Indonesian Journal of Data and Science

Volume 4 Issue 2
ISSN 2715-9936
https://doi.org/10.56705/ijodas.v4i2.76

Research Article

Comparison Analysis of Classification Model Performance in Lung Cancer Prediction Using Decision Tree, Naive Bayes, and Support Vector Machine

Dewi Widyawati [1,*], Amaliah Faradibah [2], Poetri Lestari Lokapitasari Belluano [3]

[1] Universitas Muslim Indonesia, Makassar, Indonesia, dewiwidyawati@umi.ac.id
[2] Universitas Muslim Indonesia, Makassar, Indonesia, amaliah.faradibah@umi.ac.id
[3] Universitas Muslim Indonesia, Makassar, Indonesia, poetrilestari@umi.ac.id

Correspondence should be addressed to Dewi Widyawati; dewiwidyawati@umi.ac.id

Received 10 June 2023; Accepted 18 June 2023; Published 31 July 2023

Abstract:

This research aims to analyze the performance of three classification models, namely Decision Tree Classifier, Support Vector Machine, and Naive Bayes Classifier, in predicting lung cancer using the "Lung Cancer Prediction" dataset. The performance evaluation metrics used include accuracy, precision weighted, recall weighted, and F1 weighted. As a preliminary step, exploratory data analysis (EDA) and dataset preprocessing, including feature selection, data cleaning, and data transformation, were conducted. The test data results showed that the Decision Tree Classifier and Naive Bayes Classifier had similar performances with high accuracy, precision, recall, and F1 values. Meanwhile, the Support Vector Machine also exhibited competitive performance, although its precision weighted value was slightly lower. Additionally, an outlier analysis was conducted using box plots, revealing that the Decision Tree Classifier had 2 outlier values, while the Support Vector Machine had 4 outlier values, and Naive Bayes had no outlier values. In conclusion, all three classification models demonstrated good potential in lung cancer prediction. However, selecting the best model requires consideration of relevant evaluation metrics for the application and accommodating the limitations of each model. Further evaluation and in-depth analysis are needed to ensure the reliability of the models in predicting lung cancer cases more accurately and consistently.

Keywords: *decision tree classifier*, *Naïve Bayes Classifier*, *support vector machine*, *classification*, perbandingan performa, prediction.
Dataset link: Lung Cancer

1. Introduction

Lung cancer is a serious and highly impactful public health issue. The high rates of morbidity and mortality caused by lung cancer highlight the importance of prevention efforts, early detection, and appropriate treatment [1], [2]. In the current digital and information technology era, the use of machine learning-based classification models is becoming increasingly attractive in the medical field, particularly in lung cancer prediction. Classification models enable us to identify risk factors and classify individuals based on specific features to predict the risk of developing lung cancer.

The main objective of this research is to compare the performance of three classification algorithms, namely Decision Tree, Naive Bayes, and Support Vector Machine, in predicting lung cancer. By comparing these three algorithms, it is expected that the most suitable algorithm for classifying clinical data in the context of lung cancer prediction will be identified.

In this research, we have several research questions to be answered. First, we want to know which algorithm, Decision Tree, Naive Bayes, or Support Vector Machine, is more effective in predicting lung cancer based on the clinical data we have. We will compare the performance of these three algorithms to determine which one provides the best prediction results in the context of lung cancer prediction. Additionally, we are interested in identifying specific clinical features that have a significant influence on lung cancer prediction. These features may provide additional insights into risk factors contributing to lung cancer and may be crucial knowledge for early diagnosis and more effective treatment. Lastly, we will compare the performance of the three algorithms in terms of accuracy, precision, recall, and F1 score. The results of this comparison will provide guidance to medical professionals in selecting the most suitable method to support efforts in lung cancer prevention and management more efficiently and accurately.

This research will use the "Lung Cancer Prediction" dataset, which contains clinical data and features related to lung cancer. However, this research has limitations in terms of the number of samples and features available in the dataset. Additionally, the analysis is conducted retrospectively using existing data, so the results may depend on the quality and availability of the data.

