Identifying Academically At-Risk Student using Predictive Analysis Model

Joshua Reyes
San Carlos College
Galarin Urbiztondo
Pangasinan, Philippines

Roy Wilhem Ferrer
San Carlos College
Paitan Panoypoy San Carlos City
Pangasinan, Philippines

Reymart Jay Epan
San Carlos College
Dumpay, Basista,
Pangasinan, Philippines

M.A. Lourdes Villapando
San Carlos College
Mestizo Norte
Pangasinan, Philippines

Maynard Gel F. Carse
San Carlos College
Abanon, San Carlos City
Pangasinan, Philippines

Aldrich Michael B. Garcia
San Carlos College
Rizal St., San Carlos City
Pangasinan, Philippines

ABSTRACT
The increasing dropout rates in higher education institutions underscore the critical need for proactive strategies to identify academically at-risk students. This study presents the development and evaluation of a predictive analysis model leveraging machine learning—specifically the Random Forest algorithm—to accurately identify students at risk of academic failure. The model integrates both academic indicators (e.g., GPA, attendance, exam scores) and non-academic factors (e.g., socio-economic status, behavioral patterns, family dynamics) to provide a holistic assessment of student performance. A dataset of 100,256 student records from the Australian Student Performance Dataset was preprocessed, with key features selected to enhance model accuracy. The model achieved a predictive accuracy of 69% and was deployed through a web-based application developed using the Flask framework. Functionality includes real-time prediction, risk classification, and user-friendly visualization. Stakeholder evaluation involving 40 respondents showed 88% user satisfaction, confirming the system's reliability, usability, and practical value. The findings demonstrate the model's effectiveness in enabling early interventions, thereby contributing to reduced attrition rates and more inclusive, data-informed educational practices.

Keywords
Predictive analysis, academically at-risk students, machine learning, data mining, student retention, academic performance, grades, test scores, socio-economic factors, behavioral patterns, real-time data, personalized learning support, educational interventions, attrition rate

1. INTRODUCTION
The rising dropout rates in higher education institutions have become a pressing concern, with attrition rates reaching 40.98% in the 2022-2023 academic year (Sarao, 2023). This trend highlights the urgent need for proactive measures to identify and support academically at-risk students before they disengage from their studies. Traditional methods of identifying struggling students often rely on reactive measures, such as poor exam performance or attendance records, which may be too late for effective intervention. Predictive analytics, powered by machine learning, offers a transformative approach by analyzing historical and real-time data to forecast academic risks early.

This study leverages the **Random Forest algorithm** to develop a predictive model that integrates both **academic** (e.g., GPA, attendance, exam scores) and **non-academic factors** (e.g., socio-economic status, behavioral patterns, family dynamics). The model was trained on the **Australian Student Performance Dataset**, comprising **100,256 student records**, to ensure robustness and generalizability. A **web-based application** was developed using the **Flask framework**, allowing educators to input student data and receive real-time risk assessments. The system's effectiveness was evaluated through stakeholder feedback, with **88% of respondents** reporting satisfaction with its accuracy and usability.

The primary objectives of this study are:

- To develop a **highly accurate predictive model** for identifying at-risk students.
- To **integrate academic and non-academic factors** for a holistic risk assessment.
- To **deploy a user-friendly web application** for real-time predictions.
- To **evaluate the model's effectiveness** through stakeholder feedback.

By enabling early intervention, this model aims to **reduce dropout rates**, **optimize resource allocation**, and **enhance student success** through data-driven decision-making

2. RELATED LITERATURE
Student dropout is a persistent and multifaceted challenge in education systems worldwide, with far-reaching consequences for individuals and society. Research consistently demonstrates that academic struggles alone do not fully explain dropout patterns; rather, they intersect with complex socio-economic and demographic factors to create varying levels of risk. Studies reveal significant disparities in dropout rates among different racial and ethnic groups, with Hispanic students facing nearly four times the risk of Asian students (American Community Survey, 2010), while students from low-income families or single-parent households show markedly higher attrition rates (Whitcomb et al., 2021). These systemic inequalities are compounded by academic warning signs like chronic absenteeism, which doubles failure risk (Glavin, n.d.), and poor performance in foundational courses that often

