Implementation of Feature Selection and Data Split using Brute Force to Improve Accuracy

Mahmud Mustapa[1], Ummiati Rahmah[2], Pandu Adi Cakranegara[3], Winci Firdaus[4], Dendi Pratama[5] and Robbi Rahim[6*]

[1]Associate Professor, Department Electronic Engineering Education, Universitas Negeri Makassar, Makassar, Indonesia. mahmud.mustapa@unm.ac.id, Orcid: https://orcid.org/0000-0001-8974-9728

[2]Associate Professor, Department Electronic Engineering Education, Universitas Negeri Makassar, Makassar, Indonesia. ummiati.rahmah@unm.ac.id, Orcid: https://orcid.org/0009-0006-5102-4119

[3]Assistant Professor, Universitas Presiden, Jakarta, Indonesia. pandu.cakranegara@president.ac.id Orcid: https://orcid.org/0000-0001-8754-3646

[4]Badan Riset dan Inovasi, Jakarta, Indonesia. wincifirdaus1@gmail.com Orcid: https://orcid.org/0000-0002-8261-4211

[5]Lecture Politeknik Bina Madani, Indonesia. dendi@poltekbima.ac.id Orcid: https://orcid.org/0000-0003-2002-6358

[6*]Lecturer Sekolah Tinggi Ilmu Manajemen Sukma, Medan, Indonesia. usurobbi85@zoho.com Orcid: https://orcid.org/0000-0001-6119-867X

Received: December 06, 2022; Accepted: January 12, 2023; Published: March 30, 2023

Abstract

This study seeks to classify data using feature selection and brute force. The dataset contains irrelevant characteristics, therefore feature selection influences computing time and the classification model. UCI's YouTube Spam Collection was used for testing. This dataset contains five datasets with 1,956 legitimate messages from five popular videos (Shakira, Katy Perry, Psy, Eminem, and LMFAO). Using weight information gain, the feature selection technique finds the best attributes. The dataset will then be separated into two parts: training with a 70:30 ratio and testing with a 30:70 ratio. Comparing using C4.5 and Nave Bayes. The FS+BF+C4.5 approach has an accuracy of 69.90%, 63.37%, 98.32%, 50.89%, and 91.75 for five videos (Psy, Katy Perry, LMFAO, Eminem and Shakira). Standard C4.5 technique accuracy is 66.99%, 59.41%, 95.80%, 50.89%, and 88.66%. Naive Bayes accuracy is 61.17, 51.49, 89.08, 50.00, and 79.38. FS+BF+C4.5 obtains an overall average accuracy of 74.85%, 2.5% and 8.6% higher than C4.5 and Naive Bayes (72.35 percent and 66.22 percent). Using feature selection and brute force with the C4.5 approach can reduce classification error compared to the normal C4.5 and Naive Bayes methods.

Keywords: Classification, Feature Selection, Brute Force, YouTube Spam Collection, Naive Bayes.

1 Introduction

Over the past few decades, researchers in computational intelligence have come up with a lot of ideas for data mining algorithms that can be used to solve classification problems in the real world (Bardab,

Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications (JoWUA), volume: 14, number: 1 (March), pp. 50-59. DOI: *10.58346/JOWUA.2023.I1.004*

*Corresponding author: Lecturer Sekolah Tinggi Ilmu Manajemen Sukma, Medan, Indonesia.

S.N., 2021) (Othman, N.A., 2021) (Wahab, A., 2020). In general, the function of classification in data mining is to describe and distinguish between data classes or concepts (Jatnika, H., 2021) (Widyastuti, M., 2019) (Supriyadi, B., 2018) (Sudarwanto, A.S., 2020) (Amar, M.Y., 2021) (Fasihah, E., 2023). The goal of classification is to accurately predict the class label of instances whose attribute values are known but whose class values are not (Fang, J., 2022) (Muhdi, Buchori, A., 2019). Decision Tree (Cheng, R., 2021) (Al-Hawari, A., 2021) (Sunandar, 2016) is an example of a data mining algorithm that is often used to sort data. Decision Tree (DT) is a classification algorithm that is often used in data mining, where one method is C4.5 (Novita, R., 2021) (Mardiansyah, H., 2021).

