Application of the pessimistic pruning to increase the accuracy of C4.5 algorithm in diagnosing chronic kidney disease

M A Muslim1*, A J Herowati1, E Sugiharti1 and B Prasetyo1
1Department of Computer Science, FMIPA, Universitas Negeri Semarang

*Corresponding author: a212muslim@yahoo.com

Abstract. A technique to dig valuable information buried or hidden in data collection which is so big to be found an interesting patterns that was previously unknown is called data mining. Data mining has been applied in the healthcare industry. One technique used data mining is classification. The decision tree included in the classification of data mining and algorithm developed by decision tree is C4.5 algorithm. A classifier is designed using applying pessimistic pruning in C4.5 algorithm in diagnosing chronic kidney disease. Pessimistic pruning use to identify and remove branches that are not needed, this is done to avoid overfitting the decision tree generated by the C4.5 algorithm. In this paper, the result obtained using these classifiers are presented and discussed. Using pessimistic pruning shows increase accuracy of C4.5 algorithm of 1.5% from 95% to 96.5% in diagnosing of chronic kidney disease.

1. Introduction
All health organizations around the world store health data in electronic format. Health data can contain all the information about the patient and the parties involved in the healthcare industry. The storage of such type of data is increased at a very rapidly [1]. Health data becomes complex because the data are large. By using traditional methods becomes difficult to extract meaningful information from the data. Science that can process large amounts of data into knowledge is called data mining. Data mining is the technique of extracting valuable information buried or hidden in a collection of data (database) is so great to be able to find interesting patterns that were previously unknown. Data mining has been applied in many areas because of its ability quickly analyze large amounts of data. Several methods are often mentioned in the literature of data mining include classification, clustering, association rules mining, neural networks, genetic algorithms, and more [2] Classification is a data mining techniques that can be used to predict group membership to a data instance [3]. Classification is the process of finding a model or a function that describes or distinguish the concept or class of data, with the aim to estimate the class of an object [4].

One of the tasks in classification is to whether a particular disease is present. Common data mining methods used for classification are k-nearest neighbor (KNN), decision trees and neural network [5]. Decision tree is a very powerful and well-known classification and prediction method [6]. The decision tree is regarded as one of the most popular approaches to represent classifiers. The decision tree is widely used by many researchers in the field of health [1]. The decision tree provides the advantages of visualization suggestion (in the form of a decision tree) makes predictions can be
observed [7]. The tree consists of rules for taking decision at every node [8]. In the research that has been done [9], his research has been tested on intelligent systems Intelligent Heart Disease Prediction System (IHDDS) by using three algorithms, namely the decision tree algorithm, naive bayes and artificial neural network, results from these research decision tree algorithm shows the highest accuracy levels. There are several algorithms used in decision tree including ID3, CART, CHAID, and C4.5 Algorithm [10]. C4.5 is a famous algorithm used for generating the decision tree. In late 1970 and early 1980, J. Ross Quinlan develop a decision tree algorithm known as ID3 (Iterative dichotomiser) then presented C4.5 as a substitute ID3, which is a measure of supervised learning algorithm newer [4]. Decision trees generated by C4.5 algorithms can be used for classification [11]. Additionally, the C4.5 performed consistently better in terms of accuracy [12].

Classification of data mining algorithms can be utilized in the field of health in diagnosing a disease, one of which is chronic kidney disease. Chronic Kidney Disease (CKD) is a pathophysiological process with various causes (etymology) are varied and result in a progressive decline in renal function which generally end up with kidney failure. Laboratory test is required in the process of early detection Chronic Kidney Disease. Creatinine serum levels, urea plasma levels and value of glomerulus filtration rates became a strong indicator expresses a patient diagnosed with the Chronic Kidney Disease or not [13].

