JoCoSiR
Journal of Computer Science Research ISSNXXXX-XXXX (Online)

Implementation of data mining with the c4.5 algorithm for student majors (Case Study: SMA N 1 Bp.Mandoge)

[1]Nabilah Sapira Putri, 2M. Safii
1.2 Department of STIKOM Tunas Bangsa Pematang Siantar Information Systems
e-mail: anabilahsapiraputri2001@gmail.com , bm.safii@amiktinasbangsa.ac.id

ARTICLEINFO

Article history:
Received Jan 02, 2023
Revised Jan 16, 2023
Accepted Jan 30, 2023
Available online

Keywords:
C4.5 Algorithm;
Data Mining;
Decision Tree;
Student Majors;
Classification;

IEEE style in citing this article:
Nabilah Sapira Putri, M. Safii, "Article Title," JoCoSiR: Scientific Journal of Information Systems Technology, vol. 1, no. 1, pp. 18-28, 2023.

ABSTRACT

Classification of student majors is the process of grouping students according to abilities (values), talents and interests that are relatively the same so that the lessons that will be given to students will be more focused and directed. The process of classifying student data can be explored for patterns in the field of data mining, namely the process of obtaining relationships or patterns from large data so as to provide useful indications. SMA N 1 Bp.Mandoge is one of the educational institutions that started introducing majors and divided them into two choices of majors, namely "IPA" and "IPS". The curriculum currently used by SMA Negeri 1 Bp.Mandoge is Curriculum 2013, which regulates the process of sorting majors for class X (ten) students based on average junior high school report cards, junior high school national exam scores, and MTK, IPA, and social studies test scores. One method that can be used to solve data mining classification problems is the C4.5 Algorithm method. Algorithm C4.5 is used to construct a decision tree that divides a large data set into smaller record sets by applying a series of decision rules to classify the data. In this study, student majors were classified based on MTK, IPA, and Social Sciences academic test scores, average junior high school report cards for MTK, Science, and Social Studies subjects, SMP National Examination scores for MTK and Science subjects, and student interests. Based on the results of the research, the results of the classification of student majors that have been tested correspond to an accuracy rate of 89.74%. 5 is used to form decision trees that divide large data sets into smaller record sets by applying a series of decision rules to classify data. In this study, student majors were classified based on MTK, IPA, and Social Sciences academic test scores, average junior high school report cards for MTK, Science, and Social Studies subjects, SMP National Examination scores for MTK and Science subjects, and student interests. Based on the results of the research, the results of the classification of student majors that have been tested correspond to an accuracy rate of 89.74%. 5 is used to form decision trees that divide large data sets into smaller record sets by applying a series of decision rules to classify data. In this study, student majors were classified based on MTK, IPA, and Social Sciences academic test scores, average junior high school report cards for MTK, Science, and Social Studies subjects, SMP National Examination scores for MTK and Science subjects, and student interests. Based on the results of the research, the results of the classification of student majors that have been tested correspond to an accuracy rate of 89.74%. and Social Studies, the average junior high school report cards for MTK, Natural Sciences, and Social Sciences subjects, the SMP National Examination scores for MTK and Natural Sciences subjects, and student interest. Based on the results of the research, the results of the classification of student majors that have been tested correspond to an accuracy rate of 89.74%. and Social Studies, the average junior high school report cards for MTK, Natural Sciences, and Social Sciences subjects, the SMP National Examination scores for MTK and Natural Sciences subjects, and student interest. Based on the results of the research, the results of the classification of student majors that have been tested correspond to an accuracy rate of 89.74%.

1. Introduction

Majoring is the process of placing or distributing in the selection of teaching programs to students. High School (SMA) is one of the educational institutions that has begun to introduce majors and divide them into several choices of majors. Majoring is very important to group students according to abilities (values), talents and interests that are relatively the same so that the lessons given to students are more focused and directed. In this major system, students are given the opportunity to choose a major, be it the "IPA" or "IPS" major, before later classifying the major's decision according to each student's grades and interests, so that later the student's choice and the final decision on the student's major can be different because according to the ability (value) of the student.

ISSN 2502-3357 (Online)

SMA Negeri 1 Bp.Mandoge is one of the leading schools that implements a majoring process for its students, to be divided into 2 majors, namely "IPA" and "IPS". The curriculum currently used by SMA Negeri 1 Bp.Mandoge is the 2013 Curriculum, which regulates the majors process using report cards and SMP National Examination scores, as well as majors test scores. Majoring begins when students are in class X (ten). The majors test was taken by all students of class X which would then be sorted along with each student's junior high school report card and National Examination scores.

In the majors process, students are given the opportunity to choose majors, be it Science or Social Sciences majors, before later classifying their major decisions according to each student's grades and interests, so that later the student's choice and the final outcome of the student's major decision can be different because it adjusts abilities (grades) the student.

