ISSN: 2096-3246
Volume 58, Issue 02, 2026

Predicting Student Performance with Principal Component Analysis (PCA)

M. Rama Moorthi[1], Dr. R. Kalaimagal[2], Dr. S. Mary Vennila[3]

[1]Research Scholar, Government Arts College, Nanadanam, Chennai.

[2]Associate Professor, Department of Computer Science, Government Arts College, Nandanam, Chennai.

[3]Associate Professor, Department of Computer Science, Presidency College, Chennai.

Abstract

Predicting student academic success from big, high-dimensional institutional datasets is hampered by substantial multicollinearity among behavioral, engagement, and continuous-assessment factors, which can weaken the stability and interpretability of downstream classifiers. This work studies Principal Component Analysis (PCA) as a dimensionality-reduction technique for multi-class student performance prediction on the dataset (15,000 students, one universities, six academic years). After removing cumulative-grade-point-average (CGPA)-derived features to prevent target leakage, 38 leakage-aware predictors were standardized and one-hot encoded into a 78-dimensional feature space. PCA was applied to this space, and 46 principle components were determined to jointly explain 95% of total variance — a 41.0% reduction in dimensionality. Seven classifiers (Naive Bayes, Logistic Regression, Decision Tree, k-Nearest Neighbors, Support Vector Machine, Random Forest, and XGBoost) were trained and evaluated, under an identical held-out test protocol, on both the original 78-dimensional feature space and the 46-dimensional PCA-reduced representation. The PCA-reduced Logistic Regression model produced the best overall result (87.07% accuracy, 87.55% macro F1-score, macro ROC-AUC = 0.981), barely beating its own non-reduced counterpart. More significantly, Naive Bayes improved from 77.13% to 83.63% accuracy following PCA — a 6.50-percentage-point gain attributable to PCA's de-correlation of the transformed feature space, which better meets Naive Bayes conditional-independence premise. A similar, smaller gain was observed for the Decision Tree. Loading analysis and a two-dimensional PCA projection further demonstrate that the first main component is dominated by continuous-assessment and engagement indicators (Project_Score, Attendance_Pct, Assignment_Score) and visually distinguishes the four performance classes along a single axis. These findings show that PCA provides a viable and understandable method for reducing the dimensionality of the data in the prediction of student performance with comparable or better accuracy at reduced dimensions, with the largest improvements found in classifiers whose assumptions are vulnerable to feature correlation..

Keywords

Principal Component Analysis, Dimensionality Reduction, Feature Engineering, Machine Learning, Educational Data Mining, Explained Variance, Multicollinearity.

1. Introduction

Institutional student-information systems are capturing dozens of per-student attributes on behavior, engagement, socio-economic status, and continuous assessment, spurred by the promise of richer feature sets for early-warning prediction of academic achievement. However, many of these attributes are strongly intercorrelated; for example, attendance, assignment timeliness, quiz performance, and learning-management-system engagement all tend to move together for a given student. This property is called multicollinearity, and it can inflate the effective dimensionality of a predictive model without adding proportionate genuine information, destabilize the coefficient estimates of linear classifiers, and violate the conditional-independence assumptions underlying probabilistic classifiers such as Naive Bayes [1], [4].

Principal Component Analysis (PCA) is a classical unsupervised dimensionality-reduction technique that transforms a set of potentially correlated variables into a smaller set of linearly uncorrelated variables, called principal components, ordered such that the first component captures the maximum possible variance in the data and each subsequent component captures the maximum remaining variance subject to being orthogonal to the preceding components [9], [17]. In recent years, a number of research in educational data mining have employed PCA as either a pre-processing step to improve the accuracy of classifiers and training efficiency [4], [8], [17] or a tool for visualization and comprehension of the latent structure of student-performance datasets [3], [13]. Sabri et al. coupled PCA with a Support Vector Machine to predict continuous student performance using weekly behavioral and self-efficacy data, showing that PCA-derived composite scores can serve as effective, lower-dimensional predictors of academic result [8]. Al-tameemi employed PCA with K-Means clustering as an unsupervised pre-processing step before the supervised multi-class prediction and reported enhanced visualization and downstream classification efficacy [13]. Wusu, Manuel and Olabanjo examined the particular effect of PCA on Radial Basis Function Neural Network for the prediction of secondary school performance. They discovered that PCA-transformed inputs impacted the sensitivity-specificity trade-off of the resultant classifier [4].

Yet, even with this growing body of work, few studies have systematically compared a large panel of classifiers with and without PCA under an equal leakage-aware evaluation process, or specifically quantified which specific classifiers benefit most from the decorrelation property of PCA and why. This study directly fills that hole. This study contributes the following from the dataset using restricted leakage-aware, non-grade predictors.

• A thorough comparison of seven classifiers (Naive Bayes, Logistic Regression, Decision Tree, k-Nearest Neighbors, SVM, Random Forest, XGBoost), each evaluated with and without PCA-based dimensionality reduction, using the same train/test process.

• A quantified explained variance analysis of minimum number of principal components to

maintain 90%, 95% and 99% of total variance in a 78 dimensional encoded feature space.

• A loading interpretation of the leading principal components, with 2D PCA projection and biplot, to show which original predictors dominate the directions of maximum variance, and how they connect to the four ordinal performance classes.

•A training-time analysis assessing the computational trade-offs of PCA-based dimensionality reduction across algorithmically diverse classifier families.

The rest of this paper is organized as follows. Section 2 evaluates the research on PCA and hybrid machine learning techniques for student performance prediction from 2020 to 2025. In Section 3 we describe the dataset, preparation, PCA algorithm and evaluation protocol. Experimental data are presented and discussed in Section 4. The paper is concluded in Section 5 and future work is discussed.