C4.5 is a supervised learning classification algorithm for constructing a decision tree from data with a high level of accuracy and interpretability (Chokkanathan, K., 2018). In the past decade, numerous studies have developed strategies (Budi, I.N., 2019) (Nasution, M.Z.F., 2018) to boost the classification's accuracy (Thohari, A.H., 2020), hence maximizing C4.5's capability.

As demonstrated by Nasution (Nasution, M.Z.F., 2018), feature reduction with the Principal Component Analysis (PCA) approach is used to the C4.5 classification algorithm. The dataset included for the evaluation is the cervical cancer dataset from UCI-machine learning. The experimental results indicate that PCA+C4.5 is 4.65% more accurate than C4.5 without PCA. It has been demonstrated that including PCA into the C4.5 approach improves accuracy. Mohanty (Mohanty, M., 2018) Additional research was undertaken on the C4.5 method's usage of feature selection and classification. Time-frequency and statistical features are employed for feature selection. The databases utilized for testing include The CU Ventricular Tachyarrhythmia Database (CUDB) and the MIT-BIH Malignant Ventricular Ectopy Database (VFDB). The experimental results demonstrate that the classification employing Time-frequency and statistical features in the C4.5 technique is more accurate than SVM, with respective accuracy values of 97.02 and 92.23 percent (an increase of 4.79 percent). Then, study was undertaken Aziz & Lawi (Aziz, F., 2022) utilizing ensemble bagging for the C4.5 and CART classifications on a benchmark power grid stability simulation dataset derived from UCI-machine learning. In terms of accuracy, the ensemble bagging strategy improved the performance of both algorithms (C4.5 and CART) by 5.6% and 5.3%, respectively, according to the experimental findings. Luo (Luo, J., 2021) did additional study in which the Bagging methodology was applied to the C4.5 method in the event of course failure. In comparison to the normal C4.5 method, the Bagging methodology in the C4.5 method shown a greater ability to predict course failure.

Based on these benefits, the objective of this study is to optimize the C4.5 classification approach by developing feature selection and brute force strategies to improve classification outcomes. In cases when the results will be compared to a classification method without development (standard).

2 Research Methodology

The dataset used for testing in research on the use of feature selection and split data on the brute force method to raise the accuracy value is the YouTube Spam Collection from the UCI Machine Learning Repository (https://archive.ics.uci.edu/ml/datasets /YouTube+Spam+Collection). This dataset contains public comments gathered for spam study. This dataset contains five datasets, each of which contains 1,956 genuine messages collected from five videos (Shakira, Katy Perry, Psy, and Eminem's LMFAO) that were among the top ten most viewed during the collecting period. As metadata, all samples include the author's name, publishing date, and time. The individual dataset information is represented in the table below, which includes the YouTube ID, number of samples in each class, and total number of samples.

Table 1: Dataset Description

Dataset	YouTube ID	Spam	Not Spam (Ham)	Total
Psy	9bZkp7q19f0	175	175	350
Katy Perry	CevxZvSJLk8	175	175	350
LMFAO	KQ6zr6kCPj8	236	202	438
Eminem	uelHwf8o7 U	245	203	448
Shakira	pRpeEdMmmQ0	174	196	370

In helping with the research, a computer with Intel(R) Core (TM) i5-4980HQ 2.80 GHz, 8 GB RAM, and the Windows 10 Pro operating system was used. For the process of analysis, the software Rapid Miner Studio 9.10 was used. This study suggests using feature selection and split data with the classification method to improve the accuracy of the brute force method. Accuracy will measure the data that comes out of the validation process. Figure 1 shows how the proposed method is set up as a model.

