Proposed Approach for Analysis and Visualization of Educational Data, Based on the Concept of Big Data

Gabrijela Dimić[1], Boško Bogojević[1], Ljiljana Pecić[1], Ivan Tot[2] and Petar Spalević[3]

[1]Academy of Technical and Art Applied Studies, School of Electrical and Computer Engineering, Vojvode Stepe 283, 11010 Belgrade, Serbia; gdimic@viser.edu.rs; bbosko@viser.edu.rs; ljiljanap@viser.edu.rs

[2]Military Academy, University of Defense, Veljka Lukića Kurjaka 33, 11042 Belgrade, Serbia; ivan.tot@va.mod.gov.rs

[3]Faculty of Technical Sciences, University of Pristina, Knjaza Miloša 7, 38220 Kosovska Mitrovica, Serbia; petar.spalevic@pr.ac.rs

Abstract: This paper introduces a framework for data analytics in a blended learning environment, based on Big Data technologies. The framework integrates heterogeneous educational data sources and enables real-time processing using Apache Spark. The methodology applies Z-score and Min–Max normalization and uses Principal Component Analysis (PCA) for dimensionality reduction. K-means clustering is employed to identify patterns in student behavior. The comparison of normalization methods shows that Min–Max normalization produces more compact clusters than Z-score. The analysis also indicates consistent relationships between students' activity on the Moodle platform and their academic outcomes. The study contributes a Spark-based procedure for descriptive, diagnostic, and cluster analytics. It also offers an empirical evaluation of normalization methods. In addition, it provides visual analytics that support early identification of at-risk students. These tools can help teachers improve course organization.

Keywords: big data; data analytics; normalization; dimensionality reduction; clustering

1 Introduction

The digitalization of education and the rapid development of information and communication technologies have led to the generation of significantly larger and more diverse datasets, than previously encountered in educational practice. Many of these datasets exceed the capabilities of traditional analytical methods.

The term Big Data refers to large, heterogeneous, and dynamic datasets that require advanced analytical techniques and specialized processing tools [1] [2]. The widespread adoption of digital platforms, learning management systems (LMS), and various online learning resources has further increased the availability of student activity data, creating opportunities for more comprehensive analyses of learning processes. Student interactions within digital learning environments produce substantial amounts of data that can offer meaningful insights into learning behavior and support instructional improvement. In such environments, data are generated continuously, vary in complexity, and require scalable analytical systems capable of processing information from multiple heterogeneous sources. Their analysis is further complicated by differences in structure, format, and origin.

The case study presented in this paper was conducted in an Apache Spark environment on a cluster with 64 GB of RAM and eight processing cores. The dataset integrates Moodle log files, Google Sheets records and data from the institution's information system. The contribution of this study lies in the development of a real-time Spark-based analytical process and a visual analytics approach that enables instructors to identify relationships between student activities and learning outcomes, recognize clusters of students with similar behavioral patterns, and detect indicators of potential risk in the learning process. The obtained results may support improvements in course organization and overall instructional effectiveness.

In addition to the technological context, the study is grounded in the fields of Educational Data Mining (EDM) and Learning Analytics (LA), which examine how digital traces of learning behavior can be used to understand and enhance educational processes. Although numerous studies analyze LMS generated data, far fewer integrate heterogeneous data sources into a unified Big Data environment capable of near real-time analytics. This gap highlights the need for scalable analytical frameworks that can process diverse datasets while providing pedagogically meaningful interpretation.

2 Related Work

Numerous studies have examined the use of Big Data technologies, learning analytics, and educational data mining in higher education. Earlier work explored the visualization of large datasets and tools for processing high-volume data streams [3]. Learning analytics has been widely applied to LMS data for instructional improvement [4], for examining the pedagogical value of built-in analytic tools [5], for visualizing multimodal feedback [6], and for managing large educational datasets [7]. Log-based analyses have addressed student behavior [8], learning patterns [9], and online testing environments [10].

Research also points to the need for scalable learning analytics architectures. Several studies propose frameworks for data integration and processing in higher education [11] [12]. Additional reviews outline key educational data mining techniques used in blended and online learning [13]. More recent work combines LMS logs with institutional and behavioral data to build predictive and diagnostic models [14] [15].

Advanced visualization methods support the interpretation of complex and heterogeneous educational data. Current approaches include multimodal dashboards, network representations of student interactions, and interactive visual-analytics systems that enable real-time decision-making [16] [17].

This study builds on earlier contributions by integrating Big Data methods with established educational theories. It proposes a framework for identifying and visualizing behavioral patterns in blended learning. The framework integrates heterogeneous data sources, descriptive and diagnostic analytics, dimensionality reduction, clustering, and visualization techniques to support interpretable representations of learner behavior.

Ethical considerations remain essential, particularly with respect to privacy, fairness, and responsible interpretation in Big Data based educational systems.

3 Background

3.1 Learning Theories

Interpreting educational data benefits from grounding in established learning theories. Self-regulated learning (SRL) describes learning as a process in which students set goals and regulate their strategies [18]. Higher levels of self-regulation are often reflected in LMS data through consistent access and timely participation. Cognitive load theory offers further insight. It assumes limited working memory and stresses the need to avoid excessive cognitive demands [19]. Large content volumes or multiple simultaneous tasks may overload learners, a pattern that can be observed in how students navigate course materials. Student engagement models add a complementary perspective. Behavioral engagement visible through assignments, quizzes, and forum activity is directly traceable in LMS data and is linked to improved academic outcomes [20] [21].

These frameworks support clearer interpretation of identified patterns. Students who rarely access supplementary materials may show lower self-regulation or engagement, while frequent and purposeful use of resources often aligns with effective learning strategies. Together, these theories provide a coherent lens for explaining how data patterns relate to underlying learning behaviors [18-20].

