IIAI Open Conference Publication Series
IIAI Letters on Institutional Research
Vol. 005, LIR473
DOI: https://doi.org/10.52731/lir.v005.473

Predicting Performance in First-Year Required Courses Using Machine Learning: An Analysis of Students' Learning Outcomes Based on At-Enrollment Data

Shintaro Tajiri [*], Kunihiko Takamatsu [†], Naruhiko Shiratori [‡], Tetsuya Oishi [§], Masao Mori [†], Masao Murota [†]

Abstract

In response to the growing importance of data literacy across disciplines, this study explores the potential of machine learning to predict student performance in first-year information literacy courses using only at-enrollment data. Conducted at Hokuriku University in Japan, the study utilizes a rich dataset encompassing students' academic background, standardized test scores, cognitive skills assessments, and self-reported academic habits collected at the time of admission. This research uses Random Forest, Support Vector Machine, and Logistic Regression models to identify at-risk students early in the academic year. Our findings reveal that Random Forest achieved the highest accuracy in binary classification with an AUC score of 0.878, highlighting key predictors such as English proficiency, high school GPA, and conceptual skills. This predictive approach demonstrates the feasibility of early intervention for at-risk students, offering insights into student preparedness and support enhancement. By identifying critical factors influencing success in mandatory data science education, this study contributes to the global dialogue on improving foundational data science courses and proposes scalable methods to foster equitable academic outcomes.

Keywords: Data science education, student performance prediction, machine learning, academic analytics, early intervention

1 Introduction

1.1 The New Information Literacy Course

In today's digital society, universities must equip first-year students with essential digital literacy skills for academic and career success, fostering adaptability to fast-evolving technologies. The Japanese government has recognized this imperative, establishing its first national AI strategy in 2019. This initiative set an ambitious goal to provide basic literacy in mathematics, data science, and AI to approximately 600,000 students annually – equivalent to the country's annual university enrollment. Through the Approved Program for Mathematics, Data Science, and AI Smart Higher Education (MDASH), the strategy aims to ensure that all students, regardless of their field of study, acquire essential skills in statistical concepts, data-driven thinking, problem-solving, information technology literacy, and data ethics. This research extends previous work by

[*] Hokuriku University, Ishikawa, Japan
[†] Institute of Science Tokyo, Japan
[‡] Tokyo City University, Tokyo, Japan
[§] Kyushu Institute of Technology, Fukuoka, Japan

analyzing quantitative outcomes to assess collaborative teaching effectiveness in data science. Furthermore, the rapid evolution of technology necessitates educational approaches that can readily adapt to changing industry standards and tools.

In response to these challenges, Hokuriku University launched an innovative data science education program in 2022, integrating Tableau, a professional business intelligence tool, into a mandatory first-year information literacy course [1]. This program represents a novel approach to modern data science education by combining traditional information literacy topics with hands-on data visualization experience using real-world datasets from campus facilities. The implementation provides valuable insights into effective methods for incorporating data science education across multiple academic disciplines while maintaining student engagement through practical, real-world applications.

This approach demonstrates universities' potential to prepare students for digital society demands. Through the integration of professional tools and real-world data, the program bridges the gap between theoretical knowledge and practical application, while simultaneously addressing the growing demand for data literacy across all fields. The experience of Hokuriku University offers important lessons for other institutions seeking to enhance their data science education programs in response to the accelerating pace of technological innovation.

1.2 Grade Distributions

Following these educational imperatives, Tajiri et al. (2024) study presents a comparative analysis of grade distributions in team-taught introductory data science courses for first-year students [2]. This research represents a significant extension of our previous work, moving beyond the initial implementation phase to examine the quantitative outcomes of our educational approach. Through detailed analysis of student performance across multiple classes and teaching teams, we examine the effectiveness of collaborative teaching methods in data science education.

The comparative analysis of grade distributions provides crucial insights into how different teaching approaches and team compositions affect student learning outcomes. This research is particularly relevant given the increasing adoption of team-teaching models in data science education, where instructors with complementary expertise collaborate to provide comprehensive coverage of both theoretical foundations and practical applications. Our analysis focuses specifically on first-year courses, as this represents a critical period for establishing fundamental data science competencies and shaping students' attitudes toward the field.

