Implementation of the C45 Algorithm in Classifying Classes

Nellysa Putri Denila Nasution, Fristi Riandari
[1,3]Informatics Engineering, STMIK Pelita Nusantara, Sumatera Utara, Indonesia

A R T I C L E I N F O

ABSTRACT

Mts Swasta YPII Kotarih is one of the educational institutions which in its implementation classifies several classes for students, one of which is the superior class. However, in practice, each person is still selected based on the rank in each class, making the classification process less accurate and efficient, a computerized classification technique is needed in its implementation, the classification technique used by Data Mining is the c4.5 algorithm that works with the Decision tree to find the highest gain value from each criterion existing criteria, the criteria used in this study are the subjects of Q. Hadith, Akidah, Fiqh, SKI, Arabic, Tahfizh, General research results show that from There are 3 criteria that are most influential in the classification process, namely General, Arabic, Fiqh, implemented in the form of a computerized system. Based on the research results, the highest gain value is the highest gain is General with a value of 0.3629, so that the general attribute becomes node 1 or the root of the decision tree. second highest gain B.Arabic attribute, with a gain value of 1.7189, and the third highest gain is Fiqh with a value of 0.3486.

Corresponding Author:

Fristi Riandari,
Informatics Engineering,
STMIK Pelita Nusantara, Sumatera Utara, Indonesia
Jl. Iskandar Muda No.1 Medan, Sumatera Utara, Indonesia
Email: fristy.rianda@ymail.com

1. Introduction

Mts Swasta YPII Kotarih is one of the educational institutions which in its implementation classifies several classes for students, one of which is the superior class. The industrial revolution 4.0 is the latest advancement in the world of technology that changes the way human resources work, putting aside the role of humans and replacing them with digital technology (Arta et al., 2017)(Asroni et al., 2018). Data mining is often called knowledge discovery in database (KDD), an activity that includes collecting, using historical data to find regularities, patterns or relationships in large data sets (Suntoro, 2019). C4.5 algorithm is a tree structure in which nodes describe the attributes, each branch describes the results of the tested attributes, and each leaf describes the class(Saputra et al., 2016).

The qualifications of each school are different in terms of the things they do, so the YPII Kotarih private MTS must specify which students will be classified based on class. This research was conducted to classify students based on the scores achieved from each student, using the C45 algorithm classification method, which uses a decision tree (SAWITRI, 2019), (Sutoyo, 2018).

Based on previous research (Septiani, 2017)(Handoko, 2016)[15]. Analysis and Application of the C4.5 Method for Predicting Customer Loyalty. Essay. shows that the accuracy reaches 97.5%, which indicates that the C4.5 algorithm is suitable for measuring cellular data customer loyalty

(Putra & Wadisman, 2018)(Renhoran et al., 2018). Meanwhile, Algorithm classification method C4.5 for Credit Eligibility Prediction at Bank Mayapada Jakarta Industry from this research, the model formed with the C4.5 algorithm itself already has a good accuracy of 83.67% with the selection process attributes by the C4.5 algorithm (Karsito & Sari, 2019)(Nasrullah, 2018).

The goal is to classify excellent students who will enter the superior class at the Private MTS YPII Kotarih, improve the accuracy of the selection of superior students according to existing criteria (ISWATI, n.d.). of this research As an alternative in classifying superior classes at YPII Kotarih Private MTS and can be used as a reference source for other researchers (Al Syahdan & Sindar, 2018).

2. Method
2.1 Research Framework

The framework discusses the research implementation model in classifying superior classes, the following stages:
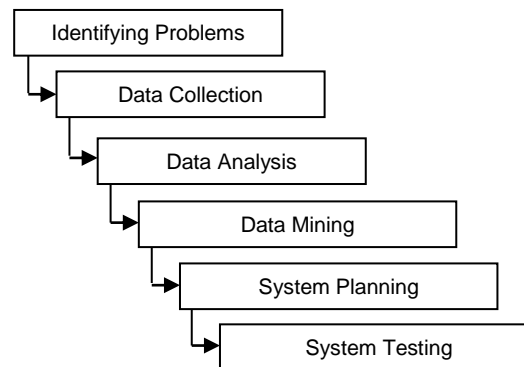

Fig 1. Research Framework

a. Problem Identification

At this stage, identification of the problem will be carried out, which is found, namely how to make the right decisions in classifying students into superior classes.

b. Data Collection

In this study the authors get data directly from the object of research through observations made, these observations are made frankly to the data source, that the author is conducting research, the author also uses passive particion observations, namely the researcher comes to the researcher's place but is not involved in the activity. at the research site to obtain complete data, the results of the data obtained are in the form of student names, NISN, value of subjects.

c. Data Analysis

At this stage, the inspection and cleaning process is carried out, with the aim of finding useful information, using qualitative data analysis, namely after the data obtained through interviews, the data needed are grouped, at this stage the data obtained is the data of grade 7 students in the odd semester of the 2020 academic year. /2021.