This research is expected to contribute to the development of more effective lung cancer prediction approaches using classification models. The results of the performance comparison of the three algorithms can provide guidance to medical professionals in selecting the most suitable method to support the diagnosis and treatment of lung cancer patients.

Method:

Our research is designed in five well-structured main stages, and their aspects are illustrated in **Figure 1**.

Figure 1. General Research Design Stages

Exploratory Data Analysis

Firstly, the initial step in this research is to conduct exploratory data analysis. At this stage, lung cancer data will be analyzed descriptively to understand the characteristics, distributions, and relationships between variables in the dataset. Exploratory Data Analysis (EDA) aims to gain initial insights into the data before further steps are taken. **Table 1** shows general information on the dataset used in this study.

Table 1. Dataset Information

Dataset	***Number of cases***	***Number of attribute***	***Attribute characteristics***	***Missing values***
Lung-cancer	309	16	*Numeric, String*	*No*

Dataset Preprocessing

This stage is part of the process where data is processed according to the methods and tools used. Dataset preprocessing involves several steps, including feature selection, data cleaning, and data transformation. **Figure 2** visualizes the steps of dataset preprocessing.

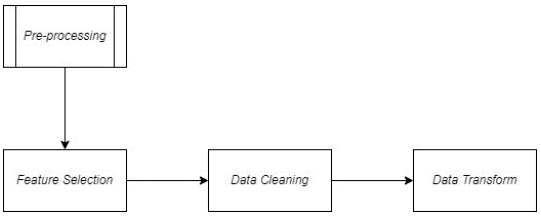


Figure 2. Data Preprocessing

a. *Feature Selection*
   At this stage, irrelevant or non-significant features or variables in predicting lung cancer will be identified and removed from the dataset. The feature selection process aims to simplify the dataset and improve the model's performance.
b. *Data Cleaning*
   Incomplete, duplicate, or noisy data will be cleaned at this stage. Data cleaning is essential to ensure the quality of the data used in modeling.
c. *Data Transform*
   Data transformation processes include feature normalization or standardization, so that feature values are in a similar scale. This step helps avoid dominance of features with large value ranges and can enhance model stability and convergence.

Support Vector Machine

SVM (Support Vector Machine) is a machine learning algorithm used for data modeling and classification. This algorithm can be applied to binary and multiclass classification problems [3]-[7], as shown in **Figure 3**.

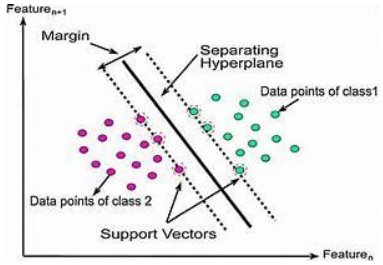


Figure 3. Support Vector Machine (SVM) Algorithm

The SVM algorithm involves several main stages. First, the training data undergoes preprocessing such as normalization and noise removal [8]. Next, a kernel is selected to map the data into a higher feature space to better separate the data. Then, the training data is used to train SVM by finding the optimal hyperplane that maximizes the margin [9]-[11]. The margin is the distance between the hyperplane and the nearest supporting vectors, and the parameter C is used to control the trade-off between margin and classification errors. After the training process, SVM is ready to classify new data based on its position relative to the hyperplane determined during training, where data on one side of the hyperplane is considered a member of one class, while data on the other side is considered a member of a different class.

$$Dataset = (x_I, y_i) \quad (1)$$

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

Naïve Bayes Classifier

The Naive Bayes Classifier is a probabilistic classification algorithm that uses the assumption of independent features [12], [13]. Despite this assumption rarely being met, Naive Bayes remains a frequently used and effective choice, as seen in **Figure 4**.

