Predicting Student Performance: A Machine Learning Approach to Forecasting Pass/Fail Outcomes

Nawal Mohamed Bahy Eldin,
Ayda Elsemsar,
Ramy Kamal Amin

Teaching Assistant
Management Information Systems & Basic Science Department
Egyptian Institute of Alexandria Academy for Management & Accounting

(ISSN: 3062 – 5548)

Predicting Student Performance: A Machine Learning Approach to Forecasting Pass/Fail Outcomes

التنبؤ بأداء الطلاب :نهج تعلم الآلة للتنبؤ بنتائج النجاح/الرسوب

Nawal Mohamed Bahy Eldin
Ayda Elsemsar
Ramy Kamal Amin
Teaching Assistant, Management Information Systems& Basic Science Department
Egyptian Institute of Alexandria Academy for Management & Accounting
Email:

Abstract

This study aims to predict student performance using machine learning models based on the Open University Learning Analytics Dataset (OULAD). The dataset includes various features such as student assessments, virtual learning environment (VLE) interactions, and demographic data. Several machine learning models were applied, including Logistic Regression, Linear Discriminant Analysis, Random Forest, and Neural Networks, to predict student outcomes (Pass, Fail, or Distinction). The results demonstrate that models incorporating both weighted grades and pass rates outperformed those relying on single features. Although Neural Network models achieved the highest accuracy, they faced challenges in predicting failure cases. This paper offers insights into the performance of different models and proposes directions for future improvements.

Keywords: Student Performance Prediction, Machine Learning, Neural Networks, Educational Data Mining, Pass/Fail Classification.

المستخلص

تهدف هذه الدراسة إلى التنبؤ بأداء الطلاب باستخدام نماذج تعلم الآلة، وذلك بالاعتماد على مجموعة بيانات Open University Learning Analytics Dataset (OULAD). تتضمن هذه البيانات مجموعة من المتغيرات المهمة، مثل تقييمات الطلاب، ومستوى تفاعلهم مع بيئة التعلم الافتراضية (VLE)، بالإضافة إلى البيانات الديموغرافية. تم تطبيق مجموعة من نماذج تعلم الآلة، شملت الانحدار اللوجستي، وتحليل التمييز الخطي (LDA)، والغابات العشوائية، والشبكات العصبية، بهدف التنبؤ بالنتائج الأكاديمية للطلاب (نجاح، رسوب، أو تميز). وقد أظهرت النتائج أن النماذج التي دمجت بين الدرجات المرجحة ونسب النجاح حققت أداءً أفضل مقارنةً بتلك التي اعتمدت على أحد المتغيرين فقط. وعلى الرغم من تفوق نماذج الشبكات العصبية في تحقيق أعلى معدلات الدقة، إلا أنها واجهت تحديات في التنبؤ بحالات الرسوب. تقدم هذه الدراسة رؤى مهمة حول كفاءة النماذج المختلفة، وتقترح اتجاهات مستقبلية لتحسين دقة التنبؤ في هذا المجال.

الكلمات المفتاحية :التنبؤ بأداء الطلاب، تعلم الآلة، الشبكات العصبية، تنقيب البيانات التعليمية، تصنيف النجاح والرسوب.

1. Introduction

Student academic performance in higher education (HE) is researched extensively to tackle academic underachievement, increased university dropout rates, graduation delays, among other tenacious challenges [1]. In simple terms, student performance refers to the extent of achieving short-term and long-term goals in education [2]. However, academicians measure student success from different perspectives, ranging from students' final grades, grade point average (GPA), to future job prospects [3]. The literature offers a wealth of computational efforts striving to improve student performance in schools and universities, most notably those driven by data mining and learning analytics techniques [4]. However, confusion still prevails regarding the effectiveness of the existing intelligent techniques and models. The timely prediction of student performance enables the detection of low performing students, thus, empowering educators to intervene early during the learning process and implement the required interventions. Fruitful interventions include, but are not limited to, student advising, performance progress monitoring, intelligent tutoring systems development, and policymaking [5].This study seeks to contribute to the

ongoing discourse by leveraging machine learning models to predict student performance, with the ultimate goal of enhancing educational outcomes and reducing attrition rates in higher education.

