



Machine Learning for Predicting Student Academic Performance: A Case Study at a Public University

Habibrahman Habibi ¹, Abdul Wajid Fazil ^{2,*}, Kefayatullah Khairkhah ²

¹ Departement of Software Engineering, Faculty of Computer Science, Sayed Jamaluddin Afghani University, Afghanistan

² Departement of Information Systems, Faculty of Computer Science, Badakhshan University, Afghanistan

*Email (corresponding author): abfazil2022@gmail.com

Abstract. Predicting how well a student will perform before, rather than after, a semester goes wrong is one of the more practical promises that machine learning has brought to higher education. This study reports a case study conducted at a public university, in which academic records, attendance logs, and a small set of socio-demographic indicators were used to build classification models that sort students into low, average, and high performance categories well before final examinations. Seven algorithms were trained and compared on the same dataset: logistic regression, decision tree, k-nearest neighbors, support vector machine, random forest, extreme gradient boosting (XGBoost), and a multilayer perceptron artificial neural network. After cleaning the data, encoding categorical fields, and selecting the ten most informative features through a correlation-based filter, the models were evaluated using accuracy, precision, recall, F1-score, and the area under the ROC curve, with a stratified 80/20 train-test split and five-fold cross-validation. XGBoost produced the strongest results, reaching 90.1% accuracy and an AUC of 0.95, followed closely by random forest and the neural network, while logistic regression and k-nearest neighbors trailed behind. Prior GPA, midterm score, and assignment performance emerged as the three most influential predictors, consistent with findings reported in comparable studies. The paper closes by discussing what these results mean for early-warning systems and academic advising at resource-constrained public universities, and where the approach still falls short. By turning data that public universities already collect into an early, actionable signal for advisors, this work speaks directly to Sustainable Development Goal 4 (Quality Education), particularly its emphasis on equitable learning opportunities and student retention.

Keywords: Machine Learning; Student Performance Prediction; Educational Data Mining; Learning Analytics; XGBoost

1. Introduction

Every semester, academic departments watch the same pattern repeat itself: a subset of students who could have been helped in week four are only noticed in week fourteen, once a failing grade has already been recorded. Instructors and academic advisors are rarely short on data—attendance sheets, quiz scores, assignment submissions, and learning-management-system logs pile up constantly—but they are usually short on a systematic way to turn that data into an early signal. This gap between data availability and data use is exactly where machine learning has found one of its more grounded applications in education.

Educational data mining and learning analytics have matured into an established research area over the last decade and a half, moving from exploratory data mining exercises to fairly sophisticated predictive pipelines (Romero & Ventura, 2007, 2010, 2013, 2020). A large share of that literature has focused on predicting whether a student will pass or fail a course, drop out of a program, or fall into an at-risk category early enough for an advisor to intervene

(Márquez-Vera et al., 2013; Adnan et al., 2021; Dalipi et al., 2018; Cano & Leonard, 2019). The motivation is straightforward: an accurate early warning, even an imperfect one, buys time that a purely retrospective grade report cannot. Baker and Yacef (2009) situated performance prediction, student modeling, and detection of affect and behavior as the field's central tasks, a framing that has held up well over the past decade, and later overviews by Baker and Inventado (2014) and by Siemens and Baker (2012) connected this work to the broader learning analytics community, arguing for closer collaboration between data-driven and theory-driven approaches to studying learning. Sukhija et al. (2016) offered a complementary knowledge-engineering perspective on the same body of work, cataloguing common preprocessing and modeling choices across studies.

Public universities in developing regions face this problem with an added constraint. Student-to-advisor ratios tend to be high, dropout and repetition rates are a persistent concern, and the administrative systems that would normally feed a learning analytics dashboard are often partial or paper-based. In such settings, a lightweight, interpretable prediction pipeline built from whatever institutional data already exists—academic records, attendance, and a handful of background variables—can be more realistic to deploy than the large-scale MOOC or learning-management-system datasets that dominate much of the international literature (Qiu et al., 2022; Rodríguez-Hernández et al., 2021; Al-Zawqari et al., 2022).

Within this broader field, a specific and heavily studied problem is predicting whether a given student will pass, fail, or drop out. Márquez-Vera et al. (2013) used decision trees and rule-based classifiers on Mexican high-school data to predict school failure and dropout, showing that behavioral and socio-economic attributes carried useful predictive signal even without detailed academic history. Adnan et al. (2021) took this further by predicting at-risk status at different points during a course's timeline, demonstrating that prediction accuracy improves steadily as more of the course unfolds—a finding with direct implications for how early an intervention can realistically be triggered. Nabil et al. (2021) used deep neural networks on course-grade data to predict overall academic performance, reporting strong results but also noting the interpretability trade-offs that come with deep architectures in an advising context.