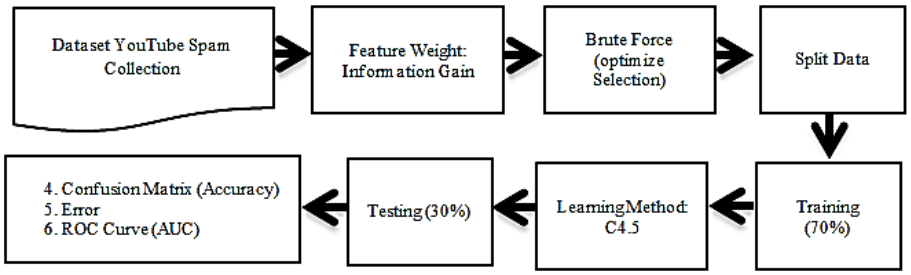


Figure 1: The Proposed Model

In Figure 1, the proposed application of feature selection is feature weight, also known as information gain and brute force approach with optimal selection employing the C4.5 classification algorithm. The dataset will be split between training (70 percent) and testing (30 percent). This investigation will yield precision and AUC value. These findings will be compared without optimization to other classification techniques.

3 Results and Discussion

RapidMiner 9.10 software is used in the experiment. The YouTube Spam Collection dataset was used for model testing. A total of 1,956 genuine messages have been retrieved from five videos (Shakira, Katy Perry, Psy, and Eminem LMFAO) that were among the ten most popular throughout the time period of the data collection. The proposed model will be evaluated on each set of data. Figures 2 and 3 below show the suggested model in contrast to two standard models (C4.5 and Naive Bayes).

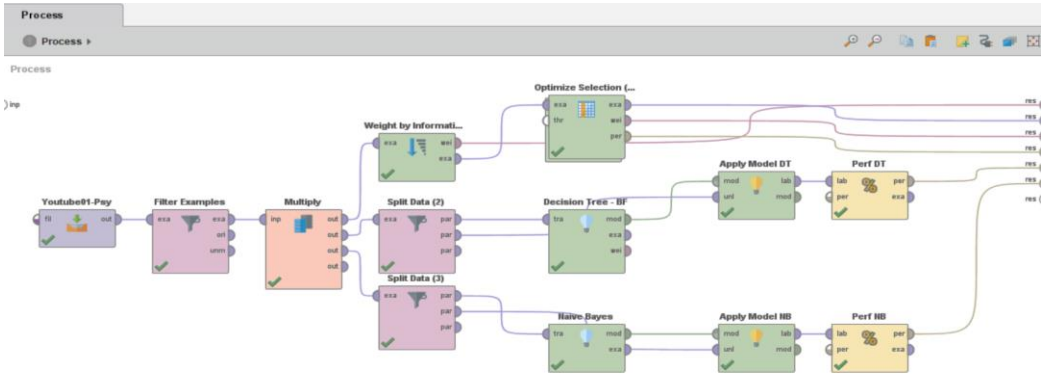


Figure 2: Proposed Model as a Whole with RapidMiner 9.10

Figure 2 shows the proposed model and two other standard classification models (C4.5 and Nave Bayes). The process starts with choosing a dataset, which is followed by choosing a "filter example." Then use feature selection to choose which attributes to use. This needs to be done because the dataset also has features that aren't important. "Weighted Information Gain" (WIG) was employed in Feature Selection. WIG is used to give each attribute a weight value because it is better for choosing the best attribute. Once you've chosen the best attribute, you can move on to optimizing the choice (Brute Force), which is a nested operator with subprocesses. This subprocess must always give a performance vector as a return value. This operator picks the set of features that gives the best performance vector. The C4.5 method is used by the subprocess to split the dataset into two parts: 70% for training and 30% for testing. In contrast, the standard C4.5 and Naive Bayes classification models do not use feature selection and brute force optimization. But the dataset is still split into two parts: 70% for training and 30% for testing.

