© MAR 2026 | IRE Journals | Volume 9 Issue 9 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I9-1715129

Student Performance Prediction Using Machine Learning Algorithms

S SANJEEV[1], S GUHAN[2], V VEERAKUMARAN[3]
[1,2]*UG Student, Department of Computer Science and Data Science, Nehru Arts and Science College, Coimbatore, Tamil Nadu, India.*
[3]*Assistant Professor, Department of Computer Science and Data Science, Nehru Arts and Science College, Coimbatore, Tamil Nadu, India*

Abstract- ***Education plays a vital role in building a productive life by fostering self-confidence and providing essential knowledge and resources. With the rapid advancement of technologies such as artificial intelligence, higher education institutions are increasingly integrating technological tools into traditional teaching and learning methods. Predicting students' academic success has become an important research focus, as strong academic performance not only enhances a university's reputation and ranking but also improves graduates' employment prospects. However, modern educational institutions face several challenges, including analyzing student performance, maintaining high-quality education, developing effective evaluation strategies, and anticipating future educational needs. E-learning has emerged as a rapidly expanding and advanced mode of education, allowing students to participate in online courses and learning platforms. Technologies such as Intelligent Tutoring Systems (ITS), Learning Management Systems (LMS), and Massive Open Online Courses (MOOCs) utilize Educational Data Mining (EDM) to support automated grading systems, recommendation systems, and adaptive learning environments. Despite these advantages, e-learning environments remain challenging due to the limited direct interaction between students and instructors. Machine learning (ML) plays an important role in the development of intelligent and adaptive systems capable of performing complex tasks that often exceed human capabilities. ML algorithms are widely applied in fields such as cluster analysis, pattern recognition, image processing, natural language processing, and medical diagnostics. In this study, the K-means clustering technique, combined with the Davies–Bouldin index, was used to identify clusters and determine significant features affecting student performance. The research evaluated several classification algorithms, including Support Vector Machine (SVM), Decision Tree, Naïve Bayes, and K-Nearest Neighbors (KNN). The results indicate that the SVM algorithm achieved the highest prediction performance after hyperparameter tuning, reaching an accuracy of 96%. Parameter optimization significantly improved the accuracy of all four prediction models. Among them, the Naïve Bayes classifier demonstrated the lowest predictive performance, primarily due to its assumption of strong independence among features.***

I. INTRODUCTION

Education is vital for a productive life. It builds self-confidence and provides essential resources. With the rise of technology, especially artificial intelligence, colleges and universities are integrating technology into traditional teaching.

Student academic performance is an important measure of educational progress. It is influenced by factors like gender, age, teaching staff, and learning environments. There is growing interest in predicting academic success. A strong academic record boosts a university's ranking and improves student job prospects since employers often consider this a key factor.

Modern learning institutions face challenges in assessing performance, delivering quality education, creating evaluation strategies, and identifying future needs. They implement intervention plans at the entry level and throughout the academic journey. These plans help universities improve their approaches effectively. E-learning is a fast-growing type of education that allows students to enroll in online courses. Platforms like intelligent tutoring systems, learning management systems, and massive open online courses use educational data mining to build automatic grading systems, recommenders, and adaptive systems. Even though e-learning is cheaper and more flexible, it can be a challenging learning

environment due to limited direct interaction between students and instructors.

Three main challenges of e-learning include the lack of standardized assessment measures, high dropout rates, and difficulties in understanding students' specific needs due to limited communication. Long-term log data from e-learning platforms can help with student and course evaluations.

Several machine-learning algorithms have proven effective for specific learning tasks. They are particularly useful in areas where people may lack the expertise to create efficient knowledge-engineering algorithms. Generally, machine learning examines algorithms that learn from external examples to develop general hypotheses for making predictions about future instances. Data mining plays a critical role in sifting through large amounts of data to find relevant information, helping in decision-making. Data mining has several important applications in education.

Learning analytics focuses on collecting and analyzing data from learners to improve learning materials and experiences. This need can be addressed by categorizing students based on their profiles, which can also suggest enhancements in course design and delivery. The main goal is to identify significant indicators or metrics in a learning context and to explore the relationships among these metrics using concepts from learning analytics and educational data mining. Spotting meaningful patterns in educational databases is called "educational data mining." It helps educators anticipate, improve, and assess students' academic performance. Understanding these activities allows management to improve system performance. Educational data mining has greatly influenced recent advancements in education, creating new opportunities for enhanced learning systems tailored to student needs.

