International Journal of Engineering, Science and Information Technology
Volume 5, No. 4 (2025) pp. 124-131
eISSN: 2775-2674
Website: http://ijesty.org/index.php/ijesty
DOI: https://doi.org/10.52088/ijesty.v5i4.1183
Research Paper, Short Communication, Review, Technical Paper

Smart Campus Dropout Prediction: Hybrid Features and Ensemble Approach

M. Safii[1]*, Adli Abdillah Nababan[2], Husain[3]

[1]*Department of Informatics, STIKOM Tunas Bangsa, Indonesia*
[2]*Department of Information Systems, School of Information Systems, Bina Nusantara University, Indonesia*
[3]*Department of Information Systems, Universitas Bumigora, Lombok, Indonesia*

**Corresponding author Email:m.safii@amiktunasbangsa.ac.id*

The manuscript was received on 24 February 2025, revised on 10 May 2025, and accepted on 23 August 2025, date of publication 4 November 2025

Abstract

The issue of the high number of students dropping out of college is a major concern in higher education, especially in the smart campus ecosystem. This research aims to design a prediction system for students who are at risk of dropping out by integrating hybrid feature selection methods and ensemble learning that leverage academic data and students' digital footprints. The initial process of model development involves data cleaning and the selection of important features through a combination approach using filter-based methods (mutual information) and recursive feature elimination. A classification model is then designed using the XGBoost and Random Forest algorithms. The testing was conducted using a secondary dataset that included variables such as participation in discussions, attendance rates, interaction with learning materials, and academic achievement. The results of testing with the XGBoost model showed a satisfactory accuracy level, with an F1 score of 0.77 and a ROC AUC of 0.89. The confusion matrix recorded 67 correct predictions for students who graduated and 17 correct predictions for students who dropped out, with a total of 12 misclassifications. These findings suggest that the combination of hybrid feature selection strategies and XGBoost can produce sufficiently accurate predictions of student dropouts and has the potential to be utilized as an early warning system in the governance of a more flexible and responsive smart campus.

Keywords: *Classification, Dropout, Hybrid Feature Selection, Ensemble Learning, XGBoost.*

1. Introduction

Student dropout or discontinuation of studies is one of the main challenges in the management of higher education that has a significant impact on the efficiency, effectiveness, and reputation of higher education institutions. This phenomenon not only causes personal losses for students but also leads to inefficiencies in the utilisation of educational resources. In Indonesia, based on the report from the Higher Education Database (PDDIKTI), the percentage of students who do not complete their studies on time is still quite high, even in some study programs at both public and private universities; the dropout rate exceeds 30% over the last three years [1]. Along with the development of the smart campus concept, educational institutions are required to implement technology-based approaches in the learning process, data management, and strategic decision-making. Smart campuses emphasise the importance of utilising big data and artificial intelligence to enhance the quality of educational services, including aspects of monitoring and intervention for potential student study failure [2]. One relevant approach in this context is the application of educational data mining (EDM) to build early prediction models for students at risk of dropping out based on academic data and digital activities.

Various previous studies have applied machine learning methods for classifying student dropout risk, such as Decision Tree, Random Forest, Support Vector Machine, and Artificial Neural Network [2][3][4][5]. However, the accuracy of the models is greatly influenced by the quality of the features used. Therefore, feature selection techniques become a crucial step to reduce redundancy, improve interpretability, and optimise model performance. Hybrid feature selection techniques, which combine filter-based approaches like Mutual Information and wrapper-based approaches like Recursive Feature Elimination (RFE), have been shown to produce more representative feature subsets [6][7]. The use of ensemble learning algorithms such as Random Forest and Extreme Gradient Boosting (XGBoost) empirically shows superior performance in classification problems compared to single methods [8][9][10]. The XGBoost excels in handling complex and imbalanced data, resulting in higher accuracy in academic prediction studies [11]. However, research that combines hybrid feature selection techniques and ensemble learning in the context of predicting student dropout based on academic data and digital activity in Indonesia is still very limited. Based on the exploration results of the secondary dataset, variables such as attendance rate, participation in online discussions, frequency of accessing learning resources, and academic scores were found to

influence student success. This dataset represents the common data structure used by academic information systems and learning management systems (LMS) in many higher education institutions in Indonesia.

