Comparison of Supervised Learning Methods for Spatial User Clustering in Downlink NOMA

Hurianti Vidyaningtyas[1,2*], **Iskandar**[1], **Hendrawan**[1], **Aloysius Adya Pramudita**[2]
[1]Bandung Institute of Technology,ITB
[1]Jl. Ganesa, Bandung, Indonesia
[2]University Center of Excellence for Intelligent Sensing-IOT, Telkom University
[2]Jl. Telekomunikasi, Bandung, Indonesia
Email*: huriantividya@telkomuniversity.ac.id

ARTICLE INFORMATION

Received on 21 February 2024
Revised on 05 May 2024
Accepted on 21 June 2024

Keywords:
Clustering
PDNOMA
Supervised learning
Spatial
NOMA

ABSTRACT

The performance of Power Domain Non-Orthogonal Multiple Access (PD-NOMA) is affected by the performance of Successive Interference Cancellation (SIC) in decoding user data. The large number of users will cause error propagation in SIC, which results in decreased SIC performance. This research aims to optimize the performance of SIC in PD-NOMA by applying spatial concepts to classify users. This research applies various supervised machine learning classification algorithms, including Decision Tree, K-Nearest Neighbors (K-NN), Support Vector Machine (SVM), Random Forest, Logistic Regression, and Naive Bayes. The experimental results show that Random Forest achieves the highest accuracy in classifying users, followed by Decision Tree. In addition, in performance measurement using ROC (Receiver Operating characteristic) and AUC (Area under the Curve) curves, the Random Forest method achieved the best results. In terms of experimentation process time, a decision tree has a faster time compared to a random forest. Overall, the Random Forest algorithm is suitable for the task of user clustering in the context of PD-NOMA, which utilizes the spatial concept from user to base station (BS).

1. Introduction

Power Domain Non-Orthogonal Multiple Access (PD-NOMA) is one of the non-orthogonal multiple access techniques that utilizes power differences in the user access process. In this research, PD-NOMA will be referred to as non-orthogonal multiple access (NOMA). NOMA is one of the proposed technologies for 5G and beyond to address the challenges of increasing data rates, massive connectivity, very low latency, and reliable communication (Benjebbour, 2017; Islam et al., 2017; Song et al., 2017). NOMA is a promising solution for wireless communication systems where multiple users share the same time-frequency resources. NOMA can increase the effectiveness of orthogonal access systems due to its compatibility with other multiple access methods. NOMA also has capabilities in terms of spectral efficiency and adaptability in optimizing system resources (Islam et al., 2017). However, due to the strong co-channel interference among mobile users introduced by NOMA, it poses significant challenges for system design and resource management (Song et al., 2017).

In the NOMA system (Higuchi, 2015), each user is allocated power based on the distance to the BTS. This system uses Superposition Coding (SC) at the transmitter to combine multiple users and uses Successive Interference Cancellation (SIC) at the receiver side to separate the user signals. This system is simpler in its detection and decoding process compared to other non-orthogonal multiple access systems. However, if the number of users is large, the interference between users will also be greater. This is shown in research (Vidyaningtyas et al., 2023, 2024) discussing the optimal maximum number of users in NOMA using sequential power allocation by looking at sum rate and BER performance. Therefore, system management is needed to reduce interference in order to maximize user performance in NOMA.

One system design done to reduce interference is user grouping. In Kang & Kim (2018), users are paired based on the user's channel gain value. Strong users in the first group are sequentially swapped with weak users

in the second group until the last group if the resulting sum rate value increases. This user pairing principle was also developed by Zhu et al. (2019). Users who have the largest channel gain will be paired with users who have the smallest channel gain. The user who has the second-largest channel gain will be paired with the user who has the second-smallest channel gain, and so on. User pairing research by iterating to find users who are close in space but have large channel gain differences is also discussed in research (Bui et al., 2019). The user pairing principle is developed into a user grouping that can consist of more than two users. User grouping using genetic algorithms provides good grouping results. (You et al., 2020). However, the disadvantage of this genetic algorithm is that it takes a long time to find the optimal user group. In Prabha Kumaresan et al. (2020), user grouping uses Artificial Neural Network (ANN) classification. In this study, the number of users for each cluster is unknown. In addition, there is no further study regarding the performance of NOMA in the resulting group.