Previous research related to data mining of student majors has been carried out by Obbie Kristianto, namely the student majors of SMAN 6 Semarang with the application of the ID3 algorithm. The application was made with the Java programming language and succeeded in determining student majors according to needs [1]. Other research was also carried out by Eka Budi Rahayu, with the research title "C4.5 Algorithm for the Majors of SMA Negeri 3 Pati Students" in which the study used the RapidMiner software as a tool for modeling to produce rules that would be used for classification of majors. Applications for classifying student majors that are made can classify students for majors into IPA and Social Sciences classes [2].

Based on this brief explanation, further research will create an application designed to be able to classify student majors with a case study at SMA Negeri 1 Bp.Mandoge. Majoring applications will be created with the C4.5 algorithm which will be processed automatically in the application without using certain machine learning software to create models and classification rules.

2. Method

3.1. Data Mining

Data Mining is a term used to describe the discovery of knowledge in databases. Data mining is a process that uses statistical, mathematical, artificial intelligence, and machine learning techniques to extract and identify useful information and assembled knowledge from large databases.

Important things related to data mining are:

1. Data mining is an automated process of existing data.
2. The data to be processed is very large data.
3. The goal of data mining is to get relationships or pattern Which may provide a useful indication

classification. Classification is the process of finding models or functions that explain or differentiate data concepts or classes, with the aim of being able to estimate the class of an object whose label is unknown. The model itself can be an “if-then” rule, a decision tree, a mathematical formula or a neural network.

The classification process is usually divided into two phases: learning and testing. In the learning phase, some data whose data class is known is fed to form an approximate model. Then in the test phase the model that has been formed is tested with some other data to determine the accuracy of the model. If the accuracy is sufficient, this model can be used to predict unknown data classes.

3.2. C4.5 Algorithm

Algorithm C4.5 is one of the algorithms used to form a decision tree. The decision tree method turns very large facts into decision trees that represent rules. Rules can be easily understood with natural language. And they can also be expressed in the form of database languages such as Structured Query Language to find records in certain categories.

Decision Tree Algorithm C4.5 or Classification version 4.5 is the development of the ID3 algorithm. Because of this development, the C4.5 algorithm has the same basic working principle as the ID3 algorithm.

In general, the C4.5 algorithm process for building a decision tree is as follows.

2. Select attribute as root
3. Create a branch for each value
4. Split cases in a branch
5. Repeat the process for each branch until all cases on the branch have the same class [4].

In particular, the C4.5 Decision Tree algorithm uses a modified split criterion called Gain Ratio in the attribute split selection process. Split attribute is the main process in forming a decision tree (Decision Tree) in C4.5 [5].

The stages of the C4.5 algorithm are as follows.

a. Calculating the Entropy value,
b. Calculating the Gain Ratio value for each attribute,
c. The attribute that has the highest Gain Ratio is selected to be the root (root) and the attribute that has a lower Gain Ratio value than the root (root) is selected to be branches (branches).
d. Recalculating the Gain Ratio value for each attribute by not including the attribute that was selected to be the root in the previous stage,
e. Attributes that have the highest Gain Ratio are selected to be branches (branches),
f. Repeat steps 4 and 5 until Gain = 0 is generated for all the remaining attributes.

ISSN 2502-3357 (Online)

To calculate the Entropy value can be calculated by the equation:

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

Where:
S = set of cases
A =features
n = number of partitions S
pi =the proportion of S1 to S

Meanwhile the value of information gain (Gain) can be calculated using the equation:

$Gain(S,A)=Entropy(S)-\sum^{n}$

Where:

6. S = set of cases
7. A = attribute
8. n = number of partitions attribute A
9. $|Si|$ = number of cases on the i-th partition

|S| = number of cases in S

Furthermore, the value of Split Info can be calculated by the equation:

Then the Gain Ratio value that determines an attribute can be used as the root or branch of a decision tree can be calculated by the equation:

GainRatio(S,A) = (S,A)

Where:
S = set of cases
A = attribute
Gain(S,A)=gain info on attribute A
SplitInfo(S,A)=split info on queue A

2.3. Decision Tree

Decision tree is a classification algorithm that is often used and has a simple structure and is easy to interpret [6].

The tree that is formed resembles an upside down tree, where the roots are at the top and the leaves are at the bottom. The decision tree is a classification model that is shaped like a tree, where the decision tree is easy to understand even by users who are not experts and is more efficient in inducing data. Decision trees are well used for classification or prediction [7]. The decision tree graph display can be seen in Figure 1.

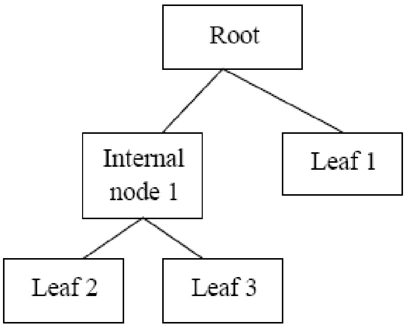


Figure 1. Basic Concepts of a Decision Tree Graph

The process in the decision tree is to change the shape of the data (table) into a tree, change the tree model into a rule, and simplify the rule [8].

The concept of a decision tree in outline can be seen in Figure 2.

