
Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2025 SCPE. Volume 26, Issues 3, pp. 1551–1558, DOI 10.12694/scpe.v26i3.4336

EDUCATIONAL DATA MINING FOR STUDENT PERFORMANCE PREDICTION

LINQIANG TANG* AND CHEN SIAN†

Abstract. The topic of Educational Data Mining (EDM) has gained significant traction in improving the quality of education by identifying patterns and insights through the analysis of data gathered from diverse educational settings. In order to discover important elements that affect educational achievement and to give educators and policymakers with useful insights, this study investigates the use of machine learning techniques in predicting student performance. We use a variety of machine learning methods, such as decision trees, support vector machines, and neural networks, to create predictive models by utilizing past educational information, demographics, and behavioral tendencies. The study assesses these models' efficacy and accuracy while also emphasizing how important choosing features and data preparation are to enhancing prediction results. Our results show that applying machine learning approaches can greatly improve the prediction of pupil achievement, which in turn allows for more focused interventions and individualized learning plans. This study highlights the possibilities of machine learning in promoting a data-driven method to educational improvement and adds to the expanding body of knowledge in EDM.

Key words: Educational Data Mining, student performance, prediction, machine learning methods, Educational.

1. Introduction. The abundance of data available in today's educational environment has made it possible for creative methods to improve student learning outcomes [8]. The discipline of Educational Data Mining (EDM) is gaining importance as it uses advanced data mining methods to examine educational data and derive important insights. EDM looks for patterns and trends in the vast amounts of data collected by educational institutions in order to anticipate student performance [10]. This allows teachers to customize interventions and methods for the best possible learning outcomes. EDM is important because it can convert unprocessed data into useful knowledge. Big data and advanced analytics have made it possible for educators and academics to have a deeper knowledge of the many variables influencing students' achievement [15].

A wide range of factors, including behavioral data, socioeconomic indicators, academic records, and demographic information, are included in EDM and contribute to a thorough knowledge of how students perform [6]. A fundamental component of EDM is predictive modeling, which makes use of statistical methods and algorithms to project future results from past data. These algorithms have the ability to pinpoint kids who are at danger, forecast grades, and even provide individualized learning plans [16]. Through the utilization of artificial intelligence and machine learning, EDM not only improves the precision of predictions but also offers insights into the fundamental elements influencing the achievement of students [18].

The increasing awareness of data-driven approaches' potential to transform teaching methods is motivated this research. Large volumes of data are produced when educational institutions use digital tools and systems more frequently, recording numerous facets of student behavior, academic achievement, and demographic traits. Even with this wealth of data at their disposal, many educational institutions nevertheless depend on antiquated, one-size-fits-all strategies that underutilize the insights this data may provide for enhancing student results.

In order to forecast and evaluate student performance, Educational Data Mining (EDM) for Student Performance Prediction applies data mining techniques to educational data. With the goal of assisting educators, administrators, and policymakers in enhancing the educational process and results, EDM seeks to reveal hidden patterns, trends, and insights from educational data [13]. The purpose of this introduction is to highlight the revolutionary possibilities of Educational Data Mining in the field of predicting student performance. Institutions can cultivate a data-driven culture that encourages academic success, reduces dropout rates, and supports students' holistic development by methodically examining educational data [17]. As we learn more about this

*Zhejiang Institute of Communications, Hangzhou, Zhejiang, 311112, China (tanglinqianglearni@outlook.com)
†Zhejiang Institute of Communications, Hangzhou, Zhejiang, 311112, China.

area, it becomes clearer how EDM has the potential to completely transform education and provide hope for a better educated and functioning educational system. The main contribution of proposed method is given below:

1. This research's primary contribution to Educational Data Mining (EDM) for student performance prediction is the creation of a solid, data-driven framework that makes use of cutting-edge machine learning algorithms to predict educational results with high accuracy.
2. In order to produce a comprehensive prediction model, this study takes into account a variety of behavioral, demographic, and socioeconomic characteristics in addition to standard academic indicators.
3. Through a methodical examination of extensive datasets, the study pinpoints significant trends and indicators of student achievement, providing educators and decision-makers with valuable perspectives.
4. This work stands out for its innovative integration of multifaceted data sources and state-of-the-art analytical approaches, which has a substantial positive impact on the development of analytics for prediction in education.

