



Predicting Mandarin Vocabulary Learning Outcomes Using Data-Driven Machine Learning

Yan Qin¹, Joseph Teguh Santoso², Agus Wibowo²

¹Department of International Education Center, Faculty of General Education, Fuzhou Polytechnic China, Fuzhou, China

²Dept. of Computer Science, Faculty of Study Academic, Universitas Sains dan Teknologi Komputer, Semarang, Indonesia

*[gjyxy@fvti.edu.cn](mailto:gjjyxy@fvti.edu.cn)

Abstract

The rapid expansion of language learning in higher education highlights the need for data-driven approaches to monitor student progress and provide timely instructional support. This study aims to develop a predictive framework for Mandarin vocabulary mastery using supervised machine learning. A dataset of 147 undergraduate students was analyzed, incorporating study hours, number of exercises, pre-test scores, and attendance as predictors of learning outcomes. Logistic Regression, Random Forest, and XGBoost algorithms were trained and evaluated, with XGBoost achieving the highest performance (accuracy 88%, F1-score 0.88), demonstrating its superior ability to capture complex learning patterns. Analysis of feature importance revealed that pre-test scores and the number of exercises were the most influential predictors of student success. Furthermore, a prototype graphical user interface (GUI) was developed to visualize predictions in real time, enabling instructors to identify at-risk students and adjust teaching strategies accordingly. The novelty of this study lies in integrating predictive analytics with pedagogical applications, bridging machine learning and educational practice. Beyond its technical contributions, this research provides practical insights for higher education stakeholders, showing how predictive models can support early intervention, enhance curriculum design, and promote evidence-based decision-making in Mandarin vocabulary instruction.

Keywords: educational data mining; machine learning; mandarin vocabulary learning; predictive analytics; XGBoost

How to Cite: Y. Qin, J. T. Santoso, and A. Wibowo, "Predicting Mandarin Vocabulary Learning Outcomes Using Data-Driven Machine Learning", *J. RESTI (Rekayasa Sist. Teknol. Inf.)*, vol. 10, no. 3, pp. 825 - 833, Jun. 2026.
Permalink/DOI: <https://doi.org/10.29207/resti.v10i3.7143>

Received: September 2, 2025

Accepted: May 23, 2026

Available Online: June 24, 2026

*This is an open-access article under the CC BY 4.0 License
Published by Ikatan Ahli Informatika Indonesia*

1. Introduction

Mandarin Chinese has become increasingly important in global education, business, and intercultural communication. Its expanding role has increased the demand for effective instructional strategies for non-native learners [1]. However, vocabulary acquisition remains a persistent challenge in second-language learning because it requires not only memorization but also repeated exposure, contextual understanding, and sustained practice [2]. Conventional teaching approaches often rely on delayed assessment and teacher observation, making it difficult to identify students at risk of poor learning outcomes at an early stage [3]. This limitation highlights the need for predictive approaches that can support more timely and data-informed instructional decisions.

Data-driven prediction has been widely adopted in educational research to monitor student progress and

estimate academic outcomes. Previous studies have shown that machine learning can support the prediction of student success, graduation, and broader academic performance using structured educational data [4], [5]. In parallel, research on artificial intelligence in education has emphasized the growing role of predictive and adaptive systems in improving feedback, learner support, and self-regulated learning processes [6], [7]. These developments suggest that predictive analytics is increasingly relevant for educational decision-making, but its application remains uneven across specific learning domains.

In language education, predictive analytics has been applied to several areas, including reading performance, pronunciation assessment, general language proficiency, and learner behavior modeling [8-11]. Some studies have used logistic regression and classification-based methods to estimate language proficiency or identify learners at risk [9], [12], while

others have shown that combining multiple predictors can improve predictive accuracy [13], [14]. Ensemble methods such as Random Forest have also been reported to capture nonlinear relationships in learning outcomes more effectively than simpler baseline models [15], [16], and broader evidence suggests that ensemble approaches can reveal hidden patterns that are not easily detected by conventional statistical techniques [17], [18].

Recent studies have also highlighted the importance of prior knowledge, behavioral indicators, and adaptive modeling in language-related learning contexts. Research on Mandarin and related educational settings has emphasized the contribution of prior ability to subsequent learning performance [19], [20]. Meanwhile, machine learning and deep learning approaches have shown potential for distinguishing learner profiles and supporting early intervention [10], [11], [21].

Other studies have discussed the broader value of AI-based prediction systems for personalized feedback, early warning mechanisms, and targeted educational support [22-24]. Taken together, these studies confirm the potential of predictive analytics in language learning, but they also show that the field has focused more strongly on general proficiency or broad learning behavior than on vocabulary-specific mastery in Mandarin. While these studies demonstrate strong predictive capabilities, most of them focus primarily on model accuracy rather than on interpretability or practical classroom integration.

Although prior studies have provided useful insights into modeling student performance through various machine learning algorithms [25-29]. Most have been conducted in English-learning or general academic contexts rather than in Mandarin vocabulary learning contexts. Similarly, studies on ensemble and deep learning models in education have demonstrated strong predictive capability [30-34], yet many of them place greater emphasis on model performance than on how predictions can be translated into practical classroom use. As a result, two gaps remain evident: first, limited work has specifically addressed predicting Mandarin vocabulary mastery; second, practical implementation for teacher-facing instructional support remains underdeveloped.

Therefore, this study aims to develop and evaluate supervised machine learning models for predicting Mandarin vocabulary learning outcomes using behavioral and academic indicators, namely study hours, completed exercises, pre-test scores, and attendance. In addition to comparing Logistic Regression, Random Forest, and XGBoost, this study introduces a prototype graphical user interface (GUI) to demonstrate how predictive results can be used in real time to support education. The novelty of this research lies in integrating predictive modeling with practical pedagogical application, thereby transforming model

outputs into actionable insights for early intervention and data-informed Mandarin vocabulary instruction. Beyond predictive performance, this study also emphasizes methodological transparency and practical applicability. The proposed framework is designed not only to generate accurate predictions but also to support interpretable and actionable use in real educational settings.

