



Available Online at :
<http://ejournal.amikompurwokerto.ac.id/index.php/telematika/>

Telematika

Accredited SINTA “2” Kemenristek/BRIN, No. 85/M/KPT/2020



Students Grade Grouping to Optimize On-Time Graduation Predictions by Combining K-Means and C4.5 Algorithms (Case Study: University Potensi Utama)

Bob Subhan Riza¹, Sarjon Defit²

¹Sistem Informasi, University Potensi Utama

²Teknik Informatika University Putra Indonesia YPTK Padang

e-mail: ¹bob.potensi@gmail.com, ²sarjond@yahoo.co.uk

ARTICLE INFO

History of the article:

Received September 4, 2020

Revised November 9, 2020

Accepted February 20, 2021

Keywords:

K-Means,
C4.5,
Graduated on time,
Student Grades

Correspondence:

Telephone: +628126013912

E-mail:

bob.potensi@gmail.com

ABSTRACT

Graduating on time is the dream of every student who studies in universities. Some factors that can lead to failure in graduating on time, such as grades, though students are sometimes careless and underestimating this factor, despite knowing that problematic Grade will hinder the student from graduating on time. This research helps the study program to predict which students will graduate on time. There are 2 stages in the research, first is the process of clustering students' data using the K-Means algorithm, while the second stage predicts students' graduation using the C4.5 algorithm. Variable used are Grade, Failing Grade, Specialization, Internship, Thesis, Undergraduate Thesis 1, Undergraduate Thesis 2, and Passing Grade. Using RapidMiner and processing these data using this software can predict students that graduate on time.

INTRODUCTION

The problem that occurs today is students who do not realize nor care about the problematic grades (failed Grade). Problematic grades can affect students in taking specialization, street vendors, thesis 1 and thesis 2 not on the right time, where it impacts graduation. In the 2015 information system study program of students' data showed that there were 0.49% of students who had failed or problematic grades that did not pass on time, and so were students who could not complete internship and thesis on time, thus resulting in graduating not on time with a percentage value of 100%. So the criteria that will be included in this case study are Failing Grade, Grade, Specialization, Internship, Undergraduate Thesis 1 and 2. This case study also uses RapidMiner software. The clustering method uses K-Means, which functions to cluster students' data to make it easier for the next process, and C4.5 is used to process predictive decisions of students who can graduate on time. This research only discusses the prediction of students graduating on time, where the benefits are that students are more concerned with grades and courses to be taken next to graduate in time, and the study program can find out which students have problematic grades.

The research (Priyatman, 2019) explains that the prediction of students who will graduate on time aims to provide knowledge to know and be able to predict students who will graduate on time only by paying attention to students who participate or are members of special clusters that have parameters in accordance with the provisions, including GPA and student attendance. This research uses the K-Means

method. The K-Means method divides existing data into clusters, data that has separate or special characteristics will be combined into the same cluster, if there are data that have different characteristics, they will be combined into different clusters as well. With the resulting system, it will certainly be able to help universities or students to be able to estimate the graduation of each student so that it is on time, it can also increase the credibility of the college, and students who graduate on time will be of benefit to themselves in completing their studies. Also, universities can take steps to predict each student who will not graduate in time. The same is the case with research (Rosmini, Fadlil & Sunardi, 2018), which examines student groups' mapping through lecture activity data.

While research of (Pradnyana, 2018) has examined the division of classes where sometimes there are problems that occur in each class division, namely the different abilities possessed by each student for the student class, it will impact the learning process that students will follow. The division of student groups based on each student's same ability is a parameter used in this division. This is very important, especially in improving the quality of each learning process carried out. If student groups' division is carried out properly, each student can certainly help each other in the teaching and learning process. Also, student class divisions must be made according to these students' abilities so that lecturers can easily determine the teaching and learning process method or method. In the learning process, effective and good strategies and methods must be used to improve the teaching and learning process. This research combines the K-Means and K-Nearest Neighbors (KNN) methods for each division of class groups. The class division of student lectures using the K-Means method according to the existing value components and is taken from the courses that each student as a prerequisite has followed. The criteria used for this group's division are assignments grade, the midterm exam grade, the final exam grade, and the cumulative grade point average (GPA). In predicting a student's graduation for a course, the KNN method is used, and the data used is taken from existing data. Based on the tests that have been carried out, it can be concluded that the number of classes and the amount of data used can affect the quality of the clusters that will be formed using the K-Means and KNN methods. The highest Silhouette Index value is obtained when the data used is 100 data with 10 clusters of 0.534, which are included in the medium structure quality class. Related to this (Ridwan, Suyono & Sarosa, 2013) also discussed the evaluation of student academic performance to see the extent of student academic relevance in improving graduation.