Figure 2. The Concept of a Decision Tree

The main benefits of using trees

Decision making is the ability to make complex decision-making processes simpler so that decision makers will better interpret solutions to problems

3.4. Confusion Matrix

ISSN 2502-3357 (Online)

For problems in classification, measurements commonly used are precision, recall and accuracy. This value can be calculated with the confusion matrix. The Confusion Matrix is a table consisting of the number of rows of test data that are predicted to be true (positive) and incorrect (negative) by the classification model [9]. The confusion matrix table can be seen in Table 1.

Table 1. Confusion Matrix

	classified Negative	*classified Positive*
actual Negative	a	b
actual Positive	c	d

Entry “a” contains the same number of negative initial results as the negative results of the classification. Entry "b" contains the amount of negative initial result data that turns into a positive result in the classification results. Entry "c" contains the amount of positive data from the initial results that turn into negative classification results. Entry "d" contains the number of positive data from the same initial result as a positive result in the classification results. The total number of fields "b" and "c" is the difference in the number of differences in the results produced after the classification process. From the results of these differences, the accuracy value can be calculated.

1. Precision

Precision is the part of the data that is taken according to the information needed. The Precision formula is:

$P = ( ) x$ 100%

2. Recall

Recall is data retrieval that has been successfully carried out on parts of the data that are relevant to the query. The Recall formula is:

$R= ( ) x 100\%$

3. accuracy

Accuracy is the percentage of the total test data that is correctly identified. Accuracy formula is:

3. Results and Discussion

This study uses a research methodology that includes literature study, then observation and data collection is carried out, then needs analysis is carried out, then design is carried out, and finally system testing is carried out.

3.1 Research Data

The research data used is data from class X students of SMA Negeri 1 Bp. Mandoge Academic Year 2015/2016. The data was obtained from the Deputy Principal (WaKa) of SMA Negeri 1 Curriculum, Bp. Mandoge. The data obtained is then carried out by the process of data analysis so that data is obtained in the form of case data tables that are ready for data mining. Furthermore, data that is numerical in nature is transformed into categorical with a range of value classifications in Table 2.

Table 2. Value Classification

Mark	Value Classification
86 - 100	A
71-85	B
56 - 70	C
41 - 55	D
≤ 40	E

The data amounted to 392 data. The data will then be divided into 2 groups, namely testing data and training data. The training data functions for the modeling or learning process of the C4.5 algorithm decision tree method, while the testing data is used to test the model that has been formed.

4.2. Research variable

The research variables that will be used as data attributes for the classification data mining process are MTK test scores, Science test scores, IPS test scores, MTK average report cards, average Science report cards, average IPS report cards, MTK UN scores , UN Science scores, student interests, and class decisions. The decision class is the research target variable which contains 2 class values, namely "IPA" and "IPS".

ISSN 2502-3357 (Online)

4.3. Software Flow Chart

The method proposed for the application of student majors is a data classification method using the C4.5 decision tree algorithm. The flowchart of the application is shown in Figure 3

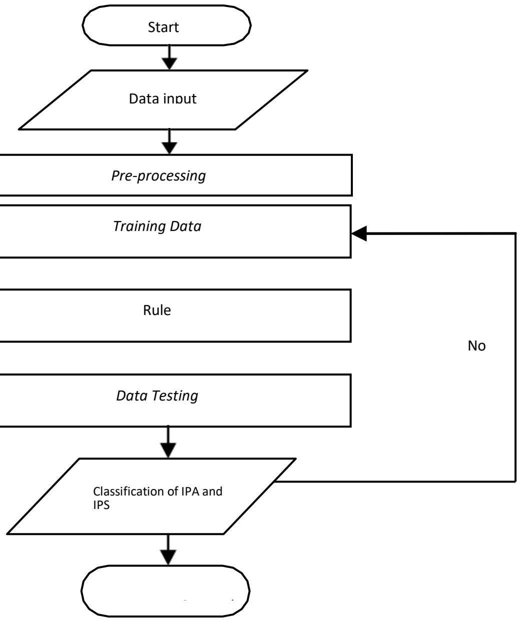


Figure 3. Software Flowchart

The following is an explanation of the flow chart

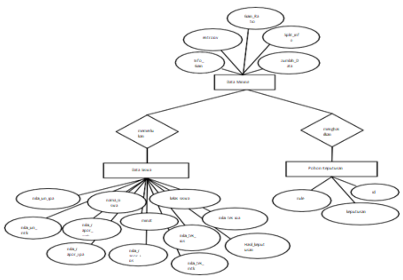
in Figure 3:

1. The process starts.
2. Input student data to be processed.
3. Data pre-processing is done by transforming attribute data that is numerical to categorical, and dividing the data to be processed into 2 groups, namely, training data and testing data.
4. Data training is carried out as a data mining learning model to get decision trees and rules or rules are formed.
5. The rules are successfully formed from the perfect training data modeling process.
6. Data testing is carried out by testing the process of majoring students with the rules formed by data testing.
7. Classification of IPA and IPS is the expected result. If the process is not successful, then the Data Training process is repeated. If successful, then the process is complete.
8. The process is complete.

