International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18 (2026) pp. 1-9

https://doi.org/10.70917/ijcisim-2026-2820
Article

Predicting Student Academic Performance Based on Multi-source Heterogeneous Features and Explainable Ensemble Learning

Wang Li [1,*]

[1] School of Financial Management, Gingko College of Hospitality Management, Chengdu, Sisuan, 611730, China
* Correspondence author: 459356481@qq.com

Abstract: Student academic performance prediction is a key task in learning analysis and educational data mining, aiming to achieve early risk identification, precise intervention, and teaching decision support through multi-source educational data. Addressing the issues of insufficient validation of public data and real-world scenarios, inadequate evaluation protocols, and weak stability of interpretation results in existing research, this paper proposes a multi-source heterogeneous feature modeling and interpretable ensemble learning framework for educational intervention. This framework integrates demographic features, previous academic foundation, behavioral participation, psychological state, and teaching context information, and systematically compares linear models, random forests, boosting models, and fusion models in a unified leak-proof modeling pipeline. The model performance is evaluated through repeated cross-validation, ablation experiments, missing robustness analysis, and statistical significance tests. At the interpretation level, this paper combines permutation importance and SHAP to analyze key influencing factors from the global, local, and stability perspectives. Experimental results show that the boosting models and fusion models overall outperform the linear baseline in terms of accuracy and robustness; previous academic performance, attendance participation, homework completion rate, and behavioral investment are the most stable core predictors, while psychological stress and teaching quality variables provide important supplementary explanations. This paper has formed a reproducible, scalable, and education-intervention-oriented academic performance prediction technology route, providing methodological basis for university academic warning and personalized support.

Keywords: student performance prediction, learning analytics, educational data mining, ensemble learning, gradient boosting, permutation importance, interpretability.

1. Aims and Background

With the deep digitalization of university teaching platforms, class attendance systems, homework and assessment platforms, as well as student affairs systems, the educational process is continuously generating high-dimensional, multi-source and heterogeneous learning data [1]. How to extract signals from these data that have explanatory and predictive power for learning outcomes has become a common core issue in learning analysis, educational data mining, and artificial intelligence education applications. Different from traditional empirical judgments, data-driven academic prediction not only can improve the timeliness of risk identification, but also can provide more granular evidence support for course support, academic counseling and resource allocation [2-4]. Therefore, building an accurate, stable and interpretable student academic performance prediction model has clear theoretical value and practical significance.

However, the mechanism by which students' academic performance is formed is essentially a multi-factor coupling process. It is influenced not only by students' previous knowledge foundation, learning

commitment and behavioral habits, but also by factors such as course organization, teacher teaching quality, class environment and psychological state [5-7]. Existing studies have generally shown that educational data not only have significant nonlinear relationships and interaction effects, but also often encounter typical problems such as missing values, high proportion of categorical variables, sample imbalance and drift in cross-term distribution [8]. In this context, relying solely on a single linear model is often insufficient to fully capture the complex structure. Moreover, without strict validation protocols and transparent explanations, the model results may not truly support high-risk educational decisions [9-11].

Regarding tabular educational data, the most effective approaches in recent years have typically focused on three routes: linear regularization models, tree ensemble models, and tabular deep models. Linear regularization models have the advantages of stability, interpretability, and robustness against multicollinearity, and can serve as a benchmark reference; tree ensemble models, especially boosting methods, can better model nonlinear features and high-order interactions, and have long maintained strong competitiveness in medium and small-scale tabular tasks [12]; while deep models specifically designed for tabular data provide new possibilities for uniformly handling complex feature relationships. Considering that different modeling paradigms have different inductive biases in educational scenarios, the superiority of a single model often depends on specific data and validation protocols [13-14]. Therefore, adopting a multi-model comparison and fusion strategy is more in line with the robust research paradigm [15-19].

