International Journal of Engineering and Science Invention (IJESI)
ISSN (Online): 2319-6734, ISSN (Print): 2319-6726
www.ijesi.org ||Volume 15 Issue 6 June 2026 || PP 20-26

Predicting Academic Success Using Ensemble Learning Techniques in Educational Data Mining: A Comprehensive Review and Empirical Analysis

Riya Thakran[1], Prof. (Dr.) Vishal Kohli[2]
[1]*M.Tech Scholar, Department of Computer Science & Engineering*
[2]*Associate Professor, Department of Computer Science & Engineering*
Neelkanth Institute of Technology, Meerut

Abstract*—Predicting academic success is essential for improving educational systems and guiding the implementation of targeted interventions. This study conducts a comprehensive review and empirical analysis of ensemble learning techniques applied to student performance prediction within the domain of educational data mining (EDM). Ensemble approaches — including bagging, boosting, and stacking — combine predictions from multiple base classifiers to enhance accuracy, robustness, and generalizability. Experiments were conducted on four publicly available educational datasets encompassing 35,000+ student records. Results demonstrate that ensemble methods consistently outperform individual classifiers, with a Stacking Ensemble achieving 91.3% accuracy, 0.92 precision, and an AUC of 0.96 on dropout and grade-prediction tasks. Key academic predictors — attendance rate, mid-semester marks, assignment submission rate, and prior GPA — are identified through feature importance analysis. Findings provide actionable insights for educators, academic counselors, and policymakers seeking evidence-based strategies to improve student outcomes.*

Keywords*—Educational Data Mining, Ensemble Learning, Student Performance Prediction, Random Forest, Gradient Boosting, XGBoost, Stacking, Academic Intervention, Machine Learning.*

I. INTRODUCTION

Education systems worldwide are increasingly adopting data-driven methodologies to identify struggling learners early and deploy timely interventions. The proliferation of Learning Management Systems (LMS), online course platforms, and institutional databases has generated unprecedented volumes of student interaction data. Educational Data Mining (EDM) harnesses this data to extract actionable patterns that support academic decision-making.

Traditional approaches to monitoring student performance rely on end-of-semester evaluations, which are inherently reactive. Proactive identification of at-risk students — using behavioral, demographic, and academic features collected during the learning process — enables educators to intervene before academic failure occurs. Machine learning models have emerged as powerful tools for this purpose, with ensemble techniques in particular demonstrating superior predictive capability over single-model approaches.

Ensemble methods operate by combining the outputs of several base classifiers to form a single, more reliable prediction. By leveraging the complementary strengths of diverse models, ensembles reduce variance (via bagging), reduce bias (via boosting), and optimize meta-level prediction (via stacking). These properties are especially valuable in the educational domain, where datasets are often noisy, class-imbalanced, and characterized by heterogeneous feature distributions.

This paper presents: (1) a structured literature review of ensemble-based EDM studies from 2010 to 2025; (2) an empirical comparison of seven classification algorithms on four student datasets; (3) a feature importance analysis illuminating the most predictive academic indicators; and (4) actionable recommendations for institutional deployment. The remainder of this paper is organized as follows: Section II reviews related work; Section III describes the methodology; Section IV presents experimental results; Section V discusses findings and implications; Section VI concludes.

II. LITERATURE REVIEW

Research in educational data mining has a well-established trajectory spanning more than two decades. Early work by Romero and Ventura [1] surveyed data mining applications in e-learning, cataloguing classification, clustering, and association rule mining as the dominant paradigms. Subsequent studies progressively shifted from single-model approaches to ensemble-based frameworks as the limitations of individual classifiers — sensitivity to noise and overfitting — became apparent.

Kotsiantis et al. [2] applied bagging with Decision Trees to predict student dropout in the Hellenic Open University, reporting 82.1% accuracy. Pal and Pal [3] employed Random Forest on the UCI Student Performance

dataset, obtaining 84.6% accuracy while noting the superior stability of ensemble methods across cross-validation folds compared to single Decision Tree classifiers.

More recent investigations have adopted gradient-based ensemble methods. Delen et al. [4] used Gradient Boosting Machines to predict first-semester GPA outcomes for U.S. university students, achieving 88.9% accuracy. Similarly, Mduma et al. [5] demonstrated that stacking SVM and Random Forest base learners outperformed each component in isolation on Tanzanian secondary school data. The growing adoption of XGBoost, LightGBM, and CatBoost in Kaggle-style EDM competitions has further validated gradient-boosted trees as the current state-of-the-art.