Although various studies have examined the prediction of student dropout using machine learning algorithms and several feature selection approaches, there are still several important gaps that have not been widely explored. Most previous research only used a single approach in feature selection, such as filter or wrapper alone, without combining them into a hybrid framework that can maximise the relevance and effectiveness of features [6][12][13]. The dropout prediction approach generally only utilises academic data, without considering the digital behaviour of students, which is becoming increasingly important in the era of online learning and smart campuses. The utilisation of ensemble learning algorithms such as XGBoost in the context of predicting student dropout in Indonesia is still very limited, especially when combined with a hybrid feature selection approach. The lack of local research that integrates these three components, namely academic data and digital activities, hybrid feature selection techniques, and ensemble algorithms, shows a significant and strategic research gap that can be filled through this study.

This research focuses on developing a prediction model for student dropout using a hybrid feature selection and ensemble learning approach (Random Forest and XGBoost) by utilising academic data and digital activity. This model is expected to serve as a foundation for the development of an adaptive early warning system to effectively support smart campus management.

2. Literature Review

2.1. Data Mining

Data mining is an important process in data exploration aimed at discovering hidden patterns, relationships among attributes, and meaningful information from large datasets [14]. This process is a core part of Knowledge Discovery in Databases (KDD), which includes stages of cleaning, integration, transformation, and evaluation of patterns. Some main methods in data mining include classification, clustering, association, regression, and anomaly detection. As the complexity and volume of data increase, data mining approaches continue to evolve, including through integration with machine learning, real-time big data processing using Apache Spark and Flink, automated modelling (AutoML), and explainable AI (XAI) approaches to enhance algorithm transparency. Data mining has been widely applied in various fields, such as disease predictions in healthcare, fraud detection in the financial sector, student dropout predictions in education, as well as traffic and energy management in smart city systems. Although it has been widely developed, there are several research gaps that present opportunities for further development, such as the efficiency of processing large-scale data, the resilience of models to imbalanced data, the need for optimal feature selection, and the assurance of data privacy and security. One of the cutting-edge approaches that has emerged is quantum-assisted feature selection, which utilises quadratic unconstrained binary optimisation models to efficiently select subsets of features, showing significant potential in improving predictive model performance on large and complex data.

2.2. Random Forest algorithms

First proposed by Breiman in 2001, Random Forest is an ensemble learning technique that makes use of bagging and has become one of the most reliable classification and regression methods in data mining due to its ability to produce accurate models, resistance to overfitting, and capability to handle large-scale and high-dimensional data[15]. Using the bootstrap aggregating method, this algorithm constructs a collection of decision trees trained on randomly chosen subsets of data and features; the results from each tree are then merged using majority voting (for categorisation) or averaging (for regression). Mathematically, the final prediction of the Random Forest classification model can be written as [16]:

$$\hat{y} = majority\_vote\{h_t(x)\}_{t=1}^{T} \quad \text{..... (1)}$$

wherw $h_t(x)$ is a prediction function from the tree to t, x is the feature input, and T is the total number of trees in the forest. For the regression case, the formula is[17]:

$$\hat{y} = \frac{1}{T}\sum_{t=1}^{T} h_t\,(x) \quad \text{..... (2)}$$

This, therefore, suggests the predicted value is the average output from all trees. Random Forest's working mechanism is as follows: (1) bootstrapping samples from the training data; (2) creating trees on each sample with arbitrary feature selection at each node; and (3) averaging the prediction outputs of all trees. The main advantages of this algorithm include its ability to handle non-linear data, tolerance to outliers and missing data, as well as its capability to evaluate the importance of features through measurements such as Gini Importance or Mean Decrease in Impurity. Random Forest is also very effective in handling imbalanced data when combined with techniques such as SMOTE. Its use has spread widely in various fields such as fraud detection, disease diagnosis, student dropout prediction, and classification in intelligent systems. Recent research also points to the enhancement of Random Forest performance through hyperparameter optimisation, integration with metaheuristic-based feature selection methods (such as genetic algorithms and PSO), as well as hybrid approaches with quantum computing technologies like quantum-assisted feature selection based on QUBO, making it one of the most flexible and adaptive algorithms in the modern data mining domain.

