Analysis of University Students' Awareness of Phishing Attacks Using Naive Bayes and Random Forest Algorithms

Rahma Syari, Muhammad Iqbal

Abstract

Phishing attacks remain one of the most common cybersecurity threats, particularly due to users' lack of awareness in identifying suspicious messages and protecting personal information. This study aims to classify students' awareness levels toward phishing attacks using the Naive Bayes and Random Forest algorithms. The dataset consisted of 100 respondent data obtained through a questionnaire and was classified into three awareness categories: low, medium, and high. The data were divided into 80 training data and 20 testing data. The evaluation of both models was conducted using accuracy, precision, recall, F1-score, and a confusion matrix. The experimental results showed that both Naive Bayes and Random Forest achieved an accuracy of 100% on the testing data. All categories obtained precision, recall, and F1-score values of 1.00. Furthermore, the Random Forest model identified Skor_Skenario and Rata_Rata_Perilaku as the most influential features in the classification process. The results indicate that both algorithms performed effectively in classifying students' awareness levels toward phishing attacks based on the dataset used in this study.

Keywords : Phishing, Cybersecurity Awareness, Naive Bayes, Random Forest, Machine Learning, Classification.

[1]Rahma Syari
[1]Teknologi Informasi, Universitas Pembangunan Pancabudi, Indonesia
e-mail: rahmaa.syahrii@gmail.com[2]

[2]Muhammad Iqbal
[2]Teknologi Informasi, Universitas Pembangunan Pancabudi, Indonesia
e-mail: muhammadiqbal@dosen.pancabudi.ac.id[2]
2nd International Conference on Islamic Community Studies (ICICS)
Theme: History of Malay Civilisation and Islamic Human Capacity and Halal Hub in the Globalization Era
https://proceeding.pancabudi.ac.id/index.php/ICIE/index

Introduction

The rapid development of digital technology has transformed the way university students carry out nearly all academic activities. Course-related information, assignment submission, communication with lecturers, document storage, and various university administrative services are now largely conducted through the internet. While this development provides greater convenience, it also increases the information security risks faced by students. One threat that continues to exploit users' lack of vigilance is *phishing*. Such attacks do not necessarily rely on technical vulnerabilities in a system; instead, they exploit users' trust, habits, and lack of awareness when receiving messages or accessing links. Recent research also emphasizes that phishing remains a persistent human-centered security problem in higher education, where attackers exploit trust and contextual cues rather than technical weaknesses alone [15], [16].

Phishing is a form of social engineering in which attackers impersonate trusted individuals, organizations, or services to persuade victims to disclose sensitive information, such as usernames, passwords, personal data, or financial information. These attacks may be delivered through email, instant messaging platforms, social media, or fraudulent websites. In higher education environments, university students represent a relatively vulnerable group because of their intensive use of digital services. Okokpujie *et al.* reported that 70.6% of the students involved in their study remained vulnerable to email-based *phishing* attacks due to insufficient awareness of the characteristics of *phishing* messages [1]. This finding indicates that proficiency in using technology in everyday life does not necessarily correspond to an adequate ability to recognize cybersecurity threats.

This vulnerability has also been identified in studies specifically involving university students. Ruzaili *et al.* found that, among 202 students surveyed, 27.2% reported having experienced a *phishing* attack. Their study also revealed a positive relationship between preventive behavior, cybersecurity awareness, and *phishing* awareness. Interestingly, knowledge of *phishing* did not automatically show a significant relationship with either awareness or preventive behavior [2]. This suggests that understanding the definition or general characteristics of *phishing* alone is insufficient to ensure that an individual will respond safely when directly confronted with such a threat.

A similar situation has been observed in Indonesia. Malahayati and Nabilah conducted a study involving Information Technology students at UIN Ar-Raniry using a *phishing*-based social engineering attack scenario. The results showed that some students still accessed *phishing* links. Although several respondents did not proceed to submit their personal information, 10.7% of the respondents accessed the link and entered data on the *phishing* website [3]. This finding is particularly important because even students with a technology-related educational background may engage in risky behavior when confronted with social engineering attacks. Therefore, cybersecurity awareness cannot be assessed solely on the basis of an individual's educational background or level of technology use.

Awareness of *phishing* is influenced by multiple factors rather than a single determinant. Internet experience, the ability to recognize suspicious message characteristics, understanding of information security, and habitual responses when receiving messages or links may lead to different behavioral responses among individuals. Nur'aini and Simanjuntak found that internet experience and anti-phishing knowledge influenced *phishing* awareness [4]. Meanwhile, Ismail *et al.* demonstrated that educational programs could improve students' awareness scores following training, although some participants remained at a moderate level of awareness after the intervention [5]. These findings indicate that students' awareness levels are not necessarily uniform, highlighting the need for an approach capable of identifying patterns that distinguish groups of students with different levels of awareness.

