https://jurnal.seaninstitute.or.id/index.php/jukomi
Jurnal Komputer Indonesia (JU-KOMI)
Volume 2 No 02, 2024

JU-KOMI
E-ISSN: 2963-0460

Implementation of C4.5 Algorithm for Disease Classification: Literature Review

Sari Siregar
Program Diploma, Program Studi Manajemen Informatika, Akedemi Manajemen Informatika & Komputer, Universal

ARTICLE INFO

Keywords:
C4.5 Algorithm, Disease Classification

Email :
siregar09@gmail.com

ABSTRACT

Several studies have shown that the C4.5 algorithm is capable of achieving a high level of accuracy in diagnosing various diseases. For example, a study by Konstas et al. (2010) showed that the C4.5 algorithm was capable of achieving an accuracy level of 89% in diagnosing coronary heart disease. This study aims to discuss the application of the C4.5 Algorithm in disease classification as has been done by previous studies. This study is a literature review, which includes references related to the cases or problems identified. The author uses data from a literature study, a method used to collect data or sources related to the topic discussed in a study. The results of this Algorithm literature review can help doctors diagnose diseases more quickly, accurately, and efficiently. However, it is important to remember that the C4.5 algorithm is only a tool, and cannot replace the diagnosis of a professional doctor.

INTRODUCTION

The C4.5 algorithm, developed by J. Ross Quinlan, is one of the popular classification algorithms and is widely used in various fields, including disease classification. This algorithm has several advantages compared to other classification methods, such as the ability to process data quickly and efficiently, the ability to learn from new data, and the ability to be easily interpreted (Quinlan, 1993).

Several studies have shown that the C4.5 algorithm is capable of achieving a high level of accuracy in diagnosing various diseases. For example, a study by Konstas et al. (2010) showed that the C4.5 algorithm was capable of achieving an accuracy level of 89% in diagnosing coronary heart disease. Another study by Polat et al. (2014) showed that the C4.5 algorithm was capable of achieving an accuracy level of 95% in diagnosing breast cancer.

Despite its many advantages, the application of the C4.5 algorithm in disease classification also has several challenges. One of the main challenges is data quality. The C4.5 algorithm can only produce an accurate diagnosis if the data used is of good quality and complete. Incomplete or inaccurate data can cause the C4.5 algorithm to produce a wrong diagnosis.

Another challenge is the potential for bias in the data. If the training data is biased, then the C4.5 algorithm will also produce biased diagnoses. Therefore, it is important to ensure that the training data used is unbiased and represents the actual patient population. Therefore, this article aims to discuss the application of the C4.5 Algorithm in disease classification as has been done by previous studies.

METHOD

This research is a literature review, which includes references related to the case or problem identified. The author uses data from a literature study, a method used to collect data or sources related to the topic discussed in a study. The data obtained is then analyzed using descriptive analysis.

RESULTS

Implementation of C4.5 Algorithm for Mobile-Based Diagnosis of Diarrhea in Toddlers

1. Method
   Experimental, implementation of C4.5 algorithm. Software Hardware used in this application are all mobile devices such as Handphone, Smartphone, or Tablet that have operating system from Android version 2.2 Froyo to Android version 5.1.1 Lollipop
2. Background
   In this study, data analysis of diarrhea disease in toddlers will be conducted using data mining classification, namely the C4.5 algorithm using four parameters, namely gender, age (months), frequency of defecation and stool consistency. Therefore, in order to obtain faster and more flexible information values, this expert system will be applied in the form of an android-based mobile application describing the attributes, each branch describes the results of the attributes being tested [10]. This algorithm recursively visits each decision node, selects the optimal division, until it cannot be divided anymore. The concept used to select the optimal entropy is information gain or entropy reduction.
3. Results
   This expert system application provides various knowledge about diarrhea, including various types of diarrhea, common symptoms that appear when a child is infected with diarrhea, and is equipped with early actions/treatment, so that parents will be more responsive in handling diarrhea that attacks their child.

