Data Mining Model for Designing Diagnostic Applications Inflammatory Liver Disease

Amrin[1)*], Omar Pahlevi[2)]
[1)2)]Universitas Bina Sarana Informatika, Indonesia
[1)]amrin.ain@bsi.ac.id, [2)]omar.opi@bsi.ac.id

Submitted : Aug 23, 2020 | **Accepted** : Oct 6, 2020 | **Published** : Oct 6, 2020

Abstract: The liver is a vital organ for humans. Liver disease is a disorder of every liver function**.** Early diagnosis of liver disease is very important so that it can be treated and treated quickly. In the medical field, diagnosing inflammatory liver disease has become a rather difficult thing to do. However, there are medical records that store the patient's symptoms. This is certainly very beneficial for medical personnel or doctors. They can use previous medical records as material for making decisions about the patient's disease diagnosis. The conventional manual analysis technique that has been used so far is no longer effective for diagnosis. Along with the development of medical knowledge-based systems, the demand for the use of computer-based knowledge systems as an analytical technique in diagnosing diseases is becoming increasingly important. In this study, researchers will apply and compare several data mining classification methods, including the C4.5 algorithm, Naïve Bayes, and k-Nearest Neighbor to diagnose inflammatory liver disease, then compare which of the three methods is the most accurate. Based on the results of measuring the performance of the three models using the Cross Validation, Confusion Matrix and ROC Curve methods, it is known that the C4.5 method is the best method with an accuracy of 70.99% and an under the curva (AUC) value of 0.950, then the k-Nearest Neighbor method with accuracy of 67.19% and the value under the curve (AUC) 0.873, then the naïve Bayes method with an accuracy rate of 66.14% and a value under the curve (AUC) of 0.742.

Keywords: C4.5, naïve bayes, k-Nearest Neighbor, confusion matrix, ROC Curva

INTRODUCTION

The liver is a vital organ for humans. This organ is located in the right abdominal cavity, precisely below the diaphragm. There are several functions of the liver, including as an antidote and neutralizer of toxins, regulating hormone circulation, regulating the composition of blood containing fat, sugar, protein, and other substances. The liver also makes bile, a substance that helps digest fat. Liver disease is a disorder of every liver function. The liver is responsible for critical functions in the body, where the loss of these functions can cause significant damage to the body. The liver is the only organ in the body that can easily replace damaged cells, but if these cells are lost, the liver may not be able to meet the body's needs. Liver disease is often referred to as the silent killer because it may not develop symptoms (Pusporani et al., 2019).

Recently, in the medical field, diagnosing inflammatory liver disease has become a rather difficult thing to do. However, there are medical records that store the patient's symptoms. This is certainly very beneficial for medical personnel or doctors. They can use previous medical records as material for making decisions about the patient's disease diagnosis.

The conventional manual analysis technique that has been used so far is no longer effective for diagnosis. Along with the development of medical knowledge-based systems, the demand for the use of computer-based knowledge systems as an analytical technique in diagnosing diseases is becoming increasingly important. Therefore, now is the right time to develop a modern, effective and efficient computer-based knowledge system in diagnosing diseases (Neshat & Yaghoobi, 2009).

This research is in accordance with research conducted by (Nahar & Ara, 2018) who examined Liver Disease Prediction By Using Different Decision Tree Techniques. The study employed some decision tree algorithm such as J48, LMT, Random Forest, Random tree, REPTree, Decision Stump and Hoeffding Tree to predict the liver disease at an earlier stage. From the analysis, Decision Stump outperforms well than other algorithms and its achieved accuracy is 70.67%.

*Corresponding Author

This research is related to previous research conducted by (Thirunavukkarasu et al., 2018) who examined The Prediction of Liver Disease using Classification Algorithms. From the experiment, it’s found out that k-nearest neighbour model, accuracy has been calculated and it's coming out to be 73.97%. Logistic Regression model, accuracy has been calculated and it's coming out to be 73.97%. Support Vector Machine model, accuracy has been calculated and it's coming out to be 71.97%.

The research conducted (Setiawati et al., 2019) who examined the diagnosis of liver disease with the Decision Tree method. From the accuracy test, it is known that the accuracy of the Naïve Bayes method is 72.67%.

