Journal of Sisfotek Global
ISSN (Online): 2721 - 3161, ISSN (Print): 2088 – 1762
DOI: http://dx.doi.org/10.38101/sisfotek.v12i2.551
Vol. 12, No. 2, September 2022, pp. 138-142

# The Utilization of Decision Tree Algorithm In Order to Predict Heart Disease

**Mia[1], Anis Fitri Nur Masruriyah[2], Adi Rizky Pratama[3]**

[1,2,3] Universitas Buana Perjuangan Karawang, Indonesia, 41316
E-mail: [1]if19.mia@mhs.ubpkarawang.ac.id, [2]anis.masruriyah@ubpkarawang.ac.id, [3]adi.rizky@ubpkarawang.ac.id

## ARTICLE HISTORY

*Received : September 9th, 2022*
*Revised : September 27th, 2022*
*Accepted : September 28th, 2022*

## KEYWORDS

*ADASYN*
*C45*
*Heart Disease*
*Random Forest*
*SMOTE*



## ABSTRACT

The data on heart disease patients obtained from the Ministry of Health of the Republic of Indonesia in 2020 explains that heart disease has increased every year and ranks as the highest cause of death in Indonesia, especially at productive ages. If people with heart disease are not treated properly, then in their effective period a patient can experience death more quickly. Thus, a predictive model that is able to help medical personnel solve health problems is built. This study employed the Random Forest and Decision Tree algorithm classification process by processing cardiac patient data to create a predictive model and based on the data obtained, showing that the data on heart disease was not balanced. Thus, to overcome the imbalance, an oversampling technique was carried out using ADASYN and SMOTE. This study proved that the performance of the ADASYN and SMOTE oversampling techniques on the C45 algorithm and the Random Forest Classifier had a significant effect on the prediction results. The usage of oversampling techniques to analyze data aims to handle unbalanced datasets, and the confusion matrix is used for testing Precision, Recall, and F1-SCORE, as well as Accuracy. Based on the results of research that has been carried out with the K-Fold 10 testing technique and oversampling technique, SMOTE + RF is one of the best oversampling techniques which has a greater Accuracy of 93.58% compared to Random Forest without SMOTE of 90.51% and SMOTE + ADASYN of 93.55%. The application of the SMOTE technique was proven to be able to overcome the problem of data imbalance and get better classification results than the application of the ADASYN technique.

## 1. Introduction

Based on a collection of data obtained from the Ministry of Health of the Republic of Indonesia [1] Heart disease had increased every year and ranks as the highest cause of death in Indonesia, especially at productive ages. During the COVID-19 pandemic, patients with congenital heart disease have a greater safety risk because it caused exacerbations and even death. If people with heart disease are not treated properly, then in their productive age a patient can experience death more quickly. Thus, the need for a predictive model that is able to help medical personnel solve health problems. Based on the results of research conducted by the British Heart Foundation (BHF) [2] attached to Figure 1 the graph describes the most common premature death from heart and circulatory diseases (before the age of 75) in the north of England. Mortality rates take the local age structure (demography) into account to reveal significant differences in statistics.

Furthermore, coronary heart disease (CHD) is the most common type of heart and circulatory disease. It occurs when the coronary arteries become narrowed by the buildup of atheroma, a fatty material within their walls [2]. Attached to Figure 2 Coronary heart disease chart in 2020 is quite large, around 65.644%.

