electronics

MDPI

Article

Unpacking Prediction: Contextualized and Interpretable Academic Risk Modeling with XAI for Small Cohorts

Di Sun [1,2], Pengfei Xu [3], Gang Cheng [4] and Ping Zhang [5,*]

[1] Graduate School of Education, Dalian University of Technology, Dalian 110624, China; sundi@dlut.edu.cn
[2] Engineering Research Center of Integration and Application of Digital Learning Technology of Ministry of Education in China, Beijing 100039, China
[3] School of Artificial Intelligence, Beijing Normal University, Beijing 100875, China; xupf@bnu.edu.cn
[4] Digitalization Department, The Open University of China, Beijing 100039, China; chenggang@ouchn.edu.cn
[5] School of Health and Social Care, Shanghai Urban Construction Vocational College, Shanghai 201415, China
* Correspondence: zhangping2006@fudan.edu.cn

Abstract

Effective prediction of academic risk is vital in higher education to enable timely intervention and support student retention. While the introduction of Educational Data Mining (EDM) has enhanced prediction effectiveness, existing research often focuses only on single factors or large scale samples, and is notably deficient in providing transparent explanations for prediction results. To address these gaps, this study proposes an Explainable Artificial Intelligence (XAI) framework for predicting and interpreting academic risk within a high-dimensional, small sample context. Based on a dataset from a specific student cohort, we employed an ML model combined with SHapley Additive exPlanations (SHAP) method as the XAI framework. The findings provide two major contributions to the "Data-Related Challenges in ML" discussion. Firstly, by leveraging the XAI framework, it successfully enhances data interpretability, revealing the out-of-class peer support as the feature with the strongest association with academic risk, which is a complex and often underestimated data dimension, surpassing traditional academic metrics. Specifically, learning support from peers is identified as the most critical feature in mitigating risk at both the group and individual levels. Secondly, methodologically, this framework validates a reliable approach for extracting meaningful, trustworthy, and interpretable knowledge from limited and specific cohort data, offering a solution for applications with highly contextualized and precise interventions, where large, generalizable datasets are impractical. In conclusion, this study enhances the transparency and trustworthiness of ML in EDM, ensuring responsible intervention strategies in academic risk prediction.

Keywords: prediction of academic risk; explainable AI; learning peers; small datasets; high-dimensional features

Check for updates

Academic Editor: Ping-Feng Pai

Received: 30 November 2025
Revised: 28 January 2026
Accepted: 28 January 2026
Published: 2 February 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Helping students complete their courses and successfully graduate is one of the pursuits of our higher education systems. However, based on the report of the Organization for Economic Co-operation and Development (OECD), despite registration rises, the drop-out rate of higher education in the countries of OECD generally averages 30%: the drop-out rate in Nordic countries is relatively low; however, the drop-out rate in the United States is relatively high, especially in two-year community colleges, which exceeds 40%; in addition, drop-out rates in Southern European countries are even higher, with some

regions approaching 50% [1]. Increasing graduation rates continues to be a high priority for higher education administrators [2]. Predicting students' academic risk is critical for higher education to identify low-performing students early enough to help them overcome difficulties, improve their academic performance, and finally prepare them for graduation [3].

Academic risk is closely related to academic performance; they are two sides of the same coin. When academic risk is high, academic performance is low; conversely, the same is true. With the rapid advancement of educational big data and machine learning (ML) techniques, Educational Data Mining (EDM) has demonstrated substantial potential for identifying risk factors and predicting student performance [4]. However, a notable imbalance persists in current EDM research: scholars often prioritize predictive performance (e.g., accuracy, F1 score, or AUC) while paying comparatively less attention to model interpretability [5]. This "outcome-oriented, process-neglecting" tendency has been repeatedly documented in systematic reviews, which show that much of the literature concentrates on bench-marking algorithmic performance, whereas how educational stakeholders (e.g., instructors and administrators) can understand the rationale behind model decisions remains underexamined [6,7]. In pursuit of marginal performance gains, many studies have increasingly adopted complex deep neural networks; although such models may raise the predictive ceiling, they frequently function as opaque "black boxes" [8,9]. In human-centered educational contexts, this lack of transparency can erode practitioners' trust limits the capacity of predictive systems to yield pedagogically meaningful guidance for individualized intervention [6]. Against the backdrop of growing concerns regarding fairness, accountability, and ethics in educational artificial intelligence, Explainable Artificial Intelligence (XAI) has emerged as a pivotal avenue to address this gap [7]. By translating complex model outputs into transparent and actionable insights, XAI offers a pathway to couple predictive effectiveness with scientifically grounded and transparent educational decision-making [10,11].

Moreover, EDM often focuses on large datasets; however, within a university, there is considerable variation across different majors, and the number of students in each major is typically limited [12]. This "large-data" orientation is closely tied to the rise in online platforms: a prominent EDM review notes a shift from traditional offline classrooms with small datasets toward "massive educational data from online education" (e.g., Coursera/edX), which has catalyzed a large body of work built around large-scale logs and massive participation contexts [13]. Related learning-analytics scholarship further argues that many machine-learning-based analytics techniques "require big data" to produce reliable and high-accuracy prediction models, whereas small-scale courses and bounded cohorts (e.g., specialized programs) remain comparatively under-addressed despite similar institutional needs [14]. This presents a challenge when dealing with small datasets, making it crucial to explore effective methods for predicting students' academic risk in such contexts. Additionally, while researchers have identified several key factors influencing academic risk, such as prior academic achievement, student demographics, e-learning activity, psychological attributes, and social information [15–17], peer interactions, and support out of class have yet to be thoroughly examined and warrant further investigation [3,18].

In this study, we focus on a small dataset from a specific major at a technology university, combining information from learning peers with other influential factors to predict academic risk using XAI method, SHapley Additive exPlanations (SHAP). Our workflow mainly includes data preparation with unified preprocessing, imbalance-aware hyperparameter tuning for candidate models via five-fold stratified cross-validation, robustness assessment with uncertainty quantification, optimal model selection, and SHAP-based global and individual explanations. This approach is designed to support reliable early

warning in small-cohort settings while providing actionable explanations for administrators and instructors. It may help policy-makers, administrators, and instructors better understand academic risk prediction models in small datasets and be more confident in making decisions to improve current educational procedures and learning resources.

2. Related Work

2.1. Academic Risk Prediction

Previous research on academic risk prediction used to employ traditional qualitative or quantitative methods to identify critical factors that influence students' performance. For example, Troll et al. investigated the link between students' self-control and academic performance [17]. Goh et al. examined the relationship between emotional intelligence (EI) and academic performance among hospitality master students by regression analysis [19]. With longitudinal data from 4489 college students, Spight conducted logistic regression to examine the potential relationship between matriculating with or without a major declared and degree completion [2]. The factors influencing students' academic risk identified in these studies are usually limited to one or a few variables, and the relationships between these factors and academic risk are generally interpreted linearly. However, the factors influencing academic risk are diverse, and their effects may follow non-linear patterns [20]. Consequently, describing a high-dimensional landscape that captures the interplay of these various factors has long been an area of interest for many researchers.

In recent years, EDM has gradually become more and more popular. As one of the challenging topics in EDM, academic risk prediction has moved into data-intensive computation. Some main categories of factors have been summarized and reported. Abu Saa et al. classified four categories of influential factors based on their review: (1) students' e-learning activity, the activity logs of students in e-learning systems, such as the number of logins, the number of assignments performed, number of quizzes performed. (2) Students' previous grades and class performance, the grades or other performance indicators of students in previous courses, semesters, or years. (3) Students' demographics, such as gender, age, nationality, and ethnicity. (4) Students' social information, information related to social life, like the number of friends, if s/he smokes or not [3]. Later, Alyahyan et al. reported their review of influential factors: (1) prior academic achievement, which includes high school background, and pre-admission data. (2) Student demographics, which means gender, age, race/ethnicity, and socioeconomic status. (3) Students' environment, which includes class type, semester period, and program type. (4) Psychological attributes, which include student interest, behavior of study, stress, anxiety, time of preoccupation, self-regulation, and motivation. (5) Student e-learning activity, which includes the number of login times, number of tasks, number of tests, assessment activities, number of discussion board entries, and number/total time material viewed [18].

A particular focus in EDM research for predicting academic risk is to verify the efficiency of approaches and the accuracy of models. For example, Li et al. demonstrated that the Sequential Prediction based on deep network model outperforms the baselines and significantly improves early academic warning [21]. Sokkheyet et al. introduced an optimization approach for improving the classification performance of a deep learning framework called deep belief networks (DBNs), which is more accurate and effective than other proposed algorithms [22]. Because the conventional single classifier-based predictive analysis is not efficient in providing accurate results, Ramanathan et al. introduced a novel technique, Minkowski Sommon Feature Map-based Densely connected Deep Convolution Network with LSTM; the outcomes inferred that this technique performs well with higher precision, recall, f-measure, and lesser time consumption than the state-of-the-art methods [23]. Pallathadka et al. used various ML algorithms to analyze performance

metrics including accuracy and error rates, and emphasized the effectiveness of EDM in predicting and classifying student performance [24]. Arashpour et al. used a teaching-based optimization (TLBO) algorithm to predict individual student test scores and demonstrated the utility of these hybrid models in improving the accuracy of student achievement predictions [25]. More recently, Casillano and Cantilang employed a comparative approach to predict drop-out risks among programming students by testing multiple classifiers, including k-Nearest Neighbor (kNN), Decision Trees, Logistic Regression, and Neural Networks. Their findings indicated that kNN achieved the highest predictive accuracy, followed closely by Decision Trees, Logistic Regression, and Neural Networks. Also, the study identified assignment completion, participation in laboratory exercises, and attendance as critical predictors of drop-out risk [26]. Complementing these findings, Malik et al. advanced EDM research by introducing a stacking ensemble model that integrates Random Forest, XGBoost, Gradient Boosting, and feedforward neural networks to enhance the prediction of student drop-out in higher education. Their study demonstrated that ensemble learning approaches consistently outperformed individual classifiers in terms of both accuracy and AUC, and underscored the importance of incorporating high-dimensional features such as socioeconomic status, environmental conditions, and family background [27].

