INTENSIF, Vol.8 No.2 August 2024
ISSN: 2580-409X (Print) / 2549-6824 (Online)
DOI: https://doi.org/10.29407/intensif.v8i2.22280

Enhancing the Decision Tree Algorithm to Improve Performance Across Various Datasets

Received:
21 May 2024
Accepted:
1 June 2024
Published:
1 August 2024

[1]Pandu Pratama Putra, [2*]M. Khairul Anam, [3]Sarjon Defit, [4]Arda Yunianta
[1]Informatics Engineering, Universitas Lancang Kuning
[2]Informatics, Universitas Samudra
[3]Information Technology, Universitas Putra Indonesia YPTK Padang
[4]Faculty of Computing and Information Technology in Rabigh, King Abdulaziz University
E-mail: [1]pandupratamaputra@unilak.ac.id, [2]khairula210@gmail.com, [3]sarjon_defit@upiyptk.ac.id, [3]ayunianta@kau.edu.sa
*Corresponding Author

Abstract— Background: The Village Fund is an initiative by the central government to promote equitable regional development. However, it has also led to corruption. Many Indonesians share their opinions on the Village Fund on social media platforms like X, and news coverage is extensive on portals like detik.com. **Objective**: This study aims to classify data from social media and news coverage to enhance understanding. **Methods**: The research improves the decision tree algorithm by integrating other algorithms and techniques such as XGBoost and SMOTE. Ensuring high accuracy is vital for the credibility of machine learning classifications among the public. The study uses two different datasets, necessitating varied testing approaches. For the news portal dataset, a single test with seven labels is conducted, followed by enhancement with XGBoost. The X dataset undergoes two tests with datasets of 1200 and 3078 entries, using three labels. **Conclusion**: The evaluation results indicate that the highest accuracy achieved with the news portal data was 82%, thanks to a combination of decision tree algorithms with various parameters and the balancing effect of SMOTE. For the Twitter dataset with 3078 entries, the highest accuracy reached 95%, attributed to the application of ensemble techniques, particularly boosting.
Keywords— Village Funds, News Portal, Tweet, Ensemble

Corresponding Author:
M. Khairul Anam,
Department Informatics,
Universitas Samudra,
Email: khairula210@gmail.com,
Orchid ID: https://orcid.org/0000-0003-4295-450X

ISSN: 2580-409X (Print) / 2549-6824 (Online)
DOI: https://doi.org/10.29407/intensif.v8i12.22280

I. INTRODUCTION

Village funds provide an opportunity for rural development and the economic empowerment of communities. This involves training and marketing local crafts, developing livestock and fisheries businesses, and promoting tourism through the Village-Owned Enterprises (BUMDES) [1]. In the planning stage, the utilization of village funds tends to align with programs outlined in the village head's plan [2]. Financial management in the village encompasses planning, execution, record-keeping, reporting, and accountability [3]. Based on research findings, direct community oversight of Village Fund Allocation (ADD) management is limited. This is due to a lack of understanding among the community members regarding the ADD program, necessitating public awareness campaigns and transparent fund utilization [4].

However, challenges arise in preparing village financial reports, mainly stemming from a lack of human resources capability, absence of guidelines, and insufficient training. The low accountability in ADD financial administration has become a corruption target by village heads and officials [5]. Previous research also discusses the management and supervision of village funds, including studies on the role of the Village Consultative Body (BPD) [6]. community empowerment through Village-Owned Enterprises (BUMDES) [7]. formulation and approval of village regulations, and performance monitoring of village heads [8].

News related to Village Funds is often highlighted on news portals. In this study, the news portal used for data collection is detik.com. According to similarweb.com, detik.com is ranked second in terms of the most visited websites, with 145 million visits. Data for this research is gathered using web scraping, an automated method for extracting information from websites [9], [10]. The use of web scraping techniques was also carried out by previous researchers, including applying web scraping to search media and automatically saving scientific articles based on keywords [11]. and web scraping with sentiment analysis on news [12]. Implementation of web scraping and sentiment analysis on news by [13].

The collected data, labelled manually, will be processed using the Decision Tree algorithm, which is one of the classification algorithms using the tree structure representation [14], [15] to determine the status of village fund allocations. Classification using decision trees involves constructing a decision tree that tests attribute decision nodes and produces branches directed to other nodes until a decision is reached [16].

The use of the decision tree algorithm has also been done by previous researchers in classification, including a study by [17] classifying tweets about COVID-19, using the DECISION TREE, obtaining an accuracy of 70.13%, then improving with ID3 to achieve 88.82% accuracy. Then [18] improved the DECISION TREE using bagging and achieved an accuracy of

ISSN: 2580-409X (Print) / 2549-6824 (Online)
DOI: https://doi.org/10.29407/intensif.v8i2.22280

97% on the breast test set. Decision tree [19] was also used to classify the Heart dataset and achieved an accuracy of 90.8%. Based on these studies, this research uses the decision tree to classify village funds in news portals and uses Twitter datasets with two different dataset sizes.

