KAMPUS AKADEMIK PUBLISING
Jurnal Sains Student Research
Vol.3, No.6 Desember 2025
e-ISSN: 3025-9851; p-ISSN: 3025-986X, Hal 461-470
DOI : https://doi.org/10.61722/jssr.v3i6.6557

Comparison Naïve Bayes and SVM to Classify Drought-Infected Rice Plants Based on Morphological Characteristics in Supporting National Food Security

Damaris Easter Nugrahita Christi
The Republic of Indonesia Defense University, INDONESIA
Angelia Melisa Hutapea
The Republic of Indonesia Defense University, INDONESIA
Fulkan Kafilah Al Husein
The Republic of Indonesia Defense University, INDONESIA
Nadiza Lediwara
The Republic of Indonesia Defense University, INDONESIA
Sembada Denrineksa Bimorogo
The Republic of Indonesia Defense University, INDONESIA
Alamat: Kawasan Indonesia Peace and Security Center (IPSC) Sentul, Sukahati, Kec. Citeureup, Bogor, Jawa Barat, 16810, Indonesia.
Korespondensi penulis: damarischristi@gmail.com

Abstrak. *Data mining is part of the Knowledge Discovery in Database (KDD) process. The use of data mining serves to classify, predict, and extract other useful information from large data sets. This study aims to classify rice plants under treatment (drought stress and control) using data mining, focusing on the analysis of the variables of Leaf Area (LA), Root Length (RL), and Shoot Length (SL). Each classification algorithm has different characteristics, resulting in varied performance results. After testing both classification algorithms, the accuracy results were 71.70% for Naïve Bayes and 73.85% for SVM. This shows that the SVM algorithm performs better than Naïve Bayes algorithms to determine best treatment of rice to support national food security further. Furthermore, It also can be concluded that using a machine learning approach can solve problems in the classification of rice plants affected by drought threats is fairly effective with the maximum score obtained is only 73.85%.*
Keywords: *Rice Plant, Modelling, SVM, Naïve-Bayes.*

INTRODUCTION

Rice (*Oryza Sativa*) is a staple food that sustains nearly half of the global population (Mohidem, et al., 2022). Especially in Indonesia, the significance of rice by stating that rice is a daily staple for about 95% of Indonesia's population (Paiman, et al., 2020). The increasing global food demand, driven by population growth, exerts significant pressure on agricultural systems to enhance productivity of rice and ensure sustainable food availability (Wang, 2022). As a water-dependent crop, rice cultivation is highly susceptible to environmental stressors, especially drought, which can drastically reduce yields (Toulotte, et al., 2022). Drought stress limits the plant's ability to carry out vital physiological processes, such as nutrient uptake and photosynthesis, thus hindering growth and leading to lower production (Toulotte, et al., 2022). Given these challenges, understanding how rice plants respond to drought conditions is critical for improving crop resilience and developing better agricultural management practices.

Received Setember 08, 2025; Revised Oktober 01, 2025;Desember 01, 2025
** Damaris Easter Nugrahita Christi, damarischristi@gmail.com*

In recent years, technological advancements in agriculture have increasingly relied on data-driven approaches, such as machine learning, to analyze plant responses to environmental factors. Machine learning models can be employed to classify plant responses to different treatments, helping farmers and agricultural managers make informed decisions to mitigate the effects of stressors like drought (Rico-Chávez , et al., 2022). Specifically, the application of classification algorithms, such as Naïve Bayes and Support Vector Machine (SVM), has gained attention in agriculture for their ability to predict outcomes based on plant growth data (Riyadi et al., 2022). These algorithms utilize input features like plant physiological traits—such as Leaf Area (LA), Root Length (RL), and Shoot Length (SL)—to classify the treatment conditions (control or drought) that have been applied to the plants. The plant materials used in this study were an F9 population consisting of 90 recombinant inbred lines (RILs), which originated from a cross between the rice varieties IR64 and Hawara Bunar, referred to as IRH (Satrio et al., 2021). The parent lines have distinct traits in response to drought stress. IR64 is a high-yielding lowland variety but is vulnerable to drought, whereas Hawara Bunar (HB) is a local upland variety that is well suited to drought-prone environments (Miftahudin et al., 2020; Satrio et al., 2019).