This paper is organized into four main sections to provide a comprehensive study on predicting student performance using machine learning techniques. **Section II** reviews previous literature on student performance prediction, focusing on the methodologies used and the accuracy rates achieved by various models. **Section III** outlines the proposed methodology, including data collection, preprocessing, feature engineering, model selection, and evaluation metrics. **Section IV** presents the obtained results, compares them with prior research, and provides a detailed analysis of the models' accuracy and performance, particularly in predicting failure cases. Finally, **Section V** concludes the study and offers recommendations for future research.

2. Literature Review

Predicting student performance is one of the most significant areas of research in education, as it directly contributes to enhancing the quality of the educational process and enables academic institutions to intervene early to support struggling students. With the expansion of e-learning technologies and the availability of vast amounts of student data, employing data analysis and artificial intelligence techniques has become essential for understanding the factors influencing academic performance and developing accurate and effective predictive models. Many recent studies have focused on developing predictive models based on machine learning algorithms and big data analysis to identify students at risk of failure or academic decline. These studies vary in terms of data processing methods, types of data used, and developed models. The data considered include demographic information, interaction records with virtual learning environments, assessment and exam results, as well as social and economic factors. The following is a review of the most prominent studies in this field, highlighting their methodologies and key findings, which serve as the foundation for this research in developing a comprehensive predictive model using artificial intelligence and data analysis techniques.

Charlotte Van Petegem et al. (2023) [6] In their study introduced a privacy-friendly early-detection framework designed to identify students at risk of failing in introductory programming courses at the university level. The framework was rigorously validated across two different courses, each with annual editions, involving a total of 2,080 higher education students. The results demonstrated that the framework is highly accurate and robust against variations in course structures, teaching and learning styles, programming exercises, and classification algorithms.

In the study by V. Balachandar and K. Venkatesh (2025) [7] a new method was developed to address common challenges in educational datasets, especially issues related to imbalance and temporal settings. The model is also explainable through AI features. By applying adaptive hyper-parameter tuning and advanced graph neural network layers in the MSPP model, the approach provides a denser representation of predictions, leading to more accurate classification. Experimental results showed that the MSPP model outperformed other models such as EAI&ML, MTSDA, XAI, DGNN, and DLM, achieving a high accuracy of 76%.

In Mikhail Dorrer's (2025) [8] study, data analysis tools were used to predict student performance based on their previous academic achievements. The project analyzed historical educational data from over 35,000 students collected over seven years, covering 1.24 million grades. Neural network regression models were developed to predict future grades and improve educational processes. The model's performance was evaluated using the coefficient of determination (R^2) and root mean square error (RMSE) with 10-fold cross-validation. Results showed that over 70% of the models achieved an R^2 greater than 0.7.

In the study by Xiaoyi Zhang et al. (2025) [9] the researchers proposed the GNN-Transformer-InceptionNet (GNN-TINet) model to predict student performance in complex multi-label scenarios, where students may belong to multiple performance categories simultaneously. Utilizing the California Student Performance Dataset

containing 97,000 records, the model combined InceptionNet, transformer architectures, and graph neural networks (GNNs) to enhance prediction accuracy. Advanced preprocessing techniques, including Contextual Frequency Encoding (CFI) and Contextual Adaptive Imputation (CAI), were applied. The model achieved outstanding results, with a Predictive Consistency Score (PCS) of 0.92 and an accuracy of 98.5%.

