A Student Performance Prediction Model Using Machine Learning Models in Multimodal Learning Analytics

Uoc Tran Van[1,2], Binh Hoang Tieu[3], and Dang Hung Tran[4(✉)]

[1] Hanoi National University of Education, Hanoi, Vietnam
[2] Institute for Transport Administration and Management Cadres, Hanoi, Vietnam
[3] CMC University, Hanoi, Vietnam
htbinh@cmcu.edu.vn
[4] Hanoi University of Industry, Hanoi, Vietnam
hungtd@haui.edu.vn

Abstract. Multimodal learning analytics in educational data mining is an increasingly popular research field aimed at improving student learning outcomes by applying techniques to a diverse range of data sources. This study proposes a multimodal learning analytics framework using machine learning for early student performance prediction, leveraging data from various sources. Such early-stage insights provide actionable intelligence, allowing educators to implement timely interventions and personalized support strategies to address challenges proactively. Our proposed model classified student performance into four distinct outcomes: Distinction, Pass, Fail, and Withdraw and was validated using the Open University Learning Analytics Dataset (OULAD), achieving a notable accuracy of 81.67% and an F1-score of 81.40% for multi-class classification. For binary classification tasks, the model attained an impressive accuracy of 97.45%. This study demonstrates how predictive analytics can be used to improve teaching methods, encourage proactive educational interventions, and raise overall student success rates by focusing on multimodal data integration and early prediction.

Keywords: Multimodal Learning Analytics · Data Fusion · Machine Learning · OULAD · Oversampling

1 Introduction

The growing use of digital tools in education has made data-driven decisions essential for improving teaching and helping students succeed [1]. One key application is the early prediction of learning outcomes, allowing for timely interventions to reduce dropout rates, refine teaching methods, and optimize resource allocation [2]. This study utilizes the Open University Learning Analytics Dataset (OULAD) [3] and applies machine learning models to classify student performance into four categories: Distinction, Pass, Fail, and Withdraw. By analyzing weekly data from various sources including demographics, virtual learning interactions, and assessment scores, our model predicts these outcomes before the final exam.

P. Meesad et al. (Eds.): IC2IT 2025, LNNS 1390, pp. 11–20, 2025.
https://doi.org/10.1007/978-3-031-90295-6_2

The use of multimodal data significantly enhances the accuracy and robustness of predictive models by capturing the complex interplay of behavioral, cognitive, and contextual factors that influence individual student learning. This methodology, often referred to as Multimodal Learning Analytics (MLA) [4, 5] combines diverse data sources, to provide a bigger understanding of student learning processes throughout a course. While this approach enables personalized interventions, most studies using the OULAD dataset rely on only one or two data sources [6, 7] primarily demographics and online interactions, often excluding student scores. Only a few studies [8] have incorporated three or more data sources to model individual grade progression, which is essential for early and accurate prediction of student outcomes.

Our study addresses this gap in literature by exploring data fusion techniques, including early and late fusion [4], for early performance prediction while incorporating student scores into the model. To tackle class imbalance, we compare results before and after applying SMOTE [9], which generates synthetic samples for minority classes without discarding original data. Our findings show that SMOTE improves class distribution and enhances the model's predictive accuracy across all performance categories. Additionally, we focus on the most relevant features to improve generalizability to new data. In summary, this study makes three key contributions:

- First, it proposes a multimodal machine learning framework that fuses data from diverse sources to early predict student performance before the final exam.
- Second, the framework validates the effectiveness of an oversampling technique called SMOTE in addressing class imbalance and identifies key features contributing to prediction accuracy.
- Finally, the proposed method aims to capture the complex interactions of behavioral, cognitive, and contextual factors that influence individual student learning trajectories.

2 Related Work

Several studies have emerged that leverage data from multiple sources to construct models for predicting learner skills [6, 10–12]. Machine learning and deep learning are the dominant methodological approaches employed in multimodal learning analytics (MLA) research.

Chago *et al.* [10] tackled student performance prediction using data fusion techniques on self-collected datasets. They explored four fusion methods: merging all attributes, selecting the best attributes, using ensemble models, and combining ensembles with feature selection, evaluating them with six classification algorithms. The authors findings showed that the best-performing approach, combining ensembles and feature selection, achieved an average accuracy of 83.29%.

