Indonesian Journal of Data and Science

Volume 5 Issue 1
ISSN 2715-9930
https://doi.org/10.56705/ijodas.v5i1.126

Research Article

Estimating Obesity Levels Using Decision Trees and K-Fold Cross-Validation: A Study on Eating Habits and Physical Conditions

Fadhila Tangguh Admojo [1,*], Nurul Rismayanti [2]

[1] Universiti Kuala Lumpur, Malaysia, fadhila.tangguh@s.unikl.edu.my
[2] Universitas Negeri Malang, Malang, Indonesia, nurulrismayanti.labfik@umi.ac.id

Correspondence should be addressed to Fadhila Tangguh Admojo; fadhila.tangguh@s.unikl.edu.my

Received 12 February 2024; Accepted 20 March 2024; Published 31 March 2024

© Authors 2024. CC BY-NC 4.0 (non-commercial use with attribution, indicate changes).
License: https://creativecommons.org/licenses/by-nc/4.0/ — Published by Indonesian Journal of Data and Science.

Abstract:

This study harnesses the predictive capabilities of machine learning to explore the determinants of obesity within populations from Mexico, Peru, and Colombia, using a Decision Tree algorithm bolstered by 5-fold cross-validation. Our comprehensive analysis of 2111 individuals' lifestyle and physical condition data yielded accuracy, precision, recall, and F1-scores that notably peaked in the third and fifth folds. The findings affirmed the significance of dietary habits and physical activity as substantial predictors of obesity levels. The variability in model performance across the folds underscored the importance of robust cross-validation in enhancing the model's generalizability. This research contributes to the burgeoning field of data science in public health by providing a viable model for obesity prediction and laying the groundwork for targeted health interventions. Our study's insights are pivotal for public health officials and policymakers, serving as a stepping stone towards more sophisticated, data-driven approaches to combating obesity. The study, however, recognizes the inherent limitations of self-reported data and the need for broader datasets that encompass more diverse variables. Future research directions include the analysis of longitudinal data to establish causal relationships and the comparison of various machine learning models to optimize predictive performance.

Keywords: Obesity Prediction, Decision Tree, Public Health, Lifestyle Factors, Health Interventions.
Dataset link: https://www.kaggle.com/datasets/aravindpcoder/obesity-or-cvd-risk-classifyregressorcluster/versions/1

1. Introduction

The prevalence of obesity has escalated into a global health crisis, with a marked increase in developing countries, posing significant challenges to public health systems. This alarming trend is attributed to a complex interplay of factors, including sedentary lifestyles, poor eating habits, and genetic predispositions. The consequences of obesity extend far beyond individual health, impacting societal well-being and economic stability through increased healthcare costs and reduced productivity. Thus, understanding the determinants of obesity and accurately estimating its levels in populations are critical for designing effective interventions. The research presented herein focuses on the countries of Mexico, Peru, and Colombia, regions where rapid urbanization and lifestyle changes have contributed to the rising obesity rates, making them ideal subjects for this study.

Addressing the obesity epidemic requires innovative approaches to identify and mitigate the contributing factors. Traditional methods of assessing obesity risk factors often rely on direct measurements and self-reported data, which, while useful, may not fully capture the complex interactions between lifestyle choices and obesity outcomes. This research aims to solve this problem by applying machine learning techniques, specifically Decision Trees, to analyse and predict obesity levels based on a comprehensive set of variables including eating habits, physical activity, and other relevant demographic data. By leveraging the predictive power of machine learning, this study seeks to uncover hidden patterns and relationships that contribute to obesity, offering a more nuanced understanding of its drivers.

The objectives of this research are multifaceted: to apply a Decision Tree algorithm to estimate obesity levels accurately; to evaluate the model's performance using 5-fold cross-validation [1]–[5]; and to analyse the influence of various lifestyle and demographic factors on obesity. Through this approach, the study aims to provide valuable insights into the most significant predictors of obesity, thereby informing targeted public health interventions and policies designed to combat this issue.

This study poses several research questions: What are the key predictors of obesity among the populations of Mexico, Peru, and Colombia? How accurately can a Decision Tree model, validated through 5-fold cross-validation, estimate obesity levels based on these predictors [6]–[9]. And what practical implications do the findings have for public health strategies and interventions aimed at reducing obesity rates?