4.4 Entity Relationship Diagram (ERD)

The logical structure of the database can be described in a graph using an Entity Relationship Diagram (ERD). ERD is a relationship between entities used in the system to describe relationships between entities or data structures and relationships between files.

The application design requires a relationship between Data Mining entities, Student Data, and Decision Trees. The relationship needed is that the Data Mining process will require Student Data which will then produce a Decision Tree. In the Decision Tree entity there is a key attribute for the classification process, namely the rule attribute. The application ERD design can be seen in Figure 4.



ISSN 2502-3357 (Online)

Figure 4. Entity Relationship Diagram

4.5 Data Flow Diagrams

System development uses Data Flow Diagrams (DFD) as a medium to explain all data flows and the processes contained in the system.

a. Context Diagram

Context Diagram is the highest level in data flow diagrams and contains only one process, showing the system as a whole. The application context diagram can be seen in Figure 5.

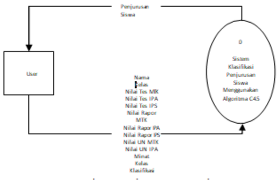


Figure 5. Context Diagram

Data flow originates from input by the user into the system, namely, attributes and attribute values. The required attributes are name, class, MTK test score, Science test score, IPS test score, average MTK report card, average Science report card, average IPS report card, MTK UN score, Science UN score, interest students, and decision outcomes. Then enter the system process and produce student majors output.

b. DFD Level 1

The process of context diagrams or level 0 DFD will be further broken down into level 1 DFDs, shown in Figure 6.

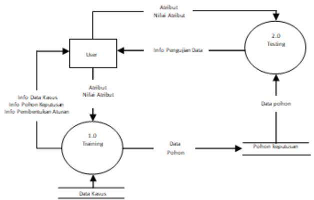


Figure 6. DFD Level 1

The data flow originates from input attributes and required attribute values into process 1.0 and then data training is carried out and produces output in the form of case data info, decision tree info, rule formation info. The input of the same attributes and attribute values is also required for the 2.0 process, to then carry out the data testing process using the rules that have been formed from the decision tree. The output is data testing info.

DFD Level 2

The DFD level 1 process is broken down again and divided into two for the Training and Testing processes. For DFD Level 2 the training process can be seen in Figure 7

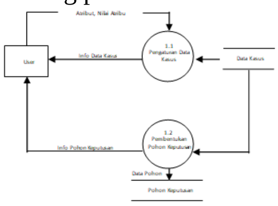


Figure 7. DFD Level 2 Training Process

Process 1.1 is the process of setting case data that will be processed in data mining to look for classification patterns according to the C4.5 algorithm. Process 1.2 shows the process of forming a decision tree which will later become the final output of the process in the form of a decision tree and rules/rules. Furthermore, for DFD Level 2 the Testing Process can be seen in Figure 8.

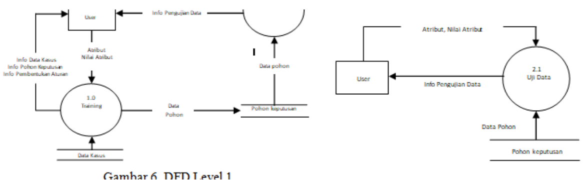


Figure 8. DFD Level 2 Testing Process

http://doi.org/10.XXXXX/JoCoSiR.v1iss1.pp18-28

ISSN 2502-3357 (Online)

Furthermore, in Figure 8, process flow 2.1 is explained, namely data testing. The inputs are attributes and data attribute values and decision tree rules which will generate data testing info in the form of the results of the majors classification process.

4.4 Interface Design

Interface design aims to make it easier to design the appearance of the application to be made to determine the stages of making an application such as menu layouts, buttons, color display, and the required forms look better and are structured according to user needs in using the application system. The menu structure designed for the majoring application can be seen in Figure 9

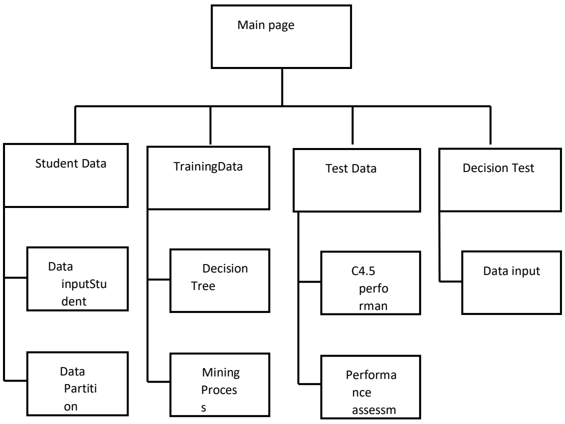


Figure 9. Menu Structure

Data Testing Info

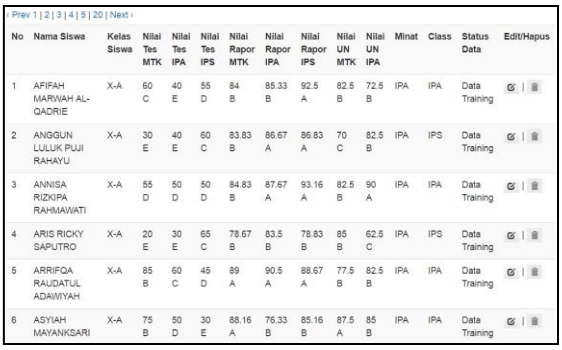
10. Results and Discussion

Application testing aims to determine the functioning of the application that has been made as expected. Application testing is done on a Web Browser. Tests are carried out for each main menu in the application. The main display of the application can be seen in Figure 10.