This research significantly contributes to the field of educational data mining by improving the prediction of student performance using machine learning techniques. By tackling the challenges faced by contemporary learning institutions and using innovative methods, the study provides essential insights into better academic results. It examines how machine learning algorithms can be integrated into traditional teaching to enhance student performance analysis and educational outcomes. It applies K-means clustering with Davies' Bouldin method to identify clusters and significant features affecting student performance, giving a deeper understanding of the factors behind academic success. The study also compares various machine learning algorithms, including Support Vector Machine, Decision Tree, Naïve Bayes, and K-Nearest Neighbors, to assess their predictive performance regarding student outcomes. It addresses technical gaps in predicting student performance by emphasizing alternative algorithms rather than artificial neural networks. The research employs rigorous methods like repeated k-fold cross-validation and hyperparameter optimization to ensure reliable prediction results. The proposed model features an innovative clustering technique, a thorough comparative analysis, and practical applications for forecasting student performance, highlighting the relevance and impact of the findings in educational practice.

II. RELATED WORKS

Educational Data Mining (EDM) plays an important role in improving modern learning environments through effective analytical methods and application techniques. This field focuses on exploring, researching, and applying data mining (DM) approaches by integrating various methods to achieve good outcomes. EDM extracts valuable insights from raw educational data, helping to identify patterns that improve students' learning experiences and the performance of institutions. The processed data is further analyzed using different machine learning techniques to enhance usability and to develop interactive tools within learning platforms.

Machine learning (ML), a significant branch of artificial intelligence (AI), allows systems to learn from data, recognize patterns, and predict future outcomes. The rapid increase in data volumes, lower storage costs, and better computational power have led to the development of machine learning from traditional pattern-recognition techniques to more advanced Deep Learning (DL) methods.

Several studies have looked into applying machine learning techniques to predict student performance. For example, the University of Córdoba used a grammar-guided genetic programming algorithm called G3PMI to foresee student success or failure in courses, achieving an accuracy of 74.29%. Similarly, research published in the Vishwakarma Engineering Research Journal created a platform for predicting student performance using machine learning algorithms based on factors such as attendance and subject marks. Another model from Somiya College in Mumbai analyzed the links between past academic performance and future outcomes. As the dataset grew, its neural network model achieved better performance, reaching a precision of 70.48%. In another study, artificial neural networks (ANNs) were used by Talwar et al. to predict student success in exams, achieving about 85% accuracy.

Kotsiantis et al. evaluated multiple machine learning techniques for predicting student success and found that the Naïve Bayes classifier had the highest average accuracy at 73%. Likewise, Eindhoven University of Technology examined various machine learning methods for predicting student dropout rates and found that the J48 classifier yielded the best results. Researchers from three Indian universities analyzed university student datasets using different algorithms and compared their accuracy and recall values. Their findings indicated that the ADT decision tree architecture delivered the most accurate outcomes. In another study at the University of Minho in Portugal, researchers evaluated algorithms like decision trees, random forests, support vector machines, and neural networks to predict student performance in mathematics and Portuguese. This study also forecasted student success at the beginning of the academic cycle, achieving an accuracy of 85%.

A review presented in one study examined machine learning methods used to predict student performance in higher education. An analysis of 29 research studies identified six major machine learning models: decision tree, artificial neural networks (ANNs), support vector machines (SVM), k-nearest neighbor (KNN), linear regression, and Naïve Bayes (NB). Among these models, ANN showed the highest accuracy. The review also noted a growing trend in this research domain, with many machine learning algorithms being applied. These findings suggest that machine learning techniques can significantly help identify factors influencing academic performance and improve educational outcomes.

Another study focused on predicting student performance using AI techniques to help students avoid poor academic results and prepare for exams. By understanding course dependencies and requirements, teachers can offer suitable guidance and support. The proposed system enables instructors to track students' progress and provide personalized help, thereby closing learning gaps. This study achieved an accuracy rate of 94.88%, showing the effectiveness of AI-based prediction systems in assisting both students and teachers.

