Predicting Student Performance Using Data Mining Technology in E-Learning: A Review

Haiam Hamed[1], Lamiaa Fattouh[1,2*] Hesham A. Salman[3], Hadeer Mahmoud[4]

[1]Faculty of Graduate Studies for Statistical Research, Cairo University, Egypt

[2] Modern Academy for Computer Science and Management Technology in Maadi, Cairo, Egypt

[3]Higher Institute of Computer and Information Technology, Alshrouk Academy, Cairo, Egypt

[4]Faculty of Computers and Artificial Intelligence, Modern University for Technology & Information, Cairo, Egypt

***Corresponding Author**

DOI: https://doi.org/10.51244/IJRSI.2024.11120029

Received: 24 December 2024; Accepted: 28 December 2024; Published: 10 January 2025

ABSTRACT

Education data mining is analyzing educational data to improve decision-making and learning outcomes that are important for educational organizations. Utilizing educational data mining techniques is important for improving modern education. During the COVID-19 pandemic, E-learning has become more prevalent, and predicting student performance in this context has become a significant challenge. Studying and analyzing educational data is important, especially when predicting student performance. There are several factors theoretically assumed to affect student performance. These factors include the student's prior academic achievement, study habits, access to resources, quality of teaching, class size, and the learning environment. Additionally, factors such as student engagement, support services, and institutional policies can also have an impact on student performance. After surveying, we found that random forest (RF) and recurrent neural networks (RNN) are the best models to classify and predict student performance. They perform well on educational data and can be efficiently used to predict student performance and in early warning systems.

Keywords: Student performance; performance prediction; educational data mining

INTRODUCTION

Predicting student performance in E-learning through data mining is a significant area of research that aims to enhance the educational experience for students and educators (Sakız, 2021). By leveraging data mining techniques, researchers and educators can analyze student data to identify patterns and trends that may influence academic outcomes (Kariyana, 2012). This approach allows for the early detection of at-risk students and the implementation of targeted interventions to support student success (Miguéis, 2018). In this context, predicting student performance involves exploring factors contributing to academic achievement, such as engagement, study habits, and learning preferences (Amir, 2020). By understanding these factors, E-learning platforms can be optimized to better cater to individual student needs, leading to improved learning outcomes (Zoric, 2019). Currently, numerous techniques are being proposed for evaluating student performance, with data mining standing out as one of the most widely embraced methods for analyzing students' performance (Khang, 2023). This survey highlights the potential benefits and challenges associated with this field.

The rest of the paper is organized as follows: We first discuss background in Section (2). Section (3) presents a discussion and challenge for limitations in Predicting Student Performance in E-learning using Data Mining. In Section (4) Methodology is used in data mining techniques for E-learning. Finally, concludes the paper and points to future work are discussed in section (5).

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XI Issue XII December 2024

BACKGROUND

Educational analytics has become increasingly important for anticipating learner achievement in online learning environments. Learning management systems accumulate extensive behavioral data from student interactions, participation levels, assessments, and learning patterns. By employing analytical techniques and algorithms, to investigate these sizable datasets, educators, and institutions can gain useful insights (Seyedan, 2020)**.** Educational data mining involves discovering patterns and extracting meaningful information from large collections of student data. When utilized in digital education, it provides an important understanding for forecasting learner outcomes. Predictive modeling applies statistical and machine learning algorithms to examine past information and project future academic performance. These models can recognize early signs of success or risk factors for underperformance, permitting initiative-taking assistance for struggling learners and improving general learning results (Fischer, 2020).

Educational data mining in online learning can also contribute to personalized curricula, adaptive course content, and targeted interventions tailored to diverse student needs and styles. By capitalizing on analytical capabilities, instructors can optimize pedagogical design, curriculum development, and resource allocation to enhance learner involvement and achievement in digital environments (Alam, 2023).

Deep learning techniques have been increasingly utilized to predict student performance in educational settings. These methods utilize artificial neural networks, which can learn from extensive data to recognize intricate patterns and connections. Deep learning models can accurately predict a student's future performance by examining a range of factors, including the student's demographics, academic background, and behavioral patterns (Waheed H. H., 2020)**.**

One common approach is to use recurrent neural networks (RNNs) or long short-term memory (LSTM) networks to analyze sequential data, such as a student's academic progress over time. This allows the model to capture temporal dependencies and trends in the student's learning journey (He, 2020). Additionally, convolutional neural networks (CNNs) can be employed to process visual data, such as handwritten assignments or diagrams, to gain insights into a student's understanding and engagement (Mubarak, 2022).

Furthermore, deep learning models can leverage natural language processing (NLP) techniques to analyze text-based data, such as essays or responses to open-ended questions. This enables the model to assess the quality of written work and provide insights into a student's communication and critical thinking skills (Karasavvidis, 2022).

By integrating these diverse sources of data, deep learning models can provide educators with valuable insights into individual student needs, identify at-risk students, and tailor personalized interventions to support student success. However, it is important to ensure that using these techniques is ethical and respects student privacy, while also considering potential biases in the data and model predictions (Baker, 2021).

In summary, deep learning can be a powerful approach to analyzing and extracting meaningful insights from the vast amount of data generated by E-learning platforms (Waheed H. H., 2020) By using techniques such as neural networks and natural language processing (Moubayed, 2018), deep learning algorithms can help to Personalized learning experiences by analyzing student data. Deep learning can identify patterns and preferences, allowing for the customization of learning materials and recommendations based on individual needs (Shemshack, 2021).

Predict student performance: Deep learning models can analyze student behavior and performance to predict future outcomes, allowing educators to provide targeted interventions and support (Namoun, 2020). Improved content recommendation: By understanding student preferences and learning styles, deep learning can enhance the recommendation of relevant and engaging learning materials (Murtaza, 2022).

Enhance assessment and feedback: Deep learning can automate the grading and feedback process, providing immediate and personalized responses to students (Waheed H. H., 2020).

Data mining techniques, on the other hand, can be used to **extract valuable insights and patterns from E-**

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XI Issue XII December 2024

learning data, such as identifying common learning pathways, understanding student behavior, and improving course design (Lee, 2023).