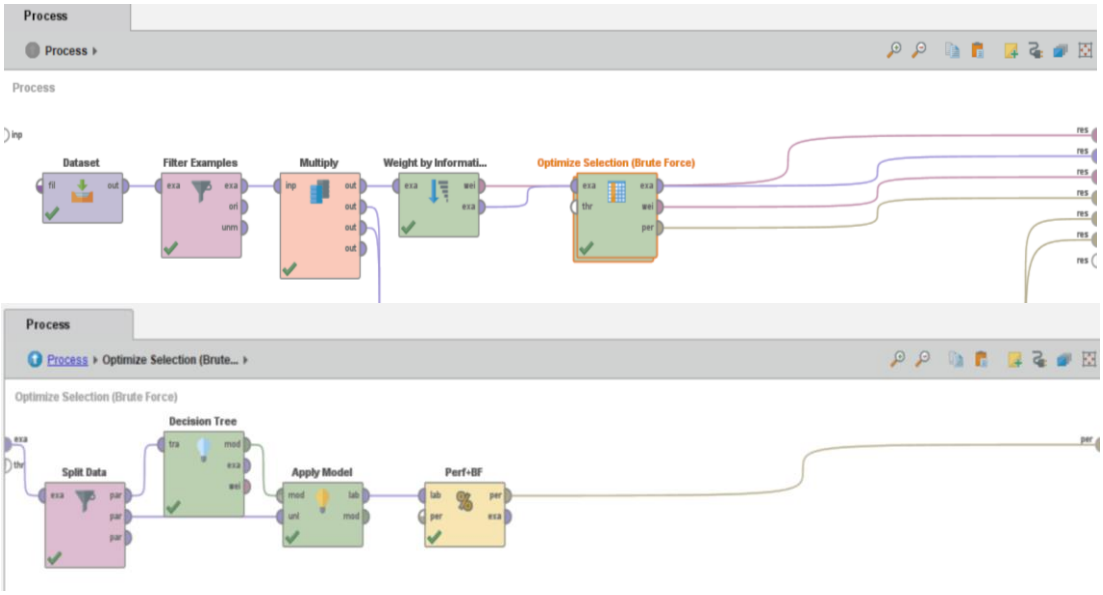


Figure 3: Explanation of Details of the Proposed Model

In Figure 3, all models are judged by their accuracy, which is measured with a confusion matrix that looks at classes in general. For AUC, the ROC Curve is used. From the "Psy" video data shown in Figure 4, here are the results of several accuracy tests using a confusion matrix.

accuracy: 69.90%

	true 1	true 0	class precision
pred. 1	33	13	71.74%
pred. 0	18	39	68.42%
class recall	64.71%	75.00%	

(a)

accuracy: 66.99%

	true 1	true 0	class precision
pred. 1	29	12	70.73%
pred. 0	22	40	64.52%
class recall	56.86%	76.92%	

(b)

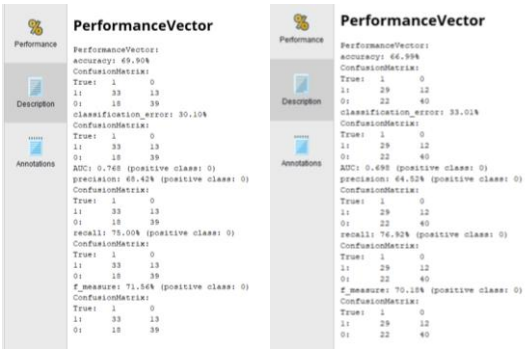
accuracy: 61.17%

	true 1	true 0	class precision
pred. 1	18	7	72.00%
pred. 0	33	45	57.69%
class recall	35.29%	86.54%	

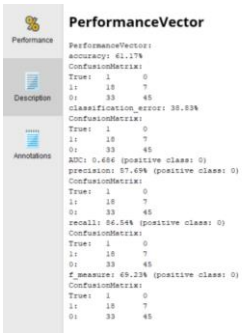
(c)

Figure 4: The Results of the Data Accuracy Analysis "Psy" is (a) FS+BF+C4.5; (b) C4.5; and (c) Naive Bayes

Figure 4 shows that combining the C4.5 method with feature selection and brute force (FS+BF+C4.5) has a higher accuracy rate of 69.9% than the simple classification method, which has an accuracy rate of 66.99% (C4.5) and 61.10%. (Naive Bayes). In the "Psy" data, the AUC can be seen in Figure 5 below.