2.2. XGBoost algorithms

Derived from the gradient boosting decision tree developed by Chen and Guestrin in 2016, Xtreme Gradient Boosting (XGBoost) is an ensemble algorithm that has become among of the most used predictive techniques due to its great speed, scalability, and great accuracy performance across many data science contests and commercial applications [18]. XGBoost works by building models in an additive manner, where each new decision tree attempts to correct the prediction errors of the previous model by minimizing the loss function using a gradient descent approach. Mathematically, XGBoost optimizes the objective function as follows [19]:

$$\mathcal{L}(\emptyset) = \Sigma_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t)}\right) + \Sigma_{k=1}^{t} \Omega(f_k) \quad \text{..... (3)}$$

With $l\left(y_i, \hat{y}_i^{(t)}\right)$ is the loss function (for example, log-loss for classification or MSE for regression) between the actual labels y_i, and prediction $\hat{y}_i^{(t)}$ whereas $\Omega(f_k) = \Upsilon T + \frac{1}{2} \lambda ||\omega||^2$ is a regularization function controlling tree complexity f_k with T number of leaves, w predicted weight at each leaf, Υ penalty structure, and λ L2 regularization. This algorithm also applies tree pruning techniques, automatic handling of missing values, and supports parallel processing, making it superior in big data scenarios. XGBoost has various advantages, such as being resistant to overfitting, able to handle sparse and skewed data, and providing informative feature importance estimates. This algorithm has been widely applied in the financial sector for credit scoring and fraud detection, in healthcare for disease prediction, as well as in education, marketing, and smart city sectors. Recent studies also show that the integration of XGBoost Utilizing smart feature selection methods such Recursive Feature Elimination, metaheuristics, and quantum-assisted methods can further improve the efficiency and interpretability of the model, making it one of the most adaptive and robust algorithms in various classification and regression problems[20], [21].

3. Research Method

This study uses an exploratory quantitative approach with educational data mining methods Construct a student dropout forecast model based on digital activity and academic information. This process can be seen in the following figure 1:

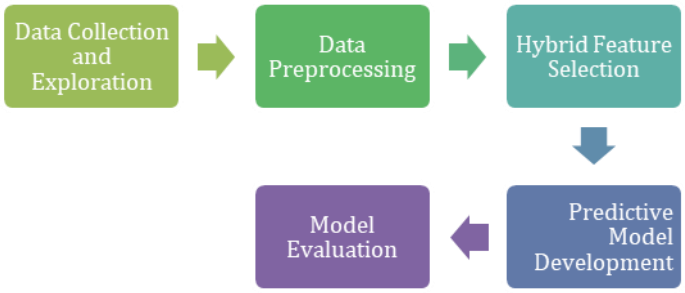

Fig 1. Flowchart of Research Methodology

The methodology process consists of five main stages: (1) data collection and exploration, (2) data preprocessing, (3) feature selection using a hybrid approach, (4) development of an ensemble learning-based classification model, and (5) model evaluation using classification performance metrics.

3.1 Data Collection and Exploration

The secondary dataset xAPI-Edu-Data is used in this study. This dataset contains 480 student data that includes academic attributes and digital activities, such as attendance levels, participation in class discussions, access to learning materials, gender, and final study results classified into three levels of success: low, middle, and high.

Table 1. Dataset

No	Gender	Nationality	Topic	Semester	Vis Resources	Annou View	Discussion	Absen	Class
1	1	KW	IT	F-	16	2	20	Under-7	M
2	1	KW	IT	F-	20	3	25	Under-7	M
3	1	KW	IT	F-	7	0	30	Above-7	L
4	1	KW	IT	F-	25	5	35	Above-7	L
5	1	KW	IT	F-	50	12	50	Above-7	M
6	0	KW	IT	F-	30	13	70	Above-7	M
7	1	KW	Math-	F-	12	0	17	Above-7	L
8	1	KW	Math-	F-	10	15	22	Under-7	M
9	0	KW	Math-	F-	21	16	50	Under-7	M
	…	…	…	…	…	…	…	…	…
480	0	Jordan	History	S-	14	23	62	Above-7	L

3.2 Data Preprocessing

The data preprocessing stage includes data cleaning, handling categorical values, and normalization. Several irrelevant or redundant columns, such as Nationality and Place of Birth, are removed from the dataset. Next, all categorical features are encoded into numerical form using label encoding techniques. To ensure the uniformity of the scale of numerical features, normalization is performed using Min-Max Scaling, so that all feature values are within the range [0, 1].

3.3 Hybrid Feature Selection

Feature selection was done using a hybrid approach that is, a combination of filter-based and wrapper-based techniques in order to increase the model's accuracy and efficiency[9][22]. First stage: the Mutual Information (MI). The ten most significant characteristics were found using an algorithm based on their correlation coefficients with the target variable. Next, further selection was performed, getting the best subset of features using the Recursive Feature Elimination (RFE) method with a Random Forest estimator[23][24]. This combination is expected to balance speed and accuracy in feature selection.