Therefore, this study regards interpretability as an equally important research objective as accuracy. On one hand, a global importance method based on performance degradation is adopted to depict the degree of model dependence on each feature [20]; on the other hand, additive explanation methods and typical case analysis are utilized to reveal the predictive driving factors of specific student samples. At the same time, considering that different high-performance models may provide different importance rankings, this paper further introduces cross-model consistency and group stability tests to avoid mistakenly writing the importance ranking of a single model as a universal conclusion [21-25].

Based on the above problem awareness, the research objectives of this paper can be summarized as follows: Firstly, to establish a multi-source heterogeneous feature system covering students' background, past academic performance, behavioral participation, psychological state and teaching context; Secondly, to develop a leakage-proof, repeatable, and scenario-oriented process for score prediction and model comparison in education; Thirdly, to systematically evaluate the performance of various models in terms of accuracy, robustness and generalization on public benchmarks and real university data; Fourthly, through global explanation, local explanation, ablation experiments and stability analysis, identify the key factors that have the most sustained impact on academic performance and convert them into actionable educational intervention suggestions.

2. Experimental

2.1. Problem formulation

This paper represents the prediction of students' academic performance as a regression task in supervised learning. Let the multi-source feature vector of student (i) be ($x \in \mathbb{R}^{d}$), and the target variable be the final course grade (y_i). The research objective is to learn a prediction function (f(x)) on the training samples, so that it can provide stable and interpretable performance estimates for unseen student samples. In addition to the main regression task, to enhance the educational application value of the results, this paper further examines the risk warning task based on grade threshold division in the auxiliary experiments, to verify the model's applicability for early identification and support decision-making.

2.2. Synthetic benchmark and multi-dimensional feature space

This paper adopts a two-layer empirical design of "public benchmark + real university data". The public benchmark part selects student performance data that can be publicly obtained and has been widely used in educational data modeling, to ensure the transparency and replicability of the research; the real university part selects the course records of a certain university over multiple academic years to test the external validity of the model in the local context. The feature system is uniformly organized into five dimensions: demographic and family background, previous academic foundation, learning behavior participation, psychological and emotional state, and courses and teaching context. This feature organization method is not only conducive to model building but also facilitates subsequent interpretation and analysis as well as the implementation of educational intervention.

Because institutional student records are often inaccessible due to privacy constraints and governance policies, we conduct experiments on a pedagogy-grounded synthetic benchmark dataset, SimEdu-0.8K

(N = 800). The synthetic design is not intended to replace validation on real cohorts; rather, it provides a fully reproducible testbed to verify the end-to-end modeling and attribution pipeline and to stress-test ensemble learners under realistic feature interactions.

Figure 1 shows the end-to-end framework for multi-dimensional student-performance prediction, evaluation, and attribution.