The research conducted (Prayoga, 2018) who examined the diagnosis of liver disease with the Naïve Bayes method. From the accuracy test, it is known that the accuracy of the Naïve Bayes method is 87.50%.

In this study, researchers will apply and compare several data mining classification methods, including C4.5 Algorithm, Naïve Bayes, and k-nearest neighbor to diagnose inflammatory liver disease. Which of the three methods is most accurate in diagnosing inflammatory liver disease.This research is expected to help health workers to early diagnose liver trafficking disease, and by using this application, it is expected that the general public can predict whether an inflammatory liver disease is diagnosed or not. If diagnosed, people as soon as possible consult and contact a doctor.

LITERATURE REVIEW

Liver Disease

Liver disease is a disease that has become a national problem in all countries, both in developing countries such as Indonesia and in developed countries. This disease can occur in all age groups, from children, adolescents, adults to the elderly. Statistics show that liver disease is one of the leading causes of death, both in the United States and worldwide. People with liver disease are difficult to detect, especially in the early stages of the disease. This is because the patient does not feel any symptoms of the disease and it is as if the liver is functioning normally, even though part of the liver has been damaged (Hannan et al., 2010).

Data Mining

According to Larose (2005) in (Pahlevi et al., 2018) in his book entitled "Discovering Knowledge in Data: An Introduction to Data Mining", data mining is divided into several groups based on tasks / jobs that can be done, such as:

1. Description

Sometimes researchers and analysts simply want to try to find ways to describe the patterns and trends contained in the data. The description of a trend pattern often provides possible explanations for a pattern or trend.

2. Estimation

Estimation are almost the same as classification, except that the estimated target variables are more numerical than categorical. The model is built using a complete line of data (record) that provides the value of the target variable as a predictive value. Furthermore, in the next review, the estimated value of the target variable is made based on the value of the predicted variable.

3. Prediction

Prediction is almost the same as classification and estimation, except that in the prediction the value of the results will be in the future. Several methods and techniques used in classification and estimation can also be used (under appropriate circumstances) for prediction.

4. Classification

In classification, there are target categorical variables. For example, the classification of income can be separated into three categories, namely high income, medium income, and low income.

5. Clustering

Clustering is a grouping of records, observations, or attention and forms a class of objects that have similarities. A cluster is a collection of records that are similar to one another and have different records in other clusters. Unlike the classification, the clustering there is no target variable. Clustering does not classify, estimate, or predict the value of the target variable, however, the clustering algorithm tries to divide the entire data into groups that are similar (homogeneous), where the similarity of records in one group will be the maximum value, while the similarity with records in other groups will be of minimal value.

6. Association

The task of association in data mining is to find attributes that appear at one time. One implementation of the association is market basket analysis or a priori algorithm.

Algorithm C4.5

According to Han and Kamber in (Amrin, 2018a) the C4.5 algorithm is a tree structure in which there are nodes describing the attributes, each branch describes the results of the attributes being tested, and each leaf describes the class. The C4.5 algorithm recursively visits each decision node, selecting the optimal division, until it cannot be subdivided. The C4.5 algorithm uses the concept of information gain or entropy reduction to select the optimal distribution. As for according to (Prasetyo, 2014) in (Handrianto & Farhan, 2019) explained that, "C4.5 algorithm was introduced by J. Ross Quinlan (1996) as an improved version of ID3 (Iterative Dichotomiser 3). The improvements that differentiate the C4.5 algorithm from ID3 are that it can handle numeric type features, pruning decision tree, and deriving rule set. The C4.5 algorithm is used for the decision tree.

Naïve Bayes

According to (Kusrini & Luthfi, 2009) Bayes Classification is a statistical classification that can be used to predict the probability of a class membership.