Furthermore, research (Evi) revealed that the enrichment of a material is one of the preparations needed by high school children to take the National Examination (*Ujian Nasional/UN*). At SMA NEGERI 4 TASIKMALAYA, there has not been any special group formation to explore the UN's material. Because there is no such group, it results in not precise goals or targets in designing each subject schedule that students need. Therefore, a solution is made which can certainly overcome the problems at hand. The group division method by clustering each student in this study uses the K-Means Algorithm and the subjects of each group of students using the FP-Growth Algorithm. The number of clusters used is $k = 2$, and the maximum number of item sets to be made is 3 item sets. The dataset to be used is 190 datasets. This application can help the school by displaying each group of students on the UN's existing subjects and can also display the UN subjects for each group. In line with research (Saleh & Nasari, 2018) which also discusses the importance of determining student majors to facilitate the teaching and learning process with several criteria that are used as references in determining the majors such as the Grade of Natural Science, Social Sciences, questionnaires, and recommendations.

The K-Mean method has been widely applied in various scientific fields (Elizawati & Lesmana, 2017), utilizing the K-Means method to determine student competence by analyzing Student Report Card scores on student interests. (Hairani, Saputra & Fadli, 2020) also applies this method to deal with class imbalances in diabetes classification. (Halilintar & Farida, 2018) predicts student academic achievement using the K-Means method, and (Muliono & Sembiring, 2019) uses the application of K-Means has been carried out in clustering the level of suitability achievement of lecturer performance with the Tridarma of Lecturer. (Mulyani et al., 2018) also implemented the K-Means method to recommend appropriate tutoring based on student academic segmentation. Furthermore, (Nanda, Pramono & Hartati, 2020) conducted a study to determine the level of student satisfaction with Academic services by implementing the K-Means method. The same thing was done (Nasari & Sianturi, 2016) to classify the spread of diarrhea, and (Putra & Wadisman, 2018) also applied the K-Means method for Selection of Potential Customers Using the K Means Algorithm.

RESEARCH METHODS

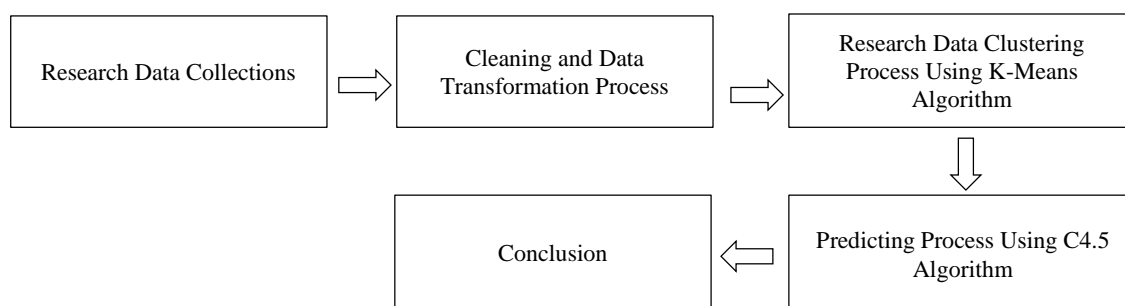


Figure 1. Research Stage

Data Collections

This study's data sources are student transcript data, student data taking specialization courses, student internship data, undergraduate thesis 1 student data, undergraduate thesis 1 student data, and judicial data for the 2015 IS (Information System) study program of University Potensi Utama. The amount of data used is 85 records with 7 variables (Grade, Failing Grade, Specialization, Internship, Undergraduate Thesis 1, Undergraduate Thesis 2, and Passing Grade). This is attached to table 1.

