Improving the Accuracy of the C45 Classification Algorithm Using Information Gain Ratio Feature Selection for Classification of Type 2 Diabetes Mellitus Disease

Ivandari [1], Much. Rifqi Maulana[2], Ichwan Kurniawan[3], M. Adib Al Karomi[4*]
[1-3] *Computer science, STMIK Widya Pratama Pekalongan, Indonesia*

Abstract— Diabetes is a disease that can cause death. Diabetes can cause heart failure, chronic kidney disease, glaucoma that attacks the eyes and several other diseases. WHO data states that there were more than 2 million deaths due to diabetes in 2019. Data from the International Diabetes Federation shows that around 537 adults are recorded as living with diabetes. This condition must be treated immediately, considering that diabetes is one of the most deadly non-communicable diseases in the world. Patient registration is mostly done in hospitals. A lot of data will only become digital waste if it does not have more benefits. In 2020 Diabetes and Hospital in Sylhet donated patient data for further research. This data contains 520 patient records with 17 attributes that have been validated by specialist doctors. Early stage diabetes risk prediction data is released by the uci repository as public data and can be used for research testing. Research using this dataset has been widely carried out with the previous best accuracy level of 95.96%. In previous studies, all attributes were used in the classification process. The number of irrelevant attributes can affect the performance of the classification algorithm. This study uses the information gain ratio for feature selection of the early stage diabetes risk prediction dataset. The C45 algorithm is used for classification, evaluation using confusion matrix and validation using 10 folds cross validation. The results of this study improve the performance of C45 so that it obtains an accuracy level of 96.15%. This study also produces a decision tree for diabetes..

Index Terms— information gain ratio, decision tree, diabetes type 2.

1. Inrtoduction

There are many changes in the lifestyle of modern society, especially related to diet and exercise. This reduces the endurance of some people and causes several diseases. One of the non-communicable diseases that kills many people is diabetes [1].. Diabetes is characterized by increased blood glucose levels so that the body cannot control excess glucose in the blood [2]. Significant and sustained increases in blood glucose can cause serious damage to blood vessels, heart, kidneys, eyes and nerves.

It was recorded that in 2019 there were 2 million deaths worldwide due to diabetes [3]. From the International Diabetes Federation (IDF) it was recorded that 537 million adults (aged 20-79 years) were living with diabetes [4]. The IDF also predicts a significant increase in 2045. In Indonesia, cases of diabetes in adolescents under the age of 15 have been found [5]. Handling the disease will be easier if the diagnosis is made earlier.

Current technological developments allow artificial intelligence to analyze data to obtain useful information in the future. Machine learning in artificial intelligence is often done using existing data. This process is also known as data mining. Data mining can search for new knowledge patterns from existing data sets. The patterns formed from the mathematical calculation process can then also be used as a model or output from an algorithm, for example in a decision tree [6]. In this data mining process, an evaluation is then carried out to determine the best algorithm that can be used for a data set.

The trend of data processing using data mining is widely used by researchers to solve various problems. Various processes in data mining include association, estimation, classification, clustering and prediction. Some of these processes can handle different data and problems. Classification is one of the data mining processes that utilizes existing datasets to complete new datasets. In the classification of existing datasets, it is mandatory to have a label attribute or a purpose attribute. The use of data mining classification in the health sector is widely used, several studies of data mining classification in the health sector include early detection of chronic kidney disease [7], breast cancer detection [8] and detection of diabetes [9].

Classification of type 2 diabetes using data mining has been done before. The data used is the PIMA dataset which is public data. This dataset is published by the uci repository. The uci repository is a provider of public datasets and is widely used by researchers for algorithm testing. In 2023, the Gradient Boost Machine (GBM) algorithm outperformed other algorithms with an accuracy rate of 80.4% in a comparison of diabetes classification using the PIMA dataset [10]. In previous studies, the classification of type 2 diabetes was also carried out using

the C45 algorithm C45 [11]. In this study, the C45 algorithm obtained an accuracy rate of 95.96%.

Decision tree is an algorithm output model that is widely used and proven to be good for classification. C45 is the basis for calculating algorithms with decision tree output. In its calculation, the C45 algorithm uses the gain value to determine the first node in the decision tree. The attribute with the largest gain value is the attribute with the highest importance in a classification. Another advantage of using C45 is that the output produced can be easily understood by human language.

One of the determinants of the success of an algorithm's classification is the type of data used [6]. The number of data attributes does not guarantee better classification. Data attributes that have small gains can reduce the performance of an algorithm. In addition, the number of irrelevant data attributes can increase the computational process [12]. The selection of relevant attributes in the classification process can be done during pre-processing. The pre-processing model that is widely used and proven to be good is feature selection [13]. One of the best feature selection algorithms is the information gain ratio [14]. The information gain ratio is a development of the previous model, namely information gain.

