Research Article

Performance Comparison of K-nearest Neighbor, Decision Tree, and Random Forest Methods for Classification of Cyber Defense Master Scholarship Recipients

[1,*]Dinisfusya'ban, [2]Bambang Suharjo, [3]Richardus Eko Indrajit

[1,2] Republic of Indonesia Defense University, Bogor, West Java, Indonesia
[3] Pradita University, Serpong, Banten, Indonesia

* Corresponding Author: dinisfuyaban@tp.idu.ac.id

Abstract: Cyber defense education is essential for developing a workforce capable of addressing evolving cyber threats, particularly in the military sector, where interconnected systems are vital for secure communication and command. This research aims to enhance the selection process for the Cyber Defense Master Scholarship at the Republic of Indonesia Defense University by employing machine learning algorithms. The study compares the performance of K-Nearest Neighbor (KNN), Decision Tree, and Random Forest for classifying eligible scholarship candidates. The results reveal a clear performance hierarchy: KNN achieves a moderate accuracy of 80.68%, offering simplicity and interpretability but lacking the precision of other models. The decision Tree performs with high accuracy (98.86%) but shows vulnerability to overfitting, which may impact generalizability to unseen data. Random Forest emerges as the most robust model, achieving the highest precision and overall stability, with minimal compromise on other metrics. Given the scholarship's selection requirements, Random Forest is recommended for tasks needing high accuracy and resilience against overfitting, while KNN and Decision Tree offer suitable alternatives for simpler, more interpretable applications.

Keywords: K-Nearest Neighbor; Decision Tree; Random Forest; Cyber Defense; Scholarship Recipients; Classification.

check for updates

Citation: Dinisfusya'ban, Suharjo., B, Indrajit., R.E. Performance Comparison of K-nearest Neighbor, Decision Tree, and Random Forest Methods for Classification of Cyber Defense Master Scholarship Recipients. Iota, 2024, ISSN 2774-4353, Vol.04, 04. https://doi.org/10.31763/iota.v4i4.823

Academic Editor : Adi, P.D.P

Received : September, 25 2024

Accepted : Oktober, 16 2024

Published : November, 05 2024

Publisher's Note: ASCEE stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Introduction

Cyberdefense education holds profound significance within the sphere of national defense, serving as a linchpin in safeguarding a nation's security, critical infrastructure, and sensitive data. In an era dominated by technological integration, the evolving threat landscape necessitates a skilled and informed cyber workforce. The interconnected nature of critical infrastructure, spanning sectors like energy, healthcare, and finance, underscores the importance of robust cyber defense measures. Moreover, as nation-states increasingly engage in cyber espionage to pilfer intellectual property and sensitive information, cyber defense education becomes instrumental in equipping individuals with the knowledge and skills needed to counter such threats (Das et al., 2023).

The military, reliant on interconnected systems for communication and command, is particularly susceptible to cyber-attacks. Cyber defense education becomes indispensable in preparing military personnel to navigate the digital battlefield, ensuring the integrity of communications, and thwarting potential cyber warfare threats. Furthermore, the economic security of a nation is intricately linked to its cyber defense posture. Cyber attacks can result in substantial economic losses, impacting businesses, financial institutions, and overall economic stability. Here, cyber defense education contributes to the development of a skilled workforce capable of securing digital assets, promoting innovation, and fostering a secure environment for economic growth (Mehanović & Kevrić, 2020).

In the realm of global cyber diplomacy, collaborative efforts are imperative, as cyber threats transcend national borders. Cyber defense education plays a pivotal role in shaping professionals who can engage in global cyber diplomacy, sharing best practices, collaborating on threat intelligence, and contributing to international efforts to combat cybercrime and cyber warfare. Beyond the professional sphere, cyber defense education extends to public awareness and resilience. Educating the general populace about cyber hygiene, safe online practices, and the potential risks of cyber threats enhances overall societal resilience. An informed citizenry is better equipped to identify and report potential threats, contributing to the collective defense against cyber attacks (Sengsri & Khunratchasana, 2023).

The objectives of this research aim to enhance the accuracy and efficiency of the selection process for the Cyber Defense Master Scholarship by applying and comparing three machine learning algorithms namely K-Nearest Neighbor (KNN), Decision Tree, and Random Forest. Through a comparative analysis, this research seeks to provide insights into the strengths and limitations of each algorithm, enabling more precise, effective, and context-specific model selection tailored to the scholarship selection process requirements.

2. Theory

One of the machine learning models is to classify objects that have certain criteria. One of the applications of the use of this machine learning model is in terms of scholarship acceptance, the use of classification algorithms to improve selection accuracy. These models analyze historical data to identify key attributes that affect scholarship eligibility, streamlining the selection process.

