Interpretable machine learning for academic risk analysis in university students

Mukti Ratna Dewi, Mochammad Reza Habibi, Bassam Babgei, Lovinki Fitra Ananda, Brodjol Sutijo Suprih Ulama
Department of Business Statistics, Faculty of Vocational Studies, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia

Article Info

Article history:

Received Nov 12, 2024
Revised Jun 11, 2025
Accepted Jul 10, 2025

Keywords:

Academic risk
Interpretable machine learning
LightGBM
Shapley additive explanations
SMOTE

ABSTRACT

Higher education institutions often grapple with issues related to academic risk among their students. These academic risks encompass low academic performance, study delays, and dropouts. One approach to address these challenges is to predict students' academic performance as accurately as possible by leveraging advanced computational techniques and utilizing academic and non-academic student data. This research aims to develop a model that accurately identifies students with high potential for academic risk while explaining the contributing factors to this phenomenon in the Faculty of Vocational Studies, Institut Teknologi Sepuluh Nopember (ITS). The prediction model is constructed using the light gradient boosting machine (LightGBM) method and is subsequently interpreted using the Shapley additive explanations (SHAP) value. Additionally, an oversampling method, based on synthetic minority oversampling technique (SMOTE), is implemented to address imbalances in the dataset. The proposed approach achieves 96% and 97% accuracy and specificity rates, respectively. Analysis based on SHAP values reveals that extracurricular activities, choice of major, smoking habit, gender, and friendship circle are among the top five factors impacting students' academic risk.

Corresponding Author:

Mukti Ratna Dewi
Department of Business Statistics, Faculty of Vocational Studies, Institut Teknologi Sepuluh Nopember
Teknik Kimia Street, Keputih, Sukolilo, Surabaya 60111, Indonesia
Email: mukti_ratna@its.ac.id

1. INTRODUCTION

The academic risk of a university student refers to the risk of experiencing delays in one's educational journey or postponing or failing to obtain a degree, ultimately resulting in an incomplete degree [1]. Delays in studies may occur when students need additional time to complete coursework due to factors like repeating courses, changing majors, or taking semesters off for personal reasons. Besides, dropping out of university can stem from many factors [2], [3], including academic difficulties, financial constraints, or a shift in personal priorities. Those issues can significantly impact students' educational and future career prospects. From an institutional perspective, student dropout and delayed completion potentially damage the university's image, thus negatively impacting accreditation assessments. This situation emphasizes the importance of identifying and mitigating academic risk early to provide students with the support and resources they need to succeed and achieve their educational goals. Addressing these risks benefits individual students and contributes to educational institutions' effectiveness and success.

Monitoring student performance is a challenging task for several reasons, such as the difficulty of identifying students at academic risk [4], limited access to certain parts of the curriculum [5], and the challenge of obtaining information related to teaching methods, interventions, and academic support from the educational institution in question. One approach that can be used is utilizing academic and non-academic information from students, such as high school academic track records, performance during university studies, study habits, lifestyle behaviors, social and environmental conditions, to analyze student academic performance. This approach is based on research initiated by [6], which asserts that students' backgrounds and fundamental abilities in their academic journey can be used to predict their academic performance. Furthermore, Yang [1] suggests that factors outside of academics also influence student performance.

In recent years, research on interpretable machine learning techniques within the educational sector has increased rapidly [1], [7]–[12]. It cannot only predict student performance but also reveal the key factors influencing individual predictions. In this study, we construct a predictive model using the light gradient boosting machine (LightGBM) method, which offers high accuracy and faster, more efficient computational times [13]. This method's prediction results are then expounded and interpreted using Shapley additive explanations (SHAP) [14]. These interpretation results will serve as the foundation for stakeholders to formulate strategies to improve the performance of students at risk.

The primary objectives of this work can be summarized as follows: i) using LightGBM to predict the student's academic performance (no-risk or at-risk) and ii) utilizing SHAP to identify key factors affecting academic performance. In educational institutions, relevant data generally have the characteristics of high dimensions, small samples, and imbalanced classification. Recent studies suggest that class imbalance negatively impacts the interpretability of machine learning models [15]–[19]. Thus, we first perform special treatment to address data imbalance before analyzing and evaluating performance using synthetic minority oversampling technique (SMOTE).