In Hosam A. Althibyani's (2025) [10] study, demographic data, assessment scores, and student interaction records were used to build predictive models for student outcomes in online courses. The study applied logistic regression and random forest classification to predict performance based on four possible outcomes and a simple pass/fail result. The findings showed that even basic indicators like daily activity levels were effective in predicting success. The logistic regression model achieved 72.1% accuracy for the four-class prediction and 92.4% for pass/fail, while the random forest model performed better with 74.6% and 95.7% accuracy.

A. Maanas Sai Surya Chandra et al. (2025) [11] conducted a study aiming to predict students' pass/fail outcomes without relying on final grades (G3). The dataset included various features, including socioeconomic factors. The researchers developed machine learning models using ensemble classification algorithms and cross-validation techniques. The results showed that the proposed models achieved good classification accuracy, confirming the effectiveness of machine learning in predicting student success based on early indicators. The study emphasized the importance of such predictive models in supporting at-risk students and enhancing overall academic performance through early intervention strategies.

Building on the insights gained from the reviewed studies, this research aims to develop a comprehensive predictive model for student performance using artificial intelligence and data analysis techniques. By leveraging a rich dataset that captures various aspects of student demographics, assessments, and interactions with virtual learning environments, the study seeks to enhance prediction accuracy and provide actionable insights for academic institutions. The following section outlines the

methodology adopted to achieve these objectives, detailing each stage of the data analysis and machine learning pipeline.

3. Methodology

This study employs a systematic data analysis and machine learning pipeline to predict student performance. The methodology comprises several sequential phases: data collection and description, data preprocessing, exploratory data analysis (EDA), feature engineering, model development, and evaluation. Each phase is elaborated below.

3.1 Data Collection and Description

The study utilizes the Open University Learning Analytics Dataset (OULAD), which consists of multiple interrelated tables capturing various aspects of students' demographic, academic, and behavioral data. These tables are briefly described as follows:

Table 1: Dataset Description and Key Attributes

Table Name	Description	Key Attributes
Student Info	Contains demographic and academic details of students.	gender, age_band, highest_education, region, disability, num_of_prev_attempts, final_result
Assessments	Includes information about course assessments.	assessment_type, date, weight, module_code, code_presentation
Student assessment	Records students' performance in each assessment.	id_assessment, id_student, date_submitted, score
StudentVLE	Captures students' interactions with the	id_student, id_site, date, sum_click

(ISSN: 3062 – 5548)

	Virtual Learning Environment (VLE).	
VLE	Provides details of learning activities available within the VLE platform.	id_site, activity_type, week_from, week_to

The Student Info table serves as the primary source for demographic and academic records, including the target variable, final_result, categorized as Pass, Fail, or Distinction. The Assessments and StudentAssessment tables offer granular insights into students' assessment types, weights, and achieved scores. Meanwhile, StudentVLE and VLE tables provide behavioral data by tracking engagement levels within the virtual learning environment through activity clicks and interaction types. Figure 1 presents the Entity-Relationship Diagram (ERD), illustrating the relationships among the dataset tables and highlighting the data flow necessary for constructing a comprehensive student performance prediction model.