The Naïve Bayes algorithm, rooted in Bayes' Theorem, is a probabilistic classifier that operates under the assumption of feature independence. This simplicity in design, combined with its effectiveness, makes Naïve Bayes a popular choice in various classification tasks, especially in text and sentiment analysis (Lubis et al., 2020). Despite the often unrealistic assumption of feature independence, Naïve Bayes frequently delivers strong performance in practice, as it calculates the posterior probability of each class based on the prior probabilities and the likelihood of the observed data. On the other hand, Support Vector Machine (SVM) is a powerful supervised learning algorithm known for its robustness in handling complex, high-dimensional datasets. SVM works by finding the optimal hyperplane that separates data points into distinct classes, making it particularly well-suited for binary classification tasks where the goal is to separate data into two categories (Abdullah, et al., 2021).

Several studies have demonstrated the efficacy of Naïve Bayes and SVM in classification problems across different domains. Riyanto (2019) compared the performance of Naïve Bayes and SVM in classifying online readership, reporting that the SVM algorithm achieved an accuracy of 63.39%, outperforming Naïve Bayes. Similarly, Narayan (2020) conducted a study about comparative analysis of two classifiers—Support Vector Machine (SVM) and Naive Bayes—for classifying surface electromyography (sEMG) signals, finding that SVM produced a higher accuracy (95.8%) than Naïve Bayes. Another study by Apriyani & Kurniati (2020) compared Naïve Bayes and SVM in the classification of diabetes mellitus at Siti Khadijah Islamic Hospital in Palembang, with SVM achieving the highest accuracy of 96.27%. These findings suggest that while Naïve Bayes is a strong baseline classifier, SVM often provides superior performance in more complex classification tasks.

In the context of rice cultivation, classification models serve as essential tools for predicting how rice plants will respond to varying treatment conditions (Alfred, 2021). Given the critical role that rice plays in global food security, especially in countries like Indonesia where it serves as a staple food, optimizing crop management strategies is imperative (Verma, et al., 2021). One way to achieve this optimization is through the development of machine learning models that can accurately predict the treatment needs of rice plants based on specific physiological characteristics. These models can utilize input features such as Leaf Area (LA), Root Length (RL), and Shoot Length (SL), which are key indicators of plant health and growth. By leveraging such features, machine learning algorithms can classify whether rice plants are under drought stress or in a controlled environment.

The present study seeks to investigate which algorithm—Naïve Bayes or SVM—yields the highest classification accuracy for rice plant responses to drought conditions. This research is based on data collected in 2021 from a study conducted by Satrio et al. (2021), involving 90 rice varieties subjected to two treatments: control and drought stress. The dataset contains observations on three key physiological variables: Leaf Area (LA), Root Length (RL), and Shoot Length (SL). These variables are used as features in the machine learning models, with the treatment condition (control or drought) serving as the target variable.

The importance of this research lies in its potential to contribute to the growing body of literature on data mining and machine learning in agriculture. By comparing the performance of Naïve Bayes and SVM in classifying rice plant responses to treatment, this study aims to provide insights into which algorithm is more effective for this specific application. Additionally, this research highlights the broader applicability of machine learning techniques in agricultural decision-making, offering a foundation for future technological developments aimed at improving crop resilience and productivity. More over, the purpose of this research is to contribute to national food security by improving the understanding of how rice plants, a staple crop for millions of Indonesians, respond to drought conditions. By utilizing machine learning techniques, specifically the Naïve Bayes and Support Vector Machine (SVM) algorithms, this study aims to accurately classify rice plant treatments and responses under drought stress. The insights gained from this classification can help optimize agricultural management strategies, ensuring that rice plants receive the most effective treatments during adverse environmental conditions. This, in turn, enhances rice productivity and resilience, directly supporting Indonesia's efforts to secure a stable food supply amidst growing challenges like climate change and water scarcity. Through this research, technological advancements in data-driven decision-making for crop management can be fostered, ultimately strengthening t

METHOD

This study aims to classify rice plant growth based on treatments (drought and control) using data mining techniques. The detailed stages applied in this research include several steps as shown in Figure 1.