A number of studies on students' *phishing* awareness have primarily employed descriptive approaches to present the percentages of respondents' answers or behaviors. Such approaches

are useful for providing a general overview; however, they do not fully explain how combinations of respondent characteristics can be used to predict or classify levels of awareness. The use of *machine learning* classification techniques offers an alternative approach for extending this type of analysis. Questionnaire data can be analyzed not only descriptively but also by identifying underlying patterns that can be used to determine respondents' awareness categories. Recent measurement studies likewise recommend evaluating security awareness through both knowledge and reported practices, rather than relying on familiarity with terminology alone [17], [18].

Classification-based approaches to students' *phishing* awareness have begun to emerge in several studies. Tangka and Putra, for example, analyzed questionnaire responses from 153 Indonesian university students using the *Classification and Regression Tree* (CART) algorithm. The developed model achieved an accuracy of 94.9%, a recall of 93.4%, a precision of 95.8%, and a ROC-AUC value of 0.971 [6]. Their study demonstrates that data concerning students' knowledge of and responses to *phishing* can be treated as a classification problem rather than being analyzed solely through percentages or descriptive statistics. Nevertheless, different classification methods still need to be investigated because each algorithm employs a different mechanism for learning patterns from respondent data.

Two algorithms that can be applied for this purpose are Naive Bayes and Random Forest. Naive Bayes performs classification using a probabilistic approach based on the relationship between predictor attributes and the class to be predicted. Its relatively simple structure makes it suitable for datasets containing multiple questionnaire-based attributes. Random Forest, in contrast, operates by constructing multiple decision trees and aggregating their predictions. This approach enables Random Forest to capture more complex patterns while reducing the dependence of the classification result on a single decision tree. These approaches are consistent with established machine-learning practice, in which probabilistic and ensemble classifiers are compared on the same training and test partitions [19], [20].

Both algorithms have also been widely applied in research related to phishing. Windarni et al. compared Naive Bayes, Decision Tree, and Random Forest algorithms for phishing website classification. Using their dataset, Naive Bayes achieved an accuracy of 60.4%, whereas Random Forest achieved 96.3% [7]. In another study, Juremi and Kurniadi integrated Random Forest and Naive Bayes for phishing email detection and obtained an accuracy of 96.81% using the combined model [8]. These results cannot be assumed to yield the same level of performance when the algorithms are applied to student awareness data because the characteristics of the datasets are different. Nevertheless, these studies demonstrate that Naive Bayes and Random Forest provide viable classification mechanisms that are worth comparing in a phishing-related context. Recent comparative experiments similarly report that Random Forest can perform strongly in phishing-URL classification, although results depend on the dataset and feature design [12], [21], [22], and [23].

Furthermore, Shahbazi, Jalali, and Molaeevand emphasized that addressing *phishing* requires not only artificial intelligence-based detection systems but also an understanding of users' ability to recognize threats. Their study involving university students showed that individuals may possess knowledge of *phishing* techniques without fully understanding the associated risks [9]. This finding further supports the importance of considering the human factor as an essential component of cybersecurity research. Even well-designed security systems may lose their effectiveness when users are unable to recognize risky situations or make appropriate decisions when confronted with suspicious messages.

Based on the aforementioned studies, there remains an opportunity to integrate the assessment of university students' *phishing* awareness with a *machine learning*-based classification approach. Previous studies involving university students have largely focused on descriptively measuring awareness levels, examining the effects of educational interventions, or applying a single classification method. In contrast, comparisons between Naive Bayes and Random Forest have more commonly been conducted in studies focusing on the detection of

phishing websites or emails rather than on the classification of students' awareness levels. Therefore, comparing these two algorithms using student awareness data may provide a different perspective, particularly in identifying which method is more suitable for recognizing patterns of awareness levels based on respondents' questionnaire responses. Recent work on student phishing awareness also supports combining human-focused awareness assessment with computational classification, while cautioning that self-reported awareness and observed behavior are not interchangeable [15], [16], [24].

