Proceedings of Machine Learning Research vol 278:1–9, 2025 2025 2nd International Conference on Machine Learning and Intelligent Computing

Analysis of Learning Factors and Academic Performance of Non-Elite Students Using Machine Learning Models

Guangda Yang* 20211134@LISE.EDU.CN
College of Sciences, Liaoning Institute of Science and Engineering, Kunming Street, High-tech Industrial Park, Jinzhou, Liaoning Province, China

Hongfei Zhang
College of Sciences, Liaoning Institute of Science and Engineering, Kunming Street, High-tech Industrial Park, Jinzhou, Liaoning Province, China

Yongjiao Pang
College of Sciences, Liaoning Institute of Science and Engineering, Kunming Street, High-tech Industrial Park, Jinzhou, Liaoning Province, China

Suning Luo
College of Sciences, Liaoning Institute of Science and Engineering, Kunming Street, High-tech Industrial Park, Jinzhou, Liaoning Province, China
Corresponding author

Editors: Nianyin Zeng, Ram Bilas Pachori and Dongshu Wang

Abstract

This study applies machine learning models, including logistic regression, random forest, support vector machine (SVM), and XGBoost, to analyze and predict the final grades of non-elite university science students. The data includes attendance, note-taking scores, homework scores, quiz scores, and screen-cutting behavior. The results indicate that quiz scores, homework scores, and note-taking scores are the key factors for predicting final grades, with a particular impact from midterm and pre-final quiz scores. SVM performs well in predicting students at risk of failing, while random forest and XGBoost show stronger stability in handling complex data. Analysis of the importance analysis reveals that students' engagement, such as quiz and homework scores, is strongly correlated with final grades. The findings suggest that machine learning can effectively identify students at academic risk, providing data support for educational interventions. Future research could integrate more student behavior data and sentiment analysis to further improve prediction accuracy.

Keywords: Machine learning; Grade prediction; Academic risk; Non-elite universities; Educational intervention

1. Introduction

In today's educational field, predicting student academic performance and identifying students at risk of academic failure early have become key tasks for improving educational quality. As the education environment becomes more complex, students' academic performance is influenced not only by their intellectual abilities but also by various factors such as motivation and self-discipline. This is especially true in non-elite universities, where students often face challenges such as low entrance scores, poor motivation, and weak self-discipline, making academic performance more difficult to predict (Zhao and et al., 2024). These students often lack sufficient motivation and have

weak self-management skills, which makes them prone to academic difficulties and increases the complexity of educational management and intervention.

Effectively predicting student performance, especially identifying and intervening with students at academic risk, has become an important topic in education. Many studies suggest that early warning systems are crucial for improving students' academic success rates. By using scientific prediction models, educators can identify struggling students early in the semester and provide necessary support to help them improve their academic performance (Chung and Lee, 2019). In this context, machine learning and data mining technologies have become widely used tools in education. By analyzing students' historical grades, behavioral data, and other relevant information, researchers can develop accurate prediction models that provide data support to educators, helping them create more personalized teaching strategies (Kardan and et al., 2013; Nachouki and et al., 2023; Hai and et al., 2023).

However, most existing studies focus on students from elite schools or large-scale online education platforms, while research on academic performance prediction for students from non-elite universities, particularly in the field of science, is relatively scarce. Students from non-elite universities often face challenges such as low entrance scores, poor motivation, and weak self-discipline, leading to greater uncertainty in their academic performance. Therefore, developing prediction models suited to this specific group and identifying students at academic risk has become a major challenge in current educational research.

2. Literature Review

With the rapid development of machine learning technology, these techniques are increasingly being used in education to analyze student performance and learning behaviors, particularly in the research of academic performance prediction and early warning systems. Machine learning can analyze large amounts of student data to help educators identify potential academic issues, providing personalized interventions for students.

In the field of student performance prediction, many researchers have employed various machine learning algorithms. Kardan and et al. (2013) used a neural network model to analyze students' course selection behavior and successfully predicted academic performance and course enrollment numbers. Chung and Lee (2019) used methods like random forests to develop an early warning system for student dropout risks. These studies show that machine learning models can handle complex relationships within educational data and have achieved good results in predicting student performance.