Application of C4.5 Algorithm for Hepatitis Disease Prediction

1. Method
   The method used to solve the problem is the C4.5 Algorithm by testing the performance of the method. The method testing was carried out using the confusion matrix method, testing with 10-Fold Cross Validation and ROC curves and using the RapidMiner tool. From the initial processing of the data above, 155 data were obtained with 123 data with the 'ALIVE' class and 32 data with the 'DEAD' class.
2. Background
   Along with the development of science and information technology, the presence of a new branch of science in the field of computer data mining has attracted much attention in the world of information systems. Literature on the discussion of hepatitis prediction has been done with several methods. Decision trees have been widely used to classify and predict in various fields. The C4.5 algorithm is one method in decision trees. Decision trees change very large facts into decision trees that represent rules (Suhartinah, 2010). For this reason, this study will be calculated using the C4.5 algorithm method for predicting hepatitis disease.
3. Results
   The data mining classification method of the C4.5 algorithm produces an accuracy of 77.29% and an AUC value of 0.846 which is included in Good Classification. Out of 155 data, 103 data were predicted accordingly, namely 103 'LIFE' data and 15 data predicted 'LIFE' but turned out to be 'DIE'. And as many as 20 data predicted 'DIE' turned out to be included in the 'LIFE' classification and as many as 17 data were predicted accordingly, namely 'DIE'. Thus, it can be concluded that this method is accurate in predicting hepatitis disease.

Classification of Factors Causing Diabetes Mellitus at UNHAS Hospital Using the C4.5* Algorithm

1. Method
   The analysis method used in this paper is the data mining technique with the C4.5 algorithm with R software. C4.5 algorithm with R software. The steps taken are data collection, attribute selection, and training data determination. are collecting data, selecting attributes, determining training data and test data with ten-fold cross validation, applying

the C4.5 algorithm with R software. and test data with ten-fold cross validation, applying the C4.5 algorithm, interpreting the results of the analysis. C4.5 algorithm, interpretation of results and validation of results. Data mining with C4.5 algorithm with R software. Test data with 10-fold cross validation. Attribute selection using Chi-Square Test (χ 2).
Cross-validation is a statistical method that is used to perform an evaluation and comparison of a dataset by dividing the data into two parts, namely training data and test data. data into two parts, namely training data and test data. One type of cross-validation
This validation is done by dividing the dataset into ten segments. dataset into ten segments d 1 - d 10 that are equal in size by randomising the data. randomising the data. Then d 1 will be used first for training process and validated using the rest of the data other than d 1. After that d 2 will be used for training, while the rest of the data other than d 2 is used for validation, and so on. validation, and so on. By doing validation like this, the accuracy will be will be higher (Refaeilzadeh et al., 2009). Interpretation of accuracy values can be classified into five different parts, namely 50%-60% (very weak accuracy), 60%-70 very weak), 60%-70% (weak accuracy level), 70%-80% (medium accuracy level), 80%-90% (high accuracy level), and 90%-100% (very high accuracy level).
(Gorunescu, 2011).

2. Background
   In the medical field, there are many records of DM sufferers. The large amount of data cannot be used if there is no information or conclusion from the data. Even a lot of data can actually be garbage and useless. Therefore, an extraction process is needed to find information in previously unknown data. One method that can be used for this extraction process is through machine learning. In relation to DM, the DM status of the sufferer is important to know before the DM sufferer experiences serious complications. The C4.5 algorithm has been popularly used to predict disease status. Therefore, in this paper, the C4.5 algorithm will be used as one of the data mining implementations to classify DM.
3. Results
   Based on the rules obtained with the C4.5 algorithm, there are four attributes that can be used to identify substantial factors that influence someone suffering from DM, namely GDP cholesterol, LDL, age and body weight. The accuracy value ranges from 50% to 100% with an average prediction accuracy of 98.5%. This means that the model obtained is very good with a very high level of accuracy.