2. Related Work

As more and more features are being added to institutional datasets, educational data mining research has been using dimensionality-reduction strategies to cope with this situation. Sokkhey and Okazaki [1] proposed a hybrid architecture for predicting academic achievement that incorporated numerous classical classifiers. They reported that careful handling of features and the integration of the model enhanced accuracy over separate baselines. Albreiki, Zaki, and Alashwal [2] in their systematic literature review on machine learning based student performance prediction, concluded that the classification accuracy is highly dependent on the feature selection and dimensionality reduction strategy, hence, further investigation of methods such as PCA is justified. Al-Zawqari, Peumans, and Vandersteen developed a flexible feature-selection technique for the prediction of online course success, demonstrating that decreasing the predictor set can lead to both better accuracy and interpretability relative to the complete raw feature set [3].

In several recent works, PCA has been directly used in the student-performance-prediction pipeline. Olabanjo, Wusu and Manuel measured the specific effect of PCA on a Radial Basis Function Neural Network trained on secondary-school academic, cognitive and psychomotor records, reporting a sensitivity of 93.49% for pass prediction and an overall accuracy of 86.59%, explicitly noting the fact that PCA changed the network's sensitivity-specificity balance [4]. Sabri et al. used continuous, weekly self-efficacy and learning-behavior data in a PCA-based Support Vector Machine model and showed that low-dimensional PCA-derived composite scores were sufficient to retain a wide enough predictive signal to be useful for forecasting performance over the course of an academic term [8]. The hybrid architecture proposed by Al-tameemi first uses PCA to reduce the dimensions, then uses K-Means to cluster, and finally uses multi-class supervised prediction. It is suggested that this pipeline of unsupervised then supervised will improve the visualization and the effectiveness of subsequent classification on multi-class educational datasets [13].

Hybrid and ensemble techniques without explicit PCA have also grown in the literature for 2020-2025. Asselman, Khaldi and Aammou improved student performance prediction using the XGBoost algorithm. They showed that gradient boosted trees performed better than numerous classical baselines using data from an interactive learning environment [5]. Kukkar, Mohana, Sharma and Nayyar [6] created a SAPP system which combines stacked Long Short-Term Memory networks with Random Forest and Gradient Boosting and showed that layered multi-model architectures can capture complementing behavioral patterns. Farzanehfar and Arashpour [7] hybridized Support Vector Machine and Artificial Neural Network models with a teaching-learning-based optimization technique and reported better prediction reliability than either constituent model. Ali et al. examined many machine-learning approaches for student performance prediction, and discussed the sensitivity of the reported accuracy to the choices of preprocessing [9]. Pallathadka et al. [10] used multiple machine learning techniques for classifying and predicting student-performance data and showed that ensemble and boosting-based approaches performed better than individual decision-tree or naïve probabilistic classifiers for tabular educational data.

More recent hybrid designs have integrated deep and classical learners without a specified PCA stage. A hybrid model for prediction of student performance was suggested by Rajaram, Shetty and Vaidya which incrementally improves over the single model baselines [11]. Alharbi and Allohibi [12] presented a new hybrid classification algorithm and tested it against conventional classifiers using educational benchmark data. Within the framework of virtual learning environments, Adnan et al. coupled explainable AI techniques with early-prediction models and the authors' focus was on both global and local interpretability [14]. Choi, Lam, and Mendes improved the prediction of learning-performance through explainable machine learning by including the interpretation of models directly into the prediction pipeline [15]. The work of Pham Xuan Lam et al. [16] investigated the efficacy of different machine learning algorithms in predicting student performance using interactive learning-management-system data, and emphasized the need for pretreatment of data and feature representation for accurate grade predictions.

The latest 2025 literature still explored hybrid deep-classical architectures with PCA-based visualization together. Ben George used explainable AI to forecast student grades and increase academic success by providing clear and actionable feedback [18]. In [19], Johora, Hasan, Rajbongshi, Ashrafuzzaman and Akter proposed an explainable AI-based approach for forecasting the academic performance of undergraduate students by combining ensemble learning with post-hoc interpretation. The feedback from stakeholders corroborated a hybrid AI strategy developed by Gonzales et al. to predict academic achievement of students in project-based learning [20]. To visualize the resultant student clusters in two-dimensional component space, Khan et al. [17] explicitly used PCA and reported 88% accuracy which is higher than the individual constituent models. They proposed a novel hybrid architecture combining Convolutional Neural Networks and Random Forests with XGBoost as a meta-learner. Sawarkar, Pinjarkar, Agrawal and Motghare presented a hybrid predictive model that combined logistic regression, decision tree models, and random forests to find at-risk students in a gamified education environment. The authors showed

that the ensemble was better than any of the individual constituent classifiers [21].

Taken together, these recent bodies of literature provide two convergent findings that inspire the present investigation. Finally, PCA is increasingly used in student-performance prediction, but mainly as an unsupervised visualization or clustering aid [13], [17] or as a pre-processing step evaluated for a single family of classifiers [4], [8] and not as a dimensionality-reduction strategy systematically compared across a broad panel of algorithmically distinct classifiers. Second, no reviewed 2020-2025 study explicitly quantifies, by a controlled with/without-PCA comparison, which classifier families benefit most from PCA's decorrelation property and by how much, nor relates this benefit to the underlying statistical assumptions of each algorithm – the specific gap addressed by this paper.

3. Materials and Methods

3.1 Dataset and Leakage-Aware Feature Selection

This study uses the (Student Academic Performance & Engagement) dataset which consists of 15,000 anonymized student records from the university across six academic years (2019-2024). The target variable is a four-class ordinal variable, Performance_Class (At-Risk, Second Class, First Class, Distinction), determined by cumulative grade-point-average (CGPA) thresholds. The initial correlation study showed that the target label is a near deterministic function of CGPA and its directly constituent attributes, six semester GPAs, Internal_Avg, Lab_Score, Mid_Term_Avg, End_Term_Avg, Prev_Year_CGPA and Class_Rank. These thirteen attributes, along with Student_ID and the secondary Dropout indicator, were dropped from the predictor set to avoid target leakage. This resulted in 38 leakage-aware predictors covering demographic, socio-economic, institutional, behavioral /engagement and continuous - assessment categories.