By combining deep learning and data mining, E-learning platforms can gain a deeper understanding of student learning behaviors, improve the effectiveness of teaching and learning, and provide a more personalized and engaging educational experience (Manikandan, 2021).

Challenges

Several challenges and limitations in predicting student performance in E-learning using data mining are associated with predicting student performance. Some of these problems include Data quality: The accuracy and reliability of the data used for prediction can be a significant challenge. When the data used for predictions is incomplete, inaccurate, or outdated, it can result in unreliable forecasts (Jain, 2020) Data privacy: There are ethical and legal considerations related to the collection and use of student data for predictive purposes. Ensuring the privacy and security of student information is crucial (Jones, 2020). Over-reliance on data: Relying solely on data-driven predictions may overlook important contextual and qualitative factors that can influence student performance (Ali, 2022).

Interpretability of models: The complex nature of data mining models can make it difficult to interpret the reasons behind a particular prediction. Understanding the factors that contribute to student performance predictions is crucial for educators to provide targeted support (Sahlaoui, 2021).

Bias and fairness: Data mining models can inadvertently perpetuate biases present in the data, leading to unfair predictions and potentially exacerbating existing inequalities (Varona, 2022).

Dynamic nature of learning: Student performance is influenced by a wide range of dynamic and evolving factors, such as motivation, engagement, and external circumstances, which may not be fully captured in static data (Sarker, 2021).

Addressing these challenges requires a careful and thoughtful approach to the use of data mining in E-learning, including the development of transparent and ethical practices as well as the integration of qualitative insights and contextual understanding alongside quantitative data analysis.

METHODOLOGIES USED IN DATA MINING TECHNIQUES FOR E-LEARNING.

Deep learning has become an advanced method in recent years, with applications in many different domains (Khan M. J.-4., 2019). Deep learning has proven to be a very effective technique since neural networks can extract higher-level concepts by learning the features of the input.

A deep learning technique derived from machine learning imitates the neural network’s structure and mode of operation in the human brain. Using model training, the recognition and classification jobs are realized. In contrast to traditional machine learning techniques, deep learning can handle more complex tasks and datasets, resulting in improved presentation and generalization capabilities (Gordan, 2022) Since the development of multilayer perceptron (MLP), convolutional neural networks (CNN), and recurrent neural networks (RNN), deep learning has undergone numerous stages of evolution before reaching the state of use that it is in today. Deep learning is becoming widely used in fields like vision recognition and speech recognition because of advances in processing power and algorithms (Singh, 2021).

This section will present a comprehensive analysis of the application of data mining techniques, organized by algorithms, for predicting student performance.

Supervised learning, unsupervised learning, and reinforcement learning are three types of machine learning algorithms employed for analyzing E-learning data.

Supervised Learning

Supervised learning can be applied to knowledge tracing and result prediction, but it requires a significant

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XI Issue XII December 2024

amount of labeled historical data. The supervised learning models can be trained using these data to create input-to-output mappings and predict potential achievement and knowledge mastery among students (Kotsiopoulos, 2021).

In the domains of computer vision and image processing, CNN serves as a prime example of deep learning models. Convolution and pooling techniques are used in CNN theory to extract features and reduce image size. The highly dimensional image is transformed into vector data that is one-dimensional and suitable for use in tasks such as classification and regression (O'Mahony, 2020). In the process of activating the CNN network, the ReLU algorithm is a nonlinear function that helps the network learn features more effectively and enhances the model's nonlinearity (You, 2019). Furthermore, expertise in natural language processing (NLP) using deep learning is valuable for analyzing textual data in E-learning (Borakati, 2021), such as student essays, forum discussions, and feedback. Understanding how to preprocess, represent, and model textual data using deep learning techniques is essential (Bernius, 2022).

Additionally, familiarity with unsupervised learning methods such as autoencoders and generative adversarial networks (GANs) can be beneficial for tasks like recommendation systems, anomaly detection, and data augmentation in E-learning data mining (Tran, 2023).

Overall, a solid background in deep learning techniques, including neural network architectures, optimization algorithms, and model evaluation, is crucial for effectively applying data mining in E-learning using deep learning methods (Soui, 2023).

Here are some common supervised learning methods that could be used to predict student performance:

- Logistic regression: This is a popular method for classification problems like predicting if a student will pass or fail. It can identify important factors that influence performance (Hashim, 2020).

- Decision trees: The decision tree technique is widely used for prediction due to its simplicity and ability to reveal patterns in both small and large datasets (Charbuty, 2021). Decision tree models are easily understandable because of their logical reasoning process and can be directly converted into a set of IF-THEN rules." Predicting the academic performance" of MCA students in their third semester and determining career paths based on student behavioral patterns were among the evaluations conducted (Silva, 2023). These assessments utilized features extracted from an education web-based system, including students' final grades (Sokkhey, 2020), final cumulative grade point average (CGPA), and marks obtained in specific courses. Researchers analyzed these datasets to identify the key attributes or factors influencing student performance and explored suitable data mining algorithms for predicting student performance (Tatar, 2020).

- Random forests can capture complex interaction effects between input variables and provide estimates of what variables are important in the classification or regression. It delivers state-of-the-art predictive performance for both classification and regression problems.

- Neural networks: are a widely used and popular technique in educational data mining. One advantage of using neural networks in predicting student performance is their ability to handle complex, non-linear relationships within the data. Neural networks can capture intricate patterns and dependencies in the input features, making them effective for modeling the diverse and often interconnected factors that influence student performance (Tsiakmaki, 2020). Furthermore, neural networks can learn from extensive datasets and adjust to new information, which can be advantageous in predicting student outcomes based on evolving academic and behavioral patterns. Deep neural networks can discover complex patterns in student data to predict performance. Requires large, labeled datasets to be effective (Rivas, 2021).

- Support vector machines (SVM): An SVM can per- form classification by finding the optimal boundary between pass/fail groups. Works well for problems with many attributes (H. Alamri, 2020).

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XI Issue XII December 2024

- K-nearest neighbors: Uses similarity between student profiles to classify new students based on closest matches in the training data (Arcinas, 2021).

