Clean Energy and Smart Technology
Volume 01, Number 02, April 2023
DOI : https://doi.org/10.58641/cest.v1i2.39

e-ISSN : 2964-2647

DATA MINING TO PREDICTE THE TIME OF KHATAMAN TAHFID STUDENTS TAHFID PPTQ AL-ASY'ARIYYAH USING C4.5 ALGORITHM

Muhamad Maulana Ikhsanul Khafidli[1)], Muhamad Fuat Asnawi[2)*], Adi Suwondo[3)], Sukowiyono[4)], Dimas Prasetyo Utomo[5)]
[1)2)3)4)5)] Universitas Sains Al-Qur'an, Indonesia
[1)]maulanaikhsanul@gmail.com, [2)]fuatasnawi@unsiq.ac.id, [3)]alethadhelvyaaa@gmail.com,
[4)]suko34497@gmail.com, [5)]vikiran.dpu@gmail.com

*fuatasnawi@unsiq.ac.id

Submitted : 2 Februari 2023 | **Accepted** : 2 Maret 2023 | **Published** : 30 April 2023

Abstract: PPTQ Al-Asy'ariyyah is an Al-Qur'an-based Islamic boarding school where most students memorize the Al-Qur'an. This study uses data mining to calculate how many students are late in completing their memorization and the factors that influence it by using the C4.5 algorithm. Algorithm C4.5 has a decision that can give the desired result. The results of this study found several factors that influence the punctuality of khatam and several factors. Of the 250 students located, 96 could complete their memorization on time, and 154 were off target or on time. Based on what was obtained from the study results, the timeliness of completing the completion is sufficient. Students who have little activity outside the boarding school become students who have a high level of punctuality.

Keywords: PPTQ Al Asy'ariyyah, Data Mining, Algorithm C4.5, Decision Tree

1. INTRODUCTION

Islamic boarding schools are one of the many educational institutions that have begun to develop following technological advances. Technology in a boarding school is very helpful in solving various problems related to computerization. PPTQ Al-Asy'ariyyah is an Islamic educational institution in Wonosobo, Central Java, Indonesia which focuses on teaching the Al-Quran and Tahfidz. Khataman santri Tahfid PPTQ Al-Asy'ariyyah is a final activity carried out by the santri after completing the process of studying the Koran for several years. The Khataman is an important moment for the students and also for the PPTQ Al-Asy'ariyyah institution. In this context, data mining can be used to predict the timeliness of completing Tahfid PPTQ Al-Asy'ariyyah students. One algorithm that can be used is C4.5. This algorithm is one of the popular decision tree algorithms and is used for data classification.

Following are some previous studies that are related to the use of the algorithm that we use, such as research conducted by Fauzan and Yustina (2021) which uses the C5.0 algorithm to predict high school student graduation based on variables such as test scores, attendance, and the level of student activity in class. The results of the study show that the C5.0 algorithm has high prediction accuracy and can assist schools in taking appropriate actions to increase student graduation rates. In addition, research conducted by Nandita and Putri (2020) uses the Decision Tree algorithm to predict student learning progress based on variables such as the level of student activity in class, the number of hours studied, and test results. The results of the study show that the Decision Tree algorithm has fairly good predictive accuracy and can assist teachers in giving special attention to students who need help.

Clean Energy and Smart Technology
Volume 01, Number 02, April 2023
DOI : https://doi.org/10.58641/cest.v1i2.39
e-ISSN : 2964-2647

In its use to predict the timeliness of completing Tahfid PPTQ Al-Asy'ariyyah students, the C4.5 algorithm will use historical data which includes various variables of timeliness, inaccuracy, status, origin, experience, and organization. By utilizing this historical data, the C4.5 algorithm will build a predictive model that can be used to predict the timeliness of completing Tahfid PPTQ Al-Asy'ariyyah students. Thus, the use of data mining and the C4.5 algorithm can help PPTQ Al-Asy'ariyyah to monitor the learning progress of students and predict the timeliness of completing Tahfid students, so that appropriate actions can be taken to help students who have difficulty learning and increase the success rate of completing Students.