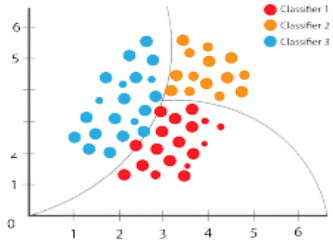


Figure 4. Naive Bayes Classifier

The Naive Bayes algorithm is a statistical method in machine learning for classification with the following steps: calculating prior probabilities, likelihood probabilities, and posterior probabilities using Bayes' rule [14]-[16] as shown in Equations 3, 4, 5, and 6. Based on posterior probabilities, the algorithm predicts the most likely class based on the observed features as shown in Equation 8. This algorithm is efficient and commonly used in various data classification applications.

$$P(Y = y) = \frac{count(Y = y)}{count(Y)} \tag{3}$$

$$P(Xi = xi|Y = y) = \frac{count(Xi - xi, Y = y)}{count(Y = y)} \tag{4}$$

$$P(Y = y|X = x) = \frac{P(X = x|Y = y)P(Y = y)}{P(X = x)} \tag{5}$$

$$y\,pred = argmax\, yP(Y = y|X = x) \tag{6}$$

Decision Tree Classifier

The Decision Tree is a classification method in the form of a tree structure with nodes representing decisions or predictions [17]-[19]. At each node, the algorithm divides the data based on the most informative input variables, as shown in **Figure 5**, explaining the concept of the Decision Tree algorithm.

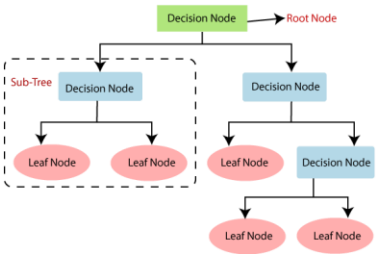


Figure 5. Decision Tree Algorithm

The Decision Tree Classifier algorithm is a classification method that uses a tree structure with nodes representing decisions or predictions. This decision tree is built by dividing the data into smaller subsets based on the value of the most informative input variable [20], [21]. The tree-building process uses impurity measures such as Gini Index and Entropy to measure data impurity at each node. The goal is to minimize impurity at each node by selecting the most relevant input variables, allowing the decision tree to provide accurate and easily interpretable predictions. Equations 7 and 8 can be observed.

$$Gini\ Index: Gini(t) = 1 - \sum_{i=1}^{c} (pi)^2 \quad (7)$$

$$Entropy: Entropy(t) = -\sum_{i}^{c} p_i \log_2(pi) \quad (8)$$

Performance Comparison Analysis

The performance analysis results from the previous steps are used to compare the performance of three different classification models, namely Decision Tree, Support Vector Machine, and Naive Bayes, in predicting lung cancer based on predetermined evaluation metrics [22]-[24]. The evaluation metrics used include accuracy, which measures how well the model can classify data overall (Equation 9); precision, which measures how well the model can correctly identify positives compared to all its positive predictions (Equation 10); recall (Sensitivity or True Positive Rate), which measures how well the model can correctly classify true positives (Equation 11); and F-Measure, which is the harmonic mean of precision and recall and is used to combine these two metrics into a comprehensive value (Equation 12). By considering the evaluation results through these metrics, we can identify the model that provides the best performance in lung cancer prediction [25], [26].

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \quad (9)$$

$$Pericision = \frac{TP}{(TP + FP)} \quad (10)$$

$$Recall = \frac{TP}{(TP + FN)} \quad (11)$$

$$F - measure = \frac{2(presisi \times recall)}{(presisi + recall)} \quad (12)$$

The above formulas explain:
True Positive (TP): The number of cases correctly predicted as positive by the model.
True Negative (TN): The number of cases correctly predicted as negative by the model.
False Positive (FP): The number of cases incorrectly predicted as positive by the model.
False Negative (FN): The number of cases incorrectly predicted as negative by the model.

Using evaluation metrics such as accuracy, precision, recall, and F-Measure, we can evaluate the performance of each classification model (Decision Tree, Support Vector Machine, and Naive Bayes) and select the best model that can provide optimal accuracy and reliability in lung cancer prediction. To perform the performance evaluation, a confusion matrix is used as a tool to compare the model's predictions with the true labels on test data, resulting in four main components: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). Detailed information about these components can be found in **Figures 6**, **7**, and **8**. Thus, through performance analysis using evaluation metrics and information from the confusion matrix, we can identify the most suitable and reliable model in predicting lung cancer based on the tested data. Selecting the appropriate model will contribute to the development of more accurate and efficient diagnostic methods in the management and treatment of lung cancer.