3.2 Big Data Analytical Process (BDA)

The BDA process forms an analytical framework used to extract relevant information, identify patterns, and reveal relationships in large and heterogeneous datasets [2] [22]. Such datasets require scalable and efficient processing methods due to their volume, diversity, and complexity.

As described in [2] [22] this framework is typically divided into three phases. The first phase, data collection, is demanding but fundamental. Data are gathered from multiple sources, and each source must be assessed for reliability and relevance. The second phase, data preprocessing, includes operations that detect irregularities, clean and transform raw data, and prepare a consistent dataset suitable for further analysis. High quality preprocessing improves the accuracy and interpretability of results. The third phase, data analysis, identifies patterns, extracts insights, and supports the formulation of conclusions.

Different analytical approaches may be used depending on the study objectives. Descriptive analysis summarizes historical data through statistical measures and visualizations. Predictive analysis applies statistical models and machine learning to estimate future outcomes. Diagnostic analysis investigates causes behind observed events. Prescriptive analysis extends predictive insights by proposing potential actions through modeling and optimization.

Given the size and heterogeneity of contemporary educational datasets, implementing such an analytical framework remains challenging. Each phase plays a critical role in ensuring that large volumes of data can be transformed into reliable and meaningful insights.

3.3 Cluster Analysis

Cluster analysis is an unsupervised learning technique used to group data instances based on similarity [23]. Instances within the same cluster share more characteristics with each other than with instances in other clusters. Among clustering methods, K-means is widely applied due to its simplicity and efficiency on large datasets [23] [24]. Selecting an appropriate number of clusters is an important step. The elbow method [25] is a common heuristic, allowing the examination of how distortion decreases as k increases. The point at which this decrease slows noticeably is typically taken as a suitable value for k. Cluster quality is often assessed using the silhouette coefficient, which measures cohesion within clusters and separation between clusters. Silhouette values range from −1 to +1, with higher values indicating more compact and well-separated clusters. The overall silhouette score represents the average across all instances [26].

3.4 Z-score and Min-Max Normalization Methods

The Z-score method [27] [28] is a data normalization strategy that avoids issues with outliers but does not produce normalized data with exactly the same scale. This method is commonly used in statistics and data analysis to facilitate comparison and interpretation of different data distributions. If a value is exactly equal to the mean of all values for a given attribute, it will be normalized to zero. If it is below the mean, it will be a negative number, and if it is above the mean, it will be a positive number. The magnitude of the negative or positive value is determined by the standard deviation of the original attribute. For unnormalized data with a large standard deviation, the normalized values will be closer to zero.

The Min-Max method [27] [28] is one of the most common ways to normalize data. The procedure is based on scaling the data values to a specified range (often 0 to 1). For each attribute, the minimum value is transformed to 0, the maximum value is transformed to 1, and every other value is transformed into a decimal number between 0 and 1. The Min-Max method ensures that all attributes share the same scale but does not handle extreme values well.

3.5 Principal Components Analysis (PCA)

The PCA method [23] [29] is widely used to reduce the complexity and dimensionality of large datasets. It relies on the eigenvalues and eigenvectors of the covariance matrix to project data into a new coordinate system. Eigenvalues indicate the amount of variance captured in a particular direction, while eigenvectors define those directions. In this way, PCA retains most of the information while representing the dataset in fewer dimensions. Selecting only the components that account for the largest share of variability simplifies analysis and enables visualization in two or three dimensions. As a linear and deterministic method, PCA can process very large datasets without distorting their underlying structure [30]. Jolliffe and Cadima [31] note that PCA reduces dimensionality with minimal information loss, leading to clearer and more interpretable data representations. They also emphasize its adaptability, particularly when working with large and complex datasets. Chang [32] argues that PCA is fast and reliable because it preserves global variance and produces interpretable low-dimensional projections, making it well suited for large datasets. Jeon et al. [33] further highlight that PCA preserves global distances between points and treats them as continuous values, which facilitates comparison.

In contrast, nonlinear methods such as t-SNE and UMAP often distort global relationships and are highly sensitive to hyperparameters. These characteristics make them less suitable for high-dimensional educational data. t-SNE is used mainly for visualization and does not provide stable global distances, while UMAP offers a better balance of local and global structure but still depends strongly on parameter settings.

Given these characteristics, PCA was selected in this study as the method for dimensionality reduction. It reliably preserves variance structure and provides linear components that clarify relationships among engagement dimensions.

3.6 Data Visualization

Data visualization plays a crucial role in the data analysis process [34]. Its purpose is to reveal trends, relationships, and potential irregularities that are not easily detectable through numerical inspection alone [35]. With increasing data volume and complexity, visualization serves as an important link between large datasets and their practical interpretation. Effective visual analysis requires an understanding of the underlying data so that identified patterns can be interpreted accurately and in the appropriate context [36] [37]. These considerations form the basis of the visualization approach in this study, where graphical representations are used to interpret engagement patterns and clustering outcomes.

3.7 Apache Spark Framework

Apache Spark [38] is an open-source framework that is meant for processing large amounts of data. It focuses on speed, ease of development, and advanced analytical capabilities. The system was built using Spark's in-memory processing features, which let it keep intermediate results in memory instead of writing them to disk over and over again. This made analysis take less time. Reading data from AWS S3 at the same time made it possible to quickly process large datasets. The cloud cluster was set up so that the whole dataset could be processed in memory. This made latency even lower and made it possible to see the data almost in real time.