2.2. Reflection of the Existing Literature

First, given the advantages in data analysis and the development of artificial intelligence technology, EMD has gradually become the mainstream method to predict academic risk, and verification of model accuracy holds most of the attention in these EDM-based studies. However, the issue of how to achieve a proper explanation of the model is often overlooked. Generally, EDM involves two main topics to pursue: one is the "what," and the other is the "why." The former means the factors or features predicting certain tasks or problems; model accuracy is the essential guarantee of what; the latter is how the model comes to the prediction, which means model interpretability. Interpretability is the degree to which a human can understand the cause of a decision and predict the model's result [28]. The higher the interpretability of a prediction model, the easier to comprehend why predictions have been made [29]. The single focus on model accuracy is an incomplete description of tasks; convincing explanations are necessary, which may strongly support the stakeholders to make confident decisions in real-world practice [30]. Therefore, model interpretability should not be ignored in the research on academic prediction.

Second, although many studies have investigated the main categories of factors predicting academic risk, few have paid particular attention to peer support indicating interactive behavior between students. Researchers listed factors like students' social information or students' environment, which seem to be close to collaboration or interaction [3,31]; however, social information always means information related to a social life instead of a learning process, and students' environment only includes the information of class type, semester duration, and type of program. However, such factors cannot reveal students' interaction with peers. In fact, sociocultural and social learning theories have long emphasized that learning is not an isolated individual activity but a socially situated process [32,33]. Student persistence and academic success are strongly influenced by both academic and social integration within the learning community, and peers serve not only as collaborators but also as sources of motivation, comparison, and emotional support in learning environments [34–36]. Nevertheless, most of these studies focus on structured collaboration within courses or online systems, while informal, out-of-class peer learning among university students has received limited attention [37]. Higher education research often assumes that college students are independent learners [38,39]; however, recently,

emerging evidence suggests that out-of-class learning peers can offer cognitive scaffolding and socio-emotional support that are crucial for academic success [40].

Third, previous studies pay more attention to large datasets than small ones to pursue generalization. However, in higher education, each major is different, and students' performance varies greatly. Also, the enrollment of each major at each university is limited, which often falls short of handling large datasets and cannot effectively support generalization to a broader range. Therefore, rather than overemphasizing large-scale promotion, it is necessary to focus on the specific learning background and small datasets, as well as to explore effective methods for predicting academic risk.

In summary, while substantial progress has been made in identifying cognitive, behavioral, and demographic predictors of academic risk, significant gaps remain in three areas: (1) despite the widespread use of machine learning for risk prediction, interpretability is often neglected, limiting stakeholders' understanding of the mechanisms behind predictions; (2) prior research has rarely examined how peer support, especially informal, out-of-class learning relationships, impacts on academic risk; (3) most studies emphasize large, generalized datasets, overlooking small, domain-specific cohorts that demand personalized, context-sensitive analysis. These gaps underscore the need for data-driven yet explainable approaches to reveal how and why peer factors influence academic risk, and for explainable models to uncover individualized risk patterns within small cohorts, together forming the central motivation and contribution of the present study.

2.3. Explainable Artificial Intelligence (XAI) and SHAP

Fortunately, the research community has increasingly realized the importance of interpretability, and there has been a trend seeking to investigate more transparency and explanation of EDM models [28,41]. The factors making the prediction are always called features in EDM. The best explanation of a simple model is the model itself; it perfectly represents itself and is easy to understand based on the weight (or importance) of each feature. However, for complex models, such as ensemble methods or deep networks, the original model cannot be the best explanation on its own because the analysis process is black-box and not easy to understand [42]. Therefore, researchers have been trying to use new methods to pursue explanations such as XAI [29,42].

The common methods of interpretation include Partial Dependence Plot, Individual Conditional Expectation, Accumulated Local Effects Plot, and Shapley value [41]. Lundberg et al. particularly elaborated on a unified framework for interpreting predictions, which is SHAP [42]. SHAP constructs an additive interpretation model based on the Shapley value. The Shapley value is a method inspired by coalitional game theory [29,43]. In this framework, the "game" is the prediction task for a single instance; the "gain" is the actual prediction for this instance minus the average prediction for all instances (the baseline); the "players" are the feature values of the instance that collaborate to receive the gain. The Shapley value quantifies the average marginal contribution of a feature value across all possible coalitions (feature subsets) [42,43].

Formally, for a given machine learning model f and a specific instance with feature vector $\boldsymbol{Xi}$, the SHAP explanation model g is defined as an additive linear function of binary variables:

$$g(z') = \phi_0 + \sum_{j=1}^{M} \phi_j z'_j \quad (1)$$

where $z' \in \{0,1\}^M$ represents the presence (1) or absence (0) of each of the M features, ϕ_0 is the model's expected output over the background dataset (i.e., $E[f(\boldsymbol{X})]$), and ϕ_j is the Shapley value for feature j [42]. The key property of this explanation model is local accuracy: the

sum of the feature contributions ϕ_j plus the baseline ϕ_0 equals the model's actual prediction for the instance $\boldsymbol{Xi}$:

$$f(\boldsymbol{Xi}) = \phi_0 + \sum_{j=1}^{M} \phi_j \quad (2)$$

Here, ϕ_j represents the contribution of feature j to the prediction $f(\boldsymbol{Xi})$. A positive ϕ_j indicates that the feature increases the prediction value relative to the baseline, while a negative value indicates a decrease. This formulation ensures that the prediction is fairly distributed among the features according to their marginal contributions [42,43].

Because the Shapley value is based on solid game theory and guarantees that the prediction is fairly distributed among the features, SHAP might be the method to deliver a full explanation, which means it is a solution for computing feature contributions for single predictions for any machine learning model [29,42]. Therefore, to present the internal operating mechanism with XAI methods such as SHAP, education administrators can not only obtain more accurate prediction results but also understand the reasons behind the prediction [29].

3. Materials and Methods

3.1. Research Questions

In this study, we employ ML techniques and the XAI method to investigate the important factors predicting academic risk in a specific major at a technology university in China. Specific small datasets, the explanation of the model, and peer support factors influencing risk are the three aspects we pay attention to.

The research questions are as follows:

RQ1: *At the overall level, what factors and how do they predict students' academic risk under a specific background?*

RQ2: *At the individual level, what factors and how do they predict students' academic risk under a specific background?*

3.2. Data and Features

3.2.1. Data Collection and Feature Summary

We collected desensitization data of 482 students enrolling from 2021 to 2023 in the major of computer science from a university in northern China. After filtering missing data and data redundancy, data from 466 students were finally obtained. In this study, Grade Point Average (GPA) at the end of the second year is defined as the indicator of academic risk. In China, GPA during the first two years in university is a key indicator of final graduation. A second-year GPA above 2 means the student has a qualified performance for successful graduation; otherwise, the student will receive a warning of drop-out, which indicates academic risk. Therefore, to identify students at risk early and prepare them for final graduation, this study decided to identify the features influencing the second-year GPA instead of the final GPA.

Based on the literature review and the real background of the university, five categories of features are collected: (1) demographics, including Gender, Age at entrance, Guardian type, From urban or rural, and Entrance type; (2) College entrance exam scores, regarded as the prior academic achievement, including Chinese score, Foreign language score, Math score, and Comprehensive subjects score; (3) learning activity, including Self-study, Seating choice in a classroom, Truancy level, Academic awards, Teacher–student relationship, Part-time job, and In resident; (4) learning peers, including learning support from learning peers, Social and emotional support from learning peers, Stability of learning peers, Number of learning peers, and Dorm learning climate; (5) social life, including Love relationships,

Campus loan, Smoke, and Playing video games. The details of the features are listed in Table 1.

Table 1. Summary of features.

Feature Category	Feature	Abbreviation	Value
Demographics	Gender	Gndr	Male = 1, Female = 0
	Age at entrance	Age	16, 17, 18, 19, 20, 21, 22, . . .
	Guardian type	Gurdn	Parents = 1, Father = 2, Mother = 3, Other = 4
	From urban or rural	UrbnRrl	Rural = 0, Urban = 1
	Entrance type	EntrTp	Re-taker = 0, Freshman = 1
College entrance exam scores	Chinese score	ExamCn	Mean (Std): 114.82 (7.51)
	Foreign language score	ExamFrn	Mean (Std): 110.76 (10.62)
	Math score	ExamMth	Mean (Std): 106.15 (9.67)
	Comprehensive subjects score	ExamCmpr	Mean (Std): 221.16 (19.86)
Learning activity	Self-study	SlfStdy	Never = 1, Seldom = 2, Sometimes = 3, Always = 4
	Seating choice in a classroom	Seat	Back = 1, Middle = 2, Front = 3
	Truancy level	Truant	Never = 1, Seldom = 2, Sometimes = 3, Always = 4
	Academic awards	Awrds	No = 0, Yes = 1
	Teacher–student relationship	TsRltn	Tense = 1, Neutral = 2, Harmonious = 3
	Part-time job	PtJob	No = 0, Yes = 1
	In campus resident	Rsdnt	No = 0, Yes = 1
Learning peers	Learning support from learning peers	LpLrng	From "Very low = 1" to "Very high = 5"; Mean (Std): 3.21 (1.03)
	Social and emotional support from learning peers	LpScl	From "Very low = 1" to "Very high = 5"; Mean (Std): 3.15 (0.96)
	Stability of learning peers	LpStbl	From "Very low = 1" to "Very high = 5"; Mean (Std): 2.79 (0.94)
	Number of learning peers	LpNmb	From "One = 1" to "Equal or more than 5"; Mean (Std): 3.05 (1.13)
	Dorm learning climate	DmClmt	From "Very low = 1" to "Very high = 5"; Mean (Std): 3.17 (1.16)
Social life	Love relationships	LvRltn	No = 0, Yes = 1
	Campus loan	CmpsLn	No = 0, Yes = 1
	Smoke	Smk	No = 0, Yes = 1
	Playing video games	Game	Never = 1, Seldom = 2, Sometimes = 3, Always = 4

In this study, we changed the general "E-learning activity" to "Learning activity" because face-to-face instruction but not e-learning is the main form in this traditional university. Another type of influence feature is learning peers, which we emphasized here because learning peers is a built-in feature in this specific background: each semester, students were guided to form a loose learning group to facilitate communication and collaboration outside of class. Five features of learning peers are defined: learning support from learning peers, Social and emotional support from learning peers, Stability of learning peers, Number of learning peers, and Dorm learning climate. (Please see the details in Table 1).