The rest of our research article is written as follows: Section 2 discusses the related work on various educational data Mining. Section 3 shows the algorithmprocess and general working methodology of proposed work. Section 4 evaluates theimplementation and results of the proposed method. Section 5 concludes the work anddiscusses the result evaluation.

2. Related Works. Education data mining (EDM) is a rapidly developing field that analyzes data from educational environments with the goal of improving education. Predicting student performance has become a major area of study for EDM because of its potential to enhance educational results [5]. Highlighting numerous strategies, models, and conclusions, this section examines significant contributions and methodology in this field. Research have used a variety of data sources such as educational records, social media activity, interaction logs from Learning Management Systems (LMS), and student demographic data [3, 20]. The author, for example, used LMS interaction data in conjunction with demographic information and past academic performance to forecast future performance.

For forecasting algorithms to be reliable and efficient, efficient techniques for preprocessing including choosing features, data cleaning, and normalization are essential. Numerous machine learning techniques have been used to forecast student achievement [9]. Diverse degrees of success have been shown by decision trees, artificial neural networks, and ensemble techniques like random forests [2] Models based on deep which can identify intricate patterns in big datasets, seem to be the direction of recent advancements. By comparing multiple algorithms, for instance, the author came to the conclusion that ensemble approaches perform better than single classifiers in most cases [11].

An important factor in the effectiveness of models for prediction is feature design. Numerous aspects have been investigated by researchers, such as involvement in online forums, assignment submission deadlines, attendance records, and even psychological elements like ambition and self-control [14]. Principal component analysis and correlation-based choice of features are two feature selection strategies that have been used to improve the performance of models by minimizing over fitting and complexity. Many metrics, such as accuracy, recall, F1 score, precision, and Area Under the Receiver Operating Characteristics Curve (AUC-ROC), are used to assess the efficacy of models for prediction. The significance of false positives vs false negatives and the particular educational setting are two factors that frequently influence the selection of metrics [4, 1].

In EDM, models of prediction are frequently used in conjunction with strategies for intervention meant to raise student achievement. Predictive analytics has been used to create tailored feedback, early warning systems, and adaptable learning environments [19]. For instance, the author developed a system for early detection that, by giving at-risk students targeted help, greatly increased student retention rates. Predictive model use in education brings up a number of ethical issues, mostly with regard to algorithmic bias, informed consent, and data protection [7, 12]. In order to guarantee responsible utilization of student data, the author talked on the significance of open data practices and the requirement for ethical principles. They support involving all relevant parties in the creation and application of predictive structures, such as teachers, pupils, and legislators.

This problem can be effectively solved with the help of Educational Data Mining (EDM), which makes it possible to analyze educational data systematically and find links and patterns that can guide decision-making.

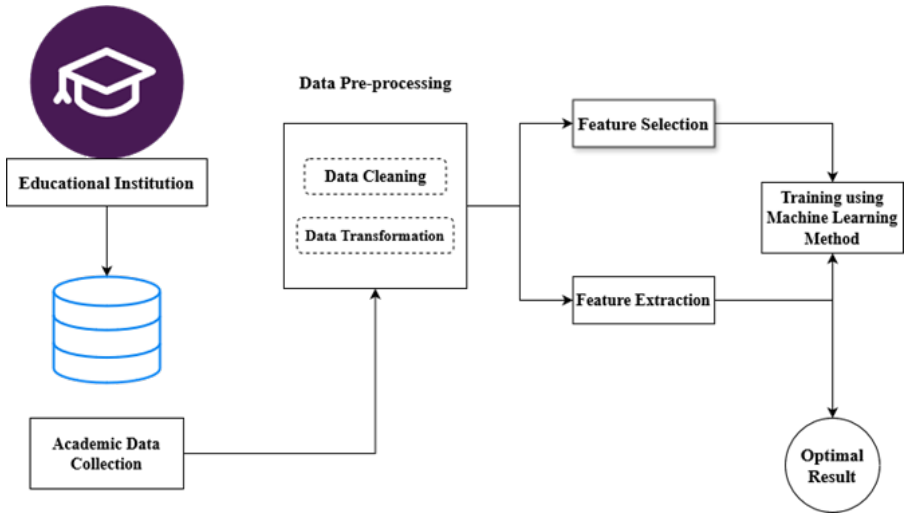

Fig. 3.1: Architecture of Proposed Method

One major development in this area is the use of machine learning algorithms to predict student performance. Educators and policymakers can improve the quality of education by implementing targeted interventions that are tailored to the specific needs of children by precisely forecasting which kids are at danger of underperforming.

