PREDICTING STUDENT ACADEMIC PERFORMANCE USING MACHINE LEARNING

[1]IFEANYI MARTINS NWANEGBO; [2]COSTA DENISON-GEORGE; [3]DANIEL JONATHAN; & [4]NONSO FREDRICK CHIOBI
[1,2&3]Department of Data Analytics and Information Systems, Texas State University. [4]Department of Management Information Systems, Lamar University

DOI: https://doi.org/10.70382/caijmasr.v8i9.015

Abstract

This work explores the use of machine learning techniques to predict student academic performance, particularly their final grades, using a variety of demographic, academic, and lifestyle factors. With data sourced from the UCI Machine Learning Repository, we implemented regression and classification models—including Linear Regression, Random Forest, Gradient Boosting, Support Vector Regression, and Support Vector Classification. Additional models, such as K-Nearest Neighbors (KNN), Neural Networks (MLPClassifier), and Naive Bayes, were also evaluated to provide a more comprehensive analysis. The regression models aimed to predict the actual grade score (G3), while the classification models categorized students into performance bands (low, medium, high). The project's novelty lies in its dual-model approach to understanding and forecasting academic outcomes, offering granular and strategic insights. Results showed that ensemble models performed best, while alternative models offered interpretability and benchmark value. Insights from this project could inform early interventions by educators and support data-driven policy decisions.

Keywords: Students' Academic Performance, Machine Learning, Linear Regression, Boosting, Classifier

Introduction

Academic performance is a pivotal factor in shaping a student's educational and professional trajectory. Educational institutions strive to identify struggling students early and tailor interventions to improve outcomes. With the advent of data-driven education, predictive analytics powered by machine learning has gained traction to personalize academic support and enhance decision-making.

According to prior literature, academic success is influenced by a blend of cognitive, socio-economic, behavioral, and institutional factors. For example, studies have shown that parental education, past academic performance, and lifestyle habits like study hours and internet usage significantly affect student outcomes. Educational data mining has emerged as a rapidly expanding discipline that combines statistics, machine learning, and pedagogy to unearth insights from student data and enhance educational strategies (Romero & Ventura, 2010).

Our project extends this literature by exploring both regression and classification approaches to predicting performance using an open-source dataset from Portugal. The dataset, sourced from the UCI Machine Learning Repository, includes various demographic, behavioral, and academic features collected from secondary school students (Cortez & Silva, 2008). This dual-modelling strategy allows for both numeric precision and categorical clarity, providing not just grade predictions but actionable insights into which students are at risk. The novelty of this study lies in combining interpretability from regression with intervention-friendly classification models, offering a comprehensive solution for academic forecasting. Ultimately, we aim to demonstrate how these models can support real-time decision-making and long-term educational planning.

Research Design

Dataset:

The dataset, sourced from the UCI Machine Learning Repository, comprises records of Portuguese secondary school students and includes 33 diverse attributes. These attributes cover demographic details (such as age, sex, and family size), parental background (such as parents' education and job types), academic indicators (including grades from the first two periods), and lifestyle factors (such as alcohol consumption, internet access, and social outings). The dataset was selected for its real-world nature and granularity, offering a rich model training and testing field. The target variable for regression is the final grade (G3). For classification purposes, we binned G3 into three categories: 'low' (0–9), 'medium' (10–14), and 'high' (15–20), to make insights more actionable for educators.

Variables:

To identify influential features, we first conducted exploratory data analysis and correlation checks. Key predictors included parental education, study time, number of past failures, previous grades (G1 and G2), and student health. We also tested the impact of more nuanced variables such as the quality of family relationships and access to the internet at home. A p-value analysis using OLS regression helped us narrow down the most statistically significant predictors. Features like G1, G2, study time, and failures emerged as strong indicators of academic performance. These were processed using one-hot encoding for categorical variables and scaling for numerical features, ensuring that the data was model-ready.

Methodology:

Our action plan was iterative and flexible. It began with data cleaning and preprocessing, which included handling missing values, encoding categorical features, and splitting the dataset into 80% training and 20% validation sets. We implemented Random Over Sampler and Random Under

Sampler techniques to account for class imbalance in classification. We then applied and fine-tuned multiple models across both predictive categories:

Regression Models:
Linear Regression, Random Forest Regressor, Gradient Boosting Regressor, Support Vector Regressor.

