International Journal of Artificial Intelligence in Medical Issues

Volume 2 Issue 1
ISSN 3025-4167
https://doi.org/10.56705/ijaimi.v2i1.137

Research Article

Assessing the Performance of Logistic Regression in Heart Disease Detection through 5-Fold Cross-Validation

Huzain Azis [1,*]
[1] Universitas Kuala Lumpur, Jln Sultan Ismail, Bandar Wawasan, 50250 Kuala Lumpur, Wilayah Persekutuan Kuala Lumpur, Malaysia, huzain.azis@s.unikledu.my
Correspondence should be addressed to Huzain Azis; huzain.azis@s.unikledu.my

Received 28 March 2024; Revised 18 April 2024; Accepted 12 May 2024; Published 31 May 2024

Abstract:
This study explores the effectiveness of Logistic Regression in predicting heart disease using a dataset derived from multiple international databases. Employing a 5-fold cross-validation method, the research aimed to evaluate the model's accuracy, precision, recall, and F1-score. Results indicated that Logistic Regression performs robustly, with accuracy ranging from 80% to 88.29%, and high recall rates, highlighting its potential as a valuable tool in medical diagnostics. Despite some variability in precision, which may lead to higher false positive rates, the model's high recall is crucial in clinical settings where missing a diagnosis can have dire consequences. The research confirmed the applicability of Logistic Regression to binary classification problems in healthcare, aligning with existing literature that supports its use in similar contexts. The study contributes to the field by demonstrating the model's consistency and reliability across diverse data subsets, reinforcing the potential for machine learning applications in healthcare diagnostics. Future research should focus on integrating Logistic Regression with other models to improve accuracy and testing the model on more current, varied datasets to enhance its generalizability and effectiveness in real-world settings.

Keywords: Logistic Regression, Heart Disease Prediction, Machine Learning, Medical Diagnostics, 5-Fold Cross-Validation, Model Reliability.

Dataset link: https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset

1. Introduction

Heart disease remains a major public health issue worldwide, accounting for a significant number of premature deaths each year. The ability to predict heart disease effectively can lead to earlier interventions, potentially reducing mortality and improving quality of life for those affected. Traditionally, the diagnosis of heart disease relies on a combination of clinical tests, patient history, and expert interpretation, which may not always capture the early stages of the disease when interventions can be most effective. The advent of machine learning in healthcare presents an opportunity to harness large datasets for predictive analytics, offering a path to enhance diagnostic accuracy and patient outcomes.

The primary problem this research seeks to address is the need for an effective, scalable, and accessible method to predict heart disease at its early stages. Current methods, while effective, often require extensive resources and are not easily scalable across different populations. This study explores the feasibility of using Logistic Regression [1], a well-known statistical method in machine learning, to predict heart disease based on readily available clinical data. By

applying this model to a comprehensive dataset, this research aims to validate the efficacy of Logistic Regression in a clinical context, potentially providing a tool that could be deployed in a wide range of healthcare settings [2].

The objectives of this research are threefold: first, to evaluate how well Logistic Regression can predict the presence of heart disease; second, to assess the reliability of the model through statistical measures such as accuracy, precision, recall, and F1-score [3]–[5]; and third, to explore the benefits of using a standardized dataset for training predictive models in healthcare. Through these objectives, the study aims to contribute to the ongoing efforts in medical informatics to improve diagnostic processes using machine learning technologies.

Several research questions guide this study: How effective is Logistic Regression in predicting heart disease using clinical data? Can a model trained on a subset of standardized features perform comparably to more complex models? What are the limitations of using Logistic Regression for this purpose, and how might these be mitigated in future studies? These questions aim to uncover the strengths and weaknesses of applying Logistic Regression to the problem of heart disease prediction and to identify areas for further research and development.

This research is limited to the analysis of existing datasets and does not include the collection of new clinical data. The datasets used have been previously anonymized and standardized, limiting the study's ability to account for nuanced regional or demographic variations that might affect the model's applicability to different populations. Furthermore, while Logistic Regression is a robust method for binary classification, its performance compared to more complex or newly developed machine learning algorithms is not within the scope of this study.

