Performance Comparison of Decision Tree J48, CART and Naïve Bayes Algorithms for Predicting Chronic Kidney Disease

[1]Ali Nur Ikhsan, [2]Alif Nur Fadilah, [3]Alifah Dafa Iftinani
[1,2]Departement of Informatics, Universitas Amikom Purwokerto
[3]Departement of Information System, Universitas Amikom Purwokerto
Email: [1]alinurikhsan@amikompurwokerto.ac.id, [2]alifnurf11@gmail.com, [3]alifahdafa1924@gmail.com

Article Info
Article history:
Received Aug 25th, 2023
Revised Nov 30th, 2023
Accepted Dec 18th, 2023

Keyword:
Chronic Kidney Disease
Decision Tree J48
CART
Naïve Bayes

ABSTRACT
Chronic Kidney Disease could be a worldwide issue that proceeds to extend with high treatment costs. Accurate diagnosis is essential for managing this disease. There is a requirement for a technique to anticipate chronic kidney disease, with prevalent use being made of Decision Tree J48, Naive Bayes, and CART algorithms which offer benefits like swift computation, ease of use, and high precision. The researchers aimed to determine the comparison results of Decision Tree J48, CART, and Naive Bayes algorithms for predicting chronic kidney disease. From the research findings, it was concluded that the CART algorithm had the highest accuracy rate of 97.25% in predicting chronic kidney disease, compared to the J48 Decision Tree algorithm and the Naïve Bayes algorithm with accuracy rates of 96.5% and 93.5% respectively. The CART algorithm can be utilized by pathologists to develop a program for predicting chronic kidney disease.

Corresponding Author:
Ali Nur Ikhsan,
Departement of Informatika, Universitas Amikom Purwokerto,
Jl. Letjend Pol. Soemarto No.127, Watumas, Purwanegara, Kec. Purwokerto Utara, Kabupaten Banyumas, Jawa Tengah 53127
Email: alinurikhsan@amikompurwokerto.ac.id

DOI: http://dx.doi.org/10.24014/ijaidm.v7i1.26472

1. INTRODUCTION

The kidney is a crucial organ responsible for eliminating toxins, balancing fluids and other substances, managing blood pressure, and producing hormones necessary for maintaining good health [1]. CKD is a type of kidney ailment characterized by structural or functional irregularities in the kidneys, which are irreversible. This disorder results in the body's inability to regulate metabolism, fluid, and electrolyte balance, consequently leading to uremia [2]. Riskesdas data from 2018 shows that the prevalence of chronic kidney disease in the Indonesian population is 0.38%, an increase from 2013 data of 0.2%. The highest incidence rate is in North Kalimantan province, accounting for 0.64% [3]. Agreeing to the 2010 Worldwide Burden of Infection comes about, Chronic Kidney Disease was the 27th driving cause of passing within the world in 1990, and this rose to 18th by 2010. In Indonesia, CKD is the moment most noteworthy restorative condition that requires treatment and gets the most elevated subsidizing assignment from BPJS healthcare, after heart malady [4]. With the high number of chronic kidney disease patients and the need for accurate diagnosis, research is necessary to find methods that provide the highest accuracy values.

Outside of statistics, the term "prediction" can often refer to an informed opinion or guess. These predictions may be based on a person's experiences, as well as various forms of reasoning such as deductive, inductive, and abductive. If a person has expertise in a particular field, their predictions may be particularly useful. The Delphi Method is a technique that allows for the generation of controlled expert-based predictions. Although the data used in this method is not statistical in nature, it can still be considered consistent with statistical approaches, as it involves the cognitive experiences of expert predictors and forms an intuitive "probability curve" [5]. Predicting may be a efficient handle of estimating what is most likely to happen within the future based on past and show data, pointing to play down blunders (contrasts between what is anticipated and what really happens) [6].