2. METHOD

2.1. Background on Indonesia higher education and vocational programs in ITS

This study is based on data from undergraduate students at the Faculty of Vocational Studies in Institut Teknologi Sepuluh Nopember (ITS). We first present briefly the curriculum of the Faculty of Vocational Studies in ITS and Indonesia's higher education system in general to make the paper easier to follow. Generally, students in Indonesia undergo a 12-year educational journey from elementary school to senior high school. Those who wish to pursue higher education at public universities can choose one of three main pathways, i.e., joint national selection of state universities (SNMPTN), computer-based writing examination of test-based national selection (UTBK-SNBT), and college entrance examination organized by each state university (MANDIRI). All graduates of senior high schools are allowed to participate in each selection. SNMPTN involves inviting eligible students who have demonstrated: i) academic excellence at their respective schools, ii) have been nominated by their schools, and iii) fall within the allocated quota as determined by the university. The other selection pathways, UTBK and MANDIRI, consist of at least three subjects being examined: scholastic aptitude, English, and academic ability, which vary by discipline [20].

ITS is one of the higher education institutions in Indonesia that applies the same student admission selection process. There are two kinds of programs in ITS: vocational and non-vocational programs that differ in academic regulations. The academic regulations governing vocational programs diverge from those applicable to non-vocational programs, resulting in distinct academic risks between these program types. The most significant distinction arises from implementing a block program and the mandatory leave of absence (not going up semester or TNS) policy as stated in ITS Chancellor Regulation Number 13 of 2019 concerning academic regulations for ITS Vocational Program.

In detail, course evaluation for students in the Faculty of Vocational Studies in ITS is regulated in article 10 paragraph (11). Article 10 paragraph (12) states that students are declared to have passed a course if they have a minimum grade of C. If a student has a grade of D or E, the student is declared TNS and must take TNS leave for one semester. The implications of this TNS policy will certainly impact overall academic performance, such as study delays and failure. Study delays if the applied undergraduate study period is more than eight semesters, while students are declared to have failed or failed their studies if the study period is more than 14 semesters (article 14 paragraph 6). Based on this description, a student is said to be at academic risk if they have one or more of the following criteria: i) currently or formerly had TNS status; ii) have a GPA ≤ 2.75; and iii) take more than 8 semesters.

2.2. Data collection and data description

This study collected data by survey, with the targeted respondents being all students in the Faculty of Vocational Studies, ITS, who enrolled between 2017 and 2021 and are still active in the 2022/2023 academic year. There are six departments in the Faculty of Vocational Studies, ITS: i) Civil Infrastructure Engineering, ii) Industrial Mechanical Engineering, iii) Automation Electrical Engineering, iv) Industrial Chemical Engineering, v) Instrumentation Engineering, and vi) Business Statistics. However, we only used four out of six departments for analysis due to insufficient respondents from the other two departments (Department of Electrical Automation Engineering and Department of Instrumentation Engineering).

After preprocessing data, a dataset containing 46 variables and 539 effective individual behaviors and group behaviors records was obtained. We also include information about students' academic interaction, learning behaviors, and social interaction. The dependent variable is a binary variable with or without academic risk (0: no-risk, 1: at-risk), whereas the independent variables are summarized in Table 1.

Table 1. Independent variables

Feature category	Variable name	Description	Type
Demography	X1	Sex	Binary
	X2	Department	Nominal
	X3	Hometown	Binary
	X4	Residence	Nominal
High school and enrollment information	X5	TEFL score	Real
	X6	Type of high school	Nominal
	X7	Academic achievements in high school	Binary
	X8	Non-academic achievements in high school	Binary
	X9	Gap year	Binary
Academic info	X10	Admission type	Nominal
	X11	Awardee of education scholarship	Binary
	X12	Tuition category	Ordinal
	X13	Enrolment age	Real
	X14	Intention to study at the current department	Binary
	X15	Interest in getting accepted into the current department	Likert
	X16	Satisfaction with major choices	Likert
	X17	Academic leave	Binary
	X18	Applied mathematics score	Ordinal
	X19	Applied chemistry score	Ordinal
	X20	Applied physics score	Ordinal
Behavior and academic experience	X21	Study hours outside class in a week	Real
	X22	Seating position in the classroom	Nominal
	X23	Truant level	Ordinal
	X24	Trouble with the lecturer	Binary
	X25	Study habits with friends	Binary
	X26	Involved in non-academic activities	Binary
	X27	Number of international achievements during college	Real
	X28	Number of national achievements during college	Real
	X29	Number of local achievements during college	Real
	X30	Extracurricular credit score	Real
Non-academic and social life	X31	Legal guardian	Nominal
	X32	Relationship with parents/guardians	Likert
	X33	Part-time job	Binary
	X34	Friendship circle	Binary
	X35	Number of close friends	Real
	X36	Romantic relationship	Binary
	X37	Smoking habit	Ordinal
	X38	Electric smoke habit	Ordinal
	X39	Addicted to alcohol	Ordinal
	X40	Hours of playing games	Real
	X41	Number of hangouts each week	Real
	X42	Physical health issue	Binary
	X43	Mental issue	Binary
	X44	Consulted to psychologist	Binary
	X45	Victim of bully	Binary
	X46	Suicidal thoughts	Binary

