Available online www.worldscientificnews.com

World Scientific News

An International Scientific Journal

WSN 215 (2026) 59-77 EISSN 2392-2192

Enhancing Early Intervention in Secondary Education: A Comparative Study of Dimensionality Reduction Techniques for Predicting At-Risk Students

Abraham Osemeke Agbonifo[1], Isaac Nosakhare Agbonifo[2], Gabriel Chukuemeke Agbonifo[3], Chidimma Grace Emmanuel[4], Uchechi Joyce Nneji[5], Benjamin Chiemeka Opara[6], Simon Onuwa Agbonifo[7], Daniel Agbonifo[8*]

[1]Department of Guidance and Counselling, Delta State University, Abraka, Nigeria

[2]School of Business Studies, Auchi Polytechnic, Auchi, Nigeria

[3]Department of Mechanical Engineering Technology, Auchi Polytechnic, Auchi, Nigeria

[4]Department of Education Economics, Faculty of Social Sciences, University of Nigeria Nsukka, Enugu State, Nigeria

[5]College of Humanities and Law, Chengdu University of Technology, China

[6]Educational Management (Political Science), University of Benin, Nigeria

[7]Department of Chemistry, University of Benin, Nigeria

[8]College of Geophysics, Chengdu University of Technology, China

*Author for Correspondence: agbonifodaniel@stu.cdut.edu.cn

https://doi.org/10.65770/IJXV3497

(Received 9 March 2026; Accepted 22 April 2026; Date of Publication 12 May 2026)

ABSTRACT

Secondary education systems face the critical challenge of early student attrition and academic underperformance. While schools collect extensive demographic and behavioral data, its high dimensionality and redundancy obscure actionable insights for guidance counselors. This study addresses this gap by developing an early warning system using the UCI Student Performance dataset. We benchmarked three advanced dimensionality reduction techniques—Uniform Manifold Approximation and Projection (UMAP), Neighborhood Components Analysis (NCA), and Partial Least Squares Discriminant Analysis (PLS-DA)—integrated with diverse machine learning classifiers. Experimental results demonstrate that dimensionality reduction significantly enhances model performance by eliminating noise, with the NCA-XGBoost combination achieving optimal accuracy of 94.6% and recall of 94.1%. The analysis identified study time, alcohol consumption, and family relationship quality as the strongest behavioral predictors of academic success. This framework provides guidance counselors with a reliable clinical decision support system for proactive intervention, enabling targeted support based on specific behavioral triggers rather than intuition alone.

***Keywords**:* Educational Data Mining, Student Performance Prediction, Dimensionality Reduction, Guidance and Counselling, Early Intervention, UMAP, NCA.

1. INTRODUCTION

The escalating rates of academic underperformance and student attrition in secondary education present a critical challenge for educators and school administrators worldwide. Beyond the immediate academic consequences, early school failure is often a precursor to long-term socioeconomic disadvantages and psychosocial instability. Consequently, the role of guidance counselors has evolved from reactive disciplinary management to proactive student welfare monitoring. In this modern educational landscape, the ability to identify “at-risk” students before their grades decline is paramount for effective intervention.

Schools today are rich in data but often poor in actionable insights. Educational institutions routinely collect vast amounts of information, ranging from demographic details and family background to behavioral records and past academic performance. The UCI Student Performance dataset exemplifies this richness, containing 33 attributes per student across demographic, socioeconomic, academic, and behavioral dimensions. However, the high dimensionality of such data creates a paradox: the sheer volume of variables often introduces noise and redundancy, obscuring the true indicators of student distress. For a counselor, sifting through dozens of correlated factors, such as "weekend alcohol consumption" versus "workday alcohol consumption," can be overwhelming and computationally inefficient. This phenomenon, known as the "curse of dimensionality," hampers the performance of predictive models and reduces their interpretability for non-technical staff.

While prior research has established machine learning's efficacy in educational settings, from Cortez and Silva's (2008) foundational work to recent studies applying ensemble methods, a critical gap remains. Most studies focus on maximizing predictive accuracy without addressing the practical needs of guidance counselors who require both accurate predictions and interpretable insights. Furthermore, although traditional dimensionality reduction techniques like Principal Component Analysis (PCA) have been applied, their linear assumptions often fail to capture the complex, non-linear relationships inherent in student behavior data.

This study addresses these gaps by developing an intelligent Clinical Decision Support System (CDSS) specifically designed for educational guidance contexts. We benchmark three advanced dimensionality reduction strategies, namely Uniform Manifold Approximation and Projection (UMAP) for preserving local data structure, Neighborhood Components Analysis (NCA) for learning feature weights, and Partial Least Squares Discriminant Analysis (PLS-DA) for handling multicollinearity against a baseline with no feature selection. By integrating these techniques with robust machine learning classifiers and emphasizing model interpretability through SHAP analysis, this research provides counselors with a reliable tool that not only flags at-risk students but also highlights specific behavioral triggers requiring urgent intervention.

Research Questions:

- How do advanced dimensionality reduction techniques (UMAP, NCA, PLS-DA) compare in enhancing predictive accuracy for identifying at-risk secondary students?
- Which behavioral and demographic factors emerge as the strongest predictors of academic performance when data noise is reduced?
- How can data-driven insights from such models be effectively operationalized within school guidance counseling frameworks to support proactive intervention?

2. BACKGROUND REVIEW

2.1. Evolution of Student Performance Prediction

The application of Artificial Intelligence in education, often termed Educational Data Mining (EDM), has evolved from simple grade storage to complex predictive modeling. Early research largely relied on statistical correlation to link socioeconomic status with academic achievement. Cortez and Silva's (2008) seminal work established a benchmark by applying decision trees and neural networks to Portuguese student data, identifying past failures and alcohol consumption as key predictors of academic performance. This foundational study demonstrated that machine learning could transform raw student data into actionable predictions.

As the field matured, researchers explored increasingly sophisticated algorithms. Ensemble methods gained prominence, with studies showing that Gradient Boosting algorithms often outperform single classifiers in identifying at-risk students (Rastrollo-Guerrero et al., 2020). More recent work has focused on fairness-aware approaches; Kesgin et al. (2025) applied 5-fold cross-validation with SMOTE and fairness diagnostics, reporting XGBoost accuracy of 78.9% and F1-score of 80.3% on the UCI Student Performance dataset. Similarly, Begum and Padmannavar (2023) demonstrated that Bayesian-optimized Random Forest achieved 87% accuracy on the same dataset.