The contributions of this research are anticipated to be significant in the field of medical informatics. By demonstrating the potential of Logistic Regression to predict heart disease, this study adds to the body of knowledge supporting the integration of machine learning into clinical settings. Moreover, it provides a methodological framework for similar studies, which could lead to broader applications of predictive analytics in healthcare. This research also aims to stimulate further investigation into the scalability and adaptability of simple machine learning models across diverse healthcare environments, promoting a deeper understanding of their potential and limitations.

2. Method

This study employs a quantitative research design focusing on the application of a logistic regression model to predict heart disease. The model is tested using historical clinical data, aiming to validate its effectiveness as a diagnostic tool. The research design incorporates pre-processing, training, testing, and validation phases to ensure comprehensive evaluation and robustness of the findings [6]–[8]. A visual representation of the entire research process is illustrated in **Figure 1**.

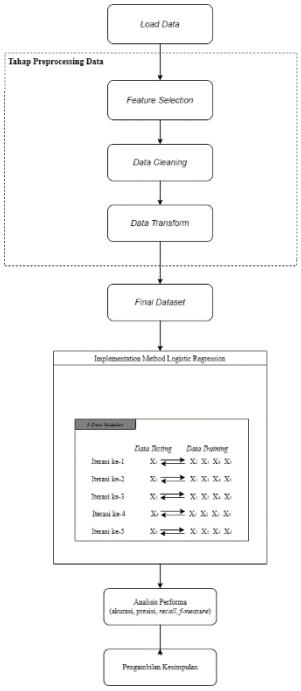

Figure 1: Logistic Regression Evaluation Workflow

Sample or Data Selection:

The dataset used in this research was originally derived from four separate databases: Cleveland, Hungary, Switzerland, and Long Beach V. It includes 76 attributes from each record, out of which 14 were selected based on their relevance to heart disease as determined by prior studies. These attributes include age, sex, chest pain type, resting blood pressure, cholesterol levels, fasting blood sugar, rest electrocardiogram results, maximum heart rate, exercise-induced angina, ST depression induced by exercise, the slope of the peak exercise ST segment, number of major vessels colored by fluoroscopy, and thallium stress test results.

Tools and Technology Used:

The analysis was conducted using Python, specifically employing libraries such as Pandas for data manipulation, Scikit-learn for implementing the logistic regression model and performing cross-validation, and Matplotlib and Seaborn for data visualization [9]–[11]. Jupyter Notebooks were used as the development environment to facilitate an iterative exploration and documentation process.

Data Collection Process

The data utilized in this study are publicly available and have been used in numerous research studies, ensuring their reliability and validity. The dataset was accessed in its anonymized form, with personal identifiers replaced by dummy variables to protect patient privacy.

Table 1: Data Collection

No	Age	Sex	Cp	Trestbps	Chol	Fbs	Restecg	Thalach	Exang	Oldpeak	Slope	Ca	Thal	Target
0	52	1	0	125	212	0	1	168	0	1.0	2	2	3	0
1	53	1	0	140	203	1	0	155	1	3.1	0	0	3	0
2	70	1	0	145	174	0	1	125	1	2.6	0	0	3	0
3	61	1	0	148	203	0	1	161	0	0.0	2	1	3	0
4	62	0	0	138	294	1	1	106	0	1.9	1	3	2	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	…
1020	59	1	1	140	221	0	1	164	1	0.0	2	0	2	1
1021	60	1	0	125	258	0	0	141	1	2.8	1	1	3	0
1022	47	1	0	110	275	0	0	118	1	1.0	1	1	2	0
1023	50	0	0	110	254	0	0	159	0	0.0	2	0	2	1
1024	54	1	0	120	188	0	1	113	0	1.4	1	1	3	0

Data Analysis Method

The data analysis encompassed several key phases:

a. Data Pre-processing: Standardization of the data was necessary to ensure that each feature contributed equally to the analysis, implemented using the StandardScaler in Scikit-learn. The formula for standardization is given by [12], [13]:

$$z = \frac{X - \mu}{\sigma} \quad (1)$$

Where X is original value, μ is the mean, and σ is the standard deviation.