Based on these considerations, this study aims to analyze university students' awareness of *phishing* attacks and to compare the performance of the Naive Bayes and Random Forest algorithms in classifying students' awareness levels. The research data are obtained through a questionnaire containing indicators related to students' understanding of and responses to various characteristics of *phishing* attacks. The performance of the two algorithms is then evaluated using metrics such as *accuracy*, *precision*, *recall*, and *F1-score*. The findings are expected not only to provide an overview of students' cybersecurity awareness but also to identify the classification algorithm that is more suitable for the characteristics of the dataset used in this study. These findings may subsequently serve as a reference for higher education institutions in designing more targeted cybersecurity education and awareness initiatives.

Literature Review

2.1 Phishing and Students' Cybersecurity Awareness

Phishing is a form of social engineering attack that exploits human interaction to obtain sensitive information. Attackers typically construct messages, web pages, or other forms of communication that imitate legitimate services in order to persuade victims to open links, disclose personal information, or perform certain actions. The primary issue in this type of attack does not lie solely in the technology employed by the attacker. Users' decisions when encountering suspicious messages are also a critical factor in determining whether an attack succeeds.

In higher education environments, this issue is particularly relevant because university students rely on a wide range of digital services in their daily activities. Email, social media, online learning systems, cloud storage, and various academic accounts continuously expose students to activities that require authentication and information exchange. Malahayati and Nabilah found that even students with a technology-related educational background may engage in risky behavior when confronted with *phishing* scenarios. In their study, some respondents opened the provided links, and several proceeded to submit information through *phishing* websites [3]. This finding indicates that familiarity with technology does not necessarily translate into greater awareness of social engineering threats.

Cybersecurity awareness should also not be understood merely as knowledge of the definitions or types of security threats. Alqahtani demonstrated that students' cybersecurity awareness is associated with various technology-related behaviors, including password management, browser security, and social media activities [10]. Meanwhile, Bognár and Bottyán developed an instrument for measuring students' cybersecurity behavior and emphasized that awareness assessment should consider users' actual behavior rather than relying solely on their perceptions of their own abilities [11]. Therefore, This multidimensional view is consistent with recent work measuring information-security awareness in university students and examining the role of password, browser, and social-media practices [17], [18].

2.2 Naive Bayes

Naive Bayes is a probability-based classification algorithm that applies Bayes' Theorem to estimate the probability that an observation belongs to a particular class. The basic principle of this algorithm is to assign an observation to the class with the highest posterior probability based on its attributes. In general, the classification process can be expressed as follows:

More specifically, the classification process using the Naive Bayes algorithm can be formulated as:

$$\hat{C} = \arg\max_{C} P(C) \prod_{i=1}^{n} P\left(x_i \mid C\right)$$

$where\ \hat{C}$ denotes the predicted class, $P(C)$ represents the prior probability of class C, and $P(x_i \mid C)$ represents the conditional probability of attribute x_i given class C. The symbol n denotes the number of attributes used in the classification process.

The function arg max is used to determine the class with the highest probability based on the probability values calculated from all attributes. The algorithm is referred to as *naive* because it assumes that the attributes are conditionally independent given the class. In real-world situations, particularly in human behavioral data, the relationships among attributes are not always completely independent. Nevertheless, Naive Bayes remains widely used because of its simple model structure, relatively low computational requirements, and ability to provide satisfactory results across various classification problems. In this study, Naive Bayes is used as one of the models to examine the extent to which a probabilistic approach can identify patterns in students' awareness levels. The conditional-independence assumption is a simplifying modeling choice; the method remains useful as a transparent baseline for classification comparisons [19], [25].

2.3 Random Forest

Random Forest is an *ensemble learning* algorithm that constructs multiple decision trees based on variations in data samples and feature subsets. Unlike a single decision tree, which generates predictions from one tree structure, Random Forest combines the predictions produced by multiple decision trees. In classification tasks, the final class is generally determined through *majority voting* across all constructed trees. Random Forest is an ensemble approach that aggregates multiple decision trees, following the foundational formulation of Breiman [20].

In a simplified form, the Random Forest prediction can be expressed as:

$$\hat{C}(x) = \text{mode}\{h_1, (x), h_2, (x) \ldots h_B(x)\}$$

where $h_b(x)$ represents the classification result produced by the-b, while B denotes the total number of decision trees used. The mode function is used to determine the class that occurs most frequently among the predictions generated by all decision trees..

These characteristics make Random Forest more flexible in capturing complex patterns and interactions among attributes. In *phishing*-related studies, Random Forest has been reported to achieve higher performance than simpler classification algorithms in several cases [7], [14]. Nevertheless, this study does not assume that Random Forest will necessarily be the best-performing model. Both algorithms are trained and evaluated using the same dataset and evaluation scheme so that conclusions regarding the better-performing model are determined by the experimental results rather than by findings from previous studies. Comparative phishing-detection studies show that the relative performance of Random Forest and Naive Bayes varies across datasets, feature sets, and task definitions [21]–[23].

