Application of Random Forest on C5.0 Algorithm for Diabetes Mellitus Disease Classification in RSUD Tora Belo Sigi District

Nur Intan[1], Mohammad Fajri[2], Hartayuni Sain[3]

[1,2,3]Department of Statistics, Faculty of Mathematics and Natural Sciences, Tadulako University
nurrintan66@gmail.com

Abstract. In 2021, Sigi Regency is the area with the highest level of diabetes mellitus in Central Sulawesi Province, as evidenced by the number of patients at RSUD Tora Belo which continues to increase where in 2020 there were 466 patients and in 2021 it increased to 526 patients. Accurate classification of people who have positive or negative laboratory test results for diabetes mellitus is important to get the right treatment. The purpose of this study is to classify the status of people who have positive or negative laboratory test results for diabetes using random forest applied to the C5.0 algorithm. The results obtained are that the Low Density Lipoprotein variable is the most important variable in the classification with a mean decrease gini value of 40.701691 so that the main factor that causes a person to suffer from diabetes mellitus is the Low Density Lipoprotein variable with a classification accuracy of 88.17%.

Keywords: C5.0, Decision Tree, Diabetes Mellitus, Classification, Random Forest.

1 Introduction

Based on data from the World Health Organization (WHO), it is estimated that 422 million adults have diabetes mellitus. The International Diabetes Federation (IDF) organization in 2019 also said that at least 463 million people in the age group of 20-79 years in the world suffer from diabetes or equivalent to a prevalence rate of 9.3%. Indonesia ranks third in the Southeast Asia region with a prevalence of 11.3%. In Central Sulawesi Province, the area with the highest number of diabetes cases is Sigi Regency. Sigi Regency experienced an increase in cases of there were 15,244 patients with diabetes mellitus and 704 people who received treatment. The comparison between treatment recipients and those suffering from diabetes mellitus is much different [1].

To be able to identify diabetes mellitus, it is necessary to know the characteristics of diabetes mellitus patients through various laboratory tests. The laboratory test results have discrete values that can be categorized, therefore a machine learning method is needed to classify diabetes mellitus disease, including random forest applied to the C5.0 algorithm. Random forest is used to build several decision trees and combine them to get more accurate and stable predictions as well as one of several ways to overcome the problem of overfitting, while C5.0 is used to maximize the level of user interpretation of the results presented in two forms, namely using a decision tree and

Y. Yuyun et al. (eds.), *Proceedings of the 5th International Seminar on Science and Technology (ISST 2023)*,
Advances in Physics Research 10,
https://doi.org/10.2991/978-94-6463-520-1_15

a set of easy-to-understand rules. This method will form a number of trees, where one tree will provide one voting unit. The final result on classification is done by majority voting, or the most votes will be taken [2].

Random Forest is a flexible and easy-to-use machine learning algorithm that produces, even without the use of many parameters, relatively good results. It is also one of the most widely used algorithms, due to its simplicity and diversity. Meanwhile, the C5.0 algorithm is one of the data mining methods with classification techniques based on the decision tree algorithm, making it easier to interpret the final results in the form of a decision tree that is easy to understand [3].

2 Data and Method

The stages of preparing and estimating using random forest are:

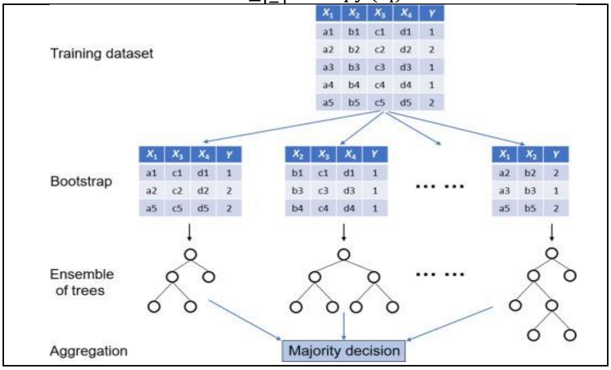
1. Random sampling is done with n sized selection from the training data cluster. This stage is called the bootstrap stage.
2. Random selection of explanatory variables m times, with m<d. This stage is carried out during the sorting process in single tree formation. This stage is called the random sub-setting stage.
3. Repeat steps a-b for k times so that k random trees are obtained.
4. Perform joint estimation based on the k trees. Note that at each time of tree formation, the candidate explanatory variables used for splitting are not all the variables involved but only some of them are randomly selected. It is conceivable that this process produces a collection of single trees of different sizes and shapes. The expected result is that the collection of single trees has a small correlation between the trees. This small correlation results in the variance of the random forest result being small and smaller than the variance of the bagging result [4].