3. Proposed Methodology. The proposed methodology is to enhance academic results and comprehend students' learning processes, educational information mining, or EDM, applies methods of data mining to educational data. This methodology's main goal is to forecast pupil achievement utilizing a variety of methods for data mining and instructional information. Using academic data analysis tools, the suggested methodology offers an extensive structure for forecasting student performance. Educational organizations can apply tactics to improve educational results and acquire helpful insights into pupil retention processes by adhering to this methodology. In figure 3.1 shows the architecture of proposed method.

Academic data is first gathered from the educational institution and pre-processed to remove discrepancies and convert the data into a format that may be used. To improve representation, the most pertinent features are found and additional features are created through feature extraction and selection. An ideal predictive model for student performance is produced by training a machine learning model using these features. To make sure this model is accurate and reliable, its performance is assessed using a variety of measures.

3.1. Data Collection. The original source of academic information, such as attendance, performance, and other pertinent data records for students. Offers unprocessed academic data for study. combines information from many educational institution sources. a thorough database with all the pertinent student data.

3.2. Data Pre-processing. Pre-processing data entails Managing missing values: removing or imputation, Eliminating duplicates, fixing mistakes in data entry. Data analysis outcomes can be distorted by irrelevant or meaningless information, which is referred to as noise in the data. Noise can originate from a number of things, including data entry mistakes, malfunctioning sensors, and anomalies that do not accurately reflect the dataset. Methods including filtering, outlier detection, and smoothing are used to get rid of noise. Predictive models become more accurate as noise is reduced because the data is more representative of the real underlying patterns.

3.2.1. Data Cleaning. Entails eliminating noise, dealing with missing data, and fixing inconsistent results.

$$Cleaned\ Data = Raw\ data - Noise \tag{3.1}$$

3.2.2. Data Transformation. In educational datasets, missing data is a prevalent problem. Records may be incomplete for a number of reasons, including student absences, incomplete grades, or mistakes in data collecting. In order to handle missing data, records with large gaps must be removed, or missing values must be imputed using statistical techniques.

$$Transformed\ Data = \frac{cleaned\ data - \mu}{\sigma} \tag{3.2}$$

Ready-to-use preprocessed data for the extraction and selection of features.

The process of transforming cleaned data into a format appropriate for analysis and modeling is known as data transformation. Making sure the data is consistent and suitable for machine learning algorithms requires taking this crucial step. Rescaling data to have a mean ( ) of zero and a standard deviation ( ) of one is the process of standardization. This approach guarantees that every feature contributes equally to the model, which is especially crucial for algorithms that rely on distance measurements, such neural networks or support vector machines. By addressing variables that may have varying scales, standardization helps keep characteristics with larger scales from unduly affecting the model.

3.3. Feature Selection. Determines the most important characteristics that go into predicting a student's achievement. methods such as feature importance from models, mutual information, and correlation analysis.

$$Selected\ Features = \arg max_{Fi}\ \ Importance\ (F_i) \tag{3.3}$$

3.4. Feature Extraction. Uses the available data to generate new features that more accurately capture the underlying patterns.

3.4.1. Principal Component Analysis (PCA). In order to simplify complex datasets and preserve as much variability (information) as possible, data analysts employ principal component analysis (PCA), a dimensionality reduction approach. Principal component analysis (PCA) aims to convert the original data into a new, uncorrelated set of features known as principal components. The arrangement of these elements ensures that the majority of the variety found in the original dataset is retained in the first few.