For predicting failing students, Chui and et al. (2020) used a support vector machine model and achieved high classification accuracy. XGBoost, a gradient-boosted decision tree algorithm, has also achieved notable results in many educational research areas, particularly in modeling complex nonlinear relationships (Walid and et al., 2022; Hazzam and et al., 2024). Maheshwari and et al. (2024) investigated the effectiveness of various machine learning models, such as Logistic Regression, Random Forest, Support Vector Machine (SVM), and Neural Networks, in predicting the early academic performance of students in an online object-oriented programming course. Dervenis and et al. (2022) studied the prediction of student academic performance using algorithms like Random Forest, Support Vector Machine, and k-Nearest Neighbors. They found that the Random Forest model performed excellently across metrics such as accuracy, F1 score, precision, and recall.

Although existing studies have achieved significant results in academic performance prediction, most of the research focuses on students from elite schools or large-scale online education platforms. Research on academic performance prediction for non-elite university students, particularly in the field of science, is relatively scarce. Students in non-elite universities often face challenges such as low entrance scores, poor motivation, and weak self-discipline (Zhao and et al., 2024). These factors can lead to various challenges in their academic performance, requiring specific prediction models to address these special circumstances.

This study focuses on non-elite university science students and uses machine learning models to predict and analyze factors influencing academic performance. Compared to existing studies, this research targets a specific, challenging student group and combines various features such as students' regular grades and classroom performance for prediction. The aim is to provide effective intervention measures for educators. This research offers new insights for educational management in non-elite universities, particularly in student intervention and academic support, providing data-driven decision support.

3. Methodology

3.1. Data Collection and Preprocessing

The dataset used in our work was collected from non-elite university science students taught by the author. It includes academic records of 177 students, covering attendance, note scores, homework scores, quiz scores, screen-switching instances during quizzes, and final exam scores. The preprocessing steps are as follows:

1. Missing value handling: The time field for the fourth quiz had significantly missing data (approximately 40% of records). Since this temporal data had a minimal direct impact on final exam performance, it was excluded from modeling. Other features (e.g., note scores, homework scores, quiz scores) contained no missing values and were directly used.

2. Feature construction and standardization: Numerical features included note scores, homework scores, quiz scores, and screen-switching instances. Z-score standardization was applied to normalize features with different scales. For tree-based models (e.g., Random Forest, XGBoost), raw feature values were retained due to their insensitivity to feature scales.

3. Feature selection: Features strongly correlated with final exam scores (e.g., homework scores, quiz scores, attendance) were selected for model training based on correlation analysis.

3.2. Model Selection and Evaluation

As shown in figure 1, the following machine learning models were used to predict final exam scores:

1. Logistic Regression: Suitable for linear classification problems. Regression coefficients provide interpretability for feature impacts.

2. Random Forest: An ensemble method combining multiple decision trees. It captures complex nonlinear relationships.

3. Support Vector Machine (SVM): Identifies optimal classification boundaries by maximizing margins. Effective for high-dimensional feature spaces.

4. XGBoost: A gradient-boosted decision tree algorithm optimized for large datasets and non-linear patterns.

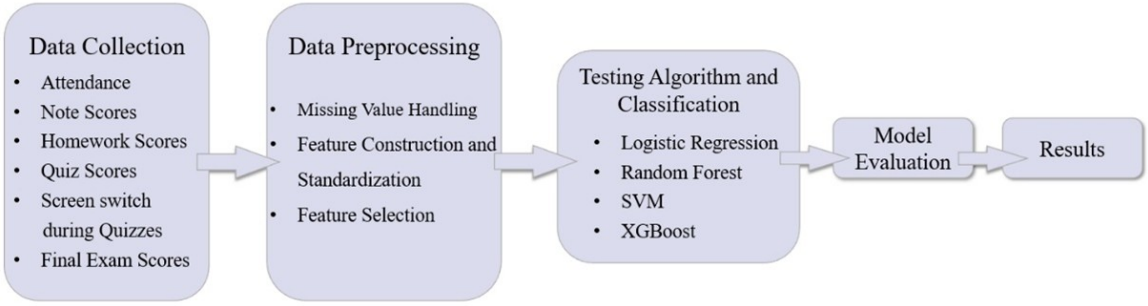

Figure 1: The procedure followed

4. Results

4.1. Data Description

The dataset in this study contains academic records of 177 students, including note scores, homework scores, quiz scores, attendance records, and final exam scores. Descriptive statistical analysis shows that the average final exam score was 64.24 points with a standard deviation of 14.05 points, indicating a dispersed score distribution. A significant proportion of students had low scores (32.2% failed). For regular assessments, homework and note scores were generally high, while quiz scores exhibited greater variability.

