VOL. 6. NO. 1 AUGUST 2020
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480 /jitk.v6i1.1403

JITK (JURNAL ILMU PENGETAHUAN DAN TEKNOLOGI KOMPUTER)

IDENTIFICATION OF CHRONIC KIDNEY DISEASE USING NAIVE BAYES, ADABOOST, AND RANDOM FOREST LEARNING METHODS

Raras Tyasnurita[1]; Shafira Widya Hapsari[2]

Information Systems Study Program[1, 2]
Institut Teknologi Sepuluh Nopember
www.its.ac.id
[1]raras@is.its.ac.id ; [2]shafira16@mhs.is.its.ac.id

Abstract— *Chronic kidney disease is a decrease in kidneys function where the condition leads to kidney damage. This disease causes damage to the body's immunity, because the body fails to maintain fluid balance. Therefore, it becomes a critical need to identify whether a patient is a sufferer of chronic kidney disease or not. The classification methods used in this study are Naive Bayes, AdaBoost, and Random Forest. The main objective of this research is to help the classification process of a patient whether classified as a patient with chronic kidney disease or not. Recently, proper early recognition is needed to detect chronic kidney disease to prevent delays in its treatment. Given the large number of chronic kidney disease cases that occur, this study is expected to be an effort to control the increase in sufferers. The results showed that the Naive Bayes approach achieved 95.4% accuracy, which increased to 98.6% after AdaBoost was implemented, and Random Forest led at 99.3%.*

Keywords*: Chronic Kidney Disease, Machine Learning, Classification, Naive Bayes, AdaBoost, Random Forest.*

Abstrak— Penyakit ginjal kronis merupakan penurunan fungsi pada ginjal dimana kondisinya mengarah pada kerusakan ginjal. Penyakit ini menyebabkan kerusakan pada kekebalan tubuh karena tubuh gagal untuk mempertahankan keseimbangan cairan. Metode klasifikasi yang digunakan dalam penelitian ini adalah *Naive Bayes, AdaBoost,* dan *Random Forest*. Penelitian ini bertujuan untuk membantu proses klasifikasi pasien apakah tergolong sebagai penderita penyakit ginjal kronis atau tidak. Saat ini dibutuhkan pengenalan dini yang tepat untuk mendeteksi penyakit ginjal kronis agar dapat mencegah keterlambatan dalam penanganannya. Mengingat besarnya kasus penyakit ginjal kronis yang terjadi, penelitian ini diharapkan dapat menjadi upaya pengendalian angka peningkatan penderita. Hasil penelitian menunjukkan bahwa pendekatan *Naive Bayes* meraih akurasi 95.4%, dimana meningkat menjadi 98.6% setelah *AdaBoost* diterapkan, dan *Random Forest* memimpin di angka 99.3%.

Kata Kunci*: Penyakit Ginjal Kronis, Pembelajaran Mesin, Klasifikasi, Naive Bayes, AdaBoost, Random Forest*

INTRODUCTION

A kidney is an organ that has a role in processing the body's metabolism [1]. The kidneys are also involved in regulating water balance in the human body. Humans need to maintain their health so that they can function correctly.

Chronic Kidney Disease (CKD) is a disorder of the kidney characterized by abnormalities in the structure or function of the organ, which lasts more than three months [2]. Several diseases also trigger and cause CKD, such as diabetes, high blood pressure, or gout.

CKD has become one of the population's health problems worldwide, where the number has increased with the rapid population growth rate [3]. About 1 in 10 people from the world's population are sufferers at a particular stage. CKD can be divided into 5 degrees, namely degrees I, II, III, IV, and V.

According to a 2006 report from the Indonesian Diatrans Kidney Foundation, there were approximately 150 thousand sufferers of chronic kidney disease. About 21% are in the age range of 15 to 34 years, 49% are aged 35-55 years, and 30% are aged over 56 years [4].

Treatment of CKD today has also made progress, one of which uses dialysis techniques or kidney transplants. Dialysis is a therapy that temporarily replaces kidney function, aiming to excrete metabolic waste [4].

With the rapid development of Artificial Intelligence, it is encouraging its use for various fields. This research develops Machine Learning methods as part of Artificial Intelligence. Machine Learning is defined as one of the disciplines that

study how a machine or computer has intelligence by learning from examples (data) [5]. The implementation of this learning has been successfully applied in various fields, such as banking, trade, and health.