3.2 Preprocessing of data

From the 38 predictors that were maintained, 24 are numeric and 14 are categorical. Numeric predictors were z-score normalized (zero mean, unit variance); categorical predictors were one-hot encoded. This preprocessing increased the feature space from 38 raw predictors to D = 78 encoded dimensions. The dataset was split into training (80%, 12,000 records) and held-out test (20%, 3,000 records) sets using stratified sampling to maintain the class proportions in both sets. The PCA transform itself (Section 3.3) was only fit on the training partition and then applied to the test partition, to avoid any data leakage from the test set into the dimensionality reduction step.

3.3 Principal Component Analysis

The principle of PCA is to search for an orthogonal linear transformation of the original D dimensional feature space to a new coordinate system where the axes (principal components) are ordered by the amount of variation they explain. In formal terms, PCA finds the eigenvectors and eigenvalues of the covariance matrix of the standardized training feature matrix. The eigenvector

with the largest eigenvalue is the first principal component (PC1) that captures the direction of maximum variance in the data and each of the subsequent principal components captures the maximum remaining variance subject to orthogonality in relation to all preceding components [9], [17]. The proportion of total variance explained by the k-th component is given by its eigenvalue divided by the sum of all eigenvalues, and the cumulative explained variance after k components is the running sum of these proportions.

PCA was fitted on the standardized training feature matrix (of dimension 78) and the number of components necessary to get 90%, 95% and 99% cumulative explained variance were calculated (Table 1). As is common in the dimensionality-reduction literature [9], [17], the 95% variance threshold was chosen as the primary operating point for the main experiments in this study, retaining N = 46 principal components — a reduction of 41.0% relative to the original 78-dimensional encoded feature space.

Table 1. Number of principal components required to reach successive explained-variance thresholds

Variance Threshold	Components Required	Dimensionality Reduction
90%	38	51.28%
95% (used in this study)	46	41.03%
99%	57	26.92%

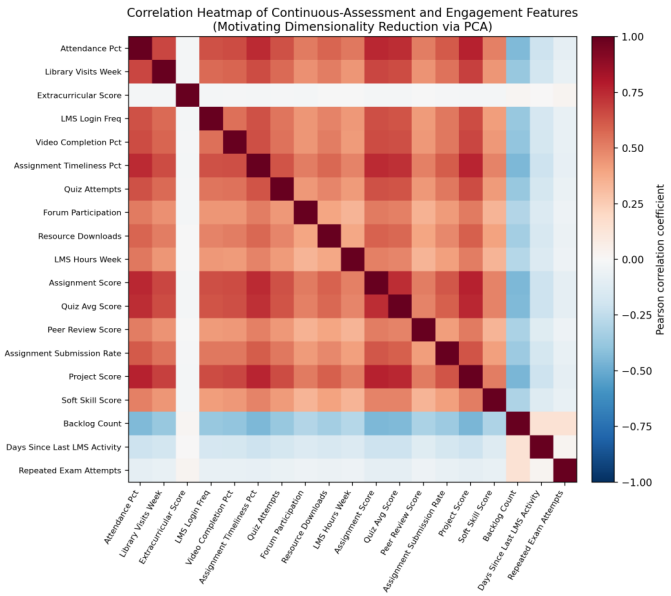

Fig. 1. Correlation heatmap of continuous-assessment and engagement predictors, motivating dimensionality reduction via PCA.

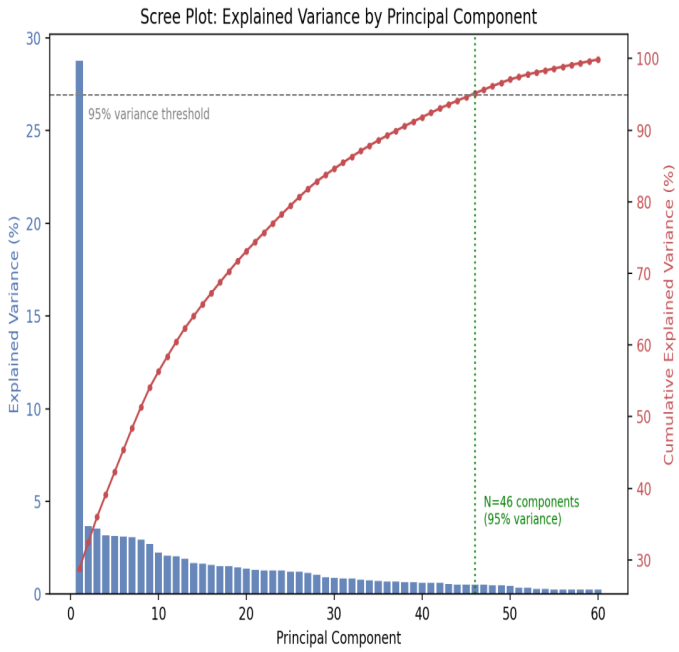

Fig. 2. Scree plot: individual and cumulative explained variance across principal components.

3.4 Proposed Methodology and Algorithm

Fig. 3 illustrates the overall methodological pipeline. After leakage-aware feature selection and preprocessing, PCA is fit on the training partition and used to project both the training and test partitions onto the retained N = 46 principal components. Seven classifiers, spanning probabilistic (Naïve Bayes), linear (Logistic Regression), tree-based (Decision Tree, Random Forest, XGBoost), instance-based (k-Nearest Neighbors), and margin-based (SVM) families, are then trained and evaluated identically on both the original 78-dimensional space and the 46-dimensional PCA-reduced space, isolating the effect of dimensionality reduction from any confounding difference in classifier configuration or data partitioning.

