International Journal of Computer Science Trends and Technology (IJCST) – Volume 14 Issue 4, Jul-Aug 2026

RESEARCH ARTICLE OPEN ACCESS

Educational Data Mining for Student Profiling of NEET-UG Aspirants: A Comparative Analysis of Principal Component Analysis and Clustering Techniques Using Synthetic Educational Data

Jayshri Patel
Department of Computer Science, Veer Narmad South Gujarat University, Surat, India

ABSTRACT

Educational Data Mining (EDM) has emerged as an important research area for analyzing educational data and supporting data-driven decision-making. This study proposes a reproducible EDM framework for computational learner profiling using a synthetic dataset of 900 simulated NEET-UG aspirants described by 13 educational attributes. The proposed methodology integrates data preprocessing, feature standardization, Principal Component Analysis (PCA) for dimensionality reduction, and three unsupervised clustering algorithms: K-Means, Agglomerative Hierarchical Clustering, and DBSCAN. Since the dataset contains no predefined class labels, clustering performance was evaluated using the Silhouette Score, Davies–Bouldin Index, and Calinski–Harabasz Index. Experimental results indicate that K-Means achieved the most suitable overall clustering solution by assigning all learner records to five clusters while obtaining the highest Calinski–Harabasz Index. Although DBSCAN demonstrated superior cluster separation according to the Silhouette Score and Davies–Bouldin Index, its large number of noise points reduced its applicability for complete learner profiling. The study demonstrates that integrating PCA with clustering techniques provides a systematic and reproducible approach for organizing multidimensional educational data into computational learner profiles. As the analysis is based entirely on synthetic data, the identified profiles represent methodological examples rather than validated categories of actual NEET-UG aspirants. Future work will focus on validating the proposed framework using authentic educational datasets.

Keywords — Educational Data Mining, Student Profiling, Principal Component Analysis, K-Means Clustering, Agglomerative Hierarchical Clustering, DBSCAN, Synthetic Educational Data, NEET-UG Aspirants.

I. INTRODUCTION

Educational institutions generate vast amounts of information through teaching, learning, assessments, attendance, and student participation. Converting these data into meaningful insights has become one of the primary goals of Educational Data Mining (EDM), which combines data mining, machine learning, and statistical techniques to support educational research and improve decision-making [1]. Over the years, EDM has expanded beyond analysing academic performance to understanding learning behaviour, identifying learner characteristics, and supporting personalized education through data-driven approaches [2]. More recently, the integration of Educational Data Mining with Learning Analytics has further strengthened the use of computational techniques for improving educational outcomes [3].

Among the various machine learning approaches used in EDM, unsupervised clustering has received considerable attention because it can discover hidden patterns in educational data without requiring predefined class labels. By grouping learners with similar characteristics, clustering helps researchers understand the diversity of learner populations and provides valuable information for educational planning and intervention [4]. Several clustering algorithms have been successfully applied in educational research, including K-Means [6], Agglomerative Hierarchical Clustering [7], and DBSCAN [8], each employing a different strategy to identify groups within multidimensional datasets. Density-based clustering methods such as DBSCAN are particularly effective in identifying irregular cluster structures and noisy observations [8], whereas hierarchical clustering provides a tree-based representation of relationships among observations [7].

Most existing studies in Educational Data Mining focus on predicting student performance, identifying at-risk learners, or analysing learning behaviour using institutional educational datasets [18]. Recent research has also explored learner profiling and personalized learning through clustering techniques and learning analytics [19]. However, comparatively fewer studies have systematically compared multiple clustering algorithms for computational learner profiling within a common analytical framework. In addition, access to real educational datasets is often restricted because of privacy, confidentiality, and ethical considerations. These challenges limit the reproducibility of experimental studies and make comparative evaluation more difficult.

