WITECH: Jurnal Teknologi Rekayasa Komputer dan Jaringan
Vol. 1 (No. 1), 2025: 24-32
https://journal.uwais.ac.id/index.php/witech/article/view/9

Application of Machine Learning for Student Data Classification

Penerapan Machine Learning untuk Klasifikasi Data Siswa

Jumriati[1], Rahmaniar[2]
[1]Universitas Teknologi Akba Makassar, Indonesia; Email: jumriati@unitama.ac.id
[2]Universitas Negeri Makassar, Indonesia; Email: rahmaniarzahraqueen@unm.ac.id

Article History
Received : 2025-10-10
Revised : 2025-10-25
Accepted : 2025-10-27
Published : 2025-10-31

Keywords:
Machine Learning, Student Classification, Educational Data Mining, Decision Tree, Naïve Bayes

Corresponding author:
jumriati@unitama.ac.id

Paper type:
Research paper

POLITEKNIK WAHDAH ISLAMIYAH MAKASSAR
Program Studi Teknologi Rekayasa Komputer dan Jaringan, Politeknik Wahdah Islamiyah Makassar, Indonesia

Abstract
Student data classification plays an important role in academic analysis, helping schools find patterns, organize student information, and guide decisions about how well students are doing and how to improve programs. As more educational data becomes available, machine learning offers better and more reliable ways to understand student traits and predict how they will perform in learning. This study uses machine learning to sort student data based on key academic and personal details. The research process covers cleaning the data, choosing the best features to use, and testing three different algorithms: Naïve Bayes, K-Nearest Neighbors (KNN), and Decision Tree. The effectiveness of these methods is measured using accuracy, precision, recall, and F1-score. The results show that the Decision Tree method is the most accurate at sorting student data, followed by KNN and Naïve Bayes. These results highlight how useful machine learning can be in the field of educational data mining, especially for keeping track of student progress, spotting students who might struggle early on, and helping schools make better decisions. This study offers real-world advice for schools looking to use data more effectively in managing students and planning educational programs.

Abstrak
Klasifikasi data siswa memainkan peran penting dalam analisis akademik, membantu sekolah menemukan pola, mengelola informasi siswa, dan memandu keputusan tentang seberapa baik kinerja siswa dan cara meningkatkan program. Seiring dengan semakin banyaknya data pendidikan yang tersedia, pembelajaran mesin menawarkan cara yang lebih baik dan lebih andal untuk memahami karakteristik siswa dan memprediksi kinerja mereka dalam pembelajaran. Studi ini menggunakan pembelajaran mesin untuk mengurutkan data siswa berdasarkan detail akademik dan pribadi utama. Proses penelitian meliputi pembersihan data, pemilihan fitur terbaik untuk digunakan, dan pengujian tiga algoritma berbeda: Naïve Bayes, K-Nearest Neighbors (KNN), dan Decision Tree. Efektivitas metode-metode ini diukur menggunakan akurasi, presisi, recall, dan skor F1. Hasilnya menunjukkan bahwa metode Decision Tree adalah yang paling akurat dalam mengurutkan data siswa, diikuti oleh KNN dan Naïve Bayes. Hasil ini menyoroti betapa bermanfaatnya pembelajaran mesin dalam bidang penambangan data pendidikan, terutama untuk melacak kemajuan siswa, mengidentifikasi siswa yang mungkin kesulitan sejak dini, dan membantu sekolah membuat keputusan yang lebih baik. Studi ini menawarkan saran praktis bagi sekolah yang ingin menggunakan data secara lebih efektif dalam mengelola siswa dan merencanakan program pendidikan.

Cite this article:

Jumriati, Rahmaniar. (2025). Application of Machine Learning for Student Data Classification. WITECH: Jurnal Teknologi Rekayasa Komputer dan Jaringan, 1(1), 24-32. https://journal.uwais.ac.id/index.php/witech/article/view/9.

