



Analysis of Campus Facility Availability and Academic Community Satisfaction Using Logistic Regression and Random Forest

Nurul Azizah¹, Moch. Rizky Ramadani², Nadia Siti Ramadhani³, Ramdhan Giovanni Syarif Rizan⁴, Najma Syahidah⁵, Jameel Ahmed⁶

^{1,2,3}Universitas Pembangunan Nasional “Veteran” Jawa Timur, Indonesia, ⁴Institut Teknologi Sepuluh Nopember (ITS), Surabaya, Indonesia, ⁵Al-Azhar University, Cairo, Egypt, ⁶Quaid-e-Awam University of Engineering, Science & Technology (QUEST), Nawabshah, Pakistan
¹25083010051@student.upnjatim.ac.id, ²25083010052@student.upnjatim.ac.id, ³25083010053@student.upnjatim.ac.id, ⁴5008251093@student.its.ac.id, ⁵najmasyahidah662@gmail.com, ⁶Jameelahmedmemon10@gmail.com

ABSTRACT

Campus facilities play an important role in supporting academic activities and student satisfaction. This study aimed to classify student satisfaction with campus facilities using Logistic Regression and Random Forest. Survey data from 150 student respondents were used, with overall satisfaction scores of 1–3 classified as lower satisfaction and scores of 4–5 as higher satisfaction. Five facility availability predictors were selected using Pearson correlation with a threshold of ($r \geq 0.20$). The dataset was divided into training and testing sets using an 80:20 stratified split. Class weighting, five-fold stratified cross-validation, and GridSearchCV were applied to address class imbalance and optimize model hyperparameters. On the testing set, the tuned Random Forest achieved 93.33% accuracy, 83.33% precision, 83.33% recall, an F1-score of 83.33%, and a ROC-AUC of 86.81%. Logistic Regression achieved 90.00% accuracy, 71.43% precision, 83.33% recall, an F1-score of 76.92%, and a ROC-AUC of 92.36%. Random Forest was selected as the final model because it achieved higher test-set accuracy, precision, and F1-score than Logistic Regression. Feature importance analysis identified parking availability as the most important predictor, with an importance value of 0.3899. These findings suggest that Random Forest can classify student satisfaction in this dataset, although further validation using larger samples is needed to assess model generalizability.

Keywords: campus facilities, student satisfaction, Logistic Regression, Random Forest, feature selection.

I. INTRODUCTION

Campus facilities are vital resources for supporting academic and non-academic activities at institutions of higher education. The availability of facilities such as classrooms, places of worship, restrooms, and internet access directly impacts students' comfort and perceived satisfaction. Effective facility management not only supports the smooth running of the learning process but also serves as a key indicator of the quality of operational services at higher education institutions in maintaining academic competitiveness. Therefore, evaluating user satisfaction levels regarding the availability of campus facilities is crucial as a basis for strategic decision-making by university management.

Although the provision of campus facilities continues to improve, students' perceptions of satisfaction often vary due to capacity limitations and differences in actual usage. Real-world issues frequently faced by students include limited parking availability and unstable internet connectivity across campus. The gap between the physical availability of facilities and students' expectations fuels dissatisfaction that has the potential to disrupt the learning process. Without objective, data-driven mapping and classification, campus administrators will struggle to identify which facilities require the most urgent improvements.

Recent studies have applied machine learning and statistical methods to assess student satisfaction. Tiftazani et al. [1] used the Random Forest algorithm to classify student satisfaction with academic services and reported an accuracy of 94%. Meanwhile, Ermawati et al. [2] applied logistic regression to model satisfaction with academic administrative services and identified "tangibles" and "reliability" as significant service dimensions. However, previous studies have primarily focused on academic and administrative services, while research specifically classifying student satisfaction based on the availability of essential physical campus facilities remains limited. Furthermore, the selection of relevant facility availability predictors using Pearson correlation and the handling of imbalanced satisfaction classes are important considerations in developing a reliable classification model. These gaps highlight the need for a comparative study that focuses on physical campus facilities, applies correlation based feature selection, and accounts for class imbalance in evaluating student satisfaction.

To address these limitations, this study aims to analyze and classify students' satisfaction levels with campus facilities using two machine learning algorithms: Logistic Regression and Random Forest. This study employs a feature selection process based on Pearson correlation coefficients, retaining only predictors with ($r \geq 0.20$) and excluding variables below this threshold, as correlation values below this level represent poor or negligible relationships [3] and retaining such irrelevant features can introduce noise and degrade machine learning model performance [4]. Class weighting was also applied to account for the imbalanced distribution of satisfaction classes, and multiple classification metrics were used to evaluate model performance. The selection process resulted in five primary facility predictor variables including parking, Wi-Fi, places of worship, restrooms, and classrooms. Through this comparative approach, the study aims to identify a suitable classification model and determine which facility-availability predictors contribute most to the model's predictions, providing preliminary insights for campus facility improvement planning.