d. Data Mining

The use of large data to find patterns, in this study using the C4.5 . algorithm Identification of problems. Includes stages : selection, preprocessing, transformation, data mining, interpretation (Karsito & Sari, 2019). There are several stages in making a decision tree in the C4.5 algorithm (Handoko, 2016), (Iriadi & Nuraeni, 2016) :

1) Prepare training data
2) Calculating the root of the tree, first calculate the entropy value, to calculate the entropy value is used the formula :

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

Information :

S   = Case set
n   = Number of partitions
pi   = Proportion pi to S

3) Calculating the Gain value using Equation 2

$$Gain(S, A) = entropy(S) - \sum_{i=1}^{n} \frac{|Si|}{|S|} Entropy (Si)$$ (2)

Information :
S   = Case set
A   = feature
n   = Number of attribute partitions A
$|Si|$ = Proportion Si to S
$|S|$ = number of case in S

4) Repeat step 2 and step 3 until all records are partitioned
5) The decision tree partitioning process will stop when:
- Records in node N have the same class and there are no attributes in the partitioned record anymore.

3.   Result and Discussion

There are 7 criteria used for the classification of the superior class students, these criteria are set by MTS Swasta YPII Kotarih, namely: Q. Hadith, Akidah, Fiqh, SKI, Tahfizh, B. Arabic, General. After the data criteria are available, the existing values will be changed into categories that have been set by MTS Swasta YPII Kotarih:

TABLE 1.
CATEGORIES

Number	Value	Categori
1	9 – 100	Sangat Baik
2	83 – 90	Baik
3	70-82	Cukup
4	≤69	Buruk

This research was conducted at MTS Swasta YPII Kotarih, this study used passive particion observations. The data obtained from the data of class VII students in the odd semester of the 2020/2021 academic year, the number of student data taken as a sample is 142 people. The output in this study is divided into two categories, namely Excellent Students and Not. Based on the output to be generated, the authors use classification techniques.

3.1 Data Selection

At this stage the data that has been obtained will be selected according to the needs of the research.

The data used is grade 7 odd semester for the 2020/2021 academic year, the data obtained is 142 data.

TABLE 2.
THE C4.5 . ALGORITHM 7TH SEMESTER STUDENT DATA 2020/2021

Num	Name	NISN	Q.Hadist	Akidah	Fiqih
1	Aldyansah Ramad	0087072669	87	85	85
2	Alfatiha Fatwa	0086122120	85	86	85
3	Andran Syahputra	0072392934	85	87	85
4	Andri Iriyanto	0075216135	85	86	85
5	Andrian sentosa	0082333829	86	88	85
6	Deniel ma'ruf	0085198547	87	87	84
7	Deswita Indah	0073094706	88	88	87
8	Diki irwandi	0089128545	81	80	80
9	Dinda priska dewi	0063241235	85	87	86
10	Efranda gunawan	0085240512	85	86	83

3.3 Pre-processing

The process of data preparation by cleaning inconsistent data and noise. At this stage all student data is used because inconsistent data and noise are not found.

3.4 Transformation

The data will be converted into an assessment according to predetermined criteria. The results of the research carried out are superior and not superior students based on the information Furthermore, after the data has been cleaned, it enters the transformation stage, where the data will be changed or combined into the assessment according to the predetermined value of the assessment indicator.

TABLE 6.
VALUE ASSESSMENT CRITERIA

Rangked		Description
1-10	≥83	Superior
>10	<83	Not superior

For the attributes of price, taste and service have the same assessment criteria. After the questionnaire data is cleaned, then the data is then transformed from each criterion which can be seen in the following table.

TABLE 7.
TRANSFORMATION DATA

Num	Name	Values		
		Q.Hadist	Akidah	Fiqih
1	Aldyansah Ramad	Good	Good	Good
2	Alfatiha Fatwa	Good	Good	Good
3	Andran Syahputra	Good	Good	Good
4	Andri Iriyanto	Good	Good	Good
5	Andrian sentosa	Good	Good	Good
6	Deniel ma'ruf	Good	Good	Good
7	Deswita Indah	Good	Good	Good
8	Diki irwandi	Enough	Enough	Enough
9	Dinda priska dewi	Good	Good	Good
10	Efranda gunawan	Good	Good	Good

3.5 Data Mining

Processing the data Abundance becomes important information [2], for each attribute To calculate the gain using the formula :

$$Gain(S, A) = entropy(S) - \sum_{i=1}^{n} \frac{|Si|}{|S|} Entropy (Si) \tag{3}$$

While to calculate entrophy :

$$Entropy (S) = \sum_{i}^{n} -1 - p_i log_2 (p_i) \tag{4}$$

Entropy (Total) with the following formula :

$$Entropy (total) = -\sum_{i=1}^{n} -Pi \times Log_2 Pi \tag{5}$$

Calculating the Entropy Value of each attribute :

$$Entropy (total) = (-\frac{27}{43} *Log_2(\frac{27}{43})) + (-\frac{16}{43} *Log_2(\frac{16}{43})) = 0,9522$$