2.3. Research design and methodology

The proposed method for academic risk analysis in university students is illustrated in Figure 1. There is an initial preprocessing phase that involves collecting and processing the data to create a proper dataset, including balancing the data. The preprocessed data are then fed into each of the objectives outlined earlier. We first performed SMOTE to address data imbalance before modeling. SMOTE is an oversampling technique that boosts the minority class by generating "synthetic" samples, rather than by duplicating existing ones through replacement [21]. The minority class is oversampled by selecting each sample from the minority class and generating synthetic samples along the line segments connecting it to one or more of its *k*-nearest minority class neighbors. The number of synthetic examples created depends on the desired level of oversampling, and neighbors are randomly chosen from the *k*-nearest neighbors [22], [23].

To address objective 1, we applied LightGBM to build a predictive model for the students' performance, as well as model evaluation. LightGBM enhances the traditional gradient boosting decision tree (GBDT) algorithm by incorporating two novel techniques: gradient-based one-side sampling (GOSS) and exclusive feature bundling (EFB) [24]. These techniques are designed to greatly improve the efficiency and scalability of GBDT [13]. For objective 2, we used SHAP to interpret the prediction results and find the student performance's contribution keys. SHAP is a game-theoretic method that explains the output of any machine learning model by linking optimal credit allocation with local explanations through the classical Shapley value from game theory and its extensions. It assigns an individual's importance value to each feature for a specific prediction. Its innovative aspects include i) the introduction of a new class of additive feature importance measures and ii) theoretical results demonstrating that there is a unique solution within this class that possesses a set of desirable properties [14].

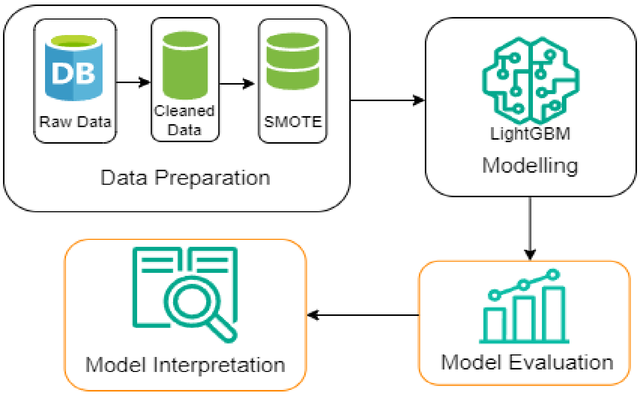

Figure 1. The workflow of this project

3. RESULTS AND DISCUSSION

3.1. Imbalanced data handling and modeling

To better understand the distribution of academic risk, we present Figure 2, which consists of two parts.

Figure 2(a) visualizes the proportion between students with and without academic risk, in which only 38 out of 539 students (7.1%) have academic risk.

Figure 2(b) breaks down the characteristics of at-risk students in the Faculty of Vocational Studies by department. Most respondents are from the Industrial Mechanical Engineering and Business Statistics Departments. However, the most at-risk students are from the Department of Automation and Electrical Engineering and the Department of Business Statistics.

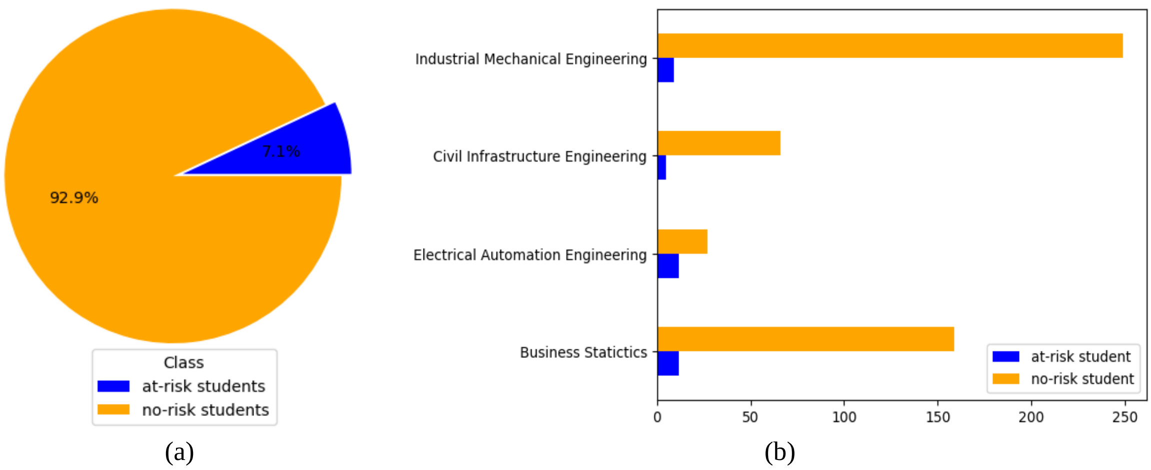

Figure 2. The ratio of students' academic risk (a) in general and (b) by department

The dataset used in this research is imbalanced, which can reduce model performance. This research aims to understand factors contributing to academic risk using SHAP, which requires good model performance. Therefore, managing imbalanced classes is necessary. We use the SMOTE to resample these categories. After SMOTE, the class proportion is balanced to 50-50, with the total number of at-risk students becoming 501. We then apply a machine learning model using LightGBM to provide predictions on academic risk status. Table 2 and Figure 3 compare the model's performance before and after SMOTE.