The scope of this research is limited to the available dataset, which encompasses individuals aged 14 to 61 from Mexico, Peru, and Colombia, with data collected through a web-based survey. While this dataset provides a rich source of information on eating habits, physical activity, and demographic characteristics, it is important to acknowledge the limitations inherent in self-reported data, including potential biases and inaccuracies. Additionally, the study focuses on the application of a single machine learning model, the Decision Tree, which, despite its advantages, represents just one of many possible approaches to analysing obesity data.

Despite these limitations, the research contributes significantly to the field of public health by demonstrating the applicability of machine learning techniques to the study of obesity. It offers a novel approach to identifying and understanding the factors that contribute to obesity, providing a foundation for future studies and interventions. Furthermore, by highlighting the key predictors of obesity, the findings have the potential to guide public health officials and policymakers in designing more effective obesity prevention and management strategies, tailored to the specific needs and characteristics of the populations studied.

2. Method:

This study employs a quantitative research design, utilizing a cross-sectional dataset to investigate the predictors of obesity levels. The Decision Tree algorithm, a machine learning technique known for its ability to handle both categorical and numerical data, is applied to model the relationship between lifestyle factors and obesity. The model's performance is evaluated using 5-fold cross-validation, ensuring robustness and generalizability of the findings. This design facilitates the identification of significant predictors and their contribution to obesity levels, enabling the development of targeted interventions. Our research is designed in five well-structured main stages, and their aspects are illustrated in **Figure 1**.

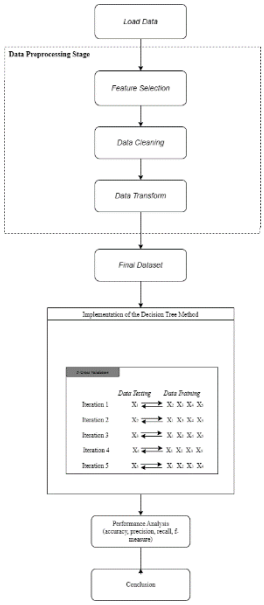


Figure 1. General Research Design Stages

Data Collection Process: **Mental Disorder Classification**

The dataset comprises 2111 individuals aged 14 to 61 from Mexico, Peru, and Colombia, selected through a web-based survey. The inclusion criteria ensured a diverse representation of eating habits, physical activity levels, and demographic variables. The data encompasses 17 attributes, including gender, age, height, weight, and specific lifestyle factors relevant to obesity, such as frequency of high caloric food consumption and physical activity frequency. Splitting dataset is presented in **Figure 2**

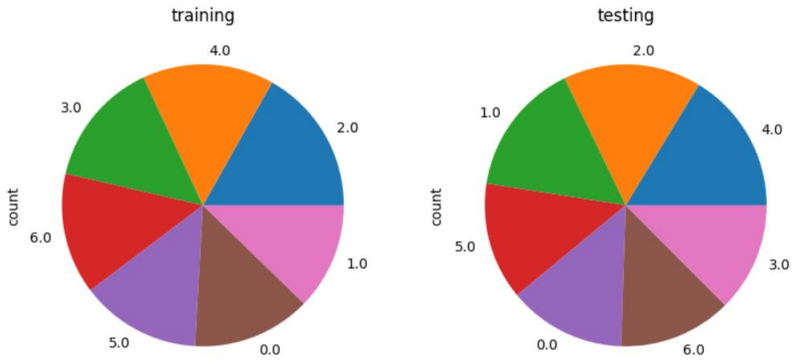


Figure 2. Splitting Dataset 10 % testing, 90% training

Tools and Technology Used

Data pre-processing, analysis, and modelling were conducted using Python, a programming language favoured for its extensive libraries and tools in data science. Key libraries employed include Pandas for data manipulation, Scikit-learn for machine learning model implementation and evaluation, and Matplotlib and Seaborn for data visualization. The Decision Tree model was developed using the DecisionTreeClassifier from Scikit-learn, while cross-validation was facilitated by the cross_val_score function.

Data Collection Process

Data was collected using an anonymous online survey, hosted on a web platform designed to capture a wide range of information related to eating habits, physical condition, and demographic details. Participants provided informed consent before participating, ensuring ethical standards were upheld. The survey was disseminated through social media and health forums, targeting a diverse population within the specified age range and countries.

