Perbanas International Conference on Economics, Business, Management, Accounting and IT (PROFICIENT) 2026
"Integrity-Driven Innovation: Reimagining Business, Finance, and Digital Transformation for Sustainable Impact"

Proceedings of the Perbanas International Seminar on Economics, Business, Management, Accounting and IT
[June 2026] [ISSN 3047-0951]

Random Forest-Based Classification for Early Detection of Student Dropout

Ines Kartika Dewi[1,a)], Chelsea Aulia Zahra[2,b)], Elliana Gautama[3,c)]

[1,2,3] *Data Science Study Program, Faculty of Information Technology, Perbanas Institute*

[a)]Corresponding author: ines.kartika05@perbanas.id
[b)] chelsea.aulia@perbanas.id
[c)] elliana@perbanas.id

Abstract. *Student dropout is one of the major challenges in higher education because it affects academic performance, graduation rates, and institutional reputation. Early identification of students who are at risk of dropping out is important so that universities can provide appropriate intervention strategies and academic support systems. Along with the growth of educational data, machine learning techniques have become increasingly useful for analyzing student behavior patterns and predicting academic outcomes more effectively. This study aims to classify student dropout using the Random Forest algorithm on the Predict Students' Dropout and Academic Success dataset obtained from the UCI Machine Learning Repository. The research process includes data preprocessing, feature selection, model training, and performance evaluation. Random Forest is used to classify students into dropout, enrolled, or graduate categories based on academic, demographic, and socioeconomic attributes. Model performance is evaluated using accuracy, precision, recall, F1-score, and confusion matrix analysis. The results of this study are expected to support universities in identifying students at risk of dropout and improving student retention strategies through data-driven decision-making. In addition, this research contributes to the application of machine learning techniques in educational data mining for academic prediction systems.*

Keywords: Student Dropout, Random Forest, Classification, Machine Learning, Educational Data Mining

INTRODUCTION

Student dropout has become one of the major challenges faced by higher education institutions because it affects academic achievement, graduation rates, and institutional reputation. High dropout rates can reduce the quality of educational performance and negatively impact student success. Therefore, early identification of students who are at risk of dropping out is important so that universities can provide appropriate intervention strategies and academic support systems. Along with the growth of educational data, machine learning techniques have become increasingly useful for analyzing student behavior patterns and predicting academic outcomes more effectively. Educational Data Mining (EDM) applies data mining and machine learning methods to discover meaningful patterns in educational datasets and support data-driven decision-making processes in education (Romero & Ventura, 2010).

Among various machine learning techniques, classification methods are widely used in student dropout prediction. Random Forest is one of the most effective supervised learning algorithms because it provides high predictive accuracy, robustness against overfitting, and the ability to handle datasets with multiple variables efficiently (Breiman, 2001). This study aims to classify student dropout using the Random Forest algorithm on the Predict Students' Dropout and Academic Success dataset obtained from the UCI Machine

Proceedings of the Perbanas International Seminar on Economics, Business, Management, Accounting and IT
[June 2026] [ISSN 3047-0951]

Learning Repository. The use of Random Forest is expected to improve prediction performance and provide valuable insights for universities in developing effective student retention strategies and academic intervention programs.

LITERATURE REVIEW

Educational Data Mining

Educational Data Mining (EDM) is a research field that applies data mining and machine learning techniques to educational datasets in order to analyze student behavior, academic performance, and learning patterns. EDM helps educational institutions discover useful information from student data and supports decision-making processes to improve educational quality (Romero & Ventura, 2010). With the increasing amount of educational data, machine learning methods have become important tools for predicting academic outcomes and identifying students who are at risk of dropout.

Student Dropout Prediction

Student dropout prediction refers to the process of identifying students who are likely to discontinue their studies before graduation. High dropout rates can negatively affect graduation performance, institutional reputation, and student success. Previous studies have shown that factors such as academic achievement, socioeconomic background, and student engagement significantly influence dropout rates (Baker & Yacef, 2009). Therefore, predictive analysis using machine learning techniques has become increasingly important in higher education institutions.

Random Forest Classification

Random Forest is a supervised machine learning algorithm that combines multiple decision trees to improve classification accuracy and reduce overfitting problems (Breiman, 2001). The algorithm generates several decision trees using random subsets of training data and combines their predictions through majority voting. Random Forest is widely applied because it provides strong predictive performance and can handle large datasets with many variables. In student dropout prediction, Random Forest can classify students into categories such as dropout, enrolled, or graduate based on academic and demographic attributes.