Figure 10. Application Main Display

a. Student Data Menu

The student data menu is a menu for entering and managing student data that is ready to be classified as a major. The student data used is that of SMA Negeri 1 Bp.Mandoge class X year 2015/2016. After the data is entered, student data can be displayed, as shown in Figure 11.



‹ Prev 1 | 2 | 3 | 4 | 5 | 20 | Next ›

No	Nama Siswa	Kelas Siswa	Nilai Tes MTK	Nilai Tes IPA	Nilai Tes IPS	Nilai Rapor MTK	Nilai Rapor IPA	Nilai Rapor IPS	Nilai UN MTK	Nilai UN IPA	Minat	Class	Status Data	Edit/Hapus
1	AFIFAH MARWAH AL-QADRIE	X-A	60 C	40 E	55 D	84 B	85.33 B	92.5 A	82.5 B	72.5 B	IPA	IPA	Data Training	|
2	ANGGUN LULUK PUJI RAHAYU	X-A	30 E	40 E	60 C	83.83 B	86.67 A	86.83 A	70 C	82.5 B	IPA	IPS	Data Training	|
3	ANNISA RIZKIPA RAHMAWATI	X-A	55 D	50 D	50 D	84.83 B	87.67 A	93.16 A	82.5 B	90 A	IPA	IPA	Data Training	|
4	ARIS RICKY SAPUTRO	X-A	20 E	30 E	65 C	78.67 B	83.5 B	78.83 B	85 B	62.5 C	IPA	IPS	Data Training	|
5	ARRIFQA RAUDATUL ADAWIYAH	X-A	85 B	60 C	45 D	89 A	90.5 A	88.67 A	77.5 B	82.5 B	IPA	IPA	Data Training	|
6	ASYIAH MAYANKSARI	X-A	75 B	50 D	30 E	88.16 A	76.33 B	85.16 B	87.5 A	85 B	IPA	IPA	Data Training	|

Figure 11. Display of Student Data

ISSN 2502-3357 (Online)

Figure 13. Display of Root Node Calculations. Furthermore, all data that has been entered is divided into 2 groups of data, namely training data and testing data. Data division is done with a ratio of 90:10. Data sharing is done in the Data Partition sub menu as shown in Figure 12.

Set Data Training (Semua Data) 90 % Ascending ▾ Proses

Status Data (392 Data)	Data Training (353 Data)	Data Testing (39 Data)
IPS (82 Data)	74	8
IPA (310 Data)	279	31

Figure 12. Data Partition Display

b. Mining Menu C4.5

Menu mining C4.5 is one of the main processes of this major application. In this menu, the mining process of all training data is carried out according to the C4.5 algorithm decision tree method.
The calculation process is done automatically in the application. The calculation includes the entropy value for each attribute, information gain, split info, and gain ratio.
The table display of the calculation results in the mining process to determine the root of the decision tree can be seen in Figure 13.