4 Methodology

4.1 Data Collection and Storage

The dataset used in this study was created by integrating three heterogeneous sources of educational data: Moodle log files, Google Sheets documents, and records from the institution’s information system. These sources were selected because together they capture the full range of student activity in a blended learning environment. Moodle logs record online engagement, including access to resources, forum participation, quiz attempts, and interaction with homework and laboratory materials. Google Sheets records reflect continuous assessment

conducted throughout the semester, while the institutional databases provide official course outcomes such as final test scores and grades. Combined, these sources provide both process-level data (learning activities and engagement) and outcome-level data (performance), which are essential for diagnostic and cluster analysis. The dataset includes all students enrolled in the course during the observed semester; no sampling or exclusion was applied. It therefore represents an exhaustive dataset for the examined cohort, reducing the potential for sampling bias. The integrated dataset was processed in the Apache Spark environment on the Databricks platform, which supports large-scale data storage and analytics [39]. All data were stored in Amazon S3, a cloud-based storage service provided by AWS. For the analysis, an AWS cloud cluster with 64 GB of RAM and eight processing cores was provisioned. A detailed description of the extracted attributes is provided in Table 1.

Table 1
Description of extracted attributes

Attribute	Description	Value
CountF	Forum access count	[1,…,60]
CountI	Course guide access count	[1,…,13]
CountLW	Number of accesses to laboratory preparation videos	[1,…,73]
CountHW	Number of accesses to homework assignment materials	[1,…,117]
CountLec	Number of accesses to lessons	[1,…,140]
AttLec	Points achieved in lectures (interaction score)	[1,…,10]
LW	Points achieved in laboratory exercises	[1,…,10]
HE	Points achieved in homework assignments	[1,…,20]
ET	Points achieved in exam test	[-1,…,70]
Grade	Finale grade	[3,5,6,7,8,9,10]

Each row in the dataset represents a record of a student's activities and points achieved. From Table 1 it can be observed that most attributes are numeric. The Grade attribute has categorical values: 3 indicates that the student did not take the exam, 5 indicates failure, while values 6 through 10 are passing grades.

4.2 Proposed Framework

This study presents an environment for processing and visually monitoring large educational datasets collected from heterogeneous sources. The environment is implemented on the Apache Spark platform, where big data analytical processes are executed using the PySpark programming interface. This configuration provides a unified workflow for preprocessing, normalization, dimensionality reduction, clustering, and visualization, which are described in the following sections.

4.2.1 Descriptive Analysis

During this phase, we have noted that several attributes contain missing values. These missing values appear only in attributes that record online actions. Moodle generates a log entry whenever a student accesses a resource. If no entry exists, this indicates that the student did not access the resource and does not represent a logging failure. Therefore, missing values in these attributes were replaced with zero, as the absence of a log entry appropriately reflects no activity. Figure 1 shows how frequently students accessed different Moodle course materials. A notably higher proportion of missing values is observed for the CountI attribute, which represents accesses to the Course Guide an informational PDF file rather than a learning resource.

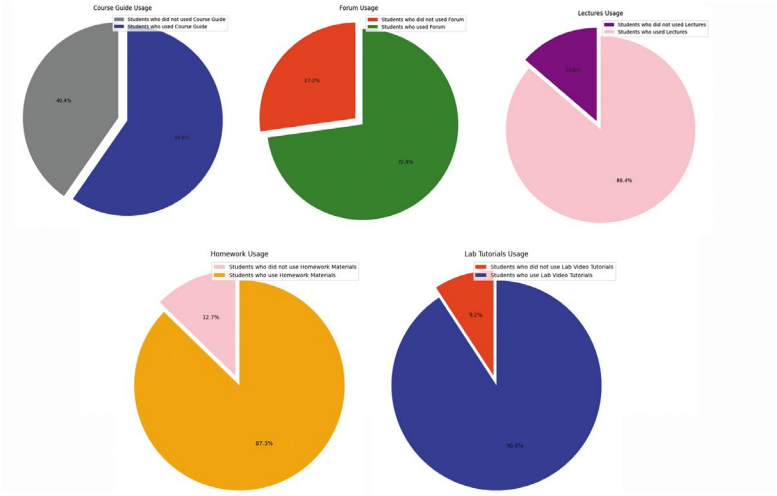
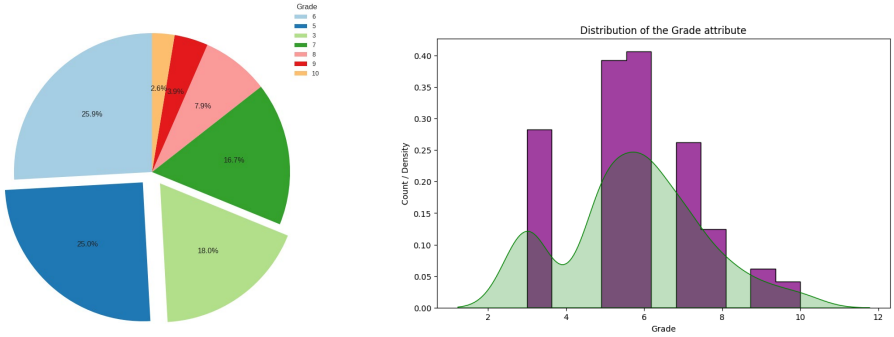


Figure 1
Distribution of Moodle course material usage



(a) percentage distribution (b) frequency with density approximation

Figure 2
Final grade distribution

Final grade distribution is shown in Figure 2. Figure 2a presents the percentage distribution of final grades, while Figure 2b shows their frequency together with a smoothed density curve.

Figure 3 illustrates the distribution of AttLec, which reflects student interactivity and participation during lectures.

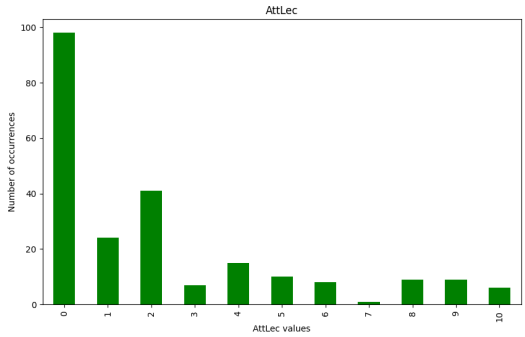


Figure 3
Distribution of AttLec values

The LW, HE, ET attributes represent scores from laboratory exercises, homework assignments, and the exam test, respectively. These distributions are shown in Figure 4 and exhibit considerable variation.