In a similar vein, another research study proposed a model for predicting students' academic performance using supervised machine learning algorithms, including support vector machines and logistic regression. This research found that the sequential minimal optimization algorithm surpassed logistic regression in prediction accuracy. The model identified important factors like teacher performance and student motivation, which significantly impact predicting student behavior and reducing dropout rates.

Another study investigated what affects student performance in final exams using Support Vector Machines (SVM) and Random Forest (RF) algorithms to predict final grades in mathematics and Portuguese courses. The findings showed that binary classification reached an accuracy rate of 93%, while regression analysis with the Random Forest model resulted in the lowest Root Mean Square Error (RMSE) of 1.13. Early prediction of academic performance can help educational institutions identify students who are struggling and implement interventions to improve outcomes.

Recent research highlights the challenges educational institutions face in evaluating student achievement, delivering quality instruction, and analyzing performance data. A review of EDM literature from 2009 to 2021 reveals that machine learning techniques are commonly used to predict student risk levels and dropout rates. Most studies rely on data gathered from online learning platforms and student

© MAR 2026 | IRE Journals | Volume 9 Issue 9 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I9-1715129

information systems. Machine learning methods play a key role in improving student performance and identifying potential risks. Researchers also suggest that future studies should focus on developing dynamic and ensemble-based prediction techniques and automated intervention systems for precise education.

Despite extensive research in this field, predicting student performance still comes with challenges. Many studies report limitations such as lower prediction accuracy and the inability to recognize certain hidden or influential features. In various EDM applications, algorithms like decision trees, support vector machines, k-nearest neighbors, and Naïve Bayes are popular due to their simplicity and ease of use. Although artificial neural networks offer better predictive accuracy, their practical use is often limited because they require advanced technical skills for successful implementation.

Therefore, this study aims to enhance prediction accuracy by reviewing and refining the performance of commonly used algorithms such as Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Decision Tree (DT), and Naïve Bayes. The research proposes a refined support vector machine model with performance improvements and compares these algorithms. Compared to existing methods, the proposed framework employs more reliable predictors for student performance. Hyperparameter tuning is also applied to uncover previously hidden features and boost the accuracy of the models.

III. MATERIALS AND METHODS

Machine learning (ML) has become a crucial tool for creating systems that can perform complex tasks beyond human capability. ML algorithms are commonly used in areas like pattern recognition, cluster analysis, image processing, natural language processing, and medical diagnostics. Among these techniques, clustering is important for finding hidden patterns in large datasets. Clustering is an unsupervised learning method that groups similar data points without predefined labels.In this study, we used the K-means clustering algorithm along with the Davies-Bouldin index to identify clusters and determine key factors that affect student academic performance.

3.1 Methodology

The proposed framework has four main components:

- Data preprocessing
- Hyperparameter tuning
- Prediction modeling
- Model evaluation

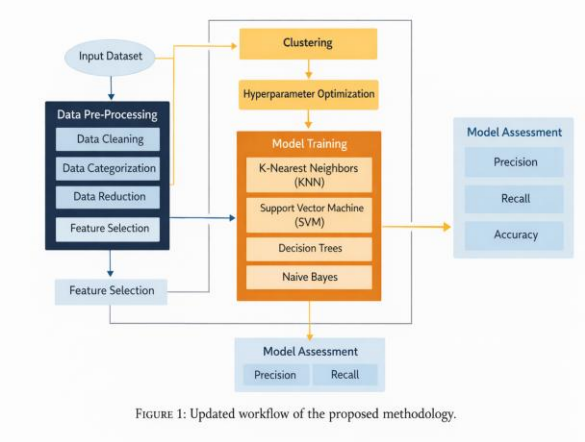

FIGURE 1: Updated workflow of the proposed methodology.

Figure 1 illustrates the overall architecture of the methodology.

We first collected the dataset from the A+ Learning Management System of Wollo University. Then we performed data preprocessing in three main stages: data cleaning, data categorization, and data reduction. These steps ensured the dataset was suitable for training machine learning algorithms.