(a) (b)



(c)

Figure 5: The Results of the Data AUC Analysis "Psy" is (a) FS+BF+C4.5; (b) C4.5; and (c) Naive Bayes

Using optimization, the model FS+BF+C4.5 in Figure 5(a) has an AUC value of 0.768%. This AUC result is superior to models (b)(c) with AUC values of 0.698 and 0.676, respectively. Following are the entire findings of the recapitulation of the accuracy values of all models from the YouTube Spam Collection dataset, as displayed in Table 2.

Table 2: Comparison of the Accuracy of all Classification Models

Dataset	Accuracy		
	FS+BF+C4.5	C4.5	Naïve Bayes
Psy	69.90%	66.99%	61.17%
Katy Perry	63.37%	59.41%	51.49%
LMFAO	98.32%	95.80%	89.08%
Eminem	50.89%	50.89%	50.00%
Shakira	91.75%	88.66%	79.38%

The overall accuracy of the FS+BF+C4.5 approach is greater than that of the standard classification method, as shown in Table 2. This is evident from all the dataset tests performed (Psy, Katy Perry, LMFAO, Eminem, and Shakira), where the FS+BF+C4.5 method has an average accuracy of 74.85 percent compared to the standard C4.5 method's 72.35 percent (an increase of approximately 2.5 percent) and the standard Nave Bayes method's 66.22 percent (an increase of about 8.62 percent). Figure 6 displays the outcomes of the comparison chart for each model.

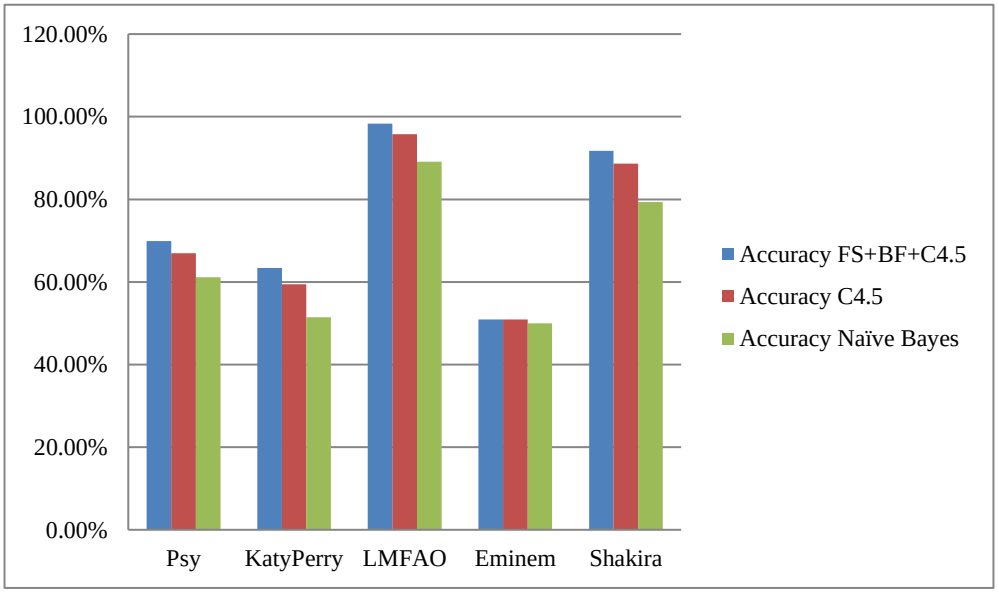


Figure 6: Accuracy Values for all Models Plotted on a Graph

Based on the results of experiments and evaluations in this study, in general it can be concluded that the application of feature selection and brute force techniques with the C4.5 (FS+BF+C4.5) method has the smallest error of all experiments where the "Psy" data is 30.10 percent (2.91 percent and 8.73 percent smaller than the C4.5 and Naïve Bayes methods); “Katy Perry” data of 36.63 percent (3.96 percent and 11.88 percent smaller than the C4.5 and Naïve Bayes methods); “LMFAO” data of 1.68 percent (2.52 percent and 9.24 percent smaller than the C4.5 and Naïve Bayes methods); “Eminem” data of 49.11 percent (0 percent and 0.89 percent smaller than the C4.5 and Naïve Bayes methods); and “Shakira” data of 8.25 percent (3.09 percent and 12.37 percent smaller than the C4.5 and Naïve Bayes methods). The following is the result of the recapitulation of the error comparison of all models as shown in Table 3 below.

