Proceedings of the 5th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/112/2024.17909

Analyzing Student Online Learning Behaviors and Academic Performance in Science Education Using Machine Learning Techniques

Cheng Feng[1,a,*]

[1]*Lingnan University, Hong Kong, China*
a. fengcheng000722@163.com
**corresponding author*

Abstract: This study investigates the factors influencing student engagement and performance in online science education through the application of machine learning models, specifically Random Forests, Decision Trees, and Support Vector Machines (SVM). With the rapid growth of online education, understanding students' adaptability and learning behaviors has become increasingly critical. A systematic analysis of features such as study duration, daily study habits, and demographic factors revealed significant insights into their impact on academic achievement in science subjects. The Random Forest model outperformed others in classification accuracy, achieving an accuracy of 81%. The findings emphasize the importance of tailored educational strategies that foster consistent study practices and address the unique needs of diverse learners, ultimately enhancing learning outcomes in online science education.

Keywords: Online Learning, Science Subjects, Machine Learning, Academic Performance.

1. Introduction

In the digital era, the evolution of information technology has catalyzed significant innovation in educational paradigms, with online learning emerging as a pivotal segment within the field of education[1]. The 50th Statistical Report on Internet Development in China, published in 2022, indicates a substantial growth in the user base of online education in China, expanding from 110 million in 2015 to 377 million by 2022. Despite the evident advantages of online education, the industry confronts numerous challenges[2].

Ability to online learning is defined as the learner's capacity to actively modify their approach in response to changes in personal and environmental conditions during the process of knowledge acquisition via the internet, thereby aligning their development with the learning context and achieving educational objectives[3]. It is one of the critical factors influencing the development of online education. Although current academic research extensively covers various aspects of adolescent online learning adaptability, including individual, familial, scholastic, and community dimensions, there is a notable absence of in-depth analysis on how these factors interrelate and collectively affect adolescents' acceptance of online education.

The rise of online education presents unique challenges in analyzing student behaviors, especially in science subjects. This study examines engagement patterns in online science courses, identifying

DOI: 10.54254/2755-2721/112/2024.17909

behaviors that enhance learning outcomes, aiming to provide educators with actionable insights for more effective, tailored teaching strategies.

2. Related work

Recent advancements in machine learning have significantly impacted science education, particularly in assessment and instructional methodologies. Zhai et al. (2020) conducted a systematic review highlighting the application of machine learning techniques in science assessments, emphasizing their potential to improve the accuracy and efficiency of scoring processes [4]. The study reveals that integrating machine learning can address traditional assessment challenges, enabling more nuanced evaluations of student performance and understanding in scientific disciplines. Additionally, Zhai, Shi, and Nehm (2021) performed a meta-analysis that focused on factors influencing the agreement between machine-generated scores and human evaluations [5]. Their findings suggest that various parameters, including the complexity of assessment tasks and the nature of student responses, play crucial roles in score alignment, which has implications for developing reliable automated assessment systems.

Further expanding on the role of artificial intelligence in education, Almasri (2024) explored empirical research on the impact of AI in teaching and learning science [6]. The systematic review indicated that AI tools enhance student engagement, facilitate personalized learning experiences, and improve learning outcomes in science education. This aligns with findings by Maestrales et al., who utilized machine learning to assess multi-dimensional evaluations in chemistry and physics [7-8]. Their work demonstrates that machine learning can effectively handle complex assessment frameworks, providing accurate and scalable scoring systems that benefit both educators and learners.

The literature also delves into specific machine learning methodologies that have shown promise in educational contexts. Breiman's seminal work on Random Forests presents a robust framework for classification and regression tasks, highlighting its ability to manage large datasets with numerous features, making it particularly relevant for analyzing educational data [9]. Biau and Scornet provided an accessible overview of Random Forests, emphasizing their practical applications in various fields, including education [10]. Similarly, Quinlan and Song & Ying discussed decision tree classifiers, illustrating their straightforward interpretability and effectiveness for classification tasks in educational settings [11-12]. These models provide valuable insights into how different factors influence student learning outcomes, especially when combined with data-driven approaches.