precede later academic collapse. In response to these challenges, educational researchers have increasingly turned to predictive analytics as a proactive solution, with machine learning models - particularly Random Forest algorithms - emerging as particularly effective tools. These models achieve remarkable accuracy (up to 96% in discipline-specific applications) by synthesizing both traditional academic metrics (grades, attendance records, assignment completion rates) and non-traditional indicators (family income, mental health status, transportation access, and peer evaluations) (Qushem et al., 2023; Namoun & Alshanqiti, 2021). The most successful implementations employ rigorous data science methodologies, including proper dataset segmentation (typically 80% training, 20% testing data), cross-validation techniques to prevent overfitting, and careful feature selection to optimize model performance (Srinivas et al., 2018). Importantly, these technological solutions are most impactful when paired with institutional support systems, as demonstrated by interventions that combine predictive analytics with targeted academic support, counseling services, and financial assistance programs (Alyahyah et al., 2020). This growing body of research underscores the transformative potential of data-driven approaches in education, while also highlighting the need for ethical considerations in data collection, model transparency, and the avoidance of algorithmic bias when implementing such systems across diverse student populations.

3. PRESENTATION OF ANALYSIS OF DATA

Figure 1. Operational Framework for the Proposed Project

The proposed project employs a systematic, data-driven approach to identify academically at-risk students through predictive modeling. The process begins with comprehensive data collection, gathering key academic indicators (grades, attendance records, assignment scores) and relevant demographic information (socio-economic status, family background) to establish a robust foundation for analysis. Following data acquisition, the framework incorporates a rigorous feature selection process to identify the most influential factors impacting student performance, utilizing statistical methods and domain expertise to prioritize variables with the strongest predictive power. The selected features then feed into a machine learning pipeline where the dataset undergoes an 80-20 split - with 80% allocated for model training and 20% reserved for testing - ensuring proper validation of results while maintaining sufficient data for algorithm development. During the training phase, the model learns complex patterns and relationships within the historical student data, while the testing phase evaluates its predictive accuracy on unseen cases, providing an objective measure of real-world applicability. This structured methodology not only facilitates early identification of at-risk students but also enables continuous model refinement through iterative testing and validation cycles, ultimately creating a dynamic tool that can adapt to evolving educational contexts and student populations. The operational framework's phased implementation - from data collection to model deployment - ensures both methodological rigor and practical utility in educational settings.

Figure 2: Overview of the Experimental Design

The experimental design follows a structured, data-driven pipeline to develop an accurate predictive model for identifying at-risk students. The process begins with comprehensive **feature selection**, where key academic indicators (attendance, test scores, grades) and non-academic factors (socioeconomic background, family dynamics, behavioral patterns) are analyzed using **Random Forest** to determine their predictive importance. Once the most relevant features are identified, the dataset is **split into training (80%) and testing (20%) subsets**, ensuring the model learns from historical data while being validated on unseen cases to prevent overfitting. The **Random Forest algorithm** is then applied, leveraging an ensemble of decision trees—each trained on random subsets of data and features—to enhance accuracy and robustness by aggregating multiple predictions. This approach not only improves generalization but also handles complex relationships within high-dimensional student data. Finally, the model's performance is rigorously evaluated using standard metrics (accuracy, precision, recall) to ensure reliable predictions, enabling educators to implement timely, data-informed interventions for at-risk students

.Figure 3. Key Features for At-Risk Student Prediction

The predictive model identifies 11 critical features that most strongly indicate academic risk, with GPA emerging as the most significant predictor of student performance. Academic metrics like Attendance Rate, Final Exam Scores, and Project/Assignment Scores provide direct insight into student engagement and achievement, while High School GPA and Entrance Exam Scores offer valuable historical context about academic preparedness. Non-academic factors such as Family Income and Distance from Home to University reveal socioeconomic challenges that may impact learning, whereas Peer Reviews and Peer Evaluations capture social and behavioral influences on academic success. Together, these features form a comprehensive profile that enables accurate risk assessment, with the Random Forest algorithm weighting each factor based on its predictive importance. By focusing on these key indicators—while excluding less relevant variables—the model maintains both high accuracy and practical interpretability for educators seeking to implement targeted interventions.