Classification of Heart Disease Using C4.5 Algorithm

1. Method
   Data processing uses Google Colab tools and also uses the python programming language. Data processing in this study uses the Knowledge Discovery in Database (KDD) method. data processing using data mining methods or algorithms. This study uses the C4.5 algorithm for the process of analyzing heart disease symptoms. The results of the analysis of heart disease symptoms using the C4.5 algorithm will produce a decision tree that can predict heart disease based on the rules of the decision tree. The evaluation produced from the model created using the C4.5 algorithm in the form of a decision tree can produce accuracy, precision, and recall values from testing using the decision tree classifier and data division using k-fold cross validation.
2. Background
   Heart disease is one of the dangerous diseases. Heart disease can endanger the lives of sufferers if there is a delay in its treatment. This problem is caused by the difficulty of early detection in heart disease sufferers because sufferers always ignore the early symptoms that arise. In addition, the costs required for heart disease examination are not cheap because examinations by specialist doctors and laboratory tests are required. The prediction system is one option that can be used to perform early detection in heart disease sufferers at a cheaper cost in its use, this is because the costs used in specialist doctor

examinations and laboratory tests can be eliminated and replaced by a prediction system. This study aims to create a prediction system using the C4.5 algorithm where this algorithm makes predictions based on historical data of patients to be examined.

3. Results
   The results of the research that has been carried out using the C4.5 algorithm are applied with experiments using k-fold cross validation with the number k = 10, therefore this test has 10 test scenarios and 10 dataset division scenarios. Dataset division produces different accuracy, precision, recall values for each data division, in other words, data division affects the results of the C4.5 algorithm. The performance of the C4.5 algorithm is measured using accuracy, precision, recall values. Testing the 10 scenarios produced an average accuracy level of 0.7592, precision got an average score of 0.7933, recall got an average value of 0.7657.

Implementation of Adaboost-Based C4.5 Algorithm for Heart Disease Prediction

1. Method
   This research is an experimental research that involves investigating the treatment of parameters and variables that all depend on the researcher himself. software and hardware. Software Operating system: Windows XP SP III 32 bit Data mining: RapidMiner Version 5 Hardware CPU: Dual Core 1.7 Ghz, 2 GB Ram, 160Gb Hdd. Using the C4.5 algorithm and the Adaboost-based C4.5 algorithm. Testing: Cross Validation. Evaluation: Confusion matrix, AUC
2. Background
   Researchers choose the Decision Tree method in predicting heart disease. In this study, the application of the Decision Tree algorithm (C4.5) using the Adaboost method was carried out by optimizing attributes derived from trusted Datasets to predict heart disease with the aim of increasing accuracy. Based on the results of previous studies, the C4.5 algorithm in predicting heart disease has not yet reached the level of excellence, so the accuracy of the C4.5 model needs to be improved with the Adaboost method in solving heart prediction problems. The purpose of this study is to optimize the C4.5 algorithm based on Adaboost by performing iterations and attribute wighting to increase accuracy in predicting heart disease.
3. Results
   In this study, model testing was carried out using the C4.5 algorithm and the Adaboost-based C4.5 algorithm using patient data with or without heart disease. The resulting model was tested to obtain the accuracy and AUC values of each algorithm. The test using C4.5 obtained an accuracy value of 86.59% with an AUC value of 0.957. while the test using Adaboost-based C4.5 obtained an accuracy value of 92.24% with an AUC value of 0.982. In addition, the researcher also compared it with the Bagging-based C4.5 algorithm, obtaining an accuracy of 91.89% and an AUC value of 0.963. So it can be concluded that testing the heart disease model using the Adaboost-based C4.5 algorithm is better than C4.5 itself, with an increase in accuracy of 6.42% and an increase in AUC value of 0.26%. Thus, from the results of the model testing above, it can be concluded that Adaboost-based C4.5 provides a more accurate solution to heart disease problems.

