IIAI Open Conference Publication Series
IIAI Letters on Informatics and Interdisciplinary Research
Vol.005, LIIR209
DOI: https://doi.org/10.52731/liir.v005.209

Using C4.5 Decision Tree to Determine the Majors of Students in SMAN 4 Banjarmasin to Reduce the Cause of DropOut from School

Amrul Hadiyanoor [*], Siti Cholifah [*],
Husnul Ma'ad Junaidi [*], Irfan Febrian [*]

Abstract

Choosing the right major is important for students. Choosing the wrong major by students can make the learning process difficult and ineffective. Ineffective learning outcomes can lead to decreased grades. The worst case can cause students to drop out of school. In this study, the C4.5 algorithm is used to generate decision trees to determine major recommendations. The data used is data from previous year prospective students such as exam results, interests and talents used in the decision tree. The results of the decision tree are used for recommendations for selecting student majors. Students with the right majors can make the learning process more effective and can get better grades. Good grades can reduce the reasons for dropping out of school and make students more enthusiastic about learning.

Keywords: C4.5, classification, decision tree, major recommendations.

1 Introduction

This Student interests and talents that are not linked to the main priorities of the school are generally not fully supported. But this affects their psyche and personality. The main reason schools do not support students' interests and talents is because there is no information about them.

Ironically, majors in schools are related to students' interests and talents. For example, a student majoring in science because he has more job opportunities, even though he is gifted in writing and does not like mathematics.

Usually, new students realize the mistake in choosing a major after several semesters have passed. These mistakes can have a negative impact on the effectiveness of their learning process. The school itself still uses the oldschool method, namely by looking at grades, rankings and recommendations. Even though there are many elements of the category that the teacher council and the school can agree on to find the best major for prospective students. One of the classification learning methods that can be used to determine student majors is Decision Trees [1] generated by Algorithm C4.5 [2].

2 Method

C4.5 is an algorithm used to form a decision tree [1][2]. The decision tree is a well-known clas-

[*] Indonesian College of Informatics and Computer Management, Banjarmasin, Indonesia

sification and prediction method, this method is useful for exploring data and finding hidden relationships between a number of candidate input variables and target variables. Besides the C4.5 algorithm, there are many algorithms that can be used in the formation of decision trees, including ID3 and CART. The C4.5 algorithm is the successor of the ID3 algorithm and can produce a better decision tree.

The C4.5 algorithm process begins by selecting the highest gain attribute as the root of the tree, then creating a branch for each value, then dividing the cases until all cases in the other branches have the same class [3].

The first thing that must be done to form a decision tree is to determine the attribute/variable that will become the root of the decision tree. The way to determine which variable is the root is by using entropy, gain, split info, and gain ratio [4].

2.1 Entropy

Entropy is a measure of the degree of diversity of information and data collections. The greater the entropy value, the greater the level of diversity of a data set [5]. The Entropy formula can be written as in Eq. (1).

$$Entropy\ (S) = \sum -xi * log_2 xi \quad (1)$$

With:

i=1 and i<=n

Where:

S : case collection
n : the number of classification classes
xi : the number of sample for class i

2.2 Gain

The effective measure of a variable in classifying data is called Gain. The gain of a variable is the difference between the total entropy value and the entropy of that variable [6]. The difference can be written as in Eq. (2).

$$Gain(S,A) = Entropy(S) - \sum |Sj|/|S| * Entropy(Sj) \quad (2)$$

With:

j=1 and j<=n

Where:

A : Variable
|Sj| : Number of samples for a variable
|S| : Number of samples for the entire data

The gain value will be used to determine which variable is the node of the decision tree. The variable that has the highest gain will become the root of the decision tree [7].

2.3 Split Information

Split Information is a formula that states entropy information [8]. The information sharing formula can be written as in Eq. (3).

$$SplitInfo\ (S,A) = -\sum |Sj|/|S| * log_2 |Sj|/|S| \quad (3)$$

With:

j=1 and j<=n

2.4 Gain Ratio

Gain Ratio is used to reduce attribute bias which has many branches. The gain ratio has a large value if the data is spread out and of small value if all the data goes into one branch [9]. The gain ratio can be written as Eq. (4).

$$GainRatio\ (S,A) = Gain(S,A)/SplitInfo(S,A) \quad (4)$$

2.5 Algorithm of C4.5

The following is the C4.5 algorithm according to [10][11][12]:

1. Attribute Selection:

   Calculate the entropy of the target attribute (the attribute we want to classify). For each attribute (except the target attribute):

   - Calculate the information gain ratio using the attribute's values to split the data.
   - Calculate the intrinsic value of the attribute.

