Proactive Analytics for Student Success: A Data-Driven Approach

Mariem Nsiri Laura Keyes
Technological University Dublin
B00177462@mytudublin.ie laura.keyes@tudublin.ie

May 21, 2025

Abstract

This study introduces an academic risk detection framework using academic, behavioral, and psychological data to identify at-risk students early. Key factors like study hours and mental health were analyzed, followed by thorough preprocessing and feature engineering. Ensemble models, especially XGBoost, showed strong accuracy and interpretability. Evaluation confirmed its practical value for early intervention, offering a scalable, data-driven tool to help educators provide timely academic support and improve student outcomes.

Keywords: Academic Risk Detection, Ensemble Learning, Prediction.

1 Introduction

The growing availability of educational data has enabled predictive modeling to support early identification of at-risk students, an increasingly vital goal for institutions aiming to reduce dropout rates and improve outcomes. While many existing models rely on academic or demographic features, they often neglect behavioral and psychological factors that may offer earlier, more actionable insights. Although recent advances in Educational Data Mining (EDM) and machine learning have improved prediction accuracy, many models still suffer from narrow feature sets, limited interpretability, and poor timing, often flagging students too late for effective intervention.

This study addresses these limitations by introducing a robust academic risk detection framework integrating academic, behavioral, lifestyle, and psychological indicators. Using advanced ensemble techniques, particularly XGBoost, with focused data exploration, feature selection, and preprocessing, the model emphasizes predictive accuracy and interpretability. Based on a synthetic dataset reflecting diverse student traits, we develop a system to identify at-risk students 6–8 weeks before final assessments. Our contributions include holistic feature integration, temporal sensitivity, and strong model performance, offering a scalable, practical early warning system for real-world educational settings.

Building on the literature review in Section 2, the framework integrates academic, behavioral, and psychological data for early risk detection. Section 3.1 identifies key drivers like study time and mental health, and Section 3.2 details preprocessing and feature engineering. Ensemble modeling techniques are presented in Section 4, with evaluation results in Section 5 showing strong performance. Section 6 concludes with future directions and system scalability.

2 Literature Review

Several studies have explored student performance prediction using Educational Data Mining (EDM), each offering distinct contributions to understanding and modeling academic risk. Authors in (Khan and Ghosh, 2021) conducted a systematic review of 140 EDM studies focused exclusively on classroom-based learning. Their work is notable for considering the temporal aspect of prediction, examining not just which factors influence performance, but also when prediction is most effective. The study found that while prediction during the course is generally effective, forecasting performance before course commencement remains a significant challenge. In a secondary education context, Authors in (Cortez and Silva, 2008) applied various data mining methods, Decision Trees, Random Forests, Neural Networks, and Support Vector Machines, to predict student success in core subjects such as Mathematics and Portuguese. The results showed that prior grades, especially from the first and second school periods, were strong predictors. However, other contextual features like absences, parental education, and alcohol consumption also contributed meaningfully to prediction accuracy, highlighting the value of holistic data integration. In (Yang and Li, 2018), the authors introduced a comprehensive framework using a Student Attribute Matrix (SAM) and Back Propagation Neural Networks to estimate performance, track progress, and assess potential. This work uniquely emphasised causal relationships among attributes and proposed metrics for understanding not just current performance, but also development potential, thereby offering a multi-dimensional view of student learning dynamics. Finally, the work in (Osmanbegovic and Suljic, 2012) explored the use of data mining in a university setting, analyzing how socio-demographic variables, high school performance, entrance exam results, and attitudes toward studying affect academic success. Their findings support the development of decision systems for higher education, underlining the importance of combining academic and behavioral data to enhance predictive reliability and inform interventions.

Our study extends prior work by integrating behavioral, lifestyle, and psychological factors with academic data to identify at-risk students early. Unlike earlier research focused on academic predictors, we use diverse, measurable indicators and prioritize early intervention. Using ensemble methods with feature selection, our model achieves high accuracy and interpretability, making it practical for educational settings. This approach addresses temporal prediction gaps and offers a scalable framework adaptable to different contexts.