KNN, Decision Trees, and Random Forests have been effectively utilized for classifying. Research indicates that Random Forest models often outperform others in accuracy, achieving up to 87% in career placement predictions (Hendri et al., 2024) and demonstrating strong performance in financial aid eligibility assessments (Ismail et al., 2024). Decision Trees serve as a reliable baseline, while KNN provides competitive results, particularly in educational contexts (Hajar et al., 2022) (Basha et al., 2023). In a comparative study, Gradient Boosting Classifiers surpassed other models, achieving an accuracy of 96% in predicting graduate admissions (Basha et al., 2023). These findings highlight the importance of selecting appropriate algorithms based on the specific dataset and classification goals, emphasizing the potential of ensemble methods in enhancing predictive accuracy in scholarship acceptance scenarios.

3. Method

The four stages of the research approach are as follows the Figure 1.

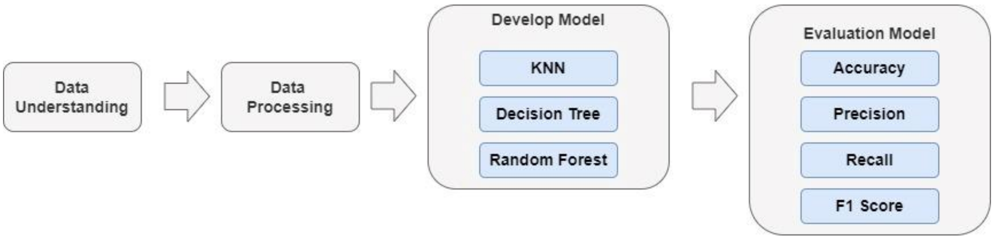

Figure 1. Research Method

3.1 Data Understanding

Data understanding encompasses initial insights into unknown datasets, requiring thorough documentation to grasp the data's meaning, it is essential for predictive modeling, where labeled data reflects domain understanding (Hajar et al., 2022) (Bhatnagar et al., 2012). The phase of data understanding in this research is foundational to comprehensively grasp the intricacies of the dataset and lay the groundwork for subsequent analyses. It encompasses a thorough exploration of the collected data sources, focusing on academic achievements, research experience, and indicators of aptitude in cyber defense. Through a detailed examination, the researchers seek to discern patterns, trends, and potential challenges inherent in the dataset. This involves identifying the distribution of values within each feature, detecting any outliers or anomalies that may impact the accuracy of the machine learning models, and gaining insights into the interrelationships among different variables (Chauhan et al., 2021). In addition, the researchers want to evaluate how well the dataset captures the range of qualifications held by candidates for the Cyber Defense Master Scholarship. The phase of data comprehension plays a crucial role in guiding the latter stages of data cleaning and preparation. This ensures that the dataset is optimized and customized to meet the unique needs of the classification task. The project seeks to improve the validity and reliability of the ensuing analyses by developing a thorough grasp of the data. This will ultimately improve the efficacy of the machine learning algorithms in finding qualified candidates for the Cyber Security Master Scholarship.

Table 1. Sample of Scholarship Recipients Dataset

ID	Male/ Female	B's Deg	GPA	TPA	TKBI	Psy	Score (total feat weight)	Result
1	Male	no	3.69	670	485	79	28536	0
2	Male	yes	2.65	707	485	50	28706	0
3	Male	no	3.07	534	455	85	24658	0
4	Female	yes	3.90	517	556	73	25998	1
5	Female	yes	3.23	669	464	58	27574	0
6	Female	yes	3.13	590	585	72	28363	1
7	Male	no	3.38	663	531	83	29393	0
8	Female	no	3.51	593	348	69	23614	0
9	Male	no	3.80	450	543	78	24145	0
10	Male	yes	3.36	414	373	64	19514	0

3.2 Data Pre-Processing

Data pre-processing refers to the systematic techniques applied to prepare raw data for analysis, ensuring its quality and suitability for model training. This process encompasses various activities, including data cleaning, which involves correcting errors, handling missing values, and removing duplicates (Gupta et al., 2025) (Kale & Pandey, 2024). Additionally, it addresses issues such as data imbalance and noise, which can significantly impair model performance (Masood & Begum, 2024). Pre-processing may involve feature extraction and transformation to enhance the dataset's relevance and fairness, ultimately leading to improved accuracy and effectiveness of machine learning algorithms (Charpentier, 2024) (Zhao et al., 2023)

This phase is a crucial step in refining and preparing the dataset for machine learning algorithms to classify Cyber defense Master Scholarship recipients. This procedure entails encoding categorical variables, standardizing numeric characteristics, and carefully cleaning data to address inconsistent or missing values to convert qualitative data into a format that machine learning algorithms can understand. Algorithms such as KNN, Decision Tree, and Random Forest can process and understand the data efficiently because of this encoding process. To address potential issues and guarantee the best possible performance of machine learning models in the classification of Cyber defense Master Scholarship recipients, the preprocessing step attempts to provide a refined and harmonized dataset.