In addition to using data from news portals, this research also incorporates data from Twitter. This data is subjected to preprocessing and word weighting using TF-IDF. After weighting the data, the next step involves modeling with the decision tree algorithm. The initial data, which was preprocessed earlier, has not undergone data balancing. The need for data balancing is crucial to ensure that the generated model does not misclassify and produces a high level of accuracy. Furthermore, this study employs the fusion method, namely ensemble. The ensemble algorithm utilized is the XGBoost algorithm. XGBoost is an enhancement of gradient tree boosting based on ensemble algorithms, effectively addressing large-scale machine learning cases [20]. The XGBoost method is chosen for its additional features that expedite computation and prevent overfitting [21]. XGBoost is capable of handling various classification, regression, and ranking scenarios. XGBoost is used on Twitter data, and the same method is also applied in this research to improve accuracy on news portal datasets.

II. RESEARCH METHOD

This study focuses on enhancing the decision tree algorithm while considering previous research as a reference for improving accuracy. Table 1 outlines the methods employed to enhance machine learning algorithm accuracy.

Table 1. Method For Improving Machine Learning Algorithm Accuracy

No	Research	Methods for Accuracy Improvement
1	[22]	Feature Selection: Chi-square, Univariate, and Information Gain
2	[23], [24]	ensemble method: Voting, Boosting, Stacking, And Bagging
3	[25], [26]	hyperparameters: Grid Search, Random Search, Bayesian Optimization, GA, PSO
4	[27]	Minimizing Overfitting: K fold Cross Validation
5	[28]	SMOTE
6	[29]	Feature Extraction: Bag of Word (BoW), TF-IDF

In Table 1, various methods for enhancing accuracy in machine learning are highlighted. Table 2 provides details on methods previously utilized to improve the decision tree algorithm.

Table 2. Increasing Accuracy in the Decision Tree Algorithm

No	Research	Methods for Decision Tree Accuracy Improvement
1	[30]	Boosting, K Fold Cross Validation, dan CART model
2	[31]	Improving algorithm based on neural network decision tree (ADT)
3	[32]	Adaboost
4	[33]	T3C
5	[34]	MSD-Splitting

ISSN: 2580-409X (Print) / 2549-6824 (Online)
DOI: https://doi.org/10.29407/intensif.v8i12.22280

In Table 2, various algorithms and methods are presented for improving the accuracy of the decision tree algorithm. This study employs boosting to enhance accuracy in the news portal dataset with 7 labels. Additionally, for the Twitter dataset, the SMOTE method is utilized to balance the data. To further improve accuracy, ensemble techniques are applied by combining decision tree with XGBoost on data with 3 labels.

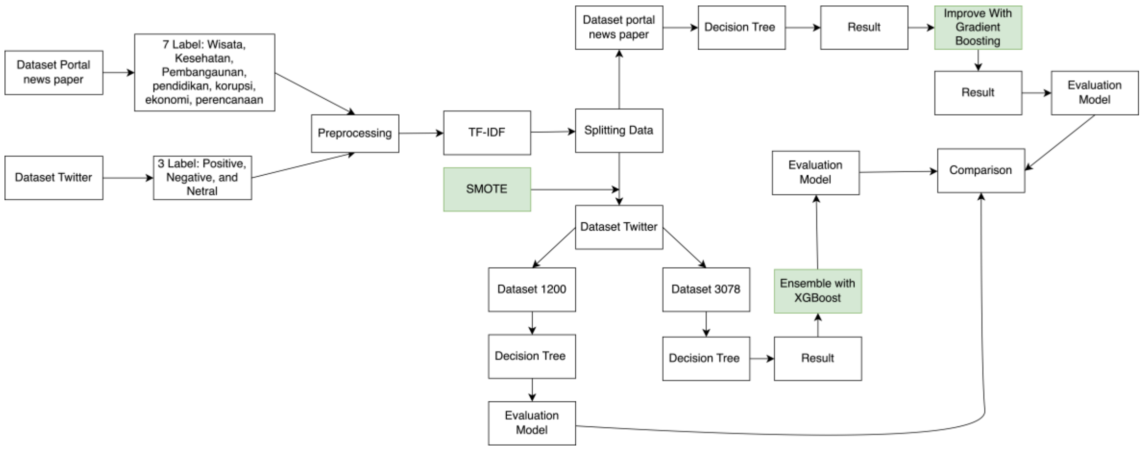

Fig 1. Research Flow for Improving the Decision Tree Algorithm

Figure 1 illustrates the workflow of the research conducted to enhance accuracy in the decision tree algorithm. The dataset for this study is sourced from two distinct platforms: news portals and Twitter, denoted as X. The initial stage of this research involves data collection using web scraping techniques, which extract information from predefined websites. Through web scraping, data such as titles, summaries, dates, and links can be obtained from the web. The selected website for data collection is detik.com, focusing on news related to Village Funds. The data extraction process involves all pages of news related to Village Funds. Once the process on the first page is completed, it continues to subsequent pages until all relevant articles or 'articles' are exhausted. The web scraping process has been executed, resulting in a dataset of 3,500 entries.