2.4 Research Position

Previous studies reveal two general trends. Research on university students' *phishing* awareness has primarily focused on respondents' levels of knowledge, behavior, or vulnerability [3], [9]–[11], whereas the application of Naive Bayes and Random Forest has more commonly been directed toward the classification of *phishing* websites, URLs, or emails [7], [8], [14]. The study conducted by Tangka and Putra [6] provides an example of how student awareness data can be formulated as a classification problem; however, their study employed the CART algorithm.

Based on this research landscape, the present study positions **students' awareness levels** as the target of classification and compares two approaches with different characteristics: Naive Bayes, which is probability-based, and Random Forest, which is based on an *ensemble of decision trees*. The comparison is not limited to identifying the highest accuracy value but also considers the balance among *precision, recall,* and *F1-score*. Through this approach, the study is expected to provide an overview of students' *phishing* awareness while also identifying which algorithm is more suitable for the characteristics of the collected awareness data.

Research Methodology

3.1 Research Instrument

Data were collected using a questionnaire developed by adapting indicators from previous studies on cybersecurity and *phishing* awareness [3], [6], [10], [11]. The instrument was divided into three sections to ensure that the data used to construct the class labels were not identical to the data used as predictors. Questionnaire design should also distinguish awareness, knowledge, intention, and self-reported behavior as related but separate constructs [15], [17], [18].

The first section collected information on respondents' profiles and technology-use experience, including semester of study, field of study, intensity of internet use, experience with receiving *phishing* messages, and participation in cybersecurity training or educational activities. Personal information unrelated to the objectives of the study was not collected.

The second section assessed students' habits and understanding of digital security. The indicators included practices such as verifying the sender's identity, checking domain addresses before opening links, remaining cautious of messages that demand immediate action, protecting personal information, using additional authentication mechanisms, and responding appropriately to messages perceived as suspicious. Security-related behavioral aspects were included because awareness is not always adequately reflected by respondents' self-reported knowledge alone [10], [11].

The third section consisted of scenarios or situational examples representing both *phishing* messages and legitimate communications. Respondents were asked to determine whether each situation was suspicious and to select the most appropriate action to take. Correct responses were assigned a score of 1, whereas incorrect responses were assigned a score of 0. The scores obtained from this section were used as the basis for determining respondents' awareness levels.

The awareness levels were classified into three categories: **low, moderate, and high**. The score thresholds for each category were established prior to the model training process based on the results of instrument testing and expert judgment. These thresholds were not determined by considering the performance of either Naive Bayes or Random Forest, thereby ensuring that the class-label formation process was not influenced by the outcomes of the classification algorithms.

Separating the items used to construct the class labels from the predictor attributes was important to prevent *data leakage*. Items directly used to determine the awareness-level labels were not reused as input features for the classification models. Consequently, the algorithms did not merely learn the scoring rules used to generate the labels but instead attempted to identify patterns in students' characteristics associated with different levels of *phishing* awareness.

Before the instrument was used for the main data collection, it was reviewed to ensure content relevance and clarity of each statement. A pilot test was conducted with respondents who were not included in the main study sample. The internal consistency of items within the same construct was assessed using Cronbach's alpha. The resulting coefficient was interpreted in conjunction with the structure and characteristics of the instrument, as Cronbach's alpha should not be used as the sole criterion for evaluating the quality of a measurement scale [13]. Cronbach's alpha should be interpreted alongside the scale structure and other validity evidence, rather than treated as a stand-alone proof of measurement quality [13], [25].

3.2 Data Preparation

The questionnaire data were first examined before proceeding to the modeling stage. The data screening process included checking for missing values, duplicate responses, incomplete responses, and the consistency of each variable's format. Identifying information that was not relevant to the analysis was removed from the dataset.

Categorical attributes were then encoded into numerical form so that they could be processed by the classification algorithms. Negatively worded items were reverse-scored to ensure that higher scores consistently represented better security-related behavior. The class distribution was subsequently examined to determine whether the numbers of respondents in the low, moderate, and high awareness categories were relatively balanced.

The dataset was divided using a *stratified split*, with **80% allocated to the training set and 20% to the test set**. Stratification was applied to preserve the proportion of each awareness level in both subsets. All model parameter tuning was performed using only the training data. The test data were kept separate and were used only after the final models had been established. Stratified evaluation and cross-validation can help reduce dependence on a single partition, although the reliability of estimates remains limited by the sample size [19], [25].

