IMPLEMENTATION OF C5.0 ALGORITHM FOR SHAPING PATTERNS OF DIAGNOSIS OF DIABETES MELLITUS DISEASE

Muhammad Latief Saputra[1], Irwan Budiman2, Radityo Adi Nugroho3, Dwi Kartini[4], and Muliadi5
FMIPA ULM Computer Science Study Program
Jl. A. Yani Km 36 Banjarbaru, South Kalimantan
Email: latiefsaputra.ilkom14@gmail.com

Abstract

This study applies the C5.0 algorithm to form a decision tree pattern for diagnosing diabetes mellitus. C5.0 algorithm is a decision tree based classification algorithm. This algorithm focuses on the acquisition of information gain on all attributes. The data used is a diabetes mellitus dataset obtained from the Kaggle database website. Data preprocessing is done and data sharing is done 4 times with the distribution of training data 60% 70% 80% and 90%. Data sharing uses stratafied random sampling methods so that the distribution of training and testing data is in accordance with its portion. Calculation of accuracy performance using confusion matrix. Classification performance using C5.0 algorithm. With 90% training data get 72.73% accuracy of rules generated as many as 70 rules. With 80% training data the accuracy value is 74.03%. The rule is 64 rules. With 70% training data get an accuracy value of 76.52% of the rules generated 59 rules. With 60% training data get an accuracy value of 74.59% of the rules generated as many as 53 rules. From all the experiments that have been done, the best accuracy is found in experiments with 70% training data.

***Keywords**: C5.0, Diabetes Mellitus, Confusion Matrix, Classification*

1. Introduction

Data mining according to Vercellis is a job that describes a process of analysis in which a job occurs repeatedly and is carried out in a large enough database, with the aim of obtaining information and knowledge related to decision making and problem solving. Data mining is the process of finding data and a large amount of information from a data base or other information repository. Usually the data mining process also finds unique patterns of information. The patterns found in the data mining process must have meaning and also provide benefits for its users.

Classification is a technique by looking at the behavior and attributes of a group that has been defined. This technique can classify new data by manipulating existing data that has been classified and by using the results to provide a number of rules. One example that is easy and popular is the Decision Tree which is one of the most popular classification methods because it is easy to interpret. Decision tree is a prediction model using tree structure or hierarchical structure.

C5.0 is a commercial version of C4.5 which is widely used in many data mining packages such as Clementine and RuleQuest. Unlike C4.5, the use of the right algorithm for C5.0 has not been revealed. The results show that C5.0 increases memory usage by around 90%, faster than C4.5.

The choice of this algorithm is because the C5.0 algorithm does not use distance vectors to classify objects so that it is suitable for observational data with numeric attributes or attributes that are of nominal value that are categorical in nature, where each

value cannot be added or subtracted, for example shapes, colors and flavors. In addition this algorithm also produces a tree with the number of branches per node varies so that it can form a more efficient decision tree.

The main benefit of using a decision tree is its ability to break down the process of complex decision making to be simpler so that decision making will become more interpretative of the problem solution. Decision trees are also useful for exploring data, finding hidden relationships between a number of potential input variables and a target variable. The decision tree combines data exploration and modeling so that it is very good as the initial step of modeling even when used as the final model of several other techniques

Diabetes Mellitus is a chronic metabolic disorder due to the pancreas not producing enough insulin or the body cannot use insulin produced effectively. Insulin is a hormone that regulates the balance of sugar levels. The result is an increase in glucose concentration in the blood. Diabetes mellitus is a chronic disease that occurs due to abnormal insulin secretion on an irregular rise in glucose. Diabetes mellitus will increase blood sugar in the body resulting in complications that can lead to several risks such as stroke, heart disease, blindness, kidney failure and death

An increasing number of people with diabetes due to diabetes is known as the silent killer. This refers to many who do not realize that they have diabetes. Patients are usually known to be infected with this disease when complications occur without any initial treatment. To reduce the number of people with diabetes who are increasing, early detection can be done by experts

2. RESEARCH METHODS

2.1. Research procedure

a. Collecting data ie Diabetes Dataset then do a literature search related to the method to be used.
b. Preprocessing Data that starts by doing Data Cleaning using the Mean Imputation method.
c. To classify the data into training data and test data. Training data for building decision trees, while test data for calculating decision tree accuracy..
d. Then a classification model is built to form a decision tree.
e. *Evaluation Model* after the model has been created, the next stage is to evaluate the results of the models formed. These models must be translated as important or high-value information. The level of accuracy is expected to have the right proximity to the actual value that exists. The level of precision is expected to have the level of information obtained with the information received can be high, and the level of recall is expected to call back information can run optimally. Measuring the level of accuracy, precision, and recall in this study is using the Confusion Matrix.