Table 1. Students Data of Information System Study Program

No.	NIM	Name	Grade	Failing Grade	Spec.	Internship	U. Thesis 1	U. Thesis 2	Passing Grade
1	1513000001	Ahmad Fariz	1	1	1	1	1	1	1
2	1513000002	Aneke Aldwi	1	1	1	1	1	1	1
3	1513000003	Arya Fatwa Gumilang	2	3	1	1	2	2	2
4	1513000004	Aulia Ulfa	1	1	1	1	1	1	1
5	1513000005	Baqizzarqoni	2	3	2	2	2	2	2
6	1513000006	Beby Faradina Anggraeny Hrp	2	3	2	2	2	2	2
7	1513000007	Bismar Fadli	1	1	1	1	1	1	1
8	1513000009	David H Christian Surbakti	2	3	2	2	2	2	2
9	1513000010	Desy Arista	1	1	1	1	1	1	1
10	1513000011	Dimas Rahmawan	2	3	2	2	2	2	2
11	1513000012	Dinda Novita Sari	2	2	1	1	1	1	1
12	1513000013	Erlina Hapni	1	1	1	1	1	1	1
13	1513000014	Gusdinda Pramiptha	2	2	1	1	1	1	1
14	1513000015	Hardiansyah	2	3	1	1	2	2	2
15	1513000016	Haris Fadillah Nasution	2	3	1	2	2	2	2

No.	NIM	Name	Grade	Failing Grade	Spec.	Internship	U. Thesis 1	U. Thesis 2	Passing Grade
16	1513000017	Kezia Veronika Sinaga	2	2	1	1	1	2	2
17	1513000018	Lestina Fraskawita Silalahi	2	3	2	2	2	2	2
18	1513000019	M. Alfin Fauzi	2	3	2	2	2	2	2
19	1513000020	M. Fauzi Syahputra	2	3	1	2	2	2	2
20	1513000024	Novia Anjelia	1	1	1	1	1	1	1
21	1513000025	Novita Sari	1	1	1	1	1	1	1
22	1513000026	NURUL HUSNAH	1	1	1	1	1	1	1
23	1513000027	Oktarina Wulandari	1	1	1	1	1	1	1
24	1513000030	Reza Riyanto	2	2	1	1	1	1	1
25	1513000031	Riani Ilyas	2	2	1	1	1	1	1
26	1513000032	Rizky Khairani Harahap	1	1	1	1	1	1	1
27	1513000033	Sherlina Dewi	1	1	1	1	1	1	1
28	1513000034	Sri Dewi Tarigan	1	1	1	1	1	1	1
.
.
.
83	1513000085	Diyah Afrita Wardani	2	2	1	1	1	1	1
84	1513000086	Dora Natalia Silalahi	2	2	1	1	1	1	1
85	1513000087	Dwi Indah Cahyani	2	3	1	2	2	2	2

Cleaning and Data Transformation Process

This stage removes inconsistent and incomplete data and transforms the data into data that can be processed using the K-Means and C4.5 algorithms. The conversion variables used are shown in table 2.

Table 2. Variables used

No	Variable	Conversion	Information
1	Grade	1	Pass
		2	Fail
2	Failing Grade	1	None
		2	$0 > x \leq 20$
		3	>20 SKS
3	Specialization	1	On-Time
		2	Not On Time
4	Internship	1	On-Time
		2	Not On Time
5	Undergraduate Thesis 1	1	On-Time
		2	Not On Time
6	Undergraduate Thesis 2	1	On-Time
		2	Not On Time
7	Passing Grade	1	On-Time
		2	Not On Time

Table 2 shows the variables that will be used to calculate on-time graduation using the K-Means Algorithm. Calculating with the K-Means Algorithm

K-Means algorithm steps:

- Determine the number of data groups
- Generate a random initial centroid
- Calculate the distance of data to each centroid.

$$d_{euclidean}(x, y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}$$

- Data group in

From the results of stage 3, data grouping is carried out.

- If there are data groups that are not the same or the group values are still changing, do step 3.

The formula below is to generate a new centroid:

$$C = \frac{\sum m}{n} \quad (1)$$

If the results of grouping the data in the first iteration are not the same as the groups initially, then the next step is to re-determine the value of the temporary data center. Use this formula :

$$C = \frac{\sum m}{n} \quad (1)$$

Calculate the distance to the new centroid for the next iteration.