Data collection from social media for this study is performed through the drone emprit academy portal. From this portal, 1,200 data points were collected in January 2023, followed by an additional 3,078 tweet data points in March 2023. The chosen topic from the drone emprit academy portal is related to Village Funds. After obtaining the data, the next step is to label the news portal dataset manually by reading article titles. The labeling phase is a crucial part of continuing the research process. The labeling process uses 3,500 data samples with 7 labels: Planning, Economic, Corruption, Education, Development, Health, and Tour. Expert verification, utilizing language and economic experts, is employed for label determination.

Subsequently, labeling for the social media X dataset is done using a lexicon-based approach. Lexicon-based labeling relies on a predefined dictionary or list of words. This research has its

ISSN: 2580-409X (Print) / 2549-6824 (Online)
DOI: https://doi.org/10.29407/intensif.v8i2.22280

lexicon of words labeled as positive, negative, and neutral. Lexicon-based labeling is chosen for its ease of implementation and interpretation, requiring no machine learning. After the labelling phase, the data enters the text preprocessing stage, preparing the data by cleaning it before further analysis. Text preprocessing involves data cleaning, case folding, tokenizing, filtering, streaming, and data transformation (label encoder).

The next step involves word weighting using TF-IDF (Term Frequency-Inverse Document Frequency), a calculation of word weights after the preprocessing stage. TF-IDF measures how important a word (term) is in a document and corpus. The TF-IDF method combines two concepts: the frequency of a word's occurrence in a document and the inverse frequency of documents containing that word. There are formulas used to calculate TF-IDF weights. Term Frequency (TF) is calculated first, with a word weight of 1. The IDF value is then formulated in the equation.

III. RESULT AND DISCUSSION

The first step involves conducting experiments on the dataset with 7 labels. Figure 2 provides a visualization of the dataset after manual labeling.

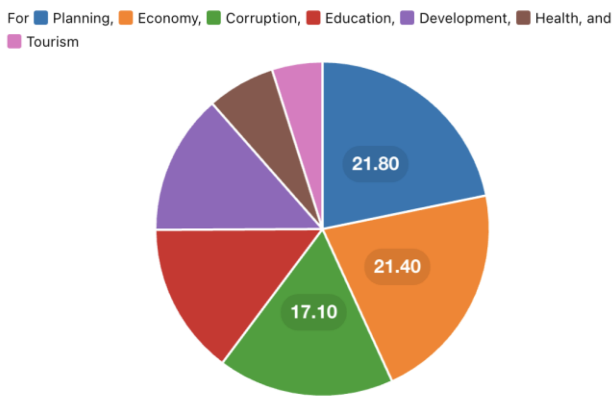

Fig 2. Visualizes Bar Chart

Figure 2 displays the distribution of data across each label. Planning emerges as the most prevalent label, comprising 746 instances, followed by the economic category with 749 instances, corruption with 598 instances, education with 514 instances, development with 475 instances, health with 230 instances, and tourism as the least represented label, amounting to 130 instances. The total data count across all labels aligns with the sample size of 3500 data points. Upon the completion of this labeling phase, the next step involves data preprocessing.

The classification results for the 7 labels using the news portal dataset are obtained through the decision tree classification method. Decision tree is a classification method that makes predictions based on a set of rules represented as a tree structure. In the initial stage, module

ISSN: 2580-409X (Print) / 2549-6824 (Online)
DOI: https://doi.org/10.29407/intensif.v8i12.22280

importation is carried out. This decision tree model utilizes two parameters: criterion='gini' and criterion='entropy'. It employs max_depth=40, determining the maximum depth of decision tree nodes, and random_state=32, ensuring result consistency to prevent variations upon rerunning the process.

After determining the parameters, the decision tree model is trained using the 'fit' method, where 'X_train' represents the training data, and 'y_train' is the subset of labels corresponding to the training subset. By invoking fit (X_train, y_train), the training subset is used to train the decision tree model, allowing it to learn decision rules from the features and labels in the training data. Once the model has acquired knowledge from the data, it proceeds to predict on the test subset 'X_test'. Upon obtaining accuracy results, a tree is constructed as part of the data processing. Figures 3 and 4 present the resulting trees based on the specified parameters.