Table 2. Metric comparison

Metric	Before SMOTE	After SMOTE
Accuracy	0.954	0.960
Sensitivity	0.990	0.948
Specificity	0.200	0.971
AUC-ROC	0.595	0.960

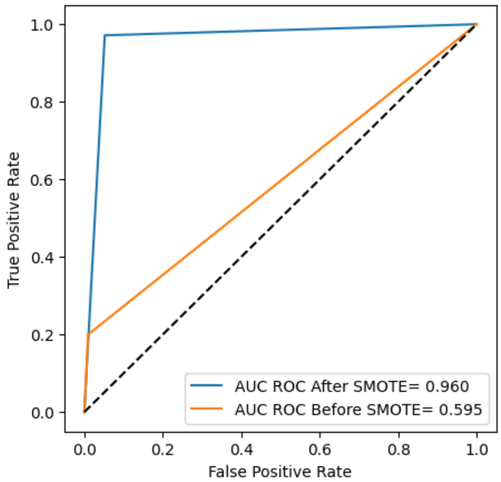

Figure 3. Performance metrics of the model before and after SMOTE

Before SMOTE, the model only managed to correctly classify one-fifth of the students with academic risk. Whereas after SMOTE, 97.1% of students at risk are correctly classified. In addition, the area under the curve-receiver-operating characteristic curve (AUC-ROC) score before SMOTE is 59.5%, which means that the model barely differentiates the two classes. However, after SMOTE, the AUC-ROC score increases significantly to 96%, which means it can now distinguish the two classes almost perfectly. These metrics show that handling imbalanced data using SMOTE can improve the model's performance.

3.2. Discussion based on feature contribution

After obtaining a predictive model, we aim to investigate the effects of the features that significantly impact students' academic performance. The 10 most important features based on SHAP global feature value are shown in Table 3. The SHAP value for the j^{th} feature indicates how much the value of that feature influenced the prediction for a particular instance, in comparison to the average prediction across the dataset. For instance, the extracurricular credit score has a mean absolute SHAP value of 2.18. It means that extracurricular credit scores contribute 2.18 to the prediction compared to the average prediction for this dataset.

Table 3. The 10 most important features according to the mean absolute SHAP value

Variable	Description	SHAP value
X30	Extracurricular credit score	2.18
X2	Department	0.77
X37	Smoking habit	0.77
X1	Sex	0.76
X34	Friendship Circle	0.74
X8	Non-academic achievements during high school	0.54
X16	Satisfaction with major choices	0.50
X21	Study hours outside class in a week	0.49
X26	Involved in non-academic activities	0.47
X41	The number of hours to hang out each week	0.45

Figure 4 shows how the features affect the student's performance. Specifically, it displays the distribution of SHAP values for each feature across all the students. In Figure 4, each student is represented by a dot in each row. The x coordinate of the dot reflects the impact of that feature on the prediction, while its color indicates the value of that feature for the student.