The J48 algorithm is an improvement of the classical ID3 algorithm which is broadly recognized for decision tree acceptance. This advanced algorithm enables the classification of data using decision trees that offer benefits such as the capability to process both continuous and discrete numerical data, accommodate absent attribute values, generate rules that are straightforward to understand, and it is the quickest algorithm among those that leverage main memory in a computer [7]. The J48 algorithm, present in the WEKA application, is a version of C4.5 algorithms. This algorithm produces a choice tree that can be utilized for classifying information and evaluating the course of given information [8]. The J48 algorithm is well-suited for predicting large amounts of data in a relatively short amount of time [9]. In the research conducted using the J48 Algorithm, it was determined whether the accuracy level was high and the required time was fast.

The Classification and Regression Tree (CART) algorithm was developed around 1980 by Leo Breiman, Jerome H. Friedman, Richard A.Olshen, and Charles J. Stone. CART is a non-parametric factual method created for the subject of categorical analysis, for both categorical and persistent reaction factors. This method is typically applied to large amounts of data with a large number of variables and mixed variable scales using a binary sorting procedure [10]. Within the classification calculation CART, a record will be classified into one of the available classifications for the target variable based on the marker variable values[11]. The CART algorithm belongs to the category of machine learning algorithms that utilize trees. Although it was initially developed in a simple form, numerous individuals have contributed to its development over time, leading to its excellent performance as an algorithm [12]. By using the CART algorithm, parameters that result in a high percentage of accuracy can be obtained [13]. The study utilized the CART algorithm to establish the necessary high precision levels and rapid prediction rates for anticipating chronic kidney disease.

The Naive Bayes Classifier algorithm is one of the calculations found in classification procedures. Naive Bayes may be a classification strategy utilizing likelihood and insights presented by the English researcher Thomas Bayes. It predicts the probability of future occasions based on past encounter, thus it is known as the Bayes' Theorem [14]. Naive Bayes is an calculation utilized for classification, based on the Bayes' theorem and expect that the values between factors are free of each other on a particular yield esteem [15]. The preferences of the Naive Bayes algorithm are that it requires moderately little sums of information to assess the parameters and is simple to actualize. In numerous cases, it moreover gives great comes about [16]. The Naive Bayes classification algorithm is suitable for handling large amounts of data and can moreover handle lost values and shifting traits within the data. However, a limitation of the Naive Bayes classification method is that it cannot accurately measure the accuracy of a classification, as it is based on probabilities [17].

In the previous research conducted by Chotimah, the Naive Bayes algorithm resulted in an accuracy of 80% in diagnosing chronic kidney disease (CKD) [18] . Within the research conducted by Pratama et al., the determination of chronic kidney disease came about within the application of highlight determination using BPSO being able to move forward the execution of C4.5 classification with an accuracy value of 96%[19]. Within the research conducted by Safuan, the accuracy of 96.25% was gotten in recognizing chronic kidney disease utilizing the ID3 algorithm [20].

From the research done, no one has applied the J48, CART and Naïve Bayes decision tree algorithms to determine the accuracy of these three algorithms in predicting chronic kidney disease. Decision Tree algorithms J48, Naïve Bayes and CART are widely used algorithms and have advantages such as fast, simple calculation and high accuracy. In this research, researchers used the CART, J48, Naïve Bayes Decision Tree algorithms to predict chronic kidney disease.

After considering the problem's context, the researcher phrased the inquiry as "How do the Decision Tree J48, CART, and Naïve Bayes Algorithms Compare in Predicting Chronic Kidney Disease's Performance?" The research goal was to evaluate how well each algorithm predicted Chronic Kidney Disease, comparing Decision Tree J48, CART, and Naïve Bayes's performance.

2. RESEARCH METHOD

In this study, the researcher used a quantitative research method. The research process flow can be seen in Figure 1.