Before applying random forest first to form a C5.0 decision tree, as follows [5].

$$\mathrm{Entropy\ (S)} = \sum_{i=1}^{n} -p_i \log_2(p_i) \qquad (1)$$

$$\mathrm{Gain\ (S, A)} = \mathrm{Entropy\ (S)} - \sum_{i=1}^{n} \frac{[S_i]}{[S]}\ \mathrm{Entropy\ (A)} \qquad (2)$$

$$\mathrm{Gain\ Ratio} = \frac{\mathrm{Gain\ (S,A)}}{\sum_{i=1}^{n} \mathrm{Entropy\ (S_i)}} \qquad (3)$$



X_1	X_2	X_3	X_4	Y
a1	b1	c1	d1	1
a2	b2	c2	d2	2
a3	b3	c3	d3	1
a4	b4	c4	d4	1
a5	b5	c5	d5	2

X_1	X_3	X_4	Y
a1	c1	d1	1
a2	c2	d2	2
a5	c5	d5	2

X_2	X_3	X_4	Y
b1	c1	d1	1
b3	c3	d3	1
b4	c4	d4	1

X_1	X_2	Y
a2	b2	2
a3	b3	1
a5	b5	2

Figure 1. Illustration Of Random Forest

The data used in this study are secondary data obtained from Torabelo Hospital, Sigi Regency. The population used in this study is data on the number of patients with diabetes mellitus in Sigi Regency while the sample used from this study is data on the number of patients with diabetes mellitus at Torabelo Hospital, Sigi Regency in 2022 which amounted to 465 patients. The following variables were used in the study:

Table 1. Description of Variables

Variables	Description
Y	Diabetes Mellitus
X_1	Gender
X_2	Age
X_3	Blood Glucose/2h
X_4	Fasting Blood Glucose
X_5	High Density Lipoprotein (HDL)
X_6	Low Density Lipoprotein (LDL)
X_7	Triglycerides

Data analysis in this study used Random Forest on the C5.0 algorithm with the help of R-studio software. The stages of the analysis carried out are as follows:

1) Performing data collection
2) Perform descriptive statistical analysis
3) Dividing data into 80% training data and 20% testing data
4) Perform boostrap sampling to take samples
5) Performing the C5.0 algorithm on training data
6) Repeat the above process until it forms a lot of trees
7) Classification is done by majority vote or taken the most votes
8) Evaluate the model by calculating the accuracy, sensitivity and specitivity values on the testing data.

3 Results and Discussion

The first step taken in the process of forming the C5.0 decision tree algorithm is to calculate the entropy, information gain and gain ratio values.

Table 2. Calculation of entropy and gain ratio of node 1

Variable X	Category	Variable Y		Total	Entropy	Gain	Gain Ratio
		0	1				
X_1	1	60	119	179	0,92	0,0002	0,0001
	2	67	126	193	0,931		
X_2	1	28	10	38	0,831	0,205	0,063
	2	34	21	55	0,959		
	3	49	59	108	0,994		

	4	16	155	171	0,448		
X_3	1	109	41	150	0,846	0,344	0,203
	2	11	142	153	0,373		
	3	7	62	69	0,474		
X_4	1	102	25	127	0,716	0,37	0,22
	2	12	87	99	0,533		
	3	13	133	146	0,433		
X_5	1	36	95	131	0,848	0,008	0,004
	2	91	150	241	0,956		
X_6	1	100	20	120	0,65	0,384	**0,237**
	2	10	92	102	0,463		
	3	17	133	150	0,51		
X_7	1	99	59	158	0,953	0,2	0,093
	2	22	158	180	0,536		
	3	6	28	34	0,672		

Based on Table 1, it is obtained information that the variable that has the highest gain ratio value is the low density lipoprotein (X_6) variable, which is equal to 0.236637378, so the low density lipoprotein (X_6) variable is used as the root node or root of the tree.