According to Han and Kamber in (Amrin, 2018b) Bayes classification, also known as Naïve Bayes, has capabilities comparable to decision trees and neural networks.

$$P(x|y) = \frac{P(y|x)\ P(x)}{P(y)} \qquad (1)$$

Remark:
y = data with unknown class
x = data hypothesis y is a specific class
P (x | y) = probability of hypothesis x based on condition y (posteriori probability)
P (x) = the probability of the hypothesis x (prior probability)
P (y | x) = probability y based on the conditions in the hypothesis x
P (y) = probability of y
Naïve Bayes is a simplification of the Bayes method. Bayes' theorem is simplified to:

$$P(x|y) = P(y|x)\ P(x) \qquad (2)$$

The K-Nearest Neighbor Algorithm

The K-Nearest Neighbor (K-NN) algorithm is a method that uses a Supervised algorithm (Han & Kamber, 2006). K-Nearest Neighbor (K-NN) is a group of instance-based learning. This algorithm is also a lazy learning technique. KNN is done by looking for groups of k objects in the training data that are closest (similar) to objects in new data or testing data (Wu & Kumar, 2009).

Confusion Matrix

The Confusion Matrix is a visualization tool commonly used in supervised learning. Each column in the matrix is an example of a prediction class, while each row represents an event in the actual class (Gorunescu, 2011).

E. ROC (Receiver Operating Characteristic) Curve

ROC curves show accuracy and visually compare classifications. ROC expresses confusion matrix. The ROC is a two-dimensional graphic with false positives as horizontal lines and true positives as vertical lines (Vercellis, 2009).

$$\theta^r = \frac{1}{mn}\sum_{j=1}^{n}\sum_{i=1}^{m}\psi\,(\mathrm{x}i^r, \mathrm{x}j^r) \qquad (3)$$

Where:

$$\boldsymbol{\psi}(\mathrm{X,Y}) = \begin{cases} 1 & Y < X \\ \frac{1}{2} & Y = X \\ 0 & Y > X \end{cases} \qquad (4)$$

According to (Gorunescu, 2011) the performance accuracy of the AUC can be classified into five groups, are as follows:
0.90 – 1.00 = Exellent Clasification
0.80 – 0.90 = Good Clasification
0.70 – 0.80 = Fair Clasification
0.60 – 0.70 = Poor Clasification
0.50 – 0.60 = Failure

F. Rapid Miner

According to (Crc & Hofmann, 2014) in (Sumpena et al., 2019) Rapid Miner is a system which supports the design and documentation of an overall data mining process. It's not only an almost comprehensive set of operators, but also structures that express the control of the process.

METHOD

This research consists of several stages as shown in the framework of Figure 1. The problem in this study is that there is no known accurate algorithm for diagnosing hepatitis. For this reason, an approach (model) is made, namely the C4.5 algorithm, Naive Bayes, and k-Nearest Neighbor to solve the problem, then testing the performance of the three methods. Testing using the Cross Validation method, Confusion Matrix and ROC curve. To develop applications (development) based on the model created, Rapid Miner is used.

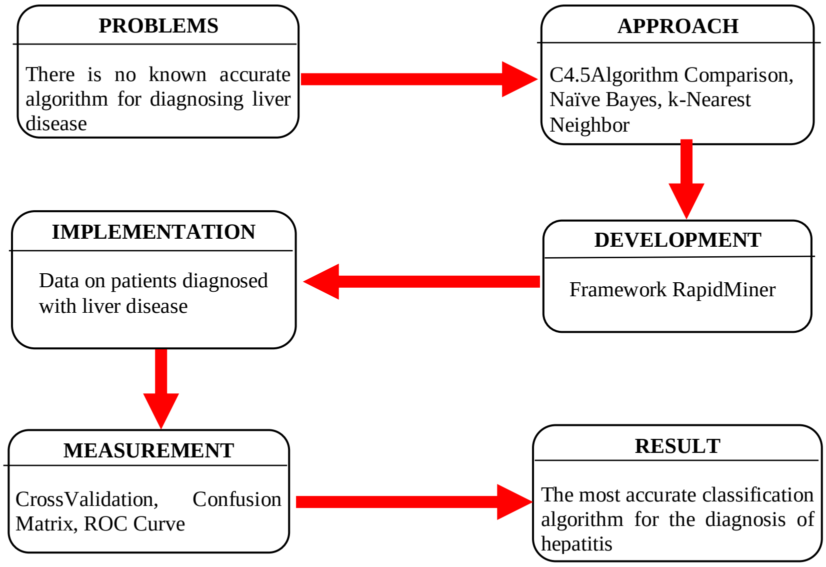
PROBLEMS
There is no known accurate algorithm for diagnosing liver disease
APPROACH
C4.5Algorithm Comparison, Naïve Bayes, k-Nearest Neighbor
DEVELOPMENT
Framework RapidMiner
IMPLEMENTATION
Data on patients diagnosed with liver disease
MEASUREMENT
CrossValidation, Confusion Matrix, ROC Curve
RESULT
The most accurate classification algorithm for the diagnosis of hepatitis