Based on Figure 1, the research process flow is as follows:

1. Problem Identity
   Amid the arrange of distinguishing the issue, the group of analysts analyzed the Chronic Kidney Disease (CKD) information obtained from the UCI Machine Learning Repository. The team leader was responsible for the direct observation of data belonging to the recipients of scholarships during this stage. Simultaneously, the other members of the research team were responsible for collecting relevant literature concerning the research.

2. Data Collection
During the stage of collecting data, the researchers accumulated information from the UCI Machine Learning Repository.

3. Pre-Processing Stage
In this stage, the researchers selected the data that would be used for the study. The data was identified and grouped according to its attributes (attribute identification and selection). Then, the missing or incomplete attribute values were handled, and the data was discretized.

4. Use of Classification Methods
The method used in this research employed Decision Tree J48, CART, and Naïve Bayes algorithms. In this stage, the accuracy values were calculated through a confusion matrix.

5. Validation and Evaluation
In this stage, it was carried out with the aim of verifying that the results obtained from the research are accurate and valid, and to proceed to draw conclusions.

6. Conclusions
The conclusion is drawn by comparing the comes about of the CART, Decision Tree J48, and Naïve Bayes algorithms utilized, which have been performed with the yield within the frame of accuracy values of the examined calculations.

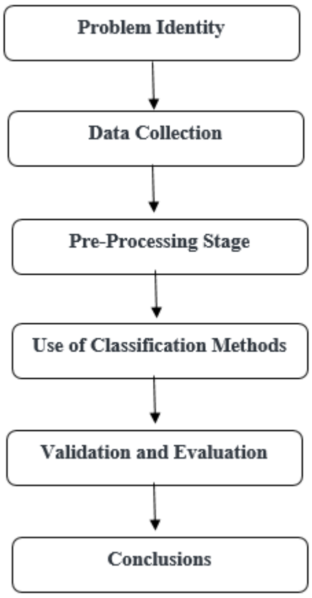

Figure 1. Research Flow

3. RESULTS AND ANALYSIS

3.1. Problem Identity

In the conducted research, a literature study has been carried out to find algorithms suitable for the research, namely the Decision Tree J48, CART and Naïve Bayes algorithm.

3.2. Data Collection

The information utilized for this study was procured from the website of the UCI Machine Learning Repository. The data set contains 400 examples that include 24 characteristics and a single label, all of which are visible in Table 1.

Table 1. Dataset Attribute

No.	Abbreviation	Attribute
1	Age	Age
2	Bp	Blood Pressure
3	Sg	Specific Gravity

No.	Abbreviation	Attribute
4	Al	Albumin
5	Su	Sugar
6	Rbc	Red Blood Cells
7	Pc	Pus Cell
8	Pcc	Pus Cell Clumps
9	Ba	Bacteria
10	Bgr	Blood Glucose Random
11	Bu	Blood Urea
12	Sc	Serum Creatinine
13	Sod	Sodium
14	Pot	Potassium
15	Hemo	Hemoglobin
16	Pcv	Packed Cell Volume
17	Wc	White Blood Cell Count
18	Rc	Red Blood Cell Count
19	Htn	Hypertension
20	Dm	Diabetes Mellitus
21	Cad	Coronary Artery Disease
22	Appet	Appetite
23	Pe	Pedal Edema
24	Ane	Anemia

3.3. Pre-Processing Stage

At this stage, the researcher conducted identification, adjustment of attributes, and selection of the dataset obtained to ensure that the data is ready to be used and processed in the Weka application.

3.4. Use of Classification Methods

The representation of the pre-processed data set can be observed in Figure 2. The subsequent step post pre-processing is the application of the Weka software for processing the data set. The aim of this step is to deduce the outcomes of the confusion matrix by utilizing the 10- fold cross-validation evaluation approach. The data set is portioned into 25 parts, consisting of 24 sections as training sets and one part as a testing set. This process is carried out ten times. The classifiers employed for this data set comprise the decision tree J48 algorithm, CART, and Naive Bayes.