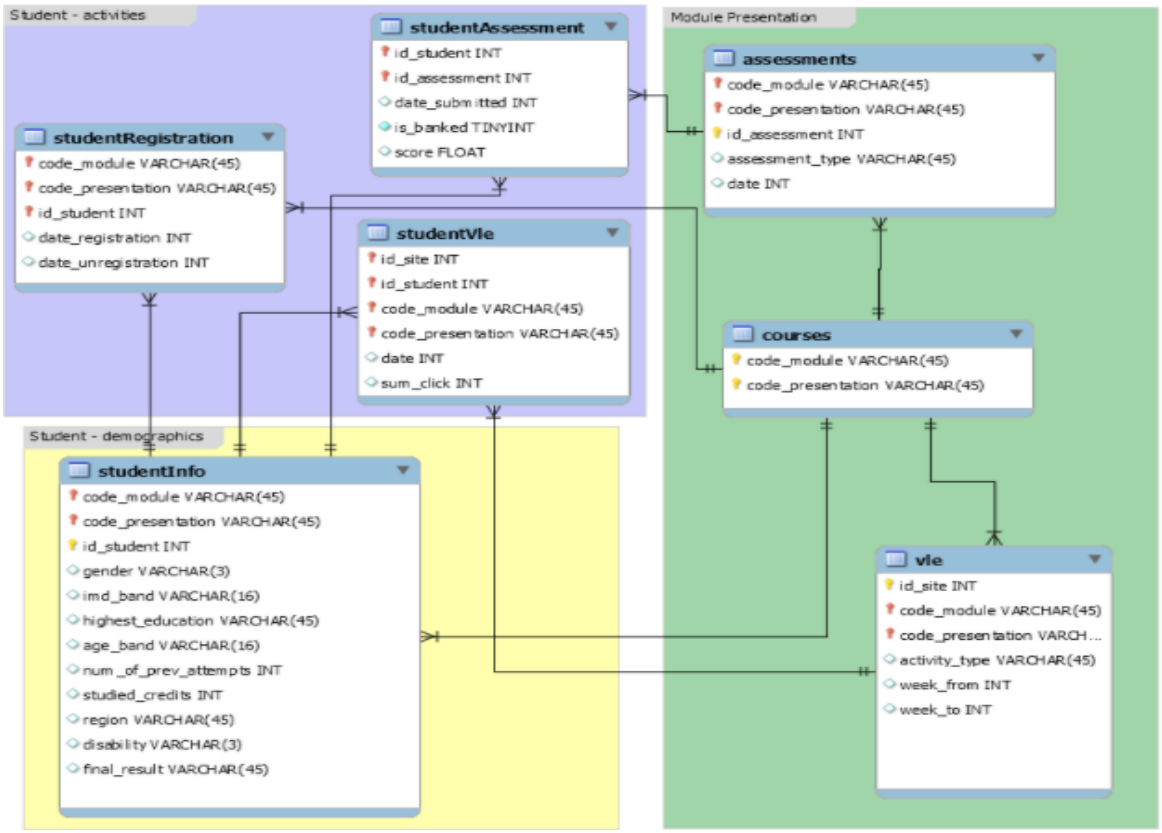

Figure 1: Entity-Relationship Diagram (ERD) of the Open University Learning Analytics Dataset

3.2 Data Preprocessing

Data preprocessing is a critical step to ensure the quality and consistency of the dataset before applying machine learning models. The following techniques were employed:

- **Handling Missing Values:** Missing data were systematically identified and treated using suitable imputation methods. For numerical features, mean imputation was applied, while mode imputation was used for categorical variables to preserve data integrity.
- **Encoding Categorical Variables:** Categorical attributes, such as *gender* and *final_result*, were transformed into numerical formats using appropriate

encoding techniques. One-Hot Encoding was applied to nominal variables, while Label Encoding was used for ordinal categories, facilitating compatibility with machine learning algorithms.

- **Feature Scaling:** To ensure that all numerical features contributed equally to the model's learning process, feature scaling was performed using the **Min-Max Scaler**. This normalization technique scales values within a defined range, typically [0, 1], enhancing model performance and convergence speed.

3.3 Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) was conducted to gain a comprehensive understanding of the dataset, identify underlying patterns, detect anomalies, and uncover relationships between variables that may influence student performance. The following analytical techniques were applied:

- **Univariate Analysis: The distribution of individual variables, such as Grade Point Average (GPA), study time, and assessment scores, was examined using histograms and boxplots. This analysis provided insights into the central tendency, dispersion, and overall distribution of each feature.**
- **Bivariate Analysis: Pairwise relationships between key variables were explored to assess potential correlations and dependencies. For instance, scatterplots and correlation matrices were utilized to analyze the relationship between study time and GPA, as well as between parental education level and student performance.**
- **Multivariate Analysis: Complex interactions among multiple variables were investigated through pair plots and heatmaps. This analysis aimed to identify significant patterns and multivariate correlations that could contribute to understanding the factors affecting academic outcomes.**