The entropy (total) calculates the total value of the leading decisions of 27 and 16 unseeded, while 43 is the total number.

a. Q.Hadist

$$Entropy(SB) = (-\frac{3}{3} *Log_2(\frac{3}{3})) + (-\frac{0}{3} *Log_2(\frac{0}{3})) = 0$$

$$Entropy(B) = (-\frac{17}{28} *Log_2(\frac{17}{28})) + (-\frac{11}{28} *Log_2(\frac{11}{28})) = 0,9666$$

$Entropy$(C)= $(-\frac{7}{12} *Log_2(\frac{7}{12})) + (-\frac{5}{12} *Log_2(\frac{5}{12}))$= 0,9798.

b. Akidah

$Entropy$(SB) = $(-\frac{1}{1} *Log_2(\frac{1}{1})) + (-\frac{0}{1} *Log_2(\frac{0}{1}))$= 0

$Entropy$(B)= $(-\frac{26}{40} *Log_2(\frac{26}{40})) + (-\frac{14}{40} *Log_2(\frac{14}{40}))$= 0,9340

$Entropy$(C) = $(-\frac{0}{2} *Log_2(\frac{0}{2})) + (-\frac{2}{2} *Log_2(\frac{2}{2}))$= 0

c. Fiqih

$Entropy$(SB) = $(-\frac{5}{6} *Log_2(\frac{5}{6})) + (-\frac{1}{6} *Log_2(\frac{1}{6}))$= 0,6500

$Entropy$(B)= $(-\frac{22}{31} *Log_2(\frac{22}{31})) + (-\frac{9}{31} *Log_2(\frac{9}{31}))$= 0,8691

$Entropy$(C) = $(-\frac{0}{6} *Log_2(\frac{0}{6})) + (-\frac{0}{6} *Log_2(\frac{0}{6}))$= 0

d. SKI

$Entropy$(SB)= $(-\frac{1}{2} *Log_2(\frac{1}{2})) + (-\frac{1}{2} *Log_2(\frac{1}{2}))$= 1

$Entropy$(B)= $(-\frac{9}{11} *Log_2(\frac{9}{11})) + (-\frac{2}{11} *Log_2(\frac{2}{11}))$= 0,6840

$Entropy$(C)= $(-\frac{17}{30} *Log_2(\frac{17}{30})) + (-\frac{13}{30} *Log_2(\frac{13}{30}))$= 0,9871

e. B. Arab

$Entropy$(SB)= $(-\frac{0}{0} *Log_2(\frac{0}{0})) + (-\frac{0}{0} *Log_2(\frac{0}{0}))$= 0

$Entropy$(B)= $(-\frac{19}{24} *Log_2(\frac{19}{24})) + (-\frac{5}{24} *Log_2(\frac{5}{24}))$= 0,7382

$Entropy$(C)= $(-\frac{8}{19} *Log_2(\frac{8}{19})) + (-\frac{11}{19} *Log_2(\frac{11}{19}))$= 0,9819

f. Tahfizh

$Entropy$(SB) = $(-\frac{3}{3} *Log_2(\frac{3}{3})) + (-\frac{0}{3} *Log_2(\frac{0}{3}))$= 0

$Entropy$(B) = $(-\frac{7}{10} *Log_2(\frac{7}{10})) + (-\frac{3}{10} *Log_2(\frac{3}{10}))$= 0,8812

$Entropy$(C)= $(-\frac{17}{30} *Log_2(\frac{17}{30})) + (-\frac{13}{30} *Log_2(\frac{13}{30}))$= 0,9871

g. Umum

$Entropy$(SB) = $(-\frac{4}{4} *Log_2(\frac{4}{4})) + (-\frac{0}{4} *Log_2(\frac{0}{4}))$= 0

$Entropy$(B) = $(-\frac{16}{17} *Log_2(\frac{16}{17})) + (-\frac{1}{17} *Log_2(\frac{1}{17}))$= 0,3227

$Entropy$(C)= $(-\frac{7}{22} *Log_2(\frac{7}{22})) + (-\frac{15}{22} *Log_2(\frac{15}{22}))$= 0,9456

After that calculate the gain of each attribute :

a. *Gain*(Total, Q.Hadist)

$= Entropy(S) - \sum_{1=1}^{n} \frac{|Q.Hadist_i|}{|Total|} * Entropy(Q.Hadist_i)$

= 0,9522 - $((\frac{3}{43} * 0) + (\frac{28}{43} * 0,9661) + (\frac{12}{43} * 0,9798))$= 0,0493

b. *Gain*(Total, Akidah)

$= Entropy(S) - \sum_{1=1}^{n} \frac{|Akidah_i|}{|Total|} * Entropy(Akidah_i)$

= 0,9522 - $((\frac{1}{43} * 0) + (\frac{40}{43} * 0,9340) + (\frac{2}{43} * 0))$= 0,0833

c. *Gain*(Total, Fiqih)