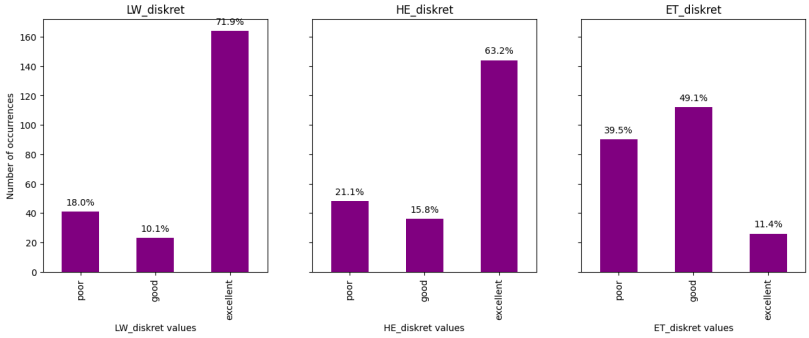


Figure 4
Score distributions of LW, HE, ET

To improve readability, a discretization procedure was applied [23]. Numerical values were grouped into categories by selecting an appropriate number of bins. Based on these intervals, LW, HE, ET were classified into three performance categories: poor, good, and excellent.

4.2.2 Diagnostic Analysis

Diagnostic analysis was applied to identify factors associated with student performance and to examine indicators that may influence the final course grade. Data on students' use of course materials and participation in various activities were analyzed to detect correlations and assess their relationship with academic outcomes. The correlations between final grades and the frequency of accessing different Moodle resources are shown in Figure 5.

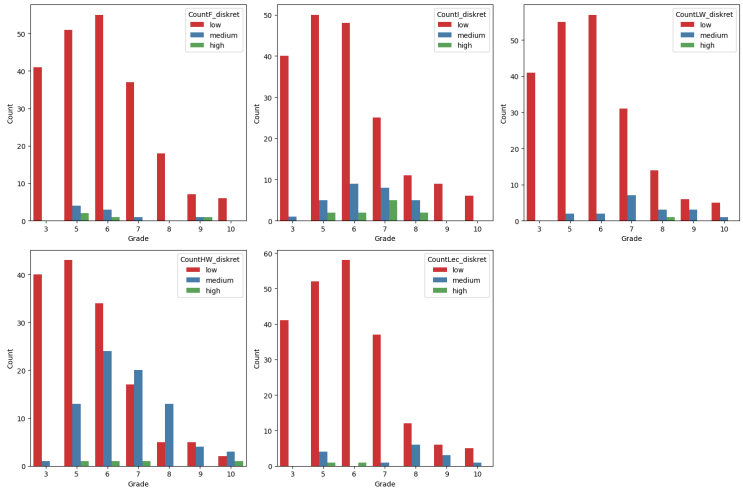


Figure 5
Correlations between Moodle course materials access and final grades

Figure 6 presents the relationship between students' participation in lecture related activities and their final grades. The visualization indicates that students who achieved the highest grades were significantly more engaged in discussions and interactive components during lectures.

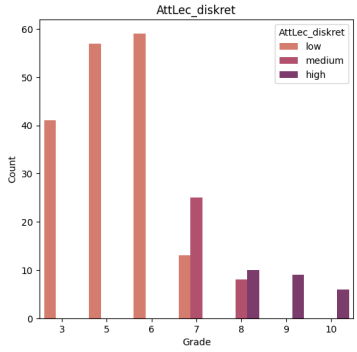


Figure 6
Correlation between lecture participation levels and final grades

Figure 7 shows the distribution of points earned on laboratory exercises, homework assignments, and the exam test, grouped by final grade. Violin plots were used to represent both the distribution and density of scores within each grade category, highlighting the shape and spread of the data.

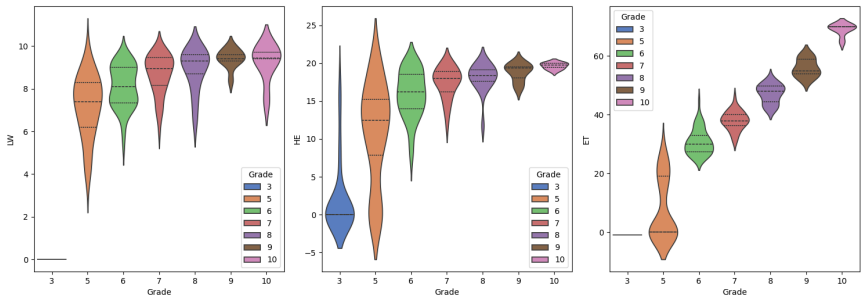


Figure 7
Score distributions (laboratory, homework, exam) by final grade categories

To further explore relationships between the attributes, a correlation matrix was created (Figure 8). Positive coefficients indicate that higher values of one attribute are associated with higher values of another, while negative coefficients indicate an inverse relationship. The heatmap highlights several strong correlations between student activities and academic outcomes. The most intense cells (correlation coefficients from approximately 0.52 to 0.92) show that frequent use of Moodle materials particularly those related to homework and laboratory preparation is strongly associated with higher exam test scores. Students who regularly accessed these resources tended to achieve higher test results, suggesting that engagement with these materials has a substantial positive effect on performance.

