Implementation of K-Nearest Neighbors Algorithm for Red Onion Crop Suitability Land Assessment in Selaawi, Indonesia

Irbah Luthfifah , Elis Hernawati* , Heru Nugroho , Bayu Rima Aditya , Pramuko Aji

Faculty of Information System, School of Applied Science, Telkom University, Bandung 40257, Indonesia

Corresponding Author Email: elishernawati@tass.telkomuniversity.ac.id

ABSTRACT

Determining land suitability is essential to ensure optimal agricultural productivity. This is because each crop has different soil, climate, hydrology, and relief requirements. By accurately assessing land suitability, farmers can strategically plant crops that are most likely to thrive in their specific conditions, thereby boosting agricultural output and increasing their income. Therefore, this research will discuss determining land suitability using a machine-learning land suitability classification system. This research uses the K-Nearest Neighbors (KNN) algorithm for shallot plants, and the dataset used is soil data taken in real-time in Selaawi, Indonesia. The evaluation results showed that the K-Nearest Neighbors algorithm achieved 98% accuracy. The proposed method can determine the suitability of land for shallot plants.

1. INTRODUCTION

Shallots are one of the most important vegetables in Indonesian society. In health, shallots have been known since ancient times as a traditional medicine to reduce fever, treat diabetes and ulcers, and reduce sugar and cholesterol levels [1]. Economically, shallots are a strategic horticultural commodity with high economic value that contributes significantly to the economic development of a region [2]. In Selaawi District, Garut Regency, the productivity of shallot farming still needs to improve. According to BPS 2023 data, shallot production in Selaawi in 2020 was only 103 quintals and decreased to 0 quintals in 2021 [3]. This decline is caused by the condition of the rainfed land, which relies only on rainfall as a water source without an irrigation supply. To overcome this problem, proper land management and an understanding of land suitability are needed to make it loose and suitable for shallots. In addition, farmers need to know whether their land is ideal for the crops they will grow. Nowadays, more and more studies are using machine learning models based on land use data as an efficient way to map land suitability [4].

The application of the KNN algorithm can help predict the suitability of shallot farming land by utilizing nearest neighbor data to determine the class or category that matches the land's characteristics. Thus, it aims to utilize machine learning technology to support the development of shallot farming.

Based on Scopus data searches from 2018 to 2023, research related to machine learning agriculture is still widely found with trends, as seen in Figure 1.

Figure 1 shows a graph of the number of agricultural machine learning researches that continue to increase yearly. This indicates that machine learning has great potential to be applied in agriculture. Machine learning research in agriculture can be used for various purposes, such as land suitability classification, crop yield prediction, and pest and disease control. Such research can help farmers to increase agricultural productivity and reduce production costs. Especially in Selaawi, Indonesia, agricultural machine learning research can be used to develop a more accurate land suitability classification system. The land suitability classification system can help farmers select crops suitable for their land conditions. This can increase agricultural productivity in the Selaawi Sub-district. Based on this explanation, agricultural machine-learning research still has excellent potential to be developed and sustained in the future. The research can provide significant benefits for farmers and society as a whole.
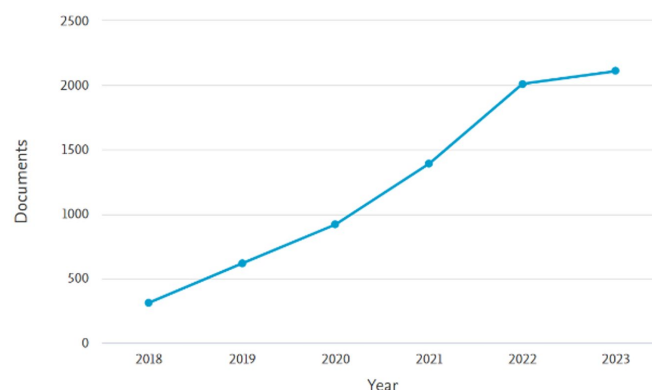
Figure 1. Research trends in machine learning utilization for agriculture 2018–2023

The main contributions of this research are as follows:

•Identification of attributes used to determine the suitability of agricultural land.

•Machine learning model for farmland classification (SMOTE + KNN).