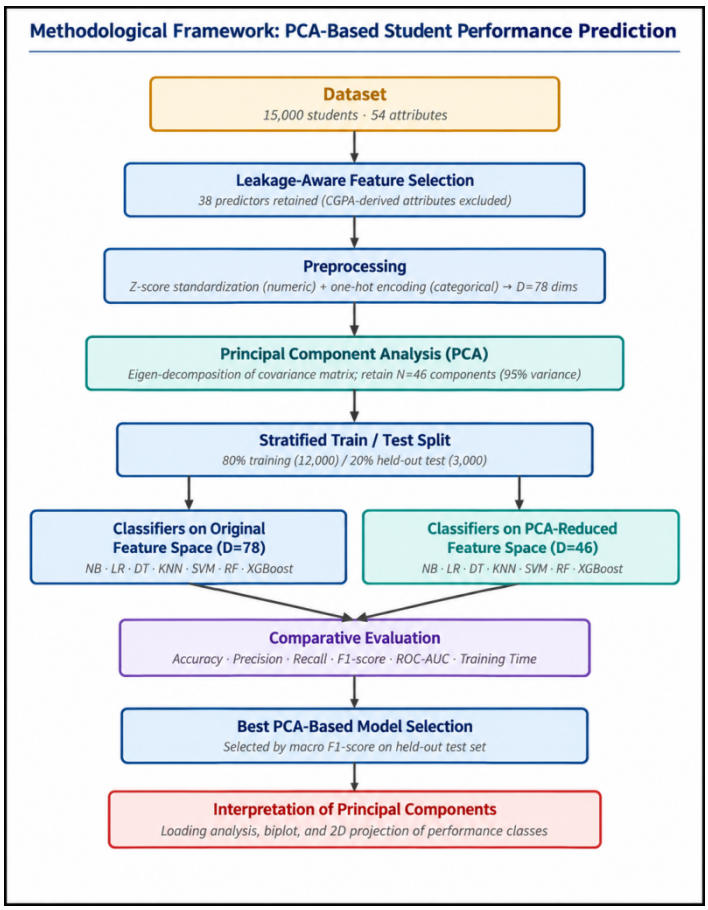

Fig. 3. Methodological framework for PCA-based student performance prediction.

Algorithm 1: PCA-Based Student Performance Prediction Pipeline

Input: Raw dataset $D = \{(x_i, y_i)\}$, $i = 1..N$, y_i {At-Risk, Second, First, Distinction}
Variance-retention threshold $\tau = 0.95$
Output: PCA-reduced classifier H and its test-set evaluation metrics

Step 1: Preprocess(D):
Exclude CGPA-derived attributes (leakage-aware protocol)
Standardize numeric features to zero mean, unit variance
One-hot encode categorical features
Step 2: Split D into D_train (80%) and D_test (20%) via stratified sampling
X_train, X_test ← encoded feature matrices of dimensionality D
// ---- Principal Component Analysis ----
Step 3: $\Sigma \leftarrow$ Covariance(X_train)
// D x D covariance matrix
$(\lambda_1, v_1), \ldots, (\lambda_D, v_D) \leftarrow$ EigenDecompose(Σ)
// eigenvalues, eigenvectors, sorted $\lambda_1 \geq \ldots \geq \lambda_D$
$\text{explained}_k \leftarrow \lambda_k / \Sigma_j \lambda_j$ for k = 1..D

variance ratio
cumulative_k ← Σ_{j=1}^{k} explained_j
Step 4: N* ← min { k : cumulative_k ≥ τ }
// smallest k reaching 95% variance
W ← [v_1, v_2, ..., v_{N*}]
// D x N* projection (loading) matrix
Step 5: Z_train ← X_train · W ; Z_test ← X_test · W
// project onto top N* components
// ---- Classifier training on PCA-reduced representation ----
Step 6: for each candidate classifier C in {NB, LR, DT, KNN, SVM, RF, XGB} do
Train C on (Z_train, y_train)
ŷ_test ← C.predict(Z_test)
Compute Accuracy, Precision, Recall, F1-score, ROC-AUC on (ŷ_test, y_test)
end for
Step 7: H ← argmax_C F1_macro(C)
// best PCA-based model by macro F1-score
Step 8: return H, {N*, W, explained_k, cumulative_k}, evaluation metrics for all C

3.5 Baseline Classifiers and Evaluation Metrics

Seven classifiers were implemented using scikit-learn and XGBoost (Python 3): Gaussian Naïve Bayes, Logistic Regression, Decision Tree (max depth 12), k-Nearest Neighbors (k=15), Support Vector Machine (RBF kernel, C=8), Random Forest (300 trees, max depth 14), and XGBoost (300 estimators, max depth 6, learning rate 0.1). For each classifier, we trained twice, using the same stratified train/test split and fixing all random-state parameters to ensure reproducibility: (1) on the original 78-dimensional encoded feature space, and (2) on the PCA-reduced 46-dimensional representation. The model performance was evaluated using overall accuracy, macro-averaged precision, recall and F1-score, a detailed per-class classification report, a confusion matrix for the best PCA-based model, multi-class (one-vs-rest) ROC-AUC, and wall-clock training time.

4. Results and Discussion

4.1 Classification Performance Without PCA

Table 2 shows the test-set performance of all seven classifiers in the original, non-reduced 78-dimensional encoded feature space. Logistic Regression had the greatest baseline performance (86.90% accuracy, 87.34% macro F1-score), followed by Random Forest (86.33% / 86.85%) and XGBoost (86.03% / 86.65%). The worst performing algorithms were Naïve Bayes (77.13% / 77.27%) and Decision Tree (78.50% / 79.39%). This is in line with the Naïve Bayes requirement of conditional-independence being broken by the high inter-feature correlation observed in Fig. 1, and with the Decision Tree's tendency to vary by correlated, duplicated splits.