- Naive Bayes: The Naive Bayes algorithm is commonly utilized in educational data mining for predicting student performance. hence the" naive" designation. It excels at managing large datasets and is recognized for its straightforwardness and quickness in making predictions. In educational data mining, Naive Bayes can be used to analyze various factors such as student demographics, academic history, and behavioral patterns to predict student performance (Agarwal, 2021). Its ability to handle numerous attributes and provide accurate predictions makes it a valuable tool for educators and researchers to understand and address student needs. Makes independent assumptions but works well for problems with multiple predictive attributes. Can predict categorical outcomes like pass/fail (Alshareef, 2020).

- Regression methods: Linear or logistic regression is commonly used to predict continuous outcomes like exam scores, or GPA based on student attributes and behaviors (Rahman, 2023). The data available, desired accuracy, and interpretability needs would determine the best method.

- Ensembles often outperform individual algorithms for predictive power.

Unsupervised learning

Unsupervised learning, on the other hand, deals with unlabeled data, where only the input features are available (McAlpine, 2022) The goal of unsupervised learning is to discover pat- terns, structures, or relationships within the data without any predefined labels. This can be useful for tasks like clustering similar students together, identifying hidden patterns in student behavior, or uncovering trends in E-learning data (Nikitina, 2020). Unsupervised learning techniques, such as clustering, can be applied to analyze E-learning data. Here are some commonly used algorithms for this purpose:

Clustering algorithms: These algorithms group similar students or course materials based on their characteristics or attributes. Examples include k-means clustering and hierarchical clustering (Choi, 2024).

K-means clustering: This algorithm partitions the data into k clusters based on similarity, allowing for the grouping of students or course materials with similar characteristics (Vankayalapati, 2021).

Fuzzy c-means: Like (k-means), this algorithm allows data points to belong to multiple clusters with varying degrees of membership. It can capture overlapping characteristics in E-learning data (Rayala, 2020).

Gaussian mixture models: This algorithm models the data as a combination of Gaussian distributions and can identify clusters with non-linear boundaries (Khan T. I., 2024)

Hidden Markov models: This algorithm models sequential data, such as student behavior over time, by capturing underlying states and transitions. It can uncover patterns in E-learning data (Tsutsumi, 2023)

Spectral clustering: By leveraging eigenvalues and eigenvectors of a similarity matrix, this algorithm partitions the data into clusters. It is effective for data with non-linear structures (Abdolali, 2021)

Applying these unsupervised learning techniques to E-learning data enables researchers and educators to gain insights into student behavior, performance, and engagement. This information can inform instructional strategies and personalized learning approaches to enhance the overall E-learning experience (Al Nagi, 2020)

Association rule mining: This algorithm identifies patterns and relationships between different items or attributes in the data. It is commonly used for market basket analysis and can be applied to E-learning data to identify frequent patterns of student behavior or course material usage (Aguilera-Hermida, 2020)

Principal component analysis (PCA): This algorithm reduces the dimensionality of the data by identifying the most important features or variables that explain the majority of the variance in the data. It can be useful for

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XI Issue XII December 2024

visualizing complex data and identifying underlying patterns (Reddy, 2020)

Anomaly detection: This algorithm identifies unusual or unexpected patterns in the data that deviate from the norm. It can be used to detect potential fraud or identify students who may be struggling with the course material (Ma, 2021).

These algorithms are just a few examples of the many unsupervised learning techniques that can be used to analyze E-learning data and gain insights into student behavior, performance, and engagement.

Reinforcement learning

Reinforcement learning is a different type of machine learning algorithm that is not typically used for analyzing E-learning data directly. Reinforcement learning is more commonly applied in scenarios where an agent learns to take actions in an environment to maximize a reward signal (Morgan, 2020).

However, in the context of E-learning, reinforcement learning could potentially be used indirectly to optimize certain aspects of the learning experience. For example, it could be employed to develop intelligent teaching systems that provide personalized feedback and recommendations to students based on their performance and interactions (Janardhanan, 2023)

In such cases, the specific reinforcement learning algorithms used would depend on the design and goals of the intelligent tutoring system. Some popular reinforcement learning algorithms include Q-learning, Deep Q-Networks (DQN) (Lapan, 2020), and Proximal Policy Optimization (PPO) (Dubé, 2022).

It is important to note that while reinforcement learning can have applications in E-learning, it is not a direct method. for analyzing E-learning data. Other techniques, such as supervised learning or unsupervised learning, are typically more commonly used for analyzing and gaining insights from E-learning data (Al Nagi, 2020).

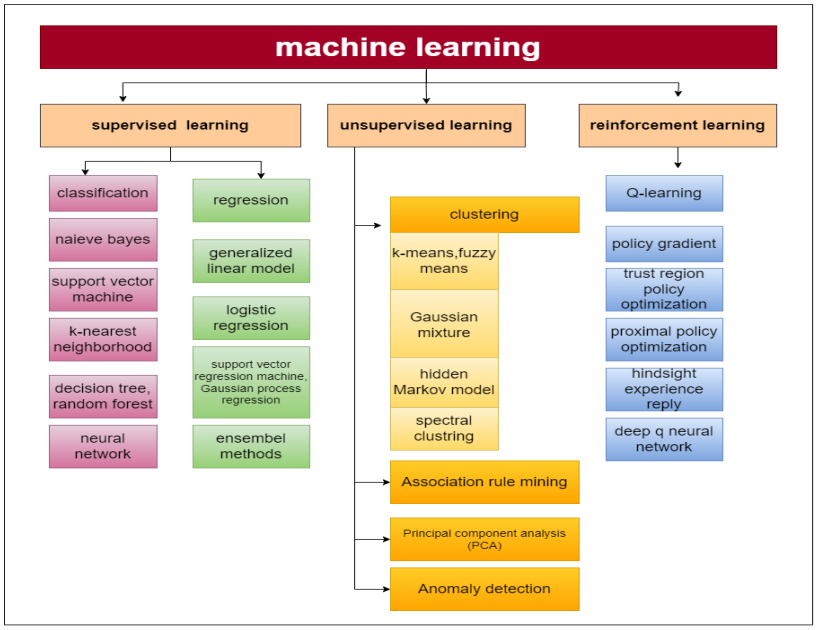

Fig. 1. The methodology used in data mining techniques for E-learning.