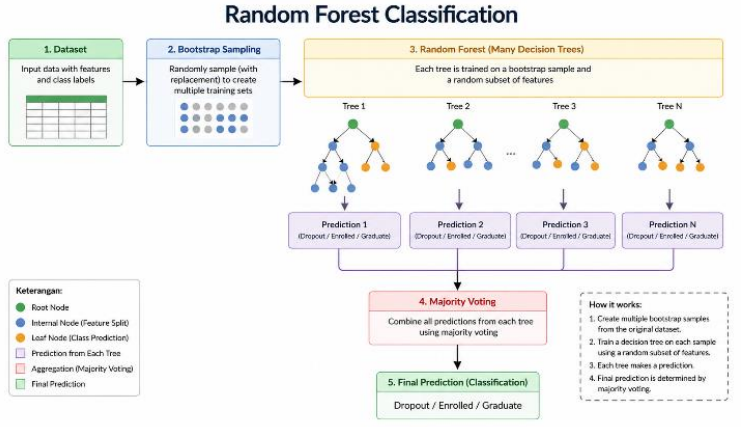


FIGURE 1. Random Forest Classification Model

The figure illustrates the workflow of the Random Forest classification algorithm for predicting student academic outcomes. The process begins with the input dataset containing student features and class labels such as dropout, enrolled, and graduate. The dataset is then divided into multiple bootstrap samples, where

Proceedings of the Perbanas International Seminar on Economics, Business, Management, Accounting and IT
[June 2026] [ISSN 3047-0951]

each sample is used to train different decision trees. During training, each tree uses a random subset of features to improve model diversity and reduce overfitting problems (Breiman, 2001). After all decision trees generate predictions, the results are combined through a majority voting process to determine the final classification output. The final prediction represents the student category predicted by the model. This approach improves classification accuracy and robustness because the prediction is based on the aggregation of multiple decision trees rather than a single classifier, making Random Forest effective for educational prediction tasks and student dropout analysis (Delen, 2010).

Previous Research

Several previous studies have demonstrated that machine learning techniques are effective for predicting student dropout. Random Forest has been recognized as one of the most effective classification algorithms for educational prediction tasks because of its robustness and high accuracy (Delen, 2010). Machine learning methods have helped universities identify students at risk of dropout and improve academic intervention systems. Therefore, the implementation of Random Forest is considered suitable for predicting student academic outcomes using educational datasets.

Research Gap

Previous studies on student dropout prediction have widely applied machine learning algorithms such as Decision Tree, Logistic Regression, Support Vector Machine, and Random Forest to classify student academic outcomes. Among these methods, Random Forest is recognized as one of the most effective classification algorithms because it provides high predictive accuracy, robustness against overfitting, and the ability to handle complex educational datasets with multiple variables (Breiman, 2001; Delen, 2010). However, many previous studies focused only on general academic prediction and did not specifically analyze the Predict Students' Dropout and Academic Success dataset from the UCI Machine Learning Repository using Random Forest classification. In addition, several studies have not fully explored the influence of academic, demographic, and socioeconomic variables on student dropout prediction. Therefore, this study aims to apply the Random Forest algorithm to classify students into dropout, enrolled, or graduate categories and evaluate the effectiveness of the model in predicting student academic outcomes.

Conceptual Framework

This study assumes that student academic, demographic, and socioeconomic data contain important patterns related to student dropout behavior. The research process begins with collecting the *Predict Students' Dropout and Academic Success* dataset from the UCI Machine Learning Repository. The dataset then undergoes preprocessing, feature selection, and Random Forest model training to classify students into dropout, enrolled, or graduate categories. Finally, the model performance is evaluated using accuracy, precision, recall, F1-score, and confusion matrix analysis to support universities in identifying students at risk of dropout through data-driven decision-making.