3.3 Develop Model

In machine learning, a model is defined as a mathematical representation that maps input data to output predictions or classifications. This mapping can take various forms, from simple linear regressions to complex neural networks. The model learns from data, improving its predictions over time without explicit programming, which distinguishes it from traditional programming paradigms (Ghosh & Dasgupta, 2022) (Palaparthi, 2023).

The development of the machine learning model for classifying Cyber defense Master Scholarship recipients is a pivotal phase in this research, marked by the implementation and optimization of three prominent algorithms:

3.3.1 K-Nearest Neighbor (KNN)

The KNN algorithm, a well-liked machine learning method for classification and regression applications, was presented by (Christopher, 2021). To make predictions or judgments on data points that are not visible, KNN finds the k-nearest data points inside a dataset. Applications for this straightforward yet effective method include medical diagnosis, recommendation systems, and picture identification.

Step 1: Decide which neighbor's number is K. The K value represents the number of closest neighbors.

Step 2: Determine the K number of neighbors' Euclidean distance. The distance between two points is known as the Euclidean distance, which we have already covered in geometry. Euclidean Distance between A and B = $\sqrt{(X_2 - X_1)^2 + (Y_2 - Y_1)^2}$, is illustrated in Figure 2.

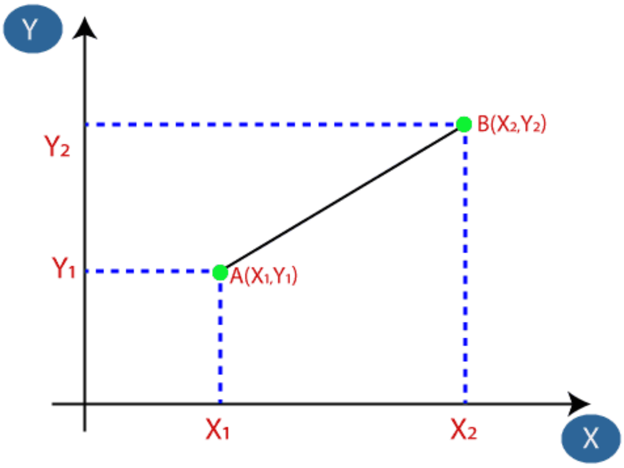

Figure 2. Euclidean Distance Formula

Step 3: Choose the K nearest neighbors based on the calculated Euclidean distance.
Step 4: Determine how many data points each of these k neighbors has in each category.
Step 5: Put the recently obtained data points in the category with the highest neighbor count.
Step 6: The model is prepared.

Several benefits of the KNN method include its interpretability, simplicity, and capacity to handle data that is not linearly separable. Its high processing cost, sensitivity to the distance metric selected and the value of k, and unsatisfactory performance with huge datasets or high-dimensional data are some of its drawbacks (Murphy, 2012).

3.3.2 Decision Tree

A popular machine learning model is the decision tree, which provides a graphical depiction of choices and their potential outcomes, such as utility, resource costs, and chance event outcomes. It works especially well for classification tasks where the branches represent decision rules based on the input features and each leaf node represents a class label (Lombardi et al., 2017).

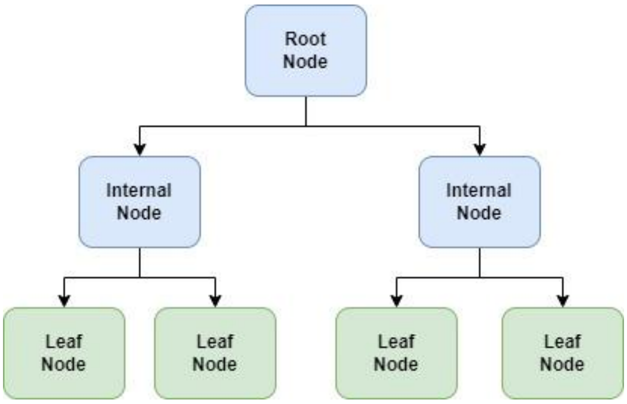

Figure 3. Decision Tree

Decision tree models choose the optimum characteristic at each node using two methods: Gini impurity and information gain. By assessing the caliber of every test condition, these techniques group samples into classes. Entropy is a notion from information theory that quantifies sample values' impurity and is defined by the following equation 1.