Figure 1 Problem Solving Framework

RESULT

Data Analysis

In this research, the dataset used was from the UCI Machine Learning Repository website, namely the Indian Liver Patient Dataset (ILPD). This dataset contains data collected from patients present in northeast Andhra Pradesh, India. After the data preprocessing technique was performed, the dataset contained 414 liver patients, while 165 patients were not liver sufferers. This dataset has 10 attributes where 9 attributes are input attributes while 1 attribute is output or class, and has 579 records. The attribute description is as shown in table 1 below:

Table 1
Attribute Description

Attribute	Remarks
Age	Patient's age
TB	Total patient bilirubin
DB	Direct the patient's bilirubin
Alkphos	Alkaline phosphotase
Sgpt	Almine Aminotransferase
Sgot	Aspartate Aminotransferase
TP	Total protiens
ALB	Albumin
A/G Ratio	Albumin dan Globulin ratio
Class	Class whether the patient is liver positive or not

Model Testing

This research was conducted by testing experiments on the proposed model. Then evaluate and validate the model to produce accuracy and AUC values. Testing uses Rapidminer with a 10-fold cross-validation operator to get the accuracy and AUC results for each algorithm being tested. The evaluation is done with the Confusion Matrix and ROC Curve or Area Under Curve (AUC).

1. Confusion Matrix

Table 2 is the confusion matrix for the C4.5 algorithm. It is known that from 579 data, 407 data classified as "Yes" were predicted according to the actual data, then 7 data were predicted "No" but it turned out to be "Yes". Then 4 data classified as "No" were predicted accordingly, and 161 data predicted "Yes" turned out to be "No".

Table 2
Confussion Matrix Model for the C4.5 Algorithm

Accuracy : 70.99% +/-1.98%(mikro:70.98%)

	True Yes	True No	Class Precission
Pred. Yes	407	161	71.65%
Pred. No	7	4	36.36%
Class Recall	98.31%	2.42%	

Table 3 is the confusion matrix for the Naïve Bayes algorithm. It is known that from 579 data, 270 data were classified as "Yes" exactly according to the actual data, then 144 data were predicted to be "No" but it turned out to be "Yes". Then 113 data classified "No" were predicted accordingly, and 52 data predicted "Yes" turned out to be "No".

Table 3
Confussion Matrix Model for Naïve Bayes Algorithm

Accuracy : 66.14% +/-4.55%(mikro:66.15%)

	True Yes	True No	Class Precission
Pred. Yes	270	52	83.85%
Pred. No	144	113	43.97%
Class Recall	65.22%	68.48%	

Table 4 is the confusion matrix for the k-Nearest Neighbor (k-NN) algorithm. It is known that from 579 data, 318 data were classified as "Yes" exactly according to the actual data, then 96 data were predicted to be "No" but it turned out to be "Yes". Then 71 data classified "No" were predicted accordingly, and 94 data predicted "Yes" turned out to be "No".

Table 4
Confussion Matrix Model for the K-Nearest Neighbor (K-NN) Method

Accuracy : 67.19% +/-3.79%(mikro:67.18%)

	True Yes	True No	Class Precission
Pred. Yes	318	94	77.18%
Pred. No	96	71	42.51%
Class Recall	76.81%	43.03%	

DISCUSSIONS

In this section, the researchers explained the research discussion that the calculation results are visualized with the ROC curve. The comparison of the three comparison methods can be seen in Figure 2 which is the ROC curve for the C45 algorithm.