Proceedings of the Perbanas International Seminar on Economics, Business, Management, Accounting and IT
[June 2026] [ISSN 3047-0951]

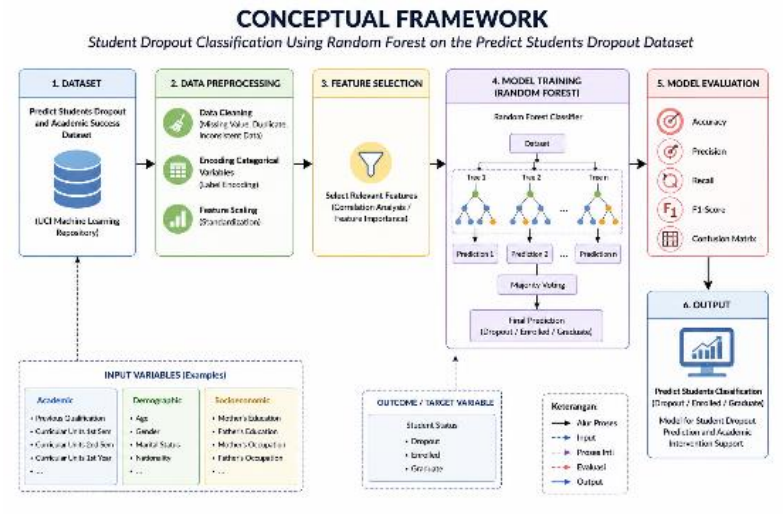


FIGURE 2. Conceptual Framework

The figure illustrates the conceptual framework of student dropout classification using the Random Forest algorithm. The process begins with collecting the Predict Students' Dropout and Academic Success dataset from the UCI Machine Learning Repository, followed by data preprocessing and feature selection to improve data quality and identify important variables. The processed data are then used to train the Random Forest model to classify students into dropout, enrolled, or graduate categories. Finally, the model performance is evaluated using accuracy, precision, recall, F1-score, and confusion matrix analysis to measure the effectiveness of the prediction model (Breiman, 2001; Delen, 2010).

METHODS

Research Design

This study uses a quantitative research approach with a machine learning technique to classify student academic outcomes. The research applies the Random Forest classification algorithm to predict student status categories, namely dropout, enrolled, or graduate. The purpose of this study is to analyze student academic, demographic, and socioeconomic data in order to identify patterns related to student dropout.

Data Source

The dataset used in this study is the Predict Students' Dropout and Academic Success dataset obtained from the UCI Machine Learning Repository. The dataset contains student demographic, academic, and socioeconomic information, including previous qualification grades, admission grades, age at enrollment, scholarship status, tuition fee status, curricular unit performance, unemployment rate, inflation rate, and GDP. The target variable consists of three categories: dropout, enrolled, and graduate.

Data Preprocessing

Data preprocessing is conducted to improve the quality and consistency of the dataset before applying the machine learning model. The preprocessing stage includes data cleaning, encoding categorical variables into numerical form, and feature scaling using standardization techniques. These processes help improve the performance and accuracy of the Random Forest classification model.

Feature Selection

Feature selection is performed to identify the most relevant variables affecting student academic outcomes. Variables with low contribution or redundant information are excluded to improve model

Proceedings of the Perbanas International Seminar on Economics, Business, Management, Accounting and IT
[June 2026] [ISSN 3047-0951]

efficiency and reduce overfitting problems. Important features include academic performance, curricular unit grades, tuition fee status, scholarship status, and demographic information.

Model Development

The processed dataset is divided into training data and testing data using an 80:20 ratio. The Random Forest classification algorithm is then trained using the training dataset to classify students into dropout, enrolled, or graduate categories. Random Forest is selected because it provides strong predictive performance, high accuracy, and robustness against overfitting problems (Breiman, 2001).

Model Evaluation

The performance of the Random Forest model is evaluated using several classification metrics, including accuracy, precision, recall, and F1-score. In addition, confusion matrix analysis is used to compare the actual class labels and predicted labels generated by the model. These evaluation metrics are used to measure the effectiveness of the classification model in predicting student academic outcomes.

RESULTS AND DISCUSSION

System Implementation Overview

The student dropout classification system was implemented using the Python programming language within the Anaconda and Jupyter Notebook environments. Several supporting libraries were utilized throughout the research process, including Pandas for data manipulation, Scikit-learn for machine learning modeling, and Matplotlib and Seaborn for visualization and evaluation purposes.

TABLE 1. Libraries Used in the Research

Library	Function
Pandas	Data manipulation and preprocessing
Matplotlib	Data visualization
Seaborn	Visualization of confusion matrix
Scikit-learn	Machine learning modeling and evaluation

Before conducting the modeling process, all required libraries were successfully imported to support data processing, model training, and evaluation activities. The implementation environment enabled efficient analysis and reproducible experimentation.

Dataset Overview

The dataset used in this study is the *Predict Students' Dropout and Academic Success* dataset obtained from the UCI Machine Learning Repository. The dataset contains demographic, academic, and socioeconomic information related to student academic performance and dropout prediction. Several important attributes include admission grade, age at enrollment, scholarship holder status, tuition fee status, curricular unit performance, unemployment rate, inflation rate, and GDP.