Table 2. Classification performance without PCA (original 78-dimensional feature space, n = 3,000)

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Naive Bayes	77.13	79.94	80.27	77.27
Logistic Regression	86.90	87.59	87.11	87.34
Decision Tree	78.50	79.58	79.21	79.39
KNN	84.07	86.61	83.23	84.65
SVM (RBF)	84.53	85.37	85.28	85.32
Random Forest	86.33	88.18	85.78	86.85
XGBoost	86.03	87.15	86.18	86.65

4.2 Classification Performance With PCA

Table 3 shows the performance of the same seven classifiers on the 46 dimensional PCA reduced representation on the test set. The PCA-reduced Logistic Regression model presented the greatest overall performance among all configurations in this investigation (87.07% accuracy, 87.55% macro F1-score), which is a slight improvement above its non-reduced version. The biggest improvement was for Naïve Bayes, with an increase in accuracy from 77.13% to 83.63% and in F1-score from 77.27% to 83.50% - an improvement of 6.50 percentage points in accuracy. The Decision Tree also increased in accuracy from 78.50% to 81.63%. XGBoost showed a slight improvement, but Random Forest and SVM exhibited minimal, statistically marginal changes in either direction, and KNN was practically the same.

Table 3. Classification performance with PCA (46-dimensional reduced feature space, n = 3,000)

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Naive Bayes	83.63	85.71	82.15	83.50
Logistic Regression	**87.07**	**87.75**	**87.36**	**87.55**
Decision Tree	81.63	82.77	82.03	82.39
KNN	84.20	86.47	83.19	84.58
SVM (RBF)	84.50	85.20	84.98	85.09
Random Forest	86.20	87.49	85.97	86.67
XGBoost	86.40	87.20	86.75	86.97

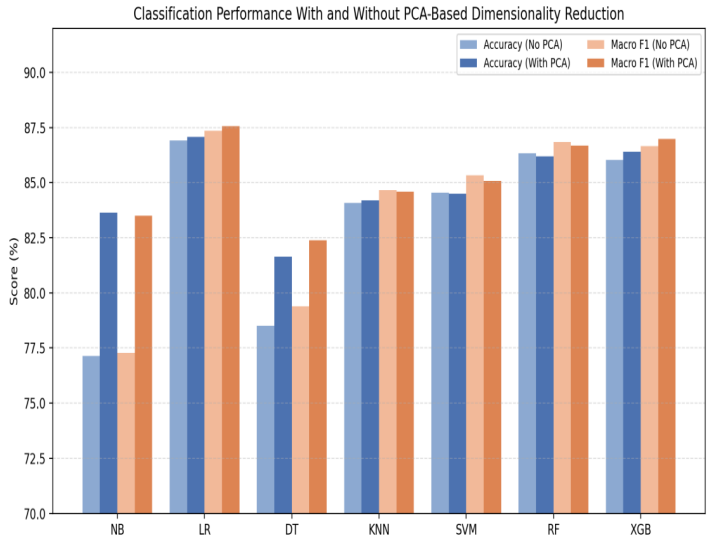

Fig. 4. Classification performance with and without PCA-based dimensionality reduction, across seven classifiers.

4.3 PCA Effect per Classifier Family (Quantified)

Table 4 lists the change in accuracy, macro F1-score and training duration by PCA for each classifier. The results show a clear algorithm-dependent pattern: Classifiers that are directly sensitive to feature correlation (Naive Bayes, whose conditional-independence assumption is directly violated by correlated raw features; and, to a lesser extent, the single Decision Tree, whose axis-aligned recursive splits benefit from a decorrelated, variance-ranked input space) benefit substantially from PCA, while classifiers that already model feature interactions robustly (Random Forest, XGBoost, SVM with a non-linear kernel) show negligible or slightly mixed change, as these algorithms are comparatively insensitive to the specific coordinate system of the input features.

Table 4. Change in accuracy, macro F1-score, and training time attributable to PCA (With PCA minus Without PCA)

Model	Δ Accuracy (pp)	Δ Macro F1 (pp)	Δ Training Time
Naive Bayes	+6.50	+6.23	-0.05s
Logistic Regression	+0.17	+0.21	-0.20s
Decision Tree	+3.13	+3.00	+0.41s
KNN	+0.13	-0.07	-0.00s
SVM (RBF)	-0.03	-0.23	-2.16s
Random Forest	-0.13	-0.18	+13.20s
XGBoost	+0.37	+0.33	+5.56s

4.4 Training Time Analysis

Figure 5 compares the training time with versus without PCA. PCA improved the training time of Naive Bayes and KNN, which is expected due to their known sensitivity to input dimensionality. Logistic Regression and SVM also learned a little faster with PCA. In contrast, a Random Forest and XGBoost were slower to train on the PCA-reduced representation, although its dimensionality was smaller. This counter-intuitive finding is because PCA replaces the original sparse, mostly binary, one-hot-encoded feature matrix with a dense matrix of continuous values with the same or fewer columns, and tree-based ensembles have to evaluate more candidate split thresholds for each dense continuous feature than for each sparse binary indicator, partially negating the benefit of the reduced dimensionality. The result points out that the computational benefit of PCA is not homogeneous across method families and has to be evaluated per-classifier rather than a priori.