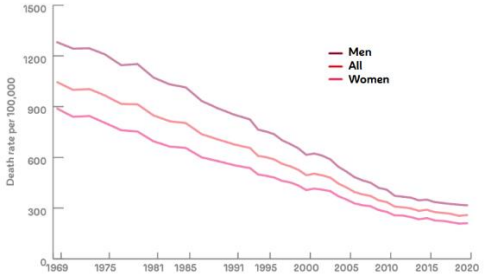


Figure 1. Heart Disease Chart 1969-2020

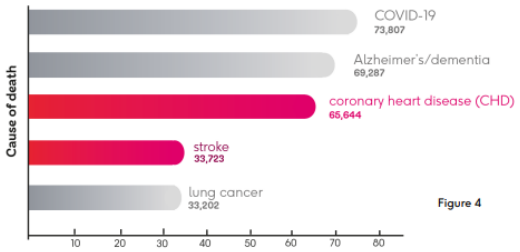


Figure 2. Coronary Heart Disease Chart 2020

An earlier study conducted by Pangaribuan [3] compared the C45 algorithm method and extreme machine learning which provided excellent heart disease diagnostic results up to 99.05%. Furthermore, a previous study conducted by Rohman and Rochcham [4] predicted the heart disease model using the C45 algorithm got a value of 86.59%, the AUC value obtained was 0.957 and included in the excellent group category. Meanwhile, in skin cancer research, the processes utilized for feature extraction include histogram, haralick and hue moments. Of the three feature extractions employed, the best Random Forest algorithm accuracy results were obtained by extracting hue moments with an accuracy value of 0.842 [5]. Hereinafter, research conducted by Ath [6] to predict heart disease employed Machine Learning (ML) algorithms as an early preventive effort in desktop-based information systems. The accuracy value obtained utilising the Random Forest and Logistic Regression methods of 84.48%, an increase of 1.32%. Furthermore, research conducted by El-Hasnony [7] constructed a model for the prevention of stroke and heart disease and employed a machine algorithm. Active learning was applied to research and determine the most significant factors in heart disease.

Based on these influencing factors, medical personnel were able to take suitable actions to treat and control stroke and heart disease. There were several approaches to overcome the imbalance in the heart disease data sample utilizing the original data sampling method, both in the majority class (under sampling) and minority class (oversampling). Oversampling is a method for balancing class distribution by randomly replicating instances across a small fraction of the class [8], [9]. This study aimed to find the best model for heart disease cases by applying several algorithms by comparing the results of the performance classification of the SMOTE and ADASYN methods in dealing with the imbalance in a data case [10]. In addition, the use of feature extraction to determine the variables that most influence heart disease was based on calculations.

## 2. Material and Method

This study employed data on heart disease patients with a total object of more than 300,000 data with seven variables and one target class. The data set was obtained from medical records that had been authorised by the World Health Organization and accessed at the Centers for Disease Control and Prevention (CDC) [11]. In general, the data analysis process began with preprocessing and feature extraction until a new understanding was generated. This study employed four analytical stages (data quality analytics, descriptive analytics, diagnostic analytics, and predictive analytics). The first stage of data pre-processing included data quality analytics and descriptive analytics. Furthermore, the results of data pre-processing were processed to get the results of diagnostic analytics and predictive analytics.

### 2.1 SMOTE

The Synthetic Minority Oversampling Technique (SMOTE) method is the primary way to overcome class imbalance. This technique synthesizes a new sample from the minority class to balance the data set by resampling the minority class sample [12]. Corrected unbalanced data operating oversampling in the minority class or undersampling in the majority class. This technique balances the dataset by creating new instances of the minority class by synthesizing new samples of the minority class to form the combinatorial convexity of adjacent models. SMOTE is a better way to increase the number of rare cases than simply duplicating existing cases. In the previous case, SMOTE is able to remove noise and solve the imbalance problem**.**

### 2.2 ADASYN

ADASYN is a sampling approach for training operating unbalanced data sets. ADASYN manages distribution weights for minority class data based on the difficulty of training the data with the model. Synthesis data is generated from minority classes that are difficult to learn and minority data that are easy to learn. Research conducted by Fico [13] ADASYN was capable generating samples adaptively in synthetic data against minority classes. ADASYN is formed by the distribution of data to reduce inequality in the majority class label data. Furthermore, research conducted by Rahayu [14] on oversampling changed the sample data by adding sample data contained in the minority class by making a replica of the sample data so that the distribution of sample data became more balanced.