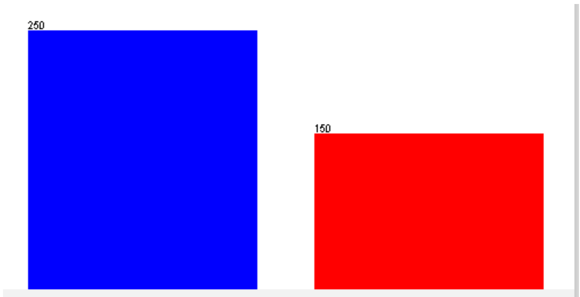

Figure 2. Data visualization after pre-processing

In Figure 2, there are 250 Chronic Kidney Disease (ckd) data with blue color graph and 150 non-Chronic Kidney Disease (notckd) data with red color graph.

In J48 Decision Tree algorithm, the obtained precision value is 96.5%, from the precision, recall and F-measure values. Time required to test the dataset using weka application 3.8.6 is 0.03 seconds. Table 2 shows details of the accuracy of the J48 Decision Tree algorithm test results.

Table 2. Accuracy Details of J48 Decision Tree Algorithm test results

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.980	0.060	0.965	0.980	0.972	0.925	0.962	0.960	ckd
Weighted Avg.	0.940	0.020	0.966	0.940	0.953	0.925	0.962	0.956	notckd
	0.965	0.045	0.965	0.965	0.965	0.925	0.962	0.958	

In CART algorithm, the obtained precision value is 97.25%, from the precision, recall and F-measure values. Time required to test the dataset using weka 3.8.6 application is 0.16 seconds. Table 3 shows details of the accuracy of the CART algorithm test results.

Table 3. Accuracy Details of CART Algorithm test results

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.984	0.047	0.972	0.984	0.978	0.941	0.964	0.963	ckd
Weighted Avg.	0.953	0.016	0.973	0.953	0.963	0.941	0.964	0.950	notckd
	0.973	0.035	0.973	0.973	0.972	0.941	0.964	0.958	

In the Naïve Bayes algorithm, the precision value obtained is 93.5%, from the precision, recall and F-measure values. Time required to test the dataset using weka 3.8.6 application is 0.01 seconds. Table 4 shows details of the accuracy of the Naïve Bayes algorithm test results.

Table 4. Accuracy Details of Naïve Bayes Algorithm test results

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.916	0.033	0.979	0.916	0.946	0.867	0.991	0.994	ckd
Weighted Avg.	0.967	0.084	0.873	0.967	0.918	0.867	0.991	0.989	notckd
	0.935	0.052	0.939	0.935	0.936	0.867	0.991	0.992	

Based on the testing results, the CART algorithm obtained the highest accuracy of 97.25% compared to the J48 Decision Tree algorithm with an accuracy value of 96.5% and Naïve Bayes algorithm with an accuracy value of 93.5%.

3.5. Validation dan Evaluation

To measure the accuracy level of the J48 Decision Tree, CART, and Naïve Bayes algorithms, confusion matrices obtained from the data processing process in the Weka application were used.

Table 5. Comparison of Confusion Matrix of J48 Decision Tree, CART and Naïve Bayes Algorithm

Algoritme Decision Tree J48			Algoritme CART		Algoritme Naïve Bayes	
	a	b	a	b	a	b
a	245	5	246	4	229	21
b	9	141	7	143	5	145

In Table 5, it can be seen that "a" refers to Chronic Kidney Disease (CKD) data, while "b" refers to non-Chronic Kidney Disease (notCKD) data. The J48 Decision Tree algorithm generated rules that predicted CKD in 245 cases that were consistent with the testing data. It also had five cases where the generated rules predicted notCKD but the testing data showed CKD, and an additional 141 cases where both the generated rules and testing data predicted notCKD.

Within the CART algorithm, there were 246 instances where the generated rules predicted CKD in the testing data. Additionally, there were four cases where the rules predicted notCKD, but the testing data showed CKD, and another 143 cases where both the generated rules and testing data predicted notCKD.