3.4. Development of Prediction Models

Two ensemble learning approaches, namely Random Forest and Extreme Gradient Boosting (XGBoost), are used to build the forecasting model. Both algorithms were chosen for their high capability in handling small to medium-sized data, as well as their ability to reduce overfitting through bagging and boosting mechanisms. The dataset that has undergone the feature selection stage is divided into two subsets using the stratified train-test split method with an 80:20 training and testing model ratio. The solution to the regression problem using the random forest algorithm is modelled as follows [25]:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(fi - yi)^2 \quad \text{..... (4)}$$

Where N represents the total count of data points, fi indicates the output provided by the model, and yi corresponds to the actual value of the data point i. The distance of each node from the real expected value is calculated by this formula, which helps to identify the branch representing the best option for the forest. yi represents the value of the data point evaluated at a certain node, while fi denotes the value provided by the decision tree. Calculating how the branches of the decision tree are set up as follows depends on the Gini index or formula while executing Random Forest based on classification information [26]:

$$Gini = 1 - \sum_{i=1}^{c} p_{i^2} \quad \text{..... (5)}$$

Using class and probability, this equation determines which branch at a node has the Gini, thereby highlighting which one is most likely to occur. In this situation p_i indicates the category seen's relative frequency in the database, and c signifies the total number of categories. To establish how nodes are branched in the decision tree, entropy can be utilized using the following equation[27]:

$$Entropy = \sum_{i=1}^{C} -p_i * log2(p_i) \quad \text{..... (6)}$$

Entropy utilizes the likelihood of specific results to influence decisions regarding node distribution. In contrast to the Gini index, this method involves more complex mathematics because it utilizes a logarithmic function in its computations.
The next formula in the regression solution using Extreme Gradient Boosting (XGBoost) is as follows[28], [29]:

a. Set the starting likelihood of forecasting $(Pr_i^1), with\ i = 1, 2, \ldots, n.$
b. Compute the residuals using the formula below:

$$Residual_i^t = y_i - Pr_i^t \quad \text{..... (7)}$$

c. Compute the coverage value of the attribute using this formula:

$$Cover\ a = \sum_{i=1}^{n}(Pr_i^t\ (1 - Pr_t^i)) \quad \text{..... (8)}$$

d. Compute the similarity score (SS) using the formula below:

$$SSnode = \frac{(\sum_{i-1}^{n} Residual_i)^2}{\sum_{i-1}^{n}(Pr_i^t(1-Pr_i^t))+\lambda} \quad \text{..... (9)}$$

e. Determine the attribute gain value using the formula below:

$$Gain(A) = SS_{left} + SS_{right} - SS_{root} \quad \text{..... (10)}$$

f. Compute the leaf output value using the given formula:

$$output(A)_i = \frac{(\sum_{i-1}^{n} Residual_i)}{\sum_{i-1}^{n}(Pr_i(1-Pr_i))+\lambda} \quad \text{..... (11)}$$

g. Compute the log odds value in this manner:

$$Log\ odds_i^t = \log\left(\frac{Pr_i^t}{1-Pr_i^t}\right) \quad \text{..... (12)}$$

h. Adjust the probability value to be standardized using the subsequent equation:

$$Pr_i^{t+1} = Log\ odds_i^t + (n\ x\ output\ (A)_i) \quad \text{..... (13)}$$

i. Adjust the probability value using the binary sigmoid function in this manner:

$$Sigmoid(Pr_i^{t+1}) = \frac{exp Pr_i^{t+1}}{1+exp Pr_i^{t+1}} \quad \text{..... (14)}$$

j. Reiterating step b up to i.
k. Assess the effectiveness of the classification algorithm.

3.5. Model Evaluation

The model performance evaluation is conducted using classification evaluation metrics, namely:

1. The accuracy of the proportion of correct predictions against the entire data set.
2. Precision, Recall, and F1-Score: utilized to assess the effectiveness and responsiveness of the model in identifying dropout.
3. ROC AUC (Receiver Operating Characteristic - Area Under Curve): utilized to assess how well the model can differentiate between various classes in general.

