p-ISSN : 2086-7867
e-ISSN : 2808-9170

APPLICATION OF DATA MINING FOR PREDICTION OF STUDENTS OUT OF COLLEGE ALGORITHM C4.5

Suandi Daulay[1], Wira Apriani[2], Yuda Perwira[3]

[1]Sekolah Tinggi Teknologi Pekanbaru, Jl. Dirgantara No.4 Riau, Indonesia
[2,3]STMIK Pelita Nusantara, Jl. Iskandar Muda No. 1 Medan, Indonesia

Email : suwandidulay90@gmail.com, wiraaprianii@gmail.com,
yudaperwira25@gmail.com

Abstract

This research was conducted to predict students dropping out of private universities, the student department needs to pay attention to students who have the potential to drop out so that they can be detected faster to make an approach with students so they don't drop out of college, with the help of data mining so that data -The data collected is useful information and with the C4.5 method so that predictions become accurate to detect students who have the potential to drop out of college. As for the results of this study, it is known that the most influential variable for students dropping out of college is marked by UKT Not Current Then Often Absent Then Gender Male whose graduation year is not recently graduated (not fresh graduate).

***Keywords** : DataMining; Prediction; Students Drop Out; C4.5*

1. Introduction

Students are a valuable asset owned by this country, students are considered as the future for this country, before they have a career and work in society and their audiences have to take formal education on campus, but there are always students who drop out of college either at state universities or colleges. private sector, this will harm the student, his family and the campus for that the campus feels the need to pay attention to students so they don't drop out of college.

To become the best private university, it is deemed necessary to pay attention to students in various aspects, the student department is expected to be able to take many approaches to students, both high achievers and students who have the potential to drop out of college in order to detect early students who have the potential to drop out of college and reduce the number of students who drop out of college. so that the problem can be solved.

Data mining is a term used to find hidden knowledge in databases. Data mining is a semi-automatic process that uses statistical, mathematical, artificial intelligence and machine learning techniques to extract and identify potentially useful and useful knowledge information stored in large databases (Turban et al, 2005).

The C.45 algorithm is a method related to Data Mining C.45 is a classification algorithm based on the linear classifier principle and has been widely used for various studies, the method is considered suitable for prediction problems for students who have the potential to drop out of college.

2. Methodology

Decision tree with the C.45 algorithm is a classification method that uses a tree structure representation where each node represents an attribute, the branch represents the value of the attribute, and the leaf represents the class. The top node of the decision tree is called the root. In the decision tree there are 3 types of nodes, namely: 1. Root Node, is the top node, at this node there is no input and can have no output or have more than one output. 2. Internal Node, is a branching node, at this node there is only one input and has a minimum of two outputs. 3. Leaf node or terminal node, is the end node, at this node there is only one input and no output.

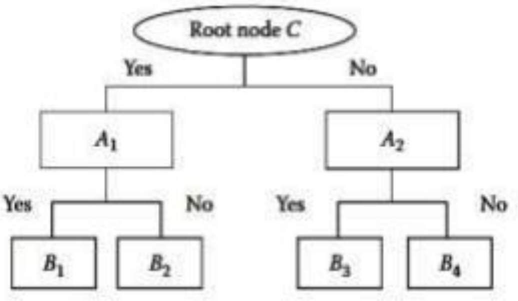


Figure 1 Decision Tree

Decision trees depend on if-then rules, but do not require parameters and metrics. The simple and interpretable structure allows the decision tree to solve multi-type attribute problems. Decision trees can also manage missing values or noise data.