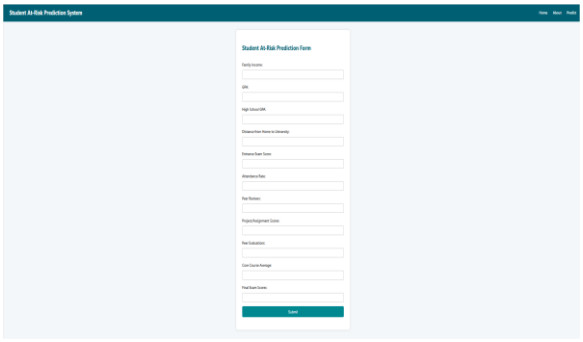

Figure 4. Prediction Forn Page Overview

Prediction Form Page of the Student At-Risk Prediction System, designed as an intuitive interface for educators to input critical academic and personal data—including **GPA, attendance rate, family income, entrance exam scores, high school GPA, and peer evaluations**—to assess a student's likelihood of academic struggle. Upon submission, the system processes these inputs using the trained predictive model to generate a **Risk Score**, classifying students as either **"At Risk"** (flagged by indicators like poor attendance, low scores, or socioeconomic challenges) or **"Not at Risk"** (reflecting stable performance and positive engagement). This real-time assessment enables early identification of vulnerable students, guiding targeted interventions such as tutoring, counseling, or financial support to mitigate dropout risks and enhance academic success. The form's structured design ensures ease of use while maintaining comprehensive evaluation, bridging data-driven insights with actionable educational support.

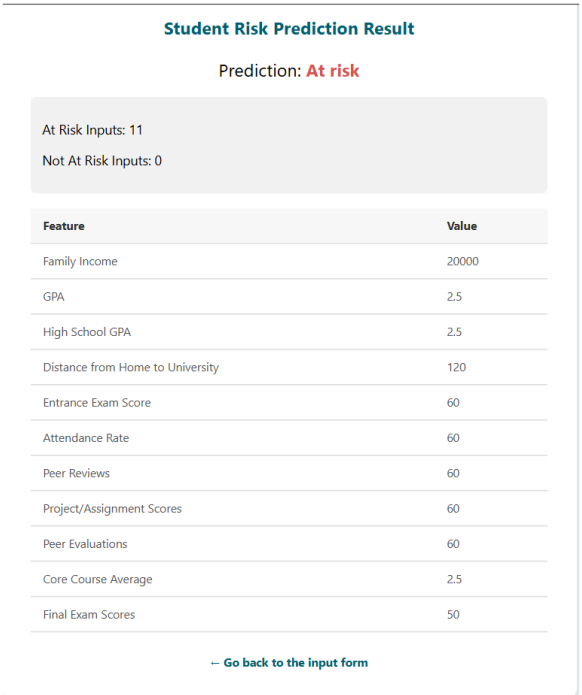

Figure 5.Prediction Result Page (At Risk)

Result Page of the Student At-Risk Prediction System, displaying a comprehensive risk assessment outcome when a student is identified as **"At Risk."** The dashboard clearly highlights this classification, providing educators with immediate visibility into the student's academic vulnerability. Below the result, a detailed **Feature Inputs table** summarizes the contributing factors—including low **GPA**, poor **attendance**, concerning **project/assignment scores**, financial constraints (family income), or weak peer evaluations—that drove the high-risk determination. This transparent breakdown enables educators to pinpoint specific academic weaknesses (such as consistently poor project submissions) while maintaining the flexibility to revisit the input form for reassessment. By visually connecting these risk indicators to actionable insights, the page transforms predictive analytics into practical tools for student support, facilitating targeted interventions like assignment remediation, academic counseling, or additional tutoring to address the identified deficiencies.

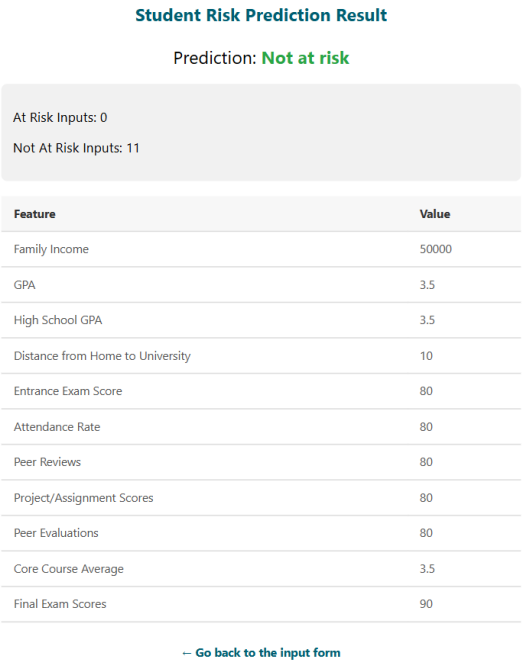

Figure 6: Prediction Result Page (Not At-Risk)