The flowchart of the research carried out can be seen below:

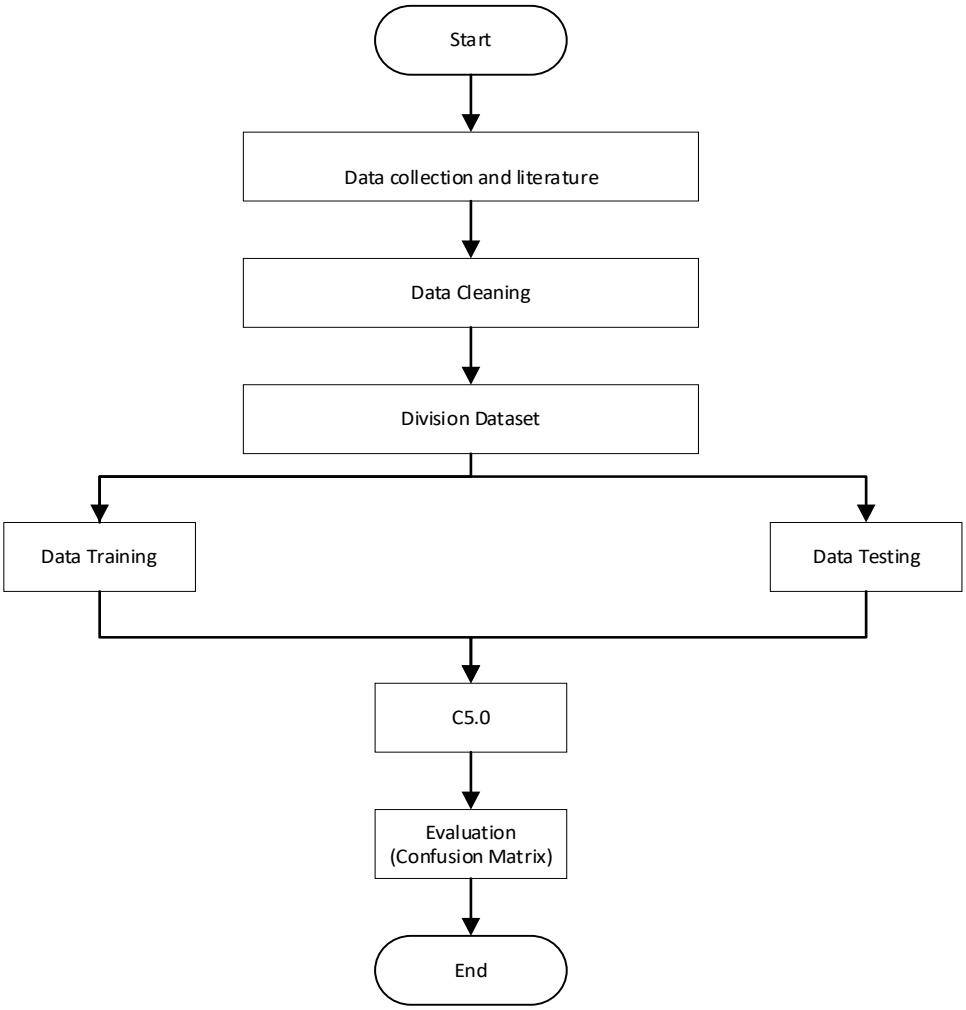

Figure 1. Research Flow

3. RESULTS AND DISCUSSION

3.1. Results

3.1.1 Data collection

Diabetes Mellitus dataset is obtained from Kaggle. This dataset consists of 768 data and has a missing value. This dataset consists of 8 numeric attributes and 1 class attribute. The target class of this dataset is twofold, positive and negative. 500 negative diabetes patients (negative class) and 268 positive patients (positive class). Literature search is a method of collecting data obtained from literature books and references about data mining, classification, C5.0 algorithm, and diabetes mellitus.

3.1.2 Preprocessing

The preprocessing stages are performed:

a. Data Cleaning

At this stage the mean imputation method is used at the data cleaning stage. You do this by replacing the missing value of the numeric attribute with the mean or average. Mean or mean is the average value of the data obtained by adding up all the values divided by the number of items in the data.

b. Data *Transformation*

The data used is changed to the form that is most suitable for the C5.0 algorithm, that is, the continuous attribute data is converted into category data. The following results data transformation data

