Learning data analysis using educational data mining techniques

Taufik Hidayat 1[a], Rahmawati Fauziah 2[b], Aan 3[c], Tegar Siswa 4[d], Suciana Wijirahayu [e a]
Universitas Muhammadiyah Prof. DR. HAMKA, DK Jakarta, Indonesia
*Correspondence: sucianawijirahayu@uhamka.ac.id

Abstract
The main purpose of this research paper is to analyze Moodle data and identify the most influencing features to develop a predictive model. The research applies a wrapper-based feature selection method called Boruta for selecting the best predicting features. Data were collected from 81 students enrolled in the Human Computer Interaction (COMP341) course offered by the Department of Computer Science and Engineering. The dataset contained eight features: Assignment.Click, Chat.Click, File.Click, Forum.Click, System.Click, Url.Click, and Wiki.Click as independent features, and Grade as the dependent feature. Five classification algorithms, namely K Nearest Neighbour, Naïve Bayes, Support Vector Machine (SVM), Random Forest, and CART decision tree, were applied to the Moodle data. The findings show that SVM has the highest accuracy compared to other algorithms. It suggested that File.Click and System.Click were the most significant features. This type of research helps in the early identification of students' performance. The growing popularity of the teaching-learning process through an online learning system has attracted researchers to work in the field of Educational Data Mining (EDM)

Article History:

Keywords:
Moodle data
Human Computer Interaction
Educational Data Mining

1. Introduction

In today's digital age, data storage has become virtually limitless, spanning across various media including files, hard drives, CDs, databases, and numerous applications. This vast volume of information creates a challenge in retrieving specific, valuable data efficiently. Data Mining emerges as the solution to extract meaningful information from these enormous datasets.

Educational Data Mining (EDM) represents an innovative branch within the broader fields of data mining and Knowledge Discovery in Databases (KDD). This specialized approach focuses on uncovering valuable patterns and extracting relevant information from educational management systems. These systems encompass various aspects of academic administration, including:

- Student admissions
- Registration processes
- Course management
- Curriculum planning

- Other educational operations

EDM applications span across different educational levels, serving the diverse needs of schools, colleges, and universities. By analyzing data at various academic stages, EDM helps institutions make informed decisions and improve their educational processes.

Recent advancements in data mining have introduced various sophisticated methodologies. These key approaches include:

- Generalization
- Characterization
- Classification
- Clustering
- Association analysis
- Evolution tracking
- Pattern recognition
- Visual data representation
- Meta-rule guided mining

The field has evolved to accommodate knowledge extraction from diverse database structures, such as:

- Relational databases
- Transactional systems
- Object-oriented frameworks
- Spatial information systems
- Active databases
- Global information networks

Through a comprehensive analysis of academic literature spanning from 2000 to 2011, this research explores the significant developments in Data Mining Technology (DMT). The study examines both theoretical applications and practical implementations, while also highlighting emerging research opportunities and challenges in the field.

2. Methodology

a) Data Collection

Researchers gathered data from 81 students participating in the COMP341 course, focusing on their engagement within the Moodle learning platform. The dataset consisted of eight independent variables representing different types of student interactions, including:

- Assignment.Click: Interactions with assignment submissions.
- Chat.Click: Participation in chat activities.
- File.Click: Access to course files.
- Forum.Click: Engagement in forum discussions.
- System.Click: General navigation within Moodle.
- Url.Click: Visits to external links shared within the course.
- Wiki.Click: Contributions or views in collaborative wikis.

The primary focus of the analysis was the dependent variable, **Grade**, to understand how these interaction metrics correlated with academic performance.

b) Feature Selection

The Boruta algorithm is used for feature selection. This wrapper-based method evaluates the importance of features iteratively by comparing them with random versions (shadow features) to determine their significance.

c) Classification Algorithms

Five classification algorithms are applied:

- K Nearest Neighbor (KNN): A nonparametric method used for proximity-based classification.
- Naïve Bayes is a classification algorithm rooted in probability theory, which utilizes Bayes' theorem to predict outcomes. It operates under the assumption that all features in the dataset are independent of one another, simplifying the computation while maintaining effectiveness in various scenarios.
- Support Vector Machine (SVM): A supervised machine learning algorithm designed to handle classification tasks by identifying the optimal hyperplane that best separates data points from different classes in a high-dimensional space.
- Random Forest: An ensemble learning technique that combines the predictions of multiple decision trees. By aggregating the outputs of individual trees, it enhances predictive accuracy and reduces the risk of overfitting.
- CART Decision Tree: A classification method based on the "Classification and Regression Tree" (CART) algorithm, which constructs a binary tree structure to divide data into subsets based on feature values, facilitating efficient and interpretable classification.