II. METHODOLOGY

A. Research Design

This study employed a quantitative research approach using survey data and machine learning methods to classify the overall satisfaction of student respondents based on perceived campus facility availability [1], [2]. The research process consisted of data collection, data cleaning, data preprocessing, target variable transformation, stratified data splitting, Pearson correlation-based feature selection, multicollinearity assessment, model development, hyperparameter tuning, final testing, and interpretation of results.

Two classification algorithms were investigated: Logistic Regression [2] and Random Forest [1]. Logistic Regression was implemented using a pipeline incorporating StandardScaler, whereas Random Forest was used as a tree-based ensemble classifier [1]. Both algorithms were trained using the selected predictor variables, with class weighting applied to address the imbalanced target distribution. Hyperparameter tuning was performed separately for each algorithm using GridSearchCV with five-fold stratified cross-validation and F1-score as the scoring metric. The best hyperparameter configuration for each algorithm was determined using cross-validation on the training set, and the resulting tuned models were evaluated on the held-out testing set. Model performance before and after tuning was compared using accuracy, precision, recall, F1-score, and ROC-AUC. The overall research workflow is presented in Fig. 1.

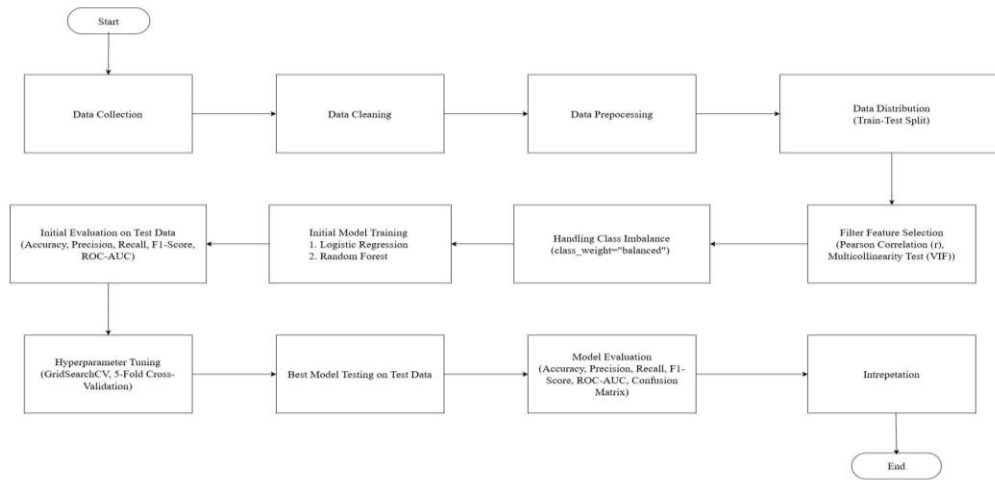


Fig. 1. Research Workflow

B. Data Collection

The data were obtained from a survey entitled Analysis of Availability and Satisfaction of the Academic Community with Campus Facilities, which collected information on respondents' perceptions of campus facility availability, experiences in using the facilities, and overall satisfaction. After the data-cleaning process, the dataset consisted of 150 respondents and 45 columns, covering respondent information, facility availability assessments, facility-related experiences, facility usage information, and overall satisfaction as the target variable. Although the survey was designed for the academic community, the respondents analyzed in this study were students; therefore, the findings reflect their perceptions and experiences and should not be directly generalized to other groups, such as lecturers or administrative staff. Personal identification variables, including respondents' names, student identification numbers, WhatsApp numbers, and timestamps, were excluded from the predictor set, as was the overall satisfaction variable because it served as the target variable.

C. Data Preprocessing

Data preprocessing was performed to prepare the survey data for statistical analysis and machine learning modeling. The preprocessing stage included data cleaning, transformation of survey responses into numerical values, construction of the binary target variable, and preparation of the predictor variables. Facility availability variables were represented on a numerical scale of 1–5. Overall satisfaction was converted into a binary target, with scores of 1–3 classified as class 0 and scores of 4–5 as class 1. The resulting target variable consisted of 118 observations (78.67%) in class 0 and 32 observations (21.33%) in class 1. The predictor set was restricted to the five facility-availability variables. Variables measuring satisfaction with individual facilities and personal identification information were not included as predictors because the objective was to examine whether facility availability could classify overall satisfaction.

D. Data Splitting

Data splitting is the process of dividing a dataset into a training set and a testing set to build a model and evaluate its performance on data that was not used during training [5]. The mathematical formulation of data splitting is presented in Equations (1) and (2).

$$n_{train} = r \times n \quad (1)$$

$$n_{test} = (1 - r) \times n \quad (2)$$

Description:

n : total number of data points.

r : proportion of training data.

n_{train} : number of training data points.

n_{test} : number of test data points.

The dataset was divided into training and testing sets using an 80:20 ratio. The training set was used to develop the classification models and optimize their hyperparameters, while the testing set was reserved for evaluating the final model performance on unseen observations. Stratified splitting was applied to preserve the approximate proportion of the two target classes in both subsets. With 150 observations, the split produced 120 observations for training and 30 observations for testing. The same data partition was used for both classification algorithms to ensure a fair comparison.