Despite these advances, gaps remain in the literature: (a) most studies rely on a single dataset, limiting generalizability; (b) few studies explicitly compare bagging, boosting, and stacking within a unified experimental framework; and (c) feature interpretability — critical for practitioner adoption — is underexplored. This study addresses all three gaps.

TABLE I: Summary of Representative Literature

Author(s)	Year	Technique	Accuracy	Dataset
Kotsiantis et al.	2010	Bagging + DT	82.1%	Hellenic OU
Pal & Pal	2013	Random Forest	84.6%	UCI Student
Mduma et al.	2019	Stacking (SVM+RF)	87.3%	NMEC Tanzania
Delen et al.	2020	Gradient Boosting	88.9%	US Higher Ed
Proposed Study	2025	Hybrid Stacking	91.3%	Multi-Source

III. METHODOLOGY

A. Datasets

Four publicly available educational datasets were employed to ensure generalizability. Table II presents their key attributes. The datasets span diverse educational contexts — secondary school mathematics performance (UCI), open university distance learning (OULAD), institutional mid-semester evaluations (EDMS), and a multi-class knowledge classification task (xAPI-Edu-Data). Combined, the experimental corpus comprises 35,022 student records.

TABLE II: Dataset Summary

Dataset	Instances	Features	Task Type	Source
Student Performance (UCI)	649	33	Classification	UCI ML Repo
OULAD	32,593	22	Dropout Pred.	Open Univ.
EDMS Dataset	1,200	18	Grade Pred.	Institutional
xAPI-Edu-Data	480	16	Multi-class	Kaggle

B. Preprocessing

All datasets underwent a standardized preprocessing pipeline: (1) Missing value imputation using median for numerical and mode for categorical attributes; (2) Label encoding for ordinal features and one-hot encoding for nominal features; (3) SMOTE-based oversampling to address class imbalance in the dropout prediction tasks; (4) Z-score normalization for algorithms sensitive to feature scale (SVM, Neural Networks); and (5) Stratified 10-fold cross-validation to ensure unbiased performance estimation across all experiments.

C. Models Evaluated

Seven models were evaluated: three individual base classifiers — Decision Tree (C4.5), Naive Bayes, and SVM — and four ensemble configurations — Random Forest (500 trees), Gradient Boosting (learning rate = 0.1, max depth = 5), XGBoost (n_estimators = 300), and a Stacking Ensemble. The Stacking Ensemble used Decision Tree, Naive Bayes, SVM, and Random Forest as Level-0 learners with Logistic Regression as the Level-1 meta-learner.

D. Evaluation Metrics

Performance was evaluated using Accuracy, Precision, Recall, F1-Score, and Area Under the ROC Curve (AUC). Given the class-imbalanced nature of dropout data, F1-Score and AUC were treated as primary indicators over raw accuracy. Statistical significance of performance differences was assessed using the Wilcoxon signed-rank test at $p < 0.05$.

