Iraqi Journal of Information and Technology Systems
DOI:
ISSN 3079-580X (Online), Vol.1, No.1, P.P. 46-53
Received May 3, 2025 , accepted May 15, 2025

An Analysis of Datasets for Student Performance Evaluation Based on Machine Learning

Husam Kadhim Gharkan[1], Mustafa Radif [2], Ali Hakem Alsaeedi[2*]

[1] *Ministry of Education, Al-Qadisiyah Directorate of Education, Al-Qadisiyah,5009 Iraq*
[2]*College of Computer Science and Information Technology, University of Al-Qadisiyah, Al-Qadisiyah,5009 Iraq*
Corresponding email Email: ali.alsaeedi@qu,edu,iq
…

Received May 3, 2025, Revised May 15, 2025, Accepted May 18, 2025, Published May 22, 2025

Abstract: The prediction of student academic performance through data analysis and machine learning has become increasingly significant for improving educational systems and student outcomes. This study explores three publicly available datasets—Open University Learning Analytics Dataset (OULAD), xAPI-Edu-Data, and the UCI Student Performance dataset—to identify critical factors influencing academic success across secondary and higher education levels. These datasets include diverse data sources such as institutional records, behavioral engagement logs, and survey responses. The analysis employs various machine learning and deep learning techniques to evaluate the predictive performance of traditional models, ensemble methods, and hybrid architectures. An experimental validation was performed using a Random Forest classifier on the xAPI-Edu-Data dataset, achieving 80.56% accuracy and supporting the reliability of ensemble approaches. Results show that advanced ensemble and hybrid models like RF-SVM and ECNN-ResNet outperform classical methods in terms of accuracy, precision, recall, and F1-score. This research underscores the potential of intelligent systems to enable early identification of at-risk students, facilitate personalized learning interventions, and inform strategic academic decisions. These findings contribute to a growing body of work that promotes data-driven, equitable, and effective educational practices.

Keywords: Student performance prediction, data analysis, machine learning ; data mining; student performance dataset

1. Introduction

Analyzing data sets is an important process in which data is examined, cleansed, transformed, and modeled to gain meaningful insights, draw conclusions, and make informed decisions [1]. It involves various methods to uncover data patterns, trends, and relationships, including statistical analysis, data visualization, and machine learning. Analyzing data sets breaks complex information into manageable components, transforming raw data into actionable insights [2-4].

Dataset analysis plays a critical role in improving student achievement and the overall quality of education by providing actionable insights that support personalized learning and targeted interventions [5, 6]. By analyzing student behavior, attendance, and academic outcomes, educators can identify at-risk students and provide timely support to prevent academic failure. In addition, insights into students' strengths and weaknesses enable the development of customized learning experiences that increase the effectiveness of instruction and ensure that students receive the support they need to succeed. Beyond personalized learning, analyzing data sets helps to improve teaching methods and curriculum design. By analyzing patterns in student achievement data, educators can evaluate the effectiveness of different teaching strategies and identify areas of the curriculum that need adjustment. This evidence-based approach to teaching allows schools to continually refine their practices and focus on the methods and content that achieve the best outcomes for learners[7]. This leads to more informed and strategic decisions in resource allocation, programmed development, and policy making. In addition, analyzing data sets promotes equity and accountability within education systems. It enables the identification of disparities in outcomes between different demographic groups, supporting efforts to create a more inclusive learning environment [8, 9]. At the same time, data-driven insights' transparency fosters a culture of accountability, ensuring that educational strategies are evaluated and adjusted based on clear, evidence-based metrics. Analyzing data sets supports more effective, equitable, and data-driven educational practices that ultimately improve student outcomes and overall system performance.

The primary goal of dataset analysis is to address key questions, validate hypotheses, or refute theories by systematically inspecting, preparing, and modeling the data [1, 10]. In doing so, the analysis process follows several key steps:

1- **Understanding the data**: Exploring the characteristics, patterns, and relationships inherent in the dataset.

2- **Data cleaning and preparation**: Resolving any inconsistencies, missing values, or errors that could impact the validity of the analysis.

3- **Application of statistical and computational methods**: Utilizing descriptive and inferential statistics and machine learning techniques to analyze and interpret the data.