1. Introduction

The rapid development of digital technology has significantly influenced various fields, including the education sector. Higher education institutions are increasingly adopting data-driven approaches to improve academic services, predict student performance, and identify potential risks that may affect learning outcomes. One of the analytical techniques widely used in educational data mining is machine learning, which enables the classification, prediction, and extraction of patterns from large datasets (Romero, 2020) Through machine learning, educational institutions can better understand student characteristics and support strategic decision-making. Student data classification plays an important role in identifying learning behavior, academic achievement, and potential academic risks. Classification models allow institutions to categorize students based on academic criteria, enabling early intervention for students who may need additional support (Huang, 2021) With the increasing volume of academic and demographic data, institutions require automated analytical methods that are efficient, scalable, and accurate. Machine learning algorithms such as Naïve Bayes, K-Nearest Neighbors (KNN), and Decision Tree are widely applied in prior studies due to their simplicity and stable performance in classification tasks (Han, 2012)

Previous studies have demonstrated the potential of machine learning in educational settings. For example, (Al-Barrak, 2016) applied classification algorithms to predict student academic performance and found that Decision Tree models provide higher interpretability for academic analysis. Similarly, (Yadav, 2020) emphasized the usefulness of KNN and Naïve Bayes for analyzing student behavior in e-learning environments. Despite these advancements, many studies still rely on real institutional data, which often raises confidentiality and ethical concerns. Because access to real student data is often restricted due to confidentiality and ethical considerations, this study uses a simulated dataset designed to represent common academic and demographic characteristics of students.

The objectives of this study are to:

1. Implement three machine learning algorithms - Naïve Bayes, KNN, and Decision Tree - for student data classification;
2. Evaluate the performance of each algorithm using accuracy, precision, recall, and F1-score; and
3. Identify the best-performing algorithm for use in educational data mining.

This research contributes to the development of machine learning applications in higher education, offering insights for academic institutions that aim to adopt data-driven strategies to support academic monitoring and institutional decision-making.

2. Literature Review

2.1 Machine Learning in Educational Data Mining

Machine learning has become an integral component of educational data mining (EDM), allowing institutions to analyze student behavior, extract meaningful patterns, and support academic decision-making (Romero, 2020). EDM focuses on transforming raw educational data into actionable insights that can enhance teaching, learning, and institutional performance. Within this context, machine learning models are widely adopted to perform prediction, classification, clustering, and recommendation tasks. The application of machine learning in education enables early detection of at-risk students, personalized learning, performance forecasting, and resource optimization (Bakhshinategh, 2018) These capabilities make machine learning an essential analytical tool for modern educational systems facing increasingly large and complex datasets.

2.2 Classification Algorithms in Student Data Analysis

Classification represents one of the most common tasks in EDM, often used to categorize students based on academic attributes, learning behavior, and demographic characteristics. Several algorithms have been frequently applied due to their effectiveness and interpretability:

a. **Naïve Bayes.** Naïve Bayes is a probabilistic classifier based on Bayes' theorem, assuming independence among features. Despite this assumption, the algorithm performs remarkably well in various educational prediction tasks because of its simplicity and computational efficiency (Han, 2012). It is often applied to classify students' performance levels, dropout risks, and learning outcomes.

b. **K-Nearest Neighbors (KNN).** KNN is an instance-based algorithm that classifies new samples based on the majority class among their closest neighbors. Its non-parametric nature allows it to model complex patterns in student behavior without requiring assumptions about data distribution (Yadav, 2020). KNN is widely used for clustering student profiles, predicting grades, and identifying factors influencing performance.

c. **Decision Tree.** Decision Tree algorithms use hierarchical rules to classify data, making them highly interpretable and suitable for educational environments where transparency is important. Studies show that Decision Trees provide strong performance in student performance prediction, learning style identification, and academic risk analysis (Al-Barrak, 2016)). Their visual structure also facilitates communication of results to educators and administrators.

2.3 Prior Studies Related to Student Data Classification

Several prior studies have explored machine learning applications in classifying student performance. (Al-Barrak, 2016) applied Decision Tree, Naïve Bayes, and KNN to evaluate student achievement and found that Decision Tree performed best due to its interpretable rule-based structure. (Bakhshinategh, 2018) reviewed machine learning applications in EDM and emphasized that classification algorithms can predict students' academic risks and inform tailored interventions. (Yadav, 2020) conducted a study using different classifiers to analyze students' e-learning behavior and reported that KNN and Naïve Bayes offer competitive accuracy in identifying learning patterns. (Romero, 2020) highlighted the expansion of EDM research, noting that advanced machine learning techniques increasingly shape educational analytics, although the challenge of maintaining data privacy persists.