Support Vector Machines (SVMs) also feature prominently in the literature as a powerful classification tool. Jakkula provided a tutorial on SVMs, detailing their theoretical underpinnings and practical applications, while Schuldt et al. demonstrated SVMs' effectiveness in recognizing patterns, which can be analogously applied to educational data analysis [13-14]. This review highlights the transformative potential of machine learning in science assessments, showcasing various approaches and their applications. As the field continues to evolve, the integration of these advanced methodologies will likely pave the way for more effective educational practices and improved student outcomes in STEM disciplines.

3. Data and Model

3.1. Feature Description

Features are key data attributes that help machine learning models analyze learning behaviors and predict academic outcomes. In this study, features like gender, age, education level, internet access, and study habits reveal insights into students' backgrounds and learning environments, enhancing model accuracy in predicting academic performance.

DOI: 10.54254/2755-2721/112/2024.17909

The target variable, or dependent variable, is the outcome that the model aims to predict or classify. In this case, academic performance in subjects like mathematics, physics, chemistry, and biology is the target variable, representing students' achievement levels in online learning. The model's task is to predict or classify students' performance based on the selected features, enabling targeted recommendations and actionable insights for educators to enhance online learning strategies and support academic success.

3.2. Random Forest model

Random Forest is an ensemble learning algorithm that improves classification by combining multiple decision trees trained on random subsets of samples and features. Using "bagging" and "feature randomness," it reduces overfitting and enhances accuracy. In this study, Random Forest is valuable for analyzing student performance in online STEM courses, as it minimizes errors, handles high-dimensional data, and uncovers complex patterns in factors like gender, age, study habits, and internet quality, providing insights into learning outcomes.

Another critical advantage of Random Forest is its interpretability, as it can output feature importance scores, indicating which factors are most significant for predicting performance. This is invaluable for educational research, as it provides actionable insights into the key influences on student success. Random Forest's resilience to missing values and outliers is also beneficial, given the variability common in educational datasets. Overall, this model's robust classification capabilities make it well-suited for offering data-driven recommendations to optimize online learning strategies.

3.3. Decision Tree model

A Decision Tree is a machine learning model that splits data into subsets based on feature values, creating a tree-like structure. Each node tests a feature, with branches leading to further splits until a final classification decision is made. Its simplicity and interpretability make it valuable for classification tasks. In this study, Decision Trees are used to analyze factors influencing student performance in online STEM courses. Their structure handles both numerical and categorical data, capturing relationships between features like gender, age, study habits, and educational settings. Decision Trees are non-parametric and can model complex, non-linear relationships, making them effective in understanding the varied factors affecting online learning success. Their interpretability is a major advantage, allowing educators to visualize how features impact outcomes and refine learning strategies. Although prone to overfitting, techniques like pruning improve generalizability, making Decision Trees a clear and effective approach to predicting student performance in online courses.

3.4. SVM model

Support Vector Machines (SVM) are a supervised learning algorithm used for classification. SVMs find the optimal hyperplane to separate classes in high-dimensional space. When data is not linearly separable, SVMs use the "kernel trick" to map data to a higher-dimensional space, enabling effective classification of complex boundaries. In this study's context of classifying factors that impact student performance in online STEM courses, SVMs offer several strengths. First, SVMs are robust in handling high-dimensional data, allowing them to model complex relationships between features like gender, age, study habits, and internet access. Additionally, by maximizing the margin between classes, SVMs often achieve high classification accuracy, even with smaller datasets, making them suitable for analyzing diverse educational datasets with intricate patterns.

Another key advantage of SVMs is their resistance to overfitting, particularly when using regularization techniques. This quality is beneficial in educational research, where noise in data can

DOI: 10.54254/2755-2721/112/2024.17909

arise from a variety of external factors. Although SVMs can be computationally intensive, their effectiveness in capturing subtle class distinctions offers reliable insights, making them a valuable tool for developing tailored educational strategies based on student behavior and performance data.