Next, we conducted feature extraction to identify the most relevant attributes affecting student performance. To improve model performance, we applied hyperparameter tuning. Hyperparameters are settings of machine learning algorithms that remain fixed during training but greatly influence model behavior and performance.

In this research, we optimized hyperparameters to find parameter combinations that minimize prediction errors on the validation dataset. We performed clustering to group students based on characteristics like gender, region, entrance exam results, number of previous attempts, studied credits, and disability status.

© MAR 2026 | IRE Journals | Volume 9 Issue 9 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I9-1715129

Several prediction models were trained and evaluated to find the most accurate method for predicting students' final academic outcomes.

3.2 Dataset

The dataset for this study was collected from Wollo University and the Kombolcha Institute of Technology, including student records from 2017 to 2022. It initially included eight attributes:

- Student ID
- Gender
- Region
- Entrance result
- Number of previous attempts
- Studied credits
- Disability
- Final result

After removing missing and inconsistent records, the dataset had 32,005 valid student entries.
Raw data often contains noise, missing values, and duplicates, so preprocessing was necessary for data quality. We used Python for preprocessing and analysis.

3.3 Data Preprocessing

We conducted data preprocessing to enhance model accuracy and ensure consistency. The process involved three stages:

- Data Cleaning

We identified and removed missing values, noisy data, and duplicates to improve dataset reliability.

- Data Categorization

Many machine learning algorithms require numerical inputs, so we encoded categorical variables using label encoding. For instance, final results such as distinction, pass, withdrawn, and fail were converted into numeric values.

Categorical variables were classified into:

- Ordinal data, where categories have a specific order (e.g., entrance results)
- Nominal data, where categories lack a natural order (e.g., region, disability)

- Data Reduction

We removed redundant and irrelevant attributes to simplify analysis and cut down on computational complexity.

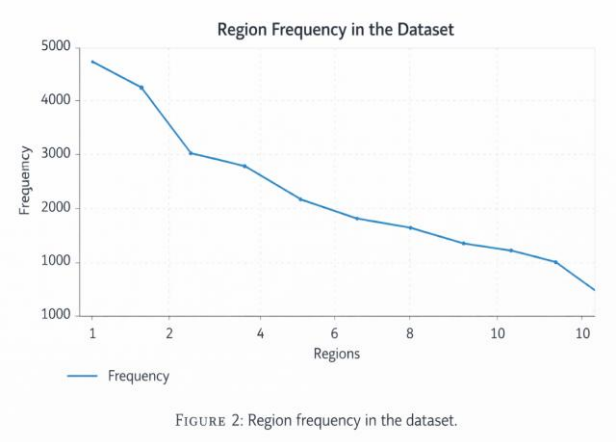

FIGURE 2: Region frequency in the dataset.

3.4 Feature Selection

The dataset included both numerical and categorical features. We used the Random Forest algorithm for feature selection to find the most informative attributes.

A 5-fold cross-validation approach was employed to assess feature importance. This method divides the dataset into five subsets, using four for training and one for validation. The process is repeated five times for a robust evaluation.

From this analysis, we identified the most influential features for predicting student performance:

- Gender
- Region
- Entrance result
- Number of previous attempts
- Studied credits
- Disability
- Final result

Selecting these features improved model efficiency and predictive accuracy.

© MAR 2026 | IRE Journals | Volume 9 Issue 9 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I9-1715129

3.5 Cluster Analysis

We used clustering to find hidden patterns in the student dataset. The K-means clustering algorithm was chosen for its effectiveness in dividing datasets into similar groups.

This algorithm splits observations into k clusters, where each observation belongs to the cluster with the closest centroid.

To determine the best number of clusters, we applied the Elbow Method using the Within-Cluster Sum of Squares (WCSS) metric. The ideal k value appears where the WCSS reduction begins to slow significantly.

According to the clustering results, we categorized student groups into three performance levels:

- Grade A – highest performance
- Grade B – moderate performance
- Grade C – lowest performance

This clustering method helped reveal patterns in student achievement across demographic and academic attributes.

To decrease sampling bias, we used 10-fold cross-validation repeated three times during model training and evaluation.