4.2. Correlation Analysis

As shown in table 1, the correlation analysis between academic indicators and final exam scores revealed positive relationships between regular homework/quiz scores and final exam performance. Specifically, Quiz 2 and Quiz 4 scores showed the strongest correlations with final exam scores (correlation coefficients: 0.25 and 0.27, respectively), suggesting midterm and pre-final quizzes strongly predict final outcomes. Note scores and attendance frequency also exhibited weaker positive correlations with final exam scores (correlation coefficients: 0.24 and 0.23).

4.3. Model Evaluation

As shown in figure 2 and figure 3, logistic regression, random forest, support vector machine, and XGBoost models demonstrated varying capabilities in predicting final exam scores:

1. Logistic Regression: Accuracy is 0.61, F1-score is 0.58. The model performed well in predicting passing students (Category 0) but showed low recall for failing students (Category 1), failing to identify all at-risk students.

2. Random Forest: Accuracy is 0.64, F1-score is 0.60. It achieved strong performance for Category 0 predictions but exhibited under detection (missed cases) for Category 1.

3. Support Vector Machine (SVM): Accuracy is 0.67, F1-score is 0.65. The SVM model outperformed others in recall for Category 1, indicating better identification of failing students. However, its precision was relatively low.

4. XGBoost: Accuracy is 0.64, F1-score is 0.58. While stable for large datasets, it underperformed SVM in predicting failing students.

Table 1: Correlation coefficients between indicators and final exam scores

Parameter	Correlation Coefficient	Parameter	Correlation Coefficient
Note 2	0.289	Homework 2	0.210
Quiz 4	0.271	Quiz 3	0.070
Homework 1	0.259	Quiz 1	0.069
Note 1	0.255	Screen exits during Quiz 2	0.052
Quiz 2	0.250	Screen exits during Quiz 4	0.006
Attendance count	0.240	Screen exits during Quiz 3	-0.056
Homework 3	0.231	Screen exits during Quiz 1	-0.137
Homework 4	0.217		

Note: The table is sorted in descending order by the absolute values of correlation coefficients. A positive correlation indicates that students with higher scores on an indicator tended to achieve higher final exam scores; a negative correlation indicates that higher values of an indicator were linked to lower final exam scores. "Screen exits" refer to the number of times a student exited the exam interface during online quizzes.

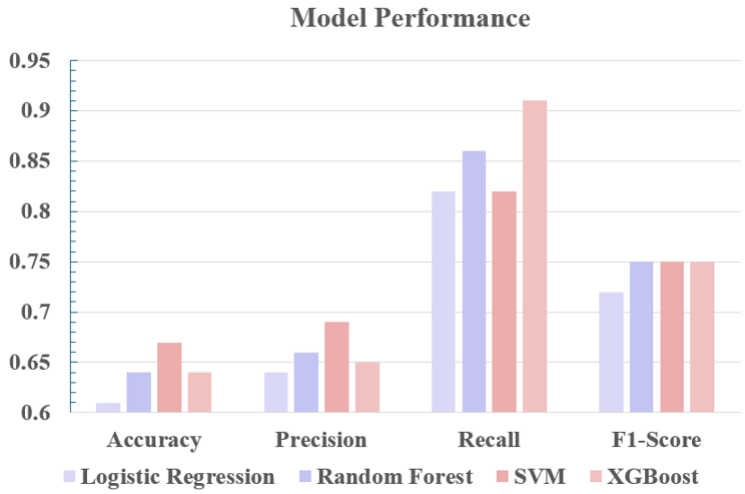

Figure 2: Logistic Regression, Random Forest, SVM and XGBoost Model Performance

4.4. Feature Importance Analysis

By analyzing feature importance across models as shown in figure 4, key factors influence final exam score predictions are identified:

Logistic Regression: Quiz scores (especially Quiz 2 and Quiz 4) had the strongest impact, followed by homework scores and Note Check Score 1.

Random Forest: Quiz scores (notably Quiz 4 and Quiz 2) and homework scores demonstrated higher importance.

SVM: Attendance records (particularly Attendance 11) and quiz scores (especially Quiz 4) received significant weights.

XGBoost: Attendance records and quiz scores (notably Quiz 4) played critical roles in predictions.