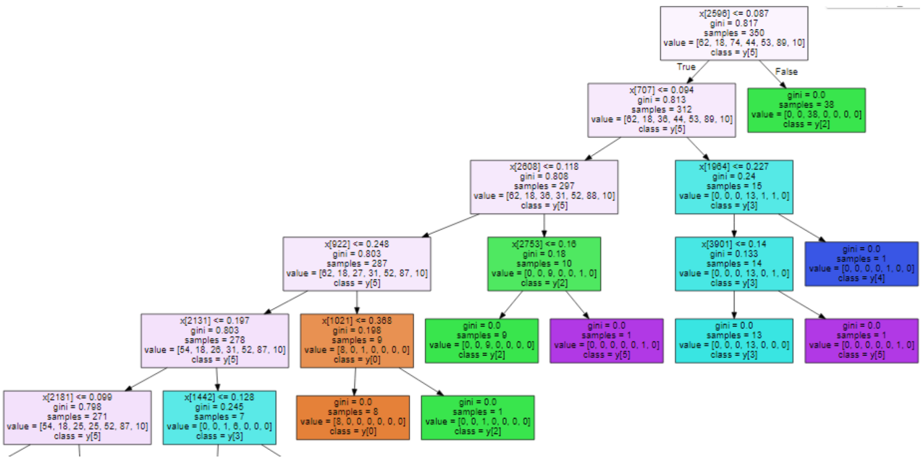

Fig 3. Tree Parameters Gini Index

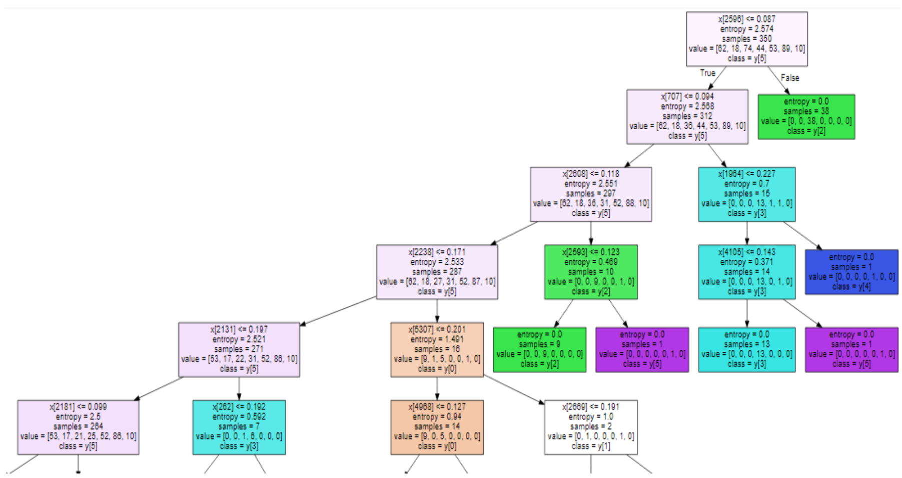

Fig 4. Tree Parameters Entropy

ISSN: 2580-409X (Print) / 2549-6824 (Online)
DOI: https://doi.org/10.29407/intensif.v8i2.22280

Subsequently, the Gradient Boosting method in machine learning is employed to construct a robust predictive model by combining simple predictive models using decision trees to rectify errors from previous predictions. In the Gradient Boosting model, parameters such as 'max_depth=7' function to control the maximum depth of decision tree nodes, and 'random_state=32' ensures the consistency of the generated model when the code is rerun.

The next step involves analyzing the 'gbc_clf' object using the appropriate parameters and training the model on the training data using the '.fit()' method. 'X_train' represents the matrix with training data examples, while 'y_train' is the target vector with corresponding labels in the training examples. The Gradient Boosting model sequentially builds trees and corrects prediction errors by learning from 'X_train' and target 'y_train' to make predictions on the data.

The final stage in this research process is model evaluation with the aim of ensuring that the constructed model aligns with the predetermined objectives. Model evaluation is conducted to determine the performance of the model using the previously established gini index and entropy parameters. The study conducts four comparative experiments with 90% training data and 10% test data, 80% training data and 20% test data, 70% training data and 30% test data, 60% training data and 40% test data. Figures 4.24 and 4.25 depict the model evaluation comparison with 90% training data and 10% test data using the gini index and entropy parameters.

With the accuracy values derived from all model comparisons using the gini index and entropy parameters, the model with a 90% training data and 10% test data comparison using the gini index parameter has the highest accuracy, reaching 53% or 0.5343. Table 3 presents the results of each comparison data.

Table 3. Result Comparison

Model/ Parameters	Splitting Data	Accuracy	Precision	Recall	F-1 Score
Gini Index	60:40	50	61	45	48
	70:30	52	65	45	48
	80:20	53	66	46	48
	90:10	53	55	41	43
Entropy	60:40	49	61	45	48
	70:30	49	57	44	47
	80:20	50	57	44	47
	90:10	52	52	43	44

With the accuracy results obtained from the four model comparisons, additional testing will be conducted using the Gradient Boosting algorithm to further enhance accuracy. Table 4 presents the testing results using the Gradient Boosting algorithm.

ISSN: 2580-409X (Print) / 2549-6824 (Online)
DOI: https://doi.org/10.29407/intensif.v8i12.22280

Table 4. Comparison of Data Splitting for Decision Tree Using Gradient Boosting

Splitting Data	Result
90:10	58.28%
80:20	53.85%
70:30	55.42%
60:40	54.07%

The best result obtained in testing using Gradient Boosting was for the 90% training data and 10% test data split, achieving an accuracy of 58.28%. The accuracy obtained using the Gini index and entropy parameters has not been able to address the issues encountered, even with the use of boosting techniques. Data imbalance might be the problem in this experiment. Therefore, this research attempts to add another parameter, namely presort, and then applies another ensemble technique, which is the voting technique. The type of voting used is soft voting. Soft voting is an ensemble learning method used in machine learning to improve prediction performance by combining the prediction results of several models [35]. In soft voting, the final prediction is based on the prediction probabilities generated by each model, rather than on the majority result [36]. The following is a combination of tree algorithms with several parameters.