Data Analysis Methods

Data Pre-processing

Data pre-processing involved encoding categorical variables into numeric formats to facilitate analysis with the Decision Tree algorithm [10]–[14]. This was achieved using the LabelEncoder and OneHotEncoder from Scikit-learn, converting text-based attributes into a form amenable to machine learning. Data pre-processing involved several key steps:

a. Encoding Categorical Variabels: Categorical variables were encoded to numerical values using one-hot encoding, facilitating their interpretation by the KNN algorithm. This is critical for attributes such as fruit type, which are inherently categorical.

$$X_{encoded} = one_hot_encode(X_{categorical}) \quad (1)$$

b. Handling Missing Values: Missing values were imputed based on the attribute type, ensuring no data point was left unutilized.

$$X_{imputed} = impute(X_{missing}) \quad (2)$$

c. Normalization: Attributes were normalized to ensure uniformity in scale, enhancing the KNN algorithm's efficiency.

$$X_{normalized} = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (3)$$

Classification Algorithm: Decision Tree

The Decision Tree model was trained to classify individuals into obesity levels based on their responses [15]–[20]. The algorithm constructs a tree-like model of decisions and their possible consequences, employing the formula [21]:

$$Gini\ Impurity = 1 - \sum_{i=1}^{n} P_i^2 \quad (4)$$

Where p_i is the proportion of samples that belong to class i for a particular mode.

Cross-Validation with K-Fold (K=5)

5-fold cross-validation was used to evaluate the model's performance, splitting the dataset into five equal parts, with each part serving as the test set while the remaining data constituted the training set [22]–[26]. This process was repeated five times, ensuring each fold served as the test set exactly once. The performance metrics, including accuracy, precision, recall, and F-measure, were computed using the formulas [27]–[32]:

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

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

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

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

Where TP, TN, FP and FN represent the numbers of true positives, true negatives, false positives, and false negatives, respectively.

3. Results and Discussion

Results

The analysis of obesity levels using a Decision Tree model augmented by 5-fold cross-validation yielded significant insights into the predictive power of lifestyle and physical condition factors. The model's performance varied across different folds, demonstrating an overall strong capability in accurately classifying individuals into correct obesity levels based on the predetermined factors. The accuracy metrics ranged from 81.09% in the first fold to 95.26% in the third and fifth folds, indicating the model's robustness in most scenarios. Precision, recall, and F1-score metrics followed a similar pattern, showcasing the model's efficiency in not only predicting obesity levels accurately but also in minimizing false positives and false negatives, essential for reliable health assessments.

The highest performance metrics observed in the third and fifth folds highlight the model's potential effectiveness in certain data distributions, suggesting that the Decision Tree algorithm can adapt well to the complexities of lifestyle and obesity data. These findings are encapsulated in a performance table that succinctly summarizes the model's efficacy across the five folds of cross-validation.

A summarization of the performance metrics across the 5 folds of cross-validation is presented in **Table 1** below, showcasing the accuracy, precision, recall, and F1-Score obtained:

Table 1. Performance Metrics Across 5-Fold Cross-Validation for the Decision Tree

K-n	Metrics			
	Accuracy	***Precision***	***Recall***	***F-Measure***
K-1	81%	81%	80%	81%
K-2	95%	94%	94%	94%
K-3	95%	96%	95%	95%
K-4	93%	94%	93%	93%
K-5	95%	95%	95%	95%
$\sum$ ***Avg***	92%	92%	91%	92%

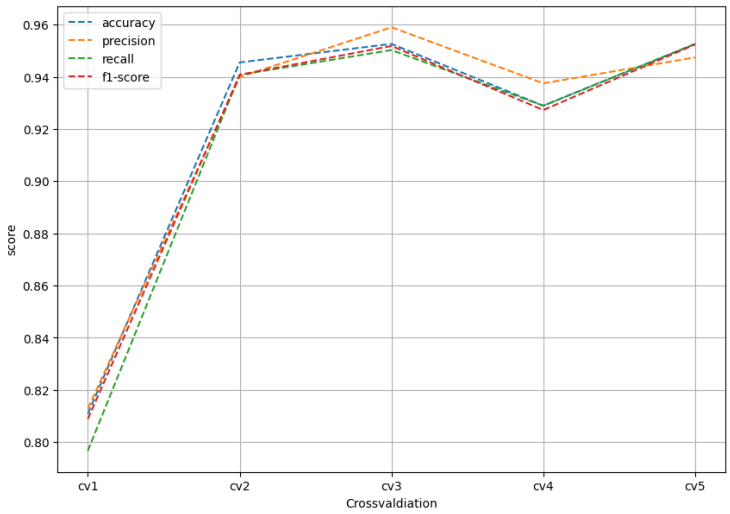


Figure 3. Visualisation Performance Metrics Across 5-Fold Cross-Validation for the Decision Tree.