The steps for the decision tree method with the C.45 algorithm in building a decision tree are as follows:

1. Forming a decision system consisting of condition attributes and decision attributes. Shows an example of a decision system in this study. It only consists of n objects, E1, E2, E3, E4, ..., En and condition attributes, namely sales, purchases, warehouse stock, and operating expenses. While profit is a decision attribute.
2. Count the number of column data, the number of data based on the result attribute members with certain conditions. For the first process the conditions are still empty.
3. Select attribute as Node.
4. Create a branch for each member of the Node.
5. Check if the entropy value of any Node member is zero. If so, determine which leaves are formed. If all the entropy values of the Node members are zero, then the process stops
6. If any Node member has an entropy value greater than zero, repeat the process from the beginning with Node as a condition until all members of the Node are zero.

Node is the attribute that has the highest gain value of the existing attributes. To calculate the gain value of an attribute, a formula as shown in the following equation is used:

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

Information:
S = Case Set
A = Attribute
n = number of attribute partitions A
|Si| = Proportion of Si to S
|S| = number of cases in S
Meanwhile, to calculate the value of Entropy can be seen in the following equation:

$$Entropy(S) = \sum_{i=1}^{n} - pi * \log_2 pi$$ ………………………………………………………(2)

Keterangan :

S= Case Set
n= number of partitions S
Pi = proportion of Si to S

3. Results

a. Data Selection
The datasets obtained were collected and then the criteria were selected into 8 criteria, namely:

1. (A) Class there are two sub-criteria in the class, namely morning and afternoon classes,
2. (B) Origin of School there are 6 Sub-criteria from school, namely Madrasah Aliyah Negri, Madrasah Aliyah Private, State High School, Private High School, State Vocational School and Private Vocational School
3. (C) Graduation Year There are 2 Sub-criteria for Graduation Year, namely recently graduated or not recently graduated
4. (D) Working Status there are 2 sub-criteria for Working Status, namely Working and not working
5. (E) Marital Status there are 2 sub-criteria for marital status, namely married and unmarried
6. (F) Absence there are 3 sub-criteria for absence, namely often rarely and never.
7. (G) UKT there are 2 sub-criteria for UKT, namely Current and Non-current
8. (H) Active Status, there are 2 sub-criteria for Active Status, namely Active and Inactive

The selection of these criteria is carried out to be used as the determining criteria in data classification

b. Data Classification

Selection Remove duplication of data, check for inconsistent data, and correct errors in data, such as typographical errors. An encrihment process is also carried out, namely the process of "enriching" existing data with other relevant data or information needed for KDD, such as external data or information.

c. Data Transformation

Transforming data forms that do not have clear entities into valid data forms or ready for data mining processes.

The following data have been completed in the selection, classification and transformation stages:

Table 1.
Knowledge Database

No Reg	Class	Gender	School Origin	Graduation year	Working Status	Marital status	Roll call	UKT	Status
176	Morning	Man	Mas	Not	Not	Single	There isn't any	Fluent	Active
171	Afternoon	Man	Mas	Just Finished	Working	Single	Often	Fluent	Active
167	Afternoon	Man	Public High School	Just Finished	Not	Single	Seldom	Fluent	Active
68	Afternoon	Man	Senior High School	Just Finished	Working	Single	There isn't any	Fluent	Active
202	Afternoon	Man	Senior High School	Just Finished	Not	Single	There isn't any	Fluent	Active
91	Afternoon	Man	Smk S	Just Finished	Working	Single	There isn't any	Fluent	Active
217	Afternoon	Man	N high school	Just Finished	Working	Single	There isn't any	Fluent	Active
95	Afternoon	Man	Smk S	Just Finished	Working	Single	Often	Fluent	Active
11	Afternoon	Man	Smk S	Just Finished	Working	Single	There isn't any	Fluent	Active
137	Afternoon	Man	Smk S	Just Finished	Not	Single	There isn't any	Fluent	Active
132	Morning	Man	Public High School	Just Finished	Not	Single	There isn't any	Fluent	Active
15	Morning	Man	Mas	Just Finished	Not	Single	There isn't any	Fluent	Active
61	Morning	Man	Mas	Just Finished	Not	Single	There isn't any	Fluent	Active
120	Morning	Man	Public High School	Just Finished	Not	Single	There isn't any	Fluent	Active
59	Morning	Man	Smk S	Just Finished	Not	Single	There isn't any	Fluent	Active
189	Morning	Man	N high school	Just Finished	Not	Single	There isn't any	Fluent	Active
26	Morning	Man	Smk S	Just Finished	Not	Single	There isn't any	Fluent	Active