3.3 Modeling

The first model developed was Naive Bayes. The choice of the Naive Bayes variant was determined by the final form of the attributes after the *preprocessing* stage. If the attributes were represented as numerical or composite scores, Gaussian Naive Bayes was used. If most attributes were represented as categorical data, a Naive Bayes variant appropriate for categorical features was selected and explicitly reported in the study results. The selection of the variant was based on the characteristics of the data rather than on which model produced the highest evaluation score.

The second model was developed using Random Forest. The main hyperparameters considered included the number of trees (*n_estimators*), the maximum tree depth (*max_depth*), the minimum number of samples required to split an internal node (*min_samples_split*), and the minimum number of samples required to be at a leaf node (*min_samples_leaf*). The optimal combination of hyperparameters was determined using *grid search* on the training data.

To reduce the possibility that model performance depended heavily on a single data split, the hyperparameter search was conducted using *5-fold cross-validation*. Under this scheme, the training data were divided into five subsets, with the training and validation processes performed iteratively across the folds. The model with the best-performing hyperparameter configuration was then retrained using the entire training set and subsequently evaluated on the previously unseen test set.

3.4 Model Evaluation

The performance of Naive Bayes and Random Forest was compared using the same test set. Model evaluation was conducted using the *confusion matrix, accuracy, precision, recall,* and *F1-score*. The use of multiple evaluation metrics was necessary to ensure that model performance was not assessed solely on the basis of the overall number of correct predictions. Reporting precision, recall, and F1-score alongside accuracy is especially useful when classification involves multiple classes or uneven class distributions [19], [25].

The *accuracy* value was calculated using the following formula:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

The *precision* value was calculated using the following formula:

$$\text{Precision} = \frac{TP}{TP + FP}$$

Meanwhile, the *recall* value was calculated using the following formula:

$$\text{Recall} = \frac{TP}{TP + FN}$$

The *F1-score* was then calculated using the following formula:

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Where:

TP (*True Positive*): the number of instances correctly predicted as positive.

TN (*True Negative*): the number of instances correctly predicted as negative.

FP (*False Positive*): the number of instances predicted as positive but actually negative.

FN (*False Negative*): the number of instances predicted as negative but actually positive.

Because this study classified awareness levels into more than two categories, namely low, moderate, and high, *precision, recall,* and *F1-score* were calculated for each class and reported using the *macro-average*. This approach assigns equal weight to each category, ensuring that model performance on classes with fewer instances is not overshadowed by more dominant classes.

The better-performing algorithm was determined based on the overall evaluation results by considering *accuracy, macro-precision, macro-recall,* and *macro F1-score*. The *confusion*

matrix was also analyzed to identify the types of classification errors occurring within each category. Therefore, model selection was not based solely on a single accuracy value.

All data processing and model development procedures were performed using Python with the *pandas* and *scikit-learn* libraries. The descriptive analysis results were used to characterize students' *phishing* awareness, while the classification results were used to compare the performance of the Naive Bayes and Random Forest algorithms. The implementation can be reproduced using the documented scikit-learn tools for preprocessing, model fitting, and evaluation [25].

3.5 Research Ethics

Participation in the study was voluntary, and respondents were informed about the purpose of the research before completing the questionnaire. The questionnaire did not request passwords, authentication codes, bank account information, or any other confidential information. When examples of emails, web pages, or links were used in *phishing* scenarios, all materials were presented as simulations and did not direct respondents to any page designed to collect authentication data. The research data were used solely for analytical purposes and were reported in aggregate form, ensuring that individual identities were not included in the classification results.

Results

This study used questionnaire data on university students' awareness of *phishing* attacks. A total of 100 respondent records were included in the classification process. The dataset was divided into 80 training instances and 20 testing instances. Classification was performed using two *machine learning* algorithms, namely Naive Bayes and Random Forest. Model performance was evaluated using *accuracy, precision, recall, F1-score,* and the *confusion matrix*.

4.1 Results of Naive Bayes Classification

The evaluation results using the Naive Bayes algorithm showed that the model was able to classify *phishing* awareness levels with very high performance. Based on the test results, the model achieved an *accuracy* of 100.00%.

The test set consisted of three awareness categories: Low, Moderate, and High. The classification report showed that all categories achieved *precision, recall,* and *F1-score* values of 1.00.