Classification Models:
Random Forest Classifier, AdaBoost Classifier (based on decision trees), and Support Vector Classifier.
We used Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared (R^2) as primary regression metrics. For classification, we evaluated performance using accuracy, precision, recall, and F1-score to ensure a comprehensive model quality assessment, especially across imbalanced class labels.

Data Analysis

The results derived from the model are analyzed as follows:

1. Linear Regression
   Linear Regression provided a baseline with RMSE around 1.71. With an RMSE of 1.71, the model's predictions are average 1.71 grade points away from the actual values. This is reasonable but leaves room for improvement, especially given the simplicity of the model and its inability to capture interactions between features.
2. Random Forest Regressor
   Random Forest Regressor reduced RMSE to around 0.97 and achieved an MAE of approximately 0.63. These results indicate the model predicted student grades with an average error of under 1 point. The model's ability to handle non-linear relationships and rank feature importance made it a standout performer.
3. Random Forest Classifier:

Random Forest Classifier		**Accuracy**	**Precision**	**Recall**
Original	High (15 - 20)	86	83	94
	Low (0 - 9)		63	80
	Medium (10 - 14)		93	84
Oversampled	High (15 - 20)	82	79	94
	Low (0 - 9)		55	80
	Medium (10 - 14)		93	78
Undersampled	High (15 - 20)	83	78	100
	Low (0 - 9)		54	93
	Medium (10 - 14)		98	75

The Random Forest Classifier performed well across all sampling strategies, but the undersampled dataset offered the most balanced results. While the original dataset had the highest overall accuracy (86%), the undersampled version achieved perfect recall for high-performing students (100%), strong recall for low performers (93%), and excellent precision for medium performers (98%). Despite a lower precision for the low category (54%), this setup provided the most reliable identification of students at both ends of the performance spectrum, making it the most actionable model configuration for early intervention and academic planning.

4. Gradient Boosting

   Gradient Boosting Regressor had a comparable RMSE of 1.02 and an MAE around 0.66, slightly trailing Random Forest. This consistent model offered better control over overfitting, especially with tuning.

5. Boosting Classifier:

Boosting Classifier		Accuracy	Precision	Recall
Original	High (15 - 20)	84	86	97
	Low (0 - 9)		50	67
	Medium (10 - 14)		92	82
Oversampled	High (15 - 20)	82	83	94
	Low (0 - 9)		50	67
	Medium (10 - 14)		91	81
Undersampled	High (15 - 20)	74	70	97
	Low (0 - 9)		40	67
	Medium (10 - 14)		90	66

The Boosting Classifier delivered strong performance overall, with the original dataset yielding the highest accuracy (84%) and the best balance of precision and recall, especially for high-performing students (precision = 86%, recall = 97%). While the oversampled dataset maintained comparable accuracy (82%) and recall, it did not significantly improve the precision or recall for low performers (precision = 50%, recall = 67%). The undersampled dataset showed the lowest overall accuracy (74%) and experienced a drop in recall for medium-performing students (66%), although it maintained high recall for the high category (97%). These results suggest that the original dataset provided the most stable and consistent performance for boosting, especially in identifying high and medium performers effectively without compromising overall model accuracy.

6. *Support Vector Regressor*

   Support Vector Regressor (SVR) yielded an RMSE of 1.25 and MAE of 0.83, showing moderate improvement over Linear Regression. An R² of 0.85 signified strong explanatory power, meaning the model could explain 85% of the variance in grades.