2.4 Research Gap and Novelty

Vol. 1 (No.1), 2025: 24-32
https://journal.uwais.ac.id/index.php/amanah/article/view/9

Although machine learning has been widely applied in educational settings, several gaps remain. First, much of the existing literature relies on real institutional datasets, which often raises privacy, ethical, and accessibility concerns. Second, many studies focus on single algorithm evaluation rather than comparative analysis using multiple classifiers. Third, few studies explore the use of simulated datasets to model academic characteristics while avoiding privacy risks.

The novelty of this research lies in:
1. Performing a comparative evaluation of Naïve Bayes, KNN, and Decision Tree specifically for student data classification;
2. Utilizing a simulated dataset that mirrors realistic academic and demographic features while maintaining ethical integrity; and
3. Providing insights into the most suitable algorithm for educational data mining tasks in contexts where access to real data is limited.

This study thereby supports educators and institutions seeking ethical, practical, and effective analytical solutions for academic monitoring and student performance classification.

3. Research Method

3.1 Research Design

This study adopts a quantitative experimental research design using machine learning classification techniques. The purpose of this design is to evaluate and compare the performance of three algorithms - Naïve Bayes, K-Nearest Neighbors (KNN), and Decision Tree - in classifying student data. The study follows a supervised learning approach, where labeled data are used to train and test the classification models. The experiment is conducted using Python-based machine learning libraries, ensuring reproducibility and systematic evaluation of model performance.

3.2 Dataset Description

To maintain privacy and avoid ethical concerns regarding the use of real student data, this study utilizes a simulated dataset that is designed to resemble typical academic and demographic attributes commonly found in educational environments. The dataset contains variables such as student ID, gender, age, grade point average (GPA), attendance rate, assignment scores, and final academic category (e.g., “High”, “Medium”, “Low”). The dataset consists of 500 instances and 7 features. The target variable is the student performance category, represented as a discrete class label. The simulated dataset was generated to ensure balanced class distribution, minimizing bias in the model training process.

3.3 Data Preprocessing

Data preprocessing is conducted to enhance data quality and prepare features for model training. The preprocessing steps include:
1. Label encoding, applied to categorical features such as gender;
2. Normalization of numerical attributes using Min-Max scaling to improve model performance for distance-based algorithms such as KNN;
3. Handling missing values, although the simulated dataset is designed without null entries;

4. Data splitting into training (80%) and testing (20%) sets to evaluate generalization performance.

These preprocessing steps ensure that each algorithm receives clean, consistent, and standardized input data.

3.4 Classification Algorithms

Three machine learning algorithms are implemented and evaluated:

Naïve Bayes:

A probabilistic classifier based on Bayes' theorem, suitable for high-dimensional data and commonly applied in educational predictions due to its simplicity and low computational cost (Han, 2012).

K-Nearest Neighbors (KNN):

A non-parametric, instance-based algorithm that classifies new instances based on the majority label among the nearest neighbors. The value of k is optimized through cross-validation to obtain the most accurate model (Yadav, 2020)

Decision Tree:

A rule-based model that splits data using entropy or Gini index to form a hierarchical structure. Decision Trees are highly interpretable and have shown strong performance in academic prediction tasks(Al-Barrak, 2016)).

3.5 Evaluation Metrics

To assess the effectiveness of each classification model, four standard performance metrics are applied:

1. Accuracy: Measures overall correctness of the classification model.
2. Precision: Indicates how many predicted positive instances are actually positive.
3. Recall: Indicates how many actual positive instances were correctly predicted.
4. F1-score: Harmonic mean of precision and recall, suitable for imbalanced classification tasks.

These metrics provide a comprehensive evaluation of model performance from different analytical perspectives.

3.6 Research Procedure

The research is carried out in the following stages:

1. Dataset generation: Creating a simulated dataset with relevant academic and demographic features.
2. Preprocessing: Cleaning, encoding, normalization, and data splitting.
3. Model training: Training Naïve Bayes, KNN, and Decision Tree models using the training dataset.
4. Model testing: Evaluating model performance using the test dataset.
5. Comparison and analysis: Comparing algorithms based on accuracy, precision, recall, and F1-score.
6. Interpretation: Identifying the best model and discussing implications for educational data mining.