This study classifies type 2 diabetes using the C45 algorithm. The dataset used is the PIMA dataset from the uci repository. The results of the study prove that the feature selection process using the information gain ratio can improve the performance of the C45 algorithm in classifying type 2 diabetes. The accuracy level of C45 is 95.96 without using feature selection. The largest increase in algorithm accuracy is obtained by applying a gain value of 0.01. In this process, one attribute, namely "itching", is eliminated because it has a gain of 0. The best accuracy level obtained is 96.15% using only 15 regular attributes.

2. Literature review

2.1. Related research

The discussion of diabetes classification using the PIMA dataset has been discussed several times in research. In 2021, a comparison of several algorithms was conducted. The result was that the soft voting classifier was the algorithm with the best accuracy rate, namely 79.08% [15]. In 2023, a similar classification was carried out by Carpinteiro from Portugal. From the results of his research, the Gradient Boost Machine (GBM) algorithm obtained an accuracy rate of 80.4%. Furthermore, in the same year, a type 2 diabetes classification test was carried out using the PIMA dataset. The result was that K-NN had an accuracy rate of 92.5% [16], and C45 had an accuracy rate of 95.96% % [11].

2.2. Data Mining Classification

Digital data is very abundant today. Existing digital data will only become digital waste if the data is not very useful. Data mining answers this problem by processing data into new knowledge [17].. This new knowledge is obtained from the existence of patterns or models created by an algorithm. These patterns or models can be used as a reference in a prediction, estimation or classification [18].

Classification is an important part of data mining that is widely developed. The classification model is basically a mathematical and statistical calculation. In classification, there is training data that is used as a reference for the next process. If the model or pattern has been created from the training data, the next step is to calculate the testing data. The dataset used in classification must have a label or target attribute. The label attribute of the testing data is what will later be searched for based on the model or pattern that was formed previously. The many calculation methods that can be used in the classification process make researchers continue to search for and develop new methods. In a classification calculation, the suitability of the dataset and algorithm greatly affects the classification results. One of the best and most widely used classification algorithms is the decision tree.

2.3. C45 algorithm

The C45 algorithm is one of the best classification algorithms. C45 is widely used in classification and has been proven to be able to handle high-dimensional datasets. In addition, C45 can also process data with numeric or nominal types. The model output from C45 is a decision tree. This makes C45 one of the most popular and used algorithms in research.

The calculation process in the C45 algorithm is by calculating the importance value (gain) of all existing attributes. The first branch of the existing decision tree is the attribute with the highest gain value. The number of branches adjusts to the content or variation of the attribute. The number of attributes in the dataset can affect the size of the decision tree formed. In addition, the number of attributes also affects the number of calculation iterations. This iteration will be completed if all records in the dataset have received labels or results. Overall, the C45 algorithm is very good for nominal data types. Figure 1 is an example of a decision tree from a dataset with 3 data attributes.

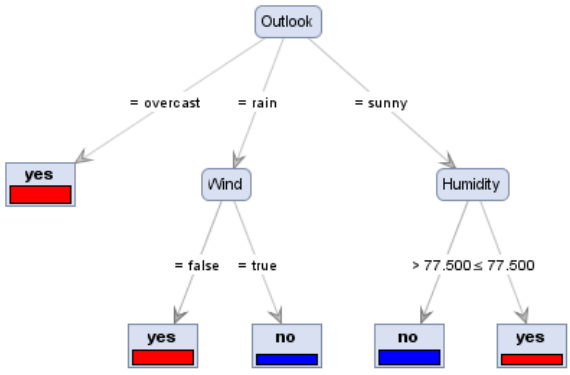


Figure 1. Example of a Decision Tree

From Figure 1, it can be seen that there are 3 attributes, namely outlook, wind and humidity. The outlook attribute is the first branch point. This means that in the calculation of the outlook attribute, it has the highest gain value of the three attributes. In the outlook attribute, there are 3 data variants, namely overcast, rain and sunny. All three are used as branches of the outlook attribute. From the existing data, the overcast variant in the outlook attribute all have the YES label. This means that for the overcast branch in the outlook, the classification results have been found. For the results that have been found, the tree branch will stop. Next, for the rain branch in the outlook attribute, the calculation is repeated. This repetition only uses the remaining attributes other than the outlook attribute, the result is that the wing attribute has a greater gain value than the humidity in the rain branch. From the wind attribute, there are only 2 variants, namely false and true. This means that there are only 2 branches formed from the wind attribute. This repetition is continued until all records in the dataset have their respective classification results. This repetition can also stop if the data attribute no longer exists.

2.4. Information gain ratio feature selection

Feature selection is a process of selecting attributes in a dataset that are considered relevant in the data mining process. The number of attributes used will slow down the computational process. If many of the attributes mentioned are irrelevant, the accuracy of the classification will decrease [17]. Basically, feature selection is removing irrelevant and redundant attributes in a dataset. This process will clearly reduce the data dimension which can later speed up the computational process. With feature selection, it also allows the classification algorithm to work faster and more effectively and removing irrelevant features allows for increased accuracy of an algorithm..