As part of the Naive Bayes algorithm, 229 instances where the generated rules predicted CKD matched with the testing data. Additionally, 21 cases emerged where the generated rules predicted notCKD, but the testing data indicated CKD. Furthermore, there were 145 instances where the generated rules predicted notCKD, and the testing data also predicted notCKD.

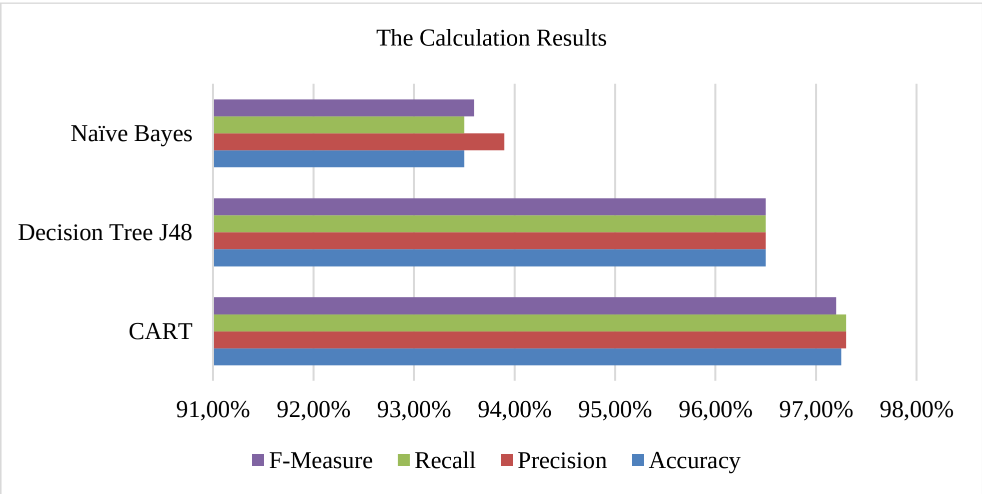

Figure 3. Algorithm Calculation Results

3.6. Conclusion

The outcome of the CART algorithm proved to be the most accurate with an accuracy performance of 97.25%; precision, recall, and F-measure values were measured at 0.973, 0.973, and 0.972, respectively. In comparison, both the J48 Decision Tree algorithm and the Naive Bayes algorithm failed to produce the same level of performance. The former exhibited an accuracy score of 96.5% for precision, recall, and F-measure values of 0.965 each, while the latter displayed an accuracy rating of 93.5%, with precision, recall, and F-measure values of 0.939, 0.935, and 0.936, respectively. The data can be seen in Figure 3.

From Figure 3, it can be observed that the calculations show the CART algorithm with the highest accuracy compared to the J48 Decision Tree and Naïve Bayes algorithms in predicting chronic kidney disease.

4. CONCLUSION

The research revealed that the algorithm used to predict chronic kidney disease was the CART algorithm, boasting the highest accuracy of 97.25%, surpassing the Decision Tree J48 algorithm with 96.5% accuracy and the Naïve Bayes algorithm with 93.5% accuracy. Pathology experts can employ the CART algorithm to develop a program for predicting chronic kidney disease. For future research, it is suggested to add optimization with PSO, Genetic Algorithm, and others to algorithms that still have low accuracy values to determine whether they can surpass the CART algorithm with the highest accuracy value in this research.

ACKNOWLEDGEMENT

Unlimited gratitude goes to Amikom Purwokerto University for organizing the Penelitian Dosen Muda Amikom (PDMA), LPPM Amikom Purwokerto University, and related parties so that this research runs well and smoothly. Hopefully, the results of this study can be useful for the development of science and technology, especially in the health sector. Thank you also to colleagues who have helped in the research process.

