# Effect of Fuzzy Criteria on the Performance of Decision Tree Models for Heart Disease Classification

Endang Sri Kresnawati[1], Des Alwine Zayanti[2], Ali Amran[3], Yulia Resti[4], Nada Aisyah[5], Anggraini Salsabila[6]

{eskresna@unsri.ac.id[1], desalwinez@unsri.ac.id[2], ali_amran@mipa.unsri.ac.id[3], fyresti@yahoo.com[4], nadaaisyah01@gmail.com[5], anggrainibila12@gmail.com[6]}

Universitas Sriwijaya, Faculty of Mathematics and Natural Sciences, Department of Mathematics[1,2,3,4,5,6]

Corresponding author: eskresna@unsri.ac.id

**Abstract.** Fuzzy is a special algorithm used in developing Decision Tree Models. The decision rules obtained not only depend on the tree structure formed, but also on the determination of numerical criteria and fuzzy linguistic criteria. This study aims to measure the performance of decision tree models in diagnosis heart disease patients. Models were built using two, three, four, and five fuzzy criteria. The research stages start from discretization, building a decision tree structure, developing decision rules, and evaluating model performance. The calculation results show that model performance increases in models using two and three fuzzy criteria, then decreases in models with four and five criteria. The model with three criteria provides better performance. The best performance was obtained from a model with three criteria, namely 73.33% accuracy, 77.08% precision and 74% recall.

**Keywords:** decision tree, fuzzy criteria, accuracy, precision, recall

## 1 Introduction

Heart disease is a degenerative disease that occurs due to narrowing of blood vessels which are blocked by fat. Heart disease can be anticipated by doing a health check. The results of the health examination become a doctor's reference for determining a diagnosis. Diagnosis is the process of making decisions about a patient's clinical problems, from data collection to conclusions. A diagnostic model is a mathematical model used to predict and classify diseases in a health status with the aim of helping doctors decide on a diagnosis of a disease more quickly and accurately. The mathematical method used for this purpose is Fuzzy Decision Trees. In several studies, this classification method has provided good performance test results compared to other classification methods, [1] - [6].

SICBAS 2023, November 03, Palembang, Indonesia

DOI 10.4108/eai.3-11-2023.2347907

One of the important things in fuzzy is domain. This domain is related to established criteria and has an impact on decision making. In theory there is no provision regarding the number of fuzzy criteria. Most studies use an odd number of criteria. [7], [8], and, [13] uses three and four criteria. [9] uses four criteria. He  uses two and three criteria, while [2] uses two, three and five criteria. Most researchers use three criteria (odd).

This research aims to develop a Fuzzy Decision Tree model to classify heart disease by applying an even number of fuzzy criteria and seeing its effect on the results of model performance evaluation. Is the even criterion better, less good, or as good in accuracy as the model with odd fuzzy criteria? The research was carried out through three main stages, initial data processing, model building, and model performance evaluation. First, the data is divided randomly into 80% training data and 20% test data. Training data is used to build a decision tree model and test data to evaluate model performance. Next, discretizing the numerical variables into categorical ones using Equal Width Distance (EWD) and determine the fuzzy membership values for these variables. The second stage builds a decision tree model based on the entropy value and gain of the predictor variables. The tree structure is a reference for preparing fuzzy rules for making decisions. The third stage, evaluating the performance of the model obtained using the Confusion Matrix, by determining the Accuracy, Specificity and Sensitivity values of the model.

## 2 Materials and Method

Criteria are measures that form the basis for evaluating or determining something. Fuzzy criteria are measurements used as a basis for determining the conditions of fuzzy variables. A variable can be expressed in terms of several criteria. Criteria are expressed linguistically and numerically. Each variable has at least two criteria. Each criterion has a domain. A set is said to be fuzzy if, at least one element of the fuzzy set falls into at least two domains. The more criteria there are, the more vague the variable is described. So that later, when used for learning, it will produce a learning model with different performance qualities.