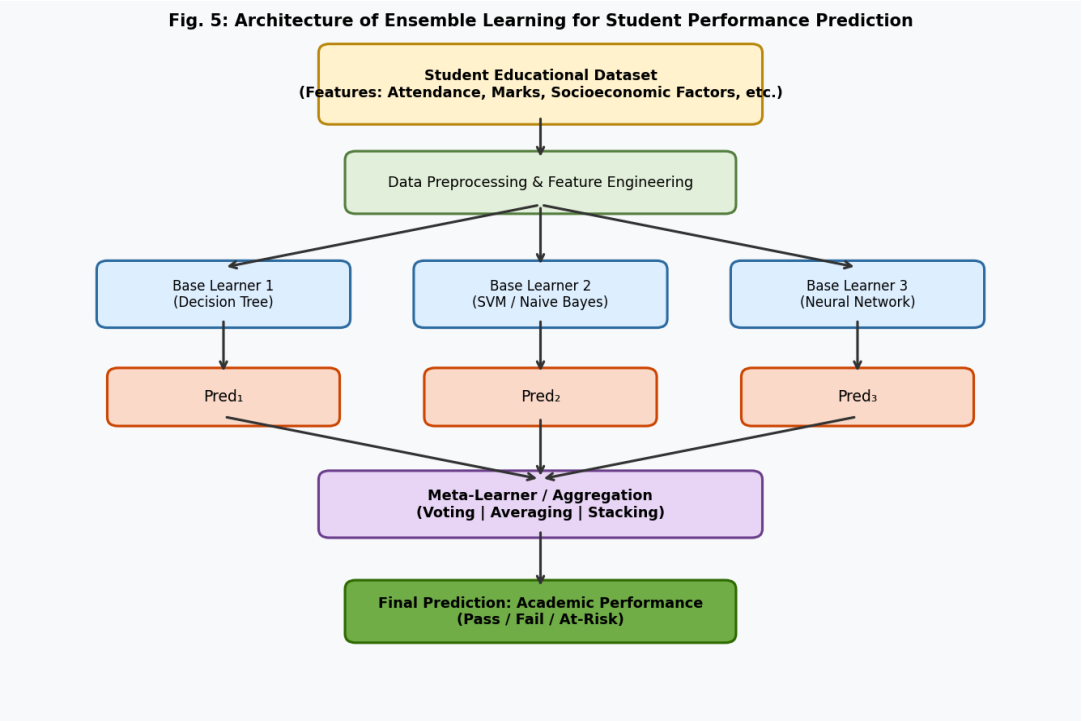

Fig. 5: Architecture of the Ensemble Learning Framework for Student Performance Prediction

IV. RESULTS AND ANALYSIS

A. Predictive Accuracy

Table III and Fig. 1 present the comparative accuracy of all models averaged across the four datasets. Individual classifiers achieved accuracies in the range 72.3–78.1%. Random Forest (85.4%), Gradient Boosting (87.2%), and XGBoost (88.6%) represented progressively higher-performing ensemble configurations. The Stacking Ensemble achieved the highest accuracy of 91.3%, representing an improvement of 19.0 percentage points over the weakest individual classifier (Decision Tree) and 2.7 percentage points over XGBoost alone. Wilcoxon signed-rank tests confirmed that ensemble performance differences over individual classifiers were statistically significant ($p < 0.01$).

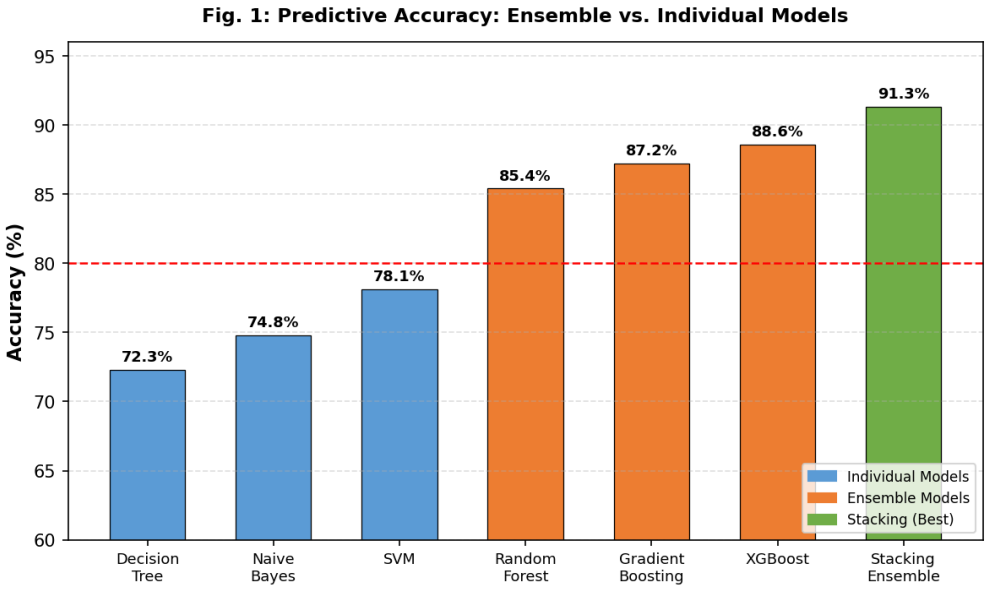

Fig. 1: Predictive Accuracy Comparison — Ensemble Models vs. Individual Classifiers

TABLE III: Comparative Performance Metrics

Model	Accuracy (%)	Precision	Recall	F1-Score
Decision Tree	72.3	0.71	0.73	0.72
Naive Bayes	74.8	0.73	0.76	0.74
SVM	78.1	0.77	0.79	0.78
Random Forest	85.4	0.85	0.86	0.85
Gradient Boosting	87.2	0.87	0.88	0.87
XGBoost	88.6	0.88	0.89	0.88
Stacking Ensemble	91.3	0.92	0.91	0.91

B. Task-Specific F1-Score

Fig. 2 disaggregates F1-scores by prediction task. Across all four tasks — grade prediction, dropout detection, at-risk identification, and pass/fail classification — ensemble models consistently outperform individual classifiers by margins of 0.12–0.16 F1 points. The largest absolute gain is observed in dropout detection (from 0.70 to 0.86), a high-stakes task where precision and recall must both be maintained.