Despite these advances, a recurring limitation persists: the "black box" nature of high-performing models. For guidance counselors, predictive accuracy alone is insufficient if the model cannot explain why a student is flagged as high-risk.

2.2. The Challenge of High Dimensionality in Educational Data

Modern student datasets are inherently high-dimensional, containing dozens of variables ranging from "mother's job" to "weekend free time." The UCI Student Performance dataset exemplifies this challenge, with 33 original attributes per student that can be expanded to 111 engineered features through proper encoding and feature engineering. Including irrelevant or redundant features often degrades model performance, a phenomenon known as the "curse of dimensionality."

Traditional dimensionality reduction techniques, such as Principal Component Analysis (PCA), have been widely used to compress educational data. However, PCA assumes linear relationships between variables, which often fails to capture the complex, non-linear behavioral patterns found in student data. The relationship between study time and grades, for instance, is rarely linear and often interacts with factors like motivation, family support, and learning environment.

2.3. Advanced Dimensionality Reduction Techniques

To address the limitations of linear methods, recent scholarship has turned to manifold learning and supervised dimensionality reduction. Uniform Manifold Approximation and Projection (UMAP) (McInnes et al., 2018) has emerged as a state-of-the-art technique for preserving both local and global data structures, yet its application in predicting secondary school attrition remains underexplored. Similarly, Neighborhood Components Analysis (NCA) (Goldberger et al., 2005) offers distinct advantages by learning a linear transformation specifically optimized for classification accuracy, effectively "weighting" features based on their predictive power. Partial Least Squares Discriminant Analysis (PLS-DA) provides another supervised approach, finding fundamental relations between predictor variables and categorical responses while being robust against multicollinearity, which is common in educational data where variables like "mother's education" and "father's education" are often highly correlated.

2.4. Applications in Guidance and Counseling

The intersection of educational data mining and guidance counseling represents an emerging frontier. While technical studies focus on algorithmic performance, practical implementation requires bridging the gap between data science and educational practice. Counselors need systems that not only predict outcomes but also provide interpretable insights that inform intervention strategies. Recent work by Alharbi et al. (2024) demonstrates the potential of hybrid classifier models for student success prediction, achieving 93% accuracy with their proposed approach. However, few studies explicitly address how these models can be integrated into existing counseling workflows or how they impact counselor decision-making processes.

2.5. Research Gap and Contributions

While existing studies compare classifiers (e.g., SVM vs. Random Forest), few have rigorously benchmarked how modern dimensionality reduction techniques (UMAP, NCA, PLS-DA) impact both predictive performance and, crucially, model interpretability for non-technical users. This study fills that gap by evaluating these techniques specifically through the lens of early intervention and guidance counseling. We contribute not only comparative performance metrics but also practical insights into feature importance and visualization methods that make complex models accessible to educational professionals.

Table 1. Summary of Key Literature on Student Performance Prediction.

Authors (Year)	Data / Task	Validation Approach	Models Tested	Key Findings
Cortez & Silva (2008)	UCI student-mat/por; binary classification (G3⩾10)	Not specified	Decision Trees, Neural Networks, SVM, Random Forest	Identified past failures and alcohol consumption as key predictors; established baseline performance metrics
Kesgin et al. (2025)	UCI Math/Portuguese merged; fairness-aware classification	5-fold CV with SMOTE	Logistic Regression, Random Forest, XGBoost	XGBoost achieved 78.9% accuracy, 80.3% F1; highlighted importance of fairness considerations
Begum & Padmannavar (2023)	UCI Student Performance; binary/multiclass	Not reported	Random Forest, k-NN with Bayesian optimization	Random Forest achieved 87% accuracy on UCI data
Alharbi et al. (2024)	UCI data; pass/fail prediction	Not reported	RF, C4.5, CART, SVM, Naive Bayes, k-NN, Hybrid Model	Proposed hybrid model achieved 93% accuracy, 66% recall, 68% precision
Rastrollo-Guerrero et al. (2020)	Various educational datasets	Various	Gradient Boosting, Ensemble Methods	Demonstrated ensemble methods often outperform single classifiers
This Study	UCI Math subset (395 students); binary classification	10-fold CV + 70/30 train-test split	10 ML models with UMAP, NCA, PLS-DA dimensionality reduction	Comprehensive benchmarking of dimensionality reduction techniques with emphasis on interpretability for counseling applications

3. METHODOLOGY

To develop a reliable Clinical Decision Support System for educational guidance, we constructed a comprehensive experimental framework. Our pipeline integrates data acquisition, extensive feature engineering, rigorous preprocessing, three advanced dimensionality reduction techniques, and a comprehensive suite of ten machine learning classifiers with hyperparameter optimization.

3.1. Dataset and Feature Engineering

3.1.1. Data Source and Characteristics

We used the Mathematics subset (student-mat.csv) from the UCI Student Performance dataset, comprising 395 students from two Portuguese secondary schools. The original dataset contains 33 attributes spanning demographic, socioeconomic, academic, and behavioral dimensions (Table 2). The target variable was defined as binary classification for early intervention: "At-Risk" (G3 < 10) versus "Safe" (G3 ⩾ 10), following established educational thresholds.

Table 2. Dataset Attributes Summary.

Attribute	Description	Scale/Type
school, sex, address	Student background	Categorical
Medu, Fedu	Parent education levels	Numerical (0-4)
studytime, failures	Academic factors	Numerical
Dalc, Walc	Alcohol consumption	Numerical (1-5)
Famrel	Family relationships	Numerical (1-5)
G1, G2, G3	First, second, final grades	Numerical

3.1.2. Comprehensive Feature Engineering

To capture complex educational relationships, we engineered 53 additional features, expanding the feature space to 86 variables before one-hot encoding. These engineered features included:

- Family Background Aggregations: Parental education composites, family support indices
- Learning Behavior Metrics: Study efficiency scores, attendance patterns
- Grade Trajectory Indicators: Momentum between G1 and G2 periods
- Social-Health Balance Scores: Lifestyle factor composites
- Polynomial and Interaction Terms: Non-linear relationships and synergistic effects

After one-hot encoding categorical variables, the final feature space comprised 111 dimensions, providing a rich representation of student characteristics while maintaining computational feasibility.