126	Morning	Man	Public High School	Just Finished	Not	Single	There isn't any	Fluent	Active
106	Morning	Man	N high school	Just Finished	Not	Single	There isn't any	Fluent	Active
163	Afternoon	Man	Mas	Not	Working	Single	Seldom	Fluent	Active
5	Afternoon	Man	Mas	Not	Working	Marry	There isn't any	Fluent	Active
20	Afternoon	Man	Public High School	Not	Working	Single	Seldom	Fluent	Active
1	Afternoon	Man	Senior High School	Not	Working	Single	There isn't any	Fluent	Active
17	Afternoon	Man	N high school	Not	Working	Single	There isn't any	Fluent	Active
169	Afternoon	Man	N high school	Not	Working	Single	Often	Fluent	Active
32	Afternoon	Man	N high school	Not	Working	Single	There isn't any	Fluent	Active
173	Afternoon	Man	Smk S	Not	Working	Single	There isn't any	Fluent	Active
170	Afternoon	Man	Smk S	Not	Working	Single	There isn't any	Fluent	Active
221	Afternoon	Man	Smk S	Not	Working	Single	Seldom	Fluent	Active
92	Morning	Man	Mas	Not	Not	Single	There isn't any	Fluent	Active
151	Morning	Man	man	Not	Not	Single	There isn't any	Fluent	Active
125	Morning	Man	Public High School	Not	Not	Single	There isn't any	Fluent	Active
13	Morning	Man	N high school	Not	Not	Single	There isn't any	Fluent	Active
43	Morning	Man	Smk S	Not	Not	Single	There isn't any	Fluent	Active
211	Morning	Man	Public High School	Not	Not	Single	There isn't any	Fluent	Active
204	Morning	Man	Senior High School	Not	Not	Single	There isn't any	Fluent	Active
87	Morning	Man	N high school	Not	Not	Single	There isn't any	Fluent	Active
83	Morning	Man	N high school	Not	Not	Single	Seldom	Fluent	Active
181	Afternoon	Woman	Mas	Just Finished	Working	Single	Seldom	Fluent	Active
44	Afternoon	Woman	Public High School	Just Finished	Working	Single	There isn't any	Fluent	Active
144	Afternoon	Woman	Public High School	Just Finished	Not	Single	Seldom	Fluent	Active

112	Afternoon	Woman	Public High School	Just Finished	Working	Single	Seldom	Fluent	Active
113	Afternoon	Woman	Senior High School	Just Finished	Not	Single	There isn't any	Fluent	Active
50	Afternoon	Woman	N high school	Just Finished	Working	Single	There isn't any	Fluent	Active
152	Afternoon	Woman	Smk S	Just Finished	Working	Single	There isn't any	Fluent	Active
149	Afternoon	Woman	Smk S	Just Finished	Not	Single	There isn't any	Fluent	Active
22	Morning	Woman	Mas	Just Finished	Not	Single	There isn't any	Fluent	Active
29	Morning	Woman	man	Just Finished	Not	Single	There isn't any	Fluent	Active
54	Morning	Woman	Public High School	Just Finished	Not	Single	There isn't any	Fluent	Active
55	Morning	Woman	Public High School	Just Finished	Not	Single	Seldom	Fluent	Active
175	Morning	Woman	Senior High School	Just Finished	Not	Single	There isn't any	Fluent	Active
10	Morning	Woman	N high school	Just Finished	Not	Single	There isn't any	Fluent	Active
56	Morning	Woman	Smk S	Just Finished	Not	Single	There isn't any	Fluent	Active
73	Morning	Woman	Senior High School	Just Finished	Not	Single	There isn't any	Fluent	Active
35	Morning	Woman	N high school	Just Finished	Not	Single	Seldom	Fluent	Active
84	Morning	Woman	N high school	Just Finished	Not	Single	There isn't any	Fluent	Active
215	Morning	Woman	Smk S	Just Finished	Not	Single	There isn't any	Fluent	Active
192	Afternoon	Woman	man	Not	Working	Single	There isn't any	Fluent	Active
18	Afternoon	Woman	man	Not	Working	Single	Seldom	Fluent	Active
33	Afternoon	Woman	Package C	Not	Working	Single	Often	Fluent	Active
4	Afternoon	Woman	Public High School	Not	Working	Marry	Seldom	Fluent	Active
200	Afternoon	Woman	Public High School	Not	Working	Single	Seldom	Fluent	Active
183	Afternoon	Woman	Public High School	Not	Working	Single	There isn't any	Fluent	Active