C4.5 Algorithm

C4.5 stages:

- Select the attribute as the rootMemuat cabang untuk setiap nilai.
- Divide the cases into each branch.
- Then do the process is repeated for each branch until all branch cases can have the same class.

In selecting the attribute as the root, it is used based on the highest gain value obtained from the existing attributes. In calculating the gain value, equation 1 is used as follows:

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

Where:

S : Case set

A : Attribute

n : The number of partitions in Attribute A

$|S_i|$: The number of cases in the i-th partition

$|S|$: The number of cases in S.

Meanwhile, equation 2 is for calculating the entropy value, which can be seen in the following equation:

$$Entropy(S) = \sum_{i=1}^n -p_i * \log_2 p_i \quad (3)$$

Where:

S : Case set

A : Attribute

n : Number of Partitions S

p_i : The proportion of S_i to S

Table 3 presents a new group's creation in calculating the new centroid. Table 4 is the result data after the formation of a new cluster or the first iteration of 85 student data records. Table 5 is the result of iteration 2 calculations.

Table 3. Calculating Distance and Creating Groups

<i>Distance to C1</i>	<i>Distance to C2</i>	<i>Distance to C3</i>	<i>Group</i>
0.00	2.83	1.41	1
0.00	2.83	1.41	1
2.83	0.00	2.00	2
0.00	2.83	1.41	1
3.16	1.41	2.45	2
3.16	1.41	2.45	2
0.00	2.83	1.41	1
3.16	1.41	2.45	2
0.00	2.83	1.41	1
3.16	1.41	2.45	2
1.41	2.00	0.00	3
0.00	2.83	1.41	1

<i>Distance to C1</i>	<i>Distance to C2</i>	<i>Distance to C3</i>	<i>Group</i>
1.41	2.00	0.00	3
2.83	0.00	2.00	2
3.00	1.00	2.24	2
2.00	1.41	1.41	3
3.16	1.41	2.45	2
3.16	1.41	2.45	2
3.00	1.00	2.24	2
0.00	2.83	1.41	1
0.00	2.83	1.41	1
0.00	2.83	1.41	1
0.00	2.83	1.41	1
1.41	2.00	0.00	3
1.41	2.00	0.00	3
0.00	2.83	1.41	1
0.00	2.83	1.41	1
0.00	2.83	1.41	1
.	.	.	.
.	.	.	.
.	.	.	.
1.41	2.00	0.00	3
1.41	2.00	0.00	3
3.00	1.00	2.24	2

Table 4. New Cluster Calculation Results (Iteration 1)

<i>No.</i>	<i>NIM</i>	<i>Name</i>	<i>Grade</i>	<i>Failing Grade</i>	<i>Specialization</i>	<i>Internship</i>	<i>U. Thesis 1</i>	<i>U. Thesis 2</i>	<i>Passing Grade</i>	<i>Initial Cluster</i>	<i>New Cluster</i>	<i>Matching</i>
1	1513000001	Ahmad Fariz	1	1	1	1	1	1	1	0	1	-1
2	1513000002	Aneke Aldwi	1	1	1	1	1	1	1	0	1	-1
3	1513000003	Arya Fatwa Gumilang	2	3	1	1	2	2	2	0	2	-2
4	1513000004	Aulia Ulfa	1	1	1	1	1	1	1	0	1	-1
5	1513000005	Baqizzarqoni Beby	2	3	2	2	2	2	2	0	2	-2
6	1513000006	Faradina Anggraeny Hrp	2	3	2	2	2	2	2	0	2	-2
7	1513000007	Bismar Fadli David H	1	1	1	1	1	1	1	0	1	-1
8	1513000009	Christian Surbakti	2	3	2	2	2	2	2	0	2	-2
9	1513000010	Desy Arista	1	1	1	1	1	1	1	0	1	-1
10	1513000011	Dimas Rahmawan	2	3	2	2	2	2	2	0	2	-2
11	1513000012	Dinda Novita Sari	2	2	1	1	1	1	1	0	3	-3
12	1513000013	Erlina Hapni	1	1	1	1	1	1	1	0	1	-1
13	1513000014	Gusdinda Pramiptha	2	2	1	1	1	1	1	0	3	-3
14	1513000015	Hardiansyah Haris	2	3	1	1	2	2	2	0	2	-2
15	1513000016	Fadillah Nasution Kezia	2	3	1	2	2	2	2	0	2	-2
16	1513000017	Veronika Sinaga Lestina	2	2	1	1	1	2	2	0	3	-3
17	1513000018	Fraskawita Silalahi	2	3	2	2	2	2	2	0	2	-2
18	1513000019	M. Alfin Fauzi	2	3	2	2	2	2	2	0	2	-2
19	1513000020	M. Fauzi Syahputra	2	3	1	2	2	2	2	0	2	-2
20	1513000024	Novia Anjelia	1	1	1	1	1	1	1	0	1	-1
21	1513000025	Novita Sari	1	1	1	1	1	1	1	0	1	-1
22	1513000026	NURUL HUSNAH	1	1	1	1	1	1	1	0	1	-1
23	1513000027	Oktarina Wulandari	1	1	1	1	1	1	1	0	1	-1
24	1513000030	Reza Riyanto	2	2	1	1	1	1	1	0	3	-3
25	1513000031	Riani Ilyas Rizky	2	2	1	1	1	1	1	0	3	-3
26	1513000032	Khairani Harahap	1	1	1	1	1	1	1	0	1	-1
27	1513000033	Sherlina Dewi	1	1	1	1	1	1	1	0	1	-1