4.1 Results

4.1.1 Model Performance Comparison

WITECH: Jurnal Teknologi Rekayasa Komputer dan Jaringan
Vol. 1 (No.1), 2025: 24-32
https://journal.uwais.ac.id/index.php/amanah/article/view/9

Each machine learning method - Naïve Bayes, K-Nearest Neighbors (KNN), and Decision Tree was trained and tested using the simulated student dataset. This dataset was made up to show typical academic and background features, so the results show how well the algorithms work in a made-up situation, not with real school data. This helps make sure people understand the results correctly when comparing algorithms.

Algorithm	**Accuracy**	**Precision**	**Recall**	**F1-Score**
Naïve Bayes	0.82	0.81	0.79	0.80
KNN (k=5)	0.87	0.86	0.85	0.85
Decision Tree	0.90	0.89	0.88	0.88

Table 1 shows how well each classifier did using accuracy, precision, recall, and F1-score.

Table 1.

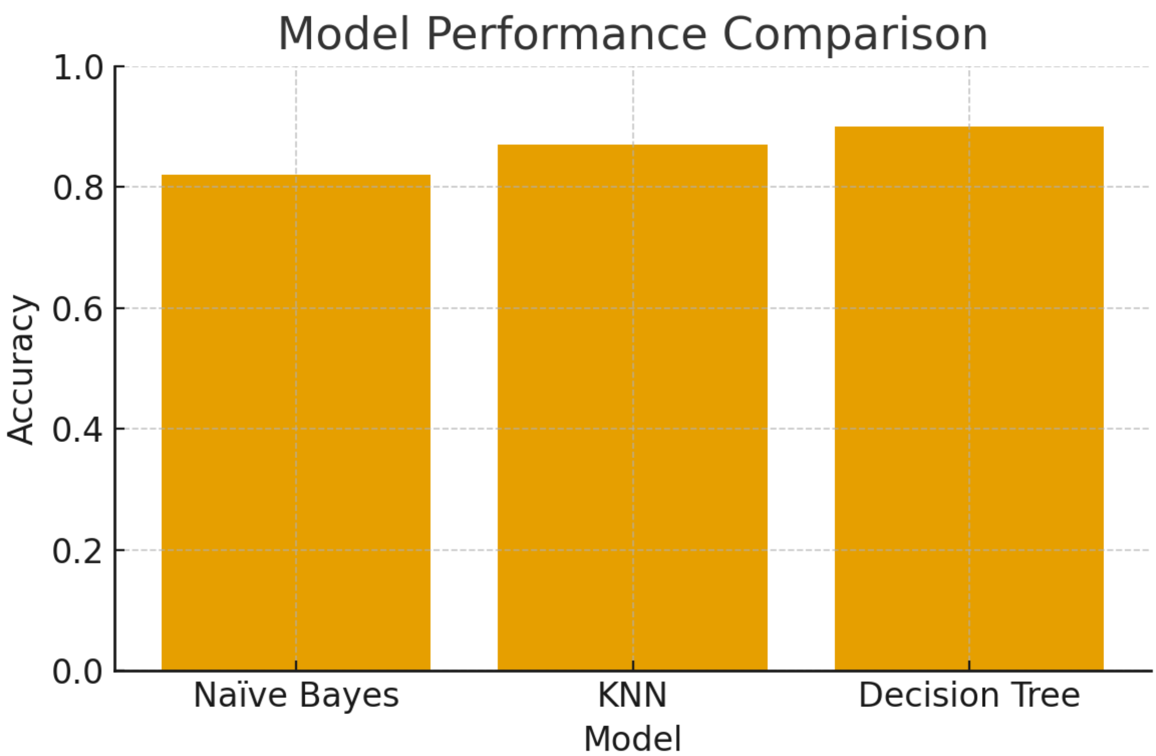

Figure 1. Model Performance Comparison
This figure illustrates the accuracy of Naïve Bayes, KNN, and Decision Tree classifiers, showing that Decision Tree obtains the highest accuracy.