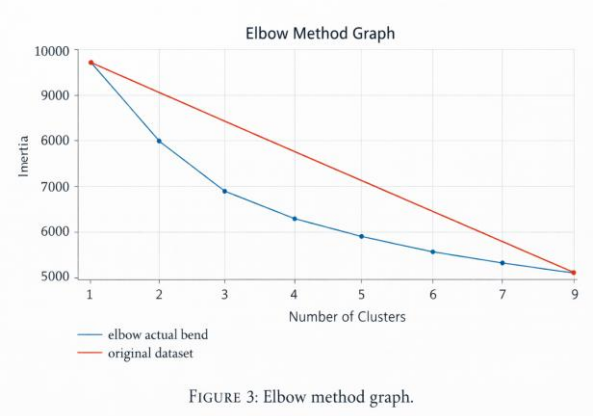

FIGURE 3: Elbow method graph.

3.6 Hyperparameter Optimization

Hyperparameter optimization is essential for enhancing machine learning model performance. We used grid search to explore different combinations of hyperparameter values.

Grid search systematically evaluates multiple parameter settings and selects the combination with the highest predictive accuracy.

The optimization process included several steps:

- Choosing suitable machine learning algorithms.
- Defining hyperparameter search spaces.
- Applying grid search to test parameter combinations.
- Performing cross-validation to avoid overfitting.
- Selecting the configuration that gave the best validation performance.

This approach ensured that each model was trained with optimal parameters.

3.7 Prediction Methods

We utilized four classification algorithms to predict student performance:

- K-Nearest Neighbors (KNN)
- Support Vector Machine (SVM)
- Decision Tree (DT)
- Naïve Bayes (NB)

We selected these algorithms for their strong performance in classification tasks.

- K-Nearest Neighbors (KNN)

KNN classifies new data based on the majority class of its k nearest neighbors in the dataset. We calculated similarity between data points using cosine similarity.

Using 10-fold cross-validation, we found the optimal value of k to be 8.

- Support Vector Machine (SVM)

SVM creates a decision boundary (hyperplane) to separate data into different classes. The algorithm

aims to maximize the gap between classes to improve classification accuracy.

Initially, we used a linear kernel for its efficiency and clarity. Later, we applied hyperparameter tuning to explore other kernel configurations and enhance model performance.

- Decision Tree

Decision Trees classify data by splitting features into branches based on information gain or entropy. They are popular due to their easy interpretation and ability to transform into decision rules.

- Naïve Bayes

Naïve Bayes is a probabilistic classifier based on Bayes' theorem, assuming independence between features. Despite this assumption, it often performs well due to its simplicity and speed.

3.8 Performance Evaluation

We evaluated model performance using a confusion matrix and the following metrics:

- Precision
- Recall
- Accuracy
- Cohen's Kappa Statistic

Accuracy measures the proportion of correct predictions. Precision and recall provide more detailed insights into classification performance.

IV. RESULTS AND DISCUSSION

We conducted experiments on a computer with an 11th Generation Intel Core i7 processor, 8GB RAM, and a 64-bit operating system.

In this study, we used the following Python tools:

- NumPy
- Pandas
- Scikit-learn
- Jupyter Notebook

These libraries aided in data preprocessing, visualization, and implementing machine learning models.

After preprocessing the dataset and training the models, we compared the algorithms.

Performance Before Hyperparameter Tuning

ALGORITHM	ACCURACY
SVM	95.4%
Decision Tree	90.9%
Naïve Bayes	77.3%
KNN	85.3%

Performance After Hyperparameter Tuning

ALGORITHM	ACCURACY
SVM	96.0%
Decision Tree	93.4%
Naïve Bayes	83.3%
KNN	87.3%

The results indicate that SVM achieved the highest accuracy, followed by Decision Tree, while Naïve Bayes had the lowest performance.

The lower performance of Naïve Bayes is primarily due to its strong independence assumption between features.

SVM performed best because it:

- Maximizes classification margins
- Effectively handles high-dimensional data
- Is robust against overfitting

Decision Trees also performed well because they model nonlinear relationships and reveal feature importance.

V. CONCLUSION

This study explored the effectiveness of machine learning techniques in predicting student academic performance. We implemented a structured methodology involving data preprocessing, clustering, hyperparameter optimization, and predictive modeling.