E. Feature Selection and Multicollinearity Checking

Feature selection was performed to determine the predictor variables used in the model. In this study, feature selection was conducted using Pearson's correlation with the target variable, while the Variance Inflation Factor (VIF) was used to check for potential multicollinearity among the predictor variables

1. Pearson's Correlation

The Pearson correlation coefficient is used to measure the direction and strength of the linear relationship between a predictor variable and a measured variable. The mathematical formula for the Pearson correlation coefficient is presented in Equation (3).

$$r_{XY} = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2 \sum_{i=1}^n (Y_i - \bar{Y})^2}} \quad (3)$$

Description:

r_{XY} : Pearson's correlation coefficient between X and Y.

X_i : the value of the predictor variable at the i-th observation.

Y_i : the value of the target variable at the i-th observation.

$\bar{X}$: the mean of variable X.

$\bar{Y}$: the mean of variable Y.

n : the number of observations analyzed.

A Pearson correlation threshold of ($r \geq 0.20$) was adopted as an operational criterion for feature selection [3], [4] to prevent weak predictors from introducing noise into the models. Shen et al. [6] applied a similar threshold by selecting predictors with Pearson correlation coefficients greater than 0.20 in their study of life satisfaction. In this study, only predictors with positive correlation coefficients meeting the specified threshold were retained, while negatively correlated predictors and predictors below the threshold were excluded. Accordingly, the threshold of ($r \geq 0.20$) was adopted as a criterion specific to this study and should not be interpreted as a universal standard. The selected predictors were subsequently used in the Logistic Regression and Random Forest models. To prevent data leakage, Pearson correlation-based feature selection was performed using the training set only. The selected predictors were then used to train the models, while the testing set was reserved for final performance evaluation.

2. Variance Inflation Factor (VIF)

VIF is used to identify potential multicollinearity, which is a condition where one predictor variable has a linear relationship with another predictor variable. The mathematical formulation of VIF is presented in Equation (4).

$$VIF_j = \frac{1}{1 - R_j^2} \quad (4)$$

Description:

VIF_j : the VIF value for the jth predictor variable.

R_j^2 : the coefficient of determination for the regression of the j-th predictor variable on the other predictor variables.

VIF values below 5 were used as the practical criterion for indicating the absence of severe multicollinearity [7]. This assessment was conducted to determine whether the selected predictors could be used together in the classification models.

F. Class Imbalance Handling

The target variable exhibited class imbalance, with class 0 accounting for 78.67% of the observations and class 1 accounting for 21.33%. To reduce the influence of the majority class during model training, class weighting was applied using the `class_weight="balanced"` option, which automatically adjusts the class weights based on the frequency of each class [8]. The class weights are calculated as shown in Equation (5).

$$w_c = \frac{n}{K n_c} \quad (5)$$

Description:

w_c : weight for class c.

n : number of training data points.

K : number of classes

n_c : number of observations in class c.

G. Logistic Regression

logistic regression is a classification method used to predict the probability that an observation belongs to one of two classes based on predictor variables [5].

1. Logistic Regression Model

The Logistic Regression method was used in this study because the target variable consists of two classes: students who are dissatisfied (class 0) and those who are satisfied (class 1) with campus facilities. This method can model the probability that a student falls into the satisfied class based on predictor variables such as the availability of parking areas, Wi-Fi, places of worship, restrooms, and classrooms. Mathematically, the logistic function is expressed in Equation (6) [5].

$$z_i = \beta_0 + \sum_{j=1}^p \beta_j X_{ij} \quad (6)$$

Description:

z_i : the linear value for the i-th observation.

β_0 : the model constant.

β_j : the regression coefficient for the j-th predictor.

X_{ij} : the value of the j-th predictor for the i-th observation.

p : the number of predictor variables.

2. Logistic Function

A mathematical function used to map the results of linear combinations of independent variables to probability values between 0 and 1. The mathematical formulation of the logistic function is presented in Equation (7).

$$p_i(Y_i = 1 | X_i) = \frac{1}{1 + e^{-z_i}} \quad (7)$$

Description:

$p_i(Y_i = 1 | X_i)$: the probability that the target Y is 1 based on the input variable (X)

e : Euler's number.

z_i : the linear coefficient of the logistic regression equation.

3. Classification Rule

Classification rules are used to convert the probabilities generated by the model into class 0 or 1 based on a specified probability threshold. The classification rule is mathematically expressed in Equation (8).

$$\begin{aligned} \hat{Y}_i &= 1, \hat{p}_i \geq 0.5 \\ &= 0, \hat{p}_i < 0.5 \end{aligned} \quad (8)$$

Description:

$\hat{Y}_i$: predicted class

$\hat{p}_i$: probability of predicting class 1.