REFERENCES

[1] A. Aditya, A. Udiyono, L. Dian Saraswati, and H. Setyawan, "SCREENING FUNGSI GINJAL SEBAGAI PERBAIKAN OUTCOME PENGOBATAN PADA PENDERITA DIABETES MELLITUS TIPE II (Studi di Wilayah Kerja Puskesmas Ngesrep)," *J. Kesehat. Masy.*, vol. 6, no. 1, pp. 191–199, 2018, [Online]. Available: http://ejournal3.undip.ac.id/index.php/jkm

[2] A. A. Eka Cahyani, D. Prasetya, M. F. Abadi, and D. Prihatiningsih, "Gambaran Diagnosis Pasien Pra-Hemodialisa Di Rsud WangayaTahun 2020-2021," *J. Ilm. Hosp.*, vol. 11, no. 1, pp. 32–40, 2022.

[3] Kemenkes RI, "Laporan Riskesdas 2018 Kementrian Kesehatan Republik Indonesia," *Laporan Nasional Riskesdas 2018*, vol. 53, no. 9. pp. 154–165, 2018. [Online]. Available: http://www.yankes.kemkes.go.id/assets/downloads/PMK No. 57 Tahun 2013 tentang PTRM.pdf

[4] A. Thofiq Madani, H. Sunandar, and S. Adelina Hutabara, "Bulletin of Data Science Diagnosis Dan Prediksi Penyakit Ginjal Kronis Dengan Menggunakan Pendekatan Stacked-Generalization," *Media Online*, vol. 2, no. 1, pp. 35–43, 2022, [Online]. Available: https://ejurnal.seminar-id.com/index.php/bulletinds

[5] B. Baskoro, S. Sriyanto, and L. S. Rini, "Prediksi Penerima Beasiswa dengan Menggunakan Teknik Data Mining di Universitas Muhammadiyah Pringsewu," *Pros. Semin. Nas. …*, pp. 87–94, 2021, [Online]. Available: https://jurnal.darmajaya.ac.id/index.php/PSND/article/view/2918

[6] M. Kafil, "Penerapan Metode K-Nearest Neighbors Untuk Prediksi Penjualan Berbasis Web Pada Boutiq Dealove Bondowoso," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 3, no. 2, pp. 59–66, 2019, doi: 10.36040/jati.v3i2.860.

[7] N. S. Pakpahan, "Implementasi Data Mining Menggunakan Algoritma J48 Dalam Menentukan Pola Itemset Belanja Pembeli (Study Kasus: Swalayan Brastagi Medan)," *J. Comput. Informatics Res.*, vol. 1, no. 1, pp. 7–13, 2021, [Online]. Available: https://journal.fkpt.org/index.php/comforch/article/view/111%0Ahttps://journal.fkpt.org/index.php/comforch/article/download/111/83

[8] S. Agustiani, A. Mustopa, A. Saryoko, W. Gata, and S. K. Wildah, "Penerapan Algoritma J48 Untuk Deteksi Penyakit Tiroid," *Paradig. - J. Komput. dan Inform.*, vol. 22, no. 2, pp. 153–160, 2020, doi: 10.31294/p.v22i2.8174.

[9] M. Defriani and I. Jaelani, "Algoritma J48 Dan Logistic Model Tree Untuk Memprediksi Predikat Kelulusan Mahasiswa: Studi Kasus STT XYZ," *INTECOMS J. Inf. Technol. Comput. Sci.*, vol. 3, no. 2, pp. 129–140, 2020, doi: 10.31539/intecoms.v3i2.1478.

[10] I. W. Misshuari and Chairunisah, "Penerapan Metode Classification and Regression Trees (Cart) Untuk Menentukkan Faktor-Faktor Yang Mempengaruhi Pembayaran Kredit Oleh Nasabah Di Pt Bprs Gebu Prima Medan," *Karismatika*, vol. 6, no. 3, pp. 10–20, 2020.

[11] S. Susanto and D. Suryadi, *Pengantar Data Mining – Menggali Pengetahuan dari Bongkahan Data*. ANDI, 2010.