2. RELATED WORKS

Agricultural land suitability refers to the suitability of certain types of land for specific uses, such as crop production, animal production, or forestry. Land suitability is evaluated to determine whether it is suitable for use in a particular cultivation, such as paddy rice or onion crops [5].

Understanding the suitability of agricultural land is important for maximizing its potential to support agricultural activities. This will ensure that agricultural land can be optimally utilized for the agricultural needs of an area [6]. This section will present some similar research that can help with this research process.

A study has been carried out predicting crop development results and identifying suitable agricultural land using the K-NN algorithm. The data used in the study include actual factors such as weather events, humidity, and soil type to predict crop development results and land suitability [7]. The results of this study show that the K-NN model can accurately predict crop development results and identify suitable agricultural land. This model achieved an accuracy of 63.63% in predicting crop development results and land suitability when three parameters were given as input.

Khadiza's research has developed a model for classifying rice fields using the K-Nearest Neighbor (KNN) algorithm to assist farmers in managing rice fields and increasing agricultural productivity. The data used in the study is derived from one of the previous studies, namely data monitoring of rice plants in the Lubuk Pakam area. The parameters include air humidity, soil moisture, light intensity, and water level. This research produces an effective and accurate model to help farmers manage rice fields and increase agricultural productivity. In this study, the model's accuracy reached 92.5%, showing that the KNN model can be used effectively to classify rice fields based on their physical characteristics [8].

The KNN algorithm has been used to identify suitable agricultural land. The data used in the study [9] are derived from various sources, such as online portals and crop reports from multiple districts in Karnataka. This research uses the KNN algorithm to identify suitable agricultural land based on several environmental parameters such as soil type, rainfall, etc. In this study, the model's accuracy reached 56.66%, which shows that the KNN model can be used to identify suitable agricultural land.

Based on the analysis of previous research (Table 1), this paper will be carried out using the stages of the method used, namely, the KNN algorithm. Seven parameters are used in this paper, including soil moisture, nitro, phosphorus, potassium, soil pH, ambient temperature, and soil temperature. The dataset used is data taken directly from Selaawi soil, Indonesia.

Table 1. Previous research

Ref.	Method	Result	Lack
[7]	The K-NN Algorithm	The K-NN model can predict crop development yields and identify suitable agricultural land.	- Using 3 parameters - Accuracy of 63.63%
[8]	The K-NN Algorithm	The accuracy of the K-NN model reached 92.5% and can be used as an effective tool to classify paddy fields based on their physical characteristics.	- Dataset taken from previous research - Using 4 parameters
[9]	The K-NN Algorithm	The K-NN algorithm is simple and easy to implement.	- Datasets were taken from various sources from online portals - Accuracy of 56.66%

3. MATERIAL AND METHODS

The following explains the materials and methods used in this research.

3.1 Dataset

The dataset used in this study is soil and environmental data taken in real-time from agricultural land in Selaawi, Indonesia. Sampling locations were selected based on variations in topography, soil type, and environmental conditions to ensure data representativeness. A stratified random sampling method was used to ensure that each part of the farm was well represented. At each site, soil samples were taken from 0-20 cm depth using a soil drill.

Sampling was conducted over four months, from August, September, October, and November 2023. This period was chosen to cover all phases of plant growth, from the seedling phase to the generative phase to the harvest phase. Sampling was conducted every two weeks to capture variations in soil conditions throughout the growing season.

To collect data, various sensors are installed at sensor stations on the farm. These sensors include sensors for soil moisture, soil pH, nitro content, phosphorus, potassium, soil temperature, ambient temperature, ambient humidity, light intensity, wind speed, and rainfall. Data from these sensors is collected continuously and transmitted to the server over the Internet using the MQTT protocol the dataset obtained has seven main parameters directly related to plant growth and environmental factors, namely: (1) Soil moisture, (2) Nitro, (3) Phosphor, (4) Potassium, (5) Soil pH, (6) Ambient temperature, and (7) Soil temperature (Figure 2).

This comprehensive data capture ensures that the dataset used is of high quality and representativeness, enabling accurate and reliable analysis to classify onion farmland suitability in Selaawi, Indonesia.