In addition, data sources differed across feature categories. Demographics and original College entrance exam scores were obtained from institutional administrative records, and the same was true for the outcome variable (the second-year average grade point). Because scoring standards and full-mark schemes for College entrance exams can vary across provinces in China, after obtaining the original subject scores, we converted them to a unified scoring scheme before model training by rescaling Chinese, Math, and Foreign Language to a 150-point maximum each and the Comprehensive subjects to a 300-point maximum, followed by integer rounding. These harmonized subject scores were then used as model inputs. In contrast, variables under "Learning Activity," "Learning Peers," and

"Social Life" in Table 1 reflect students' self-reported behaviors and perceptions. These self-reported variables were collected through a structured questionnaire that included an ordered Likert scale and binary indicators. Self-reported data may be susceptible to social desirability and recall biases; however, this approach was necessitated by the nature of the variables studied, which represent internal psychological perceptions and informal social interactions. Currently, self-reporting remains the only viable method to access these critical dimensions. To minimize potential biases, the survey was conducted anonymously and then de-identified before analysis, consistent with the IRB approval and informed consent procedures. From a pedagogical perspective, perceived support, even if subjective, is often a more significant driver of academic resilience than objective metrics alone, providing human-centered insights that administrative datasets typically overlook [44].

3.2.2. Data Characteristics and Associated Methodological Challenges

The dataset used in this study presents a multifaceted methodological challenge, which extends beyond a simple characterization of sample size. The key characteristics and their corresponding analytical implications are as follows: (1) high dimensionality relative to sample size: With $p = 25$ predictive features for $n = 466$ samples, the dataset exhibits a non-negligible p/n ratio. This setting increases the risk of model overfitting, demanding careful model selection. (2) Severe class imbalance: the target variable is highly skewed, with only 14.8% of students labeled as being at academic risk. This extreme imbalance necessitates the use of evaluation metrics focused on the minority class and may require algorithmic adjustments to prevent the model from ignoring the critical risk group. (3) Context-specific, homogeneous cohort: the data originates from a single major within a single university, representing a highly specific educational context. While this ensures internal consistency for localized intervention, it also defines a narrow domain with limited sample diversity. The core challenge is to extract reliable and interpretable patterns from this limited, context-bound sample to inform precise interventions, where large, generalizable datasets are neither available nor appropriate.

These characteristics collectively define our research context and directly motivate the methodological choices detailed in the subsequent sections.

3.3. Explainable Prediction Model Based on SHAP

There are two steps to construct the explainable prediction model here: first, different supervised ML algorithms are carried out to construct and simulate the feature data X and the academic risk data y, and the optimal performance model emerges by specific evaluation indicators. Second, SHAP method is employed to build an explainable model based on the optimal performance model obtained from the first step and give a visualization explanation of the prediction of academic risk. Please see the pipeline in Figure 1.

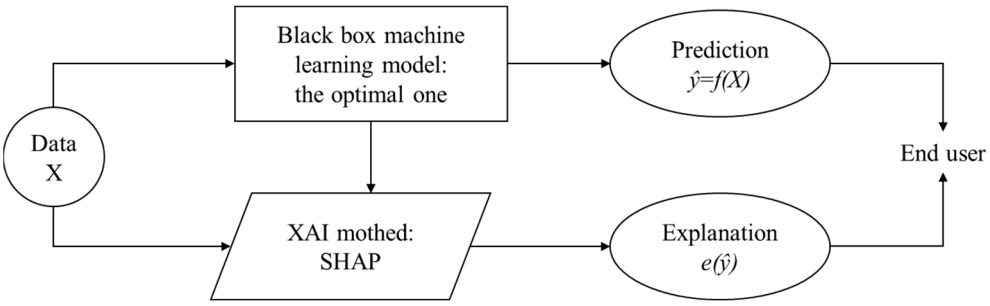


Figure 1. Explainable prediction model pipeline.

3.3.1. The Optimal ML Model

In this study, we carried out common supervised ML algorithms, including Logistic Regression (LR), Support Vector Machine (SVM), Gradient Boosting Decision Tree (GBDT), eXtreme Gradient Boosting Decision Tree (XGBoost), Categorical Boosting Decision Tree (CatBoost), Light Gradient Boosting Machine (LightGBM), K-Nearest Neighbor (KNN), Multinomial Naive Bayes (MNBayes), Random Forest (RF), and Decision Tree (DT).

For the evaluation of machine learning models, common indicators include Accuracy, Precision, Recall, F1-Score, ROC-AUC. However, the ratio between non-academic risk and academic risk in this study is 0.851:0.148, which belongs to an unbalanced classification; PR-AUC and Balanced Accuracy were adopted as the primary indicators for evaluating the performance of the classification model. PR-AUC (Area Under the Precision-Recall Curve) is a performance evaluation metric specifically designed for imbalanced classification. Unlike ROC-AUC, PR-AUC focuses on the identification ability of the minority class (positive class) and is more sensitive to performance variations when positive samples are scarce. It directly reflects the trade-off between precision (the proportion of correctly predicted positives among all predicted positives) and recall (the proportion of correctly identified positives among all actual positives), providing a more informative assessment in scenarios where the class distribution is highly skewed [45,46]. Also, balanced accuracy is an important performance metric used to evaluate classification models, particularly in situations with imbalanced datasets, by averaging the true positive rate and true negative rate [47,48].

We address methodological rigor through (1) stratified 70/15/15 data partitioning scheme (training/validation/test); (2) identical preprocessing (RobustScaler) for all models; (3) algorithm-level class-imbalance handling (class weighting, no oversampling); (4) training and hyperparameter tuning via repeated stratified 5-fold cross-validation with 3 repetitions, yielding optimal configurations including linear models with strong regularization, gradient boosting variants with conservative learning rates (0.05), and tree-based models with depth constraints; (5) variance estimation via standard deviations across cross-validation folds (Table 2); and (6) uncertainty quantification through bootstrap confidence intervals for test-set performance (Table 3).

Table 2. Summary of model performance in cross-validation.

Model	PR-AUC	Balanced Accuracy	Hyperparameters
Logistic Regression	0.5398 ± 0.1330	0.7714 ± 0.0855	{'C': 1, 'penalty': 'l2'}
SVM	0.5430 ± 0.1378	0.7828 ± 0.0573	{'C': 0.1}
Gradient Boosting Decision Tree (GBDT)	0.6616 ± 0.1457	0.6975 ± 0.0867	{'n_estimators': 200, 'max_depth': 3}
eXtreme Gradient Boosting (XGBoost)	0.6266 ± 0.1546	0.7058 ± 0.0919	{'n_estimators': 100, 'max_depth': 7}
Categorical Boosting (CatBoost)	**0.6931 ± 0.1378**	**0.7358 ± 0.0719**	**{'iterations': 200, 'depth': 7}**
Light Gradient Boosting Machine (LightGBM)	0.6092 ± 0.1566	0.7050 ± 0.0918	{'n_estimators': 200, 'max_depth': 7}
K-Nearest Neighbor (KNN)	0.6776 ± 0.1178	0.6635 ± 0.0720	{'n_neighbors': 11, 'weights': 'distance'}
Multinomial Naive Bayes (MNB)	0.5023 ± 0.1295	0.5275 ± 0.0264	{'alpha': 0.1}
Random Forest (RF)	0.6729 ± 0.1459	0.6437 ± 0.0741	{'n_estimators': 200, 'max_depth': 7}
Decision Tree (DT)	0.4465 ± 0.1440	0.7612 ± 0.1017	{'max_depth': 5}

Table 3. Summary of model performance on independent test set.

Model	PR-AUC [CI]	Balanced Accuracy [CI]
Logistic Regression	0.3406 [0.1143, 0.6085]	0.7220 [0.5523, 0.8811]
SVM	0.3637 [0.1213, 0.6271]	0.6258 [0.4653, 0.7857]
Gradient Boosting Decision Tree (GBDT)	0.3329 [0.1274, 0.6334]	0.6057 [0.4575, 0.7721]
eXtreme Gradient Boosting (XGBoost)	0.4423 [0.1779, 0.7753]	0.6841 [0.5463, 0.8447]
Categorical Boosting (CatBoost)	**0.4661 [0.1918, 0.7991]**	**0.6688 [0.5166, 0.8397]**
Light Gradient Boosting Machine (LightGBM)	0.4154 [0.1522, 0.7516]	0.6750 [0.5210, 0.8414]
K-Nearest Neighbor (KNN)	0.3815 [0.1353, 0.7198]	0.5422 [0.4761, 0.6667]
Multinomial Naive Bayes (MNB)	0.3318 [0.1472, 0.5849]	0.5 [0.5, 0.5]
Random Forest (RF)	0.4015 [0.1633, 0.7629]	0.6298 [0.4841, 0.7846]
Decision Tree (DT)	0.3059 [0.1024, 0.5517]	0.6140 [0.4419, 0.8043]

Based the information from Tables 2 and 3, CatBoost holds the best reuslts of model performance (Please see the bold text), therefore, it is the optional model under the situation of unbalanced classification, small sample size, and high dimension in this study.

3.3.2. Configuration and Implementation Details of SHAP

To ensure the reliability and reproducibility of the SHAP-based interpretable framework, the following key implementation details were explicitly configured and are documented here: (1) SHAP explainer: We used SHAP's TreeExplainer, which is designed for tree-based models. It calculates Shapley values exactly or via a fast, tree-specific approximation, eliminating the need for the sampling-based approximations (nsamples) required by the model-agnostic KernelExplainer. (2) Background dataset: To balance computational efficiency and representativeness, we constructed the background dataset by randomly sampling 100 instances from the preprocessed training set, using a fixed random seed (random_state = 42) to ensure full reproducibility of all SHAP results. (3) Feature standardization: The same RobustScaler-standardized values were fed into SHAP. For SHAP results, these values were presented on a Z-score scale (mean = 0, std = 1) to ensure stable and comparable explanations while fully preserving the model's decision logic. (4) Stability Considerations: The TreeExplainer provides deterministic and stable explanations for tree-based models, as it does not rely on random sampling. The stability of our SHAP interpretations is therefore inherent to the explainer choice given our optimal model. We acknowledge that for other, non-tree models in our comparative analysis, approximation stability would be a concern, but our core interpretations are based on the stable outputs from the TreeExplainer applied to CatBoost.