No.	NIM	Name	Grade	Failing Grade	Specialization	Internship	U. Thesis 1	U. Thesis 2	Passing Grade	Initial Cluster	New Cluster	Matching
28	1513000034	Sri Dewi Tarigan	1	1	1	1	1	1	1	0	1	-1
.
83	1513000085	Diyah Afrita Wardani	2	2	1	1	1	1	1	0	3	-3
84	1513000086	Dora Natalia Silalahi	2	2	1	1	1	1	1	0	3	-3
85	1513000087	Dwi Indah Cahyani	2	3	1	2	2	2	2	0	2	-2

Tabel 5. Hasil Perhitungan Cluster Baru (Iterasi 2)

No	NIM	Nama	Nilai	Nilai Gagal	Peminatan	PKL	Skripsi 1	Skripsi 2	Lulus	Cluster Awal	Cluster Baru	Pencocokan
1	1513000001	Ahmad Fariz	1	1	1	1	1	1	1	1	1	0
2	1513000002	Aneke Aldwi	1	1	1	1	1	1	1	1	1	0
3	1513000003	Arya Fatwa Gumilang	2	3	1	1	2	2	2	2	2	0
4	1513000004	Aulia Ulfa	1	1	1	1	1	1	1	1	1	0
5	1513000005	Baqizzarqoni	2	3	2	2	2	2	2	2	2	0
6	1513000006	Beby Faradina Anggraeny Hrp	2	3	2	2	2	2	2	2	2	0
7	1513000007	Bismar Fadli	1	1	1	1	1	1	1	1	1	0
8	1513000009	David H Christian Surbakti	2	3	2	2	2	2	2	2	2	0
9	1513000010	Desy Arista	1	1	1	1	1	1	1	1	1	0
10	1513000011	Dimas Rahmawan	2	3	2	2	2	2	2	2	2	0
11	1513000012	Dinda Novita Sari	2	2	1	1	1	1	1	3	3	0
12	1513000013	Erlina Hapni	1	1	1	1	1	1	1	1	1	0
13	1513000014	Gusdinda Pramiptha	2	2	1	1	1	1	1	3	3	0
14	1513000015	Hardiansyah	2	3	1	1	2	2	2	2	2	0
15	1513000016	Haris Fadillah Nasution	2	3	1	2	2	2	2	2	2	0
16	1513000017	Kezia Veronika Sinaga	2	2	1	1	1	2	2	3	3	0
17	1513000018	Lestina Fraskawita Silalahi	2	3	2	2	2	2	2	2	2	0
18	1513000019	M. Alfin Fauzi	2	3	2	2	2	2	2	2	2	0
19	1513000020	M. Fauzi Syahputra	2	3	1	2	2	2	2	2	2	0
20	1513000024	Novia Anjelia	1	1	1	1	1	1	1	1	1	0
21	1513000025	Novita Sari	1	1	1	1	1	1	1	1	1	0
22	1513000026	NURUL HUSNAH	1	1	1	1	1	1	1	1	1	0
23	1513000027	Oktarina Wulandari	1	1	1	1	1	1	1	1	1	0
24	1513000030	Reza Riyanto	2	2	1	1	1	1	1	3	3	0
25	1513000031	Riani Ilyas	2	2	1	1	1	1	1	3	3	0
26	1513000032	Rizky Khairani Harahap	1	1	1	1	1	1	1	1	1	0
27	1513000033	Sherlina Dewi	1	1	1	1	1	1	1	1	1	0
28	1513000034	Sri Dewi Tarigan	1	1	1	1	1	1	1	1	1	0
.
.
.
83	1513000085	Diyah Afrita Wardani	2	2	1	1	1	1	1	3	3	0
84	1513000086	Dora Natalia Silalahi	2	2	1	1	1	1	1	3	3	0
85	1513000087	Dwi Indah Cahyani	2	3	1	2	2	2	2	2	2	0

