Recursive Journal of Informatics
Vol. 1, No. 1, March 2023
p-ISSN 2963-5551 https://journal.unnes.ac.id/sju/index.php/rji e-ISSN XXXX-XXXX

C4.5 Algorithm Optimization and Support Vector Machine by Applying Particle Swarm Optimization for Chronic Kidney Disease Diagnosis

Lisa Ariyanti[1], Alamsyah[2]
[1,2]Department of Computer Science, Faculty of Mathematics and Natural Sciences,
Universitas Negeri Semarang, Indonesia

Abstract. Kidneys are one of the organs of the body that have a very important function in life. The main function of the kidneys is to excrete metabolic waste products. Chronic kidney disease is a result of the gradual loss of kidney function. Chronic kidney disease occurs when the kidneys are unable to maintain an internal environment consistent with life and the restoration of useless functions.
Purpose: Data mining is one of the fastest growing technologies in biomedical science and research. In the field of medicine, data mining can improve hospital information management and telemedicine development. In the first stage of data mining process, data processing is done with pre-processing by handling missing values and data transformation. Then, the feature selection stage is carried out using the Particle Swarm Optimization algorithm to find the best attributes. Next, it is done by classifying the dataset.
Methods/Study design/approach: The algorithm used for classification is the C4.5 Algorithm and the Support Vector Machine. Both classifications are known as algorithms that have a fairly good level of accuracy. This study uses the chronic kidney disease dataset from the UCI Machine Learning Repository.
Result/Findings: The purpose of this study was to determine the level of accuracy of the comparison between the C4.5 Algorithm and the Support Vector Machine after applying the Particle Swarm Optimization algorithm.
Novelty/Originality/Value: This research increases the accuracy by 100% for the C4.5 Algorithm and 98.75% for the Support Vector Machine by using 24 attributes and 1 class attribute.

Keywords: Chronic Kidney Disease, C4.5 Algorithm, Support Vector Machine, Particle Swarm Optimization
Received January 18, 2023 / **Revised** January 27, 2023 / **Accepted** February 05, 2023

INTRODUCTION

Along with the rapid development of technology, the amount of data is increasing at an incredible speed. In the field of computer science. Data Mining refers to datasets that cannot be sensed, acquired, managed and processed even presented in a tolerable time using hardware. Data mining can also refer to datasets that go beyond the scope of simple databases in everyday life [1]. Data mining, also known as knowledge discovery, is a technique that analyzes data to find patterns and is particularly helpful in solving problems related to large data warehouses. There are different types of methods for data mining, and choosing the appropriate method depends on the intended purpose and the process involved. Classification is one such method in data mining [2].

Data mining is one of the fastest growing technologies in biomedical science and research [3]. Utilizing data mining techniques can enable the processing of a large number of medical record data. In bioinformatics, data mining has gained wide application for analyzing biomedical data. Researchers can extract valuable information from these medical records by employing techniques such as classification, which predicts or describes data classes. Medical experts can use the classification algorithm to diagnose diseases, including hepatitis, which poses a challenge due to the complexity of factors that need to be considered and analyzed. This study will use classification to predict patient data for hepatitis diagnosis [4], [5]. One of the diseases that can be predicted by data mining is chronic kidney disease. Chronic kidney disease is the result of a gradual loss of kidney function. Chronic kidney disease occurs when the kidneys are not able to maintain an internal environment that is consistent with life and restore useless functions

[1]*Corresponding author.
Email address: lisaariyanti@students.unnes.ac.id (Ariyanti)
DOI: 10.15294/rji.v1i1.65196

[6]. Support Vector Machine is a classification method with supervised learning which is useful for performing data analysis and pattern recognition. SVM works by searching for the best hyperplane by maximizing the distance between each class [7]. In a previous study, SVM affected the output and a high degree of accuracy for the prediction of chronic kidney disease [8].

The C4.5 algorithm is a machine learning algorithm. This algorithm is used to give a group of data to be studied as a learning dataset. Furthermore, the results of the C4.5 algorithm will be used to process new data called the test dataset because the C4.5 algorithm is used for classification [9]. Particle Swarm Optimization is one of several feature selection algorithms or the best attribute search, which is used to find random solutions and one of the optimization algorithms used for decision making. PSO is used for the selection of the optimum attribute features to be used in the classification process, and PSO requires a fitness value as a feature selection to find the best solution attribute.