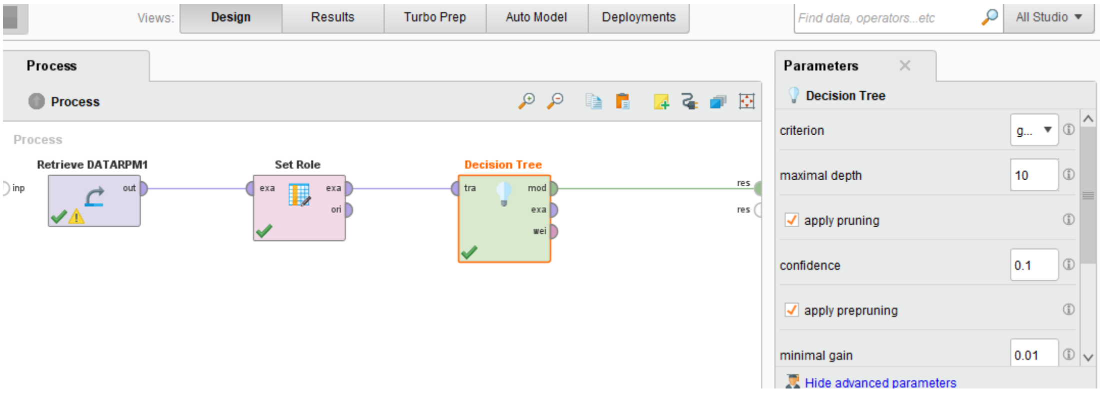
185	Afternoon	Woman	Senior High School	Not	Working	Single	There isn't any	Fluent	Active
159	Afternoon	Woman	Senior High School	Not	Not	Single	There isn't any	Fluent	Active
131	Afternoon	Woman	Senior High School	Not	Working	Marry	There isn't any	Fluent	Active
97	Afternoon	Woman	Smk S	Not	Working	Single	There isn't any	Fluent	Active
158	Afternoon	Woman	N high school	Not	Working	Single	There isn't any	Fluent	Active
64	Afternoon	Woman	N high school	Not	Working	Single	Seldom	Fluent	Active
24	Afternoon	Woman	N high school	Not	Working	Single	Often	Fluent	Active
51	Afternoon	Woman	N high school	Not	Working	Single	There isn't any	Fluent	Active
143	Afternoon	Woman	Smk S	Not	Working	Single	There isn't any	Fluent	Active
52	Afternoon	Woman	Smk S	Not	Working	Single	Often	Fluent	Active
225	Afternoon	Woman	Smk S	Not	Working	Single	There isn't any	Fluent	Active
188	Morning	Woman	man	Not	Not	Single	Seldom	Fluent	Active
213	Morning	Woman	Public High School	Not	Not	Single	There isn't any	Fluent	Active
207	Morning	Woman	Senior High School	Not	Not	Single	There isn't any	Fluent	Active
49	Morning	Woman	Smk S	Not	Not	Single	There isn't any	Fluent	Active
104	Morning	Woman	N high school	Not	Not	Single	There isn't any	Fluent	Active
86	Morning	Woman	Smk S	Not	Not	Single	There isn't any	Fluent	Active
34	Morning	Woman	N high school	Not	Not	Single	There isn't any	Fluent	Active
146	Morning	Woman	N high school	Not	Not	Single	Seldom	Fluent	Active
14	Morning	Woman	Smk S	Not	Not	Single	There isn't any	Fluent	Active
136	Afternoon	Man	N high school	Just Finished	Working	Single	Seldom	Not smooth	Not active
39	Morning	Man	Public High School	Just Finished	Not	Single	There isn't any	Not smooth	Not active
164	Afternoon	Man	Mas	Not	Not	Single	Often	Not smooth	Not active
82	Afternoon	Man	Public High School	Not	Working	Single	Often	Not smooth	Not active
209	Afternoon	Man	N high school	Not	Working	Single	Often	Not smooth	Not active