Waheed *et al.* [6] investigated the potential of Artificial Neural Networks (ANN) for predicting at-risk students using multi source data available in the OULAD. Their objective was to develop a model that could facilitate early intervention strategies. The proposed ANN architecture generated results in the form of four binary classifications based on student achievement: Pass - Fail, Distinction - Fail, Distinction - Pass and Withdraw - Pass. The model achieved high classification accuracy, ranging from 84%

to 94%, which surpassed the performance of the Logistic Regression models (79.82%–85.60%) and SVM (79.95%–89.14%).

Al-Zawqari *et al.* [13] aimed to improve assessment of learner comprehension in online and blended learning environments. Their study proposed a novel approach for automatic feature extraction from raw data, eliminating the need for manual feature identification common in previous work. This method aimed to enhance model interpretability. To evaluate this approach, they constructed a random forest binary classifier for four student ability groups on the OULAD. Additionally, they incorporated techniques to identify attributes most influential in predicting at-risk and dropout students. Furthermore, the authors developed an artificial neural network model to classify learner abilities, achieving high accuracy results (81%–93%).

Al-Azazi *et al.* [12] contribute to the growing body of research on effective decision tree-based machine learning models by proposing a novel approach. They leverage Recurrent Neural Networks (RNNs) with Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) architectures to predict student performance using frequently updated data. Their model treats OULAD as a time series dataset, allowing it to capture temporal dependencies. The results indicate that the proposed method achieves superior performance compared to a combination of individual students, with the Accuracy score reaching up to 95%. This study suggests the model's ability to learn generalizable features from the data. However, limitations exist, including potentially lengthy training times and complex installation procedures.

Similar to Al-Azazi *et al.* [12] and Zhaoyu Shou et al. [8] proposed a Long Short-Term Memory and Artificial Neural Network based model for predicting learner outcomes over time. Their approach transforms OULAD into a multidimensional time series format and utilizes binary classification to identify students at high risk of failing. This allows for early intervention by instructors. The reported accuracy of this model ranges from 74% to 99.08%, with an F1 score following a similar trend from 73% to 99%.

3 Datasets

3.1 Description

This study utilizes the OULAD dataset, which contains aggregated clickstream data, demographic details, and assessment scores, to evaluate the proposed model. The dataset spans 22 classes across 7 subjects, with courses offered in different semesters between 2013 and 2014, covering 32,592 students and over 10 million online interactions. Using this data, machine learning models are developed to classify student performance into four categories—Distinction, Pass, Fail, and Withdrawn—addressing both multi-class and binary classification tasks.

3.2 Preprocessing

Given that OULAD is composed of individual comma-separated values files, merging these files is necessary to enable effective analysis. We employed the *pandas* library's *merge* function, utilizing *student_id* as the primary key table. Following this data integration process, the resulting dataset is structured into nine distinct CSV files. Each line in these files represents information pertaining to a single student.

3.2.1 Demographics

The Open University collected demographic information on, including their id, major, gender, birthplace, GPA, enrollment history, and any declared special needs. As this data was initially provided in string format, we employed Label Encoding method [14] to transform it into a numerical representation suitable for machine learning model computations.

3.2.2 Assessment Scores

Our analysis revealed missing assessment scores in many courses, likely due to students not submitting assignments, which we addressed by imputing zeros. OULAD categorizes courses into seven main subjects from AAA to GGG, each containing multiple classes by year, with three test types: TMA (Tutor Marked Assessment), CMA (Computer Marked Assessment), and Exam. To enable data merging, we computed average component scores (0–100) from weighted individual test scores, focusing on early performance prediction for actionable insights before the final exam.

3.2.3 Virtual Learning Interactions

Moodle's log files statistically record each student's total interactions (number of clicks) with various items. We normalize each feature within the three data sources to a [0,1] range. While OULAD has an imbalanced class distribution: 42% Pass, 23% Fail, 10% Distinction, and 25% Withdraw, we apply the SMOTE oversampling technique [15] to balance the training data, as shown in Fig. 1.