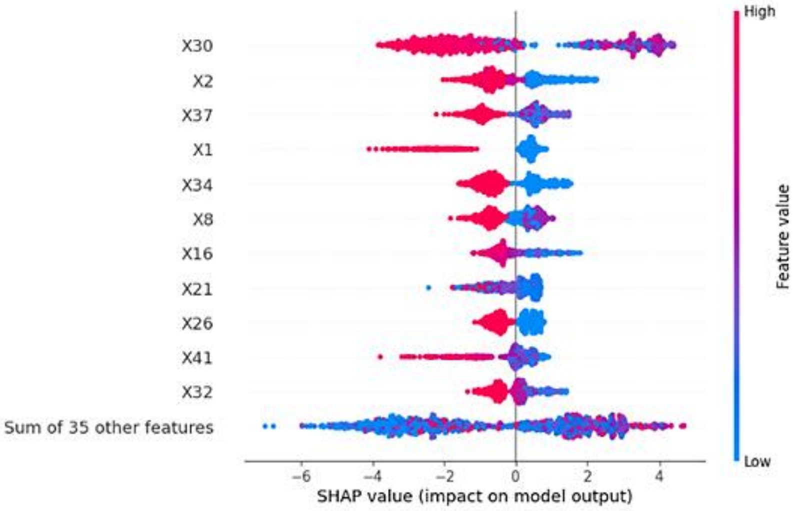

Figure 4. Impact of the features on students' academic performance

To better understand how each factor contributes to academic performance, we break them down into SHAP dependence contribution plots, as shown in Figure 5. Figure 5(a) shows that a higher extracurricular credit score leads to better academic performance. The extracurricular score has a scale of 0 to 4. Students with extracurricular credit scores below 3.0 are more likely to have academic risks than those above 3.0. Therefore, academic policymakers need to provide more support in funding, mentoring, and rewards to motivate students to be more involved in extracurricular activities. Another finding, based on Figure 5(i), shows that students participating in non-academic activities, such as student organizations, also had a lower chance of academic risk. Research by [25], [26] supports these findings, stating that the students participating in extracurricular and non-academic activities have the benefits of attending school more regularly, leading to better grades and higher standardized test scores.

Figure 5(b) shows that the students of the Industrial Mechanical Engineering Department and the Civil Infrastructure Engineering Department have a lower chance of having academic risk. On the other hand, the Business Statistics and Automation Electrical Engineering Department students have a higher chance of having academic risks. Characteristics of survey data might cause the result that the Business Statistics and Automation Electrical Engineering Department has more students with academic risk, as shown in Figure 2(b). Although several studies suggest that sex does not significantly influence academic performance, Figure 5(d) indicates that male students may be more susceptible to academic risk. It is likely because a significant proportion of the students enrolled in the vocational faculty of ITS are male. On the other hand, 49% of male students are smokers, while only 5% of female students are smokers. Figure 5(c) shows that active smokers, despite their smoking frequency, have a higher chance of academic risk than non-smoking students. Therefore, the university should enforce non-smoking zones on campus.

Figure 5(e) shows the effect of the friendship circle on academic risk. A friendship circle may provide support to students, thus lowering their academic risk. It is supported by research conducted by [27]. Their study found that social support was negatively associated with loneliness and positively associated with academic persistence decisions. Figure 5(j) also shows that the students with more hangouts perform better academically. Based on the research by [28], it is noted that students with higher levels of self-esteem and greater peer support tended to have better social adjustment and academic performance, which in turn fostered a stronger commitment to their studies.

Figure 5(f) shows an interesting result where the academic risk is lower for students without non-academic achievements during high school. We presume students who excel in non-academic activities tend to neglect their academic performance. Some colleges, including ITS, consider non-academic high school achievement in admissions. When these students are accepted into college without sufficient academic ability, they have lower retention rates and are more likely to face academic risk.

Satisfaction with major choices also affects academic risk. Students satisfied with their major choice are less likely to have academic risk than those who are not, according to Figure 5(g). Another finding that is likely related to it, based on Figure 5(h), is that students with higher study hours outside class tend to have a lower chance of having academic risk. We presume that students who are more satisfied with their major choices are more likely to have higher study hours outside of class, leading to a lower academic risk.