The target variable consists of three categories:

- Dropout

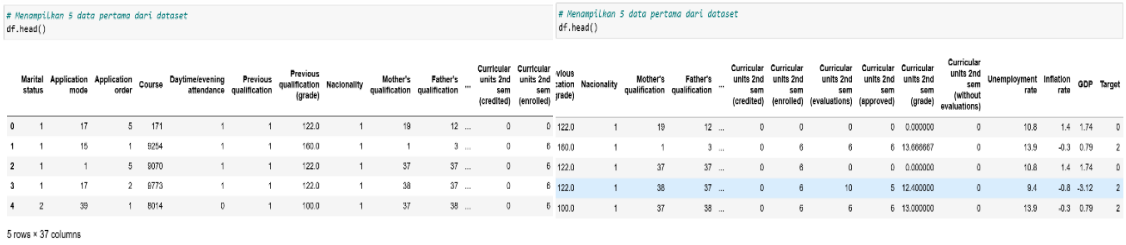
- Enrolled
- Graduate

TABLE 2. Dataset Overview

Marital Status	Admission Grade	Age at Enrollment	Scholarship Holder	Tuition Fees Up to Date	Target
1	127.3	18	1	1	Graduate
1	142.5	19	0	1	Dropout
2	124.8	20	1	0	Enrolled
1	136.0	18	1	1	Graduate
1	118.4	21	0	0	Dropout

The dataset was explored using several data analysis functions such as head(), columns, and isnull().sum() to understand the dataset structure and identify missing values before preprocessing.

The following figure shows the first 5 rows of the dataset used in this research. The data is displayed using the df.head() function in Jupyter Notebook to observe the initial structure of the dataset and its available attributes.



```
# Menampilkan 5 data pertama dari dataset
df.head()
```

	Marital status	Application mode	Application order	Course	Daytime/evening attendance	Previous qualification	Previous qualification (grade)	Nacionality	Mother's qualification	Father's qualification	...	Curricular units 2nd sem (credited)	Curricular units 2nd sem (enrolled)
0	1	17	5	171	1	1	122.0	1	19	12	...	0	0
1	1	15	1	9254	1	1	160.0	1	1	3	...	0	6
2	1	1	5	9070	1	1	122.0	1	37	37	...	0	6
3	1	17	2	9773	1	1	122.0	1	38	37	...	0	6
4	2	39	1	8014	0	1	100.0	1	37	38	...	0	6

5 rows × 37 columns

```
# Menampilkan 5 data pertama dari dataset
df.head()
```

vious ation rade)	Nacionality	Mother's qualification	Father's qualification	...	Curricular units 2nd sem (credited)	Curricular units 2nd sem (enrolled)	Curricular units 2nd sem (evaluations)	Curricular units 2nd sem (approved)	Curricular units 2nd sem (grade)	Curricular units 2nd sem (without evaluations)	Unemployment rate	Inflation rate	GDP	Target
122.0	1	19	12	...	0	0	0	0	0.000000	0	10.8	1.4	1.74	0
160.0	1	1	3	...	0	6	6	6	13.666667	0	13.9	-0.3	0.79	2
122.0	1	37	37	...	0	6	0	0	0.000000	0	10.8	1.4	1.74	0
122.0	1	38	37	...	0	6	10	5	12.400000	0	9.4	-0.8	-3.12	2
100.0	1	37	38	...	0	6	6	6	13.000000	0	13.9	-0.3	0.79	2

FIGURE 3. Dataset Overview

Based on the displayed output, the dataset contains various variables related to student information, such as marital status, application mode, previous qualification grade, and second semester academic performance. In addition, the dataset also includes a target variable that is used as the label for the data analysis and modeling process.

Data Preprocessing

Data preprocessing was conducted to improve dataset quality before the machine learning modeling stage. The preprocessing activities included removing unnecessary spaces in column names, encoding categorical target variables using LabelEncoder(), and standardizing numerical features using StandardScaler().

TABLE 3. Data Preprocessing Activities

Activity	**Method**
Remove unnecessary spaces	str.strip()
Categorical Encoding	LabelEncoder()
Feature Scaling	StandardScaler()

Proceedings of the Perbanas International Seminar on Economics, Business, Management, Accounting and IT
[June 2026] [ISSN 3047-0951]

The preprocessing stage successfully transformed the raw student dataset into a structured format suitable for machine learning analysis. The target variable was converted into numerical representations to support the Random Forest classification model. In addition, feature scaling improved dataset consistency and reduced the possibility of errors during model training.