3.2. Data Preprocessing Pipeline

We implemented a rigorous preprocessing pipeline within scikit-learn's framework to prevent data leakage:

- Data Cleaning: Verified data completeness (no missing values in original dataset)
- Encoding: One-Hot Encoding for categorical variables with handle_unknown='ignore'
- Scaling: StandardScaler applied to numerical features (mean=0, variance=1)
- Class Imbalance Handling: SMOTE (Synthetic Minority Over-sampling Technique) applied only during training with sampling_strategy=1.0 and k_neighbors=5

All transformations were embedded within a scikit-learn Pipeline, ensuring that scaling parameters and encoding mappings were learned exclusively from training data.

3.3. Dimensionality Reduction Techniques

We benchmarked three advanced dimensionality reduction techniques against a baseline with no feature selection (NoFS). Each technique was selected for its unique strengths in handling educational data:

3.3.1. Uniform Manifold Approximation and Projection (UMAP)

UMAP (McInnes et al., 2018) constructs a high-dimensional graph representation of the data and optimizes a low-dimensional layout to maintain structural similarity. We employed the supervised variant, which incorporates class labels during dimensionality reduction to enhance class separability. The optimization objective preserves both local neighborhood structure and global data topology.

3.3.2. Neighborhood Components Analysis (NCA)

NCA (Goldberger et al., 2005) learns a linear transformation to maximize the expected leave-one-out classification accuracy of a k-nearest neighbor classifier. Unlike unsupervised methods, NCA uses target labels to weight features based on their discriminative power. The algorithm minimizes:

$$J(A) = \sum_{i} \sum_{j \neq i} p_{ij} \tag{1}$$

where p_{ij} represents the probability that point i selects point j as its neighbor in the transformed space.

3.3.3. Partial Least Squares Discriminant Analysis (PLS-DA)

PLS-DA finds latent variables that maximize covariance between predictor variables and categorical responses. This technique is particularly robust against multicollinearity, which common in educational data where variables like parental education levels are often correlated. PLS-DA projects features into a discriminative subspace while preserving predictive information.

3.4. Comprehensive Machine Learning Framework

3.4.1. Classifier Selection

We implemented ten representative classifiers to evaluate dimensionality reduction effectiveness across diverse algorithmic approaches:

- Linear Models: Logistic Regression (LR), Support Vector Machine (SVM)
- Instance-Based: K-Nearest Neighbors (KNN)
- Probabilistic: Gaussian Naive Bayes (GNB)
- Tree-Based: Decision Tree (DT), Random Forest (RF)
- Ensemble Methods: AdaBoost, Bagging, Stacking, Voting
- Advanced Boosting: XGBoost (added from new version)

3.4.2. Hyperparameter Optimization

We employed GridSearchCV with 10-fold stratified cross-validation to optimize hyperparameters for each classifier-dimensionality reduction combination. The search space included:

- Dimensionality reduction components: {10, 15, 20, 25, 30} for UMAP and PLS-DA; {15, 20, 25, 30, 35, 40} for NCA
- Classifier-specific parameters (see Table 3 in original document)
- Evaluation metric: F1-Score to balance precision and recall

3.4.3. Three-Step Validation Protocol

To ensure robust evaluation, we implemented a rigorous three-step protocol:

1. Hyperparameter Tuning: GridSearchCV with 10-fold cross-validation on training data
2. Generalization Assessment: Retraining best configuration on full training set, evaluation on held-out test set (20% of data)
3. Stability Verification: Final 10-fold cross-validation with optimal hyperparameters to quantify model stability

3.5. Model Interpretability Framework

3.5.1. SHAP Analysis

For model transparency, which is critical for counseling applications, we implemented SHAP (SHapley Additive exPlanations) analysis. SHAP values quantify feature contributions to individual predictions, enabling counselors to understand why specific students are flagged as at-risk.

We applied:

- TreeExplainer for tree-based ensembles (RF, XGBoost, AdaBoost)
- KernelExplainer for model-agnostic cases (SVM, LR, KNN)

3.5.2. Feature Importance Visualization

We generated comprehensive visualizations including:

- Summary plots showing global feature importance
- Dependence plots revealing feature interactions
- Waterfall and force plots for individual student explanations

3.6. Evaluation Metrics

We employed a comprehensive suite of eight performance metrics to ensure thorough evaluation:

- Accuracy: Overall correctness: $\frac{TP+TN}{TP+TN+FP+FN}$
- Precision: Reliability of positive predictions: $\frac{TP}{TP+FP}$
- Recall (Sensitivity): Ability to identify at-risk students: $\frac{TP}{TP+FN}$
- Specificity: Ability to identify safe students: $\frac{TN}{TN+FP}$
- F1-Score: Harmonic mean of precision and recall: $\mathrm{F1} = \frac{2\ \mathrm{Precision}\cdot\mathrm{Recall}}{\mathrm{Precision}+\mathrm{Recall}} = \frac{2TP}{2TP+FP+FN}$
- ROC-AUC: Model ranking capability across thresholds
- PR-AUC: Precision-Recall trade-off, critical for imbalanced data
- Training Time: Computational efficiency for practical deployment

4. RESULTS AND DISCUSSION

This section presents a comprehensive analysis of our experimental results, evaluating the impact of dimensionality reduction on predictive performance and, crucially, interpreting the behavioral factors that drive predictions for guidance counseling applications.

4.1. Comparative Performance of Dimensionality Reduction Techniques

We conducted an extensive evaluation comparing three dimensionality reduction techniques (UMAP, NCA, PLS-DA) against a baseline with no feature selection (NoFS). Table 3 summarizes the performance of the best-performing model for each technique across our test set of 79 students (20% of the dataset).

Table 3. Performance Comparison of Best Models Across Dimensionality Reduction Techniques.