4. Results

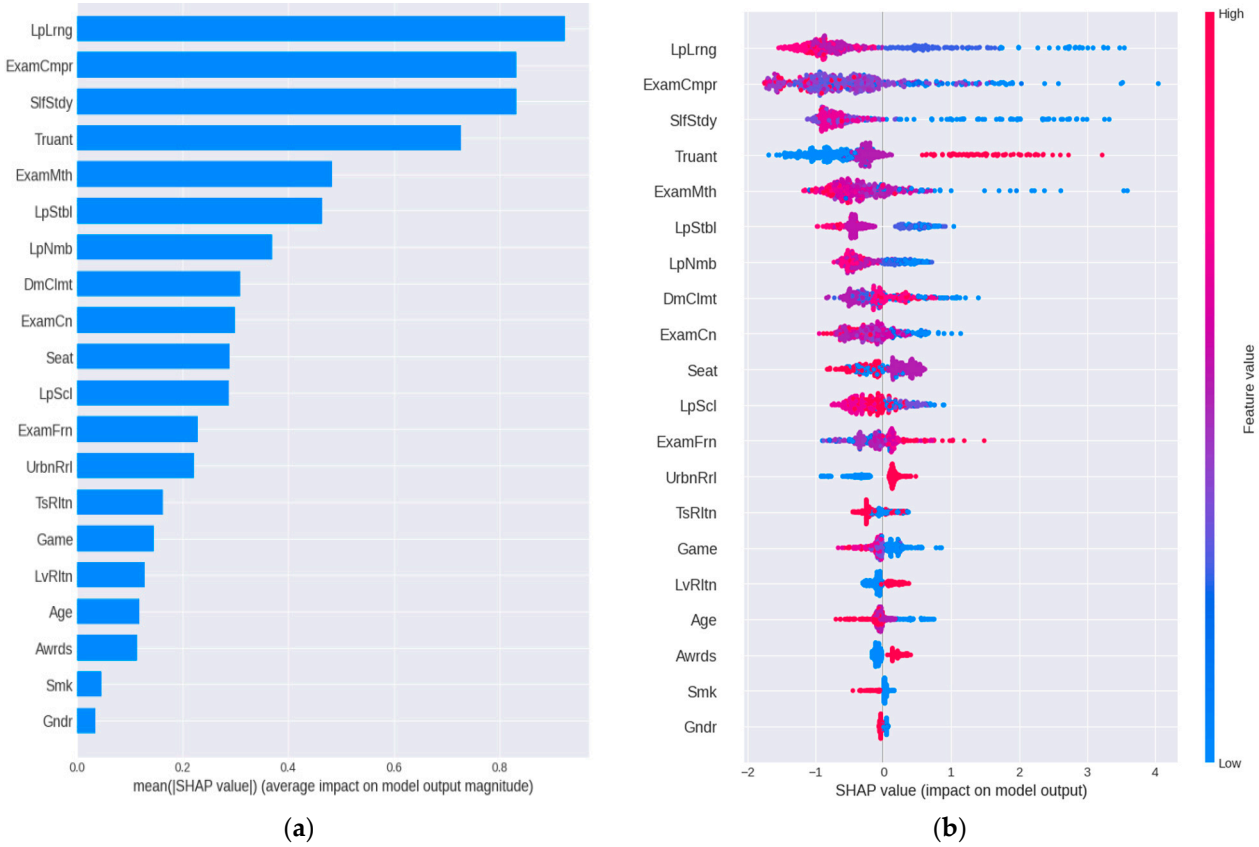
4.1. Factors Predicting Academic Risk at the Overall Level

In this study, the first research question is, at the overall level, what factors, and how, predict students' academic risk within a specific background? We discuss "what" and "how" in the following sections.

4.1.1. Factors Predicting Academic Risk

The SHAP analysis provides crucial insights into how multiple features contribute to predicting academic risk. The bar plot in Figure 2a clearly shows the mean absolute SHAP value for each feature, indicating their overall importance to the predictions. The higher the ranking of a feature, the stronger its predictive and explanatory power for academic risk. In this study, we focus on the first 11 features: LpLrng (learning support from learning peers), ExamCmpnr (Comprehensive subjects score in College entrance exam), SlfStdy (Self-study),

Truant (Truancy), ExamMth (Math score in College entrance exam), LpStbl (Stability of learning peers), LpNmb (Number of learning peers), DmClmt (Dorm learning climate), ExamCn (Chinese score in College entrance exam), Seat (Seating choice in a classroom), LpScl (Social and emotional support from learning peers).



(**a**) (**b**)

Figure 2. SHAP values: (**a**) The bar plot of the mean of absolute SHAP value. (**b**) The beeswarm plot of SHAP value.

The beeswarm plot in Figure 2b extends the results in Figure 2a by showing the distribution of SHAP values for each feature across all observations. In Figure 2b, each scatter point represents a student in the sample, and the color of each dot indicates the actual value of that feature for the specific observation it belongs to; red typically signifies high values, while blue typically signifies low. The horizontal axis (SHAP value) indicates the degree of impact of the feature on individual academic risk; the effect of the same feature may vary across different students. Additionally, the horizontal axis illustrates not only the magnitude but also the direction of each feature's impact: the right side (+SHAP) means the feature pushes the prediction higher, while the left side means lower. If most red dots are on the left and blue ones on the right, it means that high values of this feature decrease the prediction, and low values increase it, which is a negative relationship between the feature and the prediction, and vice versa. Taking LpLrng (the most important feature) as an example, high LpLrng values (red dots) are generally associated with negative SHAP values, which means greater learning support from peers correlates with a lower predicted risk. Conversely, low LpLrng values are associated with positive SHAP values, indicating that less peer support is linked to a higher predicted risk score. The vertical axis (Features) has no practical significance, but due to the repetition of SHAP values across multiple samples, a visual effect of fluctuation is generated, forming a vertical distribution.

4.1.2. How Specific Factors Influence the Prediction of Academic Risk

Additionally, a SHAP dependence plot illustrates how a specific feature influences the model's predictions. The horizontal axis represents the actual value of the feature, while the vertical axis shows its SHAP value, indicating both the direction and magnitude of the feature's contribution to academic risk: positive SHAP values suggest a positive effect on academic risk, and negative values indicate a negative one. The color of each point corresponds to another related feature, revealing potential interaction effects between variables. The first 11 important features are discussed with more details based on SHAP dependency plots as below.

- Learning support from learning peers (LpLrng)

Figure 3 illustrates how learning support from learning peers (LpLrng) influences its SHAP value for academic risk and how it interacts with Age at entrance (Age) of the student, represented in the color gradient. A clear non-linear, negative association is observed between LpLrng and its SHAP value for academic risk. When learning support from peers is relatively low (around −2 to −1), the SHAP values are high (approximately 0.1 to 0.3), indicating that lower values of this feature correlate with a higher predicted risk score. As LpLrng increases, its SHAP values approach zero and then become negative, suggesting that higher reported peer support aligns with a lower predicted risk. Additionally, the color gradient representing Age appears evenly distributed across the LpLrng range, indicating a very weak interaction between learning and Age. This suggests that Age does not appear to substantially modify the model-derived association between peer learning support and the predicted risk.

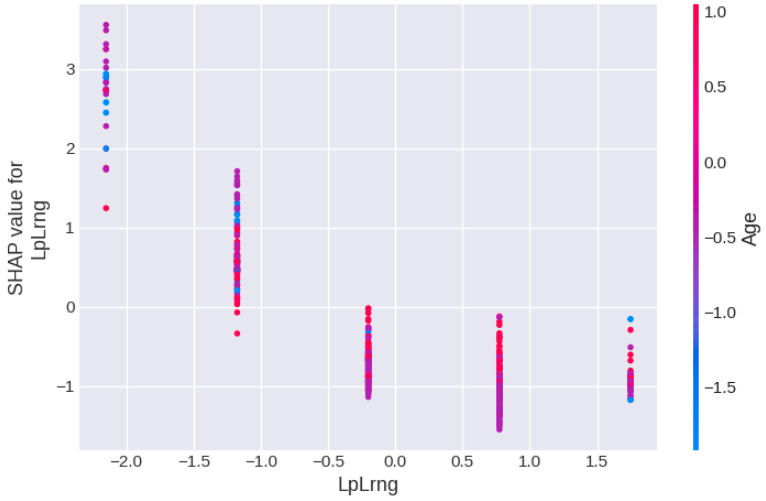


Figure 3. SHAP dependence plot for "Learning support from learning peers".

- Comprehensive subjects score of College entrance exam (ExamCmpr)

Figure 4 illustrates the relationship between Comprehensive subjects score (ExamCmpr) and its corresponding SHAP value for academic risk, as well as its potential interaction with Teacher–student relationship (TsRltn), represented by the color gradient. Overall, a predominantly negative association is observed between ExamCmpr and its SHAP value, though this association is not strictly linear. When the comprehensive score is relatively low (approximately −3 to −1), the SHAP values range from about 0 to 3 and show some fluctuation, indicating that lower scores in this range are linked to a higher predicted risk. Notably, across a middle range of scores (approximately −1 to 0.5), the SHAP values form a tight cluster (roughly −1.5 to 1.5), suggesting that for these scores, ExamCmpr shows a relatively stable and modest association with the model's prediction. As ExamCmpr

increases further into a higher range (from around 0.5 to 3), the SHAP values then exhibit a more consistent range (approximately 0 to −2). This pattern suggests that, according to the model, substantially higher scores are consistently associated with a lower predicted risk. Additionally, the color gradient representing TsRltn varies across the ExamCmpr range without a clear directional pattern, implying that TsRltn does not substantially modify the model-derived association between ExamCmpr and predicted academic risk.

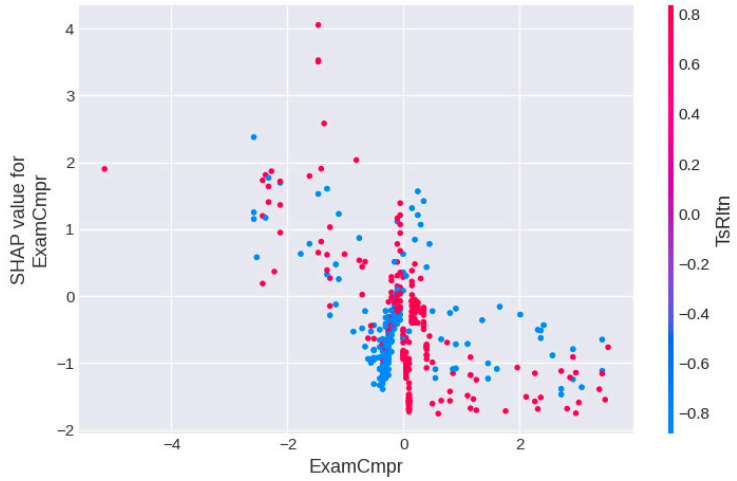


Figure 4. SHAP dependence plot for "ExamCmpr".