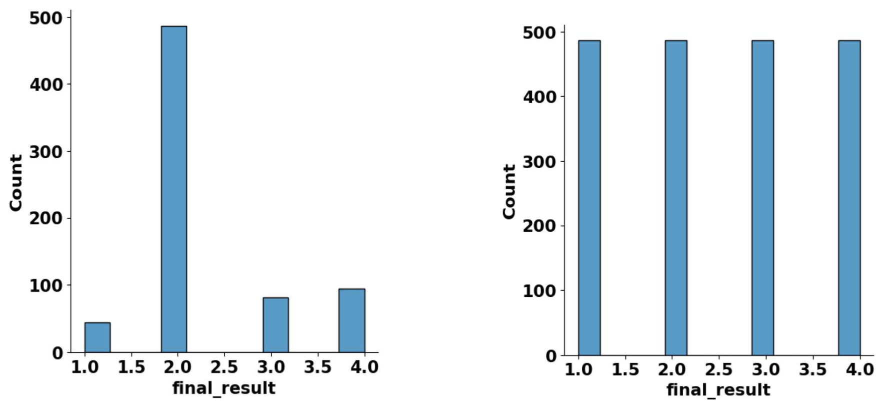

Fig. 1. Example distribution of OULAD before and after applying an oversampling technique.

Furthermore, we employed a feature selection technique involved calculating the Pearson correlation coefficient (Eq. 1) between features to identify and remove redundant attributes with minimal impact on the model's output.

$$\sigma_{X,Y} = \frac{E[(X - \mu_X)(Y - \mu_Y)}{\sigma_X \sigma_Y}. \tag{1}$$

where μ_X and μ_Y are the mean, σ_X and σ_Y are the standard deviation of X and Y respectively.

We calculate feature importance using the *feature_importances* function from the *sklearn* library, which builds a RF model across the dataset and computes the Gini index for each feature. The Gini index (2), a measure introduced in the CART decision tree algorithm, assesses the purity of data at tree nodes, with higher values signifying greater feature importance.

$$Gini = 1 - \sum_{i=1}^{C} (p_i)^2 \tag{2}$$

where: C is the number of class and $p_i = \frac{n_i}{N}$ with n_i is the number of samples belong to class *i*, N is the total number of samples.

4 Methodology

This study proposes a multimodal learning analytics model utilizing different fusion techniques to early predict student performance based on four categories: Distinction, Pass, Fail and Withdraw. We leverage OULAD, extract three key data sources: demographic, online interactions, and assessment scores. The resulting dataset, following normalization detailed in Sect. 3, comprises 25,793 rows and 31 columns. This includes a single label column indicating student performance and a student column. We then calculate feature correlations and feature importance within the merged dataset to remove irrelevant attributes. After data rebalancing, we employ three machine learning models - Random Forest (RF), Support Vector Machine (SVM), and k-Nearest Neighbors (kNN) - to address the multi-class classification and binary classification problems. Finally, the model achieving the highest accuracy is subsequently selected. The implementation process is visually summarized in Fig. 2.

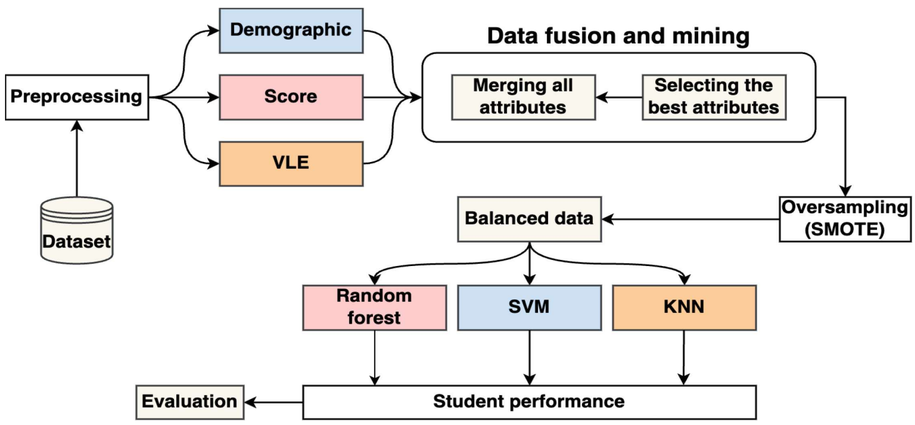

Fig. 2. The proposed student performance prediction model in multimodal learning analytics.