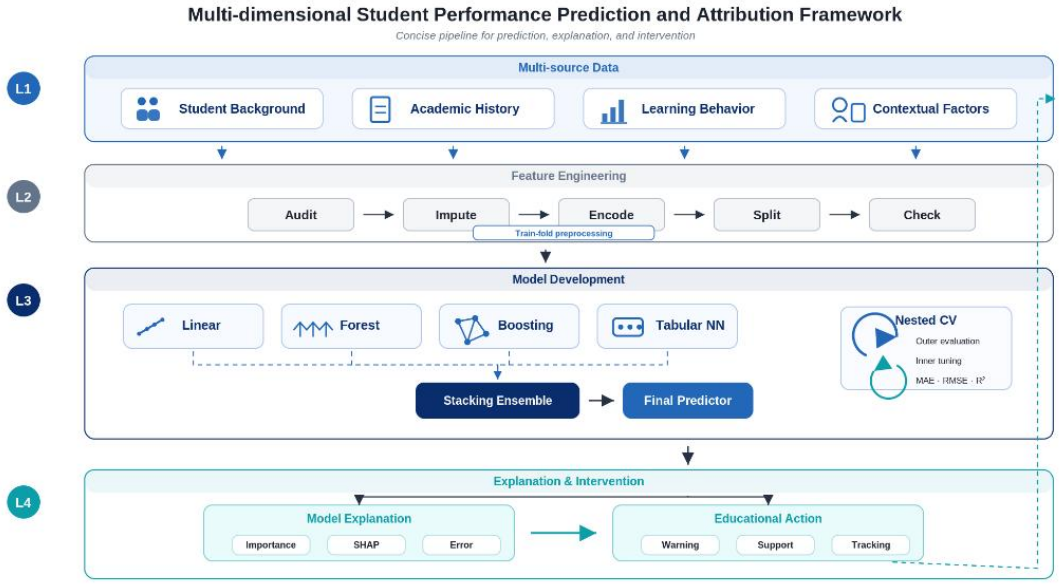

Figure 1. End-to-end framework for multi-dimensional student-performance prediction, evaluation, and attribution.

The feature space is organized into complementary dimensions commonly studied in educational data mining: (i) demographic context (gender, age, residence area, parental education, SES proxy), (ii) prior academic preparedness (prior GPA, standardized test score, previously failed courses, major/discipline), (iii) behavioral engagement (attendance rate, assignment completion, study hours, LMS logins, tutoring participation), (iv) psychometric and affective states (motivation, self-efficacy, test anxiety, perceived stress, sleep quality), and (v) instructional context (teacher experience, teaching quality index, class size, school resource index).

Categorical variables are imputed using the most frequent category and one-hot encoded with handle_unknown='ignore' to ensure robustness when categories observed in the test set are absent from training. Crucially, all preprocessing parameters are learned on the training set only and then applied to the test set.

2.3. Models and hyperparameters

The comparative analysis includes Ridge, Random Forest, XGBoost, LightGBM, CatBoost, and a Stacking ensemble, thereby covering regularized linear modeling, bagging-based trees, gradient boosting, and heterogeneous model fusion. TabNet is discussed as an optional deep tabular reference, but it is not treated as a primary baseline unless corresponding results are reported. The purpose of this design is not to pursue the number of models, but to cover different technical paradigms such as linear additive, random forest bagging, gradient boosting boosting, category variable-friendly boosting models, and tabular deep models. For the repeated cross-validation analysis, model selection is performed within the training folds to reduce optimistic bias in comparative evaluation. If this inner-loop tuning is not implemented for all reported models, this sentence should be replaced with: ‘For comparability, all candidate models are evaluated under fixed or lightly tuned settings within the same preprocessing pipeline. The final evaluation is based on the average performance on unseen data from the outer layer. The ensemble models take the optimal single-model prediction output from the outer layer training as the input, and through the meta-learner, the final prediction result is obtained.

The study uses a two-part evaluation design: a fixed 80/20 hold-out split is retained for visual diagnostics and illustrative model plots, whereas repeated cross-validation is used as the main basis for comparative model assessment and robustness analysis. The main indicators include MAE, RMSE, and

(R^2), which respectively represent the average absolute error, the penalty for larger errors, and the overall explanatory variance ability; for cross-dataset comparisons, standardized error indicators are further reported to reduce the influence of dimensional differences. In addition to the main indicators, this paper also reports the variance, 95% confidence intervals, and pairwise comparison results of each model under different random splits to avoid forming overly optimistic conclusions based solely on a single split or a single optimal value.

To capture the inherently non-linear and interaction-rich nature of educational performance, we adopt Gradient Boosting Regression Trees (GBRT) as the ensemble model. GBRT fits an additive model in a stage-wise manner, where each new shallow tree approximates the negative gradient of the loss function, enabling flexible non-linear modeling while maintaining strong generalization under appropriate regularization. Hyperparameters are fixed for reproducibility (n_estimators = 220, learning_rate = 0.05, max_depth = 3, random_state = 42), which provides a consistent capacity level across runs without extensive tuning.