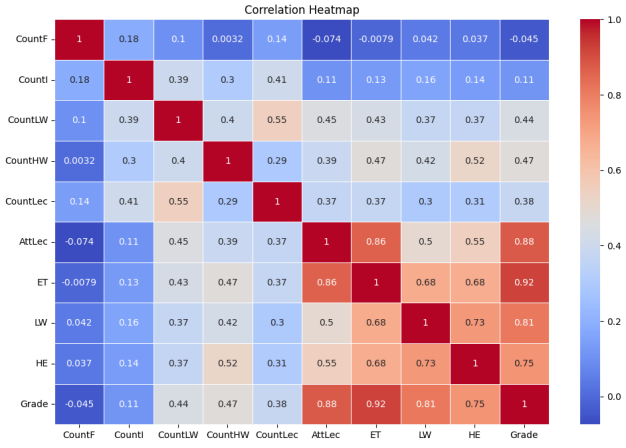


Figure 8
Correlation matrix of the analyzed attributes

4.2.3 Proposed Cluster Analysis Approach

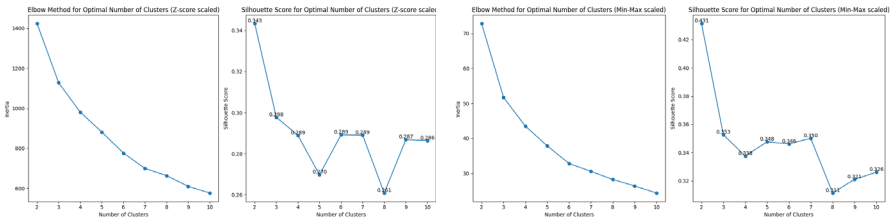
The proposed cluster analysis is based on a normalization procedure designed to accommodate heterogeneous data sources. Clusters were formed by identifying patterns in students' use of Moodle course materials combined with the points earned for activities completed in the face-to-face component of the course. A substantial number of missing values was observed in the attributes related to access counts for Moodle materials and resources (Table 2).

Table 2

The number of missing values per attribute

Attribute	Description	NaN
CountF	Forum access count	62
CountI	Course guide access count	92
CountLW	Number of accesses to laboratory preparation videos	21
CountHW	Number of accesses to homework assignment materials	29
CountLec	Number of accesses to lessons	31

After replacing the missing values with zeros, it became evident that the attributes in the dataset differed considerably in their value ranges and scales. Z-score and Min-Max methods were used depending on each variable's statistical qualities to compare features with heterogeneous magnitudes and distributions. For qualities with essentially normal distributions, Z-score normalization rescaled values around zero with unit variance. Min-Max scaling reduced extreme values in highly skewed activity measurements, such as resource-access frequencies. Dual-scaling delivers a balanced transformation suited to specific attributes, improving downstream clustering stability and interpretability. To evaluate how normalization influences clustering performance, the elbow method and the silhouette coefficient were calculated for each normalized dataset. Applying Z-score and Min–Max scaling resulted in two normalized versions. The elbow and silhouette analyses were used to estimate the optimal number of clusters for both datasets. Figure 9a presents the evaluation results for the Z-score normalized dataset, while Figure 9b shows the corresponding results for the dataset normalized using the Min–Max method.



a. df_KM_zscore_scaled b. df_KM_minmax_scaled

Figure 9

Clustering evaluation for the normalized datasets

Table 3 shows that the Min–Max normalization consistently yielded higher silhouette scores than the Z-score method. The trends indicate that silhouette values decrease from two to five clusters for the Z-score normalization, and from two to four clusters for the Min–Max normalization. Beyond these points, both methods show only minor fluctuations with small alternating increases and decreases.

Table 3
Optimal number of clusters (k) and silhouette scores (Min-Max vs Z-score)

k	Silhouette Score (Min-Max)	Silhouette Score (Z-score)
2	0.431	0.343
3	0.353	0.298
4	0.338	0.289
5	0.348	0.270
6	0.346	0.289
7	0.350	0.289
8	0.311	0.261
9	0.321	0.287
10	0.326	0.286

Figure 10 shows the distributions of instances in the PCA-transformed space, for two, three, and four clusters in both normalized datasets. Clusters formed using Min–Max normalization show clearer boundaries and less overlap, consistent with the higher silhouette scores. In contrast, Z-score normalization produces more diffuse and less distinct clusters, especially as the number of clusters increases. These PCA-based visuals indicate that normalization has a substantial impact on cluster quality. In this study, Min–Max scaling produced more coherent and better-separated clusters.

PCA was applied to obtain a lower-dimensional representation suitable for visual inspection of cluster structure. Complete separability of clusters is not expected due to the complexity of educational behavioral data; thus, PCA serves primarily as a visualization aid rather than a measure of clustering performance. Tables 4 and 5 present the centroid vectors for the selected cluster solutions obtained from the Z-score and Min–Max normalized datasets. Each centroid represents the characteristic profile of a cluster, expressed through the mean normalized values of the original attributes. These profiles provide insight into differences in student engagement patterns and performance indicators.