4- **Data visualization**: Employing charts, graphs, and other visual tools to communicate the findings effectively.

5- **Concluding and making recommendations**: Using the insights obtained from the analysis to inform decision-making, answer research questions, and provide actionable solutions

A. Problem Statement

Despite growing interest in the use of artificial intelligence and data analytics in education, many institutions still struggle to effectively apply these technologies to predict student performance and enhance academic outcomes. Traditional evaluation methods often fail to capture the complexity of learning behaviors, particularly in blended or digital environments. Moreover, there is limited integration between diverse data sources such as institutional records, LMS usage logs, and behavioral indicators, which hinders the development of robust predictive models. Addressing this gap requires a systematic analysis of available educational datasets and the implementation of machine learning models capable of delivering actionable insights.

B. Motivation

In recent years, the use of data analysis in education has gained significant attention due to its potential to enhance learning outcomes and institutional effectiveness. Educational institutions now collect vast amounts of data related to students' academic records, behavioral patterns, and engagement with learning platforms. Analyzing this data plays a crucial role in identifying students who may be at risk of academic failure, allowing for timely interventions and personalized support. Through data-driven strategies, educators can better understand individual learning needs, adapt teaching methods, and implement targeted improvements to curricula. Furthermore, analyzing datasets enables school administrators and policymakers to make informed decisions that promote fairness, accountability, and continuous improvement. By leveraging educational data effectively, institutions can create more equitable and responsive learning environments that benefit all students.

C. Contribution

This study investigates the application of dataset analysis to predict and improve student academic performance in both secondary and higher education contexts. The research aims to identify key factors that influence student success by examining publicly available datasets containing demographic, academic, and behavioral information. Through the use of statistical analysis and machine learning techniques, the study uncovers patterns and trends that provide valuable insights into student learning processes. These insights can support educators and decision-makers in designing more effective learning interventions and policies. By combining traditional data analysis with modern computational approaches, the study contributes to the growing field of educational data mining and offers practical recommendations for enhancing student achievement through evidence-based practices.

D. Organization of the Paper

The remainder of this paper is organized as follows:

- **Section 1** introduces the importance of dataset analysis in the context of student performance evaluation and outlines the research motivation and contributions.
- **Section 2** describes various data collection scenarios relevant to academic environments, including institutional data, LMS activity, behavioral indicators, and survey-based data.
- **Section 3** presents a detailed overview of three publicly available student performance datasets—OULAD, xAPI-Edu-Data, and the Student Performance dataset—highlighting their key features and attributes.
- **Section 4** discusses experimental validation by using the RF algorithm
- **Section 5** discusses the application of machine learning and deep learning techniques to these datasets, summarizing state-of-the-art results achieved in recent research, including experimental evaluation.
- **Section 6** provides the conclusion and outlines directions for future work, emphasizing the importance of integrating diverse data sources and advancing predictive analytics in education.

2. DATA COLLECTION SCENARIO

Collecting data from a college to predict student performance involves several steps. This study can suggest several scenarios that can be used in collecting data for student evaluation:

A. Institutional Data

- **Scenario**: Gather data directly from the college's administrative systems. This could include information like student demographics (age, gender, ethnicity), admission scores (SAT, ACT), course enrollment history, and academic records (grades, credits earned).
- Data Sources:
  1- Student enrollment databases
  2- Registrar's office records
  3- Admission office data
- Use Case: Predicting overall GPA or graduation rates based on historical data

B. Learning Management System (LMS) Data:

- **Scenario:** Utilize data from the college's LMS platform. LMSs track student interactions with course materials, assignments, quizzes, and discussion forums.
- Data Sources:
  1- Clickstream data (e.g., time spent on course materials)
  2- Assignment submissions

3- Quiz scores

- **Use Case**: Predicting course-specific performance (e.g., final exam scores) based on LMS activity.

C. Survey Data:

- **Scenario:** Conduct surveys or questionnaires to collect additional information directly from students. These surveys can cover various aspects of student life, including study habits, well-being, and social interactions.
- **Data Sources:**
  1- Custom-designed surveys
  2- Existing validated surveys (e.g., measuring stress levels, study motivation)
- **Use Case:** Incorporating survey responses into predictive models to understand non-academic factors affecting performance.