This investigation contributes to the global efforts to standardize and improve data science education. As more institutions implement similar programs, understanding the relationship between teaching methodologies and student outcomes becomes increasingly important for developing evidence-based best practices in data science education. By examining grade distributions across different teaching teams and student cohorts, we identify successful pedagogical strategies that can be replicated across various educational contexts.

However, the successful implementation of data science education programs faces a critical challenge: student failure rates can significantly undermine the effectiveness of these educational initiatives. When students fail to complete introductory data science courses, it not only disrupts their academic progression but also potentially affects their attitude toward the entire field of data science. This is particularly concerning in mandatory first-year courses, where early academic setbacks can have long-lasting implications for students' educational trajectories. The presence

of failing grades suggests that despite the careful design of curricula and the integration of modern tools like Tableau, some students struggle to meet the course requirements. Understanding and addressing the factors that contribute to student failure is therefore crucial for maximizing the impact of data science education programs and ensuring that these innovative educational approaches truly benefit all students.

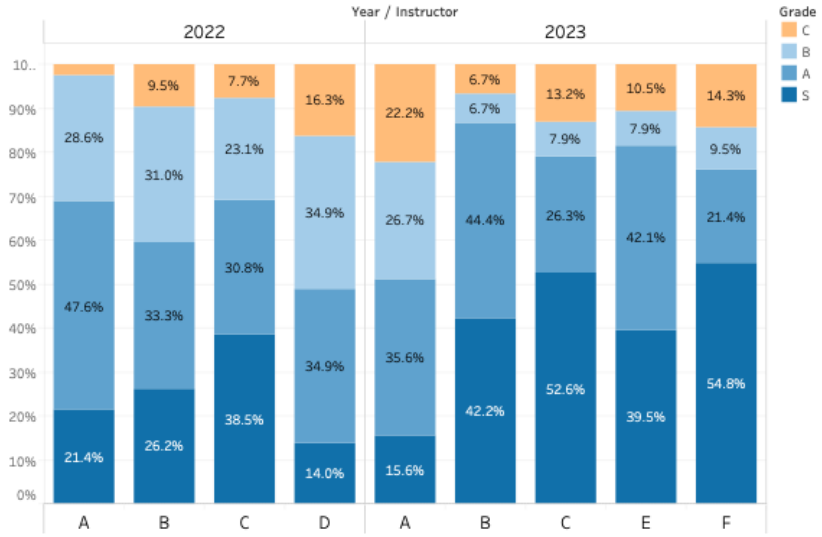

Figure 1: Class-Specific Grading Result from [2]

The study examines pass and fail rates in the Information Literacy course at the Faculty of Economics and Business Administration from 2021 to 2023, as shown in Figure 2. The course structure is divided into two semesters: the spring semester is offered exclusively to first-year students, while the fall semester serves as a remedial class for both students who failed in the previous spring and upper-class students who had not yet passed the course from earlier years.

Figure 2 illustrates the variability in pass rates for the Information Literacy course. Spring pass rates ranged from 75.91% to 87.31% between 2021 and 2023 (noting that the Japanese academic year begins in spring), while fall pass rates for retakers were consistently lower, reaching as low as 67.50% in 2023. This gap suggests that students retaking the course face more challenges than first-time takers.

These findings from Figure 2 suggest that while the majority of first-year students are able to successfully complete the course in their initial attempt during the spring semester, there is a persistent challenge in helping students who fail their first attempt to succeed in subsequent attempts. This disparity between spring and fall semester performance indicates a need for targeted interventions and potentially different pedagogical approaches for students retaking the course.

Figure 3 shows a bimodal distribution, with scores clustered below 10 and above 60 points, highlighting a divide between students who excel and those who struggle. Over the years, scores show improvement, with fewer students scoring below 10 and more achieving high marks.The 2022 distribution maintained a similar bimodal pattern but showed more even distribution among passing scores between 60 and 90 points. The number of severely struggling students remained

high, with 21 students scoring in the 0-5 range. The peak frequencies among passing students were lower than in 2021, with the highest numbers occurring in the 70-75 range (33 students), still well above the passing threshold.