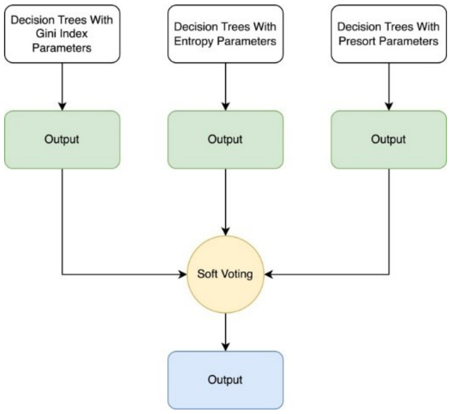

Fig 5. Combining Decision Trees with Different Parameters Using Soft Voting Techniques

From the model in Figure 5, SMOTE is also added to address the issue of class imbalance. Table 5 presents the performance results of combining decision trees with various parameters using the soft voting technique.

Table 5. Hasil Soft Voting+SMOTE

Model	Splitting Data	Accuracy	Precision	Recall	F-1 Score
Soft Voting	60:40	80%	81%	80%	80%
	70:30	82%	81%	82%	81%
	80:20	81%	81%	81%	81%
	90:10	80%	80%	81%	80%

The results in Table 5 show that the use of soft voting and SMOTE significantly improves the performance of this soft voting model. Subsequent testing involves data from social media X with a total of 1200 data. The testing results yield good accuracy without the need for data balancing.

ISSN: 2580-409X (Print) / 2549-6824 (Online)
DOI: https://doi.org/10.29407/intensif.v8i2.22280

However, after implementing the data balancing method, namely SMOTE, the accuracy decreases. Table 6 presents the accuracy results obtained using a dataset with 1200 entries.

Table 6. Comparison Result on the Twitter Dataset with a Total of 1200 Records

NO	Model	Amount of Data	Accuracy
1	DECISION TREE	1200	75%
2	DECISION TREE+SMOTE	1200	54%

Table 6 provides information regarding the performance of two experiments conducted, namely with Decision Tree and Decision Tree with SMOTE. The Decision Tree achieves an accuracy of 75%, while the Decision Tree with SMOTE experiences a decrease in accuracy, reaching 54%. This occurs because SMOTE creates new samples by combining data from existing minority classes. If the creation of synthetic samples is inaccurate or does not represent the true distribution of the minority class, it can lead to poor generalization of the model to actual data. The experiment was conducted with 1200 data using a 70:30 data split.

The following are the classification results using four experiments: with the decision tree algorithm alone, then decision tree using smote for data balancing. Additionally, this research employs ensemble techniques by adding another algorithm, namely XGBoost, applied to decision tree and decision tree + Smote.

Table 7. Comparison Result on the Twitter Dataset with a Total of 3078 Records

NO	Model	Amount of Data	Accuracy
1	DECISION TREE	3078	79%
2	DECISION TREE+ Smote	3078	45%
3	DECISION TREE+SMOTE+XGBoost (ensemble)	3078	95%
4	DECISION TREE+XGBoost (ensemble)	3078	84%

Table 7 reveals that the SMOTE method has not yet improved accuracy; instead, it has decreased accuracy compared to the basic algorithm. The Decision Tree achieves an accuracy of 79%, while the Decision Tree + Smote experiences a decrease in accuracy, reaching 45%. Unexpectedly, the use of Decision Tree + Smote + XGBoost significantly increases accuracy to 95%. Without using Smote, this accuracy decreases to 84%. The accuracy comparison with previous research is presented in Table 8 below.

Table 8. Comparison With Previous Research

NO	Researcher	Accuracy Improvement	Accuracy
1	[37]	-	86%
2	[38]	Testing with different data amounts	92%
3	[39]	SMOTE	88%
4	[40]	Ensemble with voting technique (Decision Tree, Naïve Bayes, and Random Forest)	93%
5	[41]	Testing with different data splits	77%
6	**This Study**	**SMOTE and boosting**	**95%**

ISSN: 2580-409X (Print) / 2549-6824 (Online)
DOI: https://doi.org/10.29407/intensif.v8i12.22280

From several studies in Table 8, it is evident that this study excels in accuracy, reaching 95%. Meanwhile, previous research in Table 8 shows only 93%, indicating a 2% increase.

IV. CONCLUSION

Based on the discussions conducted, the use of a dataset with 7 labels and the Decision Tree algorithm resulted in a suboptimal accuracy of 53%. Although there was an improvement when employing Gradient Boosting, the increase was only 5%, reaching 58%. After conducting the testing process using SMOTE and soft voting, the result increased to 82%.

Furthermore, when using 3 labels (positive, negative, and neutral), the research achieved relatively good accuracy. However, the use of SMOTE with the Decision Tree algorithm alone experienced a significant decrease. For the dataset with 1200 entries, there was a reduction of up to 21%, while for the dataset with 3078 entries, a substantial decrease of 34% was observed. Nevertheless, when XGBoost was added to the Decision Tree, there was an increase of 5% from the base algorithm. The performance of the Decision Tree algorithm also improved when SMOTE was added, achieving a 16% increase. From the various experiments conducted, it is evident that the fusion method used in the research can enhance accuracy, whether with 7 labels or 3 labels.