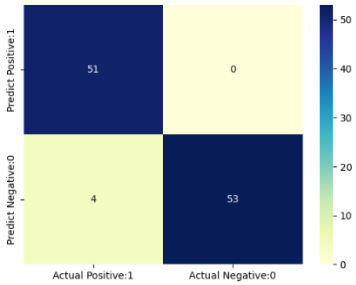


Figure 6. Confusion Matrix for Test Data of Lung Cancer Prediction using Decision Tree Classifier

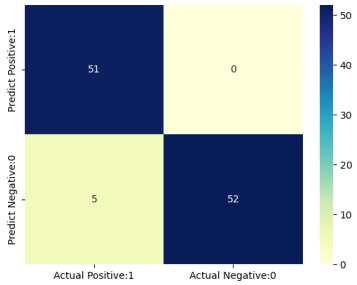


Figure 7. Confusion Matrix for Test Data of Lung Cancer Prediction using Support Vector Machine

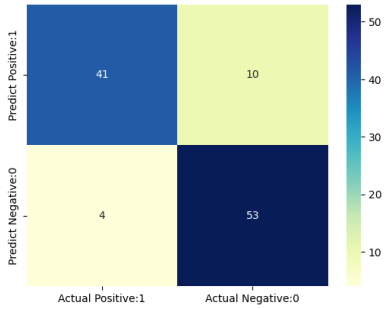


Figure 8. Confusion Matrix for Test Data of Lung Cancer Prediction using Naive Bayes Classifier

Based on the Confusion matrix results of the three classification models (Decision Tree Classifier, Support Vector Machine, and Naive Bayes Classifier) on testing using test data for lung cancer prediction (Lung Cancer Prediction), we conclude that the Decision Tree Classifier shows good performance with high True Positive (TP) and True Negative (TN) values, accurately identifying positive (lung cancer) and negative (non-lung cancer) cases. The Support Vector Machine also shows promising results, with significant TP and TN values, demonstrating good ability in classifying test data. Meanwhile, the Naive Bayes Classifier also provides competitive performance, with adequate TP and TN values for recognizing positive and negative cases. All three classification models show good potential in lung cancer prediction based on the Confusion matrix results. However, to choose the best model among them, further evaluation with more comprehensive performance metrics such as accuracy, precision, recall, and F1-score needs to be done. Additionally, selecting the best model should also consider the application context and practical needs in lung cancer management and treatment.

Decision Making

Based on the results of the performance comparison analysis, a decision will be made to select the best classification model that is most suitable for lung cancer prediction. This decision will provide guidance to medical practitioners or other stakeholders in choosing an effective model to support the diagnosis and management of lung cancer more effectively.

3. Results and Discussion

In this study, we conducted a performance comparison of three types of classification models, namely Decision tree classifier, Support vector machine (SVM), and Naïve Bayes Classifier, in predicting a specific target classification. The results of these models' performance evaluations were measured using several evaluation metrics, including accuracy, precision, recall, and F1-score using the "weighted" method to account for class imbalance. **Table 2** shows the performance comparison results of the datasets used.

Table 2. Performance Comparison Results

∑ *Rata-rata*	*Decision tree classifier*	*Support vector machine*	*Naïve Bayes Classifier*
Accuracy	0.89	0.87	0.89
Precision weighted	0.88	0.75	0.9

∑ *Rata-rata*	*Decision tree classifier*	*Support vector machine*	*Naïve Bayes Classifier*
Recall weighted	0.86	0.87	0.89
F1 weighted	0.86	0.81	0.89