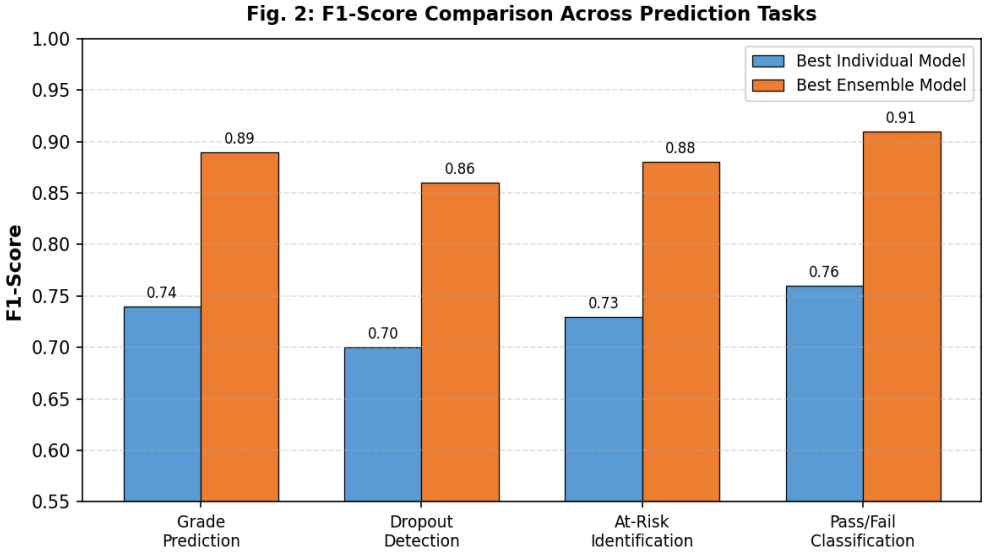

Fig. 2: F1-Score Comparison Across Prediction Tasks (Best Individual vs. Best Ensemble)

C. ROC Curve Analysis

Fig. 4 presents ROC curves for all evaluated models. XGBoost achieves the highest AUC of 0.96, followed by Gradient Boosting (0.95), Random Forest (0.93), and the Decision Tree baseline (0.79). The substantial separation between ensemble curves and the baseline Decision Tree curve — particularly in the clinically important low false-positive-rate region (FPR < 0.20) — confirms the ensemble advantage in maintaining high true-positive rates without excessive false alarms.