One of the widely used feature selection algorithms is Gain Ratio. Gain ratio is a development of Information gain which was first introduced by Quinlan in the C4.5 system. One of the shortcomings of information gain is bias towards data with many variants while in gain ratio to overcome the bias of information gain a type of normalization is used, namely split information. Gain ratio is the result of dividing the information gain value by the split information value. The following equation is a formula for calculating the value of the gain ratio:

$$\textbf{Gain Ratio}\ (A) = \frac{Gain\ (A)}{SplitInfo\ (A)}$$

Description:
Gain (A) : *Information gain* attribute A
Gain ratio (A) : *Gain ratio* attribute A
Split Info (A) : *Split Info* attribute A

Before calculating to find out the gain ratio value, the information gain value and the split information value of an attribute are calculated. Information gain is one of the feature selection methods widely used by researchers to determine the limits of the importance of an attribute [19] [20]. The information gain value is obtained from the entropy value before separation minus the entropy value after separation [13]. In calculating the previous information gain value, the total entropy value and the entropy value of the attribute in question must first be known. The information gain value is the difference between the total entropy value and the entropy of the attribute to be calculated.

The measurement of the importance of an attribute was first pioneered by Claude Shannon in information theory and is written according to the following equation:

$$info(D) = -\sum_{i=1}^{m} pi\ log_2(pi)$$

Description:
D : Case set
M : Number of partitions D
pi : Proportion of Di to D

While pi is the probability of a tuple in D that falls into class Ci and is estimated by |Ci,D| / |D|. The log function in this case uses base 2 log because the information is encoded based on bits. The calculation of the entropy value after separation can be done using the following formula equation:

$$info_A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|}\ \text{x}\ I\ (D_j)$$

Description:
D : Case set
A : Attribute
v : Number of partitions of attribute A
|Dj| : Number of cases in partition j
|D| : Number of cases in D
I (Dj) : Total entropy in partition

Meanwhile, to find the information gain of attribute A, the following formula equation can be used:

$$\textbf{Gain (A) = I (D) – I (A)}$$

Description:
Gain (A) : Information gain of attribute A
I (D) : Total entropy
I (A) : entropy A

After the information gain value is known, the split info value of attribute A can be found using the following formula equation:

$$SplitInfo_A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \times log_2\left(\frac{|D_j|}{|D|}\right)$$

Description:
D : Case set
A : Attribute
V : Number of partitions of attribute A
|Dj| : Number of cases in partition j
|D| : Number of cases in D

Furthermore, the split info value of attribute A can be used as a divisor of the info value of attribute A to obtain the gain ratio value of attribute A as previously explained in equation. The same calculation is carried out for each existing attribute to obtain the gain ratio value for all existing attributes. After the overall gain ratio value is known, the next stage is to select the gain ratio value limit for all attributes. In this case, the gain ratio value that does not meet the criteria will not be included in the classification process.

In general, the feature selection stage using the gain ratio is carried out in 3 stages, namely:

1. Calculate the gain ratio value for each attribute in the original dataset.
2. Determine the desired threshold. This will allow attributes that are weighted the same as the limit or greater to be retained and discard attributes that are below the limit.
3. The dataset is repaired by only using attributes with a gain ratio value above the specified limit.

2.5. 10 folds cross validation

Validation is one of the most important processes in classification. The validation process is carried out to ensure that the division of training data and testing data is more focused and measurable. There are several validation models that are commonly used in the classification process. One validation model that has been proven to be good for use in the classification process is cross validation [21]. Cross validation works by dividing the data into two parts, namely testing data and training data. The division of training data and testing data is carried out according to the needs of the data proportion. The division of data with the best proportion is to divide the data into 10 parts. Of the 10 parts, 9 parts are used as training data and 1 part as testing data. This process is widely used in classification research and is hereinafter referred to as 10 folds cross validation [22]. Figure 2 is the validation process of early stage diabetes risk prediction data using 10 folds cross validation.

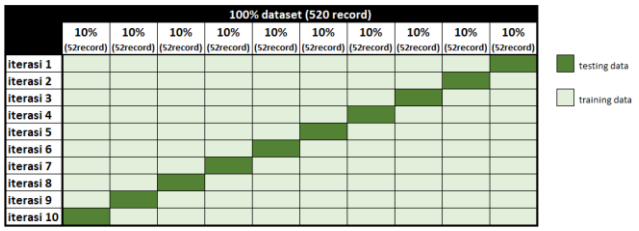


Figure 2. Representation of 10 folds cross validation

Figure 2 represents 10 iterations in the 10-fold cross validation process. The early stage diabetes risk prediction dataset has 520 records. The dataset is divided randomly by considering several rules. For testing data records must be used once. may not be used more than 1 time. In this study, the testing data used was 52 records (10% of the total data records).