(Rebecca, 2020) The aim is to improve the teaching process. Through data collection, the Inductive Miner (IM) algorithm identified the most suitable models for both pass and fail students. The IM algorithm's high-efficiency values enable it to accurately simulate student interactions on the Moodle platform. Furthermore, better results were obtained when the data was separated by units, which is expected as datasets with fewer records tend to yield more accurate measurements.

Also, the work in (Wang, 2021) suggests utilizing a data-mining approach for online English instruction to analyze student behavior. By collecting data on student behavior, a learning behavior for online English

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XI Issue XII December 2024

learning is developed. The behavior data of students in online English education is then mined using a fuzzy neural network, following the application of the apriori algorithm to find association rules and determine data similarity. The experiment's results demonstrate the high efficiency of this strategy in processing data.

In (Nguyen, 2021), Researchers have recognized the significance of leveraging technological advancements and their potential impact on education. They conducted a study to evaluate a novel PFA strategy that incorporates various ensemble learning techniques, including Random Forest, AdaBoost, and XGBoost, to enhance student performance prediction. The experimental results revealed that the scalable XGBoost algorithm significantly outperformed the original PFA algorithm and the other tested models, demonstrating its effectiveness in improving performance prediction. The objective of the study described in (Aremu, 2022) was to predict the exam performance of students. The researchers employed K-nearest neighbor and decision tree algorithms for modeling the study. Based on their findings, the decision tree algorithm demonstrated superior performance in predicting whether students would pass or fail a particular academic course.

In the study conducted by researchers in (Palacios, 2021) data mining techniques were utilized to predict student dropout. The findings indicated the potential for dropout, with average false-positive values ranging from 0.10 to 0.15 and high accuracy rates exceeding 0.80 in many instances. Various machine learning approaches, including logistic regression, Naive Bayes, K-nearest neighbors, random forests, support vector machines, and decision trees, were compared. Among these approaches, random forests outperformed others in terms of precision, F-measure, and accuracy.

In a study by (Begum, 2022) researchers aimed to predict student performance in online sessions using data from E-learning platforms. They monitored student participation and used five commonly used classifiers: logistic regression, naive Bayes, random forest, support vector machines, and multi-layer perception. Three evaluation techniques, including random data splitting and five-fold cross-validation, were employed for training and testing. Results indicated that the random forest (RF) classifier model had the highest accuracy, demonstrating its effectiveness in predicting student performance.

In their study, (Brahim, 2022) examined several classifier algorithms to predict secondary school student performance in mathematics and Portuguese classes. They employed K-nearest neighbor (KNN), support vector machine (SVM), and linear discriminant analysis (LDA) for classification. The results showed that SVM outperformed other approaches for the unbalanced class distribution problem.

In (Ouyang, 2023) proposed the integration of learning analytics techniques with an artificially intelligent (AI) performance prediction model to enhance student learning through collaborative learning. The integrated approach resulted in improved student satisfaction with learning, increased student engagement, and enhanced collaborative learning performance. In this study, (Ahmad, 2021) evaluated artificial neural network (ANN) and random forest (RF) machine learning models for predicting student performance using evaluation and demographic data. They applied various analytical methods to examine the Open University Learning Analytics Dataset (OULAD). and com- pared the performance of the two models. The results showed that the ANN model outperformed the RF model, achieving accuracy ranging from 91.08% to 81.35%. The study highlights the strong performance of artificial neural networks (ANNs) on educational data and their effectiveness in early warning systems and predicting student performance.

Table I: - Various Approaches to Predict Student Performance

Author of Paper	**Model Used**	**Results**
Rebeca Cerezo et al (2020) (Rebeca, 2020)	The Inductive Miner algorithm identified the most	The IM algorithm's high- efficiency values enable accurate simulation of student interactions on the Moodle platform
C.Wang, (2021). (Wang, 2021)	a fuzzy neural network with an apriori algorithm	results demonstrate the high efficiency of this strategy in processing data.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XI Issue XII December 2024

Nguyen et al (2021) (Nguyen, 2021)	Random Forest, and XGBoost	the experimental results revealed that the scalable XGBoost algorithm significantly outperformed the original PFA algorithm
Safira Begum et al (2022) (Begum, 2022)	Support vector machine (SVM)	(SVM) demonstrated the best overall performance
R. Aremu et al (2022) (Aremu, 2022)	K-nearest neighbor and decision tree algorithms	The decision tree algorithm demonstrated superior performance in predicting whether students would pass or fail a particular academic course.
C.Palacios, et al (2021) (Palacios, 2021)	logistic regression, Naive Bayes, K-nearest neighbors, random forests, support vector machines and decision trees,	Different approaches were evaluated based on their performance metrics, including average false-positive values ranging from 0.10 to 0.15 and high accuracy rates exceeding 0.80 in many instances. The results showed that random forest performed better than the other approaches.
Ahmadet al. (2021) (Ahmad, 2021)	Random forest (RF) and Artificial neural network (ANN),	The results indicate that. The ANN model outperformed the RF model. The accuracy achieved by the ANN model ranged from 91.08% to 81.35%.
Ghassen Ben Brahim et al (2022) (Brahim, 2022)	The evaluated classifiers included logistic regression, support vector machines, naive Bayes, random forest, and multi-layer perception.	Results indicated that the random forest (RF) classifier model had the highest accuracy,
S. Begum. S. Padmannavar, et al. (2022) (Begum, 2022)	For classification purposes, the evaluated methods included support vector machine (SVM) and linear discriminant analysis (LDA), K-nearest neighbor (KNN)	The results showed that SVM outperformed other approaches
Ouyang et al. (2023) (Ouyang, 2023)	integration of learning analytics techniques with an AI performance prediction model	The integrated approach resulted in improved student satisfaction with learning, increased student engagement, and enhanced collaborative learning performance.

