Atlas Journal of Science Education 2025, pp. 113–119
https://doi.org/10.5147/ajse.263

Published in the United States of America
https://www.atlas-publishing.org

Using Machine Learning for Early Identification of At-Risk Students in Undergraduate Biology Courses

Danielle Graham*, Justin Graham, Willietta Gibson, Khalid Lodhi, Jiazheng Yuan, Lieceng Zhu, and My Abdelmajid Kassem

Plant Genomics and Bioinformatics Lab, Department of Biological and Forensic Sciences, Fayetteville State University, Fayetteville, NC 28301, USA

Received: July 4, 2025 / Accepted: August 2, 2025

Abstract

The increasing availability of educational data has created new opportunities to apply machine learning (ML) for predicting student outcomes, particularly in STEM disciplines where early identification of academic risk is essential for improving retention and performance. This study investigates the use of supervised ML algorithms to predict final exam performance in undergraduate biology courses, leveraging earlier assessment scores—Exam 1, Midterm, and Exam 3—as predictive features. The dataset comprises 500 student records drawn from five biology courses (BIOL150, BIOL210, BIOL310, BIOL330, and BIOL499), representing a spectrum of instructional levels from introductory to advanced capstone experiences. Four ML models were implemented and compared: Linear Regression, Random Forest, Support Vector Regressor (SVR), and XGBoost. These models were evaluated using standard regression metrics, including Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and R^2. Among baseline models, Linear Regression demonstrated the highest explanatory power ($R^2 = 0.39$), while tree-based models showed competitive performance and further improvement after hyperparameter tuning. Feature importance analysis using both tree-based measures and SHAP (SHapley Additive exPlanations) values revealed that Midterm and Exam 3 scores were consistently the strongest predictors of final exam performance, whereas Exam 1 had lower predictive influence. The findings suggest that mid-course assessments provide a valuable window for identifying students at risk of underperformance, allowing for timely, targeted interventions. The use of interpretable ML models further enables actionable feedback for educators, aligning predictive outcomes with pedagogical decisions. By focusing specifically on biology—a domain underrepresented in educational data mining—this study contributes a subject-specific framework for academic early warning systems. The results support broader adoption of data-driven approaches in higher education and provide a scalable model for integrating predictive analytics into biology instruction and curriculum planning.

Keywords: Machine learning (ML) in education, Educational data mining (EDM), Biology performance prediction, STEM analytics, Student outcome forecasting, SHAP interpretability, Academic early warning systems, Predictive modeling in higher education, interpretable machine learning, course-level analytics, STEM retention.

* Corresponding author: jwgraham01@uncfsu.edu

Atlas Journal of Science Education - ISSN 2158-9224. Published By Atlas Publishing, LLC (www.atlas-publishing.org)

1. Introduction

Predicting student performance is a cornerstone of modern educational research, particularly as institutions strive to foster academic success and improve retention rates. Early identification of at-risk students provides opportunities for timely intervention, allowing educators to tailor instructional strategies and support systems to individual learner needs. In the context of higher education, such predictions have far-reaching implications for improving learning outcomes, personalizing instruction, and optimizing institutional resources (Romero and Ventura, 2020).

In recent years, the interdisciplinary field of Educational Data Mining (EDM) has gained momentum, leveraging machine learning (ML) techniques to extract actionable insights from educational datasets. These approaches have been applied to predict academic achievement, detect disengagement, and personalize learning pathways (Ahmad et al., 2015; Baker et al., 2016; Costa et al., 2017; Roy and Garg, 2017; Dutt et al., 2017; He et al., 2018). Supervised learning methods such as decision trees, support vector machines, and neural networks have shown effectiveness in modeling student success and dropout risk (Aman et al., 2019), while unsupervised clustering has been used to group learners by behavioral patterns. Moreover, recent developments in natural language processing (NLP) and reinforcement learning have enabled advanced applications in adaptive learning and real-time feedback systems (Aggarwal et al., 2024).

Despite these advances, many applications of EDM have focused on general academic contexts or quantitative STEM fields like mathematics, engineering, and computer science (Elbadrawy and Karypis, 2016; Huang and Fang, 2013). In contrast, biology education remains underrepresented in predictive modeling research, even though it presents distinct pedagogical challenges. Biology courses often require both conceptual understanding and practical laboratory skills, delivered in a hierarchical progression from introductory principles to advanced applications. Success in these courses is contingent upon a solid foundation in earlier topics, which can compound difficulties as students' progress.

Traditional academic support systems in biology—such as manual grade tracking or reactive feedback—often fail to detect struggling students until after critical assessments, such as final exams. These methods are insufficient for scalable, real-time intervention and may miss key opportunities to support learners during pivotal moments in the course. Furthermore, most existing models fail to incorporate domain-specific variables, such as exam scores from biology topics, which are more indicative of students' evolving understanding than generic features like overall GPA or attendance (Badr et al., 2016).

To address these limitations, this study applies multiple ML algorithms to predict final exam performance based on prior assessment scores (Exam 1, Midterm, and Exam 3) in a series of biology courses: BIOL150 (Principles of Biology I), BIOL210 (General Botany), BIOL310 (Principles of Genetics), BIOL330 (Microbiology and Immunology), and BIOL499 (Senior Capstone Experience). These courses span a range of complexity and learning outcomes, from foundational knowledge to advanced synthesis and application.

