Evaluating Students' Academic Proficiency using Unsupervised Learning approach for Performance Prediction

A. Sankar Gomathi

Assistant Professor, Department of Commerce (CA), Sri Kaliswari College (A), Sivakasi

Open Access

Manuscript ID:
BIJ-SPL4-Jan26-MD-057

Subject: Commerce (CA)

Received : 30.08.2025
Accepted : 22.01.2026
Published : 31.01.2026

DOI: 10.64938/bijsi.v10si4.26.Jan057

Abstract

This research reports the use of unsupervised learning approach, specifically nearest centroid and connectivity clustering, with Principal Component Analysis (PCA) to analyze student academic and behavioral datasets. By revealing hidden patterns without labelled outcomes, this research illustrates the potential combinations of clustering and labelling to identify at-risk learners, moderate learners and those with higher levels of success. The results suggest that unsupervised learning serves as a valuable complement to supervised learning within educational applications. By leveraging data mining strategies, these methods enable timely interventions and promote personalized learning pathways.

Keywords: educational data mining, unsupervised learning, K-means, hierarchical clustering, PCA, student performance

Introduction

Recent advancements in data science and machine learning (ML) have significantly transformed the landscape of educational research, creating both new opportunities and methodological challenges for traditional pedagogical approaches. In contemporary academic environments, educational institutions are progressively shifting toward data-driven paradigms to better monitor student engagement, evaluate learning outcomes, and implement timely remedial strategies that foster improved academic performance. Within this evolving research framework, Artificial Intelligence (AI) has gained substantial traction; however, ML has emerged as a more specialized and adaptive subset of AI, wherein algorithms are designed to learn autonomously from historical data, identify latent patterns, and generate reliable predictions and insights without requiring explicit rule-based programming. This capacity to derive knowledge from vast, complex datasets has positioned ML as a crucial instrument in advancing evidence-based decision-making within the educational domain.

Among the diverse applications of ML, student performance analytics remains one of the most extensively explored areas in educational data mining. Traditionally, supervised learning models such as Support Vector Machines (SVM), Decision Trees (DT), and Neural Networks (NN) have been employed to forecast student performance, classify learners, or identify individuals at academic risk. Despite their demonstrated predictive efficiency, these models inherently depend on the availability of well-labeled datasets in which each data instance corresponds to a known output or target value. In real-world educational contexts, however, the collection of such labeled data often proves to be challenging due to inconsistencies in assessment methods, incomplete student records, and subjectivity in grading practices. These constraints introduce bias and data imbalance, thereby impeding the scalability, transferability, and generalizability of supervised models across diverse learning environments and institutional frameworks.

In contrast, unsupervised learning offers an attractive and robust avenue for analyzing educational datasets that lack explicit labels. This approach facilitates the discovery of latent structures, hidden relationships, and natural groupings among students based on similarities in their behavioral, cognitive, and performance-related attributes. By employing clustering algorithms, unsupervised learning enables the segmentation of student populations into homogeneous subgroups, each representing distinct learning patterns, engagement styles, and academic tendencies. Such segmentation supports the development of personalized learning strategies, allowing educators and administrators to tailor instruction, monitor progress dynamically, and design targeted interventions that address specific learner needs. Furthermore, incorporating dimensionality-reduction techniques such as Principal Component Analysis (PCA) enhances clustering performance by filtering out noise, minimizing redundancy, reducing computational complexity, and improving the interpretability of high-dimensional educational data.

The present research builds upon these developments by utilizing unsupervised ML techniques, specifically K-Means clustering and hierarchical clustering, in conjunction with PCA, to analyze comprehensive student datasets containing both educational and socio-behavioral indicators. These datasets encompass variables such as attendance frequency, assessment performance, participation levels, and engagement metrics, collectively representing the multifaceted nature of student learning behavior. The primary objective of this research is to classify students into distinct and interpretable profiles that reflect varying degrees of academic performance, learning style, and engagement intensity. By uncovering these hidden patterns, the study aims to provide valuable insights that can guide educational institutions toward data-informed decision-making, the design of adaptive learning frameworks, and the implementation of proactive interventions to enhance student success and overall learning experiences. Ultimately, this research contributes to the growing body of knowledge in educational data mining by demonstrating how unsupervised ML techniques can complement traditional supervised methods, thus offering a more comprehensive and flexible analytical foundation for modern educational systems.

The primary achievements of this research include the following:

1. The implementation of combined PCA and clustering approaches to extract significant arrangements from educational data without prior labelling.
2. Comparative evaluation of K-Means and hierarchical clustering in segmenting students based on performance and behavioral features.
3. Generation of actionable insights for educators to identify at-risk students and implement targeted pedagogical involvements.