2. Methods

This study uses a quantitative experimental design with supervised machine learning to predict learning outcomes of Mandarin vocabulary. This methodological approach was designed to ensure that each analytical step, from data preprocessing to model evaluation, directly supports the study's objective of predicting Mandarin vocabulary learning outcomes and providing actionable insights for instructors. Such a design allows for systematic investigation of patterns between predictors of learning and performance in a reproducible and unbiased fashion.

The workflow process involves dataset preparation, preprocessing, exploratory analysis, model building, and performance evaluation. Use of a systematic method in this way not only assures methodological validity but also aligns with best practices of learning analytics and educational data mining, as asserted by [35].

2.1 Research Framework

The research framework provides a clear, sequential procedure that covers all stages of the study. It begins with dataset preparation, followed by preprocessing and exploratory analysis, and continues with model development and evaluation. The prediction stage is the final step, in which the trained model generates results for new input data. This stage also incorporates the graphical user interface (GUI) as a practical tool to visualize and apply the model's real-time predictions for educational use, as illustrated in Figure 1.

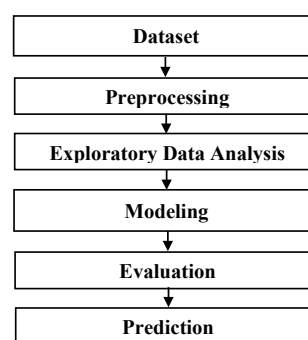


Figure 1. Research Framework of the Study

2.2 Dataset

The dataset for this study came from undergraduate students in an introductory Mandarin course. Each record contained four predictor variables: weekly study

hours (Hours_Study), number of completed exercises (Exercises), pre-test score (PreTest), and class attendance (Attendance), as well as the post-test score (PostTest). The dependent variable, Status, was derived from the post-test score, with scores of 70 or higher labeled *Pass* and scores below 70 as *Fail*. The class distribution was skewed toward the Pass category, indicating a moderate imbalance between passing and failing students. This distribution reflects the overall improvement observed after the course and was taken into account during model selection and evaluation. After removing three incomplete records, 147 valid samples remained, with descriptive statistics presented in Table 1.

Table 1. Descriptive Statistics of the Dataset

Variable	Min	Max	Mean	Std. Dev
Hours_Study	2	10	5.1	1.8
Exercises	15	60	34.7	9.3
PreTest Score	30	85	56.4	12.7
Attendance (%)	60	100	85.2	7.5
PostTest Score	40	95	71.6	11.2

2.3 Data Preprocessing

Data preprocessing was conducted to ensure data consistency and readiness for model analysis. Incomplete records were removed, resulting in 147 valid entries for further processing. All numerical features were normalized to a range between 0 and 1, and the output variable Status was encoded as binary (Pass = 1, Fail = 0). The dataset was then divided into 70% for training and 30% for testing, with the post-test score used solely to define the target label and excluded from the predictors.

Because the dataset contained more passing than failing students, model evaluation did not rely solely on accuracy. Precision, recall, and F1-score were also used to provide a more balanced assessment of classification performance in the presence of class imbalance. To improve robustness, cross-validation was applied during model development to reduce dependence on a single data partition and to support more reliable model comparison prior to final testing.

The cross-validation procedure was conducted within the training data, while the testing set was reserved for final evaluation. Due to the exploratory nature of the study and the absence of complete experimental logs, specific fold configurations were not retained.

2.4 Exploratory Data Analysis

The exploratory data analysis began with examining the cumulative distribution of students' pre-test and post-test scores to capture overall learning progress. Figure 2 shows that the post-test distribution shifts to the right compared to the pre-test, indicating measurable improvement in vocabulary acquisition. The cumulative curves reveal that a higher proportion of students reached scores above the passing threshold after the course. These findings suggest that the learning

intervention was effective in enhancing vocabulary mastery.

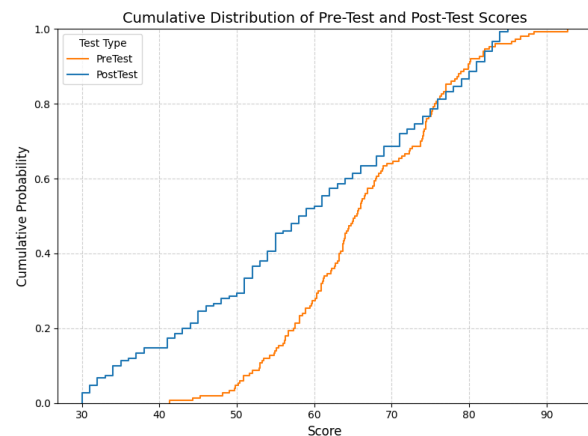


Figure 2. Cumulative Distribution of Pre-Test and Post-Test Scores

A correlation analysis was then conducted to identify interrelationships among the study variables influencing learning outcomes. As illustrated in Figure 3, study hours, number of exercises, and pre-test scores show moderate positive correlations with post-test performance. Attendance demonstrates a weaker but still meaningful relationship, underscoring the importance of consistent participation. These relationships guided the feature selection process for predictive modeling.