The primary objective of this research is to apply machine learning techniques to predict final exam performance among undergraduate biology students using data from earlier assessments, including Exam 1, Midterm, and Exam 3. By doing so, the study aims to identify students at risk of underperforming early in the academic term, thereby enabling instructors and academic support teams to implement timely and targeted interventions. Early prediction not only supports more efficient allocation of resources but also empowers educators to personalize instruction and improve learning outcomes. Furthermore, this work seeks to demonstrate the viability and effectiveness of scalable, data-driven tools that can be integrated into undergraduate STEM education to enhance academic planning and decision-making processes.

Beyond its practical objectives, this study makes several novel contributions to the field of educational data mining. First, it offers subject-specific insights by applying machine learning algorithms within the context of biology education—a domain that has received comparatively less attention in predictive modeling research. Second, it provides a comparative evaluation of multiple machine learning (ML) algorithms, including Linear Regression (LR, Kenney and Keeping, 1962), Random Forest (RF, Breiman, 2001), Support Vector Machines (SVM, Cortes and Vapnik, 1995), and Gradient Boosting (XGBoost, Chen and Guestrin, 2016), allowing for a nuanced understanding of their respective strengths and limitations in educational contexts. Third, the study delivers actionable analytics by linking model predictions to pedagogical strategies, thereby moving beyond mere forecasting to inform real-world educational practice. Finally, the methodological framework developed in this research is inherently scalable and can be adapted to other STEM disciplines or courses with similar assessment structures, broadening its applicability and impact.

Collectively, these aims and contributions position the study at the intersection of data science and discipline-specific pedagogy. By demonstrating how predictive analytics can be used not only to forecast academic outcomes but also to inform teaching strategies and support systems, this research addresses key gaps in both the educational data mining literature and the practical realities of undergraduate biology instruction.

2. Methodology

2.1 Dataset Description

This study draws on data collected from five undergraduate biology courses spanning a progression from foundational to advanced topics. The courses—BIOL150 (Principles of Biology I), BIOL210 (General Botany), BIOL310 (Principles of Genetics), BIOL330 (Microbiology and Immunology), and BIOL499 (Senior Capstone Experience)—represent key milestones in the biology curriculum, providing a comprehensive sample of students at various academic levels. A summary of these courses, including content and annual enrollment, is presented in Table 1.

The dataset consists of 500 anonymized student records, with 100 students sampled per course. Each record includes individual scores from Exam 1, Midterm, Exam 3, and the Final Exam (Exam 4). These assessments are structured to evaluate students' cumulative understanding of course material, with Exam 1 typically testing early foundational concepts, Midterm covering intermediate topics, and Exam 3 focusing on more advanced content. The Final Exam is comprehensive in scope and serves as the primary target variable for prediction. Additional contextual features such as student demographics or attendance were not included in this version of the analysis, although they may offer additional predictive power in future studies (Costa et al., 2017).

The grade distributions across courses vary, reflecting differences in course difficulty and learning expectations. For example, students in BIOL499, a research-intensive capstone course, tend to show higher consistency in performance, whereas foundational courses like BIOL150 display greater variability. This diversity in the dataset provides a robust foundation for predictive modeling, allowing algorithms to generalize across multiple levels of content complexity and student ability.

2.2 Feature Selection

The model features were selected based on their pedagogical relevance and their potential to serve as reliable indicators of student learning progression. The three input variables used are Exam 1, Midterm, and Exam 3 scores. These exams are typically spaced throughout the semester and collectively assess students' grasp of course content at multiple stages. Prior research has shown that such performance-based features are valuable predictors in educational models, particularly when structured assessments are used to monitor academic progress over time (Huang and Fang, 2013; Aman et al., 2019).

The output variable for all predictive models is the Final Exam (Exam 4) score. As a cumulative measure of course mastery, the final exam provides a valid target for evaluating students' overall performance and readiness to advance in the curriculum.

2.3 Machine Learning Models

To evaluate the potential of machine learning for predicting final exam performance, this study implemented and compared four supervised regression models. Each model was selected based on its strengths in handling educational data and its balance between interpretability and predictive performance.

Linear Regression was used as a baseline model to quantify the linear relationship between input features and the final exam score. Due to its simplicity and transparency, linear regression is often favored in educational contexts for its interpretability by instructors and stakeholders (Badr et al., 2016).

Random Forest Regression, an ensemble-based method that aggregates

Atlas Journal of Science Education - ISSN 2158-9224. Published By Atlas Publishing, LLC (www.atlas-publishing.org)

Table 1. Descriptions and enrollment of undergraduate biology courses included in this study.