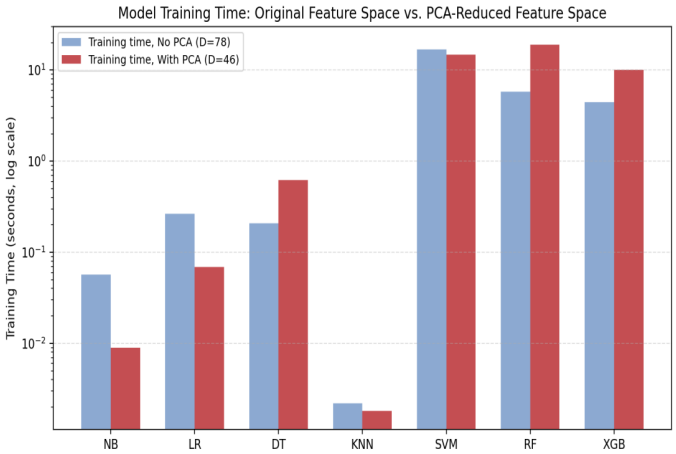

Fig. 5. Model training time comparison: original feature space versus PCA-reduced feature space (log scale).

4.5 Confusion Matrix and Per-Class Performance

Table 5 and Fig. 6 reports per-class performance and confusion matrix of the best PCA-based model (Logistic Regression, N = 46 components). All four classes get an F1-score above 85%. The best results are obtained for At-Risk (F1 = 89.19%) and Distinction (F1 = 88.76%), while the two middle neighboring classes First Class (F1 = 85.94%) and Second Class (F1 = 86.31%) perform comparatively lower, but still good. The confusion matrix shows that most misclassifications are between adjacent ordinal categories and that there is essentially little confusion between the extreme classes (At-Risk and Distinction), confirming that the faults of the PCA reduced model are ordinally graceful rather than random.

Table 5. Per-class precision, recall, and F1-score of the best PCA-based model (Logistic Regression)

Performance Class	Precision (%)	Recall (%)	F1-score (%)	Support
At-Risk	89.65	88.73	89.19	488

Second Class	86.23	86.40	86.31	1000
First Class	85.31	86.59	85.94	999
Distinction	89.82	87.72	88.76	513

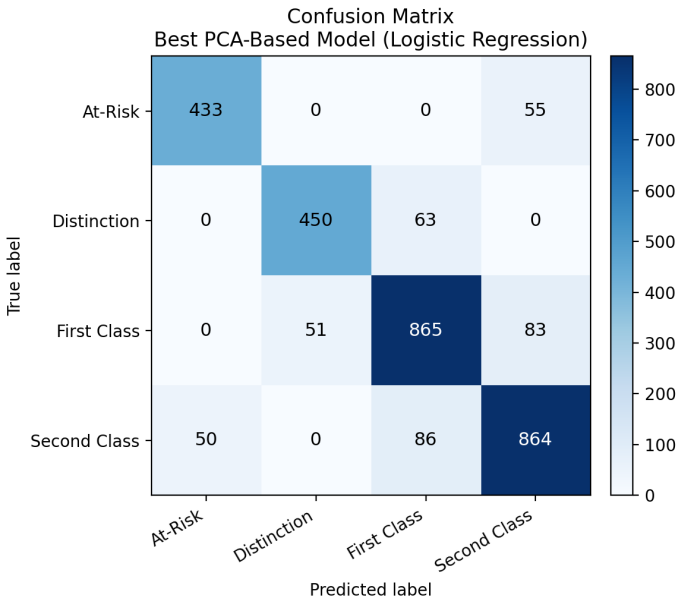

Fig. 6. Confusion matrix of the best PCA-based model on the held-out test set.

4.6 ROC-AUC Analysis

The best PCA based model produced a macro-averaged ROC-AUC of 0.981 on the held-out test set. The highest per-class AUC was achieved for At-Risk and Distinction (0.994 and 0.992, respectively) while Second Class and First Class had comparatively lower AUC (0.971 and 0.967, respectively), mirroring the pattern of F1-score in Section 4.5. This confirms that the model's class-probability estimates (not just its hard classifications) are best separated for the two extreme performance categories.

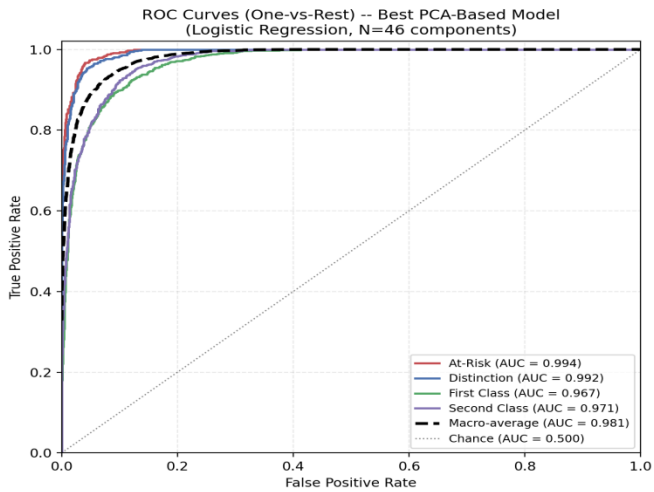

Fig. 7. One-vs-rest ROC curves and macro-average ROC curve for the best PCA-based model (Logistic

Regression, N = 46 components).

4.7 Interprettion of Principal Components

To interpret the latent structure recorded by PCA, the coefficients (loadings) of each original predictor on the first two principal components were analyzed. The five largest contributors to PC1 by absolute loading magnitude are Project_Score, Attendance_Pct, Assignment_Score, Assignment_Timeliness_Pct, and Quiz_Avg_Score, the same continuous-assessment and engagement indicators that were found to be most predictive in the correlation analysis of Section 3.3 (Fig. 1), all loading with the same sign, confirming PC1 as a single, coherent composite axis of overall academic engagement and continuous-assessment performance. Conversely, PC2 is dominated by Distance_km and Hostel_Resident status, a mainly orthogonal composite axis associated with commuting and residential situations, rather than academic participation.