2.6.Confusion matrix

The evaluation process in classification is a factor in measuring the level of accuracy of an algorithm. In this evaluation process, the performance of an algorithm in classifying data can be known. In this study, a confusion matrix was used for the evaluation process. The confusion matrix is basically comparing the actual labels of the testing data with the classification algorithm recommendation labels [23]. This process allows a comparison of the number of classification algorithm recommendations with the actual labels of the testing data. The number of these comparisons will later be converted into a percentage which is a measure of the level of accuracy of a classification algorithm.

Figure 3 is a picture of the calculation process in the confusion matrix [24]. The calculation process of the confusion matrix can be done using the following formula.

Classification		Predicted Class	
		Class = Yes	Class = No
Observed Class	Class = Yes	a (true positive-TP)	b (false negative-FN)
	Class = No	c (false positive-FP)	d (true negative-TN)

Figure 3. *confusion matrix* [24]

$$\boldsymbol{accuracy} = \frac{\boldsymbol{a+d}}{\boldsymbol{a+b+c+d}} = \frac{\boldsymbol{TP+TN}}{\boldsymbol{TP+FN+FP+TN}}$$

a = (*true positive*) = correct classification
b = (*false negative*) = incorrect classification
c = (*false positive*) = incorrect classification
d = (*true negative*) = correct classification

3. Reseach Methods

This study uses an experimental research method. In the process of calculating the C45 algorithm and calculating the information gain ratio, the rapid miner application is used. Rapid miner is used because it is an open source that is widely used in data mining research. In addition to being easy to use, the tools in rapid miner can also read many types of existing datasets. This research process is divided into several parts as follows:

3.1. Data Collection

In conducting a measurement of a study, the same object can be used. The similarity of research objects can differentiate the results of one study from another. This study uses a public dataset released by the uci repository. This diabetes dataset can be used as a comparison of one research process with another. To access the dataset, you can use the following link: https://archive.ics.uci.edu/dataset/529/early+stage+diabetes+risk+prediction+dataset. The early stage diabetes risk prediction dataset has 520 data records. This dataset has a total of 17 data attributes. One of the 17 attributes is a label

attribute or a destination attribute. Table 1 is metadata from the early stage diabetes risk prediction dataset.

3.2. Validation

The validation process in this study uses 10 folds cross validation. This process is carried out because cross validation has been proven to be good for the classification process. The calculation uses the rapid miner application. The validation application process in rapid miner can be done by dragging and dropping in the worksheet.

3.3. Information gain ratio

This process is carried out before the classification process is carried out. The information gain ratio algorithm is one of the best feature selection algorithms today [25]. The calculation process in the information gain ratio is to divide the gain value of all attributes by the split information. By using the information gain ratio, the importance of all existing data attributes can be known. The attribute with a gain ratio value of 0 can then be removed. This process can speed up the calculation of the classification algorithm. In research, it is also possible to use a gain ratio value limit. This limit is done by estimating how many attributes will be used in the next classification process. In this study, the limit used is 0.001. This limit is used considering that the second smallest gain ratio value is 0.004. The complex process and calculations are carried out using rapid miner.

3.4. C45 Calculation

This study uses the C45 algorithm for the data classification process. The C45 algorithm is used because it has proven to be good and strong for classification cases. The output of the C45 algorithm can be in the form of a decision tree. This decision tree can be easily understood and is in accordance with human language. The calculation process in C45 is to find the largest gain value from all existing data attributes. The attribute with the largest gain value is used in the decision tree hermit node. This node has as many branches as the variations of the attribute. This process is repeated until all data has a classification result. In this study, the C45 algorithm is implemented using rapid miner.

3.5. Algorithm Evaluation

The evaluation process in this study uses a confusion matrix. The confusion matrix is used to measure the level of algorithm accuracy. The level of accuracy displayed is the average value of the overall accuracy level of the research iteration. The percentage is the result of dividing the appropriate classification by the total number of trials. This process is carried out using rapid miner.

4. Result and Discussion

4.1. Data Collection

This study uses the early stage diabetes risk prediction dataset. This data is a public dataset from the uci repository. The public dataset is deliberately used in research to compare one method with another. This dataset can be used by all researchers. The complete dataset can be downloaded at: https://archive.ics.uci.edu/static/public/529/early+stage+diabetes+risk+prediction+dataset.zip. This dataset has been validated and approved by specialist doctors and has been donated in 2020 [26]. The number of records in this dataset is 520 according to the number of patients from the previous recording results. In the recording, there are 16 attributes used. More detailed metadata from the dataset can be seen in table 1 below. The entire dataset can be seen in table 2.

From the early stage diabetes risk prediction metadata above, 17 attributes can be seen, one of which is the target attribute or class attribute. The proportion in the class attribute shows that 320 records are positive and 200 records are negative. From the dataset, there are also no missing values for all existing attributes. In table 2, the dataset display is cropped. This is done considering the number of data records as many as 520 which can take up many journal pages.