The dataset in the machine learning model is divided into two sets: a 30% testing set and a 70% training set. The training set is used to build the model, while the testing set is used to evaluate its accuracy.
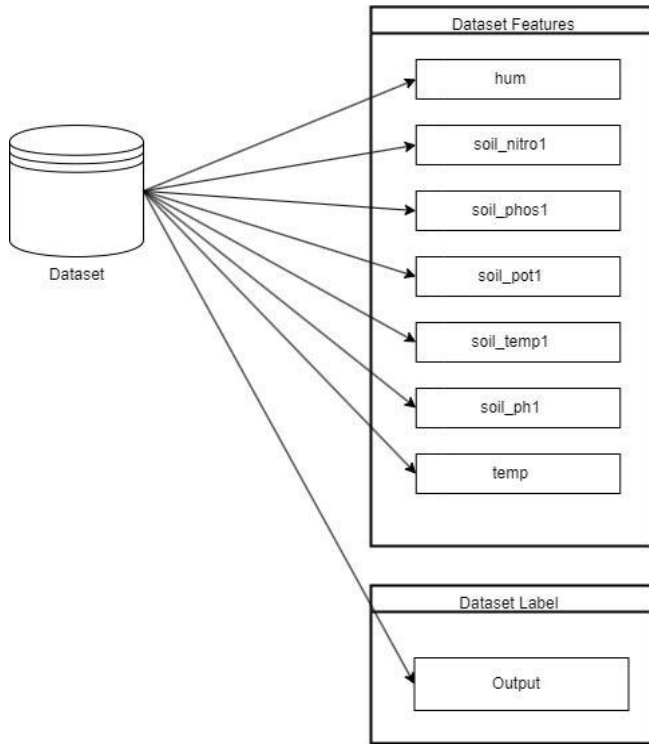
Figure 2. Dataset details

3.2 Proposed method

Figure 3 is a picture of the proposed methodology used in the machine learning model.
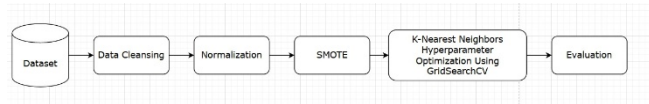

Figure 3. Proposed method

This research uses a method consisting of several main steps: Data Cleansing, Normalization, SMOTE, K-Nearest Neighbors Hyperparameter Optimization Using GridSearchCV, and Evaluation. The following is an explanation of the flow:

Data cleansing: Data cleansing is correcting errors and incomplete data to be used properly. In this process, missing or invalid data is identified, removed, or corrected. One example of data that is removed is climate data in the dataset because the climate data read by the sensor is not valid. This can prevent inaccurate or incomplete data from spreading and have a negative impact on the use of subsequent data, which can lead to errors in decision-making. With the application of data cleansing, it can improve data quality through data cleaning [10].

Normalization: Normalization is a method used in the preprocessing process of each data sample. This method is performed on datasets that have many 0 values with attributes of different scales. Normalization is used to overcome data outliers [11]. Outlier data refers to observations that have numerical values that are significantly far from other data or look very different from other sample members when the observation occurs [12]. The MinMax Scaler method is used in the normalization process carried out by the author. The MinMax Scaler method is a method in preprocessing used for feature transformation, which scales each feature individually

with a specific range. MinMax Scaler is done by subtracting the sample with the smallest sample value on the feature. It will be divided by the largest sample value on the feature that has been reduced by the smallest sample value on the feature [13]. The following is the formula for the Min-Max Scaler method.

$$Y = \frac{Y - Ymin}{Ymax - Ymin} \tag{1}$$

where, Y = sample value.

SMOTE: Synthetic Minority Oversampling Technique (SMOTE) is one of the derivatives of Oversampling. The SMOTE technique works by creating a replication of the minority data. The replication is known as synthetic data. This method works by finding the nearest neighbors for each data set in the minority class. After that, synthetic data is created as much as the desired percentage of duplication between the minor data and the k-nearest neighbors chosen randomly [14]. This technique improves the representation of the minority class so that the model can learn better in handling class imbalance. The synthesized data was created based on the K-Nearest Neighbor. All variables used in this research dataset are numeric variables, and the distance between the minor classes is calculated using the Value Difference Metric (VDM) formula as follows [15].

$$\Delta(A, B) = \sum_{i=1}^{n} \delta(V_{1i}, V_{2i}) \tag{2}$$

where,

$\Delta(A,B)$: Distance between observation A and observation B.
N: Number of independent variables.
$\delta(V_{1i}, V_{2i})$: Distance between observations A and B for each calculated variable.