148	Afternoon	Man	Smk S	Not	Working	Single	Often	Not smooth	Not active
198	Morning	Man	man	Not	Not	Single	Often	Not smooth	Not active
128	Morning	Man	Public High School	Not	Not	Single	There isn't any	Not smooth	Not active
184	Afternoon	Woman	Mas	Just Finished	Not	Single	Often	Not smooth	Not active
203	Afternoon	Woman	Public High School	Just Finished	Working	Single	Seldom	Not smooth	Not active
119	Afternoon	Woman	N high school	Just Finished	Working	Single	Often	Not smooth	Not active
12	Afternoon	Woman	Smk S	Just Finished	Working	Single	Seldom	Not smooth	Not active
74	Morning	Woman	Public High School	Just Finished	Not	Single	Seldom	Not smooth	Not active
23	Morning	Woman	N high school	Just Finished	Not	Single	Seldom	Not smooth	Not active
102	Afternoon	Woman	Public High School	Not	Working	Single	Seldom	Not smooth	Not active
107	Afternoon	Woman	Senior High School	Not	Not	Single	Often	Not smooth	Not active
142	Afternoon	Woman	N high school	Not	Working	Single	Seldom	Not smooth	Not active
205	Afternoon	Woman	Smk S	Not	Working	Single	Seldom	Not smooth	Not active
94	Afternoon	Woman	Smk S	Not	Working	Single	Often	Not smooth	Not active
98	Morning	Woman	Public High School	Not	Not	Single	There isn't any	Not smooth	Not active

d. C4.5 . Algorithm Implementation

C4.5 Algorithm Test Results With Rapid Miner

Implementation of the Decision Tree Model with Rapid Miner



Decision Tree

criterion g...

maximal depth 10

☑ apply pruning

confidence 0.1

☑ apply prepruning

minimal gain 0.01

Fig 1 C4.5 Algorithm Design With Rapid Miner

After the design is complete, the decision tree model will be run to get a decision tree