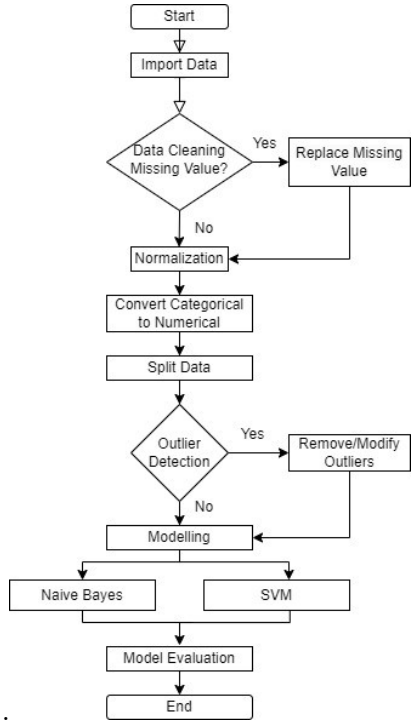

Figure 1. Research Framework

The process begins with importing the data using the Read CSV operator in RapidMiner. Next, missing values are handled with the Replace Missing Values operator, and data normalization is performed using the Normalize operator to standardize the variable scales (Chen, et al., 2024). The target variable is then converted into a numerical format through the Nominal to Numerical operator. The dataset is divided into training data (80%) and testing data (20%) using the Split Data operator (Rácz , et al., 2021). The process continues with outlier detection using the Outlier Detection operator to ensure the data is clean and ready for modeling. With these preprocessing steps, the dataset becomes structured and prepared to build an effective machine learning model for predicting the appropriate treatment for rice plants under stress conditions, as shown in Figure 2.

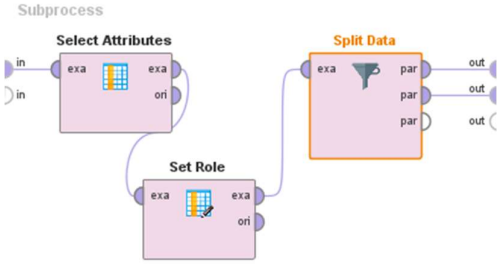

Figure 2. Data Preprocessing with RapidMiner

After the data has been successfully split and quantified, modeling is conducted using the machine learning modeling method applied for classification, which is Support Vector Machine (SVM) and Naive Bayes as shown in Figure 3.

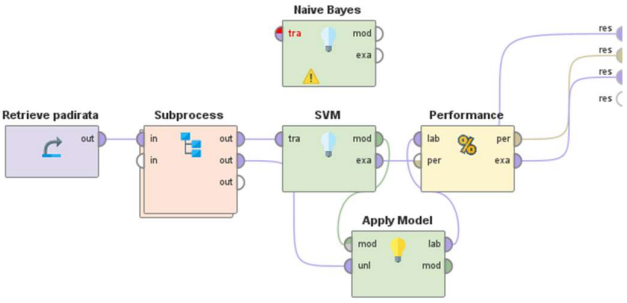

Figure 3. Application of Classification Model in RapidMiner

Dataset

This research employs direct observation methods to obtain primary data. The dataset consists of 175 observation points with four explanatory variables: leaf width, root length, plant height, and developmental stage, which are used to predict the target variable, namely the type of treatment. Table 1 presents a sample dataset used. In addition to the variable 'environment,' the other three dataset variables are numeric.