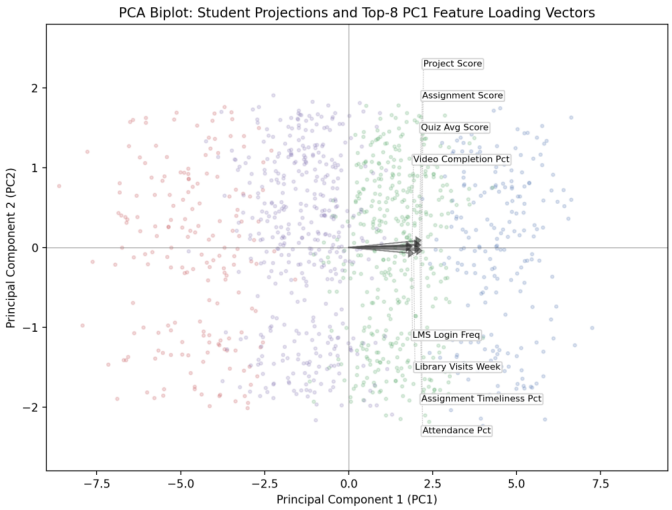

Fig. 7. PCA biplot: student projections in the PC1-PC2 plane with the top-8 PC1 feature loading vectors overlaid.

Fig. 8. Projected held-out test-set students into the first two primary components, colored by true performance class. The four classes are visually and almost monotonically separated along PC1, from At-Risk (leftmost, lowest PC1 values) to Second Class and First Class to Distinction (rightmost, highest PC1 values), and show little separation along PC2, which is a direct visual confirmation that the dominant source of variance captured by PCA in this dataset is closely aligned with the ordinal academic-performance target, although PCA itself is an unsupervised technique with no access to the class labels during fitting.

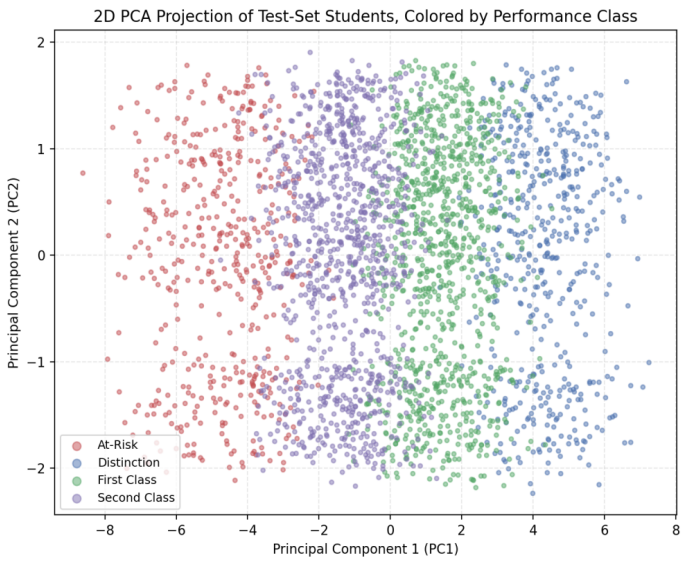

Fig. 8. Two-dimensional PCA projection of held-out test-set students, colored by true performance class.

4.8 Discussion

This investigation leads to three conclusions. First, the PCA-based dimensionality reduction, in which only 46 of 78 encoded dimensions were retained (a 41.0% reduction), did not degrade, and in the case of the best performing classifier, marginally improved, predictive accuracy on this dataset. Thus, the lower variance components that were excluded do not carry much unique information relevant to the classification task, beyond what is already captured by the retained components. Second, the advantage of PCA is highly algorithm-dependent: Naive Bayes gained 6.50 percentage points in accuracy, and the Decision Tree gained 3.13 points, whereas ensemble and margin-based approaches (Random Forest, XGBoost, SVM) had insignificant or mixed change. This pattern is directly explainable from the underlying statistical assumptions of each algorithm: The orthogonalization of the feature space performed by PCA is most valuable for exactly those classifiers – such as Naïve Bayes – that rely on the validity of approximately uncorrelated inputs, and least valuable for classifiers such as Random Forest and XGBoost that already model arbitrary feature interactions and are relatively invariant to the exact coordinate system of the input space. Third, as shown by the loading and projection analysis (Section 4.7), PCA is entirely unsupervised, but recovers a first principal component very well aligned with the ordinal performance-class structure, and driven by the same continuous-assessment and engagement predictors identified as most important in the previous, independent feature-importance analyzes of this dataset, thus providing a sort of external, cross-methodological validation of the substantive conclusion that continuous assessment and engagement are the dominant drivers of student performance.

This study has certain drawbacks. Dataset is collected from one university, and the exact number of components required to achieve a certain variance threshold may vary greatly for datasets with various feature compositions or correlational structure. PCA is a linear method and

hence cannot capture non-linear correlations among predictors. Non-linear methods such as kernel PCA or autoencoders were not investigated in the current study, but are suggested for future research. Lastly, the 95% variance threshold, a standard convention in the dimensionality-reduction literature, was not itself optimized on downstream classification accuracy; a supervised or classifier-specific selection of the number of retained components may lead to additional improvement.

5. Conclusion and Future Work

In this study, we investigate the use of Principal Component Analysis as a dimensionality reduction technique for multi-class student performance prediction on the dataset. The encoded feature space was reduced by PCA from 78 to 46 dimensions under a leakage-aware, 38-predictor feature methodology, preserving 95% of total variance, and seven classifiers were systematically tested with and without this reduction. The best PCA-based model (Logistic Regression) achieved 87.07% accuracy and 87.55% macro F1-score. Naive Bayes showed the largest benefit from PCA (6.50-percentage-point accuracy gain). This is because PCA more closely satisfies Naive Bayes' conditional-independence assumption by decorrelating the input feature space. The loading analysis and the two-dimensional projection reveal that the first principal component is dominated by continuous-assessment and engagement predictors, and distinguishes the four performance classes in an almost monotonous fashion, despite the unsupervised formulation of PCA. In future work we will explore non-linear dimensionality-reduction alternatives (kernel PCA, autoencoders), a supervised or classifier-specific criterion for the number of components to retain, and validation of the PCA-based methodology on external datasets from other disciplines and educational systems.