Feature Selection and Data Splitting

Feature selection was conducted by separating the target variable from the predictor variables. The target variable used in this study was "Target", while all remaining variables were used as input features for the model. After feature selection, the dataset was divided into training and testing sets using the train_test_split() function from Scikit-learn with an 80:20 ratio for model training and evaluation.

TABLE 4. Data Splitting Configuration

Configuration	Value
Training Data	80%
Testing Data	20%
Random State	42

The splitting process ensured that the model could be trained and evaluated fairly using separate datasets. The use of random_state=42 also guaranteed experiment reproducibility.

Random Forest Model Training

The Random Forest algorithm was implemented as the main classification model for predicting student academic outcomes using the training dataset after the preprocessing stage was completed. The Random Forest classifier was selected because of its ability to handle complex datasets, reduce overfitting risks, and generate stable prediction performance. During the training stage, the algorithm learned student academic patterns and generated classification rules used to predict student status categories.

Feature Importance

Feature importance analysis was conducted to identify which variables contributed the most to student dropout prediction. The Random Forest algorithm automatically measures the influence of each feature during the tree-building process.

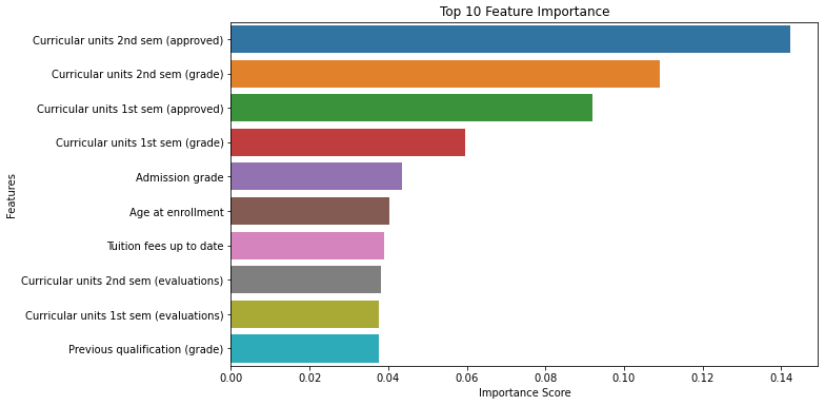


FIGURE 3. Top 10 Feature Importance

Proceedings of the Perbanas International Seminar on Economics, Business, Management, Accounting and IT
[June 2026] [ISSN 3047-0951]

The figure illustrates the importance scores generated by the Random Forest model. Several variables such as curricular unit grades, admission grades, and tuition fee status showed strong influence on student academic outcome prediction. These findings indicate that academic performance significantly contributes to student dropout risks.

Model Evaluation Results

Model evaluation was conducted to measure the effectiveness of the Random Forest algorithm in predicting student academic outcomes. Several evaluation metrics were used in this study, including accuracy, precision, recall, and F1-score. These metrics help determine the classification performance of the model in identifying dropout, enrolled, and graduate categories based on the testing dataset.

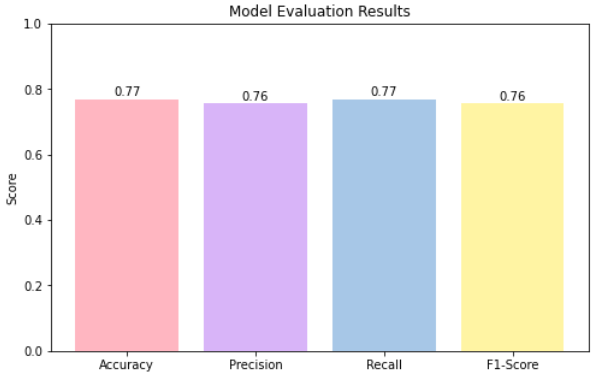


FIGURE 4. Model Evaluation Results

The figure shows that the Random Forest model achieved good classification performance with an accuracy value of 0.77, precision of 0.76, recall of 0.77, and F1-score of 0.76. These results indicate that the model was effective in classifying student academic outcomes on the Predict Studentson the Predict Students Dropout dataset#x2019; Dropout and Academic Success dataset.
Overall, the Random Forest algorithm demonstrated stable and reliable performance for student dropout prediction.

