Clustering-Based Academic Performance Assessment Using Educational Data

Saraswathi P
Department of Computer Science and Applications
SRM Institute of Science and Technology, Faculty of Liberal Arts and Business Studies, Vadapalani Campus, Chennai, Tamil Nadu, India

Abstract — Educational institutions generate large amounts of student data that can be analyzed to improve academic outcomes. This study applies clustering techniques to identify performance patterns among students based on factors such as attendance, assignment completion, assessment scores, and learning engagement. The K-Means clustering algorithm is used to group students with similar academic characteristics into distinct clusters. The analysis helps classify students into categories such as high-performing, average-performing, and low-performing groups. The discovered clusters provide valuable insights for educators to design targeted interventions, enhance student support, and improve overall academic performance. The results demonstrate that clustering is an effective educational data mining technique for identifying student learning patterns and supporting data-driven decision-making in higher education.

Keywords: Educational Data Mining, Clustering, K-Means Algorithm, Student Performance Analysis, Academic Achievement, Learning Analytics

I. INTRODUCTION

Educational institutions continuously collect large volumes of student-related data, including attendance records, assignment scores, examination results, and learning activities. Analyzing this data can provide valuable insights into student learning behavior and academic performance. Educational Data Mining (EDM) has emerged as an important research area that applies data mining techniques to discover meaningful patterns from educational datasets.

Clustering is one of the most widely used unsupervised learning techniques in data mining. It groups similar data objects into clusters based on their characteristics without requiring predefined class labels. In the educational domain, clustering helps identify groups of students with similar academic performance, learning habits, and engagement levels. Such information enables educators to provide personalized guidance and targeted interventions for improving student outcomes.

Among various clustering techniques, the K-Means algorithm is popular due to its simplicity, efficiency, and effectiveness in handling large datasets. The algorithm partitions data into a predefined number of clusters by minimizing the distance between data points and their respective cluster centroids. By applying K-Means clustering to student performance data, students can be categorized into groups such as high-performing, average-performing, and low-performing learners.

This study aims to analyze student academic data using the K-Means clustering algorithm to identify performance patterns and classify students into meaningful groups. The results can assist educators and academic administrators in monitoring student progress, identifying at-risk students, and implementing appropriate academic support strategies to enhance overall educational performance.

II. LITERATURE SURVEY

Educational Data Mining (EDM) has emerged as an effective approach for analyzing student academic data and discovering meaningful learning patterns. Merceron and Yacef demonstrated that data mining techniques can identify relationships between student activities and academic performance [1]. Romero and Ventura provided a comprehensive survey of EDM applications and highlighted the importance of clustering techniques in education [2]. Zaiane applied data mining methods to educational environments and showed their usefulness in understanding student behavior [3]. Kotsiantis et al. discussed various data mining approaches for educational systems and emphasized their role in performance analysis [4]. Pandey and Pal employed K-Means clustering to classify students based on academic achievement and obtained meaningful performance groups [5]. Baradwaj and Pal analyzed student records using data mining methods and identified factors affecting academic success [6]. Ramesh et al. utilized clustering algorithms to evaluate student performance and support academic decision-making [7]. Kabakchieva investigated predictors of student success and demonstrated the effectiveness of data mining techniques in educational datasets [8]. Hussain et al. applied K-Means clustering to categorize students according to learning outcomes and performance levels [9]. Asif et al. developed a framework for student performance analysis using educational data mining techniques [10]. Kumar and Vijayalakshmi used clustering methods to group students based on academic indicators and learning behavior [11]. Al-Radaideh et al. explored student data using classification and clustering approaches to improve educational outcomes [12]. Han and Kamber highlighted the significance of clustering algorithms, particularly K-Means, for discovering patterns in large datasets [13]. Romero et al. applied data mining techniques in e-learning systems and identified learner profiles through clustering [14]. Thakar et al. demonstrated the effectiveness of K-Means clustering in categorizing students for personalized academic support and performance improvement [15]. The reviewed literature indicates that clustering techniques, especially K-Means, are widely used for student performance analysis and provide valuable insights for improving academic outcomes.

III. METHODOLOGY

The objective of this study is to analyze student performance and identify groups of students with similar academic characteristics using the K-Means clustering algorithm. The proposed methodology consists of six major phases: data collection, data preprocessing, feature selection, clustering, cluster analysis, and result interpretation. The overall methodology is designed to discover hidden patterns in student academic data and support educational decision-making.

Figure 1. Methodological Framework for Student Performance Clustering and Evaluation illustrates the workflow of the study, including data preprocessing, feature selection, K-Means clustering, cluster formation, and performance evaluation to analyze student academic performance.