Tables

Summary of Dataset Features

Table 1

Main Information

Featur	Description	Type
Assignment.Click	Number of assignment clicks	Independent
Chat.Click	Number of chat interactions	Independent
File.Click	Number of file resource clicks	Independent
Forum.Click	Number of forum interactions	Independent
System.Click	General Moodle system usage	Independent
Url.Click	Number of external URL clicks	Independent
Wiki.Click	Number of wiki page accesses	Independent
Grade	Final course grade	Dependent

3. Results and Discussion

3.1 Results

The analysis revealed that the Support Vector Machine (SVM) algorithmachieved the highest accuracy among all methods tested, demonstrating its reliability in predicting student performance. This finding highlights SVM's strength in effectively processing and classifying data related to student interactions within an online learning platform. Its superior performance indicates that SVM is particularly well-suited for educational data mining tasks, especially those involving intricate patterns and complex relationships within the dataset. The results suggest that SVM could play a crucial role in identifying students who might need extra assistance, paving the way for targeted interventions to improve learning outcomes and overall academic success.

3.2 Discussion

The analysis reveals that certain interaction metrics, particularly File.Click and System.Click, play a pivotal role in predicting student academic performance. These results underscore the critical influence of these specific activities within the online learning environment, positioning them as strong indicators of student success. Recognizing the importance of these metrics allows educators to focus on strategies that enhance student engagement with file resources and system interactions. For example, creating more dynamic and engaging file materials or simplifying system navigation could foster increased student participation. Such targeted interventions have the

potential to improve the effectiveness of online learning platforms, ultimately supporting students in achieving their academic objectives and enhancing overall educational outcomes.

4. Conclusion

This study highlights the potential of Educational Data Mining (EDM) techniques in revealing meaningful patterns within online learning data to predict student performance with precision. By examining student interactions on digital learning platforms, the research identifies critical features—such as specific actions taken by students—that significantly affect academic outcomes. These insights provide educators with valuable tools for crafting targeted interventions, particularly for students exhibiting low engagement or behaviors that may place them at risk academically.

For instance, when students show limited interaction with essential resources, such as course files or the overall system, educators can respond by offering additional guidance, redesigning course materials, or incorporating more interactive elements to enhance engagement. This strategic approach enables instructors to address specific issues affecting student participation and improve their overall learning experience.

Leveraging EDM insights in this way ensures that students who are struggling receive timely and focused support, increasing their likelihood of success. Beyond its immediate applications, this research underscores the transformative potential of data-driven methods in advancing online education. By personalizing the learning process and optimizing teaching strategies, EDM can significantly elevate the quality and effectiveness of digital education, paving the way for innovative and impactful learning solutions.

References

Romero, C., & Ventura, S. (2008). Data mining in course management systems: Moodle case study and tutorial.

Alagib Alsuddig Hamza, H., & Kommers, P. (2018). A Review of Educational Data Mining Tools & Techniques. International Journal of Educational Technology and Learning, 3(1), 17-23.

Aslan, A. (2021). Problem-Solving in Educational Data Mining: A Review.
Donoghue, J., et al. (2021). Active Learning Strategies for Teaching Big Data Analytics.

Shi, Y. (2022). The Role of Big Data in Education: Opportunities and Challenges.
Kularbphettong, K., et al. (2012). Comparison of Classification Algorithms for Predicting Student Performance.

Waskom, M.L. (2021). Seaborn: Statistical Data Visualization.

David J. Lemay , Clare Baek b , Tenzin Doleck. Computers and Education: Artificial Intelligence. Volume 2, 2021, 100016. www.elsevier.com/locate/caeai

International Journal of Intelligent Systems and Applications in Engineering Team (2024). Analysis of Student's Education Data Based on Data Mining Techniques.

Sazol Sarker, Mahit Kumar Paul , Sheikh Tasnimul Hasan Thasin, Md. Al Mehedi Hasan. Analyzing students' academic performance using educational data mining. Artificial Intelligence 7 (2024) 100263. www.sciencedirect.com/journal/computers-and-education-artificial-intelligence

Luisa Barbeiro , Anabela Gomes , Fernanda Brito Correia, Jorge Bernardino. A Review of Educational Data Mining Trends. www.sciencedirect.com

Lamya F, Daghestani Lamiaa F, Ibrahim , Reem S. Al-Towirgi, Hesham, Salman. Adapting gamified learning systems using educational data mining techniques. 1 March 2020. Research Article

Akarshita Tripathi, Mr. Amit Kumar. IJCSMC, Analysis of Educational Data Mining Techniques. Vol.8 Issue.1, January- 2019, pg. 8-15. www.ijcsmc.com

Anduela Lile, Epoka University, Tirana, Albania. Analyzing E-Learning Systems Using Educational Data Mining Techniques. Vol. 2 (3) September 2011. www.mcser.org

Agung Triayudi and Wahyu Oktri Widyarto. Educational Data Mining Analysis Using Classification Techniques. Science and Technology (ViCEST) 2020. IOP Publishing

Syifa Faradilla Fabrianne, Agung Triayudi, Ira Diana Sholihati. Data mining using filtering approaches and ensemble methods. Annual Conference on Computer Science and Engineering Technology (AC2SET) 2020. https://iopscience.iop.org/article/10.1088/1757-899X/1088/1/012012

P. Bachhal1, S. Ahuja1 and S. Gargrish. Educational Data Mining: A Review. ICMAI 2021. https://iopscience.iop.org/article/10.1088/1742-6596/1950/1/012022