Table 1. Results of data transformation

No	Attribute	Category	Range
1	Pregnant	Nulligravida	0
		Primigravida	1
		Secungravida	2
		Multigravida	> 2
2	Glucose	Normal	<140
		High	≥140
3	Blood pressure	Low	40-60
		Normal	60.1-80
		Rather high	80.1-90
		High	90.1-100
4	Skin Triceps	Low	<20
		Normal	20-40
		High	> 40
5	Insulin	Normal	<200
		High	> 200
6	Body Mass Index	Deficiency	≤ 17.0 –18.4
		Normal	18.5 -25.0
		Exaggerated	25.1-27
		Very excessive	> 27
7	Age	Deficiency	≤ 17.0 –18.4
		Productive	15 –64
		Old	> 64
8	Diabetes pedigree function	Very low	0 -2444
		Low	0.245 - 0.525
		Sedamg	0.526 –0.805
		High	0.806 –1.11
		Very High	> 1.11

Source: Implementation of C5.0 Algorithm for Establishing Decision Tree Patterns for Diabetes Mellitus, 2020

c. Distribution of Training Data and Data Testing

In this study the total amount of data is 768 where the negative value is 500 data and the positive value is 268. Tested with 4 trials with training data sharing 60% 70% 80% and 90% and with testing data 40% 30% 20% 10% Tested with 4 times to find the best results. For the distribution of data raining and testing data using stratafied random sampling method, for example with 70% training data training data and 30% testing data. Then the training data has 538 data with 350 class negative data division and 188 positive data. As for the testing data as many as 230 data with 150 negative data class division and 230 positive data

3.1.3 Data Mining and Evaluation

a. Algorithm C5.0 algorithm dataset 1

In this search using 70% training data and 30% testing data. There are several stages in carrying out the C5.0 calculation so that a decision tree is

formed. The stages of C5.0 calculation in this study include calculating the entropy in the dataset, calculating the information gain value. After that form a decision tree and determine the rules of the decision tree. Calculating the entropy on the pregnant attribute uses the following equation.

$$Entrophy\ (pregnant = Nulligravida) = (-\frac{54}{84} * log2(\frac{54}{84})) + (-\frac{30}{84}) * log2(\frac{30}{84}))$$

$$Entrophy\ (pregnant = Nulligravida) = 0.940285959$$

$$Entrophy\ (pregnant = Primigravida) = (-\frac{72}{94} * log2(\frac{72}{94})) + (-\frac{22}{94}) * log2(\frac{22}{94}))$$

$$Entrophy\ (pregnant = Primigravida) = 0.784992089$$

$$Entrophy\ (pregnant = Secungravida) = (-\frac{55}{68} * log2(\frac{55}{68})) + (-\frac{13}{68}) * log2(\frac{13}{68}))$$

$$Entrophy\ (pregnant = Secungravida) = 0.703926068$$

$$Entrophy\ (pregnant = Multiravida) = (-\frac{169}{292} * log2(\frac{169}{292})) + (-\frac{123}{292}) * log2(\frac{123}{292}))$$

$$Entrophy\ (pregnant = Multiravida) = 0.982023501$$

The calculation of the entropy value is carried out for each attribute in the diabetes mellitus dataset. After completing all entropy counts, continue to the next step, which is to calculate the infromation gain value from the dataset.

Table 2. Number of Entropies in Each Attribute

		JML CASE	NO (S1)	YES (S2)	ENTROPY
TOTAL					
TOTAL		538	350	188	0.933569
Pregnant					
	Nulligravida	84	54	30	0.940286
	Primigravida	94	72	22	0.784992
	Secungravida	68	55	13	0.703926
	Multigravida	292	169	123	0.982024
Glucose	Normal	390	299	91	0.783777
	High	148	51	97	0.929148
Preesure	Low	86	70	16	0.693127
	Normal	328	207	121	0.949826
	High	28	18	10	0.940286
	Rather high	96	55	41	0.984604
Triceps	Normal	392	242	150	0.959894
	Low	97	82	15	0.621329
	High	49	26	23	0.997294
				0	
Insulin	Normal	472	318	154	0.911071
	High	66	32	34	0.999338
Mass	Deficiency	3	3	0	0
	Normal	76	72	4	0.297472
	Exaggerated	38	32	6	0.629249
	Very excessive	421	243	178	0.982736

				0	
Pedigree	Very low	131	97	34	0.826054
	Low	222	148	74	0.918296
	Is	110	67	43	0.965384
	High	43	24	19	0.990225
	Very high	32	11	21	0.928362
Age	Productive	527	342	185	0.934997
	Old	11	9	2	0.684038

Source: Implementation of C5.0 Algorithm for Establishing Decision Tree Patterns for Diabetes Mellitus, 2020

To calculate the information gain value on the pregnant attribute using the following equation.

$Gain\ Information\ (pregnant)$

$$Gain\ (S,A) = Entrophy\ (S) - \sum_{n=1}^{n} \frac{Ai}{S} * Entrophy\ (Ai)$$

$$Gain\ (S,A) = 0.933569067 - ((\frac{84}{538}) * 0.940285959) + (\frac{94}{538}) * 0.784992089) + (\frac{68}{538}) * 0.703926068) + (\frac{292}{538}) * 0.982023501)$$

$$Gain\ (S,A) = 0.027637632$$

As for the overall table, the calculation of information gain values for each attribute is seen in the following table.

