Early Detection of Type II Diabetes Mellitus with Random Forest and Classification and Regression Tree (CART)

Mira Kania Sabariah, MT
Informatic Engineering, School of Computing, Telkom University
Bandung, Indonesia
mira_ljuan@yahoo.com

Aini Hanifa, ST
Informatics Engineering, School of Electrical Engineering and Informatics, ITB, Bandung, Indonesia
aini_hanifa@ymail.com

Siti Sa'adah, MT
Informatic Engineering, School of Computing, Telkom University
Bandung, Indonesia
tisataz@gmail.com

Abstract— **Diabetes Mellitus (DM) is the third deadliest disease in Indonesia, and type II DM is more dangerous because it is caused by the combination between genetic and lifestyle factors. The high rate of patients infected with type II DM is caused by late diagnosis, therefore, early detection of disease is necessary to classify the detected patients with type II diabetes mellitus, and undetected patients. Moreover, analyzing the determinant and major attributes are highly recommended. In this research is implemented the combined Classification methods between Regression Tree method (CART) and Random Forest (RF) to build the classification model that is used in the early detection of diabetes mellitus type II disease. Those methods are selected based on the characteristics of the dataset used in medical records that consist of complex attributes consisting of several categorical attributes and continuous attributes, besides the advantages of the CART models are easy to implement, and it can explore the structure of complex medical records, while the RF method can handle the problem in accuracy. This research has tested a different number of trees and numbers of candidate attributes splitter. Based on the test results, it shows that the addition of trees and attributes splitter can improve the accuracy and reduce the error rate, with the optimal inputs are 50 numbers of trees and 3 number of attributes splitter with 83,8% average accuracy. The important attribute of early detection of diabetes mellitus type II is heredity, age, and body mass index.**

Keywords— Diabetes Mellitus type II, CART, Random Forest, classification, the early detection of disease, the important attribute

I. INTRODUCTION

Diabetes Mellitus (DM) is the third deadliest disease in Indonesia. There are two common types of diabetes, namely diabetes type I and type II diabetes. Type I diabetes is unpredictable and hard to prevent, because of the genetic disorder from birth. In contrast, type II diabetes can be prevented, because besides hereditary factors, the main cause of type II diabetes is an unhealthy lifestyle. Increasing number of DM caused by late diagnosis, so it deserves special attention from the public and healthcare workers. Based on these problems, it is necessary to have an early detection of the disease classification for detected patients with type II diabetes mellitus and undetected patients of diabetes mellitus type II. This is related to the purpose of national social security agency in Indonesia called *Badan Penyelenggaraan Jaminan Sosial (BPJS)* that is to detect some factors of chronic diseases in order to encourage participants to easily realize the symptoms, doing an early detection, and doing an early prevention the risk of chronic disease.

The dataset that used in this study are medical records of chronic diseases from Public Health Center, Banjarnegara, Indonesia which consists of complex attributes that are categorical and continuous attributes, so it is necessary to find out the main attributes that are dominant to the early detection of the disease. According to the advantages offered by the CART method that is easy to implement data structures, and it can explore the structure of complex medical records, the CART method is expected to correspond to the early detection of problems of type II diabetes mellitus.

In the study conducted by [1] the CART method produces the best classification accuracy at the same proportion of the division between training data and test data that is equal to 84.83%. And the study by [9] produces the greatest accuracy of the training data 800 number of data is 76%. This shows that small changes in the training set will create large changes in the learned classifier. To overcome the accuracy problems, both studies suggested combining CART with other methods (ensemble methods). Therefore, in this study the CART algorithm becomes the classifier ensemble method, while the ensemble approach used the Random Forest (RF).

II. LITERATURE REVIEW

A. Basic Concepts of Diabetes

Diabetes mellitus (DM is a disease characterized by high blood sugar levels caused by interference with the insulin secretion or disorders insulin function or both. [12]

DM current classification is as follows:

1. Type 1 diabetes (Insulin Dependent Diabetes Mellitus)
2. Diabetes type 2 (diabetes by genetic factors and lifestyle).
3. Another type of diabetes (genetic defect disorder, exocrine pancreas, hormonal abnormalities, medications, infections, immunological factors, etc.)
4. Pregnancy diabetes (occurring during pregnancy).