7. *Support Vector Classifier:*

Support Vector Classifier		Accuracy	Precision	Recall
Original	High (15 - 20)	90	97	97
	Low (0 - 9)		67	53
	Medium (10 - 14)		91	94
Oversampled	High (15 - 20)	83	78	100
	Low (0 - 9)		55	80
	Medium (10 - 14)		96	77
Under sampled	High (15 - 20)	78	70	100
	Low (0 - 9)		50	87
	Medium (10 - 14)		97	67

The Support Vector Classifier showed excellent performance, particularly on the original dataset, where it achieved the highest accuracy of 90%, along with near-perfect precision and recall for high (97%, 97%) and medium (91%, 94%) performers. However, it struggled to identify low-performing students, with a recall of just 53% despite decent precision (67%). The oversampled dataset improved recall for the low class to 80% and achieved perfect recall for high performers (100%), though overall accuracy dropped to 83%, and precision across all classes slightly declined. The undersampled dataset maintained strong recall across all categories (87%–100 %) but experienced a drop in precision, particularly for the low class (50%), and overall accuracy dipped to 78%. These results suggest that while the original dataset yielded the most balanced and accurate results, oversampling may be preferable if the goal is to ensure low and high performers are not missed, even at the cost of some precision.

Additional Models

8. *KNN - Regressor*

K-Nearest Neighbors (KNN) was also evaluated, given its history of effective use in educational distance learning prediction tasks (Kotsiantis, Pierrakeas, & Pintelas, 2004).

KNN Regressor had varied performance depending on the value of k. For k=3, RMSE hovered around 1.50, showing that predictions were further off the mark. While intuitive and straightforward, its sensitivity to local data patterns made it inconsistent.

9. KNN - Classifier

Original

Model - KNN Classifier - Original		Accuracy	Precision	Recall
N = 1	High (15 - 20)	80	77	84
	Low (0 - 9)		55	73
	Medium (10 - 14)		88	80
N = 3	High (15 - 20)	86	89	97
	Low (0 - 9)		55	73
	Medium (10 - 14)		93	84
N = 5	High (15 - 20)	90	94	100
	Low (0 - 9)		62	67
	Medium (10 - 14)		94	90

Oversampled

Model - KNN Classifier – Oversampled		**Accuracy**	**Precision**	**Recall**
N = 1	High (15 - 20)	80	76	91
	Low (0 - 9)		52	73
	Medium (10 - 14)		90	77
N = 3	High (15 - 20)	83	80	100
	Low (0 - 9)		52	80
	Medium (10 - 14)		96	77
N = 5	High (15 - 20)	82	80	100
	Low (0 - 9)		50	73
	Medium (10 - 14)		94	77

Undersampled

Model - KNN Classifier – Undersampled		**Accuracy**	**Precision**	**Recall**
N = 1	High (15 - 20)	79	73	100
	Low (0 - 9)		50	73
	Medium (10 - 14)		94	72
N = 3	High (15 - 20)	80	73	100
	Low (0 - 9)		52	93
	Medium (10 - 14)		98	70
N = 5	High (15 - 20)	81	74	100
	Low (0 - 9)		52	87
	Medium (10 - 14)		97	72

KNN Classifier showed varying performance across different sampling strategies.

- In the original dataset, KNN achieved a peak accuracy of 75%, but struggled significantly with the ‘medium’ class. This is due to its sensitivity to class imbalance—students in the medium range often share overlapping features with both low and high performers, making it difficult for KNN to find clear neighbor boundaries.
- In the oversampled dataset, KNN improved markedly, reaching up to 79% accuracy. Precision and recall were better balanced across all categories. The synthetic increase in minority class instances helped KNN form more reliable neighbor groups, boosting overall classification quality. This version provided the best result among KNN runs.
- In the undersampled dataset, accuracy dropped to around 70%. The model's neighborhood-based logic became erratic with fewer data points available, misclassifying edge cases. Undersampling tends to remove nuanced samples that help KNN find meaningful proximity, which contributed to reduced model effectiveness.

While intuitive and easy to implement, KNN's reliance on local data patterns makes it highly sensitive to data distribution and sampling techniques. It is best suited for well-balanced datasets or used alongside strong preprocessing pipelines.