Based on the description above, this study focuses and aims to improve accuracy in the diagnosis of chronic kidney disease using the C4.5 classification algorithm and the Support Vector Machine with Particle Swarm Optimization feature selection technique.

METHODS

Data Mining

Data Mining is a process of searching for interesting and hidden patterns from a large data set stored in databases, data warehouses or other storage places. According to Rudianto [10] Data Mining is an analysis of data to find clear relationships and conclude that was not previously known. As long as the data is useful for the target system or application, data mining can be applied to all types of data. Forms of Data Mining in general are database data, data warehouse, and transactional data. Particle Swarm Optimization.

Particle Swarm Optimization is one of the known optimization methods where this method is inspired by a group of animals that provide food in groups and PSO is described where each particle with a good value will be updated according to the position of each particle, PSO works to find the value of each particle will move towards the best previous position (pbest) and the best global position (gbest) and each particle has a value that is evaluated using the fitness function to be optimal. The position and speed will control the movement of each particle [11]. PSO is used as a tool to solve a problem in optimization and for feature selection [12].

C4.5 Algorithm

C4.5 algorithm is one of the classification algorithms by creating a decision tree that is used to predict a decision in accordance with the decision-making rules [13]. According to Umam [14] Decision trees are very powerful and well-known classifications and predictions, decision trees are useful for exploring data, finding hidden relationships between a number of potential input variables and a target variable.

Support Vector Machine

Support Vector Machine is a technique for carrying out predictions in the case of Classification or regression. Support Vector Machine has the basic principle of a linear classifier, namely Classification cases that are linearly distinguishable. However, SVM was developed to be able to work on non-linear problems by including the kernel concept in the workspace giving it high dimensions. The kernel has a function to map the initial dimensions or dimensions with lower to new dimensions or relatively higher dimensions [15].

SVM is a fast and effective method for text classification, one of the problems in text classification or text data processing is the number of features/attributes used on a dataset that will degrade the performance of the classifier. To optimize the work of the classifier needs to be done by selecting relevant features using feature selection. Feature selection is used to reduce feature/attribute dimension by removing irrelevant words to improve classification accuracy [8].

Confusion Matrix

Confusion matrix is one of the methods used to perform good level analysis. For the evaluation process with the confusion matrix, the precision, recall, and accuracy values obtained are shown in Table 1 and shown in Equation 1.

Table 1. Representation Confusion Matrix

Actual	Predicted	
	Positive	Negative
Positive	True Positive (TP)	False Negative (FN)
Negative	False Positive (FP)	True Negative (TN)

$$\text{Accuracy} = \frac{TP+TN}{P+N} \times 100\% \quad (1)$$

Will be carried out in this research is by implementing the selected algorithm features so that they can run normally. The framework of thinking in this study is to compare the level of accuracy in a classification algorithm, namely C4.5 Algorithm and Support Vector Machine by applying Particle Swarm Optimization in diagnosis of chronic kidney disease. The model framework to be carried out can be seen in Figure 1.

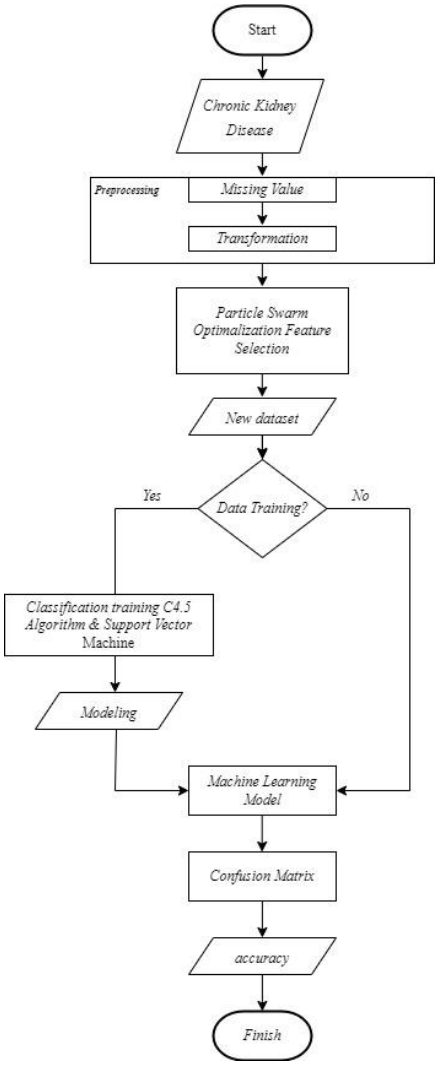

Figure 1. Flowchart Research Design

RESULT AND DISCUSSION

This section is divided into two parts, results and discussion. The result is a description of the data and findings obtained using the methods and procedures described in the data collection method. The discussion is an explanation of the results that answer research questions more comprehensively.