B. Random Forest

Random Forest (RF) is a classifier consisting of a few decision trees. Each decision tree is built by using a random vector. The general approach used to insert random vectors in the formation of the tree is to choose a random value F, as F attributes (features) input to be split at each node in the decision tree to be formed. By choosing a random value F then it should not have to check all the attributes that exist, just look at the selected F attributes. Parameters used to adjust the strength of random forest in the selection of F value and the number of trees that will be built in the forest [3]. If the F value is too small, then the tree has a tendency to have a very small correlation, and it applies the same for the other way round. Therefore, the value of F can be determined using the formula:

$$F = \log_2(M+1) \quad (1)$$

Where M is the total number of features. Besides the selection of attributes, also conducted a random way when selecting the training set. Bagging (bootstrap aggregating) is a technique that can be used to form a bootstrap sample. Each decision tree is built using a bootstrap sample of the data and candidate attributes to be split in each node is derived from the set of random attributes of the data results of bagging.

Here is the flow of the random forest algorithm:

1. Choose a value of n which indicates the number of trees that will be raised in a forest.
2. Generate n bootstrap samples with bagging technique of the training set.
3. At each node in a tree, select the F value obtained from equation (1).
4. Take the set as much as the F attribute that will be the candidates of splitting attribute of each node. Split tree using that set. Attributes that becomes the next node is determined based on specific criteria (based on a decision tree algorithm selected). During the process of the formation of the tree, the value of F is constant.
5. Random forest continues to be formed without any pruning. It is shown to eliminate bias in the percentage of predicted results.
6. The prediction results obtained from the model (most frequent class) of each decision tree in the forest.

C. CART

CART algorithm is a method of regression trees and classification trees that will produce a classification tree if the class consists of categorical attributes, and will produce a regression tree if the class consists of continuous attributes [2]. CART will select a number of attributes and interactions between attributes were most dominant in determining the outcome of the dependent attribute with binary sorting procedure.

In selecting the best splitter, CART attempted to maximize the average purity of the two child nodes. The way to measures of purity can be chosen freely, can be called a splitting criterion or splitting functions. The most common splitting functions are Gini index [7].

Calculation of Gini index is obtained by the formula:

$$Gini(t) = 1 - \sum_{i=0}^{c-1} [p(i|t)]^2 \quad (2)$$

Where P (i | t) is the relative frequency of class i at node t, and c is the number of classes. In these calculations will achieve the highest value if the distribution of uniform class, and has the smallest value if it contains all the records with the same class.

D. CART and Random Forest

Random forest is built from a series of decision trees. One of the methods for decision tree is CART. When combined with the Random Forest algorithm, CART has some modification.

Modification lies in the determination of candidate attributes that will be used as a splitter. If the CART algorithm uses all the attributes as a candidate splitter, then the ensemble algorithm CART and RF splitter candidates is limited, as in equation (1).

In addition to the determination of candidate attributes, modifications are also found in the stopping building tree. Tree pruning does not occur because the estimation error has been made in the construction of many bootstrap samples from the original dataset and the formation of n-tree in the forest. While the CART algorithm itself prune using cross validation calculation [6].

Furthermore, for the selection of both categorical and numerical attributes have not been modified.

E. Out of Bag (OOB)

Data out-of-bag (OOB) is the training data that were not selected at the time of a random number for the construction of a decision tree. With the data OOB, then random forest does not need to run repetitive test data to obtain an estimate of the internal error every time a tree is formed in the woods. In general, the OOB error rate is defined as follows:

$$\text{OOB } error\ rate = \frac{\text{Amount of data OOB predicted wrong}}{\text{The total number of data}} \quad (3)$$

III. METHODOLOGY

A. Data Characteristics

Job	Gender	Age	BMI	Sistole	Diastole	Heredity	Diagnosis
5	1	48	20.44674	180	90	0	1
5	0	51	18.90204	160	90	1	1
5	0	50	19.8791	140	80	1	1
4	0	50	22.22222	150	90	0	0
2	0	59	30.22222	150	100	0	0

Figure 1 Overview of Type II diabetes dataset

The characteristics of the data used in this study are described in Figure 1 with the following details:

1. Attribute values of job are 1. Military / Police, 2. Civil / Retired, 3. Private employees, 4. Wife / husband military / police / civil, 5. Farmers / Labor.
2. Attribute values of gender are male (M) and female (F).
3. Attribute values of age are 20-80 years old.
4. Attribute values of Body Mass Index (BMI) are a calculation of BMI = weight / (height x height).
5. Attribute values of systole are 90-230mm
6. Attribute values of diastole are 60-130Hg
7. Attribute values of diastole are 1 (yes) and 0 (no)