10. *Neural Networks*

Model - Neural Networks		Accuracy	Precision	Recall
Original	High (15 - 20)	86	89	100
	Low (0 - 9)		53	53
	Medium (10 - 14)		91	87
Oversampled	High (15 - 20)	85	84	100
	Low (0 - 9)		53	67
	Medium (10 - 14)		93	82
Undersampled	High (15 - 20)	79	76	100
	Low (0 - 9)		46	73
	Medium (10 - 14)		94	72

Neural Networks delivered consistently strong results across all sampling strategies, with the original dataset yielding the highest overall accuracy at 86%. It showed perfect recall for high performers (100%), strong precision for medium (91%) and high (89%) classes, but underperformed in detecting low performers (precision and recall = 53%). The oversampled dataset maintained high recall for high-performing students (100%) while slightly boosting recall for the low-performing group (67%), though precision for the low-performing group remained unchanged. The undersampled dataset showed the most balanced precision-recall tradeoff for medium performers (precision = 94%, recall = 72%), and again achieved 100% recall for high performers, though precision dipped for the low class (46%). Overall, while Neural Networks reliably identify high and medium achievers, their precision for low performers remains a limitation, highlighting the need for complementary models or enhanced resampling strategies when the goal is early identification of struggling students.

11. *Naive Bayes*

Model - Naive Bayes		Accuracy	Precision	Recall
Original	High (15 - 20)	44	34	97
	Low (0 - 9)		55	80
	Medium (10 - 14)		82	17
Oversampled	High (15 - 20)	45	34	94
	Low (0 - 9)		57	80
	Medium (10 - 14)		81	20
Undersampled	High (15 - 20)	44	34	97
	Low (0 - 9)		54	87
	Medium (10 - 14)		87	16

Naive Bayes consistently produced the lowest overall accuracy across all datasets, with 44–45% scores. While it demonstrated strong recall for high performers (94–97%) and decent recall for low performers (80–87%), it significantly underperformed in identifying medium-performing students, with recall dropping as low as 16%. Precision for high performers remained low (34%) across the board, indicating a high false positive rate, while medium performers had better precision (81–87%) but were rarely correctly identified. These results reflect the limitations of Naive Bayes in handling overlapping feature distributions and interdependent variables, which are common in educational datasets. Despite its speed and simplicity, its inconsistent performance makes it better suited as a lightweight benchmark than a standalone solution for academic prediction tasks.

In summary, the models explored in this study revealed key strengths and trade-offs across sampling strategies and prediction objectives. Ensemble models like Random Forest and Gradient Boosting consistently delivered high accuracy and balanced performance, particularly in predicting high and medium achievers. Neural Networks proved effective in capturing complex patterns, especially for top-performing students, though less reliable for the low category. While KNN benefited from oversampling, it remained sensitive to data distribution, and Naive Bayes, despite its speed, struggled with accuracy and recall for the majority class. These results underscore the importance of model selection based on the prediction target and the real-world implications of misclassification, especially in educational contexts where early and accurate intervention can significantly impact student outcomes.

Discussion

This study aimed to predict student academic performance using regression and classification models to support early intervention, academic monitoring, and resource allocation.

Discussion of Regression Models

Among the regression models tested, Random Forest Regressor clearly outperformed the others, achieving an RMSE of 0.97 and an MAE of 0.63. These values imply that the model could predict a student's final grade with an average error of less than 1 point, making it a powerful tool for GPA forecasting or personalized academic advising. Gradient Boosting followed closely with slightly higher error margins but provided a more regularized, stable alternative. While SVR and KNN Regressor offered moderate improvements over Linear Regression, their higher RMSE (around 1.25 and 1.50, respectively) reduced their viability for high-stakes academic decision-making.

The Random Forest Regressor's ability to rank feature importance also aligns well with the project's goal of understanding drivers of performance, highlighting prior grades (G1, G2), study time, and failures as key predictors. Its interpretability and minimal tuning needs make it practical for academic institutions to implement.

In contrast, Linear Regression, while simple and transparent, lacks the complexity needed to capture the interactions between student behaviors and outcomes. It may serve as a baseline, but it falls short for predictive intervention.

Conclusion - Regression Models:

The Random Forest Regressor is the most effective model for this project for forecasting exact student grades to inform GPA trends or graduation readiness.

Discussion of Classification Models

For classification tasks, the models were evaluated based on how well they identified students in each performance band (low, medium, high) — a more actionable format for schools aiming to flag and support specific groups. The Support Vector Classifier (SVC) achieved the highest overall accuracy at 90%, with near-perfect precision and recall for high and medium performers on the original dataset. However, its lower recall for low performers (53%) suggests it might overlook struggling students, a major risk in academic settings.

