JOURNAL MEDICAL INFORMATICS TECHNOLOGY
medinftech.org

Performance Comparison of Three Classification Algorithms for Non-alcoholic Fatty Liver Disease Patients

Adi Octaviantara[1], Moch Anwar Abbas[2], Ahmad Azhari[3], Dwiza Riana[4], and Alya Shafira Hewiz[5]*

[1,2,3,4] *Universitas Nusa Mandiri Jakarta, Indonesia*
[5]*Universitas Airlangga, Surabaya*

ARTICLE HISTORY

Received: 13 March 23

Final Revision: 16 March 23

Accepted: 30 March 23

Online Publication: 31 March 23

KEYWORDS

Non-alcoholic Fatty Liver Disease (NAFLD); Decision Trees; Naïve Bayes; k-NN.

CORRESPONDING AUTHOR

14220010@nusamandiri.ac.id

DOI

10.37034/medinftech.v1i1.2

A B S T R A C T

This study aims to carry out a comparative analysis of the three classification algorithms used in research on Non-alcoholic Fatty Liver Disease (NAFLD) Patients. NAFLD is a liver condition associated with the accumulation of fat in the liver in individuals who do not consume excessive alcohol. The algorithms used in the analysis are Decision Tree, Naïve Bayes, and k-Nearest Neighbor (k-NN), with data processing using RapidMiner software. The data used is sourced from Kaggle which comes from the Rochester Epidemiology Project (REP) database with research conducted in Olmsted, Minnesota, United States. The measurement results show that the Decision Tree algorithm has an accuracy of 92.56%, a precision of 93.24%, and a recall of 99.08%. The Naïve Bayes algorithm has an accuracy of 89.93%, a precision of 95.40% and a recall of 93.56%. While the k-Nearest Neighbor algorithm has an accuracy of 91.33%, a precision of 91.94%, and a recall of 99.27%. ROC curve analysis, all algorithms show "Excellent" classification quality. However, only the k-NN algorithm reached 1.0, showing excellent classification results in solving the problem of classifying Non-alcoholic Fatty Liver Disease patients. This study concluded that the k-NN algorithm is a better choice in solving the problem of classifying Non-alcoholic Fatty Liver Disease patients compared to the Decision Tree and Naïve Bayes algorithms. This study provides valuable insights in the development of classification methods for the early diagnosis and management of NAFLD.

1. Introduction

Non-alcoholic fatty liver disease (NAFLD) is a common cause of chronic liver disease worldwide. NAFLD is a spectrum of diseases characterized by accumulation of fat in the liver (hepatic steatosis) when no other cause can be identified for the accumulation of fat in the liver (eg, excessive alcohol consumption) [1]. NAFLD is a significant health problem worldwide, with an increasing prevalence [2]. In an effort to more effectively diagnose and manage NAFLD, the use of data analysis techniques and artificial intelligence is increasingly required.

In this study, we propose to compare the performance of three popular classification algorithms, namely Decision Tree, Naïve Bayes, and k-Nearest Neighbor in classifying patients with Non-alcoholic Fatty Liver Disease. Data processing is carried out using the RapidMiner application, a data analysis tools that can be implemented in models and prepared data [3].

The purpose of this research is to evaluate and compare the effectiveness of the three algorithms in classifying NAFLD cases. Algorithm performance will be evaluated based on standard evaluation metrics such as accuracy, precision and recall [4]. The results of this comparison will provide valuable insights in selecting the most appropriate algorithm for classifying Non-alcoholic Fatty Liver Disease patients.

In addition, this study will also introduce the NAFLD dataset used in the experiment. This dataset consists of various clinically relevant attributes, such as age, sex, weight, height, body mass index (BMI), case id, time of death or last follow-up and status (alive or dead). This dataset will be used as a basis for training and testing prediction models using the selected algorithms.

It is hoped that this research will make a significant contribution in the development of the NAFLD prediction method and the selection of the right algorithm for the task. The results of this study can provide practical guidance for health professionals in diagnosing and managing Non-alcoholic Fatty Liver Disease.

