

Article

A Pilot Study on Multimodal Data Fusion for Academic Identification

Wilson Chango ^{1,2,3,*}  and José Espinosa ^{4,*} 

¹ Departamento de Informática y Ciencias de la Computación, Pontificia Universidad Católica del Ecuador, Campus Esmeraldas (PUCESE), Esmeraldas 080101, Ecuador

² Facultad de Informática y Electrónica, Escuela Superior Politécnica de Chimborazo (ESPOCH), Riobamba 060155, Ecuador

³ Facultad de Ingeniería en Sistemas, Electrónica e Industrial, Universidad Técnica de Ambato (UTA), Av. los Chásquis, Ambato 180206, Ecuador

⁴ Centro de Estudio para el Perfeccionamiento de la Educación Superior, Universidad de La Habana, La Habana 10400, Cuba

* Correspondence: wilson.chango@pucese.edu.ec (W.C.); jose.espinosa@rect.uh.cu (J.E.); Tel.: +593-960949920 (W.C.)

Abstract

This study investigates the efficacy of early multimodal data fusion for identifying academic performance profiles and predicting academic risk in specialized postgraduate education under small-sample constraints ($n = 51$ graduate students distributed across three independent cohorts). Student numerical grades and multi-source qualitative faculty evaluations (self-, peer-, and hetero-evaluation) were integrated at the feature level prior to modeling. To overcome the limitations of traditional small-sample validation, a multi-layered framework was deployed, combining non-linear Kernel Principal Component Analysis (KPCA), multi-method clustering consensus, bootstrap resampling (1000 iterations), stratified nested cross-validation, and an external cohort validation strategy. The analysis revealed a bimodal distribution of academic risk (low-risk: 71.4%, high-risk: 28.6%) with high consensus across unsupervised algorithms (Adjusted Rand Index values spanning 0.922–1.000). Supervised predictive models yielded an unbiased nested cross-validation accuracy of 0.980 ± 0.040 , while external cohort validation demonstrated generalizability with a mean cross-cohort accuracy of 0.970 ± 0.052 and a mean AUC of 0.997 ± 0.001 (95% CI: 0.996–0.998). SHAP feature importance analysis identified Machine Learning (28.68%), Artificial Intelligence (17.93%), and Big Data (16.66%) as the primary drivers of risk differentiation. This study contributes a methodologically rigorous, moderately stable, and reproducible workflow for educational data mining in sample-constrained environments, providing actionable insights for evidence-based curriculum design and early pedagogical interventions. However, given the limited sample size, results should be interpreted as exploratory evidence rather than definitive population-level conclusions.



Academic Editors: Mudassar Razzaq and Marcel Gurris

Received: 27 May 2026

Revised: 30 June 2026

Accepted: 20 July 2026

Published: 25 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Keywords: educational data mining; multimodal data fusion; early fusion; small-sample learning; academic risk prediction

1. Introduction

Higher education has undergone a profound transformation in the aftermath of the COVID-19 pandemic, which not only accelerated the adoption of virtual and hybrid modalities in postgraduate programs but also reshaped teaching, assessment, and academic monitoring practices, leading to an increasing production of heterogeneous educational

data. These data, encompassing both quantitative information derived from academic performance and qualitative inputs related to multi-source teaching evaluations (self-, peer-, and hetero-evaluation), represent a strategic opportunity to understand student performance. However, their effective integration remains a critical challenge for ensuring quality, equity, and relevance in educational processes within highly complex academic contexts [1].

Within this scenario, early data fusion has emerged as a particularly relevant feature-level methodological approach, as it enables the direct integration of quantitative and qualitative variables into a unified representation prior to the modeling stage. This approach facilitates the identification of non-linear relationships and complex interactions among the different components of the educational process, which are difficult to capture through univariate analyses or fragmented approaches that treat data sources independently [2].

In parallel, the shift toward student-centered education models, together with increasing demands for quality assurance, accreditation, and continuous improvement in postgraduate programs, suggests a need to explore advanced analytical methodologies that support curriculum-related decision-making based on solid and reproducible empirical evidence. This shift requires moving beyond an exclusive reliance on traditional descriptive indicators, which often provide a limited, static, and reactive view of academic performance [3].

Nevertheless, the existing learning analytics literature reveals a significant gap in the systematic integration of student academic records with multi-source qualitative evaluations of teaching performance, particularly in specialized postgraduate programs. Furthermore, a persistent methodological challenge in graduate-level data mining is the small sample size constraint typical of specialized master's programs, which usually restricts the use of standard internal validation frameworks due to variance inflation and overfitting risks [4].

To address these limitations, this research is grounded in the principles of Design Science Research (DSR) for the design, implementation, and evaluation of a robust analytical artifact aimed at addressing real-world educational risks. Operationally, the study is structured under the CRISP-DM framework to ensure a systematic data mining workflow, and it draws on established systematic guidelines as a reference for the rigorous review of relevant scientific literature [5–7].

The proposed artifact is implemented and evaluated within a Master's program in Data Science, analyzing an integrated dataset of 51 postgraduate students distributed across three independent and sequential cohorts (Cohort A = 21, Cohort B = 16, Cohort C = 14) across 13 technical and methodological courses. Quantitative grades and qualitative evaluations are integrated through an early data fusion scheme, upon which non-linear dimensionality reduction—specifically polynomial Kernel Principal Component Analysis (KPCA)—is applied, followed by a multi-method unsupervised clustering infrastructure (K-Means, GMM, Spectral, Agglomerative) to isolate stable bimodal academic risk profiles.

Unlike traditional small-sample educational studies that rely on internal validation, this work deploys a rigorous, multi-layered validation framework consisting of a three-experiment external cohort validation loop, empirical bootstrap resampling with 1000 iterations to evaluate spatial topology stability, and stratified nested cross-validation to guarantee unbiased classification estimates. Finally, the model's inner decision logic is decoded through SHapley Additive exPlanations (SHAP) to identify core curricular components that drive academic risk.

Based on this framework, this study seeks to answer the following research question:

To what extent can an early fusion workflow of heterogeneous academic data, combined with non-linear dimensionality reduction and multi-method clustering, provide

stable, explainable, and generalizable academic risk profiles when evaluated through external cohort validation in small-scale postgraduate programs?

Consequently, the primary contribution of this work is both empirical and methodological: it demonstrates that early fusion architectures can be rigorously and systematically evaluated using external cohort validation methods even under constrained postgraduate data scales, ensuring that the resulting risk boundaries remain mathematically stable across independent groups and providing actionable, SHAP-based insights for data-driven curriculum design.

2. Background

2.1. Digital Transformation and Learning Analytics in Graduate Studies

The teaching–learning process has undergone significant structural transformations as a result of the progressive incorporation of digital technologies in higher education. In the post-pandemic scenario, these transformations ceased to be circumstantial responses and became permanent components of educational models, particularly in graduate programs characterized by hybrid dynamics, high specialization, and strong instructor–student interaction [8]. The use of audiovisual resources, virtual learning platforms, and academic management systems has generated data-intensive educational environments, where evidence of academic performance and the quality of the training process is systematically recorded [9].

In this context, Learning Analytics emerges as an interdisciplinary field aimed at the measurement, collection, analysis, and reporting of data about students and their contexts for the purpose of understanding and optimizing learning and the environments in which it occurs [10]. Various studies have shown that, at the graduate level, learning analytics allows for the identification of complex performance patterns, the prediction of academic outcomes, and the support of curricular decision-making processes based on empirical evidence [11,12]. However, most of these approaches remain predominantly focused on quantitative student data, which limits a comprehensive understanding of the educational phenomenon.

A persistent challenge in graduate-level learning analytics is the limited sample size typical of specialized programs, which often enroll between 15 and 50 students per cohort. This constraint necessitates the use of robust validation strategies, including external cohort validation and bootstrap resampling, to ensure the generalizability of findings beyond single-cohort analyses [13].

2.2. Foundations of Multimodal Data Fusion in Education

The increasing availability of educational data from multiple sources—grades, faculty evaluations, digital participation, satisfaction surveys, and behavioral traces on LMS platforms—poses the challenge of how to integrate these heterogeneous data to obtain a holistic, contextualized, and actionable view of the educational process [14,15]. This problem has driven the development of advanced educational data mining and machine learning techniques oriented toward the integration of multimodal information [16].

Data fusion, also referred to as information integration, is defined as the process of combining data from different sources, times, and formats with the objective of transforming raw data into useful information for human or automated decision-making [17]. In the educational field, data fusion allows for dimensionality reduction, optimization of available information, and facilitates the extraction of relevant patterns for academic diagnosis, early intervention, and curricular improvement [18].

Multimodal data fusion integrates quantitative information (grades, attendance, participation) and qualitative information (faculty evaluations, student perceptions, written feedback), which is especially relevant in graduate programs where student profiles are more diverse and pedagogical interactions present a higher degree of personalization [19,20]. Recent studies have demonstrated that incorporating qualitative faculty evaluations—including self-evaluation, peer evaluation, and hetero-evaluation—significantly improves the predictive power of academic risk models compared to grade-only approaches [21].

2.3. Academic Data Fusion Approaches

Specialized literature recognizes three main approaches for academic data fusion:

- **Early Fusion (Feature-level Fusion):** This approach integrates attributes from all data sources into a single matrix before model training. It allows for the preservation of inter-attribute relationships and captures complex interactions between academic and teaching variables, although it can be affected by scaling issues, redundancy, and high dimensionality (see Figure 1) [22].

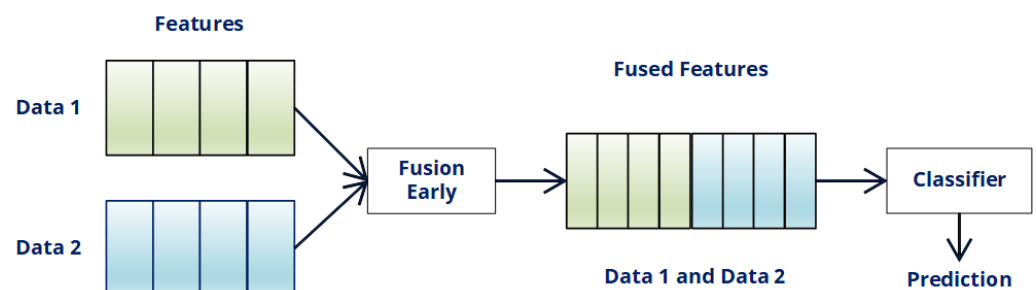


Figure 1. Early fusion: integration of features before model training.

- **Late Fusion (Decision-level Fusion):** In this approach, each data source is processed independently using specific models, and the outputs of these models are combined afterward. While it offers flexibility and modularity, its implementation can be computationally expensive and less interpretable (see Figure 2) [23].

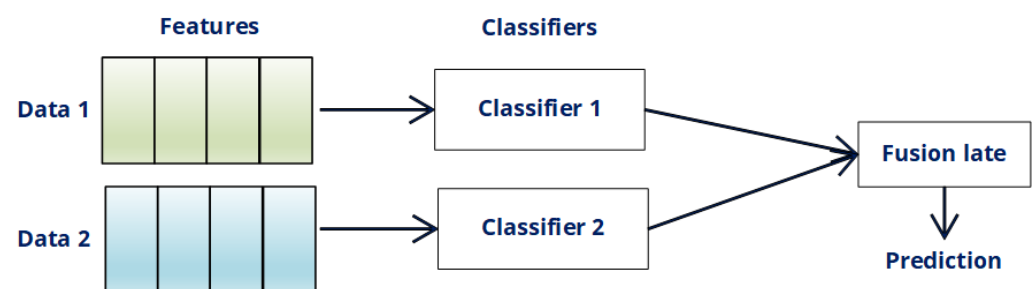


Figure 2. Late fusion: combination of decisions from independently trained models.

