# Can we predict student performance based on tabular and textual data

[1] Kandhati Saivarsha[2] Mandala Saikumar[3] Kati Caleb[4] K.Raveendra kumar
*[1, 2, 3,] UG Scholars, [4] Assistant Professor*
*[1,2,3, 4] Department of Computer Science and Engineering*
*[1,2,3,4]Guru Nanak Institutions Technical Campus, Hyderabad, Telanagana, India*

**Abstract- With the rise of teaching systems like MOOCs, massive amounts of educational data, including student behavior metrics and course comments, are being generated but remain underutilized for uncovering useful models for school management. To address this, we collected a multi model data set combining tabular student behavior data with textual course comments and proposed a Transformer-based framework to fuse these data types into a uniform vector representation for predicting student performance. Empirical results show that our approach significantly improves prediction accuracy, with F1-scores increasing by up to 3.33% and AUC by up to 4.37% compared to existing methods. Validation on an open data set confirmed the framework's strong generalization capability. Using the SHAP method for interpret-ability, we found that textual features have a greater influence on the classification model, further demonstrating the value of integrating text data. These findings suggest that fusing behavioral and textual features not only improves model performance but also provides actionable insights for educational data mining.**

## INTRODUCTION

Educational Data Mining (EDM): Transforming Education with Data

Educational Data Mining (EDM) leverages data mining techniques such as clustering, classification, and text mining to analyze diverse datasets from educational environments. These datasets come from traditional classrooms, computer-based systems, and online learning platforms, including MOOCs. With the advent of web-based education, data from various sources, including student behaviors, textual comments, and multi model inputs like audio and video, have grown exponentially. This expansion presents immense opportunities to enhance educational management and improve learning outcomes.

EDM focuses on thirteen primary tasks, including predicting student performance, detecting undesirable behaviors, profiling and grouping students, planning and scheduling, generating reports and alerts, and developing adaptive systems. Sentiment analysis, a key technique in EDM, uses text mining to assess students' emotional states and their relationship to learning outcomes. For instance, research on MOOCs posts identified a correlation between negative sentiments and higher dropout rates. Beyond sentiment analysis, multi model data from various sources, such as behavioral logs, textual data, and even brainwave signals, enable deeper insights into student experiences.

Datasets like Data Shop and tools like GISMO have facilitated EDM research by offering structured data and interactive monitoring tools for online education systems such as Moodle. Data Shop, a pioneering data set, logs student-tutor interactions, while GISMO visualizes student activities, including course attendance and assignment submissions. The MUTLA data set, another milestone, incorporates multi model inputs such as brainwave data, video records, and question-level logs for comprehensive analysis. Despite these advancements, access to rich multi model educational datasets remains a challenge.

To bridge this gap, researchers have collected multi model datasets from MOOCs, focusing on structured behavioral data and unstructured textual data like course comments. Course comments were chosen due to their widespread availability across platforms, ensuring cost-effective data collection. By aligning these datasets manually, researchers have developed multi model data fusion techniques, integrating diverse data types into unified semantic representations. This approach improves predictions of student performance and other educational outcomes.

Empirical studies validate these methods, showing significant improvements in classification metrics like recall, F1, and AUC. By addressing the challenges of heterogeneous data mining, EDM continues to unlock transformation insights, paving the way for more personalized and effective educational experiences

## RELATED WORK

Research in Educational Data Mining (EDM) has focused on applying machine learning and data analysis techniques to uncover patterns and insights from educational datasets. Early studies explored the use of clustering and classification techniques to predict student performance and identify at-risk students. Tools like Data-shop is one of the world largest repository of learning interaction data , provided a framework for analyzing student interactions with intelligent tutoring systems, enabling the development of predictive models for learning outcomes.

GISMO, a tool integrated with the Moodle learning management system, allows educators to visualize student engagement through metrics like attendance, material access, and assignment submissions. Such tools have helped instructors gain actionable insights into online student behavior. The emergence of MOOCs further expanded EDM's scope by providing large-scale datasets. Yang et al. applied sentiment analysis to students' posts on MOOCs, revealing a negative correlation between positive sentiment ratios and dropout rates.

Recent efforts have focused on multi model data fusion to integrate heterogeneous datasets. The MUTLA data set is notable for its combination of textual, behavioral, and physiological data, including brainwave signals, enabling deeper analysis of learning processes. However, MUTLA's unavailability has highlighted the need for more open multi model datasets.

Cano et al. developed multi tier early-warning systems for under performing students using genetic programming and multi-view learning. These systems leverage data from multiple sources to improve prediction accuracy. Similarly, multi model datasets collected from MOOCs now combine behavioral data and textual comments for unified semantic analysis. Such integration has proven effective in improving classification metrics like recall and F1, as demonstrated in recent empirical studies.

Overall, related work in EDM underscores the evolution from analyzing simple tabular data to integrating multi model, large-scale datasets, enabling a richer understanding of student behavior and performance.