$= Entropy(S) - \sum_{1=1}^{n} \frac{|Fiqih_i|}{|Total|} * Entropy(Fiqih_i)$

= 0,9522 - $((\frac{6}{43} * 0,6500) + (\frac{31}{43} * 0,8691) + (\frac{6}{43} * 0))$= 0,2349

d. *Gain*(Total, SKI)

$$= Entropy\text{(S)} - \sum_{1=1}^{n} \frac{|\text{SKI}_i|}{|\text{Total}|} * Entropy(\text{SKI}_i)$$

$$= 0{,}9522 - ((\tfrac{2}{43} * 1) + (\tfrac{11}{43} * 0{,}6840) + (\tfrac{30}{43} * 0{,}9871)) = 0{,}0420$$

e. *Gain*(Total, B.Arab)

$$= Entropy\text{(S)} - \sum_{1=1}^{n} \frac{|\text{B.Arab}_i|}{|\text{Total}|} * Entropy(\text{B.Arab}_i)$$

$$= 0{,}9522 - ((\tfrac{0}{43} * 0) + (\tfrac{24}{43} * 0{,}7382) + (\tfrac{19}{43} * 0)) = 0{,}9819$$

f. *Gain*(Total, Tahfizh)

$$= Entropy\text{(S)} - \sum_{1=1}^{n} \frac{|\text{Tahfizh}_i|}{|\text{Total}|} * Entropy(\text{Tahfizh}_i)$$

$$= 0{,}9522 - ((\tfrac{3}{43} * 0) + (\tfrac{10}{43} * 0{,}8812) + (\tfrac{30}{43} * 0{,}9871)) = 0{,}0586$$

g. *Gain*(Total, Umum)

$$= Entropy\text{(S)} - \sum_{1=1}^{n} \frac{|\text{Umum}_i|}{|\text{Total}|} * Entropy(\text{Umum}_i)$$

$$= 0{,}9522 - ((\tfrac{4}{43} * 0) + (\tfrac{17}{43} * 0{,}3227) + (\tfrac{22}{43} * 0{,}9456)) = 0{,}3408$$

From these calculations, the highest gain is General with a value of 0.3629, then node 1 is General, following Decision Tree node 1:
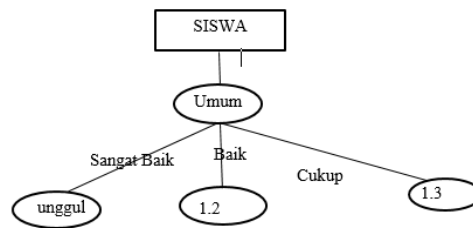
Fig 2. Decision Tree node 1.2

Next, node 1.2 is calculated as the root for the Entropy, Enough value of all Q. Hadith, Akidah, Fiqh, SKI, B.Arabic, and Tahfiz attributes, the same as the previous calculation, then calculates Entropy (Sufficient General Value), with the formula:

$$Entropy \text{ (total)} = -\sum_{i=1}^{n} -p_i log_2(p_i)$$

$$= (-\tfrac{22}{31} * Log_2(\tfrac{22}{31})) + (-\tfrac{9}{31} * Log_2(\tfrac{9}{31})) = 0{,}8691$$

a. Atribut Q.Hadist

$$Entropy \text{ (SB)} = (-\tfrac{2}{2} * Log_2(\tfrac{2}{2})) + (-\tfrac{0}{2} * Log_2(\tfrac{0}{2})) = 0$$

$$Entropy \text{ (B)} = (-\tfrac{18}{25} * Log_2(\tfrac{18}{25})) + (\tfrac{7}{25} * Log_2(\tfrac{7}{25})) = 0{,}855$$

$$Entropy \text{ (C)} = (-\tfrac{2}{2} * Log_2(\tfrac{2}{2})) + (-\tfrac{2}{2} * Log_2(\tfrac{2}{2})) = 1$$

b. Atribut Akidah

$$Entropy \text{ (SB)} = (-\tfrac{0}{0} * Log_2(\tfrac{0}{0})) + (-\tfrac{0}{0} * Log_2(\tfrac{0}{0})) = 0$$

$$Entropy \text{ (B)} = (-\tfrac{22}{29} * Log_2(\tfrac{22}{29})) + (-\tfrac{7}{29} * Log_2(\tfrac{7}{29})) = 0{,}7973$$

$$Entropy \text{ (C)} = (-\tfrac{0}{2} * Log_2(\tfrac{0}{2})) + (-\tfrac{0}{2} * Log_2(\tfrac{0}{2})) = 0$$

c. Atribut Fiqih

$$Entropy \text{ (SB)} = (-\tfrac{5}{6} * Log_2(\tfrac{5}{6})) + (-\tfrac{1}{6} * Log_2(\tfrac{1}{6})) = 0{,}6500$$

$$Entropy \text{ (B)} = (-\tfrac{17}{19} * Log_2(\tfrac{17}{19})) + (-\tfrac{2}{19} * Log_2(\tfrac{2}{19})) = 0{,}4854$$

$$Entropy \text{ (C)} = (-\tfrac{0}{6} * Log_2(\tfrac{0}{6})) + (-\tfrac{6}{6} * Log_2(\tfrac{6}{6})) = 0$$

d. Atribut SKI

Entropy (SB) = $(-\frac{0}{0} * Log_2(\frac{0}{0})) + (-\frac{1}{1} * Log_2(\frac{1}{1})) = 0$