Therefore, by utilizing technology, namely machine learning and based on available data, an algorithm is developed to classify whether someone is a sufferer of chronic kidney disease. The learning algorithm for the data implemented in this study is Naive Bayes, AdaBoost, and Random Forest.

There have been several previous studies where researchers tried various methods to classify chronic kidney disease. Fadilla et al. (2018) use the Extreme Machine Learning (EML) method, where this method increases the speed of learning on Artificial Neural Networks (ANN). EML produces an accuracy of 96.7% [3]. Kriplani et al. (2019), implementing Deep ANN with an accuracy of 97% [6]. The same results were obtained by Saha et al. (2019) [7].

Yunus (2018) applies the K-Nearest Neighbor (k-NN) approach. The accuracy value obtained by k-NN is 78.8% [8]. Particle Swarm Optimization (PSO) is added in the method to get better results. The merger of K-NN and PSO obtained 97.3% accuracy results [8]. In contrast to Arifin and Ariesta (2019), who added PSO to the Naive Bayes method. The concept of weighting this attribute achieves 98.8% accuracy [9]. In a recent study, Ilham (2020) added the Bootstrap strategy to k-NN. Bootstrap and k-NN have an accuracy rate of 97.9% [10]. A comparison of several classification algorithms from 2012 to 2017 can be seen in the tabulations presented in a recent scientific article by Dharmarajan [11].

This study aims to assist health agencies in classifying patients, whether organized as patients with chronic kidney disease or not by using all three methods. This research is expected to analyze the application of these three methods to identify sufferers of chronic kidney disease. With the right prediction, sufferers will get proper treatment, and hopefully, in the future, it will minimize the number of sufferers. This research is also expected to be a consideration for health agencies in classifying patients.

METHODOLOGY

This research aims to utilize the existing attributes in the current dataset to be then processed as a form of business to improve decision-making capabilities. In this study, several stages are conducted, namely, data collection, data cleaning, proposed methods implementation, and data testing. The steps of the research methodology are shown in Figure 1.

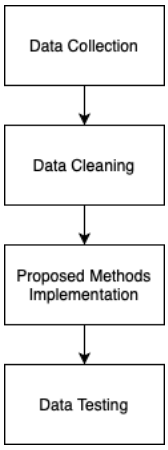

Figure 1. Research Methodology Stages

The first step taken in this research is to collect data that will be managed. In this study, a dataset regarding Chronic Kidney Disease was obtained from the *University of California Irvine* (*UCI) Machine Learning repository* (https://archive.ics.uci.edu/ml/datasets/chronic_kidney_disease) [12].

The dataset has 400 instances, 25 attributes, and two classes, chronic and non-kidney disease. Attributes involved in the study are numerical and nominal, which discuss the health of the patient, starting from age, blood pressure, and health history, as in Table 1.

Table 1. Data Attributes

No	Attributes	Value
1.	Age	Numerical
2.	Blood Pressure	Nominal
3.	Specific Gravity	Nominal
4.	Albumin	Nominal
5.	Sugar	Nominal
6.	Red Blood Cell	Nominal
7.	Pus Cell	Nominal
8.	Pus Cell Clumps	Nominal
9.	Bacteria	Nominal
10.	Blood Glucose Random	Numerical
11	Blood Urea	Numerical
12	Serum Creatinine	Numerical
13	Sodium	Numerical
14	Potassium	Numerical
15	Hemoglobin	Numerical
16	Packed Cell Volume	Numerical
17	White Blood Cell	Numerical
18	Red Blood Cell Count	Numerical
19	Hypertension	Nominal
20	Diabetes Mellitus	Nominal
21	Coronary Artery Disease	Nominal
22	Appetite	Nominal

Accredited Rank 3 (Sinta 3) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti
No. 28/E/KPT/2019, September 26, 2019. Published by PPPM STMIK Nusa Mandiri

No	Attributes	Value
23	Pedal Edema	Nominal
24	Anemia	Nominal
25	Class (Output)	Nominal

The data cleaning stage is intended to clean up the dataset to be a consistent data collection so that it can be processed at a later stage to produce valid data [13]. Replace is performed on missing data or missing value by replacing it with the mean value for numeric attributes and the mode for nominal attributes [14]. Mean means the average amount of information, while the mode value in data is a value that often appears.