To determine the efficacy of our proposed model, we performed an experiment employing a late fusion multimodal learning analysis approach visualized in Fig. 3. This technique involves constructing separate machine learning models for each data source, followed by the integration of their outputs for the final prediction.

The outputs of each model are merged using the major voting algorithm as follows: Each model assigns an integer label between 1 and 4 to each data sample. The final output is determined by the most frequent label predicted by RF, SVM and KNN models.

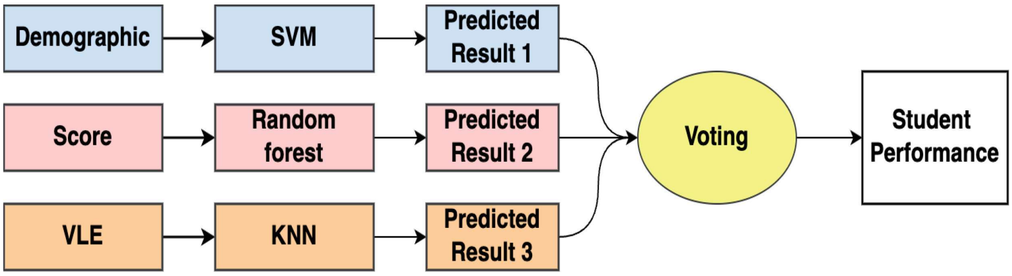

Fig. 3. Late fusion method with major voting approach for OULAD

5 Experiment Results

We utilize accuracy and F1-score to evaluate the model's performance. The formula for calculating accuracy and F1-score is presented in Eq. (3) and (4) respectively.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{3}$$

$$F1\ score = 2 * \frac{Precision\ *\ Recall}{Precision + Recall} \tag{4}$$

5.1 Result of the Proposed Model

This section summarizes the experimental results on OULAD using 5-fold cross-validation. Table 1 shows the accuracy of RF, SVM, and KNN models before and after oversampling across seven datasets. RF exhibited the highest improvement, with accuracy increasing from 63.93%–85.82% range (avg. 69.97%) to 77.12%–93.06% range (avg. 81.67%). KNN also improved significantly, from 50.88%–80.19% range, with the overall dataset accuracy rising from 62.28% to 74.62%. While SVM improved for most datasets, accuracy dropped for CCC and GGG due to their unique characteristics.

Table 1. Accuracy results for 8 datasets. The table compares the performance of the models before and after applying oversampling techniques and removing unimportant attributes.

Model	Oversampling	AAA	BBB	CCC	DDD	EEE	FFF	GGG	Total
RF	Before	85.82	68.13	63.93	73.96	68.26	72.89	68.24	69.97
	After	**93.06**	**83.17**	**77.12**	**84.90**	**85.12**	**85.56**	**80.89**	**81.67**
SVM	Before	78.72	67.55	59.68	69.40	68.26	70.59	67.06	67.39
	After	80.21	68.82	62.18	71.83	69.37	73.52	55.87	67.42
kNN	Before	75.88	63.91	65.14	64.34	64.78	64.23	55.45	62.28
	After	84.06	77.23	50.88	74.92	80.19	73.90	68.27	74.62

Table 2 compares the F1 scores of RF, SVM, and kNN models before and after oversampling across seven datasets and a combined dataset. RF showed the highest improvement, with the overall F1 score rising from 61.32% (42.78%–72.55%) to 81.40% (76.80%–93.08%). The GGG dataset saw the largest gain, with RF's score increasing by 38.41%. SVM improved from 55.08% (36.28%–61.50%) to 66.12% (54.64%–79.97%), with the AAA dataset showing notable gains. kNN's overall F1 score increased significantly from 51.96% (41.17%–55.25%) to 73.68% (64.23%–83.17%), highlighting major improvements, especially in the AAA dataset.

Table 2. F1-scores achieved by our model on 8 datasets. The table compares the performance before and after applying two data pre-processing techniques: oversampling and attribute selection.