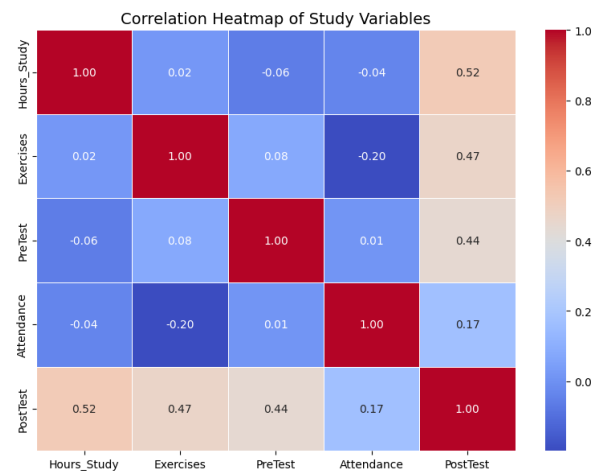


Figure 3. Correlation Heatmap of Study Variables

Finally, distribution comparisons were performed to explore how learning behaviors differ between passing and failing students. Figure 4 shows that those who passed generally had higher study hours, completed more exercises, scored better on the pre-test, and attended classes more regularly. These behavioral patterns confirm that the selected variables have strong discriminative power. The results provide a clear rationale for their inclusion in the machine learning models.

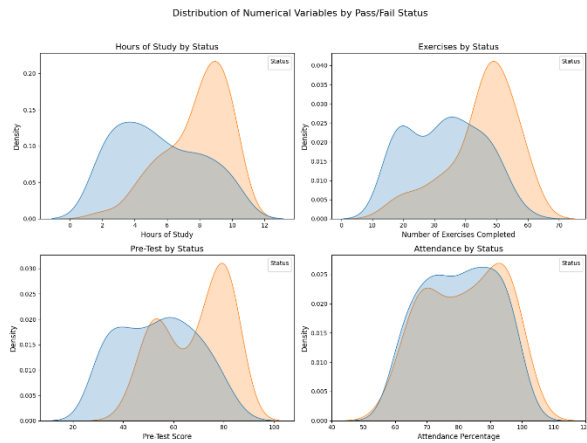


Figure 4. Distribution of Numerical Variables Based on Pass and Fail Categories

2.5 Machine Learning Models

This study employed three supervised machine learning algorithms to predict Mandarin vocabulary mastery: Logistic Regression, Random Forest, and XGBoost. Logistic Regression was used as a baseline classifier because of its interpretability and effectiveness in binary classification tasks. Random Forest was applied to capture nonlinear relationships and reduce overfitting through ensemble learning. XGBoost was selected because of its strong performance on structured educational data and its ability to model complex interactions among learning-related variables.

During model development, parameter tuning and cross-validation were performed to improve robustness before final testing. Because the present study was designed with a practical educational focus and a relatively modest dataset size, the tuning process was kept limited, and model configurations were not exhaustively optimized. Instead, the analysis prioritized comparative model performance, interpretability of the main predictors, and practical feasibility for classroom-oriented deployment.

To improve methodological transparency and reproducibility, the final model configurations used in this study are summarized in Table 2. The configurations represent the main parameters applied during model implementation. Although the tuning process was limited due to the modest dataset size, documenting these settings provides a clearer basis for interpreting the reported model performance and supports partial replication of the experimental procedure.

These configurations were selected to balance predictive performance and model stability, particularly because the dataset consisted of only 147 valid observations. Although the models were not exhaustively optimized, reporting the main parameters improves transparency and provides a clearer basis for partial replication. Future studies should retain complete tuning logs and report all experimental

settings more comprehensively to support full reproducibility.

Table 2. Summary of Final Model Configuration

Model	Main Configuration
Logistic Regression	Solver: liblinear; penalty: L2; maximum iteration: 1000; regularization strength (C): 1.0; random state: 42
Random Forest	Number of estimators: 100; criterion: Gini; maximum depth: default; minimum samples split: 2; minimum samples leaf: 1; random state: 42
XGBoost	Number of estimators: 100; learning rate: 0.10; maximum depth: 3; subsample: 0.80; colsample_bytree: 0.80; evaluation metric: logloss; random state: 42

The selected algorithms provide complementary strengths for the study objective. Logistic Regression offers a simple and interpretable baseline, Random Forest captures nonlinear relationships through ensemble learning, and XGBoost provides stronger predictive capability for structured educational data. This combination supports a comparative evaluation of supervised machine learning approaches for Mandarin vocabulary outcome prediction.

2.6 Evaluation Metrics

The first evaluation metric used in this study is accuracy, which measures the proportion of correctly classified observations among all cases. Accuracy provides a general assessment of model performance and is commonly used as a baseline metric in binary classification tasks; its computation is summarised in Equation 1.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

Here, TP denotes the number of true positives, TN the true negatives, FP the false positives, and FN the false negatives. Higher accuracy indicates better overall classification, though it may not fully reflect performance on imbalanced datasets.

The second metric is precision, which evaluates the model's ability to correctly identify positive cases among all predicted positives. Precision is critical in educational prediction, as it reflects confidence in classifying students as passing, as defined in Equation 2.

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

Precision emphasizes minimizing false positives, ensuring that students are not incorrectly predicted to pass. A higher precision score indicates stronger reliability in positive classifications.

The third metric is recall, which measures the proportion of correctly identified positive cases out of all actual positives. Recall is crucial when missing at-risk students could have serious educational consequences, as shown in Equation 3.

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

Recall penalizes models that fail to identify true positives, ensuring that most students who pass are correctly identified. High recall indicates that the model captures most of the students who actually passed.

The fourth metric is the F1-score, representing the harmonic means of precision and recall. This measure balances false positives and false negatives, especially in the presence of uneven class distributions, as shown in Equation 4.

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4)$$

The F1-score provides a single value capturing the trade-off between precision and recall. Higher F1 values indicate a better balance in correctly identifying both passing and failing students.

2.7 Prediction Process

Once the models were trained and evaluated, the selected classifier was used to predict student learning outcomes based on observed behavioral and academic indicators. New data, including study duration, exercise count, pre-test score, and attendance, were processed through the same preprocessing pipeline as the training data.

The model generated Pass or Fail predictions with probability scores indicating confidence. This prediction phase was implemented through a Python-Tkinter Graphical User Interface (GUI), enabling real-time use of the model for classroom monitoring and instruction. The same preprocessing procedure used during model training was applied to new input data before prediction to maintain methodological consistency.