Course #	Course Title	Course Description	Enrolled/Year
BIOL150	Principles of Biology I	The study of the major principles relating to the nature of organisms, with emphasis on molecular, cellular, genetic, and evolutionary concepts, and with two (2) hours of lab consisting of experiments on the analysis of the chemistry of cellular and related materials. Prerequisite: MATH 121 or higher level of MATH.	638
BIOL210	General Botany	An introduction to the morphology, anatomy, physiology, reproduction, taxonomy, and ecology of higher plants, fungi, and algae, with two (2) hours of lab consisting of observation interpretation of the morphology and structure relating to the function, identification, and adaptation of higher plants, fungi, and algae. Prerequisite: BIOL 200 and BIOL 200L.	116
BIOL310	Principles of Genetics	A study of the fundamental concepts and principles of Mendelian genetics, cytogenetics, molecular genetics, and the application of genetic technologies, with two-hour laboratory experiments and exercises designed to reinforce and deepen students' understanding of basic concepts and principles of genetics and to provide an opportunity to obtain hands-on experimental and problem-solving skills. Prerequisite: BIOL 200 and BIOL 200L.	95
BIOL330	Microbiology & Immunology	An introduction to the structure, physiology, ecology, and immunological host relationships of prokaryotes and other microorganisms, with two (2) hours of lab consisting of applications of microbiological and immunological techniques. Prerequisite: BIOL 200 and CHEM 141 and CHEM 141L and CHEM 161 and CHEM 161L.	88
BIOL499	Senior Capstone Experience	Senior Capstone Experience is a course involving guided scientific research, field studies, and other special assignments. The course serves to give students guided experience in scientific research. Students will be trained to critically review literature, design and conduct experiments, and present their findings in a research paper and presentation. Prerequisite: Senior status or instructor permission.	54

Table 2. Descriptive statistics of exam scores across all courses.

	Exam1	Midterm	Exam3	Final Exam
Count	500.00	500.00	500.00	500.00
Mean	88.94	85.94	88.09	88.34
Std	12.13	11.48	12.69	10.86
Min	0.00	23.50	0.00	0.00
25%	85.92	82.00	84.00	84.00
50%	92.46	88.14	92.00	90.00
75%	96.00	92.21	96.00	95.50
Max	105.00	108.00	105.00	106.00

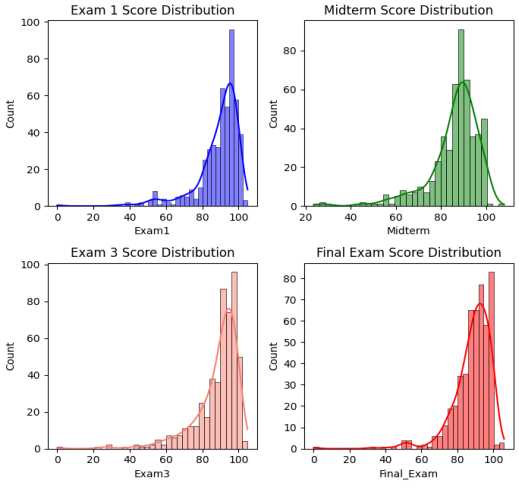

Figure 1. Distribution of Exam Scores Across All Courses.

multiple decision trees, was employed to capture complex, non-linear relationships among features. It is particularly effective at modeling feature interactions and is robust against overfitting when tuned appropriately (Breiman, 2001).

Support Vector Regression (SVR) was selected for its ability to model both linear and non-linear relationships in high-dimensional feature spaces. SVR works by finding the optimal hyperplane that minimizes error within a specified margin, making it suitable for datasets with complex and less obvious feature patterns (Cortes and Vapnik, 1995).

Finally, XGBoost (Extreme Gradient Boosting) was included due to its superior performance in many regression tasks involving structured data. XGBoost builds trees sequentially, correcting errors from prior iterations, and includes regularization parameters that improve model generalizability (Chen and Guestrin, 2016). Its speed and predictive accuracy have made it a popular choice in educational data mining competitions and real-world deployments.

Together, these four models provide a comprehensive basis for evaluating how different algorithmic strategies perform in predicting final biology exam outcomes.

2.4 Evaluation Metrics

Model performance was assessed using three standard regression metrics: Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and R-squared (R^2). MAE offers an intuitive measure of average prediction error, treating all deviations equally. RMSE penalizes larger errors more heavily, making it useful when outliers are particularly impactful in decision-making. R^2, or the coefficient of determination, quantifies the proportion of variance in the dependent variable that is explained by the model, providing an overall measure of model fit (Geron, 2019). The combination of these metrics allows for a balanced assessment of both accuracy and explanatory power across models.

2.5 Tools and Frameworks

All analyses were conducted using Python 3.11 (Van Rossum and Drake, 1995), leveraging several open-source libraries tailored for machine learning and data analysis. Scikit-learn (Pedregosa et al., 2011) was used for model training, cross-validation, and performance evaluation, offering consistent APIs across all models. Pandas (McKinney, 2010) and NumPy (Harris et al., 2020) were employed for data manipulation and feature engineering. XGBoost, an optimized gradient boosting framework, was used for high-performance model training. Matplotlib (Hunter, 2007) and Seaborn (Waskom, 2021) were utilized for data visualization, including distribution plots, correlation heatmaps, and feature importance graphs. Jupyter Notebook (Kluyver et al., 2016) was used as the interactive environment. These tools collectively ensured reproducibility, flexibility, and scalability throughout the research workflow.

3. Results

4.1 Descriptive Statistics and Distribution of Scores

The dataset comprised 500 student records across five undergraduate biology courses: BIOL150, BIOL210, BIOL310, BIOL330, and BIOL499 (Table 1). These courses span introductory to advanced content, offering a diverse range of student performance metrics. Descriptive statistics for all assessments—Exam 1, Midterm, Exam 3, and Final Exam—are summarized in Table 2. Scores were generally high across assessments, with mean values clustering in the mid- to high-80s. Exam 1 had the highest mean (88.94), while Midterm scores were slightly lower on average (85.94).