Model	Oversampling	AAA	BBB	CCC	DDD	EEE	FFF	GGG	Total
RF	Before	72.55	56.05	53.84	68.26	53.74	66.43	42.78	61.32
	After	**93.08**	**82.99**	**76.80**	**84.94**	**85.02**	**85.35**	**81.19**	**81.40**
SVM	Before	42.12	50.85	46.80	49.58	43.36	61.50	36.28	55.08
	After	79.97	67.19	60.85	71.56	68.53	72.57	54.64	66.12
kNN	Before	55.25	52.38	47.10	51.24	48.75	52.61	41.17	51.96
	After	83.17	76.58	64.23	74.50	79.08	72.93	66.98	73.68

The results outlined in Tables 1 and 2 demonstrate that the RF model outperforms the other models evaluated, achieving the highest level of accuracy. These findings emphasize the importance of employing oversampling techniques when addressing imbalanced datasets. OULAD is predominantly classified under the Pass category, resulting in training bias and diminished effectiveness in predicting minority classes. Additionally, the removal of non-essential attributes from the dataset contributes to enhanced model performance. Therefore, the application of oversampling techniques combined with attribute selection can markedly improve the performance of machine learning models in real-world scenarios.

5.2 Results Comparison

- Between different data fusion methods

Our proposed early fusion model achieved superior performance on 8 datasets compared to the late fusion method. These results were shown in Table 3 with accuracy, and the F1 score metrics respectively.

- With previous studies

Our proposed model was compared with prior studies in multi-class and binary classification settings for predicting student performance using similar datasets. Al-Azazi et al. [12] used an ANN-LSTM on the OULAD dataset, achieving 72% accuracy and a 66% F1-score, while Shou [8] combined LSTM with multi-head self-attention, reaching 74% accuracy and a 73% F1-score. Both methods lacked strategies to handle data imbalance and had high computational demands. Our approach mitigates these issues by balancing data using SMOTE and removing irrelevant features. As shown in Table 4, our RF model achieves 81.67% accuracy and an 81.40% F1-score, outperforming the ANN-LSTM and MTAPSP models by 7.67% and 8.4%, respectively.

Table 3. Comparison between early fusion method and late fusion method

Module	Sample size	Early fusion Accuracy	Late fusion Accuracy	Early fusion F1 score	Late fusion F1 score
AAA	704	**93.06**	78.40	**93.08**	78.55
BBB	6,054	**83.17**	68.85	**82.99**	65.24
CCC	3,410	**77.12**	61.86	**76.80**	61.17
DDD	4,932	**84.90**	63.47	**84.94**	63.31
EEE	2,296	**85.12**	69.28	**85.02**	68.39
FFF	6,287	**85.56**	69.27	**85.35**	68.39
GGG	2,109	**80.89**	52.74	**81.19**	51.00
Total	25,793	**81.67**	65.22	**81.40**	64.41

Table 4. Comparison results between proposed model and ANN-LSTM, MTAPSP in multi-class classification setting.

Metric	ANN-LSTM (2023)	MTAPSP (2024)	Our model
Accuracy	72	74	**81.67**
F1 Score	66	73	**81.40**

To further assess our approach, we also compare it with previous studies utilizing binary classification models. Waheed et al. [6] reduced the feature set from 54 to 30 before training an ANN with three hidden layers. Al-Zawqari [13] developed four independent binary classifiers for Pass-Fail, Distinction-Fail, Distinction-Pass, and Withdraw-Pass using OULAD, producing predictions across four quarters (Q1–Q4). Building on these studies, we created four separate binary classifiers using 29 features from the entire dataset, as outlined in Sect. 3. We employed three machine learning models along with the data processing techniques described in Sect. 3.

Table 5 shows that our model achieves competitive accuracy scores, demonstrating its effectiveness. Our proposed model attained a peak classification accuracy of 97.45% for distinguishing between “Withdraw” and “Pass” classes, surpassing Waheed’s ANN method by 4.2%. The lowest accuracy observed in our study was 88.73% for classifying “Distinction” and “Pass” classes, which still exceeds the accuracy achieved by both Waheed’s ANN method and Ali Al-Zawqari’s model. In all four binary classification tasks evaluated, our model consistently outperforms previous studies.

Table 5. Comparison results between proposed model and Simulated ANN and RF in binary classification setting.