By addressing the limitations of supervised ML in academic performance prediction, this research demonstrates how unsupervised approaches can complement existing educational analytics frameworks and contribute to the advancement of student success strategies in data-rich learning environments.

Literature Review

Supervised Learning in Academic Performance Prediction

A huge body of literature is focuses on academic performance prediction has examined supervised ML approaches. There are number of typical methods are DT, Random Forests (RF), SVM, and Artificial NN. For example, Al-Barrak and Al-Razgan used DT models to predict final exam results. They also found high accuracy, but they needed all tagged datasets, which may not be possible for all datasets. Adekoya and Olatunbosun noted that SVMs are predictive of student success in blended learning environments. Again, these techniques depend on labelled datasets,

Which may not be consistently available across datasets. In consequence the supervised models are hindered with regard to cross-institutional contexts.

Unsupervised Learning in Educational Contexts There is an increasing interest in unsupervised learning, particularly clustering, to provide alternate solutions to the limitations of supervised methods. Clustering algorithms do not require predefined labels, but instead, group students based on the similarities of their features to highlight hidden learning behavior and performance patterns. K-Means clustering, one of the highly used clustering methods, has been developed to identify students based on attendance and assessment patterns, which can support educators to identify high- and low-performing clusters. Agglomerative clustering has also employed to explore multi-level group relationships, particularly in studies aiming to categories learners into behavioral archetypes.

Fuzzy C-Means (FCM) clustering has gained attention for handling cases where students exhibit overlapping characteristics, as demonstrated by Li et al., who used FCM to evaluate examination performance and engagement in online courses. Earlier studies have implemented hybrid approaches that integrate clustering with association rule mining to uncover interpretable decision rules for target predictions.

Dimensionality Reduction in Educational Data Analysis

Educational datasets often contain a high number of structures, around of which can be irrelevant or redundant, potentially degrading clustering efficiency. Dimensional reduction methods, such as PCA, address this challenge by projecting data into a reduced feature space while holding the maximum variance. Several studies have combined PCA with clustering to improve efficiency and visual interpretability. For example, Asif et al. applied PCA prior to K-Means clustering for student grouping, achieving better cluster separation and reduced computational overhead. Similarly, Chiu et al. demonstrated that PCA could enhance clustering accuracy by minimizing noise from irrelevant features.

Gaps and Research Opportunities While Earlier research indicated the utility of unsupervised methods in education, several gaps remain. Many existing studies rely on a limited set of features, often focusing solely on grades or attendance, neglecting other important behavioral indicators. Furthermore, comparative evaluations between clustering algorithms in educational datasets are limited, with few studies incorporating multiple methods alongside dimensionality reduction. Finally, the potential of unsupervised learning for early academic risk detection is underexplored, particularly in the context of combining engagement, behavioral, and performance metrics.

This research aims to bridge the existing gaps through applying both K-Means and hierarchical clustering, in conjunction with PCA, to a comprehensive set of student performance and behavioral features. The aim is to generate actionable insights for personalized learning strategies and proactive academic support.

Methodology

The research employed a structured process to analyze student educational and personality traits using unsupervised ML methods. The methodology consisted of data collection, pre-processing, dimensionality reduction, clustering, validation, and interpretation.

Data Acquisition

Data was Collected from institutional records, comprising:

- Attendance percentage per semester
- Assessment scores from internal and external examinations
- Participation metrics in classroom and online activities
- Behavioral indicators such as punctuality, discipline records, and involvement in extracurricular activities

Data Pre-processing

The raw dataset underwent several pre-processing steps:

- Handling missing values: Missing numerical values were replaced using mean imputation, while categorical characteristics were handled using mode imputation to preserve data consistency.
- Normalization: Features were standardized to zero mean and unit variance to ensure equal weight in clustering.
- Feature encoding: Behavioral indicators were numerically encoded to facilitate ML processing.

Feature Reduction with Principle Component Analysis

PCA was applied to reduce dimensionality, retaining components that explained at least 90% of the variance. PCA transformed correlated features (e.g., attendance and participation) into orthogonal principal components, thereby reducing redundancy.

Clustering Approaches

Two clustering algorithms were implemented:

- K-Means Clustering: Partitioned students into k clusters by minimizing intra-cluster distances. The optimal k was determined using the Elbow Method and Silhouette Score.
- Hierarchical Clustering: Constructed a dendrogram to explore nested groupings of students. Agglomerative clustering with Ward’s linkage was applied.

Cluster Validation

Clustering performance was evaluated using:

- Silhouette Coefficient – to measure separation between clusters.
- Davies-Bouldin Index (DBI) – to assess cluster compactness.
- Visualization – PCA scatterplots for interpretability.