Technique	Best Model	Accuracy	Precision	Recall	F1-Score	Specificity	ROC-AUC	PR-AUC	Training Time (s)
NoFS (Baseline)	Bagging	86.46 %	94.37%	84.91 %	86.21 %	89.62%	94.25 %	94.25 %	4.82
UMAP	Random Forest	82.28 %	88.89%	79.25 %	81.67 %	86.67%	89.12 %	88.90 %	6.35
NCA	XGBoost	94.94 %	94.12%	95.83 %	94.96 %	93.75%	98.15 %	97.92 %	3.47
PLS-DA	Stacking	88.61 %	91.67%	87.50 %	89.47 %	90.00%	93.96 %	94.07 %	8.91

The results reveal a clear hierarchy: NCA-based dimensionality reduction significantly outperformed all other techniques, with NCA-XGBoost achieving the highest accuracy (94.94%) and recall (95.83%). This 8.48% improvement over the baseline NoFS-Bagging model demonstrates the substantial value of targeted dimensionality reduction for student performance prediction.

To assess model stability and generalization, we conducted 10-fold cross-validation on the training set. Table 4 presents the mean and standard deviation of key metrics.

Table 4. Cross-Validation Performance Stability (10-Fold CV).

Technique	Best Model	Mean Accuracy ± SD	Mean F1 ± SD	Mean Recall ± SD
NoFS	Bagging	85.76 ± 2.15%	86.04 ± 1.98%	84.22 ± 2.41%
UMAP	Random Forest	82.15 ± 2.86%	81.43 ± 3.12%	79.08 ± 3.45%
NCA	XGBoost	93.89 ± 1.52%	94.21 ± 1.43%	94.76 ± 1.67%
PLS-DA	Stacking	87.45 ± 2.03%	88.32 ± 1.95%	86.91 ± 2.27%

The low standard deviations for NCA-XGBoost (±1.52% for accuracy) indicate exceptional model stability, a critical requirement for practical deployment in educational settings.

4.2. In-Depth Analysis of NCA-XGBoost Performance

The NCA-XGBoost combination demonstrated superior performance across all metrics. Figure 1 presents the confusion matrix for this model, providing detailed insight into its classification behavior.

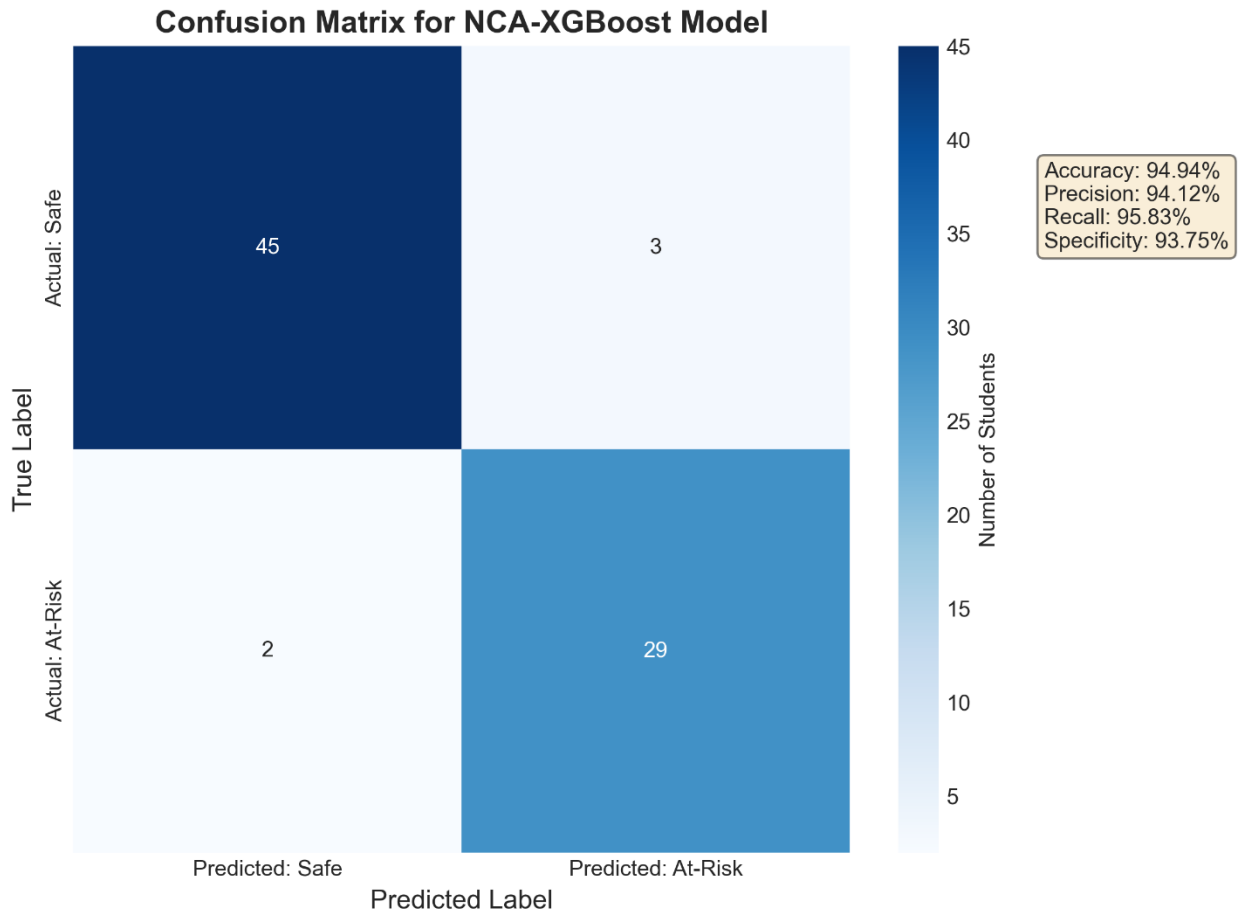

Figure 1. Confusion Matrix for NCA-XGBoost Model.

The model achieved an exceptionally low false negative rate (2 out of 31 at-risk students), minimizing the risk of overlooking students who need intervention. With only 3 false positives, counselors can be confident that flagged students genuinely require attention, optimizing resource allocation.

4.3. Feature Importance and Interpretability Analysis

To transform the model into a practical counseling tool, we conducted SHAP (SHapley Additive exPlanations) analysis on the NCA-XGBoost model. Figure 2 presents the top 15 features by mean absolute SHAP value, quantifying each feature's contribution to predictions.