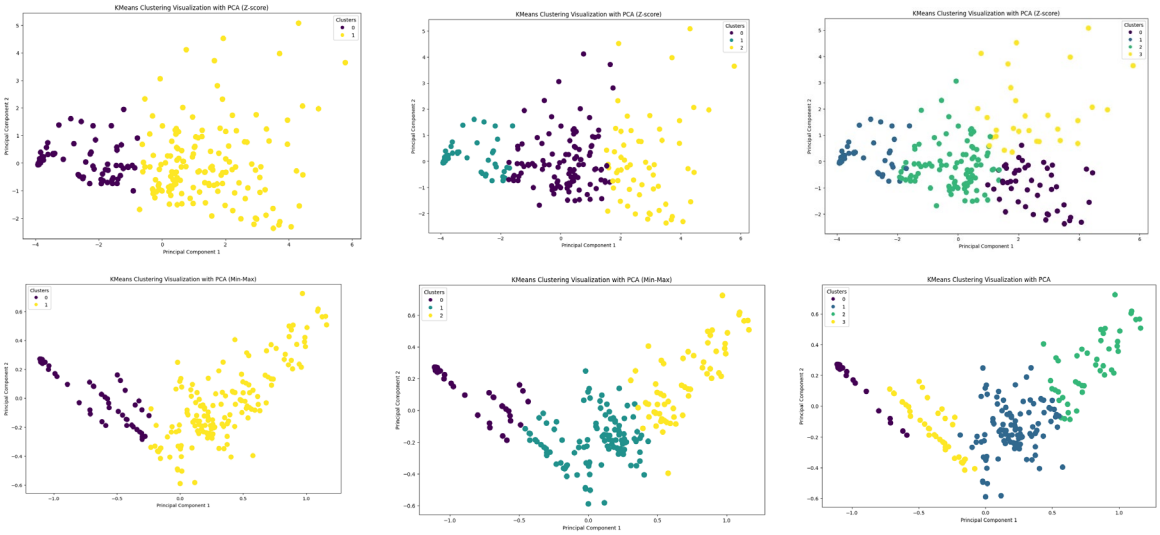


Figure 10
Normalized dataset instance distributions for 2, 3, and 4 clusters

Table 4
Centroid vectors for clusters in *df_KM_zscore_scaled* (k=4)

k	***CountF***	***CountI***	***CountLW***	***CountHW***	***CountLec***
0	-15.531	-0.151	37.164	71.259	70.934
1	-21.143	-6.954	-59.090	-135.987	-96.850
2	6.206	-3.791	-13.903	3.484	-36.627
3	42.627	25.896	95.221	121.145	187.055
k	***AttLec***	***LW***	***HE***	***ET***	***Grade***
0	14.947	6.926	15.689	84.403	12.040
1	-7.745	-14.778	-31.197	-86.182	-6.214
2	-4.143	3.107	6.269	-6.726	2.672
3	4.072	5.176	10.188	35.813	6.306

Table 5
Centroid vectors for clusters in *df_KM_minmax_scaled* (k=4)

k	***CountF***	***CountI***	***CountLW***	***CountHW***	***CountLec***
0	2.611	0.537	2.293	4.927	1.609
1	7.524	2.602	12.243	37.495	17.573
2	6.022	3.422	20.133	45.111	31.489
3	5.000	0.744	7.333	15.051	5.513
k	***AttLec***	***LW***	***HE***	***ET***	***Grade***
0	-2.776e-16	-3.331e-15	1.992	-1.000	3.000
1	1.748e+00	8.193e+00	16.112	29.697	6.078

2	7.222e+00	8.933e+00	18.449	50.003	8.200
3	2.776e–16	6.928e+00	8.641	4.410	5.000

Using the centroid vectors from the Z-score normalized dataset (Table 4), the behavioral characteristics of each cluster can be identified. Cluster 0 shows large positive centroid values across all engagement attributes (CountLW, CountHW, CountLec, ET), indicating very high activity. This group has the highest grade centroid (12.04 in Z-score units) and represents students with strong engagement and strong performance. Cluster 1 exhibits consistently negative centroid values, indicating low engagement in lectures, homework, and online learning materials. The group also shows weak performance, with a grade centroid of –6.21. Cluster 2 has centroid values near zero, reflecting moderate engagement and average achievement (grade centroid ≈ 2.67). Cluster 3 shows high values for structured academic activities (CountHW, CountLW, CountLec), suggesting strategically engaged students who focus on formal course requirements and achieve above-average results (grade centroid = 6.31). These patterns are consistent with findings in learning analytics research, where higher behavioral engagement is typically associated with stronger academic performance.

Using the centroid vectors from the Min–Max normalized dataset (Table 5), a similar cluster structure appears. Cluster 0 shows very low engagement, with values near zero across most attributes (e.g., CountLW = 2.29, ET = –1.00). Cluster 1 represents students with moderate engagement, reflected in lower-to-mid activity levels (CountHW = 37.49, HE = 16.11, ET = 29.70) and mid-range grades (6.08). Cluster 2 displays the highest engagement and performance, with the largest centroids for homework (45.11), lecture participation (31.49), overall interaction (50.00), and the highest grade centroid (8.20). Cluster 3 is smaller and shows moderate activity (CountHW = 15.05, ET = 4.41) and average performance (grade = 5.00).

We evaluated K-means models trained on raw, Z-score, and Min–Max normalized data to assess how normalization affects clustering performance. Z-score scaling reduced within-cluster variance by 99.27% (from 557.37 to 4.07) and slightly lowered the silhouette score (0.281 vs. 0.316). Min–Max scaling achieved an even stronger variance reduction (to 0.183) and produced the highest silhouette score (0.338), indicating more compact and better-separated clusters. Both normalization methods improved cohesion relative to unscaled data, but Min–Max was the most effective for this dataset.

Although Min–Max produced the most coherent clusters, it remains sensitive to outliers and shifts in data range; this can be mitigated through outlier filtering or adaptive scaling. We also tested alternative algorithms. DBSCAN detected some local density patterns but was highly unstable across ε and MinPts settings. Hierarchical clustering revealed structure but produced unbalanced clusters and was slower to compute. K-means yielded the most stable and interpretable behavioral groups. To validate the four-cluster solution, we performed a

bootstrapped silhouette analysis (1000 resamples), which showed minimal variability and strong stability. Repeated 80% subsampling with Adjusted Rand Index (ARI) evaluation also produced consistently high agreement with the original solution. These results confirm that the four-cluster structure is robust and reliable for this dataset.

4.2.4 Statistical and Comparative Analysis

To complement the descriptive and diagnostic results, several statistical tests were conducted to assess the significance and robustness of the observed behavioral patterns (Table 6). The analysis examined correlations between engagement indicators and academic performance, differences in final grades across the four K-means clusters, and post-hoc comparisons to identify which cluster pairs exhibited significant differences.

Pearson and Spearman correlations showed moderate to strong positive relationships between engagement indicators and final grades, with all results significant at $p < 0.01$. These findings confirm that higher engagement is consistently linked to better academic performance. Significant performance differences across the four K-means clusters were supported by both the Kruskal–Wallis test ($H = 189.19$, $p < 0.001$) and one-way ANOVA ($F = 424.52$, $p < 0.001$). Post-hoc tests (Tukey HSD and Dunn with Bonferroni correction) indicated that all cluster pairs differed significantly ($p < 0.001$), confirming clear performance separation.