2. Research Method

2.1. Related Work

In previous studies, the classification of liver disease has been studied using various data mining algorithms. A study by Hartatik, et al., entitled Prediction for Diagnosing Liver Disease in Patients using KNN and Naïve Bayes Algorithms Based on the results of calculations and analysis it is known that the area under the curve (AUC) performance evaluation value for the naïve Bayes algorithm is 72.5% and k-nearest neighbor (KNN) of 63.19% [5]. Furthermore, research conducted by Fadilah and Murnawan entitled Performance Comparison of K-Nearest Neighbor and Decision Tree C4.5 by Utilizing Particle Swarm Optimization for Prediction of Liver Disease. Shows that Decision Tree C4.5 with PSO has a better level of performance than KNN with PSO, so Decision Tree C4.5 with PSO can be used in predicting disease. The results obtained are the accuracy value of Decision Tree C4.5 with a PSO of 91.26%, and an AUC value of 0.935. Then, Decision Tree C4.5 with PSO in processing data only takes 25 seconds to execute [6]. Finally, research conducted by Muflikhah, et al., entitled Prediction of Liver Cancer Based on DNA Sequence Using Ensemble Method compared several classifier methods including Naïve Bayes, GLM, KNN, SVM, and C5.0 Decision Tree. The results show that the ensemble method achieves high evaluation performance values with an accuracy rate of 88.4%, a sensitivity rate of 88.4%, and a specificity level of 91.4% [7].

2.2. Proposed Method

Type of research is an experimental research type, aiming to make comparisons to the Decision Tree, Naïve Bayes and k-Nearest Neighbor (k-NN) algorithms. This comparative experimental research is based on a problem-solving framework as shown in Figure 1 below:

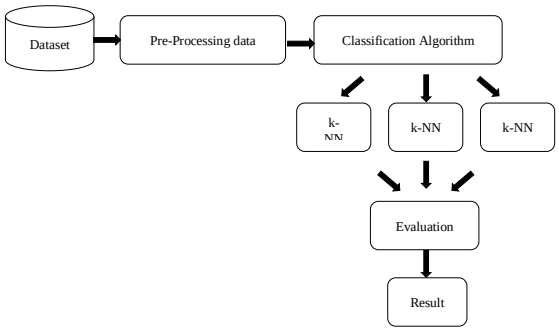

Figure 1. Research steps

2.3. Dataset

The dataset obtained is secondary data because it is obtained from data originating from Kaggle (https://www.kaggle.com/). This data is sourced from the Rochester Epidemiology Project (REP) database while the research location is in Olmesed Minnesota, United States of America , where this place is famous for being the location of the Mayo Clinic Research , the study subjects are a population of all adult Non-alcoholic Fatty Liver Disease (NAFLD) subjects from 1997 to 2014 [1] showed in Table 1.

Table 1. Attribute Description

Attribute	Information
id	subject id
age	subject's age at entry to the study
male	0 = female , 1 = male
weight	subject's weight in kg
heights	subject's height in cm
bmi	body mass index
case.id	NAFLD case id that matches the subject
futime	time of death or last follow-up
status	0 = alive at last follow-up, 1 = dead

2.4. Pre-processing data

Pre-processing is an important part of classification to improve prediction accuracy [8]. The performance of any prediction algorithm largely depends on the quality of the data, if the dataset contains too many missing values, outliers, and irrelevant attributes [9], then the overall prediction accuracy for classifying Non-alcoholic Fatty Liver Disease (NAFLD) will decrease. In this research, data preprocessing includes data cleaning, removing missing values, and data transformation.

2.5. Algorithm Classification

Classification is a process for finding a data class of an object that is not yet known based on previous data [10], classification is included in the learning or supervised method because it requires learning from previous data to determine the results of new data.

Classification has 4 basic components, namely: 1) class is a variable that becomes the label or result of an object; 2) predictor, is a variable that is an attribute of the data that will be used in classification; 3) training dataset, which is data that already has a previous label; 4) dataset testing, is new data that will be carried out by the classification process.

2.6. Decision Tree

Decision tree provide an easy way to represent the impact of each event or decision through classification. Data elements assist in the mining process and facilitate predictive modeling by dividing the data set into smaller and smaller, more precise groups [11]. The use of the Decision Tree algorithm as a classification problem solver is very good because we can find out only based on the pattern of the tree shape. The advantages of this algorithm are that it is easy to understand, flexible and has an attractive appearance because it is described in the form of a tree. While the drawbacks of this algorithm are that there is often overlap if the amount of data is very large, determining the optimal decision tree design is still difficult to produce, the quality depends on the decision tree design.