In general, in carrying out classifications, several methods can be applied. This study chose three classification methods, namely Naive Bayes, AdaBoost, and Random Forest. Naive Bayes is the most straightforward algorithm [15]. This algorithm is based on Bayes' Theorem, which predicts based on probability, with a strong assumption of independence. Therefore, Naive Bayes is a model that has independent features.

The Bayes formulas is shown in Equation 1 [16].

$$P(H|E) = \frac{P(E|H) \times P(H)}{P(E)} \quad \text{..... (1)}$$

Where,

P(H|E): The conditional probability that Hypothesis H occurs if E occurs

P(E|H): Probability of E which will affect Hypothesis H

P(H): The initial probability of hypothesis H occurs in general

P(E): The initial probability E without calculating Hypothesis H

Naive Bayes works based on probabilities that make it possible to consider all features for consideration. It causes Naive Bayes to have several characteristics such as:

1. Easy to understand
2. Calculations made are relatively faster
3. Able to overcome the value of the attribute that is wrong or missing during models development and predictions
4. Able to overcome irrelevant attributes

In addition to the advantages possessed above, some deficiencies can affect the results of classification by this method, namely the assumption of independence between attributes that might reduce the value of accuracy. Therefore in overcoming these shortcomings, AdaBoost was involved. AdaBoost was added to Naive Bayes as an adaptive boosting algorithm to improve classification performance [17].

Furthermore, besides Naive Bayes, a trial was also conducted using the Random Forest method. Random Forest is a decision tree technique that combines several trees to form a forest [18]. Consequently, the prediction is based on the class that most often appears or the most votes of several classification trees. This technique can be used for classification and regression problems.

This technique has the advantage of being able to overcome if there are missing values and can still be implemented on extensive data. Random Forests also tend to be less affected by outlier data values [19]. However, the drawback of this technique is that it takes a proper tuning model in the data to produce a reasonable classification.

The final step is the testing phase. Evaluating the dataset is done by considering the percentage accuracy of the method used in classifying data [20]. Outputs are data models, confusion matrix, and class predictions from input data.

RESULTS AND DISCUSSION

The test is carried out to measure the accuracy of each model's results by calculating the value of accuracy, precision, recall, and F1-score. The aim is to determine the difference in performance between the Naive Bayes method and AdaBoost for handling cases of chronic kidney disease prediction. Accuracy is used to predict the level of closeness between the predicted value and the actual value. Precision is used for true positive evidence on real data. The recall is used for the proportion of positive cases that are correctly predicted. Whereas F1-score as the average value of the Precision and Recall values.

The Naive Bayes classification with the Cross-validation folds 10 test is applied. The results obtained are from a total of 400 patients, 250 patients are predicted to suffer from CKD, but in fact, only 229 patients suffer from CKD, while the remaining 21 patients do not suffer from CKD. The data that has been accurately predicted as the Not CKD class is 149. There is one data foreseen as the Not CKD class but entered into the CKD class. Therefore the accuracy value of the Confusion Matrix produced by Naive Bayes is 95.4%. The results of the Confusion Matrix can be seen in Table 2.

Table 2. Confusion Matrix for Naive Bayes

	True CKD	*True NOT CKD*	*Class Precision*
Pred CKD	229	21	99.6%
Pred NOT CKD	1	149	87.6%
Class Recall	91.6%	99.3%	

The AdaBoost algorithm is used as the second method. Experiments carried out using ten repetitions. This repetition aims to see the behavior of the algorithm used in the same conditions.

The results of the Confusion Matrix can be seen in Table 3. The accuracy value of the addition of AdaBoost to Naive Bayes is 98.6%. The increase achieved was 3.2%. The addition of the AdaBoost algorithm gives quite good results in increasing the value of Accuracy, Precision, Recall, and F1-score.

Table 3. Confusion Matrix for *AdaBoost*

	True CKD	*True NOT CKD*	*Class Precision*
Pred CKD	244	1	99.6%
Pred NOT CKD	6	149	96.1%
Class Recall	97.6%	99.3%	

Some changes occurred during the experiment after the data were pre-processed. The dataset without any data cleaning process is shown by the NB (Naive Bayes) and NBA (Naive Bayes + Adaboost) notations, while the dataset that is performed for data cleaning has the NBdc (Naive Bayes with data cleaning) and NBAdc (Naive bayes + Adaboost with data cleaning) notations. The increase occurred in the value of accuracy in the data cleaning process, where initially NBdc = 94.5% became NBAdc = 98.25%. Here there is an increase of 3.75%.

