Clasification of Heart Disease using Decision Tree Algorithm

A Ishak¹, A Ginting², K Siregar³, C Junika⁴

¹,³,⁴Industrial Engineering Department, Engineering Faculty, Universitas Sumatera Utara, Medan, Indonesia
²Mechanical Engineering Department, Engineering Faculty, Universitas Sumatera Utara, Medan, Indonesia

E-mail: christine.yunika@gmail.com aulia.ishak@usu.ac.id

Abstract. Data mining is the process of manipulating data by extracting information that was previously unknown from a large dataset. The Tree Model in relation to data structures is a data type that simulates a hierarchical tree structure with root values and sub-trees with the parent node represented as a series of recursively defined node links as a means of presenting complex data analysis. Every year an estimated 17.3 million people die from cardiovascular disease. As many as 7.3 million of them occur due to heart disease and 6.2 million due to stroke. Risk factors for heart disease are age, gender, heredity or genetics, smoking habits, lack of physical activity, obesity, diabetes mellitus, stress and diet. The purpose of the dataset in this study is to classify data to determine categories of heart disease based on the presence or absence of narrowing of the arteries.

1. Introduction

Data mining is the process of manipulating data by extracting information that was previously unknown from a large dataset [1]. In solving classification problems, the use of methods or techniques aims to simplify the classification process. Some techniques used in the case of classification are decision tree, Support Vector Machine, Ada Boost, Naïve Bayes, Constant, KNN, Tree, Random Forest, Stochastic Gradient Descent, and CN2 Rule.

The Tree Model in relation to data structures is a data type that simulates a hierarchical tree structure with root values and sub-trees with the parent node represented as a series of recursively defined node links as a means of presenting complex data analysis[2]. In its application to classification in this case, using a decision tree is a conditional control statement that uses a tree-like graph or decision model and is likely to have consequences for machine learning.

Coronary Heart Disease (CHD) is a disease that attacks the heart. The organ has the function of pumping blood throughout the body. Abnormalities in these organs can cause narrowing of the arteries that drain blood to the heart muscle, resulting in reduced supply of oxygen and nutrients to move the heart optimally. Narrowing of the arteries is caused by the deposition of calcium and yellow fatty deposits known as atherosclerosis. According to the World Health Organization (WHO), cardiovascular disease is a cause of death and disability throughout the world. Every year an estimated 17.3 million people die from cardiovascular disease. A total of 7.3 million of them occur due to heart disease and 6.2 million due to stroke[3]. Risk factors for heart disease are age, gender, heredity or genetics, smoking habits, lack of physical activity, obesity, diabetes mellitus, stress and diet (habits or eating patterns).
Dietary factors such as intake of monounsaturated fatty acids, water soluble fiber, complex carbohydrates and vegetarian diets will positively influence the increase in HDL cholesterol levels. These factors are thought to have an effect on cholesterol in the blood [4]. On the off chance that the consequences of QC tests can't satisfy the acknowledgment models, the aftereffects of examination of the entire arrangement of the estimations on that day must be eliminated or should be re-dissected, and an incomplete or full re-approval of the strategy considered [11].

Momentous advances in information technology results in the enormous growth of data in health care informatics [5]. Data mining appliance like Orange, Rapid miner and Weka [6][7][8] are used to predict and evaluate the better result for medical management data. The classification results of Decision Tree produce two rules of supplier selection model to classify suppliers so that companies can easily choose suppliers according to the criteria desired by the company [9, 12, 13]. The purpose of the dataset in this study is to classify data to determine categories of heart disease based on the presence or absence of narrowing of the arteries.

2. Methodology
The dataset used for this work is from UCI Machine Learning repository. The dataset consists of 13 main data, divided into 8 numerical data types (age, gender, rest SBP, cholesterol, max HR, ST by exercise and major vessels) and 5 categorical type data (narrowing diameter, gender, chest pain, fasting blood sugar, ECG rest, exercise, slope peak, major vessels colored).

The algorithm used in this study is to use the Decision Tree algorithm. Decision tree is a flow chart that is shaped like a tree structure where each internal node states the test of an attribute, each branch states the output of the test and the leaf node states the classes or class distributions. The top node is called the root node or the root node. A root node will have several exit edges but no entry edge, internal nodes will have one entry edge and several exit edges, while a leaf node will only have one entry edge without having an exit edge. Decision tree is used to classify a sample of data that is not yet known class into classes that already exist. The data testing path is first through the root node and the last is through the leaf node which will conclude the class prediction for the data. The data attribute must be in the form of categorical data, if it is continuous then the attribute must be discretized first [10].