D. Behavioral Data:

- **Scenario:** Capture behavioral data related to student engagement, attendance, and participation**.**
- **Data Sources:**
  1- Attendance records
  2- Library usage
  3- Extracurricular activities
- **Use Case:** Predicting student success based on their level of involvement in campus life.

E. Social Network Analysis:

- **Scenario:** Explore social connections among students. Who interacts with whom? Are there influential students within specific groups?
- **Data Sources:**
  1- Social network data (e.g., friendship networks, study groups)
  2- Communication patterns (emails, messages)
- **Use Case: Investigating how social ties impact academic performance.**

F. Multi-Source Heterogeneous Data with Deep Learning:

- **Scenario:** Combine data from various sources (institutional, LMS, surveys) using deep learning techniques.
- **Data Sources:**
  All of the above scenarios
- **Use Case:** Building an end-to-end deep learning model that automatically extracts features from diverse student behavior data to predict academic performance**.**

3. PUBLIC STUDENT PERFORMANCE DATASETS

In this section, we will discuss three important and most widely used databases in student performance, available for use in scientific research. We will explore them and show their most important components.

A. Open University Learning Analytics dataset

The Open University Learning Analytics dataset (OULAD) contains data from courses at the Open University (OU) [11], a large distance-learning university. The dataset includes demographic data about students and clickstream data logging their interactions with the university's virtual learning environment (VLE). This allows analysis of student behavior and learning. The dataset covers 22 courses taken by 32,593 students, including their assessment results and 10,655,280 daily activity logs in the VLE. The dataset was anonymized to protect student privacy, with quasi-identifying attributes like gender, age, and location generalized using the ARX anonymization tool. The OULAD dataset is freely available for research purposes under a CC-BY 4.0 license and has been certified by the

Open Data Institute. The dataset enables research on predictive models for student assessment and learning outcomes, as well as analysis of how the VLE structure impacts learning.

B. Students' Academic Performance Dataset

The dataset, called xAPI-Edu-Data, was collected from a learning management system (LMS) called Kalboard 360. It includes 480 student records with 16 features. Features fall into three categories, demographic, academic background, and behavioral. Demographic features include gender and nationality. Students come from various countries, such as Kuwait, Jordan, Palestine, Iraq, and Lebanon. Academic Background Features Include educational stage (e.g., lower level, middle school, high school), Grade level (e.g., G-01, G-02), Section ID (classroom), Behavioral Features Includes raised hand in class, visited resources (course content), Viewing announcements. Participation in discussion groups. Parent participation (answering surveys, school satisfaction), Students are classified based on absence days (above 7 or under 7)

C. Student pro-mat Dataset

The dataset focuses on student achievement in secondary education in two Portuguese schools, It includes student grades[12], demographic, social, and school-related features, Two datasets are provided for distinct subjects: Mathematics (mat) and Portuguese language (por), The target attribute is G3, representing the final year grade (issued at the 3rd period), G3 has a strong correlation with attributes G2 (second-period grade) and G1 (first-period grade), Features include information such as school type, student sex, age, family size, parental education, mother's and father's jobs, travel time, study time, and more, Quality of family relationships, free time, going out with friends, alcohol consumption, health status, and school absences are also part of the dataset. Table 1 shows a summary of the datasets.

Table 1 : summary of student performance datasets

No	Dataset	# of students	Features	Attributes	Criteria
1	**(OULAD)**	32,593	**-Demographic Features**: Gender Age Band Region Highest Education Level IMD Band (Index of Multiple Deprivation) Disability Status **-Academic Features:** Module ID (code_module(Presentation ID (code_presentation) Number of Previous Attempts Studied Credits Final Result **-Interaction Features:** Clickstream Data (e.g., number of interactions with VLE materials)	**-Demographic Attributes:** Gender (Male, Female) Age Band (e.g., 18-24, 25-34, etc). Region (various geographical areas) Highest Education Level (e.g., High School, Bachelor's, etc). IMD Band (various categories based on deprivation levels) Disability (Yes, No) **Academic Attributes:** Module Codes (specific identifiers for each module) Presentation Codes (identifiers for module presentations) Final Result (e.g., Pass, Fail) **-Interaction Attributes:** Click counts for various VLE materials (e.g., number of clicks on specific resources)	**Assessment Types:** Tutor Marked Assessments (TMA) Computer Marked Assessments (CMA) Final Exam **Performance Metrics:** Scores on assessments (ranging from 0 to 100) Final results (Pass, Fail)
2	**xAPI-Edu-Data**	480	-User ID -Timestamp -Activity type -Resource type -Interaction metrics	-Engagement levels -Time spent on activities -Completion status	-Engagement analysis -Performance prediction based on interactions
3	**Student-por**	649	-Demographic details -School information -Grades in various subjects -Behavioral metrics	-Academic performance -Learning behaviors	-Secondary education achievement