Figure 3 accompanying visualization provides a clear comparative overview of the Decision Tree model's performance metrics across the 5-fold cross-validation process. The graph delineates the scores for accuracy, precision, recall, and F1-score for each fold, offering a graphical representation of the model's predictive prowess. The upward trend from the first to the third fold and the subsequent maintenance of high-performance levels highlight the model's robustness and reliability in estimating obesity levels, affirming the potential of machine learning applications in public health domains.

Discussion

The obtained results corroborate the hypothesis that lifestyle and physical condition factors are significant predictors of obesity levels. This aligns with previous research indicating the importance of eating habits, physical activity, and other lifestyle choices in determining obesity risk. The Decision Tree model's ability to discern intricate patterns within the data underscores the potential of machine learning techniques in public health research, offering a nuanced approach to understanding obesity determinants compared to traditional statistical methods.However, the variance in model performance across different folds warrants a discussion on the heterogeneity of data and its impact on predictive modelling. This variability underscores the importance of cross-validation in assessing a model's generalizability, a crucial step in ensuring the reliability of predictive models in healthcare settings.

One significant finding of this study is the model's heightened performance in specific folds, which may reflect the presence of more distinct or pronounced patterns in certain subsets of the data. This observation suggests that the effectiveness of interventions could be maximized by tailoring them to specific demographic or lifestyle profiles, a practical implication that could inform public health strategies aimed at combating obesity.

Nevertheless, the research is not without limitations. The reliance on self-reported data for lifestyle and physical condition factors introduces the possibility of bias and inaccuracies, potentially affecting the model's predictions.

Additionally, the study's focus on a Decision Tree model, while beneficial for interpretability, may overlook the potential of more complex models or ensemble methods that could offer improved predictive performance.

Future research should explore the integration of additional variables, such as genetic factors or socioeconomic status, to enhance the model's predictive accuracy. Moreover, comparing the performance of Decision Trees with other machine learning algorithms could yield insights into the most effective techniques for obesity level prediction. Lastly, longitudinal studies could provide a deeper understanding of the causality between lifestyle factors and obesity, offering a more dynamic perspective on this complex health issue.

4. Conclusion

In conclusion, the study's results from the application of a Decision Tree algorithm using 5-fold cross-validation provide insightful revelations into the predictive relationships between lifestyle factors and obesity levels. The analysis yielded high accuracy, precision, recall, and F1-scores, particularly in folds three and five, suggesting that the model is adept at capturing the underlying patterns within the data. The performance, as highlighted in the confusion matrix, indicates strong predictive capabilities, albeit with some variability that warrants further investigation. The research successfully addressed the posed questions, demonstrating that eating habits and physical activity are indeed significant predictors of obesity and that machine learning can effectively model these relationships.

The study makes a notable contribution to the intersection of public health and data science, illustrating how machine learning models like Decision Trees can be applied to complex health issues such as obesity. It opens the door for future research to explore other predictive models and incorporate additional variables that may offer deeper insights into obesity prediction. In practice, the findings advocate for the integration of predictive modeling into public health strategies to tailor interventions more precisely and effectively. Continued research in this domain is recommended, focusing on longitudinal data to establish causation and utilizing a broader array of machine learning techniques to enrich the predictive accuracy and reliability of health assessments.

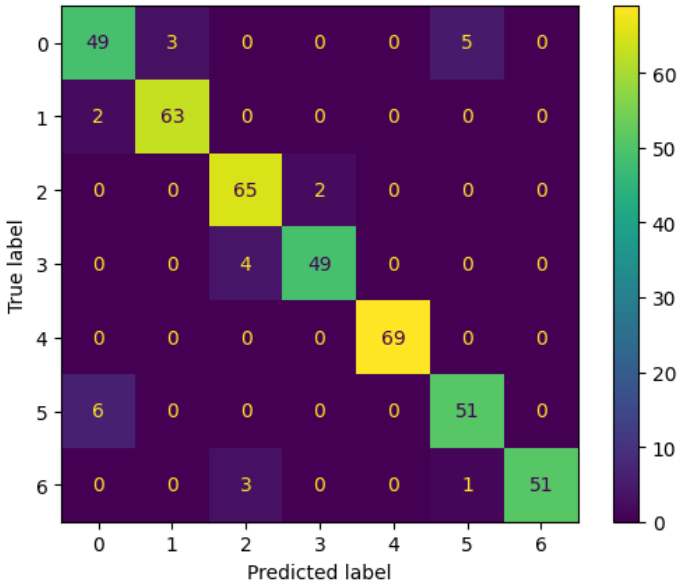


Figure 4. Confusion Matrix.