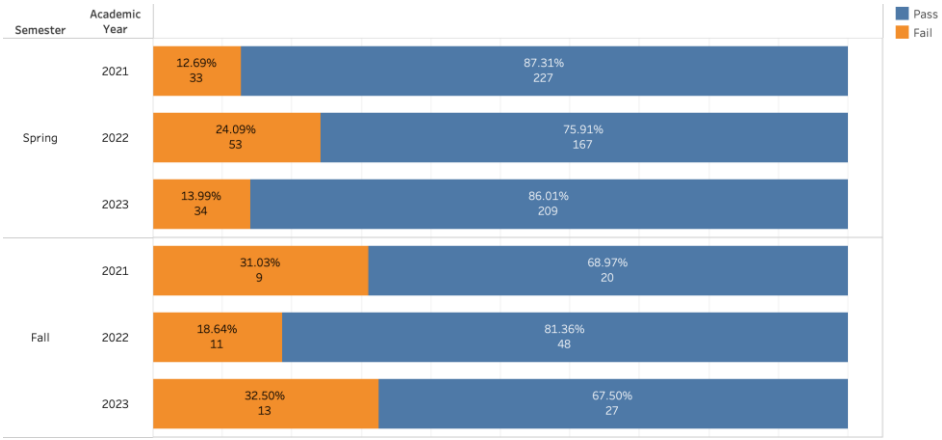

Figure 2: Pass and Fail Rates in Information Literacy Course (2021-2023).

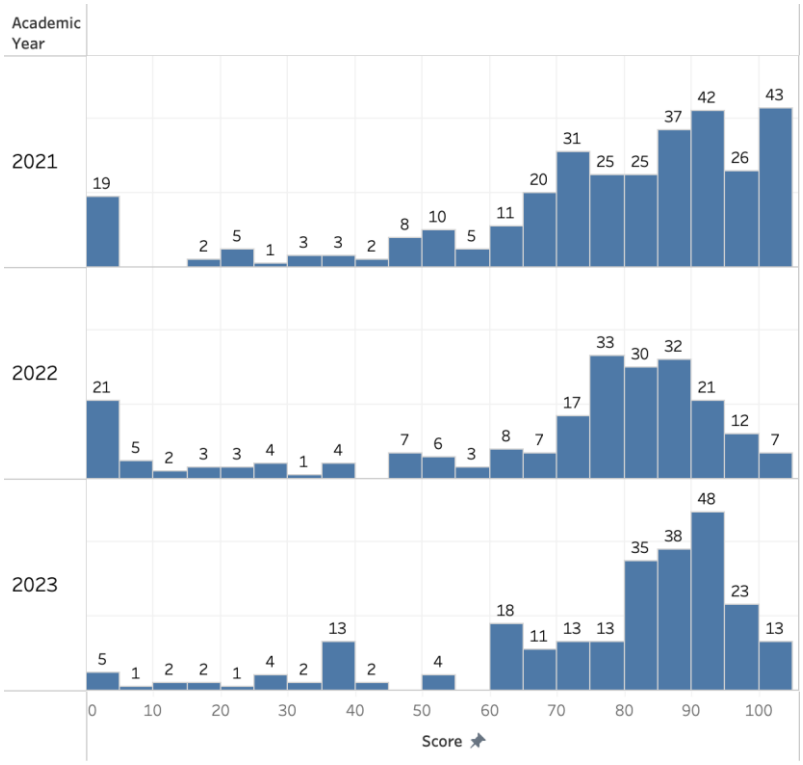

Figure 3: Distribution of Final Scores in Information Literacy Course During Spring Semesters (2021-2023)

The 2023 data shows encouraging improvements, particularly in reducing the number of severely failing students. Only 5 students scored in the 0-5 range, compared to 19 and 21 in previous years. Among passing students, there was a strong concentration in the 85-90 range (48 students),

suggesting that successful students were not merely passing but achieving high scores. This improvement in both reducing severe failures and maintaining high achievement among passing students might indicate the effectiveness of new teaching strategies or support systems implemented in 2023.

This detailed score distribution complements the pass/fail rates presented in Figure 2 by revealing not just whether students passed or failed, but the degree of their success or failure. The persistent gap between high achievers and severe underperformers suggests that while the course effectively serves engaged students, there remains a critical need for early identification and intervention strategies for at-risk students before they disengage completely from the course.