User grouping using reinforcement learning in NOMA power allocation is reviewed in the paper (S. Rezwan, 2020), where the research focuses on allocating power efficiently to maximize the amount of data in the NOMA system using reinforcement learning elaborated with a user grouping algorithm. User grouping is done by dividing into several areas based on the distance to the BS. Furthermore, users with the largest channel gain in each area are grouped into one group. This allows users in one group to be in a position that is physically far apart from each other.

Based on previous research, most of the research related to user grouping using machine learning does not pay attention to the position of users based on the signal direction (Ding et al., 2016) from the base station (BS), thus allowing interference between users from different groups. Such interference can cause large propagation errors, so that the signal decoding process at the receiver side becomes more difficult. This will result in high error values in the data received by the user. Therefore, this research performs user grouping using user channel gain and spatial user position to minimize interference between users from different groups. User grouping uses several supervised machine learning algorithms, namely Decision Tree, K-NN, Support Vector Machine (SVM) Random Forest, Logistic Regression, and Naive Bayes.

This paper consists of five sections. The first section is an introduction regarding grouping users using machine learning on the NOMA system, which has been studied previously. Section II presents an illustration of user grouping in NOMA and the supervised learning algorithm that will be used. Section III is dedicated to outlining the research methodology employed in this study. Experimental results and discussion are discussed in Section IV. The final section is the conclusion and future network.

2. Supervised Learning for User Grouping on NOMA

2.1 User Grouping with Spatial Concept in NOMA

The downlink NOMA system is the process of sending data from the base station (BS) to the user equipment (UE). In the NOMA system, the amount of power allocated for each user determines the receiver's ability to decode user data. The determination of power allocation is based on the user's distance to the BS. On the receiving side, SIC is used to separate the signals from all users. SIC works by carrying out a signal reduction process that has large power so that it can then process the user's signal. If the number of users is very large, the SIC work process will take a long time and allow error propagation to occur. Therefore, user grouping is carried out so that SIC can work optimally.

User clustering in NOMA is performed following an aligned user pattern (Ding et al., 2016). Users located in the same direction will minimize interference from users in other groups. This is achieved through spatial concepts using beamforming. Additionally, users within the same group must have varying distances from the base station (BS). Power allocation is based on the user's distance from the BS, with greater power allocated to users farther from the BS. An illustration of user clustering is presented in Figure 1.

In Figure 1, the grouping of users in one cell is divided into N spatial areas, namely beam 1, beam 2, beam 3, up to beam N. Each beam is divided into M areas that describe the different distances of users to the BS. In this research, the number of N and M is designed as 8 beams and 4 distance areas.

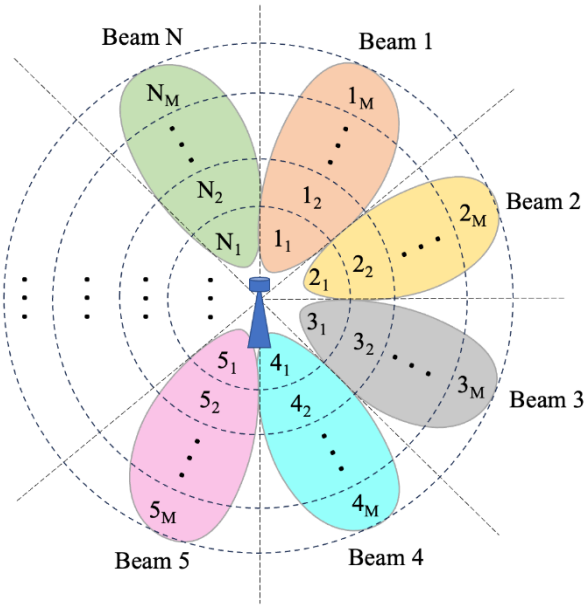

Figure 1. Illustration of user grouping using the spatial concept

2.2 Supervised Learning Algorithm for User Classification

According to Oladipupo (2010), there are various types of supervised machine learning algorithms for classification. In this research, the supervised machine learning algorithms used in relation to user grouping classification include Logistic Regression, Naive Bayes Classifier, Support Vector Machine, Decision Tree, Random Forest (RF), and K-Nearest Neighbors (K-NN).

