Table 5. Comparison of Clustering Algorithms using Silhouette Score

Algorithm	Type	Metrics used	Silhouette Score
K-Means	Partitioning	Sum of Squared Distance between points	0.5926
BIRCH	Hierarchical	Euclidian Distance between points	0.49
DBSCAN	Density Based	Euclidian Distances between nearest points	0.0253

The scatter plot given Figure 4 visualizes the clustered data based on K- Means clustering with the respective centroids and the values are given in Table 6.

Table 6. Centroids of clusters

Clusters	No. of Data Points	Centroids (CA,CT)
1 (Violet)	46	63.19565217, 51.54347826
2 (Yellow)	27	79.48148148 50.18518519
3 (Blue)	27	75.7037037, 69.62962963

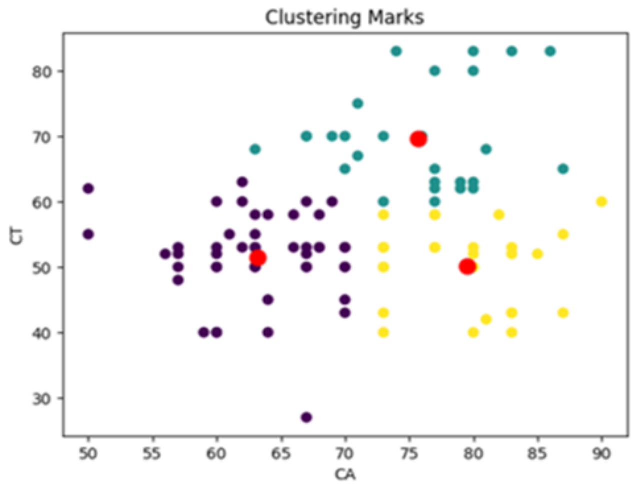

Fig 4. Applying K-Means Clustering with Centroid values

3.2.3 Recommendations based on Cluster Analysis

Students were categorized into 3 clusters based on their CA and CT marks and the proposed recommendations for improving their performance as per the cluster are given below.

- **Violet Cluster**: 46% of the students are with low CA marks and low CT Marks. This category of students are slow learners who need extra care.
- **Yellow Cluster**: 27% of the students are with low CT marks but a High CA marks. This category of students can perform well with minimum portions but not with the content as a whole. They are recommended to improve their performance in CT by giving more attention to cumulative Test.
- **Blue C luster**: 27% of the students are with High CA Marks and High CT Marks. This category of students is Good Performers.

This study investigated the impact of marks scored by the students in CA and CT on Academic Performance. Based on the interpretations of EDA, Continuous Assessment is recommended as the prevalent evaluation method. This study also identified that the k-means algorithm produces the best clustering compared to the BIRCH and DBSCAN algorithms based on the Silhouette score which also performs well on a data set of smaller size. Applying the Elbow method demonstrated that the optimal number of clusters for this dataset is 3. Hence, implementing k-means produced reliable and consistent clusters of students as per CA, and CT marks and recommendations are made for improving their academic performance based on their level.

4 Conclusion

This work contributes to the applications of Statistical Analysis and Clustering Analysis to determine their potential within the educational context. According to the findings of this study, Continuous Assessment (CA) is recommended as the most prevalent assessment method as it offers students a consistent stream of opportunities to demonstrate mastery of the material and has the potential to capture and recognise the whole range of learners' performance, which promotes good learning. The more frequently we examine students, the more they will discover about themselves, like how to approach problems differently, what their blind spots are, and how to eliminate them. As an outcome, it is a learning assessment. Thus, it assists academic planners in assessing students with optimal methods and tracking students' progress during their course of study. This work also recognises the student's performance and indicates areas that require attention using clustering analysis. In this work, three clusters were identified based on students' performance. This can help educators discover the students with a high risk of underperforming as well as the cause and characteristics of the drop in students' performance. A system like this can help students identify their strengths and weaknesses, focus on relevant study resources, and enhance their academic performance. A future extension of this work will be the evaluation of the assessment methods and cluster analysis to determine their potential impact on the major courses and the performance of the students in all dimensions.

References

1) Alawi SJS, Shaharanee INM, Jamil JM. Clustering student performance data using k-means algorithms. *Journal of Computational Innovation and Analytics*. 2023;2(1):41–55. Available from: https://doi.org/10.32890/jcia2023.2.1.3.
2) Khan A, Ghosh SK. Student performance analysis and prediction in classroom learning: A review of educational data mining studies. *Education and Information Technologies*. 2021;26(1):205–240. Available from: https://doi.org/10.1007/s10639-020-10230-3.
3) Inyang UG, Umoh UA, Nnaemeka IC, Robinson SA. Unsupervised Characterization and Visualization of Students' Academic Performance Features. *Computer and Information Science*. 2019;12(2):103–116. Available from: https://doi.org/10.5539/cis.v12n2p103.
4) Feng G, Fan M, Chen Y. Analysis and Prediction of Students' Academic Performance Based on Educational Data Mining. *IEEE Access*. 2022;10:19558–19571. Available from: https://doi.org/10.1109/ACCESS.2022.3151652.
5) Swiecki Z, Khosravi H, Chen G, Martinez-Maldonado R, Lodge JM, Milligan S, et al. Assessment in the age of artificial intelligence. *Computers and Education: Artificial Intelligence*. 2022;3:1–10. Available from: https://doi.org/10.1016/j.caeai.2022.100075.
6) Liu R. Data Analysis of Educational Evaluation Using K-Means Clustering Method. *Computational Intelligence and Neuroscience*. 2022;2022:1–10. Available from: https://doi.org/10.1155/2022/3762431.
7) Chi D. Research on the Application of K-Means Clustering Algorithm in Student Achievement. In: 2021 IEEE International Conference on Consumer Electronics and Computer Engineering (ICCECE). IEEE. 2021;p. 435–438. Available from: https://doi.org/10.1109/ICCECE51280.2021.9342164.
8) Said RA, Mwitondi KS. An integrated clustering method for pedagogical performance. *Array*. 2021;11:1–10. Available from: https://doi.org/10.1016/j.array.2021.100064.
9) Omar T, Alzahrani A, Zohdy M. Clustering Approach for Analyzing the Student's Efficiency and Performance Based on Data. *Journal of Data Analysis and Information Processing*. 2020;08(03):171–182. Available from: https://doi.org/10.4236/jdaip.2020.83010.
10) Oppong SO. Predicting Students' Performance Using Machine Learning Algorithms: A Review. *Asian Journal of Research in Computer Science*. 2023;16(3):128–148. Available from: https://doi.org/10.9734/ajrcos/2023/v16i3351.
11) Okolichukwu UV, Sunday BA, Onuodu FE. Review and Comparative Analysis of Data Clustering Algorithms. *International Journal of Research and Innovation in Applied Science (IJRIAS)* |. 2020;V(II):112–116. Available from: https://www.rsisinternational.org/journals/ijrias/DigitalLibrary/Vol.5&Issue2/112-116.pdf.
12) Shutaywi M, Kachouie NN. Silhouette Analysis for Performance Evaluation in Machine Learning with Applications to Clustering. *Entropy*. 2021;23(6):1–17. Available from: https://doi.org/10.3390/e23060759.
13) Souabi S, Retbi A, Idrissi MK, Bennani S. A Recommendation Approach in Social Learning Based on K-Means Clustering. *Advances in Science, Technology and Engineering Systems Journal*. 2021;6(1):719–725. Available from: https://www.astesj.com/publications/ASTESJ_060178.pdf.