Visual inspection of score distributions (Figure 1) revealed moderate right-skewness, with most students scoring in the 80–100 range, but a small number of low-performing outliers. Notably, Final Exam scores exhibited less dispersion than earlier assessments, suggesting performance convergence by the end of the course.

4.2 Correlation Among Assessments

To explore the relationships between earlier assessments and final exam performance, a Pearson correlation matrix was computed (Figure 2). Moderate correlations were observed among all exams. The Midterm exhibited the strongest correlation with Exam 1 (r = 0.64) and with the Final Exam (r = 0.54). Exam 3 showed a slightly higher correlation with the Final Exam (r = 0.55) compared to Exam 1 (r = 0.49). These values suggest that while prior assessments are informative, they do not fully explain Final Exam outcomes, supporting the case for more complex predictive modeling.

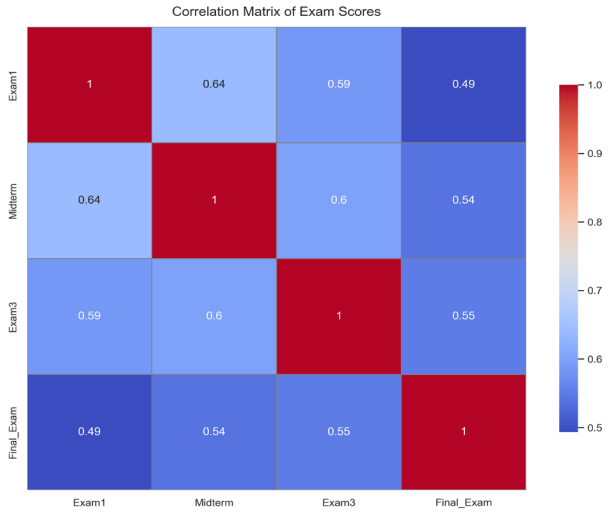

Figure 2. Correlation Matrix of Exam Scores

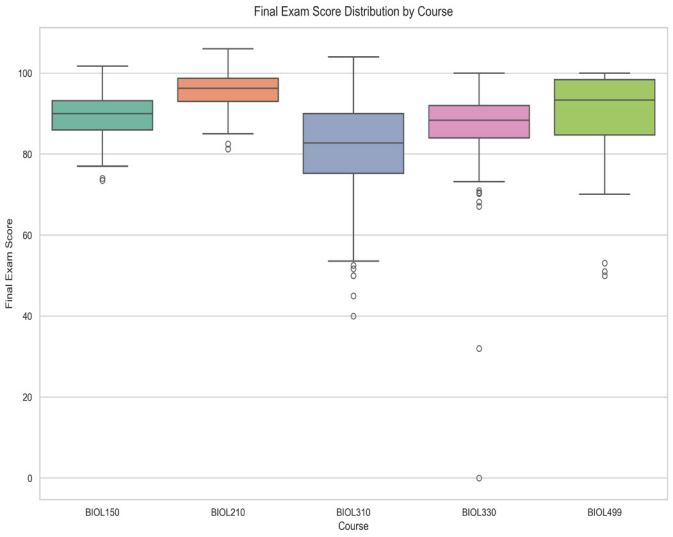

Figure 3. Final Exam Score Distribution by Course.

4.3 Course-Level Performance Trends

When disaggregated by course, substantial variation in student performance emerged (Table 3). BIOL210 and BIOL499 had the highest median Final Exam scores (96.25 and 93.33, respectively), while BIOL310 displayed the widest score spread and the lowest median (82.74), indicating greater difficulty or variability in student understanding. These differences are further visualized in the Final Exam boxplot by course (Figure 3), where BIOL310 and BIOL330 displayed greater dispersion and more outliers. The pairplot in Figure 4 further underscores the positive, but non-linear, relationships between early assessments and Final Exam scores across all courses.

4.4 Baseline Model Performance

Initial predictive modeling using four machine learning algorithms—Linear Regression, Random Forest, Support Vector Regressor (SVR), and XGBoost—produced mixed results (Table 4). Among these, Linear Regression yielded the highest R^2 score (0.39), followed by Random Forest (0.32). SVR and XGBoost performed comparatively worse, with R^2 values of 0.25 and 0.12, respectively.

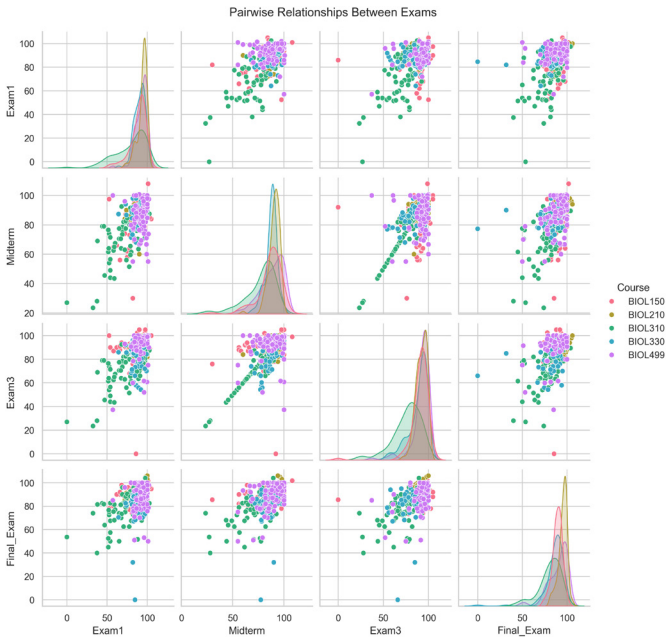

Figure 4. Pairwise Scatter and Density Plots of Exam Scores by Course.