The mathematical formula in determining the distance between observations A and B for each variable is as follows [15].

$$\delta(V_1, V_2) = \sum_{i=0}^{n} \frac{C_{1i}}{C_1} - \frac{C_{2i}}{C_2} \tag{3}$$

where,

N: Number of categories in the i-th variable.
C_{1i}: The number of the 1st category included in the i-th variable.
C_{2i}: Number of 2nd categories included in the i-th variable.
C_1: The number of times the 1st category occurs.
C_2: The number of times the 2nd category occurs.

K-Nearest neighbors hyperparameter optimization using GridSearchCV: Hyperparameter optimization is a step in machine learning models that involves finding the best values for parameters. This process involves minimizing the function $\Psi(\lambda)$ with respect to $\lambda \in \Lambda$, also called the response surface. The function $\Psi(\lambda)$ is a function that represents the response surface or search space Λ, which must be optimized to find the best fit λ. The variable λ is indexed by Λ, which is the set of configurations that must be optimized in the learning process [16]. Grid search is an approach to parameter tuning that methodically builds and evaluates a model for each combination of algorithm parameters specified in a grid [17]. GridSearchCV is part of the scikit-learn module that validates

multiple models automatically and systematically, providing hyperparameters for each model.

In this case, the GridSearchCV KneighborsClassifier is used where the method is a step to find the optimal value of the parameter K (number of nearest neighbors) in the KNN algorithm. The process involves varying the value of K and evaluating the performance of the model using cross-validation techniques. The result is the K parameter that provides the best performance for the KNN model. In this study, various values of K for the n_neighbors parameter in the K-NN algorithm were evaluated. The K values tested were 3, 5, 7, 9, 11, 13, and 15. In addition, several other parameters, such as weights and metrics, were also tested. Weights have two options, namely 'uniform' and 'distance,' while metric has several options, such as 'Euclidean,' 'Manhattan,' 'Chebyshev,' and 'Minkowski.' Based on the evaluation results, here are the accuracy values for various K values (Table 2).

Table 2. Predicted parameters in the model

No.	Parameters	Values
1	weights	'uniform','distance'
2	n_neighbors	3, 5, 7, 9, 11, 13, and 15
3	metric	'Euclidean','Manhattan','Chebyshev', 'Minkowski','Wminkowski','Seuclidean'

From Table 3, it can be seen that the best K value that provides optimal performance in the K-NN model for land suitability classification in Selaawi is K = 15 with Euclidean metric and accuracy of 0.9221. This value provides the best balance between bias and variance and avoids overfitting the training data.

Table 3. Parameter accuracy results

K	Euclidean	Manhattan	Chebyshev	Minkowski
3	0.8967	0.8901	0.8675	0.8573
5	0.9098	0.9076	0.8920	0.8602
7	0.9112	0.9132	0.8963	0.8710
9	0.9143	0.9151	0.9015	0.8723
11	0.9185	0.9175	0.8954	0.8732
13	0.9207	0.9189	0.8921	0.8735
15	0.9221	0.9200	0.8984	0.8745

Evaluation: Evaluation is needed to measure the performance of a model. There are several matrices to estimate performance, one of which is the confusion matrix. Table 4 shows the confusion matrix for agricultural land suitability where each entry has the following meaning: TP is true positive, FN is false negative, FP is false positive, and TN is true negative.

Table 4. Suitability of agricultural land

	Prediction: Suitable	Prediction: Not Suitable
True Label: Suitable	TP	FN
True Label: Not Suitable	FP	TN

Besides the confusion matrix, there are also recall, accuracy, precision, and F1 scores, which are used as tools to evaluate machine learning methods. Recall, accuracy, precision, and F1 score are evaluation metrics that are calculated based on the values in the confusion matrix. The goal is to provide a deeper insight into the model's performance. The following is an explanation and formula for recall, accuracy, precision, and F1 score:

Accuracy: Accuracy is an evaluation matrix to measure how well the model makes correct predictions (True Positive and True negative) from the total predictions that have been made. To calculate the accuracy value of the model can use the following mathematical equation [18].