Table 1. The sample of Dataset

Genotype	Leaf Width	Root Lenght	Plant Height	Environment
1	6.585039683	28.51666667	64.96133333	Control
2	12.12059524	34.353	73.698	Control
3	15.7498571	29.34833333	83.56233333	Control
4	7.0599	18.479	58.8935	Drought
5	7.753833333	13.3195	60.0015	Drought

Table 1 presents a sample from a dataset consisting of 175 observations, where each observation corresponds to a different genotype of rice plants and includes four variables. The three numeric variables—leaf width, root length, and plant height—are measured in centimeters. These variables provide information about the plant's physical characteristics and are used to explain growth under different environmental conditions. The fourth variable, "Environment," is categorical and indicates whether the plant was grown under "Control" (normal growth) or "Drought" (water-stressed) conditions. For example, the first genotype has a leaf width of 6.59 cm, a root length of 28.52 cm, and a plant height of 64.96 cm, and was grown under "Control" conditions. In contrast, the fifth genotype has a leaf width of 7.75 cm, a root length of 13.32 cm, and a plant height of 60.00 cm, and was grown under "Drought" conditions. The aim of the dataset is to use the numeric variables (leaf width, root length, and plant height) to predict the environmental condition (either "Control" or "Drought") in which the plant was grown. This helps in assessing how plant morphology responds to different growing conditions. Next, the classification modeling methods that will be used to classify plants in "Drought" and "Control" conditions will be explained. Two methods, Support Vector Machine and Naïve-Bayes, will be used to compare the accuracy of both models.

Support Vector Machine Algorithm

Support Vector Machine (SVM) is a well-known machine learning algorithm used for solving various classification problems (Tarek H. M. Abou-El-Enien et al, 2015). It operates by selecting a subset of features from the training samples, such that the classification of these

features is equivalent to dividing the entire dataset (Abdullah, et al., 2021). The primary objective of SVM is to create an optimal decision boundary between the existing data classes. SVM's aim is to find the individual hyperplane with the highest margin that can divide the classes linearly (Abdullah, et al., 2021). SVM can handle both linear and nonlinear data through techniques like soft margin hyperplane and feature space transformation (Anggrawan, et al., 2023). This hyperplane maximizes the margin, defined as the distance between the hyperplane and the nearest data points from each class. In cases of non-linear classification, SVM employs a kernel technique to transform the data into a higher-dimensional space where the classes become linearly separable. To maximize the margin, SVM minimizes the following function:

$$\frac{1}{2}\|w\|^2 \quad (1)$$

subject to the constraints:

$$y_i(w \cdot x_i + b) \geq 1 \quad \forall i \quad (2)$$

Where y_i is a class label (1 or -1) from x_i

Equation (1) is illustrated in Figure 4.

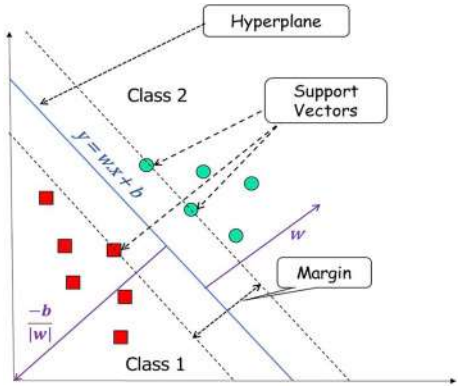

Figure 4. Hyperplane SVM Ilustration

Naïve Bayes Algorithm

Naïve Bayes is a machine learning classification algorithm based on Bayes' Theorem, with the "naive" assumption that all features are independent (Lubis et al., 2020). Despite this assumption rarely holding true in real-world situations, the algorithm often performs well across various applications, particularly in tasks like text classification and sentiment analysis (Saptadi, et al., 2023). The strength of Naïve Bayes lies in its ability to calculate both prior and posterior probabilities, which are then used to make classification decisions (Riyadi et al., 2022). Bayes' Theorem (Equation (2)) forms the basis of this approach, providing a way to update the probability of a hypothesis C given new evidence X. Specifically, the posterior probability $P(C|X)$ and the likelihood $P(X|C)$, making it a powerful tool for decision-making under uncertainty.

$$\boldsymbol{P(C|X) = P(C|X).\frac{P(C)}{P(X)}} \quad (3)$$

Evaluation Methods

The evaluation stage serves to measure the performance of the model with test data. The evaluation results will show how well the model can predict the optimal treatment for rice plants. At this stage, evaluation metrics such as accuracy, precision, and recall are used to assess model performance (Tharwat, 2021).