- **Outlier Detection: Outliers within critical features, such as the number of virtual learning environment (VLE) clicks and the number of previous course attempts, were detected and addressed. The identification and treatment of outliers were essential to ensure data quality, reduce noise, and enhance the robustness of the predictive models.**

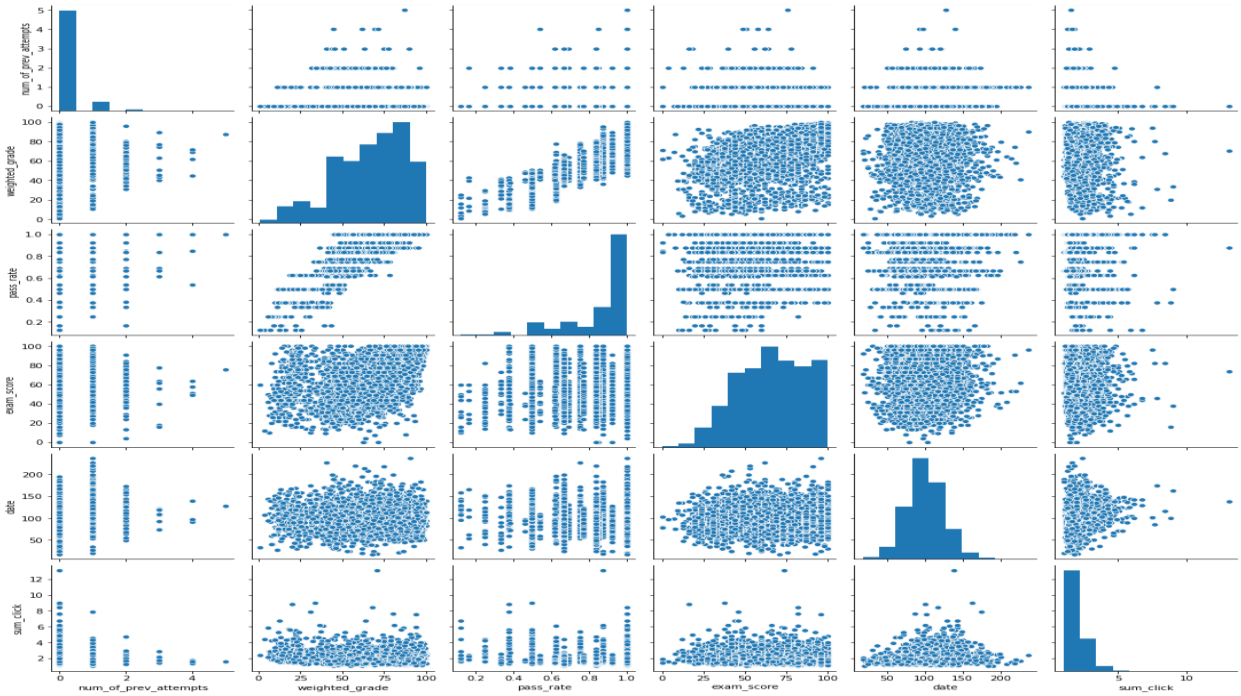

Figure 2: Pair plot showing the relationships and distributions of the dataset features.

3.4 Feature Engineering

In this phase, new features were engineered to enhance the predictive power of the dataset and improve model performance. The process included the following:

- **Assessment-Based Features: A weighted grade was calculated for each student by combining assessment scores with their respective weights to reflect the overall academic performance. A pass rate was computed, representing the percentage of assessments a student successfully passed.**

(ISSN: 3062 – 5548)

A passing threshold was set at 40%. Exam scores were extracted and treated as a distinct feature due to their significant influence on the final grade.