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (4)$$

Precision: Precision is a metric that evaluates how well the model predicts positive classes from the total optimistic predictions (True Positive and False Positive) made. The following is a mathematical equation to calculate precision [19].

$$Precision = \frac{TP}{TP + FP} \quad (5)$$

Recall: Recall is an evaluation measure that describes how well a model correctly identifies positive classes. The following mathematical equation can be used to calculate the value of recall [14].

$$Recall = \frac{TP}{TP + FN} \quad (6)$$

F1 Score: F1 score is an evaluation metric that shows the balance between precision and recall in the classification model. The value of F1 score will provide information on how well the model has been made in combining precision and recall capabilities so that the F1 score value can also provide an understanding of how effective the model is in classifying. The following mathematical formula generates the F1 score value [20].

$$F1\ Score = 2 \times \frac{Recall \times Precission}{Recal + Precission} \quad (7)$$

4. RESULT AND DISCUSSION

4.1 Normalization transformation

The following is the implementation of Min-Max Scaler normalization, a simple but effective method for scaling features to a certain range, generally 0 to 1. The features of the dataset before normalization using Min-Max Scaler are shown in the Table 5.

Table 5. Features data before normalization

Features	Values
Hum	37–99.7
Soil_nitro1	0–74
Soil_phos1	0–217
Soil_pot1	0–210
Soil_temp1	20.1–36.7
Soil_ph1	3–8.0
Temp	15.1–38.3

After normalization is used, the following results are obtained (Table 6).

**Table 6.** Features data after normalization

Features	Values
Hum	0.571429–0.938776
Soil_nitro1	0.054054–0.216216
Soil_phos1	0.253456–0.382488
Soil_pot1	0.233333–0.361905
Soil_temp1	0.084337–0.512048
Soil_ph1	0.381818–0.709091
Temp	0.137931–0.426724

4.2 SMOTE testing

SMOTE testing aims to find the accuracy value before and after using SMOTE.

4.2.1 Testing before SMOTE

The dataset used in this study, especially in the dataset label column, experiences data imbalance. The class between 'No' and 'Match' has very unbalanced data, this can be seen in the Figure 4.
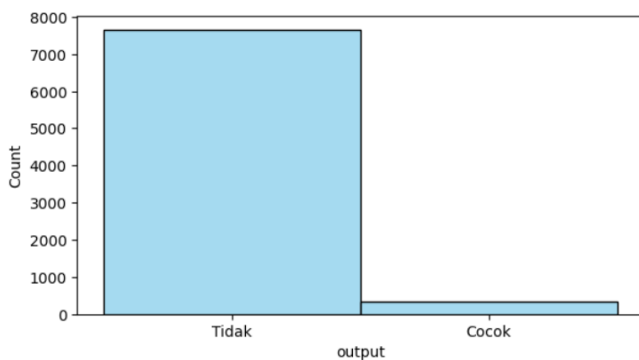

Figure 4. Bar plot before SMOTE

The total number of datasets before SMOTE oversampling is 7,983 rows.

4.2.2 Testing after SMOTE

Since the classes in the label column of the dataset are not balanced, the SMOTE algorithm is used to solve the problem. By adding new minority class instances, the data will become more balanced, so the machine learning algorithm can better handle imbalance classification. The figure below shows the imbalance in the output column between the 'Not and 'Suitable' classes.

To balance the above classes, an oversample algorithm using SMOTE is used. Figure 5 is the bar plot image after using oversample on the output class.
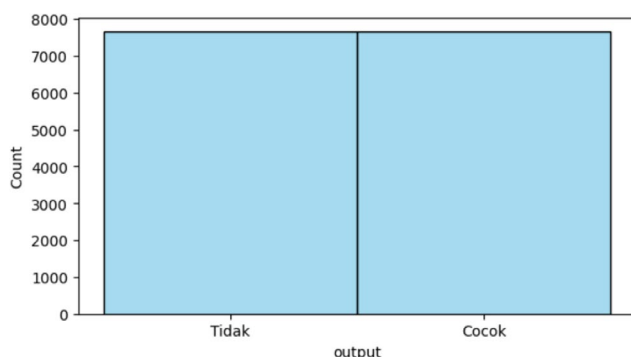
Figure 5. Bar plot after SMOTE

After using the SMOTE algorithm, the data that was originally unbalanced becomes balanced. The total number of datasets after oversampling the data is 15,312 row data.