Table 2. Confussion Matrix

		Actual	
		Control	**Drought**
Prediction	**Control**	*True Positive (TP)*	*False Negative (FN)*
	Drought	*False Positive (FP)*	*True Negative (TN)*

$$Accuration = \frac{TP+TN}{TP+TN+FP+FN} .100\% \quad (4)$$

$$P = \frac{TP}{TP+FP} .100\% \quad (5)$$

$$Recall(R) = \frac{TP}{TP+FN} .100\% \quad (6)$$

$$F1\ score = \frac{2PR}{P+R} .100\% \quad (7)$$

The evaluation process focuses on improving the scores by modifying existing features, adjusting model parameters, and further exploring the properties of the data (Chen, et al., 2020). The goal is to identify the most suitable method and achieve the highest possible performance scores.

RESULT AND DISCUSSION

The results of the comparison between the two methods, SVM and Naïve Bayes, are presented separately in the form of a confusion matrix evaluation. The accuracy of both models will be evaluated and discussed one by one in the following subsections. Further implementation and interpretation will also be discussed in more detail in the subsections below

Results of the SVM Algorithm

The first classification modeling is performed using the SVM algorithm. The performance of the SVM algorithm is presented in Table 6, with an accuracy of 73.85%, precision of 70%, and recall of 80.77%. Below, Table 3 presents the confusion matrix to clarify the accuracy that has been explained above.

Table 3. Confussion Matrix SVM

		Actual	
		Control	**Drought**
Prediction	**Control**	21	9
	Drought	5	18

The confusion matrix illustrates the performance of a Support Vector Machine (SVM) model in classifying rice plants as either under drought conditions or in a control (non-drought) condition. The matrix consists of the predicted classifications compared to the actual conditions. Out of the actual control plants, 21 were correctly identified as being in the control condition, while 5 were incorrectly predicted as being in drought conditions. Conversely, for the actual drought plants, 18 were correctly classified as being in drought, while 9 were mistakenly classified as being in the control condition. This matrix reveals the strengths and limitations of the SVM model in distinguishing between the two classes. While the model performs well in many cases, there are still errors, particularly when it misclassifies some drought-affected plants as being in the control group. From this matrix, additional performance metrics like accuracy,

precision, recall, and F1-score can be calculated to provide a more detailed assessment of the model's classification ability.

Results of the Naïve Bayes Algorithm

The first experiment involves the process using the Naïve Bayes algorithm. The Naïve Bayes algorithm's approach is based on probabilities, so there is no manual input of parameters. The performance of the Naïve Bayes model is presented in the table, with an accuracy of 71.70%, precision of 70.37%, and recall of 73.08%. Below, Table 4 presents the confusion matrix to clarify the accuracy that has been explained above.

Table 4. Confussion Matrix Naïve Bayes

		Actual	
		Control	**Drought**
Prediction	**Control**	19	8
	Drought	7	19

The confusion matrix shows the performance of a Naïve Bayes classification model that aims to classify rice plants as either being in a "Control" condition (no drought) or in a "Drought" condition (experiencing drought). The matrix has four key values: 19 instances where the model correctly predicted "Control" when the actual condition was "Control" (true positives), 8 instances where the model incorrectly predicted "Control" when the actual condition was "Drought" (false positives), 7 instances where the model incorrectly predicted "Drought" when the actual condition was "Control" (false negatives), and 19 instances where the model correctly predicted "Drought" when the actual condition was "Drought" (true negatives). Overall, the model performs reasonably well, with correct predictions in most cases, but it has a moderate number of misclassifications, particularly in predicting the "Control" condition for plants actually under "Drought." This can be an indication that the model might slightly overestimate the "Control" condition.