Table 6
Summary of statistical tests on engagement features and cluster differences

Test	Statistic	p-value	Interpretation
Pearson correlations	r = moderate–strong	< 0.01	Engagement positively associated with grade
Spearman correlations	ρ = moderate–strong	< 0.01	Confirms monotonic relationship
ANOVA	F = 424.52	4.25×10^{-92}	Significant differences between clusters
Kruskal–Wallis	H = 189.19	9.11×10^{-41}	Significant non-parametric cluster differences
Tukey HSD	All pairs significant	< 0.001	Clusters differ in mean grade
Dunn test (Bonferroni)	All pairs significant	< 0.05	Strong pairwise separation

Mean grade differences (Table 7) ranged from 0.83 (Cluster 0 vs. Cluster 1) to 4.93 (Cluster 2 vs. Cluster 3). Dunn's test with Bonferroni correction (Table 8) confirmed that all differences remained statistically significant. These results indicate that the four clusters form clearly distinct learner profiles that differ consistently in both engagement and performance, with no overlap in confidence intervals.

Table 7
Tukey HSD Post-hoc Comparison

Group 1	Group 2	Mean Diff	p-adj	Lower CI	Upper CI	Reject
0	1	0.8302	0.0000	0.4302	1.2302	True
0	2	2.9298	0.0000	2.4993	3.3604	True
0	3	-2.0000	0.0000	-2.4547	-1.5453	True
1	2	2.0996	0.0000	1.8090	2.3902	True
1	3	-2.8302	0.0000	-3.1556	-2.5048	True
2	3	-4.9298	0.0000	-5.2921	-4.5675	True

Table 8
Dunn Post-hoc Test (Bonferroni-corrected p-values)

Cluster	0	1	2	3
0	1.0000	0.0080	5.33×10^{-15}	0.0187
1	0.0080	1.0000	3.99×10^{-13}	3.98×10^{-15}
2	5.33×10^{-15}	3.99×10^{-13}	1.0000	2.20×10^{-39}
3	0.0187	3.98×10^{-15}	2.20×10^{-39}	1.0000

5 Discussion

This study examined student engagement and performance in a blended learning course by integrating data from heterogeneous educational data sources. Increased interaction with instructional materials, especially laboratory preparation resources and homework was associated with higher academic outcomes. Participation in lecture related activities also showed a positive relationship with final grades, reflecting well-known links between behavioral involvement and performance. Cluster analysis showed that engagement varies across distinct behavioral profiles. High-engagement clusters displayed patterns consistent with effective self-regulation and proactive study habits. Low-engagement clusters may signal motivational issues, weaker learning strategies, or external constraints. These interpretations remain tentative, as strong students may naturally interact more with course materials, meaning that digital traces can reflect consequences rather than causes of performance differences.

The results should be viewed within the context of a single course at one institution. Several limitations apply. Replacing missing log entries with zeros assumes non-engagement and may mask technical or logging inconsistencies. The analytical framework depends on cloud-based distributed computing, which may not be available in all educational settings. Furthermore, log level behavioral data cannot capture cognitive or motivational dimensions of learning; combining them with self-report or qualitative data would provide a broader perspective.

Despite these constraints, the framework offers practical value. Integrating heterogeneous data sources and enabling near real time analytics supports early identification of at-risk students and helps instructors interpret engagement patterns. Overall, the study shows that combining descriptive, diagnostic, and clustering methods within a scalable computational environment can reveal meaningful behavioral structures in blended learning.

Future work should test the approach in additional courses, refine the handling of missing data, and incorporate learner-centered measures to deepen the understanding of engagement and learning processes.

Conclusions

This study presented a scalable framework for analyzing and visualizing educational data collected from heterogeneous digital sources. By combining descriptive, diagnostic, and clustering methods, the system provides real-time insight into student behavior in blended learning environments. The analytical workflow covering preprocessing, normalization, dimensionality reduction, and clustering showed that Min–Max scaling combined with PCA improves cluster compactness and interpretability. The resulting cluster profiles revealed clear differences in engagement and performance, giving instructors a clearer picture of how student groups interact with course materials. Visualizations supported interpretation by highlighting resource-use patterns and potential indicators of academic risk.

Several limitations should be noted. Digital traces offer only partial insight into cognitive and motivational processes. Reliance on cloud-based infrastructure may limit adoption in settings with restricted technical resources. Missing or incomplete log data can obscure important behaviors. Because the analysis was conducted on a single course, the findings should not be generalized without further validation. Future research should examine additional courses and contexts, apply more extensive validation procedures, and explore alternative clustering and predictive models. Ethical issues related to privacy, transparency, and fairness also remain essential.

With further refinement, the proposed framework can support early identification of at-risk students, improve evaluation of learning resources, and contribute to more effective course design in blended learning environments.

References

[1] Y. Shi, “Big Data and Big Data Analytics”, in *Advances in Big Data Analytics*, Springer Nature Singapore, 2022, Ch. 1

[2] V. Rajaraman, “Big Data Analytics”, *Resonance*, 2016, 21(7), pp. 695-716

[3] E. Caldarola and A. Rinaldi, “Big Data Visualization Tools: A Survey – The New Paradigms, Methodologies and Tools for Large Data Sets

Visualization", Proc. of the 6th Int. Conf. on Data Science, Technology and Applications, 2017, pp. 296-305

[4] A. Nikolovska, A. Velinov, S. Spasov and Z. Zdravev, "Framework for big data analytics of Moodle data using Hadoop in the Cloud", Proc. of the 18th International Scientific Conference Computer Science, 2018, pp. 3-8

[5] O. G. Glazunova et al., "Moodle tools for educational analytics of the use of electronic resources of the university's portal", Proc. of the Symposium on Advances in Educational Technology (AET 2020), 2020, pp. 444-451