RESULTS AND DISCUSSION

Testing the Information Systems Study Program's data is as many as 85 records using the RapidMiner tool.

1. Analysis with the K-Means Algorithm

The first data analysis was performed to group data or the data clustering process. The data clustering process uses the K-means method. There are 3 clusters to be formed by randomly taking the initial centroid. The results of observations using the K-Means algorithm obtained the center of the data cluster according to the image below, and Figure 2 is the result of the calculation of the last centroid value of the student's Grade using RapidMiner, the first cluster center for Grade = 2, Failing Grade = 3, Specialization = 1.6, internship Grade = 1.8, Undergraduate Thesis 1 = 2, Undergraduate Tesis 2 = 2 and Passing Grade = 2, while for the second cluster has the center of the cluster = 1 for all variables, while in the third cluster the center of the cluster Grade = 2, Failing Grade = 2, Specialization = 1, Internship = 1, Undergraduate Thesis 1 = 1, Undergraduate Thesis 2 = 1.2 and Passing Grade = 1.2.

Attribute	Cluster_0	Cluster_1	Cluster_2
Grade	2	1	2
Failing Grade	3	1	2
Specialization	1.600	1	1
Internship	1.800	1	1
Undergraduate Thesis 1	2	1	1
Undergraduate Thesis 2	2	1	1.200
Passing Grade	2	1	1.200

Figure 2. The last Centroid value from the testing process with the RapidMiner software

Figure 2 describes the graph of the student grade clusters' results, where the blue color shows the first cluster, the red color shows the third cluster, and the green shows the second cluster.

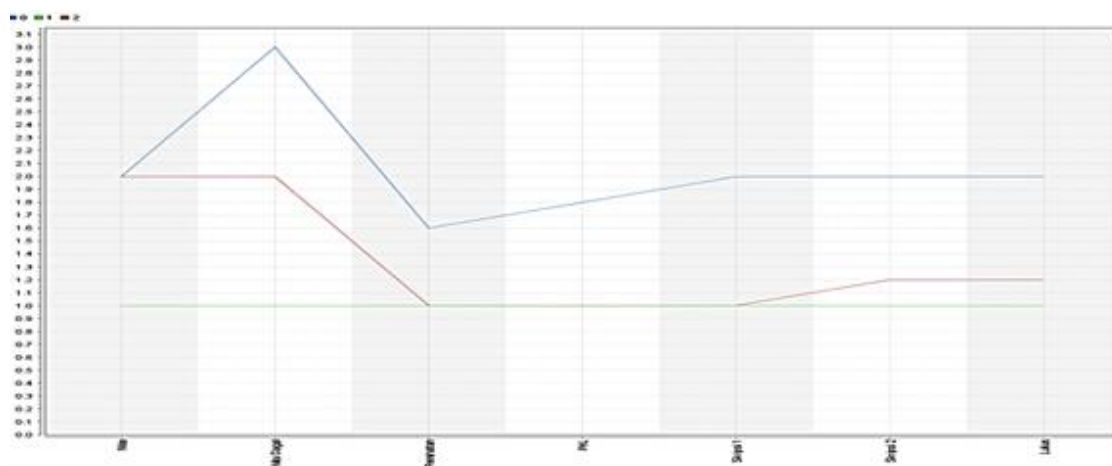


Figure 3. Clustering graph of the test results with the RapidMiner software

Figure 4 describes the process of retrieving data or the process of retrieving research data and installation using a clustering operator. In this section, the number of clusters is determined = 3.