Figure 6 presents the system's assessment outcome when a student is classified as "Not at Risk", displaying the key factors contributing to this positive determination. The dashboard clearly indicates the student's stable academic status through evaluated metrics including satisfactory GPA, strong attendance records, solid exam scores, and favorable socioeconomic indicators (family income). While the prediction probability percentage remains hidden in this view, the comprehensive Feature Inputs table details all contributing elements - from academic performance (high school GPA, entrance exam scores) to behavioral metrics (peer evaluations, project scores) - providing educators with transparent insights into the student's academic health. This intuitive interface not only confirms the student's current stability but also serves as a monitoring tool, with direct navigation options allowing for periodic reassessments or immediate intervention planning if future indicators change. The design emphasizes clarity and actionability, enabling educators to quickly verify positive outcomes while maintaining awareness of all performance dimensions that contributed to the classification.

Figure 7: User Trying the Academically At-Risk System.

Captures a key evaluation phase where users actively engage with the predictive system, demonstrating its real-world application. In this hands-on session, participants explore the interface's full functionality - from data input to result interpretation - guided by the development team who provide clear explanations of each feature. The image shows users entering sample student data (GPA, attendance records, socioeconomic indicators) while observing how the system processes this information to generate risk assessments. This interactive testing approach serves dual purposes: it validates the system's usability through direct feedback while educating stakeholders about the underlying predictive analytics. The proponents facilitate the session by breaking down complex concepts into understandable terms, ensuring participants can accurately evaluate both the technical performance and practical value of the tool. Such user testing sessions proved critical for refining the interface and confirming the system's effectiveness in real educational settings, as evidenced by subsequent survey responses showing high user comprehension and satisfaction rates.

4. RESULT, CONCLUSIONS, AND RECOMMENDATIONS

Table 1:Survey Results on System Effectiveness and User Satisfaction

The evaluation of the predictive analysis system's effectiveness and user satisfaction was conducted through a comprehensive survey of 40 respondents, generating 400 total responses across 10 key performance indicators. Results revealed overwhelmingly positive feedback, with 88% of responses affirming the system's accuracy in identifying at-risk students, usefulness for early intervention, and overall user-friendliness. Only 12% of responses indicated areas needing improvement, primarily related to real-time tracking features. The strong approval ratings demonstrate the system's success in transforming complex predictive analytics into practical tools for educators, with particular strengths in its clear risk visualizations and actionable intervention suggestions. These findings validate the system's readiness for implementation in educational settings, while the small percentage of critical feedback provides valuable insights for future refinements to enhance its impact on student support initiatives.

The survey results demonstrate strong validation of the system's effectiveness, with **88% of users confirming** that the Academically At-Risk Student Prediction model successfully identifies vulnerable students and supports timely interventions. This overwhelming positive response indicates that educators find the tool accurate, actionable, and valuable for proactive student support. The high approval rating reflects the system's ability to translate predictive analytics into practical educational solutions, with users particularly appreciating its clear risk assessments and data-driven intervention guidance. These findings substantiate the platform's real-world utility in enhancing student retention strategies

The study successfully developed and implemented a predictive analysis model that effectively identifies academically at-risk students by integrating both academic and non-academic indicators. Key findings demonstrate that factors like **GPA, attendance rates, family income**, and **peer evaluations** significantly influence student performance, with the model achieving **88% user approval** in real-world testing. The Flask-based web application provides educators with an intuitive, data-driven tool for early intervention, addressing the critical challenge of student attrition. By transforming raw data into actionable insights, this project bridges the gap between predictive analytics and practical educational support, ultimately fostering more inclusive and proactive learning environments.