Table 4. Baseline model performance metrics.

Model	MAE	RMSE	R2
Linear Regression	5.46	6.81	0.39
Random Forest	5.42	7.18	0.32
Support Vector Regressor	5.67	7.52	0.25
XGBoost Regressor	5.94	8.14	0.12

Table 3. Course-wise exam performance: Mean, median, and standard deviation.

Course	Exam1			Midterm			Exam3			Final Exam		
	Mean	Median	Std	Mean	Median	Std	Mean	Median	Std	Mean	Median	Std
BIOL150	88.47	92.00	10.26	86.35	88.00	11.72	90.69	92.00	11.46	89.26	90.00	5.43
BIOL210	93.82	96.00	5.98	90.40	92.00	5.51	92.22	93.00	5.98	95.28	96.25	4.87
BIOL310	79.43	87.00	19.30	76.99	82.00	15.15	76.44	79.00	15.62	80.90	82.74	12.56
BIOL330	90.29	92.00	6.73	87.65	88.50	5.80	88.80	92.00	10.61	86.13	88.38	12.31
BIOL499	92.69	95.00	7.11	88.31	90.07	11.10	92.28	95.00	10.41	90.13	93.33	10.88

Atlas Journal of Science Education - ISSN 2158-9224. Published By Atlas Publishing, LLC (www.atlas-publishing.org)

Model performance is illustrated in Figure 5. While the prediction errors (MAE and RMSE) were relatively similar across models, the differences in R^2 suggest variability in explanatory power rather than raw prediction error.

4.5 Tuned Model Performance

Hyperparameter tuning led to notable performance gains for tree-based models (Table 5). Both Tuned Random Forest and Tuned XGBoost achieved improved R^2 scores of 0.34 and reduced RMSE values compared to their untuned versions. These improvements are visualized in Figure 6, where the tuned models nearly matched the baseline performance of Linear Regression. Although the gains were modest, they demonstrate the potential of tuning to optimize non-linear models in educational prediction tasks.

Table 5. Performance of tuned Random Forest and XGBoost models.

Model	MAE	RMSE	R2
Tuned Random Forest	5.57	7.07	0.34
Tuned XGBoost	5.48	7.07	0.34

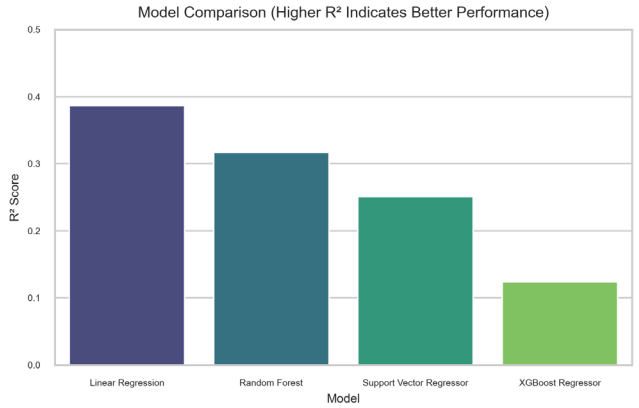

Figure 5. Baseline Model Comparison Based on R^2 Scores.

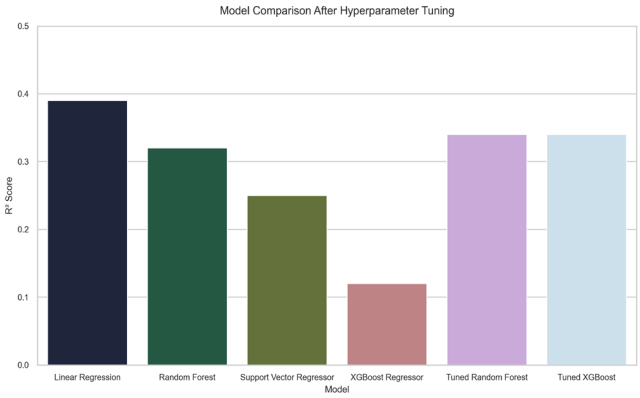

Figure 6. Model Performance After Hyperparameter Tuning.

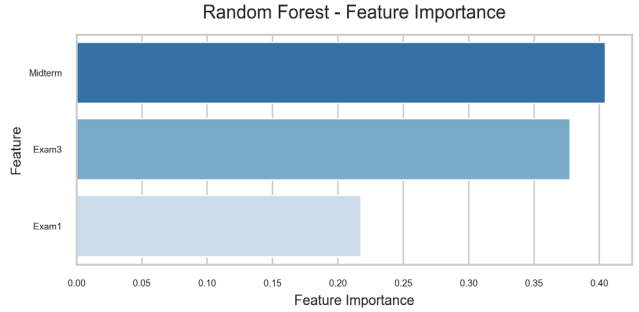

Figure 7. Feature Importance According to Random Forest Regressor.