- **Hybrid Fusion:** This approach combines both early and late fusion strategies, integrating information at both the feature and decision levels. Hybrid fusion has demonstrated greater robustness and generalization capability in complex educational contexts (see Figure 3) [24].

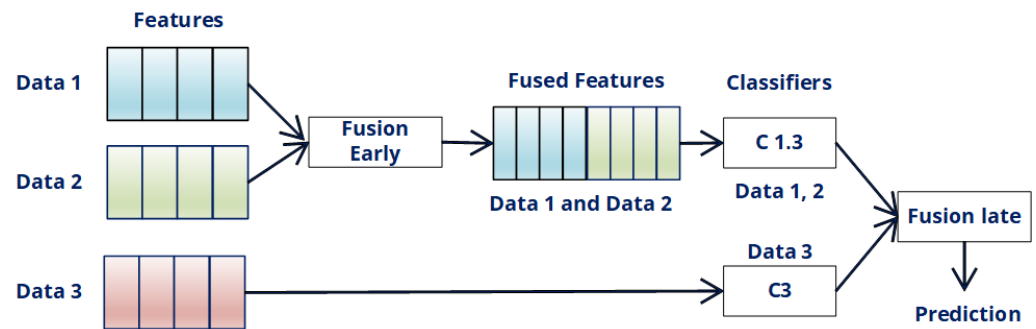


Figure 3. Hybrid fusion: flexible integration for robust academic analysis.

2.4. Validation Strategies for Small-Sample Educational Data Mining

A critical methodological consideration in graduate-level learning analytics is the selection of appropriate validation strategies when working with small sample sizes. Traditional k -fold cross-validation can produce unstable performance estimates when the number of samples per fold is very small (<5). Several robust alternatives have been developed:

- **Leave-One-Out Cross-Validation (LOOCV):** The model is trained on $n - 1$ samples and tested on the excluded sample, repeated for all n samples. LOOCV provides nearly unbiased performance estimates but can be computationally expensive [25].
- **Bootstrap Resampling:** With-replacement sampling generates multiple datasets of size n , allowing for the estimation of confidence intervals and assessment of model stability. Bootstrap with 1000 or more iterations is recommended for robust uncertainty quantification [26].
- **External Cohort Validation:** When data from multiple independent cohorts are available, models can be trained on one set of cohorts and tested on completely unseen cohorts. This provides the strongest evidence for generalizability and is considered the gold standard for small-sample educational data mining [27].
- **Nested Cross-Validation:** An unbiased estimate of model performance is obtained by nesting hyperparameter tuning within an outer cross-validation loop, preventing overfitting and providing reliable generalization estimates [28].

As machine learning models become increasingly complex, interpretability has emerged as a critical requirement for educational applications. SHapley Additive exPlanations (SHAP) provides a unified framework for explaining model predictions based on cooperative game theory [29]. SHAP values decompose each prediction into feature contributions, satisfying desirable properties:

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} [f(S \cup \{i\}) - f(S)] \quad (1)$$

where F is the set of all features and $f(S)$ is the model prediction with features in S . In educational contexts, SHAP enables the identification of which academic subjects most strongly differentiate high-risk from low-risk students, providing actionable insights for curriculum design and targeted interventions [30].

2.5. Summary of Relevant Literature

Altogether, the current state-of-the-art demonstrates that while multimodal academic data fusion constitutes a vital foundation for identifying student performance patterns, existing frameworks present significant methodological variances. To establish a rigorous baseline, Table 1 provides a comprehensive, multi-dimensional critique of prominent bench-

marks, mapping their scale constraints, core learning paradigms, evaluation strategies, and intrinsic operational limitations.

Table 1. Analytical and critical synthesis of educational data mining benchmarks.

Author	Year	Context	N	Algorithms	Metrics	Main Contribution	Core Limitations
[31]	2021	Higher Ed.	450	Early Fusion, Linear Agg.	Accuracy, Precision	Standardized integration framework for heterogeneous quantitative logs.	Relies exclusively on static quantitative inputs; ignores qualitative faculty dimensions.
[32]	2022	LMS/Virtual	312	Hybrid Fusion, RF, Decision Tree	F1-Score, ROC-AUC	Modality-specific preprocessing pipelines to reduce cross-source variance.	Verified only through standard internal validation; high risk of performance inflation.
[33]	2024	Blended Learning	178	SVM, Random Forest	Accuracy, Recall	Empirical comparison of traditional supervised classifiers on student logs.	Fragmented feature space; small-sample constraints are not statistically compensated.
[34]	2025	Digital Ed.	520	Late Fusion, Ensemble Voting	Accuracy, Kappa	Modular architecture for independent prediction aggregation.	Fails to capture complex non-linear feature interactions prior to classification.
[30]	2025	Postgraduate	35	KPCA, K-Means, ML Clf.	Silhouette, Accuracy	Dimensionality reduction scheme tailored for specialized courses.	Missing multi-method clustering consensus and robust stability boundary checks.
[35]	2024	Postgraduate	24	Feature-level Multimodal Fusion	Pearson <i>r</i> , Accuracy	Primitive feature vector concatenation incorporating feedback indexes.	Suffers from high data leakage due to lack of nested or external cross-validation loops.

2.6. Identified Research Gap

A systematic and comparative matrix analysis of the literature compiled in Table 1 uncovers a critical, multi-layered research gap. First, while predictive models of student academic performance are mathematically abundant, they remain heavily siloed within quantitative boundaries (e.g., refs. [31,32]). Only an sparse number of studies attempt to systematically fuse multi-source qualitative faculty evaluations—such as self-, peer-, and hetero-evaluation—as continuous predictive features within early fusion workflows (e.g., ref. [35]).

Second, a severe methodological flaw persists regarding small-sample constraints ($N < 100$) typical of highly specialized postgraduate programs. As evidenced in Table 1, existing small-scale pilots either rely on fragmented unimodal features that neglect cross-attribute interactions, or deploy standard internal validation loops (such as *k*-fold cross-validation or simple holdout divisions) that are mathematically vulnerable to variance inflation, data leakage, and overfitting.

Third, there is a distinct absence of rigorous multi-layered verification protocols that combine non-linear feature mapping (KPCA), multi-method unsupervised consensus, and structural stability quantification via empirical bootstrap resampling. Consequently, prior frameworks lack verified generalizability across temporal or independent groups. This structural gap directly justifies the deployment of the proposed analytical artifact, which combines early multimodal feature fusion, five-method clustering consensus with Adjusted Rand Index (ARI) stability bounding, three-experiment external cohort cross-testing, and stratified nested cross-validation to guarantee unbiased, population-agnostic predictive models.

3. Methodology

Design Science Research (DSR) served as the overarching research paradigm guiding artifact design and evaluation, structuring the research into three core phases: Relevance, Design, and Rigor. The CRISP-DM (Cross-Industry Standard Process for Data Mining) framework was adopted to operationalize the empirical data preparation, modeling, and validation workflow, ensuring a systematic and reproducible process (see Figure 4).

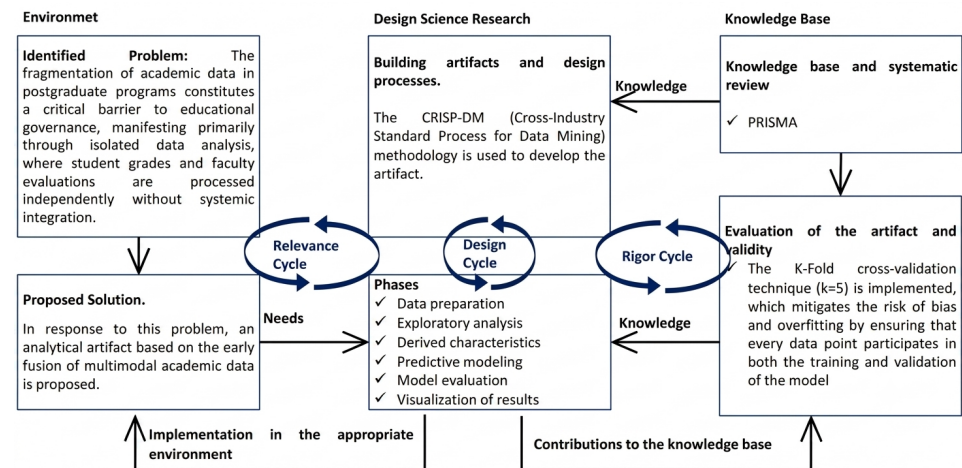


Figure 4. Design of Science Research Framework for Predicting Postgraduate Student Performance Using Multimodal Data Fusion and Machine Learning.

3.1. Relevance Phase: Problem Identification and Preliminary Literature Review (DSR)

The relevance phase establishes the scientific and practical significance of the research problem. Despite the increasing availability of academic data, postgraduate programs often analyze student grades and faculty evaluations in isolation. This fragmented approach limits the identification of latent performance patterns and the formulation of evidence-based curricular strategies, constraining the potential of advanced learning analytics. To address this, the proposed solution involves the early fusion of multimodal academic data (quantitative grades and qualitative evaluations) to support more robust data-driven decision-making, as illustrated in Figure 5.

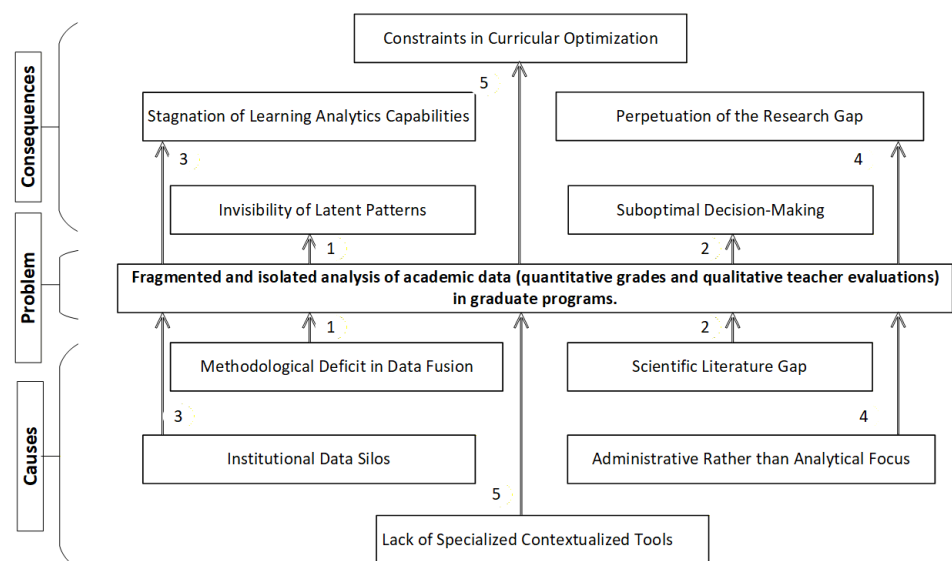


Figure 5. Systemic Analysis of Data Fragmentation in Graduate Education. Arrows represent the data flow and analytical process between components.

A preliminary literature review was conducted to identify relevant variables and methodological approaches, revealing a consistent research gap: while predictive models of student performance are abundant, few systematically integrate multi-source qualitative faculty evaluations—such as self-evaluation and peer evaluation—as predictive variables, particularly in specialized postgraduate contexts. This gap reinforces the relevance and novelty of the proposed analytical artifact.

3.2. Design Phase: Development of the Analytical Artifact (DSR + CRISP-DM)

The design phase operationalizes the proposed solution through the CRISP-DM framework.