Scaling the characteristics to have a mean of 0 and a standard deviation of 1 is the process of standardizing the data.

$$Z = \frac{X - \mu}{\sigma} \tag{3.4}$$

The degree of collective feature variation is measured by the covariance matrix. The covariance matrix for a dataset with n features is an n×n matrix.

$$\Sigma = \frac{1}{n-1} Z^T Z \tag{3.5}$$

The covariance matrix's eigenvalues and eigenvectors are calculated. The new feature space's direction is determined by the eigenvectors, while its magnitude, or relevance, is determined by the eigenvalues.

$$\Sigma v = \lambda v \tag{3.6}$$

The eigenvalues have been arranged in descending order by their corresponding eigenvectors. The first principal component is the eigenvector with the highest eigenvalue. To create a new feature space, select the top k eigenvectors. The amount of variation (e.g., 95%) that is desired to be retained determines the value of k.

Reduces the number of features while keeping the most crucial information, which simplifies the dataset. lowers the processing expense of ensuing data processing jobs. Generates uncorrelated characteristics that have the potential to enhance the efficiency of specific machine learning techniques. Projects high-dimensional data into two or three dimensions to aid in its visualization.

3.5. Machine Learning Methods. Applying data mining techniques to educational data in order to predict academic performance and study student behavior is known as educational data mining, or EDM. Institutions can deliver individualized learning experiences, enhance educational achievements, and spot patterns by utilizing machine learning. The procedures for creating a machine learning model to forecast student performance are described in this framework.

3.5.1. Decision Trees. A decision tree is an arrangement that resembles a flowchart, with each internal node denoting a choice made in response to a feature (attribute), each branch denoting the decision's result, and each leaf node representing a class label (in this case, student performance). The routes from the root to the leaf show the guidelines for classification. Collect data about students, covering a range of aspects such as personal and academic history, attendance, behavior, and other pertinent characteristics.

Starting with the complete dataset and choose the feature (e.g., pass or fail) that divides the data into the most distinct classes. Information gain, entropy, and Gini impurity are frequently used criteria to determine the optimal split. The dataset was iteratively divided into subsets according to the chosen characteristic. The goal of each split is to produce subsets that are purer, which means that the subsets are supposed to only include data points from one class.

Student data from the past is used to train the decision tree. The framework discovers connections and patterns among the goal variable—such as grades or pass/fail status—and the characteristics of the inputs. By navigating the tree according to the student's characteristics, the decision tree system can forecast a new student's success.

4. Result Analysis. The efficacy evaluation of forecasting algorithms created with instructional data mining approaches is the main objective of the outcome analysis. Employing a variety of academic and demographic variables, these models seek to predict the performance of students. For a broad range of students, the collection contains educational records, records of attendance, information on demographics, and indicators of socioeconomic status. Academic results (tests, assignments, and final examinations), attendance records, involvement in extracurricular endeavors, educational levels, and other key factors are all examined.

The linear link between the observed and expected values is measured by the correlation coefficient, which also indicates its direction and intensity. There is a significant positive association when the value is 0.86. The average absolute difference between the expected and actual values is represented by the mean absolute error, or MAE. An average deviation of 18.92 units between the predicted and actual values indicates that the predictions are not accurate.

The square root of the average squared discrepancies between the expected and actual values is indicated by the Root-Mean-Squared Error (RMSE). The residuals' (prediction errors') standard deviation is displayed with a value of 24.31. A larger average error is implied by a higher value. The square root of the mean absolute variations between the expected and actual values is what is measured by the root-absolute error. The degree of prediction mistakes is represented by the value of 17.16; a smaller value indicates better results. Root Relative Squared Error (RMSE): Provides a normalized measure of error by reflecting the RMSE in relation to the range of observed values. The RMSE in relation to the data's magnitude is represented by the value of 19.51. In figure 4.1 shows the overall performance metrics of educational data prediction.

Out of the four classifiers, the Decision-Tree classifier has the highest accuracy, just over 95%. At about 85%, the K-NN classifier has the lowest accuracy.While it does not perform as well as the Decision-Tree or GA classifiers, the GA+K-NN classifier outperforms K-NN. The Decision-Tree classifier is the most accurate, followed by the GA, GA+K-NN, and K-NN classifiers, as this figure 4.2 graphically illustrates.

With the lowest RMSE, the GA+Decision-Tree model performs best with the least amount of error. The K-NN model performs the worst and has the most errors, as evidenced by its highest RMSE.The RMSE values of the Decision-Tree (DT) and GA+K-NN models are in the middle, with the Decision-Tree model outperforming GA+K-NN. The GA+Decision-Tree model is the most accurate, followed by the Decision-Tree, GA+K-NN, and K-NN regression models, as this chart 4.3 illustrates clearly.