In this research using dataset chronic kidney disease obtained from the UCI repository of machine learning. The decision tree is built based on the object data contained in the training dataset. The number of branches is strongly influenced by anomalies may lie in the training dataset, because the 'noise' or some outliers that pass the initial screening process the data. The decision tree is built with the purpose applied in area situations with different datasets, so it is necessary to avoid the phenomenon of 'overfitting' or the size of the decision tree becomes greater in the dataset used to train [7]. In this case, use a pruning method known as tree pruning process. Pessimistic pruning is one of the pruning process in which the tree pruning done to identify and remove branches that are not needed. Pessimistic pruning based on the estimation error derived from the training data, therefore it does not require a separate trimming. procedures Pessimistic pruning was top-down. If the internal node pruned, then all of his descendants are removed from the process of pruning, resulting in a relatively fast pruning [12].

2. Methods

This research was developed from several previous studies that reference is linked to the method. The use of this reference is intended to provide limits to the methods that will be developed further [9], doing research with a test on intelligent systems Intelligent Heart Disease Prediction System (IHDDS) by using three algorithms, namely the decision tree algorithm, naive bayes, and ANN. Results from these studies showed predictive decision tree algorithm with the highest level of accuracy [14] conducted a study carried out by applying the method of decision tree algorithm ID3, C4.5, C5.0 J48 on benefits and its use to classify and predict the disease. Here the decision tree classification techniques chosen for its simplicity and accuracy [6]. This research uses classification C4.5 decision tree algorithm. Implementation of data mining techniques C4.5 decision tree algorithm can generate predictive information such as hypertension in pregnancy and can produce good accuracy [15] to conduct their research by comparing the classification decision tree using ID3 and C4.5 algorithms. This research aimed to find a comparison based on the accuracy of the two algorithms using datasets weather. Results from these research show the accuracy of the algorithm C4.5 higher than in ID3 [16] on the performance of induction tree algorithm on a set of medical data tumor in terms of accuracy and time complexity is the ID3 algorithm, C4.5 and CART. This research resulted in that third among the algorithm C4.5 is the best in improving the primary tumor dataset and to improve colon tumor dataset that is with both algorithms ID3 and C4.5 show the same classification accuracy [17] on research using the classification decision tree (C4.5) which is based on the gain ratio and reduce error pruning. Accuracy is more accurate than the results of the decision tree without pruning [18] in a study using the Decision Tree (C4.5) for the diagnosis of heart disease. The decision tree is one of the successful
data mining is used in the diagnosis of heart disease, but the fit is not perfect. Therefore, in this study using a decision tree based on the gain ratio, binary discretization, voting and reduce error pruning for the diagnosis of heart disease. Results higher accuracy than simply using decision tree [19]. This study discusses various methods of pruning that is different. Also the evaluation of the effectiveness of pruning both the complexity and accuracy of the classification. Classification algorithms used in this research is the algorithm C4.5 used for the classification of credit card database either pruning or without pruning, pessimistic pruning one using this method is much faster than the reduced error pruning and also provides higher accuracies. Pessimistic pruning resulting in a relatively fast pruning [12].

In general, the scheme of this research method is illustrated in Figure 1 and the classification process uses an C4.5 algorithm classifier and pessimistic pruning method shown in Figure 2.

![Diagram of research methods](image)

**Figure 1.** Diagram of research methods

![Working process C4.5 classifier algorithm and pessimistic pruning method](image)

**Figure 2.** Working process C4.5 classifier algorithm and pessimistic pruning method

### 2.1.1. Data Preprocessing

This research uses the method of application *pessimistic pruning* the C4.5 algorithm to diagnose chronic kidney disease using chronic kidney disease dataset obtained from the UCI repository of machine learning. The number of record data in the dataset are 400 records, consisting of 24 attributes and one class attribute. These attributes including age, blood pressure, specific gravity, albumin, sugar, red blood cells, pus cells, pus cell clumps, bacteria, blood glucose random, blood urea, serum creatinine, sodium, potassium, hemoglobin, packed cell volume, white blood cell count, red blood cell count, hypertension, diabetes mellitus, coronary artery disease, appetite, pedal edema and anemia. In the class attribute have two values, ckd and notckd. ckd class indicates the patient is suffering from chronic kidney disease by 250 the number of records, while notckd class indicates patients is not suffering from chronic kidney disease by 150 the number of records.
attribute categories