a. Logistic Regression

Logistic regression states the existence of boundaries between classes and the probability of a class depending on the distance from that boundary, in a certain approximation. It moves towards the extremes (0 and 1) faster when the data set is larger. This statement about probabilities is what makes logistic regression more than just a classifier.

b. K-Nearest Neighbors (KNN)

KNN uses distance measurements (distance metrics), such as Euclidean distance, to calculate how similar a new user is to existing users in the data set. This distance will be used to determine the K closest users. The Euclidean distance formula can be written as follows:

$$\text{dist}(a_1,a_2)= \sqrt{\sum_{i=1}^{n}(a_{1i}\text{-}a_{2i})^2} \quad \text{.....} \quad 1)$$

$$dist = \sqrt{\sum_{i=1}^{n}(a_{1i} - a_{2i})^2 + (b_{1i} - b_{2i})^2 + \cdots} \quad \text{.....} \quad 2)$$

The formula (1) can be used if there is only one independent variable. However, if the number of independent variables is more than one, the difference in distance between the variables is added up using the formula (2).

c. Naïve Bayes

Naive Bayes uses Bayes' theorem to calculate class probabilities based on input features (Rish I, 2001). In this research, this event can represent the user area among the 32 classified areas. The general formula for Bayes' theorem is as follows:

$$P(C|X) = \frac{P(X|C)\,P(C)}{P(X)} \quad \text{3)}$$

$P(C|X)$is the probability of class 'C' given variable 'X', $P(X|C)$ is the probability of variable 'X' given class 'C'. P(C) and P(X) are the prior probability of class C and the prior probability of variable X, respectively. This algorithm operates by determining the probability of each occurrence of a variable and categorizing the variable according to the outcome that has the highest probability.

d. Support Vector Machine (SVM)

SVM is a supervised learning technique employed for tackling classification tasks. In the classification process, it establishes a linear boundary to distinguish between the sample classes and identifies the optimal hyperplane. It assigns sample data points to their respective classes based on this determination. The hyperplane equation with n variables is as follows.

$$y = w_0 + \sum_{i=1}^{n} w_i x_i \quad \text{4)}$$

Where w_0 is the bias or shift, w_i is the weight vector that must be learned during the training process, and x_i is the variable.

e. Decision Tree

Decision trees take a series of decisions based on input features to determine the target class. The concept works based on the selection of separation rules based on the entropy or impurity of the data set at each node. The formula for calculating entropy is as follows:

$$Entropy(A) = \sum_{i=1}^{n} -p_i log_2(p_i) \quad \text{5)}$$

This formula measures the level of uncertainty or confusion in the data set S. The higher the entropy, the more disorganized or confused the data, which means it is more difficult to make decisions based on the features.

f. Random Forest

Random Forest is a type of supervised learning algorithm that incorporates the concept of a large number of randomly constructed decision trees (Liaw & Wiener, 2002). In the context of user clustering for NOMA, Random Forest can be used to classify users based on relevant features. Random Forest works in two phases. The first phase is to combine a number of N decision trees to create a Random Forest. Then the second phase is to make predictions for each tree created in the first phase.

3. Methodology

This research focuses on classifying users into groups spatially using Machine Learning algorithms and determining the most efficient algorithm with the highest accuracy. The process of applying supervised machine learning in this study is described in Figure 2.