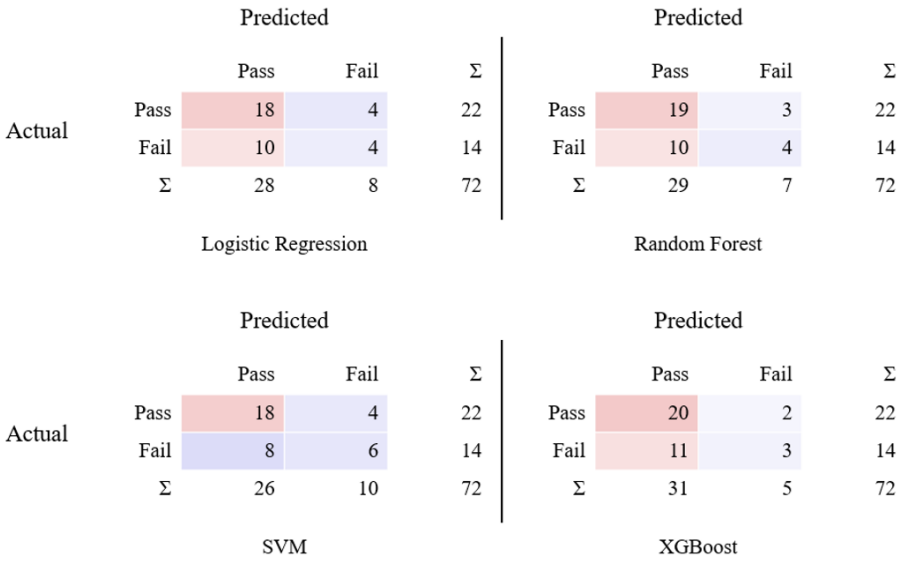

Figure 3: Logistic Regression, Random Forest, SVM and XGBoost Confusion matrix

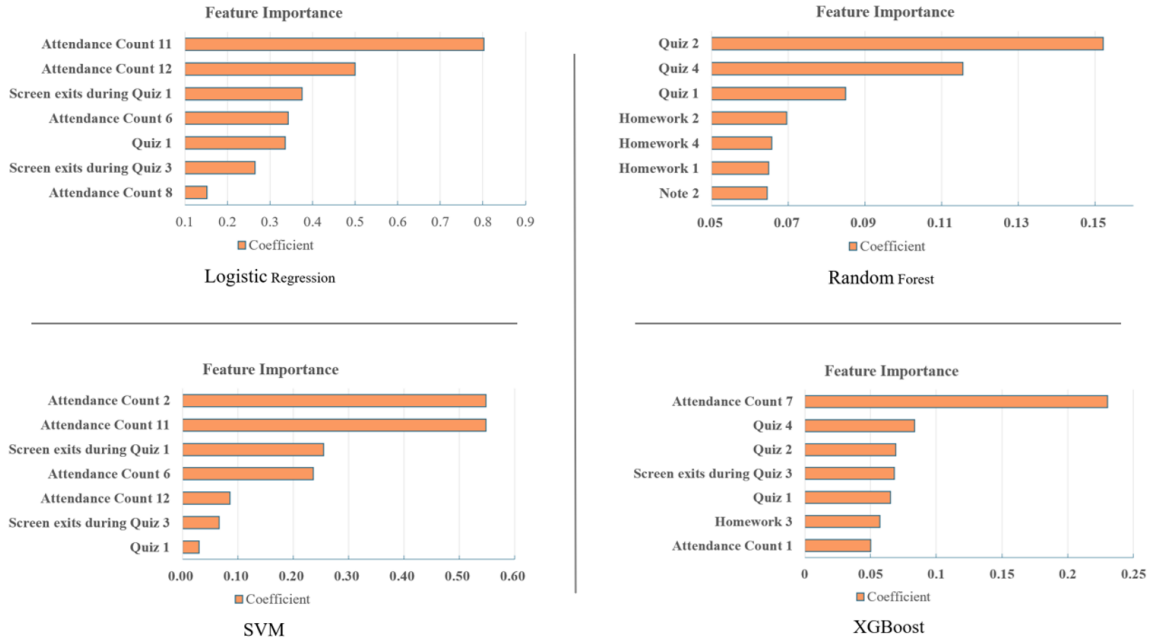

Figure 4: Top seven features with the highest coefficients for Logistic Regression, Random Forest, SVM and XGBoost

5. Discussion

5.1. Summary of Results

The model evaluation results indicate that all machine learning models demonstrated some capability in predicting final exam scores. However, they generally lacked sufficient recall, particularly in identifying failing students. The logistic regression model achieved higher precision but failed to effectively detect all at-risk students due to its low recall. Random Forest and XGBoost models, despite handling nonlinear relationships, still missed some failing cases. The Support Vector Machine (SVM) model showed more balanced performance with a higher recall.

5.2. Analysis of Influencing Factors

The results of feature importance analysis indicate that quiz scores, especially the mid-term quiz (Quiz 2) and pre-final quiz (Quiz 4), play a crucial role in predicting the final exam scores.