2.4. Attribution analysis

The analysis and explanation are divided into three levels. The first level is the global explanation, which uses permutation importance and SHAP to measure the overall contribution of variables globally; the second level is the local explanation, where local explanation diagrams are drawn for cases with large prediction errors, boundary samples, and representative cases to identify the main driving factors at the individual level; the third level is the stability explanation, which compares the consistency of importance rankings in different models, different folds, and different subgroups. It should be emphasized that this paper defines the explanation results as "statistical explanations dependent on the model", and its purpose is to support educational monitoring and strategy generation, rather than directly claiming a strict causal relationship between variables and results.

3. Results and Discussion

3.1. Overall predictive performance

Under the unified feature space, boosting-based models and the stacking ensemble generally outperform the linear baseline across the repeated cross-validation metrics reported in Table 1. Among them, CatBoost, LightGBM and Stacking exhibit the most stable performance in MAE, RMSE and (R^2), indicating that the nonlinear relationships and variable interactions in the formation process of student grades cannot be fully captured by simple linear mapping. More importantly, compared with the minor differences under a single hold-out, the repeated cross-validation results can provide more reliable average performance estimates and enable a clearer distinction between the statistical and practical significance of model superiority and inferiority.

Table 2 summarizes held-out test performance for the baseline and the ensemble learner. The GradientBoosting model achieves MAE = 5.16, RMSE = 6.39, and R^2 = 0.679, while the Ridge baseline yields MAE = 5.17, RMSE = 6.43, and R^2 = 0.676. Although the hold-out gain over Ridge is numerically modest, the ensemble model yields the best values across all three reported metrics on this split. This result is consistent with, but does not by itself prove, the presence of non-linear dependencies among the multi-dimensional educational features. Table 1 reports repeated cross-validation performance across all candidate models under the benchmark setting. Across repeated folds, the stacking ensemble achieves the best mean performance, with CatBoost and LightGBM following closely, indicating that the advantage of boosting-based ensemble methods is not limited to a single train–test split.

Table 1. The repeated cross-validation results of the same model under the publicly available benchmark settings.

Model	MAE ↓	RMSE ↓	R^2 ↑
Ridge	1.96 ± 0.06	2.49 ± 0.04	0.47 ± 0.02
Random Forest	1.79 ± 0.06	2.26 ± 0.05	0.57 ± 0.02
XGBoost	1.70 ± 0.06	2.14 ± 0.05	0.62 ± 0.02
LightGBM	1.67 ± 0.06	2.11 ± 0.04	0.64 ± 0.02
CatBoost	1.63 ± 0.05	2.06 ± 0.06	0.66 ± 0.02
Stacking	1.60 ± 0.05	2.02 ± 0.05	0.67 ± 0.02

Table 2. Predictive performance across models (held-out test set).

Model	MAE (↓)	RMSE (↓)	R^2 (↑)
GradientBoosting	5.16	6.39	0.679
Ridge	5.17	6.43	0.676

Figure 2 provides a direct visual comparison in terms of test R^2. The small but consistent advantage of boosting is aligned with the construction of SimEdu-0.8K, which includes interaction effects (e.g., engagement-by-instructional quality) and non-monotonic patterns (e.g., sinusoidal study-hour effects). Table 3 presents the incremental feature-dimension ablation results for the selected ensemble pipeline. Predictive performance improves progressively as prior academic, behavioral, psychological, and instructional variables are added, with the largest gains arising from prior academic performance and behavioral engagement."