Binary classification	Simulated ANN Waheed *et al.* (2020)	RF Ali Al-Zawqari *et al.* (2022)	kNN Our model	SVM Our model	RF Our model
Pass - Fail	85.75	85.83	84.46	84.99	**90.57**
Distinction - Fail	88.95	87.29	92.29	93.25	**95.99**
Distinction - Pass	81.21	81.26	80.23	80.38	**88.73**
Withdraw - Pass	93.28	92.90	94.29	95.06	**97.45**

6 Conclusion

This study explores the development of a student performance prediction model within the framework of multimodal learning analytics. Utilizing the publicly available OULAD dataset, the model employs early fusion techniques to integrate multiple data modalities from diverse sources. Experimental results indicate that the Random Forest model achieves competitive accuracy compared to existing approaches, demonstrating its effectiveness in student performance prediction. The findings underscore the potential of multimodal learning analytics as a robust tool for comprehensive learner assessment, enabling educators to implement targeted interventions and enhance individualized learning experiences. Moreover, the model's capacity to predict student performance at an early stage facilitates timely identification of at-risk students, allowing for proactive interventions to improve learning outcomes before final assessments. Future work will involve collecting real-world data to further validate the model's applicability. Additionally, the proposed approach is adaptable to various domains, such as healthcare and finance, provided that similar multimodal data sources are available.

References

1. Hakimi, M., Katebzadah, S., Fazil, A.W.: Comprehensive insights into e-learning in contemporary education: analyzing trends, challenges, and best practices. J. Educ. Teach. Learn. **6**(1), 86–105 (2024)
2. Said, M.B., Kacem, Y.H., Algarni, A., Masmoudi, A.: Early prediction of student academic performance based on Machine Learning algorithms: a case study of bachelor's degree students in KSA. Educ. Inf. Technol. **29**, 13247–13270 (2024)
3. Kuzilek, J., Hlosta, M., Zdrahal, Z.: Data descriptor: open University learning analytics dataset. Sci. Data **4**, 1–8 (2017)
4. Chango, W., Lara, J.A., Cerezo, R., Romero, C.: A review on data fusion in multimodal learning analytics and educational data mining. WIREs Data Min. Knowl. Discov. **12**(4), 1–19 (2022)
5. Ouhaichi, H., Spikol, D., Vogel, B.: Research trends in multimodal learning analytics: a systematic mapping study. Comput. Educ.: Artif. Intell. **4**, 100136 (2023)
6. Waheed, H., Hassan, S.U., Aljohani, N.R., Hardman, J., Alelyani, S., Nawaz, R.: Predicting academic performance of students from VLE big data using deep learning models. Comput. Hum. Behav. **104** (2020)

7. Nadar, N., Kamatchi, R.: A novel student risk identification model using machine learning approach. Int. J. Adv. Comput. Sci. Appl. **9**(11), 305–309 (2018)
8. Shou, Z., Xie, M., Mo, J., Zhang, H.: Predicting student performance in online learning: a multidimensional time-series data analysis approach. Appl. Sci. **14**(6), 2522 (2024)
9. van den Goorbergh, R., van Smeden, M., Timmerman, D., Van Calster, B.: The harm of class imbalance corrections for risk prediction models: illustration and simulation using logistic regression. J. Am. Med. Inform. Assoc. **29**(9), 1525–1534 (2022)
10. Chango, W., Cerezo, R., Romero, C.: Multi-source and multimodal data fusion for predicting academic performance in blended learning university courses. Comput. Electr. Eng. **89**, 106908 (2021)
11. Renò, V., Stella, E., Patruno, C., Capurso, A., Dimauro, G., Maglietta, R.: Learning analytics: analysis of methods for online assessment. Appl. Sci. **12**(18), 1–10 (2022)
12. Al-Azazi, F.A., Ghurab, M.: ANN-LSTM: a deep learning model for early student performance prediction in MOOC. Heliyon **9**(4), e15382 (2023)
13. Al-Zawqari, A., Peumans, D., Vandersteen, G.: A flexible feature selection approach for predicting students' academic performance in online courses. Comput. Educ.: Artif. Intell. **3**, 100103 (2022)
14. Molin, S.: Hands-On Data Analysis with Pandas: A Python Data Science Handbook for Data Collection, Wrangling, Analysis, and Visualization. Packt Publishing Ltd. (2021)
15. Mohammed, R., Rawashdeh, J., Abdullah, M.: Machine learning with oversampling and undersampling techniques: overview study and experimental results. In: 2020 11th International Conference on Information and Communication Systems (ICICS), pp. 243–248. IEEE (2020)