Implementation and Interpretation

The classification of rice plants affected by drought and those that are not, based on morphological traits such as leaf area, root length, and shoot length, holds several key expectations. Early drought detection is anticipated, enabling farmers and researchers to take quicker preventive measures to minimize its impact on crop yields (S. Pulwarty & Sivakumar, 2014). Additionally, distinguishing between drought-affected and unaffected plants allows for more efficient resource management, such as water and nutrients, potentially boosting agricultural productivity, particularly in drought-prone areas (Iqbal et al., 2020). This classification could also offer valuable insights for breeding programs, prioritizing traits like longer roots for the development of drought-tolerant rice varieties. Clear data on drought-affected plants would support better land management strategies, helping farmers choose suitable areas or adjust irrigation systems. Furthermore, the system could mitigate production risks by providing more accurate information to guide decisions on irrigation, fertilizer use, and harvest timing (Zhai et al., 2020). Ultimately, this approach could enhance the sustainability of agriculture through wiser resource use and more stable yields, even under drought stress, contributing to food security and better adaptation to climate-related water fluctuations.

The results from the SVM and Naïve Bayes algorithms demonstrate their effectiveness in classifying rice plants under drought and control conditions, with accuracy rates of 73.85% and 71.70%, respectively. The SVM model, with its higher recall (80.77%), shows a stronger ability

to correctly identify drought-affected plants, making it useful for early drought detection. Meanwhile, the Naïve Bayes model exhibits balanced precision (70.37%) and recall (73.08%), offering reliable performance in both identifying control and drought conditions. These results reveal the models' strengths and limitations in predicting crop stress, which can be applied in agricultural decision-making. In the context of national food security, these models can play a vital role in managing rice crops, a key food source. Accurate prediction of drought conditions allows for timely interventions in irrigation, resource management, and disaster response. By implementing these models in an agricultural monitoring system using remote sensing data or in-field sensors, authorities can continuously monitor crop health and act swiftly in high-risk drought areas. This AI-driven approach can help enhance food security by reducing crop losses, optimizing resource use, and ensuring better preparedness for drought conditions across regions.

CONCLUSION

The conclusion drawn from this research is that the machine learning approach to selecting the appropriate treatment for rice plants is quite effective. Each classification algorithm has distinct characteristics, leading to varying performance results. After testing both classification algorithms, the accuracy results were 71.70% for Naïve Bayes and 73.85% for SVM. SVM demonstrated the highest performance. Therefore, the most effective algorithm for the rice plant treatment selection case, using numerical data models with a machine learning approach, is the Support Vector Machine, as its performance is well-suited for classifying two different classes. Additionally, the workflow developed in this research is a significant achievement, as it can be applied to any type of data in the future using the established workflow. In general, based on the study's findings, plant height, leaf width, and root length are significant metrics for estimating the appropriate treatment for rice plants.

DAFTAR PUSTAKA

Abdullah, D. M., & Abdulazeez, A. M. (2021). Machine Learning Applications based on SVM Classification: A Review. *Qubahan Academic Journal*, *1*(2), 81–90. https://doi.org/10.48161/qaj.v1n2a50

Alfred, R., Obit, J. H., Chin, C. P. Y., Haviluddin, H., & Lim, Y. (2021). Towards paddy rice smart farming: A review on big data, machine learning, and rice production tasks. *IEEE Access*. Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/ACCESS.2021.3069449

Anggrawan, A., Hairani, H., & Satria, C. (2023). Improving SVM Classification Performance on Unbalanced Student Graduation Time Data Using SMOTE. *International Journal of Information and Education Technology*, *13*(2), 289–295. https://doi.org/10.18178/ijiet.2023.13.2.1806

Chen, F., Yu, L., Mao, J., Yang, Q., Wang, D., & Yu, C. (2024). A novel data-characteristic-driven modeling approach for imputing missing value in industrial statistics: A case study of China electricity statistics. *Applied Energy*, *373*, 123854.