4. Model Establishment and Evaluation

4.1. Data Preprocessing

In data preprocessing, label encoding was used to convert categorical features into numerical codes, enabling the model to process them effectively. The dataset was split into training and testing sets, with 70% (843 samples) for training and 30% (362 samples) for testing. This partitioning ensures the model has enough data for both training and evaluation, helping assess performance effectively.

4.2. Cross-Validation Results and Comparison of Models

In this study, the three selected machine learning models underwent 5-fold cross-validation, and their average accuracy on the validation set was calculated. Figure 1-3 illustrates the results of the three models. The results indicate that the random forest model performed best in cross-validation, achieving an accuracy of 0.81, followed by SVM models with 0.79 , while the decision tree model showed relatively poor performance at 0.69. This suggests that the random forest model may be more appropriate for constructing the online science subject education prediction model.The random forest model demonstrates excellent performance in handling high-dimensional features and large-scale data, making it effective for prediction.

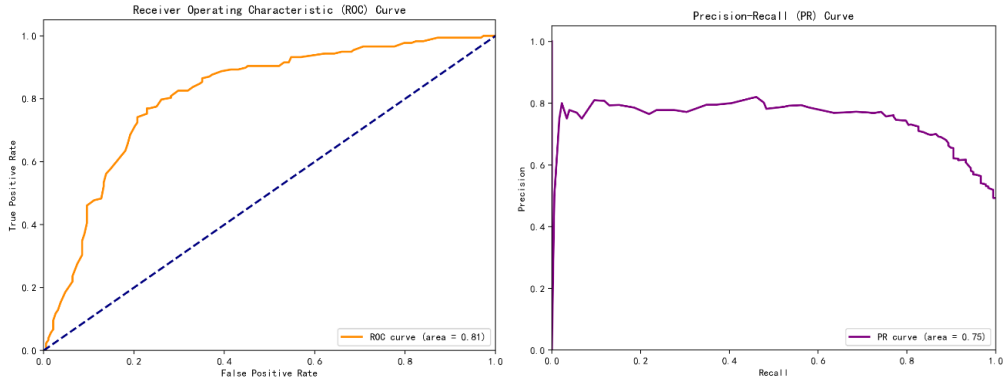

Figure 1: RF model results

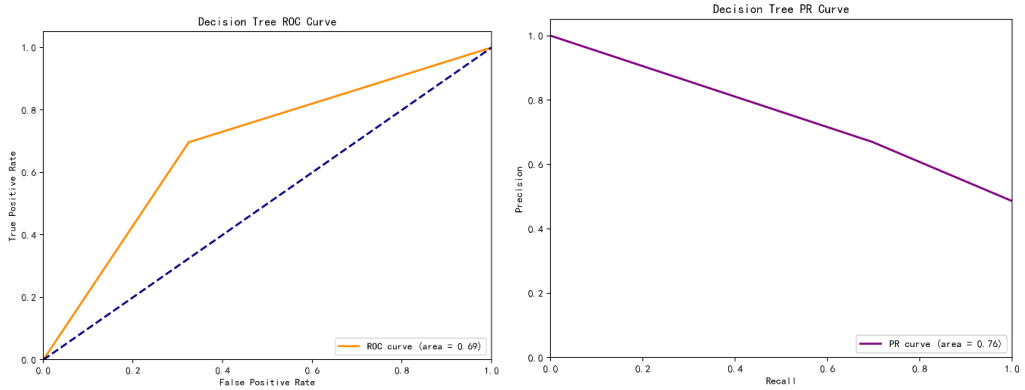

Figure 2: DT model results

DOI: 10.54254/2755-2721/112/2024.17909

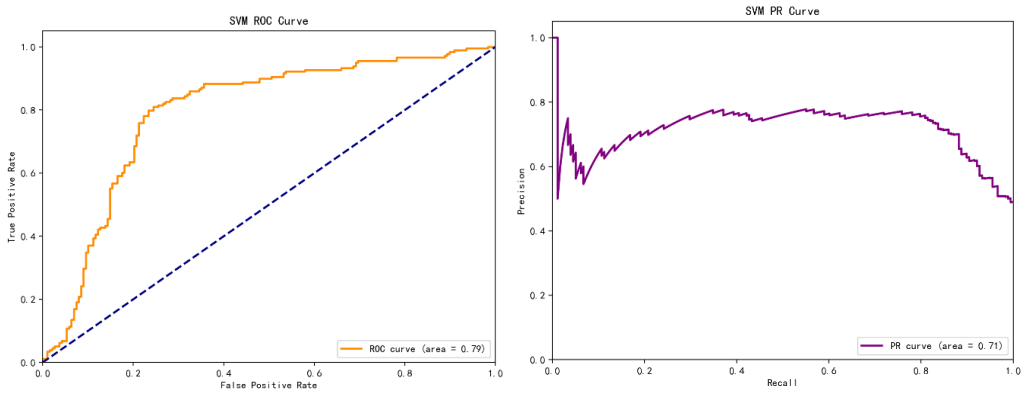

Figure 3: SVM model results

5. Discussion and Analysis of Results

Figure 4 presents the ranking of feature importance based on the random forest model, allowing for an analysis of the influence of various features on adaptability. The results reveal that "srudy duration," "insist on daily study," and "gender" have the most significant impact on science subject achievement. This suggests thatstudents who dedicate more time to their studies and maintain consistent daily study habits tend to perform better in science subjects. Additionally, the results indicate that gender also plays a role in influencing achievement, with potential differences in learning styles or preferences between males and females that could contribute to varying outcomes.

