DECISION TREE BASED DATA MODELLING FOR FIRST DETECTION OF THALASSEMIA MAJOR

Yohanes Setiawan[1], Oktavia Ayu Permata[2]*, M. Pradata Yuda[3]
Department of Information Technology, School of Computing[1], [2], [3],
Telkom University, Surabaya Campus,
Jl. Ketintang no. 156, Surabaya, East Java, Indonesia
yohanessetiawan@telkomuniversity.ac.id[1], oktapermata@telkomuniversity.ac.id[2], m.pradata@student.ittelkom-sby.ac.id[3]

***Abstract*— Thalassemia is an inherited blood disease which lacks hemoglobin, the protein that is carrying oxygen to the body. The severe one is called Thalassemia Major which needs special care about blood transfusion. Several studies carried out early detection of thalassemia based on neural networks and the formation of fuzzy rules. However, neural networks require large amounts of data and the formation of fuzzy rules takes a long time to research. This research aims to create a model based on decision tree for first detection of Thalassemia Major. The dataset is obtained by interview of Thalassemia symptoms and primary data of medical records from a hospital. Classical decision tree models used are ID3, C4.5 and CART. The models are evaluated by Train-Test Split consists of 70% training and 30% testing data and k-Fold Validation for checking model's overfitting or underfitting. The output of this research is a final tree model from the best performance of decision tree models. The final result shows that C4.5 has the best performance with accuracy 100% and not overfitting or underfitting. Also, C4.5 performs feature selections to its tree modeling to simplify the inference. In brief, decision tree based modeling is effective to be used as first detection of Thalassemia Major by interview symptoms with generating automatic rules from its tree model.**

Keywords—Data Mining, Decision Tree, Thalassemia Major

I. INTRODUCTION

Thalassemia is a blood disorder inherited from parents characterized by a lack of hemoglobin which is the protein that carries human blood to the body's organs [1]. Thalassemia is recorded as having a high proportion of chronic diseases in developing countries, one of which is Indonesia [2]. This disease causes the sufferers to experience anemia and acute fatigue. One type of severe thalassemia is Thalassemia Major. Thalassemia Major sufferers require blood transfusions throughout their lives due to the severe anemia they experience. In general, the diagnosis is made using a Complete Blood Count (CBC) and requires a follow-up Medical Check Up to see the presence of thalassemia. This makes the diagnosis has been difficult for people to carry out early detection, even though Thalassemia Major can cause other deadly chronic diseases, which developing countries have contributed to 300,000 to 400,000 deaths of newborn babies due to Thalassemia Major. [3]. Therefore, it is necessary to carry out initial detection before continuing with the further testing process. Various patient data processing techniques have been used for initial diagnosis of Thalassemia. [2] conducted a literature study in the form of a survey regarding the use of Machine Learning for the detection and classification of Thalassemia. Several popular algorithms are used, such as Decision Tree, k-Nearest Neighbor, and Artificial Neural Network. [4] also made a comprehensive review regarding various implementations of Artificial Intelligence in the diagnosis of Thalassemia. The difference lies in the dataset used in the form of classified patient CBC data. [5] using image processing of erythrocytes in blood to classify Thalassemia. Convolutional Neural Network and Multi Layer Perceptron are used as detection tools. Overall, these studies require relatively large amounts of data. [3] implementing fuzzy logic to detect Thalassemia so that it does not require large amounts of data. However, the formation of membership functions obtained through a long interview process with medical personnel accompanied by the establishment of rules took a long time to develop. So, we need a method that can automate these rules so that we only need to learn from patient data.Then, [2] conducted research survey in detecting Thalassemia Major patients using popular machine learning approaches, such as Decision Tree, Naïve Bayes, Support Vector Machine, and Neural Network.

The use of Decision Trees can shorten the formation of rules through the formation of trees that lead to decisions/classification results directly. The three conventional methods of Decision Trees are ID3, C4.5 and CART. [6] utilize ID3 to identify children with special needs with binary questions (Yes/No). Something similar was also done by [7] and [8] in applying ID3 to classify typhoid fever and stroke in the medical world. C4.5 is a modification of ID3 that utilizes information entropy [9]. [10] use C4.5 to predict the procurement of office machine equipment so that it is effective according to conditions and needs. Then, [11] implement C4.5 for customer satisfaction. The dataset used is still categorical data for each column. However, the popularity of ID3 and C4.5 is not comparable to Classification and Regression Tree (CART). CART is implemented in the Scikit-Learn library

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v13i1.1949, Copyright ©2024
Submitted : September 8, 2023, Revised : January 20, 2024, Accepted : January 25, 2024, Published : February 15, 2024

which is easily accessible via the Python programming language [12]. The implementation of CART can be broader than categorical data, namely it can be used for continuous data as well. [13] applied CART to recommend majors for high school students based on grades in the subject areas taught. In the health sector, [13] compared between CART and C4.5 for breast cancer detection. Based on experimental results, C4.5 is superior to CART even with an accuracy difference of 1%. However, [14] succeeded in classifying diabetes using CART with 100% accuracy. On [15], The decision tree from CART can be used as a reference for classification without having to carry out a deployment process into systems such as websites or mobile. The research above has generated confidence for researchers to utilize Decision Tree-based algorithms in solving problems.