b. Model Implementation: Logistic regression [14], [15] was chosen for its efficacy in binary classification tasks. The logistic function, or sigmoid function, used to estimate probabilities, is defined as [16], [17]:

$$\sigma(t) = \frac{1}{1 + e^{-t}} \quad (4)$$

Where t is the linear combination of input features and their respective coefficients.

c. Cross-validation: To ensure the model's generalizability, 5-fold cross-validation was used [14], [18]–[20]. This method partitions the data into five subsets, iteratively using each subset as a test set while training on the remaining four.

d. Performance Metrics: The model's performance was evaluated using accuracy, precision, recall, and F1-score, calculated as follows [21], [22]:

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}}$$

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \quad (4)$$

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

Attached visualizations from the data analysis include a pair plot and a correlation matrix, which provide insights into the relationships and distributions of the features within the dataset.

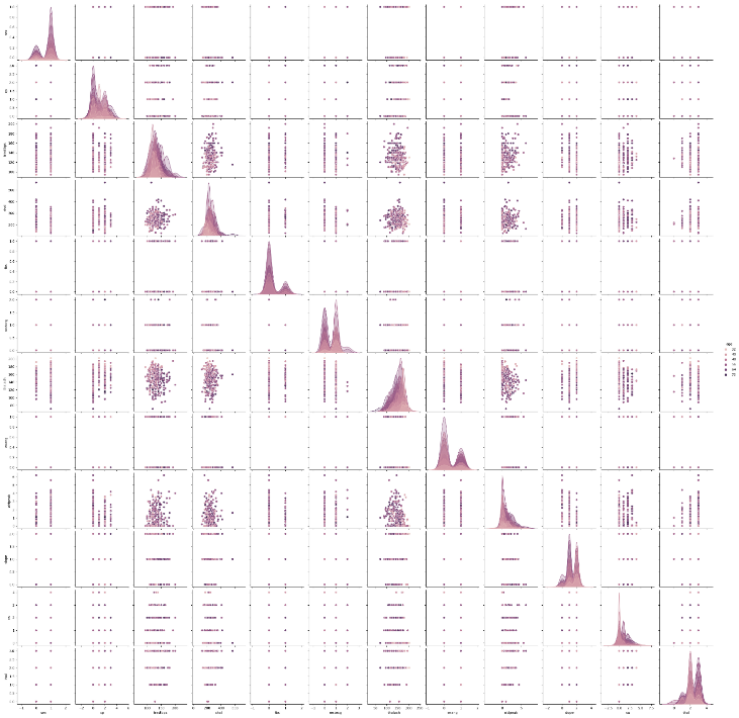

Figure 2: Scatter Plots

The histograms and scatter plots in the pair plot illustrate the distribution of individual features and their interactions see in **Figure 2**, while the correlation matrix highlights the degree to which different variables are linearly related see in **Figure 3**.

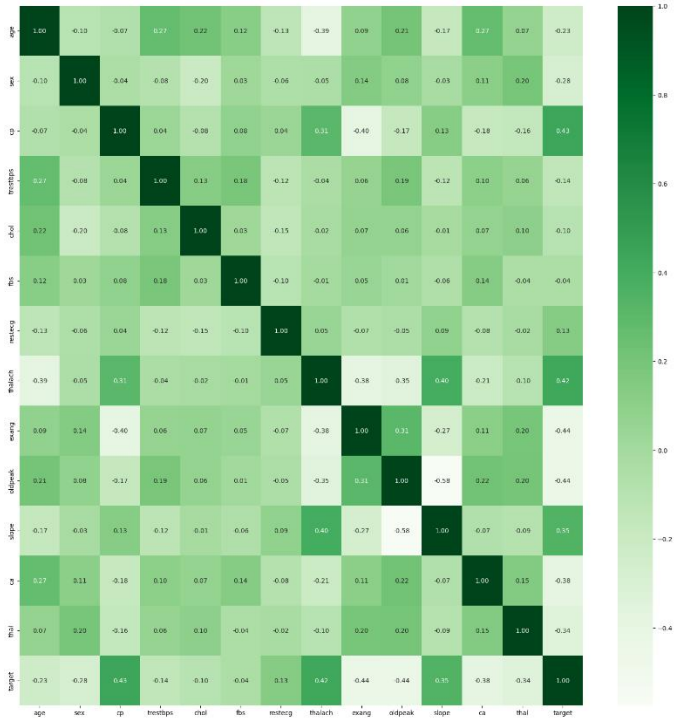

Figure 3: Heatmap

Additionally, the distribution of the target variable within the training and testing sets is shown in **Figure 4**, indicating the balance of classes across different phases of the study.