4.6 Feature Importance and Interpretability

Feature importance was analyzed for both Random Forest and XGBoost models (Figures 7 and 8). In both cases, Midterm emerged as the most influential feature, followed by Exam 3 and then Exam 1. This hierarchy suggests that assessments conducted closer to the Final Exam (e.g., Exam 3 and Midterm) offer more predictive power, possibly due to their proximity to the comprehensive material assessed at the end of the course.

Further insights were obtained using SHAP (SHapley Additive exPlanations), which quantifies the impact of individual features on model output. The global SHAP summary plot (Figure 9) confirmed Exam 3 as the top contributor to model predictions, followed closely by Midterm. The SHAP beeswarm plot (Figure 10) visualizes the effect of each feature on individual predictions, showing that higher Exam 3 and Midterm scores consistently increased predicted Final Exam grades. These visualizations enhance interpretability and support actionable use of the model in real-world academic settings.

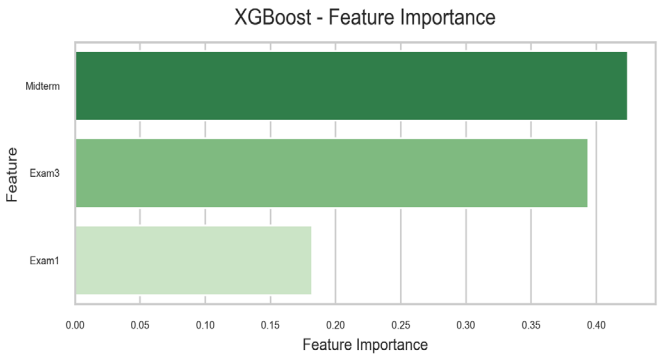

Figure 8. Feature Importance According to XGBoost Regressor.

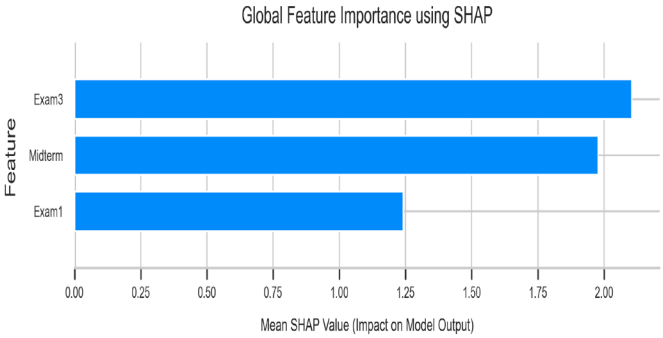

Figure 9. Global SHAP Feature Importance Plot (XGBoost).

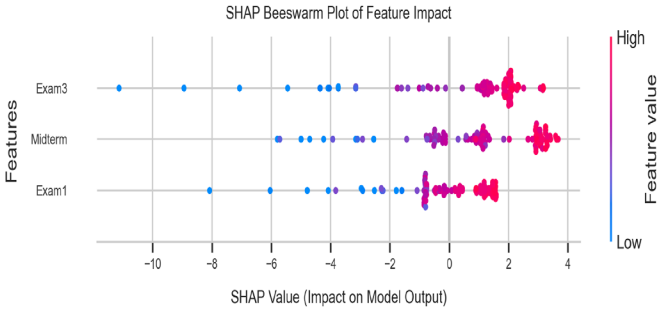

Figure 10. SHAP Summary Plot of Individual Feature Impacts (XGBoost).

5. Discussion

This study applied multiple machine learning algorithms to predict undergraduate biology students' final exam performance based on earlier assessments. The results demonstrate that prior exam scores—particularly Midterm and Exam 3—are moderate predictors of final exam performance, and that models such as Linear Regression, Random Forest, and XGBoost can capture meaningful patterns in this educational context.

Among the models evaluated, Linear Regression yielded the best overall predictive performance ($R^2 = 0.39$), with Tuned Random Forest and Tuned XGBoost following closely ($R^2 = 0.34$). While all models demonstrated relatively low root mean square errors (RMSE ≈ 7), the modest R^2 scores suggest that a substantial portion of the variance in Final Exam scores remains unexplained by prior assessments alone. These findings highlight the predictive value—but also the limitations—of relying solely on exam performance for early risk identification.

The feature importance analyses, including SHAP values and tree-based rankings, consistently identified Midterm and Exam 3 as the most influential predictors of Final Exam outcomes. This aligns with pedagogical expectations, as assessments closer to the end of the term are more likely to reflect cumulative understanding and readiness for the final assessment. The consistent lower importance of Exam 1 across models suggests that early performance may have limited utility in isolation, reinforcing the need for continuous monitoring throughout the course.

Course-level analysis revealed that performance patterns differ substantially across courses. For example, BIOL310 exhibited greater score variability and a lower median Final Exam score compared to other courses, pointing to differences in course complexity, grading practices, or student preparedness. This variation further underscores the need for context-aware predictive models that account for course-specific characteristics.

These findings have several implications for teaching practice and academic support strategies. First, the ability to predict final outcomes using midterm and pre-final assessments offers a critical window for early intervention (Zhang et al., 2014; Gorddanier et al., 2019). Educators and advisors could use predictive tools to proactively identify students at risk and provide targeted support before the end of the term, such as tutoring, review sessions, or individualized feedback.