Entropy (B) = $(-\frac{8}{11} * Log_2(\frac{8}{11})) + (-\frac{3}{11} * Log_2(\frac{3}{11})) = 0,8453$

Entropy (C) = $(-\frac{14}{19} * Log_2(\frac{14}{19})) + (-\frac{5}{19} * Log_2(\frac{5}{19})) = 0,8314$

e.   Atribut .B.Arab

Entropy (SB) = $(-\frac{0}{0} * Log_2(\frac{0}{0})) + (-\frac{0}{0} * Log_2(\frac{0}{0})) = 0$

Entropy (B) = $(-\frac{16}{17} * Log_2(\frac{16}{17})) + (-\frac{1}{17} * Log_2(\frac{1}{17})) = 0,7382$

Entropy (C) = $(-\frac{6}{14} * Log_2(\frac{6}{14})) + (-\frac{8}{14} * Log_2(\frac{2}{8\frac{6}{14}})) = 0,9852$

f.   Atribut Tahfizh

Entropy (SB) = $(-\frac{1}{0} * Log_2(\frac{1}{0})) + (-\frac{0}{0} * Log_2(\frac{0}{0})) = 0$

Entropy (B) = $(-\frac{4}{5} * Log_2(\frac{4}{5})) + (-\frac{1}{5} * Log_2(\frac{1}{5})) = 0,7219$

Entropy (C) = $(-\frac{17}{25} * Log_2(\frac{17}{25})) + (-\frac{8}{25} * Log_2(\frac{8}{25})) = 0,904$

Calculating the gain value for each attribute:

a.   *Gain*(Total, Q.Hadist)

    = *Entropy*(S) $\sum_{1=1}^{n} \frac{|Q.Hadist_i|}{|Total|} Entropy(Q.Hadist_i)$

    = 0,8691 - $((\frac{2}{31} * 0) + (\frac{25}{31} * 0,855) + (\frac{4}{31} * 1)) = 0,0522$

b.   *Gain*(Total, Akidah)

    = *Entropy*(S) - $\sum_{1=1}^{n} \frac{|Akidah_i|}{|Total|} * Entropy(Akidah_i)$

    = 0,8691 - $((\frac{0}{31} * 0) + (\frac{29}{31} * 0,7973) + (\frac{2}{31} * 0)) = 0,12325$

c.   *Gain*(Total, Fiqih)

    = *Entropy*(S) - $\sum_{1=1}^{n} \frac{|Fiqih_i|}{|Total|} * Entropy(Fiqih_i)$

    = 0,8691 - $((\frac{6}{31} * 0,6500) + (\frac{19}{31} * 0,4854) + (\frac{6}{31} * 0)) = 0,4457$

d.   *Gain*(Total, SKI)

    = *Entropy*(S) - $\sum_{1=1}^{n} \frac{|SKI_i|}{|Total|} * Entropy(SKI_i)$

    = 0,8691 - $((\frac{1}{31} * 0) + (\frac{11}{31} * 0,8453) + (\frac{19}{31} * 0,8314)) = 0,0595$

e.   *Gain*(Total, B.Arab)

    = *Entropy*(S) - $\sum_{1=1}^{n} \frac{|B.Arab_i|}{|Total|} Entropy(B.Arab_i)$

    = 0,8691 - $((\frac{0}{31} * 0) + (\frac{17}{31} * 0,7382) + (\frac{14}{31} * 0,9852)) = 1,71894$

f.   *Gain*(Total, Tahfizh)

    = *Entropy*(S) - $\sum_{1=1}^{n} \frac{|Tahfizh_i|}{|Total|} Entropy(Tahfizh_i)$

    = 0,8691 - $((\frac{1}{31} * 0) + (\frac{5}{31} * 0,7219) + (\frac{25}{31} * 0,9043)) = 0,0233$

The highest gain attribute is used as node 1.2, namely. From the table above, it is known that the highest gain attribute is used as node 1.2, namely the B.Arab attribute, with a gain value of 1.7189. Therefore B.Arabic becomes node 1.2. The decision tree can be seen in the following:
After successfully determining Node 1.2 then look for Node 1.3, as follows:

Entropy (total) = $(-\frac{25}{44} * Log_2(\frac{25}{44})) + (-\frac{19}{44} * Log_2(\frac{19}{44})) = 0,986$

The entropy (total) calculates the total value of the leading decisions of 25 and 19 unseeded, while 44 is the total number.