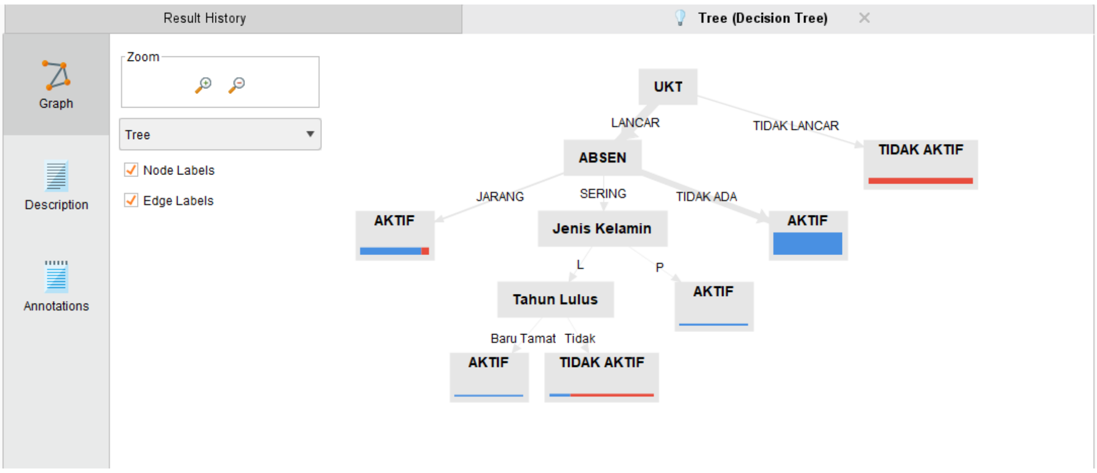
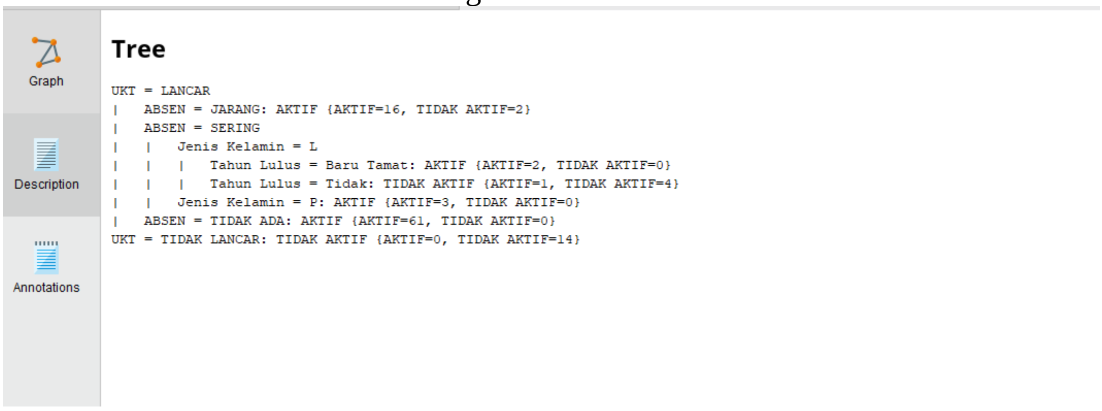