B. Classification

In general, the system that will be built in this research study is a system for early detecting the disease Diabetes Mellitus (DM) type II using the Random Forest (RF) and Classification and Regression Tree (CART). The detection results are shown using data test from the medical records of patients at the public health center. The system will also be built to test the accuracy of the constructed by CART and RF methods in detecting disease type II diabetes, and evaluating the major attributes that most influential in the development of the methods. General overview of the system is shown in Figure 2.

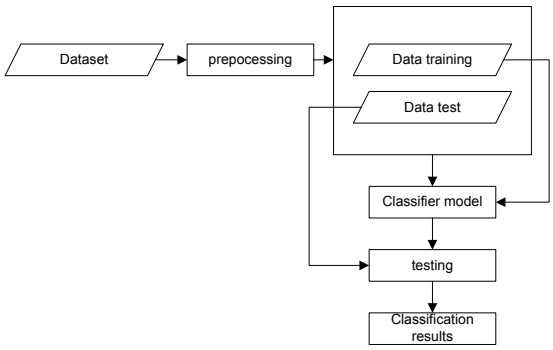

Figure 2 Overview of the system

At first the data type II diabetes disease detection is still a raw data. Therefore, the data preprocessing needs to be done. There are several stages of preprocessing that performed in this process, including feature selection and transformation of attribute values into a numeric text. After the preprocessing stage is completed, the data are divided into two types of training data and test data, then performed on the modeling of the training data and generate a classifier model. The next step will be testing the classification of the test data with the model classifier that has been formed. Next classification results obtained are detected patients with type II diabetes mellitus or undetected diabetes mellitus type II, and the accuracy of the classifier modeling.

The development process model (figure 3) is repeated as many times as the user input number of trees. After detection model formed, then the testing for model that has been constructed will be done using the test data. In the testing process, there is a voting process, which is majority vote whether the corresponding class of objects detected output from the process of testing how the results of the classification accuracy rate will be calculated by comparing the test data that has been labeled a class.

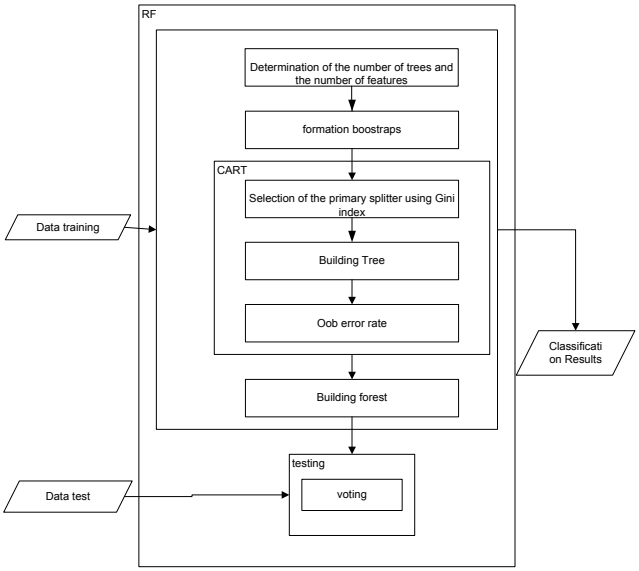

Figure 3 Classification Model

In the process of merging the CART and RF method, a dataset which consist of n instant data is formed with a number of bootstrap instant x and y number of attributes. Formation bootstrap done by random with replacement so that in the bootstrap allows the same data, and each bootstrap possibility of having different data attributes.

Then for each bootstrap that has been established, the tree constructed using the CART method (Figure 4)

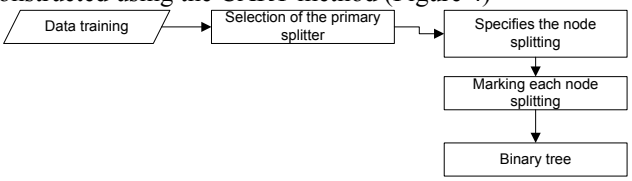

Figure 4 building the CART tree

The first stage in the construction of the tree is to determine the independent attribute being the best splitter (primary splitter). Primary splitter is a splitter that decreases the diversity of set record with the largest decrease number. In this step is calculated using the Gini index value for each attribute and issued a number of F attribute with the best value of Gini index.