Table 3. The Amount of Calculation for Each Attribute

		JML CASE	NO (S1)	YES (S2)	ENTROPY	GAIN INFORMATION
TOTAL		538	350	188	0.933569	
Pregnant						0.027638
	Nulligravida	84	54	30	0.940286	
	Primigravida	94	72	22	0.784992	
	Secungravida	68	55	13	0.703926	
	Multigravida	292	169	123	0.982024	
Glucose	Normal	390	299	91	0.783777	0.109802
	High	148	51	97	0.929148	
Preesure	Low	86	70	16	0.693127	0.019068
	Normal	328	207	121	0.949826	
	High	28	18	10	0.940286	
	Rather high	96	55	41	0.984604	
Triceps	Normal	392	242	150	0.959894	0.031311
	Low	97	82	15	0.621329	
	High	49	26	23	0.997294	
				0		

Insulin	Normal	472	318	154	0.911071	0.01167
	High	66	32	34	0.999338	
Mass	Deficiency	3	3	0	0	0.078084
	Normal	76	72	4	0.297472	
	Exaggerated	38	32	6	0.629249	
	Very excessive	421	243	178	0.982736	
				0		
Pedigree	Very low	131	97	34	0.826054	0.070213
	Low	222	148	74	0.918296	
	Is	110	67	43	0.965384	
	High	43	24	19	0.990225	
	Very high	32	11	21	0.928362	
Age	Productive	527	342	185	0.934997	0.031675
	Old	11	9	2	0.684038	

Source: Implementation of C5.0 Algorithm for Establishing Decision Tree Patterns for Diabetes Mellitus, 2020

This calculation phase is repeated until the leaves in the decision tree are fully formed. For an overview of the decision tree can be seen in the following picture.