3 Academic Risk Detection Framework

This study pursued two primary objectives: (1) a *business objective* to develop an early warning system capable of identifying at-risk students 6 to 8 weeks before final assessments, and (2) a *data mining objective* to build interpretable classification models that leverage behavioral data to guide timely academic interventions. To this end, we utilized a recent dataset from Kaggle (Nath, 2025), containing 1,000 student records with 16 academic, behavioral, and psychological features.

3.1 Data exploration and key performance drivers

The analysis revealed strong behavioral indicators associated with academic performance. Study time was the most predictive factor, with students who studied over three hours daily achieving significantly better outcomes. The Pearson correlation coefficient (r) between study hours and exam scores was high ($r = 0.83$), indicating a strong positive linear relationship. Mental health ratings, though self-reported, also showed a moderate positive correlation ($r = 0.32$). Students with low well-being scores (below 4 out of 10) experienced substantially higher failure rates compared to peers with higher mental health ratings. These patterns suggest that both academic effort and mental well-being contribute meaningfully to student success.

Digital behavior also influenced academic risk. Netflix usage demonstrated a mild negative correlation with academic performance ($r = -0.17$), stronger than that of general social media use. Although the correlation was weak, students who spent more than 3.5 hours on media daily faced nearly twice the risk of failure, underscoring the potential cumulative impact of excessive screen time. In contrast, conventional engagement metrics such as attendance (mean $\approx 84\%$) and part-time employment status had limited predictive value ($r = 0.09$, $p > 0.68$). Statistical validation confirmed the significance of media usage ($F = 26.17$, $p < 0.001$) and mental health ($F = 51.22$, $p < 0.001$) as key performance drivers. Demographic variables like gender and parental education were not significant. Internal consistency checks showed no implausible time use or behavioral reports, supporting the dataset's structural validity despite its synthetic origin.

3.2 Data preparation and feature engineering

Before preprocessing, a baseline decision tree model was built using all features, including proxy attributes such as exam score and grade, which directly influence the risk status label. This model achieved perfect performance, highlighting the strong predictive power of these proxies and their direct alignment with the target variable. To obtain a realistic evaluation, these proxy features were removed, leading to more moderate and credible results. Irrelevant columns like student ID were also dropped to reduce noise and prevent overfitting during model training and validation.

The remaining features were analyzed for correlations, revealing no significant multicollinearity, so all were retained. Missing values in categorical attributes such as parental education level were imputed using the mode, which improved predictive accuracy more than KNN imputation. Outliers detected via Isolation Forest (Liu et al., 2008) were kept to preserve natural data variance and maintain the integrity of the modeling process.

To improve feature representation and impact, Principal Component Analysis (PCA) (Pearson, 1901) was applied to related variables like social media hours and netflix hours, creating a composite media usage hours PCA feature. Continuous variables were discretized using a supervised decision tree-based binning method to capture nonlinear patterns, while categorical variables were ordinally encoded after imputation. Borderline SMOTE (Han et al., 2005) addressed the class imbalance between at risk and not at risk groups by generating synthetic samples near the decision boundary, enhancing classifier sensitivity.

Embedded feature selection using a decision tree classifier was applied but did not remove any features at this stage of preprocessing, indicating potentially that all variables carry potential predictive value. This technique is expected to play a more significant role during the modeling phase with more powerful algorithms, enhancing interpretability and generalization. The entire preprocessing and initial modeling pipeline was validated through randomized hyperparameter search combined with cross-validation, resulting in an optimized decision tree model with a depth of 6, 19 leaves, and moderate pruning (`ccp_alpha` = 0.0067) to effectively balance complexity and prevent overfitting. This model achieved around 90% accuracy and demonstrated strong, reliable performance with good precision and recall across both at-risk and not-at-risk classes.

Building on this foundation, the next phase focuses on a broader evaluation of modeling techniques and their performance.

4 Modeling, Performance Comparison

Building on the preprocessing pipeline, two ensemble classifiers, Random Forest (Breiman, 2001) and XGBoost (Chen and Guestrin, 2016), were developed to predict students at risk. Feature selection was standardized using XGBoost's embedded importance method, resulting in eight key features that capture both academic and lifestyle factors. This approach leverages multivariate interactions beyond simple correlations, enhancing model focus and robustness.