$$\text{Entropy (S)} = -\sum_{c\in C} p(c)log_2 p(c) \quad (1)$$

Moreover, the variables S and c represent the data set and classes that comprise the set, respectively, and p(c) represents the ratio of class c data points to all of the data points in the set S. Entropy values, measuring uncertainty in a dataset, range between 0 and 1. A dataset where all samples belong to one class yields zero entropy, while an equal split between two classes results in maximum entropy at 1. To construct an optimal decision tree, the attribute with the smallest entropy is chosen for splitting, minimizing uncertainty in the resulting subsets. Information gain, representing the reduction in entropy post-split, determines the attribute's effectiveness. The attribute with the highest information gain is prioritized for the split, ensuring accurate classification of training data based on target classes.

The information gain formula, Information Gain is equal to Entropy before splitting minus Weighted average entropy after splitting, quantifies the improvement in classification purity. This systematic approach enables the selection of the most informative attributes, contributing to effective decision tree construction. By prioritizing features with minimal entropy and maximal information gain, decision trees efficiently classify data, enhancing their utility in various applications, including the selection of optimal scholarship candidates in education contexts, where the Information Gain formula is shown in Equation 2.

$$\text{Information Gain (S, } a\text{)} = Entropy(S) - \sum_{c\in C} \frac{|S_v|}{|S|} Entropy(S_v) \quad (2)$$

- A stands for a particular characteristic or class label.
- The entropy of dataset S is denoted by Entropy(S).
- The ratio of values in Sv to values in the dataset, S, is denoted by the expression |Sv|/ |S|.
- Entropy (Sv) represents the dataset's entropy, Sv.

3.3.3 Random Forest

In machine learning, a random forest is an ensemble learning technique that is mostly employed for classification and regression tasks. During training, it builds several decision trees and outputs the mean prediction for regression tasks or the mode of their classification predictions. This approach enhances predictive accuracy and controls overfitting compared to individual decision trees (Kuptsova & Ramazanov, 2020).

As its name suggests, a Random Forest is a tree-based ensemble approach in which every tree in the ensemble is dependent upon a set of random variables. The real-valued response is denoted by a random variable Y, whereas the actual values input or predictor factors are expressed by a p-randomized vector X = (X1,..., Xp)T. The fundamental premise of the study is that the correlations between the predictor and response variables are determined by a new joint distribution, PXY (X, Y).

Essentially, several decision trees are generated to create the Random Forest ensemble. Every tree adds to the overall forecast that the Random Forest makes. Formulating a prediction function, f(X), that accurately predicts the true-valued response variable Y is the main goal of the model. The model attempts to minimize the predicted value of the loss in the loss function, denoted as L (Y, f(X)), which is used to evaluate the effectiveness of this prediction function. By ensuring that the anticipated and actual values nearly match, this optimization procedure improves the Random Forest model's accuracy and dependability.

This model's reliance on a set of random variables introduces an element of diversity within the ensemble, contributing to its robustness and mitigating overfitting. By aggregating predictions from multiple decision trees, the Random Forest leverages ensemble learning to create a more stable and accurate overall model. This approach enhances the model's generalization capabilities, making it well-suited for predicting real-valued responses within the framework of the unknown joint distribution. Overall, the Random Forest model stands as a versatile and powerful tool in predictive modeling, finding applications in diverse domains, including but not limited to, the classification of Cyber defense Master Scholarship recipients.

$$E_{XY}\,(L\;(Y, f(X\;))) \tag{3}$$

Where the expectation regarding the joint distribution of X and Y is indicated by the subscripts. The fact that the L (Y, f(X)) as equation 3, penalizes f(X) values that are far from Y makes sense as a measure of the proximity of f(X) to Y. Squared error loss is a typical L selection. For regressed with zero-one loss for classification, L (Y, f(X)) = (Y −f (X))2 as equation 4.

$$L\ (Y, f(X)) = I\ (Y\ /= f(X)) = 0, \text{ if } Y = f(X), 1 \text{ otherwise} \tag{4}$$

The conditional expectation can be obtained by minimizing EXY (L(Y, f(X))) for squared error loss as the equation 5.

$$f(x) = E(Y\ |X = x) \tag{5}$$

Sometimes known as a regression function. When minimizing EXY (L(Y, f(X))) with zero-one loss in the classification scenario, where the set of potential values of Y is represented by Y, the result is equation 6.

$$f\ (x) = \arg\max P(Y = y \mid X = x),\ y \in Y \tag{6}$$

Moreover, the Illustration of a Random Forest can be seen in the illustration in Figure 4.