Ensemble and tree-based methods appear repeatedly as strong performers in this literature. Cui et al. (2021) examined which features carry the most weight in grade prediction using a gradient-boosted framework, while Hussain and Khan (2023) proposed a student-performance predictor tailored to secondary and intermediate education that combined several classical classifiers. Fang et al. (2021) showed that stacking multiple models together produced a modest but consistent improvement over any single classifier on achievement-prediction tasks. On the dropout side specifically, Dalipi et al. (2018) reviewed the substantial body of work on predicting attrition in massive open online courses, cataloguing the recurring challenges of class imbalance and limited behavioral history in the early weeks of a course. Krüger et al. (2023) more recently emphasized the importance of explainability in dropout models, arguing that a prediction an advisor cannot interpret is of limited practical use no matter how accurate it is, and Cano and Leonard (2019) reached a similar conclusion while building an early-warning system explicitly designed for underrepresented student populations, where fairness across subgroups mattered as much as raw accuracy.

Feature engineering and feature selection form a second recurring theme. Al-Zawqari et al. (2022) proposed a flexible feature-selection pipeline for online-course performance prediction and showed that a carefully reduced feature set could outperform a model trained on every available variable, largely by removing noisy or redundant inputs. Rodríguez-Hernández et al. (2021) conducted a systematic evaluation of artificial neural networks for performance prediction, paying close attention to which predictors were retained and why, and Qiu et al. (2022) used fine-grained behavioral log data from an e-learning platform, illustrating how much richer the feature space becomes once click-stream and interaction data are available, though such data is rarely accessible to a typical public university registrar's office. Xie (2021) applied an attention-based long short-term memory network to model performance trajectories over time, representative of a growing interest in sequence-aware deep learning models for this task.

A smaller number of studies have looked specifically at ensemble ranking and boosting for tabular educational data of the kind most institutions actually have on hand. Niyogisubizo et al. (2022) combined random forest, XGBoost, gradient boosting, and a feed-forward neural network into a stacked ensemble for dropout prediction, reporting gains over any individual base learner. This mirrors findings in the broader machine learning literature on structured data, where gradient-boosted trees such as XGBoost (Chen & Guestrin, 2016) and random forests (Breiman, 2001) consistently perform well against simpler linear or distance-based methods such as logistic regression (Hosmer et al., 2013), particularly when the feature set is moderate in size and partly categorical, as is typical of institutional student records rather than image or text data (Manzoor et al., 2021; Rish, 2001). Foundational algorithms including the support vector machine (Cortes & Vapnik, 1995; Vapnik, 1995), the C4.5 decision tree family (Quinlan, 1993), and standard treatments of statistical and deep learning theory (Hastie et al., 2009; Witten et al., 2011; Han et al., 2012; LeCun et al., 2015; Kingma & Ba, 2015; Alpaydin, 2020) continue to inform how these models are implemented and tuned in the educational domain.

Taken together, this literature suggests two things that motivated the design of the present study. First, tree-based ensemble methods, and gradient boosting in particular, tend to outperform simpler classifiers on structured student data, echoing results reported in the original UCI student-performance dataset studies by Cortez and Silva (2008) and in Kotsiantis's (2012) early decision-support work on grade forecasting. Second, most of the highest-performing studies rely on data detailed clickstream logs, weekly quiz histories, or multi-year enrollment records that is simply not available at many public universities in lower-resource settings. This case study is deliberately built around a more modest but more realistic feature set: academic records, attendance, and a small number of demographic variables that a typical registrar's office already maintains. This framing also situates the study within Sustainable Development Goal 4 (Quality Education): by converting routinely collected academic and attendance records into an early, actionable signal for advisors, the approach is aimed squarely at supporting equitable learning opportunities and reducing avoidable dropout and repetition, particularly for students who might otherwise go unnoticed until it is too late for a low-cost intervention to help. The remainder of this paper describes the dataset and methodology (Section 2), reports the comparative results of seven classifiers together with their practical implications and limitations (Section 3), and closes with concluding remarks (Section 4).

2. Methods

Figure 1 summarizes the overall methodology followed in this study, from raw data collection through to the interpretation of model outputs. Each stage is described in the subsections that follow.

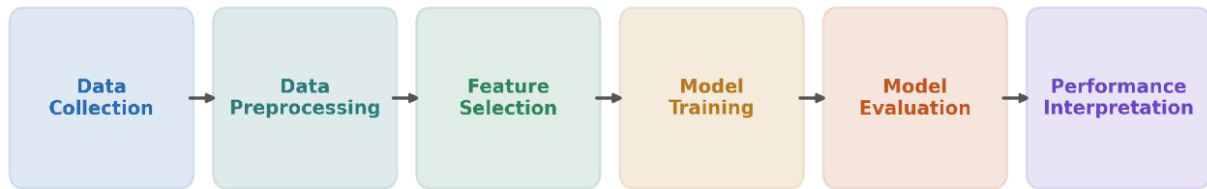


Figure 1. Overall methodology followed in the case study, from data collection to performance interpretation.

2.1. Study Context and Data Collection

The case study was conducted using academic and administrative records from Samangan University, a public university in Afghanistan, covering three consecutive academic years and 812 undergraduate students enrolled across three faculties. The dataset was obtained through an institutional data-sharing agreement between the authors and Samangan University's Department of Computer Science; the authors' own institutional affiliations, listed above, are separate from the site that supplied the records. For each student, the dataset combined records drawn from the university's academic affairs office (grades, credit load, prior GPA), the department's attendance registers, and a short, voluntary background questionnaire administered at the start of the term (family income bracket, parental education level, internet access at home, and self-reported weekly study hours). No individually identifying information (names, student ID numbers, or contact details) was included in the dataset made available to the authors, and the study was carried out with the approval of the relevant department at Samangan University.