Table 3: Comparison of Error Values in all Classification Models

Dataset	Error		
	FS+BF+C4.5	C4.5	Naïve Bayes
Psy	30.10%	33.01%	38.83%
Katy Perry	36.63%	40.59%	48.51%
LMFAO	1.68%	4.20%	10.92%
Eminem	49.11%	49.11%	50.00%
Shakira	8.25%	11.34%	20.62%

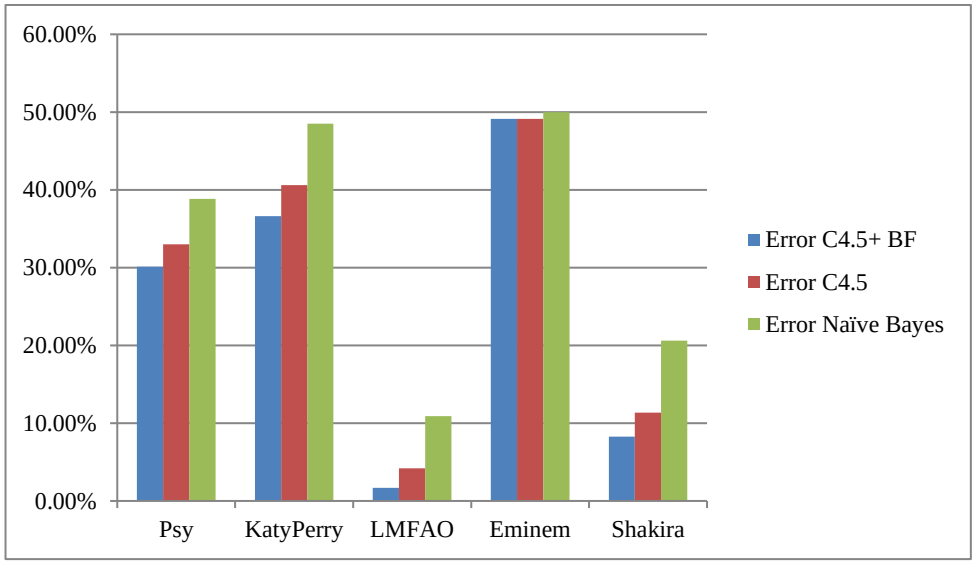


Figure 7: Error Values for all Models Plotted on a Graph

Based on the results of the experiments and evaluations in this study, it can be said that using feature selection and brute force techniques to classify the YouTube Spam Collection dataset can make it more accurate. This is shown by the fact that it does better than all standard classification models.

4 Conclusion

The application of feature selection and data split techniques, along with brute force optimization, can be used to improve the classification's accuracy. The C4.5 method is used to optimize the classification, and its accuracy results are compared to those of other classification techniques, including C4.5 and Nave Bayes. The experimental results outperformed all other results based on a sample of 1,956 actual messages extracted from five videos (Shakira, Katy Perry, Psy, Eminemand LMFAO). In every experiment, the standard C4.5 and Nave Bayes methods improved accuracy by 2.50 and 8.62 percent, respectively.

References

[1] Al-Hawari, A., Najadat, H., Shatnawi, R. (2021). Classification of application reviews into software maintenance tasks using data mining techniques, *Software Quality Journal*, 29(3), 667–703.

[2] Amar, M.Y., Syariati, A., Ridwan, R., Parmitasari, R.D.A. (2021). Indonesian hotels' dynamic capability under the risks of covid-19, *Risks*, 9(11), 1-18.