Both models demonstrated strong predictive performance, with XGBoost slightly outperforming Random Forest across accuracy (94.79% vs. 93.84%), precision, recall, and F1-score. Notably, XGBoost improved detection of at-risk students, benefiting from regularization and subsampling techniques that mitigate overfitting. The higher AUC-ROC for XGBoost (0.95 vs. 0.94) confirms its superior generalization, positioning it as the preferred model. Moreover, XGBoost's consistent performance across folds reinforces its reliability for real-world educational deployment.

5 Evaluation Results

The final evaluation results confirmed the model's ability to generalize well to unseen data, achieving high precision and recall in detecting at-risk students. While not flawless, its balanced performance across key metrics demonstrates practical utility in educational settings, particularly for early intervention. Importantly, the model's interpretability allowed for meaningful insights into contributing risk factors, enabling informed decision-making by educators. These outcomes highlight the system's potential as a valuable complement to human judgment, supporting timely and data-driven academic guidance.

Beyond accuracy metrics, the model offers actionable value by flagging students who may benefit from targeted academic support, counseling, or engagement strategies. This predictive capacity can help educators allocate resources more effectively and intervene before academic failure occurs. With ongoing data collection and monitoring, the system can adapt over time to reflect changing student behaviors and institutional contexts. As such, it holds promise not only for static analysis but as a dynamic tool integrated into continuous learning support systems.

6 Conclusion

This study presents an effective early warning system that integrates behavioral, lifestyle, psychological, and academic data to identify students at risk of academic difficulties well before final grades are available. Through careful data preprocessing, feature selection, and the application of ensemble learning methods such as XGBoost, the model achieves strong predictive performance while maintaining interpretability. Our approach addresses key limitations found in prior research by emphasizing early intervention and incorporating diverse, real-time indicators beyond traditional academic metrics. Although challenges remain, such as dataset size and scalability, the results demonstrate the potential for data-driven tools to support timely, targeted educational interventions.

Future work will focus on expanding datasets, improving model fairness, and validating the system across broader educational contexts to further enhance its practical impact. Additionally, incorporating association rule mining could uncover hidden patterns and relationships among behavioral and academic variables that are not captured through classification models alone. These insights can complement predictive outputs by highlighting co-occurring risk factors, enriching intervention strategies with more nuanced, explainable rules. This hybrid approach can offer educators deeper, actionable understanding to inform more personalized support mechanisms.

References

Breiman, L. (2001). Random forests. *Machine learning*, 45:5–32.

Chen, T. and Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794.

Cortez, P. and Silva, A. M. G. (2008). Using data mining to predict secondary school student performance. In *Proceedings of the 5th Annual Future Business Technology Conference*, pages 5–12, Porto, Portugal. EUROSIS-ETI.

Han, H., Wang, W.-Y., and Mao, B.-H. (2005). Borderline-SMOTE: a New Oversampling Method in Imbalanced Data Sets Learning. In *International conference on intelligent computing*, pages 878–887. Springer.

Khan, A. and Ghosh, S. K. (2021). Student performance analysis and prediction in classroom learning: A review of educational data mining studies. *Education and information technologies*, 26(1):205–240.

Liu, F. T., Ting, K. M., and Zhou, Z.-H. (2008). Isolation Forest. In *2008 eighth ieee international conference on data mining*, pages 413–422. IEEE.

Nath, J. (2025). Student habits vs academic performance. Dataset. URL: https://www.kaggle.com/datasets/jayaantanaath/student-habits-vs-academic-performance.

Osmanbegovic, E. and Suljic, M. (2012). Data mining approach for predicting student performance. *Economic Review: Journal of Economics and Business*, 10(1):3–12.

Pearson, K. (1901). Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572.

Yang, F. and Li, F. W. (2018). Study on student performance estimation, student progress analysis, and student potential prediction based on data mining. *Computers & education*, 123:97–108.