   Select the attribute with the highest information gain ratio.

2. Create a Decision Node by creating a decision node with the selected attribute as the node's test condition.

3. Split Data and Recursion:

   For each value of the selected attribute:

   - Create a branch from the decision node for that attribute value.
   - Recursively call the C4.5 algorithm on the subset of data that matches the selected attribute value.
   - Attach the resulting subtree to the corresponding branch of the decision node.

4. Stopping Criteria:

   - If all instances in the subset belong to the same class, create a leaf node with that class label.

- If there are no more attributes left to split, create a leaf node with the majority class label.

3 Research Results and Discussions

This study uses data from prospective SMA4 Banjarmasin students. In Table 1, a sample of prospective student data is shown.

Table 1: Sample Data of Prospective Students

Name	Interest	Talent	UN Math	UN Indonesian	UN English	UN Sport	Choice
Hero	IPA	Art	90	90	77	80	1
Devy	IPS	Computer	80	80	90	77	1
Deni	IPA	Art	90	91	90	88	2
Rindu	IPS	Computer	88	78	90	88	1
Lina	IPA	Sport	90	87	78	90	2
Erik	IPS	Sport	90	88	88	80	1
Ivy	IPA	Art	90	87	79	79	1
Rino	IPA	Computer	88	87	88	83	2
Kinan	IPS	Computer	88	88	96	85	2
Santo	IPS	Art	87	90	84	89	1

3.1 Data Conversion

In Table 2. data conversion is shown to classify the value of prospective students.

Table 2: Grade Classification

Grades	Classification
86-100	A
76-85	B
66-75	C
< 66	D

In Table 3. data conversion is shown to classify the major interests of prospective students.

Table 3: Major Interest Classification

Interest	Classsification
IPA	A
IPS	B

In Table 4. data conversion is shown to classify the talents of prospective students.

Table 4: Talent Classification

Talent	Classification
Computer	A
Art	B
Sport	C

The converted data will be used as an attribute to train Algorithm C4.5. The converted data is shown in Table 5.

Table 5: Converted Data

Code	K1 (Interest)	K2 (Talent)	K3 (UN Math)	K4 (UN Ind)	K5 (UN Eng)	K6 (UN Sports)	K7 (Choice)
001	A	B	A	A	C	B	1
002	B	A	B	B	A	C	1
003	A	B	A	A	A	A	2
004	B	B	B	C	A	A	1
005	A	C	A	B	C	A	2
006	B	C	A	A	A	B	1
007	A	B	A	A	B	B	1
008	A	A	B	A	A	A	2
009	B	A	A	A	A	B	2
010	B	B	A	A	A	A	1

3.2 Classifications

Table 6 shows the entropy and gain calculation results as well as the classification results.

Table 6: Classification Results

No	Total			Result
	Entropy	Gain	Total	
001	0.15	0.1125	0.625	IPA
002	0.15	0.075	0.775	IPA
003	0.1125	0.075	0.5875	IPA
004	0.15	0.1125	0.8	IPA
005	0.075	0.0375	0.4125	IPS
006	0.15	0.1125	0.6	IPA
007	0.15	0.075	0.5875	IPA
008	0.1125	0.075	0.4242	IPS
009	0.15	0.075	0.3	IPS
010	0.075	0.1125	0.214	IPS

A sample of prospective student data as shown in table 1 is converted using several classifications such as grade classification shown in table 2, major interest classification shown in table 3 and talent classification shown in table 4. The results of the data conversion are shown in table 5.

From the conversion data in table 5, entropy calculations were carried out using equation 1 and gain calculations using equation 2. The classification results were obtained in table 6.

3.3 Results

In the testing, all prospective new student data at SMAN4 was used, all students' major choices were compared with the results of calculations using the C4.5 decision tree algorithm [13].

The following are the results of the comparison between student choices and the C4.5 calculation results which are shown in table 7.

Table 7: Testing Results

Data No	Classification Results	Actual Choices
1	A	A
2	B	A
3	A	A
4	A	A
5	B	B
6	B	B
…		
181	A	A
182	A	A
183	B	A

Where:

A=IPA (Natural Science)

B=IPS (Social Science)

The data used was 183 prospective student data, of all the data there were 131 decision tree results that matched the student's choice, the remaining 52 results did not match.

It can be seen in Figure 1 the comparison between the Decision Tree results and prospective student choices.