The Fuzzy Decision Tree Model is built based on the fuzzy membership function values. Determining the type of membership function, domain, and number of criteria provides different calculation and decision results. The root node in the decision tree is determined by the highest gain of all predictor variables. Entropy and Information Gain are obtained using [14]

$$H_f(S,A) =- \sum_{i=1}^{C} \frac{\sum_j^N \mu_{ij}}{S} log_2 \frac{\sum_j^N \mu_{ij}}{S} \tag{1}$$

$$G_f(S,A) = H_f(S) - \sum_{v\subseteq A}^{C} \frac{|S_v|}{|S|} H_f(S_v, A) \tag{2}$$

$\mu_j$, membership value of the *j*th pattern for the *i*th class. $H_f(S)$, Set $S$ entrophy. $|S_v|$, size of subset $|S_v| \subseteq S$, from training data $x_j$ with variable $v$. $|S|$, size of set $S$.

The predictor variable that has the highest gain value will become the root node in the fuzzy decision tree. The variables are additionally recalculated entropy and gain to determine which will be the leaf and branch nodes. This step is carried out until all predictor variables have answers of one class. Figure 1 is the structure of the Decision Tree Model with two fuzzy criteria. This study used data from heart disease patients with thirteen predictor variables and one target variable, namely heart and non-heart.

## 2 Result and Discussion

Five of the thirteen predictor variables are numerical variables. The ID3 algorithm on decision trees can only be applied to categorical data. Therefore, data transformation was first carried out on these five variables. The transformation uses fuzzy [15] combined with the Equivalent Weight Distance (EWD) technique. EWD divides the data sequence into equal intervals based on criteria, and each domain element matches only one criterion. This does not meet the fuzzy set requirements. Therefore, after all domain elements are divided equally based on the number of criteria, the intervals are adjusted so that there are domain elements that fall into two or more criteria. Determining elements as members of a criterion uses a linear fuzzy membership function. All results of data transformation are in Table 1. The criteria for young, old, low or high are obtained by fuzzyifying numerical variables into fuzzy membership values. Numerical variables are converted into fuzzy membership values, which are then assigned to a linguistic criterion. After data transformation, the training data and test data are divided into a composition of 80% and 20%.

**Table 1.** Discretization of Predictor Variables with Two Criteria

| No | age | | trestbps | | cholestrol | | thalach | |
|---|---|---|---|---|---|---|---|---|
| | numeric | linguistic | numeric | linguistic | Numeric | linguistic | numeric | linguistic |
| 1 | 70 | old | 130 | low | 322 | normal | 109 | normal |
| 2 | 67 | old | 115 | low | 564 | high | 160 | up normal |
| 3 | 57 | old | 124 | low | 261 | normal | 141 | up normal |
| 4 | 64 | old | 128 | low | 263 | normal | 105 | normal |
| 5 | 74 | old | 120 | low | 269 | normal | 121 | normal |
| 6 | 56 | old | 130 | low | 256 | normal | 142 | up normal |
| 7 | 59 | old | 110 | low | 239 | normal | 142 | up normal |
| 8 | 60 | old | 140 | low | 293 | normal | 170 | up normal |
| 9 | 63 | old | 150 | high | 407 | high | 154 | up normal |
| ⋮ | | | | | | | | |
| 214 | 56 | old | 140 | low | 294 | normal | 153 | up normal |
| 215 | 57 | old | 140 | low | 192 | normal | 148 | up normal |
| 216 | 67 | old | 160 | high | 286 | normal | 108 | normal |

Numerical to categorical data transformation is carried out on numerical predictor variables, so that they can be processed in the Decision Tree Model which only reads categorical data. This discretization technique is different from [9] - [11], which only use a fuzzy approach. The decision tree structure is prepared by determining the variables that are the roots, branches and leaves based on the values in Table 1 and Equations (1) and (2). The variable that has the highest gain becomes the root, then another calculation is carried out to determine the variable that occupies the leaf node. This is done repeatedly until there are no more predictor variables that can be expanded. Figure1 is a picture of a decision tree structure that uses two fuzzy criteria. This tree produces 23 decision rules. Tree structures using three, four, and five fuzzy criteria have a similar shape, but have a greater number of branches. As a result, these trees produce more decision rules as well. The complete number of decision rules is shown in Table 3.