Tabel 1. Metadata early stage diabetes risk prediction

Role	Name	Type	Statistics	Range	Missings
label	class	binominal	mode = Positive (320), least = Negative (200)	Positive (320), Negative (200)	0.0
regular	Age	integer	avg = 48.029 +/- 12.151	[16.000 ; 90.000]	0.0
regular	Gender	binominal	mode = Male (328), least = Female (192)	Male (328), Female (192)	0.0
regular	Polyuria	binominal	mode = No (262), least = Yes (258)	No (262), Yes (258)	0.0
regular	Polydipsia	binominal	mode = No (287), least = Yes (233)	Yes (233), No (287)	0.0
regular	sudden weight loss	binominal	mode = No (303), least = Yes (217)	No (303), Yes (217)	0.0
regular	weakness	binominal	mode = Yes (305), least = No (215)	Yes (305), No (215)	0.0

regular	Polyphagia	binominal	mode = No (283), least = Yes (237)	No (283), Yes (237)	0.0
regular	Genital thrush	binominal	mode = No (404), least = Yes (116)	No (404), Yes (116)	0.0
regular	visual blurring	binominal	mode = No (287), least = Yes (233)	No (287), Yes (233)	0.0
regular	Itching	binominal	mode = No (267), least = Yes (253)	Yes (253), No (267)	0.0
regular	Irritability	binominal	mode = No (394), least = Yes (126)	No (394), Yes (126)	0.0
regular	delayed healing	binominal	mode = No (281), least = Yes (239)	Yes (239), No (281)	0.0
regular	partial paresis	binominal	mode = No (296), least = Yes (224)	No (296), Yes (224)	0.0
regular	muscle stiffness	binominal	mode = No (325), least = Yes (195)	Yes (195), No (325)	0.0
regular	Alopecia	binominal	mode = No (341), least = Yes (179)	Yes (179), No (341)	0.0
regular	Obesity	binominal	mode = No (432), least = Yes (88)	Yes (88), No (432)	0.0

Tabel 2. Early stage diabetes risk prediction dataset

No	Class	Age	Gender	Poly uria	Poly dipsia	Sudden weight loss	Weak ness	Poli phagia	Genital thrush	Visual blurring	Irrita bility	Delay healing	Partial paresis	Muscle stiffness	Alo pecia	Obecity
1	Positive	40	Male	No	Yes	No	Yes	No	No	No	No	Yes	No	Yes	Yes	Yes
2	Positive	58	Male	No	No	No	Yes	No	No	Yes	No	No	Yes	No	Yes	No
3	Positive	41	Male	Yes	No	No	Yes	Yes	No	No	No	Yes	No	Yes	Yes	No
4	Positive	45	Male	No	No	Yes	Yes	Yes	Yes	No	No	Yes	No	No	No	No
5	Positive	60	Male	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes
6	Positive	55	Male	Yes	Yes	No	Yes	Yes	No	Yes	No	Yes	No	Yes	Yes	Yes
7	Positive	57	Male	Yes	Yes	No	Yes	Yes	Yes	No	No	Yes	Yes	No	No	No
8	Positive	66	Male	Yes	Yes	Yes	Yes	No	No	Yes	Yes	No	Yes	Yes	No	No
9	Positive	67	Male	Yes	Yes	No	Yes	Yes	Yes	No	Yes	No	Yes	Yes	No	Yes
10	Positive	70	Male	No	Yes	Yes	Yes	Yes	No	Yes	Yes	No	No	No	Yes	No
11	Positive	44	Male	Yes	Yes	No	Yes	No	Yes	No	Yes	Yes	No	Yes	Yes	No
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
509	Negative	58	Male	No	No	No	Yes	No	No	No	No	Yes	No	Yes	Yes	No
510	Negative	54	Male	No	No	No	No	No	No	No	No	No	No	No	No	No
511	Negative	67	Male	No	No	No	Yes	No	No	No	No	Yes	No	No	Yes	No
512	Negative	66	Male	No	No	No	Yes	Yes	No	Yes	No	Yes	Yes	Yes	Yes	No
513	Negative	43	Male	No	No	No	No	No	No	No	No	No	No	No	Yes	No
514	Positive	62	Female	Yes	Yes	Yes	Yes	No	No	Yes	No	No	Yes	No	No	Yes
515	Positive	54	Female	Yes	Yes	Yes	Yes	Yes	No	No	No	No	Yes	No	No	No
516	Positive	39	Female	Yes	Yes	Yes	No	Yes	No	No	No	Yes	Yes	No	No	No
517	Positive	48	Female	Yes	Yes	Yes	Yes	Yes	No	No	Yes	Yes	Yes	No	No	No
518	Positive	58	Female	Yes	Yes	Yes	Yes	Yes	No	Yes	No	No	Yes	Yes	No	Yes
519	Negative	32	Female	No	No	No	Yes	No	No	Yes	No	Yes	No	No	Yes	No
520	Negative	42	Male	No	No	No	No	No	No	No	No	No	No	No	No	No

4.2. Information gain ratio results

The next process after obtaining the dataset is the calculation of the information gain ratio. Figure 4 is the process carried out in rapid miner. The left part is a dataset with 16 regular attributes. The middle part is the process of calculating the information gain ratio. There are two outputs from the information gain ratio. The first output is wei, which is the weight of all existing attributes. Next is exa, which is the overall dataset output. The output from the next information gain ratio is selected based on the specified limits. From this limit, validation and evaluation of the algorithm are then carried out. Table 3 is the result of the calculation of the information gain ratio. Of the 16 regular attributes, there is one attribute with a gain ratio value of 0, namely "itching".