In addition, the confusion matrix is used to evaluate the classification distribution of the model against the dropout and non-dropout classes. The Confusion Matrix is a technique employed to assess the accuracy of the constructed model. The Confusion Matrix includes details comparing the model's output with the true classification outcomes. The four components that depict the results of classification in CM are True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN), as displayed in Table 2.[30], [31]:

Table 2. Confusion Matrix (CM)

Confusion Matrix		Actual	
		Positive	**Negative**
Predicted	+	TP	FP
		Correct result	Unexpected result
	-	FN	TN
		Missing result	Correct absence of result

Table 2 indicates that True Positive refers to a genuinely positive instance, True Negative denotes a genuinely negative instance, False Positive indicates a negative instance incorrectly classified as positive, and False Negative signifies a positive instance inaccurately classified as negative. The equations used to determine the accuracy, precision, sensitivity, and specificity metrics in evaluating the performance of classification algorithms are as follows[32]:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} x100\% \quad (9)$$

$$Precision = \frac{TP}{TP+FP} x100\% \quad (10)$$

$$Sensitivity = \frac{TP}{TP+FN} x100\% \quad (11)$$

$$Specificity = \frac{TN}{TN+FP} x100\% \quad (12)$$

4. Results and Discusssion

This research aims to develop a prediction model for student dropout based on academic data and digital activities in a smart campus environment using a hybrid feature selection approach and ensemble learning algorithms. The experiments conducted based on the implementation of the model on the secondary xAPI-Edu-Data dataset are presented step by step, starting from the evaluation results of the main model, interpretation of performance metrics, to a critical discussion regarding the effectiveness of the used approach. The main focus is directed towards the analysis of the model's accuracy, dropout detection capability, and the relevance of the selected features to the classification results. A comparison of the research findings with previous studies demonstrates the scientific contribution and superiority of the proposed approach. All analyses were conducted using the Extreme Gradient Boosting (XGBoost) algorithm as the best model selected based on the testing results of several other ensemble models, such as Random Forest. Model evaluation was carried out using standard classification metrics, namely accuracy, precision, recall, F1-score, and confusion matrix. This approach is designed to address the main issue in the research, which is how to identify at-risk students for dropout earlier and accurately in the context of smart campus management.

4.1. Results of the Random Forest Model Evaluation

The Random Forest model was trained with basic parameters using 100 estimators. The evaluation results of the model's performance are shown in the following Table 3:

Table 3. Classification Report Model Random Forest

Label (Dropout)	Precision	Recall	F1-score	Support
0 (Tidak Dropout)	0.90	0.93	0.91	74
1 (Dropout)	0.77	0.77	0.77	22
Accuracy			0.88	96
Macro Avg	0.83	0.85	0.84	96
Weighted Avg	0.88	0.88	0.88	96

The Random Forest model provides stable and balanced results. With a precision of 0.90 and a recall of 0.93 for the non-dropout class, the model demonstrates a high ability to recognise students who persist. However, for the dropout class, the precision and recall are only 0.77 each, indicating room for improvement in capturing patterns of students at risk of dropping out.

4.2. Results of XGBoost Model Evaluation

The XGBoost model provides better results compared to Random Forest. Below is the classification report of XGBoost, summarised in the following Table 4:

Table 4. Classification Report Model XGBoost

Label (Dropout)	Precision	Recall	F1-score	Support
0 (Tidak Dropout)	0.94	0.91	0.92	74

Label (Dropout)	Precision	Recall	F1-score	Support
1 (Dropout)	0.72	0.82	0.77	22
Accuracy			0.89	96
Macro Avg	0.83	0.86	0.85	96
Weighted Avg	0.89	0.89	0.89	96

The XGBoost model shows that its accuracy increased to 89% with an F1-score of 0.77 for the dropout class, higher than Random Forest. The lower precision in the dropout class (0.72) indicates that there are still several false positive predictions, but the high recall (0.82) shows that most of the students who truly dropped out were successfully identified by the model.

4.3. Confusion Matrix

The outcomes from the confusion matrix related to the XGBoost model are displayed in the following Figure 2:

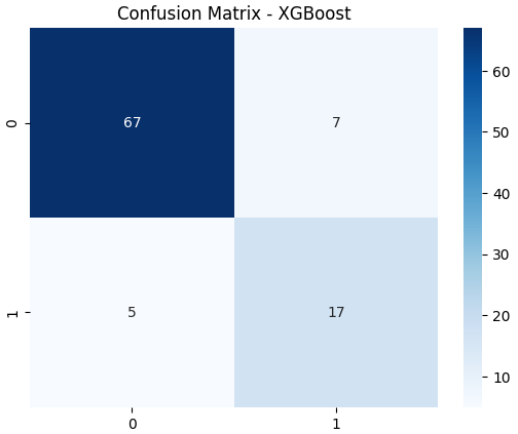

Fig 2. Confusion Matrix (CM)