    a.   Q.Hadist

$Entropy$(SB)= $(-\frac{3}{3} * Log_2(\frac{3}{3})) + (-\frac{0}{3} * Log_2(\frac{0}{3}))$ = 0

$Entropy$(B)=$(-\frac{14}{27} * Log_2(\frac{14}{27})) + (-\frac{13}{27} * Log_2(\frac{13}{27}))$ = 0,9990

$Entropy$(C) = $(-\frac{8}{14} * Log_2(\frac{8}{14})) + (-\frac{6}{14} * Log_2(\frac{6}{14}))$ = 0,9852

b.   Akidah

$Entropy$(SB) = $(-\frac{1}{1} * Log_2(\frac{1}{1})) + (-\frac{0}{1} * Log_2(\frac{0}{1}))$ = 0

$Entropy$(B)= $(-\frac{24}{41} * Log_2(\frac{24}{41})) + (-\frac{24}{41} * Log_2(\frac{24}{41}))$ = 0,9788

$Entropy$(C) = $(-\frac{0}{2} * Log_2(\frac{0}{2})) + (-\frac{2}{2} * Log_2(\frac{2}{2}))$ = 0

c.   Fiqih

$Entropy$(SB) = $(-\frac{5}{6} * Log_2(\frac{5}{6})) + (-\frac{1}{6} * Log_2(\frac{1}{6}))$ = 0,6500

$Entropy$(B)= $(-\frac{20}{28} * Log_2(\frac{20}{28})) + (-\frac{8}{28} * Log_2(\frac{8}{28}))$ = 0,8631

$Entropy$(C)= $(-\frac{0}{10} * Log_2(\frac{0}{10})) + (-\frac{10}{10} * Log_2(\frac{10}{10}))$ = 0

d.   SKI

$Entropy$(SB) = $(-\frac{0}{2} * Log_2(\frac{0}{2})) + (-\frac{2}{2} * Log_2(\frac{2}{2}))$ = 0

$Entropy$(B) = $(-\frac{7}{11} * Log_2(\frac{7}{11})) + (-\frac{4}{11} * Log_2(\frac{4}{11}))$ = 0,9456

$Entropy$(C)= $(-\frac{18}{31} * Log_2(\frac{18}{31})) + (-\frac{13}{31} * Log_2(\frac{13}{31}))$ = 0,9811

e.   B. Arab

$Entropy$(SB) = $(-\frac{0}{0} * Log_2(\frac{0}{0})) + (-\frac{0}{0} * Log_2(\frac{0}{0}))$ = 0

$Entropy$(B)= $(-\frac{20}{27} * Log_2(\frac{20}{27})) + (-\frac{7}{27} * Log_2(\frac{7}{27}))$ = 0,8256

$Entropy$(C)= $(-\frac{5}{17} * Log_2(\frac{5}{17})) + (-\frac{12}{17} * Log_2(\frac{12}{17}))$ = 0,8739

f.   Tahfizh

$Entropy$(SB)= $(-\frac{3}{3} * Log_2(\frac{3}{3})) + (-\frac{0}{3} * Log_2(\frac{0}{3}))$ = 0

$Entropy$(B)= $(-\frac{5}{12} * Log_2(\frac{5}{12})) + (-\frac{7}{12} * Log_2(\frac{7}{12}))$ = 1,0262

$Entropy$(C)= $(-\frac{17}{29} * Log_2(\frac{17}{29})) + (-\frac{13}{29} * Log_2(\frac{13}{29}))$ = 0,9705

g.   Umum

$Entropy$(SB)= $(-\frac{4}{4} * Log_2(\frac{4}{4})) + (-\frac{0}{4} * Log_2(\frac{0}{4}))$ = 0

$Entropy$(B)= $(-\frac{15}{18} * Log_2(\frac{15}{18})) + (-\frac{3}{18} * Log_2(\frac{3}{18}))$ = 0,6500

$Entropy$(C)= $(-\frac{6}{22} * Log_2(\frac{6}{22})) + (-\frac{16}{22} * Log_2(\frac{16}{22}))$ = 0,8453

After that calculate the gain of each attribute:

a.   *Gain*(Total, Q.Hadist)

=*Entropy(*S) - $\sum_{1=1}^{n} \frac{|Q.Hadist_i|}{|Total|} * Entropy(Q.Hadist_i)$

= 0,9865 - $((\frac{3}{44} * 0) + (\frac{27}{44} * 0,9990) + (\frac{14}{44} * 0,9852))$ = 0,0600

b.   *Gain*(Total, Akidah)

=*Entropy(*S)- $\sum_{1=1}^{n} \frac{|Akidah_i|}{|Total|} * Entropy(Akidah_i)$

= 0,9865 - $((\frac{1}{44} * 0) + (\frac{41}{44} * 0,9788) + (\frac{2}{44} * 0))$ = 0,0744

c.   *Gain*(Total, Fiqih)