To address these limitations, this study proposes a reproducible Educational Data Mining framework for profiling synthetic NEET-UG aspirants. A synthetic educational dataset was employed to provide a controlled experimental environment while avoiding the use of sensitive student information. The proposed framework integrates data preprocessing, feature standardization, Principal Component Analysis (PCA) for dimensionality reduction, and comparative clustering using K-Means, Agglomerative Hierarchical Clustering, and DBSCAN. The clustering solutions are evaluated using the Silhouette Score [11], Davies–Bouldin Index [12], and Calinski–Harabasz Index [13] to determine the most suitable algorithm for computational learner profiling. The resulting learner profiles are intended to demonstrate the effectiveness of the proposed methodology and should be interpreted as computational clusters rather than validated categories of actual NEET-UG aspirants.

II. LITERATURE REVIEW

Educational Data Mining (EDM) has become an important interdisciplinary research area that applies data mining, machine learning, and statistical techniques to educational data for improving teaching, learning, and academic decision-making. Romero and Ventura [1] presented one of the earliest comprehensive surveys of EDM, highlighting its applications in student modelling, performance analysis, and knowledge discovery. Their later review demonstrated the rapid evolution of the field and emphasized the increasing integration of EDM with Learning Analytics to support personalized education and evidence-based decision-making [3]. Peña-Ayala [2] further analysed recent developments in EDM and reported that advances in machine learning have significantly expanded its educational applications.

Clustering is one of the most widely used unsupervised learning techniques in Educational Data Mining because it groups learners with similar characteristics without requiring predefined class labels. Baker and Yacef [4] noted that clustering enables the discovery of hidden learner patterns and supports educational decision-making through exploratory data analysis. More recently, Kalita *et al.* [22] reviewed a decade of Educational Data Mining research and identified learner segmentation as one of the major emerging applications of unsupervised learning techniques.

Among clustering algorithms, K-Means remains one of the most popular methods because of its computational simplicity, scalability, and effectiveness in partitioning multidimensional datasets [6]. Agglomerative Hierarchical Clustering provides an alternative approach by progressively merging observations according to similarity, allowing hierarchical relationships among learner groups to be explored [7]. In contrast, DBSCAN identifies clusters based on data density and is capable of detecting irregular cluster structures while distinguishing sparse observations as noise [8]. Since these algorithms employ different clustering strategies, their performance varies depending on the characteristics of the dataset, making comparative evaluation essential for selecting an appropriate learner profiling approach.

Educational datasets often contain correlated variables describing academic performance, learning behaviour, attendance, and resource utilization. Such correlations may increase computational complexity and reduce clustering efficiency. Principal Component Analysis (PCA) is widely used to transform correlated variables into a smaller set of uncorrelated principal components while preserving most of the original information [9]. Jolliffe and Cadima [10] reported that PCA improves computational efficiency and frequently enhances the performance of distance-based machine learning algorithms by reducing redundancy in high-dimensional data.

Several recent studies have demonstrated the usefulness of clustering for learner profiling and educational analytics. Zhang *et al.* [18] reviewed machine learning techniques for student performance analysis and highlighted the importance of clustering for identifying learner groups with similar characteristics. Yang *et al.* [19] applied Educational Data Mining and Learning Analytics to identify learning patterns from digital educational environments, while Araka *et al.* [20] used clustering techniques to analyse profiles of self-regulated learners. Pecuchova and Drlik [21] further demonstrated that clustering analysis can improve the identification of meaningful student groups for educational analytics and intervention planning.

Although previous studies have confirmed the value of clustering in educational applications, most have focused on learner performance prediction, dropout analysis, or specific institutional datasets. Comparatively fewer studies have performed a systematic comparison of multiple clustering algorithms within a unified Educational Data Mining framework using standardized preprocessing and dimensionality reduction. In addition, the use of real educational datasets is frequently constrained by privacy and ethical considerations, limiting reproducibility and comparative evaluation. To address these limitations, the present study proposes a reproducible Educational Data Mining framework that integrates feature standardization, Principal Component Analysis, and comparative clustering using K-Means, Agglomerative Hierarchical Clustering, and DBSCAN on a synthetic educational dataset. The resulting computational learner profiles provide a methodological demonstration of learner segmentation while avoiding the use of sensitive educational records.