3. Results and Discussions

The results show measurable improvement in students' Mandarin vocabulary learning, as reflected by higher post-test scores than pre-test performance. Three supervised machine learning algorithms, namely Logistic Regression, Random Forest, and XGBoost, were trained and evaluated to predict these outcomes. Among them, XGBoost achieved the best performance and was further examined through confusion matrix analysis, feature importance, ROC evaluation, and prediction output.

3.1 Initial Data Analysis

Descriptive analysis revealed an improvement in learning outcomes, with mean scores increasing from 56 on the pre-test to 72 on the post-test, indicating a measurable vocabulary gain. Correlation analysis revealed a moderate correlation between pre- and post-test scores ($r \approx 0.52$) and between practice intensity (number of exercises) and post-test scores ($r \approx 0.47$), which suggests that prior knowledge and practice intensity do matter. The distribution of outcomes showed that 120 students (81.6%) passed and 27 (18.4%) failed, indicating that most students met the

minimum mastery threshold after the instructional period. These results suggest that frequent practice and prior ability play important roles in Mandarin vocabulary development and may inform curriculum planning and early instructional support.

3.2 Model Performance

To compare predictive performance, three supervised models were trained and tested (Random Forest, Logistic Regression, and XGBoost). As shown in Table 3, XGBoost provided the best result (accuracy 0.88, precision 0.89, recall 0.87, F1-score 0.88), followed by Random Forest (accuracy 0.85) and Logistic Regression (accuracy 0.81). These findings indicate that ensemble-based boosting methods more effectively capture the complexity of learning indicators in predicting Mandarin vocabulary outcomes. They also suggest that machine learning can support earlier monitoring of student performance and more timely instructional intervention.

Table 3. Performance of Machine Learning Models

Model	Accuracy	Precision	Recall	F1-score
Logistic Regression	0.81	0.82	0.79	0.80
Random Forest	0.85	0.86	0.84	0.85
XGBoost	0.88	0.89	0.87	0.88

The superior performance of XGBoost in this study is consistent with prior findings that ensemble boosting methods are effective at capturing complex, nonlinear relationships in educational data [15], [16]. Compared with Logistic Regression, which assumes a more direct linear relationship between predictors and outcomes, XGBoost is better able to model interactions among prior knowledge, learning practice, attendance, and study behavior. This is particularly relevant in vocabulary learning, where performance is unlikely to depend on a single factor in isolation.

Compared with Random Forest, which also accommodates nonlinear patterns, XGBoost benefits from sequential learning that iteratively corrects prediction errors. This likely contributed to its stronger performance in the present dataset, where the predictive signal appears to emerge from the combined influence of multiple behavioral and academic indicators. These findings support the use of boosting-based ensemble approaches when learner outcomes are shaped by interacting variables rather than by a single dominant feature.

3.3 Evaluation of the Best Model XGBoost

The confusion matrix in Figure 5 shows that the XGBoost model correctly classified most students in both Pass and Fail categories, with only a small number of misclassifications. This indicates that the model achieved balanced classification performance and effectively distinguished learner outcomes. The relatively low number of errors suggests that the model is sufficiently reliable to identify learner profiles in the

present dataset and may be useful for early instructional support.

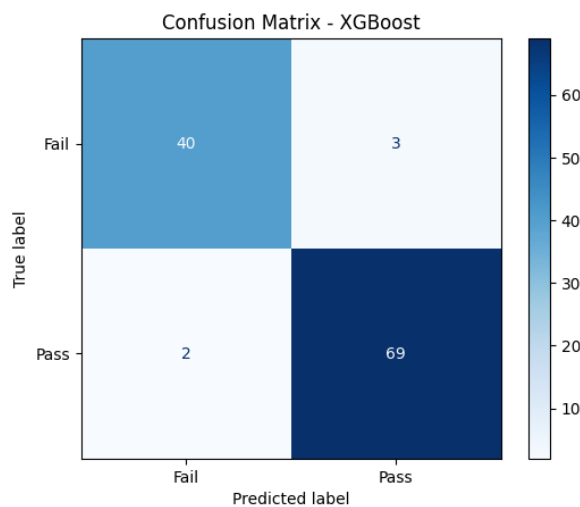


Figure 5. Confusion Matrix of XGBoost

Feature importance analysis was conducted to examine the model's decision-making process. Pre-test scores contributed the most to the predictions, accounting for 42% of the model weight, followed by the number of exercises at 31%, attendance at 17%, and weekly study time at 10%. These results indicate that prior knowledge and consistent practice are the strongest predictors of vocabulary mastery. Such insights can guide instructors in designing adaptive learning activities that reinforce these key factors, as illustrated in Figure 6.

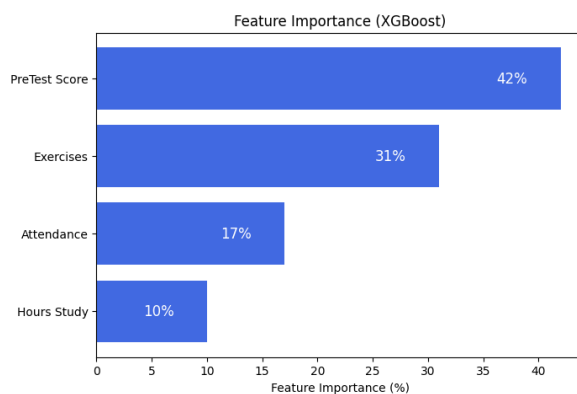


Figure 6. Feature Importance of XGBoost

From a language-learning perspective, these findings reflect vocabulary acquisition as a cumulative and reinforcement-based process. The strong contribution of pre-test scores indicates that prior lexical knowledge provides a critical foundation for subsequent learning, as learners with stronger initial vocabulary are better positioned to recognize, retain, and integrate new words. The importance of completed exercises further highlights the role of repeated practice and retrieval in strengthening vocabulary mastery, while attendance supports sustained exposure and continuity of learning. These findings align with previous studies highlighting the interaction between prior knowledge, practice

intensity, and learner engagement in shaping language learning outcomes [19-21]. Together, the findings suggest that vocabulary mastery is not driven by a single factor, but emerges from the combined influence of cognitive readiness and sustained learning behavior.