[6] Y. Cui, X. Song, Q. Hu, Y. Li, A. Shanthini and T. Vadivel, "Big data visualization using multimodal feedback in education", *Computers & Electrical Engineering*, 2021, 96, 107544

[7] P. Wang, Z. Pengfei and L. Yingji, "Design of education information platform on education big data visualization", *Wireless Communications and Mobile Computing*, 2022

[8] R. Krishnan, S. Nair, B. S. Saamuel, S. Justin, C. Iwendi, C. Biamba and E. Ibeke, "Smart Analysis of Learners' Performance Using Learning Analytics for Improving Academic Progression: A Case Study Model", *Sustainability*, 2022, 14(7), 3378

[9] K. Dobashi et al., "Learning pattern classification using Moodle logs and the visualization of browsing processes by time-series cross-section", *Computers and Education: Artificial Intelligence*, 2022, 3, 100105

[10] G. Dimić, I. Milošević and Lj. Pecić, "Big Data Analytics Process Implementation on an Educational Data Set Extracted from Online Testing System", Proc. of the 9th Int. Scientific Conf. Technics and Informatics in Education, 2022, pp. 229-237

[11] Z. Papamitsiou and A. A. Economides, "Learning analytics and educational data mining in practice: A systematic literature review of empirical evidence", *Educational Technology & Society*, 2014, 17(4), pp. 49-64

[12] A. Nguyen, L. Gardner and D. Sheridan, "Data analytics in higher education: An integrated view", *Journal of Information Systems Education*, 2020, 31(1), pp. 61-71

[13] C. Romero and S. Ventura, "Educational Data Mining: A Review of the State of the Art", 2020

[14] M. Saqr and S. López-Pernas, "Learning analytics in blended learning: A systematic review," *Computers & Education*, Vol. 163, 104099, 2021

[15] M. Hlosta, Z. Zdráhal, and J. Zendulka, "Beating the baseline: Predicting student success using LMS data and diverse learner attributes," *Journal of Learning Analytics*, Vol. 7, No. 3, pp. 120-139, 2020

[16] S. Charleer, A. Vande Moere, K. Verbert, et al., "Visual analytics for learning dashboards: A systematic review," *Computers & Education*, Vol. 180, 104431, 2022.

[17] M. Saqr, J. Nouri, and U. Fors, "Network analytics for studying student interactions in digital learning environments," *Journal of Learning Analytics*, Vol. 7, No. 1, pp. 20-38, 2020

[18] B. J. Zimmerman, "Becoming a Self-Regulated Learner: An Overview", *Theory Into Practice*, 2002, 41(2), pp. 64-70

[19] J. Sweller, "Cognitive Load During Problem Solving: Effects on Learning", *Cognitive Science*, 1988, 12(2), pp. 257-285

[20] J. A. Fredricks, P. C. Blumenfeld and A. H. Paris, "School Engagement: Potential of the Concept, State of the Evidence", *Review of Educational Research*, 2004, 74(1), pp. 59-109

[21] G. D. Kuh, "What We're Learning About Student Engagement from NSSE", *Change: The Magazine of Higher Learning*, 2003, 35(2), pp. 24-32

[22] P. Russom, "Big Data Analytics", *TDWI Best Practices Report*, 2011, 19(4), pp. 1-34

[23] J. Han, M. Kamber and J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed., Morgan Kaufmann, 2012

[24] B. Mirkin, *Clustering for Data Mining: A Data Recovery Approach*, Chapman & Hall/CRC, 2005

[25] P. Dangeti, *Statistics for Machine Learning*, Packt Publishing, 2017

[26] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis", *J. Comput. Appl. Math.*, 1987, 20, pp. 53-65

[27] S. Patro and K. Sahu, "Normalization: A preprocessing stage", *arXiv preprint*, 2015, arXiv:1503.06462

[28] G. Ciaburro, V. Ayyadevara and A. Perrier, *Hands-On Machine Learning on Google Cloud Platform*, Packt Publishing, 2018

[29] A. Müller and S. Guido, *Introduction to Machine Learning with Python*, O'Reilly Media, 2017

[30] "Performance Comparison of Dimension Reduction Implementations", *UMAP Documentation*, Available: https://umap-learn.readthedocs.io/en/latest/performance.html (Accessed: November 2025)

[31] I. T. Jolliffe and J. Cadima, "Principal component analysis: a review and recent developments", *Philos. Trans. A Math. Phys. Eng. Sci.*, 2016, 374(2065), 20150202

[32] Y.-C. Chang, "A Survey: Potential Dimensionality Reduction Methods", *arXiv preprint*, 2025, arXiv:2502.11036

[33] H. Jeon, J. Park, S. Shin and J. Seo, "Stop Misusing t-SNE and UMAP for Visual Analytics", *arXiv preprint*, 2025, arXiv:2506.08725

[34] P. Bednár, J. Ivančáková and M. Sarnovský, "Semantic Composition of Data Analytical Processes", *Acta Polytechnica Hungarica*, 2024, 21(2), pp. 133-151

[35] C. Chen, W. Härdle and A. Unwin, *Handbook of Data Visualization*, Springer, 2007

[36] S. M. Ali, N. Gupta, G. K. Nayak and R. K. Lenka, "Big data visualization: Tools and challenges", Proc. of the 2nd Int. Conf. on Contemporary Computing and Informatics (IC3I), 2016, pp. 656-660

[37] S. R. Midway, "Principles of effective data visualization", *Patterns*, 2020, Article 100141

[38] "Apache Spark – A unified engine for large-scale data analytics", *Spark Documentation*, Available: https://spark.apache.org/docs/latest/ (Accessed: November 2025)

[39] "The Databricks Data Intelligence Platform", *Databricks*, Available: https://www.databricks.com/ (Accessed: November 2025)