III. RESEARCH GAP

The literature indicates that Educational Data Mining has been widely applied for student performance prediction, learning analytics, and learner behaviour analysis using various machine learning techniques [3]. Several studies have employed clustering algorithms for learner segmentation; however, most investigations focus on a single clustering approach or specific institutional datasets [18]. Comparative evaluation of multiple clustering algorithms within a unified framework that incorporates Principal Component Analysis (PCA) for dimensionality reduction has received comparatively limited attention, particularly for methodological studies using synthetic educational data [22].

Furthermore, privacy and accessibility constraints associated with real educational datasets often limit the reproducibility of clustering experiments. To address these gaps, the present study proposes a reproducible Educational Data Mining framework that integrates PCA with K-Means, Agglomerative Hierarchical Clustering, and DBSCAN to evaluate computational learner profiling using a synthetic dataset of NEET-UG aspirants.

IV. RESEARCH METHODOLOGY

This study adopts a quantitative Educational Data Mining (EDM) methodology to develop a reproducible framework for profiling NEET-UG aspirants using unsupervised machine learning techniques. The proposed framework integrates data preprocessing, feature standardization, Principal Component Analysis (PCA), clustering algorithms, and cluster validation within a unified analytical workflow. Since the objective of this research is methodological evaluation, all experiments were conducted using a synthetic educational dataset. No human participants were involved, no institutional records were collected, and no student surveys were conducted. Consequently, the identified learner profiles represent computational clusters for methodological demonstration rather than validated categories of actual NEET-UG aspirants.

The overall research workflow is illustrated in Fig. 1.

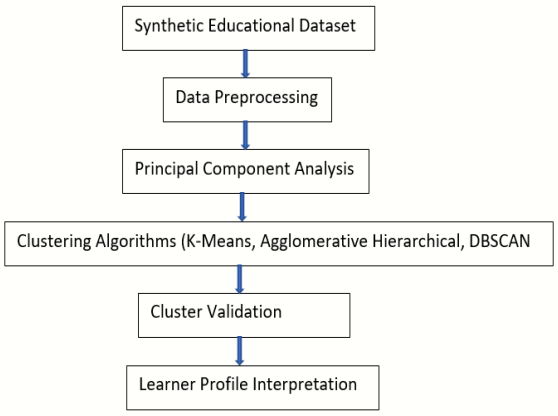

Fig. 1. Proposed Educational Data Mining Framework for Student Profiling

The proposed framework was evaluated using a synthetic educational dataset comprising 900 simulated NEET-UG aspirant records described by 13 educational variables. The dataset was generated exclusively for methodological evaluation and does not contain information from real students, educational institutions, or surveys. The selected variables represent learner characteristics, including academic performance, study behaviour, attendance, learning mode, coaching status, stress level, and resource accessibility. Since the dataset contains no predefined target variable, the study focuses on unsupervised learner profiling through the comparative evaluation of clustering algorithms.

Data preprocessing was performed to ensure data consistency and improve the effectiveness of the clustering algorithms. The preprocessing procedure included dataset inspection, categorical feature encoding, duplicate verification, and feature standardization. Categorical variables were transformed into numerical values using label encoding, while feature scaling was performed using Z-score normalization, where the standardized feature value is calculated as:

$$z = (x - \mu) / \sigma \quad (1)$$

where x represents the original feature value, μ denotes the mean of the feature, and σ represents the standard deviation. The standardized dataset was subsequently used for Principal Component Analysis (PCA) and clustering, ensuring that all features contributed equally to the distance-based learning algorithms.

Principal Component Analysis (PCA) was applied to reduce the dimensionality of the standardized educational dataset by transforming the original correlated variables into a smaller set of uncorrelated principal components. This transformation preserves the major variance of the dataset while reducing feature redundancy, thereby improving the efficiency of the clustering algorithms. The PCA transformation is represented as:

$$Z = X \times W \quad (2)$$

where X is the standardized feature matrix, W is the matrix of principal component vectors (eigenvectors), and Z is the transformed feature matrix. The first five principal components (PC1–PC5) were selected for clustering because they retained 48.49% of the total variance while substantially reducing the dimensionality of the original feature space.