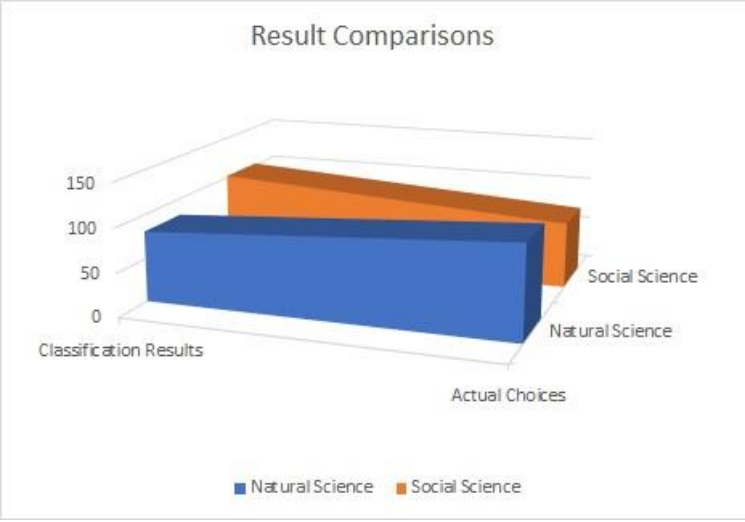

Figure 1: Result Comparisons

With this data, 52 prospective students were given recommendations to choose a more appropriate major so that the learning process could be more effective.

4 Conclusions

With the results of this study using the C4.5 Algorithm, SMAN4 Banjarmasin can provide better major recommendations and can maintain effective communication with parents of students, as well as assist in managing activities related to these students and can manage resources more effectively and provide better learning experience for students.

Research can be continued with the same data, after new students graduate to see their grades and how many students will drop out.

References

[1] J.R. Quinlan, “Induction of Decision Trees MachineLearning”, pp. 81-106, 1986.

[2] J.R. Quinlan, “C4.5: Programs for Machine Learning”, San Mateo, CA:Morgan Kaufmann Publishers, 1993.

[3] Y R Pratama, S Atin and I Afrianto, "Predicting Student Interests Against Laptop Specifications Through Application of Data Mining Using C4.5 Algorithms", IOP Conf. Series: Materials Science and Engineering, vol. 662, 2019.

[4] Mittal, D. Khanduja, and P. Tewari, “An Insight into ‘Decision Tree Analysis”, Int. J. Peer Rev. J. Ref. J. Index. J. UGC Approv. J. Impact Factor, vol. 3, no. 12, pp. 111–115, 2017, [Online]. Available: www.wwjmrd.com

[5] Al-Barrak and M. Al-Razgan, “Predicting Students Final GPA Using Decision Trees: A Case Study”, Int. J. Inf. Educ. Technol., vol. 6, no. 7, pp. 528–533, 2016, doi: 10.7763/ijiet.2016.v6.745.

[6] Hamoud, A. S. Hashim, and W. A. Awadh, “Predicting Student Performance in Higher Education Institutions Using Decision Tree Analysis”, Int. J. Interact. Multimed. Artif. Intell., vol. 5, no. 2, p. 26, 2018, doi: 10.9781/ijimai.2018.02.004.

[7] Patel and P. Prajapati, “Study and Analysis of Decision Tree Based Classification Algorithms”, Int. J. Comput. Sci. Eng., vol. 6, no.10, pp. 74–78, 2018, doi: 10.26438/ijcse/v6i10.7478.

[8] N. Tryfona and C. S. Jensen, “Conceptual data modeling for spatiotemporal applications,” Geoinformatica, vol. 3, no. 3, pp. 245– 268, 1999.

[9] A. Neeraj, G. Bhargava and M. Manish, "Decision Tree Analysis on J48 Algorithm for Data Mining", International Journal of Advanced Research in Computer Science, vol. 3, no. 6, pp. 22-45, 2013.

[10] Sinam and A. Lawan, “An Improved C4.5 Model Classification Algorithm Based on Taylor’s Series”, Jordanian Journal of Computers and Information Technology (JJCIT), Vol. 05, No. 01, April 2019.

[11] R. Rahim et al., "C4.5 Classification Data Mining for Inventory Control", International Journal of Engineering & Technology, vol. 7 no. 2.3 , 2018), Special Issue 3, pp. 68-72.

[12] M. A. Riantika, R. Arifudin, "Optimization of the C4.5 Algorithm by Using a Genetic Algorithm for the Diagnosis of Life Expectancy for Hepatitis Patients", Journal of Advances in Information Systems and Technology, vol. 3, no. 1, 2021, pp. 25-32

[13] Fen Yang, "Decision Tree Algorithm Based University Graduate Employment Trend Prediction", Informatica , An International Journal of Computing and Informatics, vol. 43, no. 4, 2019