The confusion matrix presents the outcomes in this manner: True Negative (TN) equals 67, False Positive (FP) is 7, False Negative (FN) amounts to 5, and True Positive (TP) stands at 17. Based on these results, several important evaluation metrics can be calculated as follows. :

$$\blacktriangleleft\ accurasi = \frac{TP + TN}{TP + TN + FP + FN} = \frac{17 + 67}{17 + 67 + 7 + 5} = \frac{84}{96} \approx 87{,}5\%$$

$$Precision = \frac{TP}{TP + FP} = \frac{17}{17 + 7} \approx 70{,}83\%$$

$$Recall = \frac{TP}{TP + FN} = \frac{17}{17 + 5} \approx 77{,}27\%$$

$$F1 - Score = 2x\ \frac{Precision\ x\ Recall}{Precision + Recall} \approx 73{,}89\%$$

$$Specificity = \frac{TN}{TN + FP} = \frac{67}{67 + 7} \approx 90{,}54\%$$

$$False\ Positive\ Rate(FPR) = \frac{FP}{FP + TN} = \frac{7}{74} \approx 9{,}46\%$$

The results show that the XGBoost model is capable of providing good classification performance in predicting student dropout potential. With an accuracy of 87.5% and an F1-score of 73.89%, this model is quite balanced in detecting students who drop out and those who do not.

5. Conclusion

This study seeks to develop a model that forecasts student dropouts by utilising a combined method for feature selection along with ensemble learning techniques, particularly focusing on Random Forest and XGBoost, on academic data and digital activity within the smart campus ecosystem. Based on the testing results that have been conducted, both algorithms demonstrate fairly good classification performance, with evaluation metrics including accuracy, precision, recall, and F1-score. The XGBoost model delivers the best performance result with an accuracy of 89%, a weighted average precision of 0.89, and a weighted average F1-score of 0.89. Meanwhile, the Random Forest model shows an accuracy of 88%, with a weighted average precision and f1-score of 0.88, respectively. From the obtained Confusion Matrix, the XGBoost model is more effective in identifying students at risk of dropping out (label 1), with a correct

prediction count of 17 out of 22 cases, indicating the model's ability to handle class imbalance. This result suggests that the integration of appropriate feature selection strategies with ensemble learning models can enhance the accuracy and efficiency of dropout prediction systems. Thus, this approach can serve as an effective tool for higher education institutions in early detection and data-driven decision-making to reduce the number of students who drop out. From the obtained Confusion Matrix, the XGBoost model is more effective in identifying students at risk of dropout (label 1), with 17 correct predictions out of 22 cases, demonstrating the model's ability to handle class imbalance. This result indicates that the integration of appropriate feature selection strategies with ensemble learning models can enhance the accuracy and efficiency of dropout prediction systems. Therefore, this approach can serve as an effective tool for higher education institutions in early detection and data-driven decision-making to reduce the number of students who drop out.

Acknowledgement

Thank you to the leadership of the Muhammad Nasir Foundation, AMIK and STIKOM Tunas Bangsa for their support in the implementation of the tridharma of higher education for lecturers and all academic community.

Author Contribution

The overall design of the research methodology, including the selection and extraction of hybrid features from the campus academic information system and the data dropout labelling process, was carried out by M. Safii. The focus on the development and implementation of the ensemble model for predicting dropout, including programming, hyperparameter tuning, and performance validation of the model using relevant evaluation metrics, was conducted by Adli Abdillah Nababan. Analysis of the results, writing of the scientific manuscript, and ensuring the quality and coherence of the manuscript's content overall, including the preparation of the discussion section and implications for the smart campus system, were conducted by Husian. The three authors actively contributed to the revision and final editing process of the article.

References

[1] Nurmalitasari, Z. Awang Long, and M. Faizuddin Mohd Noor, "Factors Influencing Dropout Students in Higher Education," *Educ. Res. Int.*, vol. 2023, 2023, doi: 10.1155/2023/7704142.

[2] S. A. Khairullah, S. Harris, H. J. Hadi, R. A. Sandhu, N. Ahmad, and M. A. Alshara, "Implementing artificial intelligence in academic and administrative processes through responsible strategic leadership in the higher education institutions," *Front. Educ.*, vol. 10, 2025, doi: 10.3389/feduc.2025.1548104.

[3] O. Jimenez, A. Jesús, and L. Wong, *Model for the Prediction of Dropout in Higher Education in Peru applying Machine Learning Algorithms: Random Forest, Decision Tree, Neural Network and Support Vector Machine*, vol. 33. 2023. doi: 10.23919/FRUCT58615.2023.10143068.