5. Conclusion. The substantial potential of using data analytics to improve educational outcomes has been shown by the research on educational data mining for pupil achievement prediction. Through the implementation of diverse machine learning techniques and statistical methods on student data, this study has

Fig. 4.1: Performance metrics

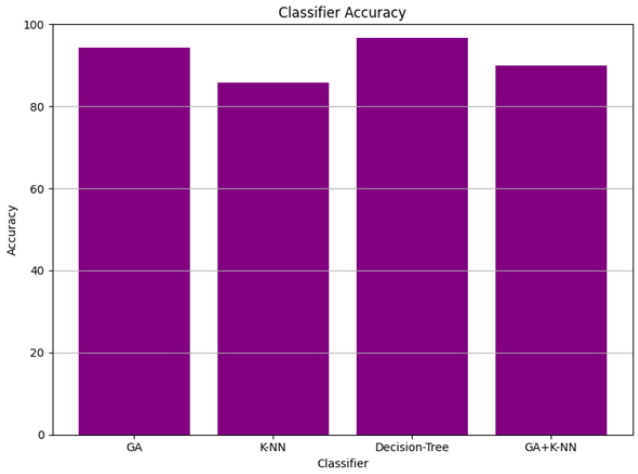

Fig. 4.2: Classification accuracy

effectively discovered pivotal aspects that impact academic achievement. With the help of the models for prediction created in this study, teachers will be able to proactively address any academic challenges that may arise and customize their lesson plans to meet the requirements of each unique student. The findings highlight the value of ongoing data gathering and analysis in learning environments. The knowledge gathered from these studies aids in the creation of individualized education programs in addition to providing insight into the behavior and learning behaviors of students. Additionally, the creation of curriculum, allocation of resources, and institutional decision-making can all be aided by the use of educational data mining, which will ultimately promote a more encouraging and productive learning environment. To further improve the models for prediction, future studies should concentrate on incorporating a wider range of data sources, such as behavioral and socioeconomic variables. To ensure that the advantages of data mining in education are realized without jeopardizing student privacy, ethical concerns about data security and privacy must also be taken into account. To sum up, data mining in education presents a viable way to use information-driven knowledge to raise the

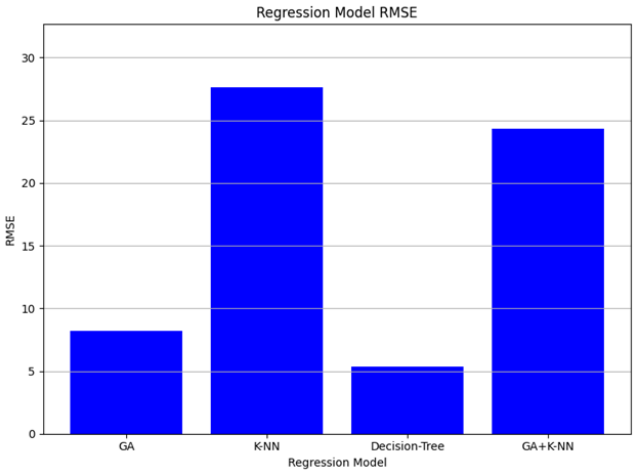

Fig. 4.3: Regression Model RMSE

achievement of students. Schools can raise academic expectations, improve learning experiences, and provide greater assistance for their students by utilizing analytics to predict outcomes.

Subsequent investigations can concentrate on investigating and incorporating increasingly sophisticated machine learning models, like deep learning architectures, ensemble techniques like Extreme Gradient Boosting (XGBoost) and Gradient Boosting Machines (GBM), and hybrid models that blend the advantages of many algorithms. These algorithms might perform better when managing intricate, large-scale educational statistics, resulting in more precise forecasts and profound understanding of student achievement.

REFERENCES

[1] F. A. Al-Azazi and M. Ghurab, *Ann-lstm: A deep learning model for early student performance prediction in mooc*, heliyon, 9 (2023).

[2] A. Alam, *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), IEEE, 2023, pp. 249–257.