1. Impact of Quiz Scores

In all machine learning models, quiz scores (especially Quiz 2 and Quiz 4) ranked highly in terms of feature importance. This phenomenon can be explained in several ways: the mid-term quiz is often an important checkpoint in the course, allowing students to assess their mastery of the course material. The pre-final quiz, on the other hand, provides feedback on the content covered in the final stages of the course, helping students adjust their study strategies in a timely manner. Therefore, quiz scores reflect students' understanding of the material and have a direct impact on their final exam scores.

2. Impact of Homework and Note-Taking Scores

Homework scores reflect students' ability to learn independently outside of class, as well as their ongoing attention to the course material. Note-taking scores indirectly reflect students' concentration and engagement in class. Through these features, machine learning models can identify students who actively participate in class, complete their homework independently, and take thorough notes. These students typically have strong motivation and self-discipline, which helps them perform well on the final exam. Specifically, students with higher homework and note-taking scores demonstrate higher levels of engagement, which positively impacts their academic performance.

3. Impact of Attendance

Although the impact of attendance is relatively weaker compared to quiz and homework scores, it still reflects students' learning attitude and participation. High attendance is often associated with active participation in class and higher levels of classroom interaction, both of which help students understand and master the course content. Therefore, students with higher attendance rates tend to perform better on the final exam.

4. Impact of Screen-Cutting Behavior During Quizzes

The frequency of screen-cutting during quizzes is a feature we explored as a potential influencing factor. Although its correlation is relatively low, it still holds some predictive value. Screen-cutting behavior typically indicates that students are distracted during the exam, possibly due to insufficient preparation, consulting with peers for answers, or feeling anxious and nervous during the exam. Studies have shown that students who frequently cut the screen during quizzes tend to score lower, suggesting that their level of focus is lower or they have not effectively mastered the exam content. Therefore, screen-cutting behavior can serve as an indirect indicator of students' engagement and preparation for the exam.

6. Conclusion

This study analyzed and predicted final exam scores for non-elite university science students using logistic regression, random forest, support vector machine (SVM), and XGBoost models. The results indicate that quiz scores, homework scores, and note scores are key predictors of final exam performance, with midterm (Quiz 2) and pre-final (Quiz 4) quiz scores exerting the strongest influence. The SVM model performed better in identifying failing students, while random forest and XGBoost demonstrated stronger stability in handling complex data. Feature importance analysis revealed a significant positive correlation between students' learning engagement (e.g., quiz and homework performance) and final exam outcomes. This study provides educators with a predictive tool for early identification of at-risk students and targeted interventions. Future research could integrate additional student behavioral data to optimize model accuracy and further refine educational support strategies.

Acknowledgments

This work was supported by the Fundamental Research Project of the Education Department of Liaoning Province (LJ212413217009).

References

K. T. Chui and et al. Predicting at-risk university students in a virtual learning environment via a machine learning algorithm. *Computers in Human Behavior*, 107:105584, 2020.

J. Y. Chung and S. Lee. Dropout early warning systems for high school students using machine learning. *Children and Youth Services Review*, 96:346–353, 2019.

C. Dervenis and et al. Predicting students' performance using machine learning algorithms. In *Proceedings of the 6th international conference on algorithms, computing and systems*, 2022.

T. Hai and et al. Academic performance prediction using machine learning algorithms. In *International Conference on Advances in Communication Technology and Computer Engineering*. Springer Nature Switzerland, 2023.

J. Hazzam and et al. The influence of linkedin group community on postgraduate student experience, satisfaction and grades. *Computers & Education*, 216:105052, 2024.

A. A. Kardan and et al. Prediction of student course selection in online higher education institutes using neural network. *Computers & Education*, 65:1–11, 2013.

A. Maheshwari and et al. Comparative analysis of machine learning models in predicting academic outcomes: insights and implications for educational data analytics. In *2024 International Conference on Smart Systems for applications in Electrical Sciences (ICSSES)*. IEEE, 2024.

M. Nachouki and et al. Student course grade prediction using the random forest algorithm: Analysis of predictors' importance. *Trends in Neuroscience and Education*, 33:100214, 2023.

M. A. A. Walid and et al. Analysis of machine learning strategies for prediction of passing undergraduate admission test. *International Journal of Information Management Data Insights*, 2(2): 100111, 2022.

J.H. Zhao and et al. Impact of pre-knowledge and engagement in robot-supported collaborative learning through using the icapb model. *Computers & Education*, 217:105069, 2024.