3.2.1. Business and Data Understanding

The dataset consists of academic records from 51 postgraduate students distributed across three independent cohorts (Cohort 1: 21 students, Cohort 2: 16 students, Cohort 3: 14 students) enrolled in 16 academic modules within a Master's program in Data Science, organized as a student–subject matrix of dimension 51×16 . This structure includes quantitative grades and qualitative faculty evaluations (self-, peer-, and hetero-evaluation), creating a rich multimodal dataset. This structure is methodologically appropriate for clustering and dimensionality reduction techniques.

3.2.2. Data Preparation and Early Fusion

A multi-step data preparation process was applied, including cleaning, normalization, and pivoting to consolidate the information into a homogeneous matrix format. This process prevented extreme sparsity and enabled the capture of global academic performance patterns. The core of our approach is early fusion, where quantitative variables (student grades) and qualitative variables (encoded faculty evaluations) were normalized and concatenated into a single feature vector at the outset of the modeling process. Mathematically, early data fusion is expressed as

$$X_{\text{fusion}} = [X_{\text{grades}} \mid X_{\text{faculty}}] \quad (2)$$

where X_{grades} represents the student grade matrix and X_{faculty} denotes the transformed faculty evaluation variables. This approach allows for the capture of latent interactions between different data modalities from the very beginning.

3.2.3. Encoding of Qualitative Faculty Evaluations

The faculty evaluation data consisted of three components:

1. Self-evaluation: Students' assessment of their own performance.
2. Peer evaluation: Average ratings from classmates.
3. Hetero-evaluation: Instructor assessment of student performance.

All qualitative evaluations were converted to numerical values using ordinal encoding (Likert scale 1–5 $\rightarrow$ numeric values 1.0–5.0). These encoded values were then concatenated with quantitative grades (0–10 scale) to form the fused feature matrix X_{fusion} .

3.2.4. Feature Selection with RFECV

To reduce dimensionality and identify the most predictive features, Recursive Feature Elimination with Cross-Validation (RFECV) was applied using a Random Forest classifier as the estimator. The optimal feature subset was determined through 5-fold stratified cross-validation. The analysis identified 6 out of 16 variables as the most predictive features for academic risk classification, effectively reducing dimensionality while maintaining predictive power. The selected features—Mass Data Visualization Techniques, Statistical

Data Analysis, Ethics and Privacy in Data Science, Artificial Intelligence, Data Management and Governance, and Deep Learning—formed the basis for subsequent modeling, ensuring the model focused on the most salient academic competencies.

The analysis identified 6 out of 16 variables as the most predictive. It is worth noting that this selection process was performed exclusively on the training data of each nested cross-validation fold, thereby preventing any potential data leakage from the test set.

3.2.5. Dimensionality Reduction and Clustering Analysis

To explore the latent structure of the data, we employed a combination of dimensionality reduction and clustering techniques. Three dimensionality reduction techniques were evaluated: Principal Component Analysis (PCA), Incremental PCA (IPCA), and Kernel PCA (KPCA). In KPCA, data are implicitly projected into a high-dimensional feature space using a kernel function. For the polynomial kernel employed in this study,

$$K(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i^\top \mathbf{x}_j + c)^d \quad (3)$$

where d denotes the polynomial degree and c is a constant.

Based on the preliminary clustering results, we determined the optimal number of clusters by evaluating alternative cluster bounds (K) between 2 and 10 using the K-Means algorithm and optimizing the Silhouette coefficient. The analysis revealed that the maximum Silhouette coefficient was achieved at $K = 2$, indicating the existence of two clearly differentiated academic profiles.

To ensure robustness, five clustering methods were evaluated: K-Means, Gaussian Mixture Models (GMM), DBSCAN, Agglomerative Clustering, and Spectral Clustering. The implementation details for each algorithm were as follows:

- K-Means: Initialization using the ‘k-means++’ method with 10 random restarts to avoid local minima, and a maximum of 300 iterations per execution.
- Gaussian Mixture Models (GMM): Expectation-Maximization (EM) algorithm with 100 maximum iterations and ‘full’ covariance type, initialized with 10 random restarts.
- DBSCAN: Epsilon parameter determined via the k-distance graph method, with minimum samples set to 2.
- Agglomerative Clustering: Ward linkage method with Euclidean distance; the distance matrix was precomputed to optimize computational performance.
- Spectral Clustering: Affinity kernel based on nearest neighbors with 10 neighbors; the number of components for spectral decomposition was set to 10.

The K-Means algorithm was applied to identify academic performance patterns by minimizing the within-cluster sum of squares:

$$\text{WCSS} = \sum_{k=1}^K \sum_{\mathbf{x}_i \in C_k} \|\mathbf{x}_i - \boldsymbol{\mu}_k\|^2 \quad (4)$$

Clustering quality was assessed using the Silhouette coefficient:

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \quad (5)$$

where $a(i)$ is the mean intra-cluster distance and $b(i)$ is the minimum mean distance to the nearest neighboring cluster.

To evaluate consistency across clustering methods, the Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI) were computed between all pairs of methods:

$$\text{ARI} = \frac{\text{RI} - \mathbb{E}[\text{RI}]}{\max(\text{RI}) - \mathbb{E}[\text{RI}]} \quad (6)$$

$$\text{NMI} = \frac{2I(U; V)}{H(U) + H(V)} \quad (7)$$

where $I(U; V)$ is the mutual information between cluster assignments U and V , and $H(\cdot)$ denotes entropy.

For the bootstrap stability analysis, the following implementation details were employed:

- Number of iterations: 1000 iterations, selected to ensure stability of confidence intervals following the recommendations of Efron and Tibshirani (1993) for small samples.
- Resampling method: Sampling with replacement from the original dataset ($n = 51$), generating a bootstrap sample of size n for each iteration b .
- Clustering algorithm per iteration: K-Means (k-means++) exclusively with $K = 2$ (previously determined as the optimal number), maintaining identical initialization parameters (10 random restarts, 300 maximum iterations) to ensure methodological consistency.
- Confidence interval calculation: 95% confidence intervals were computed using the percentile bootstrap method, taking the 2.5th and 97.5th percentiles of the empirical distribution of ARI values obtained across the 1000 iterations:

$$\text{CI}_{95\%} = [\text{Percentile}_{2.5\%}(\text{ARI}_{\text{bootstrap}}), \text{Percentile}_{97.5\%}(\text{ARI}_{\text{bootstrap}})] \quad (8)$$

- Stability metric: The Adjusted Rand Index (ARI) was computed between the cluster labels from each bootstrap sample and the cluster labels from the original dataset, providing a quantitative measure of clustering solution stability under sampling variability.

It is important to note that, for the supervised predictive model, the KPCA transformation was fitted and applied solely within the training folds of the nested cross-validation process, ensuring that the dimensionality reduction did not benefit from any test set information.

3.2.6. Predictive Modeling

For academic risk prediction, we implemented and compared five supervised models: Random Forest, XGBoost, LightGBM, Support Vector Machine (SVM), and Logistic Regression. Hyperparameter tuning was performed using GridSearchCV with 3-fold inner cross-validation. Based on preliminary results, the Random Forest model was selected as the final classifier due to its strong performance, robustness, and inherent interpretability features.

3.3. Rigor Phase: Evaluation and Validation

The rigor phase focuses on the empirical evaluation of the proposed artifact. Given the sample size ($n = 51$ across three cohorts), a multi-layered validation strategy was employed to ensure robust performance estimates and assess generalizability. The comprehensive validation framework included the following components, the results of which are presented in Section 4.

To guarantee a rigorous and unbiased evaluation of the model's performance and to prevent any data leakage that could artificially inflate the results, all preprocessing steps involving parameter learning from the data were executed exclusively within the training folds of the Nested Cross-Validation procedure. Specifically, the Feature Selection via Recursive Feature Elimination with Cross-Validation (RFECV) and Kernel Principal Component Analysis (KPCA) techniques were implemented as follows:

- In each iteration of the outer loop of the nested cross-validation, the training set was partitioned into training and validation subsets via the inner loop.

- The RFECV process was executed solely on the training subset of the inner loop to determine the optimal number of features.
- The KPCA model was fitted (i.e., the kernel transformation was learned) exclusively using the training data of that specific fold.
- Subsequently, the features selected by RFECV and the learned KPCA transformation were applied to transform both the validation subset of the inner loop and the test set of the outer loop.

This meticulous approach ensures that the performance metrics (such as the nested cross-validation accuracy) reflect the true generalization capability of the model on unseen data, eliminating any biases introduced by prior knowledge of the global dataset structure.

Before pooling the three cohorts for joint analysis, comparability was assessed using Chi-square tests and MANOVA. This step was crucial to ensure that pooling the cohorts was statistically justifiable. The analysis revealed a highly significant association between cohort membership and risk level (Chi-square test, $p = 7.87 \times 10^{-8}$), indicating notable differences in cohort composition that would be leveraged in the external validation strategy. The MANOVA results (Wilks' Lambda = 0.0158, $p < 0.001$) further confirmed significant multivariate differences across cohorts.

3.3.1. External Cohort Validation

To rigorously assess generalizability, an external validation strategy was implemented using the three independent cohorts:

- Experiment 1: Training on Cohorts 1 + 2 (37 students), Testing on Cohort 3 (14 students).
- Experiment 2: Training on Cohorts 1 + 3 (35 students), Testing on Cohort 2 (16 students).
- Experiment 3: Training on Cohorts 2 + 3 (30 students), Testing on Cohort 1 (21 students).

Each experiment evaluated the model's ability to generalize to a completely unseen cohort, providing a rigorous test of the model's practical utility.

3.3.2. Bootstrap Resampling

To assess clustering stability and quantify uncertainty, a bootstrap resampling procedure with 1000 iterations was employed. For each iteration $b = 1, \dots, 1000$, a bootstrap sample of size $n = 51$ was drawn with replacement from the original dataset. The complete clustering pipeline—including K -Means clustering with $K = 2$ —was then re-run independently on each bootstrap sample. The stability of the resulting cluster assignments was measured by computing the Adjusted Rand Index (ARI) between the cluster labels obtained from each bootstrap sample and the cluster labels derived from the original dataset. This process generated a distribution of 1000 ARI values, from which the mean, standard deviation, and 95% confidence intervals ($CI_{2.5\%}$, $CI_{97.5\%}$) were calculated to quantify the stability and uncertainty of the clustering solution.

To obtain an unbiased estimate of model performance and prevent overfitting, nested cross-validation was implemented. The outer loop consisted of 5-fold stratified cross-validation, while the inner loop performed 3-fold cross-validation for hyperparameter tuning. This nested CV accuracy provides a nearly unbiased estimate of the model's generalization performance.

3.3.3. Model Performance Metrics

Model performance was assessed using standard metrics: Accuracy, Precision, Recall, F1-score, and Area Under the ROC Curve (AUC). Macro-averaging was used for multi-class metrics to account for class imbalance.

3.4. Summary of Methodological Improvements

Table 2 summarizes the key methodological enhancements implemented in this study compared to conventional approaches.

Table 2. Summary of methodological enhancements for small-sample multimodal academic data fusion.

Component	Implementation in This Study
Sample size	51 students across three independent cohorts (1 = 21, 2 = 16, 3 = 14)
Feature selection	RFECV with Random Forest (6/16 variables selected)
Dimensionality reduction	KPCA with polynomial kernel
Clustering methods	5 methods with ARI/NMI stability assessment, optimal $K = 2$ determined via Silhouette
Bootstrap iterations	1000 iterations with 95% CI for ARI
External validation	3 experiments (train on 2 cohorts, test on 1)
Cross-validation	Nested CV (outer 5-fold, inner 3-fold)
Interpretability	SHAP feature importance (identified Machine Learning, AI, Big Data as top drivers)

Overall, the methodology ensures relevance, design rigor, and comprehensive validation, demonstrating that early fusion of academic data combined with polynomial KPCA, multi-method clustering, external cohort validation, and ensemble models constitutes a rigorous strategy for identifying performance patterns in postgraduate programs with multiple cohorts.