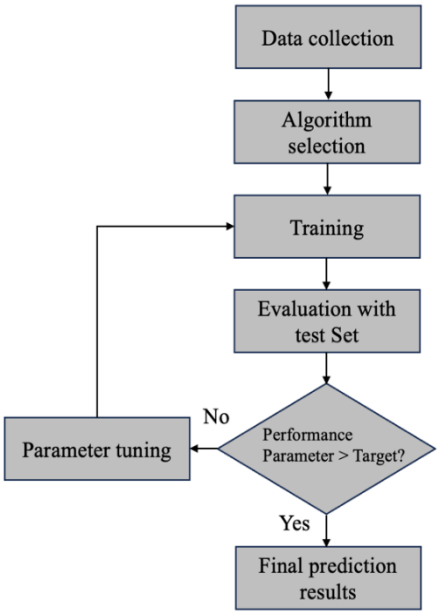

Figure 2. Diagram of user classification process

3.1 Data Collecting

In this study, the data used comes from random user generation using MATLAB software with as many as 1000 users. User attributes are obtained from formulas used in relevant wireless communication standards. User positions are randomly distributed within a radius of 50–1000 m. The user channel gain value is calculated based on the user's distance to the BS using the Rayleigh channel.

The simulated user grouping using supervised learning has a total of 32 classes in 1 cell, which describes 8 spatial areas with each area divided into 4 parts. The number of users in each class is different. This study will compare classification using cross-validation and without cross-validation. The cross-validation used is 10-fold cross-validation. In the experiment without cross-validation, the data is divided into 700 for training data and 300 for testing data. The following is a picture of the user generation used in this research dataset.

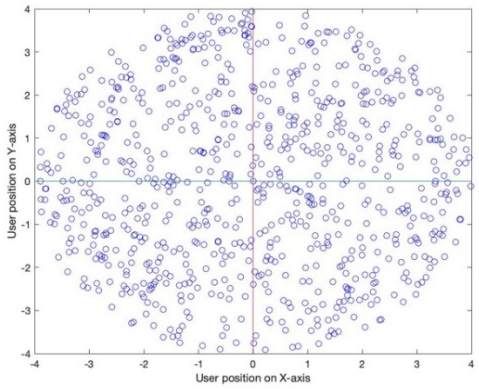

Figure 3. Results of 1000 user distribution

3.2 Model Testing

Testing will be carried out in several experimental scenarios, namely:

a. The first experiment is conducted in several stages. First, the algorithms are trained on a training dataset, and the parameters are compared. Then, the algorithms are validated on a validation dataset to find the optimal values for the hyperparameters of the model under consideration. Finally, the algorithms are tested on a testing dataset to evaluate their performance. In addition, another experiment can be conducted by testing the algorithms on training and testing data only, without using a validation dataset.
b. The second experiment involves performing hyperparameter optimization for all algorithms used. Hyperparameter optimization is an important step in machine learning to improve the performance of models. For example, in the Random Forest algorithm, the number of trees in the forest, the maximum depth of the trees, and the minimum number of samples required to split an internal node can be optimized using cross-validation. Similarly, for the Decision Tree algorithm, the maximum depth of the tree and the minimum number of samples required to split an internal node can be optimized to avoid overfitting. For K-Nearest Neighbors (K-NN), the number of neighbors to consider and the distance metric used to calculate the distance between points can be optimized. In logistic regression, the regularization parameter and the solver used to optimize the logistic regression objective can be optimized. Finally, in Support Vector Machine (SVM), the regularization parameter and the kernel used to transform the data can be optimized. These hyperparameter optimization techniques can help improve the performance of the algorithms and make them more accurate and robust.