Before fitting the Logistic Regression model, StandardScaler was applied to standardize the predictor variables within a modeling pipeline. Data scaling using StandardScaler is a commonly applied preprocessing step to bring predictor variables to a comparable scale and improve model convergence [9].

H. Random Forest

Random Forest is an ensemble learning classification method that combines the predictions of multiple decision trees to determine the final class through majority voting. This method was selected because it can capture nonlinear relationships and interactions among predictor variables without requiring a linear relationship between the predictors and the target variable. In this study, Random Forest was used as a comparative classification model alongside Logistic Regression to classify satisfaction with campus facilities into two classes. The model was trained using the training dataset and evaluated using the testing dataset to assess its classification performance [10]. The majority voting mechanism is formulated in Equation (9).

$$\hat{Y} = \text{mode}\{T_1(X_i), T_2(X_i), \dots, T_B(X_i)\} \quad (9)$$

Description:

$\hat{Y}_i$: the predicted class.

$T_b(X_i)$: Prediction result from decision tree b.

B : the number of decision trees in the Random Forest.

mode : the most frequently predicted class.

I. Initial Model Evaluation

Calculation of classification performance from each method test with Confusion Matrix to obtain accuracy, precision, recall. A confusion matrix is used to evaluate classification results and determine the model's ability to classify data into each class [10]. The detailed layout and components of the confusion matrix used to break down the positive and negative classifications are illustrated in Table I.

Table I. Confusion Matrix

Actual Value	Positive	Negative
Positive	True Positive	False Negative
Negative	False Positive	True Negative

The performance of the classification model was evaluated using several metrics: accuracy, precision, recall, specificity, and F1-score. These metrics were used to assess the model's performance in classifying data based on the predicted results and the actual classes [10]. The formulas for these metrics are presented in Equations (10)–(13).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (10)$$

$$Precision = \frac{TP}{TP + FP} \quad (11)$$

$$Recall = \frac{TP}{TP + FN} \quad (12)$$

$$F1 = \frac{2TP}{2TP + FP + FN} \quad (13)$$

Description:

Accuracy : the proportion of all correct predictions.

Precision : the proportion of correct positive predictions.

Recall : the proportion of positive data successfully identified by the model.

F1 – Score : the harmonic mean of precision and recall.

TP : positive data correctly predicted as positive.

TN : negative data correctly predicted as negative.

FP : negative data predicted as positive.

FN : positive data predicted as negative.

J. Hyperparameter Tuning and Cross-Validation

Hyperparameter optimization was performed using GridSearchCV with 5-fold cross-validation. Hyperparameter tuning using grid search and cross-validation is used to find the best combination of parameters for a Random Forest model[11]. The hyperparameter selection process is mathematically expressed in Equation (14).

$$\begin{aligned} \bar{S}_{CV}(\theta) \\ = \frac{1}{k} \sum_{k=1}^k S_k(\theta) \end{aligned} \quad (14)$$

Description:

$S_{CV}(\theta)$: the average validation score for the hyperparameter combination.

θ : the hyperparameter combination being tested.

k : the number of folds

$S_k(\theta)$: the evaluation score for the i-th fold.

K. Final Model Evaluation

The best-performing Logistic Regression and Random Forest classification models obtained through hyperparameter tuning and cross-validation were evaluated on the test dataset using accuracy, precision, recall, F1-score, ROC-AUC, and the confusion matrix. These metrics were used to compare the classification performance of both models [12]. The mathematical formulations of these measures and the ROC-AUC are presented in Equations (15)–(17).

$$TPR = \frac{TP}{TP + FN} \quad (15)$$

$$FPR = \frac{FP}{FP + TN} \quad (16)$$

$$AUC = \int_0^1 TPR(FPR) d(FPR) \quad (17)$$

Description:

AUC : the area under the ROC curve..

TPR : True Positive Rate, which is the same as recall.

FPR : False Positive Rate, calculated as $\frac{FP}{FP+TN}$.

ROC-AUC must be calculated using probabilities or prediction scores, not just class labels 0 and 1.

L. Model Interpretation

Random Forest feature importance was used to interpret the classification model by identifying the relative contribution of each predictor to the model's predictions. Feature importance provides a model-specific measure for comparing the importance of input variables, while higher importance values indicate greater contributions to the model's classification process [13]. The feature importance of each predictor in the Random Forest model is calculated as presented in Equation (18).

$$\sum_{j=1}^p FI_j = 1 \quad (18)$$

This means that the sum of the feature importance values for all features equals 1 because the values have been normalized.

Description:

FI_j : the importance value of the jth feature.

p : the number of features used in the model.

III. RESULT AND DISCUSSION

A. Descriptive Analysis of Campus Facility Availability and Satisfaction

This study used 150 respondents to analyze ratings of campus facilities and the level of satisfaction among the academic community. Descriptive statistics were used to examine the distribution of ratings for the five campus facilities that were the focus of the analysis, as presented in Table II.

