School of Executive Education and Lifelong Learning

Predicting Student Dropout Risk Using Supervised Machine Learning

An AI and ML Capstone Project: Education Domain

Submitted by: Conrado Banjo B Mamaradlo, Jr.

Program: Postgraduate Diploma in AI and Machine Learning

Advisor: Professor Emeritus Carmen Tagliente

Date of Submission: January 2026

ABSTRACT

Student dropout presents a persistent and multifaceted challenge for higher education institutions, affecting academic outcomes, financial sustainability, and long-term student success. This capstone project examines the application of supervised machine learning as a decision-support mechanism for identifying students at elevated risk of dropping out early in their academic journey. The study integrates analytical rigor with institutional strategy, demonstrating how predictive models can be used responsibly to improve prioritization, optimize limited resources, and strengthen governance while maintaining human oversight.

INTRODUCTION AND PROBLEM CONTEXT

Student attrition is rarely a sudden or isolated occurrence. Rather, it is typically the result of accumulating academic, financial, and personal pressures that manifest over time. Traditional early-warning systems often rely on lagging indicators such as failed courses or prolonged absence, resulting in interventions that are reactive, costly, and frequently ineffective. As enrollment environments become more complex, institutions require earlier and more structured decision support to intervene when support is most likely to succeed.

This project addresses the strategic need for evidence-based prioritization. Instead of replacing academic judgment, the proposed approach supports advisors and leaders by highlighting which students may require attention first, enabling proactive and equitable allocation of support services.

OBJECTIVES AND SCOPE

The primary objective of this study is to develop and evaluate a predictive framework capable of estimating student dropout risk using enrollment-time and early academic data. The scope is intentionally limited to decision support rather than automated decision-making. Model outputs are expressed as probabilities and rankings, ensuring that final intervention decisions remain with institutional professionals.

DATA AND METHODOLOGICAL APPROACH

The analysis is based on a publicly available dataset from the UCI Machine Learning Repository containing 4,424 student records and 36 predictor variables. These variables capture academic pathways, demographic characteristics, socioeconomic indicators, financial status, and early academic performance. A supervised classification framework was employed, with dropout risk modeled as a binary outcome to support intervention prioritization.

To ensure robustness and reproducibility, standardized preprocessing pipelines, stratified cross-validation, and probability-based evaluation metrics were applied consistently across models. This methodological discipline ensures that results are transparent, repeatable, and suitable for institutional review.

MODEL PERFORMANCE AND RESULTS

Multiple modeling approaches were evaluated, including logistic regression, random forests, support vector machines, and gradient boosting. Across these models, strong and consistent ranking performance was observed, with ROC-AUC values approaching 0.91. This level of performance indicates that the models are

effective at distinguishing higher-risk students from lower-risk peers, even in the presence of class imbalance.

Rather than focusing solely on classification accuracy, performance was interpreted through the lens of operational capacity. By prioritizing the top segment of highest-risk students, institutions can focus limited resources where they are most likely to have impact.

BUSINESS VALUE, ROI, AND STRATEGIC IMPLICATIONS

From a strategic perspective, the value of this project lies in its ability to convert data into actionable insight. Early identification enables targeted outreach, reducing the cost and inefficiency associated with late-stage interventions. Even modest improvements in retention among prioritized students can yield meaningful financial returns through stabilized tuition revenue and improved cohort continuity.

The framework also supports strategic planning by providing leaders with visibility into risk distribution across cohorts, enabling informed decisions about staffing, program design, and student support investments.

RISK MANAGEMENT, ETHICS, AND GOVERNANCE

The use of predictive analytics in education introduces important risks related to bias, data drift, and unintended consequences. This project incorporates fairness auditing, human-in-the-loop safeguards, and clear governance recommendations to mitigate these risks. Model outputs are framed as contextual indicators rather than deterministic judgments, preserving institutional accountability.

LIMITATIONS AND FUTURE WORK

Several limitations must be acknowledged. The reliance on historical data may not fully capture future cohort dynamics, and some proxy variables may reflect underlying structural inequalities. Future work should focus on continuous monitoring, recalibration, and institution-specific validation, as well as the integration of qualitative insights from advisors and students.

ACADEMIC TOOLS AND AI ASSISTANCE DISCLOSURE

This capstone project utilized contemporary computational and productivity tools in accordance with academic integrity standards. Google Colab was used as the primary execution environment for the Jupyter notebook to enable reproducible analysis, model training, and evaluation using Python-based machine learning libraries.

Additionally, ChatGPT Plus was used as a supportive writing and structuring aid during the preparation of narrative components, including drafting, language refinement, and organization of sections in the final report and presentation materials. The use of these tools did not replace original analysis, critical thinking, or academic judgment, but served to enhance clarity, efficiency, and documentation quality.

CONCLUSION

This capstone project demonstrates that supervised machine learning can be responsibly applied as a decision-support tool to address student dropout risk. By integrating technical rigor with business strategy and ethical governance, the study provides a practical blueprint for institutions seeking to improve student success while

managing risk and resource constraints. The findings underscore the importance of collaboration between human expertise and intelligent systems.

REFERENCES

Martins, M. V., Tolledo, D., Machado, J., Baptista, L. M. T., & Realinho, V. (2021). Early prediction of student's performance in higher education: A case study. In Trends and Applications in Information Systems and Technologies (Advances in Intelligent Systems and Computing). Springer. https://doi.org/10.1007/978-3-030-72657-7_16

Realinho, V., Vieira Martins, M., Machado, J., & Baptista, L. (2021). Predict Students' Dropout and Academic Success [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C5MC89

ENCLOSURES

Jupyter Notebook: PGD in AI and ML Capstone Project.ipynb

Microsoft Powerpoint:

Appendix A. Technical Deck.pptx

Appendix B. Business Deck.pptx

Github Repo Link: https://github.com/NeuroSage/student-dropout-risk-prediction