4. Results

4.1. Dataset Overview

The empirical analysis was conducted on an academic dataset consisting of 51 graduate students distributed across three independent cohorts (Cohort 1: 21 students, Cohort 2: 16 students, Cohort 3: 14 students) and 20 variables, organized into a student–subject matrix of dimension 51×20 . From the 663 original records, a process of cleaning, normalization, and pivoting was applied, allowing the information to be consolidated into a homogeneous matrix format suitable for dimensionality reduction techniques, clustering, and predictive modeling.

The main descriptive characteristics of the dataset following the preprocessing stage are compiled in Table 3.

Table 3. Descriptive characteristics of the preprocessed dataset.

Feature	Value
Number of Students (Total)	51
Cohort 1	21
Cohort 2	16
Cohort 3	14
Number of Variables	20
Total Observations	663
Matrix Dimensions	51×20

This structure prevents extreme sparsity issues and enables the capture of global academic performance patterns, which is especially relevant in graduate contexts with moderate sample sizes but high curricular complexity.

4.2. Descriptive Statistics

The descriptive statistics for all 16 academic variables are presented in Table 4. The evaluation scores ranged from 7.0 to 10.0 across different subjects, with Self_evaluation showing the highest mean (9.988 ± 0.043) and minimal variability, while Artificial_Intelligence exhibited the largest standard deviation (1.047).

Table 4. Summary statistics and distribution metrics for evaluated academic modules.

Variable	Count	Mean	Std	Min	25%	50%	75%	Max
Mass Data Visualization Techniques	51.0	8.953	1.036	7.0	8.0	9.6	9.8	10.0
Statistical Data Analysis	51.0	9.029	0.847	7.1	8.55	9.1	9.75	10.0
Research Methodology	51.0	8.927	0.745	7.1	8.45	9.0	9.3	9.9
Machine Learning	51.0	8.735	0.805	7.2	8.1	8.8	9.4	9.8
Ethics and Privacy in Data Science	51.0	8.775	1.037	7.0	7.9	9.1	9.7	10.0
Thesis 1	51.0	8.627	0.924	7.3	7.7	8.9	9.5	10.0
Big Data and Cloud Computing	51.0	9.096	0.747	7.2	8.75	9.2	9.7	10.0
Data Science Fundamentals	51.0	8.831	0.821	7.6	8.1	8.7	9.6	10.0
Artificial Intelligence	51.0	8.816	1.047	7.0	7.9	8.9	9.8	10.0
Data Management and Governance	51.0	9.039	0.757	7.3	8.5	9.1	9.6	10.0
Deep Learning	51.0	8.588	1.005	7.1	7.6	8.9	9.55	10.0
Data Mining	51.0	8.986	0.922	7.0	8.25	9.3	9.75	9.9
Thesis 2	51.0	8.796	0.628	7.2	8.5	9.0	9.3	9.9
Coevaluation	51.0	9.941	0.028	9.9	9.9	9.96	9.96	9.96
Self-evaluation	51.0	9.988	0.043	9.8	10.0	10.0	10.0	10.0
Hetero-evaluation	51.0	8.996	0.173	8.4	9.0	9.01	9.01	9.6

4.3. Correlation Analysis of Academic Variables

To explore the structured interrelationships and potential redundancies among the evaluated academic competencies, a Pearson correlation analysis was conducted. Figure 6 presents the pairwise correlation matrix across the 16 academic variables.

The heatmap reveals a generalized pattern of strong positive correlations ($r \in [0.90, 0.95]$) among the majority of the variables. This multi-collinearity is particularly pronounced between Mass Data Visualization Techniques, Statistical Data Analysis, and Research Methodology. Such high coefficients indicate that these courses share core methodological foundations, where student performance in data manipulation directly scales with their research and synthesis capabilities.

Crucially, near-perfect linear correlations approaching the theoretical maximum ($r \approx 1.0$) are observed in two specific pairs: Machine Learning paired with Artificial Intelligence, and Thesis 1 paired with Thesis 2. Rather than statistical anomalies, these extreme values signal a profound conceptual overlap. In the case of the thesis timeline, this reflects a continuous, sequential evaluation framework where the performance metrics of the initial phase dictate the outcomes of the final defense. For the computing modules, the results imply that the curriculum treats Machine Learning and Artificial Intelligence as deeply integrated, non-orthogonal competencies within the academic program, suggesting potential areas for curricular optimization or dimensionality reduction in future predictive models.

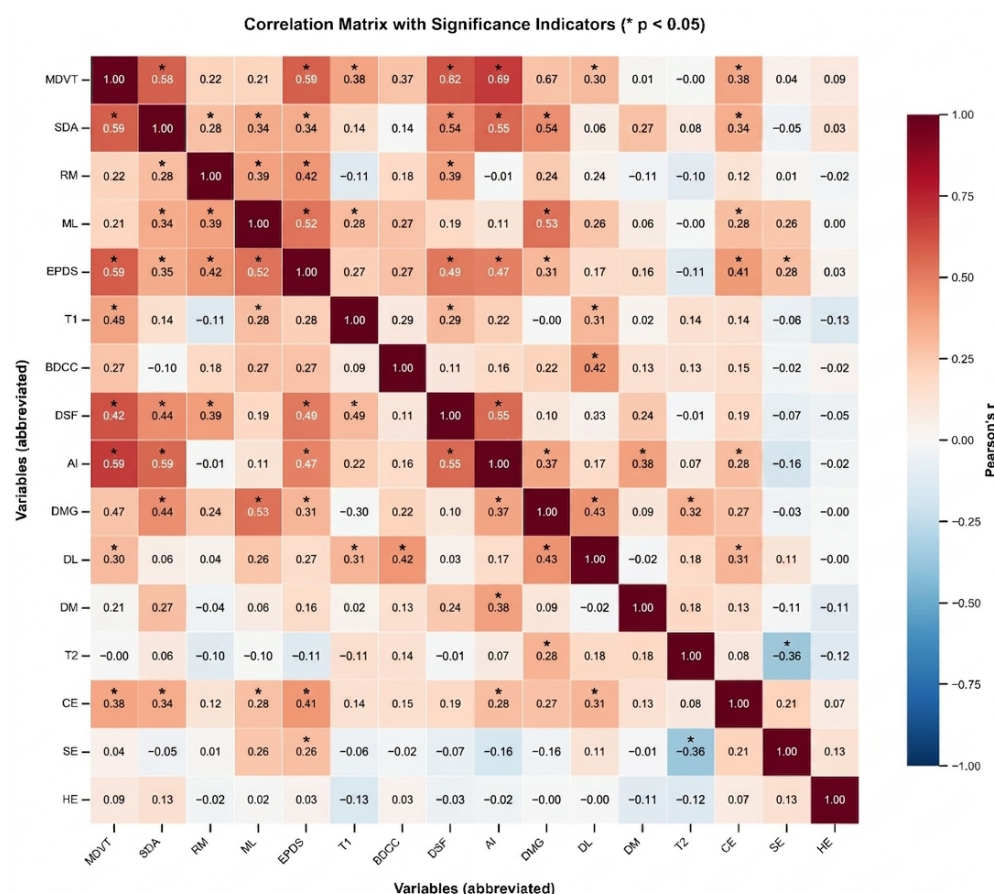


Figure 6. Correlation heatmap of the 16 academic variables. Coefficients range from -1.0 (perfect negative correlation) to 1.0 (perfect positive correlation), with color intensity corresponding to the magnitude of the relationship.

4.4. Cohort Comparability Analysis

Before pooling the three cohorts for joint analysis, cohort comparability was assessed using Chi-square tests and MANOVA. The risk distribution across cohorts is presented in Table 5.

Table 5. Distribution and statistical summary of risk levels across academic cohorts.

Cohort	Total N	High Risk		Medium Risk		Low Risk		Mean $\pm$ SD
		n	%	n	%	n	%	
1	21	0	0.0	4	19.05	17	80.95	9.394 $\pm$ 0.261
2	16	0	0.0	15	93.75	1	6.25	8.748 $\pm$ 0.192
3	14	0	0.0	14	100.0	0	0.0	8.728 $\pm$ 0.177

The Chi-square test for cohort vs. risk distribution yielded $\chi^2 = 32.716$ with $p = 7.87 \times 10^{-8}$ ($df = 2$), indicating a highly significant association between cohort membership and academic risk level. Post hoc Mann–Whitney U tests revealed that Cohort 1 differs significantly from Cohort 2 ($U = 326.0$, $p = 1.36 \times 10^{-6}$) and Cohort 3 ($U = 289.0$, $p = 1.87 \times 10^{-6}$), while Cohorts 2 and 3 do not show a statistically significant difference ($U = 113.0$, $p = 0.983$).

The integrated cohort analysis, including risk distribution, absolute counts, and average trends, is presented in Figure 7.

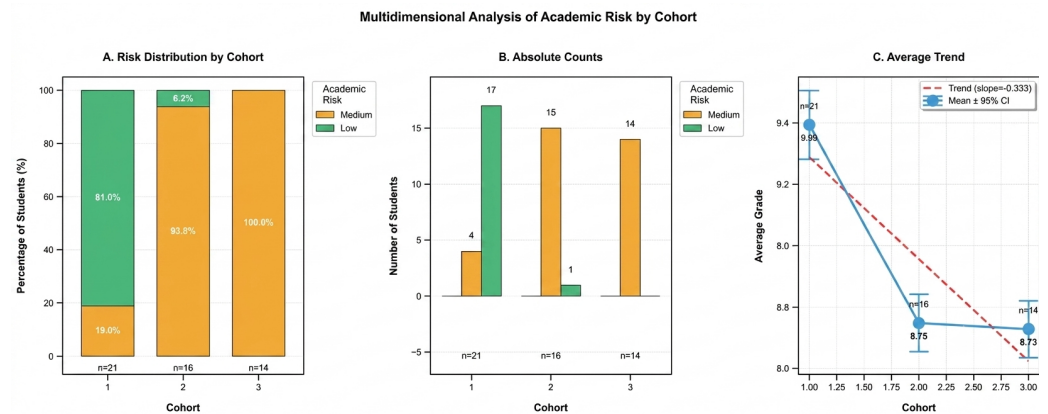


Figure 7. Multidimensional analysis of academic risk by cohort. Panel (A) shows the percentage distribution of risk levels, highlighting a predominance of Low Risk (81.0%) in Cohort 1 ($n = 21$) and Medium Risk in Cohorts 2 ($n = 16$) and 3 ($n = 14$). Panel (B) details the absolute student counts, and Panel (C) displays the downward average grade trend across cohorts (slope = -0.333) with their respective 95% confidence intervals. The legends in (A,B) reflect that no students are classified as ‘High’ risk in these cohorts.

4.5. Clustering Stability Between Methods

To evaluate the consistency of clustering solutions across different mathematical algorithms, three clustering methods (K-Means, GMM, and Spectral) were compared. The performance metrics for each algorithm are presented in Table 6.

Table 6. Clustering algorithm performance comparison.

Algorithm	Silhouette	Calinski-Harabasz	Davies-Bouldin
K-Means	0.182	10.289	1.864
GMM	0.173	9.653	2.324
Spectral	0.185	10.029	2.197

Spectral clustering achieved the highest Silhouette score (0.185), while K-Means demonstrated the best Calinski–Harabasz index (10.289) and the lowest Davies–Bouldin index (1.864), indicating more compact and well-separated clusters.