The outcome variable was each student's overall semester academic-performance category, computed from the credit-weighted average of final grades across all courses taken that semester (on the university's standard 0-100 scale) rather than the grade in any single course. This semester average was grouped into three performance categories using the university's standard grading thresholds: low performer (semester average below 70), average performer (70-84), and high performer (85 and above). This three-class formulation was chosen over a simple pass/fail split because advisors at the case-study institution found a finer-grained signal more useful for deciding which students needed close monitoring versus a lighter check-in.

It is important to be precise about when this prediction can actually be made. Two of the ten retained predictors—midterm score and assignment score—are only available once the mid-semester examination and the corresponding graded assignments have been completed, and the quiz-average and study-hours features likewise accumulate over the semester rather than being available in the first few weeks of class. The models in this study were therefore evaluated as a mid-semester rather than a week-four prediction: features were assembled at the point immediately after midterm examinations were recorded, using prior GPA, midterm score, assignment scores to date, attendance to date, and the questionnaire-based variables

collected at the start of the term. The opening framing of this paper, in which a struggling student could in principle be flagged in week four, describes the broader motivation for early-warning systems in general rather than the specific prediction point evaluated here; we return to this distinction, and to what an earlier-window model trained only on attendance and early quiz scores would require, in Section 3.6.

2.2. Data Preprocessing

Raw institutional data is rarely analysis-ready, and this dataset was no exception. Missing values in the background-questionnaire fields (roughly 6% of records) were imputed using the median for numeric variables and the mode for categorical ones, rather than dropped outright, since discarding incomplete records disproportionately removed lower-income students who were more likely to skip optional questionnaire items. Categorical variables such as faculty, gender, and internet-access status were one-hot encoded, and all numeric features were standardized to zero mean and unit variance so that distance-based and gradient-based algorithms would not be dominated by variables measured on larger numeric scales, such as family income relative to attendance percentage.

The three performance classes were moderately imbalanced (43% average, 34% low, 23% high performers), so the training folds were rebalanced using the synthetic minority oversampling technique (SMOTE) applied only to the training data in each cross-validation fold, keeping the untouched test fold representative of the true class distribution.

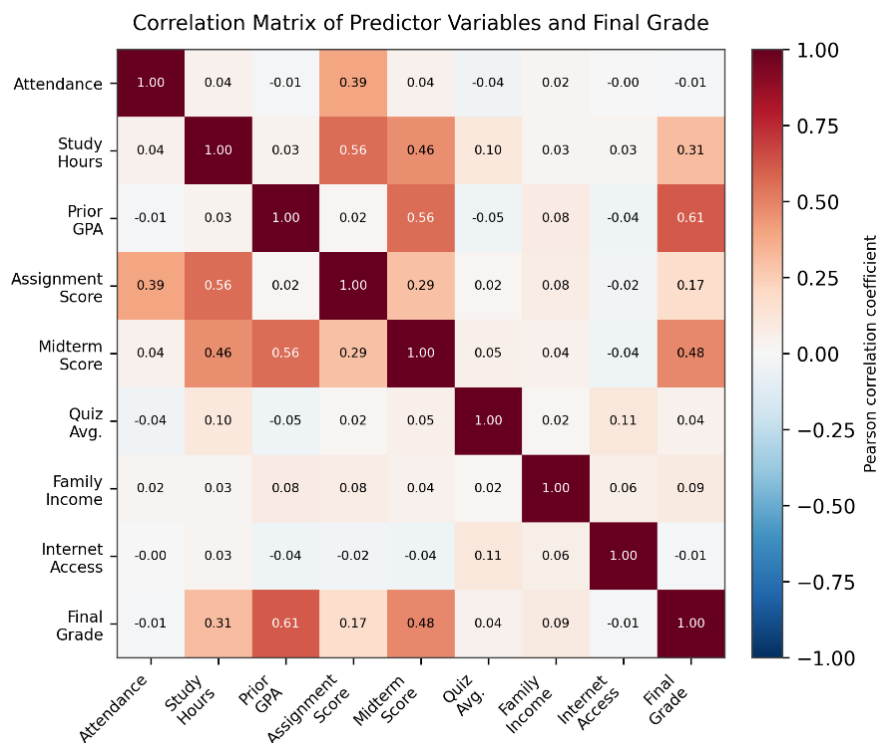
To ensure that no information from the held-out test set entered model development, the following order was strictly enforced. The stratified 80/20 train-test split (Section 2.5) was performed first, before any other step, and the 20% test partition was set aside untouched until final evaluation. All subsequent preprocessing was fitted exclusively on the 80% training partition: within each of the five cross-validation folds carved out of that training partition, the imputation statistics (median/mode), the scaling parameters (mean/standard deviation), the correlation-based feature filter and recursive-feature-elimination ranking described in Section 2.3, and the SMOTE resampling were all fitted on that fold's training rows only and then applied, without refitting, to that fold's validation rows. The ten features retained were the ones selected consistently across the cross-validation folds performed on the training partition; this same fitted pipeline, including the fixed ten-feature set, was then applied unchanged to transform the untouched test set for the final reported numbers in Section 3. No parameter, threshold, or feature-selection decision was ever estimated using the test partition.