[4] S. A. Sulak and N. Koklu, "Predicting Student Dropout Using Machine Learning Algorithms," *PLUSBASE Akad. Organ. VE DANISMANLIK LTD STI*, vol. 3, pp. 91–98, Sep. 2024, doi: 10.58190/imiens.2024.103.

[5] A. D. Riyanto, A. A. Pratiwi, C. S. Faculty, U. A. Purwokerto, C. S. Faculty, and U. A. Purwokerto, "ANALYSIS OF FACTORS DETERMINING STUDENT SATISFACTION USING DECISION TREE , RANDOM FOREST , SVM , AND NEURAL NETWORKS : A ANALISIS FAKTOR PENENTU KEPUASAN MAHASISWA MENGGUNAKAN DECISION TREE , RANDOM FOREST , SVM , DAN NEURAL NETWORKS : SEBUAH STUDI KOMPAR," vol. 5, no. 4, pp. 187–196, 2024.

[6] P. Rani, A. Jain, and S. K. Chawla, "A Hybrid Approach for Feature Selection Based on Genetic Algorithm and Recursive Feature Elimination," *Int. J. Inf. Syst. Model. Des.*, vol. 12, pp. 17–38, Apr. 2021, doi: 10.4018/IJISMD.2021040102.

[7] N. S. Mohd Nafis and S. Awang, "An Enhanced Hybrid Feature Selection Technique Using Term Frequency-Inverse Document Frequency and Support Vector Machine-Recursive Feature Elimination for Sentiment Classification," *IEEE Access*, vol. 9, no. Ml, pp. 52177–52192, 2021, doi: 10.1109/ACCESS.2021.3069001.

[8] L. Shrivastav and R. Kumar, "An Ensemble of Random Forest Gradient Boosting Machine and Deep Learning Methods for Stock Price Prediction," *J. Inf. Technol. Res.*, vol. 15, pp. 1–19, Jan. 2022, doi: 10.4018/JITR.2022010102.

[9] E. M. Ferrouhi and I. Bouabdallaoui, "A comparative study of ensemble learning algorithms for high-frequency trading," *Sci. African*, vol. 24, p. e02161, 2024, doi: https://doi.org/10.1016/j.sciaf.2024.e02161.

[10] M. Imani, A. Beikmohammadi, and H. R. Arabnia, "Comprehensive Analysis of Random Forest and XGBoost Performance with SMOTE, ADASYN, and GNUS Under Varying Imbalance Levels," *Technologies*, vol. 13, no. 3, pp. 1–40, 2025, doi: 10.3390/technologies13030088.

[11] Tao-Hongli, "Educational data mining for student performance prediction: feature selection and model evaluation," *J. Electr. Syst.*, vol. 20, pp. 1063–1074, Apr. 2024, doi: 10.52783/jes.3434.

[12] L. Sun *et al.*, "A Hybrid Feature Selection Framework Using Improved Sine Cosine Algorithm with Metaheuristic Techniques," *Energies*, vol. 15, no. 10, pp. 1–24, 2022, doi: 10.3390/en15103485.

[13] A. Roman, M. M. Rahman, S. A. Haider, T. Akram, and S. R. Naqvi, "Integrating Feature Selection and Deep Learning: A Hybrid Approach for Smart Agriculture Applications," *Algorithms*, vol. 18, no. 4, pp. 1–26, 2025, doi: 10.3390/a18040222.

[14] M. Chaudhry, I. Shafi, M. Mahnoor, D. L. Vargas, E. B. Thompson, and I. Ashraf, "A Systematic Literature Review on Identifying Patterns Using Unsupervised Clustering Algorithms: A Data Mining Perspective," *Symmetry*, vol. 15, no. 9. 2023. doi: 10.3390/sym15091679.

[15] A. Yavuz Ozalp, H. Akinci, and M. Zeybek, "Comparative Analysis of Tree-Based Ensemble Learning Algorithms for Landslide Susceptibility Mapping: A Case Study in Rize, Turkey," *Water*, vol. 15, no. 14. 2023. doi: 10.3390/w15142661.

[16] J. Mark *et al.*, "Performance evaluation of random forest algorithm for automating classification of mathematics question items," *World J. Adv. Res. Rev.*, vol. 18(02), pp. 34–43, Apr. 2023, doi: 10.30574/wjarr.2023.18.2.0762.