(i) specific gravity
(ii)Albumin
(iii) Sugar
(iv) Red blood cells
(v) pus cell
(vi) Pus cell clumps
(vii) bacteria
(viii) hypertension
(ix) Diabetes mellitus
(x) Coronary artery disease
(xi) appetite
(xii) pedal edema
(xiii) Anemia

numeric attributes

(i) Age
(ii) Blood pressure
(iii) Blood glucose
(iv) Blood urea
(v) serum creatinine
(vi) Sodium
(vii) potassium
(viii) Hemoglobin
(ix) packed cell
(x) White blood cell
(xi) Red blood cell

On the dataset to do the data cleaning, Data cleaning is the process of removing noise and inconsistent the data or the data irrelevant. In this research, the data cleaning is done by ignoring the missing values (missing value). For some data that one of its attributes is not filled will be directly substituted for the average value for the attributes that are numeric (continuous), while relatively categorical attributes, missing values are replaced with the mode of these attributes. In this data, there are several attributes that consisting of numerical value. Therefore, before using this dataset, numeric type attributes necessary data transformation process, at this stage of the discretization using Entropy-Based Discretization to obtain entropy values of numeric type attributes in order to obtain its information gain.

2.1.2. Data Mining Process
Classification algorithm applied to the training data by applying the algorithm C4.5 with pessimistic pruning. Generate rule or application of the model is applied to testing data, then get the classified dataset which is shown in Figure 2 above. Tests using 8-fold cross validation, then this test is used in the algorithm C4.5 prior to implementation pessimistic pruning and the application of pruning.
2.1.3. Algorithm C4.5
C4.5 is an evolution of ID3, which was delivered by [20]. C4.5 uses the gain ratio as a criterion for separation (split). Before we headed toward the extraction of data into a form the model tree, there are several processes that must be considered in the formation of the tree structure, namely:

1. Select the root by the largest ratio gain
2. Select the internal root / root branch ratio based on the greatest gain after removing the attribute that has been selected as the root
3. Repeat until all the attributes calculated gain value ratio.

The exact parameters used to measure the effectiveness of an attribute in the data sample classification techniques, one of which is by using a gain ratio. Before looking for the value of the gain, first seek out the likelihood of a record in the attribute (entropy).

(1) Calculation of Entropy Value
To get the gain value ratio, we first have to know the other parameters that affect the value of the gain, which is very necessary parameters to get the value of the gain. These parameters are entropy. Mathematically the value of entropy can be calculated using Eq. (1).

\[ \text{Entropy}(S) = -p_+ \log_2 p_+ - p_- \log_2 p_- \]  

(1)

(2) The calculation of information gain
After getting the value of entropy, then the next step is the calculation of information gain. Based on mathematical calculations information gain of an attribute A can calculated using Eq. (2).

\[ \text{Gain}(S, A) = \text{Entropy}(S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} \cdot \text{Entropy}(S) \]  

(2)

(3) gain ratio
To calculate the gain ratio, we need to know a new term called split information. Split information is calculated using Eq. (3).

\[ \text{Split information}(A) = -\sum_{i=1}^{n} \frac{|S_i|}{|S|} \cdot \log_2 \left( \frac{|S_i|}{|S|} \right) \]  

(3)

Furthermore, the gain ratio is calculated by using Eq. (4).

\[ \text{Gain Ratio} (A) = \frac{\text{Gain}(S, A)}{\text{split information}(A)} \]  

(4)