On the other hand, the Random Forest Classifier, while slightly lower in accuracy (86%), offered the best balance across all classes, especially with the undersampled dataset: it achieved 100% recall for high performers, 93% for low, and 98% precision for medium. This makes it the strongest model for practical, equity-focused use, ensuring no high-risk students are missed.

The Boosting Classifier and Neural Networks also performed well, particularly with high performers. Neural Networks showed strength across datasets, but like SVC, had lower precision and recall for the low category. Naive Bayes, while fast, was inconsistent and underperformed significantly in identifying medium performers (recall as low as 16%).

Conclusion - Classification Models:

For identifying at-risk or high-achieving students for timely support, Random Forest Classifier (undersampled) provides the best performance-to-action tradeoff. It maximizes inclusivity while maintaining reliable accuracy.

In practical terms, the Random Forest Regressor offers a powerful tool for academic forecasting by enabling advisors, administrators, and policy makers to predict individual student grades with high precision. This can support personalized academic planning, early identification of at-risk students falling below GPA thresholds, and strategic scholarship or graduation readiness evaluations. Meanwhile, the Random Forest Classifier, with its strong ability to categorize students into low, medium, and high performers, provides a clear framework for designing intervention programs, allocating resources, and implementing targeted support systems. Together, these models enable institutions to move beyond reactive measures and toward a proactive, data-driven approach to student success, making early intervention, customized support, and outcome tracking not only possible but highly actionable.

Conclusion

This project set out to explore and compare machine learning models for predicting student academic performance, both in terms of forecasting exact grade outcomes and classifying students into performance bands. Using a diverse set of regression and classification models trained on student demographic, behavioral, and academic data, we assessed each model's suitability for educational forecasting and intervention.

In the regression task, the Random Forest Regressor emerged as the most accurate and reliable model, consistently predicting final grades with minimal error (RMSE = 0.97, MAE = 0.63). Its ability to rank feature importance made it not just a predictive tool but also an explanatory one, highlighting prior grades and failure history as critical factors. Compared to simpler models like Linear Regression, ensemble methods demonstrated a superior ability to capture complex relationships and improve the precision of academic forecasts.

For classification, which framed student performance in more actionable categories (low, medium, high), the Random Forest Classifier, especially when trained on an undersampled dataset, offered the most balanced and inclusive results. It achieved perfect recall for high-performing students, strong recall for those at risk, and high precision for medium achievers, making it ideal for real-world academic intervention strategies. While models like the Support Vector Classifier and Neural Networks also delivered strong results, they were slightly less consistent in identifying students who needed the most support.

The findings underscore that model selection should depend on the specific educational objective. Regression models are ideal for continuous forecasting, such as GPA prediction and individualized progress tracking. Classification models, meanwhile, are better suited for early warning systems, scholarship eligibility, or targeted interventions.

Looking ahead, future work could explore additional variables like attendance patterns, engagement metrics, and emotional well-being data to enrich prediction accuracy. Cross-institutional datasets could also improve generalizability. More sophisticated models, such as XGBoost or deep learning architectures, may offer incremental performance boosts and could be valuable in real-time academic decision systems.

Ultimately, this project demonstrates the power of data-driven insights in education. With thoughtful implementation, these models can serve as essential tools to understand student performance and proactively shape more equitable and responsive educational environments.

References

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. https://link.springer.com/article/10.1023/A:1010933404324

Cortez, P., & Silva, A. (2008). Student Performance Data Set. UCI Machine Learning Repository. https://archive.ics.uci.edu/dataset/320/student+performance

Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences, 55(1), 119–139. https://www.sciencedirect.com/science/article/pii/S002200009791504X?via%3Dihub

Kotsiantis, S. B., Pierrakeas, C., & Pintelas, P. (2004). Predicting students' performance in distance learning using machine learning techniques. Applied Artificial Intelligence, 18(5), 411–426. https://www.tandfonline.com/doi/full/10.1080/08839510490442058

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830.

Romero, C., & Ventura, S. (2010). Educational data mining: A review of the state of the art. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 40(6), 601–618. https://ieeexplore.ieee.org/document/5524021