This research models a Decision Tree for early detection of Thalassemia Major. Decision Tree-based modeling was chosen because it is easy to interpret and carries out automatic feature selection in the learning process on the data. The main contribution of this research is developing a tree model for early detection of Thalassemia Major using Decision Tree based modelling. The algorithms used are ID3, C4.5, and CART. Then, a metric evaluation will be carried out between the three models and the model with the best metrics will be selected. Next, a decision tree model is created using the best model that has been selected, making it easier to interpret symptoms and early detection results of the disease suffered by Thalassemia patients

II. METHODOLOGY

In this section, the Thalassemia Major detection research method used is explained. This research uses CRISP-DM (The Cross Industry Standard Process for Data Mining) as a standard reference for systematic Data Mining methodology that is used to extract and recognize patterns contained in information through solving algorithms [16]. The evaluation metrics used in this research are also explained in more detail along with an understanding of their similarities. The complete research method is visualized in the flow diagram in Fig. 2.

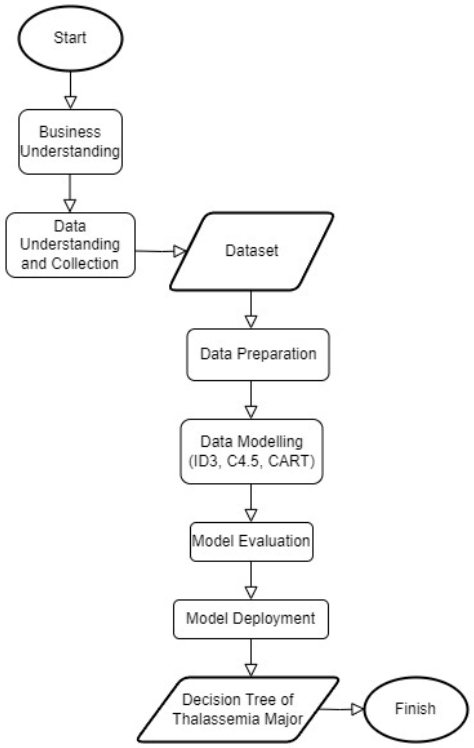

Figure 2. Flowchart of Research Methodology

A. *Business Understanding*

This research begins with the required business understanding. The business referred to in the context of this research is understanding domain knowledge related to Thalassemia Major. Related literature studies and medical consultations were carried out to obtain an overview of the symptoms of Thalassemia Major and non-Major. Medical consultations are carried out through interviews with specialist internal medicine doctors.

B. *Data Understanding and Collection*

Next, the data collection process was carried out at a private hospital in Surabaya, East Java, Indonesia. Data collection was carried out by recording data anonymously through the patient's medical record data with the help of symptoms obtained from the list of symptoms resulting from previous interviews. In medical records containing the listed symptoms, the anonymous copy for research data will be written as "Yes". Meanwhile, if it does not contain symptoms, it will be written as "No". Obtained 30 medical record data from Thalassemia patients, consisting of Thalassemia Major and non-Major.

C. *Data Preparation*

Before the data enters the model, the data must go through a pre-processing process first. The data preparation stage in this research includes pre-processing of data that has been taken and understood in the previous process. Checking for missing values and duplicate data as well as changing categorical Yes/No data into numeric (encoding) needs to be done so that it can be processed by the Decision Tree algorithm. Data division is carried out with details of 70% training data and 30%

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v13i1.1949, Copyright ©2024
Submitted : September 8, 2023, Revised : January 20, 2024, Accepted : January 25, 2024, Published : February 15, 2024

test data so that the training data is not too much (enough) and the test data is not too little if the data size is not large.

D. Data Modelling (ID3, C4.5, CART)

After going through the pre-processing stage, modeling is carried out based on Decision Tree. Decision Tree-based methods are data mining algorithms that have a structure similar to a tree [8]. Decision Tree based modelling has been known for its best to handle categorical features than other machine learning algorithms, which is impotant to train Yes/No questions to create a simple tree model in detecting Thalassemia Major earlier through tree visualization. The tree data structure in a Decision Tree consists of internal nodes, branches, and leaves. Internal nodes perform tests on an attribute, the results of which are represented by branches and the classes are displayed via leaves. There is also a root node which is the starting node (top node) of the decision tree model. Three popular Decision Tree models include ID3, C4.5, and CART. This research will compare the performance of the three and choose one model to deploy. An example image of a decision tree is visualized in Fig 1.