No	Atribut	Nilai Atribut	Jumlah Kasus Total	Jumlah Kasus IPA	Jumlah Kasus IPS	Entropy	Information Gain	Split Info	Gain Ratio
1	Total	Total	353	279	74	0.7408			0
2	nilai_tes_mtk	A	30	29	1	0.2108	0.4216	2.1096	0.1998
3	nilai_tes_mtk	B	125	123	2	0.1184	0.4216	2.1096	0.1998
4	nilai_tes_mtk	C	101	99	2	0.1403	0.4216	2.1096	0.1998
5	nilai_tes_mtk	D	66	25	41	0.9572	0.4216	2.1096	0.1998
6	nilai_tes_mtk	E	31	3	28	0.4587	0.4216	2.1096	0.1998
7	nilai_tes_ipa	A	0	0	0	0	0.0683	1.6169	0.0422
8	nilai_tes_ipa	B	4	3	1	0.8113	0.0683	1.6169	0.0422
9	nilai_tes_ipa	C	96	89	7	0.3767	0.0683	1.6169	0.0422
10	nilai_tes_ipa	D	156	129	27	0.6647	0.0683	1.6169	0.0422
11	nilai_tes_ipa	E	97	58	39	0.9721	0.0683	1.6169	0.0422
12	nilai_tes_ips	A	0	0	0	0	0.0012	1.5949	0.0008
13	nilai_tes_ips	B	7	6	1	0.5917	0.0012	1.5949	0.0008
14	nilai_tes_ips	C	59	45	14	0.7905	0.0012	1.5949	0.0008
15	nilai_tes_ips	D	131	103	28	0.7486	0.0012	1.5949	0.0008
16	nilai_tes_ips	E	156	125	31	0.7194	0.0012	1.5949	0.0008
17	nilai_rapor_mtk	A	203	186	17	0.4152	0.1008	1.0328	0.0976
18	nilai_rapor_mtk	B	148	92	56	0.9569	0.1008	1.0328	0.0976
19	nilai_rapor_mtk	C	1	1	0	0	0.1008	1.0328	0.0976
20	nilai_rapor_mtk	D	1	0	1	0	0.1008	1.0328	0.0976
21	nilai_rapor_mtk	E	0	0	0	0	0.1008	1.0328	0.0976
22	nilai_rapor_ipa	A	201	177	24	0.5276	0.052	1.0541	0.0493
23	nilai_rapor_ipa	B	149	101	48	0.9067	0.052	1.0541	0.0493
24	nilai_rapor_ipa	C	2	1	1	1	0.052	1.0541	0.0493
25	nilai_rapor_ipa	D	0	0	0	0	0.052	1.0541	0.0493
26	nilai_rapor_ipa	E	1	0	1	0	0.052	1.0541	0.0493
27	nilai_rapor_ips	A	184	156	28	0.6153	0.0211	1.0237	0.0206
28	nilai_rapor_ips	B	168	123	45	0.8384	0.0211	1.0237	0.0206
29	nilai_rapor_ips	C	0	0	0	0	0.0211	1.0237	0.0206
30	nilai_rapor_ips	D	0	0	0	0	0.0211	1.0237	0.0206
31	nilai_rapor_ips	E	1	0	1	0	0.0211	1.0237	0.0206
32	nilai_un_mtk	A	140	126	14	0.469	0.0777	1.4924	0.0521
33	nilai_un_mtk	B	176	136	40	0.7732	0.0777	1.4924	0.0521
34	nilai_un_mtk	C	26	9	17	0.9306	0.0777	1.4924	0.0521
35	nilai_un_mtk	D	4	2	2	1	0.0777	1.4924	0.0521
36	nilai_un_mtk	E	7	6	1	0.5917	0.0777	1.4924	0.0521
37	nilai_un_ipa	A	89	83	6	0.3562	0.0586	1.4129	0.0415
38	nilai_un_ipa	B	207	162	45	0.7554	0.0586	1.4129	0.0415
39	nilai_un_ipa	C	55	34	21	0.9593	0.0586	1.4129	0.0415
40	nilai_un_ipa	D	2	0	2	0	0.0586	1.4129	0.0415
41	nilai_un_ipa	E	0	0	0	0	0.0586	1.4129	0.0415
42	minat	IPA	339	279	60	0.6735	0.094	0.2407	0.3905
43	minat	IPS	14	0	14	0	0.094	0.2407	0.3905

Atribut **minat** memiliki nilai gain terbesar

minat = IPA (IPS = 60, IPA = 279) : ?
minat = IPS (IPS = 14, IPA = 0) : IPS

Figure 13. Display of Root Node Calculations

From the calculation process of determining the roots of the tree decision,Then repeat the calculation process continuously to determine the branches of the decision tree. The data mining process will be complete if all the attributes have entered their respective class or target variable. The process will take quite a long time according to the amount of data being processed. The "Mining Process Complete" notification will appear at the end of the mining process, indicating that the mining process is running perfectly.
After the data mining process is complete, a decision tree and a series of rules will be formed that will be used for the data testing process. The resulting rules are 162 rules.

ISSN 2502-3357 (Online)

The rules displayed are in the form of "IF" and "THEN" rules. In the rules there is also an id rule, namely the order number for the formation of a branch of the decision tree. The rules obtained are quite a lot because the data that is processed is quite a lot and for 1 data there are many attributes and attribute values.

The following shows 3 rules or rules out of a total of 162 major classification rules which can be seen in Figure 13.

```
Rule:
1. if (minat == IPA AND nilai_tes_mtk == A) then IPA (id = 2)
2. if (minat == IPA AND nilai_tes_mtk == B) then IPA (id = 3)
3. if (minat == IPA AND nilai_tes_mtk == C) then IPA (id = 4)
```

Figure 13. Display of 3 Rules

c. Performance menu

The performance menu is a menu for testing rules or rules that have been formed from training data modeling using the C4.5 algorithm decision tree method. Data testing used as many as 39 data from all student data used for the application testing process. Testing data testing can run well marked by all data can be classified into class "IPA" or "IPS".

From the 39 student data, it was found that there were 29 students who were classified in the Science class and 10 students who were classified in the Social Studies class. For each student data, the id rule number is also included. This rule plays a role in classifying data into classes according to the C4.5 algorithm in the data testing process. The results of testing data testing are shown in Figure 14.