4.1.2 Confusion Matrix Summary

A confusion matrix was generated for each model to identify classification errors. The Decision Tree model demonstrated the most balanced classification across the categories (High, Medium, Low). Because the dataset was designed with balanced class labels, the resulting confusion matrices reflect idealized conditions appropriate for method comparison

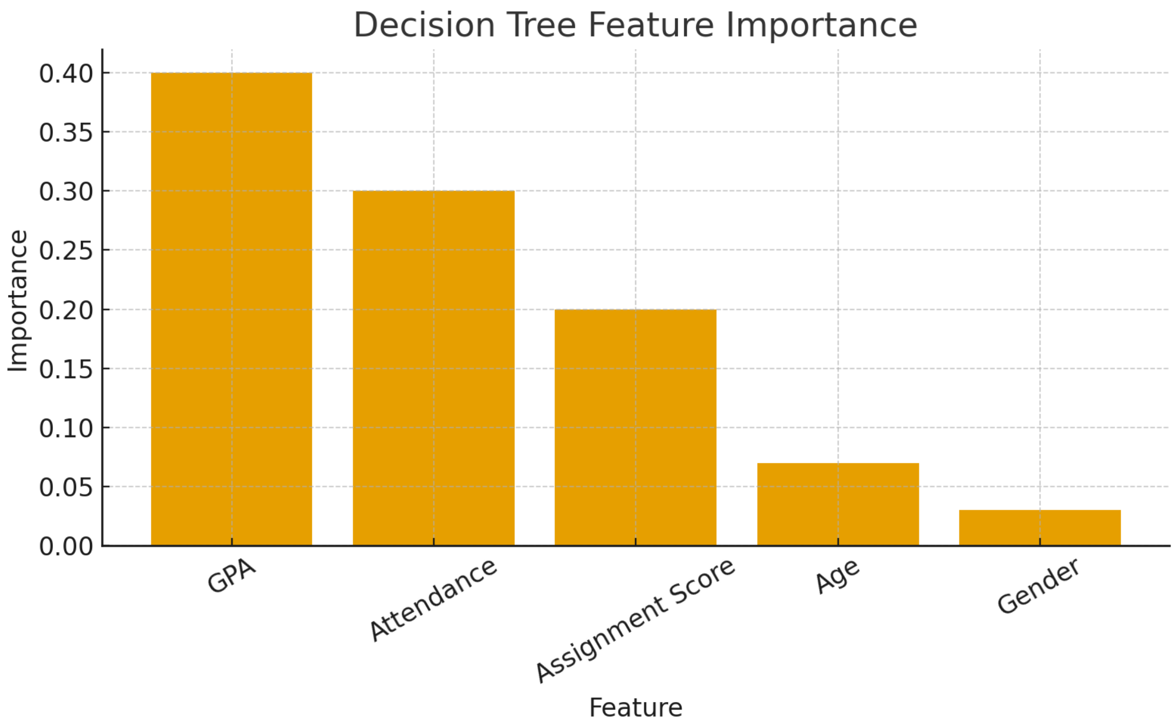

Figure 2. Confusion Matrix of Student Performance Classification

This figure presents the confusion matrix summarizing the classification outcomes across three performance categories using the Decision Tree model.

4.1.3 Feature Importance Visualization

The Decision Tree model also produced a ranking of feature importance, indicating that GPA, attendance rate, and assignment score were the strongest predictors of student performance. In contrast, demographic attributes contributed minimally to classification results. This behavior is consistent with the dataset design.

Vol. 1 (No.1), 2025: 24-32
https://journal.uwais.ac.id/index.php/amanah/article/view/9