- **Virtual Learning Environment (VLE) Interaction Features: The average number of clicks per activity and the average time spent on activities were calculated to quantify student engagement levels with the VLE platform.**
- **Final Dataset Compilation:** All engineered features were merged with the student's demographic and academic data from the student information table. The result was a comprehensive dataset prepared for model training.

3.5 Model Selection and Training

In this phase, various machine learning models were developed and trained to predict student performance based on the engineered features.

- **Baseline Models:** Initial models such as Logistic Regression, Linear Discriminant Analysis (LDA), and Random Forest were trained to establish baseline performance.
- **Advanced Models:** More sophisticated models, including Neural Networks, were implemented to improve prediction accuracy.
- **Hyperparameter Tuning:** Grid search and randomized search techniques were employed to optimize hyperparameters and enhance model performance.
- **Cross-Validation:** K-fold cross-validation was utilized to validate the models, ensure generalizability, and prevent overfitting.

3.6Model Evaluation

To comprehensively assess model performance, multiple evaluation strategies were adopted:

- **Evaluation Metrics:** Models were evaluated using standard classification metrics such as accuracy, precision, recall, and F1-score.
- **Confusion Matrix:** The confusion matrix was used to visualize the performance of classification models, highlight prediction errors, and identify areas for improvement.

4. Results and discussion

Logistic Regression: Achieved an accuracy of 88%, with a precision of 0.87 for predicting failures and 0.88 for predicting passes. The model performed well but struggled with distinguishing between Distinction and Pass cases.

```
[[209   0  58]
 [  0 135  67]
 [ 35  20 961]]
```

	precision	recall	f1-score	support
Distinction	0.86	0.78	0.82	267
Fail	0.87	0.67	0.76	202
Pass	0.88	0.95	0.91	1016
accuracy			0.88	1485
macro avg	0.87	0.80	0.83	1485
weighted avg	0.88	0.88	0.88	1485

Figure 3: The Confusion Matrix and Classification Report for the Logistic Regression Model (Model 1).

Linear Discriminant Analysis (LDA): Achieved an accuracy of 88%, with a precision of 0.81 for failures and 0.92 for passes. LDA showed better performance in handling class imbalances compared to Logistic Regression.

```
[[229   0  38]
 [  0 162  40]
 [ 64  38 914]]
```

	precision	recall	f1-score	support
Distinction	0.78	0.86	0.82	267
Fail	0.81	0.80	0.81	202
Pass	0.92	0.90	0.91	1016
accuracy			0.88	1485
macro avg	0.84	0.85	0.84	1485
weighted avg	0.88	0.88	0.88	1485

(ISSN: 3062 – 5548)

Figure 4: Confusion Matrix and Classification Report of the Linear Discriminant Analysis (LDA) Model (Model 2).

Random Forest: Achieved the highest accuracy of 90%, with a precision of 0.87 for failures and 0.91 for passes. This model outperformed the others,

```
[[222   0  45]
 [  0 152  50]
 [ 33  22 961]]
```

	precision	recall	f1-score	support
Distinction	0.87	0.83	0.85	267
Fail	0.87	0.75	0.81	202
Pass	0.91	0.95	0.93	1016
accuracy			0.90	1485
macro avg	0.88	0.84	0.86	1485
weighted avg	0.90	0.90	0.90	1485

demonstrating its robustness in handling complex interactions between features.

Figure 5: Confusion Matrix and Classification Report of the Neural Network Classifier (Model 3).

Neural Network: Achieved an accuracy of 95% for binary classification (combining Pass and Distinction into one class). However, it struggled to predict Fail cases, likely due to the class imbalance and the simplification of the target

```
[[ 133   69]
 [   9 1274]]
              precision    recall  f1-score   support

           0       0.94      0.66      0.77       202
           1       0.95      0.99      0.97      1283

    accuracy                           0.95      1485
   macro avg       0.94      0.83      0.87      1485
weighted avg       0.95      0.95      0.94      1485
```

variable.