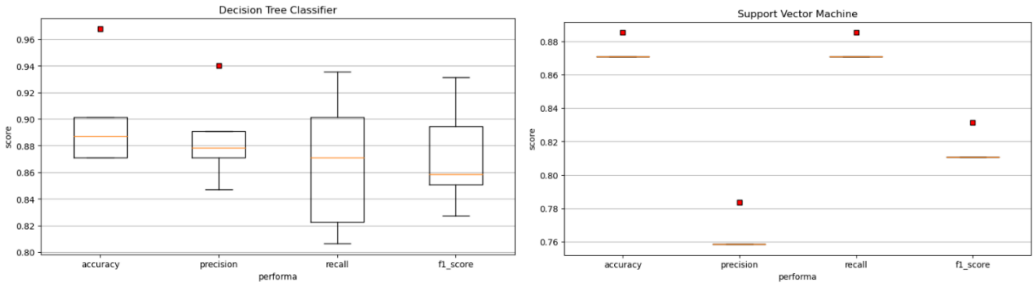
The performance evaluation results of the three classification models (Decision tree classifier, Support vector machine, and Naïve Bayes Classifier) for lung cancer prediction are displayed in the table above. In testing using test data, the average accuracy values were 0.89 for both Decision tree classifier and Naïve Bayes Classifier, while Support vector machine achieved 0.87. As for the weighted precision metric, Naïve Bayes Classifier showed the highest performance with a score of 0.9, followed by Decision tree classifier with 0.88, and Support vector machine with 0.75. However, in the weighted recall metric, both Decision tree classifier and Naïve Bayes Classifier had the same value of 0.86, while Support vector machine achieved 0.87. Similarly, for the F1 weighted metric, Naïve Bayes Classifier showed the best performance with a score of 0.89, followed by Decision tree classifier with 0.86, and Support vector machine with 0.81.

Based on the performance evaluation results of the three classification models, it can be concluded that both Decision tree classifier and Naïve Bayes Classifier have similar performance and show good predictive ability for lung cancer. Both models have relatively high values for accuracy, weighted precision, weighted recall, and weighted F1, indicating their good ability to accurately classify lung cancer and non-lung cancer cases. On the other hand, the Support vector machine also showed good performance, especially in the weighted recall and F1 metrics. However, its weighted precision performance was lower compared to the other models, indicating its limitations in identifying lung cancer cases.

The best model selection for lung cancer prediction needs to consider the most relevant evaluation metrics in the application context. If the primary focus is on detecting lung cancer cases as accurately as possible, the Naïve Bayes Classifier could be chosen because it has the highest weighted precision value. However, if a balance between precision and recall is prioritized, the Decision tree classifier could be a good alternative with a competitive weighted F1 value.

Additionally, these results also highlight the importance of evaluating model performance using various performance metrics to obtain a comprehensive understanding of the model's capability in predicting lung cancer cases. To enhance model performance, research can be conducted by optimizing parameters and features used, and combining various ensemble techniques to improve prediction accuracy and stability.

The obtained performance results indicate that the Decision tree classifier has 2 outlier values, which significantly differ from most of the other data. Outliers can affect model performance since these extreme values could be noise or unusual information that disrupts the learning process. In the Support vector machine model, there are 4 outlier values that need attention as they can affect the model's accuracy and stability. However, unlike the previous two models, the Naive Bayes did not have outlier values when visualized using a boxplot, indicating more homogeneous data in classification, which could be one reason why this model showed competitive performance in predicting lung cancer. Nonetheless, further analysis is still needed to understand the impact of outlier values and other factors that can affect the performance results of each model. The data visualization results using a boxplot can be seen in **Figure 9**.



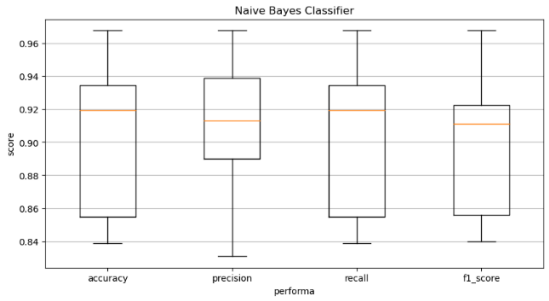


Figure 9. Bloxplot Visualization Results

4. Conclusion

Based on the entire performance evaluation results of the three classification models (Decision tree classifier, Support vector machine, and Naïve Bayes Classifier) in predicting lung cancer, it can be concluded that all three models show good potential in this task. Both Decision tree classifier and Naïve Bayes Classifier have similar performance with high values for accuracy, precision, recall, and F1. Meanwhile, the Support vector machine also produced competitive performance, even though its weighted precision value was slightly lower. The best model selection needs to consider the most relevant evaluation metrics for application needs and accommodate the limitations of each model. Further evaluation with more comprehensive performance metrics and advanced analysis needs to be conducted to gain a deeper understanding and ensure the model's reliability in predicting lung cancer cases.