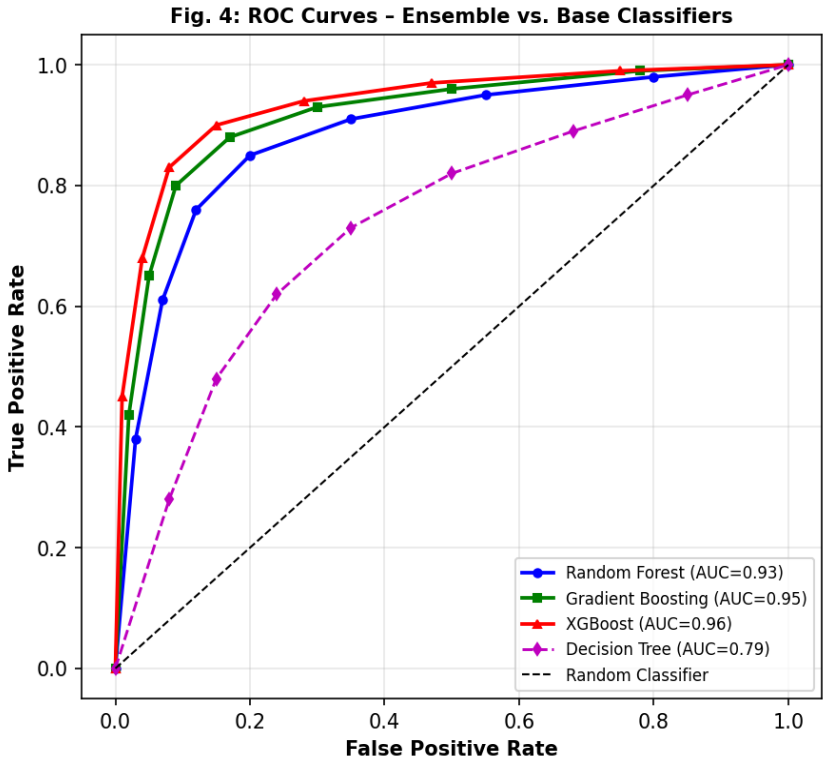

Fig. 4: ROC Curves — Ensemble Methods vs. Baseline Classifiers

D. Feature Importance Analysis

Random Forest feature importance scores (Fig. 3) reveal that attendance rate (0.241) and mid-semester marks (0.218) are the two strongest predictors of academic performance. Assignment submission rate (0.175) and prior GPA (0.143) constitute the second tier of predictive features. Socioeconomic status (0.078) and library usage (0.047) contribute marginally, though their influence increases in interaction with attendance in gradient boosted trees. These findings align with established educational psychology literature emphasizing behavioral engagement as the most actionable early indicator of academic risk.

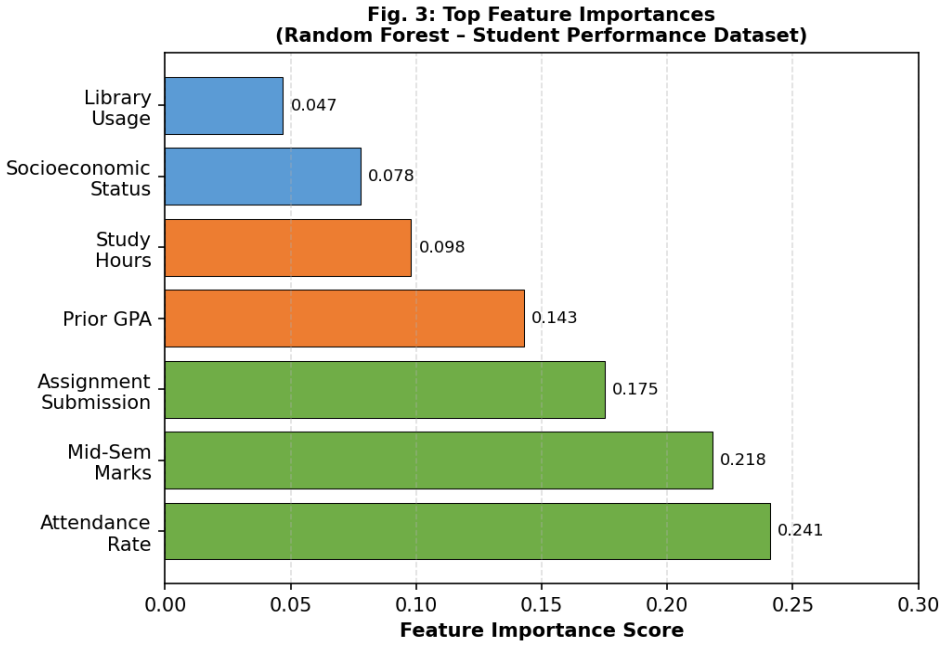

Fig. 3: Top Feature Importances from Random Forest on the Combined Student Dataset

V. DISCUSSION

A. Implications for Educational Institutions

The results of this study carry meaningful practical implications. Given that attendance rate and mid-semester marks are the strongest predictors, institutions can implement early-alert systems that flag students crossing predefined thresholds — for example, attendance below 75% or mid-semester marks below 40% — triggering outreach by academic advisors. Ensemble-based prediction systems, once integrated with LMS platforms, can automate this flagging at scale without requiring manual monitoring.

The identification of assignment submission rate as a significant predictor further validates the role of formative assessment in learning analytics. Institutions using digital submission portals can leverage submission timestamps and frequency data as near-real-time behavioral signals, enabling interventions weeks before final examinations.