Given these findings, minimizing the failure rate in Information Literacy, a mandatory first-year course, emerges as a critical educational priority. The stark contrast in performance patterns shown in Figures 2 and 3 underscores the importance of early intervention. This leads us to our primary research question: Can we identify students at risk of failing the Information Literacy course using data available at the time of enrollment? If such identification is possible, appropriate support mechanisms must be implemented to prevent academic failure and ensure these students receive the foundation in data science education they need for their future academic success.

This study is structured as follows. Section 2 (Method) reviews existing research on academic performance prediction and describes our dataset, which includes various student characteristics and academic background information available at enrollment. We examine how previous studies have approached similar predictive challenges in mandatory first-year courses and detail the specific variables and analytical approaches employed in our investigation. Section 3 (Results) presents the outcomes of our machine learning models, evaluating their effectiveness in predicting course failure risk using only at-enrollment data. Finally, Section 4 (Discussion) addresses the remaining challenges in implementing these predictive insights into practical support systems and considers the broader implications for data science education in higher education.

Through this investigation, we aim not only to develop a predictive model for identifying at-risk students but also to contribute to the broader discourse on how universities can better support student success in foundational data science education. This work represents a crucial step toward ensuring that the benefits of modern data science education reach all students, regardless of their initial preparedness or background.

2 Methods

2.1 Prior Research

Research on predicting student academic performance in higher education has a long and extensive history. In a significant study, Pike and Saupe (2002) analyzed records of 8,764 freshmen at a research university, comparing three prediction models: traditional regression, high-school-effects, and hierarchical linear models [3]. They found that incorporating high school characteristics improved prediction accuracy, explaining about 40% of the variance in first-year grades. However, their study was limited by its focus on a single institution and only included in-state students from high schools with substantial enrollment numbers, potentially restricting the generalizability of their findings.

Peña-Ayala (2014) conducted a comprehensive survey and data mining-based analysis of educational data mining (EDM) research between 2010 and 2013 [4]. Through analysis of 240

EDM works, the study discovered that most EDM approaches rely on three basic elements: educational systems (primarily intelligent tutoring systems), disciplines (mainly probability and machine learning), and algorithms (notably Bayesian methods and decision trees). The analysis revealed two distinct patterns in EDM approaches - descriptive models focused on clustering and pattern discovery, and predictive models aimed at student performance classification. However, the study noted limitations in that it only covered publications from a specific time period and primarily examined works from journals and conferences, potentially missing relevant research from other sources.

Costa et al. (2017) conducted a comprehensive study evaluating the effectiveness of educational data mining techniques for early prediction of student failure in introductory programming courses [5]. Their analysis compared four machine learning methods (Neural Networks, Decision Trees, Support Vector Machines, and Naive Bayes) using data from both distance and on-campus courses at a Brazilian university. The study demonstrated that Support Vector Machines achieved the highest prediction accuracy, identifying at-risk students with 92% accuracy in distance education and 83% in on-campus courses when students had completed 50% and 25% of their respective courses. However, their study was limited by focusing on data from a single university and relied primarily on manual fine-tuning of some algorithms, which could affect reproducibility.

Research on student performance prediction has significantly evolved with the emergence of Educational Data Mining (EDM). Albreiki et al. (2021) conducted a systematic literature review of machine learning techniques used for predicting student performance between 2009 and 2021 [6]. Their analysis revealed that various machine learning methods effectively predicted student performance and dropout rates using both institutional databases and online learning platform data. The study highlighted the increasing importance of early intervention strategies based on machine learning predictions. However, their review was limited by focusing only on published academic literature, excluding grey literature that might contain valuable practical implementations, and primarily examined historical data rather than real-time prediction approaches.

Sandra et al. (2021) conducted a systematic literature review of machine learning algorithms used for predicting student performance, analyzing 11 research articles selected from 2,753 papers published between 2019-2021 [7]. Their analysis revealed that classification algorithms were most commonly used, with Artificial Neural Networks (ANN), Naïve Bayes, Logistic Regression, Support Vector Machine (SVM), and Decision Trees being the most prevalent techniques. The study found these algorithms were primarily used to classify students into binary (pass/fail) or tertiary (fail/pass/excellent) categories. The research highlighted that machine learning approaches were effective for early identification of at-risk students and supporting instructional decision-making. However, the study was limited by its narrow time frame and reliance on only two databases (IEEE Access and Science Direct).