Chen, R. C., Dewi, C., Huang, S. W., & Caraka, R. E. (2020). Selecting critical features for data classification based on machine learning methods. *Journal of Big Data*, *7*(1). https://doi.org/10.1186/s40537-020-00327-4

https://doi.org/10.18201/ijisae.2019252786

Lubis, C. P., Rosnelly, R., Roslina, R., Situmorang, Z., & Wanayumini, W. (2021). PENERAPAN METODE NAÏVE BAYES DAN C4.5 PADA PENERIMAAN PEGAWAI DI UNIVERSITAS POTENSI UTAMA. *CSRID (Computer Science*

Research and Its Development Journal), *12*(1), 51. https://doi.org/10.22303/csrid.12.1.2020.51-62

Mohidem, N. A., Hashim, N., Shamsudin, R., & Man, H. C. (2022, June 1). Rice for Food Security: Revisiting Its Production, Diversity, Rice Milling Process and Nutrient Content. *Agriculture (Switzerland)*. MDPI. https://doi.org/10.3390/agriculture12060741

Narayan, Y. (2020). Comparative analysis of SVM and Naive Bayes classifier for the SEMG signal classification. In *Materials Today: Proceedings* (Vol. 37, pp. 3241–3245). Elsevier Ltd. https://doi.org/10.1016/j.matpr.2020.09.093

Paiman, Ardiyanta, Ansar, M., Effendy, I., & Sumbodo, B. T. (2020). Rice cultivation of superior variety in swamps to increase food security in Indonesia. *Reviews in Agricultural Science*. Gifu University - United Graduate School of Agricultural Science. https://doi.org/10.7831/ras.8.0_300

Rácz, A., Bajusz, D., & Héberger, K. (2021). Effect of dataset size and train/test split ratios in qsar/qspr multiclass classification. *Molecules*, *26*(4). https://doi.org/10.3390/molecules26041111

Rico-Chávez, A. K., Franco, J. A., Fernandez-Jaramillo, A. A., Contreras-Medina, L. M., Guevara-González, R. G., & Hernandez-Escobedo, Q. (2022, April 1). Machine Learning for Plant Stress Modeling: A Perspective towards Hormesis Management. *Plants*. MDPI. https://doi.org/10.3390/plants11070970

Riyadi, S., Siregar, M. M., Margolang, K. fadhli F., & Andriani, K. (2022). ANALYSIS OF SVM AND NAIVE BAYES ALGORITHM IN CLASSIFICATION OF NAD LOANS IN SAVE AND LOAN COOPERATIVES. *JURTEKSI (Jurnal Teknologi Dan Sistem Informasi)*, *8*(3), 261–270. https://doi.org/10.33330/jurteksi.v8i3.1483

Saptadi, N. T. S., Suyuti, A., Ilham, A. A., & Nurtanio, I. (2023). Modeling of Organic Waste Classification as Raw Materials for Briquettes using Machine Learning Approach. *International Journal of Advanced Computer Science and Applications*, *14*(3), 577–585. https://doi.org/10.14569/IJACSA.2023.0140367

Tharwat, A. (2021). Classification assessment methods. *Applied Computing and Informatics*, *17*(1), 168–192.https://doi.org/10.1016/j.aci.2018.08.003

Toulotte, J. M., Pantazopoulou, C. K., Sanclemente, M. A., Voesenek, L. A. C. J., & Sasidharan, R. (2022, February 1). Water stress resilient cereal crops: Lessons from wild relatives. *Journal of Integrative Plant Biology*. John Wiley and Sons Inc. https://doi.org/10.1111/jipb.13222

Verma, V., Vishal, B., Kohli, A., & Kumar, P. P. (2021, November 1). Systems-based rice improvement approaches for sustainable food and nutritional security. *Plant Cell Reports*. Springer Science and Business Media Deutschland GmbH. https://doi.org/10.1007/s00299-021-02790-6

Wang, X. (2022, April 1). Managing Land Carrying Capacity: Key to Achieving Sustainable Production Systems for Food Security. *Land*. MDPI. https://doi.org/10.3390/land11040484

Yasar, A., & Saritas, M. M. (2019). Performance Analysis of ANN and Naive Bayes Classification Algorithm for Data Classification. *International Journal of Intelligent Systems and Applications in Engineering*, *7*(2), 88–91.