References:

[1] G. Sruthi, C. L. Ram, M. K. Sai, B. P. Singh, and ..., "Cancer prediction using machine learning," ... *in Technology and ...*, 2022, [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9754059/

[2] Y. Zhou, C. Zhang, and S. Gao, "Breast cancer classification from histopathological images using resolution adaptive network," *IEEE Access*, 2022, [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9745527/

[3] H. Hamdani, H. R. Hatta, N. Puspitasari, and ..., "Dengue classification method using support vector machines and cross-validation techniques," ... *Journal of Artificial ...*, 2022, [Online]. Available: https://search.proquest.com/openview/a607c8361a7aac70dfc0dabf2b63f41b/1?pq-origsite=gscholar&cbl=1686339

[4] A. Roy and S. Chakraborty, "Support vector machine in structural reliability analysis: A review," *Reliability Engineering &System Safety*, 2023, [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0951832023000418

[5] H. Hafdaoui, A. Chahtou, S. Bouchakour, and ..., "Analyzing the performance of photovoltaic systems using support vector machine classifier," ... *Energy, Grids and ...*, 2022, [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352467721001508

[6] T. A. Mutiara and Q. N. Azizah, "Klasifikasi Tumor Otak Menggunakan Ekstraksi Fitur HOG dan Support Vector Machine," *Jurnal Infortech*, 2022, [Online]. Available: https://ejournal.bsi.ac.id/ejurnal/index.php/infortech/article/view/12813

[7] H. N. Mahendra and ..., "An efficient classification of hyperspectral remotely sensed data using support vector machine," *International Journal of ...*, 2022, [Online]. Available: https://yadda.icm.edu.pl/baztech/element/bwmeta1.element.baztech-17005b9c-52b2-4dc3-9618-036c7b97d6f9

[8] F. A. Satria, A. Abdiansah, and A. S. Utami, *Deteksi Domain Tidak Relevan (Out-Of-Domain) Pada Chatbot Berbahasa Indonesia Menggunakan Algoritma Support Vector Machine*. repository.unsri.ac.id, 2022. [Online]. Available: https://repository.unsri.ac.id/72916/

[9] W. Sun and J. Zhang, "A novel carbon price prediction model based on optimized least square support vector machine combining characteristic-scale decomposition and phase space …," *Energy*, 2022, [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0360544222010702

[10] T. Adugna, W. Xu, and J. Fan, "Comparison of random forest and support vector machine classifiers for regional land cover mapping using coarse resolution FY-3C images," *Remote Sens (Basel)*, 2022, [Online]. Available: https://www.mdpi.com/2072-4292/14/3/574

[11] A. Fatihin, D. Khairani, S. U. U. Masruroh, and ..., "Public Sentiment on User Reviews about Application in Handling COVID-19 using Naive Bayes Method and Support Vector Machine," *... on Science and ...*, 2022, [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9829068/

[12] B. Imran, H. Hambali, A. Subki, and ..., "Data Mining Using Random Forest, Naïve Bayes, and Adaboost Models for Prediction and Classification of Benign and Malignant Breast Cancer," *Jurnal Pilar Nusa ...*, 2022, [Online]. Available: http://ejournal.nusamandiri.ac.id/index.php/pilar/article/view/2912

[13] N. Deepa, J. S. Priya, and T. Devi, "Towards applying internet of things and machine learning for the risk prediction of COVID-19 in pandemic situation using Naive Bayes classifier for improving …," *Mater Today Proc*, 2022, [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2214785322016868