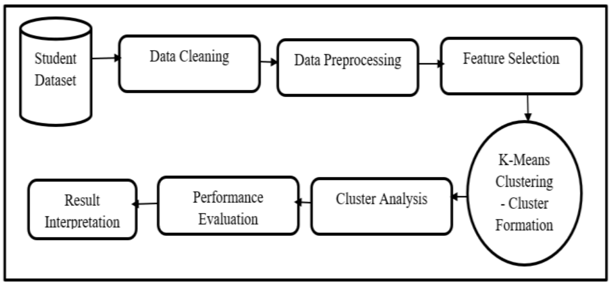

Fig. 1: Methodological Framework for Student Performance Clustering and Evaluation

A. Data Collection

The dataset used in this study consists of student academic records collected from educational institutions. The collected data contain various performance-related attributes that influence academic achievement. The primary attributes considered for analysis include attendance percentage, assignment completion percentage, assessment score percentage, and overall academic performance. These attributes provide a comprehensive representation of student learning behavior and academic progress.

B. Data Preprocessing

Data preprocessing is an essential step to improve the quality of the dataset before applying the clustering algorithm. Raw educational data may contain missing values, duplicate records, inconsistent entries, and irrelevant information that can affect clustering accuracy. The preprocessing phase includes the following activities:

1) Removal of duplicate student records.
2) Identification and handling of missing values.
3) Correction of inconsistent data entries.
4) Conversion of categorical values into numerical form when necessary.
5) Normalization of attribute values to maintain uniformity among different data scales.

After preprocessing, the dataset becomes suitable for clustering analysis and produces more reliable results.

C. Feature Selection

Feature selection is performed to identify the most relevant attributes influencing student performance. Selecting appropriate features improves clustering efficiency and reduces computational complexity.

The selected attributes are:

– Attendance Percentage
– Assignment Completion Percentage
– Assessment Score Percentage

These attributes are considered significant indicators of academic performance and are used as input variables for the K-Means clustering algorithm.

D. K-Means Clustering Algorithm

K-Means is an unsupervised machine learning algorithm used to partition data into K distinct clusters. The algorithm groups student records based on similarity and minimizes the distance between data points and their respective cluster centroids.

The K-Means clustering process involves the following steps:

1) Select the number of clusters (K = 3) representing high, average, and low-performing students.
2) Initialize cluster centroids randomly.
3) Calculate the Euclidean distance between student records and centroids.
4) Assign each student to the nearest cluster.
5) Update cluster centroids using the mean of assigned records.
6) Repeat the process until the centroids stabilize.

E. Cluster Formation

After the execution of the K-Means algorithm, students are grouped into three clusters based on their academic characteristics.

1) Cluster 1 – High Performance

Students with high attendance, high assignment completion rates, and excellent assessment scores.

2) Cluster 2 – Average Performance

Students with moderate attendance, average assignment completion, and satisfactory assessment scores.

3) Cluster 3 – Low Performance

Students with low attendance, poor assignment completion, and lower assessment scores.

The clustering process enables the identification of students belonging to each performance category.

F. Cluster Analysis

The generated clusters are analyzed to determine the characteristics of each student group. Statistical measures such as average attendance, average assignment completion, and average assessment scores are calculated for each cluster.

The analysis helps educators:

– Identify academically strong students.
– Detect students at risk of poor performance.
– Understand learning behavior patterns.
– Design targeted academic interventions.

G. Result Interpretation

The final clusters are interpreted to provide meaningful educational insights. The clustering results assist academic administrators and faculty members in making data-driven decisions for improving student outcomes. By identifying performance-based student groups, institutions can implement personalized support strategies, enhance learning effectiveness, and improve overall academic success.

IV. EXPERIMENTAL RESULT

The proposed K-Means clustering model was implemented to analyze student performance based on attendance percentage, assignment completion percentage, and assessment scores.

After preprocessing the dataset, the K-Means algorithm was applied with K = 3 to classify students into three distinct performance groups. The clustering process successfully grouped students according to their academic characteristics and learning behavior.

A. Clustered Student Performance Data

The clustering results generated three performance categories: High Performance, Average Performance, and Low Performance. The average values of the attributes in each cluster are presented in Table 1.

Cluster	Attendance (%)	Assignment Completion (%)	Assessment Score (%)	Performance Level
C1	90	92	88	High
C2	81	79	76	Average
C3	72	68	65	Low

Table 1. Performance Analysis Based on Student Clusters

The results indicate that students in Cluster 1 achieved the highest attendance, assignment completion, and assessment scores. Cluster 2 consisted of students with moderate academic performance, while Cluster 3 represented students requiring additional academic support.

Figure 2. Performance Level Analysis Using K-Means Clustering shows the grouping of students into different performance levels based on academic factors such as attendance, assignments, and assessment scores. The clusters help identify low-, medium-, and high-performing student groups for better academic analysis and support.