Results

The results of this study used Particle Swarm Optimization, C4.5 Algorithm and Support Vector Machine to diagnose chronic kidney disease. Particle Swarm Optimization is used as feature selection in chronic kidney disease, and the C4.5 Algorithm and Support Vector Machine are used for the classification process in chronic kidney disease. The research was carried out in several stages, namely the data preprocessing stage, the feature selection stage, and the data mining stage. The following is a more complete explanation of the research results.

Preprocessing Results

Before entering into the data mining, the data is processed in the preprocessing. This process is used to prepare data so that it can be processed for mining and produce high accuracy. The preprocessing in this research is handling missing values and data transformation.

Missing Value Handling

Handling Missing Value is done to handle missing values in the chronic kidney disease dataset. Handling missing values is done by correcting the kidney data before the mining. Because, in the dataset found data that has the value "?" or can be referred to as missing data as shown in Table 2. which displays some of the missing value contained in the dataset.

Table 2. Missing Value Handling

No	Attribute	Modus
1	Age	60
2	Blood Pressure	80
3	Specific Gravity	1.020
4	Albumin	0
5	Sugar	0
6	Red Blood Cells	Normal
7	Pus Cell	Normal
8	Pus Cell Clumps	Not present
9	Bacteria	Not present
10	Blood Glucoses Random	99
11	Blood Urea	46
12	Serum Creatinine	1.2
13	Sodium	135
14	Potassium	3.5
15	Hemoglobin	15.0
16	Packed Cell Volume	41
17	White Blood Cell Count	9.800
18	Red Blood Cell Count	5.2
19	Hypertension	No
20	Diabetes Mellitus	No
21	Coronary Artery Disease	No
22	Appetite	Good
23	Pedal Edema	No
24	Anemia	No

Data Transformation

Data transformation aims to create a standard format for the datasets used in the research. The transformation carried out on the chronic kidney disease dataset shown in Table 3.

Table 3. Data Transformation

No	Attribute	Transformation
1	Rbc (Red Blood Cells)	Changing the Rbc attribute to 0 for negative abnormal and 1 positive normal
2	Pc (Pus Cell)	Changing the Pc attribute to 0 for negative abnormal and 1 to positive normal
3	Pcc (Pus Cell Clumps)	Changing the Pcc attribute to 0 for negative not present and 1 for positive present
4	Ba (Bacteria)	Changing the Ba to 0 for negative not present and 1 for positive present
5	Htn (Hypertension)	Changing the Htn to 0 for negative no and 1 for positive yes
6	Dm (Diabetes Mellitus)	Changing the Dm to 0 for negative no and 1 for positive yes
7	Cad (coronary artery disease)	Changing the Cad to 0 for negative no and 1 for positive yes
8	Appet (Appetite)	Changing the Appet to 0 for negative good and 1 for positive poor
9	Pe (Pedal Edema)	Changing the Pe to 0 for negative no and 1 for positive yes
10	Ane (Anemia)	Changing the Ane to 0 for negative no and 1 for positive yes
11	Class	Changing the Class to 0 for negative notckd and 1 for positif ckd

Feature Selection Results

At this stage, the PSO algorithm will be used as a feature selection process or attribute selection which is then continued with classification using one of the algorithms, namely Algorithm C4.5. PSO is used to find feature set which will be used in the classification process. The selected feature will be symbolized by the number 1 and the feature that is not selected is symbolized by the number 0. The selected feature from the feature selection process is shown in Table 4 and Table 5.

Table 4. Selected Attributes of the PSO Process on the C4.5 Algorithm

Attributes	Representation
Age	1
Blood Pressure	1
Specific Gravity	1
Albumin	1
Sugar	1
Red Blood Cells	1
Pus Cell	1
Pus Cell Clumps	0
Bacteria	1
Blood Glucoses Random	1
Blood Urea	1
Serum Creatinine	1
Sodium	0
Potassium	1
Hemoglobin	0
Packed Cell Volume	1
White Blood Cell Count	1
Red Blood Cell Count	1
Hypertension	1
Diabetes Mellitus	1
Coronary Artery Disease	1
Appetite	1
Pedal Edema	1
Anemia	1