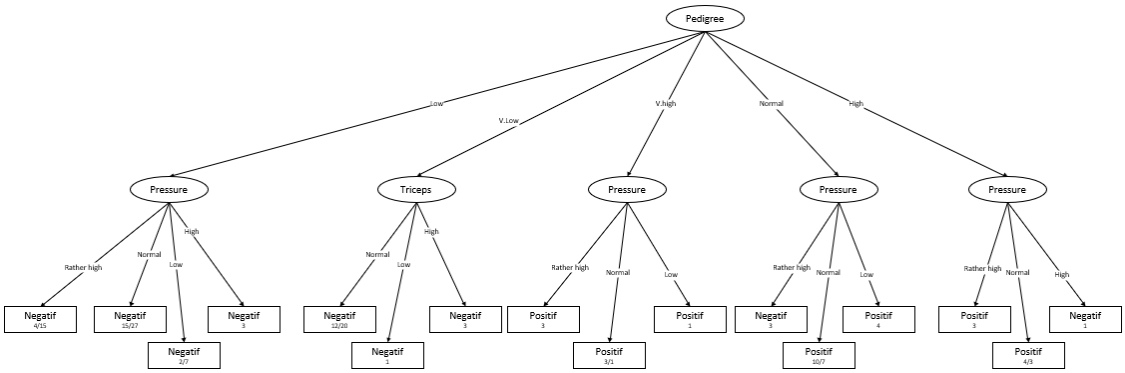

Figure 2. Decision tree level 4

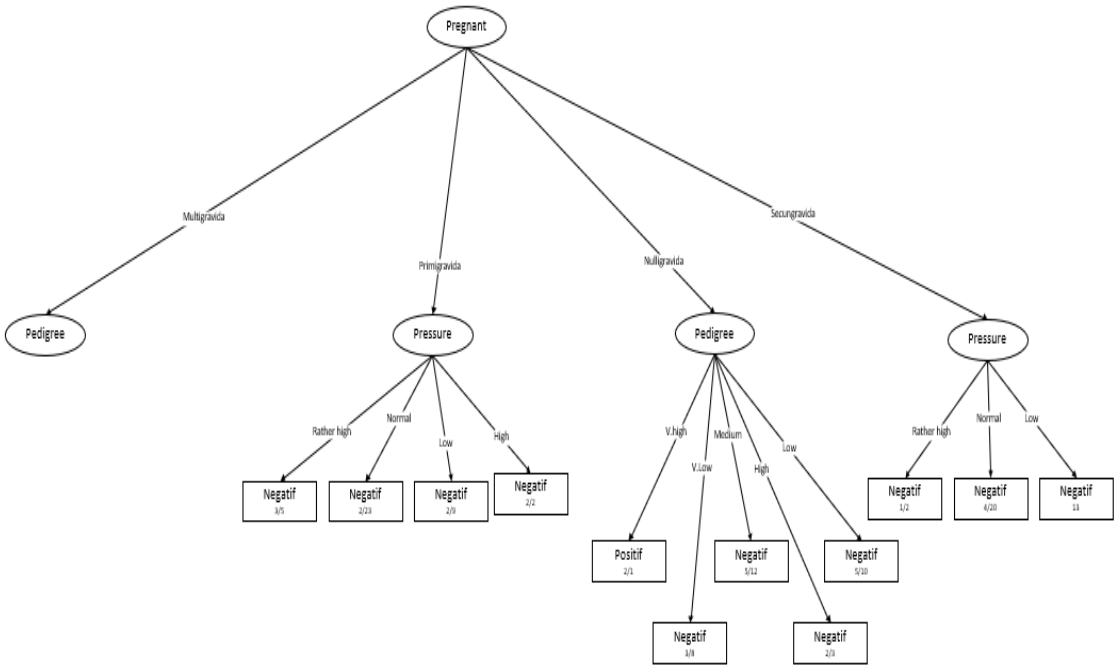

Figure 3. Level 3 decision tree on the left

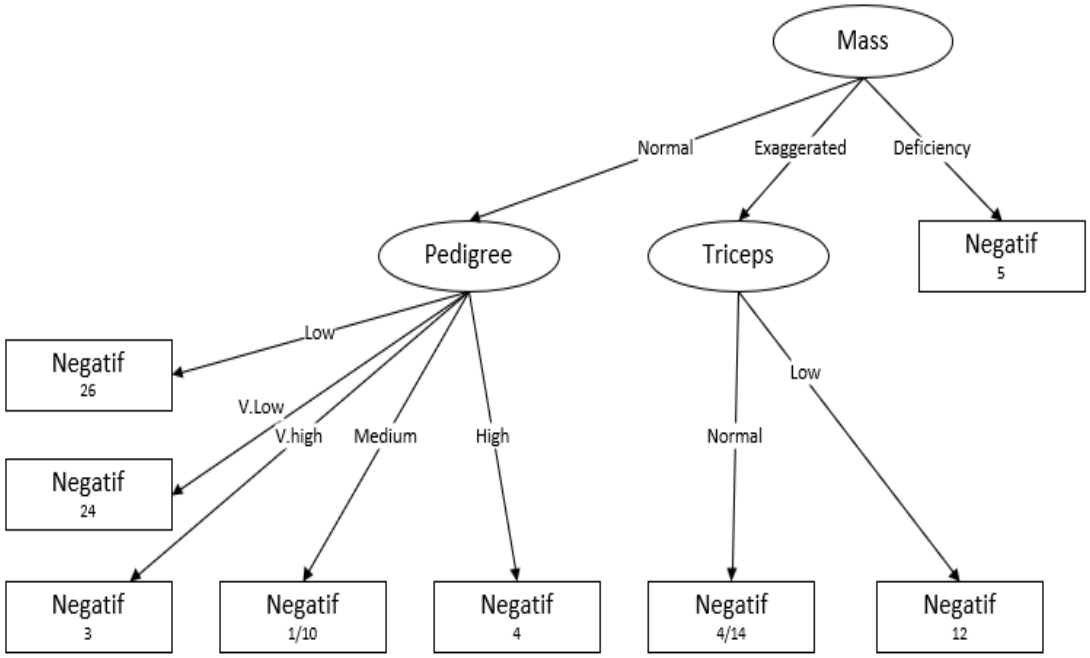

Figure 4. Level 2 decision tree on the left

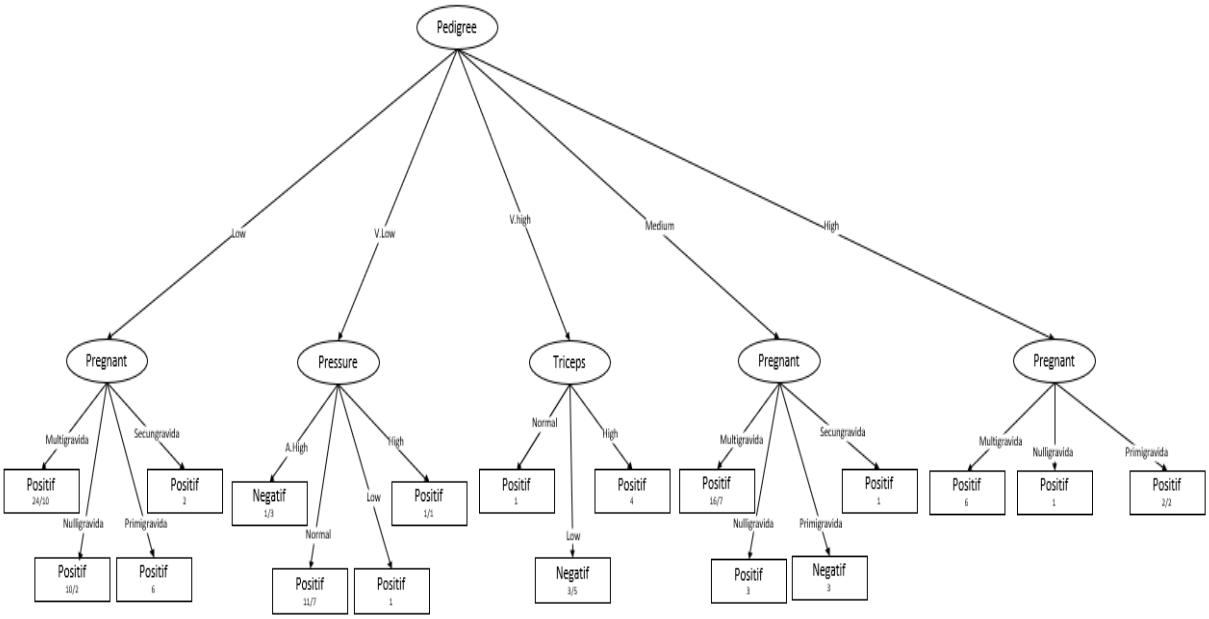