3.3 Performance Parameter

Evaluating the performance of machine learning models involves the utilization of various metrics tailored to the specific task at hand. In this study, the assessment included evaluation of accuracy, recall, precision, AUC-ROC curves, and F1 score metrics with a target of 90% for each parameter. Each parameter highlights a different aspect of the model's effectiveness. Accuracy provides a measure of overall precision, while recall and precision offer insight into the model's ability to correctly identify positive cases and minimize false positive cases. The AUC-ROC curve serves as a graphical representation of the model's ability to distinguish classes, and the F1 score is a weighted comparison of recall and precision. For a deeper understanding of the mathematical fundamentals behind these metrics, see Equations 6, 7, 8, and 9, which provide the specific formulas used in this evaluation (Pek et al., 2023).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad \text{6)}$$

$$Recall = \frac{TP}{TP+FN} \quad \text{7)}$$

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

$$F1\ score = 2\ x \frac{(Recall\ x\ Precision)}{Recall+Precision} \quad \text{9)}$$

4. Simulation Results

Various classification algorithms, such as KNN, SVM, Random Forest, Logistic Regression, Naive Bayes, and Decision Tree in supervised learning, are used in the process of clustering users based on spatial concepts. In the proposed methodology, we apply the 10-fold cross-validation method to these models. The purpose of

using cross-validation is to check its effect on the performance of the model as well as to compare it with the results obtained without the use of cross-validation.

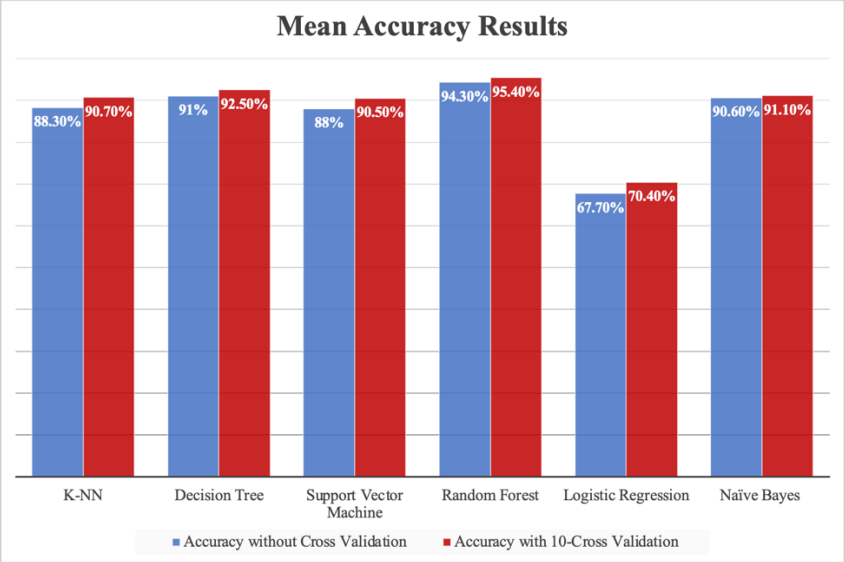

Figure 4. Comparison diagram of accuracy values with cross-validation and without cross-validation

Successfully utilizing cross-validation is essential to prevent deviations and errors when dividing the dataset into training and testing data. The bar chart in Figure 4 visualizes the comparison of model performance based on accuracy. Although the experimental results show the positive impact of cross-validation on model performance, the improvement in accuracy is not significant, and it should be noted that these results were obtained with the default parameters. Table I was compiled to examine the effect of hyperparameter optimization on the models, which shows the improvement in accuracy for all models after parameter adjustment. Before parameter optimization, Random Forest showed the highest performance with 95.4% accuracy. After the optimization was performed, Random Forest still achieved the highest accuracy with 96.1%, followed by Decision Tree at 93.1%. Decision trees proved to be more reliable, especially for handling non-linear relationships between features and classes without certain assumptions about the data distribution or the relationship between features and target variables. in multi-class classification problems with data sets containing many classes. This reliability can be attributed to the decision tree's ability to partition the data based on different features with non-parametric properties. Random forest is a combination of several decision trees, so its reliability is higher than that of a decision tree.