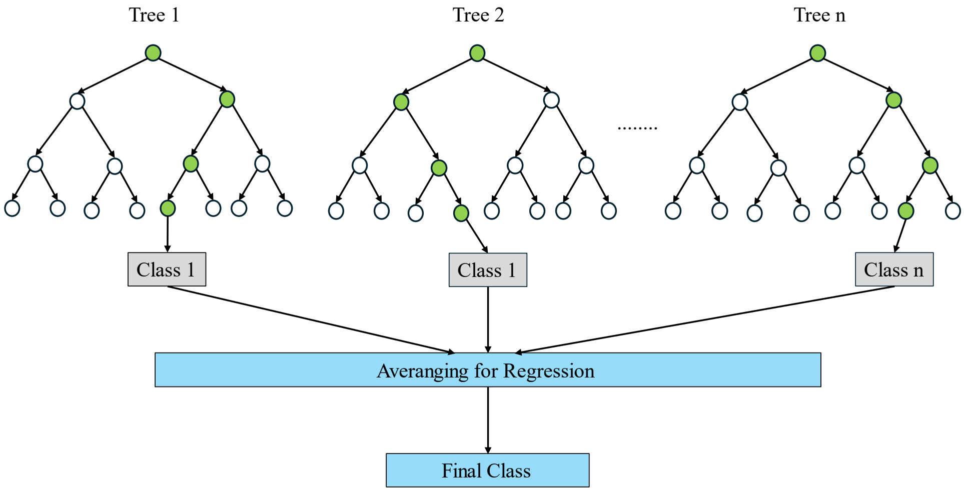

Figure 4. Illustration of Random Forest

3.3.4 Evaluation Model

At this point, performance model calculations on KNN, Decision Tree, and Random Forest are performed on the algorithm models used in the learning classification approach. The classification model's performance is calculated by testing the true and incorrect objects. In this study, the classification performance calculation is a confusion matrix, which incorporates calculations of predictable real classification results (Kasanah et al., 2019). Accuracy, precision, recall, and F1-score are frequently used metrics in machine learning to assess how well classification models perform. While precision and recall offer information about the model's performance on positive class predictions, accuracy gauges the model's overall correctness. The F1-score is a harmonic mean of the two, balancing the trade-offs between the two (Geng, 2024) (Vickers et al., 2023). Table 2 depicts two class matrix confusions.

Table 2. Confusion Matrix

	Positive Prediction	**Negative Prediction**
Positive Actual	TP	FN
Negative Actual	FP	TN

Table 2 displays TP (True Positive), TN (True Negative), FP (False Positive), and FN (False Negative). The four classification performance metrics in this study are recall, accuracy, precision, and F1 score.cThe calculation formulas for accuracy, precision, recall, and F1 score are shown in equations 7, 8, 9, and 10.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{7}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{8}$$

$$\text{Recall} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN}} \tag{9}$$

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{10}$$

3.3.5 Pseudocode

In this research, researchers write a program with phyton to compare the model. The program first loads a dataset from the file dataset-fix.csv into a DataFrame. Once the data is loaded, it separates the data into features (X) and targets (y), where the features include the columns Bachelor's Degree, GPA, TPA, TKBI, and psy, while the column Hasil serves as the target. Next, the categorical feature Bachelor's Degree is encoded using LabelEncoder to make it suitable for modeling. The program then splits the data into training and testing sets, with 80% used for training and 20% for testing, ensuring the test data is kept separate during model training. Three classification models are built: K-Nearest Neighbors (KNN) with n_neighbors=5, Decision Tree, and Random Forest with 100 estimators. Each model is trained on the training data, followed by predictions on the test data. After making predictions, the program calculates evaluation metrics (accuracy, precision, recall, and F1 score) for each model. These metrics are stored in dictionaries for each model: knn_metrics, dt_metrics, and rf_metrics. All evaluation results are then combined into a data frame, metrics_df, for easy comparison of model performance. Finally, the data frame with the comparison of model performance is displayed as the output. The pseudocode of the program can be seen in Pseudocode 1.

```
LOAD dataset from 'dataset-fix.csv' into DataFrame `data`
#Separate Features and Target
SET `X` = ['Bachelor's Degree', 'GPA', 'TPA', 'TKBI', 'psy']
SET `y` = 'Hasil'
#Encode Categorical Feature
INITIALIZE `LabelEncoder` as `le`
TRANSFORM `X['Bachelor's Degree]` using `le`
#Split the dataset into training and testing sets
SET `X_train`, `X_test`, `y_train`, `y_test` = train_test_split(X, y, test_size=0.2, random_state=42)
#Initialize and Train Models
#KNN Model
INITIALIZE `KNeighborsClassifier` with `n_neighbors=5` as `knn`
TRAIN `knn` on `X_train` and `y_train`
PREDICT with `knn` on `X_test`
CALCULATE classification report for `knn`
STORE `knn` metrics (accuracy, precision, recall, F1 score) in `knn_metrics`
#Decision Tree Model
INITIALIZE `DecisionTreeClassifier` with `random_state=42` as `dt`
TRAIN `dt` on `X_train` and `y_train`
PREDICT with `dt` on `X_test`
CALCULATE classification report for `dt`
STORE `dt` metrics (accuracy, precision, recall, F1 score) in `dt_metrics`
#Random Forest Model
INITIALIZE `RandomForestClassifier` with `n_estimators=100` and `random_state=42` as `rf`
TRAIN `rf` on `X_train` and `y_train`
PREDICT with `rf` on `X_test`
CALCULATE classification report for `rf`
STORE `rf` metrics (accuracy, precision, recall, F1 score) in `rf_metrics`
#Combine Results
CREATE DataFrame `metrics_df` from `knn_metrics`, `dt_metrics`, and `rf_metrics`
DISPLAY `metrics_df`
```