2.1.4. Pruning
Pessimistic pruning uses the pessimistic statistical correlation test instead [20]. The basic idea is that the error ratio that was estimated using the training set is not reliable enough. Instead, a more realistic measure, known as the continuity correction for the binomial distribution, should be used Eq. (5).

\[ \varepsilon'(T, S) = \varepsilon(T, S) + \frac{|\text{leaves}(T)|}{2 \cdot |S|} \]  

(5)

However, this correction still produces an optimistic error rate. Consequently, [20] suggests an internal node pruning if its error rate is within one standard error from a reference tree items, namely with Eq. (6).

\[ \varepsilon'(\text{pruned}(T, t), S) \leq \varepsilon'(T, S) + \sqrt{\frac{\varepsilon'(T, S) \cdot (1 - \varepsilon'(T, S))}{|S|}} \]  

(6)

The last condition is based on the statistical confidence interval for proportions. Usually, the last condition is used such that T refers to a subtree who is the internal root node t and S denotes the portion of the training set that Refers to the node t. The pessimistic pruning procedure performs a top-down traversal over the internal nodes. If an internal node is pruned, then all its Descendants are removed from the pruning process, the resulting in a relatively fast pruning[12].
2.1.5. Evaluation
The evaluation was done by analyzing the results of the classification of the use methods the application pessimistic pruning of the C4.5 algorithm in diagnosing chronic kidney disease, Which determine the value of the resulting accuracy. Calculation accuracy use confusion matrix which is shown in Table 1.

| Classifications | ckd | Notckd |
|-----------------|-----|--------|
| ckd             | TP  | FN     |
| notckd          | FP  | TN     |

Information:
(1) TP (True Positive), Observed Class suffering from chronic kidney disease class predicted result of chronic kidney disease.
(2) TN (True Negative), Observed Class does not suffer from chronic kidney disease predicted result class does not suffer from chronic kidney disease.
(3) FP (False Positive), Observed Class does not suffer from chronic kidney disease class predicted result of chronic kidney disease.
(4) FN (False Negative), Observed Class suffering from chronic kidney disease predicted result class does not suffer from chronic kidney disease.

\[
\text{Accuracy rate (\%) } = \frac{TP + TN}{TP + TN + FP + FN} \times 100\%
\]  

3. Result and discussion
The data used in this research was obtained from the UCI repository of machine learning using chronic kidney disease dataset, In these data there were 400 data records.

3.1.1. Testing and Results C4.5 Classification Algorithm
The results of the evaluation of the data using the classifier tree algorithm C4.5 produce 19 leaves and 28 trees. Evaluation C4.5 algorithm using 8-Fold Cross Validation is presented in Table 2.

| Measurement Specification | Value |
|---------------------------|-------|
| Correctly Classified Instances | 380 or 95% |
| Classified incorrectly Instances | 20 or 5% |
| kappa statistic | 0.8945 |
| Mean absolute error | 0.0507 |
| Root mean squared error | 0.2082 |
| Relative absolute error | 10.8107% |
| Root relative squared error | 43.0034% |
| Total Number of Instances | 400 |

Correctly classified instances is how many rows of the data are classifiable Correctly. Accuracy Obtained is 95% with a record that are classified as much as 380. Correctly incorrectly classified instances is how many rows of the data are classifiable not true. The number of records that are incorrectly classified as much as 20 or 5%. Kappa statistics were generated 0.8945. Mean absolute error is generated 0.0507. With the results of root mean squared error 0.2082. Relative absolute error is generated 10.8107%. Root relative squared error is generated 43,0034%. Using C4.5 algorithm, generates detailed accuracy by class are presented in Table 3.
Table 3. Detailed Accuracy by Class C4.5 Algorithm

| Class   | TP Rate | FP Rate | Precision | Recall | F-measure | ROC Area |
|---------|---------|---------|-----------|--------|-----------|----------|
| ckd     | 0.944   | 0.04    | 0.975     | 0.944  | 0.959     | 0.969    |
| not Ckd | 0.96    | 0.056   | 0.911     | 0.96   | 0.935     | 0.969    |
| Weight Avg. | 0.95 | 0.046   | 0.951     | 0.95   | 0.95      | 0.969    |

Explanation of Table 3 that the positive value is 0.944 True, False positive 1 negative 0.056 True, False negative and 0.04. True Positive Rate is 0.95, while the False Positive Rate is 0.002. Precision value is 0.951 to 0.95 recall and F-measure 0.95.