Figure 4 presented here illustrates the performance of the Decision Tree classifier across the different categories of obesity levels. Each cell of the matrix represents the number of observations from the actual classes (true label) and the predicted classes by the model (predicted label). This matrix is a powerful tool for understanding the model's classification accuracy for each category, showcasing where the model performs well and where it may confuse between classes. The diagonal elements represent the number of correct predictions, which are predominant in most classes, indicating a commendable level of accuracy in the model's predictions.

References:

[1] A. Fitria and H. Azis, "Analisis Kinerja Sistem Klasifikasi Skripsi menggunakan Metode Naïve Bayes Classifier," *Pros. Semin. Nas. Ilmu Komput. dan Teknol. Inf.*, vol. 3, no. 2, pp. 102–106, 2018.

[2] M. M. Baharuddin, T. Hasanuddin, and H. Azis, "Analisis Performa Metode K-Nearest Neighbor untuk Identifikasi Jenis Kaca," *Ilk. J. Ilm.*, vol. 11, no. 28, pp. 269–274, 2019, doi: 10.33096/ilkom.v11i3.489.269-274.

[3] 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, doi: 10.33096/ilkom.v12i2.507.81-86.

[4] 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, doi: 10.33633/tc.v19i3.3646.

[5] A. Nurul, Y. Salim, and H. Azis, "Analisis performa metode Gaussian Naïve Bayes untuk klasifikasi citra tulisan tangan karakter arab," *Indones. J. Data Sci.*, vol. 3, no. 3, pp. 115–121, 2022, doi: 10.56705/ijodas.v3i3.54.

[6] 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, doi: 10.56705/ijaimi.v1i2.98.

[7] N. Rismayanti, A. Naswin, U. Zaky, M. Zakariyah, and D. A. Purnamasari, "Evaluating Thresholding-Based Segmentation and Humoment Feature Extraction in Acute Lymphoblastic Leukemia Classification using Gaussian Naive Bayes," *Int. J. Artif. Intell. Med. Issues*, vol. 1, no. 2, 2023, doi: 10.56705/ijaimi.v1i2.99.

[8] A. Naswin and A. P. Wibowo, "Performance Analysis of the Decision Tree Classification Algorithm on the Pneumonia Dataset," *... Artif. Intell. Med. ...*, 2023, doi: 10.56705/ijaimi.v1i1.83.

[9] 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, doi: 10.56705/ijaimi.v1i1.87.

[10] A. Tuppad and S. D. Patil, "Data Pre-processing Issues in Medical Data Classification," *2023 Int. Conf. ...*, 2023, doi: 10.1109/NMITCON58196.2023.10275855.

[11] G. Ketepalli and P. Bulla, "Data Preparation and Pre-processing of Intrusion Detection Datasets using Machine Learning," *2023 Int. Conf. ...*, 2023, doi: 10.1109/ICICT57646.2023.10134025.

[12] J. Zhao, K. S. Chong, W. Shu, and ..., "A Data Pre-Processing Module for Improved-Accuracy Machine-Learning-based Micro-Single-Event-Latchup Detection," *2023 IEEE 9th Int. ...*, 2023, doi: 10.1109/SMC-IT56444.2023.00009.

[13] B. D. Finley, *Optimizing Data Pre-Processing Transformations with Reinforcement Learning*. search.proquest.com, 2022, doi: 10.3390/a17010037.

[14] N. Rezova, L. Kazakovtsev, G. Shkaberina, and ..., "Data Pre-Processing for Ecosystem Behavior Analysis," *2022 Int. ...*, 2022, doi: 10.1109/InfoTech55606.2022.9897105.

[15] P. S. Kumar, "Classification of skin cancer using convolutional neural network in comparison with decision tree classifier," *AIP Conf. Proc.*, vol. 2822, no. 1, 2023, doi: 10.1063/5.0173035.