### 2.3 Decision Tree Algorithm

The C4.5 algorithm is the development of the ID3 decision tree algorithm proposed by Quinlan in 1983 [15]. The way these algorithm works begins by assessing the weight of each attribute by calculating entropy (Equation 1), and then calculating the relationship between attributes using information gain

(Equation 2, 3) [11][12]. Furthermore, the attribute which has the highest relationship to other attributes intention serve as the root of the decision tree. Moreover, other attributes which have a lower gain value become branches or leaves. This method is done by employing the split algorithm (Equation 4) so that the attribute that maximizes the information acquisition ratio is selected as the best split feature.

$$\text{Entropy}(s) = \sum_{i=1}^{e} -p_i \log_2 p_i \quad (1)$$

$$\text{Information}_{\text{Atribute}}(D) = \sum_{j=1}^{v} \left|\frac{D_i}{D}\right| \text{x Info}(D_i) \quad (2)$$

$$\text{Information Gain (Atribute)} = \text{info}(D) - \text{Info}(D_i) \quad (3)$$

$$\text{Splitln o}_{\text{Atribute}}(D) = \sum_{j=1}^{v} \left|\frac{D_i}{D}\right| \text{x} \log_2 \left|\frac{D_i}{D}\right| \quad (4)$$

## 2.4 Random Forest Algorithms

A Random Forest (RF) classifier is one of the methods utilised for classification and regression [18]. Random Forest is capable of interpreting as formed from a set of decision trees or decision trees as well. It is proficiency to make categorical predictions operating multiple possible values and adjustable output probabilities. One thing to watch out for is overfitting. Random Forests can be overfitting, especially when working with relatively small data sets. The advantage of the Random Forest algorithm is that it can classify data with incomplete attributes. Used for classification, but not very suitable for regression, more suitable for classifying data, and handling large sample data. Attached in Figure 1 using several stages of the Random Forest classifier algorithm, namely, importing telco customer churn data, Training data and Testing Data (Towards the Model), Performing data visualization, Predicting Random Forest Classifier, Model, Determining Accuracy Value, Variable Importance.

Based on the modelling process carried out in this study, the application of ADASYN and SMOTE Techniques was employed to overcome the imbalance of the dataset. The significant difference in the amount of data between classes resulted in the classification model being often unable to predict the minority class correctly so a lot of test data that should have been in the minority class was predicted wrongly by the classification model. The oversampling method modifies the distribution of data between the majority and minority classes in the training dataset to balance the amount of data for each class. This study then applies a classification technique using Random Forest and C45. The following is a flowchart of the flow in Figure 3, the focus of the research is more on solving problems and achieving research objectives..

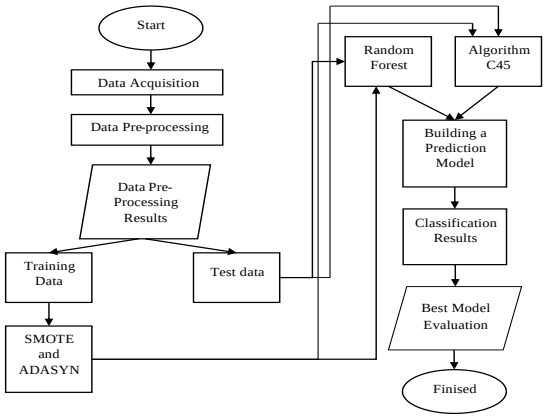


Figure 3. Flowchart Diagram

In the testing process, this study compared the implementation of Random Forest, ADSYN RF, SMOTE RF, C45, ADSYN C45 and SMOTE C45. Furthermore, performance evaluation in this study utilised the K-Fold Cross Validation technique. The workings of the K-Fold Cross Validation technique are to divide the data into test data and training data as much as K.

## 3. Result and Discussion

The results of the data preprocessing stage in this study were carried out by removing data with incomplete components to avoid further data entry manipulation since the data employed was more than 1000. This was done so the data was ideal to employ at a later stage. Furthermore, after data with incomplete components has been removed, normalization was carried out on data that has more than 4 categories. This was done to reduce repetition and map similar events. Shown in the pie chart Figure 4 explains that the heart disease data in 2020 is not balanced.