Second, the use of interpretable models (e.g., SHAP-enhanced visualizations) ensures that educators can trust and understand model outputs. Rather than functioning as black-box predictions, these insights help instructors see why a student is predicted to struggle—whether due to declining trends, inconsistencies, or underperformance on specific content areas.

Finally, by implementing such tools at scale—particularly in foundational courses like BIOL150—institutions could build robust early alert systems. These systems could complement existing retention programs and support data-informed decision-making in curriculum design and academic advising.

This study contributes to a growing body of literature on educational data mining (EDM) and predictive modeling in STEM education. Previous studies have successfully applied ML techniques to predict student success in domains like engineering (Huang and Fang, 2013; He et al., 2028), computer science (Elbadrawy and Karypis, 2016), and mathematics (Aman et al., 2019). However, the application of predictive analytics to biology education remains limited, despite the discipline's complexity and centrality in STEM curricula.

Our findings align with studies that emphasize the predictive power of intermediate assessments (e.g., midterms) over earlier scores (Costa et al., 2017), and reinforce calls for feature-aware models that go beyond demographic or behavioral data to include content-specific academic indicators. Unlike studies that rely solely on black-box models like neural networks, our approach also prioritizes interpretability, making it more applicable to real-world educational settings.

Moreover, the study adds value by examining course-specific variation, an often-overlooked factor in predictive modeling. Biology courses, which combine theoretical knowledge with lab-based practical skills, pose unique challenges not found in more formulaic disciplines. By focusing on this subject area, the study highlights the importance of domain-specific predictive research within the broader EDM community.

Despite promising results, this study has several limitations. The dataset includes only performance metrics from exams and does not incorporate additional features such as student engagement data, attendance, assignment performance, or demographics. Including these could enhance model accuracy and fairness.

Second, the models assume uniform grading standards and exam difficulty across instructors and semesters, which may not hold true in diverse academic settings. Future studies could explore hierarchical or multi-level models that adjust for course- or instructor-level variability.

Finally, the current deployment is based on a single-institution dataset. To generalize findings, further validation across multiple universities, biology subdisciplines, and institutional types is recommended.

6. Conclusion

This study demonstrates the feasibility and practical value of using machine learning models to predict final exam outcomes in undergraduate biology courses. By leveraging prior assessment data, specifically Exam 1, Midterm, and Exam 3 scores, predictive models achieved moderate accuracy in forecasting final exam performance. Among the models tested, Linear Regression offered strong baseline performance, while tree-based models such as Random Forest and XGBoost showed measurable gains after tuning.

The results underscore the importance of mid-semester assessments as indicators of overall course success and highlight the utility of interpretable ML techniques, such as SHAP, in identifying the most influential predictors at both the individual and global levels. These insights can empower educators to make timely and targeted interventions, offering a proactive alternative to traditional, reactive academic support systems.

By focusing specifically on biology courses, this research addresses a gap in educational data mining, where most prior work has concentrated on quantitative STEM fields like engineering and computer science. The methodological framework developed here is both scalable and adaptable, with potential applications across other subject areas and institutional contexts.

Future work may expand on this foundation by incorporating behavioral and engagement features, extending the dataset to multiple institutions, and exploring longitudinal impacts of predictive interventions on student retention and success. Overall, the integration of machine learning into higher education presents a promising avenue for enhancing both instructional effectiveness and student outcomes.

Data Availability Statement

The dataset used in this study contains academic records and is not publicly available due to institutional policies regarding student data confidentiality. However, the Jupyter notebooks used for data analysis and model development are available from the corresponding author upon reasonable request.

Ethics Statement

This study involved the analysis of anonymized, pre-existing academic performance data collected as part of standard university assessment procedures. No personally identifiable information was used, and all data were handled in accordance with institutional data privacy policies. Ethical approval was not required as the research did not involve any intervention or interaction with human participants, and data analysis was conducted solely for educational research purposes. The authors affirm that the study complies with the ethical standards of the institution and relevant national guidelines.

Author Contributions

DG: Validation, Writing – review and editing. JG: Validation, Writing – review and editing. WG: Validation, Writing – review and editing. KL: Validation, Writing – review and editing. JY: Investigation, Validation, Writing – review and editing. LZ: Validation, Writing – review and editing. MAK: Conceptualization, Writing – original draft.

Funding

This research received no funding.

Conflict of Interest

The authors declare no conflict of interest.

References

Aggarwal D., D. Sharma, and A. B. Saxena (2024). A Comprehensive Analysis on the Application of Natural Language Processing (NLP) in Higher Education. The 8th International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), Kirtipur, Nepal, 2024, pp. 897-904. https://doi.org/10.1109/I-SMAC61858.2024.10714892.

Ahmad F., N.H. Ismail, and A.A. Aziz (2015). The Prediction of Students' Academic Performance Using Classification Data Mining Techniques. Applied Mathematical Sciences 9(129): 6415-6426. http://dx.doi.org/10.12988/ams.2015.53289.