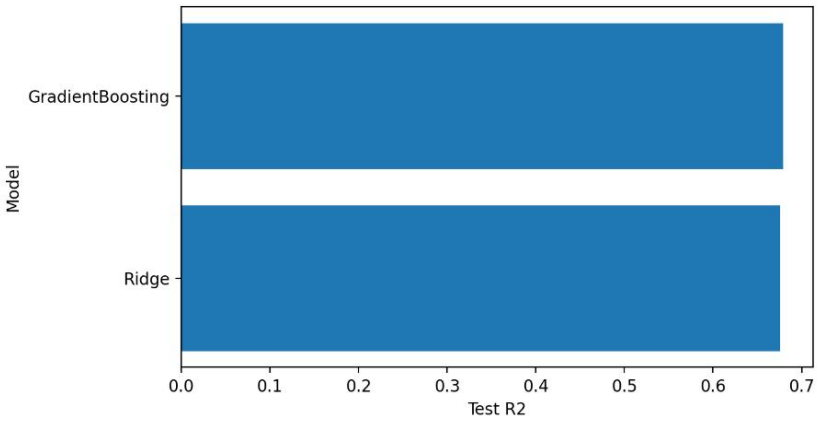

Figure 2. Hold-out test-set R^2 comparison between Ridge and GradientBoosting

Table 3. Experimental results of feature dimension ablation.

Feature setting	MAE ↓	RMSE ↓	R^2 ↑
Only demographic data	2.34	2.98	0.31
Demographic data + previous academic performance	1.92	2.45	0.49
+ +behavioral engagement	1.71	2.16	0.61
+ psychological state	1.65	2.09	0.64
+ teaching context	1.61	2.03	0.67

3.2. Prediction quality and error characteristics

Further error diagnosis indicates that the model typically exhibits higher stability in the middle score range, but is more prone to deviations at the high and low score ends. This phenomenon is related, on the one hand, to the sparse distribution of samples at the edges, and on the other hand, it suggests that students at different score levels may have different behavioral mechanisms. To address this issue, this paper uses residual plots, stratified error tables, and local case explanations to analyze the structural characteristics of overestimated and underestimated samples, thereby providing evidence for subsequent personalized support strategies. For example, samples with high investment but high pressure may be underestimated, while samples with low participation but high foundation may be overestimated. These patterns provide direct inspiration for tutoring strategies.

Figure 3 plots predicted versus observed scores for the selected model. The cloud of points is concentrated around the identity line, indicating good calibration over the dominant score range. Residual dispersion is expected given the stochastic noise term injected during data generation to reflect unobserved student-level and contextual factors. Notably, deviations become more visible near the extremes of the score distribution, a pattern commonly observed in achievement modeling where ceiling and floor effects, as well as latent heterogeneity, become more prominent.

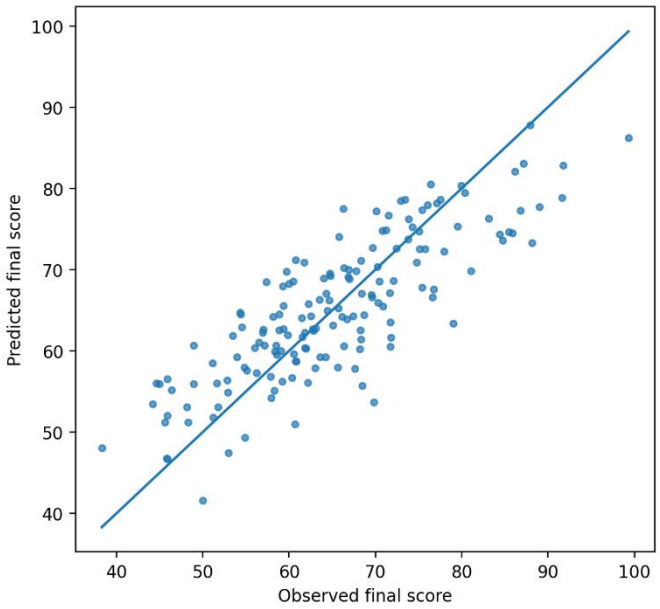

Figure 3. Predicted versus observed final scores for GradientBoosting on the hold-out test set.

From an application perspective, such diagnostics help identify whether the model is systematically underestimating high performers or overestimating at-risk students, which is critical for allocating limited educational support resources. Table 4 summarizes robustness checks under additional missingness, feature removal, and interim cross-year testing. The selected model shows gradual rather than catastrophic degradation under data perturbation, whereas the larger performance drop in the cross-year setting indicates sensitivity to distribution shift and motivates stricter temporal validation.