Figure 4. Results of data retrieving and data installation with Clustering operators

2. Analyzing Data with C4.5

Based on the results of data grouping using the K-Means algorithm, then analyzed with the C4.5 algorithm to obtain knowledge that can be used to predict graduation. The data used were as many as 85 records with 3 sub-output variables, namely Cluster 1, cluster 2, and Cluster 3. It can be seen in Figure 5 that it explains the results of the existing data records to produce a decision tree to determine the criteria for students who pass on time, where the output variable is obtained from the results of processing with the K-Means algorithm.

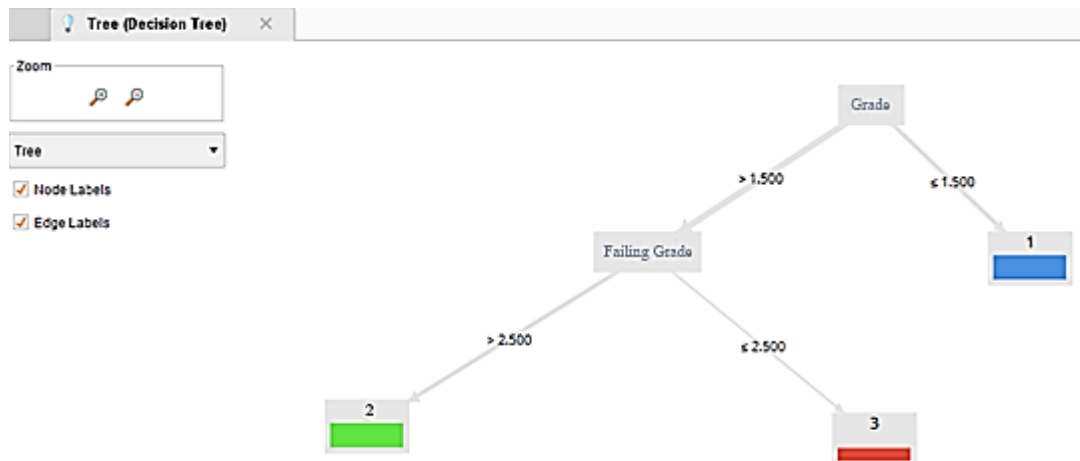


Figure 5. Result of Decision Tree Display in RapidMiner software

Knowledge generated:

- a. If $\text{Grade} \leq 1.5$, Then Cluster 1 (Students Graduate On Time With Good Grade)
- b. If $\text{Grade} > 1.5$ and $\text{Failing Grade} > 2.5$ Than Cluster 2 (Students Graduate Not On-Time with Problematic Grade)
- c. If $\text{Grade} > 1.5$ and $\text{Failing Grade} \leq 2.5$, Then Cluster 2 (Students Graduate On Time With Problematic Grade)

CONCLUSION AND RECOMMENDATION

Based on research conducted in data grouping, the results of data grouping are more optimal for the prediction process using the C4.5 algorithm. The K-Means algorithm can be used simultaneously with the C4.5 algorithm in research has many variables and many sub-variables. (i) The K-means algorithm can group student graduation data. In this study, they were grouped into 3 groups: students graduating on time with passing Grade, the second group of students not graduating on time due to problematic grades, and the

third group of students graduating on time with problematic grades. (ii) The C4.5 algorithm can produce knowledge that can be used as a basis for predicting student graduation. Based on the processing of the C4.5 algorithm, the main criterion for determining student graduation is Grade, followed by the Failing Grade. The percentage of testing the suitability of knowledge with test data is obtained $\geq 90\%$.

It's recommended to determine the initial centroid before performing calculations using additional algorithms and determine the number of groups to be generated so that the resulting grouping results are more optimal. For further research, another tool can be used, not just RapidMiner.