By using the C4.5 algorithm, generating Confusion Matrix as shown in Table 4.

Table 4. Confusion Matrix of C4.5 Algorithm

| Classified as | a | b |
|---------------|---|---|
| a = ckd       | 236 | 14 |
| b = not ckd   | 6   | 144 |

Known from 400 the data 250 the data is classified as a class ckd (with 236 the data is Correctly classified as ckd and 144 class ckd defined as class notckd), 150 the data is classified as a class notckd (with 14 the data is Correctly classified as a notckd class and 6 Data notckd defined as class ckd).

3.1.2. Testing and Results by applying Pessimistic Pruning In C4.5 Classification Algorithm

The results of the evaluation of the data by applying the pessimistic pruning in C4.5 Classification Algorithm produce 11 leaves and 16 trees. Evaluation using 8-Fold Cross Validation is presented in Table 5.

Table 5. Evaluation by applying pessimistic pruning in C4.5 Classification Algorithm

| Measurement Specification | Value |
|---------------------------|-------|
| Correctly Classified Instances | 386 or 96.5% |
| Classified incorrectly Instances | 14 or 3.5% |
| kappa statistic | 0.9257 |
| Mean absolute error | 0.049 |
| Root mean squared error | 0.1851 |
| Relative absolute error | 10.4437% |
| Root relative squared error | 38.225% |
| Total Number of Instances | 400 |

Correctly classified instances is how many rows of the data are classifiable Correctly. Obtained accuracy is 96.5% with a record that are classified Correctly as much as 386. incorrectly classified instances is how many rows of the data are classifiable not true. The number of records that are incorrectly classified as much as 14 or 3.5%. Kappa statistics were generated 0.9257. Mean absolute error is generated 0.049. With the results of root mean squared error 0.1851. Relative absolute error is generated 10.4437%. Root relative squared error is generated 38.225%. Generates detailed accuracy by class are presented in Table 6.
Table 6. Detailed Accuracy by Class

| Class      | TP Rate | FP Rate | Precision | recall | F-measure | ROC Area | Class   |
|------------|---------|---------|-----------|--------|-----------|----------|---------|
| ckd        | 0.964   | 0.033   | 0.98      | 0.964  | 0.972     | 0.954    | ckd     |
| not Ckd    | 0.967   | 0.036   | 0.942     | 0.967  | 0.954     | 0.954    | not Ckd |
| Weight Avg.| 0.965   | 0.034   | 0.965     | 0.965  | 0.965     | 0.954    |         |

Explanation of Table 6 that the positive value ie 0.964 True. True Positive Rate is 0.965, while the False Positive Rate is 0.034. Precision value is 0.965 to 0.965 recall and F-measure 0.965. By applying pessimistic pruning in C4.5 Classification Algorithm, generating Confusion Matrix as shown in Table 7.

Table 7. Confusion Matrix of C4.5 Algorithm

| a | b | classified as |
|---|---|---------------|
| 241 | 9  | a = ckd      |
| 5  | 145 | b = not ckd  |

Known from 400 the data 250 the data is classified as a class ckd (with 241 the data is Correctly classified as ckd and 9 class ckd defined as class not ckd), 150 the data is classified as a class not ckd (with 145 the data is Correctly classified as a not ckd class and 5 Data not ckd defined as class ckd).