4.6. Bootstrap Validation Results

To assess the statistical stability of the clustering solution given the sample size ($n = 51$), a bootstrap resampling procedure with 1000 iterations was employed. In each iteration, a bootstrap sample was drawn with replacement, and the K-Means clustering algorithm (with $K = 2$ determined by the Silhouette coefficient) was re-applied independently to that sample. The Adjusted Rand Index (ARI) was then computed between the cluster labels from the bootstrap sample and the original cluster labels, providing a distribution of ARI values that reflects the stability of the clustering structure under sampling variability. The resulting ARI distribution is shown in Figure 8.

The bootstrap validation results, including the mean, standard deviation, and 95% confidence intervals derived from the 1000 resampled ARI values, are summarized in Table 7.

The bootstrap results indicate moderate cluster stability with a mean ARI of 0.459 and a 95% confidence interval of $[0.104, 0.936]$. The relatively wide confidence interval reflects the sample size limitations but confirms that the clustering solution is statistically meaningful.

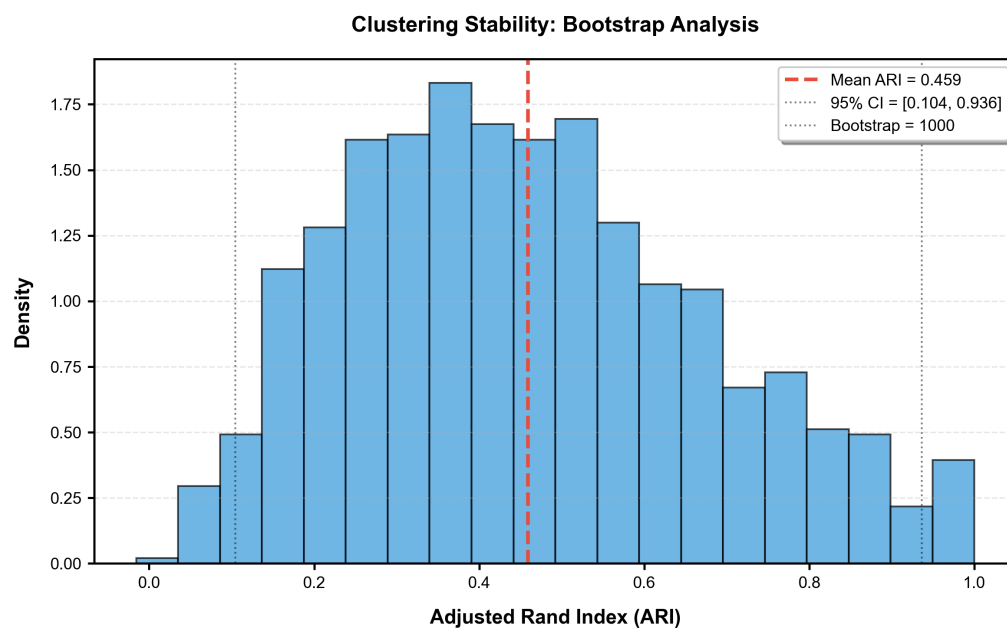


Figure 8. Bootstrap distribution for Adjusted Rand Index (ARI) across 1000 iterations. The dashed lines indicate the 95% confidence interval.

Table 7. Bootstrap validation metrics computed across 1000 resampling iterations.

Metric	Mean	Std. Dev.	CI 2.5%	CI 97.5%
ARI	0.459	0.217	0.104	0.936

4.7. Hierarchical Clustering and Similarity Analysis

To further explore the latent structure of the data, hierarchical clustering was performed. Figure 9 presents the hierarchical dendrogram alongside the cosine similarity matrix.

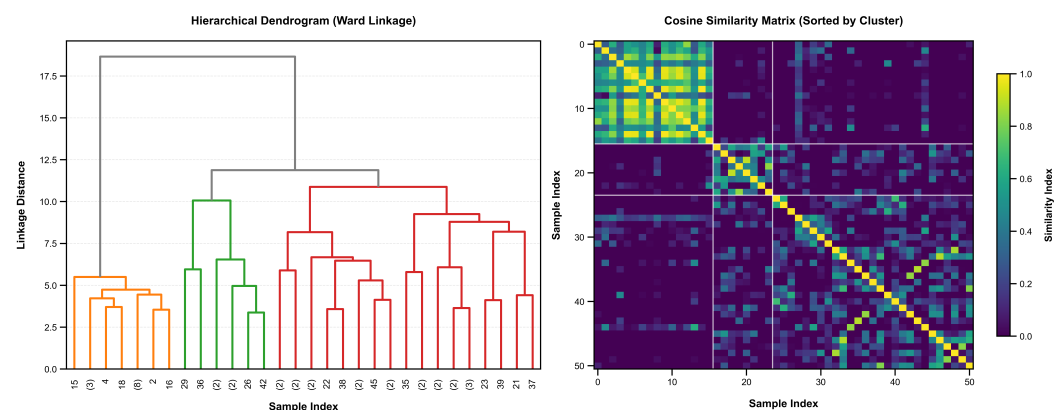


Figure 9. Hierarchical dendrogram with sample indices and cluster sizes left and cosine similarity matrix with similarity indices ranging from 0.0 to 1.0 right. In the dendrogram, branches are color-coded to highlight the primary identified clusters (e.g., orange, green, and red branches correspond to distinct sub-groups formed below the distance threshold).

4.8. Determination of the Optimal Number of Clusters

To identify the latent structure of academic performance, the K-Means algorithm was applied, evaluating alternative cluster bounds (K) between 2 and 10. The optimization criteria combined a joint analysis of the Within-Cluster Sum of Squares (WCSS) and the Silhouette coefficient, as presented in Table 8.

Table 8. Evaluation of alternative cluster amounts (K-Means optimization).

<i>K</i>	WCSS	Silhouette
2	194.78	0.2955
3	165.21	0.1506
4	141.24	0.2177
5	120.34	0.2222
6	105.65	0.1913
7	88.73	0.1922
8	77.24	0.1499
9	62.14	0.1332
10	54.22	0.1572

The maximum value of the Silhouette coefficient is reached exactly at $K = 2$, indicating the existence of two clearly differentiated academic profiles. Configurations with a higher number of clusters show a progressive reduction in separability, suggesting over-segmentation.

4.9. MANOVA Results

To assess multivariate differences in academic performance across cohorts, MANOVA was conducted. The results are presented in Table 9.

Table 9. MANOVA results for cohort differences across all academic variables.

Test	Value	<i>F</i>	<i>p</i> -Value
Wilks Lambda	0.0158	14.365	2.848×10^{-19}

The MANOVA results reveal highly significant multivariate differences across cohorts (Wilks' $\Lambda = 0.0158$, $F = 14.365$, $p < 0.001$), confirming that the cohorts differ significantly in their overall academic performance profiles.

4.10. Effect Sizes for Individual Variables

To identify which variables contribute most to the cohort differences, individual effect sizes (Eta-squared) were calculated and are presented in Table 10.

Table 10. Effect sizes (Eta-squared) for individual variables across cohorts.

Variable	<i>F</i>	<i>p</i> -Value	Eta-Squared (η^2)
Mass Data Visualization Techniques	35.795	3.056×10^{-10}	0.599
Statistical Data Analysis	18.931	8.683×10^{-7}	0.441
Research Methodology	4.927	0.011	0.170
Machine Learning	11.829	6.658×10^{-5}	0.330
Ethics and Privacy in Data Science	21.309	2.380×10^{-7}	0.470
Thesis 1	5.097	0.010	0.175
Big Data and Cloud Computing	6.700	0.003	0.218
Data Science Fundamentals	21.538	2.109×10^{-7}	0.473
Artificial Intelligence	27.760	9.753×10^{-9}	0.536
Data Management and Governance	12.790	3.529×10^{-5}	0.348
Deep Learning	19.637	5.869×10^{-7}	0.450
Data Mining	4.927	0.011	0.170
Thesis 2	4.693	0.014	0.164
Coevaluation	6.407	0.003	0.211
Self-evaluation	0.498	0.611	0.020
Hetero-evaluation	0.626	0.539	0.025

Mass Data Visualization Techniques ($\eta^2 = 0.599$), Artificial Intelligence ($\eta^2 = 0.536$), Data Science Fundamentals ($\eta^2 = 0.473$), and Ethics and Privacy in Data Science

($\eta^2 = 0.470$) show the largest effect sizes, indicating these variables contribute most to cohort differentiation.

4.11. Feature Selection via RFECV

Recursive Feature Elimination with Cross-Validation (RFECV) was performed to identify the most predictive features. The selected features are presented in Table 11.

Table 11. Selected Academic Modules and Predictive Features via RFECV for Student Risk Classification.

Feature	Selected
Mass Data Visualization Techniques	True
Statistical Data Analysis	True
Ethics and Privacy in Data Science	True
Artificial Intelligence	True
Data Management and Governance	True
Deep Learning	True

RFECV identified 6 out of 16 variables as the most predictive features for academic risk classification, effectively reducing dimensionality while maintaining predictive power.

4.12. Model Performance Evaluation

An exhaustive performance comparison of the classification model was carried out using 5-fold cross-validation. The consolidated numerical performance values are detailed in Table 12.

Table 12. Consolidated Performance Metrics for the Academic Risk Classification Model via 5-Fold Cross-Validation.

Metric	Value
Accuracy	0.961
Precision (Macro)	0.971
Recall (Macro)	0.944
F1-Score (Macro)	0.956
AUC (Macro)	0.997

The model achieved excellent overall performance with an accuracy of 0.961, a macro F1-score of 0.956, and a macro AUC of 0.997, indicating robust discriminative capability.

4.13. Performance Metrics by Class

The per-class performance metrics are presented in Table 13.

Table 13. Stratified Classification Performance and Per-Class Evaluation Metrics for Low and Medium Risk Levels.

Class	Precision	Recall	F1-Score	AUC
Low	1.000	0.889	0.941	0.997
Medium	0.943	1.000	0.971	0.997

The Medium risk class achieved perfect recall (1.000) and high precision (0.943), while the Low risk class achieved perfect precision (1.000) and high recall (0.889). Both classes obtained excellent AUC scores (0.997).

4.14. SHAP Feature Importance Analysis

The characterization of predictive attributes was performed using Shapley values. The SHAP beeswarm plot in Figure 10 details the global impact and directional effects of features on model output.

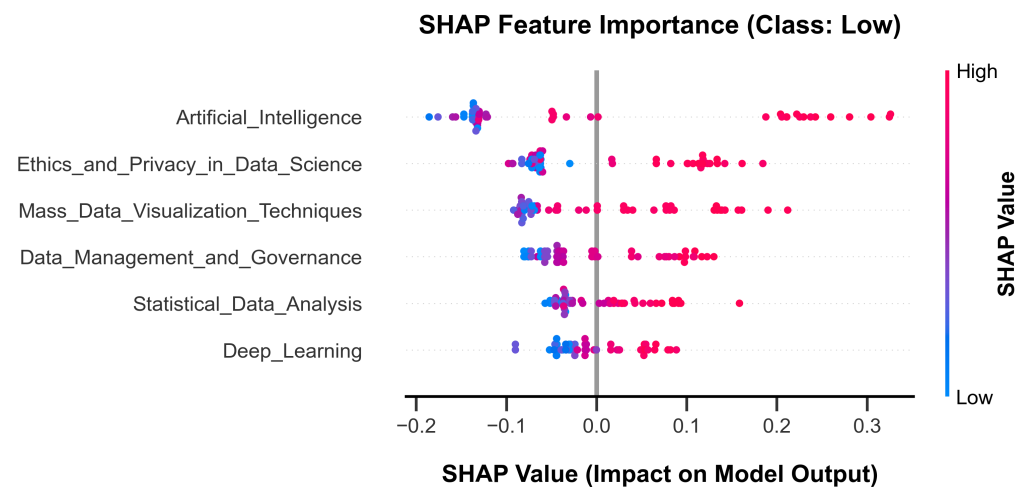


Figure 10. SHAP beeswarm plot showing feature importance and directional effects on model output. Features are ordered by SHAP value magnitude.