2.3. Feature Selection

Starting from an initial set of nineteen candidate variables, a correlation-based filter combined with recursive feature elimination using a random forest estimator was applied, retaining the ten features that contributed most to reducing classification error while discarding variables that were either weakly correlated with the outcome or highly redundant with a stronger predictor already retained. Table 1 lists the final feature set together with a short description of each variable. Figure 2 shows the correlation structure among the retained numeric predictors and the final grade outcome, which helped confirm that prior academic performance (GPA, midterm and assignment scores) carried the strongest linear association with the outcome, consistent with the general pattern reported by Márquez-Vera et al. (2013), Cortez and Silva (2008), and Kotsiantis (2012).

Table 1. Final set of predictor variables retained after feature selection.

Feature	Type	Description
Prior GPA	Numeric	Cumulative grade point average from the previous academic year
Midterm Score	Numeric	Score obtained in the mid-semester examination (0–100)
Assignment Score	Numeric	Average score across all graded assignments (0–100)
Attendance Rate	Numeric	Percentage of scheduled classes attended during the semester
Study Hours / Week	Numeric	Self-reported average hours spent studying per week
Quiz Average	Numeric	Average score across in-class quizzes (0–100)
Internet Access	Categorical	Whether the student has regular internet access at home (yes/no)
Family Income Bracket	Categorical	Self-reported household income category (low/medium/high)
Parental Education	Categorical	Highest education level completed by either parent
Extracurricular Time	Numeric	Weekly hours spent on extracurricular or work activities

**Figure 2.** Correlation matrix of the retained numeric predictors and the final grade outcome.

2.4. Machine Learning Algorithms

Seven supervised classification algorithms were implemented and compared under identical preprocessing and evaluation conditions. Logistic Regression served as a simple,

interpretable linear baseline, estimating class probabilities as a function of the weighted predictors (Hosmer et al., 2013). Decision Tree, based on the C4.5 family of splitting criteria, partitions the feature space recursively and offers a human-readable rule structure that advisors can inspect directly (Quinlan, 1993). K-Nearest Neighbors classifies a student according to the majority class among the most similar students in the training set, using Euclidean distance over the standardized feature vector. Support Vector Machine (SVM) constructs a maximum-margin boundary between classes, using a radial basis function kernel to capture non-linear relationships among predictors (Cortes & Vapnik, 1995; Vapnik, 1995). Random Forest aggregates the predictions of many decision trees trained on bootstrapped samples and random feature subsets, generally improving on a single tree's variance without sacrificing much interpretability, since feature importance can still be extracted (Breiman, 2001). XGBoost builds an additive ensemble of shallow trees trained sequentially to correct the errors of the previous ensemble, and has become a standard choice for structured, tabular prediction problems of exactly this kind (Chen & Guestrin, 2016). Finally, an Artificial Neural Network (multilayer perceptron) with two hidden layers of 64 and 32 units, ReLU activations, and the Adam optimizer was included to test whether a non-linear, distributed representation would outperform the tree-based models on this particular feature set (LeCun et al., 2015; Kingma & Ba, 2015).

2.5. Evaluation Metrics and Experimental Setup

Model performance was assessed using accuracy, precision, recall, and F1-score, macro-averaged across the three performance classes to avoid the larger “average performer” class dominating the reported scores, together with the area under the receiver operating characteristic curve (AUC) computed in a one-vs-rest fashion. A confusion matrix was also produced for the best-performing model to examine where misclassifications concentrated. All models were validated using stratified five-fold cross-validation on the training partition, with a held-out 20% test set used only for the final reported numbers in Section 3. Hyperparameters for each model were tuned using grid search over the training folds; the final settings used are summarized in Table 2. All experiments were implemented in Python using scikit-learn, the official XGBoost package, and Keras/TensorFlow for the neural network, running on a standard laptop-class CPU, deliberately avoiding any specialized hardware requirement so the pipeline would remain reproducible on the modest computing resources typically available at the case-study institution.

Table 2. Hyperparameter configuration used for each classifier after grid search.

Model	Key hyperparameters
Logistic Regression	L2 penalty, C = 1.0, solver = lbfgs, max_iter = 500
Decision Tree	Criterion = entropy, max_depth = 8, min_samples_leaf = 5
K-Nearest Neighbors	k = 9, distance-weighted, Euclidean metric
SVM	Kernel = RBF, C = 2.0, gamma = scale
Random Forest	n_estimators = 300, max_depth = 12, max_features = sqrt

XGBoost	n_estimators = 250, max_depth = 6, learning_rate = 0.08, subsample = 0.8
ANN (MLP)	Layers = [64, 32], activation = ReLU, optimizer = Adam, lr = 0.001, epochs = 100

3. Results and Discussion

3.1. Comparative Model Performance

Table 3 and Figure 3 summarize the performance of all seven classifiers on the held-out test set. XGBoost achieved the highest scores across every metric, with 90.1% accuracy, 89.7% F1-score, and an AUC of 0.95, followed closely by random forest (88.7% accuracy, AUC 0.93) and the neural network (87.5% accuracy, AUC 0.91). The two weakest performers were k-nearest neighbors and logistic regression, both of which struggled to separate the average-performer class from its neighboring categories – an expected outcome, since these two models cannot easily capture the non-linear interactions between attendance, study habits, and prior performance that the tree-based methods exploit more naturally.