The error rate of the model can be found by using the voting against all the training data on each tree that is not involved in forming bootstrap tree.

In the evaluation phase, the classification results data matched the test data, thus will get a percentage value of modeling accuracy of the classifier that has been created, and can be obtained major determinant attributes of type II diabetes disease detection.

IV. DISCUSSION

A. Testing Models

System testing performed on 600 training data and 100 test data. Tests performed on the data using the parameter number of different trees and the number of different splitter attributes.

1) Scenario Testing Models Against Number of Trees

One of the inputs of the Random Forest is the number of trees or classifiers that are used in a model. According to research conducted by [3] have performed testing the effect of number of trees. In that study found that increasing numbers of trees will not overfit, but will generate an error value limit in general. In this research, will be doing the test to prove that the addition of tree only produces error thresholds in general and had little influence on the value of the model's accuracy.

In the random forest, there is an optimal number of trees, where in this condition, the value of accuracy and the error rate is stable by considering the time required. However, the minimum number of trees is not fixed for each dataset characteristics. This is due to the random function that rose, that is the random function to select a row of data and random function to choose attributes. Therefore, at this research will be established models with four option of the number of trees that is 5, 25, 50, and 100. Then each model recorded the values of accuracy, error rate, time of performance, and importance of attributes to be analyzed.

Related to a random function, for a dataset with the same input parameters, when executed more than once trial can produce different accuracy values. The resulting accuracy in the next execution, it could be better, or could be worse. Therefore, in this test, was observed by testing as much as five times experiment.

2) Scenario Testing Models Against Number of Attribute Splitter

In addition to the number of trees used in the model, Random Forest also has a number of splitter attributes input to build a model. The number of trees on testing in this scenario set 50 trees. While the value of splitter attributes is 1 to 4 attributes. The number of candidate attributes is obtained from the calculation splitter equation (1) which produces 3 attribute values; the 4 value is to prove that the equation (1) is true.

B. Results

1) Testing Results of the Number of Tree

Tests will be done to measure the level of accuracy of the prediction and to determine whether there is an influence based on the number of trees. The accuracy describes the correctness of classification which has been built. Based on the test results it finds that the result of accuracy continues to increase while the addition of 5 to the 50 trees number of trees, while the value of accuracy between 50 trees to 100 trees showed no significant change.

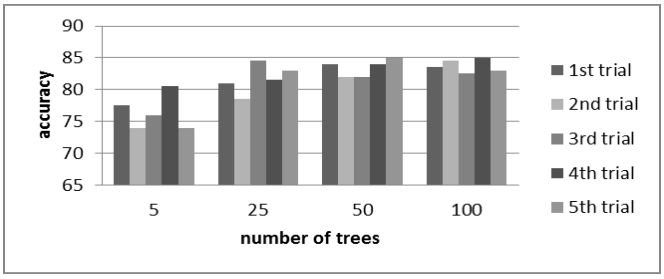

Figure 5 accuracy value based the number of trees

In the test results also showed that adding the number of trees would create the value of the classifier error rate decreases.

While testing the performance of the time shows that a rule generated from the training data, when tested on the test data has an increasing number of trees, it will require increased testing time. Time difference between models with 50 trees with 100 trees appear significant, while the average difference in accuracy between 50 and 100 trees only 0.3% it was determined that in this case study the formation of the optimal model that uses is 50 trees.

Furthermore, based on testing by varying the amount of trees obtained value of attribute importance heredity always occupies the highest value, followed by age and BMI. The addition of tree after 50 trees showed no significant change.

2) Testing Results on the Number of Attribute Splitter

Based on the test results showed that at the time of testing, the average value of the 3 splitter attributes accuracy is 83, 8%, while in the 4 splitter attributes is 84%. It is proved that the equation (1) in accordance with the applied methods.