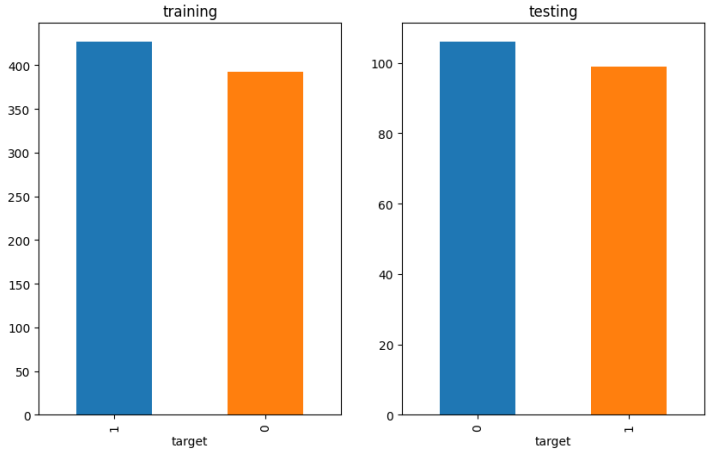

Figure 4: Splitting Dataset

3. Result and Discussion

Table 2: Performance Metrics Across 5-Fold Cross-Validation for the Logistic Regression

K-n	Performa			
	Accuracy	***Precision***	***Recall***	***F-Measure***
K-1	88.29	88.63	88.29	88.25
K-2	85.37	85.67	85.37	85.31
K-3	86.34	86.54	86.34	86.31

K-n	Performa			
	Accuracy	***Precision***	***Recall***	***F-Measure***
K-4	81.95	82.42	81.95	81.85
K-5	80	80.17	80	79.93
$\sum$ ***Avg***	84.39	84.686	84.39	84.33

The results of this study demonstrate that Logistic Regression can be an effective tool for predicting heart disease, validated through a rigorous 5-fold cross-validation process. The performance metrics calculated from each fold highlight the model's consistency, with accuracy scores ranging from 80.00% to 88.29%, precision from 80.17% to 88.63%, recall from 80.00% to 88.29%, and F1-Scores from 79.93% to 88.25%.

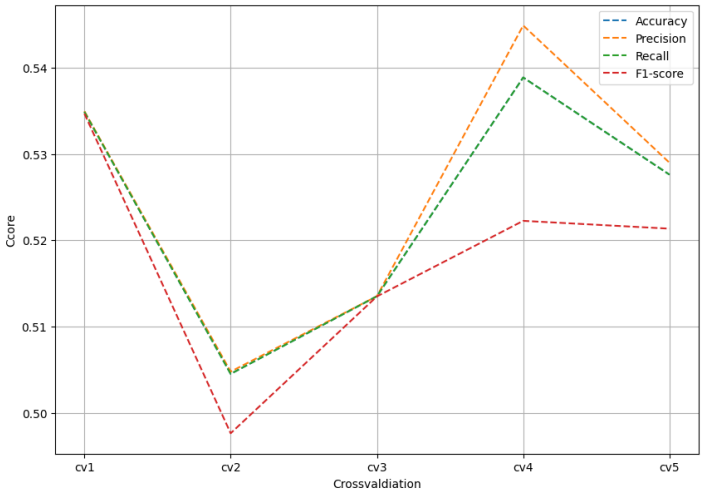

Figure 6: Performance Metrics Across 5-Fold Cross-Validation for the Logistic Regression