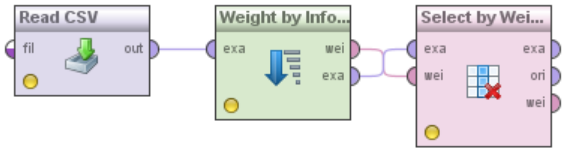


Figure 4. Gain ratio worksheet process

Tabel 3. Gain ratio values for all attributes

Attribute name	Gain ratio
Itching	0.0
delayed healing	0.00404705734656098
Obesity	0.010278293437522311
Genital thrush	0.024623363914229954
muscle stiffness	0.029946379622923937
Age	0.06148485683721434
weakness	0.11746635591463185
visual blurring	0.1283458061651502
Alopecia	0.14093030028329173
Irritability	0.2008816914574173
Polyphagia	0.24221999998261234
partial paresis	0.3991022428656577
sudden weight loss	0.41047853676228374
Gender	0.4509279896648905
Polydipsia	0.9911776302313886
Polyuria	1.0

4.3. Validation and Evaluation of Results

The validation process using 10 folds cross validation is shown in Figure 5. The input of cross validation is a dataset that has been previously selected for features. This process is carried out using the rapid miner tool. The cross validation attribute is given a value of 10. This value of 10 means that the dataset is divided into 10 parts. The process is carried out 10 times so that all data records have one chance to become testing data. This process is called 10 folds cross validation.

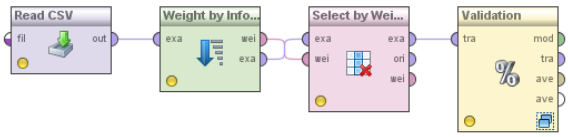


Figure 5. Rapid miner cross validation process

Evaluation is carried out using cross validation to calculate the accuracy of C45. The evaluation process is carried out in the validation process. Figure 6 is a process carried out in one validation process. The evaluation process is carried out with 10 iterations as the value filled in the cross validation attribute.

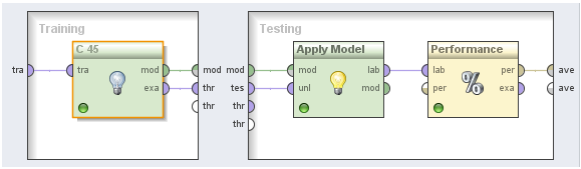


Figure 6. Evaluation process in rapid miner

Table 4 is the result of the confusion matrix from the research conducted. The C45 accuracy level of this study is 96.15%. This value is obtained from the calculation of the classification that matches the original label divided by the total number of data records. From table 4, it can be seen that there are 500 data records that match the actual label. And there are only 20 data records that do not match.

Tabel 4. Results of the C45 confusion matrix

	True positive	True negative	Class precission
Pred. positive	307	7	97,77%
Pred. negative	13	193	93,69%
Class recall	95,94%	96,50%	

The best level of accuracy was obtained by conducting several experiments changing the limits on the information gain ratio algorithm. Table 5 is the level of accuracy of C45 with several attribute limits used.

Tabel 5. Accuracy of C45 with several limits

Batasan gain ratio	Jumlah atribut	Akurasi (%)
0.0	16	95,96
0.00404705734656098	15	96,15
0.010278293437522311	14	95,58
0.024623363914229954	13	95,58
0.029946379622923937	12	94,81
0.06148485683721434	11	95
0.11746635591463185	10	93,63
0.1283458061651502	9	93,46