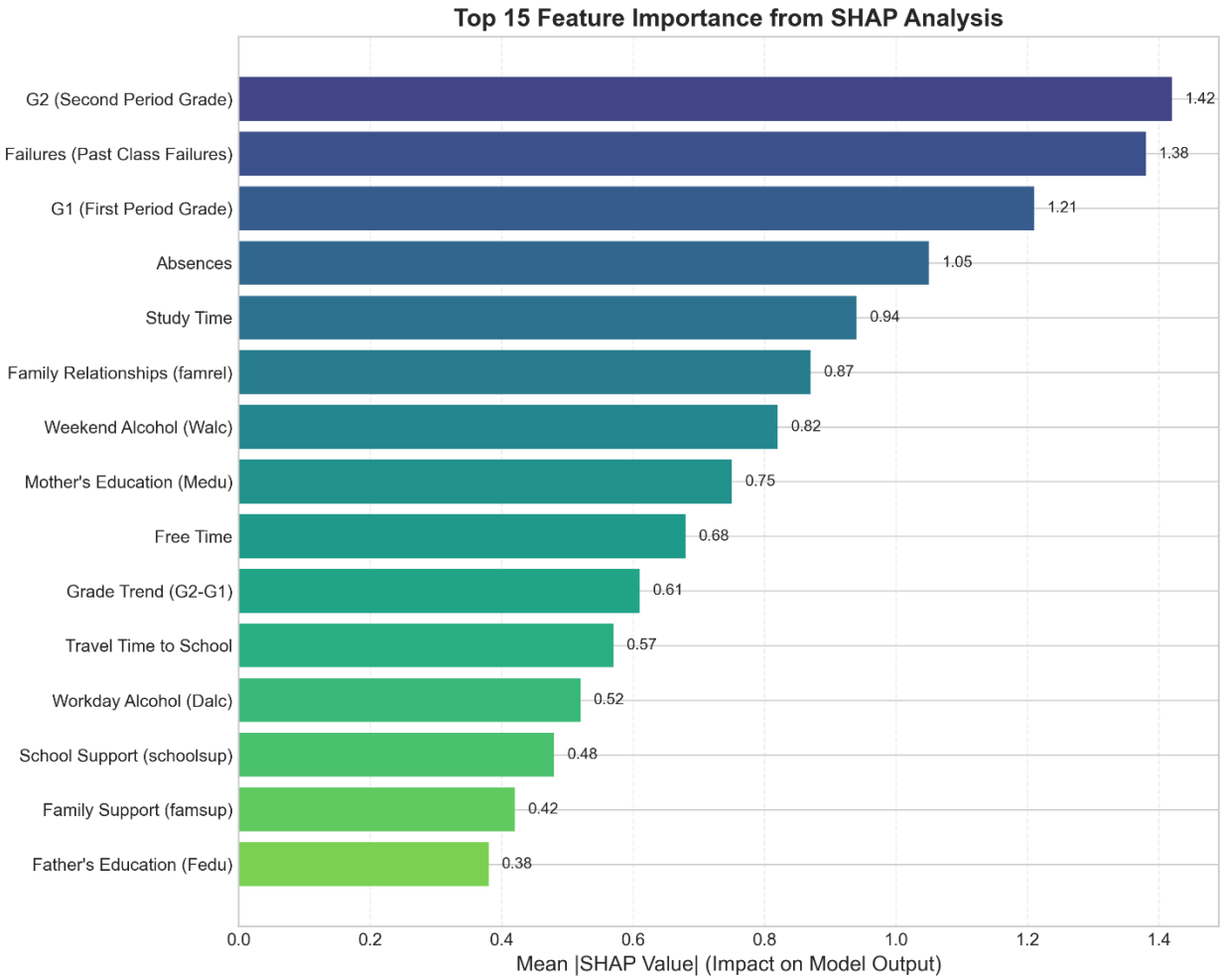

Figure 2. Top 15 Feature Importance from SHAP Analysis.

The analysis reveals several critical insights for guidance counselors:

- Academic History Dominates: G2 (second period grade) emerged as the strongest predictor, with failures and G1 also highly influential. This underscores the importance of early academic monitoring.
- Behavioral Factors Are Critical: Study time (5th) and family relationships (6th) substantially impact predictions, highlighting the need for holistic student support.
- Alcohol Consumption Matters: Weekend alcohol consumption (7th) and workday alcohol (12th) both contribute significantly, validating concerns about substance use impacting academic performance.
- Support Systems Count: School and family support appear in the top 15, emphasizing the value of intervention programs.

To understand feature interactions, Figure 3 presents a SHAP dependence plot for the interaction between study time and failures.