2.2 Data Collection and Features

The main research question is whether students' pass/fail outcomes in a mandatory first-year Information Literacy course can be predicted, using data available at the time of enrollment. This prediction capability would enable early intervention strategies for potentially struggling students before they encounter academic difficulties. While previous studies such as Pike and Saupe (2002) and Costa et al. (2017) focused primarily on academic performance data and high school characteristics, and recent reviews by Albreiki et al. (2021) and Sandra et al. (2021) highlighted

the effectiveness of various machine learning approaches, our study uniquely incorporates comprehensive at-enrollment assessment data including standardized generic skills measurements through the PROG Test, which is distinctive to the Japanese higher education context.

The data collected for this study, as shown in Table 1, encompasses five major categories. The Basic Student Information was extracted from the university's academic information system database, which serves as the primary repository for student records and academic administration. This data includes fundamental information such as academic year of admission, gender, high school rank, high school GPA, and entrance examination type. These variables have been consistently identified in previous studies as significant predictors of academic performance and are particularly valuable due to their objective and standardized nature.

For Information Literacy course-related data, we collected three key metrics: grade scores (0-100 points), grade points, and first-week typing test scores. The typing test is a standardized five-minute assessment conducted during each class session, with particular attention paid to the first week's performance as an early indicator of basic computer literacy skills. This weekly typing assessment provides a consistent measure of students' progress throughout the course, but our study specifically focuses on the first week's results as they represent baseline capabilities at course commencement.

Table 1: Data Collection

Category	Feature Measures
Basic Student Information	Academic year of admission, Gender, High school rank, High school GPA, Entrance examination type
Information Literacy Course-related data	Grade score (0-100 points), Grade point, First week typing test score
Enrollment Student Survey	Achievement level of university admission policy, Number of universities applied to, Preference ranking for Hokuriku University, High school activities, Level of knowledge and skills at time of admission, Grit Score
Placement Test	Japanese, Mathematics, English
PROG Test	Literacy level consisting of problem-solving ability, verbal processing ability and non-verbal processing ability, 33 types of competency scores

The Enrollment Student Survey, administered to all new students during the orientation week before classes began, provides crucial self-reported data that complements the objective measures from other sources. This comprehensive survey collected various data points including students' self-assessment of their alignment with university admission policy, their application history (number of universities applied to and preference ranking for Hokuriku University), and their self-evaluation of high school activities and current knowledge/skill levels. The survey also includes the innovative Grit Score measurement, which assesses students' perseverance and passion for long-term goals.

The Placement Test scores in Japanese, Mathematics, and English represent standardized assessments of academic proficiency at enrollment. While all three subjects are tested, the English scores are particularly significant as they are used for class placement purposes. These placement tests provide objective measures of academic preparedness across key subject areas and serve as important baseline indicators of academic capability.

The PROG Test represents a unique aspect of our data collection that distinguishes this study from previous research in the field. This widely-used assessment in Japanese higher education measures generic skills through two distinct dimensions: Literacy and Competency. The Literacy component evaluates problem-solving skills in verbal and non-verbal processing. The Competency component assesses students' ability to build positive relationships and adapt to their environment, comprising three fundamental areas: interpersonal skills, problem-solving abilities, and self-management capabilities. Both components use sophisticated scoring systems (1-7 for comprehensive assessment, 1-5 for sub-items), providing a multifaceted view of students' capabilities beyond traditional academic metrics.

Unlike previous studies that primarily relied on academic performance data or online learning behavior, our research incorporates a uniquely comprehensive dataset that combines traditional academic indicators with standardized assessments of generic skills, detailed self-reported measures, and specific computer literacy indicators. This rich combination of data sources allows for a more nuanced understanding of the factors that might influence success in information literacy education, particularly in the Japanese higher education context.

2.3 Data Analysis Methods

Based on recent systematic literature reviews by Albreiki et al. (2021) and Sandra et al. (2021), which identified the most effective machine learning algorithms for student performance prediction, we selected three widely-used machine learning approaches: Random Forest, Logistic Regression, and Support Vector Machine (SVM). These algorithms were particularly chosen because of their demonstrated effectiveness in educational contexts, as shown by Costa et al. (2017) who achieved 92% accuracy in predicting student performance using SVM, and their ability to handle both classification and regression tasks.