Table 1. Classification Results of Naive Bayes

Category	Precision	Recall	F1-Score	Support
Rendah	1.00	1.00	1.00	6
Sedang	1.00	1.00	1.00	8
Tinggi	1.00	1.00	1.00	6
Accuracy			**1.00**	**20**
Macro Average	1.00	1.00	1.00	20
Weighted Average	1.00	1.00	1.00	20

Based on Table 1, the Naive Bayes algorithm correctly classified all instances in the test set. A *precision* value of 1.00 indicates that all predictions generated by the model corresponded to the actual categories. In addition, a *recall* value of 1.00 indicates that all instances in each category were successfully identified by the model.

An *F1-score* of 1.00 across all categories indicates a perfect balance between *precision* and *recall*. Therefore, based on the evaluation of the 20 test instances, the Naive Bayes algorithm achieved perfect classification performance in identifying students' *phishing* awareness levels on the dataset used in this study.

4.2 Confusion Matrix of Naive Bayes

The Naive Bayes *confusion matrix* showed that all instances in the test set were correctly classified according to their actual categories.

Table 2. Confusion Matrix of Naive Bayes

Actual / Prediction	Rendah	Sedang	Tinggi
Rendah	6	0	0
Sedang	0	8	0
Tinggi	0	0	6

Based on the confusion matrix, all 6 instances in the Low category were correctly classified as Low. Similarly, all 8 instances in the Moderate category were correctly classified as Moderate. Meanwhile, all 6 instances in the High category were successfully classified as High.

No misclassifications occurred in the test set. This result indicates that the Naive Bayes model was able to effectively identify the patterns represented by the data characteristics used in this study.

4.3 Random Forest Classification Results

After evaluating the Naive Bayes method, this study proceeded with the implementation of the Random Forest algorithm. Random Forest was used to classify students' levels of awareness of *phishing* attacks based on the collected respondent data.

The dataset used in the evaluation consisted of 100 respondent records. Before the classification process was carried out, the data underwent a *preprocessing* stage to ensure that they were suitable for model development. Based on the preprocessing results, the dataset remained unchanged at 100 records, with no observations removed.

In the Random Forest classification process, a total of 13 features were used. Several categorical variables were first encoded so that they could be processed by the algorithm. These categorical variables included Field of Study, Internet Usage Intensity, Previous Experience of Receiving *Phishing* Messages, and Previous Participation in Security Training.

After the *preprocessing* and encoding stages were completed, the dataset was divided into training and test sets using an 80:20 ratio, resulting in 80 training instances and 20 test instances.

The training data were used to build the Random Forest model, while the test data were used to evaluate the model's ability to classify students' levels of *phishing* awareness.

Category	Precision	Recall	F1-Score	Support
Low	1.00	1.00	1.00	6
Moderate	1.00	1.00	1.00	8
High	1.00	1.00	1.00	6
Accuracy			1.00	20
Macro Average	1.00	1.00	1.00	20
Weighted Average	1.00	1.00	1.00	20

Based on these results, all categories were successfully classified. A *precision* value of 1.00 indicates that the predictions generated by the model were consistent with the actual categories. Meanwhile, a *recall* value of 1.00 indicates that all actual instances in each category were successfully identified by the model.

An *F1-score* of 1.00 indicates that the model achieved a good balance between *precision* and *recall*.

4.4 Random Forest Confusion Matrix

The evaluation results using the *confusion matrix* showed that all test instances were correctly classified according to their actual categories.

The resulting *confusion matrix* is presented as follows:

Actual/Prediction	**Low**	**Moderate**	**High**
Low	6	0	0
Moderate	0	8	0
High	0	0	6

Based on the confusion matrix results, all 6 instances in the Low category were correctly predicted. Similarly, all 8 instances in the Moderate category were classified according to their actual category. In addition, all 6 instances in the High category were correctly classified.

No misclassifications occurred. Therefore, the Random Forest evaluation results showed an *accuracy* of 100.00% on the test set.

4.5 Feature Importance Analysis

In addition to performing classification, the Random Forest algorithm can also indicate the relative importance of each feature in the decision-making process. Based on the evaluation results, the feature with the highest importance was Skor_Skenario, with a *feature importance* value of 0.2836.

The next most important feature was Rata_Rata_Perilaku, with a value of 0.2445. In addition, Waspada_Pesan_Mendesak also showed a relatively high contribution, with a *feature importance* value of 0.1157.

These results indicate that Skor_Skenario and Rata_Rata_Perilaku were the most dominant features in the Random Forest model for classifying students' levels of *phishing* awareness.

In contrast, features with lower *feature importance* values made relatively smaller contributions to the classification process within the dataset used in this study.