4	**Student-mat**	395	-Similar to Student-por, focusing on math performance -Demographic details -Academic engagement	-Subject-specific grades -Learning behaviors	-Achievement in math subjects

4. AN EXPERIMENTAL VALIDATION BY USING THE RF ALGORITHM

This comparative analysis shows that ensemble and hybrid models often outperform individual traditional ML techniques, especially in handling complex student interaction data.

Figure 1 below summarizes performance metrics across the three datasets, emphasizing ensemble model advantages. An experimental validation was conducted using a Random Forest classifier on the xAPI-Edu-Data dataset. The model achieved an accuracy of 80.56%, with the highest F1-score of 86.67% for the medium-performing student group. This supports prior findings that ensemble models offer robust performance for educational data mining tasks. It presents the confusion matrix for the three-class prediction task (Low, Medium, High).

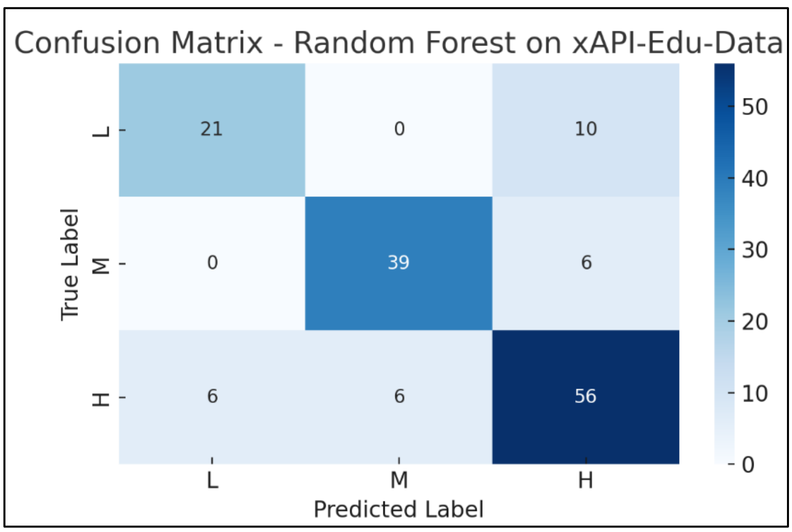

Figure 1: Confusion Matrix of Random Forest Classifier on xAPI-Edu-Data Dataset

5. RESULTS AND DISCUSSION

As a summary of the findings from prior literature and an experimental evaluation performed in this study. A comprehensive analysis of the three datasets—OULAD, xAPI-Edu-Data, and Student Performance—reveals that ensemble and hybrid models consistently outperform traditional machine learning approaches.

A. *For the OULAD dataset*

The highest accuracy (97.88%), recall (98.88%), F1-score (98.88%), and precision (98.88%) were achieved using an ensemble model (RF-SVM) [18]. Deep learning studies using this dataset achieved accuracy up to 93.23% with Artificial Neural Networks (ANN) [20], and F1-score of 91% using Fully Connected Networks (FCN) [22]. The hybrid model combining ECNN and ResNet with Butterfly Optimization achieved 95.67% accuracy and balanced high scores across other metrics [23].