Table 5. Selected Attributes of the PSO Process on the SVM

Attributes	Representation
Age	0
Blood Pressure	0
Specific Gravity	1
Albumin	1
Sugar	1
Red Blood Cells	1
Pus Cell	0
Pus Cell Clumps	1
Bacteria	1
Blood Glucoses Random	0
Blood Urea	0
Serum Creatinine	1
Sodium	0
Potassium	1
Hemoglobin	1
Packed Cell Volume	0
White Blood Cell Count	0
Red Blood Cell Count	1
Hypertension	1
Diabetes Mellitus	1
Coronary Artery Disease	1
Appetite	1
Pedal Edema	1
Anemia	1

Data Mining

At this stage, the researcher carried out the mining. First, the classification process using the C4.5 algorithm on the chronic kidney disease dataset. Second, the classification process using the Support Vector Machine on the chronic kidney disease dataset. Third, the classification process of the C4.5 Algorithm on the chronic kidney disease dataset that has been carried out feature selection with the Particle Swarm Optimization.

Fourth, the classification process of the Support Vector Machine in Chronic Kidney Disease which has been carried out feature selection with the Particle Swarm Optimization algorithm.

C4.5 Algorithm Classification Results

At this stage, the chronic kidney disease dataset is classified using the C4.5 Algorithm without using feature selection. Then, the training data is processed using the C4.5 Algorithm for model testing. Performance evaluation of the C4.5 algorithm is calculated using a confusion matrix and can be seen in Table 6.

Table 6. Confusion Matrix C4.5 Algorithm

	True Positive	True Negative	Total
Pred Positive	28	1	29
Pred Negative	2	49	51
Total	30	50	80

$$Accuracy = \frac{TP+TN}{P+N}\ x\ 100\% \quad (2)$$

$$Accuracy = \frac{28+49}{29+51}\ x\ 100\% = 96{,}25\%$$

Support Vector Machine Classification Results

At this stage, the chronic kidney disease dataset is classified using SVM without using feature selection. Then, the training data is processed using SVM for model testing. SVM performance evaluation is calculated using a confusion matrix and can be seen in Table 7.

Table 7. Confusion Matrix SVM

	True Positive	True Negative	Total
Pred Positive	0	29	29
Pred Negative	0	51	51
Total	0	80	80

$$Accuracy = \frac{TP+TN}{P+N}\ x\ 100\% \quad (3)$$

$$Accuracy = \frac{0+51}{29+51}\ x\ 100\% = 63{,}75\%$$

Classification Results of C4.5 Algorithm by Applying PSO

Chronic Kidney Disease dataset was classified using the C4.5 Algorithm by applying the Particle Swarm Optimization. Then, the training data is processed using C4.5 Algorithm for model testing. The evaluation of C4.5 Algorithm performance is calculated using a confusion matrix and can be seen in the table 8 for the highest accuracy and table 9 for the lowest accuracy. The process of calculating the C4.5 Algorithm classification using the PSO feature selection was experimented with 10 times. And from 10 times of testing, the best and lowest accuracy were taken as the final result of this research. The following are the results of the accuracy of 10 experiments, each of which can be seen in Table 8.

Table 8. Confusion Matrix Highest Accuracy C4.5 + PSO Algorithm

	True Positive	True Negative	Total
Pred Positive	29	0	29
Pred Negative	0	50	50
Total	29	50	79

Table 9. Confusion Matrix Lowest Accuracy C4.5+PSO Algorithm

	True Positive	True Negative	Total
Pred Positive	29	0	29
Pred Negative	4	47	51
Total	33	47	80

$$Accuracy = \frac{TP+TN}{P+N} \; x \; 100\% \tag{4}$$

$$Accuracy = \frac{29+50}{29+50} \; x \; 100\% = 100\%$$

Table 10. Results 10 Trials C4.5+PSO

Execution	Accuracy
1	97.5 %
2	96.25 %
3	97.5%
4	97.5%
5	98.75%
6	96.25%
7	95%
8	98.75%
9	100%
10	95%

$$Accuracy = \frac{TP+TN}{P+N} \; x \; 100\% \tag{5}$$

$$Accuracy = \frac{29+47}{29+51} \; x \; 100\% = 95\%$$

Classification Results of SVM by Applying PSO

Chronic Kidney Disease dataset was classified using the SVM by applying the Particle Swarm Optimization. Then, the training data is processed using SVM for model testing. The evaluation of SVM performance is calculated using a confusion matrix and can be seen in the Table 12 for the highest accuracy and Table 13 for the lowest accuracy. The process of calculating the SVM classification using the PSO feature selection was experimented with 10 times. And from 10 times of testing, the best and lowest accuracy were taken as the final result of this research. The following are the results of the accuracy of 10 experiments, each of which can be seen in Table 11.