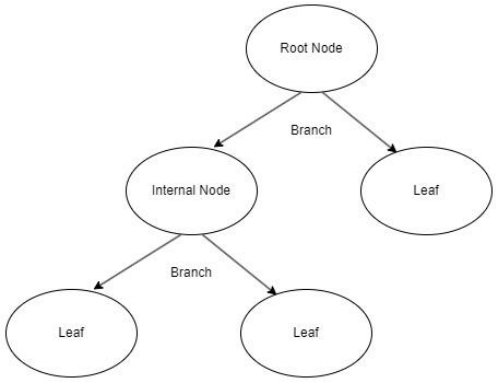

Figure 1. General Structure of Decision Tree

ID3, short for Iterative Dichotomize 3, is a type of decision tree that works on discrete classes and cannot be used on missing data. [17]. ID3 is the basic algorithm that has been developed on the C4.5 and CART methods. This algorithm is a standard decision tree which is formed by using information directly to carry out the selection process for the required features [18]. The formation of a decision tree with ID3 is carried out comprehensively on all possible trees from top to bottom. First of all, form a root node through the selected attribute and calculate the Entropy value. Entropy is a metric used to measure uncertainty in a dataset S. The entropy equation can be seen in Eq. (1).

$$Entropi(S) = -\sum_{i=1}^{n} p_i * \log_2 p_i \quad (1)$$

where n represents the number of partitions in the dataset and p_i represents the proportion of S_i towards S. Then, the Information Gain (IG) calculation is carried out for each existing feature. IG calculations are based on the calculated entropy of the dataset. The IG equation can be seen in Eq. (2).

$$IG(S,a) = Entropi(S) - \left(\sum_{i=1}^{n} \left|\frac{S_i}{S}\right| * Entropi(S_i)\right) \quad (2)$$

where a states the attributes/features contained in the dataset, S_i is the number of data samples and S is the total amount of data. Next, a branch is formed for each value. These processes are repeated on each branch with the same class until a decision tree is formed. The need for calculating entropy and IG information in the ID3 algorithm is needed for feature selection based on the largest IG value as a splitter of the attributes [19].

Then, C4.5 is a modification of ID3 by expanding the scope of ID3 which only supports categorical data to support continuous data and missing values [9]. Basically, C4.5 works the same way as ID3. The basic difference lies in the selection of features/attributes which are calculated using the Gain Ratio (GR) given in Equation (3).

$$GR(S,a) = \frac{IG(S,a)}{SI(S,a)} \quad (3)$$

where $IG(S,a)$ is the Information Gain contained in ID3 in Equation (2), and $SI(S,a)$ states Split Info which is the potential information entropy expressed in Equation (4).

$$SI(S,a) = -\sum_{i=1}^{n} \frac{S_i}{S} \log_2 \frac{S_i}{S} \quad (4)$$

The feature with the highest GR value will be selected as the test feature for the node in the Decision Tree. In contrast to ID3 which uses IG as a feature selection metric. Through differences in feature selection compared to ID3, C4.5 can handle missing values, process discrete (categorical) and continuous (numerical) data, and produce interpretable branching decision tree rules. In addition, C4.5 can also overcome model overfitting, algorithm computational efficiency, and has a high success rate on balanced datasets [20].

Lastly, CART is an abbreviation for Classification and Regression Trees. It is called CART because if the target or dependent variable is continuous, the resulting tree is a regression tree. Meanwhile, if the target is categorical, the resulting tree is a classification tree [21]. The advantage of CART compared to ID3 and C4.5 is the construction of a tree with only two children per node [17]. Establishing a CART tree is done through the process of collecting datasets at the root node of the tree with each node divided into two child nodes with a separation variable which is one of the predictor variables [22]. There are three steps in implementing CART [15]. First, tree formation. All candidate branches of the predictor variable are given, and the split attribute selection is calculated using the impurity value formulated in the Gini Index. Gini Index formulation on attributes a for data with class c and probability p a row of data that has a class attribute K_i and calculated by dividing the number of class attributes K_i on the data on the total number of data rows can be seen in Equation (5).