Table II. Distribution of Campus Facility Ratings

Facility	Score 1	Score 2	Score 3	Score 4	Score 5
Parking Area	28	58	44	16	4
Wi-Fi	14	44	62	24	6
Place of Worship	11	24	61	37	17
Restroom	2	21	48	49	30
Classroom	1	8	50	41	50

Based on Table II, the parking area received the most scores of 2, while Wi-Fi and places of worship received the most scores of 3. Restrooms received the most scores of 4, while classrooms had the highest number of ratings tied at scores of 3 and 5. These results indicate that the distribution of ratings varies across facilities.

Next, the target satisfaction variable was grouped into two classes: Class 0 for satisfaction scores of 1–3 and Class 1 for scores of 4–5. The distribution of the number and percentage of respondents in each class is presented in Table III.

Table III. Distribution of Target Satisfaction Classes

Class	Satisfaction Category	Count	Percentage
0	score 1 – 3	118	78.67%
1	score 4 – 5	32	21.33%
Total		150	100%

Table III shows that Class 0 dominates the data with a proportion of 78.67%, while Class 1 accounts for only 21.33%. This distribution indicates a class imbalance that must be taken into account in the modeling process. Therefore, model evaluation should consider metrics such as precision, recall, and F1-score in addition to accuracy. To provide further context on students' perceptions of campus facilities, Table IV presents the mean availability ratings for each facility.

Table IV. Descriptive Statistics of Campus Facility Availability

Variable	Availability
Parking Area	2.40
Wi-Fi	2.76
Place of worship	3.17
Restroom	3.56
Classroom	3.87

Table IV shows that classrooms received the highest mean availability rating (3.87), followed by

restrooms (3.56) and places of worship (3.17). Meanwhile, Wi-Fi (2.76) and parking areas (2.40) received the lowest ratings. These findings indicate that parking areas and Wi-Fi should be prioritized for improvement to better meet students' needs. However, the scores reflect respondents' perceptions of availability rather than actual facility capacity.

B. Feature Selection

1. Results of Feature Selection Using Pearson's Correlations

Feature selection was performed to identify predictor variables that have a linear relationship with the target variable. Based on a Pearson correlation coefficient of ($r \geq 0.20$) five variables were selected, as shown in Table V.

Table V. Feature Selection Results Using Pearson's Correlation

Variable	Pearson Correlation	Decision
Parking Area	0.4791	Selected
Wi-Fi	0.3034	Selected
Place of Worship	0.2935	Selected
Restroom	0.2675	Selected
Classroom	0.2141	Selected

Based on Table V, the parking area had the highest correlation value of 0.4791, while the classroom had the lowest value of 0.2141 among the selected features. All five variables showed a positive correlation with the target variable and met the selection criteria used. These variables were then used as predictors in Logistic Regression and Random Forest models.

2. Results of the Multicollinearity Test Using VIF

After determining the features to be used, a multicollinearity check was performed to assess the degree of linear relationship among the predictors. The VIF results for the five selected variables are presented in Table VI.

Table VI. VIF Results for Selected Variables

Variable	VIF Value
Parking Area	1.2357
Wi-Fi	1.5785
Place of worship	1.9693
Restroom	1.2647
Classroom	1.7266

Based on Table VI, all five selected predictors had VIF values below 5, which was adopted as a practical threshold for assessing multicollinearity [7]. These results indicate no evidence of severe multicollinearity among the selected predictors according to this criterion. Therefore, the five predictors were retained for use in the Logistic Regression and Random Forest models.

C. Hyperparameter Tuning and Cross-Validation Results

The next step is to determine the optimal combination of hyperparameters for Logistic Regression and Random Forest using stratified 5-fold cross-validation. The tuning results and the

average validation F1-score for both models are presented in Table VII.

Table VII. Hyperparameter Tuning Results

Model	Best Parameters	Average CV F1-score
Logistic Regression	C = 1	0.5790
Random Forest	max_depth = 5, min_samples_leaf = 2, min_samples_split = 5, n_estimators = 200	0.5573

Table VII shows that Logistic Regression achieved a higher average cross-validation F1-score (0.5790) than Random Forest (0.5573). These scores represent average performance across five validation folds within the training data. Thus, based on cross-validation F1-score, Logistic Regression performed better during model selection. The final comparison also considers performance on the held-out test set, as reported in Table VIII.

D. Comparative Classification Performance

The classification performance of Logistic Regression and Random Forest was evaluated before and after hyperparameter tuning using the held-out testing set. The evaluation considered accuracy, precision, recall, F1-score, and the Area Under the Receiver Operating Characteristic Curve (ROC-AUC). The results are presented in Table VIII.