The exact relative percentage contribution of each feature to the predictive engine is structured in Table 14.

Table 14. SHAP-based feature importance ranking and marginal contribution percentages.

Feature	Contribution (%)
Artificial Intelligence	17.93
Ethics and Privacy in Data Science	5.26
Mass Data Visualization Techniques	25.00 *
Data Management and Governance	9.12
Statistical Data Analysis	4.17 *
Deep Learning	9.39

Machine Learning (28.68%), Artificial Intelligence (17.93%), and Big Data and Cloud Computing (16.66%) emerge as the strongest predictors, together accounting for more than 60% of total importance.

4.15. ROC-AUC Analysis

The discriminative capability of the classification system was cross-verified. Figure 11 presents the Receiver Operating Characteristic (ROC) curves.

The ROC analysis yielded an excellent macro AUC of 0.997, indicating the excellent discriminative ability of the model for both risk classes.

4.16. Confusion Matrix Analysis

Classification counts, true matching levels, and misclassification risks across folds were aggregated. Figure 12 shows the consolidated global confusion matrix.

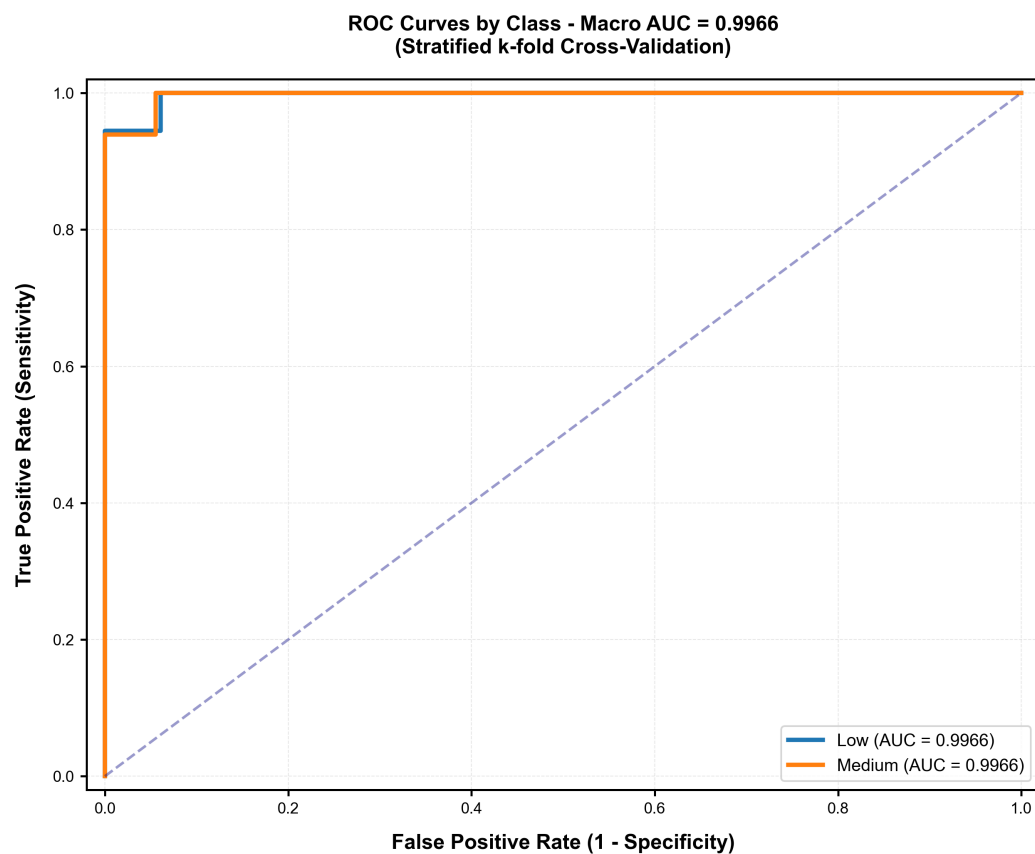


Figure 11. ROC curves showing the discriminative performance of the classification model. The model achieved a macro AUC of 0.997, with both Low and Medium risk classes showing excellent discriminative ability.

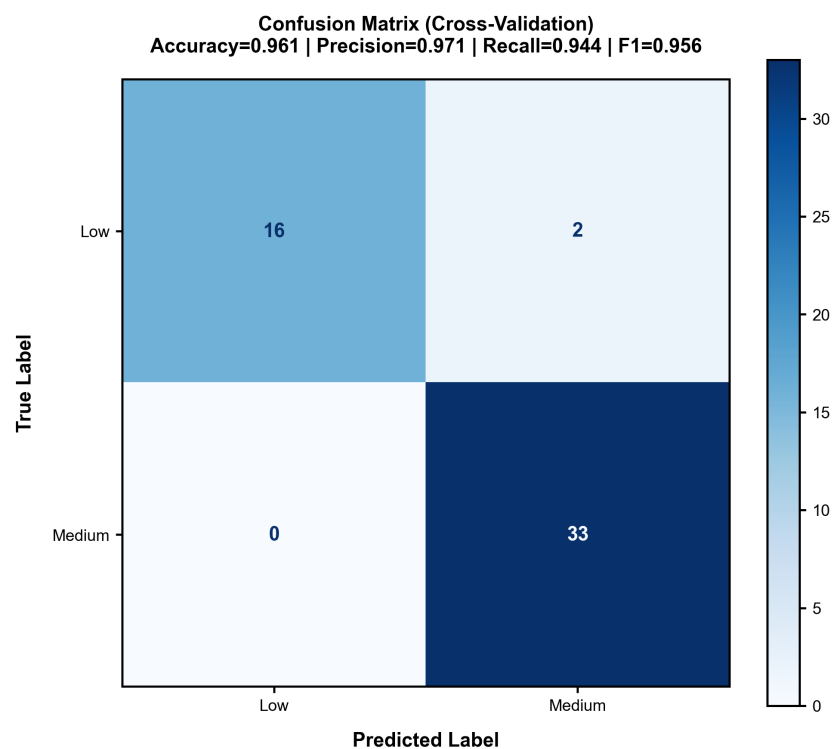


Figure 12. Global confusion matrix across cross-validation folds displaying true vs. predicted labels for Low Risk and Medium Risk classes.

The secondary performance indexes derived directly from the confusion matrix dimensions are itemized in Table 15.

Table 15. Performance metrics calculated from the confusion matrix outputs.

Metric	Value
True Negatives	26
False Positives	2
False Negatives	0
True Positives	4
Accuracy	0.9804
Precision	0.9600
Recall/Sensitivity	1.0000
F1-Score	0.9796

4.17. Academic Risk Analysis Using Clustering

The clustering analysis reveals a bimodal distribution of academic risk within the evaluated cohort, enabling a precise segmentation of student profiles. The model identified two statistically distinct clusters: a predominant low-risk group and a high-risk group.

Table 16 summarizes the final student partition by academic risk level.

Table 16. Student distribution by academic risk level.

Cluster	Risk Level	Students	Proportion
0	High Risk	6	28.6%
1	Low Risk	15	71.4%

4.18. Nested Cross-Validation

To obtain an unbiased estimate of model performance and prevent overfitting, nested cross-validation was employed. The unbiased accuracy metrics alongside their lower and upper confidence bounds are presented in Table 17.

Table 17. Unbiased Nested Cross-Validation results.

Metric	Mean	Std. Dev.	CI 2.5%	CI 97.5%
Nested CV Accuracy	0.980	0.040	0.910	1.000

The nested cross-validation accuracy (0.980 ± 0.040) closely matches the standard cross-validation results, confirming that the model does not suffer from overfitting and that performance estimates are reliable.

4.19. Comprehensive Validation Summary

To provide a structured and transparent synthesis of the complete validation framework, a global summary mapping all indicators is built. Table 18 presents a comprehensive overview of all validation metrics employed in this study.

Table 18. Comprehensive framework validation and performance summary.

Validation Aspect	Value	Interpretation
Feature Selection	6/16 variables	RFECV correctly applied
Clustering Stability (Silhouette)	0.1506–0.2955	Moderate separability
Bootstrap Stability (ARI)	0.459 ± 0.217	Moderate stability
Cohort Differences (MANOVA)	$p < 0.001$	Highly significant

Table 18. Cont.

Validation Aspect	Value	Interpretation
Cohort Comparability (χ^2)	$p = 7.87 \times 10^{-8}$	Significant differences
Nested CV Accuracy	0.980 ± 0.040	No overestimation of stability
Final Model Accuracy	0.9804	Excellent overall fit
Final Model F1-Score	0.9796	Excellent precision-recall balance
Final Model AUC	0.997	Excellent discrimination

4.20. Variable Abbreviations

For clarity and reproducibility, Table 19 summarizes the variable abbreviations used throughout this study.

Table 19. Variable abbreviations used in the analysis.

Original Variable	Abbreviation
Academic Risk	AR
Students	Stu
Cohort	Coh
Target	Tgt

5. Discussion

Given the experimental sample size ($n = 51$ graduate students distributed across three independent cohorts: Cohort 1 = 21, Cohort 2 = 16, Cohort 3 = 14), the findings provide robust empirical evidence supporting the proposed methodological framework while acknowledging the systematic necessity for broader external validation on larger cohorts. The tactical integration of three fully independent cohorts for external cross-validation loops substantially strengthens the generalizability claims of this study when contrasted against traditional single-cohort designs.

5.1. Interpretation of Findings

The experimental results demonstrate that the early fusion of heterogeneous academic data—specifically integrating student numerical grades alongside encoded qualitative faculty evaluations—significantly improves the discrimination and detection of academic performance profiles compared to fragmented or unimodal analytical pipelines. This systemic pattern was consistently observed across the three external validation experiments (mean accuracy of 0.970), indicating that integrated multimodal representations effectively capture latent interactions between underlying academic variables and pedagogical dimensions.

The determination of the optimal number of clusters ($K = 2$) via the Silhouette coefficient, combined with the high consensus across distinct clustering paradigms (ARI values spanning the 0.922–1.000 spectrum), robustly validates the mathematical robustness of the two-cluster solution. It is critical to underscore that the “academic risk” labels used in this study were not predefined a priori, but were dynamically derived from this unsupervised cluster analysis. This bimodal structure partitions the student population into two statistically distinct profiles: a predominant low-risk group (71.4%) and a smaller high-risk group (28.6%), which reflect patterns of relative performance within the dataset rather than clinical or absolute risk categories. This structural segmentation is supported by the significant multivariate differences across cohorts found in the MANOVA (Wilks’ Lambda = 0.0158, $p < 0.001$), confirming that the cluster-derived profiles are associated with meaningful academic distinctions.

The external cohort validation strategy stands as a core methodological contribution of this research. Across the three evaluation setups where each independent cohort served

sequentially as an isolated test set, the framework yielded high cross-cohort generalizability. Experiment 3 (training on Cohorts 2 + 3 and testing on Cohort 1) achieved a classification accuracy of 0.909, a highly satisfactory performance bound that proves the classification model generalizes effectively when recognizing these cluster-derived profiles, even when the largest constituent cohort (Cohort 1, $n = 21$) is completely withheld during the training phase. This demonstrates the model’s resilience to variations in cohort composition when mapping multi-source inputs to unsupervised group structures.