Figure 5. Level 3 decision tree on the right

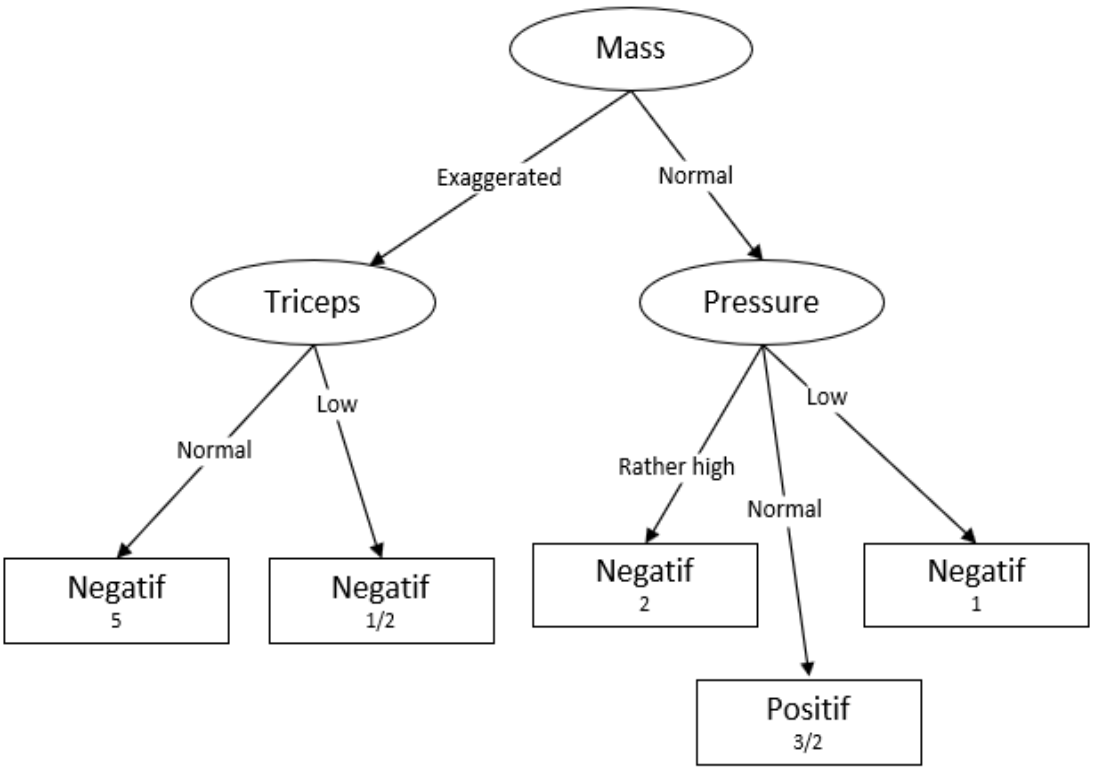

Figure 6. Right hand level 2 decision tree

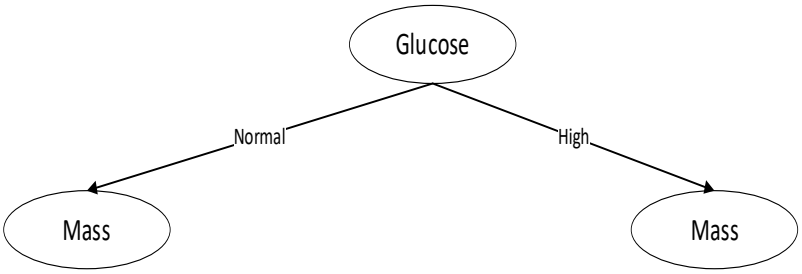

Figure 7. Decision tree level 1

After that a rule is formed which will be compared to the testing data. For rules can be seen in the following table.