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v13i1.1949, Copyright ©2024
Submitted : September 8, 2023, Revised : January 20, 2024, Accepted : January 25, 2024, Published : February 15, 2024

$$Gini(a) = 1 - \sum_{i=1}^{c} p_i^2 \quad (5)$$

The data is divided into two by determining the lowest Gini Index value. Gini Index as a result of sorting data into two subsets, for example a_1 and a_2, shown in Equation (6).

$$Gini = \left|\frac{a_1}{a}\right| Gini(a_1) + \left|\frac{a_2}{a}\right| Gini(a_2) \quad (6)$$

The three methods are compared with the same proportion and distribution of training and test data. So that the algorithm with the best performance will be selected for the model deployment stage.

E. Model Evaluation

Model evaluation will be carried out in two ways, namely through metrics on the division of training and test data and based on K-Fold Validation. The training and test data division metrics use Accuracy, Precision, Recall, and F1-Score which are related to the Confusion Matrix concept in the model. The Confusion Matrix has evaluation components such as True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN). The similarities of the Accuracy, Precision, Recall, and F1-Score metrics can be seen in Equations (7), (8), (9), and (10), respectively.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (7)$$

$$Precision = \frac{TP}{TP + FP} \quad (8)$$

$$Recall = \frac{TP}{TP + FN} \quad (9)$$

$$F1 - Score = \frac{2 * Precision * Recall}{Precision + Recall} \quad (10)$$

In K-Fold Validation based evaluation, we will see whether the model is overfitting/underfitting or not so that the model is ready to be deployed using the Mean Absolute Error (MAE) error metric of n data which is the average of the absolute reduction between each actual data x_{akt_i} with prediction data x_{pred_i} which is stated in Equation (11).

$$MAE = \frac{1}{n} \sum_{i=1}^{n} \left| x_{akt_i} - x_{pred_i} \right| \quad (11)$$

F. Model Deployment

The model deployment process is the process of forming a model that has been trained into a model that is ready to be used to predict new data. The model with the best evaluation will be deployed through the decision tree visualization results. The results of decision tree visualization from the optimal model can provide automatic rules regarding early detection of Thalassemia Major without having to create complicated computer programs on certain platforms.

III. RESULTS AND DISCUSSIONS

Interviews with internal medicine specialists produced 15 (fifteen) question-type symptoms with Yes/No answers which were used to carry out initial detection whether they were related to the symptoms of Thalassemia Major or not. A list of symptoms is shown in Table 1.

TABLE I. LIST OF SYMPTOMS RESULTING FROM THE INTERVIEW

No	Symptom
1	Facial changes
2	Blue toe nails
3	Black stools
4	Anemia
5	Yellowing of the skin
6	Pale
7	Frequent headaches
8	Frequent falls
9	Unintended weight loss
10	Excessive fatigue
11	Abdominal bloating
12	Body pain
13	Fever
14	Weakness
15	Difficulty breathing

Then, 30 lines of data were obtained from 30 Thalassemia patients taken from the hospital. The proportion of Thalassemia Major and non-Thalassemia Major patients is shown in Figure 3. It can be seen that the proportion is 60:40 (60% Thalassemia Major and 40% non-Thalassemia Major). There is an imbalance class, but it is not severe, so treatment is carried out by focusing on metrics other than accuracy, namely the Precision, Recall and F1-Score metrics to see the reliability of the model for detecting positive (detected as Thalassemia Major) and negative (detected as not Thalassemia Major).

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v13i1.1949, Copyright ©2024
Submitted : September 8, 2023, Revised : January 20, 2024, Accepted : January 25, 2024, Published : February 15, 2024

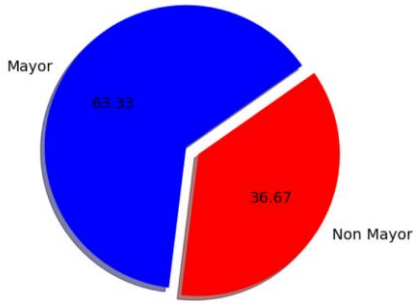

Figure 3. Pie Chart of the Proportion of Thalassemia Major and Non-Thalassemia Major

The Decision Tree Model with the ID3, C4.5, and CART algorithms is implemented using the Python programming language. The metric results of Accuracy, Precision, Recall, and F1-Score from the model are shown in Fig. 4. It can be seen that all models produce perfect metrics, namely 100%. In this case, perfect Accuracy inevitably results in perfect other metrics (Precision, Recall, and F1-Score) as well. This shows that the model captures patterns from the available dataset very well, which should later be re-evaluated further regarding its reliability. The model can differentiate Thalassemia Major from non-Major perfectly from testing results taken from unseen data, namely a set of data that has never been "seen" by the model through the training process.

```
Classification report ID3 Classifier :
              precision    recall  f1-score

           0       1.00      1.00      1.00
           1       1.00      1.00      1.00

    accuracy                           1.00


Classification report C4.5 Classifier :
              precision    recall  f1-score

       Mayor       1.00      1.00      1.00
   Non Mayor       1.00      1.00      1.00

    accuracy                           1.00


Classification report CART Classifier :
              precision    recall  f1-score

           0       1.00      1.00      1.00
           1       1.00      1.00      1.00

    accuracy                           1.00
```