------------ Pseudocode of Model Comparison -------------

4. Result and Discussion

The experimental results of the research are pivotal in evaluating the efficacy of each algorithm in scholarship selection. Below is a detailed analysis of the experimental results, including the confusion matrix for each performance metric.

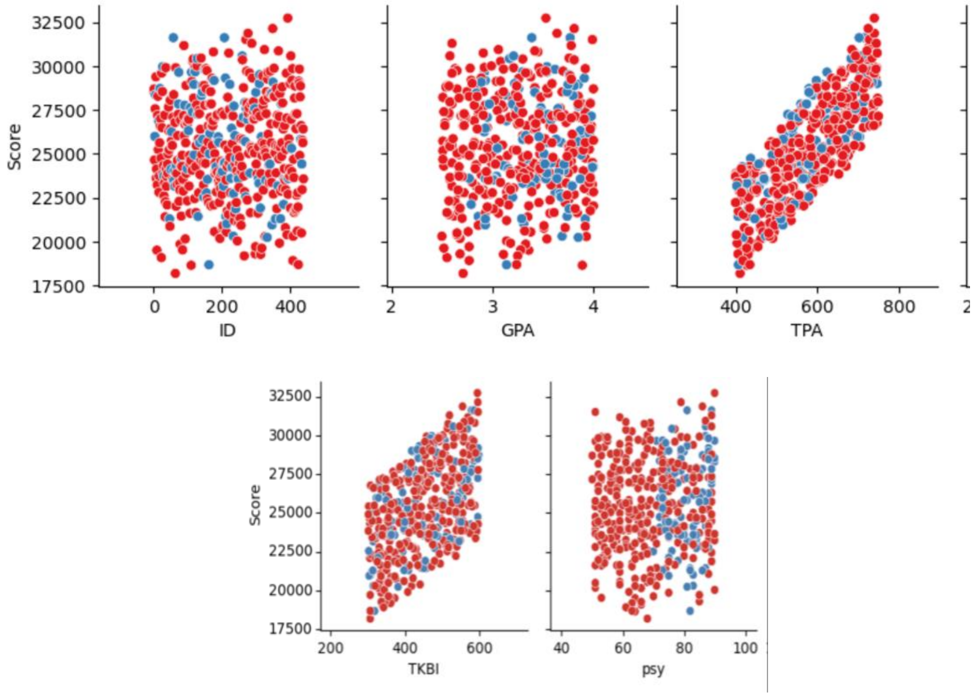

Figure 5. Prediction Data Distribution

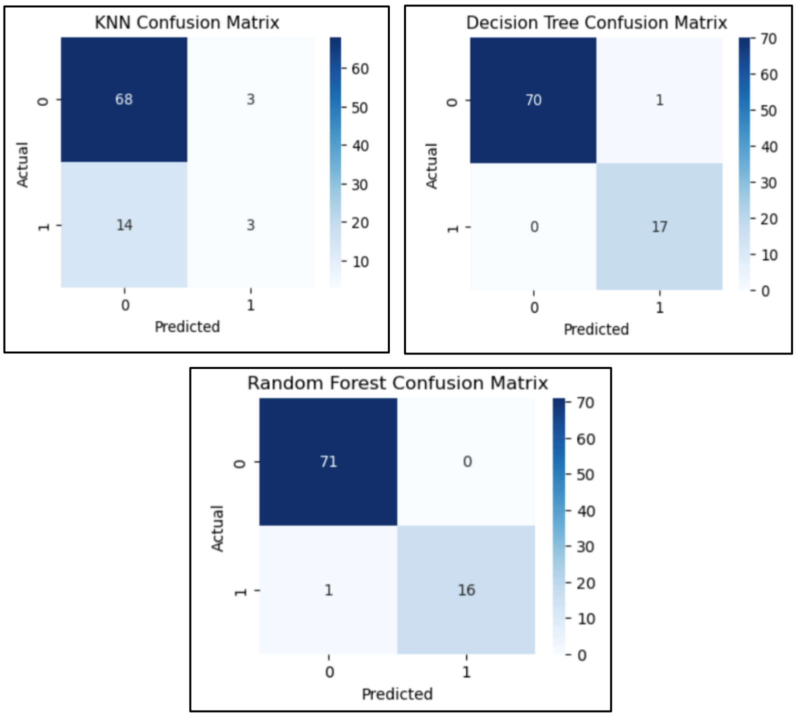

Figure 6. Confusion Matric of Algorithm

Performance measures give a thorough evaluation of the models' capacity to accurately identify worthy candidates. These metrics include accuracy, precision, recall, and F1 score. The detailed Performance Comparison model can be seen in Table 3.

Table 3. Model Performance Comparison

Model	**Accuracy**	**Precision**	**Recall**	**F1 Score**
KNN	0.806818	0.765660	0.806818	0.767567
Decision Tree	0.988636	0.989268	0.988636	0.988758
Random Forest	0.988636	0.988794	0.988636	0.988504

The results showed a clear hierarchy in performance, with KNN achieving an acceptable accuracy of 80.68% but falling behind other models due to its complexity. Decision Tree outperformed KNN with an impressive accuracy of 98.86%, but its susceptibility to overfitting and instability raises concerns for generalization to unseen data. Random Forest, leveraging ensemble learning, emerged as the most robust model, with marginally higher precision (0.05%) compared to Decision Tree and minimal sacrifice in other metrics. This superior accuracy and ability to mitigate overfitting make Random Forest the most reliable choice for tasks demanding high precision and robust performance.

5. Conclusion