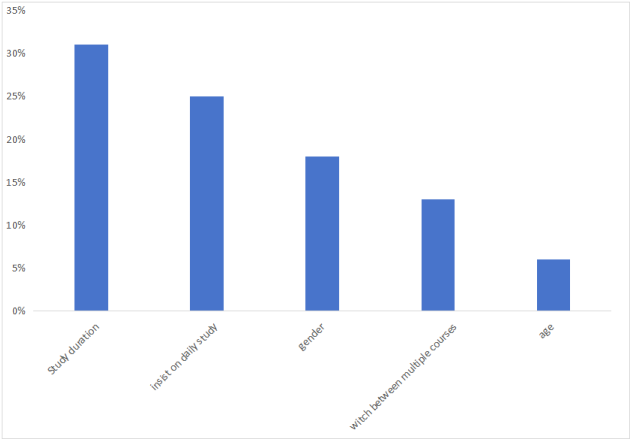

Figure 4: Ranking of Feature Importance

Delving deeper into the feature importance, further analysis can be conducted to understand how these factors interact with each other and their combined effect on student adaptability and achievement. For instance, it may be valuable to explore whether the impact of study duration is more pronounced for certain genders or if consistent study habits benefit one gender more than the other.

Moreover, additional features such as "teacher support," "learning environment," and "peer influence" also appear to have a noteworthy impact, though to a lesser extent. Understanding the role of these factors could provide a more comprehensive view of the elements that contribute to student success in science subjects.

By leveraging the insights gained from the random forest model, educators and policymakers can design targeted interventions and strategies that address the most influential factors. This could involve implementing programs that encourage consistent study habits, creating gender-sensitive learning approaches, or fostering a supportive educational environment.

DOI: 10.54254/2755-2721/112/2024.17909

6. Conclusion

In conclusion, this study highlights the transformative potential of machine learning techniques in analyzing student behaviors and academic performance in online science education. By employing Random Forests, Decision Trees, and Support Vector Machines, we identified key factors such as study duration and consistent study habits that significantly influence students' success in STEM subjects. The superior performance of the Random Forest model underscores its effectiveness in managing complex educational data and providing actionable insights for educators.