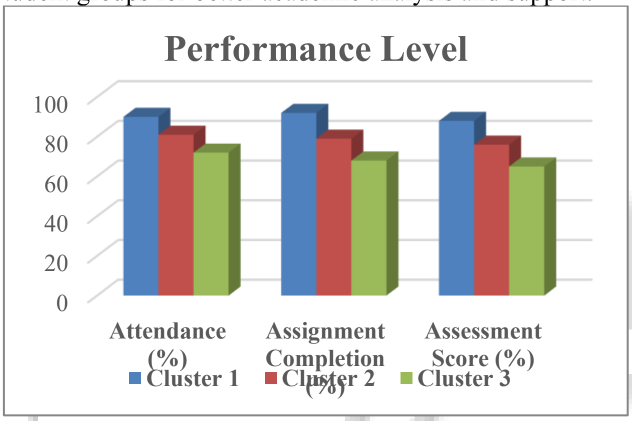

Fig. 2: Performance Level Analysis Using K-Means Clustering

B. Cluster Analysis

The clustering results reveal a strong relationship between attendance, assignment completion, and academic achievement. Students with higher attendance and consistent assignment submission obtained better assessment scores. Conversely, students with lower participation levels demonstrated weaker academic performance.

The K-Means algorithm effectively identified performance-based student groups and provided meaningful insights into student learning patterns. The identified clusters can assist educators in monitoring academic progress and implementing targeted interventions for students at risk of poor performance.

C. Discussion

The experimental results demonstrate that K-Means clustering is an effective technique for educational data analysis. The algorithm successfully categorized students into homogeneous groups based on academic indicators. The clustering outcome can support academic decision-making, personalized learning strategies, and student performance improvement initiatives.

The study confirms that attendance, assignment completion, and assessment scores are significant factors influencing student performance. The generated clusters provide valuable information for educators to enhance teaching effectiveness and improve student learning outcomes.

V. CONCLUSION

This study applied the K-Means clustering algorithm to analyze student performance data and identify groups of students with similar academic characteristics. Factors such as attendance, assignment completion, and assessment scores were used to form meaningful clusters representing high-performing, average-performing, and low-performing students. The clustering results demonstrated that K-Means is an effective technique for discovering hidden patterns in educational data and classifying students based on their performance levels. The identified clusters can assist educators in monitoring student progress, providing targeted academic support, and implementing timely interventions for students who require additional attention. In the future, the study can be extended by incorporating additional attributes and comparing the performance of different clustering algorithms to achieve more accurate and comprehensive student performance analysis.

REFERENCES

[1] O. R. Zaiane, “Building a Recommender Agent for e-Learning Systems,” Proceedings of the International Conference on Computers in Education, pp. 55–59, 2002.
[2] C. Romero and S. Ventura, “Educational Data Mining: A Survey from 1995 to 2005,” Expert Systems with Applications, vol. 33, no. 1, pp. 135–146, 2007.
[3] O. R. Zaïane, “Web Usage Mining for a Better Web-Based Learning Environment,” Proceedings of Conference on Advanced Technology for Education, pp. 60–64, 2002.
[4] S. B. Kotsiantis, C. J. Pierrakeas, and P. E. Pintelas, “Predicting Students’ Performance in Distance Learning Using Machine Learning Techniques,” Applied Artificial Intelligence, vol. 18, no. 5, pp. 411–426, 2004.
[5] U. K. Pandey and S. Pal, “Data Mining: A Prediction for Performance Improvement of Engineering Students Using Classification,” World of Computer Science and Information Technology Journal, vol. 1, no. 2, pp. 51–56, 2011.
[6] B. K. Baradwaj and S. Pal, “Mining Educational Data to Analyze Students’ Performance,” International Journal

of Advanced Computer Science and Applications, vol. 2, no. 6, pp. 63–69, 2012.
[7] V. Ramesh, P. Parkavi, and K. Ramar, "Predicting Student Performance: A Statistical and Data Mining Approach," International Journal of Computer Applications, vol. 63, no. 8, pp. 35–39, 2013.
[8] D. Kabakchieva, "Student Performance Prediction by Using Data Mining Classification Algorithms," International Journal of Computer Science and Management Research, vol. 1, no. 4, pp. 686–690, 2013.
[9] S. Hussain, A. Dahan, F. Ba-Alwi, and N. Ribata, "Educational Data Mining and Analysis of Students' Academic Performance Using WEKA," Indonesian Journal of Electrical Engineering and Computer Science, vol. 9, no. 2, pp. 447–459, 2018.
[10] R. Asif, A. Merceron, S. A. Ali, and N. G. Haider, "Analyzing Undergraduate Students' Performance Using Educational Data Mining," Computers & Education, vol. 113, pp. 177–194, 2017.
[11] V. Kumar and B. Vijayalakshmi, "Data Mining in Higher Education for Student Classification," International Journal of Computer Science Issues, vol. 8, no. 5, pp. 185–190, 2011.
[12] Q. A. Al-Radaideh, E. M. Al-Shawakfa, and M. I. Al-Najjar, "Mining Student Data Using Decision Trees," International Arab Conference on Information Technology, pp. 1–6, 2006.
[13] J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques, 3rd ed., Morgan Kaufmann Publishers, 2012.
[14] C. Romero, S. Ventura, M. Pechenizkiy, and R. S. J. D. Baker, Handbook of Educational Data Mining, CRC Press, 2010.
[15] P. Thakar, A. Mehta, and M. Manisha, "Performance Analysis and Prediction in Educational Data Mining: A Research Travelogue," International Journal of Computer Applications, vol. 110, no. 15, pp. 60–68, 2015.