2.7. Naïve Bayes

Naïve Bayes is a classification that uses probability and statistical methods. Naïve Bayes often produces better results in many complex real-world situations. Naïve Bayes is a popular model in machine learning applications because of its simplicity in allowing all attributes to contribute equally to the final decision. This simplicity goes hand in hand with computational efficiency, thus making the Naïve Bayes technique attractive and suitable for various fields [12]. The advantages of the Naïve Bayes algorithm are: a) it is easy to implement because it does not require numerical, matrix and other optimizations; b) classified as efficient in training and use; c) binary or polynomial data can be used; d) independent nature so that it can be implemented with various datasets; e) results relatively high accuracy. Meanwhile, the disadvantages of Naïve Bayes are the inaccuracy of estimating possible classes and having to determine the threshold manually.

2.8. K-Nearest Neighbor (k-NN)

K-Nearest Neighbor (k-NN) is one of the supervised learning algorithms in data mining and is the simplest algorithm for predicting any dataset with the help of Euclidean distance [13]. K-Nearest Neighbor will find the closest distance between the test data and the k nearest neighbors in the training data. The advantages of this algorithm are that training from training data is very fast, simple, easy to learn, resistant to noise-containing training data and remains effective even though the training data is large. While the drawbacks of this algorithm are that the k value is biased, the computation required is complex, memory is limited and it is easy to be fooled if there are irrelevant attributes.

3. Result and Discussion

The performance of an algorithm in solving classification problems can be known by measuring, one of the most common ways is to calculate the accuracy of the algorithm. If the accuracy of an algorithm is said to be high, it does not mean that the algorithm is said to be good for solving classification.

Algorithm Naïve Bayes and Decision Tree are able to support data classification with good accuracy when the data type is nominal or letter. Meanwhile, the K-Nearest Neighbor algorithm is able to support data classification with good accuracy when the data type is numeric or numbers. This proves that data types are very influential in solving classification problems using data mining algorithms.

The process of implementing the Non-alcoholic Fatty Liver Disease patient classification using 3 algorithms namely Decision Tree, Naïve Bayes and K-Nearest Neighbor in the form of a schematic form in RapidMiner Studio 9.10 is shown in Figure 2.

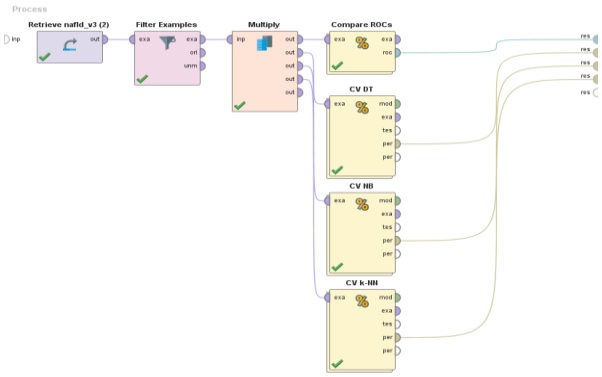

Figure 2. Schematic Classification of Non-alcoholic Fatty Liver Disease patients in RapidMiner 9.10

Implementation of Non-alcoholic Fatty Liver Disease patient classification with 3 data mining algorithms in RapidMiner 9.10 using 4 operators provided in RapidMiner. The operators and their explanations are:

Filter example, is an operator used to filter the rows in the dataset based on certain criteria. These operators allow us to filter data based on column values, logical conditions, or a combination of the two. In this case study, filter examples are used to exclude rows in the dataset that have an empty attribute / Missing Value so that these rows do not participate in data processing. The method is selected if the number of empty data is not more than 1/3 of the total data.

Multiply is operator that is used to connect many operators so that they can run simultaneously. *Compare ROCs* is the operator used to display the performance ROCs curve of each algorithm. In Compare ROCs, the 3 operator algorithms used for this research are given.

Cross Validation is an operator used to show how accurately the performance of the algorithm is. In the implementation of the Non-alcoholic Fatty Liver Disease patient classification, there are 3 cross validation operators because it only uses 3 data mining algorithms, so the number of cross validations depends on the number of algorithms used. Within this operator there is an algorithm operator, the apply model operator which is used to apply the training data model to data testing, and the performance operator is used to evaluate the algorithm. The performance results are accuracy, precision, recall and in the form of a confusion matrix.