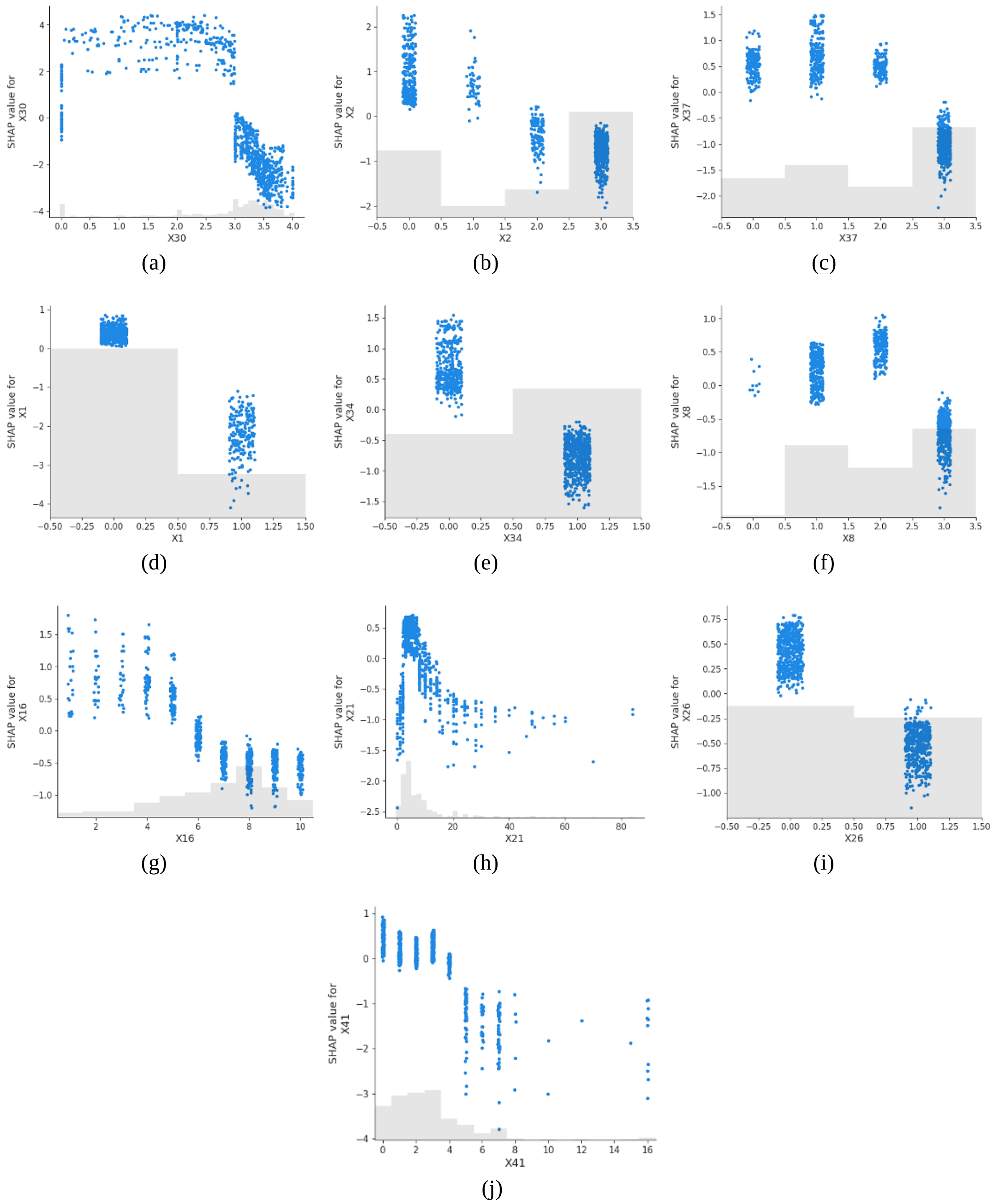

Figure 5. SHAP dependence plots of the top 10 contributing factors to students' academic risk: (a) extracurricular credit score, (b) department, (c) smoking habit, (d) sex, (e) friendship circle, (f) non-academic achievement, (g) satisfaction with major choices, (h) study hours outside class per week, (i) involvement in non-academic activities, and (j) number of hangouts each week

4. CONCLUSION

Academic risk can impact students’ educational and future career prospects, specifically in higher education. Therefore, providing a mechanism to foresee students with academic risk is necessary. This research uses a LightGBM model to construct a predictive model on an imbalanced dataset. An upsampling technique based on SMOTE has been applied to handle imbalance problems. A model obtained from the LightGBM algorithm achieved a decent accuracy and specificity of 96% and 97%, respectively. Based on the results, students from the Department of Business Statistics have the highest academic risk compared to other departments in the Faculty of Vocational Studies. Factors affecting students’ academic performance come from academic, personal, and social aspects. Stakeholders from the department and faculty members can use the information to formulate policies to mitigate academic risk among students, especially factors of academic parts. The admission team should give more attention to prospective students’ high school non-academic achievements. Given how extracurricular credit scores affect academic risk in students, providing more support in funding, mentoring, and rewards is crucial. In addition, satisfaction with major choices also plays an essential role in reducing students’ academic risk. Therefore, the departments should improve their services to students, such as teaching quality, administrative processes, and teaching and learning facilities. Students are also encouraged to study outside class with friends, be more active in non-academic activities, and avoid cigarettes. The university can support this by providing public areas that support learning activities for students, enforcing non-smoking zones inside campus areas, and encouraging students to be involved in non-academic activities.

FUNDING INFORMATION

This research was financially supported by Institut Teknologi Sepuluh Nopember through the Institutional Development Research Grant program under contract number 1799/PKS/ITS/2023.

AUTHOR CONTRIBUTION STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Mukti Ratna Dewi	✓	✓		✓	✓	✓	✓		✓	✓		✓	✓	✓
Mochammad Reza Habibi	✓	✓		✓	✓	✓	✓		✓	✓		✓	✓	✓
Bassam Babgei			✓		✓		✓	✓	✓	✓	✓			
Lovinki Fitra Ananda			✓		✓		✓	✓	✓		✓			
Brodjol Sutijo Suprih Ulama	✓								✓					✓

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest

DATA AVAILABILITY

The data supporting the findings of this study are not publicly available due to confidentiality restrictions.