From the results obtained, the accuracy of C4.5 algorithm results that can be seen in Table 8 and Figure 3 is a graph increase the resulting accuracy of the resulting accuracy comparison between C4.5 algorithm and C4.5 algorithms with applying pessimistic pruning.

Table 8. Results of Accuracy

| algorithm                        | Accuracy (%) | Error (%) |
|----------------------------------|--------------|-----------|
| C4.5 algorithms                  | 95%          | 5%        |
| C4.5 algorithms with Pessimistic| 96.5%        | 3.5%      |

Figure 3. Comparison of Results Accuracy
From Table 8 shows after testing the method will be generated that apply pessimistic pruning in C4.5 algorithm with higher accuracy by 96.5% compared with the usual C4.5 algorithm has an accuracy of 95%. C4.5 algorithm method with pessimistic pruning avoid overfitting and increased accuracy of 1.5%. It said that by applying the pessimistic pruning can increase the accuracy of the C4.5 algorithm in diagnosing chronic kidney disease.

4. Conclusion
In this research, researchers used a chronic kidney disease dataset obtained from the UCI repository of machine learning. From the experimental results by applying pessimistic pruning in C4.5 algorithm shows increase accuracy of 1.5% from 95% to 96.5%. It can be concluded that the application of the pessimistic pruning can increase the accuracy of C4.5 algorithm in diagnosing chronic kidney disease.

References
[1] Ahmad P, Saqib Q and Syed Q A R 2015 Int. J. Comput. Appl. 120 38
[2] Sugiharti E, Safit F and Feroza R D 2017 J. Theor. Appl. Inf. Technol. 95 902
[3] Wijaya K P and Muslim M A 2016 Proc. 3rd Natl. Semin. Comput. Sci. Semarang Semarang State Univ.
[4] Han J, Micheline K and Jian P 2011 Data Mining Concepts and Techniques 3rd Edition (Waltham: Elsevier Inc)
[5] Larose T D 2005 Discovering Knowledge In Data An Introduction To Data Mining (New Jersey: John Wiley & Sons, Inc)
[6] Muzakir A and Rica A W 2016 Sci. J. Inform. 3 21
[7] Gorunescu F 2011 Data mining Concept, Model and Technique (Berlin: Springer)
[8] Patil M P and Satish R K 2014 Int. J. Comput. Sci. Appl. 11 38
[9] Soni J, Ujma A, Dipesh S and Sunita S 2011 Int. J. Comput. Appl. 17 43
[10] Bramer M 2007 Principles of Data mining (London: Springer)
[11] Adhatrao K, Aditya G, Amiraj D, Rohit J and Vipul H 2013 J. Data Min. Knowl. Manag. Process 3 39
[12] Rokach L and Oded M 2015 Data mining with Decision Trees Theory and Applications 2nd Edition (Singapore: World Scientific)
[13] Muslim M A, Iin K and Endang S 2015 J. Theor. Appl. Inf. Technol. 78 70
[14] Thenmozhi K and P Deepika 2014 Int. J. Eng. Res. Gen. Sci 2(6) 6
[15] Hssina B, Abdelkarim M, Hanane E and Mohammed E 2014 Int. J. Adv. Comput. Sci. Appl. IJACSA Special Issue Adv. Veh. Ad Hoc Netw. Appl. 10 13
[16] Sujatha G and Rani K U 2013 Int. J. Emerg. Trends Technol. Comput. Sci. 2 418
[17] Pandey A K, Prabhat P, K L Jaiswal and Ashish K S 2013 J. Comput. Eng. IOSR-JCE 12 83
[18] Shouman M, Tim T and Rob S 2011 Proc. 9th Australas. Data Min. Conf. Aust. Ballarat
[19] Patil D D, Wadhai V M and Gokhale J A 2010 Int. J. Comput. Appl. 11 23
[20] Quinlan J R 1993 C45 Programs for Machine Learning (San Mateo: CA Morgan Kaufmann)