Our analysis focused on students enrolled in the Information Literacy course in the Faculty of Economics and Business Administration during the spring semesters of 2021, 2022, and 2023. The total sample comprised 716 students across these three academic years. This sample size is substantial compared to similar studies in the field and provides sufficient data for robust machine learning analysis.

We established two prediction targets: a regression task predicting the numerical course score (0-100 points) and a classification task predicting course pass/fail status. For the classification task, students scoring below 60 points were categorized as failing the course, consistent with the university's grading policy. This dual approach allows us to not only identify students at risk of failing but also to predict their likely performance level, providing more nuanced information for potential interventions.

Random Forest was chosen for its handling of diverse variables and non-linear relationships. Logistic Regression was chosen for its interpretability and proven track record in binary classification tasks, particularly in educational contexts as highlighted in previous studies. SVM was

included based on its strong performance in similar contexts, notably in Costa et al.'s (2017) study where it achieved high accuracy rates in predicting student performance in programming courses.

All analyses were conducted using Dataiku, a comprehensive data science and machine learning platform that facilitates end-to-end analytics workflows. Dataiku is an enterprise-grade platform that combines data preparation, machine learning, and deployment capabilities in a collaborative environment. The platform's visual interface and automated machine learning capabilities make it particularly suitable for educational data analysis, allowing for efficient model development and evaluation while maintaining robust analytical standards.

3 Results

We first attempted to predict students' final numerical scores (0-100 points) in the Information Literacy course using at-enrollment data. Multiple machine learning algorithms were evaluated using Dataiku's automated machine learning capabilities, with each model being trained on 792 instances. The performance of each model was assessed using the R^2 score, which indicates the proportion of variance in the dependent variable that is predictable from the independent variables.

The results showed that LightGBM achieved the highest performance with an R^2 score of 0.285 (±0.045), followed closely by Random Forest with 0.284 (±0.050). Ridge (L2) regression yielded an R^2 score of 0.149 (±0.070), while Ordinary Least Squares produced 0.080 (±0.150). The Decision Tree and SVM models performed poorly, with negative R^2 scores of -0.204 (±0.351) and -0.059 (±0.052) respectively. Key predictors included high school GPA rank and PROG Test scores in problem-solving and presentation skills.

However, the overall predictive power of these models was relatively weak, with even the best-performing model (LightGBM) explaining only 28.5% of the variance in final course scores. These results suggest that predicting precise numerical grades in Information Literacy based solely on enrollment data presents significant challenges. This suggests that course performance may depend on factors emerging during the semester, such as study habits and course engagement, which enrollment data alone cannot capture.

Given the limited success in predicting numerical scores, we shifted our focus to the binary classification of course outcomes (pass/fail). This simpler prediction task yielded substantially better results, with models evaluated using the ROC AUC score. The Random Forest classifier achieved the highest performance with an AUC score of 0.878, followed by SVM (0.799) and Logistic Regression (0.773).

To understand the factors driving these predictions, we conducted a detailed feature importance analysis using Shapley values, which provide a robust method for estimating each feature's contribution to the model's predictions. As shown in Figure 4, the absolute feature importance analysis revealed that the top 20 features accounted for 39.1% of the total feature importance, with English Score (5%), High_School_GPA x High School Rank (4%), and ES1G_Prepared_Reviewed_Homework (3%) emerging as the most influential predictors.