[14] A. Ainurrohmah and D. T. Wiyanti, "Analisis Performa Algoritma Decision Tree, Naive Bayes, K-Nearest Neighbor untuk Klasifikasi Zona Daerah Risiko Covid-19 di Indonesia," *Jurnal Teknologi Informasi dan Ilmu ...*, 2023, [Online]. Available: http://jtiik.ub.ac.id/index.php/jtiik/article/view/5935

[15] N. Attamami, A. Triayudi, and ..., "Analisis Performa Algoritma Klasifikasi Naive Bayes Dan C4. 5 Untuk Prediksi Penerima Bantuan Jaminan Kesehatan," *Jurnal Jtik (Jurnal ...*, 2023, [Online]. Available: http://journal.lembagakita.org/index.php/jtik/article/view/756

[16] T. Nugraha, *Analisis Sentimen Respons Masyarakat Terhadap Kartu Prakerja Menggunakan Algoritma K-Nn, Naïve Bayes Dan Svm*. repository.mercubuana.ac.id, 2022. [Online]. Available: https://repository.mercubuana.ac.id/70599/

[17] M. Kiguchi, W. Saeed, and I. Medi, "Churn prediction in digital game-based learning using data mining techniques: Logistic regression, decision tree, and random forest," *Appl Soft Comput*, 2022, [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1568494622000436

[18] W. Gao *et al.*, "Prediction of acute kidney injury in ICU with gradient boosting decision tree algorithms," *Computers in biology and ...*, 2022, [Online]. Available: https://www.sciencedirect.com/science/article/pii/S001048252100891X

[19] R. Guo, D. Fu, and G. Sollazzo, "An ensemble learning model for asphalt pavement performance prediction based on gradient boosting decision tree," *International Journal of Pavement ...*, 2022, doi: 10.1080/10298436.2021.1910825.

[20] L. M. Sotarjua And D. B. Santoso, "Perbandingan Algoritma Knn, Decision Tree,* Dan Random* Forest Pada Data Imbalanced Class Untuk Klasifikasi Promosi Karyawan," *... Informatika Sains dan ...*, 2022, [Online]. Available: https://journal3.uin-alauddin.ac.id/index.php/instek/article/view/31385

[21] M. H. Setiono, "A Komparasi Algoritma Decision Tree, Random Forest, Svm Dan K-Nn Dalam Klasifikasi Kepuasan Penumpang Maskapai Penerbangan," *Inti Nusa Mandiri*, 2022, [Online]. Available: https://ejournal.nusamandiri.ac.id/index.php/inti/article/view/3420

[22] R. A. Zuama, S. Rahmatullah, and ..., "Analisis Performa Algoritma Machine Learning pada Prediksi Penyakit Cerebrovascular Accidents," *Jurnal Media ...*, 2022, [Online]. Available: http://www.stmik-budidarma.ac.id/ejurnal/index.php/mib/article/view/3488

[23] E. Apriliyani and Y. Salim, "Analisis performa metode klasifikasi Naïve Bayes Classifier pada Unbalanced Dataset," *Indonesian Journal of Data and Science*, 2022, [Online]. Available: https://jurnal.yoctobrain.org/index.php/ijodas/article/view/45

[23] E. Apriliyani and Y. Salim, "Analisis performa metode klasifikasi Naïve Bayes Classifier pada Unbalanced Dataset," Indonesian Journal of Data and Science, 2022, [Online]. Available: https://jurnal.yoctobrain.org/index.php/ijodas/article/view/45.

[24] M. Khushi, "A Comparative Performance Analysis of Data Resampling Methods on Imbalance Medical Data," *IEEE Access*, vol. 9, pp. 109960–109975, 2021, doi: 10.1109/ACCESS.2021.3102399.

[25] S. Rahman, "Performance analysis of boosting classifiers in recognizing activities of daily living," *Int. J. Environ. Res. Public Health*, vol. 17, no. 3, 2020, doi: 10.3390/ijerph17031082.

[26] P. Sharma, "Performance analysis of deep learning CNN models for disease detection in plants using image segmentation," *Inf. Process. Agric.*, vol. 7, no. 4, pp. 566–574, 2020, doi: 10.1016/j.inpa.2019.11.001.