[16] M. Bhattacharya, "Diabetes Prediction using Logistic Regression and Rule Extraction from Decision Tree and Random Forest Classifiers," *2023 4th Int. Conf. Emerg. Technol. INCET 2023*, 2023, doi: 10.1109/INCET57972.2023.10170270.

[17] T. R. Sahoo, "Decision tree classifier based on topological characteristics of subgraph for the mining of protein complexes from large scale PPI networks," *Comput. Biol. Chem.*, vol. 106, 2023, doi: 10.1016/j.compbiolchem.2023.107935.

[18] A. Anitha, "Disease prediction and knowledge extraction in banana crop cultivation using decision tree classifiers," *Int. J. Bus. Intell. Data Min.*, vol. 20, no. 1, pp. 107–120, 2022, doi: 10.1504/IJBIDM.2022.119957.

[19] J. A. D. de Jesus Ferreira, "Decision tree classifiers for unmanned aircraft configuration selection," *Aircr. Eng. Aerosp. Technol.*, vol. 93, no. 6, pp. 1122–1132, 2021, doi: 10.1108/AEAT-03-2021-0074.

[20] G. Sajiv, "Machine Learning based Analysis of Histopathological Images of Breast Cancer Classification using Decision Tree Classifier," *6th Int. Conf. I-SMAC (IoT Soc. Mobile, Anal. Cloud), I-SMAC 2022 - Proc.*, pp. 989–995, 2022, doi: 10.1109/I-SMAC55078.2022.9987276.

[21] H. Azis and S. R. Jabir, "Chemical Composition and Aroma Profiling: Decision Tree Modeling of Formalin Tofu," *J. Embed. Syst. Secur. ...*, 2023.

[22] M. Rafało, "Cross validation methods: Analysis based on diagnostics of thyroid cancer metastasis," *ICT Express*, vol. 8, no. 2, pp. 183–188, 2022, doi: 10.1016/j.icte.2021.05.001.

[23] K. M. Bain, "Cross-validation of three Advanced Clinical Solutions performance validity tests: Examining combinations of measures to maximize classification of invalid performance," *Appl. Neuropsychol.*, vol. 28, no. 1, pp. 24–34, 2021, doi: 10.1080/23279095.2019.1585352.

[24] M. Stusek, "Accuracy Assessment and Cross-Validation of LPWAN Propagation Models in Urban Scenarios," *IEEE Access*, vol. 8, pp. 154625–154636, 2020, doi: 10.1109/ACCESS.2020.3016042.

[25] O. Karal, "Performance comparison of different kernel functions in SVM for different k value in k-fold cross-validation," *Proc. - 2020 Innov. Intell. Syst. Appl. Conf. ASYU 2020*, 2020, doi: 10.1109/ASYU50717.2020.9259880.

[26] T. R. Mahesh, "AdaBoost Ensemble Methods Using K-Fold Cross Validation for Survivability with the Early Detection of Heart Disease," *Comput. Intell. Neurosci.*, vol. 2022, 2022, doi: 10.1155/2022/9005278.

[27] N. Rismayanti and A. P. Utami, "Improving Multi-Class Classification on 5-Celebrity-Faces Dataset using Ensemble Classification Methods," *Indones. J. Data ...*, 2023, doi: 10.56705/ijodas.v4i2.78.

[28] D. Ratnasari, "Comparison of Performance of Four Distance Metric Algorithms in K-Nearest Neighbor Method on Diabetes Patient Data," *Indones. J. Data Sci.*, 2023, doi: 10.56705/ijodas.v4i2.71.

[29] F. T. Admojo and S. R. Jabir, "Analisis performa metode Naïve Bayesh Classifier pada Electronic Nose dalam identifikasi formalin pada tahu," *Indones. J. Data ...*, 2023, doi: 10.56705/ijodas.v4i1.67.

[30] R. F. Syam, "Performance Comparison Analysis of Classifiers on Binary Classification Dataset," *Indones. J. Data Sci.*, 2023, doi: 10.56705/ijodas.v4i2.77.

[31] R. Setiawan, H. Zein, R. A. Azdy, and ..., "Rice Leaf Disease Classification with Machine Learning: An Approach Using Nu-SVM," *Indones. J. ...*, 2023, doi: 10.56705/ijodas.v4i3.114.

[32] 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, doi: 10.1109/IMCOM60618.2024.10418246.