- Self-study (SlfStdy)

Figure 5 illustrates the association between Self-study (SlfStdy) and its SHAP value for academic risk, as well as its interaction with Foreign language score (ExamFrn), represented by the color gradient. When self-study is very low (around −1.9), the SHAP values are clearly positive (roughly −0.3 to 3), indicating that within the model, very low levels of self-study are associated with a higher predicted risk. As SlfStdy increases to moderately low levels (around −0.7 to 0.6), the SHAP values remain below zero, suggesting that moderate self-study levels correspond to a neutral or slightly lower predicted risk. When SlfStdy reaches to high levels (around 1.8), the SHAP values range from −0.5 to −1, which reflects that higher self-study reports align with a lower predicted risk. The color gradient representing ExamFrn is scattered across the entire SlfStdy range without a clear vertical separation, suggesting that ExamFrn does not substantially modify the basic associated pattern between self-study and predicted academic risk.

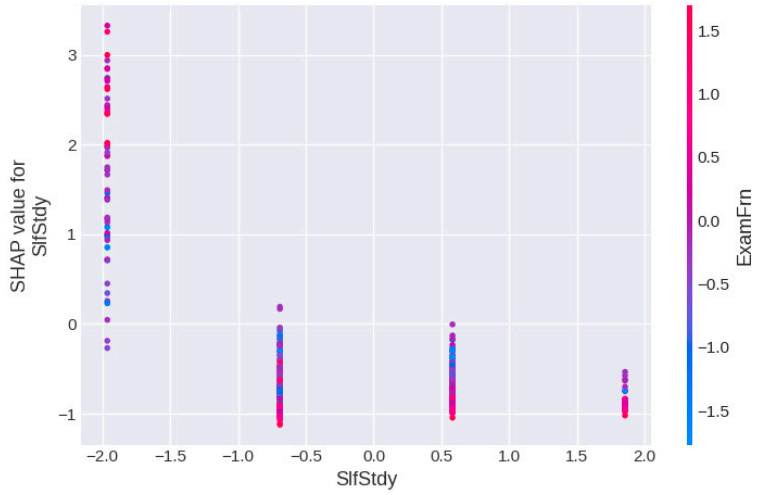


Figure 5. SHAP dependence plot for "Self-study".

- Truancy level (Truant)

Figure 6 illustrates the association between Truancy level (Truant) and its SHAP value for academic risk, as well as its interplay with Gender (Gndr), represented by the color gradient. Low-level or no truancy (around −1 to 0.5) shows SHAP values below zero, suggesting that regular class attendance is correlated with a lower predicted risk. In contrast, as truancy increases from approximately 1.5 to 3, the SHAP values are consistently positive from near 0.5 to 3, indicating that high level truancy is linked to a higher predicted risk. The color gradient representing Gender shows no systematic variation across the range of Truant values, implying that Gender does not substantially modify the model-derived relationship between truancy and predicted academic risk.

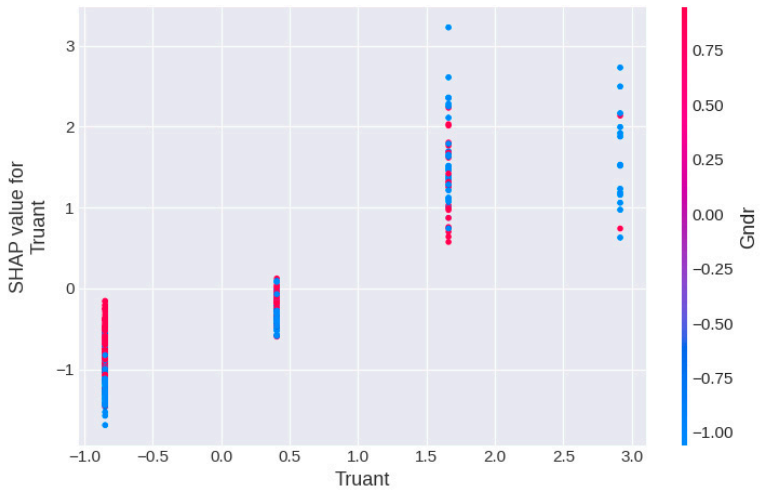


Figure 6. SHAP dependence plot for "Truancy level".

- Math score (ExamMth)

Figure 7 illustrates the association between Math score (ExamMth) and its SHAP value for academic risk, as well as its potential interaction with Dorm learning climate (DmClmt), represented by the color gradient. Overall, a predominantly negative trend is observed between ExamMth and its SHAP value, though the relationship is not strictly linear. When the Math score is relatively low (approximately −3 to −1), the SHAP values range from about −1 to 3, indicating that in this range, lower Math scores in this range are linked to a higher predicted risk. Notably, across a range of Math scores (approximately −1 to 4), the SHAP values form a tight cluster and some scatter plots (from −1 to 1). This pattern suggests that Math scores around or above the average show a weaker and more variable association with the predicted risk. Additionally, the color gradient representing DmClmt varies across the ExamMth range without a clear directional pattern, implying that DmClmt does not substantially modify the model-derived association between Math score and predicted academic risk.

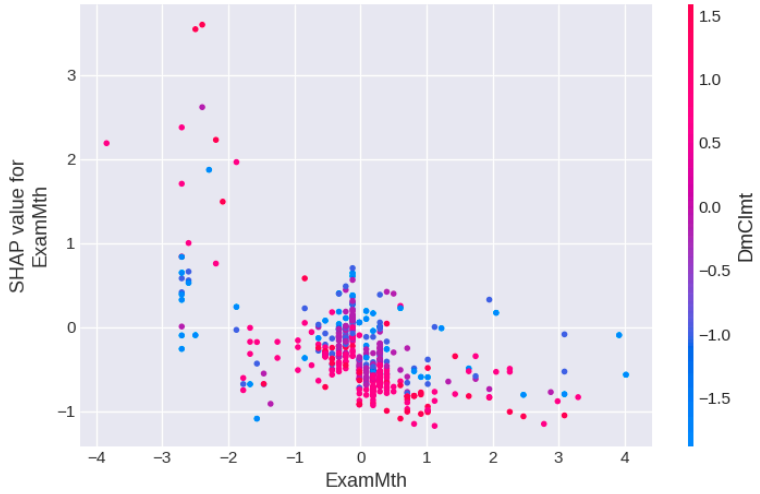


Figure 7. SHAP dependence plot for "Math score".

Because the remaining features were analyzed in the same way as the first five features, we do not present separate figures here and instead provide a brief summary.

At the overall level, features related to learning peers, including Stability of learning peers (LpStbl), Number of learning peers (LpNmb), Dorm learning climate (DmClmt), Chinese score (ExamCn) and Seating choice in a classroom (Seat), along with Social and emotional support from learning peers (LpScl), provide additional nuance to the model.

In general, higher stability of learning peers (LpStbl) is associated with a decrease in the SHAP value for predicted academic risk, indicating a negative association in the model, where higher peer stability aligns with a lower predicted risk. The number of learning peers (LpNmb) shows a non-linear relationship with its SHAP value for academic risk. Within the model, a greater number of peers is generally associated with decreased SHAP values, suggesting a link to lower predicted risk. Dormitory learning climate (DmClmt) exhibits a U-shaped pattern with its SHAP value. Both very low and very high levels are linked to higher SHAP values (and thus a higher predicted risk), while moderate levels correspond to lower or neutral SHAP values in the model's output. Chinese score (ExamCn) shows a negative trend with its SHAP value, where higher scores are associated with lower predicted risk. Seating choice (Seat) also displays a U-shaped relationship, with both low and high seating positions linked to a higher predicted risk compared to moderate positions in the model's predictions. Social and emotional support from learning peers (LpScl) shows a negative association, meaning that higher levels of support correspond to a lower predicted risk within the model. These features collectively contribute to the model's predictive structure by capturing nonlinear patterns and contextual interactions. Peer-related and environmental factors, in particular, show consistent and interpretable associations with predicted academic risk in this modeling framework.

At the overall level, through the SHAP framework, this study clearly elucidated the key predictive factors influencing SHAP value for academic risk in this specific situation. The importance ranking of these factors demonstrates a close integration of learning-peer support, learning activity, and College entrance exam scores.

4.2. Factors Predicting Academic Risk at the Individual Level

In this study, the second research question is, at the individual level, which factors, and how, predict students' academic risk within a specific background?

This section demonstrates how SHAP is used to interpret individual predictions, by examining the features associated with student academic risk within this specific model and illustrating how their values contribute to the model's output. The SHAP force plot

visualizes the contribution of each feature to the predicted risk for a given student. Features with red arrows are associated with an increase in the predicted risk value, whereas features with blue arrows are associated with a decrease, with the length of each bar representing the magnitude of this directional contribution. E[f(X)] denotes the average model prediction across all students in the sample (the baseline), which is −1.961 in this study, and f(x) indicates the specific model output for an individual student, representing the predicted log-odds (or probability) of academic risk. Each label on the left specifies the feature and its value for that student; for example, "LpLrng = 2" indicates that the student's learning support from peers corresponds to a value of 2.

1. Student with high potential academic risk

In Figure 8, Student #46 exhibits significant academic risk across multiple dimensions. The value of f(x) for this student is 3.972, indicating a substantially higher academic risk than the overall mean of −1.961. Figure 9 reveals how this risk level arises from an accumulation of learning behavior deficits, a weak learning-support ecology and relatively average College entrance exam scores.

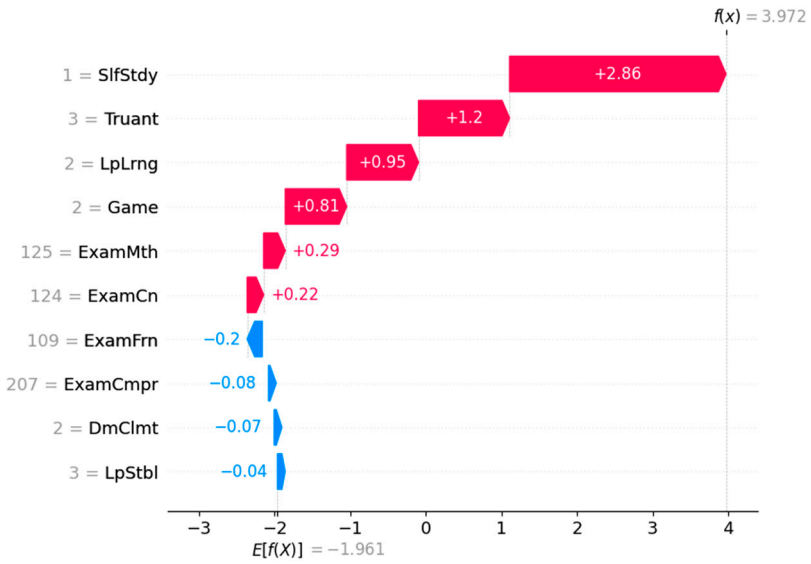


Figure 8. SHAP marginal contribution of student #46 with potential risk.