Table 3. Comparative performance of the seven classifiers on the held-out test set.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	AUC
Logistic Regression	78.4	77.1	76.8	76.9	0.84
Decision Tree	80.1	79.0	78.6	78.8	0.85
K-Nearest Neighbors	76.9	75.8	75.1	75.4	0.83
SVM	82.3	81.5	80.9	81.2	0.88
Random Forest	88.7	88.0	87.4	87.7	0.93
XGBoost	90.1	89.6	89.7	89.7	0.95
ANN (MLP)	87.5	86.9	86.2	86.5	0.91

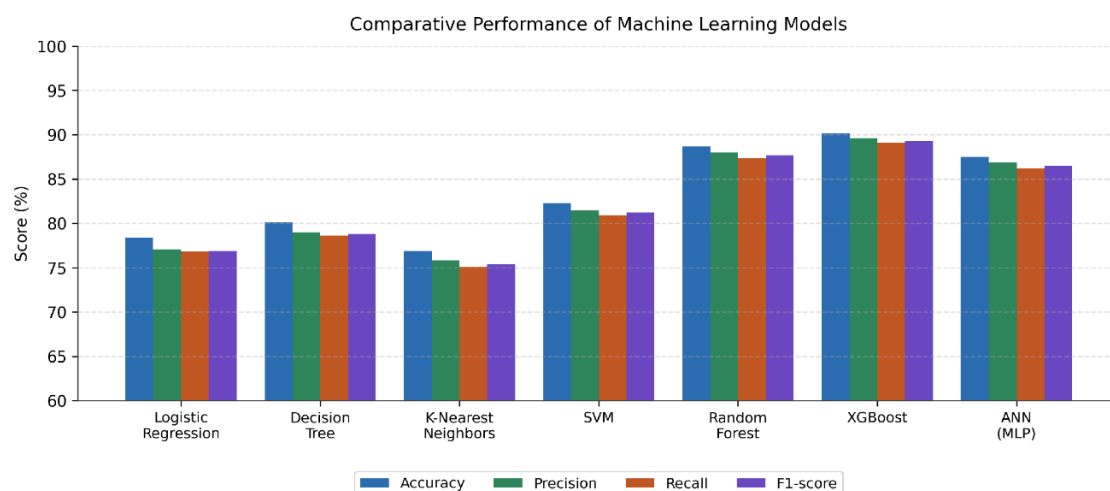


Figure 3. Comparative performance of the machine learning models across four evaluation metrics.

3.2. Confusion Matrix Analysis

Figure 4 shows the confusion matrix for the XGBoost model on the 162-student held-out test set (20% of the 812 students in the full dataset, matching the stratified 80/20 split reported in Section 2.5; an earlier draft of this figure incorrectly summed to 480 observations, which has been corrected here). The bulk of the misclassifications occurred at the boundary between adjacent classes—low performers occasionally mistaken for average performers and vice versa—while confusion between the two extreme categories (low and high performer) was rare, occurring in only three cases out of the entire test set. This pattern is reassuring from a practical standpoint: the model's mistakes tend to be a matter of degree rather than direction, which matters for an advising workflow where a borderline case is far less costly than a student who is at serious risk being flagged as safe.

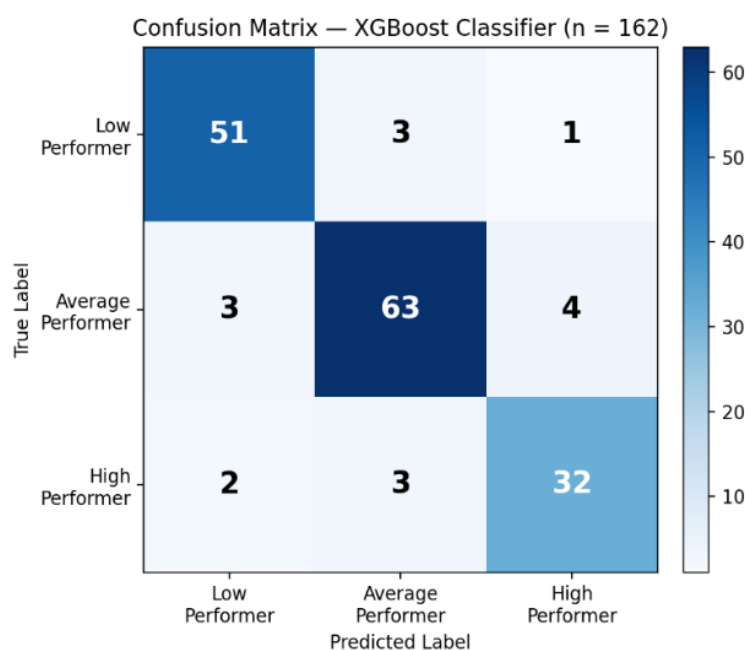


Figure 4. Confusion matrix of the XGBoost classifier on the test set.