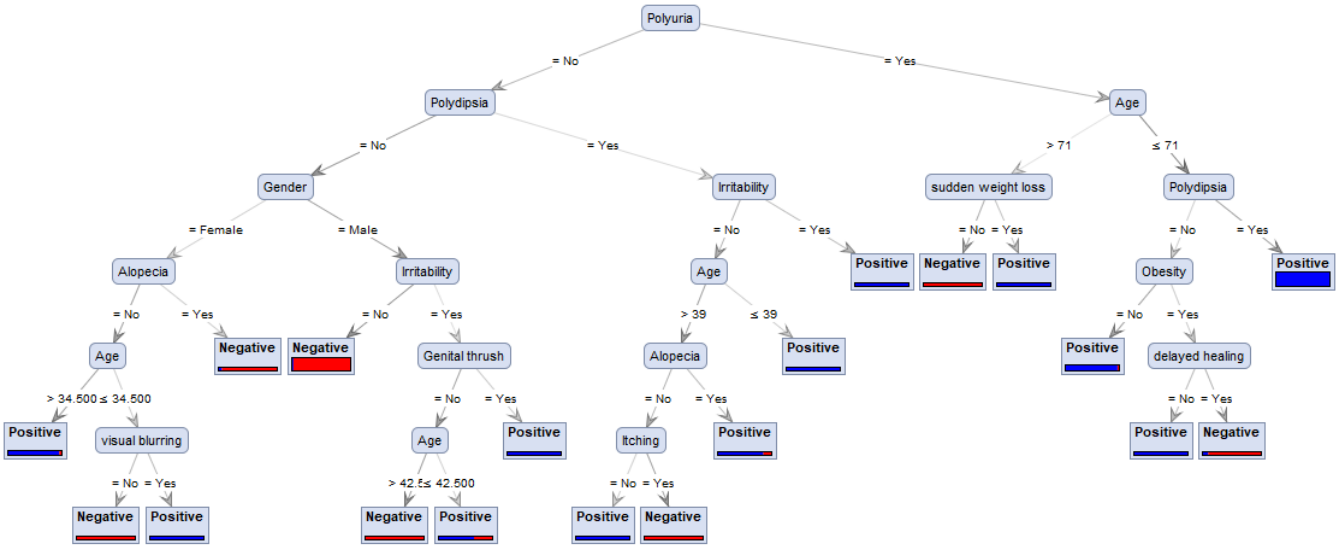
0.14093030028329173	8	91,73
0.20088169145741 73	7	89,04
0.24221999998261 234	6	89,81
0.39910224286565 77	5	90,19
0.41047853676228 374	4	89,42
0.45092798966489 05	3	89,43
0.99117763023138 86	2	86,92
1.0	1	82,31

4.4. Decision Tree Representation

After completing the C45 algorithm calculation, an accuracy level of 96.15% was obtained. Another output of the C45 algorithm is a decision tree. Figure 7 is a decision tree formed from the classification of diabetes using C45. From Figure 7, it can be seen that the first node in the decision tree is "polyuria". This result is in accordance with the calculation of the information gain ratio in table 3 which states that "polyuria is the attribute with the highest gain value. There are 2 branches of "polyuria" as the variation of the attribute. From the "no" variation group in the "polyuria" attribute, the information gain ratio value is calculated again. It was found that "polydipsia" is the attribute with the next highest gain ratio value. This branch continues to form a decision tree as in Figure 7 below.

4.5. Discussion

This study uses C45 to classify diabetes. The output of the C45 algorithm is a decision tree as shown in Figure 7. The best accuracy level obtained is 96.15%. This accuracy increases after previously conducting feature selection using the information gain ratio. This feature selection process can select several attributes with low gain ratio values. The most important stage is determining the limit of the information gain ratio. This limit value determines how many attributes will be used in the classification process. The limit value used in this study is 0.001. After this limit value is applied, one attribute is eliminated. The attribute "itching" is an attribute that is eliminated because it has a gain ratio value of 0. The attribute above it, namely "delayed healing", is still included in the classification process because it has a gain ratio value of 0.004047. This research process was carried out several times with several experiments on the gain ratio limit values. Table 5 contains several experiments with the limit values used. The best accuracy level was obtained in the second experiment using 15 attributes.



Gambar 7. Decision tree formed

5. Conclusion

Pre-processing can improve the performance of the classification algorithm. The use of information gain ratio for feature selection of early stage diabetes risk prediction dataset has proven to be good. The accuracy level of diabetes disease classification using C45 was previously 95.96%. This process is carried out using all existing data attributes. After pre-processing, it is known that there are attributes that do not have a gain value. The attribute with a gain value of 0, namely "itching", is then not included in the classification process. The accuracy level of C45 after feature selection is 96.15. This study proves that the use of information gain ratio feature selection can improve the performance of C45 for diabetes disease classification.

References

[1] WHO, “Diabetes,” 2023 World Health Organization. [Online].