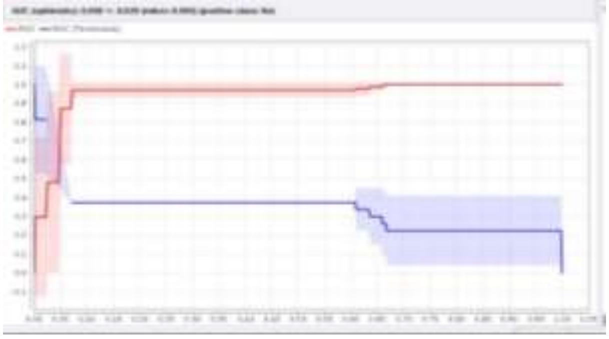

Figure 2 ROC curve with C4.5 algorithm

The ROC curve in Figure 2 expresses confusion matrix. Horizontal lines are false positives and vertical lines true positives. Furthermore, the ROC curve for the Naïve Bayes algorithm as shown in Figure 3 below.

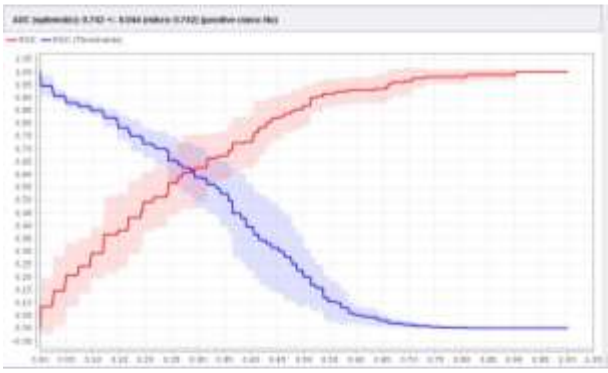

Figure 3 Naïve Bayes Algorithm ROC Curve
Figure 4

Furthermore, the ROC curve for the k-Nearest Neighbor algorithm is shown in Figure 4 below.

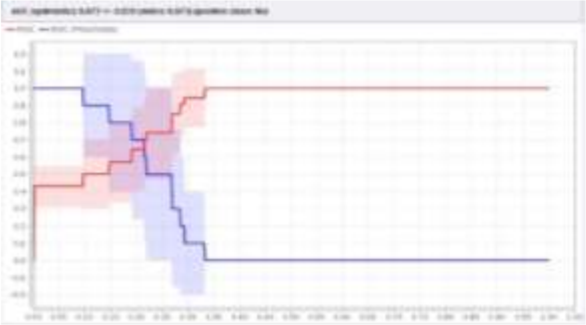

Figure 5 k-Nearest Neighbor Algorithm ROC Curve
Figure 6

The comparison of the results of the calculation of the AUC value for the C4.5, naïve Bayes, and k-Nearest Neighbor methods can be seen in Table 5.

Comparison of accuracy and ROC Curve or AUC values for the C4.5, Naïve Bayes, and k-Nearest Neighbor algorithms is shown in table 5 below. C4.5 method has the highest accuracy value, followed by the k-Nearest Neighbor method, and the lowest is the Naïve Bayes method.

Table 5
Comparison of Accuracy and AUC values

Value	C4.5	Naïve Bayes	k-NN
Accuracy	70.99%	66.14%	67.19%
AUC	0.950	0.742	0.873

Table 5 compares the accuracy and AUC of each algorithm. It can be seen that the C4.5 accuracy value for the algorithm is the highest as well as the AUC value. For data mining classification, the AUC value can be divided into several groups.

a. 0.90-1.00 = very good classification
b. 0.80-0.90 = good classification
c. 0.70-0.80 = moderate classification
d. 0.60-0.70 = poor classification
e. 0.50-0.60 = incorrect classification

Based on the grouping above and Table 5, it can be concluded that the C4.5 algorithm model is included in the very good classification category, the Naïve Bayes algorithm is in the enough classification category, and k-Nearest Neighbor is in the good classification category.