Three widely used unsupervised clustering algorithms were evaluated on the PCA-transformed dataset to compare their effectiveness for computational learner profiling. K-Means partitions observations into clusters by iteratively assigning each record to the nearest cluster centroid. The optimal number of clusters was determined using the Elbow Method, resulting in K = 5. Agglomerative Hierarchical Clustering follows a bottom-up strategy in which similar observations are progressively merged using Ward linkage until five clusters are formed. DBSCAN (Density-Based Spatial Clustering of Applications with Noise) identifies clusters according to data density while distinguishing isolated observations as noise. The algorithm was implemented using $\varepsilon = 0.8$ and MinPts = 8.

Since the synthetic educational dataset does not contain predefined class labels, clustering performance was evaluated using internal validation metrics. The Silhouette Score was used to measure cluster cohesion and separation, where higher values indicate better clustering quality [11]. The Davies–Bouldin Index (DBI) was employed to evaluate cluster similarity, with lower values representing better cluster separation [12]. The Calinski–Harabasz Index (CHI) was used to assess the ratio of between-cluster dispersion to within-cluster dispersion, where higher values indicate more distinct and well-defined clusters [13].

The proposed Educational Data Mining framework was implemented using Python 3.x in the Google Colab environment. Data preprocessing was performed using Pandas [15], numerical computations were carried out using NumPy [16], clustering algorithms and validation metrics were implemented using Scikit-learn [14], and graphical visualizations were generated using Matplotlib [17].

The proposed methodology provides a systematic and reproducible framework by integrating standardized data preprocessing, PCA-based dimensionality reduction, comparative clustering analysis, and objective cluster validation. The effectiveness of the proposed framework is evaluated and discussed in the following experimental results section.

V. EXPERIMENTAL RESULTS

The proposed Educational Data Mining framework was experimentally evaluated using a synthetic dataset comprising 900 simulated NEET-UG aspirant records described by 13 educational attributes. The experimental workflow included data preprocessing, feature standardization, Principal Component Analysis (PCA), and comparative clustering using K-Means, Agglomerative Hierarchical Clustering, and DBSCAN. Since the dataset contained no predefined class labels, clustering quality was assessed using three internal validation metrics: Silhouette Score, Davies–Bouldin Index (DBI), and Calinski–Harabasz Index (CHI). The objective of the evaluation was to compare the clustering algorithms and identify the most suitable approach for computational learner profiling.

The PCA-transformed dataset was used as the input for all clustering algorithms. The optimal number of clusters for K-Means was determined using the Elbow Method, which indicated that five clusters provided an appropriate balance between cluster compactness and model complexity.

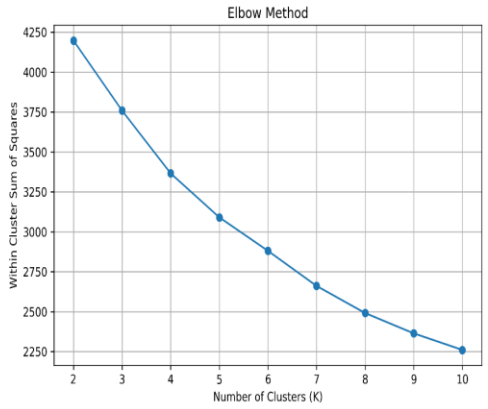

Fig. 2. Elbow Method for Selecting the Optimal Number of Clusters

The elbow curve indicates that K = 5 is an appropriate choice for K-Means clustering.

The clustering performance obtained from the three algorithms is presented in Table I.

TABLE I. PERFORMANCE COMPARISON OF CLUSTERING ALGORITHMS

Algorithm	Clusters	Silhouette Score	Davies–Bouldin Index	Calinski–Harabasz Index
K-Means	5	0.1754	1.7110	186.9647
Agglomerative	5	0.1233	2.0620	153.7071
DBSCAN	7	0.2648	0.9016	57.0188

The comparative evaluation demonstrates that each clustering algorithm exhibits different characteristics. DBSCAN achieved the highest Silhouette Score and the lowest Davies–Bouldin Index, indicating strong separation among dense clusters. However, it classified 772 observations as noise, leaving only a small proportion of learner records assigned to clusters. This behavior limits its applicability for Educational Data Mining tasks that require complete learner segmentation.