Figure 4. Evaluation Metric Results from ID3, C4.5 and CART

After knowing the accuracy of the testing data, the first stage of evaluation is then carried out through a comparison between these metrics and the accuracy of the testing data. If the accuracy of the training data is greater than the testing data, then the model experiences overfitting, which means the model fails to learn new data, resulting in low accuracy on the new data. If the accuracy of the testing data is close to the testing data with a difference of only 2-3%, then the model can be said to be good for use (an ideal model). Then, if the accuracy of the training data is low, then the model experiences underfitting, which means the model fails to learn patterns in the data. The comparison results between the accuracy of testing data and training data are shown in Table 2. Based on Table 2, all algorithms have perfect accuracy. This means that a simple decision tree model can overcome the problem of categorical data in cases of early detection of Thalassemia Major. So that the model does not experience overfitting. However, it is also necessary to check with the K-Fold Validation method, which can be used to select the best model from the three available algorithms.

TABLE II. ACCURACY METRIC OF TRAINNG AND TESTING DATA

Algorithm	Training Accuracy	Testing Accuracy
ID3	100%	100%
C4.5	100%	100%
CART	100%	100%

The second stage of evaluation is carried out using K-Fold Validation so that overfitting/underfitting of the model can be seen more clearly through the errors. This research uses MAE as the error metric of the model. In this case, the MAE value will get better as the value gets closer to zero. Parameter k on K-Fold Validation using $k = 10$, which means the model is divided into 10 consecutive parts because these values do not have high bias or high variance. Then each part will take turns being training and testing data so that the error stability of the model can be seen. A good fit model is a model which during testing experiences a decrease in errors. Because if there is an increase in error when alternating between training and testing data, it is an indication that the model cannot learn patterns well, so the model becomes overfitting. The same thing can also happen with training errors. If the training error is higher when changing training and testing data, then the model is not successful in recognizing patterns when trained, causing underfitting.

The result of k-Fold Validation for $k = 10$ shown in Fig. 5. The ID3 and CART model graphs have similar performance, namely having the highest MAE of more than 0.3 at the 10th fold. When the average error of the testing data is higher than the training data, the model is experiencing overfitting. This will result in the model being unable to detect the presence of new data later even though it had high accuracy previously. Meanwhile, C4.5 shows its best performance with an error of zero, which means there is no difference between the predicted data and the ground truth. The C4.5 model has good stability because the error is not high in the first fold and continues with an error equal to zero in the second to fifth fold, which shows that the model can provide accurate prediction results even though the training and testing data are used interchangeably. on K-Fold Validation. Therefore, the C4.5 model was chosen

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v13i1.1949, Copyright ©2024
Submitted : September 8, 2023, Revised : January 20, 2024, Accepted : January 25, 2024, Published : February 15, 2024

as the best Decision Tree model and a deployment process will be carried out to form the tree.