When evaluating the deep learning performance models in E-learning data mining, several matrix measures can be used to assess the model's effectiveness. These measures are essential for understanding how well the model is performing in tasks such as student performance prediction, dropout prediction, recommendation systems, and sentiment analysis (Al-Fraihat, 2020). Some of the key matrix measures include:

Confusion Matrix: A confusion matrix is a tabular representation of the model's predictions compared to the actual ground truth. It provides a breakdown of true positives, false positives, true negatives, and false negatives, allowing for a detailed assessment of the model's performance (Li, 2023)

Accuracy: is a metric that quantifies the percentage of instances that are correctly classified out of all the instances. It provides an overall indication of the model's correctness in its predictions (Abdelkader, 2022).

Precision: Precision evaluates the ratio of true positive predictions to all positive predictions. This measure is

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XI Issue XII December 2024

crucial in assessing the model's capability to minimize false positives (Ashraf, 2020)

Recall (Sensitivity): Recall assesses the ratio of true positive predictions to all positive instances. It plays a critical role in evaluating the model's ability to identify all relevant instances, as highlighted by Islam et al.in (Islam, 2020)

F1 Score is the harmonic average of recall and precision, offering a balancing measure of the model's performance (Miao, 2020)

Area Under the Curve (AUC-ROC) quantifies the model's capacity to differentiate between classes and is especially beneficial for binary classification tasks. (Movahedi, 2023)These equations provide a quantitative means of assessing the deep learning performance models in E-learning data mining tasks. By using these measures, researchers and practitioners can analyze the effectiveness of the models and make informed decisions about model selection, parameter tuning, and feature engineering.

$$Accuracy = \frac{\{TP + TN\}}{\{TP + TN + FP + FN\}} \quad (1)$$

$$Precision = \frac{\{TP\}}{\{TP + FP\}} \quad (2)$$

$$Recall\ (Sensitivity) = \frac{\{TP\}}{\{TP + FN\}} \quad (3)$$

$$F1\ Score = \frac{2 * (Precision * Recall)}{\{Precision + Recall\}} \quad (4)$$

where: TP denotes the count of accurately predicted in- stances, and FP denotes the count of instances falsely predicted as positive. TN denotes the count of accurately predicted negative instances, and FN denotes the count of instances falsely predicted as negative.

CONCLUSION

Educational Data Mining (EDM) has become pivotal in predicting student performance, a task that has grown increasingly complex due to the vast volumes of data available in educational systems, particularly in the E-learning landscape influenced by the COVID-19 pandemic. Research indicates that Random Forest (RF) and Recurrent Neural Networks (RNN) are among the most effective models for classifying and forecasting student outcomes. These models demonstrate strong performance, when applied to educational datasets, and are highly suitable for developing early warning systems to improve student success and retention.

Future work based on these findings can focus on several promising areas to further enhance the application and effectiveness of Educational Data Mining (EDM) models in predicting student performance such as the Exploration of Additional Machine Learning Models, Personalized Learning Pathways, Integration of Diverse Data Sources, Cross-Institutional Validation, Real-Time Predictive Systems, Incorporation of Explainable AI (XAI), Longitudinal Impact Studies, and Ethical Considerations and Data Privacy.

By pursuing these areas of future work, the effectiveness and application of EDM for predicting student performance can be significantly enhanced, leading to more refined predictive systems and more successful educational outcomes.

Declaration by Authors

Ethical Approval: Approved

Acknowledgment: None

Source of Funding: None

Conflict of Interest: The authors declare no conflict of interest.

REFERENCES

1. Abdelkader, H. E.-1.-6. (2022). Abdelkader, H. E., Gad, A. G., Abohany, A. A., & Sorour, S. E. (2022). An efficient data mining technique for assessing satisfaction level with online learning for higher education students during the COVID-19. IEEE Access, 10, 6286-6303.
2. Abdolali, M. &. (2021). Abdolali, M., & Gillis, N. (2021). Beyond linear subspace clustering: A comparative study of nonlinear manifold clustering algorithms. Computer Science Review, 42, 100435.
3. Agarwal, A. S. (2021). Agarwal, A., Sharma, P., Alshehri, M., Mohamed, A. A., & Alfarraj, O. (2021). Classification model for accuracy and intrusion detection using machine learning approach. PeerJ Computer Science, 7, e437.
4. Aguilera-Hermida, A. P.-1. (2020). Aguilera-Hermida, A. P. (2020). College students' use and acceptance of emergency online learning due to COVID-19. International journal of educational research open, 1, 100011.
5. Ahmad, M. S.-1. (2021). Ahmad, M. S., Asad, A. H., & Mohammed, A. (2021, May). A Machine Learning Based Approach for Student Performance Evaluation in Educational Data Mining. In 2021 International Mobile, Intelligent, and Ubiquitous Computing Conference (MIUCC) (pp. 187-192). I.
6. Al Nagi, E. &.-M.-5. (2020). Al Nagi, E., & Al-Madi, N. (2020, October). Predicting students performance in online courses using classification techniques. In 2020 International Conference on Intelligent Data Science Technologies and Applications (IDSTA) (pp. 51-58). IEEE.
7. Alam, A. (.-2. (2023). Alam, A. (2023, May). Improving Learning Outcomes through Predictive Analytics: Enhancing Teaching and Learning with Educational Data Mining. In 2023 7th International Conference on Intelligent Computing and Control Systems (ICICCS) (pp. 249-257). IEEE.
8. Al-Fraihat, D. J.-l.-8. (2020). Al-Fraihat, D., Joy, M., & Sinclair, J. (2020). Evaluating E-learning systems success: An empirical study. Computers in human behavior, 102, 67-86.
9. Ali, H. A.-L.-8. (2022). Ali, H. A., Mohamed, C., & Abdelhamid, B. (2022, November). Prediction Student Performance in E-learning Environment: Challenge and Opportunity. In The International Conference on Artificial Intelligence and Smart Environment (pp. 861-867). Cham: Springer.
10. Alshareef, F. A. (2020). Alshareef, F., Alhakami, H., Alsubait, T., & Baz, A. (2020). Educational data mining applications and techniques. International Journal of Advanced Computer Science and Applications, 11(4).
11. Amir, L. R.-1. (2020). Amir, L. R., Tanti, I., Maharani, D. A., Wimardhani, Y. S., Julia, V., Sulijaya, B., & Puspitawati, R. (2020). Student perspective of classroom and distance learning during COVID-19 pandemic in the undergraduate dental study program Universitas Indonesia.
12. Arcinas, M. M.-6. (2021). Arcinas, M. M., Sajja, G. S., Asif, S., Gour, S., Okoronkwo, E., & Naved, M. (2021). Role of data mining in education for improving students performance for social change. Turkish Journal of Physiotherapy and Rehabilitation, 32(3), 6519-6526.
13. Aremu, D. R. (2022). Aremu, D. R., Awotunde, J. B., & Ogbuji, E. (2022, January). Predicting Students Performance in Examination Using Supervised data mining techniques. In Informatics and Intelligent Applications: First International Conference, ICIIA 2021, Ota, Nigeria, Nov.
14. Ashraf, M. Z.-1. (2020). Ashraf, M., Zaman, M., & Ahmed, M. (2020). An intelligent prediction system for educational data mining based on ensemble and filtering approaches. Procedia computer science, 167, 1471-1483.
15. Baker, R. S.-4. (2021). Baker, R. S., & Hawn, A. (2021). Algorithmic bias in education. International Journal of Artificial Intelligence in Education, 1-41.
16. Begum, S. &. (2022). Begum, S., & Padmannavar, S. S. (2022). Genetically Optimized Ensemble Classifiers for Multiclass Student Performance Prediction. International Journal of Intelligent Engineering & Systems, 15(2).
17. Bernius, J. P. (2022). Bernius, J. P., Krusche, S., & Bruegge, B. (2022). Machine learning based feedback on textual student answers in large courses. Computers and Education: Artificial Intelligence, 3, 100081.