Table 4. Diabetes Decision Tree C5.0 Dataset1

No	*Rule*	Decision
1	if glucose is high and mass is excessive and triceps is normal	negative
2	if glucose is high and mass is excessive and triceps is low	negative
3	if high glucose and normal mass and pressure are rather high	negative
4	if high glucose and normal mass and normal pressure	positive
5	if glucose is high and mass is normal and pressure is low	negative
6	if high glucose and mass are very excessive and low pedigree and pregnant mul tigravida	positive
7	if glucose is high and mass is very excessive and pedigree is low and pregnant is nulligravida	positive
.		
.		
59	if high glucose and mass are very excessive and pedigree is very high and tricep s is low	negative

Source: Implementation of C5.0 Algorithm for Establishing Decision Tree Patterns for Diabetes Mellitus, 2020

b. Evaluation

The evaluation phase is obtained by comparing the available testing data, with the decision tree pattern that has been formed from the calculation of the algorithm in the training data. For measuring the level of accuracy, precision, and recall of data that has been tested using the confusion matrix. Confussion matrix is based on a comparison of the amount of data that is considered right and wrong by the application, with the actual data available, then calculated in the form of a matrix.

	pregnant	glucose	pressure	triceps	insulin	mass	pedigree	age	diabe
2	Primigravi	Normal	Normal	Normal	Normal	Berlebiha	Rendah	Produktif	neg
3	Primigravi	Normal	Normal	Normal	Normal	Sangat Be	Sangat Re	Produktif	neg
4	Multigravi	Normal	Normal	Normal	Normal	Sangat Be	Sangat Re	Produktif	neg
5	Multigravi	Tinggi	Normal	Normal	Normal	Sangat Be	Sedang	Produktif	pos
6	Multigravi	Normal	Normal	Normal	Normal	Sangat Be	Rendah	Produktif	pos
7	Multigravi	Normal	Agak Ting	Tinggi	Tinggi	Sangat Be	Sedang	Produktif	neg
8	Multigravi	Normal	Normal	Normal	Normal	Sangat Be	Rendah	Produktif	pos
9	Multigravi	Normal	Normal	Tinggi	Tinggi	Sangat Be	Sangat Tin	Produktif	pos
10	Multigravi	Tinggi	Normal	Normal	Normal	Sangat Be	Rendah	Produktif	neg
11	Multigravi	Normal	Agak Ting	Normal	Normal	Sangat Be	Sedang	Produktif	neg
12	Nulligravi	Tinggi	Normal	Normal	Normal	Sangat Be	Sangat Tin	Produktif	pos
13	Primigravi	Normal	Rendah	Rendah	Normal	Normal	Rendah	Produktif	neg
14	Multigravi	Tinggi	Normal	Normal	Tinggi	Sangat Be	Rendah	Produktif	pos
15	Nulligravi	Normal	Normal	Tinggi	Normal	Sangat Be	Sangat Re	Produktif	neg
16	Multigravi	Normal	Normal	Normal	Normal	Sangat Be	Rendah	Produktif	pos
17	Multigravi	Normal	Normal	Normal	Normal	Normal	Sedang	Produktif	neg
18	Secungrav	Normal	Tinggi	Normal	Normal	Sangat Be	Tinggi	Produktif	neg
19	Multigravi	Normal	Agak Ting	Normal	Normal	Sangat Be	Sedang	Produktif	pos
20	Multigravi	Normal	Agak Ting	Rendah	Tinggi	Sangat Be	Sangat Re	Produktif	neg
21	Primigravi	Normal	Rendah	Rendah	Normal	Normal	Sangat Re	Produktif	neg
22	Multigravi	Normal	Normal	Normal	Normal	Sangat Be	Sedang	Produktif	neg
23	Multigravi	Normal	Tinggi	Normal	Normal	Sangat Be	Sangat Re	Produktif	pos

Table 5. Testing Data Testing

Source: Implementation of C5.0 Algorithm for Establishing Decision Tree Patterns for Diabetes Mellitus, 2020

Table 6. Performance Assessment C5.0

	Classification	Prediction		
		Negative	**Positive**	**total**
Prediction	**Negative**	133	37	170
	Positive	17	43	60
	total	150	80	230

Source: Implementation of C5.0 Algorithm for Establishing Decision Tree Patterns for Diabetes Mellitus, 2020