4.3 Confusion matrix

A confusion matrix is data in a growth or achievement diagram used to visually assess model performance and compare the response of the built model with the results achieved by chance [21]. The matrix consists of four main components: true positive (TP), true negative (TN), false positive (FP), and false negative (FN) [19]. Figure 6 shows the confusion matrix results of the land suitability classification system.
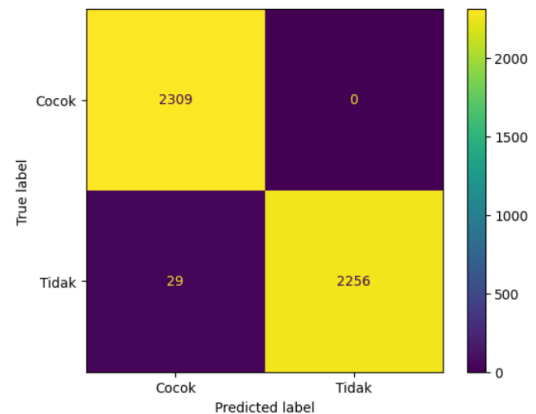
Figure 6. Confusion matrix

4.4 Classification report

Figure 7 is the classification report for the model, which outlines the precision, recall, F1 score, and accuracy metrics.

Figure 7. Classification report

The classification model shows excellent results with 98% accuracy, where precision is 97%, recall is 1.00%, and F1 score is 98%.

4.5 Comparison to other algorithms

The K-Nearest Neighbors (KNN) algorithm used in this research is compared with several other commonly used algorithms, such as the Support Vector Machine (SVM). These algorithms are selected based on their effectiveness in handling classification data. The results of comparing with these algorithms resulted in the following.

Support Vector Machine (SVM): SVM still needs to be more effective in handling unbalanced datasets where the majority class is dominant. On datasets with class imbalance, SVM tends to ignore the minority class, which causes the overall accuracy to look high but the performance on the minority class to be low.

5. CONCLUSION

The application of machine learning models to predict land suitability in Selaawi, Indonesia, is proven to help extension workers classify land suitability. The data training and testing process uses real-time data taken from the land of Selaawi, Indonesia, divided into 70% training data and 30% test data. By using Hyperparameter Optimization, the best parameters of the existing model can be determined. The calculation results using the K-Nearest Neighbors (KNN) algorithm in the machine learning model for the land suitability classification system showed a high accuracy of 98%. This shows that using the K-NN algorithm can effectively classify land suitability, especially in Selaawi farmland, Indonesia.

This data modeling can be used for further research, such as developing modeling that can be a monitoring tool for soil actions. In the future, this data modeling can also be extended by applying the K-NN model to other crops and regions and exploring further methods to improve the model's accuracy. These efforts will strengthen the significance of this paper and make greater contributions to the field of land suitability classification and agricultural soil management.

REFERENCES

[1] Sun, W., Shahrajabian, M.H., Cheng, Q. (2019). The insight and survey on medicinal properties and nutritive components of shallot. Journal of Medicinal Plants Research, 13(18): 452-457. https://doi.org/10.5897/JMPR2019.6836

[2] Dewi, T., Yustika, R.D., Arianti, F.D. (2024). Enhancement of production and food security through sustainable shallot cultivation. IOP Conference Series: Earth and Environmental Science, 1364(1): 012052. https://doi.org/10.1088/1755-1315/1364/1/012052.

[3] Central Bureau of Statistics of Garut Regency (2023). Garut Regency in Figures 2023. https://garutkab.bps.go.id/id/publication/2023/02/28/9c915d2e8b7374303606ddd5/kabupaten-garut-dalam-angka-2023.html.

[4] Møller, A.B., Mulder, V.L., Heuvelink, G.B., Jacobsen, N.M., Greve, M.H. (2021). Can we use machine learning for agricultural land suitability assessment? Agronomy, 11(4): 703. https://doi.org/10.3390/agronomy11040703

[5] Robbo, A., Galib, M. (2023). Evaluasi kesesuaian Lahan Padi sawah (Oryza sativa l.) di kabupaten luwu. Jurnal Tanah dan Sumberdaya Lahan, 10(2): 319-325. https://doi.org/10.21776/ub.jtsl.2023.010.2.15

[6] Samaniego, L., Schulz, K. (2009). Supervised classification of agricultural land cover using a modified k-NN technique (MNN) and Landsat remote sensing imagery. Remote Sensing, 1(4): 875-895. https://doi.org/10.3390/rs1040875

[7] Khadiza, R. (2019). Rice field classification using k-nearest neighbor algorithm. University of Sumatera Utara Repository. https://repositori.usu.ac.id/handle/123456789/15413.