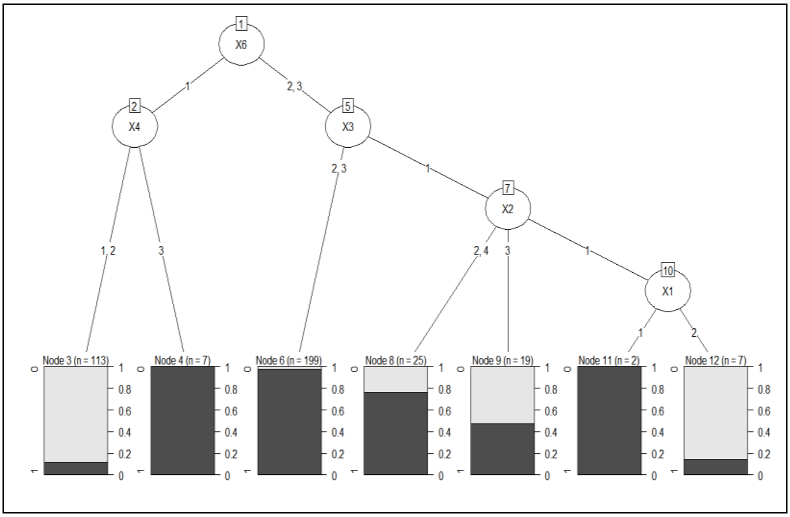


Figure 2. C.5 Decision Tree

Based on table 1, variable X_6 has the highest gain ratio value. If variable X_6 category 1, and variable X4 category 1,2, then it is classified into the category of no diabetes mellitus, whereas if variable X_6 category 1 and variable X_4 category 3, then it is classified into the category of diabetes mellitus. If variable X_6 category 2,3 and variable X_3 category 2,3 then it is classified into the category of diabetes mellitus, whereas if variable X_6 category 2,3, variable X_3 category 1, and variable X_2 category 2,4 then it is also classified into the category of diabetes mellitus, whereas if variable X_6 category 2,3, variable X_3 category 1, and variable X_2 category 3 then it is classified into the category of not diabetes mellitus. If variable X_6 category 2,3, variable X_3 category 2,3, variable X_2 category 2,4, and variable X_1 category 1, it is classified into the diabetes mellitus category, whereas if variable X_6 category 2,3 and variable X_2 category 3, and X_1 category 2, it is classified as not having diabetes mellitus. If variable X_6 category 2,3 and variable X_2 category 1, then it is classified into the category of diabetes mellitus, whereas if variable X_6 category 2,3 and variable X_1 category 2, then it is classified as not diabetes mellitus.

Table 3. Confusion Matrix of C5.0

Prediction	Actual		Total
	1	0	
1	57	6	63
0	6	24	30
Total	63	30	93

Table 4. Out of Bag Random Forest

Ntree	*Error*
100	10,48%
250	10,75%
500	10,48%
750	**10,22%**
1000	10,22%

Based on Table 4, it is obtained that the smallest error value is at ntree 500, which is 10,22%, so it can be concluded that the optimum ntree is 750.

Table 5. Confusion Matrix of Random Forest

Prediction	Actual		Total
	1	0	
1	57	5	62
0	6	25	31
Total	63	30	93

Based on Table 5, it is obtained that the amount of data used as evaluation is 93 data, where the correctly predicted class 1 is 57 data and the correctly predicted class 0 is 25 data. In addition, there are also errors in classification where there are 5 data included in class 0 but predicted as class 1 and 6 data included in class 1 but predicted as class 0.

Table 6. Comparison of C5.0 performance with ensemble random forest

Model Evaluation	C5.0	C5.0 + Random Forest
Accurasy	84,95%	88,17%
Sensitivity	83,33%	90,48%
Specitivity	85,71%	83,33%

Based on the results of the calculations carried out, the accuracy value is 88,17%, the sensitivity value is 90,48% and the specitivity value is 83,33%, so it can be concluded that the ensemble random forest has succeeded in improving the performance of C5.0.

4 Conclusion

Based on the results and discussion, it can be concluded as follows.
By using the C5.0 algorithm, a classification model is obtained where the main factor that affects a person suffering from diabetes mellitus is the Low Density Lipoprotein variable (X_6). Ensemble random forest successfully improves the performance of C5.0 with an accuracy value of 87.1%.

References

[1] Marcania, M. (2019). Prediksi pengangkatan karyawan dengan metode klasifikasi algoritma C5.0 (studi kasus pt. kiyokuni indonesia factory-2). Thesist, 0, 71.

[2] Nugraha, R. H., Yuwono, E., Prasetyohadi, L., Arief, Y. B., & Patria, H. (2022). Analisis Konsumsi Energi Listrik Pelanggan Dan Biaya Pokok Produksi Penyediaan Energi Listrik dengan Machine Learning. Jurnal Sains Komputer & Informatika (J-SAKTI, 6(1), 47–56. http://dx.doi.org/10.30645/jsakti.v6i1.424

[3] Putri, S. U., Irawan, E., & Rizky, F. (2021). Implementasi Data Mining Untuk Prediksi Penyakit Diabetes Dengan Algoritma C4.5. Kesatria: Jurnal Penerapan Sistem Informasi (Komputer dan Manajemen), 2(1), 39-46.

[4] Rahmayanti, V., Azhar, Y., & Pramudita, A. E. (2020). Penerapan algoritma c5.0 pada analisis faktor-faktor pengaruh kelulusan tepat waktu mahasiswa teknik informatika UMM. Jurnal Repositor, 1(2), 131. https://doi.org/10.22219/repositor.v1i2.545

[5] Triana, L., & Salim, M. (2017). Perbedaan Kadar Glukosa Darah 2 Jam Post Prandial. Jurnal Laboratorium Khatulistiwa, 53.

[6] WHO. World Health Organization. Classification of Dabetes Mellitus 2019. Geneva: World Health Organization; 2019. 1-40.