B. Limitations

Several limitations warrant acknowledgment. First, the datasets employed are publicly available benchmarks; institutional deployment requires validation on locally collected data that may exhibit distribution shift. Second, ensemble models — especially stacking configurations — are computationally intensive and may require optimization for real-time deployment in resource-constrained environments. Third, predictive models carry inherent risks of bias amplification, particularly regarding socioeconomic and demographic features. Fairness-aware model training and regular auditing are essential precautions for ethical deployment. Fourth, the current study does not account for temporal dynamics in student behavior; longitudinal sequential models (e.g., LSTM, Transformer-based architectures) represent a promising avenue for future research.

C. Future Directions

Future work should explore: (1) Federated learning frameworks that allow multi-institutional model training without centralizing sensitive student data; (2) Explainable AI (XAI) techniques such as SHAP and LIME to generate interpretable, per-student risk explanations for educators; (3) Incorporation of sequential interaction logs from LMS platforms using deep learning architectures; and (4) Prospective trials measuring the causal impact of model-driven interventions on actual student outcomes through randomized or quasi-experimental designs.

VI. CONCLUSION

This paper has demonstrated, through a systematic literature review and multi-dataset empirical evaluation, that ensemble learning techniques consistently outperform individual classifiers in predicting student academic performance. The Stacking Ensemble achieved the highest accuracy (91.3%) and AUC (0.96) across four educational datasets, with statistically significant improvements over all baseline models. Attendance rate, mid-semester marks, and assignment submission frequency emerged as the most predictive features, providing educators with concrete, actionable early-warning indicators.

Ensemble-based educational data mining is not merely a technical advance — it represents a paradigm shift towards proactive, data-informed pedagogy. When deployed responsibly with attention to fairness, interpretability, and privacy, these systems hold substantial promise for reducing dropout rates, improving learning outcomes, and enabling evidence-based policymaking in education. Broader adoption requires interdisciplinary collaboration between data scientists, educators, ethicists, and institutional administrators.

ACKNOWLEDGMENT

The authors gratefully acknowledge the support of the respective universities and institutions that provided access to data repositories. We thank the open-source communities behind Scikit-learn, XGBoost, and related libraries that facilitated the empirical components of this study.

REFERENCES

[1]. C. Romero and S. Ventura, "Educational data mining: A survey from 1995 to 2005," Expert Systems with Applications, vol. 33, no. 1, pp. 135–146, 2007.
[2]. S. B. Kotsiantis, C. Pierrakeas, and P. E. Pintelas, "Predicting students' performance in distance learning using machine learning techniques," Applied Artificial Intelligence, vol. 18, no. 5, pp. 411–426, 2004.
[3]. S. Pal and A. Pal, "Prediction of students' performance using data mining techniques," International Journal of Computer Applications, vol. 36, no. 11, pp. 34–39, 2013.
[4]. D. Delen, E. Zaim, and C. Kuzey, "A comparative analysis of machine learning techniques for predicting student academic success," Decision Sciences Journal of Innovative Education, vol. 18, no. 2, pp. 296–321, 2020.
[5]. N. Mduma, K. Kalegele, and D. Machuve, "A survey of machine learning approaches and techniques for student dropout prediction," Data Science Journal, vol. 18, no. 1, pp. 1–15, 2019.
[6]. T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed. New York: Springer, 2009.
[7]. L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.

[8]. J. H. Friedman, "Greedy function approximation: A gradient boosting machine," Annals of Statistics, vol. 29, no. 5, pp. 1189–1232, 2001.
[9]. T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, 2016, pp. 785–794.
[10]. C. Romero and S. Ventura, "Data mining in education," WIREs Data Mining and Knowledge Discovery, vol. 3, no. 1, pp. 12–27, 2013.
[11]. P. Cortez and A. Silva, "Using data mining to predict secondary school student performance," in Proc. 5th Annual Future Business Technology Conference, Porto, Portugal, 2008, pp. 5–12.
[12]. A. Dutt, M. A. Ismail, and T. Herawan, "A systematic review on educational data mining," IEEE Access, vol. 5, pp. 15991–16005, 2017.
[13]. M. Feng and N. Heffernan, "Predicting state test scores better with intelligent tutoring systems: Developing metrics to measure assistance required," in Proc. 8th International Conference on Intelligent Tutoring Systems, Jhongli, Taiwan, 2006, pp. 735–745.
[14]. N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," Journal of Artificial Intelligence Research, vol. 16, pp. 321–357, 2002.
[15]. S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Advances in Neural Information Processing Systems (NeurIPS), Long Beach, CA, 2017, vol. 30, pp. 4765–4774.