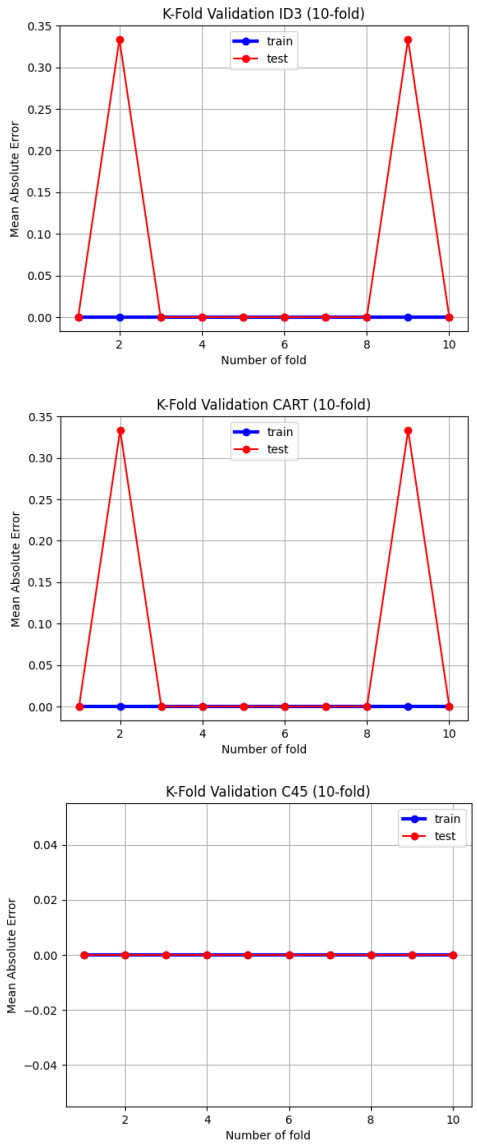

Figure 5 Results of K-Fold Validation for $k = 10$

Furthermore, C4.5 has been evaluated by performing the confusion matrix. Fig 6 has shown the confusion matrix between training and testing dataset. The confusion matrix shows the best performance from C4.5 with no mistake in predicting between two classes. For training data, 13 major Thalassemia and 8 non major Thalassemia have been detected correctly. Also, 6 major Thalassemia and 3 non major Thalassemia have been detected correctly.

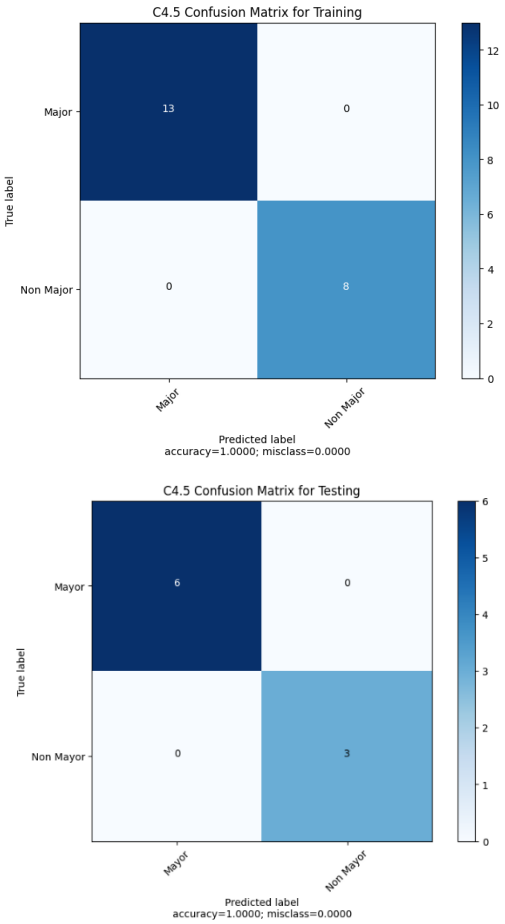

Figure 6 Confusion Matrix for Training and Testing

The selected C4.5 model is deployed through the formation of a decision tree which can form rules that are directly connected to its class. The resulting decision tree is as shown in Fig. 7. Fig. 7 shows that the C4.5 model does not use all 15 features to achieve a prediction that the patient has Thalassemia Major or not. This means that the C4.5 model performs feature selection automatically through the Gain Ratio calculation process. Important features, or what can be further referred to as feature importance, are changes in facial shape, black defecation, and anemia. These three features are what Thalassemia sufferers need to pay attention to when making suspicions about Thalassemia Major or not. When a patient experience one of these symptoms, the resulting prediction is always Thalassemia Major, which is a fatal type of Thalassemia that requires regular blood transfusions throughout his life. Through the C4.5 algorithm in Decision Trees, rules can be formed automatically like this:

IF Facial changes

OR Black stools

OR Anemia

THEN Thalassemia Major

ELSE Non-Thalassemia Major

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v13i1.1949, Copyright ©2024
Submitted : September 8, 2023, Revised : January 20, 2024, Accepted : January 25, 2024, Published : February 15, 2024

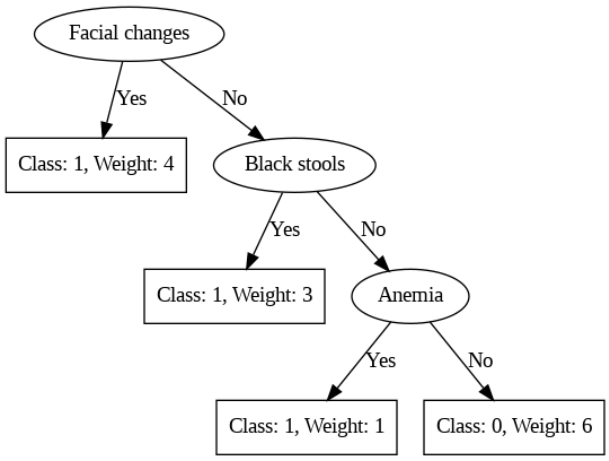

Figure 7. Decision Tree Model of Thalassemia Major Detection

One of the limitations of this conducted research is the relatively small sample size dataset, which may restrict the generalizability of the findings to a larger population. Furthermore, this model only considers Yes/No questions without additional information, i.e. image laboratory dataset. These limitations can be another opportunity for the next research project about Thalassemia.