The results show that the Support Vector Machine achieved the highest prediction accuracy of 96%, followed by Decision Tree with 93.4% accuracy. Naïve Bayes showed the lowest accuracy due to its assumption of feature independence.

The proposed model can help educational institutions identify students at risk of poor academic performance and facilitate early intervention strategies.

Despite promising results, several limitations exist, such as relying on data from one institution and having limited feature diversity. Future research should focus on:

- Using larger and more diverse datasets
- Including additional student behavioral variables
- Enhancing model interpretability
- Conducting longitudinal analyses

Addressing these points will help create more reliable predictive systems to improve student outcomes in higher education.

VI. ACKNOWLEDGMENTS

The author expresses sincere thanks to the Wollo University ICT team for their valuable help and support during the data collection process.

REFERENCES

[1] Y. Baashar, G. Alkawsi, N. Ali, H. Alhussian, and H. T. Bahbouh, “Predicting student’s performance using machine learning methods: A systematic literature review,” in Proceedings of the 2021 International Conference on Computer and Information Sciences (ICCOINS), Kuching, Malaysia, June 2021, pp. 357–362.

[2] S. K. Yadav and S. Pal, “Data mining: A prediction for performance improvement of engineering students using classification,” 2012. Available: https://arxiv.org/abs/1203.3832

[3] M. Liu and D. Yu, “Towards intelligent e-learning systems,” Education and Information Technologies, vol. 28, no. 7, pp. 7845–7876, 2023.

[4] T. M. Mitchell, “The discipline of machine learning,” Machine Learning, vol. 9, 2006.

[5] O. Fy, A. Jet, O. Awodele, J. O. Hinmikaiye, O. Olakanmi, and J. Akinjobi, “Supervised machine learning algorithms: Classification and comparison,” International Journal of Computer Trends and Technology, vol. 48, no. 3, pp. 128–138, 2017.

[6] N. Delavari, S. Phon-Amnuaisuk, and M. R. Beikzadeh, “Data mining application in higher learning institutions,” Informatics in Education, vol. 7, no. 1, pp. 31–54, 2008.

[7] S. Nunn, J. T. Avella, T. Kanai, and M. Kebritchi, “Learning analytics methods, benefits, and challenges in higher education: A systematic literature review,” Online Learning, vol. 20, no. 2, pp. 13–29, 2016.

[8] S. Bharara, S. Sabitha, and A. Bansal, “Application of learning analytics using clustering data mining for students’ disposition analysis,” Education and Information Technologies, vol. 23, no. 2, pp. 957–984, 2018.

[9] B. K. Baradwaj and S. Pal, “Mining educational data to analyze students’ performance,” 2012. Available: https://arxiv.org/pdf/1201.3417.pdf

[10] 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, vol. 101, Article ID 102814, 2023.

[11] J. M. Helm, A. M. Swiergosz, H. S. Haeberle et al., “Machine learning and artificial intelligence: Definitions, applications, and future directions,” Current Reviews in Musculoskeletal Medicine, vol. 13, no. 1, pp. 69–76, 2020.

[12] C. K. Suryadevara, “Predictive modeling for student performance: Harnessing machine learning to forecast academic marks,”

International Journal of Applied Science and Engineering, vol. 8, no. 12, 2018.

[13] S. Talwar, M. Talwar, V. Tarjanne, and A. Dhir, "Why retail investors trade equity during the pandemic? An application of artificial neural networks to examine behavioral biases," Psychology and Marketing, vol. 38, no. 11, pp. 2142–2163, 2021.

[14] S. Kotsiantis, K. Patriarcheas, and M. Xenos, "A combinational incremental ensemble of classifiers as a technique for predicting students' performance in distance education," Knowledge-Based Systems, vol. 23, no. 6, pp. 529–535, 2010.

[15] B. A. Sani and H. Badamasi, "Machine learning algorithms to predict student's academic performance," Bakolori Journal of General Studies, vol. 12, no. 2, pp. 3656–3671, 2021.

[16] Y. A. Alsariera, Y. Baashar, G. Alkawsi, A. Mustafa, A. A. Alkahtani, and N. Ali, "Assessment and evaluation of different machine learning algorithms for predicting student performance," Computational Intelligence and Neuroscience, vol. 2022, pp. 1–11, 2022.