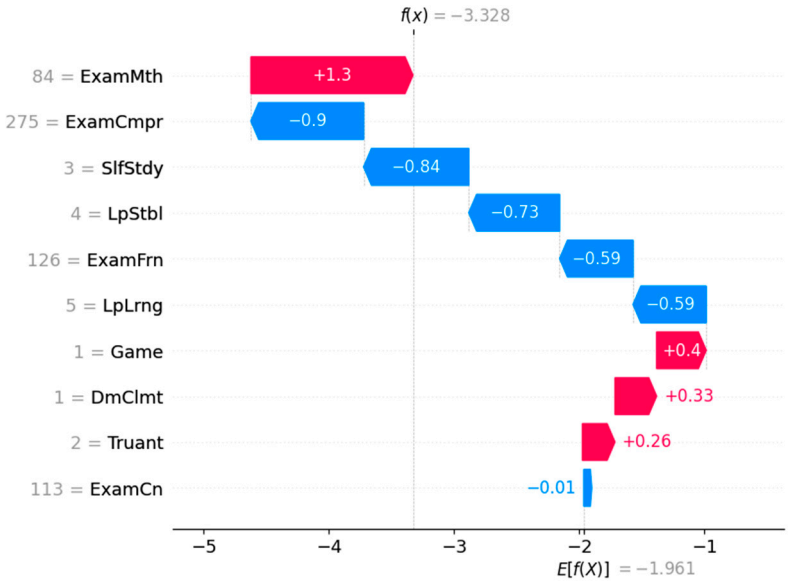


Figure 9. SHAP marginal contribution of student #466 with no potential risk.

The most significant contributors are two learning behaviors. The student reported never engaging in self-study (SlfStdy = 1), which alone contributed +2.86 to the predicted SHAP value for academic risk. In addition, the student indicated occasional truancy (Truant = 3), adding another +1.2. These two factors are strongly associated with higher risk, which aligns with the prior literature that identifies poor learning management and disengagement as correlates of academic risk. A second major feature is associated with the student's informal learning network. The student received low learning support from learning peers (LpLrng = 2), contributing +0.95. Consistent with sociocultural and peer-learning theories, insufficient academic and strategic support from learning peers reduces access to shared resources, emotional reinforcement, and productive study norms, which are associated with academic risk. The student also reported seldom playing video games (Game = 2), which contributed +0.81, indicating such a behavior was not associated with a protective effect on the predicted risk score.

Interestingly, although the student achieved relatively average scores in the College entrance exam, the contributions to SHAP value for academic risk were neutralized. This suggests that academic foundations alone were not sufficient to offset the risk signals associated with deficits in daily learning behaviors and perceived peer support.

2. Student with low potential academic risk

In Figure 9, Student #466 demonstrates a notably low overall predicted academic risk, with a predicted academic risk of -3.328, far below the cohort average of -1.961. The SHAP waterfall plot shows that this low-risk prediction is associated with a combination of features: a high Comprehensive subjects score, certain learning behaviors, and aspects of the peer environment.

Regarding academic foundation variables, the positive contribution of SHAP value (+1.3) comes from the student's low Math score (ExamMth = 84), meaning that within the model's logic, this particular score is linked to a higher predicted risk. Conversely, the student shows a strong composite exam performance (ExamCmpr = 275), contributing -0.9 and indicating solid academic preparation was associated with low academic risk. Other academic features include a high Foreign language score (ExamFrn = 126, SHAP: -0.59) and an average Chinese score (ExamCn = 113, SHAP: -0.01). Collectively, within the model, higher entrance exam scores tend to be associated with lower predicted risk. In terms of learning behaviors, the student sometimes engages in self-study (SlfStdy = 3), which reduces risk by -0.84. Although not at the highest level, this behavior is sufficient to maintain a stable academic routine. The student indicated occasional truancy (Truant = 2) adding another +0.33, which was associated with the potential risk. Complementing this, the stability of learning peers (LpStbl = 4, SHAP = –0.73) suggest consistent, reliable partnerships, which was associated with risk reduction. Similarly, another protective factor is the student's very high level of learning support from learning peers (LpLrng = 5), which was associated with risk reduction by –0.59. But, lower Dorm learning climate (DmClmt = 1) may be associated with high academic risk. According to peer-learning theories, high peer support provides richer academic resources, emotional reassurance, strategy sharing, and timely help, all contributing to a favorable learning ecology.

Taken together, Student #466 exemplifies a low-risk profile shaped by strong learning-peer support, adequate self-learning, stable peer networks, and solid academic foundations.

Overall, these two typical cases demonstrate how learning behaviors and learning peer ecology dominate the interpretive pathway for predicting academic risk. It supports our argument that XAI can uncover the behavioral and ecological mechanisms behind individual risk formation, offering actionable insights for targeted interventions.

5. Discussion

5.1. The Influence of Learning Peers on Academic Risk

One of the key findings of this study is that features related to learning peers show the strongest predictive association with academic risk in our model. In this study, learning peers refer to individuals who participate in daily academic activities outside of class. The results indicate that in our predictive model, peer support, especially learning support, emerges as the most influential feature for academic risk and consistently emerges as the top predictor. At the overall level, learning support from peers shows a stable negative association with SHAP value for academic risk and remains at the top of the feature importance in our final model. At the individual level, SHAP explanations for high-risk students repeatedly indicate that very low learning support from peers is the primary driver of risk alerts.

In addition, stability and number of learning peers, the dormitory learning climate and social–emotional support from peers shift the focus from isolated individual behaviors or background information, such as self-study or entrance exam scores, to the informal social learning environments. Stable learning peer relationships are conducive to better learning performance, and a larger number of learning peers facilitates the construction of a broader support network. This deepens the application of social learning perspectives in higher education settings. According to sociocultural theory, high-quality peers typically possess strong academic competence, positive learning attitudes, and the capacity to provide social–emotional support [49,50]. These characteristics are theorized to motivate students and help them cope with academic challenges, which could potentially reduce academic risk [51]. Our model's findings are consistent with this theoretical perspective, showing a strong negative association.

Taken together, learning support, peer stability and size, social–emotional support and a favorable dorm learning climate collectively form a "protective peer–environment factor cluster", validated across different analyses as critical variables that help reduce risk. Conversely, Truancy and Self-study constitute a "risky behavioral factor cluster". High-frequency truancy and weak self-study input are the main behavioral mechanisms that are strongly associated with higher academic risk scores. In other words, peer support and individual learning behavior are not mutually exclusive; instead, a "high-quality peer network plus good study habits" jointly construct a dual defense line for students against academic risk. At the same time, College entrance exam scores also show a non-negligible influence in the model. The impact of Comprehensive subjects scores and math scores on SHAP of academic risk ranks among the highest, indicating that academic foundations remain an important component in predicting risk.

It should be noted that the peer-related variables (e.g., learning support, social–emotional support, stability) are conceptually linked and likely correlated in measurement. While the machine learning models employed can handle correlated features, and SHAP values aim to fairly distribute contributions among them, this intercorrelation suggests that these features collectively reflect a broader, latent construct of "peer support ecology". Therefore, the identification of "learning support" as the strongest individual predictor should be interpreted as it carries the most unique and predictive information among this set of related features within our model. This finding nonetheless strongly underscores the paramount importance of the peer support dimension as a whole in the academic risk landscape.

However, the existing research has paid relatively limited attention to out-of-class learning peers [37]. One possible reason is that prior work has primarily focused on task-oriented collaborative learning within the classroom, while out-of-class learning peers tend to be overlooked because they are not directly tied to in-class academic tasks [52]. Moreover, in university settings, students typically transition from the collective learning

model of upper secondary school to a more independent learning model [53,54]. This may lead educational stakeholders to underestimate the importance of out-of-class learning peers. Our findings challenge this oversight by demonstrating that peer support and individual learning behaviors are not mutually exclusive. Instead, a high-quality peer network, when coupled with strong study habits, jointly establishes a dual defense line against academic risk.

From the perspectives of contemporary work on institutional integration and student persistence [55,56], our findings suggest academic and social integration are not confined to formal classrooms and course-based groups. Informal out-of-class peer networks also constitute an important support structure for academic persistence and performance. Seemingly trivial everyday interactions among peers can have a profound impact on students' academic outcomes. Through daily interaction, learning peers provide emotional support, share information, and exchange learning strategies, all of which play a key role in academic success.

5.2. The Role of XAI in EDM

As noted, although EDM has made significant strides in improving the accuracy of academic risk prediction, model explainability is equally vital alongside accuracy. By introducing explainable techniques, this study successfully resolves the "black-box problem" inherent in traditional machine learning models, providing trustworthy, transparent, and fair attribution for prediction results, which significantly enhances the fairness and acceptability of the early warning outcomes.

In educational practice, the value of XAI is realized through providing both local explanations for individual students and global feature importance for the overall sample [57]. This dual-level explanation allows institutions to intervene precisely at the individual counseling level while simultaneously optimizing resource allocation at the institutional and policy level [58]. For instance, the SHAP method offers personalized, actionable explanations for high-risk students, precisely identifying the primary risk drivers, whether it be critically low learning support or severe Truancy behavior, enabling counselors to design targeted interventions informed by the model's identified associations. Furthermore, based on feature importance ranking, administrators can effectively allocate resources towards high-leverage intervention points such as enhancing peer support, monitoring truancy, guiding self-study, and improving the dorm learning climate. Finally, the identification of these critical features provides a new, more insightful direction for the design of indicators in Management Information Systems (MIS) for data collection and feature engineering.

Notably, explainable models also provide a technical foundation for fairness auditing, since background variables such as demographic attributes and prior achievement can be examined to determine whether a model relies disproportionately on potentially sensitive features [59]. This capability supports the detection and mitigation of systemic bias, reducing the risk that prediction results unfairly disadvantage particular gender or background groups and contributing to more trustworthy, data-driven early warning systems and educational decision-making [60].