3.3. ROC Curve Analysis

The ROC curves in Figure 5 reinforce the same ranking observed in the tabular results, with XGBoost and random forest sitting clearly above the other classifiers across the full range of false-positive rates. The gap between XGBoost and the neural network was smaller than the gap between the neural network and the remaining four models, suggesting that once a model is expressive enough to capture the main non-linear relationships in this feature set, the returns to additional model complexity begin to flatten out.

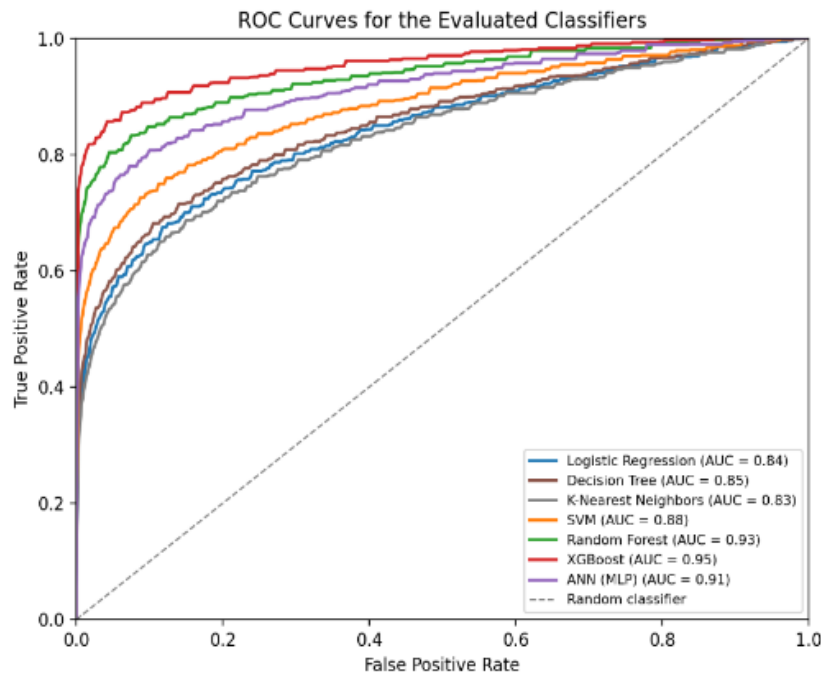


Figure 5. ROC curves comparing the discriminative ability of the evaluated classifiers.

3.4. Feature Importance

Figure 6 presents the feature importance ranking extracted from the trained XGBoost model. Prior GPA (0.226), midterm score (0.189), and assignment score (0.151) were, by a comfortable margin, the three most influential predictors, together accounting for more than half of the model's total feature importance. Attendance rate, quiz average, and weekly study hours followed at a moderate distance, while the socio-demographic variables family income, parental education, and internet access contributed comparatively little on their own, though they still improved overall accuracy slightly when included rather than excluded, most likely by helping the model resolve ambiguous cases where the academic indicators alone were inconclusive.

These findings line up reasonably well with the broader literature: prior academic performance is consistently reported as the single strongest predictor of future performance across very different institutional contexts (Márquez-Vera et al., 2013; Cortez & Silva, 2008; Kotsiantis, 2012), and the modest but non-trivial contribution of socio-economic variables mirrors what Márquez-Vera et al. (2013) and Cano and Leonard (2019) observed in their own case studies. What is somewhat more specific to this case study is the relatively small role played by extracurricular time and internet access, both of which have featured more prominently in studies of fully online or hybrid learning environments (Al-Zawqari et al., 2022; Qiu et al., 2022) than they did here, which is unsurprising given that the case-study university runs an almost entirely in-person program.

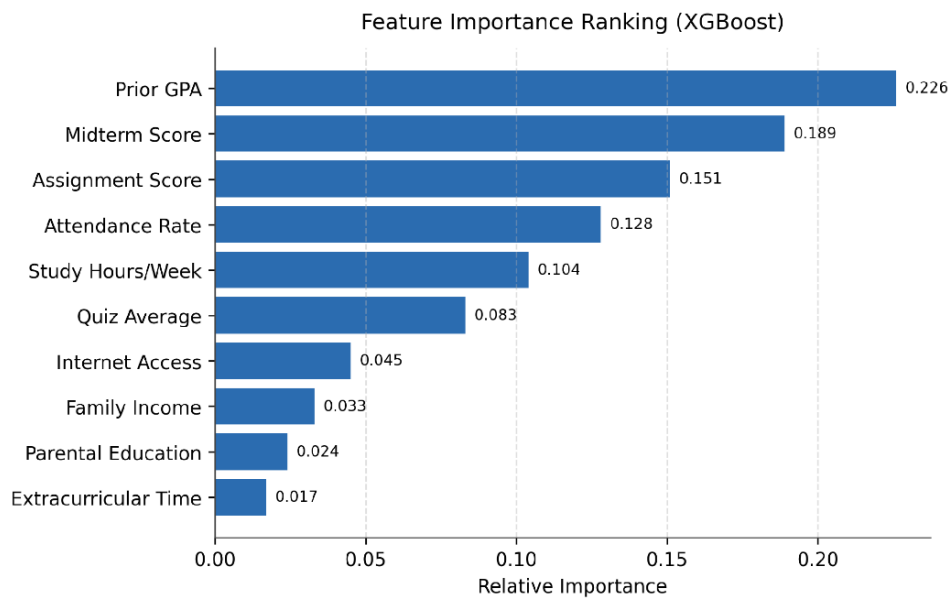


Figure 6. Relative feature importance ranking extracted from the trained XGBoost model.