Future research should focus on exploring other ensemble techniques such as voting, stacking, and bagging. Additionally, employing hyperparameter tuning is necessary to automate the search for the best parameters and enhance accuracy. Furthermore, testing with new data using a graphical user interface (GUI) is crucial to streamline the classification process.

Author Contributions: *Pandu Pratama Putra*: Software, Investigation, Data Curation, Writing - Original Draft. *M. Khairul Anam*: Conceptualization, Methodology, Writing - Original Draft, Writing - Review & Editing, Supervision. *Sarjon Defit*: Investigation, Data Curation. *Arda Yuniatna*: Investigation, Data Curation.

All authors have read and agreed to the published version of the manuscript.

Funding: This research was sponsored by Lancang Kuning University.

Conflicts of Interest: The authors declare no conflict of interest.

Data Availability: The data for this research was collected from the news portal detik.com, focusing on topics related to village funds. Additionally, the data includes Tweets from Twitter or X, collected through the Drone Emprit Academy.

Informed Consent: There were no human subjects.

Animal Subjects: There were no animal subjects.

ISSN: 2580-409X (Print) / 2549-6824 (Online)
DOI: https://doi.org/10.29407/intensif.v8i2.22280

ORCID:

Pandu Pratama Putra: http://orcid.org/0000-0003-0639-2810

M Khairul Anam: http://orcid.org/0000-0003-4295-450X

Sarjon Defit: http://orcid.org/0000-0001-7538-9274

Arda Yuniatna: http://orcid.org/0000-0002-4732-1026

REFERENCES

[1] A. Sofianto and T. Risandewi, "Mapping of Potential Village-Owned Enterprises (BUMDes) for Rural Economic Recovery during the COVID-19 Pandemic in Central Java, Indonesia," in *IOP Conference Series: Earth and Environmental Science*, IOP Publishing Ltd, Nov. 2021, pp. 1–17. **doi:** 10.1088/1755-1315/887/1/012022.

[2] Haeranah, "Village Funds Management in Improving the Development Leppangeng Village, Ajangale District, Bone Regency," *Jurnal Ilmu Pemerintahan Suara Khatulistiwa*, vol. 5, no. 1, pp. 81–91, 2020, **doi:** 10.33701/jipsk.v5i1.1126.

[3] M. Rahmadanti, G. Gamaputra, D. A. U. Yuni Lestari, and P. Pinata, "Village Financial System Management in Kebumen Regency," *KnE Social Sciences*, May 2022, **doi:** 10.18502/kss.v7i9.10992.

[4] E. Hermawan, "Community Empowerment through Management of Village Funds Allocation in Indonesia," *International Journal of Science and Society*, vol. 1, no. 3, pp. 67–79, 2019, **doi:** 10.54783/ijsoc.v1i3.30.

[5] S. Wahyudi, T. Achmad, and I. D. Pamungkas, "Prevention Village Fund Fraud in Indonesia: Moral Sensitivity as a Moderating Variable," *Economies*, vol. 10, no. 1, pp. 1–16, 2022, **doi:** 10.3390/economies10010026.

[6] B. Santoso and A. Awangga, "Village Government Implementation Based on Law Number 6 of 2014," *Hermeneutika*, vol. 7, no. 1, pp. 155–163, 2023, **doi:** 10.33603/hermeneutika.v6i3.8326.

[7] A. A. I. N. Marhaeni *et al.*, "Empowerment Of Village Owned Enterprises (BUMDes) In The Context Of Optimizing The Assets Of Nyuhtebel Village, Manggis District, Karangasem Regency," *International Journal Of Community Service*, vol. 2, no. 4, pp. 447–453, 2022, **doi:** 10.51601/ijcs.v2i4.151.

[8] M. A. Ladiku, F. U. Puluhulawa, and N. M. Nggilu, "Measuring The Evaluation And Clarification of The Implementation of The Forming of Village Regulations In The New Normal Time," *Estudiente Law Journal*, vol. 3, no. 1, pp. 56–69, 2021, **doi:** 10.33756/eslaj.v0i0.14942.

[9] J. Boegershausen, H. Datta, A. Borah, and A. T. Stephen, "Fields of Gold: Scraping Web Data for Marketing Insights," *J Mark*, vol. 86, no. 5, pp. 1–20, Sep. 2022, **doi:** 10.1177/00222429221100750.

[10] H. Hairani and T. Widiyaningtyas, "Augmented Rice Plant Disease Detection with Convolutional Neural Networks," *INTENSIF: Jurnal Ilmiah Penelitian dan Penerapan Teknologi Sistem Informasi*, vol. 8, no. 1, pp. 27–39, Feb. 2024, **doi:** 10.29407/INTENSIF.V8I1.21168.