Figure 6: Confusion Matrix and Classification Report of the Neural Network Classifier (Model 4).

5. Conclusion:

In this study, several predictive models were developed and evaluated to predict student performance. The results demonstrated that models incorporating both weighted grade and pass rate variables outperformed those that excluded them, emphasizing the significance of these factors in determining academic outcomes. Although the Neural Network Classifiers faced some challenges in predicting failure cases, they overall achieved better performance compared to other models. This superior performance is likely attributed to simplifying the classification task by merging the "Pass" and "Distinction" categories. Traditional models such as Logistic Regression and Linear Discriminant Analysis (LDA) could still serve valuable purposes, especially in headhunting programs aimed at identifying outstanding students for scholarships and employment opportunities. Despite creating multiple features for the model, there remains great potential for enhancing predictive accuracy

through the development of additional relevant features. Future work could explore these possibilities to improve model performance further. Overall, this study confirms the importance of careful feature selection and model design in educational data mining, providing useful insights for supporting decision-making processes in academic institutions.

References

1. Daniel, B. Big data and analytics in higher education: Opportunities and challenges. *Br. J. Educ. Technol.* **2015**, *46*, 904–920. [Google Scholar] [CrossRef]
2. Zohair, L.M.A. Prediction of student's performance by modelling small dataset size. Int. J. Educ. Technol. High. Educ. 2019, 16, 27. [Google Scholar] [CrossRef]
3. Hellas, A.; Ihantola, P.; Petersen, A.; Ajanovski, V.V.; Gutica, M.; Hynninen, T.; Knutas, A.; Leinonen, J.; Messom, C.; Liao, S.N. Predicting academic performance: A systematic literature review. In Proceedings of the Companion of the 23rd Annual ACM Conference on Innovation and Technology in Computer Science Education, Larnaca, Cyprus, 2–4 July 2018; pp. 175–199. [Google Scholar]
4. Baradwaj, B.K.; Pal, S. Mining educational data to analyze students' performance. Int. J. Adv. Comput. Sci. Appl. 2012, 2, 63–69. [Google Scholar] [CrossRef]
5. Zhang, L.; Li, K.F. Education analytics: Challenges and approaches. In Proceedings of the 2018 32nd International Conference on Advanced Information Networking and Applications Workshops (WAINA), Krakow, Poland, 16–18 May 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 193–198. [Google Scholar] [CrossRef]
6. Van Petegem, C., Deconinck, L., Mourisse, D., Maertens, R., Strijbol, N., Dhoedt, B., ... & Mesuere, B. (2023). Pass/fail prediction in programming courses. Journal of Educational Computing Research, 61(1), 68-95.

7. Balachandar, V., & Venkatesh, K. (2025). A multi-dimensional student performance prediction model (MSPP): An advanced framework for accurate academic classification and analysis. MethodsX, 14, 103148.
8. Balachandar, V., & Venkatesh, K. (2025). A multi-dimensional student performance prediction model (MSPP): An advanced framework for accurate academic classification and analysis. *MethodsX*, *14*, 103148.
9. Zhang, X., Zhang, Y., Chen, A. L., Yu, M., & Zhang, L. (2025). Optimizing multi label student performance prediction with GNN-TINet: A contextual multidimensional deep learning framework. *PloS one*, *20*(1), e0314823.
10. Althibyani, H. A. (2024). Predicting student success in MOOCs: A comprehensive analysis using machine learning models. *PeerJ Computer Science*, *10*, e2221.
11. Chandra, A. M. S. S., Bhoomika, G., Manisha, M., Ganesh, Y. J. S., & Shareefunnisa, S. (2024, July). Analyzing Student Performance for Early Intervention: A Binary Classification Study. In *2024 8th International Conference on Inventive Systems and Control (ICISC)* (pp. 568-574). IEEE.