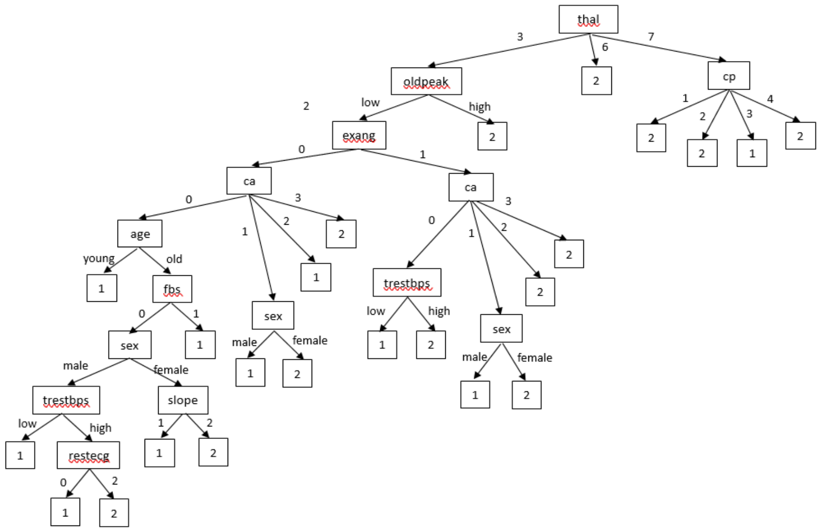

**Fig.1.** Decision Tree with Two Fuzzy Criteria

The following is an example of the decision rules that are formed

If thal = 3 and oldpeak=low and exang = 0 and ca = 0 and age = young, Then 1 (heart disease)

If thal = 7 and cp = 1, Then 2 (no heart disease)

And others

The first rule means if thal is category three and oldpeak is low, and exang is zero, and the ca is zero, and age is young, then heart disease. Second, if thal is category 7 and cp is one, then it does not heart disease.

The root of tree in Figure 1 is the variable Thal. Thal is a blood disorder called thalassemia. Value 0: Null (dropped from the previous dataset). Value 1: fixed defect (no blood flow in some part of the heart), Value 2: normal blood flow, Value 3: reversible defect (a blood flow is observed but it is not normal). As a root, thal is the dominant factor causing heart disease. After obtaining a decision tree model, the model needs to be evaluated to see how good its performance is. Evaluation using a confusion matrix.

**Table 2**. Confusion Matrix of Decision Tree Model with Two Fuzzy Critera

| | True 1 | True 2 |
|---|---|---|
| Prediction 1 | 21 | 7 |
| Prediction 2 | 8 | 18 |
| Accuracy 72.96% | | |

Table 2 is the result of evaluating the performance of the decision tree model with two fuzzy criteria. Presented in matrix form, it states the actual value and predicted value. Accuracy is 72.96%, Precision is 76.92, and Recall is 73.33%. The evaluation results of the decision tree model with three, four, and five criteria are shown in Table 3 below.

**Table 3**. Model Evaluation Results with Confusion Matrix

| Number of Fuzzy Criteria | Accuracy (%) | Precision (%) | Recall (%) | Number of Classification Rule |
|---|---|---|---|---|
| 2 | 72.96 | 76.92 | 73.33 | 23 |
| 3 | 73.33 | 77.08 | 74.00 | 35 |
| 4 | 73.33 | 76.71 | 74.67 | 37 |
| 5 | 72.22 | 75.51 | 74.00 | 37 |

Table 3 states that the lowest accuracy is in two criteria and the highest is in five criteria. It also appears that the more criteria there are, the more rules are generated by the tree structure. Model performance testing in Table 3 shows that performance increases in models with two to three fuzzy criteria, then decreases in models with four and five fuzzy criteria. This means that more criteria does not make performance better. If we look at the criteria for an even or odd number, the odd criteria provide better performance than the even criteria. The best odd criterion is three, more than three the model performance tends to decrease.

**Table 4**. Similar Work without Fuzzy Control

| Authors | Number of Fuzzy Criteria | Accuracy (%) | Year |
|---|---|---|---|
| [9] | 3 & 4 | 67.17 | 2022 |
| [12] | 2 & 3 | 55 | 2021 |
| [8] | 3 | 71.49 | 2014 |

When compared with similar studies in Table 4, the accuracy obtained from this study is greater than the previous one. However, when compared with this research, the accuracy results are smaller. It is diffirent to the similar research in Table 5. This could be due to, research [13], [4], and [1], using fuzzy controllers. The fuzzy controller functions to control tree expansion. Because, in theory, the more variables and fuzzy criteria, the more branches the tree will have, the more branches actually reduce the performance of the model. However, this research focuses on proving whether the number of criteria affects the quality of model performance or not, without involving fuzzy controller tools.