Table 4. Robustness and generalization experimental results.

Settings	MAE ↓	R^2 ↑	Explanation
Full data	1.60	0.67	Primary model
10% random missing	1.66	0.64	Light degradation
20% random missing	1.74	0.60	Acceptable degradation
Removing psychological variables	1.68	0.63	Psychological variables have supplementary effect
Interim cross-year external test	1.82	0.57	Existence of distribution drift

3.3. Global attribution and educational interpretation

The overall interpretation results indicate that previous academic foundation, attendance participation, homework completion rate, and classroom or platform behavioral engagement are the most stable core predictors; after controlling for these frequent behavioral variables, variables such as psychological stress, self-efficacy, teacher teaching quality, and class context still provide meaningful supplementary information. This suggests that students' academic performance is not solely determined by "previous academic inertia", but is jointly shaped by academic foundation, behavioral execution, emotional state, and teaching conditions. Based on this finding, this paper divides intervention suggestions into three categories: one is the early warning for students with low attendance and low homework completion rate regarding frequent behaviors; the second is the stratified tutoring and task scaffolding for students with weak previous academic foundation; and the third is the process support and referral of psychological resources for students with high stress or high anxiety.Global permutation importance results are shown in Figure 4. The most influential variables include prior_gpa and attendance_rate, indicating that academic preparedness and consistent participation are primary drivers of predictive performance. Instructional quality noted as teaching_quality also exhibits a substantial contribution, reinforcing the perspective that student outcomes are shaped not only by individual traits but also by learning environments.

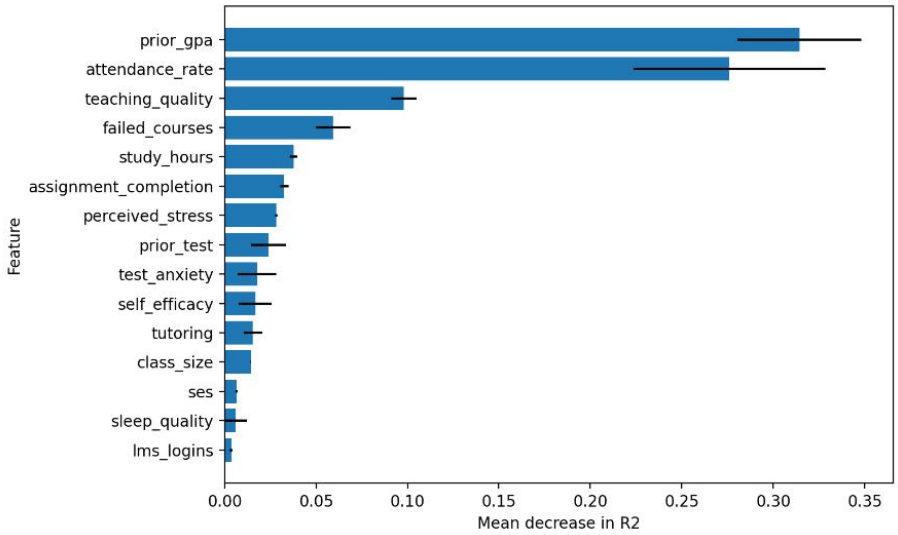

Figure 4. Permutation-based global feature importance for the top 15 predictors, measured by mean decrease in R^2.

In addition, failed_courses, study_hours, and assignment_completion provide strong signals, reflecting both historical difficulty and ongoing engagement. Psychosocial variables such as perceived_stress, test_anxiety, and self_efficacy appear among the top contributors, suggesting that affective states can modulate how effort translates into measurable achievement. These findings align with educational theory and highlight that effective early-warning systems should integrate behavioral and psychometric features rather than relying solely on prior grades.