[11] V. A. Flores, P. A. Permatasari, and L. Jasa, "Penerapan Web Scraping Sebagai Media Pencarian dan Menyimpan Artikel Ilmiah Secara Otomatis Berdasarkan Keyword," *Majalah Ilmiah Teknologi Elektro*, vol. 19, no. 2, p. 157, 2020, **doi:** 10.24843/mite.2020.v19i02.p06.

[12] S. Satriajati, S. Bagus Panuntun, and S. Pramana, "Implementasi Web Scraping Dalam Pengumpulan Berita Kriminal Pada Masa Pandemi COVID-19 (Studi Kasus: Situs Berita

ISSN: 2580-409X (Print) / 2549-6824 (Online)
DOI: https://doi.org/10.29407/intensif.v8i12.22280

detik.com)," in *Seminar Nasional Official Statistics*, 2020, pp. 300–308. **doi:** 10.34123/semnasoffstat.v2020i1.578.

[13] A. Suryadi, W. A. Syb'an, N. Alfa'inna, E. H. Hermaliani, and U. N. Mandiri, "Implementasi Web Scraping dan Sentiment Analysis Terhadap Berita Menggunakan Machine Learning," *JURNAL SWABUMI*, vol. 11, no. 1, p. 2023, 2023, **doi:** 10.31294/swabumi.v11i1.15145.

[14] M. Yusa, E. Utami, and E. T. Luthfi, "Evaluasi Performa Algoritma Klasifikasi Decision Tree ID3, C4.5, dan CART Pada Dataset Readmisi Pasien Diabetes," *InfoSys Journal*, vol. 4, no. 1, pp. 23–34, 2016, **doi:** 10.22303/infosys.4.1.2016.23-34.

[15] S. Sucipto, D. D. Prasetya, and T. Widiyaningtyas, "Educational Data Mining: Multiple Choice Question Classification in Vocational School," *Matrik: Jurnal Manajemen, Teknik Informatika, dan Rekayasa Komputer*, vol. 23, no. 2, pp. 367–376, 2024, **doi:** 10.30812/matrik.v23i2.3499.

[16] G. Katz, A. Shabtai, L. Rokach, and N. Ofek, "Confdtree: A statistical method for improving decision trees," *J Comput Sci Technol*, vol. 29, no. 3, pp. 392–407, 2014, **doi:** 10.1007/s11390-014-1438-5.

[17] F. Es-Sabery *et al.*, "A MapReduce Opinion Mining for COVID-19-Related Tweets Classification Using Enhanced ID3 Decision Tree Classifier," *IEEE Access*, vol. 9, pp. 58706–58739, 2021, **doi:** 10.1109/ACCESS.2021.3073215.

[18] Y. Q. Song, X. Yao, Z. Liu, X. Shen, and J. Mao, "An Improved C4.5 Algorthm in Bagging Integration Model," *IEEE Access*, vol. 8, pp. 206866–206875, 2020, **doi:** 10.1109/ACCESS.2020.3032291.

[19] X. Luo, X. Wen, M. C. Zhou, A. Abusorrah, and L. Huang, "Decision-Tree-Initialized Dendritic Neuron Model for Fast and Accurate Data Classification," *IEEE Trans Neural Netw Learn Syst*, vol. 33, no. 9, pp. 4173–4183, Sep. 2022, **doi:** 10.1109/TNNLS.2021.3055991.

[20] J. M. Ahn, J. Kim, and K. Kim, "Ensemble Machine Learning of Gradient Boosting (XGBoost, LightGBM, CatBoost) and Attention-Based CNN-LSTM for Harmful Algal Blooms Forecasting," *Toxins (Basel)*, vol. 15, no. 10, Oct. 2023, **doi:** 10.3390/toxins15100608.

[21] S. S. Dhaliwal, A. Al Nahid, and R. Abbas, "Effective intrusion detection system using XGBoost," *Information (Switzerland)*, vol. 9, no. 7, Jun. 2018, **doi:** 10.3390/info9070149.

[22] M. Fayaz, A. Khan, J. U. Rahman, A. Alharbi, M. I. Uddin, and B. Alouffi, "Ensemble machine learning model for classification of spam product reviews," *Complexity*, vol. 2020, pp. 1–10, 2020, **doi:** 10.1155/2020/8857570.

[23] A. Mohammed and R. Kora, "A comprehensive review on ensemble deep learning: Opportunities and challenges," *Journal of King Saud University - Computer and Information Sciences*, vol. 35, no. 2. King Saud bin Abdulaziz University, pp. 757–774, Feb. 01, 2023. **doi:** 10.1016/j.jksuci.2023.01.014.

[24] I. D. Mienye and Y. Sun, "A Survey of Ensemble Learning: Concepts, Algorithms, Applications, and Prospects," *IEEE Access*, vol. 10, pp. 99129–99149, 2022, **doi:** 10.1109/ACCESS.2022.3207287.

[25] E. Elgeldawi, A. Sayed, A. R. Galal, and A. M. Zaki, "Hyperparameter tuning for machine learning algorithms used for arabic sentiment analysis," *Informatics*, vol. 8, no. 4, pp. 1–21, Dec. 2021, **doi:** 10.3390/informatics8040079.