**Table 5**. Similar Work with Fuzzy Control

| Authors | Number of fuzzy criteria | Accuracy (%) | Year | Information |
|---|---|---|---|---|
| [13] | 3 & 4 | 95.85 | 2020 | Using fuzzy decision tree branch controller |
| [4] | 2,3, and 5 | 84.8 | 2021 | |
| [1] | 3 | 91.56 | 2021 | |

## 3 Conclusion

The greater the number of fuzzy criteria used in a decision tree, it does not guarantee that the model's performance will be better. Model accuracy increases from two criteria to three criteria, then continues to decrease to five criteria. The highest accuracy was obtained from a model with three criteria, namely 73.33%. This also shows that the  number of criteria three has a better performance than the model with number of  criteria two. For future work, use three fuzzy criteria and a fuzzy branch controller to increase the accuracy of model performance.

## References

[1] Idris, N. F. and Ismail, M. A.: Breast cancer disease classification using fuzzy-id3 algorithm with fuzzy dbd method: Automatic fuzzy database definition. PeerJ Comput. Sci. Vol. 7, pp. 1–22 (2021)

[2] Begenova, S. B. and Avdeenko,  T. V.: Building of fuzzy decision trees using id3 algorithm. J. Phys. Conf. Ser. Vol. 1015 (2018)

[3] Aggarwal, H., Arora, H. D., and Kumar, V.: Constructing a data mining model using fuzzy decision tree. Int. J. Adv. Sci. Technol. Vol. 29, pp. 3924–3934 (2020)

[4] Rabcan, J. et al: Fuzzy decision tree based method in decision-making of covid-19 patients treatment. Mathematics, vol. 9, (2021)

[5] Kadi, I. and Idri, A.: Cardiovascular dysautonomias diagnosis using crisp and fuzzy decision tree: a comparative study. Stud. Health Technol. Inform. Vol. 223, pp. 1–8 (2016)

[6] Bressan, G. M., de Azevedo, B. C. F., and de Souza, R. M. A fuzzy approach for diabetes mellitus type 2 classification. Brazilian Arch. Biol. Technol. Vol. 63 (2020)

[7] Lotfi, S. et al.: Scalable fuzzy decision tree induction using fast data partitioning and incremental approach for large dataset.  J. Adv Comp Eng Technol. Vol. 7, pp. 55–66 (2021)

[8] Gupta, V. A. and Soni, S.: review of fuzzy decision tree: an improved decision making classifier. SSRG Int. J. Compt. Sci. Eng. Vol. 1, pp. 12–17 (2014)

[9] Lotfi, S. et al.: Scalable decision tree based on fuzzy partitioning and an incremental approach.  Int. J. Electr. Comput. Eng. Vol. 12, pp. 4228–4234 (2022)

[10] Resti, Y. et al.: Diagnosis of diabetes mellitus in women of reproductive age using the prediction methods of naive bayes, discriminant analysis,  and logistic regression. Sci. Technol. Indones. Vol. 6, pp. 96–104 (2021)

[11] Resti, Y.et al.: Rain event prediction performance using decision tree method. Towar. Adapt. Res. Technol. Dev. Futur. Life. Vol. 2689, pp. 120006 (2023)

[12] Devare, V. S.: Heart disease prediction using binary classification. Calif. State  Univ. San Bernardin (2023)

[13] Santoso, H. B.: Fuzzy decision tree to predict student success in their studies. Int. J. Quant. Res. Model. Vol. 1, pp. 135–144 (2020)

[14] Wang, T. et al.: A survey of  fuzzy decison tree classifier methodology. Proceedings of the Second International Conference of Fuzzy Information and Engineering (2007)

[15] Eliyati, N. et al.: Prediction of air quality index using decision tree with discretization. Indonesian Journal of Engineering and Science. Vol. 3, pp. 61–67 (2022)