REFERENCES

[1] S. Yang, “Who will dropout from university? academic risk prediction based on interpretable machine learning,” *arXiv-Computer Science*, Dec. 2021, doi: 10.48550/arxiv.2112.01079.
[2] N. Nurmalitasari, Z. A. Long, and M. F. M. Noor, “Factors influencing dropout students in higher education,” *Education Research International*, vol. 2023, no. 1, pp. 1–13, Feb. 2023, doi: 10.1155/2023/7704142.
[3] O. Lorenzo-Quiles, S. Galdón-López, and A. Lendínez-Turón, “Corrigendum: factors contributing to university dropout: a review,” *Frontiers in Education*, vol. 8, May 2023, doi: 10.3389/feduc.2023.1191708.

[4] Y. Bengio, Y. Lecun, and G. Hinton, “Deep learning for AI,” *Communications of the ACM*, vol. 64, no. 7, pp. 58–65, Jun. 2021, doi: 10.1145/3448250.

[5] I. Koprinska, J. Stretton, and K. Yacef, “Students at risk: detection and remediation,” *Proceeding of the 8th International Conference on Educational Data Mining, EDM15*, pp. 512–515, 2015.

[6] N. Purwaningsih and D. R. Arief, “Predicting students’ performance in English class,” *AIP Conference Proceedings*, vol. 1977, no. 1, Jun. 2018, doi: 10.1063/1.5042876.

[7] M. Baranyi, M. Nagy, and R. Molontay, “Interpretable deep learning for university dropout prediction,” in *Proceedings of the 21st Annual Conference on Information Technology Education*, 2020, pp. 13–19, doi: 10.1145/3368308.3415382.

[8] B. Albreiki, “Framework for automatically suggesting remedial actions to help students at risk based on explainable ML and rule-based models,” *International Journal of Educational Technology in Higher Education*, vol. 19, no. 1, pp. 1–26, Dec. 2022, doi: 10.1186/s41239-022-00354-6.

[9] M. Katsuragi and K. Tanaka, “Dropout prediction by interpretable machine learning model towards preventing student dropout,” in *Advances in Transdisciplinary Engineering*, vol. 28, IOS Press, 2022, pp. 678–683, doi: 10.3233/ATDE220700.

[10] M. Nagy and R. Molontay, “Interpretable dropout prediction: towards XAI-based personalized intervention,” *International Journal of Artificial Intelligence in Education*, vol. 34, no. 2, pp. 274–300, 2024, doi: 10.1007/s40593-023-00331-8.

[11] P. Padmasiri and S. Kasthuriarachchi, “Interpretable prediction of student dropout using explainable AI models,” in *2024 International Research Conference on Smart Computing and Systems Engineering (SCSE)*, 2024, vol. 7, pp. 1–7, doi: 10.1109/SCSE61872.2024.10550525.

[12] T. H. Nguyen, P. Le, T. T. T. Nguyen, and A. K. Su, “A multivariate analysis of the early dropout using classical machine learning and local interpretable model-agnostic explanations,” *CTU Journal of Innovation and Sustainable Development*, vol. 16, pp. 98–106, 2024, doi: 10.22144/ctujoisd.2024.327.

[13] G. Ke *et al.*, “LightGBM: a highly efficient gradient boosting decision tree,” *Advances in Neural Information Processing Systems*, pp. 3147–3155, 2017.

[14] S. M. Lundberg and S. I. Lee, “A unified approach to interpreting model predictions,” *Advances in Neural Information Processing Systems*, pp. 4766–4775, May 2017.

[15] T. Hasanin, T. M. Khoshgoftaar, J. L. Leevy, and R. A. Bauder, “Severely imbalanced big data challenges: investigating data sampling approaches,” *Journal of Big Data*, vol. 6, no. 1, pp. 1–25, 2019, doi: 10.1186/s40537-019-0274-4.

[16] P. Kumar, R. Bhatnagar, K. Gaur, and A. Bhatnagar, “Classification of imbalanced data: review of methods and applications,” in *IOP Conference Series: Materials Science and Engineering*, 2021, vol. 1099, no. 1, doi: 10.1088/1757-899x/1099/1/012077.

[17] Y. Gao, Y. Zhu, and Y. Zhao, “Dealing with imbalanced data for interpretable defect prediction,” *Information and software technology*, vol. 151, 2022, doi: 10.1016/j.infsof.2022.107016.

[18] Y. Chen, R. Calabrese, and B. Martin-Barragan, “Interpretable machine learning for imbalanced credit scoring datasets,” *European Journal of Operational Research*, vol. 312, no. 1, pp. 357–372, 2024, doi: 10.1016/j.ejor.2023.06.036.

[19] D. Dablain, C. Bellinger, B. Krawczyk, D. W. Aha, and N. Chawla, “Understanding imbalanced data: XAI & interpretable ML framework,” *Machine Learning*, vol. 113, no. 6, pp. 3751–3769, 2024, doi: 10.1007/s10994-023-06414-w.