Agglomerative Hierarchical Clustering successfully partitioned all learner records into five clusters but produced lower Silhouette and Calinski–Harabasz scores than K-Means, indicating comparatively weaker cluster separation and compactness.

K-Means assigned all 900 learner records to five clusters and achieved the highest Calinski–Harabasz Index (186.9647), indicating a well-balanced clustering structure. Although its Silhouette Score was lower than that of DBSCAN, K-Means provided complete dataset coverage while maintaining good cluster quality, making it the most suitable algorithm for computational learner profiling.

Based on the comparative evaluation, K-Means was selected as the final clustering model. The distribution of learner records across the five clusters is presented in Table II.

TABLE II. DISTRIBUTION OF K-MEANS CLUSTERS

Cluster	Records	Percentage
Cluster 0	199	22.11%
Cluster 1	177	19.67%
Cluster 2	198	22.00%
Cluster 3	173	19.22%
Cluster 4	153	17.00%
Total	900	100%

The cluster distribution indicates that the selected model generated reasonably balanced learner groups without leaving any record unassigned. Such balanced segmentation is desirable for subsequent learner profile interpretation and comparative analysis.

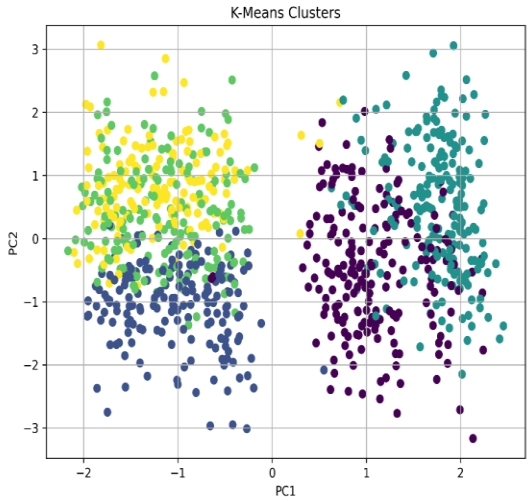

Fig. 3. K-Means Cluster Visualization in PCA Space

The two-dimensional PCA projection illustrates the distribution of the five learner clusters generated by K-Means. The visualization demonstrates clear computational grouping of learner records within the reduced feature space.

Overall, the experimental results demonstrate that the proposed PCA-assisted Educational Data Mining framework effectively organizes multidimensional educational data into meaningful computational learner groups. Among the evaluated algorithms, K-Means provided the most appropriate balance between clustering quality and complete learner assignment, making it the preferred method for subsequent learner profile interpretation. Since the experiments were conducted using a synthetic educational dataset, the identified learner profiles should be interpreted as computational clusters for methodological evaluation rather than validated categories of actual NEET-UG aspirants.

VI. Student Profile Interpretation

Based on the comparative evaluation presented in the previous section, K-Means clustering was selected for learner profile interpretation because it assigned all 900 simulated learner records to five clusters while achieving the highest Calinski–Harabasz Index among the evaluated algorithms. The identified profiles were generated from the PCA-transformed feature space and represent computational groupings of learners with similar educational characteristics. Since the dataset is entirely synthetic, these profiles are intended for methodological evaluation and should not be interpreted as validated categories of actual NEET-UG aspirants.

TABLE III. Summary of Computational Learner Profiles

Cluster	Records	Computational Interpretation
Cluster 0	199	Profile A: Similar multivariate educational characteristics
Cluster 1	177	Profile B: Distinct computational learner pattern
Cluster 2	198	Profile C: Comparable multidimensional feature representation
Cluster 3	173	Profile D: Alternative learner grouping identified through clustering
Cluster 4	153	Profile E: Compact computational learner segment

The five computational learner profiles demonstrate that the proposed PCA-assisted K-Means clustering framework effectively organizes multidimensional educational data into meaningful groups based on similarities in learner characteristics. The relatively balanced distribution of records across the clusters indicates that the selected clustering model provides comprehensive coverage of the synthetic dataset while preserving meaningful separation among learner groups. Although the identified profiles illustrate the capability of Educational Data Mining techniques for computational learner segmentation, they should be regarded as methodological examples because they were derived exclusively from synthetic data. Future research should validate the proposed framework using authentic educational datasets to assess the educational significance and practical applicability of the identified learner profiles.