However, for the same method, data cleaning does not affect or even reduce the value of accuracy. It can be seen from the NB when compared to the NBdc where, after cleaning the data, the accuracy value decreased by 0.9%. The same thing happened with the NBA vs. NBAdc, where the accuracy value reduces by 0.35%. It means that the Naive Bayes method can demonstrate its ability to overcome missing values. The results of the effect of data cleaning on the missing values in the data are shown in Table 4.

Table 4. Effect of *Data Cleaning* on *Missing Values*

	Accuracy (%)	*Precision*	*Recall*	*F1*
NB	95.40	100.0	92.0	95.8
NBdc	94.50	99.6	91.6	95.4
NBA	98.60	100.0	96.8	98.4
NBAdc	98.25	99.6	97.6	98.6

As a comparison, predictions are made using Random Forest with the results can be seen in Table 5. The accuracy of classification with Random Forest reaches the highest value compared to the two approaches previously described, which is 99.3%.

Table 5. Confusion Matrix for *Random Forest*

	True CKD	*True NOT CKD*	*Class Precision*
Pred CKD	249	1	99.2%
Pred NOT CKD	2	148	99.3%
Class Recall	99.6%	98.7%	

A comparison of the accuracy from the three classifiers applied in this study with the most recent previous studies is shown in Table 6. It can be seen that various methods have the potential to be good classifiers because the average accuracy value is above 90%, where the range of accuracy values is 70-99%. Random Forest with 99.3% accuracy is the first method recommended as a result of this study. Another alternative method is to use a simple classification method such as Naive Bayes but accompanied by modification or addition of strategies (AdaBoost or PSO) to improve performance.

Table 6. Comparison of Classifiers Performance

	References	**Accuracy**
Naive Bayes		95.4
Naive Bayes + AdaBoost		98.6
Random Forest		99.3
Extreme Machine Learning	Fadilla et al. (2018)[3]	96.7
k-NN	Yunus (2018) [8]	78.8
k-NN+PSO		97.3
Deep ANN	Kriplani et al. (2019) [6] Saha et al. (2019) [7]	97
Naive Bayes + PSO	Arifin and Ariesta (2019) [9]	98.8
k-NN + Bootstrap	Ilham (2020) [10]	97.9

Accredited Rank 3 (Sinta 3) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No. 28/E/KPT/2019, September 26, 2019. Published by PPPM STMIK Nusa Mandiri

By using the same dataset, different results are obtained. It can be caused by the various mechanisms of the methods tested. Based on the accuracy value, it gets more than 95% for the proposed methods in this study. Therefore, researcher can start with Naive Bayes, AdaBoost, and Random Forest learning for classification. Moreover, it is necessary to take into account the future impacts of both the positive and negative effects of the resulting classification.

CONCLUSION

From the results of this study, it can be concluded that the Naive Bayes Classifier can classify chronic kidney disease data, despite there are still some missing values. The selection of the best method needs to be adjusted to the objectives to be achieved from the research. In this study, the principal amount emphasized is the value of accuracy. With the Naive Bayes method, the results obtained an accuracy value of 95.4%. If the dataset is pre-processed for the missing value, the Naive Bayes method's accuracy will decrease by 0.9% to 94.5%. Besides, the Boosting (AdaBoost) algorithm method is proven to improve the performance of the Naive Bayes method, where the accuracy value is 98.6%. Random Forest achieved the highest accuracy at 99.3%. Random Forest and Naive Bayes with AdaBoost are classification methods with excellent performance compared to previous studies.

REFERENCES

[1] K. Daenen, A. Andries, D. Mekahli, A. Van Schepdael, F. Jouret, and B. Bammens, "Oxidative stress in chronic kidney disease," *Pediatr. Nephrol.*, vol. 34, no. 6, pp. 975–991, 2019.

[2] H. Widiani, "Penyakit ginjal kronik stadium V akibat nefrolitiasis," *Intisari Sains Medis*, vol. 11, no. 1, pp. 160–164, 2020.