To further strengthen model interpretability, an additional model-based feature contribution visualization was included. The visualization supports the feature importance results by showing that pre-test score and completed exercises were the dominant contributors to the XGBoost prediction output, followed by attendance and weekly study hours, as shown in Figure 7.

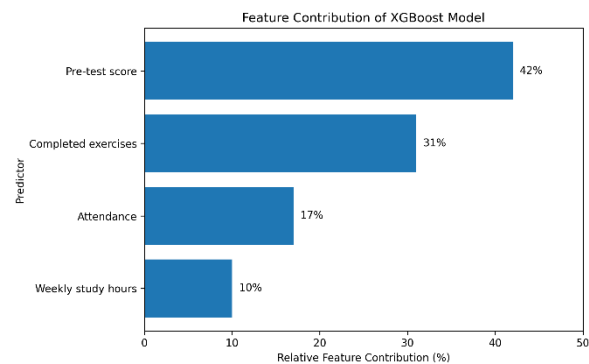


Figure 7. Model-Based Feature Contribution Visualization of XGBoost

This pattern indicates that the model's predictions were mainly influenced by pedagogically meaningful variables. In practical terms, students with stronger initial vocabulary knowledge and more consistent exercise completion tended to receive higher probability scores for the Pass category, whereas lower pre-test scores and limited practice were associated with a greater risk of being classified as Fail. Future studies may extend this analysis using SHAP values or other instance-level explainability techniques to provide a more detailed interpretation of individual student predictions.

The ROC-AUC value of 0.99 indicates very strong discriminative ability of the XGBoost model in separating Pass and Fail categories within the present dataset. The ROC curve is presented in Figure 8. However, this result should be interpreted cautiously because the dataset contains only 147 valid samples and was collected from a single learning context. In small datasets, exceptionally high AUC values may partly reflect cohort-specific patterns rather than fully generalizable predictive performance. Therefore, the high ROC-AUC should not be understood as definitive evidence of model robustness across different Mandarin learning populations.

To reduce the risk of overfitting, several precautions were applied during model development. First, the post-test score was excluded from the predictor set and used only to define the target label, preventing direct leakage from the outcome variable. Second, the dataset was

separated into training and testing subsets, with the testing set reserved for final evaluation. Third, cross-validation was applied during the development stage to reduce dependence on a single data split. Fourth, model interpretation was supported by feature importance analysis, which showed that the prediction was driven by pedagogically meaningful variables, particularly pre-test scores and completed exercises. Nevertheless, further validation using larger samples, external cohorts, and repeated experimental settings is required before the model can be claimed to have strong generalization capability.

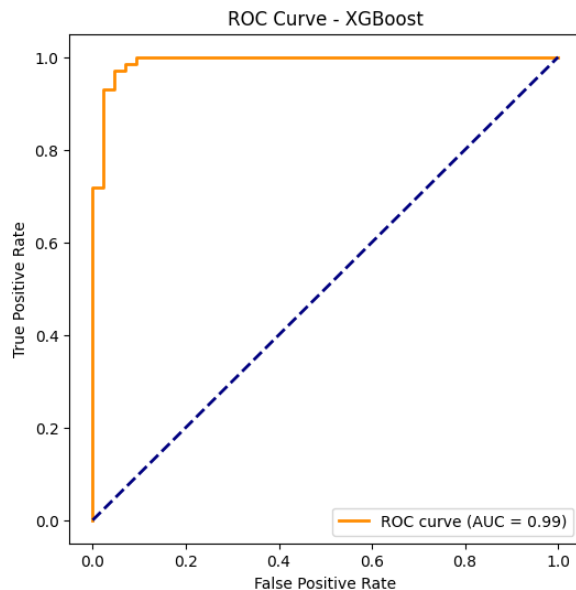


Figure 8. ROC Curve of XGBoost (AUC = 0.99)

3.4 Prediction Output

After training and validating the XGBoost model, it was used to predict new student records not used during training. The input features included weekly study hours, number of completed exercises, pre-test score, and attendance percentage, all of which were processed using the same preprocessing pipeline as the training data. The model generated Pass or Fail labels together with probability scores indicating prediction confidence. To demonstrate practical feasibility, a Python-Tkinter graphical user interface (GUI) was developed, as shown in Figure 9. The GUI functions as a prototype interface that allows instructors to enter student data and obtain real-time predictions for classroom monitoring and instructional decision-making.

From an educational practice perspective, the GUI may extend the predictive model's value beyond statistical output by making prediction results more accessible to instructors. Rather than waiting until the end of a course, teachers may use prediction scores to identify students who may need earlier support, such as additional vocabulary exercises, remedial review, or closer monitoring of attendance. In this sense, the predictive framework functions not only as a

classification tool but also as a pedagogically oriented support mechanism for more responsive teaching.

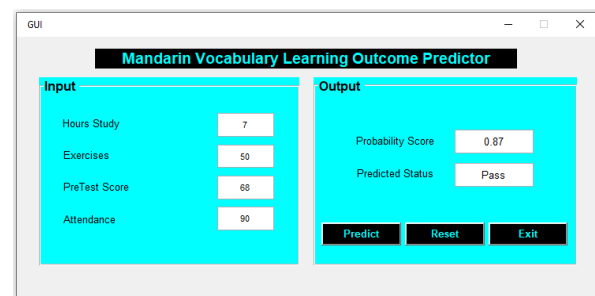


Figure 9. GUI for Mandarin Vocabulary Prediction

A small usability observation was also conducted to obtain preliminary feedback on the GUI prototype. Two Mandarin instructors were asked to review the interface workflow, including data input, prediction generation, and interpretation of the output label. Their feedback indicated that the interface was simple to operate and that the Pass/Fail prediction could help instructors identify students who may require additional support. However, they also suggested that future versions should include clearer probability interpretation, warning indicators for at-risk students, and brief instructional recommendations linked to the prediction result. This preliminary observation suggests that the GUI has practical potential, although broader usability testing is still required before classroom deployment.