[3] Aziz, F., Lawi, A. (2022). Increasing electrical grid stability classification performance using ensemble bagging of C4.5 and classification and regression trees, *International Journal of Electrical and Computer Engineering*, 12(3), 2955–2962.
[4] Bardab, S.N., Ahmed, T.M., Mohammed, T.A.A. (2021). Data mining classification algorithms: An overview, *International Journal of Advanced and Applied Sciences*, 8(2), 1–5.
[5] Budi, I.N., Ranggadara, I., Prihandi, I., Kurnianda, N.R., Suhendra. (2019). Prediction using C4.5 method and RFM method for selling furniture, *International Journal of Engineering and Advanced Technology*, 9(1), 535–541.
[6] Cheng, R., Kong, X., Yu, M., Wang, N. (2021). A classification algorithm: Data mining and mathematical model, *Journal of Physics: Conference Series*, 2068(1), 1-6.
[7] Chokkanathan, K., Koteeswaran, S. (2018). Privacy protection and perfect classification nature of C4.5 algorithm, *International Journal of Engineering and Technology(UAE)*, 7(2), 235–238.
[8] Fang, J., Li, J. (2022). Research on Classification of Primary Liver Cancer Syndrome Based on Data Mining Technology, *Journal of Healthcare Engineering*.
[9] Fasihah, E., Hidayah, N., Mubarokah, A. (2023). Validation Analysis of Learning Development on Optic Material with Core Learning Model, *SAGA: Journal of Technology and Information System*, 1(1), 1–4.
[10] Jatnika, H., Huda, M., Amelia, R.R., Manuhutu, M.A. (2021). Analysis of Data Mining in the Group of Water Pollution Areas using the K-Means Method in Indonesia Analysis of Data Mining in the Group of Water Pollution Areas using the K-Means Method in Indonesia, *Journal of Physics: Conference Series PAPER*.
[11] Lu, Y., & Ishida, T. (2020). Implementation and Evaluation of a High-presence Interior Layout Simulation System using Mixed Reality. *Journal of Internet Services and Information Security*, *10*(1), 50-63.
[12] Luo, J. (2021). Study on Prediction of Course Failure Based on Improved Bagging-C4.5 Algorithm, *Journal of Physics: Conference Series*, 1861(1), 1-7.
[13] Mardiansyah, H., Zarlis, M., Sitompul, O.S. (2021). Analysis of C4.5 Algorithm of Water Quality Dataset, *Journal of Physics: Conference Series*, 1898(1), 1-6.
[14] Mohanty, M., Sahoo, S., Biswal, P., Sabut, S. (2018). Efficient classification of ventricular arrhythmias using feature selection and C4.5 classifier, *Biomedical Signal Processing and Control*, *44*, 200–208.
[15] Muhdi, Buchori, A., Wibisono, A. (2019). Whiteboard animation for android design using think talk write model to improve the post graduates students' concepts understanding, *Journal of Advanced Research in Dynamical and Control Systems*, 11(7), 535–543
[16] Nasution, M.Z.F., Sitompul, O.S., Ramli, M. (2018). PCA based feature reduction to improve the accuracy of decision tree c4.5 classification, *Journal of Physics: Conference Series*, 978(1), 1-6.
[17] Novita, R., Zakir, S., Nur Khomarudin, A., Maiyana, E., Hasyim, H. (2021). Use of the C4.5 Algorithm in Determining Scholarship Recipients, *Journal of Physics: Conference Series*, 1779(1), 1-8.
[18] Othman, N.A., Foozy, C.F.M., Mustapha, A., Mostafa, S.A., Palaniappan, S., Kashinath, S.A. (2021). A data mining approach for classification of traffic violations types, *International Journal of Advances in Intelligent Informatics*, 7(3), 282–291.
[19] Sudarwanto, A.S., Pujiyono. (2020). Responsibilities of banks to loss of customers using mobile banking, *International Journal of Advanced Science and Technology*, 29(4), 1702–1706.
[20] Sunandar; Buchori, A., Rahmawati, N.D. (2016). Development of media kocerin (Smart box interactive) to learning mathematics in Junior High School, *Global Journal of Pure and Applied Mathematics*, 12(6), 5253–5266
[21] Supriyadi, B., Windarto, A.P., Soemartono, T., Mungad. (2018). Classification of natural disaster prone areas in Indonesia using K-means, *International Journal of Grid and Distributed Computing*, 11(8), 87–98.