Confusion Matrix

To further analyze the prediction performance, a confusion matrix was generated using the testing dataset. The confusion matrix compares the actual student categories with the predicted classifications generated by the Random Forest model.

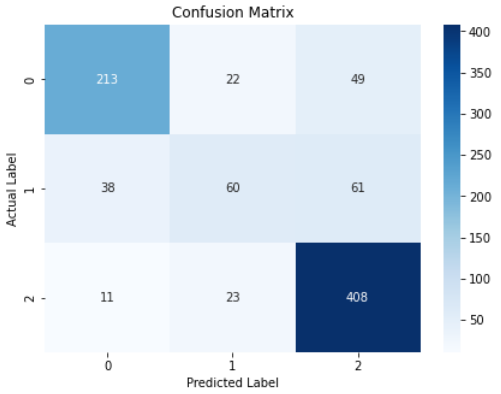


FIGURE 5. Confusion Matrix

Proceedings of the Perbanas International Seminar on Economics, Business, Management, Accounting and IT
[June 2026] [ISSN 3047-0951]

The figure illustrates the confusion matrix generated from the testing dataset. The model successfully classified the majority of student categories correctly while still producing several misclassifications between dropout, enrolled, and graduate classes. Overall, the Random Forest model demonstrated good capability in predicting student academic outcomes and minimizing classification errors.

Discussion

The results of this study indicate that the Random Forest algorithm achieved good performance in classifying student academic outcomes on the Predict Studentson the Predict Students Dropout dataset#x2019; Dropout and Academic Success dataset. Based on the evaluation results, the model obtained an accuracy value of 0.77, precision of 0.76, recall of 0.77, and F1-score of 0.76. These values demonstrate that the model was capable of predicting student categories effectively, including dropout, enrolled, and graduate classes. The balanced performance across evaluation metrics also indicates that the model was relatively stable in handling the classification process.

The feature importance analysis showed that several academic variables, such as curricular unit grades and admission grades, contributed significantly to the prediction results. This finding suggests that student academic performance strongly influences dropout risk and academic success. In addition, the confusion matrix results demonstrated that the majority of student categories were classified correctly, although several misclassifications still occurred between similar classes. Overall, the Random Forest algorithm proved to be suitable for student dropout prediction because of its strong predictive capability and robustness in handling educational datasets.

CONCLUSIONS

This study successfully implemented the Random Forest algorithm for student dropout classification using the Predict Students' Dropout and Academic Success dataset. Based on the evaluation results, the model achieved good classification performance with an accuracy value of 0.77, precision of 0.76, recall of 0.77, and F1-score of 0.76. These results indicate that the Random Forest algorithm was effective in predicting student academic outcomes, including dropout, enrolled, and graduate categories.

In addition, feature importance analysis showed that academic-related variables significantly influenced the prediction results. Overall, the Random Forest model demonstrated stable and reliable performance for educational prediction tasks and can support universities in identifying students at risk of dropout through data-driven decision-making.

Acknowledgments

The authors would like to thank the lecturers teaching the Data Mining and Machine Learning courses in the Data Science Study Program at the Faculty of Information Technology, Perbanas Institute Jakarta, for their guidance and support throughout this research. The authors also express their gratitude for the opportunity to present the results of this Final Project research at the Perbanas - International Conference (PROFICIENT) 2026.

References

Baker, R. S. J. D., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. *Journal of Educational Data Mining*, *1*(1), 3–17.

Berens, J., Schneider, K., Görtz, S., Oster, S., & Burghoff, J. (2018). Early detection of students at risk. *Predicting Student Dropouts Using Administrative Student Data and Machine Learning Methods, CESIFO WP*, *7259*.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Delen, D. (2010). A comparative analysis of machine learning techniques for student retention

Proceedings of the Perbanas International Seminar on Economics, Business, Management, Accounting and IT
[June 2026] [ISSN 3047-0951]

management. *Decision Support Systems*, *49*(4), 498–506.

Helal, S., Li, J., Liu, L., Ebrahimie, E., Dawson, S., Murray, D. J., & Long, Q. (2018). Predicting academic performance by considering student heterogeneity. *Knowledge-Based Systems*, *161*, 134–146.

Realinho, V., Martins, M. V., Machado, J., & Baptista, L. (2021). Predict students' dropout and academic success. *UCI Machine Learning Repository*, *10*, C5MC89.

Romero, C., & Ventura, S. (2010). Educational data mining: a review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *40*(6), 601–618.