[26] M. K. Anam, M. I. Mahendra, W. Agustin, Rahmaddeni, and Nurjayadi, "Framework for Analyzing Netizen Opinions on BPJS Using Sentiment Analysis and Social Network Analysis (SNA)," *Intensif*, vol. 6, no. 1, pp. 2549–6824, 2022, **doi:** 10.29407/intensif.v6i1.15870.

[27] Y. Jung, "Multiple predicting K-fold cross-validation for model selection," *J Nonparametr Stat*, vol. 30, no. 1, pp. 197–215, Jan. 2018, **doi:** 10.1080/10485252.2017.1404598.

ISSN: 2580-409X (Print) / 2549-6824 (Online)
DOI: https://doi.org/10.29407/intensif.v8i2.22280

[28] M. K. Anam *et al.*, "Sentiment Analysis for Online Learning using The Lexicon-Based Method and The Support Vector Machine Algorithm," *ILKOM Jurnal Ilmiah*, vol. 15, no. 2, pp. 290–302, 2023, **doi:** 10.33096/ilkom.v15i2.1590.290-302.

[29] R. Haque, N. Islam, M. Tasneem, and A. K. Das, "Multi-class sentiment classification on Bengali social media comments using machine learning," *International Journal of Cognitive Computing in Engineering*, vol. 4, pp. 21–35, Jun. 2023, **doi:** 10.1016/j.ijcce.2023.01.001.

[30] L. Zhao, S. Lee, and S. P. Jeong, "Decision tree application to classification problems with boosting algorithm," *Electronics (Switzerland)*, vol. 10, no. 16, Aug. 2021, **doi:** 10.3390/electronics10161903.

[31] M. Zhang, H. Peng, and X. Yan, "Improved algorithm of decision tree based on neural network," in *Journal of Physics: Conference Series*, IOP Publishing Ltd, Dec. 2020, pp. 1–8. **doi:** 10.1088/1742-6596/1693/1/012081.

[32] M. Riansyah, S. Suwilo, and M. Zarlis, "Improved Accuracy In Data Mining Decision Tree Classification Using Adaptive Boosting (Adaboost)," *SinkrOn*, vol. 8, no. 2, pp. 617–622, Apr. 2023, **doi:** 10.33395/sinkron.v8i2.12055.

[33] P. Tzirakis and C. Tjortjis, "T3C: improving a decision tree classification algorithm's interval splits on continuous attributes," *Adv Data Anal Classif*, vol. 11, no. 2, pp. 353–370, Jun. 2017, **doi:** 10.1007/s11634-016-0246-x.

[34] P. Rim and E. Liu, "Optimizing the C4.5 Decision Tree Algorithm using MSD-Splitting," *IJACSA) International Journal of Advanced Computer Science and Applications*, vol. 11, no. 10, pp. 41–47, 2020, **doi:** 10.14569/IJACSA.2020.0111006.

[35] A. R. Manga', A. N. Handayani, H. W. Herwanto, R. A. Asmara, Y. I. Sulistya, and Kasmira, "Analysis of the Ensemble Method Classifier's Performance on Handwritten Arabic Characters Dataset," *ILKOM Jurnal Ilmiah*, vol. 15, no. 1, pp. 186–192, Apr. 2023, **doi:** 10.33096/ilkom.v15i1.1357.186-192.

[36] F. Leon, S.-A. Floria, and C. Bădică, "Evaluating the Effect of Voting Methods on Ensemble-Based Classification," in *International Conference on INnovations in Intelligent SysTems and Applications (INISTA)*, 2017, pp. 1–6. **doi:** 10.1109/INISTA.2017.8001122.

[37] A. Pohon *et al.*, "The Decision Tree Algorithm on Sentiment Analysis: Russia and Ukraine War," vol. 13, no. 2, 2023, **doi:** 10.30700/jst.v13i2.1397.

[38] A. Y. Ikhsanti, Y. Fauziah, and R. I. Perwira, "Implementation of the c4.5 decision tree learning algorithm for sentiment analysis in e-commerce application reviews on google play store," *Computing and Information Processing Letters* , vol. 1, no. 1, pp. 25–30, 2021, **doi:** 10.31315/cip.v1i1.6128.

[39] F. Fersellia, E. Utami, and A. Yaqin, "Sentiment Analysis of Shopee Food Application User Satisfaction Using the C4.5 Decision Tree Method," *Sinkron*, vol. 8, no. 3, pp. 1554–1563, Jul. 2023, **doi:** 10.33395/sinkron.v8i3.12531.

[40] Y. Rianto and A. Y. Kuntoro, "Prediction Using Random Forest, Decision Tree, Naïve Bayes, And Ensemble Algorithm," *SinkrOn*, vol. 5, no. 1, pp. 9–20, Sep. 2020, **doi:** 10.33395/sinkron.v5i1.10565.

[41] I. Sabilirrasyad, Z. Hasan, and mas'ud Hermansyah, "Sentiment Analysis of Twitter Discussions on Rafael Alun: Multinomial Naïve Bayes and Decision Tree Approach," in *International Conference On Economics ,Businessand Information Technology*, 2023, pp. 803–809. **doi:** 10.31967/prmandala.v4i0.827.