References

[1] P. T. Sokkhey Okazaki, "Hybrid machine learning algorithms for predicting academic performance," International Journal of Advanced Computer Science and Applications, vol. 11, no. 1, pp. 32–41, 2020, doi: 10.14569/ijacsa.2020.0110104.

[2] B. Zaki, N. Albreiki, and H. Alashwal, "Systematic literature review of students performance prediction using machine learning techniques," Education Sciences, vol. 11, no. 9, 552, 2021, doi: 10.3390/educsci11090552.

[3] A. Al-Zawqari, D., Peumans, G. Vandersteen, "A flexible feature selection approach for predicting students' academic performance in online courses," Computers and Education: Artificial Intelligence, vol. 3, article 100103, 2022. doi: 10.1016/j.caeai.2022.100103.

[4] O. A. Olabanjo, A. S. Wusu, M. [16] Manuel, "Prediction of academic performance of secondary school students using radial basis function neural network: A machine learning approach," Scientific African, 2022.

[5] A. Asselman, M. Khaldi and S. Aammou, "Enhancing student performance prediction using the XGBoost machine learning algorithm," Interactive Learning Environments, vol. 31, no. 6, pp. 3360–3379, 2023.

[6] A. A. Kukkar, R. Mohana, A. Sharma and A. Nayyar, "Prediction of student academic performance based on their mental health and interaction on various e-learning platforms," Education and Information Technologies, vol. 28, no. 8, pp. 9655-9684, 2023.

[7] P. M. Farzanehfar & M. Arashpour, "Predicting individual learning performance using machine learning hybridized with the teaching-learning-based optimization," Computer Applications in Engineering Education, vol. 2023, 31, 83-99. doi: 10.1002/cae.22572.

[8] N. A. M. Sabri et al., "Prediction model based on continuous data for student performance using principal component analysis and support vector machine," TEM Journal, vol. 12, no. 2, pp. 1201–1210, 2023, doi: 10.18421/TEM122-66.

[9] D. A. Ali, M. Aborizka and A. Dahroug, "Forecasting student performance using machine learning techniques," in Proc. Int. Conf. Automation, Robotics and Computer Science (AIRC), 2023, doi: 10.1109/AIRC57904.2023.10303160.

[10] H. A. Wenda, E. Ramirez-Asis, M. Asis-Lopez, J. Flores-Albornoz and K. Pallathadka Phasinam, "Classification and prediction of student performance data using different machine learning algorithms," Materials Today: Proceedings, vol. 80, pp. 3782-3785, 2023.

[11] K. Rajaram M R, Shetty B, M. V. Vaidya, "Hybrid model to predict the performance of students," in Proc. 15th Int. Conf. on Advances in Computer, Control and Telecommunication Technologies (ACT), Hyderabad, India, 2024, pp. 598–604.

[12] A. A. Alharbi and J. . Allohibi, "A novel hybrid classification algorithm for predicting student performance," AIMS Mathematics, vol. 9, no. 7, pp. 18308–18323, 2024, doi: 10.3934/math.2024893.

[13] G. Al-tameemi, "A hybrid machine learning approach to predict student performance using multi-class educational datasets," Procedia Computer Science, vol. 238, pp. 888-895, 2024, doi: 10.1016/j.procs.2024.06.108.

[14] M. Adnan, M.I. Uddin, E. Khan, F.S. Alharithi, S. Amin, A. A. Alzahrani. The earliest feasible global and local interpretation of students' performance in virtual learning environment using explainable AI. Informatics in Education, 20:1–20, 24, no. 1, pp. 1–24, 2024.

[15] W. C. Choi, C.-T. Lam, and A. . J. Mendes, "Explainable machine learning for better learning performance prediction," in Proc. 2024 IEEE Frontiers in Education Conference (FIE). 2024.

[16] P. X. Lam, P. Q. H. Mai, Q. H. Nguyen, T. Pham, T. H. H. Nguyen, and T. H. Nguyen, "Predictive Student Assessment Modeling for Better Educational Assessment," Computers and Education: Artificial Intelligence, vol. 6, article 100244, 2024.

[17] S. Khan, T Mazhar, T Shahzad, M A Khan, W Waheed, A Waheed and H Hamam, "Predictive analytics in education: improving student achievement using machine learning," ScienceDirect journal article, 2025.

[18] E. Ben George, "Explainable AI methods for predicting student grades and improving academic success," Journal of Information Systems Engineering and Management, vol. 10, no. 23s, pp. 117-126, 2025, doi: 10.52783/jisem.v10i23s.3680.

[19] F. T. Johora, M. N. Hasan, A. Rajbongshi, M. Ashrafuzzaman and F. Akter, "An explainable AI-based approach for predicting undergraduate students' academic performance," Array, vol. 26, art. 100384, 2025, doi:10.1016/j.array.2025.100384.

[20] W. Z. Gonzales, C. Cordero, C. D. Abanto-Ramírez, H. Iftikhar, E. T. Susanibar Ramirez and J. L. Lopez-Gonzales, "A hybrid AI approach for predicting academic performance in RBE students," Front. Artif. Intell., vol. 8, art. 1651100, 2025. doi: 10.3389/frai.2025.1651100

[21] D. Sawarkar, L. Pinjarkar, P. Agrawal and D. Motghare, "Predictive analytics in gamified education: A hybrid model for identifying at-risk students," MethodsX, vol. 15, 103486, 2025, doi:10.1016/j.mex.2025.103486.