Table 11. Results 10 Trials SVM + PSO

Execution	Accuracy
1	92.50 %
2	92.75 %
3	95%
4	98.75%
5	96.25%
6	96.25%
7	92.5%
8	98.75%
9	92.5%
10	95%

Table 12. Confusion Matrix Highest Accuracy SVM + PSO

	True Positive	True Negative	Total
Pred Positive	29	0	29
Pred Negative	1	50	51
Total	30	50	80

$$Accuracy = \frac{TP+TN}{P+N}\ x\ 100\% \quad (6)$$

$$Accuracy = \frac{29+50}{29+51}\ x\ 100\% = 98{,}75\%$$

Table 13. Confusion Matrix Lowest Accuracy SVM+PSO

	True Positive	True Negative	Total
Pred Positive	25	4	29
Pred Negative	2	49	51
Total	27	52	79

$$Accuracy = \frac{TP+TN}{P+N}\ x\ 100\% \quad (7)$$

$$Accuracy = \frac{25+4}{29+51}\ x\ 100\% = 92.50\%$$

Discussion

In this study, the accuracy of the C4.5 Algorithm and Support Vector Machine using the Particle Swarm Optimization to increase the accuracy of chronic kidney disease taken from the UCI Machine Learning Repository. Then, the data is carried out in preprocessing, namely handling missing values and data transformation before the process is carried out using an algorithm. From the results obtained, it shows that the highest accuracy results are obtained by the application of the C4.5 Algorithm by applying the Particle Swarm Optimization with an accuracy of 100%. So, it can be concluded that the C4.5 algorithm by applying the Particle Swarm Optimization for the diagnosis of chronic kidney disease is better than Support Vector Machine feature selection Particle Swarm Optimization. To find out the method used in the study, it is better to use other methods. The results of the comparison can be seen in Table Figure. Each figure should have a brief caption describing it and, if necessary, a key to interpret the various lines and symbols on the Table 12.

Table 12. Comparison of Accuracy with Other Methods

Authors	Methods	Results
Ananta Padmanaban & Parthiban (2016)	Decision Tree	91%
	Naïve Bayes	86%
ADEM (2018)	SVM+PSO	98.25%
	KNN+PSO	95.75%
Boukenze, et al (2016)	C4.5	63%
	SVM	60.25%
	NB	57.5%
Ifraz, et al (2021)	Decision Tree	96%
	K-NN	71%
Proposed Method	SVM	63.75%
	C4.5	96.25%
	SVM+PSO	98.75%
	C4.5+PSO	100%

CONCLUSION

Based on the results and discussion of the research that has been carried out, regarding the application of Particle Swarm Optimization for optimization using the C4.5 Algorithm classification and Support Vector Machine to diagnose Chronic Kidney using the Chronic Kidney Disease Dataset obtained from the UCI Machine Learning Repository and it can be concluded that accuracy Chronic Kidney's before using feature

selection, the classification of the C4.5 Algorithm and Support Vector Machine without using feature selection got an accuracy of 96.25% and 63.75%, respectively. Then, after applying the Particle Swarm Optimization classification of the C4.5 Algorithm and Support Vector Machine, the accuracy is 100% and 98.75%, respectively. By applying the Particle Swarm Optimization for the diagnosis of chronic kidney disease and the classification of the C4.5 Algorithm after applying the PSO feature selection, the highest accuracy is 100% and the lowest is 95% with an average of 97.25% so that the accuracy of the C4.5 algorithm is better than classification Support Vector Machine for diagnosing chronic kidney disease.

REFERENCES

[1] W.-T. Wu *et al.*, "Data mining in clinical big data: the frequently used databases, steps, and methodological models," *Mil Med Res*, vol. 8, no. 1, p. 44, Aug. 2021, doi: 10.1186/s40779-021-00338-z.

[2] A. Lestari, "Increasing accuracy of C4. 5 algorithm using information gain ratio and adaboost for classification of chronic kidney disease," *Journal of Soft Computing Exploration*, vol. 1, no. 1, pp. 32–38, 2020.

[3] S. Joshi and H. Joshi, "Applications of Data Mining in health and pharmaceutical industry," *Int J Sci Eng Res*, vol. 4, no. 4, 2013.