The classification of Cyber defense Master Scholarship recipients using machine learning algorithms, i.e., K-Nearest Neighbor, Decision Tree, and Random Forest, have provided valuable insights into the strengths and weaknesses of each approach. The findings reveal that Random Forest is the most reliable and accurate model for classifying recipients of the Cyber Defense Master Scholarship, outperforming Decision Tree in its resilience to overfitting, even though Decision Tree shows high accuracy. KNN, while less accurate, remains a viable option for contexts that prioritize simpler interpretation. Therefore, Random Forest is recommended for selection processes requiring high accuracy and complex decision-making, while KNN or Decision Tree serve as alternatives in contexts where transparency and interpretability are prioritized.

Acknowledgments: With utmost gratitude, we would like to express our sincere appreciation to The Republic of Indonesian Defense University for their invaluable support throughout the entire process of this research, from inception to publication. The assistance and resources they provide play an important role in facilitating the successful completion of these studies.

Author contributions: The authors were responsible for building Conceptualization, Methodology, analysis, investigation, data curation, writing—original draft preparation, writing—review and editing, visualization, supervision of project administration, funding acquisition, and have read and agreed to the published version of the manuscript.

Funding: The study was conducted without any financial support from external sources.

Availability of data and Materials: All data are available from the authors.

Conflicts of Interest: The authors declare no conflict of interest.

Additional Information: No Additional Information from the authors.