No	Nama Siswa	Kelas Siswa	Nilai Tes MTK	Nilai Tes IPA	Nilai Tes IPS	Nilai Rapor MTK	Nilai Rapor IPA	Nilai Rapor IPS	Nilai UN MTK	Nilai UN IPA	Minat	Keputusan Asli	Keputusan C4.5 ID Rule C4.5
1	ARINTO MANGKU SAPUTRO	X-I	B	C	B	A	A	A	A	A	IPA	IPA	IPA 3
2	ASTARI RAFITA GALANTTIARA	X-I	B	E	C	A	A	A	A	B	IPA	IPA	IPA 3
3	AULIIPA NURULJANNAH	X-I	B	D	D	A	B	B	A	C	IPA	IPA	IPA 3
4	AURA AINUN INTANY	X-I	B	C	E	A	B	B	A	B	IPA	IPA	IPA 3
5	CLEMENTINEKE CERELIA AURELIA	X-I	B	E	D	B	B	B	A	B	IPA	IPA	IPA 3
6	CUT VANIA ISMARIZA	X-I	C	C	E	B	B	B	A	B	IPA	IPA	IPA 4
7	DEVINA WANDA ANGELITA	X-I	E	E	E	B	B	A	D	C	IPA	IPA	IPA 200
8	ERIKA PRATIWI	X-I	A	D	D	A	A	A	A	B	IPA	IPA	IPA 2
9	FADILA	X-I	C	E	E	B	A	B	B	B	IPA	IPA	IPA 4
10	FAKHRI RAMADHAN ISHAN	X-I	B	D	D	A	A	A	A	B	IPA	IPA	IPA 3
11	FARDHI LIRA RAMADHAN	X-I	A	C	D	A	A	A	A	B	IPA	IPA	IPA 2
12	HESTI AYUNINGTYAS	X-I	B	D	D	B	B	B	B	C	IPA	IPA	IPA 3
13	HIKMAHWAN MUHAMMAD AKMAL SESAR	X-I	D	C	E	B	B	B	B	C	IPA	IPA	IPS 84
14	HIMA MAHDA NADIA	X-I	D	E	E	B	B	B	B	B	IPA	IPS	IPS 118
15	INGGIT SEPTANIA AURELIA BAKRI	X-I	B	D	E	B	B	A	A	A	IPA	IPA	IPA 3
16	JUWITA PRATIWI	X-I	A	C	D	A	A	A	A	A	IPA	IPA	IPA 2
17	MOHAMMAD MARUF	X-I	D	D	D	B	B	B	B	A	IPA	IPS	IPA 92
18	MUHAMMAD AZHAR ZAEN	X-I	C	D	E	A	B	A	B	B	IPA	IPA	IPA 4
19	MUHAMMAD FARHAN	X-I	C	D	E	A	A	A	A	B	IPA	IPA	IPA 4
20	MUHAMMAD LIO RADIFAN	X-I	A	E	E	A	B	B	A	B	IPA	IPA	IPA 2
21	MUHAMMAD RIFQI HARYADI	X-I	D	E	E	B	B	B	D	B	IPA	IPS	IPS 144
22	NURAINI	X-I	A	C	E	A	A	B	B	A	IPA	IPA	IPA 2
23	NURTANIA AFIFAH	X-I	D	E	E	A	A	A	B	B	IPA	IPS	IPS 56
24	PUTRI KUSUMA WARDANI	X-I	A	C	C	A	A	B	A	B	IPA	IPA	IPA 2
25	QUEENTERA CANTIKA ARASANDA	X-I	D	D	E	B	A	A	C	B	IPA	IPA	IPS 133
26	RADINDA FARIDZIPA FITRIYANTO	X-I	A	C	E	A	A	A	C	B	IPA	IPA	IPA 2
27	REDHA DHIKA MAHFUSZAN	X-I	C	C	D	A	A	A	B	A	IPA	IPA	IPA 4
28	RETNO AYU CHOIR RASYIDAH	X-I	B	C	D	A	A	A	B	A	IPA	IPA	IPA 3
29	RIDHO BAGASKORO	X-I	E	E	E	B	B	B	B	B	IPA	IPS	IPS 193
30	RIFKI CHAIRULLAH	X-I	B	E	D	A	B	B	B	B	IPA	IPA	IPA 3
31	RIZKA RAMADHANTI	X-I	C	C	C	B	B	B	B	B	IPA	IPA	IPA 4
32	SADDAM HUSEIN SURAJ MUHAMMAD	X-I	E	E	E	B	B	A	B	C	IPA	IPS	IPS 193
33	SALSABILA AWANIS ZHARFA	X-I	C	D	B	A	A	A	A	B	IPA	IPA	IPA 4
34	SANDRA SRI PUSPA	X-I	E	E	E	B	B	B	B	B	IPA	IPS	IPS 193
35	SARAH SALSABILA	X-I	D	D	D	B	A	A	C	A	IPA	IPA	IPS 133
36	SEPTIANTO	X-I	C	E	D	A	A	B	B	A	IPA	IPA	IPA 4
37	SYAHWAL DESPRIADI	X-I	A	E	E	A	A	A	A	B	IPA	IPA	IPA 2
38	SYARIF MUHAMMAD FARHAN AL IDRUS	X-I	E	D	E	B	B	B	B	C	IPA	IPS	IPS 193
39	YOGAMA WISNU OKTYANDITO	X-I	C	D	E	A	A	A	B	A	IPA	IPA	IPA 4

Figure 14. Display of Testing Data Testing

d. Decision Maker Menu

The decision-making menu is used to try out1 single data will be classified to one of the science or social studies major classes. Student data is entered according to the requested attribute variables into the data input form, then the data will receive the results of the major classification.