[20] I. P. Y. Pradnyana, I. G. A. S. Darmayani, I. G. A. H. Sundariyati, and I. B. A. P. Manuaba, “Relationship between university admission selection and academic prediction in medical students,” *E-Jurnal Medika Udayana*, vol. 14, no. 3, pp. 107–112, 2025, doi: 10.24843/mu.2025.v14.i3.p17.

[21] N. V Chawla, K. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: synthetic minority over-sampling technique,” *Journal of Artificial Intelligence Research*, vol. 16, no. 1, pp. 321–357, 2002.

[22] D. Elreedy and A. F. Atiya, “A comprehensive analysis of synthetic minority oversampling technique (SMOTE) for handling class imbalance,” *Information Sciences*, vol. 505, pp. 32–64, 2019, doi: 10.1016/j.ins.2019.07.070.

[23] A. Patil, A. Framewala, and F. Kazi, “Explainability of smote based oversampling for imbalanced dataset problems,” in *2020 3rd international conference on information and computer technologies (ICICT)*, 2020, pp. 41–45, doi: 10.1109/ICICT50521.2020.00015.

[24] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–794, doi: 10.1145/2939672.2939785.

[25] A. P. Rahayu and Y. Dong, “The relationship of extracurricular activities with students’ character education and influencing factors: a systematic literature review,” *AL-ISHLAH: Jurnal Pendidikan*, vol. 15, no. 1, pp. 459–474, 2023, doi: 10.35445/alishlah.v15i1.2968.

[26] A. F. Zaki, “Investigating the impact of extracurricular activities on the academic and social skills of university students in post-covid-19: a case study,” in *Future Trends in Education Post COVID-19: Teaching, Learning and Skills Driven Curriculum*, Singapore: Springer Nature, 2024, pp. 251–262, doi: 10.1007/9789819919277_20.

[27] I. Backhaus *et al.*, “Mental health, loneliness, and social support among undergraduate students: a multinational study in Asia,” *Asia-Pacific Journal of Public Health*, vol. 35, no. 4, pp. 244–250, 2023, doi: 10.1177/10105395231172311.

[28] W. Rahardjo, M. Hermita, N. Qomariyah, and I. A. Andriani, “Is academic achievement influenced by self-esteem, loneliness, and internet addiction?,” *Psychopreneur Journal*, vol. 7, no. 1, pp. 1–14, 2023, doi: 10.37715/psy.v7i1.3403.

BIOGRAPHIES OF AUTHORS

Mukti Ratna Dewi is an assistant professor in the Department of Business Statistics, Institut Teknologi Sepuluh Nopember (ITS), Surabaya, Indonesia. In 2018, she earned a Master of Statistics degree from the University of Glasgow, UK, following her B.Sc. in statistics from ITS in 2015. She is currently pursuing a Ph.D. at the University of Edinburgh with research interests that include machine learning, computational statistics, and multivariate analysis. She can be contacted at email: mukti_ratna@its.ac.id.

Mochammad Reza Habibi received his bachelor's and master of mathematics degrees from Institut Teknologi Sepuluh Nopember (ITS), Indonesia, in 2018 and 2021, respectively. He started his career as a lecturer in the Department of Business Statistics, ITS in 2022. He is currently pursuing a Ph.D. at Kumamoto University with research interests that include machine learning and data mining. He can be contacted at email: reza.habibi@its.ac.id.

Bassam Babgei is a final student at the Department of Business Statistics, Faculty of Vocational Studies, ITS. His final project is about the development of an appraisal model based on satellite imagery data using a deep learning method for property appraisal in Surabaya City. He can be contacted at email: 2041221066@student.its.ac.id.

Lovinki Fitra Ananda received his bachelor of statistics degree from Institut Teknologi Sepuluh Nopember (ITS), Indonesia in 2024. His final project is about the academic risk prediction of Faculty of Vocational Studies ITS students using random forest. He can be contacted at email: 2043201112@student.its.ac.id.

Brodjol Sutijo Suprih Ulama held a Doctor of Statistics degree from Gajah Mada University, Yogyakarta, Indonesia, in 2008. In 2011, he received his master's degree in Statistics from Bogor Agricultural University (IPB), Bogor, Indonesia. As for the Bachelor of Statistics program, he got it from Institut Teknologi Sepuluh Nopember (ITS), Surabaya, Indonesia, in 1989. He is currently an Associate Professor in the Department of Business Statistics, ITS, Surabaya, Indonesia. His doctoral research is about radial basis function for time series data modeling. Most of his research is about time series and neural networks for data analysis. He can be contacted at email: brodjol_su@statistika.its.ac.id.