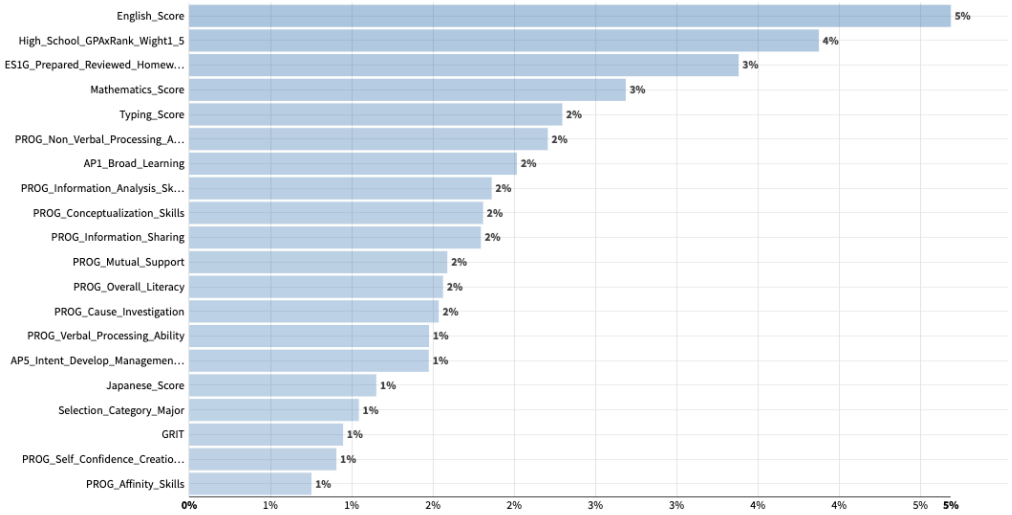

Figure 4: Absolute Feature Importance

The feature effects analysis, visualized in Figure 5, provides deeper insights into how specific feature values influence predictions. The Shapley values indicate each feature's relative impact on predicting course failure, where positive values suggest a higher likelihood of failure. Notably, English Score emerged as the most impactful feature (impact: 0.14, correlation: -0.74), with lower scores strongly associated with increased likelihood of failure. The Typing Score showed the fifth-highest impact (impact: 0.069, correlation: 0.71), where higher scores correlated with increased likelihood of failure. PROG_Conceptualization_Skills demonstrated the ninth-highest impact (impact: 0.054, correlation: -0.72), with lower scores associated with higher failure probability.

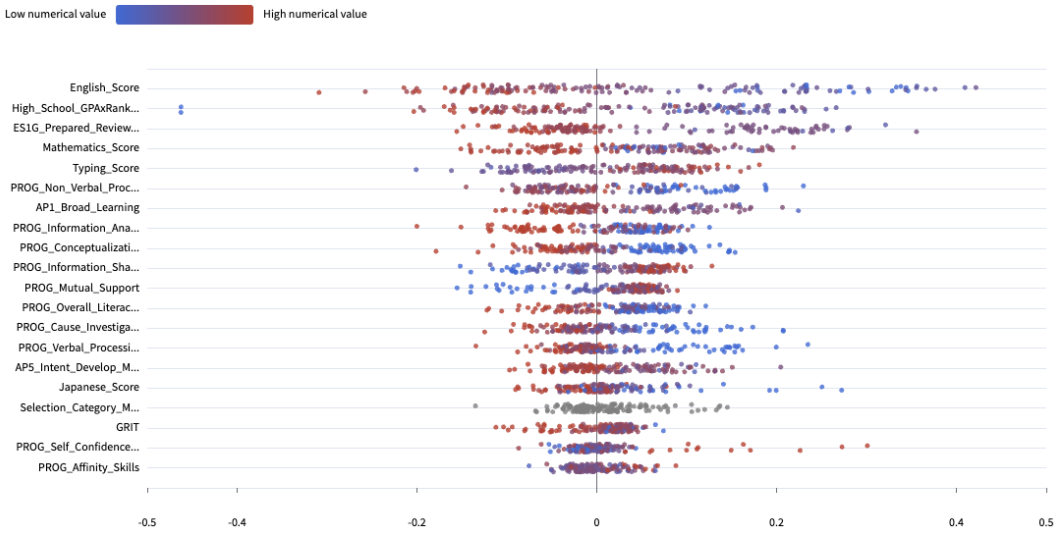

Figure 5: Feature Effects

These findings suggest that while predicting exact numerical scores proved challenging, identifying students at risk of failing the Information Literacy course using at-enrollment data is

feasible with reasonable accuracy. The analysis reveals that a combination of traditional academic indicators (English scores, high school GPA), basic computer skills (typing scores), and general cognitive abilities (PROG test components) can effectively predict course outcomes. This suggests that early intervention strategies could be targeted based on these at-enrollment indicators, particularly focusing on students with lower English proficiency and conceptual skills, even before the course begins. The relatively high AUC scores also indicate that this binary classification approach could serve as a practical tool for identifying at-risk students, enabling proactive academic support measures.