2. METHOD

According to Sugiyono, the object of research is an attribute, characteristic or value of a person. These activities have certain variations determined by the researcher to be studied and concluded. The objects in this study were student-level students who had memorized the Al-Qur'an in the past five years, which would be used as data. The data is used as a reference in data mining calculations using the C4.5 algorithm, which will later produce results as material for the management's evaluation. Data collection is a technique in the research process that aims to obtain information related to the case at hand; in this final project, we must conduct field and literature studies to get the data needed.

In field studies, the authors used two data collection methods, namely interviews and observation; in this study, the authors conducted interviews with male and female administrators and male and female tahfidz leaders to know the start memorization method. The author's observation made observations to collect data on all tahfidz students at Pptq Al-Asy'ariyyah at the same level as students in the past five years. And then, the data is used as reference data. A literature study is carried out by looking for references from books, journals, and articles related to the problems that have been planned to be researched to find solutions. The research process is essential in the research process because it refers to the stages carried out to complete the research; the research flow includes Problem Identification, Data Analysis, Data Processing, Results, and Reporting. The tools used in this study are software for processing data and hardware tools in the operation of the software that supports this research, while the details are:

1. Software
   a. Microsoft Windows 8: Used as an operating system whose function is as a liaison between the author and the hardware used.
   b. RapidMiner: Has a function to assist writers in processing, analyzing and interpreting data
   c. Microsoft Office Word: Used in writing research reports
   d. Microsoft Office Excel: Used to process numbers with a spreadsheet.
2. Hardware
   a. Acer Laptops
   b. Flashdisk
   c. Smartphones

The conceptual framework in a study is a conclusion from the concepts arranged systematically so that the research objectives can be achieved by what is expected. Data sources are data obtained from research results that will be processed later; the data collection process is divided into primary and secondary data. Population is a generalized area of objects or subjects with specific quantities and characteristics determined by the researcher to be studied and concluded. In this study, the population was all tahfidz students at the PPTQ Al Asy'ariyyah student level. Sample, in this study, using a non-probability sampling method. According to Retnawati, non-probability sampling is a sampling technique that only provides equal opportunities or opportunities for each member of the population selected to be the sample. The sample used in this study was tahfidz students at the last five-year level. The data processing technique used is using the C4.5 algorithm with the Decision Tree method,

3. RESULT

Evaluation and Validation of Algorithm Results in C4.5. The initial step in the C4.5 algorithm evaluation process is to prepare the data; in this research, the data to be processed is the tahfidz students at the PPTQ Al-Asy'ariyyah student level for the past five years. Dataset Manual Calculation. To make a decision tree from the C4.5 algorithm based on the data above, we need to look for the Entropy value and the Gain value of each attribute.

The formula for finding the entropy value is as follows

Clean Energy and Smart Technology
Volume 01, Number 02, April 2023
DOI : https://doi.org/10.58641/cest.v1i2.39
e-ISSN : 2964-2647

Entropy(S)= ∑_(i=1)^n▒- pi* 〖log〗_(2) pi
Information :
S = Case Set
A = Features
n = Number of S Partitions
pi = the proportion of S to S
While the formula for finding Gain is:

Gain(S,A)=Entropy(S)-∑_(i=1)^n▒|S_i |/S*Entropy(Si)
Information :
S = Case Set
A = Attribute
n = Number of partitions attribute A
|Si| = Number of cases in partition i
|S| = Number of cases in S