Table. 1 Comparison of accuracy values before and after applying hyperparameter optimization

Algorithms	Before Optimization	After Optimization	Hyperparameter Optimization
K-NN	90.7 %	92.4 %	n_neighbors=7,weights='distance'
Decision Tree	92.5 %	93.1 %	criterion="gini", max_depth=10
Support Vector Machine	90.5 %	92.8 %	kernel='linear', C = 10.0
Random Forest	95.4 %	96.1 %	max_depth=20, random_state=0

Algorithms	Before Optimization	After Optimization	Hyperparameter Optimization
Logistic Regression	70.4 %	87.8 %	penalty='none', C=1.0, solver="sag"
Naïve Bayes	91.1%	92.5 %	var_smoothing=1e-8

Table. 2 Comparison of accuracy values before and after applying hyperparameter optimization

Algorithms	Accuracy	Recall	Precision	F1 Score	AUC Score
K-NN	92.4 %	89 %	89.3 %	88.9 %	0.94
Decision Tree	93.1 %	**91.3 %**	90.6 %	90.6 %	**0.95**
Support Vector Machine	92.8 %	89.3 %	88.6 %	88.4 %	0.53
Random Forest	**96.1 %**	91 %	**94.5 %**	**92.2 %**	**0.95**
Logistic Regression	87.8 %	84.7 %	79 %	81.1 %	0.48
Naïve Bayes	92.5 %	90.7 %	89.9 %	89.3 %	0.55

Table 2 shows the comparison of accuracy, recall, precision, F1 score, and AUC score of each supervised learning algorithm used in this study. The Random Forest algorithm achieved the highest accuracy, precision, and F1 score values. The highest recall value is achieved by the Decision Tree algorithm at 91.3%. The highest AUC score was achieved by the Random Forest and Decision Tree algorithms with 0.95. To better observe the AUC score results, the ROC-AUC curve is shown in Figure 5.

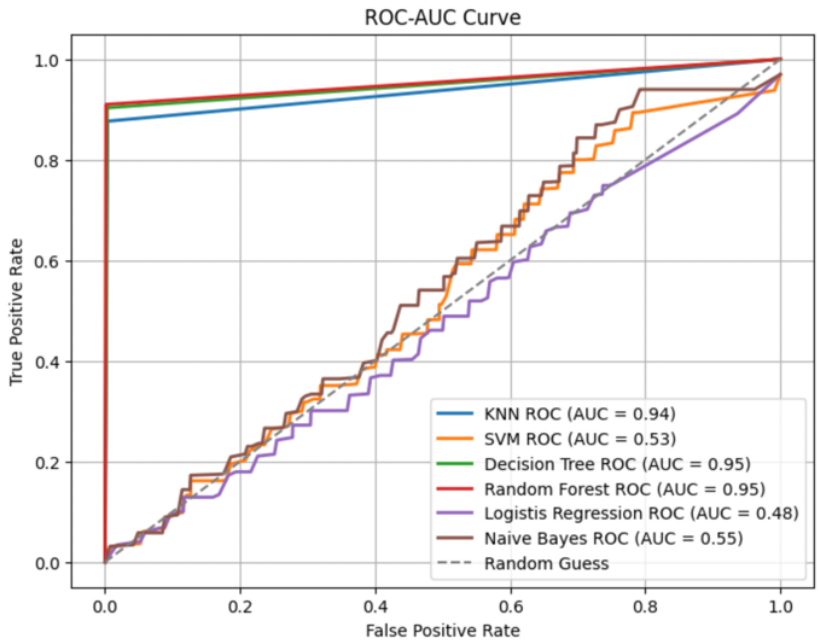

Figure 5. Comparison of ROC-AUC curves of various supervised learning algorithms

ROC-AUC curves are important tools for comparing and selecting appropriate classification models. The evaluation results using ROC-AUC curves provide significant insight into the reliability of various algorithms in the context of user classification based on the concept of beamforming with different regions to represent the user's distance to the BS in the beam. The Random Forest and Decision Tree algorithms stand out with AUC values of around 0.95, showing excellent ability to distinguish regions based on user distance. These two algorithms are solid choices for this task as they are able to cope with the non-linear relationship between features and user classes in multiclass classification problems such as this. However, keep in mind that the Random Forest algorithm tends to be more robust overall due to its ensemble nature, which can increase the stability and reliability of the model. On the other hand, the SVM, Logistic Regression, and Naive Bayes algorithms have lower AUC values of around 0.53, 0.48, and 0.55, respectively, indicating weaker performance in distinguishing regions based on user distance. This could be due to limitations in handling non-linear problems and differences in data patterns.