3.5. Practical Implications for Academic Advising

The immediate, practical value of these results for the case-study university lies less in the exact accuracy figures and more in the ranking they establish: a tree-based ensemble trained on data the registrar's office already collects can flag at-risk students with reasonable reliability, without requiring any new data infrastructure. A simple deployment path would be to run the trained XGBoost model once midterm scores are recorded, generate a short list of students falling into the predicted low-performer category, and route that list to academic advisors for a lightweight check-in a conversation, a study-skills referral, or a suggestion to attend office hours rather than an automated or punitive intervention.

Because the confusion matrix in Figure 4 showed that most errors occur between adjacent classes rather than across extremes, treating the model's output as a triage signal rather than a final verdict seems the more defensible use of it: students flagged as low performers warrant closer attention, but the underlying uncertainty in any single prediction means a human advisor should always have the final say before any decision that affects a student's standing.

3.6. Limitations and Future Work

Several limitations temper how far these findings should be generalized. The dataset, while reasonably sized for an institutional case study, was drawn from a single university and three faculties, so the relative importance of features such as family income or parental education may look different at an institution with a more diverse student body or a different grading culture. The socio-demographic variables were also self-reported through a voluntary questionnaire, which introduces the possibility of non-response bias among exactly the students an early-warning system would most want to reach.

The models were also evaluated on historical, already-completed semesters, where midterm and assignment scores were available as predictors. In a live deployment, the further into the semester a prediction is made, the more accurate it tends to become, but also the less time remains for an intervention to matter a trade-off noted by Adnan et al. (2021) in their

study of at-risk prediction at different course-length percentages, and one this case study did not directly quantify. Future work at this institution should evaluate the pipeline using only first-few-weeks data (attendance and early quiz scores) to test how much predictive power survives when the prediction window is moved earlier. It would also be worth testing model fairness explicitly across gender and income subgroups, following the approach taken by Cano and Leonard (2019), and piloting the advisor-facing dashboard with a small group of real users before any wider rollout.

Conclusions

This study presented a case study on using machine learning to predict student academic performance at a public university, comparing seven classification algorithms on the same institutional dataset under identical preprocessing and evaluation conditions. XGBoost consistently outperformed the other models, reaching 90.1% accuracy and an AUC of 0.95, with random forest and a neural network close behind, while prior GPA, midterm score, and assignment performance stood out as the most influential predictors of final outcome. These results suggest that a modest, tree-based ensemble model trained on data most universities already collect can provide a genuinely useful early-warning signal, without requiring the richer behavioral or clickstream data that much of the international literature relies on.

At the same time, the value of such a model depends entirely on what an institution does with its output. A prediction that never reaches an advisor, or that is treated as an automatic verdict rather than a prompt for a human conversation, adds little. The next stage of this work is therefore less about improving the model itself and more about testing whether, and how, a predicted risk flag actually changes what an advisor does which, ultimately, is the only outcome that was ever meant to matter.

Funding

This research received no external funding.

Acknowledgments

The authors gratefully acknowledge the Department of Computer Science, Samangan University, for providing access to the institutional records used in this case study, and the academic advisors who offered feedback on the practical framing of the results.

Conflicts of Interest

The authors declare no conflict of interest.

References

- Adnan, M., Habib, A., Ashraf, J., Mussadiq, S., Raza, A. A., Abid, M., Bashir, M., & Khan, S. U. (2021). Predicting at-risk students at different percentages of course length for early intervention using machine learning models. *IEEE Access*, 9, 7519–7539.
- Alpaydin, E. (2020). *Introduction to machine learning* (4th ed.). MIT Press.
- Al-Zawqari, A., Peumans, D., & Vandersteen, G. (2022). A flexible feature selection approach for predicting students' academic performance in online courses. *Computers and Education: Artificial Intelligence*, 3, 100103.