18. Borakati, A. (.-l.-1. (2021). Borakati, A. (2021). Evaluation of an international medical E-learning course with natural language processing and machine learning. BMC medical education, 21, 1-10.
19. Brahim, G. B.-1. (2022). Brahim, G. B. (2022). Predicting student performance from online engagement activities using novel statistical features. Arabian Journal for Science and Engineering, 47(8), 10225-10243.
20. Charbuty, B. &.-2. (2021). Charbuty, B., & Abdulazeez, A. (2021). Classification based on decision tree algorithm for machine learning. Journal of Applied Science and Technology Trends, 2(01), 20-28.
21. Choi, I. K.-b.-s. (2024). Choi, I., Koh, W., Koo, B., & Kim, W. C. (2024). Network-based exploratory data analysis and explainable three-stage deep clustering for financial customer profiling. Engineering Applications of Artificial Intelligence, 128, 107378.
22. Dubé, S. (. (2022). Dubé, S. (2022). Toward erobotics: An investigation of the relationships between stigma, personality, sexual arousal, and willingness to engage erotically with robots (Doctoral dissertation, Concordia University).
23. Fischer, C. P.-1. (2020). Fischer, C., Pardos, Z. A., Baker, R. S., Williams, J. J., Smyth, P., Yu, R., ... & Warschauer, M. (2020). Mining big data in education: Affordances and challenges. Review of Research in Education, 44(1), 130-160.
24. Gordan, M. S.-Y.-o.-t.-a. (2022). Gordan, M., Sabbagh-Yazdi, S. R., Ismail, Z., Ghaedi, K., Carroll, P., McCrum, D., & Samali, B. (2022). State-of-the-art review on advancements of data mining in structural health monitoring. Measurement, 193, 110939.
25. H. Alamri, L. S. (2020). H. Alamri, L., S. Almuslim, R., S. Alotibi, M., K. Alkadi, D., Ullah Khan, I., & Aslam, N. (2020, December). Predicting student academic performance using support vector machine and random forest. In Proceedings of the 2020 3rd International Conference on.
26. Hashim, A. S. (2020). Hashim, A. S., Awadh, W. A., & Hamoud, A. K. (2020, November). Student performance prediction model based on supervised machine learning algorithms. In IOP Conference Series: Materials Science and Engineering (Vol. 928, No. 3, p. 032019). IOP Publishing.
27. He, Y. C.-r.-G. (2020). He, Y., Chen, R., Li, X., Hao, C., Liu, S., Zhang, G., & Jiang, B. (2020). Online at-risk student identification using RNN-GRU joint neural networks. Information, 11(10), 474.
28. Islam, M. M.-1. (2020). Islam, M. M., Haque, M. R., Iqbal, H., Hasan, M. M., Hasan, M., & Kabir, M. N. (2020). Breast cancer prediction: a comparative study using machine learning techniques. SN Computer Science, 1, 1-14.
29. Jain, A. P. (2020). Jain, A., Patel, H., Nagalapatti, L., Gupta, N., Mehta, S., Guttula, S., ... & Munigala, V. (2020, August). Overview and importance of data quality for machine learning tasks. In Proceedings of the 26th ACM SIGKDD international conference on knowledge dis.
30. Janardhanan, A. K.—a. (2023). Janardhanan, A. K., Rajamohan, K., Manu, K. S., & Rangasamy, S. (2023). Digital education for a resilient new normal using artificial intelligence—applications, challenges, and way forward. Digital Teaching, Learning and Assessment, 21.
31. Jones, K. M. (2020). Jones, K. M., Asher, A., Goben, A., Perry, M. R., Salo, D., Briney, K. A., & Robertshaw, M. B. (2020). “We're being tracked at all times”: Student perspectives of their privacy in relation to learning analytics in higher education. Journal of the Associat.
32. Karasavvidis, I. P.-G.-4. (2022). Karasavvidis, I., Papadimas, C., & Ragazou, V. (2022). Student-Generated Texts as Features for Predicting Learning from Video Lectures: An Initial Evaluation. Themes in eLearning, 15, 21-45.
33. Kariyana, I. M.-c.-1. (2012). The influence of learners’ participation in school co-curricular activities on academic performance: assessment of educators’ perceptions. Journal of Social Sciences, 33(2), 137-146.
34. Khan, M. J.-4. (2019). Khan, M., Jan, B., Farman, H., Ahmad, J., Farman, H., & Jan, Z. (2019). Deep learning methods and applications. Deep learning: convergence to big data analytics, 31-42.
35. Khan, T. I. (2024). Khan, T. I., Sakib, N., Hassan, M. M., & Ide, S. (2024). Gaussian mixture model in clustering acoustic emission signals for characterizing osteoarthritic knees. Biomedical Signal Processing and Control, 87, 105510.
36. Khang, A. G.-d.-1. (2023). Khang, A., Gupta, S. K., Dixit, C. K., & Somani, P. (2023). Data-driven application of human capital management databases, big data, and data mining. In Designing Workforce Management Systems for Industry 4.0 (pp. 105-120). CRC Press.
37. Kotsiopoulos, T. S. (2021). Kotsiopoulos, T., Sarigiannidis, P., Ioannidis, D., & Tzovaras, D. (2021).