CONCLUSION

The conclusions that can be drawn based on this research are that the performance of the C4.5 model for the diagnosis of liver disease, it provides an accuracy rate of 70.99% with an area under the curva (AUC) value of 0.950. The performance of the naïve Bayes model provides a level of accuracy of truth of 66.14% with an area under the curve (AUC) of 0.742. While the performance of the k-Nearest Neighbo model provides a level of accuracy of truth of 67.19% with an area under the curve (AUC) of 0.873 and based on the level of accuracy and the value of the area under the curve (AUC), the performance of the C4.5 method is the best for diagnosing liver disease, followed by the k-nearest neighbor method, and then the naïve bayes method.

ACKNOWLEDGMENT

We would like to take this opportunity to thank all contributors of this journal. A special thanks to Dr. Mochamad Wahyudi, MM, M.Kom, M.Pd for his invaluable encouragement, guidance and support.

REFERENCES

Amrin, A. (2018a). Aplikasi Diagnosa Penyakit Tuberculosis Menggunakan Algoritma Data Mining. *Jurnal Paradigma*, *XX*(2), 91–97.

Amrin, A. (2018b). Aplikasi Diagnosa Penyakit Tuberculosis Menggunakan Algoritma Naive Bayes. *Jurikom*, *5*(5), 498–502.

Gorunescu, F. (2011). *Data Mining: Concepts, Models, and Techniques*. Springer.

Handrianto, Y., & Farhan, M. (2019). C.45 Algorithm for Classification of Causes of Landslides. *SinkrOn*, *4*(1), 120–127. https://doi.org/10.33395/sinkron.v4i1.10154

Hannan, A., Manza, R., & Remteke, R. (2010). Generalized Regression Neural Network and Radial Basis Function for Heart Disease diagnosis. *International Journal of Computer Application (0975-8887)*, *7*(13), 7–13.

Kusrini, & Luthfi, E. . (2009). *Algoritma Data Mining*. Andi Publishing.

Nahar, N., & Ara, F. (2018). Liver Disease Prediction by Using Different Decision Tree Techniques. *International Journal of Data Mining & Knowledge Management Process*, *8*(2), 01–09. https://doi.org/10.5121/ijdkp.2018.8201

Neshat, M., & Yaghoobi, M. (2009). Designing a Fuzzy Expert System of Diagnosing the Hepatitis B Intensity Rate and Comparing it with Adaptive Neural Network Fuzzy System. *Proceeding of the World Congress on Engineering and Computer Science 2009,Vol II, WCECS 2009, ISBN:978-988-18210-2-7*, 1–6.

Pahlevi, O., Sugandi, A., & Sintawati, I. D. (2018). Penerapan Algoritma Apriori Dalam Pengendalian Kualitas Produk. *SinkrOn*, *3*(1), 272–278.

Prayoga, N. D. (2018). Sistem Diagnosis Penyakit Hati Menggunakan Metode Naïve Bayes. *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer*, *2*(8), 2666–2671.

Pusporani, E., Qomariyah, S., & Irhamah. (2019). Klasifikasi Pasien Penderita Penyakit Liver dengan Pendekatan Machine Learning. *Inferensi*, *2*(1), 25–32. https://doi.org/10.12962/j27213862.v2i1.6810

Setiawati, I., Wibowo, A. P., & Hermawan, A. (2019). Implementasi Decision Tree Untuk Mendiagnosis Penyakit Liver. *JOISM : Jurnal of Information System Management*, *1*(1), 13–17.

Sumpena, Akbar, Y., Nirat, & Henky, M. (2019). Comparison of C4 . 5 Algorithm and Naïve Bayes for Last Information on ICU Patients. *SinkrOn*, *4*(1), 88–94.

Thirunavukkarasu, K., Singh, A. S., Irfan, M., & Chowdhury, A. (2018). Prediction of liver disease using classification Algorithms. *2018 4th International Conference on Computing Communication and Automation, ICCCA 2018*, *1*(1), 1–3. https://doi.org/10.1109/CCAA.2018.8777655

Vercellis, C. (2009). *Business Intelligent: Data Mining and Optimization for Decision Making*. John Willey & Sons, Ltd.

Wu, X., & Kumar, V. (2009). *The Top Ten Algorithms in Data Mining*. CRC Press.