Detection of Diabetes Mellitus Disease Using Decision Tree Algorithm C4.5 Architecture Model

1. Method
   Decision Tree C4.5 Decision Tree or decision tree is a model for predicting a tree structure or hierarchy to change data into decision trees and decision rules.
2. Background
   In diagnosing the disease can be done in other scientific fields, but with the development of an increasingly sophisticated, flexible and fast era, the diagnosis of diabetes can also be

done in the field of technology such as in the use of applications where one of them uses the Decision Tree C4.5 Algorithm [5]. The Decision tree C4.5 algorithm is used to predict classes or objects when the data class of new items is unknown, the decision tree C4.5 algorithm is a classification of the process of finding models or functions that explain and distinguish classes or data concepts [4]. Decision trees are similar to flow graphs, with each node standing for attribute values, each branch representing test results, and each leaf representing the distribution of classes or classes. A variation of the ID3 method known as Decision Tree C4.5 uses a greedy strategy with decision making based on trees created using a top-down recursive approach and a system for attacks

3. Results
   From the creation of a diabetes disease detection application with the application of the C4.5 classification algorithm can be used to help someone in making the first diagnosis before going to the doctor, it can be concluded that 1) The C4.5 algorithm can be used to facilitate decision making by projecting existing data into the form of a decision tree, based on the entropy and gain values owned by each data attribute. 2) In more accurate prediction results, large amounts of data are needed, meaning that the greater the amount of data used, the more accurate the prediction results produced. 3) The diabetes mellitus disease detection application can be predicted by utilizing the C4.5 algorithm with a percentage result having an accuracy level of 96%.

Classification of Diabetes Patients Using the C4.5 Decision Tree Algorithm

1. Method
   Using the experiments in the completion of this research. Broadly speaking, this research will perform calculation of the dataset by using the C4.5 classification algorithm, then evaluated using a confusion matrix and produces the accuracy of the Rapidminer 9.7 Application. The variables used in this study There are 17 variables with a total of 520 data This includes data about people including symptoms that can cause diabetes. diabetes. This dataset was created from a questionnaire questionnaires directly to people who have recently become diabetic, or who are still nondiabetic but had a few or more symptoms. Data were collected from patients by using questionnaires directly from the Sylhet Diabetes Hospital of Sylhet, Bangladesh. At this stage, 520 diabetes data were divided into two, namely training data and testing data with a percentage of 80% for training data and 20% for testing data. Training data acts as a pattern or model builder and the testing data as a model tester.
2. Background
   In order to analyze patients with Diabetes early, Recording of this disease is often done so that prevention can be carried out. One of the records that can be done is by utilizing classification techniques with data mining. Data mining is a method for acquiring knowledge. With data mining, implicit and valuable information from data can be extracted [5]. According to [6] data mining is a procedure for creating bonds that have meaning, patterns, and tendencies by observing large data groups in storage using pattern identification techniques. The methods that are usually operated on data mining include: description or depiction, prediction or forecast, clustering, classification and association, and estimation. In this study, classification techniques are used to predict which people are infected with diabetes and which are not infected. Several algorithms can be used to calculate the classification process. Classification algorithms include Decision Tree C4.5, Naive Bayes, and knearest neighbor (KNN). So this study will classify Diabetes Disease by utilizing the Decision Tree C4.5 algorithm.
3. Results
   Of the 16 attributes contained in the diabetes dataset, namely age, Alopecia, Gender, Polyuria, Polydipsia, sudden weight loss, weakness, olyphagia, Genital thrush, Irritability, delayed healing, partial paresis, Itching, visual blurring, muscle stiffness, and Obesity can

be used as data for the classification of diabetic patients. This study uses the C4.5 Algorithm to classify someone with diabetes or not. From 520 data divided into 416 as training data and 104 as testing data. The test results produced a fairly large accuracy of 97.12% Precision of 93.02%, and Recall of 100.00%. The ROC (Receiver Operating Characteristic) curve shows that the C4.5 algorithm has an AUC value of 0.994 which means Excellent Classification, this shows that using the C4.5 Algorithm for the classification of diabetic patients produces high accuracy.