3 data tests have been carried out on the decision maker menu. The results can be seen in Fig 15.

Figure 15. Decision Determinants Test Results

The data that has been entered can be classified properly. In the first data test the results produced were "IPS" with an id rule: 84. In the second test, the results obtained were "IPS", with an id rule: 159. In the third test, the results obtained were "IPA", with an id rule : 2.

ISSN 2502-3357 (Online)

No	Nama Siswa	Kelas Siswa	Nilai Tes MTK	Nilai Tes IPA	Nilai Tes IPS	Nilai Rapor MTK	Nilai Rapor IPA	Nilai Rapor IPS	Nilai UN MTK	Nilai UN IPA	Minat	Keputusan C4.5 ID Rule	Hapus
1	Satiu	X-A	55 D	60 C	75 B	80 B	80 B	80 B	80 B	90 A	IPA	IPS 84	
2	Lian	X-A	35 E	55 D	60 C	80 B	81,5 B	84,33 B	87,5 A	80 B	IPA	IPS 159	
3	Yaya	X-H	90 A	80 B	70 C	60 C	78 B	88 A	87 A	88 A	IPA	IPA 2	

5.2 Evaluation of Results

From the results of application testing, all datasets were successfully classified into "IPA" and "IPS" decision classes. Obtained the number of differences in the results for the initial decision before the mining process and after the mining process C4.5. Next, an evaluation of the results will be carried out.

Evaluation done by developing the results of data mining classification. Assessment of process performance results is measured by the confusion matrix. The confusion matrix assessment table is filled based on the results obtained from data testing. The initial results or the original results of the classification of student majors obtained from the school were compared with the results of the classification using the C4.5 algorithm method.

The total number of data assessed is 39 data. For negative columns, fill in as "IPS" values, while for positive columns, fill in as "IPA" values. The results of the initial decision before the C4.5 process was carried out obtained social studies class decisions as many as students while science class decisions were 31 students. For the results after the C4.5 process, there were 10 students' social studies class decisions and 29 students' science class decisions.

The results that remained the same from the results of the initial decision "IPS" to the results of the classification C4.5 "IPS" totaled 7 data, while the results that changed from the results of the initial decision "IPS" to the results of the classification C4.5 "IPA" amounted to 1 data. The results that remained the same from the results of the initial decision "IPA" to the results of the classification C4.5 "IPA" totaled 28 data, while the results that changed from the results of the initial decision "IPA" to the results of the classification C4.5 "IPS" amounted to 3 data. So that the difference in the results of the dataset before and after the mining process is obtained and the level of classification accuracy can be calculated. The evaluation table for the majors classification application can be seen in Table 3.

Table 3. Rating Table

		Classification Results C4.5	
		IPS	IPA
Preliminary Results	IPS	7	1
	IPA	3	28

From the assessment table, it was found that the difference in classification results amounted to 4 data. From this difference, the accuracy rate can be calculated, which is 89.74%.

6. Conclusion

Based on the results of research and testing, several conclusions can be drawn as follows:

The application of classifying student majors with the C4.5 decision tree algorithm can classify student majors. The results of assessing the performance level of the classification results of student majors using the confusion matrix assessment table yield an accuracy value of 89.74%, a precision value of 96.55%, and a recall value of 90.32%.

Reference

Kristianto, O., (2014). Application of the ID3 Data Mining Classification Algorithm to Determine the Majors of High School Students**6**

Semarang.http://eprints.dinus.ac.id/5398/1
/14005.pdf accessed on 16 March 2016.

Rahayu, EB, (2015). C4.5 Algorithm for Majoring Students of SMA Negeri 3 Pati. http://eprints.dinus.ac.id/15291/2/abst rak_15319.pdf accessed on 16 March 2016.

Turban, E., et al. (2005) Decision Support Systems and intelligent Systems. Yogyakarta: ANDI

Kusrini. & Lutfi.,MT (2009). Algorithm Data mining. Yogyakarta: ANDI

ISSN 2502-3357 (Online)

Quinlan, JR (1986). Induction of Decision Trees. Machine Learning, 81-106
Mantas, CJ, & Abellán, J. (2014). Credal-C4.5: Decision tree based on imprecise probability to classify noisy data. Expert Systems with Applications, 41(10), 4625–4637.
doi:10.1016/j.eswa.2014.01.017
Sammut, GWC (2011). Encyclopedia of Machine Learning. (C. Sammut & GI Webb, Eds.). Boston, MA: Springer US.
Basuki, A., & Syarif, I. (2003). Decision Tree. http://lecturer.eepis-its.edu/~basuki/lecture/DecisionTree.pdf, accessed on 16 March 2016.