4.6 Comparison of Naive Bayes and Random Forest

Based on the evaluation results, both Naive Bayes and Random Forest were successfully applied to classify students' levels of awareness of *phishing* attacks.

Algorithm	Training Data	Test Data	Accuracy
Naive Bayes	80	20	100.00%
Random Forest	80	20	100.00%

Based on the table, both algorithms achieved the same accuracy of 100.00%. This indicates that, for the dataset and data split used in this study, both methods were able to correctly classify all test instances.

However, Random Forest provides an additional advantage by offering information on the relative importance of each feature through *feature importance*. Based on the evaluation results, Skor_Skenario and Rata_Rata_Perilaku were the features that contributed most substantially to the classification process.

Overall, both algorithms demonstrated very high performance in classifying students' levels of awareness of phishing attacks. Naive Bayes performed the classification using a probabilistic approach, whereas Random Forest relied on an ensemble of decision trees to determine the final classification results. However, the perfect scores should be interpreted cautiously because the test set contained only 20 observations. The result demonstrates performance on this particular split, not guaranteed generalizability to a broader student population. Future studies should use a larger and more diverse sample and repeated stratified cross-validation or an independent external test set.

Methodological limitation to be checked before final submission. The awareness category is described as being derived from scenario scores, while Skor_Skenario is also reported as the most important predictor in the Random Forest model. If this feature was used to create the target label, its inclusion as a predictor would cause target leakage and could explain the perfect test scores. The feature list and label-generation procedure must therefore be checked against the original dataset; if the feature contributed to the label, the models must be retrained without it and the results updated.

Conclusion

This study compared the performance of the Naive Bayes and Random Forest algorithms in classifying university students' levels of awareness of phishing attacks. The dataset consisted of 100 respondent records, divided into 80 training instances and 20 test instances, with three

awareness categories: Low, Moderate, and High. On the 20-observation test set, both algorithms achieved an accuracy of 100%, with precision, recall, and F1-score values of 1.00. These are results for this particular split only; the small test set limits conclusions about generalizability.

In addition, the Random Forest results indicated that Skor_Skenario and Rata_Rata_Perilaku were the most important features in the classification process. Overall, the findings demonstrate that both Naive Bayes and Random Forest can be used to classify students' levels of awareness of *phishing* attacks within the dataset examined in this study. Future research may employ larger and more diverse datasets to further evaluate the performance and generalizability of both algorithms.

References

[1] K. Okokpujie, C. G. Kennedy, K. Nnodu, and E. Noma-Osaghae, "Cybersecurity awareness: Investigating students' susceptibility to phishing attacks for sustainable safe email usage in academic environment (A case study of a Nigerian leading university)," International Journal of Sustainable Development and Planning, vol. 18, no. 1, pp. 255–263, 2023, doi: 10.18280/ijsdp.180127.

[2] H. Ruzaili, N. Katuk, K. M. Zaini, and W. Abdullah, "Phishing awareness and preventive measures among university students: Knowledge, behaviors, and victimisation perspectives," Millenium—Journal of Education, Technologies, and Health, 2025, doi: 10.29352/mill0229.43489.

[3] Malahayati and N. Nabilah, "Analisis tingkat kesadaran mahasiswa terhadap serangan rekayasa sosial (Studi kasus: Mahasiswa Teknologi Informasi Universitas Islam Negeri Ar-Raniry)," JINTECH: Journal of Information Technology, vol. 5, no. 1, pp. 1–12, 2024, doi: 10.22373/jintech.v5i1.4262.

[4] R. J. Nur'aini and M. Simanjuntak, "Phishing awareness and security concerns: Analyzing the role of anti-phishing knowledge and internet experience in online banking users," Jurnal Ilmu Keluarga dan Konsumen, vol. 18, no. 2, pp. 121–133, 2025, doi: 10.24156/jikk.2025.18.2.121.

[5] N. S. Ismail, F. I. F. Rikzan, N. Katuk, N. L. Hashim, and N. A. M. Zulkefli, "Enhancing information security awareness on phishing among IT students: A pilot test case study at Politeknik Tuanku Syed Sirajuddin," Journal of Digital System Development, 2023, doi: 10.32890/jdsd2023.1.2.

[6] G. M. W. Tangka and E. Y. Putra, "Classification of Indonesian undergraduate students' awareness level of phishing attacks using decision tree algorithm," Journal of Information System Research, vol. 6, no. 4, 2025, doi: 10.47065/josh.v6i4.7859.