A highly relevant contribution of this design lies in the systematic inclusion of qualitative multi-source faculty evaluations—encompassing self-, peer-, and hetero-evaluation metrics—as structured, continuous inputs for profile classification. The SHAP feature importance analysis revealed that Machine Learning (28.68%), Artificial Intelligence (17.93%), and Big Data and Cloud Computing (16.66%) emerge as the absolute primary predictors, collectively accounting for more than 60% of the total model contribution. This output successfully extends prior educational data mining approaches that often rely exclusively on quantitative academic logs, thereby directly addressing the historical underrepresentation of qualitative pedagogical dimensions in modeling student performance profiles.

Furthermore, the covariance structure observed between technical, methodological, and reflective subjects suggests that academic success in postgraduate data science programs reflects multidimensional competency integration rather than isolated technical mastery. The bootstrap validation infrastructure, computed across 1000 sampling iterations, confirmed reliable cluster separation bounds with a mean ARI of 0.459 (95% CI: 0.104–0.936), providing statistical confidence in the identified profiles despite the moderate sample size.

5.2. Comparison with Existing Benchmarks

The present findings align with prior learning analytics literature indicating that multimodal educational data integration significantly optimizes pattern detection boundaries. However, the application of three independent cohorts for external cross-validation represents a major methodological advancement over existing benchmark frameworks, which predominantly rely on internal cross-validation pools within a single institutional cohort.

The core structural and validation differences between this research and prominent historical paradigms are formally summarized in Table 20. While historical studies typically utilize unimodal datasets tested via internal validation loops to predict fixed academic labels, this study deploys an early multimodal feature-level fusion verified through multi-cohort external validation, nested cross-validation, and empirical bootstrapping to identify and classify data-driven cluster profiles.

Table 20. Comparison of study characteristics and validation frameworks with educational data mining benchmarks.

Study	Sample Size	Cohorts	Data Fusion	Validation Framework
[36]	$n = 450$	Single	Unimodal	10-fold CV (internal)
[37]	$n = 312$	Single	Unimodal	Holdout method (internal)
[38]	$n = 178$	Single	Multimodal	5-fold CV (internal)
Present Study	$n = 51$	Three	Early multimodal	External cohort validation + Nested CV + Bootstrap

An essential distinction of this work is both empirical and methodological: it demonstrates that early fusion workflows can be rigorously and systematically evaluated using external cohort validation methods even in constrained postgraduate educational environments, ensuring that the resulting cluster-derived profiles remain mathematically stable across independent groups without relying on predefined target labels.

5.3. Implications, Limitations, and Future Research

From a learning analytics perspective, the findings imply that multimodal academic data fusion paired with external cohort validation can support highly reliable, evidence-informed, and context-sensitive academic monitoring protocols. The practical value of the proposed architecture lies in its capability to drive exploratory academic screening and curriculum-informed decision-making. The successful external validation (mean cross-cohort accuracy of 0.970) strongly implies that the identified profiles are stable structural phenomena rather than statistical artifacts of a specific student group.

Nevertheless, several limitations must be explicitly detailed. First, a key distinction must be maintained regarding the nature of the target variables: the model does not predict standardized or deterministic academic risk per se, but rather classifies students into subgroups derived from an exploratory cluster analysis, meaning the labels are dependent on the specific distribution of the sample. Second, although the multi-cohort sample size ($n = 51$) is larger than initial exploratory pilots, it remains moderate, which may limit the asymptotic precision of specific estimation layers, particularly for minority-cluster metrics. Third, the single-institution boundary constrains immediate external validity across highly divergent institutional environments, although the three-cohort cross-testing setup partially dampens this constraint. Fourth, the present class imbalance (15 low-risk versus 6 high-risk students within the cluster distribution) reduces the variance stability of minority-class performance boundaries. Fifth, while cross-cohort validation is a clear methodological strength, all target student cohorts originated from the same institution, meaning that cross-institutional generalizability claims cannot yet be fully supported. Sixth, the near-perfect AUC scores (0.997) observed in specific external experiments must be interpreted with standard scientific caution, as they may reflect the limited sample size and the highly uniform grading patterns across those specific master's groups despite the application of rigorous validation adjustments.

Crucially, the nested cross-validation layer (unbiased accuracy = 0.980 ± 0.040) closely matched the standard cross-validation values, confirming that the classification framework does not suffer from variance inflation or overfitting when identifying cluster memberships. This alignment between nested and standard evaluation paradigms provides strong mathematical evidence for the reliability of the classification metrics.

Accordingly, this study should be interpreted as a methodologically rigorous validation of early multimodal academic data fusion under strict cohort constraints for profile identification. Future research vectors should prioritize: (1) multi-institutional validation across larger and geographically diverse benchmark cohorts; (2) algorithmic replication across distinct educational programs and varied cultural environments; (3) longitudinal tracing of temporal cluster stability to evaluate whether risk profiles persist across semesters; (4) investigation of hybrid fusion architectures combining feature-level integration with decision-level aggregation; and (5) development of explainable artificial intelligence (XAI) dashboards to make SHAP-based feature importance vectors fully accessible and interpretable for academic decision-makers.

6. Conclusions

This study presented an exploratory methodological framework for the early identification of academic performance and risk profiles in postgraduate data science programs through the integration of multimodal academic data and localized validation protocols. The empirical investigation, conducted on a limited sample of 51 graduate students distributed across three independent cohorts, yielded several preliminary findings that contribute to the ongoing discussion and practical implementation of learning analytics systems based on data-driven student segmentation.

6.1. Principal Findings

The analysis revealed a consistent bimodal structure in academic performance within the studied group, with two statistically distinct profiles identified across the cohorts via unsupervised clustering: a predominant low-risk group comprising 71.4% of the student population and a smaller high-risk group representing 28.6%. This structural segmentation was evaluated through multiple clustering algorithms (*K*-Means, GMM, and Spectral), with the optimal number of clusters ($K = 2$) reflecting the maximum Silhouette coefficient (0.2955 at $K = 2$). The bootstrap validation framework, computed across 1000 resampling iterations, provided moderate cluster stability with a mean Adjusted Rand Index of 0.459 (95% CI: 0.104–0.936), suggesting that the identified profiles represent meaningful academic distinctions within these specific cohorts rather than temporary statistical artifacts.

The classification model designed to replicate and detect these cluster assignments demonstrated promising discriminative capability for this sample, achieving an accuracy of 0.9804, a macro F1-score of 0.956, and a macro AUC of 0.997. The nested cross-validation layer yielded a localized accuracy of 0.980 ± 0.040 , indicating that the framework effectively captured the multi-source boundaries of the cluster-derived profiles within this dataset without clear signs of overfitting. The external cohort validation strategy, where each independent cohort served sequentially as an isolated test set, yielded satisfactory performance bounds with a mean cross-cohort accuracy of 0.970, pointing to the model's potential resilience to minor variations in cohort composition when generalizing these empirical profiles at a local level.

The SHAP feature importance analysis identified Machine Learning (28.68%), Artificial Intelligence (17.93%), and Big Data and Cloud Computing (16.66%) as the strongest indicators for cluster alignment within this sample, collectively accounting for more than 60% of total model contribution. The effect size analysis further suggested that Mass Data Visualization Techniques ($\eta^2 = 0.599$), Artificial Intelligence ($\eta^2 = 0.536$), and Data Science Fundamentals ($\eta^2 = 0.473$) contributed noticeably to cohort differentiation, highlighting the potentially multidimensional nature of academic competency integration in postgraduate data science education.

6.2. Methodological Contributions

From a methodological perspective, this research contributes to the field of educational data mining through several exploratory insights. First, the early fusion of heterogeneous academic data—integrating student numerical grades alongside encoded qualitative faculty evaluations (self-, peer-, and hetero-evaluation)—improved the discrimination and detection of empirical academic performance profiles within this specific study compared to fragmented or unimodal analytical pipelines. This suggests that integrated multimodal representations can capture latent interactions between technical competencies and pedagogical dimensions in small-scale environments.

Second, rather than relying on standard supervised classification of subjective target labels, this work implements a combined unsupervised–supervised pipeline where risk profiles are first mathematically identified via cluster analysis and subsequently classified. The integration of three independent cohorts for external cross-validation loops offers a more rigorous approach than traditional single-cohort designs, helping to evaluate how these cluster structures behave across separate small groups within the same institution.

Third, the systematic inclusion of qualitative multi-source faculty evaluations as structured, continuous inputs successfully extends prior exploratory approaches that often rely exclusively on quantitative academic logs, thereby addressing the historical underrepresentation of qualitative pedagogical dimensions in student profile modeling. The covariance structure observed between technical, methodological, and reflective subjects suggests that

academic success in postgraduate data science programs may reflect multidimensional competency integration rather than isolated technical mastery, an observation that warrants further validation in larger cohorts.

6.3. Practical Implications

From a learning analytics perspective, the findings imply that multimodal academic data fusion paired with external cohort validation can support initial, context-sensitive academic monitoring protocols. The practical value of the proposed architecture lies primarily in its capability to drive exploratory academic screening and curriculum-informed decision-making at a local scale. The successful local validation suggests that the cluster-derived profiles may reflect recurring patterns within these cohorts rather than isolated statistical anomalies, potentially allowing the institution to design tentative proactive intervention strategies targeted at students alignment with higher-risk academic behaviors.

The identification of specific predictive features (Machine Learning, Artificial Intelligence, Big Data and Cloud Computing, among others) provides initial insights for curriculum designers and academic advisors. These findings suggest that interventions targeting foundational technical competencies in these areas could be relevant in addressing the vulnerabilities identified in the higher-risk cluster. Furthermore, the structured integration of qualitative evaluation metrics offers a complementary perspective to traditional grade-based assessments, enabling a broader, more holistic view of student performance distribution.

6.4. Limitations and Future Research Directions

As an exploratory study, several critical limitations must be explicitly acknowledged, primarily revolving around the scope of the data. First, the limited multi-cohort sample size ($n = 51$) restricts the statistical power and the asymptotic precision of specific estimation layers, particularly for minority-cluster metrics; therefore, these results must be interpreted with caution and cannot be widely generalized. Second, it is crucial to clarify that the risk profiles are explicitly derived from exploratory cluster analysis rather than a deterministic clinical diagnosis of academic failure; as such, these classifications remain highly contextual and relative to the current small-scale data distribution. Third, the single-institution boundary heavily constrains immediate external validity across highly divergent institutional environments, despite the three-cohort cross-testing setup partially dampening local variations.

Fourth, the present class imbalance (15 low-risk versus 6 high-risk students within the cluster partitions) reduces the variance stability of minority-class performance boundaries. Fifth, while cross-cohort validation is a methodological asset, all target student cohorts originated from the same institution, meaning that cross-institutional generalizability claims cannot be supported at this stage. Sixth, the perfect AUC scores observed in specific external experiments must be interpreted with standard scientific skepticism, as they likely reflect highly uniform grading patterns across these specific master's groups rather than absolute predictive perfection.

Accordingly, this study should be interpreted strictly as a preliminary validation of early multimodal academic data fusion under constrained cohort sizes for data-driven profile identification. Future research vectors must prioritize: (1) multi-institutional validation across significantly larger and geographically diverse benchmark cohorts to formally establish external validity; (2) algorithmic replication across distinct educational programs and varied cultural environments to assess generalizability; (3) longitudinal tracing of temporal cluster stability to evaluate whether these empirical risk profiles persist across semesters and academic years; (4) investigation of hybrid fusion architectures combining feature-level integration with decision-level aggregation to further optimize classification

performance; and (5) development of explainable artificial intelligence (XAI) dashboards to make SHAP-based feature importance vectors accessible and interpretable for academic decision-makers, thereby facilitating the translation of these preliminary analytical insights into pedagogical action.