The learner profiles demonstrate the effectiveness of integrating Principal Component Analysis with K-Means clustering for organizing multidimensional educational data into interpretable computational groups. Because the analysis was performed exclusively on a synthetic dataset, the profiles should be regarded as methodological examples rather than validated descriptions of real NEET-UG aspirants. Future research should validate the proposed framework using authentic educational datasets to assess the educational significance and practical applicability of the identified learner profiles.

VII. Discussion

The proposed Educational Data Mining framework successfully integrated Principal Component Analysis (PCA) with three unsupervised clustering algorithms to investigate computational learner profiling using a synthetic educational dataset. This study emphasizes comparative clustering analysis rather than student performance prediction, reflecting the growing role of clustering in Educational Data Mining research [3].

Among the evaluated algorithms, K-Means provided the most suitable solution by assigning all 900 learner records to five clusters and achieving the highest Calinski–Harabasz Index. In contrast, DBSCAN produced better cluster separation but classified a large number of observations as noise, reducing its practical applicability for complete learner segmentation [8].

The application of PCA reduced feature redundancy and generated a compact feature space, improving the efficiency and consistency of the clustering process. This finding is consistent with previous studies that recommend PCA as an effective preprocessing technique for high-dimensional educational data [10].

Since the experiments were conducted using a synthetic educational dataset, the identified learner profiles represent computational clusters rather than validated categories of actual NEET-UG aspirants. Future research should evaluate the proposed framework using authentic educational datasets and additional clustering techniques to improve learner profiling and educational decision-making.

VIII. CONCLUSION

This study presented an Educational Data Mining (EDM) framework for computational learner profiling using a synthetic dataset of 900 simulated NEET-UG aspirants. The proposed methodology integrated data preprocessing, Principal Component Analysis (PCA), and three unsupervised clustering algorithms to evaluate their effectiveness for learner segmentation.

The experimental results showed that K-Means provided the most appropriate clustering solution by assigning all learner records to five clusters while achieving the highest Calinski–Harabasz Index. Although DBSCAN demonstrated better cluster separation according to the Silhouette Score and Davies–Bouldin Index, its large number of noise points limited its suitability for complete learner profiling. Agglomerative Hierarchical Clustering produced meaningful clusters but yielded comparatively lower validation performance.

The proposed framework demonstrates that integrating PCA with clustering techniques provides a systematic and reproducible approach for organizing multidimensional educational data into computational learner profiles. Since the analysis was performed using a synthetic dataset, the identified profiles should be regarded as methodological examples rather than validated categories of actual NEET-UG aspirants. Future work will focus on validating the proposed framework using authentic educational datasets and exploring additional clustering techniques to enhance learner profiling and educational analytics.

ACKNOWLEDGMENT

The author sincerely thanks the Department of Computer Science, Veer Narmad South Gujarat University (VNSGU), Surat, Gujarat, India, for providing the academic environment and support necessary to carry out this research.

REFERENCES

[1] 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, doi: 10.1016/j.eswa.2006.04.005.

[2] A. Peña-Ayala, "Educational data mining: A survey and a data mining-based analysis of recent works," *Expert Systems with Applications*, vol. 41, no. 4, pt. 1, pp. 1432–1462, 2014, doi: 10.1016/j.eswa.2013.08.042.

[3] C. Romero and S. Ventura, "Educational data mining and learning analytics: An updated survey," *WIREs Data Mining and Knowledge Discovery*, vol. 10, no. 3, Art. no. e1355, 2020, doi: 10.1002/widm.1355.

[4] R. S. Baker and K. Yacef, "The state of educational data mining in 2009: A review and future visions," *Journal of Educational Data Mining*, vol. 1, no. 1, pp. 3–17, 2009.