=*Entropy(*S)- $\sum_{1=1}^{n} \frac{|Fiqih_i|}{|Total|} * Entropy(Fiqih_i)$

= 0,9865 - $((\frac{6}{44} * 0,6500) + (\frac{28}{44} * 0,8631) + (\frac{10}{44} * 0))$ = 0,3486

d.   *Gain*(Total, SKI)

=*Entropy(*S) - $\sum_{1=1}^{n} \frac{|SKI_i|}{|Total|} * Entropy(SKI_i)$

$$= 0{,}9865 - ((\tfrac{2}{44} * 0) + (\tfrac{11}{44} * 0{,}9456) + (\tfrac{31}{44} * 0{,}9811)) = 0{,}0588$$

e.   *Gain*(Total, B.Arab)

$$= Entropy(S) - \sum_{1=1}^{n} \frac{|\text{B.Arab}_i|}{|\text{Total}|} * Entropy(\text{B.Arab}_i)$$

$$= 0{,}9865 - ((\tfrac{0}{44} * 0) + (\tfrac{27}{44} * 0{,}8256) + (\tfrac{17}{44} * 0{,}8739)) = 0{,}1422.$$

f.   *Gain*(Total, Tahfizh)

$$= Entropy(S) - \sum_{1=1}^{n} \frac{|\text{Tahfizh}_i|}{|\text{Total}|} * Entropy(\text{Tahfizh}_i)$$

$$= 0{,}9865 - ((\tfrac{3}{44} * 0) + (\tfrac{12}{44} * 1{,}02626) + (\tfrac{29}{44} * 0{,}9705)) = 0{,}0669$$

g.   *Gain*(Total, Umum)

$$= Entropy(S) - \sum_{1=1}^{n} \frac{|\text{Umum}_i|}{|\text{Total}|} * Entropy(\text{Umum}_i)$$

$$= 0{,}9865 - ((\tfrac{4}{44} * 0) + (\tfrac{18}{44} * 0{,}6500) + (\tfrac{22}{44} * 0{,}8453)) = 0{,}2979.$$

From these calculations, it can be seen that the attribute with the highest gain is Fiqh with a value of 0.3486, then node 1.3 is Fiqh, as follows:

Fig 3. Decion tree node 1.3

Evaluation : correction of pattern obtained, so that it can be identified Is the pattern already representative, knowledge to be achieved [9].

The rules or rules formed based on the last decision tree as shown in Figure, 19 rules :
1) IF General = Very Good THEN decision = Excellent
2) General IF = Enough AND Fiqh = Very Good AND Arabic = Good THEN decision = Excellent
3) General IF = Enough AND Fiqh = Very Good AND Arabic = Enough THEN decision = Not Excellent
4) General IF = Enough AND Fiqh = Good AND Aqidah = Good AND Arabic = Good AND SKI = Good THEN
   decision = Excellent
5) IF General = Enough AND Fiqh = Good AND Aqidah = Good AND B.Arabic = Good AND

SKI = Enough
AND Q. Hadith = Good AND Tahfizh = Good THEN Decision = Excellent

6) IF General = Enough AND Fiqh = Good AND Akidah = Good AND B.Arabic = Good AND SKI = Enough
AND Q. Hadith = Good AND Tahfizh = Enough THEN decision = Not Excellent

7) General IF = Enough AND Fiqh = Good AND Akidah = Good AND Arabic = Good AND SKI = Enough AND
Q. Hadith = Enough THEN decision = Not Excellent

8) IF General = Enough AND Fiqh = Good AND Akidah = Good AND Arabic = Enough THEN decision = Not
Excellent

9) General IF = Enough AND Fiqh = Good AND Aqeedah = Enough THEN decision = Not superior

10) General IF = Enough AND Fiqh = Enough THEN decision = Not Superior

11) IF General = Good AND Arabic = Enough AND SKI = Very Good THEN decision = Excellent

12) General IF = Good AND Arabic = Enough AND SKI = Good THEN decision = Superior

13) General IF = Good AND B.Arabic = Enough AND SKI = Enough AND Q. Hadith = Good AND Akidah =
Good AND Fiqh = Good AND Tahfizh = Good THEN decision = Superior

14) IF General = Good AND Fiqh = Very Good THEN decision = Excellent

15) IF General = Good AND Fiqh = Good THEN decision = Excellent

16) IF General = Good AND Fiqh = Enough THEN decision = Not Superior

17) IF General = Fair AND SKI = Very Good THEN decision = Excellent

18) IF General = Enough AND SKI = Good THEN decision = Excellent

19) General IF = Enough AND SKI = Enough THEN decision = Not Excellent
IF Umum = Very Good , THEN Decision = superior

4   Conclusion

Based on the results of the decision tree above, there are 19 rules that were tested, the data testing was based on the 19 rules that had been set, from 142 data tested 73.94% : 26.06%, 105 data were appropriate and 37 data did not match. Application of the C4.5 Algorithm in classifying using a decision tree. The C45 algorithm is used to analyze student assessments in semester 7 of 2020/2021, the student data is applied to the KDD stages starting from data selection, Pre-Processing, Transformation, Data Mining, and Evaluation, by classifying which students are superior students, so that placement students who enter the superior class according to the achievements achieved.