Available: https://www.who.int/news-room/fact-sheets/detail/diabetes
[2] C. J. Ejiyi *et al.*, "A robust predictive diagnosis model for diabetes mellitus using Shapley-incorporated machine learning algorithms," *Healthc. Anal.*, vol. 3, no. December 2022, p. 100166, 2023, doi: 10.1016/j.health.2023.100166.
[3] Univercity of Washington, "Explore results from the 2019 Global Burden of Disease (GBD) study." [Online]. Available: https://vizhub.healthdata.org/gbd-results/
[4] International Diabetes Federation, "Diabetes."
[5] Databoks, "Diabetes Tipe 2 Paling Banyak Diderita Orang Indonesia pada 2023."
[6] I. H. Witten, E. Frank, M. A. Hall, and C. J, *Data Mining (Fourth Edition)*, 4th ed. Kaufmann, Morgan, 2017. doi: https://doi.org/10.1016/B978-0-12-804291-5.00004-0.
[7] ikhsan wisnuadji Gamadarenda and I. Waspada, "Implementasi Data Mining Untuk Deteksi Penyakit Ginjal Kronis (Pgk) Menggunakan K-Nearest Neighbor (Knn) Dengan Backward Elimination," vol. 7, no. 2, pp. 417–426, 2018, doi: 10.25126/jtiik.202071896.
[8] M. F. Kurniawan and Ivandari, "Komparasi Algoritma Data Mining untuk Klasifikasi Kanker Payudara," *IC Tech*, vol. I April 20, pp. 1–8, 2017.
[9] G. Aguilera-Venegas, A. López-Molina, G. Rojo-Martínez, and J. L. Galán-García, "Comparing and tuning machine learning algorithms to predict type 2 diabetes mellitus," *J. Comput. Appl. Math.*, vol. 427, p. 115115, 2023, doi: 10.1016/j.cam.2023.115115.
[10] C. Carpinteiro, J. Lopes, A. Abelha, and M. F. Santos, "A Comparative Study of Classification Algorithms for Early Detection of Diabetes," *Procedia Comput. Sci.*, vol. 220, pp. 868–873, 2023, doi: 10.1016/j.procs.2023.03.117.
[11] I. Ivandari, M. R. Maulana, and M. A. Al Karomi, "Classification of Type 2 Diabetes using Decission Tree Algorithm," *Jaict*, vol. 8, no. 2, pp. 236–241, 2023, [Online]. Available: https://jurnal.polines.ac.id/index.php/jaict/article/view/4835
[12] M. A. Alkaromi, "Information Gain untuk Pemilihan Fitur pada Klasifikasi Heregistrasi Calon Mahasiswa dengan Menggunakan K-NN," 2014.
[13] B. Azhagusundari and A. S. Thanamani, "Feature Selection based on Information Gain," no. 2, pp. 18–21, 2013.
[14] M. A. Al Karomi, M. R. Maulana, S. J. Prasetiyono, Ivandari, and Arochman, "Strengthening campus finance by analyzing attribute attributes for student registration classifications." p. 1, 2019. [Online]. Available: https://jurnal.polines.ac.id/index.php/jaict/article/view/1431
[15] S. Kumari, D. Kumar, and M. Mittal, "An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier," *Int. J. Cogn. Comput. Eng.*, vol. 2, no. January, pp. 40–46, 2021, doi: 10.1016/j.ijcce.2021.01.001.
[16] Ivandari, W. Setianto, and M. A. Alkaromi, "Klasifikasi Diabetes Tipe 2 Menggunakan Algoritma K-Nearest Neighbour," *IC-Tech*, vol. 18, no. 1, pp. 36–41, 2023, doi: 10.47775/ictech.v18i1.273.
[17] I. H. Witten, E. Frank, and M. A. Hall, *Data Mining: Practical Machine Learning Tools and Techniques 3rd Edition*. Elsevier, 2011.
[18] E. Prasetyo, *Data Mining Konsep dan Aplikasi menggunakan Matlab*. Yogyakarta: Andi Offset, 2012.
[19] H. Deng and G. Runger, "Feature Selection via Regularized Trees," Jan. 2012, Accessed: Oct. 16, 2014. [Online]. Available: http://arxiv.org/abs/1201.1587v3
[20] J. Novakovic, "The Impact of Feature Selection on the Accuracy of 1DwYH Bayes Classifier," vol. 2, pp. 1113–1116, 2010.
[21] Ian H Witten. Eibe Frank. Mark A Hall, *Data Mining 3rd*. 2011.
[22] Ivandari and M. A. Al Karomi, "Classification of Covid-19 Survillance Datasets using the Decision Tree Algorithm," *Jaict*, vol. 6, no. 1, pp. 44–49, 2021, [Online]. Available: https://jurnal.polines.ac.id/index.php/jaict/article/view/2896
[23] Ivandari and M. A. Al Karomi, "Algoritma K-NN untuk klasifikasi dataset Covid-19 survillance," *IC Tech*, vol. 16, no. 1, pp. 12–15, 2021, [Online]. Available: https://ejournal.stmik-wp.ac.id/index.php/ictech/article/view/137
[24] F. Gorunescu, *Data Mining: Concepts; Models and Techniques*. Springer, 2011.
[25] J. Gao, Z. Wang, T. Jin, J. Cheng, Z. Lei, and S. Gao, "Information gain ratio-based subfeature grouping empowers particle swarm optimization for feature selection," *Knowledge-Based Syst.*, vol. 286, no. 28 February 2024, 111380, 2024, doi: https://doi.org/10.1016/j.knosys.2024.111380.
[26] S. Diabetes and B. Hospital in Sylhet, "Early stage diabetes risk prediction dataset." [Online]. Available: https://archive.ics.uci.edu/dataset/529/early+stage+diabetes+risk+prediction+dataset