When the schema on RapidMiner is finished running it will display the results of the accuracy of the three algorithms used for the classification of Non-alcoholic Fatty Liver Disease patients. The results of measuring accuracy, precision and recall of each algorithm are:

3.1. Decision Tree

3.1.1. Accuracy

The following is a display of the results of the confusion matrix for accuracy showed in Table 2.

Table 2. Confusion Matrix Accuracy Decision Tree

	True 0	True 1	Class Precision (%)
Pred. 0	11464	831	93.24
Pred. 1	106	187	63.82
Class recall	99.08%	18.37%	

0 = alive at last follow-up, 1 = dead

Accuracy: 92.56% +/- 0.56% (micro average: 92.56%)

The formula for calculating accuracy showed in Formula (1).

$$\text{Accuracy} = \frac{11464+187}{11464+187+831+106} x100\% = 92{,}56\% \quad (1)$$

3.1.2. Precision

The following is a display of the results of the confusion matrix for precision showed in Table 3.

Table 3. Confusion Matrix Precision Decision Tree

	True 0	True 1	Class Precision (%)
Pred. 0	11464	831	93.24
Pred. 1	106	187	63.82
Class recall	99.08%	18.37%	

0 = alive at last follow-up, 1 = dead

Precision: 64.91% +/- 13.25% (micro average:63.82%)

The formula for calculating precision showed in Formula 2.

$$\text{Precision} = \frac{11464}{11464+831} x100\% = 93{,}24\% \quad (2)$$

3.1.3. Recall

The following is a display of the results of the confusion matrix for Recall showed in Table 4.

Table 4. Confusion Matrix Recall Decision Tree

	True 0	True 1	Class Precision (%)
Pred. 0	11464	831	93.24
Pred. 1	106	187	63.82
Class recall	99.08%	18.37%	

0 = alive at last follow-up, 1 = dead

Precision: 18.37% +/- 4.75% (micro average:18.37%)

The formula for calculating Recall showed in Formula (3).

$$\text{Recall} = \frac{11464}{11464+106} x100\% = 99{,}08\% \quad (3)$$

3.2. Naïve Bayes

3.2.1. Accuracy

The following is a display of the results of the confusion matrix for accuracy showed in Table 5.

Table 5. Confusion Matrix Accuracy Naïve Bayes

	True 0	True 1	Class Precision (%)
Pred. 0	10825	522	95.40
Pred. 1	745	496	39.97
Class recall	93.56%	48.72%	

0 = alive at last follow-up, 1 = dead

Accuracy :89.93% +/- 0.86% (micro average:89.93%)

The formula for calculating accuracy showed in Formula (4).

$$\text{Accuracy} = \frac{10825+496}{10825+496+522+745} x100\% = 89{,}93\% \quad (4)$$

3.2.2. Precision

The following is a display of the results of the confusion matrix for precision showed in Table 6.

Table 6. Confusion Matrix Precision Naïve Bayes

	True 0	True 1	Class Precision (%)
Pred. 0	10825	522	95.40
Pred. 1	745	496	39.97
Class recall	93.56%	48.72%	

0 = alive at last follow-up, 1 = dead

Precision: 40.14 +/- 4.37% (micro average: 48.72%)

The formula for calculating precision showed in Formula (5).

$$\text{Precision} = \frac{10825}{10825+522} x100\% = 95{,}40\% \quad (5)$$

3.2.3. Recall

The following is a display of the results of the confusion matrix for Recall showed in Table 7.

Table 7. Confusion Matrix Recall Naïve Bayes

	True 0	True 1	Class Precision (%)
Pred. 0	10825	522	95.40
Pred. 1	745	496	39.97
Class recall	93.56%	48.72%	

0 = alive at last follow-up, 1 = dead

Recall: 48.73 +/- 4.37% (micro average: 48.72%)

The formula for calculating Recall showed in Formula (6).

$$\text{Recall} = \frac{10825}{10825+745} x100\% = 93{,}56\% \quad (6)$$

3.3 k-Nearest neighbors

3.3.1. Accuracy

The following is a display of the results of the confusion matrix for accuracy showed in Table 8.