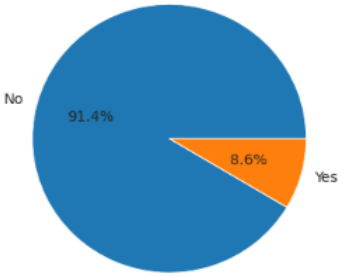


Figure 4. Heart Disease Pie Chart

Therefore, before establishing classification and modelling rules, it was necessary to divide the data into two groups, namely train and test. This data sharing aimed to analyze whether the classification rules generated by the Random Forest algorithm and the C45 algorithm were used to predict heart disease.

Since the target data for heart disease was not balanced, it was necessary to perform an oversampling technique with default parameters. Thus, the unbalanced data becomes balanced based on the oversampling process that has been carried out. Moreover, the K-Fold testing technique employed K-Fold 10 [19], [20]. The illustration of K-Fold is shown in Figure 5, where the prediction model begins by dividing all data into training data and test data with K-Fold cross-validation, and cross-testing of each - each algorithm. Performance evaluation is carried out on the model with the aim of knowing how well the model is performing using test data.

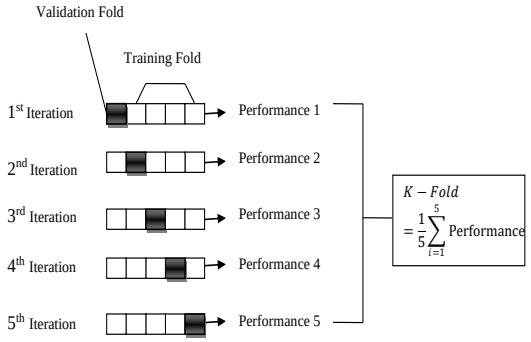


Figure 5. K-Fold Illustration

The evaluation for this study was based on accuracy, precision, sensitivity, and specificity. The evaluation process utilised the test data that has been separated in the previous process and the evaluation results employed the confusion matrix shown in Table2.

Table 1. K-Fold 10

| Algorithm | Accuracy K-Fold (%) |
|---|---|
| C45 | 86,74 |
| RANDOM FOREST | 90,56 |
| C45 + SMOTE | 91,77 |
| RF + SMOTE | 94,43 |
| C45 + ADASYN | 91,71 |
| RF + ADASYN | 94,34 |

The model generated from running Equation 1 to Equation 4 operated the Confusion Matrix on the C45 algorithm which produced an accuracy of 86.74%. On the other hand, the Random Forest Classifier algorithm had succeeded in producing a prediction model with a higher accuracy value than C4.5, which was 90.56%. Table 3 shown the results of the comparison of SMOTE and ADASYN.

Table 2 Comparison of SMOTE and ADASYN

| K-Fold Cross Validation | | |
|---|---|---|
| **Algorithm** | **SMOTE (%)** | **ADASYN (%)** |
| Random Forest | 94,43 | 94,34 |
| C45 | 91,77 | 91,71 |

By using the K-Fold calculation based on table 3, in this study the best accuracy value was obtained in the K-Fold 10 calculation for SMOTE applied to Random Forest 94.43% and ADASYN applied to Random Forest 94.34%. Thus, the combined technique of Random Forest with SMOTE and Random Forest with ADSYN had better performance than C45 with ADASYN and C45 with SMOTE. It was proven that the Random Forest algorithm with SMOTE has the best ability to predict class data compared to Random Forest with ADASYN with an accuracy of 94.43%.

## 4. Conclusion

Based on the results of the research that has been carried out, it was concluded that the SMOTE and ADASYN oversampling techniques had a significant impact on the classification results. It was proven that the increase in accuracy which occurred in the Random Forest and C.45 algorithms was quite significant. However, the highest accuracy was the combination of implementing SMOTE with Random Forest which reached 94.43%. The results of this study can be considered by experts to assist decisions in dealing with heart disease. Moreover, regarding further research, it is suggested to correlate a dashboard and visualization of the relationship between features which affect heart disease.

## References


[1] "Kementerian Kesehatan Republik Indonesia." .

[2] BHF, "UK Factsheet," *Br. Hear. Found.*, no. April, pp. 1–21, 2019.

[3] J. J. Pangaribuan, C. Tedja, and S. Wibowo, "PERBANDINGAN METODE ALGORITMA C4.5 DAN EXTREME LEARNING MACHINE UNTUK MENDIAGNOSIS PENYAKIT JANTUNG KORONER," 2019.