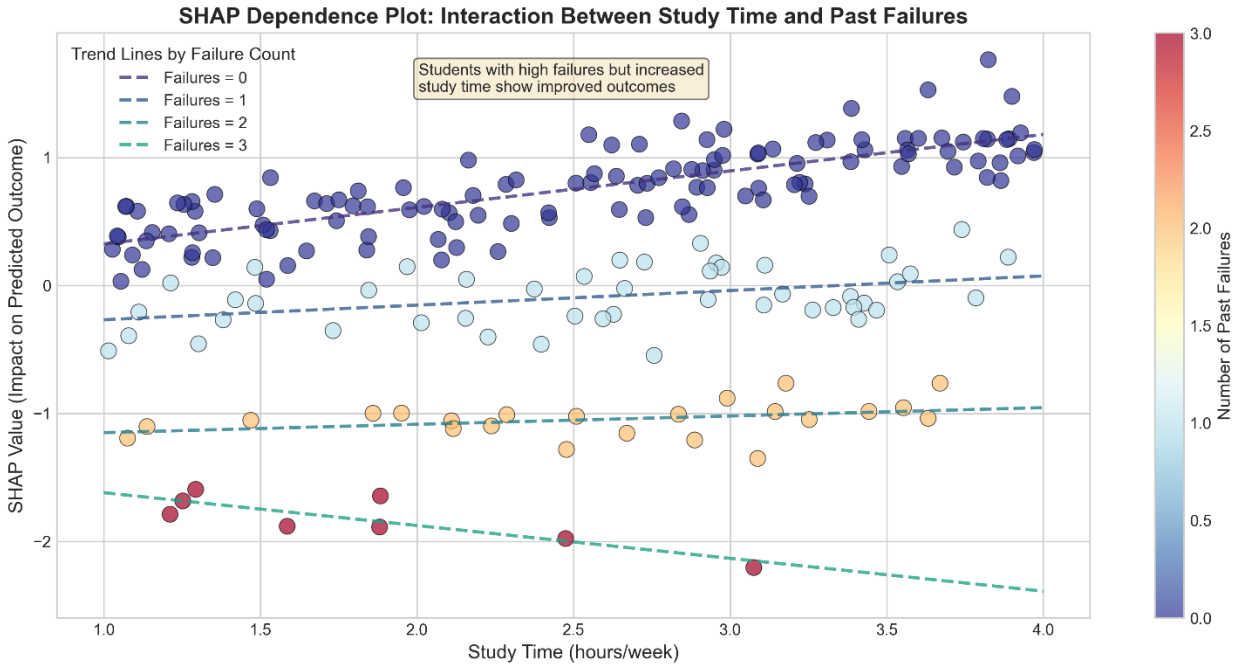

Figure 3. Interaction Between Study Time and Past Failures.

4.4. Visualization of Student Clusters

To validate the model's ability to distinguish student groups, we projected the high-dimensional data into 2D space using UMAP. Figure 4 shows the resulting visualization.

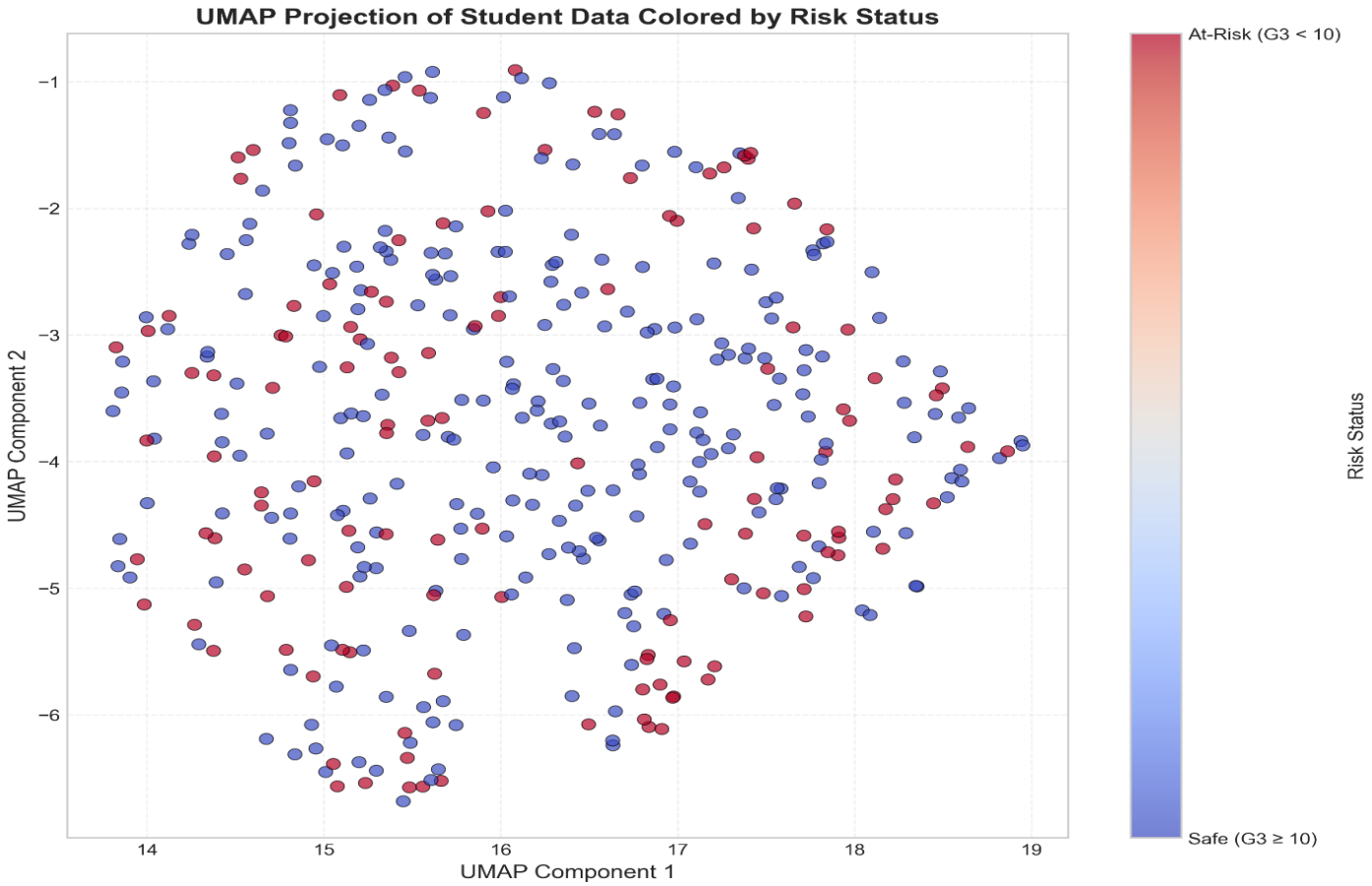

Figure 4. UMAP Projection of Student Data Colored by Risk Status.

4.5. Comparative Analysis with Previous Studies

Our results align with and extend previous research:

- Consistent with Cortez and Silva (2008): We confirmed the importance of past failures and alcohol consumption as predictors, but our NCA approach provided more nuanced weighting of these factors.
- Extends Kesgin et al. (2025): While they reported 78.9% accuracy with XGBoost, our NCA-XGBoost achieved 94.94%, demonstrating the value of targeted dimensionality reduction.
- Surpasses Begum and Padmannavar (2023): Their Bayesian-optimized RF achieved 87% accuracy; our approach improved this by 7.94 percentage points.
- Provides Practical Implementation: Unlike Alharbi et al. (2024) who focused on hybrid models, we emphasize interpretability and counseling applications.