5.3. Exploring High-Dimensional and Small Datasets in Education

Traditional academic risk studies often rely on large-scale datasets, aiming for generalizability and drawing conclusions from broad, cross-contextual samples [61]. While large sample analysis can achieve high statistical significance, it often overlooks the micro-level variances and complexities within smaller, specific cohorts [62]. Accordingly, this study utilized a high-dimensional and small dataset to analyze complex and subtle relationships between students' individual characteristics and their learning environment. In contrast

to large studies that pursue cross-institutional, cross-major universal conclusions, this research deliberately shifted its focus to the contextualized level of a specific institution and major, prioritizing the achievement of high-accuracy, explainable risk identification within the local context to support context-sensitive intervention decisions.

Employing a high-dimensional and small dataset, this study focused on academic risk prediction within a specific educational context, analyzing and explaining a series of factors affecting academic risk at both the group and individual levels. By focusing on small datasets, this study offers an innovative research methodology for academic risk prediction. Results demonstrate this methodology not only captures hidden individualized features within small datasets but also effectively identifies micro-interaction effects that might be averaged out or diluted in large sample analysis, avoiding issues like feature dilution and information loss. Furthermore, the implementation adopted methods like cross-validation and parameter regularization to control the risk of overfitting, proving that under data-sparse conditions, combining machine learning and XAI can still yield relatively robust and stable prediction and explanation results [63,64].

Therefore, this XAI-driven EDM approach is thus proven to offer robust solutions even in the data-sparse context of a specific major, providing a useful reference for related research.

5.4. Considerations of Context and Generalizability

While the insights derived from our XAI framework are robust within the studied cohort, it is crucial to situate them within their specific cultural and institutional context to properly assess their generalizability.

Cultural context: The strength of the association between peer support and academic risk in our model may be related to broader collectivist cultural norms prevalent in Chinese educational settings, where collaborative learning and group harmony are often emphasized. The concept of learning peers in our study aligns with cultural values that favor interdependence and collective academic striving. In more individualistic educational cultures, the magnitude of this effect, or the very nature of peer support, might differ.

Institutional context: The university's specific practice of informally encouraging semester-based peer learning groups provided a structural backdrop that made peer interactions a salient and measurable feature of student life. Similarly, the dormitory-based learning climate is a factor deeply embedded in the residential campus model common in Chinese universities, where students spend extensive time in assigned dormitories. In commuter schools or institutions with different housing policies, this environmental factor may play a negligible role.

Therefore, the external validity of our predictive model is inherently constrained. We do not claim that the exact same feature importance ranking (e.g., LpLrng > Truant) would replicate universally. Instead, the primary contribution of this work is methodological and demonstrative: (1) it validates an XAI framework capable of extracting contextually meaningful and interpretable insights from a high-dimensional, imbalanced dataset drawn from a specific, bounded educational setting. (2) It provides strong, data-driven evidence that informal peer dynamics and the learning environment (factors often overlooked in traditional institutional data systems) can be critical predictors of academic risk, a principle likely to hold relevance across contexts even if the specific operationalization varies.

Future research should aim to test the transferability of these insights through cross-institutional comparisons or multi-cohort studies. Applying the same XAI framework in different cultural and institutional settings will help disentangle universal patterns of academic risk from those that are context-specific.

6. Conclusions and Future Plan

Effective prediction can help identify students at risk of academic failure early, providing timely support for educational interventions. However, the explainability of data and conclusions has remained a significant challenge, making research into academic influencing factors less detailed and in-depth. Additionally, the existing research has largely overlooked high-dimensional small sample data. To address these limitations, this study explored solutions from three dimensions: methodology, theory, and practice. At the methodological level, we proposed an "ML + SHAP" framework for academic risk prediction, suitable for high-dimensional small datasets, achieving both overall and individual-level explainable warnings. At the theoretical level, we highlighted the protective role of out-of-class peer support and the dorm learning climate as informal learning environments, expanding the traditional understanding of peer and environmental factors in academic risk research. At the practical level, the study provided specific, actionable variables and interpretation pathways for institutional academic warning systems, offering a foundation for precise intervention.

At the same time, this study has several limitations. Future work will therefore proceed along three main directions. First, more rigorous causal inference methods are needed to unpack the mechanisms through which peer support and environmental factors influence academic risk and to design and experimentally evaluate interventions that enhance learning support from peers and prevent truancy. In this context, we will explore Fuzzy Cognitive Maps (FCMs) as a complementary causal modeling paradigm that offers interpretability and causal transparency through an explicit, directed concept–relation structure, thereby compensating for limited data through expert assessments while addressing the key limitations of post hoc explanation methods [65]. FCM (including learning-based FCM variants) should also be applied to predict student performance, as they have been proven to be feasible in educational prediction tasks where data may be scarce and domain knowledge is valuable [66]. Second, to address issues of external validity and generalizability, we plan to conduct cross-context validation and transfer learning with high-dimensional small sample datasets from different regions and disciplines. This will enable a more comprehensive understanding of how various factors jointly shape academic risk. Finally, we aim to integrate explainable predictive models into educational decision-making processes and embed Explainable Artificial Intelligence techniques within a broader educational intervention framework to achieve early identification of academic risk, precise intervention, and continuous monitoring.

These further explorations have the potential to promote more scientific and fine-grained management in education and to provide stronger support for students' holistic development.

Author Contributions: Conceptualization, D.S. and P.Z.; methodology, D.S. and P.X.; formal analysis, D.S., P.X. and G.C.; investigation, P.Z.; writing—original draft preparation, D.S. and P.X.; writing—review and editing, D.S., G.C. and P.Z.; funding acquisition, D.S. All authors have read and agreed to the published version of the manuscript.

Funding: This study was funded by INNOVATION FUNDING PROJECT OF ENGINEERING RESEARCH CENTER OF INTEGRATION AND APPLICATION OF DIGITAL LEARNING TECHNOLOGY OF MINISTRY OF EDUCATION OF CHINA, grant number 1331004.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and approved by the Institutional Review Board of the Biological and Medical Ethics Committee, Dalian University of Technology (DUTGSE231020-01, 20 October 2023) for studies involving humans.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The data is available from ScienceDB with DOI: 10.57760/sciencedb.32184.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. OECD. *Education at a Glance 2022: OECD Indicators*; OECD: Paris, France, 2022.
2. Spight, D.B. Undeclared versus declared: Who is more likely to graduate? *J. Coll. Stud. Retent. Res. Theory Pract.* **2022**, *23*, 945–964. [CrossRef]
3. Abu Saa, A.; Al-Emran, M.; Shaalan, K. Factors affecting students' performance in higher education: A systematic review of predictive data mining techniques. *Technol. Knowl. Learn.* **2019**, *24*, 567–598. [CrossRef]
4. Romero, C.; Ventura, S. Educational data mining and learning analytics: An updated survey. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2020**, *10*, e1355. [CrossRef]
5. Mi, J.-X.; Li, A.-D.; Zhou, L.-F. Review study of interpretation methods for future interpretable machine learning. *IEEE Access* **2020**, *8*, 191969–191985. [CrossRef]
6. Roslan, M.B.; Chen, C. Educational data mining for student performance prediction: A systematic literature review (2015–2021). *Int. J. Emerg. Technol. Learn. (iJET)* **2022**, *17*, 147–179. [CrossRef]
7. Albreiki, B.; Zaki, N.; Alashwal, H. A systematic literature review of student'performance prediction using machine learning techniques. *Educ. Sci.* **2021**, *11*, 552. [CrossRef]
8. Barredo Arrieta, A.; Díaz-Rodríguez, N.; Del Ser, J.; Bennetot, A.; Tabik, S.; Barbado, A.; Garcia, S.; Gil-Lopez, S.; Molina, D.; Benjamins, R.; et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf. Fusion* **2020**, *58*, 82–115. [CrossRef]
9. Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* **2019**, *1*, 206–215. [CrossRef]
10. Fiok, K.; Farahani, F.V.; Karwowski, W.; Ahram, T. Explainable artificial intelligence for education and training. *J. Def. Model. Simul.* **2022**, *19*, 133–144. [CrossRef]
11. Swamy, V.; Frej, J.; Käser, T. The future of human-centric eXplainable Artificial Intelligence (XAI) is not post-hoc explanations. *J. Artif. Intell. Res.* **2025**, *84*, 2–7. [CrossRef]
12. Ashfaq, U.; Booma, P.; Mafas, R. Managing student performance: A predictive analytics using imbalanced data. *Int. J. Recent Technol. Eng.* **2020**, *8*, 2277–2283. [CrossRef]
13. Zhang, Y.; Yun, Y.; An, R.; Cui, J.; Dai, H.; Shang, X. Educational data mining techniques for student performance prediction: Method review and comparison analysis. *Front. Psychol.* **2021**, *12*, 698490. [CrossRef]
14. Nguyen, N.B.C.; Karunaratne, T. Learning analytics with small datasets—State of the art and beyond. *Educ. Sci.* **2024**, *14*, 608. [CrossRef]
15. Fonteyne, L.; Duyck, W.; Fruyt, F.D.J.L.; Differences, I. Program-specific prediction of academic achievement on the basis of cognitive and non-cognitive factors. *Learn. Individ. Differ.* **2017**, *56*, 34–48. [CrossRef]
16. Zimmerman, B.J.; Kitsantas, A.J.C.E.P. Comparing students' self-discipline and self-regulation measures and their prediction of academic achievement. *Contemp. Educ. Psychol.* **2014**, *39*, 145–155. [CrossRef]
17. Troll, E.S.; Friese, M.; Loschelder, D.D. How students' self-control and smartphone-use explain their academic performance. *Comput. Hum. Behav.* **2021**, *117*, 106624. [CrossRef]
18. Alyahyan, E.; Düştegör, D. Predicting academic success in higher education: Literature review and best practices. *Int. J. Educ. Technol. High. Educ.* **2020**, *17*, 3. [CrossRef]
19. Goh, E.; Kim, H.J. Emotional intelligence as a predictor of academic performance in hospitality higher education. *J. Hosp. Tour. Educ.* **2021**, *33*, 140–146.
20. Khan, A.; Ghosh, S.K. Student performance analysis and prediction in classroom learning: A review of educational data mining studies. *Educ. Inf. Technol.* **2021**, *26*, 205–240. [CrossRef]
21. Li, X.; Zhu, X.; Zhu, X.; Ji, Y.; Tang, X. Student Academic Performance Prediction Using Deep Multi-source Behavior Sequential Network. In *Advances in Knowledge Discovery and Data Mining*; Lauw, H., Wong, R., Ntoulas, A., Lim, E., Ng, S., Pan, S., Eds.; Springer International Publishing: Cham, Switzerland, 2020; Volume 12084, pp. 567–579.
22. Sokkhey, P.; Okazaki, T. Development and optimization of deep belief networks applied for academic performance prediction with larger datasets. *IEIE Trans. Smart Process. Comput.* **2020**, *9*, 298–311. [CrossRef]
23. Ramanathan, K.; Thangavel, B. Minkowski Sommon feature map-based densely connected deep convolution network with LSTM for academic performance prediction. *Concurr. Comput. Pract. Exp.* **2021**, *33*, e6244. [CrossRef]
24. Pallathadka, H.; Wenda, A.; Ramirez-Asís, E.; Asís-López, M.; Flores-Albornoz, J.; Phasinam, K. Classification and prediction of student performance data using various machine learning algorithms. *Mater. Today Proc.* **2023**, *80*, 3782–3785. [CrossRef]