3.5 Limitations and Future Work

Several limitations should be considered when interpreting the findings of this study. First, the dataset consisted of 147 valid records collected from a single cohort of undergraduate Mandarin learners. The exceptionally high ROC-AUC value should also be considered in this context, as small and homogeneous datasets may inflate discrimination metrics and increase the possibility of overfitting. Although the predictive results are encouraging, the relatively small sample size limits statistical stability and increases sensitivity to sampling variation. As a result, the reported performance may reflect patterns that are specific to the present cohort and may not fully represent broader learner populations.

Second, the class distribution was moderately imbalanced, with more students in the Pass category than in the Fail category. To address this condition, model evaluation was based not only on accuracy but also on precision, recall, and F1-score to provide a more balanced assessment. Nevertheless, the imbalance may still affect sensitivity to minority cases, particularly students at risk of failing. Future studies may compare the present baseline approach with alternative strategies such as oversampling, synthetic minority generation, undersampling, or class-weighted learning.

Third, the analysis relied on only four predictors: study hours, completed exercises, pre-test scores, and

attendance. Although these variables represent important academic and behavioral indicators, they do not fully capture the broader factors that may influence Mandarin vocabulary acquisition. Future research should incorporate additional educational variables such as learner motivation, interaction logs from digital learning platforms, assignment completion patterns, socio-demographic background, prior exposure to Mandarin, feedback quality, and learning strategy use. Including these variables may improve predictive robustness, reduce model dependency on a narrow set of indicators, and provide a more comprehensive understanding of the factors shaping vocabulary learning outcomes.

Fourth, although cross-validation and comparative model testing were applied to improve robustness, the study does not retain a complete record of all detailed parameter configurations explored during model development. This limits full methodological reproducibility and should be acknowledged as a technical limitation of the present work. Future research should therefore document tuning settings and experimental logs more comprehensively.

Lastly, although a small usability observation with instructors offered preliminary insights into the GUI workflow, the prototype has not yet been assessed through systematic classroom trials or broader user testing. Future research should therefore evaluate its usability, instructor acceptance, and practical contribution to instructional intervention and vocabulary learning outcomes. In addition, future studies should incorporate more advanced explainability techniques, such as SHAP values or instance-level feature contribution analysis, to clarify how each predictor influences individual student predictions.

4. Conclusions

This study evaluated the effectiveness of supervised machine learning models for estimating Mandarin vocabulary learning performance using data from 147 students. The results showed a notable improvement in mean scores, from 56 on the pre-test to 72 on the post-test, with an 81.6% pass rate. Among the evaluated models, Logistic Regression achieved an accuracy of 0.81, Random Forest reached 0.85, and XGBoost attained the highest performance with an accuracy of 0.88 and an F1-score of 0.88. Although the ROC-AUC result was high, it should be interpreted as preliminary because the model was evaluated on a small and context-specific dataset. These findings indicate that predictive modeling can effectively capture meaningful relationships between learning indicators and vocabulary outcomes.

The analysis further showed that pre-test scores and the number of exercises were the strongest predictors, followed by attendance and study hours. This finding reinforces the importance of prior knowledge,

consistent practice, and learner engagement in Mandarin vocabulary learning. For a practical demonstration, a prototype GUI based on the XGBoost model was developed to support real-time prediction and instructional monitoring.

This study contributes to educational data mining by demonstrating how predictive modeling can be effectively integrated with practical pedagogical applications in Mandarin vocabulary learning. However, because the study was based on a relatively small and context-specific dataset, the findings should be interpreted with appropriate caution. Future studies should also expand the predictor set by including motivational, behavioral, demographic, and prior-language-exposure variables to improve the robustness and educational relevance of the prediction model.

Acknowledgements

The authors affirm that this study was conducted independently and without any financial or commercial influences. No external funding was obtained for this research, and the work was carried out without support from governmental, private, or non-profit funding agencies.