Machine learning and deep learning in smart manufacturing: The smart grid paradigm. Computer Science Review, 40, 100341.
38. Lapan, M. (.-O. (2020). Lapan, M. (2020). Deep Reinforcement Learning Hands-On: Apply modern RL methods to practical problems of chatbots, robotics, discrete optimization, web automation, and more. Packt Publishing Ltd.
39. Lee, A. V.-c.-1. (2023). Lee, A. V. Y., Luco, A. C., & Tan, S. C. (2023). A Human-centric automated essay scoring and feedback system for the development of ethical reasoning. Educational Technology & Society, 26(1), 147-159.
40. Li, D. C.-L.-L.-S.-A.-2. (2023). Li, D., Che, M., Meng, W., Wu, Y., Yu, Y., Xia, F., & Li, W. (2023, June). Frame-Level Multi-Label Playing Technique Detection Using Multi-Scale Network and Self-Attention Mechanism. In ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech a.
41. Ma, X. W. (2021). Ma, X., Wu, J., Xue, S., Yang, J., Zhou, C., Sheng, Q. Z., ... & Akoglu, L. (2021). A comprehensive survey on graph anomaly detection with deep learning. IEEE Transactions on Knowledge and Data Engineering.
42. Manikandan, S. D.-l.-2. (2021). Manikandan, S., Dhanalakshmi, P., Priya, S., & Teena, A. M. O. (2021). Intelligent and deep learning collaborative method for E-learning educational platform using TensorFlow. Turkish Journal of Computer and Mathematics Education, 12(10), 2669-2676.
43. McAlpine, E. D.-1. (2022). McAlpine, E. D., Michelow, P., & Celik, T. (2022). The utility of unsupervised machine learning in anatomic pathology. American Journal of Clinical Pathology, 157(1), 5-14.
44. Miao, J. &.-r. (2020). Miao, J., & Zhu, W. (2020). Precision-recall curve (prc) classification trees. arXiv preprint arXiv:2011.07640.
45. Miguéis, V. L.-5. (2018). Miguéis, V. L., Freitas, A., Garcia, P. J., & Silva, A. (2018). Early segmentation of students according to their academic performance: A predictive modelling approach. Decision Support Systems, 115, 36-51.
46. Morgan, D. &.-1. (2020). Morgan, D., & Jacobs, R. (2020). Opportunities and challenges for machine learning in materials science. Annual Review of Materials Research, 50, 71-103.
47. Moubayed, A. I.-l.-3. (2018). Moubayed, A., Injadat, M., Nassif, A. B., Lutfiyya, H., & Shami, A. (2018). E-learning: Challenges and research opportunities using machine learning & data analytics. IEEE Access, 6, 39117-39138.
48. Movahedi, F. (.-a.-L. (2023). Movahedi, F. (2023). Towards Optimizing Left Ventricular Assist Device (LVAD) Therapy for Patients with Advanced Heart Failure: Exploring Machine Learning Applications in Pre-and Post-LVAD Therapy (Doctoral dissertation, University of Pittsburgh).
49. Mubarak, A. A.-2. (2022). Mubarak, A. A., Cao, H., Hezam, I. M., & Hao, F. (2022). Modeling students' performance using graph convolutional networks. Complex & Intelligent Systems, 8(3), 2183-2201.
50. Murtaza, M. A.-b.-l. (2022). Murtaza, M., Ahmed, Y., Shamsi, J. A., Sherwani, F., & Usman, M. (2022). AI-based personalized E-learning systems: Issues, challenges, and solutions. IEEE Access.
51. Namoun, A. &. (2020). Namoun, A., & Alshanqiti, A. (2020). Predicting student performance using data mining and learning analytics techniques: A systematic literature review. Applied Sciences, 11(1), 237.
52. Nguyen, H. T.-5. (2021). Nguyen, H. T. T., Chen, L. H., Saravanarajan, V. S., & Pham, H. Q. (2021, May). Using XG Boost and Random Forest Classifier Algorithms to Predict Student Behavior. In 2021 Emerging Trends in Industry 4.0 (ETI 4.0) (pp. 1-5). IEEE.
53. Nikitina, K. (.-1. (2020). Nikitina, K. (2020). Educational game analysis using intention and process mining. In CEUR workshop proceedings (pp. 117-125).
54. O'Mahony, N. C. (2020). O'Mahony, N., Campbell, S., Carvalho, A., Harapanahalli, S., Hernandez, G. V., Krpalkova, L., ... & Walsh, J. (2020). Deep learning vs. traditional computer vision. In Advances in Computer Vision: Proceedings of the 2019 Computer Vision Conference (CVC), .
55. Ouyang, F. W. (2023). Ouyang, F., Wu, M., Zheng, L., Zhang, L., & Jiao, P. (2023). Integration of artificial intelligence performance prediction and learning analytics to improve student learning in online engineering course. International Journal of Educational Technology in .
56. Palacios, C. A.-S. (2021). Palacios, C. A., Reyes-Suárez, J. A., Bearzotti, L. A., Leiva, V., & Marchant, C. (2021). Knowledge discovery for higher education student retention based on data mining: Machine