[12] H. M. Zhafran, I. T. Bandung, J. G. Bandung, and A. Graf, "Aplikasi Algoritma CART dalam Klasifikasi Jamur Berdasarkan Kelayakan Makan," 2023.

[13] M. F. Rizalno, A. Johar, and F. F. Coastera, "Analisis Prediksi Masa Studi Mahasiswa Menggunakan Metode Decision Tree Dengan Penerapan Algoritme Cart (Classification and Regression Trees) (Studi Kasus Data Alumni Fakultas Teknik Universitas Bengkulu)," *Rekursif J. Inform.*, vol. 10, no. 1, pp. 96–106, 2022, doi: 10.33369/rekursif.v10i1.21362.

[14] M. Idris, “Implementasi Data Mining Dengan Algoritma Naive Bayes Untuk Memprediksi Angka Kelahiran,” *J. Pelita Inform.*, vol. 7, no. 3, pp. 421–428, 2019, [Online]. Available: https://ejurnal.stmik-budidarma.ac.id/index.php/pelita/article/view/1154

[15] J. Han, M. Kamber, and J. Pei, *Data Mining Concepts and Techniques 3rd Edition*. San Fransisco: Morgan Kauffman, 2012.

[16] F. Solikhah, M. Febianah, A. L. Kamil, W. A. Arifin, and Shelly Janu Setyaning Tyas, “Analisis Perbandingan Algoritma Naive Bayes Dan C.45 Dalam Klasifikasi Data Mining Untuk Memprediksi Kelulusan,” *Tematik*, vol. 8, no. 1, pp. 96–103, 2021, doi: 10.38204/tematik.v8i1.576.

[17] T. Arifin and D. Ariesta, “Prediksi Penyakit Ginjal Kronis Menggunakan Algoritma Naive Bayes Classifier Berbasis Particle Swarm Optimization,” *J. Tekno Insentif*, vol. 13, no. 1, pp. 26–30, 2019, doi: 10.36787/jti.v13i1.97.

[18] S. N. Chotimah, “KLASIFIKASI DIAGNOSIS PENYAKIT GINJAL KRONIS DENGAN Organization (WHO) yang dihimpun tahun Repository (Almustafa , 2021). Penelitian lain Component Analysis (Islam dkk ., 2023) . pernah dilakukan dengan menerapkan konsep,” vol. 4, no. 1, pp. 8–15, 2023.

[19] I. G. A. Mahardika Pratama, L. G. Astuti, I. M. Widiartha, I. G. N. A. Cahyadi Putra, C. R. Adi Pramartha, and I. D. M. B. Atmaja Darmawan, “Diagnosis Penyakit Ginjal Kronis dengan Algoritma C4.5, K-Means dan BPSO,” *JELIKU (Jurnal Elektron. Ilmu Komput. Udayana)*, vol. 10, no. 4, p. 371, 2022, doi: 10.24843/jlk.2022.v10.i04.p07.

[20] Safuan Safuan, “Deteksi Penyakit Gagal Ginjal Kronis Menggunakan Algoritma ID3,” *Elkom J. Elektron. dan Komput.*, vol. 13, no. 1, pp. 8–17, 2020, doi: 10.51903/elkom.v13i1.136.

BIBLIOGRAPHY OF AUTHORS

Ali Nur Ikhsan is a lecturer in the Informatics Department at Amikom Purwokerto University. He is a graduate of Amikom Purwokerto University for his undergraduate degree and Amikom Yogyakarta University for his postgraduate degree. His fields of expertise include Data Mining, Machine Learning, and Artificial Intelligence. He can be contacted via email at alinurikhsan@amikompurwokerto.ac.id

Alif Nur Fadilah is a third-semester undergraduate student in the Informatics Department at Amikom Purwokerto University. He can be contacted via email at alifnurf11@gmail.com

Alifah Dafa Iftinani is a third-semester undergraduate student in the Information System Department at Amikom Purwokerto University. She can be contacted via email at alifahdafa1924@gmail.com