4 Discussion

This study demonstrates both the potential and limitations of using at-enrollment data to predict student performance in mandatory first-year information literacy courses. While precise grade prediction proved challenging, with regression models achieving modest R^2 scores no higher than 0.285, the binary classification approach showed remarkable promise in identifying students at risk of course failure. The Random Forest classifier's achievement of a 0.878 AUC score suggests that institutions can effectively identify at-risk students before classes begin, enabling proactive intervention strategies. The emergence of English proficiency scores and high school academic performance as key predictors indicates that success in modern data science education relies not only on technical aptitude but also on fundamental academic capabilities.

A crucial finding is the role of the PROG test. While traditional metrics like English proficiency and high school GPA were the strongest individual predictors, the test's collective contribution was significant, with 10 of the top 20 predictors being its components. This indicates the test's value lies not in a single metric, but in providing a multifaceted view of student preparedness. Its assessment of generic skills complements traditional academic data and improves overall predictive accuracy.

These findings inform comprehensive support strategies beyond prediction alone. The strong predictive power of pre-existing study habits and academic engagement patterns, combined with the significance of language proficiency, suggests that institutions should consider implementing integrated support systems that address both technical and foundational academic skills. However, our findings also reveal that approximately 70% of the variance in student performance remains unexplained by at-enrollment factors alone, highlighting the crucial role of in-semester experiences and engagement in determining academic outcomes.

Our research indicates that the path to improving success rates in mandatory data science education requires a nuanced understanding of student preparedness. The bimodal distribution of course scores, with distinct peaks at both failing and high-performing levels, suggests that current educational approaches may not adequately address the diverse needs of all students. The significant correlation between English proficiency and course success particularly emphasizes the need to consider language support as an integral component of data science education, especially as tools and interfaces become increasingly text dependent.

Future challenges lie in developing responsive support mechanisms for at-risk students. While our study focused on a single Japanese university during a three-year period marked by global educational disruptions, the findings suggest broader implications for the design and delivery of data science education worldwide. Future research should explore the integration of real-time performance data with at-enrollment predictors, evaluate the effectiveness of various intervention strategies, and examine the long-term impact of early information literacy success on students'

academic and professional development. Through such continued investigation and refinement of support systems, institutions can work toward ensuring that the benefits of data science education reach all students, regardless of their initial preparedness.

Acknowledgment

This work was supported by Hokuriku University Special Research Grant 2023 in Fundamental Research Area and JSPS KAKENHI Grant Number JP23K02558.

References

[1] S. Tajiri, K. Takamatsu, N. Shiratori, T. Oishi, M. Mori, and M. Murota, "Integrating Tableau into a First-Year Information Literacy Course: A Practical Approach to Enhancing Data Science Education," in *16th International Conference on Data Science and Institutional Research (DSIR 2024)*, 2024, p. in press.

[2] S. Tajiri, K. Takamatsu, N. Shiratori, T. Oishi, M. Mori, and M. Murota, "Comparative Analysis of Grade Distributions in Team- Taught Introductory Data Science Courses for First-Year Students." .

[3] G. R. Pike and J. L. Saupe, "Does High School Matter? An Analysis of Three Methods of Predicting First-Year Grades," *Res. High. Educ.*, vol. 43, no. 2, pp. 187–207, 2002.

[4] A. Peña-Ayala, "Educational data mining: A survey and a data mining-based analysis of recent works," *Expert Syst. Appl.*, vol. 41, no. 4, pp. 1432–1462, Mar. 2014.

[5] E. B. Costa, B. Fonseca, M. A. Santana, F. F. de Araújo, and J. Rego, "Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses," *Comput. Human Behav.*, vol. 73, pp. 247–256, Aug. 2017.

[6] B. Albreiki, N. Zaki, and H. Alashwal, "A systematic literature review of student' performance prediction using Machine Learning techniques," *Educ. Sci. (Basel)*, vol. 11, no. 9, p. 552, Sep. 2021.

[7] L. Sandra, F. Lumbangaol, and T. Matsuo, "Machine learning algorithm to predict student's performance: A systematic literature review," *TEM J.*, pp. 1919–1927, Nov. 2021.