Table 8. Confusion Matrix Accuracy k-NN

	True 0	True 1	Class Precision (%)
Pred. 0	11485	1007	91.94
Pred. 1	85	11	11.46
Class recall	99.27%	1.08	

0 = alive at last follow-up, 1 = dead

Accuracy: 91.33% =/- 0.36% (micro average:91.33%)

The formula for calculating accuracy showed in Formula (7).

$$\text{Accuracy} = \frac{11485+11}{11485+11+1007+85} x100\% = 91{,}33\% \quad (7)$$

3.3.2. Precision

The following is a display of the results of the confusion matrix for precision showed in Table 9.

Table 9. Confusion Matrix Precision k-NN

	True 0	True 1	Class Precision (%)
Pred. 0	11485	1007	91.94
Pred. 1	85	11	11.46
Class recall	99.27%	1.08	

0 = alive at last follow-up, 1 = dead

Precision: 10.75% =/- 12.52% (micro average:11.46%)

The formula for calculating precision showed in Formula (8).

$$\text{Precision} = \frac{11485}{11485+1007} x100\% = 91{,}94\% \quad (8)$$

3.3.3. Recall

The following is a display of the results of the confusion matrix for Recall showed in Table 10.

Table 10. Confusion Matrix Recall k-NN

	True 0	True 1	Class Precision (%)
Pred. 0	11485	1007	91.94
Pred. 1	85	11	11.46
Class recall	99.27%	1.08	

0 = alive at last follow-up, 1 = dead

Recall: 1.08% =/- 1.27% (micro average:1.08%)

The formula for calculating Recall showed in Formula (9).

$$\text{Recall} = \frac{11485}{11485+85} x100\% = 99{,}27\% \quad (9)$$

The tabular form of the measurement results of accuracy, precision and recall of the three algorithms showed in Table 11.

Table 11. Algorithm Measurement Results (%)

Algorithm	Accuracy	Precision	Recall
DecisionTree	92.56	93.24	99.08
Naïve Bayes	89.93	95.40	93.56
k-Nearest Neighbor	91.33	91.94	99.27

Based on table 4, it can be seen that the results of the three algorithms for solving the problem of classifying Non-alcoholic Fatty Liver Disease patients are quite good. The Decision Tree algorithm produces better accuracy than the Naïve Bayes and k-Nearest Neighbor algorithms. Decision Tree has an accuracy of 92.56%.

The results of the ROC curve for the classification of Non-alcoholic Fatty Liver Disease patients using 3 algorithms are in Figure 3.

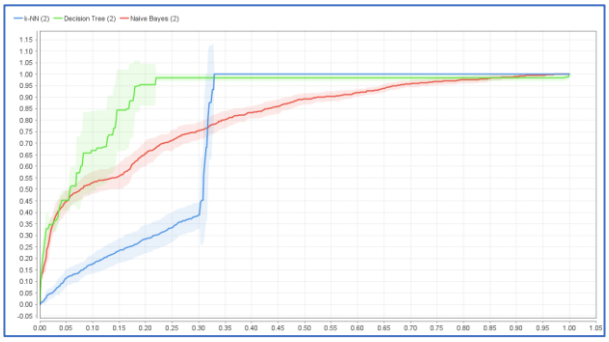

Figure 3. ROCs curve

Based on the results of the ROCs curve in Figure 3, it can be said that all algorithms have an "Excellent" classification quality where it can be seen that all the graphs of the three algorithms are in the accuracy range of 0.90 - 1.00 on the Y axis, but only the k-NN algorithm which touches 1.0 thus can it was concluded that the k-NN algorithm has good classification results in solving the classification problem of Non-alcoholic Fatty Liver Disease patients.

4. Conclusion

The findings of this study show that machine learning classification model especially the k-NN model accurately predicts Non-alcoholic Fatty Liver Disease patients using minimum clinical variables. This method may lead to greater insights in the real world clinical practice which would assist physicians to effectively identify NFLD for novel diagnosis, preventive and therapeutic purpose to mitigate the global burden of NFLD. Future studies are needed to validate our model to predict NFLD in various types of dataset.