Results and Discussion

PCA Findings

PCA reduced the feature space from 12 original variables to 4 principal components explaining 91% of variance. The first two components corresponded primarily to academic performance (exam scores, assignments) and engagement behaviors (attendance, participation).

K-Means Clustering Results

K-Means yielded an optimal clustering solution with three student groups:

1. High Performers & Engaged Learners – strong academic scores, high attendance, active participation.
2. Moderate Performers with Inconsistent Engagement – average academic results, fluctuating attendance.
   3. At-Risk Learners – low scores, poor attendance, limited engagement.

Silhouette score = 0.57 indicated reasonably distinct clusters.

Agglomerative Clustering Results

Agglomerative clustering produced a similar three-group structure, confirming K-Means results. However, hierarchical methods revealed sub-grouping within moderate performers (e.g., “hard-working but irregular” vs. “consistent but average”).

Implications for Educators

The findings demonstrate that unsupervised clustering can detect unknown groupings in data beyond raw grades. For instance:

- At-risk learners can be identified early through their low engagement profile.
- Moderate performers may require differentiated interventions depending on whether their weakness stems from attendance or exam preparation.
  - High performers' study patterns can inform peer-mentoring programs.

Conclusion

This study applied unsupervised ML approaches—K-Means and hierarchical clustering combined with PCA—to learning datasets to discover outlines in academic performance and engagement. Results demonstrated the ability of clustering to segment students into meaningful subgroups without requiring labelled outcomes.

The study highlights three key contributions:

1. Demonstrated the effectiveness of PCA-enhanced clustering for educational analytics.
2. Compared K-Means and hierarchical methods, showing complementary strengths.
3. Provided actionable insights for early intervention and personalized learning strategies.

Upcoming studies may expand on this approach by integrating temporal data (e.g., performance trends over semesters) and exploring advanced clustering techniques such as DBSCAN or deep clustering models. Incorporating explainable AI frameworks can further improve interpretability for educators.

References

1. Romero, C., & Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 10(3).
2. Siemens, G., & Baker, R. (2013). Learning analytics and educational data mining: Towards communication and collaboration. Proceedings of LAK.
3. Cortez, P., & Silva, A. (2008). Using data mining to predict secondary school student performance. European Symposium on Artificial Neural Networks. Siemens, G., & Baker, R. (2013). Learning analytics and educational data mining: Towards communication and collaboration. Proceedings of LAK.
4. Kaminˊski B, Jakubczyk M, and Szufel P. A framework for sensitivity analysis of decision trees. Cent Eur J Operat Res 2018; 26(1): 135–159.
5. Wei Y, Zhang X, Shi Y, et al. A review of data-driven approaches for prediction and classification of building energy consumption. Renew Sust Energ Rev 2018; 82(1): 1027–1047.
6. Namasudra S and Roy P. PpBAC: popularity based access control model for cloud computing. J Organizat End User Comput 2018; 30(4): 14–31. 22. RapidMiner Studio, https://rapidminer. com/products/studio/ (2019, accessed 24 Septe mber 2019).
7. Grubljeˇsiˇc T, Coelho PS, and Jakliˇc J. The shift to socioorganizational drivers of business intelligence and analytics acceptance. J Organizat End User Comput 2019; 31(2): 37–62.
8. 8.Lamichhane, C.D.: 'Understanding the education philosophy and its implications', NCC J., 2018, 3, (1), pp. 24–29.
9. Stapa, M.A.; Mohammad, N. The use of Addie model for designing blended learning application at vocational colleges in Malaysia. Asia-Pac. J. Inf. Technol. Multimed. 2019, 8, 49–62.
10. Mohamed, H.; Judi, H.M.; Jenal, R. Soft Skills Assessment Based On Undergraduate Student Perception. Asia-Pac. J. Inf. Technol. Multimed. 2019, 8, 27–35.
11. Nahar, K.; Shova, B.I.; Ria, T.; Rashid, H.B.; Islam, A.S. Mining educational data to predict students performance. Educ. Inf. Technol. 2021, 26, 6051–6067.
12. Yang, S.J.; Lu, O.H.; Huang, A.Y.; Huang, J.C.; Ogata, H.; Lin, A.J. Predicting students' academic performance using multiple linear regression and principal component analysis. J. Inf. Process. 2018, 26, 170–176.
13. Romero, C., & Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 10(3).

14. Seth Adjei, Korinn Ostrow, Erik Erickson, and Neil T Heffernan. Clustering students in assistments: Exploring system-and school-level traits to advance personalization. In The 10th International Conference on Educational Data Mining, pages 340–341, 2017
15. Lau, E.; Sun, L.; Yang, Q. Modelling, prediction and classification of student academic performance using artificial neural networks. SN Appl. Sci. 2019, 1, 982.