References

Arta, K. J. dkk, (2016). Data Mining Rekomendasi Calon Mahasiswa Berprestasi Di Stmik Denpasar Menggunakan Metode Technique For Others Reference By Similarity To Ideal Solutionmunawar. 5(2),751. https://ejournal.undiksha.ac.id/index.php/JST/article/view/8549

Asroni, dkk (2018). Penerapan Metode Clustering dengan Algoritma K-Means pada Pengelompokkan Data Calon Mahasiswa Baru di Universitas Muhammadiyah Yogyakarta (Studi Kasus: Fakultas Kedokteran dan Ilmu Kesehatan, dan Fakultas Ilmu Sosial dan Ilmu Politik). 21(1), 60. https://journal.umy.ac.id/index.php/st/article/view/2610/3639 .

Dwi Septiani Program Studi Manajemen Informatika AMIK BSI Jakarta Jl Kramat Raya No, W., & Pusat, J. (2017). Dan Naive Bayes Untuk Prediksi Penyakit   Hepatitis. Jurnal Pilar Nusa Mandiri, 13(1), 76–84.   http://archive.ics.uci.edu/ml/ .

Handoko, K. (2016). Penerapan Data Mining Dalam Meningkatkan Mutu Pembelajaran   Pada   Instansi Perguruan Tinggi Menggunakan Metode K   Means Clustering (Studi Kasus Di Program Studi Tkj Akademi Komunitas   Solok Selatan). Jurnal Teknologi Dan Sistem Informasi, 02(03), 31–40. http://teknosi.fti.unand.id/index.php/teknosi/article/view/70.

Iriadi, N, & Nuraeni, N. (2016). Kajian Penerapan Metode Klasifikasi Data Mining Algoritma C4.5 Untuk Prediksi Kelayakan Kredit Pada Bank Mayapada Jakarta. Jurnal Teknik Komputer AMIK BSI, 2(1), https://ejournal.bsi.ac.id/ejurnal/index.php/jtk/article/view/371.

Karsito & Sari, W. M. (2018). Prediksi Potensi Penjualan Produk Delifrance Dengan Metode Naive Bayes Di Pt. Pangan Lestarilumbantoruan, 9(1), 64. https://www.jurnal.pelitabangsa.ac.id/index.php/sigma/article/view/465/282 .

Nasrullah, A. H. (2018). Penerapan Metode C4.5 untuk Klasifikasi Mahasiswa Berpotensi Drop Out. ILKOM Jurnal Ilmiah, 10(2), 245. https://doi.org/10.33096/ilkom.v10i2.300.244-250.

Putra, R. R. & Wadisman, C. (2018). Implementasi Data Mining Pemilihan Pelanggan Potensial Menggunakan Algoritma K-means, 1(1), 73. https://media.neliti.com/media/publications/256104-implementasi-data-mining-pemilihan-pelan-45b1352b.pdf.

Renhoran, B. S. dkk, (2018). Penerapan Algoritma C4.5 Untuk Menentukan Data Stok Dan Target Permintaan Material Yang Paling Dibutuhkan Area Kebon Jeruk. 12(2), 16. https://ejournal.nusamandiri.ac.id/index.php/inti/article/view/1564

Saputra, N. E. dkk, (2016). Penerapan Knowledge Management System (Kms) Menggunakan Teknik Knowledge Data Discovery (Kdd) Pada Pt Pln (Persero) Ws2jb Rayon Kayu Agung, 8(2), 5. http://ejournal.unsri.ac.id/index.php/jsi/index .

Sawitri, D. (2019). Revolusi Industri 4.0 : Big Data Menjawab Tantangan Revolusi Industri 4.0, 4(3), 1. http://download.garuda.ristekdikti.go.id/article.

Suntoro, J. 2019. Data Mining : Algoritma Implementasi dengan Pemrograman PHP. Jakarta : PT Gramedia.

Sutoyo, I. (2018). Implementasi Algoritma Decision Tree Untuk Klasifikasi Data Peserta Didik. Jurnal Pilar Nusa Mandiri Vol. 14, No. 2, 218. https://ejournal.nusamandiri.ac.id/index.php/pilar/article/view/70/59 .

Syahdan, S. Al, & Sindar, A. (2018). Data Mining Penjualan Produk Dengan Metode Apriori Pada Indomaret Galang Kota. Jurnal Nasional Komputasi Dan Teknologi Informasi (JNKTI), 1(2). 56. https://doi.org/10.32672/jnkti.v1i2.771.

Tusarwenda, T. R. I. B. (2018). Penerapan data mining dengan algoritma c4.5 dalam prediksi penjualan botol pada cv. seribukilo. 10.