Comparison of Decision Tree and Naive Bayes Algorithms in Diabetes Disease Classification

1. Method
   Implementing the Knowledge Discovery in Databases (KDD) method. It is a method used to obtain valuable information from a database. The research steps include Selection, Pre-processing, Transformation, Data Mining (an important process occurs in the application of the Decision Tree and Naïve Bayes algorithms, namely the separation of the dataset into two main components, namely the dataset for training and testing), and Interpretation / Evaluation (This process involves assessing the performance of both algorithms based on relevant evaluation metrics, such as accuracy, precision and recall).
2. Background
   Although many studies have been conducted using machine learning to classify diabetes, the average accuracy is still poor, which is around 82%. This creates a significant risk of error in disease identification. Therefore, this study focuses on a comparison between the Decision Tree and Naive Bayes methods in improving the accuracy of the diabetes classifier. This study aims to improve the accuracy of diabetes disease classification by comparing the Decision Tree and Naive Bayes algorithms. The purpose of the comparison is to determine the most appropriate algorithm in improving the accuracy of diabetes diagnosis. This study will evaluate the effectiveness of both algorithms, with the main focus on measuring and comparing the level of accuracy. The results are intended for endocrinologists, pharmacists, and nurses. Doctors can improve disease diagnosis and management by utilizing research findings. Pharmacists can provide drug-related support to patients based on research information. Nurses can use research results to support patient care and education about lifestyle management and diabetes treatment. In addition, it is expected that increasing the accuracy value of research results will have a positive impact on efforts to prevent and manage diabetes.
3. Results
   Based on the results of comparative research of Decision Tree and Naive Bayes methods in classifying diabetes, two classification algorithms were evaluated to understand their performance in classifying diabetes. From the results of the study, it can be seen that Naive Bayes managed to achieve a higher accuracy rate of 91.56%, while Decision Tree had an accuracy rate of 87.01%. Based on this evaluation, it can be concluded that Naive Bayes provides superior performance in classifying diabetes in the dataset used when compared to Decision Tree. The high accuracy of Naive Bayes shows the expertise of this model in distinguishing between positive and negative classes better than Decision Tree. The suggestions submitted based on the test results in this study are that further research is expected to explore other classification models to further improve its accuracy and further research is expected to optimize hyperparameters on the Naive Bayes algorithm to improve its accuracy.

Decision Tree C4.5 Algorithm Used To Classify Stroke Data

1. Method
   Data processing using the Decision Tree C4.5 algorithm. Data testing using rapidminer. Evaluation of the decision tree algorithm or Decision Tree C4.5 uses a confusion matrix to

calculate performance metrics such as accuracy, precision, and recall. A confusion matrix is a table used to describe the performance of a classification model on a test data set whose true values are known.

2. Background
   The advantage of classifying stroke data is that it can provide valuable information to develop more effective prevention and treatment strategies. By identifying important risk factors, we can take steps to reduce the risk of stroke, such as adopting a healthy lifestyle, controlling blood pressure and cholesterol, and managing other conditions. Medications can increase the risk of stroke and cerebrovascular disease. The aim of this study was to see the differences in the average age of stroke patients by sex in a specific population. Stroke is one of the vascular diseases that causes the most morbidity and mortality worldwide. A better understanding of the factors that influence stroke, including differences in patient sex, can help in developing better methods of stroke prevention, diagnosis, and treatment [3]. The results of this study confirm that classifying stroke data using statistical analysis and machine learning is an effective approach in identifying key risk factors and developing more effective stroke prevention and treatment strategies. Thus, this study makes a significant contribution to the global effort to reduce the burden of stroke disease, through improving the understanding of risk factors and the application of technology in early detection and intervention.
3. Results
   positive stroke prediction. In this context, it is recommended to conduct further studies by considering more stroke risk factors, such as family history and other health conditions. In addition, further studies can also be conducted to compare the effectiveness of Decision Tree C4.5 with other classification algorithms, such as Random Forest, Naive Bayes, or Support Vector Machine (SVM), in the context of stroke data. This will help in developing more accurate and efficient prediction models for stroke, which may ultimately contribute to better stroke prevention and treatment efforts.

CONCLUSION

The application of the C4.5 algorithm in disease classification offers many benefits for doctors and patients. This algorithm can help doctors diagnose diseases more quickly, accurately, and efficiently. However, it is important to remember that the C4.5 algorithm is only a tool, and cannot replace a professional doctor's diagnosis. Doctors should always consider the diagnosis results produced by the C4.5 algorithm with other clinical information before making a final diagnosis decision. This algorithm has been proven to be able to produce accurate classification models and help doctors diagnose diseases earlier and more accurately. Further research is still needed to improve the performance of this algorithm and develop its applications in the medical field.