B. *For the xAPI-Edu-Data dataset*

Ensemble models such as Random Forests and Boosting with Decision Trees achieved up to 86.74% accuracy [24-26]. Recent deep learning approaches achieved even better performance, including 93.23% accuracy using Deep Neural Networks (DNN) [27] and F1-score of 84% using Spiking Neural Networks (SNN) [28]. Additionally, experimental validation conducted in this study using Random Forest on xAPI-Edu-Data resulted in 80.56% accuracy. The highest F1-score, 86.67%, was recorded for medium-performing students, affirming the robustness of ensemble models in practical applications.

C. *For the UCI Student Performance datasets (student-mat and student-por)*

Classification models such as Decision Trees and SVMs achieved a maximum of 74.87% accuracy [29], while deep learning models like BiLSTM and SNN showed improved accuracy up to 90%, with recall reaching 97% and F1-score of 95% [28, 30].

Overall, these results underline that ensemble and hybrid techniques demonstrate superior predictive capacity across various educational contexts. The experimental analysis further validates this trend and emphasizes the practical applicability of machine learning for educational data mining tasks. In the field of supervised learning (Classification) [13-18] covered these databases, and it achieved the highest accuracy 97.88 recall 98.88, F1 score 98.88, and precision 98.88 by using an ensemble model (RF-SVM) [18] In the field of deep learning, The research's paper [19-22] covered this databases, and it achieved the highest accuracy of 93.23 and precision 96 by using ANN technology, and recall 88%, and f1-score 91by using FCN technology [22]. The research paper [23] used the Hybrid Model technology and achieved an accuracy of 95.67 % recall of 96.45 %, an F1 score of 93.78%, and a precision of 94.24% by using ECNN and ResNet with butterfly optimization [23].

6. Conclusion

This study has demonstrated the critical role of dataset analysis in evaluating student performance and enhancing educational outcomes. By examining three widely used public datasets—OULAD, xAPI-Edu-Data, and Student Performance—this work highlighted how diverse data sources, including demographic, behavioral, and academic records, can be effectively utilized through machine learning and deep learning techniques. The findings indicate that ensemble models, such as RF-SVM, and hybrid deep learning approaches, including ECNN and ResNet with optimization, achieve high accuracy and reliability in predicting academic success. These results support the application of data-driven strategies for identifying at-risk students, enabling personalized learning, and informing institutional decisions.

Future work may focus on incorporating real-time and multi-modal data, including emotional and social indicators, to build adaptive and context-aware models. The integration of natural language processing for analyzing student feedback and sentiment could further enrich prediction capabilities. Additionally, expanding the scope to cover diverse educational systems and cultural settings will help in assessing the generalizability and fairness of the proposed models. Emphasis on explainable AI is also essential to increase transparency and user trust, thereby encouraging broader adoption of intelligent systems in educational environments.

Data Availability Statement

The data used are included in the manuscript

Conflict of Interest

The authors declare no conflict of interest in influencing the work in this paper

7. References

[1] J. W. Osborne, *Best Practices in Data Cleaning: A Complete Guide to Everything You Need to Do Before and After Collecting Your Data*. Thousand Oaks, CA: Sage Publications, 2012.
[2] B. Baesens, *Analytics in a Big Data World: The Essential Guide to Data Science and Its Applications*. Hoboken, NJ: John Wiley & Sons, 2014.
[3] W.-J. Chen *et al.*, *Systems of Insight for Digital Transformation: Using IBM Operational Decision Manager Advanced and Predictive Analytics*. IBM Redbooks, 2015.
[4] C. B. Davis, *Making Sense of Open Data: From Raw Data to Actionable Insight*. Next Generation Infrastructures Foundation, 2012.
[5] J. Horton, "Identifying at-risk factors that affect college student success," *Int. J.Process Educ.*, vol. 7, no. 1, pp. 83–