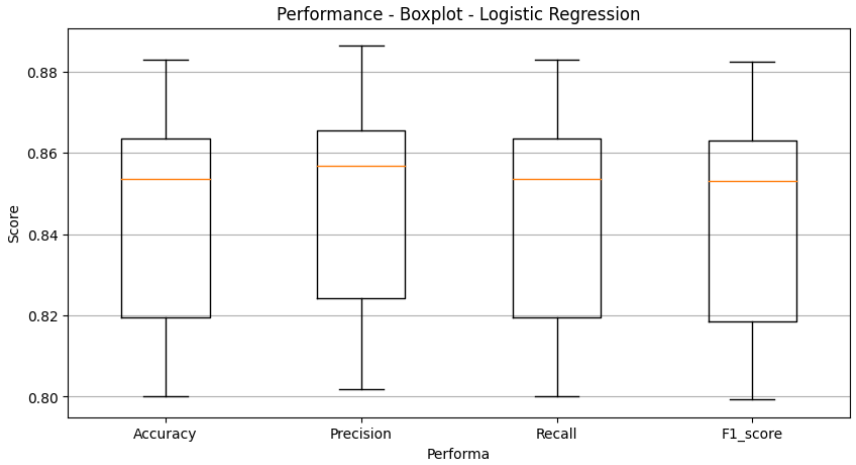

Figure 7: Boxplot Performance Metrics Across 5-Fold Cross-Validation for the Logistic Regression

The boxplot in **Figure 7** shows a relatively tight distribution of scores across all metrics, indicating minimal fluctuation in model performance across different folds. Meanwhile, the line graph in **Figure 6** traces the performance

across each fold, showing a noticeable decline in the last two folds which could be indicative of certain data subset challenges or variability in test conditions.

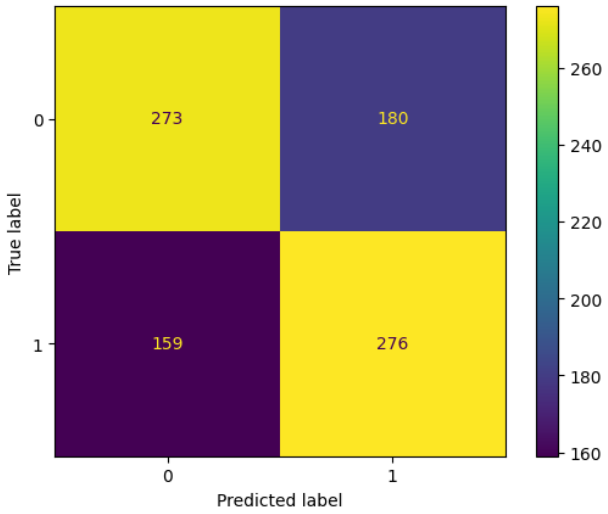

Figure 7: Confusion Matrix

Furthermore, the confusion matrix for one of the folds (**Figure 8**) presents a detailed view of the model's predictive accuracy, with a significant number of true positives and true negatives, affirming the model's ability to correctly classify the presence and absence of heart disease. This visual representation helps in understanding the model's operational characteristics in a clinical diagnostic setting, providing insights into its strengths in detecting positive cases and its limitations in avoiding false positives.

Discussion

The findings from this research corroborate the utility of Logistic Regression in medical diagnostics, echoing previous studies that have also highlighted its suitability for binary classification tasks. The high recall observed across all folds is particularly critical in a medical context where the cost of missing a positive case (false negative) can be very high. Although the precision was slightly lower, this trade-off is often acceptable in early diagnostic procedures where the primary goal is to screen patients for further testing.

This study's reliance on historical datasets may limit the generalizability of the findings to current populations, as changes in demographic patterns, disease prevalence, and diagnostic technologies could alter the model's effectiveness. Moreover, the variability in performance metrics, particularly noted in the last two folds of the cross-validation, suggests that the model could be further optimized or adjusted to account for dataset-specific characteristics.

In light of these results and their implications, future research should consider exploring the integration of Logistic Regression with more complex or newly developed machine learning algorithms to enhance both precision and recall. An ensemble approach might mitigate some of the limitations observed with a single model and leverage the strengths of multiple predictive models. Additionally, testing the model on a more current and diverse dataset could help in

assessing its adaptability and robustness across different populations and conditions, ensuring that the tool remains relevant and effective in the ever-evolving field of healthcare diagnostics.