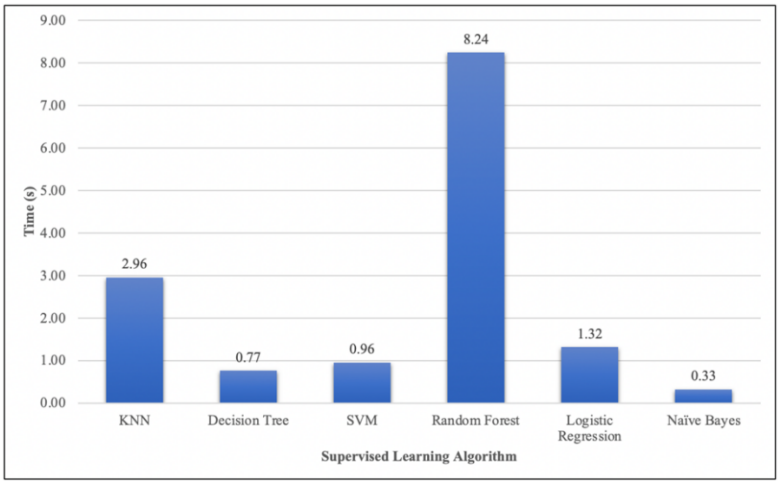

Figure 6. Processing time of each supervised learning algorithm

The processing time of each supervised learning algorithm in this paper is shown in Fig. 6. The training and prediction process time using the Naïve Bayes algorithm has the fastest time of 0.33 seconds, while the longest time is achieved by the Random Forest algorithm at 8.24 seconds. This can be caused because in the Naïve Bayes algorithm, all features are independent of each other, so the time complexity is low. In the random forest algorithm, it takes longer because it involves many decision trees, where the number of trees and tree depth also have an influence on the processing time.

Based on the experimental results, the random forest algorithm provides the highest accuracy value for user classification. This algorithm also has the highest accuracy value after hyperparameter optimization. The best algorithm in terms of model fit is also owned by the random forest and decision tree with the highest AUC value. However, the processing time required by random forest is longer than that of other algorithms, while the decision tree algorithm has a fairly fast processing time. When looking at the overall performance results, the decision tree algorithm has better performance than all the algorithms used.

5. Conclusion

This paper addresses user clustering using supervised machine learning for PD-NOMA systems, where users are clustered based on spatial concepts. The correlation between the user's position on the BS and the user's channel gain condition are the main variables in determining the dependent variable. The supervised learning methods used are K-NN, SVM, Decision Tree, Random Forest, Logistic Regression, and Naive Bayes. Based on

the simulation results, the random forest algorithm has accuracy, precision, F1 score, and AUC parameter values that outperform other algorithms. However, in terms of training and prediction process time, random forest has a longer time than other algorithms. However, the time required is still acceptable in the process of classifying users.

One promising future development of this research is the use of unsupervised learning and reinforcement learning techniques to create a more dynamic and adaptable clustering model. The clustering results obtained are then used to simulate the performance of user quality parameters, thus further enriching the scope and impact of this research.

6. Acknowledgements

Thank you for Indonesia Endowment Fund for Education (LPDP), Ministry of Finance Republic Indonesia for financial support in this research.