3. Result and Conclusion
Implementation is done using one of the Data Mining software, Orange 3.25. All input indicator indicators and destination attributes are stored in xlsx format, then imported into Orange 3.25 software. The following below are the steps.

- Open the Orange Data Mining Software, the Welcome to Orange window will appear as follows.
• Click the New button and you are presented with a new worksheet.

![Figure 2. Display of new worksheets in orange software](image1)

• Click File on the data widget, a file icon will appear on the worksheet.

![Figure 3. Display icon file on the orange software worksheet](image2)

• Browse the dataset you want to use and connect to the data table by clicking the line next to the file icon then connecting to the side, select the data table to see the contents of the dataset.
Figure 4. Display of the data table icon on the orange software worksheet

- To eliminate missing values in the dataset, you can do it by preprocess by clicking the line next to the file icon and connecting to the side, then select preprocess.

Figure 5. Display the preprocess icon on the orange software worksheet

- Click twice on the preprocess icon then choose what method to eliminate the missing value here we choose average / most frequent.
• Click the line next to the Preprocess icon to see the results of the preprocess data.

• Click twice on the data table icon to see the results of preprocess data, it can be seen that the data has no missing value.
Figure 8. Display the results of preprocess data

- Click the line next to the result icon (data table) and connect to the side, select Tree as the learning method.

Figure 9. Display icon tree

- Click twice on the Tree icon to select the learning method parameter
Figure 10. Display parameter of decision tree

- Click the line next to the tree icon and connect sideways, select Tree viewer to display the results of the decision tree. Click twice on the Tree icon to select the learning method parameter.

Figure 11. Display icon tree viewer

- Click the line next to the tree icon and select column then connect sideways, then select test and score.
Figure 12. Display icon test and score

- Click the line next to the Test and Score icon then connect sideways, then the Confusion Matrix is selected to provide information comparing the classification results made by the system (model) with the actual classification results.

Figure 13. Display icon confusion matrix

Test and Score is done by the sampling method. Obtained Area Under ROC Curve (AUC) value of 77.1%, then Classification Accuracy (AC) of 77.6%, and Precision of 77.5%. In this case the decision tree method successfully classified 77 out of 100 sample data.
Figure 14. Display of test results and decision tree classification scores

The following is a picture of the results of data classification by the Decision Tree method can be seen in the image below.

Figure 15. Display of decision tree results

Acknowledgements
The author deeply grateful and acknowledges the support from Indonesian Ministry of Research, Technology and Higher Education especially The Directorate of Research and Community

References
[1] Vijayakumar V and Nedunchezhian R 2012 A Study on video data mining International Journal of Multimedia Information Retrieval
[2] Klas W and Schrefl M 1995 Data model tailoring and database integration Springer
[3] WHO 2013 About Cardiovascular diseases World Health Organization

[4] Almatsier S 2004 Prinsip Dasar Ilmu Gizi Development (Jakarta:Gramedia Pustaka Utama)

[5] Kavakiotis I, Tsave O, Salifoglou A, Maglaveras N, Vlahavas I and Chouvarda I 2017 Machine learning and data mining methods in diabetes research *Computational and structural biotechnology journal* **15** 104-116.

[6] Kiruthika Devi S, S Krishnapriya and Dristipona Kalita 2016 Prediction of Heart Disease using Data Mining Techniques *Indian Journal of Science and Technology* pp 1-5

[7] Usha Rani D 2017 Survey on Data Mining Tools and Techniques in Medical Field *International Journal of Advanced Networking &Applications*

[8] Patil PH, Thube S, Ratnaparkhi B and Rajeswari K 2014 Analysis of Different Data Mining Tools using Classification, Clustering and Association Rule Mining *International Journal of Computer Applications* pp 35-39

[9] Aulia Ishak and Tommy Wijaya 2020 *IOP Conf. Ser.: Mater. Sci. Eng.* **801** 012118

[10] Akkaş E, Akin L, Evren Çubukçu H and Artuner H 2015 Application of Decision Tree Algorithm for classification and identification of natural minerals using SEM-EDS *Comput. Geosci* pp 38-48

[11] Indrayanto G 2018 Recent development of quality control methods for herbal derived drug preparations *Natural Product Communications*, **13** (12)

[12] Aulia Ishak et al 2019 *IOP Conf. Ser.: Mater. Sci. Eng.* **505** 012085

[13] Aulia Ishak and Tommy Wijaya 2020 *IOP Conf. Ser.: Mater. Sci. Eng.* **801** 012122