[8] Hart, P. (1968). The condensed nearest neighbor rule (corresp.). IEEE Transactions on Information Theory, 14(3): 515-516. https://doi.org/10.1109/TIT.1968.1054155

[9] Prasad, K.H., Faruquie, T.A., Joshi, S., Chaturvedi, S., Subramaniam, L.V., Mohania, M. (2011). Data cleansing techniques for large enterprise datasets. In 2011 Annual SRII Global Conference, San Jose, CA, USA, pp. 135-144. https://doi.org/10.1109/SRII.2011.26

[10] Sari, I.P., Azzahrah, A., Qathrunada, I.F., Lubis, N., Anggraini, T. (2022). Perancangan sistem absensi pegawai kantoran secara online pada website berbasis HTML dan CSS. Blend Sains Jurnal Teknik, 1(1): 8-15. https://doi.org/10.56211/blendsains.v1i1.66

[11] Muñiz, C.D., Nieto, P.G., Fernández, J.A., Torres, J.M., Taboada, J. (2012). Detection of outliers in water quality monitoring samples using functional data analysis in San Esteban estuary (Northern Spain). Science of the Total Environment, 439: 54-61. https://doi.org/10.1016/j.scitotenv.2012.08.083

[12] Hale, J. (2019). Scale, standardize, or normalize with scikit-learn. https://towardsdatascience.com/scale-standardize-or-normalize-with-scikit-learn-6ccc7d176a02.

[13] Siringoringo, R. (2018). Klasifikasi data tidak Seimbang menggunakan algoritma SMOTE dan k-nearest neighbor. Journal Information System Development (ISD), 3(1): 44-49.

[14] Permatasari, R.D., Rizki, S.W., Debataraja, N.N. (2020). Penerapan synthetic minority oversampling technique Dalam mengatasi Data Tidak seimbang pada Metode classification and regression tree. Bimaster: Buletin Ilmiah Matematika, Statistika dan Terapannya, 9(1): 231-238. https://doi.org/10.26418/bbimst.v9i1.38949

[15] Bergstra, J., Bengio, Y. (2012). Random search for hyper-parameter optimization. Journal of Machine Learning Research, 13(2): 281-305.

[16] Ranjan, G.S.K., Verma, A.K., Radhika, S. (2019). K-nearest neighbors and grid search cv based real time fault monitoring system for industries. In 2019 IEEE 5th International Conference for Convergence in Technology (I2CT), Bombay, India, pp. 1-5. https://doi.org/10.1109/I2CT45611.2019.9033691

[17] Meshref, H. (2019). Cardiovascular disease diagnosis: A machine learning interpretation approach. International Journal of Advanced Computer Science and Applications, 10(12): 258-269. https://doi.org/10.14569/IJACSA.2019.0101236

[18] Li, Y., Zhang, Y., Zhao, L., Zhang, Y., et al. (2018). Combining convolutional neural network and distance distribution matrix for identification of congestive heart failure. IEEE Access, 6: 39734-39744. https://doi.org/10.1109/ACCESS.2018.2855420

[19] Rajamhoana, S.P., Devi, C.A., Umamaheswari, K., Kiruba, R., Karunya, K., Deepika, R. (2018). Analysis of neural networks based heart disease prediction system. In 2018 11th International Conference on Human System Interaction (HSI), Gdansk, Poland, pp. 233-239. https://doi.org/10.1109/HSI.2018.8431153

[20] Das, C., Sahoo, A.K., Pradhan, C. (2022). Multicriteria recommender system using different approaches. In Cognitive Big Data Intelligence with a Metaheuristic Approach, pp. 259-277. https://doi.org/10.1016/B978-0-323-85117-6.00011-X

[21] Javaheri, S.H., Sepehri, M.M., Teimourpour, B. (2014). Chapter 6 - Response modeling in direct marketing: Adata mining-based approach for target selection. In Data Mining Applications with R, pp. 153-180. https://doi.org/10.1016/B978-0-12-411511-8.00006-2