[3] I. Fadilla, P. P. Adikara, and R. S. Perdana, "Klasifikasi Penyakit Chronic Kidney Disease (CKD) Dengan Menggunakan Metode Extreme Learning Machine (ELM)," *J. Pengemb. Teknol. Inf. dan Ilmu Komput. e-ISSN*, vol. 2548, p. 964X, 2018.

[4] A. W. Kusumo, *Perbedaan Penyebab Gagal Ginjal Antara Usia Tua Dan Muda Pada Penderita Penyakit Ginjal Kronik Stadium V Yang Menjalani Hemodialisis Di Rsud Dr. Moewardi*. Surakarta: Universitas Muhammadiyah Surakarta, 2010.

[5] R. C. Deo, "Machine learning in medicine," *Circulation*, vol. 132, no. 20, pp. 1920–1930, 2015,.

[6] H. Kriplani, B. Patel, and S. Roy, "Prediction of chronic kidney diseases using deep artificial neural network technique," in *Computer Aided Intervention and Diagnostics in Clinical and Medical Images*, Springer, 2019, pp. 179–187.

[7] A. Saha, A. Saha, and T. Mittra, "Performance measurements of machine learning approaches for prediction and diagnosis of chronic kidney disease (CKD)," *ACM Int. Conf. Proceeding Ser.*, pp. 200–204, 2019.

[8] W. Yunus, "Algoritma K-Nearest Neighbor Berbasis Particle Swarm Optimization Untuk Prediksi Penyakit Ginjal Kronik," *J. Cosphi*, vol. 2, no. 2, 2018.

[9] 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.

[10] A. Ilham, "Hybrid Metode Boostrap Dan Teknik Imputasi Pada Metode C4-5 Untuk Prediksi Penyakit Ginjal Kronis," *J. Stat. Univ. Muhammadiyah Semarang*, vol. 8, no. 1, 2020.

[11] K. Dharmarajan, "Prediction of Chronic Kidney Disease using Classification techniques," *Parishodh Journal*, Page No : 1420," no. March, 2020.

[12] D. Dua and C. Graff, "UCI machine learning repository, 2017," *URL http//archive. ics. uci. edu/ml*, vol. 37, 2019.

[13] A. Fatima, N. Nazir, and M. G. Khan, "Data Cleaning In Data Warehouse: A Survey of Data Pre-processing Techniques and Tools," *Int. J. Inf. Technol. Comput. Sci.*, vol. 9, no. 3, pp. 50–61, 2017.

[14] B. K. Khotimah, M. Miswanto, and H. Suprajitno, "Optimization of feature selection using genetic algorithm in naïve Bayes classification for incomplete data," *Int. J. Intell. Eng. Syst.*, vol. 13, no. 1, pp. 334–343, 2020.

[15] S. Zeynu and S. Patil, "Prediction of Chronic Kidney Disease Using Data Mining Feature Selection and Ensemble Method," *Journal of Data Mining in Genomics*, vol. 9, no. 1, pp. 1–9, 2018.

[16] B. Tiemens, R. Wagenvoorde, and C. Witteman, "Why every clinician should know Bayes' rule," *Heal. Prof. Educ.*, no. xxxx, 2020.

[17] Arif-Ul-Islam and S. H. Ripon, "Rule Induction and Prediction of Chronic Kidney Disease Using Boosting Classifiers, Ant-Miner and J48 Decision Tree," *2nd Int. Conf.*

Electr. Comput. Commun. Eng. ECCE 2019, no. February, 2019.

[18] D. S. Sisodia and A. Verma, "Prediction performance of individual and ensemble learners for chronic kidney disease," *Proc. Int. Conf. Inven. Comput. Informatics, ICICI 2017*, no. Icici, pp. 1027–1031, 2018.

[19] J. Ali, R. Khan, N. Ahmad, and I. Maqsood, "Random Forests and Decision Trees," *Int. J. Comput. Sci. Issues*, vol. 9, no. 5, pp. 272–278, 2012.

[20] J. Huang and C. X. Ling, "Using AUC and accuracy in evaluating learning algorithms," *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 3, pp. 299–310, 2005.

Accredited Rank 3 (Sinta 3) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti
No. 28/E/KPT/2019, September 26, 2019. Published by PPPM STMIK Nusa Mandiri