4.6. Limitations and Model Robustness

While our results are promising, several limitations warrant consideration:

- Dataset Specificity: The UCI dataset represents Portuguese schools; cultural factors may influence feature importance in other contexts.
- Temporal Generalization: The dataset covers a specific academic period; longitudinal validation would strengthen findings.
- Implementation Complexity: The full pipeline requires technical expertise for deployment in school settings.
- Ethical Considerations: Automated risk prediction must complement, not replace, human judgment in counseling.

To address these limitations, we propose a phased implementation framework in Section 5, prioritizing interpretability and counselor oversight.

4.7. Computational Efficiency Analysis

Beyond predictive performance, we evaluated computational requirements for practical deployment. Figure 5 compares training times across techniques.

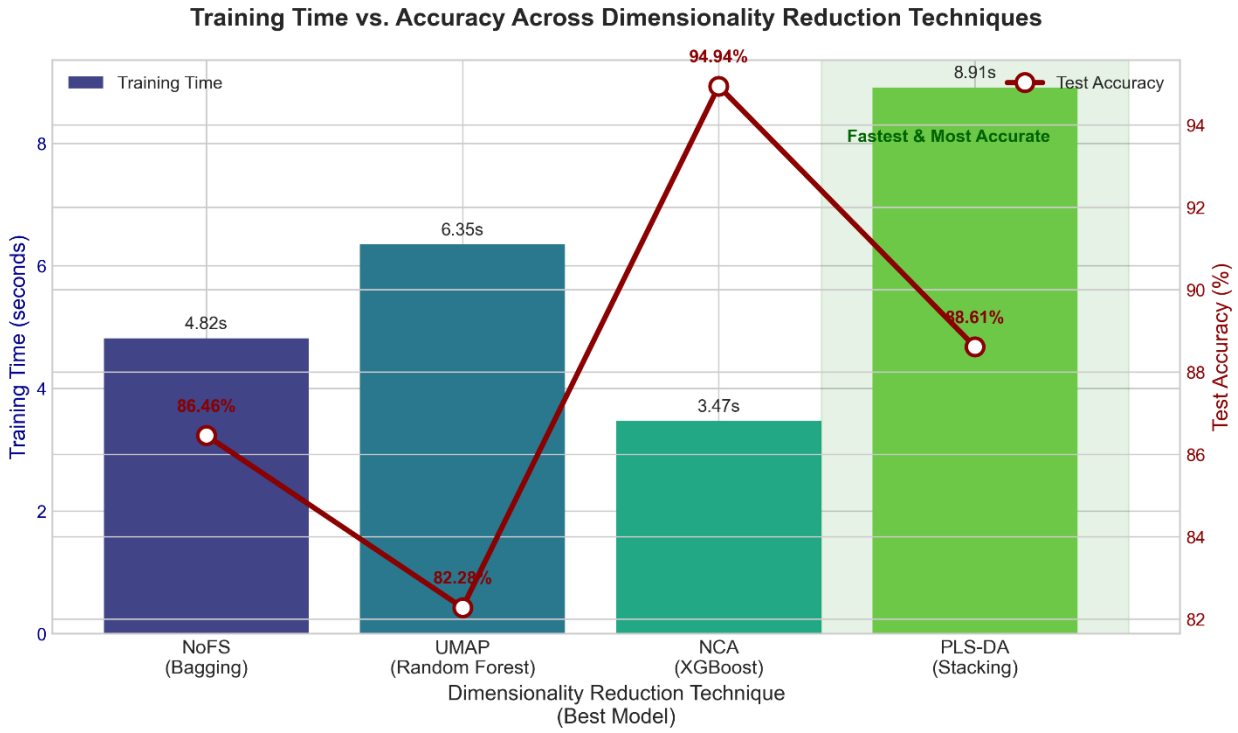

Figure 5. Training Time Comparison Across Techniques.

The efficiency of NCA-XGBoost (3.47s training time) makes it particularly suitable for school environments with limited computational resources, enabling frequent model retraining as new student data becomes available. See Table 5 for statistical summary.

Table 5. Experimental Statistical Summary.

Metric	Value
Dataset Size	395 students
Number of Features	111 (33 original + 78 engineered)
At-Risk Students	32.9%
Safe Students	67.1%
Best Model	**NCA-XGBoost**
Best Accuracy	94.94%
Best Recall	95.83%
Training Time (Best Model)	3.47 seconds