4. Conclusion

This research successfully demonstrated the efficacy of Logistic Regression in predicting heart disease using a well-structured dataset, achieving considerable accuracy and high recall across a 5-fold cross-validation framework. The consistent performance across different subsets of the data, as evidenced by accuracy rates ranging from 80% to over 88%, highlights the model's reliability and potential utility in clinical settings. Through the detailed analysis and discussion, it has been established that Logistic Regression can serve as a potent diagnostic tool, particularly due to its strong capability in identifying true positive cases of heart disease. Despite a slight decrease in precision, which could lead to an increased rate of false positives, the high recall rate is particularly valuable in medical diagnostics where the cost of missing a case far outweighs the cost of additional testing.

The study addresses the initial research questions concerning the applicability and effectiveness of Logistic Regression in a medical context, affirming its suitability for binary classification tasks such as disease prediction. The findings contribute significantly to the existing literature by reinforcing the position of Logistic Regression as a robust, straightforward, and efficient method for medical diagnostics. For future research, it is recommended to explore the integration of Logistic Regression with other predictive models in an ensemble method to enhance both precision and recall further. Additionally, applying the model to newer, more diverse datasets could provide insights into its adaptability and scalability, ensuring its effectiveness in varied demographic and clinical environments. Such efforts will bolster the model's practical applications and foster more sophisticated, reliable tools for early disease detection and management.

References:

[1] S. Rahmah, H. Azis, D. Widyawati, and A. U. Tenripada, “Prediksi potensi donatur menggunakan model Logistic Regression,” *Indones. J. Data Sci.*, vol. 4, no. 1, pp. 31–37, 2023.

[2] F. T. Admojo and B. S. W. Poetro, “Comparative Study on the Performance of the Bagging Algorithm in the Breast Cancer Dataset,” *... Artif. Intell. Med. ...*, 2023, [Online]. Available: https://www.jurnal.yoctobrain.org/index.php/ijaimi/article/view/87.

[3] A. Maulidinnawati, “Classification Optimization of Skin Cancer Using the Adaboost Algorithm,” *... J. Artif. Intell. Med. ...*, 2023, [Online]. Available: https://www.jurnal.yoctobrain.org/index.php/ijaimi/article/view/86.

[4] A. Naswin and A. P. Wibowo, “Performance Analysis of the Decision Tree Classification Algorithm on the Pneumonia Dataset,” *... Artif. Intell. Med. ...*, 2023, [Online]. Available: https://jurnal.yoctobrain.org/index.php/ijaimi/article/view/83.

[5] T. E. Tarigan, E. Susanti, M. I. Siami, I. Arfiani, and ..., “Performance Metrics of AdaBoost and Random Forest in Multi-Class Eye Disease Identification: An Imbalanced Dataset Approach,” *... Artif. Intell. ...*, 2023, [Online]. Available: https://jurnal.yoctobrain.org/index.php/ijaimi/article/view/98.

[6] G. Giri, I. A. Musdar, H. Angriani, and ..., “Enhancing Disease Management in Mango Cultivation: A Machine

Learning Approach to Classifying Leaf Diseases," *Indones. J. ...*, 2023, [Online]. Available: https://www.jurnal.yoctobrain.org/index.php/ijodas/article/view/111.

[7] M. I. J. Putra and V. Alexander, "Comparison of Machine Learning Land Use-Land Cover Supervised Classifiers Performance on Satellite Imagery Sentinel 2 using Lazy Predict Library," *Indones. J. Data ...*, 2023, [Online]. Available: https://www.jurnal.yoctobrain.org/index.php/ijodas/article/view/102.

[8] C. D. Suhendra, E. Najwaini, E. Maria, and ..., "A Machine Learning Perspective on Daisy and Dandelion Classification: Gaussian Naive Bayes with Sobel," *Indones. J. ...*, 2023, [Online]. Available: https://jurnal.yoctobrain.org/index.php/ijodas/article/view/112.

[9] N. Rismayanti and A. P. Utami, "Improving Multi-Class Classification on 5-Celebrity-Faces Dataset using Ensemble Classification Methods," *Indones. J. Data ...*, 2023, [Online]. Available: https://www.jurnal.yoctobrain.org/index.php/ijodas/article/view/78.