References

1. Basha, M. S. A., Prabhavathi, C., Khangembam, V., Sucharitha, M. M., & Oveis, P. M. (2023). Predicting Graduate Admissions using Ensemble Machine Learning Techniques: A Comparative Study of Classifiers and Regressors. 2023 2nd International Conference for Innovation in Technology (INOCON), 1–6. https://doi.org/10.1109/INOCON57975.2023.10101206
2. Bhatnagar, V., Dobariyal, R., Jain, P., & Mahabal, A. (2012). Data Understanding using Semi-Supervised Clustering. 2012 Conference on Intelligent Data Understanding, 118–123. https://doi.org/10.1109/CIDU.2012.6382192
3. Charpentier, A. (2024). Pre-processing (pp. 385–396). https://doi.org/10.1007/978-3-031-49783-4_10
4. Chauhan, D., Walia, R., Singh, C., Deivakani, M., Kumbhkar, M., & Professor, A. (2021). Detection of Maize Disease Using Random Forest Classification Algorithm. In Turkish Journal of Computer and Mathematics Education (Vol. 12, Issue 9).
5. Christopher, A. (2021). K-Nearest Neighbor. Https://Medium.Com. https://medium.com/swlh/k-nearest-neighbor-ca2593d7a3c4
6. Das, S., Bhattacharyya, K., & Sarkar, S. (2023). Performance Analysis of Logistic Regression, Naive Bayes, KNN, Decision Tree, Random Forest, and SVM on Hate Speech Detection from Twitter. International Research Journal of Innovations in Engineering and Technology, 07(03), 07–03. https://doi.org/10.47001/IRJIET/2023.703004
7. Geng, S. (2024). Analysis of the Different Statistical Metrics in Machine Learning. Highlights in Science, Engineering and Technology, 88, 350–356. https://doi.org/10.54097/jhq3tv19
8. Ghosh, S., & Dasgupta, R. (2022). Introduction to the Machine Learning Models. In Machine Learning in Biological Sciences (pp. 45–50). Springer Nature Singapore. https://doi.org/10.1007/978-981-16-8881-2_4
9. Gupta, P., Sehgal, N. K., & Acken, J. M. (2025). Practical Aspects in Machine Learning (pp. 281–330). https://doi.org/10.1007/978-3-031-59170-9_9
10. Hajar, M., Adil, J., Ali, Y., & Khalid, A. Z. (2022). Predicting Student Success in a Scholarship Program (pp. 333–341). https://doi.org/10.1007/978-3-031-02447-4_35
11. Hendri, M. N., Agung, B. H., Ali, M., & Ganda, W. (2024). Model Klasifikasi Machine Learning untuk Prediksi Ketepatan Penempatan Karir. Jurnal SAINTEKOM, 14(1), 13–25. https://doi.org/10.33020/saintekom.v14i1.512
12. Ismail, M. H., Razak, T. R., Noor, N. M., & Aziz, A. A. (2024). Evaluating Machine Learning Algorithms for Predicting Financial Aid Eligibility: A Comparative Study of Random Forest, Gradient Boosting and Neural Network. 2024 18th International Conference on Ubiquitous Information Management and Communication (IMCOM), 1–6. https://doi.org/10.1109/IMCOM60618.2024.10418450
13. Kasanah, A. N., Muladi, M., & Pujianto, U. (2019). Penerapan Teknik SMOTE untuk Mengatasi Imbalance Class dalam Klasifikasi Objektivitas Berita Online Menggunakan Algoritma KNN. Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi), 3(2), 196–201. https://doi.org/10.29207/resti.v3i2.945
14. Kuptsova, & Ramazanov. (2020). Analysis of artificial neural networks training models for airfare price prediction. Artificial Intelligence, 25(3), 45–50. https://doi.org/10.15407/jai2020.03.045
15. Lombardi, M., Milano, M., & Bartolini, A. (2017). Empirical decision model learning. Artificial Intelligence, 244, 343–367. https://doi.org/10.1016/j.artint.2016.01.005
16. Masood, S. W., & Begum, S. A. (2024). Data Collection and Pre-processing for Machine Learning-Based Student Dropout Prediction (pp. 355–367). https://doi.org/10.1007/978-981-99-3481-2_28
17. Mehanović, D., & Kevrić, J. (2020). Phishing Website Detection Using Machine Learning Classifiers Optimized by Feature Selection. Traitement Du Signal, 37(4), 563–569. https://doi.org/10.18280/ts.370403
18. Murphy, K. P. . (2012). Machine learning : a probabilistic perspective. MIT Press.
19. Palaparthi, A. (2023). Machine learning for voice and speech science. NCVS Insights, 1(1). https://doi.org/10.62736/ncvs189226
20. Prof. Arati K Kale, & Dr. Dev Ras Pandey. (2024). Data Pre-Processing Technique for Enhancing Healthcare Data Quality Using Artificial Intelligence. International Journal of Scientific Research in Science and Technology, 299–309. https://doi.org/10.32628/IJSRST52411130
21. Sengsri, S., & Khunratchasana, K. (2023). Comparison of machine learning algorithms with regression analysis to predict the COVID-19 outbreak in Thailand. Indonesian Journal of Electrical Engineering and Computer Science, 31(1), 299. https://doi.org/10.11591/ijeecs.v31.i1.pp299-304
22. Vickers, P., Barrault, L., France, M. A., Monti, E., & Aletras, N. (2023). We Need to Talk About Classification Evaluation Metrics in NLP. The Association for Computational Linguistics, 1, 498–510. https://github.com/petervickers/
23. Zhao, J., Chong, K.-S., Shu, W., & Chang, J. (2023). A Data Pre-Processing Module for Improved-Accuracy Machine-Learning-based Micro-Single-Event-Latchup Detection. 2023 IEEE 9th International Conference on Space Mission Challenges for Information Technology (SMC-IT), 1–6. https://doi.org/10.1109/SMC-IT56444.2023.00009