IV. CONCLUSION

This research uses Decision Trees to carry out early detection of Thalassemia Major. The types of decision trees used are ID3, C4.5, and CART because the models are simple and can be interpreted into simple rules that can be directly implemented as early predictions of Thalassemia Major. Datasets that have gone through pre-processing are subjected to a modeling process through evaluation using two approaches, namely the division of training and testing and K-Fold Validation. In dividing training and testing, 70% of the data is divided as training data and 30% of the data as test data (unseen data). The evaluation metrics used are Accuracy, Precision, Recall, and F1-Score. The accuracy obtained in training and testing for the three methods was 100% so that the other metrics were also perfect. However, through K-Fold Validation it was discovered that the ID3 and CART models experienced overfitting and C4.5 was an optimal model with very good error stability. So C4.5 was chosen as the optimal model for tree development from Thalassemia Major. By reviewing the C4.5 tree, it is known that there are three features that are important for early detection of Thalassemia Major, namely changes in facial shape, black defecation, and anemia. It can be concluded that the decision tree model is very suitable for datasets with simple Yes/No questions accompanied by automatic feature selection for the efficiency of building rules that are shorter and more precise in drawing conclusions without having to carry out a long interview process to obtain the rules. For further work, larger dataset with other types Thalassemia classification can be developed to provide early detection of Thalassemia patients. Also, future researchers can use multimodal dataset to support better decision making, i.e. tabular and image.

REFERENCES

[1] D. Kristanty, D. Diyah, K. Rediyanto, and M. Si, "Analisis Polimorfieme Gen CYP pada Metabolisme Obat Deteksi Dini Thalassemia," *Pratista Patologi*, vol. 8, no. 1, pp. 17–28, Jan. 2023.

[2] M. Q. Mohammed and J. M. Al-Tuwaijari, "A Survey on various Machine Learning Approaches for thalassemia detection and classification," *Turkish Journal of Computer and Mathematics Education*, vol. 12, no. 13, pp. 7866–7871, 2021, doi: 10.1182/BLOODADVANCES.2020002725.

[3] E. R. Susanto, A. Syarif, K. Muludi, R. R. W. Perdani, and A. Wantoro, "Implementation of Fuzzy-based Model for Prediction of Thalassemia Diseases," in *Journal of Physics: Conference Series*, IOP Publishing Ltd, Jan. 2021. doi: 10.1088/1742-6596/1751/1/012034.

[4] K. Ferih *et al.*, "Applications of Artificial Intelligence in Thalassemia: A Comprehensive Review," *Diagnostics*, vol. 13, no. 9, May 2023, doi: 10.3390/diagnostics13091551.

[5] Z. N. Nugroho, A. Harjoko, and M. Auzan, "Klasifikasi Eritrosit Pada Thalasemia Minor Menggunakan Fitur Konvolusi dan Multi-Layer Perceptron," *IJEIS (Indonesian Journal of Electronics and Instrumentation Systems)*, vol. 13, no. 1, Apr. 2023, doi: 10.22146/ijeis.83473.

[6] F. Hafidh, M. Arsyad Al Banjari Banjarmasin, and S. Kalimantan Indonesia, "Enhancing Special Needs Identification for Children: A Comparative Study on Classification Methods Using ID3 Algorithm and Alternative Approaches," *Journal of Engineering, Electrical and Informatics*, vol. 3, no. 2, 2023, doi: 10.55606/jeei.v3i1.1468.

[7] U. Muhammadiyah Jember, M. Yogi Firmansyah, and D. Lusiana Pater, "Penerapan Algoritma Itterative Dechotomiser 3 (ID3) Untuk Klasifikasi Penyakit Tifoid Application of Itterative Dechotomiser 3 (ID3) Algorithm for Typhoid Disease Classification," 2023. [Online]. Available: http://jurnal.unmuhjember.ac.id/index.php/JST

[8] Z. Sitorus and A. Widarma, "Data Mining Algoritma Decision Tree Itterative Dechotomiser 3 (ID3) untuk Klasifikasi Penyakit Stroke Data Mining Algorithm Decision Tree Itterative Dechotomiser 3 (ID3) for Classification of Stroke," *Journal of Computing Engineering, System and Science)*, vol. 8, no. 2, pp. 554–563, 2023, [Online]. Available: www.jurnal.unimed.ac.id

[9] M. Tohir, D. Andariya Ningsih, N. Yuli Susanti, A. Umiyah, and L. Fitria, "Comparison of the Performance Results of C4.5 and Random Forest Algorithm in Data Mining to Predict Childbirth Process," 2023.