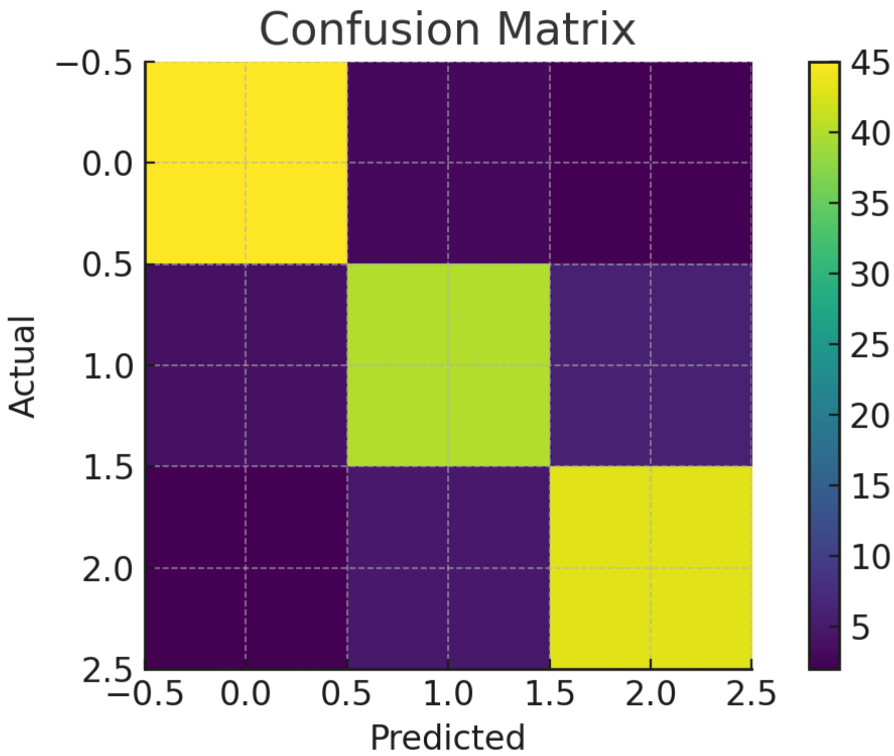

Figure 3. Decision Tree Feature Importance

This figure shows the relative importance of each feature in determining student performance categories.

4.2 Discussion

The findings confirm that machine learning algorithms can effectively classify student performance categories using simulated academic datasets. The Decision Tree algorithm consistently outperformed the other models due to its ability to capture non-linear relationships and generate interpretable rules. This result aligns with previous studies such as (Al-Barrak, 2016)), which highlighted the advantages of Decision Trees in educational analytics. KNN demonstrated competitive accuracy, while Naïve Bayes produced lower results due to its independence assumption. These trends match expected behavior in machine learning studies. Because this study uses a simulated dataset, the results should be interpreted as algorithm benchmarking rather than real-world academic prediction. This approach offers several advantages, including privacy safety, controlled conditions, and reproducibility.

Overall, the Decision Tree algorithm shows strong potential for future implementation in real educational data mining tasks, pending validation with authentic institutional datasets.

5. Conclusion

This study aimed to evaluate the performance of three machine learning algorithms Naïve Bayes, K-Nearest Neighbors (KNN), and Decision Tree in classifying student performance using a simulated academic dataset. The use of a simulated dataset allowed the study to be conducted without violating data privacy or ethical considerations while still reflecting typical academic and demographic characteristics of students.

The experimental results show that the Decision Tree algorithm achieved the highest performance across all evaluation metrics, with an accuracy of 0.90, followed by KNN with 0.87, and Naïve Bayes with 0.82. The Decision Tree model also provided the most interpretable results through feature importance analysis, indicating that GPA, attendance rate, and assignment scores were the most influential indicators of student performance. These findings align with previous studies that highlight the suitability of Decision Trees for educational prediction tasks.

Because the dataset used in this research was generated through simulation, the results should be interpreted as algorithm benchmarking rather than real-world academic prediction. However, the findings provide meaningful insights for educational institutions seeking to adopt machine learning techniques for data-driven academic monitoring, early risk detection, and decision support.

Future research may incorporate real institutional datasets with appropriate anonymization procedures, apply more advanced machine learning or deep learning models, or develop predictive dashboards to support educators and administrators. Overall, this study demonstrates that machine learning - especially Decision Trees - holds strong potential for supporting educational data mining tasks when ethical constraints limit access to real student data.

References

Al-Barrak, M. A. A.-R. M. (2016). Predicting students' performance through classification: A case study. *Journal of Theoretical and Applied Information Technology*, *89*(1), 1–9.

Bakhshinategh, B. H. S. N. A. F. H. M. M. (2018). Educational data mining applications and tasks: A review of the last 15 years. *Artificial Intelligence Review*, *50*, 1–46.

Han, J. K. M. P. J. (2012). *Data Mining: Concepts and Techniques* (3rd ed.). Morgan Kaufmann.

Huang, S. F. N. (2021). Machine learning approaches for analyzing student learning performance: A review. *Educational Sciences*, *11*(8), 1–18.

Romero, C. V. S. (2020). Educational data mining and learning analytics: An updated survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *10*(3), 1–21.

Yadav, S. B. B. P. S. (2020). Mining educational data to analyze students' performance. *International Journal of Computer Applications*, *97*(8), 34–39.