Fig. 2 Decision tree C4.5

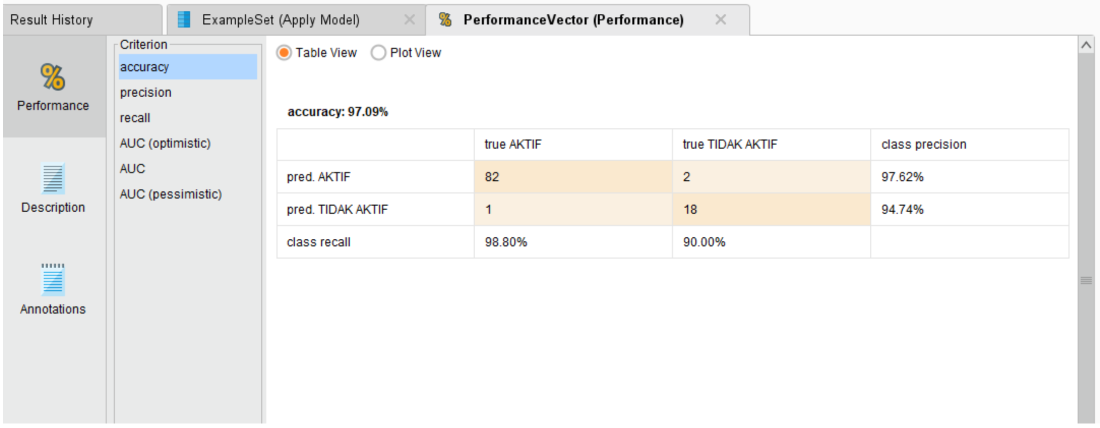


Tree

```
UKT = LANCAR
|   ABSEN = JARANG: AKTIF {AKTIF=16, TIDAK AKTIF=2}
|   ABSEN = SERING
|   |   Jenis Kelamin = L
|   |   |   Tahun Lulus = Baru Tamat: AKTIF {AKTIF=2, TIDAK AKTIF=0}
|   |   |   Tahun Lulus = Tidak: TIDAK AKTIF {AKTIF=1, TIDAK AKTIF=4}
|   |   Jenis Kelamin = P: AKTIF {AKTIF=3, TIDAK AKTIF=0}
|   ABSEN = TIDAK ADA: AKTIF {AKTIF=61, TIDAK AKTIF=0}
UKT = TIDAK LANCAR: TIDAK AKTIF {AKTIF=0, TIDAK AKTIF=14}
```

Fig. 3 Description of decision tree rules

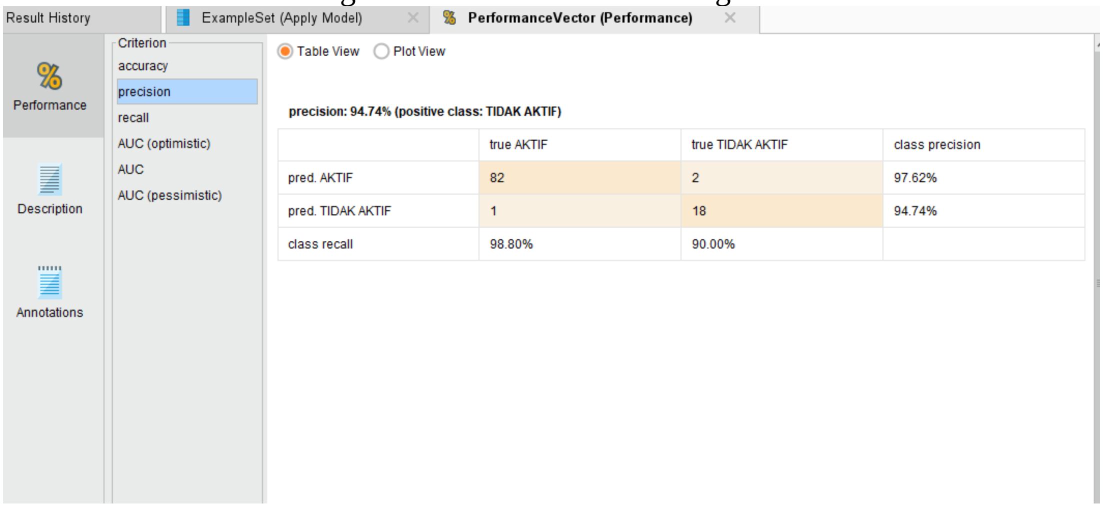
After that, test data will be made to determine the accuracy of the .c.45 . algorithm



accuracy: 97.09%

	true AKTIF	true TIDAK AKTIF	class precision
pred. AKTIF	82	2	97.62%
pred. TIDAK AKTIF	1	18	94.74%
class recall	98.80%	90.00%	

Fig. 5 Performance data Test algorithm C.45



precision: 94.74% (positive class: TIDAK AKTIF)

	true AKTIF	true TIDAK AKTIF	class precision
pred. AKTIF	82	2	97.62%
pred. TIDAK AKTIF	1	18	94.74%
class recall	98.80%	90.00%	

Fig. 6. Data Algorithm Test C4.5

The test results using Rapidminer are described through the Confusion Matrix to determine the value of accuracy, recall and precision. Classification using the Decission Tree algorithm produces an accuracy value of 97.9% and the Area Under the Curve value is calculated to measure the difference in performance, with a result of 0.742 which is included in the medium category or Fair Classification.