[5] G. Siemens and R. S. J. d. Baker, "Learning analytics and educational data mining: Towards communication and collaboration," in *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge (LAK)*, Vancouver, BC, Canada, 2012, pp. 252–254, doi: 10.1145/2330601.2330661.

[6] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley, CA, USA, 1967, pp. 281–297.

[7] L. Kaufman and P. J. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis*. New York, NY, USA: John Wiley & Sons, 1990.

[8] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD)*, Portland, OR, USA, 1996, pp. 226–231.

[9] I. T. Jolliffe, *Principal Component Analysis*, 2nd ed. New York, NY, USA: Springer, 2002.

[10] I. T. Jolliffe and J. Cadima, "Principal component analysis: A review and recent developments," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 374, no. 2065, Art. no. 20150202, 2016, doi: 10.1098/rsta.2015.0202.

[11] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987, doi: 10.1016/0377-0427(87)90125-7.

[12] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-1, no. 2, pp. 224–227, 1979, doi: 10.1109/TPAMI.1979.4766909.

[13] T. Caliński and J. Harabasz, "A dendrite method for cluster analysis," *Communications in Statistics*, vol. 3, no. 1, pp. 1–27, 1974, doi: 10.1080/03610927408827101.

[14] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[15] W. McKinney, "Data structures for statistical computing in Python," in *Proceedings of the 9th Python in Science Conference (SciPy 2010)*, Austin, TX, USA, 2010, pp. 56–61.

[16] C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. F. del Río,

M. Wiebe, P. Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, and T. E. Oliphant, "Array programming with NumPy," *Nature*, vol. 585, no. 7825, pp. 357–362, 2020, doi: 10.1038/s41586-020-2649-2.

[17] J. D. Hunter, "Matplotlib: A 2D graphics environment," *Computing in Science & Engineering*, vol. 9, no. 3, pp. 90–95, 2007, doi: 10.1109/MCSE.2007.55.

[18] Y. Zhang, Y. Yun, R. An, J. Cui, H. Dai, and X. Shang, "Educational Data Mining Techniques for Student Performance Prediction: Method Review and Comparison Analysis," *Frontiers in Psychology*, vol. 12, Art. no. 698490, 2021, doi: 10.3389/fpsyg.2021.698490.

[19] C. C. Y. Yang, I. Y. L. Chen, and H. Ogata, "Toward Precision Education: Educational Data Mining and Learning Analytics for Identifying Students' Learning Patterns with Ebook Systems," *Educational Technology & Society*, vol. 24, no. 1, pp. 152–163, 2021.

[20] E. Araka, R. Oboko, E. Maina, and R. Gitonga, "Using Educational Data Mining Techniques to Identify Profiles in Self-Regulated Learning: An Empirical Evaluation," *The International Review of Research in Open and Distributed Learning*, vol. 22, no. 4, 2021, doi: 10.19173/irrodl.v22i4.5401.

[21] J. Pecuchova and M. Drlik, "Enhancing the Early Student Dropout Prediction Model Through Clustering Analysis of Students' Digital Traces," *IEEE Access*, vol. 12, pp. 159336–159367, 2024, doi: 10.1109/ACCESS.2024.3486762.

[22] E. Kalita, S. S. Oyelere, S. Gaftandzhieva, K. N. V. P. S. Rajesh, S. K. Jagatheesaperumal, A. Mohamed, Y. M. Elbarawy, A. S. Desuky, S. Hussain, M. A. Cifci, P. Theodorou, S. Hilčenko, J. Hazarika, and T. Ali, "Educational data mining: A 10-year review," *Discover Computing*, vol. 28, Art. no. 81, 2025, doi: 10.1007/s10791-025-09589-z.

[23] R. S. Baker and P. S. Inventado, "Educational Data Mining and Learning Analytics," in *Learning Analytics: From Research to Practice*, J. A. Larusson and B. White, Eds. New York, NY, USA: Springer, 2014, pp. 61–75, doi: 10.1007/978-1-4614-3305-7_4.

[24] A. Peña-Ayala, Ed., *Learning Analytics: Fundaments, Applications, and Trends*. Cham, Switzerland: Springer, 2017.