[22] Thohari, A.H., Anita, W.S. (2020). Smart dunning to improve collection ratio in internet service provider using C4.5 algorithm, *Journal of Physics: Conference Series*, 1450(1), 1-5.
[23] Wahab, A., Abbas, N., Syariati, A., Syariati, N.E. (2020). The Trickle-Down Effect of Intellectual Capital on Banks' MacroPerformance in Indonesia, *Journal of Asian Finance, Economics and Business*, 7(12), 703–710.
[24] Widyastuti, M., Fepdiani Simanjuntak, A.G., Hartama, D., Windarto, A.P., Wanto, A. (2019). Classification Model C.45 on Determining the Quality of Custumer Service in Bank BTN Pematangsiantar Branch, *Journal of Physics: Conference Series*, 1255, 1–6.

Author Biography

Mahmud Mustapa

Mahmud mustapa received S.Pd. degree in electronics engineering education from Makassar State University, Indonesia, in 1992 and M.Pd. degree in technology and vocational education from Yogyakarta State University, Indonesia in 1999 and Dr. degree in education science 2016. Currently, he is an Associate Professor in the Department of Electronic Engineering Education, Makassar State University. His research interests include education and learning media, Internet of Things, digital electronics, audio video and vocational education.
e-mail: mahmud.mustapa@unm.ac.id
Orcid: 0000-0001-8974-9728

Ummiati Rahmah

Ummiati Rahmah received S.Pd. degree in electronics engineering from Makassar State University, Indonesia, in 1995 and M.T. degree in informatics engineering from Gajah Mada University, Indonesia in 2000 and Dr. degree. in the field of educational technology in 2015. Currently, he is an Associate Professor at the Department of Electronic Engineering Education, Makassar State University. His research interests include character education, Internet of Things, Learning Media, Educational Psychology, Student development and development of learning models.
e-mail: ummiati.rahmah@unm.ac.id
Orcid: 0009-0006-5102-4119

Pandu Adi Cakranegara

Pandu Adi Cakranegara is asisstant professor at President University, currently his research and teaching is in business related field. He received his doctorate from the Phillippine Women University, master in finance from Erasmus University, and MBA and Bachelor from Gadjah Mada.
e-mail: pandu.cakranegara@president.ac.id
Orcid: 0000-0001-8754-3646

Winci Firdaus

Winci Firdaus completed a master's program in linguistics at Padjadjaran University, currently the author is active in writing in the fields of education, linguistics, and technology related to linguistics, as well as social sciences. Worked as a researcher at the National Research and Innovation Agency.
e-mail:wincifirdaus@yahoo.com
Orcid: 0000-0002-8261-4211
Scopus ID: 57205062723

Dendi Pratama

Dendi Pratama completed Bachelor of Visual Communication Design and Master of Visual Communication Design at Trisakti University in 2001 and 2011. In 2003 completed Master of Management at STM PPM with concentration in Human Resources. His Doctoral degree was received in Visual Art at ISI Surakarta in 2019. Now, he is working as a lecture at the Bina Madani Polytechnic and active on research with a focus areas of visual communication design, visual characters, media design, visual arts, and folklore.e-mail: dendi@poltekbima.ac.id
Orcid: 0000-0003-2002-6358

Robbi Rahim

Robbi Rahim is an Indonesian academic with a Doctoral degree in Protocol Cryptography from Universiti Malaysia Perlis. He has expertise in the fields of data mining, big data, and Rapid Miner, all of which are related to the processing and analysis of large datasets. Rahim's doctoral thesis focused on the study of Protocol Cryptography, which involves securing communication protocols using cryptographic techniques. His contributions to research in various fields, including computer science and information technology, have been significant. Since 2017, Rahim has been working as a lecturer at Sekolah Tinggi Ilmu Manajemen Sukma. In his current role, he teaches and mentors students in the areas of data mining, big data, and Rapid Miner. His expertise in these fields has enabled him to bring a unique perspective to his teaching, helping students to develop the skills and knowledge needed to succeed in today's technology-driven world.
Email: usurobbi85@zoho.com
Orcid: https://orcid.org/0000-0001-6119-867X