References

Benjebbour, A. (2017). An Overview of Non-Orthogonal Multiple Access. *ZTE Communications*, *15*, 21–30. https://doi.org/10.3969

Bui, V.-P., Nguyen, P. X., Nguyen, H. V, Nguyen, V.-D., & Shin, O.-S. (2019). Optimal User Pairing for Achieving Rate Fairness in Downlink NOMA Networks. *2019 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*, 575–578. https://doi.org/10.1109/ICAIIC.2019.8669061

Ding, Z., Schober, R., & Poor, H. V. (2016). A General MIMO Framework for NOMA Downlink and Uplink Transmission Based on Signal Alignment. *IEEE Transactions on Wireless Communications*, *15*(6), 4438–4454. https://doi.org/10.1109/TWC.2016.2542066

Islam, S. M. R., Avazov, N., Dobre, O. A., & Kwak, K. S. (2017). Power-Domain Non-Orthogonal Multiple Access (NOMA) in 5G Systems: Potentials and Challenges. *IEEE Communications Surveys and Tutorials*, *19*(2), 721–742. https://doi.org/10.1109/COMST.2016.2621116

Kang, J. M., & Kim, I. M. (2018). Optimal user grouping for downlink NOMA. *IEEE Wireless Communications Letters*, *7*(5), 724–727. https://doi.org/10.1109/LWC.2018.2815683

Liaw, A., & Wiener, M. (2002). Classification and Regression by randomForest. *R News*, *2*(3), 18–22.

Oladipupo, T. (2010). Types of Machine Learning Algorithms. In Y. Zhang (Ed.), *New Advances in Machine Learning* (pp. 19–48). Intech Open. https://doi.org/10.5772/9385

Pek, R. Z., Ozyer, S. T., Elhage, T., Ozyer, T., & Alhajj, R. (2023). The Role of Machine Learning in Identifying Students At-Risk and Minimizing Failure. *IEEE Access*, *11*, 1224–1243. https://doi.org/10.1109/ACCESS.2022.3232984

Prabha Kumaresan, S., Tan, C. K., & Ng, Y. H. (2020). Efficient user clustering using a low-complexity artificial neural network (ANN) for 5G NOMA systems. *IEEE Access*, *8*, 179307–179316. https://doi.org/10.1109/ACCESS.2020.3027777

Rish I. (2001). An empirical study of the naive bayes classifier. *IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence*, *January 2001*, 41–46.

S. Rezwan, S. S. and W. C. (2020). *Efficient User Clustering and Reinforcement Learning Based Power Allocation for NOMA Systems*. https://ieeexplore.ieee.org/document/9289376

Song, L., Li, Y., Ding, Z., & Poor, H. V. (2017). Resource Management in Non-Orthogonal Multiple Access Networks for 5G and beyond. *IEEE Network*, *31*(4), 8–14. https://doi.org/10.1109/MNET.2017.1600287

Vidyaningtyas, H., Iskandar, Hendrawan, Pramudita, A. A., & Saputri, D. M. (2024). Investigating SIC Performance Using Sequential Power Allocation for Downlink NOMA. In *Lecture Notes on Data Engineering and Communications Technologies* (Vol. 186, pp. 26–34). Springer Science and Business Media Deutschland GmbH. https://doi.org/10.1007/978-3-031-46784-4_3

Vidyaningtyas, H., Kurniawan, A., Iskandar, Pramudita, A. A., & Saputri, D. M. (2023). The optimum number of users using sequential Power Allocation on PD-NOMA. *5th International Conference on Artificial Intelligence in Information and Communication, ICAIIC 2023*, 158–162. https://doi.org/10.1109/ICAIIC57133.2023.10066962

You, H., Pan, Z., Liu, N., & You, X. (2020). User Clustering Scheme for Downlink Hybrid NOMA Systems Based on Genetic Algorithm. *IEEE Access*, *8*, 129461–129468. https://doi.org/10.1109/ACCESS.2020.3009018

Zhu, L., Zhang, J., Xiao, Z., Cao, X., & Wu, D. O. (2019). Optimal User Pairing for Downlink Non-Orthogonal Multiple Access (NOMA). *IEEE Wireless Communications Letters*, *8*(2), 328–331. https://doi.org/10.1109/LWC.2018.2853741