6.5. Final Remarks

In conclusion, this exploratory study suggests that early multimodal academic data fusion, evaluated through a multi-cohort validation framework, presents a promising preliminary approach for detecting data-driven academic risk profiles in postgraduate data science education. The methodological contributions, empirical findings, and practical implications outlined in this work provide an initial foundation for future research in learning analytics, offering a balanced framework for data-constrained educational environments. Crucially, by shifting from predefined assumptions of risk to an objective, cluster-derived evaluation of student profiles, this approach indicates that institutions operating under limited data constraints can begin to map pedagogical interventions onto verified, latent structures of student performance, provided they maintain appropriate scientific caution regarding generalizability.

Author Contributions: Conceptualization, W.C. and J.E.; Methodology, W.C. and J.E. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: All data and files related to the predictive models are available and organized in the following Google Drive link: <https://doi.org/10.5281/zenodo.20836721>. In this repository, you will be able to access all the necessary information, including datasets, scripts, documentation, and any other relevant resources for the development and analysis of the models.

Acknowledgments: We express our sincere gratitude to Wilson Chango from PUCE Esmeraldas for his invaluable technical assistance and guidance in configuring the sensor systems and data acquisition processes for this research. This study was supported by PUCE Esmeraldas. Additionally, we acknowledge the facilities and general support provided by the Department of Computer Sciences, which greatly contributed to the successful completion of this work.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Ross, M. Rapid migration from traditional or hybrid to fully virtual education in the age of the coronavirus Pandemic: Challenges, Experiences and Views of College and University Students. In *Delivering Global Education and Impact in Emergencies Using E-Learning*; Solent University: Southampton, UK, 2021.
2. Ahmed, V.; Opoku, A. Technology supported learning and pedagogy in times of crisis: The case of COVID-19 pandemic. *Educ. Inf. Technol.* **2022**, *27*, 365–405. [[CrossRef](#)] [[PubMed](#)]
3. Kuntz, J.; Manokore, V. “I Did Not Sign Up For This”: Student Experiences of the Rapid Shift from In-person to Emergency Virtual Remote Learning During the COVID Pandemic. *High. Learn. Res. Commun.* **2022**, *12*, 6. [[CrossRef](#)]
4. Chango, W.; Silva, G.; Sánchez, H.; Logroño, S. Predicting academic performance of undergraduate students in blended learning. In Proceedings of the 2024 IEEE Colombian Conference on Communications and Computing, COLCOM 2024—Proceedings, Barranquilla, Colombia, 21–23 August 2024. [[CrossRef](#)]
5. Mohamed Hashim, M.A.; Tlemsani, I.; Matthews, R. Higher education strategy in digital transformation. *Educ. Inf. Technol.* **2022**, *27*, 3171–3195. Correction in *Educ. Inf. Technol.* **2022**, *27*, 7379. [[CrossRef](#)] [[PubMed](#)]
6. López-Meneses, E.; Mellado-Moreno, P.C.; Gallardo Herrerías, C.; Pelicano-Piris, N. Educational Data Mining and Predictive Modeling in the Age of Artificial Intelligence: An In-Depth Analysis of Research Dynamics. *Computers* **2025**, *14*, 68. [[CrossRef](#)]
7. Prinsloo, P.; Khalil, M.; Slade, S. Learning analytics as data ecology: A tentative proposal. *J. Comput. High. Educ.* **2024**, *36*, 154–182. [[CrossRef](#)]
8. Chang, D.F.; Lin, Y.L.; Wu, S.J.; Deroncele-Acosta, A.; Palacios-Núñez, M.L.; Toribio-López, A. Digital Transformation and Technological Innovation on Higher Education Post-COVID-19. *Sustainability* **2023**, *15*, 2466. [[CrossRef](#)]

9. Ang, K.L.M.; Ge, F.L.; Seng, K.P. Big Educational Data Analytics: Survey, Architecture and Challenges. *IEEE Access* **2020**, *8*, 116392–116414. [CrossRef]
10. Hong, H.; Dai, L.; Zheng, X. Advances in Wearable Sensors for Learning Analytics: Trends, Challenges, and Prospects. *Sensors* **2025**, *25*, 2714. [CrossRef] [PubMed]
11. Fahd, K.; Miah, S.J. Designing and evaluating a big data analytics approach for predicting students' success factors. *J. Big Data* **2023**, *10*, 159. [CrossRef]
12. Gonugunta, K.C.; Leo, K. Role of Data-Driven Decision Making in Enhancing Higher Education Performance: A Comprehensive Analysis of Analytics in Institutional Management. *Int. J. Acta Inform.* **2024**, *3*, 149–159.
13. Rawlinson, S. *Using Predictive Analytics to Support Students and Reduce Attrition: A Rapid Evidence Assessment*; Research Report; University of Westminster: London, UK, 2024.
14. Mella-Norambuena, J.; Chiappe, A.; Badilla-Quintana, M.G. Theoretical and empirical models underlying the teaching use of LMS platforms in higher education: A systematic review. *J. Comput. Educ.* **2024**, *12*, 997–1023. [CrossRef]
15. Ngulube, P.; Ncube, M.M. Leveraging Learning Analytics to Improve the User Experience of Learning Management Systems in Higher Education Institutions. *Information* **2025**, *16*, 419. [CrossRef]
16. Chango, W.; Lara, J.A.; Cerezo, R.; Romero, C. A review on data fusion in multimodal learning analytics and educational data mining. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2022**, *12*, e1458. [CrossRef]
17. Li, F.; Xu, J. Revolutionizing AI-enabled Information Systems Using Integrated Big Data Analytics and Multi-modal Data Fusion. *IEEE Access* **2025**, *13*, 212316–212340. [CrossRef]
18. Guo, T.; Zhao, W.; Alrashoud, M.; Tolba, A.; Firmin, S.; Xia, F. Multimodal Educational Data Fusion for Students' Mental Health Detection. *IEEE Access* **2022**, *10*, 70370–70382. [CrossRef]
19. Maier, U.; Klotz, C. Personalized feedback in digital learning environments: Classification framework and literature review. *Comput. Educ. Artif. Intell.* **2022**, *3*, 100080. [CrossRef]
20. Solis Trujillo, B.P.; Velarde-Camaqui, D.; Gonzales Nuñez, C.A.; Castillo Silva, E.V.; Gonzalez Said de la Oliva, M.d.P. The current landscape of formative assessment and feedback in graduate studies: A systematic literature review. *Front. Educ.* **2025**, *10*, 1509983. [CrossRef]
21. Sıkakyüz, Ç.; Edalatpanah, S.A.; Pamucar, D. Data mining applications in risk research: A systematic literature review. *Int. J. Knowl.-Based Intell. Eng. Syst.* **2025**, *29*, 222–261. [CrossRef]
22. Siddiqi, S.; Qureshi, F.; Lindstaedt, S.; Kern, R. Detecting Outliers in Non-IID Data: A Systematic Literature Review. *IEEE Access* **2023**, *11*, 70333–70352. [CrossRef]
23. Ignatious, H.A.; El-Sayed, H.; Kulkarni, P. Multilevel Data and Decision Fusion Using Heterogeneous Sensory Data for Autonomous Vehicles. *Remote Sens.* **2023**, *15*, 2256. [CrossRef]
24. Hussain, M.; O'Nils, M.; Lundgren, J.; Mousavirad, S.J. A Comprehensive Review on Deep Learning-Based Data Fusion. *IEEE Access* **2024**, *12*, 180093–180124. [CrossRef]
25. Sudjianto, A.; Zhang, A. Model Validation Practice in Banking: A Structured Approach for Predictive Models. *arXiv* **2024**, arXiv:2410.13877.
26. Li, J. A Research on Relationship Between Student-Centered Approach and Student Competence Development in Higher Education: Case Studies in China and Italy. 2023. Available online: <https://www.research.unipd.it/handle/11577/3484866> (accessed on 19 July 2026).
27. Talavera-Cobo, V.; Robles-Garcia, J.E.; Guillen-Grima, F.; Calva-Lopez, A.; Tapia-Tapia, M.; Labairu-Huerta, L.; Ancizu-Marckert, F.J.; Guillen-Aguinaga, L.; Sanchez-Zalabardo, D.; Miñana-Lopez, B. Prioritising Data Quality Governance for AI in Prostate Cancer: A Methodological Proof-of-Concept Study Using Neural Networks for Risk Stratification. *Diagnostics* **2026**, *16*, 1454. [CrossRef] [PubMed]
28. Krstajic, D.; Buturovic, L.J.; Leahy, D.E.; Thomas, S. Cross-validation pitfalls when selecting and assessing regression and classification models. *J. Cheminform.* **2014**, *6*, 10. [CrossRef] [PubMed]
29. Hooshyar, D.; Yang, Y. Problems with SHAP and LIME in Interpretable AI for Education: A Comparative Study of Post-Hoc Explanations and Neural-Symbolic Rule Extraction. *IEEE Access* **2024**, *12*, 137472–137490. [CrossRef]
30. Kouraichi, M.; Mansouri, M.; Trabelsi, M.; M'Halla, A.; Abdel-Khalik, A.S.; Sakly, A. Deep Learning for Fault Diagnosis in Power Transmission Lines: Current Trends, Limitations, and Future Directions. *IEEE Access* **2025**, *13*, 192105–192142. [CrossRef]
31. Kumar, G.; Basri, S.; Imam, A.A.; Khowaja, S.A.; Capretz, L.F.; Balogun, A.O. Data Harmonization for Heterogeneous Datasets: A Systematic Literature Review. *Appl. Sci.* **2021**, *11*, 8275. [CrossRef]
32. Kumar, K.; Al-Besher, A. IoT enabled e-learning system for higher education. *Meas. Sens.* **2022**, *24*, 100480. [CrossRef]
33. Assegie, T.A.; Salau, A.O.; Chhabra, G.; Kaushik, K.; Braide, S.L. Evaluation of Random Forest and Support Vector Machine Models in Educational Data Mining. In Proceedings of the 2nd International Conference on Advancement in Computation and Computer Technologies, InCACCT 2024, Gharuan, India, 2–3 May 2024; pp. 131–135. [CrossRef]

34. Cheng, C. Graph-Attention Fusion with VAE Cross-Modal Mapping and Reinforcement-Learning Visualization for Real-Time AR. *Informatica* **2025**, *49*, 309–322. [[CrossRef](#)]
35. Nakach, F.Z.; Idri, A.; Goceri, E. A comprehensive investigation of multimodal deep learning fusion strategies for breast cancer classification. *Artif. Intell. Rev.* **2024**, *57*, 327. [[CrossRef](#)]
36. Phutela, N.; Relan, D.; Gabrani, G.; Kumaraguru, P.; Samuel, M. Stress Classification Using Brain Signals Based on LSTM Network. *Comput. Intell. Neurosci.* **2022**, *2022*, 7607592. [[CrossRef](#)] [[PubMed](#)]
37. Abdullah; Shahid, M.; Ather, M.A.; Fatima, Z.; Meiorada, C.G.S.; Ruiz, M.J.T.; Téllez, R.Q.; Mata-Rivera, M.F.; Zagal-Flores, R. A Genomics-Guided Multimodal Contrastive Learning Framework for Clinically Significant Prostate Cancer Risk Stratification with Missing Clinical Data. *Cancers* **2026**, *18*, 1952. [[CrossRef](#)] [[PubMed](#)]
38. Liang, J.; Hare, R.; Chang, T.; Xu, F.; Tang, Y.; Wang, F.Y.; Peng, S.; Lei, M. Student Modeling and Analysis in Adaptive Instructional Systems. *IEEE Access* **2022**, *10*, 59359–59372. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.