References

[1] A. M. Allen, T. M. Therneau, J. J. Larson, A. Coward, V. K. Somers, and P. S. Kamath, "Nonalcoholic fatty liver disease incidence and impact on metabolic burden and death: A 20 year-community study," *Hepatology*, vol. 67, no. 5, pp. 1726–1736, 2018, doi: 10.1002/hep.29546.

[2] S. Pouwels *et al.*, "Non-alcoholic fatty liver disease (NAFLD): a review of pathophysiology, clinical management and effects of weight loss," *BMC Endocr. Disord.*, vol. 22, no. 1, pp. 1–9, 2022, doi: 10.1186/s12902-022-00980-1.

[3] S. Bashir, Z. S. Khan, F. Hassan Khan, A. Anjum, and K. Bashir, "Improving Heart Disease Prediction Using Feature Selection Approaches," *Proc. 2019 16th Int. Bhurban Conf.*

Appl. Sci. Technol. IBCAST 2019, pp. 619–623, 2019, doi: 10.1109/IBCAST.2019.8667106.

[4] M. Ghosh *et al.*, "A comparative analysis of machine learning algorithms to predict liver disease," *Intell. Autom. Soft Comput.*, vol. 30, no. 3, pp. 917–928, 2021, doi: 10.32604/iasc.2021.017989.

[5] M. B. T. and A. S. H. Hartatik, "Prediction for Diagnosing Liver Disease in Patients using KNN and Naïve Bayes Algorithms," *2020 2nd Int. Conf. Cybern. Intell. Syst.*, pp. 1–5, 2020, doi: 10.1109/ICORIS50180.2020.9320797.

[6] Z. Fadilah and Murnawan, "Performance Comparison of K-Nearest Neighbor and Decision Tree C4.5 by Utilizing Particle Swarm Optimization for Prediction of Liver Disease," *Int. J. Open Inf. Technol.*, vol. 9, no. 10, pp. 9–15, 2021.

[7] W. F. M. and S. L. Muflikhah, N. Widodo, "Prediction of Liver Cancer Based on DNA Sequence Using Ensemble Method," *2020 3rd Int. Semin. Res. Inf. Technol. Intell. Syst.*, pp. 37–41, 2020, doi: 10.1109/ISRITI51436.2020.9315341.

[8] S. A. D. and S. S. S. H. S. Obaid, "The Impact of Data Pre-Processing Techniques and Dimensionality Reduction on the Accuracy of Machine Learning," *2019 9th Annu. Inf. Technol. Electromechanical Eng. Microelectron. Conf.*, pp. 279–283, 2019, doi: 10.1109/IEMECONX.2019.8877011.

[9] M. A. and M. K. H. M. F. Rabbi, S. M. Mahedy Hasan, A. I. Champa, "Prediction of Liver Disorders using Machine Learning Algorithms: A Comparative Study," *2020 2nd Int. Conf. Adv. Inf. Commun. Technol.*, pp. 111–116, 2020, doi: 10.1109/ICAICT51780.2020.9333528.

[10] I. C. Husin Muhamad, Cahyo Adi Prasojo , Nur Afifah Sugianto, Listiya Surtiningsih, "Optimasi Naïve Bayes Classifier dengan Menggunakan Particle Swarm Optimization Pada Data Iris," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 4, no. 3, pp. 180–184, 2017, doi: 10.25126/jtiik.201743336.

[11] S. K. D. M. D. Ganggayah, N. A. Taib, Y. C. Har, P. Lio, "Predicting factors for survival of breast cancer patients using machine learning techniques," *BMC Med. Inform. Decis. Mak*, vol. 19, n, p. 48, 2019, doi: 10.1186/s12911-019-0801-4.

[12] A. P. Wibawa *et al.*, "Naïve Bayes Classifier for Journal Quartile Classification," *International Journal of Recent Contributions from Engineering, Science & IT (iJES)*, vol. 7, no. 2. p. 91, 2019. doi: 10.3991/ijes.v7i2.10659.

[13] R. Williams, T. Shongwe, A. N. Hasan, and V. Rameshar, "Heart Disease Prediction using Machine Learning Techniques," *2021 Int. Conf. Data Anal. Bus. Ind. ICDABI 2021*, vol. 3075, no. 5, pp. 118–123, 2021, doi: 10.1109/ICDABI53623.2021.9655783.