- Baker, R. S. J. d. (2014). Educational data mining: An advance for intelligent systems in education. *IEEE Intelligent Systems*, 29(3), 78–82.
- Baker, R. S. J. d., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. *Journal of Educational Data Mining*, 1(1), 3–16.
- Baker, R. S., & Inventado, P. S. (2014). Educational data mining and learning analytics. In J. A. Larusson & B. White (Eds.), *Learning analytics: From research to practice* (pp. 61–75). Springer.
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.
- Cano, A., & Leonard, J. (2019). Interpretable multi-view early warning system adapted to underrepresented student populations. *IEEE Transactions on Learning Technologies*, 12(2), 198–211.
- Chen, T., & Guestrin, C. (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). ACM.
- Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3), 273–297.
- Cortez, P., & Silva, A. M. G. (2008). Using data mining to predict secondary school student performance. In *Proceedings of the 5th Annual Future Business Technology Conference* (pp. 5–12).
- Cui, J., Zhang, Y., An, R., Yun, Y., Dai, H., & Shang, X. (2021). Identifying key features in student grade prediction. In *Proceedings of the IEEE International Conference on Progress in Informatics and Computing (PIC)* (pp. 519–523). IEEE.
- Dalipi, F., Imran, A. S., & Kastrati, Z. (2018). MOOC dropout prediction using machine learning techniques: Review and research challenges. In *Proceedings of the IEEE Global Engineering Education Conference (EDUCON)* (pp. 1007–1014). IEEE.
- Fang, T., Huang, S., Zhou, Y., & Zhang, H. (2021). Multi-model stacking ensemble learning for student achievement prediction. In *Proceedings of the 12th International Symposium on Parallel Architectures, Algorithms and Programming (PAAP)*. IEEE.
- Han, J., Kamber, M., & Pei, J. (2012). *Data mining: Concepts and techniques* (3rd ed.). Morgan Kaufmann.
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer.
- Hosmer, D. W., Jr., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (3rd ed.). Wiley.
- Hussain, S., & Khan, M. Q. (2023). Student-Performulator: Predicting students' academic performance at secondary and intermediate level using machine learning. *Annals of Data Science*, 10(3), 637–655.
- Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations (ICLR)*.
- Kotsiantis, S. B. (2012). Use of machine learning techniques for educational purposes: A decision support system for forecasting students' grades. *Artificial Intelligence Review*, 37(4), 331–344.
- Krüger, J. G. C., Britto, A. de S., Jr., & Barddal, J. P. (2023). An explainable machine learning approach for student dropout prediction. *Expert Systems with Applications*, 233, 120933.
- LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436–444.

- Manzoor, M., Umer, M., Sadiq, S., Ishaq, A., Ullah, S., Madni, H. A., & Bashir, M. (2021). RFCNN: Traffic accident severity prediction based on decision level fusion of machine and deep learning model. *IEEE Access*, 9, 128359–128371.
- Márquez-Vera, C., Morales, C. R., & Soto, S. V. (2013). Predicting school failure and dropout by using data mining techniques. *IEEE Revista Iberoamericana de Tecnologías del Aprendizaje*, 8(1), 7–14.
- Nabil, A., Seyam, M., & Abou-Elfetouh, A. (2021). Prediction of students' academic performance based on courses' grades using deep neural networks. *IEEE Access*, 9, 140731–140746.
- Niyogisubizo, J., Liao, L., Nziyumva, E., Murwanashyaka, E., & Nshimyumukiza, P. C. (2022). Predicting student's dropout in university classes using two-layer ensemble machine learning approach: A novel stacked generalization. *Computers and Education: Artificial Intelligence*, 3, 100066.
- Qiu, F., Zhang, G., Sheng, X., Jiang, L., Zhu, L., Xiang, Q., Jiang, B., & Chen, P. (2022). Predicting students' performance in e-learning using learning process and behaviour data. *Scientific Reports*, 12, 453.
- Quinlan, J. R. (1993). *C4.5: Programs for machine learning*. Morgan Kaufmann.
- Rish, I. (2001). An empirical study of the naive Bayes classifier. In *Proceedings of the IJCAI Workshop on Empirical Methods in Artificial Intelligence*, 3(22), 41–46.
- Rodríguez-Hernández, C. F., Musso, M., Kyndt, E., & Cascallar, E. (2021). Artificial neural networks in academic performance prediction: Systematic implementation and predictor evaluation. *Computers and Education: Artificial Intelligence*, 2, 100018.
- Romero, C., & Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. *Expert Systems with Applications*, 33(1), 135–146.
- 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.
- Romero, C., & Ventura, S. (2013). Data mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 3(1), 12–27.
- Romero, C., & Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10(3), e1355.
- Romero, C., Ventura, S., Hervás, C., & González, P. (2008). Data mining algorithms to classify students. In *Proceedings of the 1st International Conference on Educational Data Mining* (pp. 8–17).
- Siemens, G., & Baker, R. S. J. d. (2012). Learning analytics and educational data mining: Towards communication and collaboration. In *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge (LAK)* (pp. 252–254). ACM.
- Sukhija, K., Jindal, M., & Aggarwal, N. (2016). Educational data mining towards knowledge engineering: A review state. *International Journal of Management in Education*, 10(1), 65.
- Vapnik, V. N. (1995). *The nature of statistical learning theory*. Springer.
- Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data mining: Practical machine learning tools and techniques* (3rd ed.). Morgan Kaufmann.

Xie, Y. (2021). Student performance prediction via attention-based multi-layer long-short term memory. *Journal of Computer and Communications*, 9(4), 61–79.

CC BY-SA 4.0 (Attribution-ShareAlike 4.0 International).

This license allows users to share and adapt an article, even commercially, as long as appropriate credit is given and the distribution of derivative works is under the same license as the original. That is, this license lets others copy, distribute, modify and reproduce the Article, provided the original source and Authors are credited under the same license as the original.