101, 2015.
[6] S. Larose and G. M. Tarabulsy, “Academically at-risk students,” *Handb. Youth Mentoring*, vol. 7, no. 3, pp. 440–453, 2005.
[7] D. J. Moran and R. W. Malott, *Evidence-Based Educational Methods*. Elsevier, 2004.
[8] W. J. Jordan, “Defining equity: Multiple perspectives to analyzing the performance of diverse learners,” *Rev. Res. Educ.*, vol. 34, no. 1, pp. 142–178, 2010.
[9] S. Peters and L. A. Oliver, “Achieving quality and equity through inclusive education in an era of high-stakes testing,” *Prospects*, vol. 39, pp. 265–279, 2009.
[10] R. Bodily and K. Verbert, “Review of research on student-facing learning analytics dashboards and educational recommender systems,” *IEEE Trans. Learn.Technol.*, vol. 10, no. 4, pp. 405–418, 2017.
[11] J. Kuzilek, M. Hlosta, and Z. Zdrahal, “Open university learning analytics dataset,” Sci. Data, vol.4, no. 1, pp. 1–8, 2017.
[12] P. Cortez. "Student Performance," UCI Machine Learning Repository, 2008. [Online]. Available: https://doi.org/10.24432/C5TG7T.
[13] M. R. Alzahrani, “Predicting student performance using ensemble models and learning analytics techniques,” 2024. [Unpublished].
[14] S. Rizvi, B. Rienties, and S. A. Khoja, “The role of demographics in online learning: A decision tree based approach,” *Comput. Educ.*, vol. 137, pp. 32–47, 2019.
[15] J. Bravo-Agapito, S. J. Romero, and S. Pamplona, “Early prediction of undergraduate student's academic performance in completely online learning: A five-year study,” *Comput. Human Behav.*, vol. 115, p.106595, 2021.
[16] M. Adnan *et al.*, “Predicting at-risk students at different percentages of course length for early intervention using machine learning models,” *IEEE Access*, vol. 9, pp. 7519–7539, 2021.
[17] K. T. Chui *et al.*, “Predicting at-risk university students in a virtual learning environment via a machine learning algorithm,” *Comput. Human Behav.*, vol. 107, p. 105584, 2020.
[18] A. Al-Ameri *et al.*, “Student academic success prediction using learning management multimedia data with convoluted features and ensemble model,” *ACM J. Data Inf. Qual.*, 2024.
[19] N. Sharma and R. Bhardwaj, “Applying predictive analytics using clickstream data for improving the students performance.” [Unpublished].
[20] H. Waheed *et al.*, “Predicting academic performance of students from VLE big data using deep learning models,” *Comput. Human Behav.*, vol. 104, p. 106189, 2020.
[21] H. Waheed *et al.*, “Early prediction of learners at risk in self-paced education: Aneural network approach,” *Expert Syst. Appl.*, vol. 213, p. 118868, 2023.
[22] M. Hooda *et al.*, “Integrating LA and EDM for improving students success in higher education using FCN algorithm,” *Math. Probl. Eng.*, vol. 2022, no. 1, p.7690103, 2022.
[23] J. A. I. S. Masood *et al.*, “A hybrid deep learning model to predict high-risk students in virtual learning environments,” *IEEE Access*, 2024.
[24] U. Ashfaq, P. Booma, and R. Mafas, “Managing student performance: A predictive analytics using imbalanced data,” *Int. J. Recent Technol. Eng.*, vol. 8, no. 6, p. 6, 2020.
[25] A. A. Alsulami, A. S. A.-M. Al-Ghamdi, and M. Ragab, “Enhancement of e- learning student’s performance based on ensemble techniques,” *Electronics*, vol. 12, no. 6, p. 1508, 2023.
[26] W. Xiao, P. Ji, and J. Hu, “RnkHEU: A hybrid feature selection method for predicting students’performance,” *Sci. Program.*, vol. 2021, no. 1, p. 1670593, 2021.
[27] V. Vijayalakshmi and K. Venkatachalapathy, “Deep neural network for multi- class prediction of student performance in educational data,” *Int. J. Recent Technol. Eng.*, vol. 8, no. 2, pp. 5073– 5081, 2019.
[28] C. Liu *et al.*, “A predictive model for student achievement using spiking neural networks based on educational data,” *Appl. Sci.*, vol. 12, no. 8, p. 3841, 2022.
[29] N. M. Khanian, N. Sarasvathi, and T. S. Mojtaba, “Predictive modeling for student performance data using decision tree and support vector machine,” *INTI J.*, vol. 2019, no. 8, 2019.
[30] B. K. Yousafzai *et al.*, “Student-performulator: Student academic performance using hybrid deep neural network,” *Sustainability*, vol. 13, no. 17, p. 9775, 2021.