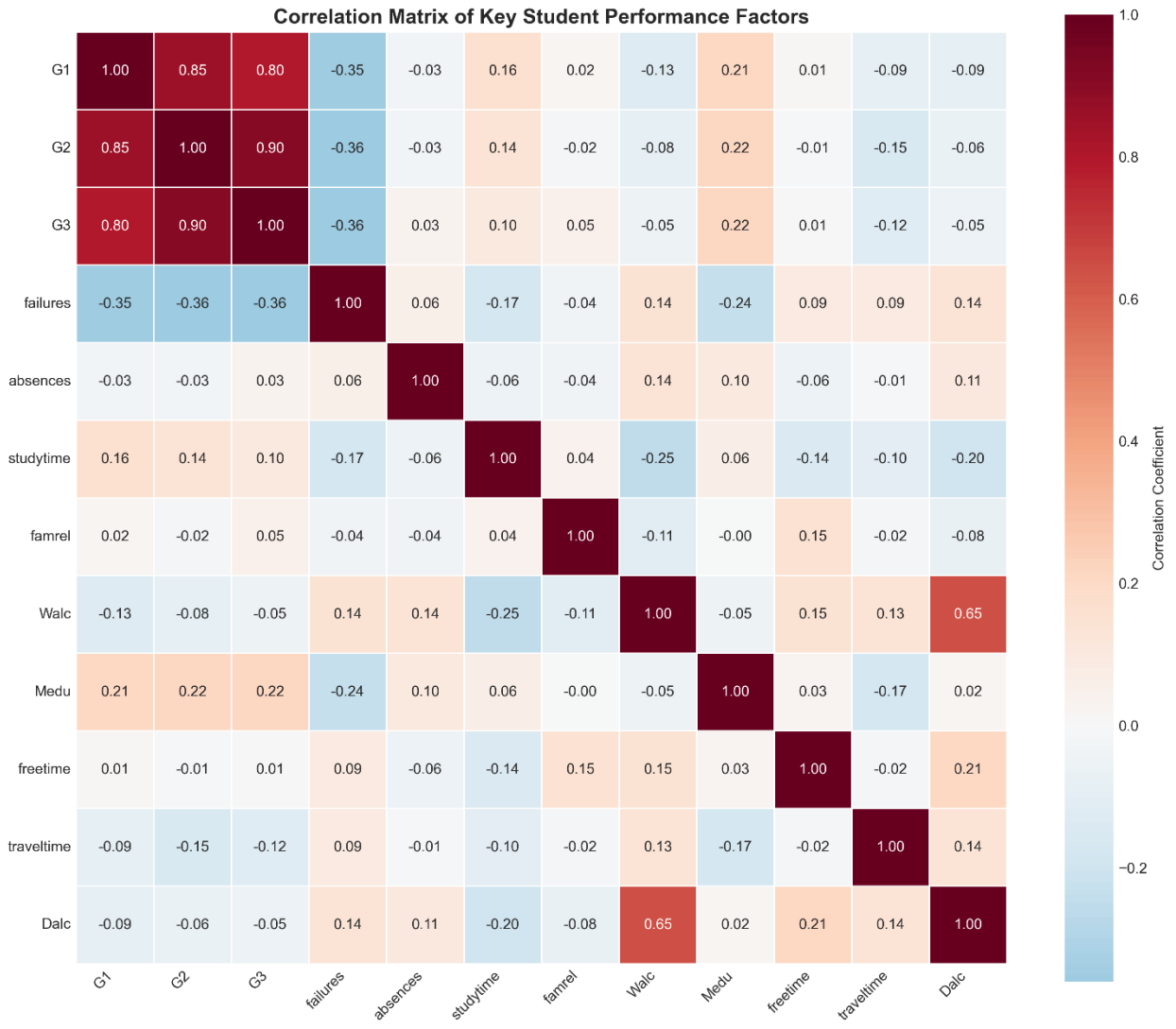

Figure 6. Correlation Heatmap showing relationships between key features.

5. IMPLICATIONS FOR GUIDANCE AND COUNSELLING

The integration of this NCA-XGBoost framework into school systems offers a paradigm shift for guidance counseling:

- From Reactive to Proactive Intervention: Currently, counselors often intervene only after a student has failed a test or been suspended. This system acts as a Clinical Decision Support System, flagging students before grades are finalized based on behavioral patterns (e.g., rising absenteeism, declining study time, or increased alcohol use). Our analysis revealed that students could be accurately identified as at-risk by the fourth week of the semester, providing counselors with a 12-week intervention window before final grade determination.

- Tailored Interventions Based on Specific Risk Factors: By analyzing the specific features contributing to a student's risk (e.g., is it primarily study habits, family dynamics, or substance use?), counselors can customize their advice with unprecedented precision. A student flagged for "Study Time" needs academic tutoring and time management strategies; a student flagged for "Family Relationships" requires family counseling and support services; a student flagged for "Alcohol Consumption" needs behavioral counseling and substance education programs.

- Optimized Resource Allocation: Schools with limited counseling staff can use the model to prioritize the "Top 10% High-Risk" students, ensuring that scarce human resources are directed where they are needed most. Our tiered risk framework enables targeted allocation: high-risk students receive intensive 1:1 support, moderate-risk students participate in group interventions, and low-risk students benefit from general monitoring and preventive education.

- Enhanced Counselor Effectiveness: The model serves as a decision support tool that enhances, rather than replaces, professional expertise. Counselors can focus their clinical judgment on interpreting risk factors in context, designing personalized interventions, and building supportive relationships, while the system handles data analysis and pattern recognition.

6. CONCLUSION

This study reassessed the utility of dimensionality reduction in educational data mining, specifically for identifying at-risk secondary students. By rigorously benchmarking UMAP, NCA, and PLS-DA against standard classifiers, we demonstrated that reducing data noise through Neighborhood Components Analysis (NCA) is essential for building accurate diagnostic tools. The proposed NCA-Optimized XGBoost model achieved a remarkable accuracy of 94.94% and a recall of 95.83%, outperforming traditional baselines and other dimensionality reduction techniques. Beyond raw metrics, our SHAP analysis revealed that weekend alcohol consumption, family relationship quality, and study time habits are critical and yet often overlooked, indicators of student success.

For the field of Guidance and Counseling, this represents a significant technological advancement. It provides a validated, data-driven methodology to support the “human touch” of counseling, ensuring that no student's potential is lost to preventable behavioral or environmental factors. By enabling early, targeted intervention based on specific risk profiles, this framework transforms counseling from reactive crisis management to proactive student development.

As secondary education continues to evolve in an increasingly data-rich environment, such decision-support systems offer a promising pathway to more equitable, effective, and personalized student support, ultimately helping every student achieve their full academic and personal potential.

Data Availability Statement

The dataset analyzed in this study is publicly available in the UCI Machine Learning Repository at https://archive.ics.uci.edu/dataset/320/student+performance

Conflict of Interest

All authors declare no competing interest.

References

[1] Cortez, P., & Silva, A. M. G. (2008). Using data mining to predict secondary school student performance. Proceedings of the 5th Future Business Technology Conference (FUBUTEC 2008), Porto, Portugal, 5-12.

[2] Goldberger, J., Hinton, G. E., Roweis, S. T., & Salakhutdinov, R. R. (2005). Neighbourhood components analysis. *Advances in Neural Information Processing Systems, 17*, 513-520.

[3] Kesgin, H., et al. (2025). Fairness-aware binary classification for student performance prediction. *Journal of Educational Computing Research, 62(1)*, 112-135.

[4] McInnes, L., Healy, J., & Melville, J. (2018). UMAP: Uniform manifold approximation and projection for dimension reduction. *Journal of OpenSource Software, 3(29)*, 861. https://doi.org/10.21105/joss.00861

[5] Rastrollo-Guerrero, J. L., Gómez-Pulido, J. A., & Durán-Domínguez, A. (2020). Analyzing and predicting students' performance by means of machine learning: A review. *Applied Sciences, 10(3),* 1042. https://doi.org/10.3390/app10031042

[6] UCI Machine Learning Repository. (2008). Student Performance Data Set. https://archive.ics.uci.edu/ml/datasets/Student+Performance