[4] J. A. M. Nugraha and Y. Kusumawati, "Data Mining dengan Metode Clustering untuk Pengolahan Informasi Persediaan Obat pada Puskesmas Pandanaran Semarang. Universitas Dian Nuswantoro," *Universitas Dian Nuswantoro*, 2014.

[5] A. Alamsyah and T. Fadila, "Increased accuracy of prediction hepatitis disease using the application of principal component analysis on a support vector machine," *J Phys Conf Ser*, vol. 1968, no. 1, p. 012016, Jul. 2021, doi: 10.1088/1742-6596/1968/1/012016.

[6] R. Rianto and N. M. S. Iswari, "Rancang bangun aplikasi pendeteksi penyakit ginjal kronis dengan menggunakan algoritma c4. 5," *Ultimatics: Jurnal Teknik Informatika*, vol. 9, no. 1, pp. 10–18, 2017.

[7] Oryza Habibie Rahman, Gunawan Abdillah, and Agus Komarudin, "Klasifikasi Ujaran Kebencian pada Media Sosial Twitter Menggunakan Support Vector Machine," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 5, no. 1, pp. 17–23, Feb. 2021, doi: 10.29207/resti.v5i1.2700.

[8] U. I. Larasati, M. A. Muslim, R. Arifudin, and A. Alamsyah, "Improve the Accuracy of Support Vector Machine Using Chi Square Statistic and Term Frequency Inverse Document Frequency on Movie Review Sentiment Analysis," *Scientific Journal of Informatics*, vol. 6, no. 1, pp. 138–149, May 2019, doi: 10.15294/sji.v6i1.14244.

[9] Y. Irawan, "Penerapan Algoritma Decision Tree C4.5 Untuk Memprediksi Kelayakan Calon Pendonor Melakukan Donor Darah Dengan Klasifikasi Data Mining," *JTIM : Jurnal Teknologi Informasi dan Multimedia*, vol. 2, no. 4, pp. 181–189, Feb. 2021, doi: 10.35746/jtim.v2i4.75.

[10] A. Rudianto, "Penerapan Data Mining Untuk Menganalisa Data Penjualan Pada Toko Bangunan Restu Bunda Menggunakan Metode Apriori," 2019.

[11] Z. Rustam, M. A. Syarifah, and T. Siswantining, "Recursive Particle Swarm Optimization (RPSO) schemed Support Vector Machine (SVM) Implementation for Microarray Data Analysis on Chronic Kidney Disease (CKD)," *IOP Conf Ser Mater Sci Eng*, vol. 546, no. 5, p. 052077, Jun. 2019, doi: 10.1088/1757-899X/546/5/052077.

[12] E. Supriyadi and D. I. Sensuse, "Optimasi Algoritma Support Vector Machine Dengan Particle Swarm Optimization dalam Mendeteksi Ketepatan Waktu Kelulusan Mahasiswa : Studi Kasus Poltek LP3I Jakarta 'Kampus Depok," 2015. [Online]. Available: http://www.nusamandiri.ac.id,

[13] R. S. Putra, E. D. Putra, M. H. Rifqo, and H. Witriyono, "Klasifikasi Penyebaran Covid-19 Menggunakan Algoritma C4. 5 Kota Pagar Alam," *JUKOMIKA (Jurnal Ilmu Komput. dan Inform., vol. 4, no. 1, pp. 23–35, 2021, doi: 10.54650/jukomika. v4i1. 346*, 2021.

[14] M. Hairul Umam, V. Wahanggara, T. A. Cahyanto, and L. A. Muharom, "Analisis Perbandingan Algoritma C4. 5 Dan Algoritma Naïve Bayes Untuk Prediksi Kelulusan Mahasiswa (Studi Kasus?: Prodi Teknik Informatika Universitas Muhammadiyah Jember)," *Jurusan Teknik Informatika Fakultas Teknik Universitas Muhammadiyah Jember*, vol. 1310651100, pp. 1–9, 2017.

[15] F. Fatmawati and M. Affandes, "Klasifikasi Keluhan Menggunakan Metode Support Vector Machine (SVM) Pada Akun Facebook Group iRaise Helpdesk," *Jurnal CoreIT: Jurnal Hasil Penelitian Ilmu Komputer dan Teknologi Informasi*, vol. 3, no. 1, p. 24, Jan. 2018, doi: 10.24014/coreit.v3i1.3552.