learning algorithms and case study in Chile. Entropy, 23(4), 485.
57. Rahman, N. H.-4. (2023). Rahman, N. H. A., Sulaiman, S. A., & Ramli, N. A. (2023). The development of a predictive model for students' final grades using machine learning techniques. Data Analytics and Applied Mathematics (DAAM), 40-48.
58. Rayala, V. &.-M.-7. (2020). Rayala, V., & Kalli, S. R. (2020). Big Data Clustering Using Improvised Fuzzy C-Means Clustering. Rev. d'Intelligence Artif., 34(6), 701-708.
59. Rebeca, C. A.-r.-l.-8. (2020). Rebeca, C., Alejandro, B., María, E., & Romero, C. (2020). Process mining for self-regulated learning assessment in E-learning. Journal of Computing in Higher Education, 32(1), 74-88.
60. Reddy, G. T.-5. (2020). Reddy, G. T., Reddy, M. P. K., Lakshmanna, K., Kaluri, R., Rajput, D. S., Srivastava, G., & Baker, T. (2020). Analysis of dimensionality reduction techniques on big data. Ieee Access, 8, 54776-54788.
61. Rivas, A. G.-B.-7. (2021). Rivas, A., Gonzalez-Briones, A., Hernandez, G., Prieto, J., & Chamoso, P. (2021). Artificial neural network analysis of the academic performance of students in virtual learning environments. Neurocomputing, 423, 713-720.
62. Sahlaoui, H. N.-1. (2021). Sahlaoui, H., Nayyar, A., Agoujil, S., & Jaber, M. M. (2021). Predicting and interpreting student performance using ensemble models and shapley additive explanations. IEEE Access, 9, 152688-152703.
63. Sakız, H. Ö. (2021). Sakız, H., Özdaş, F., Göksu, İ., & Ekinci, A. (2021). A longitudinal analysis of academic achievement and its correlates in higher education. SAGE Open, 11(1), 21582440211003085.
64. Sarker, I. H. (2021). Sarker, I. H. (2021). Deep learning: a comprehensive overview on techniques, taxonomy, applications and research directions. SN Computer Science, 2(6), 420.
65. Seyedan, M. &.-2. (2020). Seyedan, M., & Mafakheri, F. (2020). Predictive big data analytics for supply chain demand forecasting: methods, applications, and research opportunities. Journal of Big Data, 7(1), 1-22.
66. Shemshack, A. K.-5. (2021). Shemshack, A., Kinshuk, & Spector, J. M. (2021). A comprehensive analysis of personalized learning components. Journal of Computers in Education, 8(4), 485-503.
67. Silva, M. P.-1. (2023). Silva, M. P. R. I. R., Rupasingha, R. A. H. M., & Kumara, B. T. G. S. (2023). Identifying complex causal patterns in students' performance using machine learning. Technology, Pedagogy and Education, 1-17.
68. Singh, N. &.-a.-4. (2021). Singh, N., & Sabrol, H. (2021). Convolutional neural networks-an extensive arena of deep learning. A comprehensive study. Archives of Computational Methods in Engineering, 28(7), 4755-4780.
69. Sokkhey, P. &.-b.-p. (2020). Sokkhey, P., & Okazaki, T. (2020). Developing web-based support systems for predicting poor-performing students using educational data mining techniques. International Journal of Advanced Computer Science and Applications, 11(7).
70. Soui, M. &.-b.–C.-1. (2023). Soui, M., & Haddad, Z. (2023). Deep learning-based model using DensNet201 for mobile user interface evaluation. International Journal of Human–Computer Interaction, 39(9), 1981-1994.
71. Tatar, A. E. (2020). Tatar, A. E., & Düştegör, D. (2020). Prediction of academic performance at undergraduate graduation: Course grades or grade point average?. Applied sciences, 10(14), 4967.
72. Tran, H. D. (2023). Tran, H. D. (2023). Towards a new generation of deep learning based recommender systems (Doctoral dissertation, Macquarie University).
73. Tsiakmaki, M. K. (2020). Tsiakmaki, M., Kostopoulos, G., Kotsiantis, S., & Ragos, O. (2020). Transfer learning from deep neural networks for predicting student performance. Applied Sciences, 10(6), 2145.
74. Tsutsumi, E. G. (2023). Tsutsumi, E., Guo, Y., Kinoshita, R., & Ueno, M. (2023). Deep knowledge tracing incorporating a hypernetwork with independent student and item networks. IEEE Transactions on Learning Technologies.
75. Vankayalapati, R. G.-M. (2021). Vankayalapati, R., Ghutugade, K. B., Vannapuram, R., & Prasanna, B. P. S. (2021). K-Means Algorithm for Clustering of Learners Performance Levels Using Machine Learning Techniques. Revue d'Intelligence Artificielle, 35(1).
76. Varona, D. &. (2022). Varona, D., & Suárez, J. L. (2022). Discrimination, bias, fairness, and trustworthy AI. Applied Sciences, 12(12), 5826.
77. Waheed, H. H. (2020). . Predicting academic performance of students from VLE big data using deep

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XI Issue XII December 2024

learning models. Computers in Human behavior, 104, 106189.
78. Waheed, H. H. (2020). Predicting academic performance of students from VLE big data using deep learning models. . Computers in Human behavior, 104, 106189.
79. Waheed, H. H. (2020). Waheed, H., Hassan, S. U., Aljohani, N. R., Hardman, J., Alelyani, S., & Nawaz, R. (2020). Predicting academic performance of students from VLE big data using deep learning models. Computers in Human behavior, 104, 106189.
80. Wang, C. (.-1. (2021). Wang, C. (2021). Analysis of students' behavior in english online education based on data mining. Mobile Information Systems, 2021, 1-10.
81. You, W. S.-1. (2019). You, W., Shen, C., Wang, D., Chen, L., Jiang, X., & Zhu, Z. (2019). An intelligent deep feature learning method with improved activation functions for machine fault diagnosis. IEEE Access, 8, 1975-1985.
82. Zoric, A. B.-7. (2019). Zoric, A. B. (2019). Benefits of educational data mining. Economic and Social Development: Book of Proceedings, 1-7.