REFERENCE

Faust, O., Acharya, U. R., Sudarshan, V. K., San Tan, R., Yeong, C. H., Molinari, F., & Ng, K. H. (2017). Computer aided diagnosis of coronary artery disease, myocardial infarction and carotid atherosclerosis using ultrasound images: a review. *Physica Medica*, *33*, 1-15. https://www.sciencedirect.com/science/article/abs/pii/S1120179716311012

Al-Salihy, N. K., & Ibrikci, T. (2017, February). Classifying breast cancer by using decision tree algorithms. In *Proceedings of the 6th international conference on software and computer applications* (pp. 144-148). https://dl.acm.org/doi/abs/10.1145/3056662.3056716

Minton, S. (Ed.). (2014). *Machine learning methods for planning*. Morgan Kaufmann.

Munggaran, A. P., & Hidayatulloh, T. (2015). Penerapan Algoritma C4. 5 Untuk Diagnosa Penyakit Diare Pada Anak Balita Berbasis Mobile. *Swabumi*, *2*(1), 47-58. https://ejournal.bsi.ac.id/ejurnal/index.php/swabumi/article/view/1960

Septiani, W. D. (2014). Penerapan Algoritma C4. 5 untuk prediksi penyakit Hepatitis. *Jurnal Techno Nusa Mandiri*, *11*(1), 69-78. https://ejournal.nusamandiri.ac.id/index.php/techno/article/view/172

Ente, D. R., Thamrin, S. A., Arifin, S., Kuswanto, H., & Andreza, A. (2020). Klasifikasi Faktor-Faktor Penyebab Penyakit Diabetes Melitus Di Rumah Sakit Unhas Menggunakan Algoritma C4. 5. *Indonesian Journal of Statistics and Its Applications*, *4*(1), 80-88. https://journal-stats.ipb.ac.id/index.php/ijsa/article/view/330

Sepharni, A., Hendrawan, I. E., & Rozikin, C. (2022). Klasifikasi Penyakit Jantung dengan Menggunakan Algoritma C4. 5. *STRING (Satuan Tulisan Riset dan Inovasi Teknologi)*, *7*(2), 117-126. https://journal.lppmunindra.ac.id/index.php/STRING/article/view/12012

Rohman, A., Suhartono, V., & Supriyanto, C. (2017). Penerapan algoritma c4. 5 berbasis adaboost untuk prediksi penyakit jantung. *Jurnal Cyberku*, *13*(1), 2-2.

Afifuddin, A., & Hakim, L. (2023). Deteksi Penyakit Diabetes Mellitus Menggunakan Algoritma Decision Tree Model Arsitektur C4. 5. *Jurnal Krisnadana*, *3*(1), 25-33. https://ejournal.sidyanusa.org/index.php/jkdn/article/view/470

Hana, F. M. (2020). Klasifikasi Penderita Penyakit Diabetes Menggunakan Algoritma Decision Tree C4. 5. *Jurnal SISKOM-KB (Sistem Komputer dan Kecerdasan Buatan)*, *4*(1), 32-39. https://jurnal.tau.ac.id/index.php/siskom-kb/article/view/173

Maulana, R., Narasati, R., Herdiana, R., Hamonangan, R., & Anwar, S. (2023). KOMPARASI ALGORITMA DECISION TREE DAN NAIVE BAYES DALAM KLASIFIKASI PENYAKIT DIABETES. *JATI (Jurnal Mahasiswa Teknik Informatika)*, *7*(6), 3865-3870. https://ejournal.itn.ac.id/index.php/jati/article/view/8265

Sidiq, C. M., Faqih, A., & Dwilestari, G. (2024). ALGORITMA DECISION TREE C4. 5 DIGUNAKAN UNTUK MENGKLASIFIKASIKAN DATA STROKE. *JATI (Jurnal Mahasiswa Teknik Informatika)*, *8*(2), 1869-1874. https://ejournal.itn.ac.id/index.php/jati/article/view/8388