Table VIII. Classification Performance Comparison

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Logistic Regression — Before tuning	90.00%	71.43%	83.33%	76.92%	92.36%
Random Forest — Before tuning	90.00%	80.00%	66.67%	72.73%	83.68%
Logistic Regression — After tuning	90.00%	71.43%	83.33%	76.92%	92.36%
Random Forest — After tuning	93.33%	83.33%	83.33%	83.33%	86.81%

Before hyperparameter tuning, both models achieved an accuracy of 90.00%, although their other performance metrics differed. Logistic Regression achieved higher recall (83.33%), F1-score (76.92%), and ROC-AUC (92.36%), whereas Random Forest achieved higher precision (80.00%). After tuning, Logistic Regression's test-set performance remained unchanged, while Random Forest improved in accuracy (from 90.00% to 93.33%), precision (from 80.00% to 83.33%), recall (from 66.67% to 83.33%), F1-score (from 72.73% to 83.33%), and ROC-AUC (from 83.68% to 86.81%).

Despite these improvements, the two models showed different strengths. Logistic Regression achieved a higher mean five-fold cross-validation F1-score (0.5790) than Random Forest (0.5573) and retained a higher test-set ROC-AUC (92.36% versus 86.81%). In contrast, the tuned Random Forest achieved higher test accuracy, precision, recall, and F1-score. This indicates that Random Forest performed better on the test set at the applied classification threshold, whereas Logistic Regression showed stronger discrimination based on predicted probabilities and a slightly higher cross-validation F1-score. However, these differences should be interpreted cautiously because the test set contained only 30 observations, including six in the higher-satisfaction class.

To further examine the classification results, confusion matrices were used to compare the numbers of correctly and incorrectly classified observations for each model. The confusion matrices

of the tuned models are presented in Fig. II.

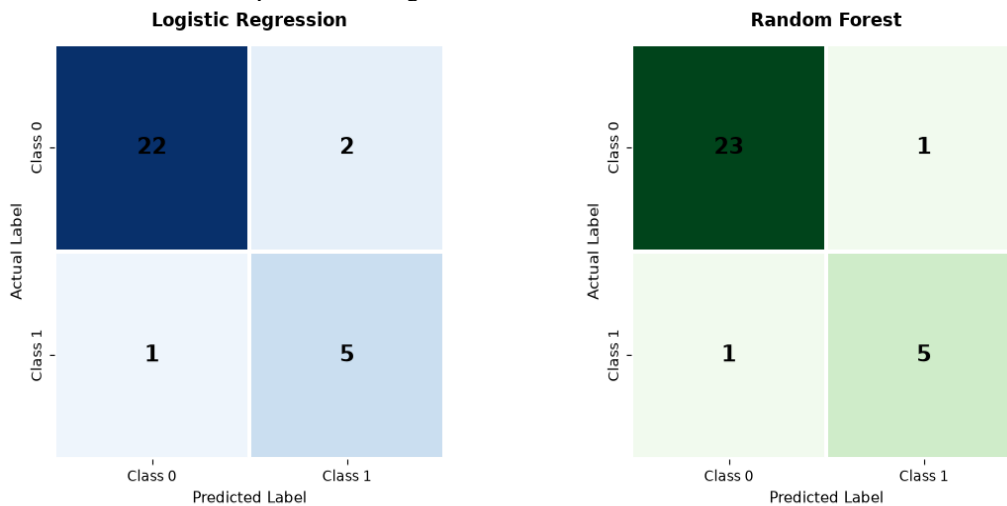


Fig. II. Confusion matrices of Logistic Regression and Random Forest after hyperparameter tuning.

The confusion matrices show that Random Forest correctly classified 28 of the 30 testing observations, while Logistic Regression correctly classified 27. Both models correctly identified five of the six observations in class 1. These results indicate that Random Forest produced fewer classification errors overall, although the small number of class 1 observations in the testing set requires cautious interpretation.

E. Final Model Interpretation: Random Forest

The tuned Random Forest model was selected as the final model based on its test-set performance. It achieved an accuracy of 93.33%, precision of 83.33%, recall of 83.33%, and F1-score of 83.33%. These results were higher than those of Logistic Regression in terms of accuracy, precision, and F1-score, while both models achieved the same recall. However, Logistic Regression obtained a higher ROC-AUC of 92.36%, compared with 86.81% for Random Forest. Therefore, the selection of Random Forest reflects its stronger performance on the threshold-based classification metrics considered in this study rather than its superiority across all evaluation metrics.

To illustrate the decision-making structure of the tuned Random Forest model, one representative decision tree from the ensemble was visualized, as shown in Fig. III. The tree begins with Parking Availability as the root node, using a threshold of 3.5 to divide the observations into two branches. The left branch contains 62 samples with parking ratings of 3 or below, while the right branch contains 12 samples with ratings above 3.5. The left branch further evaluates Classroom Availability and Toilet Availability, whereas the right branch evaluates Wi-Fi Availability and, for some observations, Toilet Availability. Each node displays the number of samples reaching that node, the weighted class values, and the predicted class. Because class weighting was applied during training, the displayed class values represent weighted counts rather than simple raw counts. The terminal nodes predict either Lower Satisfaction or Higher Satisfaction. This tree illustrates how combinations of facility ratings are used for classification; however, it represents only one component of the Random Forest ensemble and does not fully explain the predictions of the entire model.