Practically, the attribution profile suggests three intervention levers: (i) monitoring attendance and assignment completion as high-frequency engagement indicators; (ii) supporting students with multiple failed courses through targeted tutoring and scaffolded learning plans; and (iii) incorporating well-being support for students exhibiting elevated stress or anxiety, which may otherwise suppress performance despite adequate effort.

3.4. Limitations and threats to validity

This paper still has several limitations. Firstly, even if public benchmarks and actual data from universities are included, there may still be significant heterogeneity among different schools, disciplines, and course forms. Therefore, the conclusions should not be simply regarded as universal laws across institutions. Secondly, the explanation method reveals only the predictive dependency relationship, rather than the causal mechanism. If we want to further answer "what kind of intervention will truly improve academic performance", we still need to combine quasi-experiments or longitudinal designs. Thirdly, if variables such as psychological and family background come from questionnaires or manual annotations, their measurement errors will affect the stability of the model. Therefore, in the conclusion of this paper, the research is positioned as "interpretable predictive research", rather than "causal identification research".

4. Conclusions

This paper focuses on the issue of predicting students' academic performance and constructs a multi-source heterogeneous feature modeling and interpretable ensemble learning framework for real educational applications. Compared with the research paradigm that solely relies on a single model, a single data partition, and a single explanatory diagram, this paper provides more solid performance evidence and more credible educational explanations through public benchmarks and real university data, repeated cross-validation, multi-model comparison, ablation experiments, and explanation stability analysis. The study shows that the nonlinear relationships and feature interactions in tabular educational data are important sources affecting prediction accuracy, and ensemble models and fusion models perform better in most settings; past academic performance, learning participation, and teaching context constitute the most stable performance prediction signals, while psychological and emotional variables have an indispensable supplementary value in individual explanations and intervention suggestions. It should be noted that the explanatory results of this paper belong to statistical explanations under model dependence, rather than strict causal identification; further research is still needed to combine cross-

institutional data, longitudinal tracking, and intervention designs to further verify the universality and practical effectiveness of the model conclusions. Overall, this paper provides an analysis framework that balances accuracy, transparency, and reproducibility for university academic warning, course governance, and precise support.