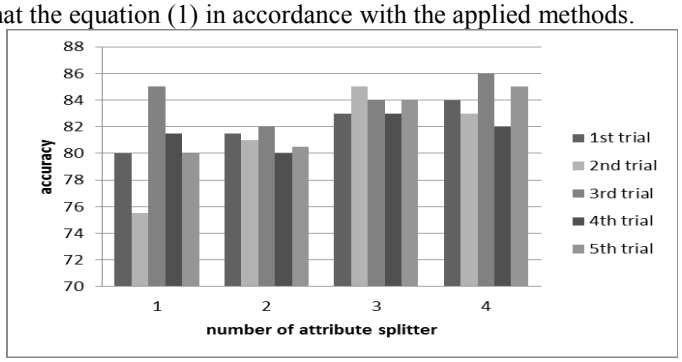

Figure 6 accuracy value based the number of attribute splitter

Based on the testing that shows the percentage of error is found that by increasing the number of attribute splitter showed a decrease in error rate.

Whereas if the number of attributes splitter added will require increased testing time. This is due to the formation of one tree, time required to analyze each of the attributes to be used as a node (best split) on the tree is longer, so that the overall time of the establishment of the model is increased.

Based on testing with the variation of the number of candidate attributes splitter was found that the value of attribute heredity always occupy the highest important value.

V. CONCLUSION

In conclusion, we found that, to get optimal models in the early detection of diabetes mellitus type II by using CART and random forest the value of the input tree number is 50 with the number of splitter attributes is 3, with the average accuracy 83, 8%. This proves that the combination of these two methods can improve the accuracy in which the single classifier CART result 77% accuracy for that dataset. Moreover, increasing the number of trees can improve the value of the accuracy and reduce the value of the resulting error rate, and by increasing the number of candidate splitter attributes give effect to the values of accuracy and error rate. However, these additions have a minimum limit according to equation (1). Another result is the most attribute importance of early detection of diabetes mellitus type II are the heredity while the other support attributes are age and BMI.

In the future, this research can be used to recommend patients when registering in the hospital to check further as checking blood sugar levels to the laboratory.

ACKNOWLEDGMENT

The authors would like to thank all those who helped in improving the quality and clarity of this paper .Without their continued efforts and support, we would have not been able to bring our work to a successful completion.

REFERENCES

[1] Aprian, Krisan Widagdo. *Formation of Binary Tree Classification Algorithm CART (Classification and Regression Tree) Case Study Diabetes Pima Indian Tribe*. Diponegoro University, 2010.

[2] Breiman, Leo., Friedman, Jerome H., Oshlen, Richard A., Stone, Charles J. 1996. *Classification and Regression Tree*. Statistics Department, University of California

[3] Breiman, Leo. 2001. *Random Forest*. Statistics Department, University of California

[4] Chen , Cheng-Mei., Hsu ,Chien-Yeh., Chiu, Hung-Wen., Rau, Hsiao-Hsien. (2011). "Prediction of Survival in Patients with Liver Cancer using Artificial Neural Networks and Classification and Regression Trees". IEEE: Seventh International Conference on Natural Computation.

[5] Duda, Richard O., Hart, Peter E., Stork, David G. (2001). *Pattern Classification*. Canada: John Wiley & Sons.

[6] Kurniawan, Fransiska A. *Analysis and Implementation of Random Forest Classification and Regression Tree and (CART) for classification in the Misuse Intrusion Detection System*. IT Telkom, 2011.

[7] Lewis, Roger J. (2000). "An Introduction to Classification and Regression Tree (CART) Analysis". Presented at the 2000 Annual Meeting of the Society for Academic Emergency Medicine in San Francisco, California.

[8] Nuwangi, S.M., Oruthotaarachchi, C.R., Tilakaratna, J., Caldera, H.A. (2010). "Utilization of Data Mining Techniques in Knowledge Extraction for Diminution of Diabetes". Second Vaagdevi International Conference on Information Technology for Real World Problems. Colombo: University of Colombo School of Computing.

[9] Puteri, Nita A. *Heart Disease Prediction Algorithm Classification and Regression Tree (CART)*. IT Telkom, 2013.

[10] Ramadhansyah, Ridha. (2011). *Analysis and Implementation of Random Forest Ensemble Method and C4.5 Algorithm for Spam Email Classification Data*. IT Telkom, 2011.

[11] Tama, Bayu Adhi, Rodiyatul, Hermansyah. (2011). "An Early Detection Method of Type-2 Diabetes Mellitus in Public Hospital". Telkomnika, Vol.9, No.2, August 2011, pp. 287-284.

[12] - . (2010). *Chronic Disease Management Program*. Jakarta: PT. Askes (persero).