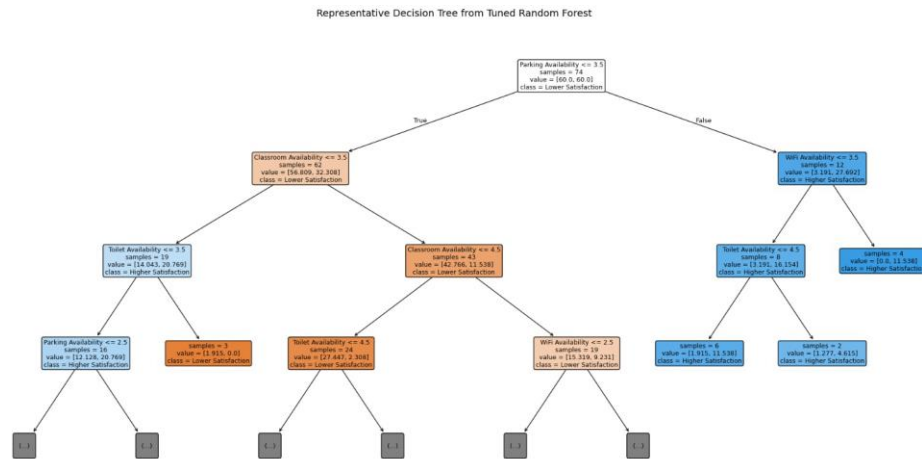


Fig. III. Representative decision tree from the tuned Random Forest model.

Feature importance analysis was subsequently performed to examine the relative contribution of the five facility-availability predictors to the predictions generated by the tuned Random Forest model. The results are presented in Fig. IV.

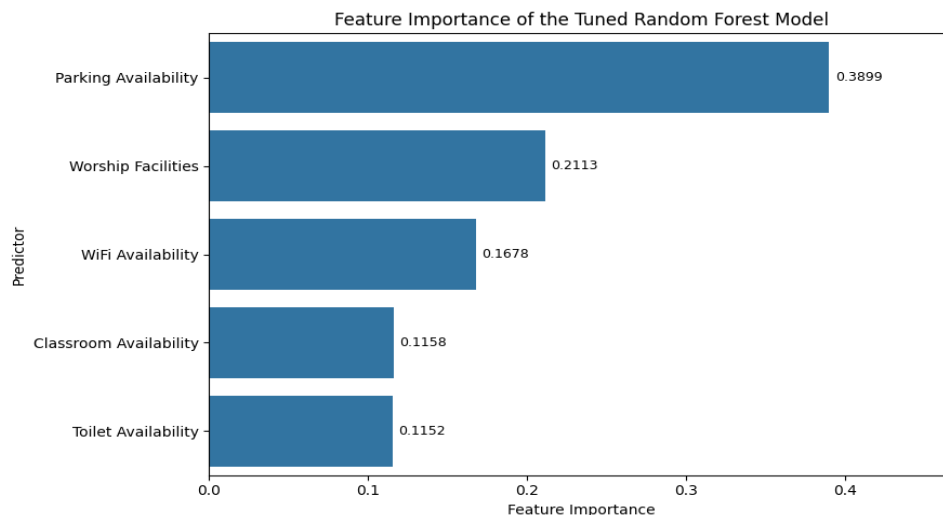


Fig. IV. Feature importance of the Random Forest model.

As shown in Fig. IV, parking availability had the highest feature importance (0.3899), followed by place-of-worship availability (0.2113), Wi-Fi availability (0.1678), classroom availability (0.1158), and restroom availability (0.1152). Parking also had the lowest mean availability rating among the five predictors, at 2.40 out of 5. Together, these findings suggest that parking availability warrants further evaluation in campus facility planning. Campus administrators could investigate parking capacity, availability during peak periods, and students' parking needs before deciding on improvements. However, feature importance represents the relative contribution of each predictor to the fitted model's predictions and does not establish a causal relationship with student satisfaction. The results should also be interpreted cautiously because the test set contained only 30 observations.

These findings are consistent with Tiftazani et al. [1], who reported 94% accuracy using Random Forest to classify student satisfaction with academic services. In this study, Random Forest achieved 93.33% accuracy and a higher test-set F1-score than Logistic Regression. However, Logistic Regression obtained a higher cross-validation F1-score, indicating that Random Forest did not perform better across all evaluation measures. The results should therefore be interpreted cautiously, particularly because the test set contained only 30 observations. Future research should use larger and more diverse samples and external validation to assess the models' performance stability and generalizability.

IV. CONCLUSION

This study classified student satisfaction with campus facilities using Logistic Regression and Random Forest. Five facility-availability predictors were selected using Pearson correlation with a threshold of ($r \geq 0.20$). To address the imbalanced distribution of satisfaction classes, class weighting was applied during model training. Multicollinearity was assessed using the Variance Inflation Factor (VIF), with a threshold of 5 treated as a practical guideline rather than an absolute criterion [7].