REFERENCES

- Elizawati, N., & Lesmana, L. S. (2017). Analisis Nilai Rapor Siswa Kelas X Jurusan Multimedia Terhadap Minat pada Pelajaran Produktif Di Kelas XII untuk Menentukan Kompetensi Siswa dengan Metode Clustering Algoritma K-Means (Studi Kasus Di SMKN 4 Padang). *Jurnal Komputer Terapan*, 3(2), 133-148.
- Hairani, H., Saputro, K. E., & Fadli, S. (2020). K-means-SMOTE untuk menangani ketidakseimbangan kelas dalam klasifikasi penyakit diabetes dengan C4.5, SVM, dan naive Bayes. *Jurnal Teknologi Dan Sistem Komputer*.
- Helilintar, R., & Farida, I. N. (2018). Penerapan Algoritma K-Means Clustering Untuk Prediksi Prestasi Nilai Akademik Mahasiswa. *Jurnal Sains Dan Informatika*. <https://doi.org/10.34128/jsi.v4i2.140>
- Muliono, R., & Sembiring, Z. (2019). Data Mining Clustering Menggunakan Algoritma K-Means Untuk Klasterisasi Tingkat Tridarma Pengajaran Dosen. *CESS (Journal of Computer Engineering, System and Science)*.
- Mulyani, E. D. S. dkk. (2018). Implementasi Algoritma K-Means Dan Fp- Growth Untuk Rekomendasi Bimbingan Belajar Berdasarkan Segmentasi Akademik Siswa. *IT Journal*.
- Nanda, A. P., Pramono, D. E. H., & Hartati, S. (2020). Menentukan Tingkat Kepuasan Mahasiswa Terhadap Pelayanan Akademik Menggunakan Metode Algoritma K-Means. *Explore: Jurnal Sistem Informasi dan Telematika (Telekomunikasi, Multimedia dan Informatika)*, 11(1), 23-28.
- Nasari, F., & Sianturi, C. J. M. (2016). Penerapan Algoritma K-Means Clustering Untuk Pengelompokan Penyebaran Diare Di Kabupaten Langkat. *CogITO Smart Journal*. <https://doi.org/10.31154/cogito.v2i2.19.108-119>
- Pradnyana, G. A., & Permana, A. A. J. (2018). Sistem Pembagian Kelas Kuliah Mahasiswa Dengan Metode K-Means Dan K-Nearest Neighbors Untuk Meningkatkan Kualitas Pembelajaran. *JUTI: Jurnal Ilmiah Teknologi Informasi*. <https://doi.org/10.12962/j24068535.v16i1.a696>
- Priyatman, H., Sajid, F., & Haldivany, D. (2019). Klasterisasi Menggunakan Algoritma K-Means Clustering untuk Memprediksi Waktu Kelulusan Mahasiswa. *Jurnal Edukasi Dan Penelitian Informatika (JEPIN)*. <https://doi.org/10.26418/jp.v5i1.29611>
- Putra, R. R., & Wadisman, C. (2018). Implementasi Data Mining Pemilihan Pelanggan Potensial Menggunakan Algoritma K Means. *INTECOMS: Journal of Information Technology and Computer Science*. <https://doi.org/10.31539/intecom.v1i1.141>
- Ridwan, M., Suyono, H., & Sarosa, M. (2013). Penerapan Data Mining Untuk Evaluasi Kinerja Akademik Mahasiswa Menggunakan Algoritma Naive Bayes Classifier. *Jurnal EECCIS*.
- Rosmini, R., Fadlil, A., & Sunardi, S. (2018). Implementasi Metode K-Means Dalam Pemetaan Kelompok Mahasiswa Melalui Data Aktivitas Kuliah. *IT JOURNAL RESEARCH AND DEVELOPMENT*. [https://doi.org/10.25299/itjrd.2018.vol3\(1\).1773](https://doi.org/10.25299/itjrd.2018.vol3(1).1773)
- Saleh, A., & Nasari, F. (2018). Penggunaan Teknik Unsupervised Discretization pada Metode Naive Bayes dalam Menentukan Jurusan Siswa Madrasah Aliyah. *Jurnal Teknologi Informasi Dan Ilmu Komputer*. <https://doi.org/10.25126/jtiik.201853705>