Aman F., A. Rauf, R. Ali, F. Iqbal and A. M. Khattak (2019). A Predictive Model for Predicting Students Academic Performance, 2019 10th International Conference on Information, Intelligence, Systems and Applications (IISA), Patras, Greece, 2019, pp. 1-4, https://doi.org/10.1109/IISA.2019.8900760.

Badr G., A. Algobail, H. Almutairi, and M. Almutery (2016). Predicting Students' Performance in University Courses: A Case Study and Tool in KSU Mathematics Department. Procedia Computer Science 82: 80–89. https://doi.org/10.1016/j.procs.2016.04.012.

Baker R. S., T. Martin and L.M. Rossi (2016). Educational Data Mining and Learning Analytics. In A.A. Rupp, J.P. Leighton (Eds.). The Wiley Handbook of Cognition and Assessment: Frameworks, Methodologies, and Applications. https://doi.org/10.1002/9781118956588.ch16.

Breiman L. (2001). Random Forests. Machine Learning 45: 5–32. https://doi.org/10.1023/A:1010933404324.

Chen T. And C. Guestrin (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794. https://doi.org/10.1145/2939672.2939785.

Cortes C. And V. Vapnik (1995). Support-vector networks. Machine Learning 20: 273–297. https://doi.org/10.1007/BF00994018.

Costa E.B., B. Fonseca, M.A. Santana, F.F. de Araujo, and J. Rego (2017). Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses. Computers in Human Behavior 73: 247–256. https://doi.org/10.1016/j.chb.2017.01.047.

Dutt A, M.A. Ismail, and T. Herawan (2017). A Systematic Review on Educational Data Mining. IEEE 5: 15991-16005. https://doi.org/10.1109/ACCESS.2017.2654247.

Elbadrawy A. and G. Karypis (2016). Domain-Aware Grade Prediction and Top-n Course Recommendation. In Proceedings of the 10th ACM Conference on Recommender Systems (RecSys '16). Association for Computing Machinery, New York, NY, USA, 183–190. https://doi.org/10.1145/2959100.2959133.

Geron A. (2019). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow (2nd ed.). ISBN-13: 978-1492032649. O'Reilly Media.

Gorddanier J., W. Hauk, and C. Sankaran (2019). Early intervention in college classes and improved student outcomes. Economics of Education Review 72: 23-29. https://doi.org/10.1016/j.econedurev.2019.05.003.

Harris CR, Millman KJ, van der Walt SJ, et al. Array programming with NumPy. Nature. 2020;585(7825):357-362. https://doi.org/10.1038/s41586-020-2649-2.

He L., R.A. Levine, A.J. Bohonak, J. Fan, and J. Stronach (2018). Predictive Analytics Machinery for STEM Student Success Studies. Applied Artificial Intelligence, An International Journal 32(4): 361–387. https://doi.org/10.1080/08839514.2018.1483121.

Huang S. and N. Fang (2013). Predicting student academic performance in an engineering dynamics course: A comparison of four types of predictive mathematical models. Computers & Education 61: 133–145. https://doi.org/10.1016/j.compedu.2012.08.015.

Hunter JD (2007). Matplotlib: A 2D graphics environment. Computing in Science and Engineering 9(3): 90-95. https://doi.org/10.1109/MCSE.2007.55.

Kenney J.F. and E.S. Keeping (1962). Linear Regression and Correlation." Ch. 15 in Mathematics of Statistics, Pt. 1, 3rd ed. Princeton, NJ: Van Nostrand, pp. 252-285.

Kluyver, T.; Ragan-Kelley, B.; Pérez, F.; Granger, B.; Bussonier, M.; Frederic, J.; Kelly, K.; Hamrick, J.; Grout, J.; Corlay, S.; Ivanov, P.; Avila, D.; Abdalla, S.; Willing, C. Jupyter Notebooks – a publishing format for reproducible computational workflows. In: Loizides F, Schmidt B, eds. Positioning and Power in Academic Publishing: Players, Agents and Agendas. IOS Press, 2016:87-90. https://doi.org/10.3233/978-1-61499-649-1-87.

McKinney W. Data structures for statistical computing in Python. In: Proceedings of the 9th Python in Science Conference. Austin, TX; 2010:51-56. https://doi.org/10.25080/Majora-92bf1922-00a.

Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine learning in Python. J Mach Learn Res. 2011;12:2825-2830. https://jmlr.org/papers/v12/pedregosa11a.html.

Romero C. and S. Ventura (2020). Educational data mining and learning analytics: An updated survey. WIREs: Data Mining and Knowledge Discovery 10(3): e1355. https://doi.org/10.1002/widm.1355.

Roy S. and A. Garg (2017). Predicting academic performance of student using classification techniques. The 4th IEEE Uttar Pradesh Section International Conference on Electrical, Computer and Electronics (UPCON), Mathura, India, 2017, pp. 568-572. https://doi.org/10.1109/UPCON.2017.8251112.

Van Rossum G. and F.L. Drake (1995). Python reference manual. Centrum voor Wiskunde en Informatica Amsterdam.

Waskom ML. Seaborn: statistical data visualization. Journal of Open-Source Software, 2021, 6(60):3021. https://doi.org/10.21105/joss.03021.

Zhang Y., Q. Fei, M. Quddus, and C Davis (2014). An Examination of the Impact of Early Intervention on Learning Outcomes of At-Risk Students. Research in Higher Education Journal 26: 1-12.