References

- [1] P. Hao and F. Li, "Exploring the Challenges of Learning and Teaching Chinese/Mandarin Language at Higher Education Institutes: Voices from Non-Chinese Speaker Teachers and Learners," *J. Psycholinguist. Res.*, vol. 53, no. 6, pp. 1–17, Dec. 2024, doi: [10.1007/s10936-024-10113-5](https://doi.org/10.1007/s10936-024-10113-5).
- [2] J. Wang and N. Halenko, "Developing Second Language Mandarin Fluency Through Pedagogic Intervention and Study Abroad: Planning Time, Speech Rate, and Response Duration," *Lang. Learn.*, vol. 74, no. 2, pp. 177–206, Dec. 2024, doi: [10.1111/lang.12694](https://doi.org/10.1111/lang.12694).
- [3] P. Xuan Lam, P. Q. H. Mai, Q. H. Nguyen, T. Pham, T. H. H. Nguyen, and T. H. Nguyen, "Enhancing Educational Evaluation Through Predictive Student Assessment Modeling," *Comput. Educ. Artif. Intell.*, vol. 6, p. 100244, Jun. 2024, doi: [10.1016/j.caeai.2024.100244](https://doi.org/10.1016/j.caeai.2024.100244).
- [4] L. R. Pelima, Y. Sukmana, and Y. Rosmansyah, "Predicting University Student Graduation Using Academic Performance and Machine Learning: A Systematic Literature Review," *IEEE Access*, vol. 12, pp. 23451–23465, 2024, doi: [10.1109/access.2024.3361479](https://doi.org/10.1109/access.2024.3361479).
- [5] S. Alturki, I. Hulpuş, and H. Stuckenschmidt, "Predicting Academic Outcomes: A Survey from 2007 Till 2018," *Technol. Knowl. Learn.*, vol. 27, no. 1, pp. 275–307, Mar. 2022, doi: [10.1007/s10758-020-09476-0](https://doi.org/10.1007/s10758-020-09476-0).
- [6] A. Mohebbi, "Enabling Learner Independence and Self-Regulation in Language Education Using AI Tools: A Systematic Review," *Cogent Educ.*, vol. 12, no. 1, Dec. 2025, doi: [10.1080/2331186x.2024.2433814](https://doi.org/10.1080/2331186x.2024.2433814).
- [7] I. T. Sanusi, S. S. Oyeler, H. Vartiainen, J. Suhonen, and M. Tukiainen, "A Systematic Review of Teaching and Learning Machine Learning in K-12 Education," *Educ. Inf. Technol.*, vol. 28, no. 5, pp. 5967–5997, May 2023, doi: [10.1007/s10639-022-11416-7](https://doi.org/10.1007/s10639-022-11416-7).
- [8] Y. Li and X. Cai, "Analysis of Factors Influencing Chinese Vocabulary Learning Level Based on Big Data Technology," *J. Intell. Fuzzy Syst.*, vol. 46, no. 2, pp. 3545–3556, Feb. 2024, doi: [10.3233/jifs-235515](https://doi.org/10.3233/jifs-235515).
- [9] A. K. H. Gujord, "Who Succeeds and Who Fails? Exploring the Role of Background Variables in Explaining the Outcomes of L2 Language Tests," *Lang. Test.*, vol. 40, no. 2, pp. 227–