[17] C. Starbuck, "Linear Regression BT - The Fundamentals of People Analytics: With Applications in R," C. Starbuck, Ed., Cham: Springer International Publishing, 2023, pp. 181–206. doi: 10.1007/978-3-031-28674-2_10.

[18] I. Cherif and A. Kortebi, *On using eXtreme Gradient Boosting (XGBoost) Machine Learning algorithm for Home Network Traffic Classification*. 2019. doi: 10.1109/WD.2019.8734193.
[19] G. Huang, Z. Liu, Y. Wang, and Y. Yang, "A Multi-Objective Prediction XGBoost Model for Predicting Ground Settlement, Station Settlement, and Pit Deformation Induced by Ultra-Deep Foundation Construction," *Buildings*, vol. 14, no. 9. 2024. doi: 10.3390/buildings14092996.
[20] A. Mehdary, A. Chehri, A. Jakimi, and R. Saadane, "Hyperparameter Optimization with Genetic Algorithms and XGBoost: A Step Forward in Smart Grid Fraud Detection," *Sensors*, vol. 24, no. 4. 2024. doi: 10.3390/s24041230.
[21] N. Linh *et al.*, "Flood susceptibility modeling based on new hybrid intelligence model: Optimization of XGboost model using GA metaheuristic algorithm," *Adv. Sp. Res.*, vol. 69, Feb. 2022, doi: 10.1016/j.asr.2022.02.027.
[22] E. Hancer, "An improved evolutionary wrapper-filter feature selection approach with a new initialisation scheme," *Mach. Learn.*, vol. 113, no. 8, pp. 4977–5000, 2024, doi: 10.1007/s10994-021-05990-z.
[23] M. Beraha, A. M. Metelli, M. Papini, A. Tirinzoni, and M. Restelli, *Feature Selection via Mutual Information: New Theoretical Insights*. 2019. doi: 10.1109/IJCNN.2019.8852410.
[24] K. Robindro, U. B. Clinton, N. Hoque, and D. K. Bhattacharyya, "JoMIC: A joint MI-based filter feature selection method," *J. Comput. Math. Data Sci.*, vol. 6, p. 100075, 2023, doi: https://doi.org/10.1016/j.jcmds.2023.100075.
[25] S.-A. Amamra, "Random Forest-Based Machine Learning Model Design for 21,700/5 Ah Lithium Cell Health Prediction Using Experimental Data," *Physchem*, vol. 5, no. 1. 2025. doi: 10.3390/physchem5010012.
[26] G. N., P. Jain, A. Choudhury, P. Dutta, K. Kalita, and P. Barsocchi, "Random Forest Regression-Based Machine Learning Model for Accurate Estimation of Fluid Flow in Curved Pipes," *Processes*, vol. 9, no. 11. 2021. doi: 10.3390/pr9112095.
[27] H. Salman, A. Kalakech, and A. Steiti, "Random Forest Algorithm Overview," *Babylonian J. Mach. Learn.*, vol. 2024, pp. 69–79, Jun. 2024, doi: 10.58496/BJML/2024/007.
[28] M. W. Dwinanda, N. Satyahadewi, and W. Andani, "Classification of Student Graduation Status Using Xgboost Algorithm," *BAREKENG J. Ilmu Mat. dan Terap.*, vol. 17, no. 3, pp. 1785–1794, 2023, doi: 10.30598/barekengvol17iss3pp1785-1794.
[29] M. Wiens, A. Verone-Boyle, N. Henscheid, J. Podichetty, and J. Burton, "A Tutorial and Use Case Example of the eXtreme Gradient Boosting (XGBoost) Artificial Intelligence Algorithm for Drug Development Applications," *Clin. Transl. Sci.*, vol. 18, Mar. 2025, doi: 10.1111/cts.70172.
[30] C. Hong and T. Oh, "TPR-TNR plot for confusion matrix," *Commun. Stat. Appl. Methods*, vol. 28, pp. 161–169, Mar. 2021, doi: 10.29220/CSAM.2021.28.2.161.
[31] M. Safii, S. Efendi, M. Zarlis, and H. Mawengkang, "Intelligent evacuation model in disaster mitigation," *Bull. Electr. Eng. Informatics*, vol. 11, no. 4, pp. 2204–2214, 2022, doi: 10.11591/eei.v11i4.3805.
[32] S. Swaminathan and B. R. Tantri, "Confusion Matrix-Based Performance Evaluation Metrics," *African J. Biomed. Res.*, vol. 27, pp. 4023–4031, Nov. 2024, doi: 10.53555/AJBR.v27i4S.4345.