On the testing set, the tuned Random Forest achieved 93.33% accuracy, 83.33% precision, 83.33% recall, and an 83.33% F1-score. It outperformed Logistic Regression in accuracy, precision, and F1-score, while both models achieved the same recall. However, Logistic Regression obtained a higher ROC-AUC (92.36% versus 86.81%) and cross-validation F1-score (0.5790 versus 0.5573). Random Forest was selected as the final model based on its stronger test-set classification performance, although the results require cautious interpretation because the testing set contained only 30 observations. Feature importance identified parking availability (0.3899) as the most important predictor, followed by worship facilities (0.2113) and WiFi availability (0.1678). Parking also had the lowest mean availability score (2.40).

These findings suggest that facility availability ratings may provide useful information for classifying students' overall satisfaction within the scope of this study. However, given the limited sample of 150 student respondents, the findings should be interpreted as preliminary and should not be generalized to broader student populations or other campuses without further validation. Parking availability is a candidate priority for further evaluation by campus administrators. However, because feature importance does not establish a causal effect, these results should be treated as indicators for further investigation. Future research should use larger and more diverse samples, additional predictors, and external validation to assess generalizability.

V. REFERENCES

- [1] A. Tiftazani, Tenriawaru, and G. A. Rahman, "Analisis klasifikasi kepuasan mahasiswa terhadap penyelenggaraan pelayanan akademik FMIPA Universitas Halu Oleo menggunakan algoritma Random Forest," *AnoaTIK: Jurnal Teknologi Informasi dan Komputer*, vol. 2, no. 2, pp. 77–83, Dec. 2024.
- [2] E. Ermawati, W. Alwi, and Arifin, "Pemodelan kepuasan mahasiswa terhadap layanan administrasi akademik di UIN Alauddin Makassar," *Teknosains: Media Informasi dan Teknologi*, vol. 17, no. 2, pp. 158–165, Aug. 2023.
- [3] H. Akoglu, "User's guide to correlation coefficients," *Turkish Journal of Emergency Medicine*, vol. 18, no. 3, pp. 91–93, Aug. 2018.
- [4] N. T. Romadloni, B. S. Prakoso, Miswanto, I. Santoso, K. Setiawan, and S. Aditya, "Comparative evaluation of statistical feature selection methods for text classification in user review mining," in *7th International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS)*, 2025, pp. 1451–1456.
- [5] G. James, D. Witten, T. Hastie, R. Tibshirani, and J. Taylor, *An Introduction to Statistical Learning: With Applications in Python*. Cham, Switzerland: Springer, 2023.
- [6] X. Shen, F. Yin, and C. Jiao, "Predictive Models of Life Satisfaction in Older People: A Machine Learning Approach," *International Journal of Environmental Research and Public Health*, vol. 20, no. 3, Art. no. 2445, 2023, doi: 10.3390/ijerph20032445.
- [7] S. G. Singh and S. V. Kumar, "Dealing with multicollinearity problem in analysis of side friction characteristics under urban heterogeneous traffic conditions," *Arabian Journal for Science and Engineering*, vol. 46, pp. 10739–10755, 2021, doi: 10.1007/s13369-020-05213-y.
- [8] B. Bakırarar and A. H. Elhan, "Class weighting technique to deal with imbalanced class problem in machine learning: Methodological research," *Turkish Journal of Clinics and Laboratory*, vol. 15, no. 1, pp. 19–29, 2023, doi: 10.5336/biostatic.2022-93961.
- [9] M. M. Ahsan, M. A. P. Mahmud, P. K. Saha, K. D. Gupta, and Z. Siddique, "Effect of data scaling methods on machine learning algorithms and model performance," *Technologies*, vol. 9, no. 3, Art. no. 52, 2021, doi: 10.3390/technologies9030052.
- [10] E. Helmud, E. Helmud, F. Fitriyani, and P. Romadiana, "Classification Comparison Performance of Supervised Machine Learning Random Forest and Decision Tree Algorithms Using Confusion Matrix", *SISFOKOM*, vol. 13, no. 1, pp. 92–97, Feb. 2024, doi: <https://doi.org/10.32736/sisfokom.v13i1.1985>.

-
- [11] R. Septia, Junadhi, S. Erlinda, and W. Agustin, "Heart failure disease classification using Random Forest algorithm with Grid Search Cross Validation technique," *The Indonesian Journal of Computer Science*, vol. 14, no. 2, Apr. 2025, doi: 10.33022/ijcs.v14i2.4765.
 - [12] Scikit-learn developers, "Model evaluation: Quantifying the quality of predictions," *scikit-learn*. [Online]. Available: https://scikit-learn.org/stable/modules/model_evaluation.html. [Accessed: Oct. 19, 2026].
 - [13] M. Saarela and S. Jauhiainen, "Comparison of feature importance measures as explanations for classification models," *SN Applied Sciences*, vol. 3, Art. no. 272, 2021, doi: 10.1007/s42452-021-04148-9.