The results reveal the need for a data-driven approach in designing educational interventions that cater to the diverse needs of learners. As online education continues to evolve, integrating machine learning into educational research and practice can lead to more personalized learning experiences, ultimately fostering improved academic outcomes for students. Future research should further explore the interactions between various features and investigate additional dimensions that contribute to student adaptability and success in online learning environments.

Given these findings, several recommendations can be made for educators, policymakers, and curriculum developers:

Promote Consistent Study Habits: Educational programs should encourage students to develop consistent daily study routines. Strategies might include implementing structured schedules, providing reminders for study sessions, and integrating gamified elements that reward consistent engagement.

Tailor Learning Approaches to Gender Differences: The influence of gender on academic performance suggests that educational strategies should be sensitive to the differing learning styles and preferences among male and female students. Educators could develop targeted resources and activities that cater to these differences, fostering a more inclusive learning environment.

Enhance Support Mechanisms: Incorporating additional supportive features, such as mentorship programs or peer tutoring, can provide students with the guidance they need to navigate online learning challenges. By fostering a community of support, students may feel more motivated to engage with their studies. Explore Additional Variables: Investigating other factors that may influence student performance, such as the quality of instructional materials, student motivation, and external environmental factors, could yield valuable insights. This comprehensive approach will enable a more thorough understanding of the dynamics at play in online learning contexts.

References

[1] *Wu Wentao, Liu Hehai, Bai Qian. Building a Learning Society: Practicing Chinese Modernization through Educational Digitization [J]. Chinese Journal of Educational Technology, 2023, (03): 17-24+45.*

[2] *Zhou Hongwei. Research on the Adaptive Learning Model of Online Education Based on Educational Big Data and Its Application [J]. Research on Continuing Education, 2023, (03): 110-114.*

[3] *Tian Lan. Review of Research on Learning Adaptability of Primary and Secondary School Students in China [J]. Psychological Science, 2004, (02): 502-504.*

[4] *Zhai X, Yin Y, Pellegrino J W, et al. Applying machine learning in science assessment: a systematic review[J]. Studies in Science Education, 2020, 56(1): 111-151.*

[5] *Zhai X, Shi L, Nehm R H. A meta-analysis of machine learning-based science assessments: Factors impacting machine-human score agreements[J]. Journal of Science Education and Technology, 2021, 30: 361-379.*

[6] *Almasri F. Exploring the impact of artificial intelligence in teaching and learning of science: A systematic review of empirical research[J]. Research in Science Education, 2024, 54(5): 977-997.*

[7] *Maestrales S, Zhai X, Touitou I, et al. Using machine learning to score multi-dimensional assessments of chemistry and physics[J]. Journal of Science Education and Technology, 2021, 30: 239-254.*

[8] *Zhai X, Shi L. Understanding how the perceived usefulness of mobile technology impacts physics learning achievement: A pedagogical perspective[J]. Journal of Science Education and Technology, 2020, 29(6): 743-757.*

[9] *Breiman L. Random forests[J]. Machine learning, 2001, 45: 5-32.*

[10] *Biau G, Scornet E. A random forest guided tour[J]. Test, 2016, 25: 197-227.*

DOI: 10.54254/2755-2721/112/2024.17909

[11] Quinlan J R. Learning decision tree classifiers[J]. ACM Computing Surveys (CSUR), 1996, 28(1): 71-72.

[12] Song Y Y, Ying L U. Decision tree methods: applications for classification and prediction[J]. Shanghai archives of psychiatry, 2015, 27(2): 130.

[13] Jakkula V. Tutorial on support vector machine (svm)[J]. School of EECS, Washington State University, 2006, 37(2.5): 3.

[14] Schuldt C, Laptev I, Caputo B. Recognizing human actions: a local SVM approach[C]//Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004. IEEE, 2004, 3: 32-36.