[4] A. Rohman and D. M. Rochcham, "MODEL ALGORITMA C4.5 UNTUK PREDIKSI PENYAKIT JANTUNG," 2018.

[5] N. Khasanah, R. Komarudin, N. Afni, Y. I. Maulana, and A. Salim, "Skin Cancer Classification Using Random Forest Algorithm," *Sisfotenika*, vol. 11, no. 2, p. 137, 2021, doi: 10.30700/jst.v11i2.1122.

[6] S. Ath *et al.*, "Jurnal Teknologi Terpadu HYBRID MACHINE LEARNING MODEL UNTUK MEMPREDIKSI PENYAKIT JANTUNG DENGAN METODE LOGISTIC REGRESSION DAN RANDOM," vol. 8, no. 1, pp. 40–46, 2022.

[7] I. M. El-Hasnony, O. M. Elzeki, A. Alshehri, and H. Salem, "Multi-Label Active Learning-Based Machine Learning Model for Heart Disease Prediction," *Sensors*, vol. 22, no. 3, 2022, doi: 10.3390/s22031184.

[8] S. Maldonado, J. López, and C. Vairetti, "An alternative SMOTE oversampling strategy for high-dimensional datasets," *Appl. Soft Comput. J.*, vol. 76, pp. 380–389, 2019, doi: 10.1016/j.asoc.2018.12.024.

[9] D. Elreedy and A. F. Atiya, "A Comprehensive Analysis of Synthetic Minority Oversampling Technique (SMOTE) for handling class imbalance," *Inf. Sci. (Ny).*, vol. 505, pp. 32–64, 2019, doi: 10.1016/j.ins.2019.07.070.

[10] R. . Nurdin, "Pernyataan Keaslian," *Digilib.Uin-Suka.Ac.Id*, no. April 2020, p. 506812, 2021.

[11] "CDC - 2020 BRFSS Survey Data and Documentation." .

[12] R. Siringoringo, "KLASIFIKASI DATA TIDAK SEIMBANG MENGGUNAKAN ALGORITMA SMOTE DAN k-NEAREST NEIGHBOR," 2018.

[13] G. Fico, J. Montalva, A. Medrano, N. Liappas, G. Cea, and M. T. Arredondo, "EMBEC & NBC 2017," *IFMBE Proc.*, vol. 65, pp. 1089–1090, 2018, doi: 10.1007/978-981-10-5122-7.

[14] S. Rahayu, T. Bharata Adji, N. Akhmad Setiawan, and D. Teknik Elektro dan Teknologi Informasi, "Penghitungan k-NN pada Adaptive Synthetic-Nominal (ADASYN-N) dan Adaptive Synthetic-kNN (ADASYN-kNN) untuk Data Nominal-Multi Kategori," *Ktrl.Inst (J.Auto.Ctrl.Inst)*, vol. 9, no. 2, p. 2017.

[15] W. Sullivan, *Machine Learning For Beginners Guide Algorithms*, vol. 4, no. 1. 2017.

[16] A. Cherfi, K. Nouira, and A. Ferchichi, "Very Fast C4.5 Decision Tree Algorithm," *Appl. Artif. Intell.*, vol. 32, no. 2, pp. 119–137, 2018, doi: 10.1080/08839514.2018.1447479.

[17] M. Kretowski, *Evolutionary Decision Trees in Large-Scale Data Mining*. 2019.

[18] A. Primajaya and B. N. Sari, "Random Forest Algorithm for Prediction of Precipitation," *Indones. J. Artif. Intell. Data Min.*, vol. 1, no. 1, p. 27, 2018, doi: 10.24014/ijaidm.v1i1.4903.

[19] T. Djatna, M. K. D. Hardhienata, and A. F. N. Masruriyah, "An intuitionistic fuzzy diagnosis analytics for stroke disease," *J. Big Data*, vol. 5, no. 1, 2018, doi: 10.1186/s40537-018-0142-7.

[20] S. Zitao, "3 min of Machine Learning: Cross Vaildation," *Zitao's Web*, 2020. .