## METHODOLOGY-ALGORITHM USED

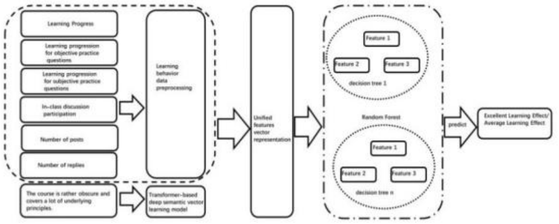
Educational Data Mining (EDM) employs various techniques to analyze and derive insights from the vast amounts of data generated in educational environments. Clustering algorithms, such as K-Means, are used to group students based on shared characteristics like behavior, performance, or demographics. This helps educators customize teaching strategies for different groups.

Classification algorithms, including decision trees and support vector machines (SVM), are applied to predict student outcomes, such as performance or dropout risks. These predictions enable early interventions to support at-risk students. Text mining techniques, such as sentiment analysis and natural language processing (NLP), analyze student feedback, discussion posts, and comments to understand emotions, engagement levels, and learning challenges.

Multi model data fusion techniques combine structured data (e.g., attendance, grades) and unstructured data (e.g., course comments, videos) to create unified representations for deeper analysis. This integration is achieved using models like Deep neural networks or multi-view learning algorithms, enhancing the accuracy of predictions and insights.

These techniques collectively address critical educational challenges, enabling improved teaching practices, personalized learning experiences, and better institutional management.

## SYSTEM ARCHITECTURE


Learning Progress
Learning progression for objective practice questions
Learning progression for subjective practice questions
In-class discussion participation
Number of posts
Number of replies
Learning behavior data preprocessing
The course is rather obscure and covers a lot of underlying principles.
Transformer-based deep semantic vector learning model
Unified features vector representation
Feature 1
Feature 2
Feature 3
decision tree 1
Random Forest
Feature 1
Feature 2
Feature 3
decision tree n
predict
Excellent Learning Effect/ Average Learning Effect

## EXPLANATION OF SYSTEM ARCHITECTURE

In this project, a data owner registers their details, logs in, and uploads documents. The data owner can send requests to data users and respond to requests by sharing a secret key. Data users can search uploaded documents, send requests to the cloud server, and download files in encrypted format. If a user enters the wrong key, they receive a warning, and repeated errors lead to permanent blocking. The cloud server manages logins, approves key requests, and monitors data, users, and stored information. It also approves or denies access requests. If a file is attacked, the system ensures data security by blocking unauthorized access.

## CONCLUSION

Swing's exceptional flexibility lies in its ability to render graphical user interface (GUI) components independently of the native host operating system's controls. Unlike traditional tool kits that rely on native OS-specific GUI controls, Swing uses Java 2D APIs to "paint" its components. This approach ensures a consistent and platform-independent appearance across different environments, making it highly adaptable for cross-platform applications. Moreover, Swing’s architecture allows developers to customize its components extensively, enhancing the user experience by tailoring the interface to specific application requirements.

In addition to GUI flexibility, Java’s thread management system plays a critical role in optimizing application performance. The Java thread scheduler employs a priority-based model, where each thread is assigned a priority value that determines its execution order. Developers can dynamically adjust these priorities using the set Priority() method, ensuring that high-priority tasks receive attention promptly while maintaining efficient multitasking. This approach is particularly valuable for applications that require concurrency, as it balances resource allocation and task execution effectively.

## REFERENCE

[1] C. Romero and S. Ventura, “Educational data science in massive open online courses,” Wiley Interdiscipl. Rev., Data Mining Know. Discovery, vol. 7, no. 1, p. e1187, Jan. 2017. [Online]. Available: https://onlinelibrary.wiley.com/doi/full/10.1002/widm.1187 and https:// onlinelibrary.wiley.com/ doi/abs/10.1002/widm.1187 and https://wires. onlinelibrary.wiley.com/ doi/10.1002/widm.1187

[2] R. S. Baker and P. S. Invent ado, “Educational data mining and learning analytics,” in Learning Analytics: From Research to Practice. Springer, Jan. 2014, pp. 61–75. [Online]. Available: https://link.springer. com/chapter/10.1007/978-1-4614-3305-7_4

[3] M. I. Baig, L. Shuib, and E. Yadegaridehkordi, “Big data in education: A state of the art, limitations, and future research directions,” Int. J. Educ. Technol. Higher Educ., vol. 17, no. 1, pp. 1–23, Dec. 2020. [Online]. Available: https://educationaltechnologyjournal.springeropen.com/articles/10. 1186/s41239-020-00223-0 [4] B. Bakhshinategh, O. R. Zaiane, S. ElAtia, and D. Ipperciel, “Educational data mining applications and tasks: A survey of the last 10 years,” Educ. Inf. Technol., vol. 23, no. 1, pp. 537–553, Jul. 2017. [Online]. Available: https://link.springer.com/article/10.1007/s10639-017-9616-z