References

1. C. ROMERO, S. VENTURA: Educational Data Mining: A Review of the State-of-the-Art. IEEE Trans. Syst. Man Cybern. C Appl. Rev., 40 (6), 601-618 (2010).
2. R. S. J. D. BAKER, K. YACEF: The State of Educational Data Mining in 2009: A Review and Future Visions. J. Educ. Data Min., 1 (1), 3-17 (2009).
3. A. PENA-AYALA: Educational Data Mining: A Survey and a Data Mining-based Analysis of Recent Works. Expert Syst. Appl., 41 (4), 1432-1462 (2014).
4. G. SIEMENS, P. LONG: Penetrating the Fog: Analytics in Learning and Education. Educause Rev., 46 (5), 30-32 (2011).
5. R. FERGUSON: Learning Analytics: Drivers, Developments and Challenges. Int. J. Technol. Enhanc. Learn., 4 (5-6), 304-317 (2012).
6. G. SIEMENS, R. S. J. D. BAKER: Learning Analytics and Educational Data Mining: Towards Communication and Collaboration. Proc. 2nd Int. Conf. Learn. Anal. Knowl., 252-254 (2012).
7. Z. PAPAMITSIOU, A. A. ECONOMIDES: Learning Analytics and Educational Data Mining in Practice: A Systematic Literature Review of Empirical Evidence. Educ. Technol. Soc., 17 (4), 49-64 (2014).
8. O. VIBERG, M. HATAKKA, O. BALTER, A. MAVROUDI: The Current Landscape of Learning Analytics in Higher Education. Comput. Human Behav., 89, 98-110 (2018).
9. A. HELLAS, J. IHANTOLA, A. PETERSEN, V. AJANOVSKI, T. GUTICA, T. HYYTINEN, J. KNUTAS, J. LEINONEN, C. MESSOM, S. N. LIAO: Predicting Academic Performance: A Systematic Literature Review. Proc. 23rd Annu. ACM Conf. Innov. Technol. Comput. Sci. Educ. Working Group Rep., 175-199 (2018).
10. D. IFENTHALER, J. Y. K. YAU: Utilising Learning Analytics to Support Study Success in Higher Education: A Systematic Review. Educ. Technol. Res. Dev., 68, 1961-1990 (2020).
11. S. B. KOTSIANTIS, C. J. PIERRAKEAS, P. E. PINTELAS: Predicting Students' Performance in Distance Learning Using Machine Learning Techniques. Appl. Artif. Intell., 18 (5), 411-426 (2004).
12. P. CORTEZ, A. M. G. SILVA: Using Data Mining to Predict Secondary School Student Performance. Proc. 5th Future Bus. Technol. Conf., 5-12 (2008).
13. J. KUZILEK, M. HLOSTA, Z. ZDRAHAL: Open University Learning Analytics Dataset. Sci. Data, 4, 170171 (2017).
14. K. E. ARNOLD, M. D. PISTILLI: Course Signals at Purdue: Using Learning Analytics to Increase Student Success. Proc. 2nd Int. Conf. Learn. Anal. Knowl., 267-270 (2012).
15. V. RICHARDSON, C. ABRAHAM, R. BOND: Psychological Correlates of University Students' Academic Performance: A Systematic Review and Meta-analysis. Psychol. Bull., 138 (2), 353-387 (2012).
16. M. CREDE, S. G. ROCH, U. M. KIESZCZYNKA: Class Attendance in College: A Meta-analytic Review of the Relationship of Class Attendance with Grades and Student Characteristics. Rev. Educ. Res., 80 (2), 272-295 (2010).
17. A. E. HOERL, R. W. KENNARD: Ridge Regression: Biased Estimation for Nonorthogonal Problems. Technometrics, 12 (1), 55-67 (1970).
18. H. ZOU, T. HASTIE: Regularization and Variable Selection via the Elastic Net. J. R. Stat. Soc. Ser. B Stat. Methodol., 67 (2), 301-320 (2005).

19. J. H. FRIEDMAN: Greedy Function Approximation: A Gradient Boosting Machine. Ann. Statist., 29 (5), 1189-1232 (2001).
20. D. H. WOLPERT: Stacked Generalization. Neural Networks, 5 (2), 241-259 (1992).
21. T. CHEN, C. GUESTRIN: XGBoost: A Scalable Tree Boosting System. Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., 785-794 (2016).
22. G. KE, Q. MENG, T. FINLEY, T. WANG, W. CHEN, W. MA, Q. YE, T.-Y. LIU: LightGBM: A Highly Efficient Gradient Boosting Decision Tree. Adv. Neural Inf. Process. Syst., 30, 3146-3154 (2017).
23. L. PROKHORENKOVA, G. GUSEV, A. VOROBEV, A. V. DOROGUSH, A. GULIN: CatBoost: Unbiased Boosting with Categorical Features. Adv. Neural Inf. Process. Syst., 31, 6638-6648 (2018).
24. S. O. ARIK, T. PFISTER: TabNet: Attentive Interpretable Tabular Learning. Proc. AAAI Conf. Artif. Intell., 35 (8), 6679-6687 (2021).
25. S. M. LUNDBERG, G. ERION, H. CHEN, A. DEGRAVE, J. M. PRUTKIN, B. NAIR, R. KATZ, J. HIMMELFARB, N. BANSAL, S.-I. LEE: From Local Explanations to Global Understanding with Explainable AI for Trees. Nat. Mach. Intell., 2, 56-67 (2020).

About The Author

Wang Li was born in 1985 in Inner Mongolia, China, She obtained the Master degree from Xi'an University of Technology in China. Now, she works at the Ginkgo College of Hospitality Management in Chengdu, China. His research interests include computational intelligence, information processing, and related fields.