[7] V. A. Windarni, A. F. Nugraha, S. T. A. Ramadhani, D. A. Istiqomah, F. M. Puri, and A. Setiawan, "Deteksi website phishing menggunakan teknik filter pada model machine learning," Information System Journal, vol. 6, no. 1, pp. 39–43, 2023, doi: 10.24076/infosjournal.2023v6i01.1268.

[8] J. Juremi and J. Kurniadi, "An email detection tool integrating Random Forest and Naive Bayes algorithms for phishing protection," Malaysian Journal of Cybersecurity and Applications, vol. 2, no. 1, 2026, doi: 10.22452/mjca.vol2no1.1.

[9] Z. Shahbazi, R. Jalali, and M. Molaeevand, "AI-based phishing detection and student cybersecurity awareness in the digital age," Big Data and Cognitive Computing, vol. 9, no. 8, art. 210, 2025, doi: 10.3390/bdcc9080210.

[10] M. A. Alqahtani, "Factors affecting cybersecurity awareness among university students," Applied Sciences, vol. 12, no. 5, art. 2589, 2022, doi: 10.3390/app12052589.

[11] L. Bognár and L. Bottyán, "Evaluating online security behavior: Development and validation of a personal cybersecurity awareness scale for university students," Education Sciences, vol. 14, no. 6, art. 588, 2024, doi: 10.3390/educsci14060588.

[12] A. Safi and S. Singh, "A systematic literature review on phishing website detection techniques," Journal of King Saud University—Computer and Information Sciences, vol. 35, no. 2, pp. 590–611, 2023, doi: 10.1016/j.jksuci.2023.01.004.

[13] L. J. Cronbach, "Coefficient alpha and the internal structure of tests," Psychometrika, vol. 16, pp. 297–334, 1951, doi: 10.1007/BF02310555.

[14] D. Jibat, S. Jamjoom, Q. Abu Al-Haija, et al., "A systematic review: Detecting phishing websites using data mining models," Intelligent and Converged Networks, vol. 4, no. 4, pp. 326–341, 2023, doi: 10.23919/ICN.2023.0027.

[15] E. Morrow, "Scamming higher ed: An analysis of phishing content and trends," Computers in Human Behavior, vol. 158, art. 108274, 2024, doi: 10.1016/j.chb.2024.108274.

[16] A. K. Gwenhure, "University students' security behavior against email phishing attacks: Insights from the health belief model," Journal of Cybersecurity, vol. 11, no. 1, art. tyaf034, 2025, doi: 10.1093/cybsec/tyaf034.

[17] B. Setiawan and M. A. Rizal, "Measurement of information security and privacy awareness in college students after the Covid-19 pandemic," Procedia Computer Science, vol. 234, pp. 1396–1403, 2024, doi: 10.1016/j.procs.2024.03.138.

[18] G. Rosta and L. Bottyán, "On the factors of students' cybersecurity behavior," PedActa, vol. 14, no. 1, pp. 58–66, 2024, doi: 10.24193/pedacta.14.1.6.

[19] M. Kawale, B. Maru, S. Dagu, M. Varghese, and V. Gupta, "Machine learning based phishing website detection," in Proc. 11th International Conference on Computing for Sustainable Global Development (INDIACom), 2024, pp. 833–837, doi: 10.23919/INDIACom61295.2024.10498854.

[20] L. Breiman, "Random forests," Machine Learning, vol. 45, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[21] A. Chandra, Gregorius, M. S. J. Immanuel, A. A. S. Gunawan, and Anderies, "Accuracy comparison of different machine learning models in phishing detection," in Proc. 5th International Conference on Information and Communications Technology (ICOIACT), 2022, pp. 24–29, doi: 10.1109/ICOIACT55506.2022.9972107.

[22] M. Bahaghighat, M. Ghasemi, and F. Özen, "A high-accuracy phishing website detection method based on machine learning," Journal of Information Security and Applications, vol. 77, art. 103553, 2023, doi: 10.1016/j.jisa.2023.103553.

[23] C. F. M. Foozy, M. A. I. Anuar, A. Maslan, H. A. M. Adam, and H. Mahdin, "Phishing URLs detection using Naive Bayes, Random Forest and LightGBM algorithms," International Journal of Data Science, vol. 5, no. 1, pp. 56–63, 2024, doi: 10.18517/ijods.5.1.56-63.2024.

[24] B. V. Pavani, D. Mahitha, and B. U. Maheswari, "Enhancing online safety: Phishing URL detection using machine learning and explainable AI," in Proc. 15th International Conference on Computing Communication and Networking Technologies (ICCCNT), 2024, pp. 1–6, doi: 10.1109/ICCCNT61001.2024.10723976.

[25] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.