25. Arashpour, M.; Golafshani, E.M.; Parthiban, R.; Lamborn, J.; Kashani, A.; Li, H.; Farzanehfar, P. Predicting individual learning performance using machine-learning hybridized with the teaching-learning-based optimization. *Comput. Appl. Eng. Educ.* **2023**, *31*, 83–99. [CrossRef]
26. Casillano, N.F.B.; Cantilang, K.W. Employing educational data mining techniques to predict programming students at-risk of dropping out. *Indones. J. Electr. Eng. Comput. Sci.* **2024**, *35*, 1219–1226. [CrossRef]
27. Malik, S.; Patro, S.G.K.; Mahanty, C.; Hegde, R.; Naveed, Q.N.; Lasisi, A.; Buradi, A.; Emma, A.F.; Kraiem, N. Advancing educational data mining for enhanced student performance prediction: A fusion of feature selection algorithms and classification techniques with dynamic feature ensemble evolution. *Sci. Rep.* **2025**, *15*, 8738. [CrossRef]
28. Miller, T. Explanation in artificial intelligence: Insights from the social sciences. *Artif. Intell.* **2019**, *267*, 1–38. [CrossRef]
29. Molnar, C. *Interpretable Machine Learning—A Guide for Making Black Box Models Explainable*; Leanpub: Victoria, BC, Canada, 2019.
30. Ribeiro, M.T.; Singh, S.; Guestrin, C. "Why should I trust You?": Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference, San Francisco, CA, USA, 13–17 August 2016; pp. 1135–1144.
31. Aljohani, O. A comprehensive review of the major studies and theoretical models of student retention in higher education. *High. Educ. Stud.* **2016**, *6*, 1–18. [CrossRef]
32. Vygotsky, L.S. *Mind in Society: The Development of Higher Psychological Processes*; Harvard University Press: Cambridge, MA, USA, 1978; Volume 86.
33. Bandura, A. Social cognitive theory of moral thought and action. In *Handbook of Moral Behavior and Development*; Psychology Press: East Sussex, UK, 2014; pp. 45–103.
34. Rienties, B.; Nolan, E.-M. Understanding friendship and learning networks of international and host students using longitudinal Social Network Analysis. *Int. J. Intercult. Relat.* **2014**, *41*, 165–180. [CrossRef]
35. Ahn, M.Y.; Davis, H.H. Four domains of students' sense of belonging to university. *Stud. High. Educ.* **2020**, *45*, 622–634. [CrossRef]
36. Tang, Y.M.; Lau, Y.-y.; Chau, K.Y. Towards a sustainable online peer learning model based on student's perspectives. *Educ. Inf. Technol.* **2022**, *27*, 12449–12468. [CrossRef]
37. Woodward, R.; Pattinson, N. Informal Peer Learning of Diverse Undergraduate Students: Some Learners Make Meaning through Collaborative Activity. *Pract. Res. High. Educ.* **2023**, *15*, 72–85.
38. Lainio, A. Independent learner as the ideal—Normative representations of higher education students in film and television drama across Europe. *Crit. Stud. Educ.* **2024**, *65*, 39–56. [CrossRef]
39. Leathwood, C. Gender, equity and the discourse of the independent learner in higher education. *High. Educ.* **2006**, *52*, 611–633. [CrossRef]
40. Geister, S.; Keser Aschenberger, F.; Çetinkaya-Yıldız, E.; Apaydın, S. The role of informal learning spaces in promoting social integration and wellbeing in higher education. *Front. Educ.* **2025**, *10*, 1637874. [CrossRef]
41. Carvalho, D.V.; Pereira, E.M.; Cardoso, J.S. Machine learning interpretability: A survey on methods and metrics. *Electronics* **2019**, *8*, 832. [CrossRef]
42. Lundberg, S.; Lee, S.-I. A unified approach to interpreting model predictions. *arXiv* **2017**, arXiv:1705.07874. [CrossRef]
43. Shapley, L.S. A value for n-person games. *Contrib. Theory Games* **1953**, *2*, 307–317.
44. Pekrun, R. The control-value theory of achievement emotions: Assumptions, corollaries, and implications for educational research and practice. *Educ. Psychol. Rev.* **2006**, *18*, 315–341. [CrossRef]
45. Marbouti, F.; Diefes-Dux, H.A.; Madhavan, K. Models for early prediction of at-risk students in a course using standards-based grading. *Comput. Educ.* **2016**, *103*, 1–15. [CrossRef]
46. Saito, T.; Rehmsmeier, M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS ONE* **2015**, *10*, e0118432. [CrossRef]
47. Tharwat, A. Classification assessment methods. *Appl. Comput. Inform.* **2021**, *17*, 168–192. [CrossRef]
48. Brodersen, K.H.; Ong, C.S.; Stephan, K.E.; Buhmann, J.M. The balanced accuracy and its posterior distribution. In Proceedings of the 20th International Conference on Pattern Recognition (ICPR 2010), Istanbul, Turkey, 23–26 August 2010; pp. 3121–3124.
49. Worley, J.T.; Meter, D.J.; Ramirez Hall, A.; Nishina, A.; Medina, M.A. Prospective associations between peer support, academic competence, and anxiety in college students. *Soc. Psychol. Educ.* **2023**, *26*, 1017–1035. [CrossRef]
50. Chen, C.; Bian, F.; Zhu, Y. The relationship between social support and academic engagement among university students: The chain mediating effects of life satisfaction and academic motivation. *BMC Public Health* **2023**, *23*, 2368. [CrossRef]
51. Zhu, Y.; Lu, H.; Wang, X.; Ma, W.; Xu, M. The relationship between perceived peer support and academic adjustment among higher vocational college students: The chain mediating effects of academic hope and professional identity. *Front. Psychol.* **2025**, *16*, 1534883. [CrossRef]
52. De Carvalho, F.C.; Geschwind, L.; Weurlander, M.; Mendonça, M. Possibilities and challenges of out-of-class interactions in the Mozambican academic context. *Cogent Educ.* **2025**, *12*, 2441057. [CrossRef]
53. Thompson, M.; Pawson, C.; Evans, B. Navigating entry into higher education: The transition to independent learning and living. *J. Furth. High. Educ.* **2021**, *45*, 1398–1410. [CrossRef]

54. Chilvers, L. The Peer-to-Peer Model: A UK Institution's Approach to Broadening and Embedding the Provision of Peer Learning and Support. *J. Peer Learn.* **2025**, *16*, 1–15. [CrossRef]
55. Cabir Hakyemez, T.; Mardikyan, S. The interplay between institutional integration and self-efficacy in the academic performance of first-year university students: A multigroup approach. *Int. J. Manag. Educ.* **2021**, *19*, 100430. [CrossRef]
56. Samoila, M.E.; Vrabie, T. First-year seminars through the lens of Vincent Tinto's theories of student departure. A systematic review. *Front. Educ.* **2023**, *8*, 1205667. [CrossRef]
57. Lundberg, S.M.; Erion, G.; Chen, H.; DeGrave, A.; Prutkin, J.M.; Nair, B.; Katz, R.; Himmelfarb, J.; Bansal, N.; Lee, S.-I. From local explanations to global understanding with explainable AI for trees. *Nat. Mach. Intell.* **2020**, *2*, 56–67. [CrossRef]
58. Mustofa, S.; Emon, Y.R.; Mamun, S.B.; Akhy, S.A.; Ahad, M.T. A novel AI-driven model for student dropout risk analysis with explainable AI insights. *Comput. Educ. Artif. Intell.* **2025**, *8*, 100352. [CrossRef]
59. Choi, W.-C.; Lam, C.-T.; Pang, P.C.-I.; Mendes, A.J. A Systematic Literature Review of Explainable Artificial Intelligence (XAI) for Interpreting Student Performance Prediction in Computer Science and STEM Education. In Proceedings of the 30th ACM Conference on Innovation and Technology in Computer Science Education, Nijmegen, The Netherlands, 30 June–2 July 2025; Volume 1, pp. 221–227.
60. Sanfo, J.-B.M.B. Application of explainable artificial intelligence approach to predict student learning outcomes. *J. Comput. Soc. Sci.* **2024**, *8*, 9. [CrossRef]
61. Lin, L.; Zhou, D.; Wang, J.; Wang, Y. A systematic review of big data driven education evaluation. *Sage Open* **2024**, *14*, 21582440241242180. [CrossRef]
62. Tang, Y.; Harvey, E.; Yao, C.; Yu, R.; Kizilcec, R.F.; Brooks, C. Understanding Predictive Models of Student Success with a Multiverse Analysis. In Proceedings of the 18th International Conference on Educational Data Mining, Palermo, Italy, 20–23 July 2025; pp. 518–525.
63. Islam, M.M.; Sojib, F.H.; Mihad, M.F.H.; Hasan, M.; Rahman, M. The integration of explainable ai in educational data mining for student academic performance prediction and support system. *Telemat. Inform. Rep.* **2025**, *18*, 100203. [CrossRef]
64. Allgaier, J.; Pryss, R. Cross-validation visualized: A narrative guide to advanced methods. *Mach. Learn. Knowl. Extr.* **2024**, *6*, 1378–1388. [CrossRef]
65. Tyrovolas, M.; Nápoles, G.; Stylios, C. Backpropagation-Based Counterfactual Explanations for Quasi-Nonlinear Fuzzy Cognitive Maps. *IEEE Trans. Syst. Man Cybern. Syst.* **2026**, 1–15. [CrossRef]
66. Mansouri, T.; ZareRavasan, A.; Ashrafi, A. A learning fuzzy cognitive map (LFCM) approach to predict student performance. *J. Inf. Technol. Educ. Res.* **2021**, *20*, 221–243. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.