[10] A. Iftitah and R. Setyadi, "Penerapan Algoritma C.45 Untuk Analisis Pengadaan Peralatan dan Mesin Kantor," *Journal of Information System Research (JOSH)*, vol. 4, no. 2, pp. 434–442, Jan. 2023, doi: 10.47065/josh.v4i2.2673.

[11] J. Prayoga, Z. Gustiana, and S. A. Rahmah, "Applying Data Mining to Classify Customer Satisfaction using C4.5 Algorithm Decision Tree," *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, vol. 17, no. 2, Apr. 2023, doi: 10.22146/ijccs.83535.

[12] R. Nicholas Reinaldo and S. Dwiasnati, "Prediction of Customer Data Classification by Company Category Using Decision Tree Algorithm (Case Study: PT. Teknik Kreasi Solusindo)," *International Journal of Advanced Multidisciplinary*, vol. 2, no. 2, Jul. 2023, doi: 10.38035/ijam.v2i2.

[13] F. Melani and Sulastri, "Analisis Perbandingan Klasifikasi Algoritma CART dengan Algoritma C4.5 Pada Kasus Penderita Kanker Payudara," *Jurnal TEKNO KOMPAK*, vol. 17, no. 1, pp. 171–183, 2023, [Online]. Available: https://www.kaggle.com/datasets/reihanenamdari/breast-cancer.

[14] F. Maisa Hana, W. Cholid Wahyudin, S. Ulya, and D. Setia Negara, "Implementasi Algoritma CART dalam Klasifikasi Penyakit Diabetes," *Jurnal Ilmu Komputer dan Matematika*, pp. 1–8, 2023.

[15] A. Jananto, S. Sulastri, E. Nur Wahyudi, and S. Sunardi, "Data Induk Mahasiswa sebagai Prediktor Ketepatan Waktu Lulus Menggunakan Algoritma CART Klasifikasi Data Mining," *Jurnal Sisfokom (Sistem Informasi dan Komputer)*, vol. 10, no. 1, pp. 71–78, Feb. 2021, doi: 10.32736/sisfokom.v10i1.991.

[16] Y. Setiawan, "Data Mining berbasis Nearest Neighbor dan Seleksi Fitur untuk Deteksi Kanker Payudara," *Jurnal Pengembangan IT (JPIT)*, vol. 8, no. 2, pp. 89–96, Jul. 2023.

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v13i1.1949, Copyright ©2024
Submitted : September 8, 2023, Revised : January 20, 2024, Accepted : January 25, 2024, Published : February 15, 2024

[17] M. Hamdi, I. Hilali-Jaghdam, B. E. Elnaim, and A. A. Elhag, "Forecasting and classification of new cases of COVID 19 before vaccination using decision trees and Gaussian mixture model," *Alexandria Engineering Journal*, vol. 62, pp. 327–333, Jan. 2023, doi: 10.1016/j.aej.2022.07.011.

[18] A. Agarwal, K. Jain, and R. K. Yadav, "A mathematical model based on modified ID3 algorithm for healthcare diagnostics model," *International Journal of System Assurance Engineering and Management*, 2023, doi: 10.1007/s13198-023-02086-w.

[19] C. Liu, J. Lai, B. Lin, and D. Miao, "An improved ID3 algorithm based on variable precision neighborhood rough sets," *Applied Intelligence*, Jul. 2023, doi: 10.1007/s10489-023-04779-y.

[20] A. Sunarto, P. N. Kencana, B. Munadjat, I. K. Dewi, A. Z. Abidin, and R. Rahim, "Application of Boosting Technique with C4.5 Algorithm to Reduce the Classification Error Rate in Online Shoppers Purchasing Intention," *J Wirel Mob Netw Ubiquitous Comput Dependable Appl*, vol. 14, no. 2, pp. 01–11, Jun. 2023, doi: 10.58346/JOWUA.2023.I2.001.

[21] A. S. R. Siregar, Y. S. Siregar, and M. Khairani, "Implementation Of The Data Mining Cart Algorithm In The Characteristic Pattern Of New Student Admissions," *Journal of Computer Networks, Architecture and High Performance Computing*, vol. 5, no. 1, pp. 263–275, Feb. 2023, doi: 10.47709/cnahpc.v5i1.1975.

[22] S. Monalisa and F. Hadi, "Penerapan Algoritma CART Dalam Menentukan Jurusan Siswa di MAN 1 Inhil," *Jurnal Sisfokom (Sistem Informasi dan Komputer)*, vol. 9, no. 3, pp. 387–394, Oct. 2020, doi: 10.32736/sisfokom.v9i3.932.

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v13i1.1949, Copyright ©2024
Submitted : September 8, 2023, Revised : January 20, 2024, Accepted : January 25, 2024, Published : February 15, 2024