4. Conclusion

Prediction of college dropout students using the Decision Tree algorithm model produces an accuracy of 97.9% and an AUC of 0.742, and the Support Vector algorithm. The most influential factor for students dropping out of college is marked by non-current UKT and then Frequent Absence Then Gender Male whose year of graduation is not recently graduated (not fresh graduate), while the variables of origin of school, time of study, working status and marital status do not really affect students who dropped out of college.

References

[1]. Andari, Shofi., Santi W, Purnami., dan Bambang W, Otok (2013) “Smooth Support Vector Machine dan Multivariate Adaptive Regression Spline untuk Mendiagnosis Kanker Payudara. Vol. 1, No 2.

[2]. Huang, Chia-Hui., Keng-CHieh, Yang., dan Kao, Han-Ying. (2014) “Analyzing Big Data With The Hybrid Interval Regression Methods” Vol. 2014.

[3]. MA Sembiring , MFL Sibuea, A Sapta, “Analisa Kinerja Algoritma C.45 Dalam Memprediksi Hasil Belajar”. **Journal of Science and Social Research ISSN 2615 – 4307 (Print) February 2018, I (1): 73 – 79 ISSN 2615 – 3262 (Online)**

[4]. Qin, Chuandong., dan Zhao, Huixia. (2014) “Selecting The Optimal Combination Model Of FSSVM For The Imbalance Datasets” Vol. 2014.

[5]. Rachman, Farizi., dan Wulan Purnami, Santi. (2012) “Perbandingan Klasifikasi Tingkat Keganasan Breast Cancer Dengan Menggunakan Regresi Logistik Ordinal dan Support Vector Machine”, **Institut Teknologi Sepuluh Nopember (ITS) Surabaya.**

[6]. Raharjo, Suwanto., dan Winarko, Edi. (2014) “Klasterisasi, Klasifikasi dan Peringkasan Teks Berbahasa Indonesia” Universitas Gunadarma. Hartama, Dedy (2012) “Model Aturan Keterhubungan Data Mahasiswa Dengan Algoritma Decision Tree”, **AMIK Tunas Bangsa Pematang Siantar.**

[7]. Ryci Rahmawati Fiska, “Penerapan Teknik Data Mining dengan Metode Support Vector Machine (SVM) untuk Memprediksi Siswa yang Berpeluang Drop Out (Studi Kasus di SMKN 1 Sutera)” **SATIN - Sains dan Teknologi Informasi, Vol. 3, No. 1, Juni 2017.**

[8]. Sembiring, M.A., 2016. Penerapan Metode Decission Tree Algoritma C.45 Untuk Memprediksi Hasil Belajar Mahasiswa Berdasarkan Riwayat Akademik. **JURTEKSI, 3(1), pp.60-65.**

[9]. Sibuea, M.F.L., 2017. Implementasi Model Pembelajaran Kooperatif Tipe Think Talk Write (Ttw) Sebagai Upaya Meningkatkan Hasil Belajar Siswa. MES **(journal of mathematics education and science), 2(2)**

[10]. Tampubolon, Kennedi., Hoga, Saragih., dan Bobby, Reza. (2013) “Implementasi Data Mining Algoritma Apriori pada Sistem Persediaan alat-alat Kesehatan”. Vol. 1, No. 1.

[11]. Widyarini, Tiananda., Budi Sentosa (2009) “Aplikasi Metode Cross Entropy Untuk Support Vector Machine”. **Institut Teknologi Sepuluh Nopember (ITS) Surabaya.**

[12]. Wulan Purnami, Santi., Mohammad Zain, Jasni., dan Heriawan, Tutut. (2011) “An Alternative Algorithm For Classification Large Categorical Dataset : k-mode Clustering Reduced Support Vector Machine”. **Universitas Malaysia Pahang.**