[3] ———, *The secret sauce of student success: Cracking the code by navigating the path to personalized learning with educational data mining*, in 2023 2nd International Conference on Smart Technologies and Systems for Next Generation Computing (ICSTSN), IEEE, 2023, pp. 1–8.

[4] S. Albahli, *Efficient hyperparameter tuning for predicting student performance with bayesian optimization*, Multimedia Tools and Applications, 83 (2024), pp. 52711–52735.

[5] A. S. Alghamdi and A. Rahman, *Data mining approach to predict success of secondary school students: A saudi arabian case study*, Education Sciences, 13 (2023), p. 293.

[6] K. Aulakh, R. K. Roul, and M. Kaushal, *E-learning enhancement through educational data mining with covid-19 outbreak period in backdrop: A review*, International journal of educational development, 101 (2023), p. 102814.

[7] C. Baek and T. Doleck, *Educational data mining versus learning analytics: A review of publications from 2015 to 2019*, Interactive Learning Environments, 31 (2023), pp. 3828–3850.

[8] S. Batool, J. Rashid, M. W. Nisar, J. Kim, H.-Y. Kwon, and A. Hussain, *Educational data mining to predict students' academic performance: A survey study*, Education and Information Technologies, 28 (2023), pp. 905–971.

[9] N. Bayes and B. A. Ningsi, *Performance comparison of data mining classification algorithms on student academic achievement prediction*, Indonesian Journal of Artificial Intelligence and Data Mining (IJAIDM), 6 (2023), pp. 29–39.

[10] Y. Chen and L. Zhai, *A comparative study on student performance prediction using machine learning*, Education and Information Technologies, 28 (2023), pp. 12039–12057.

[11] P. Guleria and M. Sood, *Explainable ai and machine learning: performance evaluation and explainability of classifiers on educational data mining inspired career counseling*, Education and Information Technologies, 28 (2023), pp. 1081–1116.

[12] C. Huang, J. Zhou, J. Chen, J. Yang, K. Clawson, and Y. Peng, *A feature weighted support vector machine and artificial neural network algorithm for academic course performance prediction*, Neural Computing and Applications, 35 (2023), pp. 11517–11529.

[13] S. Hussain and M. Q. Khan, *Student-performulator: Predicting students' academic performance at secondary and inter-*

mediate level using machine learning, Annals of data science, 10 (2023), pp. 637–655.
[14] A. Kukkar, R. Mohana, A. Sharma, and A. Nayyar, *Prediction of student academic performance based on their emotional wellbeing and interaction on various e-learning platforms*, Education and Information Technologies, 28 (2023), pp. 9655–9684.
[15] S. Mallak, M. Kanan, N. Al-Ramahi, A. Qedan, H. Khalilia, A. Khassati, R. Wannan, M. Mara'beh, S. Alsadi, and A. AlSartawi, *Using markov chains and data mining techniques to predict students' academic performance*, (2023).
[16] H. Pallathadka, A. Wenda, E. Ramirez-Asís, M. Asís-López, J. Flores-Albornoz, and K. Phasinam, *Classification and prediction of student performance data using various machine learning algorithms*, Materials today: proceedings, 80 (2023), pp. 3782–3785.
[17] M. H. B. Roslan and C. J. Chen, *Predicting students' performance in english and mathematics using data mining techniques*, Education and Information Technologies, 28 (2023), pp. 1427–1453.
[18] X. Wang, Y. Zhao, C. Li, and P. Ren, *Probsap: A comprehensive and high-performance system for student academic performance prediction*, Pattern Recognition, 137 (2023), p. 109309.
[19] T. Wongvorachan, S. He, and O. Bulut, *A comparison of undersampling, oversampling, and smote methods for dealing with imbalanced classification in educational data mining*, Information, 14 (2023), p. 54.
[20] O. R. Yürüm, T. Taşkaya-Temizel, and S. Yildirim, *The use of video clickstream data to predict university students' test performance: A comprehensive educational data mining approach*, Education and Information Technologies, 28 (2023), pp. 5209–5240.

Edited by: Rajkumar Rajavel
Special issue on: Cognitive Computing for Distributed Data Processing and Decision-Making in Large-Scale Environments
Received: Jul 23, 2024
Accepted: Aug 25, 2024