The decision of the actual data that has a "positive" value in all testing data amounts to 60 data. Decisions that were successfully identified by the C5.0 algorithm and have the same value as the decisions in the actual data obtained 43 data. While that is not the same as the C5.0 algorithm of 17 data. For data that has a "negative" value in the actual available data, 37 data are not in accordance with the C5.0 algorithm and 133 data are in accordance with the C5.0 Algorithm. For the level of accuracy, precision, and recall of data testing. For an overall assessment table the testing data can be seen as follows

Table 7. C5.0 performance results of diabetes mellitus datasets 1

Accuracy	Precision	Recall	Number of Rule
76.52%	78.24%	88.67%	59

3.2 Discussion

The same calculations are made for datasets 2, 3 and 4 as dataset 1 so that they can be made in table 8.

Table.8 C5.0 performance results

Training Data	Accuracy	Precision	Recall	Number of Rule
90%	72.73%	75.44%	86.67%	70
80%	74.03%	75.86%	88.00%	64
70%	76.52%	78.24%	88.67%	59
60%	74.59%	78.77%	83.50%	53

Source: Implementation of C5.0 Algorithm for Establishing Decision Tree Patterns for Diabetes Mellitus, 2020

This study aims to determine the performance of the classification of C5.0 to predict the risk of diabetes. The dataset used is a diabetes dataset which has 9 variables. There are two predicted classes: Positive and Negative. There are 768 total data, with 268 patients with diabetes and 500 negative. Before entering the data mining stage, the dataset is obtained first through the preprocessing stage. The missing value is filled with the mean value for each attribute that has the missing value. Furthermore the dataset is divided into training data and testing data. In this study a total of 768 data were divided into 70% training data and 30% testing data. The training data has 538 data with 350 negative class divisions and 188 positive data. Whereas for testing data as many as 230 data with 150 negative class divisions and positive data as many as 230 data Classification performance was evaluated using a confusion matrix.The researcher conducted the C5.0 classification to find the accuracy results by conducting several experiments. It can be seen in Table 8. With 90% training data and 10% testing data get an accuracy value of 72.73% and a precision value of 75.44% and a recall value of 86.67% in this experiment the rule produced 70 rules. With 80% training data and 20% testing data get an accuracy value of 74.03% and a precision value of 75.86% and a recall value of 88.00% in this experiment the resulting rule is as much64 rule. With 70% training data and 30% testing data get an accuracy value of 76.52% and a precision value of 78.24% and a recall value of 88.67% in this experiment the resulting rule is as much59 rule. With 60% training data and 40% testing data get an accuracy value of 74.59% and a precision value of 78.77% and a recall value of 83.50% in this experiment the resulting rule is as much53 rule. It can be seen in Table 8 that the best accuracy is found in experiments with 70% training data and 30% testing data. The table also shows the number of rules determined by the amount of training data. It can be concluded that the more training data, the more rules will be generated

4. CONCLUSION

Conclusions that can be taken in this study are:

a. The number of rules in trials with training data of 90% has as many as 70 rules. In trials with training data at 80%, the number of rules is 64. In trials with training data at 70%, there are 59 rules. 60% have 53 rules. It can be concluded that the more training data, the more rules will be produced

b. With 90% training data and 10% testing data get an accuracy value of 72.73% and a precision value of 75.44% and a recall value of 86.67%. With 80% training data and 20% testing data get an accuracy value of 74.03% and a precision value of 75.86% and a recall value of 88.00. With 70% training data and 30% testing data get an accuracy value of 76.52% and a precision value of 78.24% and a recall value of 88.67%. With 60% training data and 40% testing data get an accuracy value of 74.59% and a precision value of 78.77 % and recall value of 83.50%. of all experiments, the best accuracy is found in trials with training data at 70% and testing data at 30%

BIBLIOGRAPHY

[1] Latief, Muhammad. 2020. "Implementation of C5.0 Algorithm for Forming Decision Tree Patterns for Diagnosis of Diabetes Mellitus".Thesis of Computer Science Study Program, Lambung Mangkurat University, Banjarbaru.

[2] Annisa, Nabillah .2018.**"Implementation of C 5.0 Algorithm to Analyze Priority Symptoms in Children Who Have Bullying"**Informatics Engineering. Malang Muhammadiyah University

[3] Kantardzic, M. 2003. "***Data Mining, Concepts, Models, Methods, and Algorithms***". IEEE Press.

[4] Dunham, MH 2003."Data Mining Introductory and Advanced Topics". Prentice Hall: New Jersey.

[5] Silvia, Dian, et al. 2017. "**. Identification of Diabetes Mellitus Using Modified K-Nearst Neighbor (MKNN) Method.**". Informatics Engineering. Universitas Brawijaya, Vol.1, No.6, June 2017, ISSN: 2548-964X