- 248, Apr. 2023, doi: [10.1177/02655322221100115](https://doi.org/10.1177/02655322221100115).
- [10] G. Arkin, T. Abdukelim, H. Yilahun, and A. Hamdulla, "Machine Learning-Driven Acoustic Feature Classification and Pronunciation Assessment for Mandarin Learners," *Appl. Sci.*, vol. 15, no. 11, p. 6335, Jun. 2025, doi: [10.3390/app15116335](https://doi.org/10.3390/app15116335).
 - [11] Y. Zhang and J. Li, "Deep Learning-Based Model for Predicting Student Learning Behavior: A Pathway to Early Intervention and Enhanced Outcomes," *J. Comput. Methods Sci. Eng.*, vol. 25, no. 3, pp. 2822–2835, May 2025, doi: [10.1007/s11036-023-02187-7](https://doi.org/10.1007/s11036-023-02187-7).
 - [12] N. Gutiérrez, V. M. Rigobon, N. C. Marencin, A. A. Edwards, L. M. Steacy, and D. L. Compton, "Early Prediction of Reading Risk in Fourth Grade: A Combined Latent Class Analysis and Classification Tree Approach," *Sci. Stud. Read.*, vol. 27, no. 1, pp. 21–38, Jan. 2023, doi: [10.1080/10888438.2022.2121655](https://doi.org/10.1080/10888438.2022.2121655).
 - [13] J. Guo, Y. Liu, Q. Zou, L. Ye, S. Zhu, and H. Zhang, "Study on Optimization and Combination Strategy of multiple Daily Runoff Prediction Models Coupled with Physical Mechanism and LSTM," *J. Hydrol.*, vol. 624, p. 129969, Sep. 2023, doi: [10.1016/j.jhydrol.2023.129969](https://doi.org/10.1016/j.jhydrol.2023.129969).
 - [14] A. Mizumoto, "Calculating the Relative Importance of Multiple Regression Predictor Variables Using Dominance Analysis and Random Forests," *Lang. Learn.*, vol. 73, no. 1, pp. 161–196, Mar. 2023, doi: [10.1111/lang.12518](https://doi.org/10.1111/lang.12518).
 - [15] K. F. Yuen, X. Wang, T. Kyriazos, and M. Poga, "Application of Machine Learning Models in Social Sciences: Managing Nonlinear Relationships," *Encyclopedia*, vol. 4, no. 4, pp. 1790–1805, Nov. 2024, doi: [10.3390/encyclopedia4040118](https://doi.org/10.3390/encyclopedia4040118).
 - [16] F. Hatami, M. M. Rahman, B. Nikparvar, and J. C. Thill, "Non-Linear Associations Between the Urban Built Environment and Commuting Modal Split: A Random Forest Approach and SHAP Evaluation," *IEEE Access*, vol. 11, pp. 12648–12661, 2023, doi: [10.1109/access.2023.3241627](https://doi.org/10.1109/access.2023.3241627).
 - [17] E. Kalantari et al., "Evaluating Traditional Versus Ensemble Machine Learning Methods for Predicting Missing Data of Daily PM10 Concentration," *Atmos. Pollut. Res.*, vol. 15, no. 5, p. 102063, May 2024, doi: [10.1016/j.apr.2024.102063](https://doi.org/10.1016/j.apr.2024.102063).
 - [18] I. D. Mienye and Y. Sun, "A Survey of Ensemble Learning: Concepts, Algorithms, Applications, and Prospects," *IEEE Access*, vol. 10, pp. 99129–99149, 2022, doi: [10.1109/access.2022.3207287](https://doi.org/10.1109/access.2022.3207287).
 - [19] H. Li, Z. Gan, S. O. Leung, and Z. An, "The Impact of Reading Strategy Instruction on Reading Comprehension, Strategy Use, Motivation, and Self-Efficacy in Chinese University EFL Students," *SAGE Open*, vol. 12, no. 1, pp. 1–14, Mar. 2022, doi: [10.1177/21582440221086659](https://doi.org/10.1177/21582440221086659).
 - [20] M. J. Filderman, C. R. Austin, A. N. Boucher, K. O'Donnell, and E. A. Swanson, "A Meta-Analysis of the Effects of Reading Comprehension Interventions on the Reading Comprehension Outcomes of Struggling Readers in Third Through 12th Grades," *Except. Child.*, vol. 88, no. 2, pp. 163–184, Jan. 2022, doi: [10.1177/00144029211050860](https://doi.org/10.1177/00144029211050860).
 - [21] J. Zhou and N. Herencsar, "Abnormal Behavior Determination Model of Multimedia Classroom Students Based on Multi-task Deep Learning," *Mob. Networks Appl.*, vol. 28, no. 3, pp. 900–913, Jun. 2023, doi: [10.1007/s11036-023-02187-7](https://doi.org/10.1007/s11036-023-02187-7).
 - [22] J. E. Raffaghelli, M. E. Rodríguez, A. E. Guerrero-Roldán, and D. Bañeres, "Applying the UTAUT Model to Explain the Students' Acceptance of an Early Warning System in Higher Education," *Comput. Educ.*, vol. 182, p. 104468, Jun. 2022, doi: [10.1016/j.compedu.2022.104468](https://doi.org/10.1016/j.compedu.2022.104468).
 - [23] A. Hawrot, "Do School-Related Factors Affect Private Tutoring Attendance? Predictors of Private Tutoring in Maths and German among German Tenth-Graders," *Res. Pap. Educ.*, vol. 39, no. 1, pp. 1–23, Jan. 2024, doi: [10.1080/02671522.2022.2089209](https://doi.org/10.1080/02671522.2022.2089209).
 - [24] K. D. Thompson and C. Rodriguez-Mojica, "Individualized Language Plans: A Potential Tool for Collaboration to Support Multilingual Students," *J. Educ. Students Placed Risk*, vol. 28, no. 1, pp. 97–121, 2023, doi: [10.1080/10824669.2022.2123330](https://doi.org/10.1080/10824669.2022.2123330).
 - [25] P. Li and Q. Xu, "Computational Modeling of Bilingual Language Learning: Current Models and Future Directions," *Lang. Learn.*, vol. 73, no. S2, pp. 17–64, Dec. 2023, doi: [10.1111/lang.12529](https://doi.org/10.1111/lang.12529).
 - [26] V. G. Costa and C. E. Pedreira, "Recent Advances in Decision Trees: An Updated Survey," *Artif. Intell. Rev.*, vol. 56, no. 5, pp. 4765–4800, Oct. 2022, doi: [10.1007/s10462-022-10275-5](https://doi.org/10.1007/s10462-022-10275-5).
 - [27] L. Li, "A Predictive Model of English Vocabulary Teaching Effectiveness Based on Bayesian Networks," *Appl. Math. Nonlinear Sci.*, vol. 9, no. 1, pp. 1–16, Jan. 2024, doi: [10.2478/amns-2024-2545](https://doi.org/10.2478/amns-2024-2545).
 - [28] H. Chen, L. Wu, J. Chen, W. Lu, and J. Ding, "A Comparative Study of Automated Legal Text Classification Using Random Forests and Deep Learning," *Inf. Process. Manag.*, vol. 59, no. 2, p. 102798, Mar. 2022, doi: [10.1016/j.ipm.2021.102798](https://doi.org/10.1016/j.ipm.2021.102798).
 - [29] Z. Quan and L. Pu, "An Improved Accurate Classification Method for Online Education Resources Based on Support Vector Machine (SVM): Algorithm and Experiment," *Educ. Inf. Technol.*, vol. 28, no. 7, pp. 8097–8111, Jul. 2023, doi: [10.1007/s10639-022-11514-6](https://doi.org/10.1007/s10639-022-11514-6).
 - [30] P. Mahajan, S. Uddin, F. Hajati, and M. A. Moni, "Ensemble Learning for Disease Prediction: A Review," *Healthcare*, vol. 11, no. 12, p. 1808, Jun. 2023, doi: [10.3390/healthcare11121808](https://doi.org/10.3390/healthcare11121808).
 - [31] M. Cai and X. Liao, "The Relationship Between Vocabulary Depth Knowledge, Word Reading, and Reading Comprehension in Chinese," *J. Exp. Child Psychol.*, vol. 244, p. 105951, Aug. 2024, doi: [10.1016/j.jecp.2024.105951](https://doi.org/10.1016/j.jecp.2024.105951).
 - [32] H. Qiao and A. Zhao, "Artificial Intelligence-Based Language Learning: Illuminating the Impact on Speaking Skills and Self-Regulation in Chinese EFL Context," *Front. Psychol.*, vol. 14, p. 1255594, Nov. 2023, doi: [10.3389/fpsyg.2023.1255594](https://doi.org/10.3389/fpsyg.2023.1255594).
 - [33] X. Xu, Z. Shi, N. A. Bos, and H. Wu, "Student Engagement and Learning Outcomes: An Empirical Study Applying a Four-Dimensional Framework," *Med. Educ. Online*, vol. 28, no. 1, p. 2268347, Dec. 2023, doi: [10.1080/10872981.2023.2268347](https://doi.org/10.1080/10872981.2023.2268347).
 - [34] J. A. I. S. Masood, N. S. Kalyan Chakravarthy, D. Asirvatham, M. Marjani, D. Abdulkareem Shafiq, and S. Nidamanuri, "A Hybrid Deep Learning Model to Predict High-Risk Students in Virtual Learning Environments," *IEEE Access*, vol. 12, pp. 103687–103703, 2024, doi: [10.1109/ACCESS.2024.3434644](https://doi.org/10.1109/ACCESS.2024.3434644).
 - [35] C. Romero and S. Ventura, "Educational Data Mining and Learning Analytics: An Updated Survey," *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 10, no. 3, p. 1355, May 2020, doi: [10.1002/widm.1355](https://doi.org/10.1002/widm.1355).