To enhance the effectiveness of the predictive analysis system, the following measures are recommended: First, educational institutions should conduct training workshops for faculty and staff to ensure proper utilization of the system's features and accurate interpretation of results. Second, schools should establish a monitoring team to regularly review system outputs and coordinate appropriate interventions for identified at-risk students. Third, the system should be integrated with existing student information platforms to enable seamless data flow and real-time updates. Fourth, continuous model refinement is advised through periodic retraining using updated student data to maintain prediction accuracy. Finally, future development should focus on expanding the range of predictive factors to include additional behavioral and engagement metrics, while ensuring robust data privacy protections are maintained throughout the process. These steps will help maximize the system's potential to improve student outcomes while maintaining ethical standards in educational data usage

5. REFERENCES

[1] Ansa, J. (2018). *10 causes of poor academic performance in school – Most students never admit*. EduAnsa. https://www.eduansa.com/10-causes-of-poor-academic-performance-in-school-most-students-never-admit-8/#google_vignette

[2] Chapman, C., Laird, J., & KewalRamani, A. (2010). *Trends in high school dropout and completion rates in the United States: 1972–2008 (NCES 2011-012)*. National Center for Education Statistics. https://files.eric.ed.gov/fulltext/ED513692.pdf

[3] Chen, Y., Zhang, K., & Liu, X. (2019). Identifying at-risk students based on the phased prediction model. *Knowledge and Information Systems, 61*(3), 1277–1297. https://link.springer.com/article/10.1007/s10115-019-01374-x

[4] Eyman, A., Al-Khalifa, H., & Al-Salman, A. (2020). Predicting academic success in higher education: Literature review and best practices. *International Journal of Educational Technology in Higher Education, 17*, Article 52. https://educationaltechnologyjournal.springeropen.com/articles/10.1186/s41239-020-0177-7

[5] Glavin, C. (n.d.-c). The risk factor of high school dropouts. *K12 Academics*. https://www.k12academics.com/High%20School%20Dropouts/risk-factor-high-school-dropouts

[6] Jayaprakash, S., Sharma, P., & Gupta, R. (2020). Predicting students academic performance using an improved random forest classifier. *2020 IEEE International Conference on Teaching, Assessment, and Learning for Engineering (TALE)*. https://ieeexplore.ieee.org/abstract/document/9167547

[7] Laoshi. (2024). *Australian student performance dataset*. Kaggle. https://www.kaggle.com/datasets/nasirayub2/australian-student-performancedata-aspd24/data?select=Australian_Student_PerformanceData+%28ASPD24%29.csv

[8] Lim, J. (2023). Exploring the relationships between interaction measures and learning outcomes through social network analysis: The mediating role of social presence. *International Journal of Educational Technology in Higher Education, 20*(1), Article 27. https://educationaltechnologyjournal.springeropen.com/articles/10.1186/s41239-023-00384-8

[9] Murain, I. O., & Olatunji, O. O. (n.d.). Decision tree algorithm use in predicting students' academic performance in advanced programming course. *International Journal of Higher Education and Pedagogy*. https://www.diamondopen.com/journals/index.php/ijhep/article/download/274/142

[10] Ng, K., Hoo, M.-H., Nair, M., & Khor, K.-C. (2023). Predicting student performance in final year project using data mining classification techniques. *ResearchGate*. https://www.researchgate.net/publication/370961886_Predicting_student_performance_in_final_year_project_using_data_mining_classification_techniques

[11] Qushem, U. B., Khan, H. U., & AlGhamdi, J. (2023). Unleashing the power of predictive analytics to identify at-risk students in computer science. *Technology, Knowledge and Learning*. https://link.springer.com/article/10.1007/s10758-023-09674-6

[12] Sarao, Z. (2023, October 11). Dropout rate in universities, colleges at 35.15% in SY 2023-2024, says CHEd. *INQUIRER.net*. https://newsinfo.inquirer.net/1839954/dropout-rate-in-universities-colleges-at-35-15-in-sy-2023-2024-says-ched

[13] Srinivas, K., Raghunathan, R., & Parthiban, L. (2018). Predicting the academic performance of middle- and high-school students using machine learning algorithms. *Education and Information Technologies, 27*, 11355–11378. https://www.sciencedirect.com/science/article/pii/S2590291122001115

[14] U.S. Department of Education, National Center for Education Statistics. (2010). *American Community Survey (ACS) 2010*. https://nces.ed.gov/fastfacts/display.asp?id=16

[15] Whitcomb, K., Robbins, M. W., & Flaster, A. (2021). Not all disadvantages are equal: Racial/ethnic minority students have largest disadvantage among demographic groups in both STEM and non-STEM GPA. *AERA Open, 7*, 23328584211059823. https://journals.sagepub.com/doi/full/10.1177/23328584211059823