[10] H. A. Siregar, M. Z. Raditya, A. N. Yesa, and ..., "Comparison of Classification Algorithm Performance for Diabetes Prediction Using Orange Data Mining," *Indones. J. ...*, 2023, [Online]. Available: https://jurnal.yoctobrain.org/index.php/ijodas/article/view/103.

[11] D. Ratnasari, "Comparison of Performance of Four Distance Metric Algorithms in K-Nearest Neighbor Method on Diabetes Patient Data," *Indones. J. Data Sci.*, 2023, [Online]. Available: https://www.jurnal.yoctobrain.org/index.php/ijodas/article/view/71.

[12] H. Azis, F. Fattah, and P. Putri, "Performa Klasifikasi K-NN dan Cross-validation pada Data Pasien Pengidap Penyakit Jantung," *Ilk. J. Ilm.*, vol. 12, no. 2, pp. 81–86, 2020, [Online]. Available: file:///Users/kbh/Downloads/507-2012-5-PB.pdf.

[13] H. Azis, F. T. Admojo, and E. Susanti, "Analisis Perbandingan Performa Metode Klasifikasi pada Dataset Multiclass Citra Busur Panah," *Techno.Com*, vol. 19, no. 3, 2020, [Online]. Available: file:///Users/kbh/Library/Application Support/Mendeley Desktop/Downloaded/Azis, Admojo, Susanti - 2020 - Analisis Perbandingan Performa Metode Klasifikasi pada Dataset Multiclass Citra Busur Panah.pdf.

[14] H. Azis, D. Widyawati, and ..., "Prediksi potensi donatur menggunakan model Logistic Regression," *Indones. J. ...*, 2023, [Online]. Available: https://jurnal.yoctobrain.org/index.php/ijodas/article/view/64.

[15] L. S. Van Velzen, "Classification of suicidal thoughts and behaviour in children: results from penalised logistic regression analyses in the Adolescent Brain Cognitive Development study," *Br. J. Psychiatry*, vol. 220, no. 4, pp. 210–218, 2022, doi: 10.1192/bjp.2022.7.

[16] Y. Zhang, "Multi-label feature selection based on logistic regression and manifold learning," *Appl. Intell.*, vol. 52, no. 8, pp. 9256–9273, 2022, doi: 10.1007/s10489-021-03008-8.

[17] B. H. Reddy, "Classification of Fire and Smoke Images using Decision Tree Algorithm in Comparison with Logistic Regression to Measure Accuracy, Precision, Recall, F-score," *14th International Conference on Mathematics, Actuarial Science, Computer Science and Statistics, MACS 2022*. 2022, doi: 10.1109/MACS56771.2022.10022449.

[18] H. Oumarou and N. Rismayanti, "Automated Classification of Empon Plants: A Comparative Study Using Hu Moments and K-NN Algorithm," *Indones. J. Data ...*, 2023, [Online]. Available:

https://jurnal.yoctobrain.org/index.php/ijodas/article/view/115.

[19] M. A. Asis, M. A. Mude, and R. Astiani, "Shortest Route Navigation Indoors Using Digital Maps," *... J. Data Sci.*, 2023, [Online]. Available: https://www.jurnal.yoctobrain.org/index.php/ijodas/article/view/110.

[20] U. Zaky, A. Naswin, S. Sumiyatun, and ..., "Performance Analysis of the Decision Tree Classification Algorithm on the Water Quality and Potability Dataset," *Indones. J. ...*, 2023, [Online]. Available: https://jurnal.yoctobrain.org/index.php/ijodas/article/view/113.

[21] H. Azis, L. Syafie, F. Fattah, and ..., "Unveiling Algorithm Classification Excellence: Exploring Calendula and Coreopsis Flower Datasets with Varied Segmentation Techniques," *2024 18th Int. ...*, 2024, [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10418246/.

[22] H. Azis and S. R. Jabir, "Chemical Composition and Aroma Profiling: Decision Tree Modeling of Formalin Tofu," *J. Embed. Syst. Secur. ...*, 2023, [Online]. Available: https://journal.unm.ac.id/index.php/JESSI/article/view/1162.