1. total entropy
For total entropy, the formula is:
Entropy(S) = ∑ - p(I) x log2p(I)
Total on time = 96, amount not on time = 154, then:
-p(true) x log2p(true) – p(no) x log2p(no)
= -(96/250) x log2(96/250) – (154/250) x log2(154/250)
= 0.960818175
2. Student entropy
The total number of student students = 150, with the number of students being on time = 30, and students not being on time = 120, then:
-p(true) x log2p(true) – p(no) x log2p(no)
= -(30/150) x log2(30/150) – (120/150) x log2(120/150)
= 0.721928095
3. non-student entropy
The total number of non-student students = 100, with the number of non-students on time = 66, and non-students not on time = 34, then:
-p(exact) x log2p(exact) – p(not) x log2p(not)= -(66/100) x log2(66/100) – (34/100) x log2(34/100) = 0, 924818705
4. calculate the status gain
Gain status = Entropy(total) – [p(total|status=student) x Entropy(total|status=student)] – [p(total|status=non-student) x Entropy(total|status=non-student)]
= 0.960818175 – [(150/250) x 0.721928095) - (100/250) x 0.924818705)]
= 0.157733836
Then the status Gain = 0.157733836
5. calculate the entropy of students who take part in organizations on campus
The total number of students participating in organizations on campus = 35, with the number being on time = 8, and not being on time = 27, then: -p(correct) x log2p(correct) – p(not) x log2p(not)= -(8/35) x log2(8/35) – (27/35) x log2(27/35)
= 0.775512658
6. calculate the entropy of students who take part in the organization at the Islamic boarding school
The total number of students participating in the organization at the Islamic boarding school = 79, with the Number being on time = 30, and not being on time = 49, then: -p(correct) x log2p(right) – p(not) x log2p(not) = -(30/79) x log2(30/79) – (49/79) x log2(49/79)
= 0.957863024
7. calculate the entropy of students who take part in organizations at the boarding school and on campus
The total Number of students who participated in the organization at the boarding school and campus = 26, with the Number being on time = 0, and not being on time = 26, then:
-p(exact) x log2p(exact) – p(not) x log2p(not)= -(0/26) x log2(0/26) – (26/26) x log2(26/26) = 0

Clean Energy and Smart Technology
Volume 01, Number 02, April 2023
DOI : https://doi.org/10.58641/cest.v1i2.39
e-ISSN : 2964-2647

8. calculate the entropy of students who do not join the organization

The total number of students who did not join the organization = 110, with the number being on time = 58, and not being on time = 52, then:

-p(true) x log2p(true) – p(no) x log2p(no)

= -(58/110) x log2(58/110) – (52/110) x log2(52/110)

= 0.997852777

9. calculate organizational Gain

Organizational gain = Entropy(total) – [p(total|organization=campus) x Entropy(total|organization=campus)] – [p(total|organization=cottage) x Entropy(total|organization=cottage)] – [p (total|organization=both) x Entropy(total|organization=both)] – [p(total|organization=absent) x Entropy(total|organization=absent)]

= 0.960818175 – [(35/250) x 0.775512658) - (79/250) x 0.957863024)] – [(26/250) x 0) - (110/250) x 0.997852777)]

= 0.110506466

Then organizational gain = 0.110506466

10. Calculate the entropy of students from Java

The total number of students from Java = 172, with the number being on time = 64, and not being on time = 108, then: -p(correct) x log2p(correct) – p(not) x log2p(not) = -(64 /172) x log2(64/172) – (108/171) x log2(108/172) =0.952265625

11. calculate the entropy of students who come from outside Java

The total number of students coming from outside Java = 78, with the number being on time = 32, and not being on time = 46, then:

-p(exact) x log2p(exact) – p(not) x log2p(not)= -(32/78) x log2(32/78) – (46/78) x log2(46/78)

= 0.976634911

12. calculate the original Gain

Gain origin = Entropy(total) – [p(total|origin=Java) x Entropy(total|origin=Java)] – [p(total|origin=outside Java) x Entropy(total|origin=outside Java)]

= 0.960818175 – [(172/250) x 0.952265625) - (78/250) x 0.976634911)]

= 0.000949333

Then the original Gain = 0.000949333

13. calculate the entropy of students who have long experience in Islamic boarding schools

The total number of students who have long experience in Islamic boarding schools = is 56, with the Number being on time = 17 and not being on time = 39, then:

-p(exact) x log2p(exact) – p(not) x log2p(not) = -(17/56) x log2(17/56) – (39/56) x log2(39/56)

= 0.885612871

14. hit

Table 3. results of manual dataset calculations

No			Entropy	Gain
1	Total		0,960818175	
	Status	Student	0,721928095	0,157733836
		Non student	0,924818705	
	Organization	Pesantren	0,957863024	0,110506466
		College	0,775512658	
		Not participate	0,997852777	
		Both	0	

Clean Energy and Smart Technology
Volume 01, Number 02, April 2023
DOI : https://doi.org/10.58641/cest.v1i2.39
e-ISSN : 2964-2647

	From	Java	0,952265625	0,000949333
		Out of Java	0,976634911	
	Experience studying in pesantren	Long time	0,885612871	0,008205874
		A moment	0,966009606	
		Never	0,97561703	

The first step in managing data is to enter the data you want to process into RapidMiner. The data to be entered is training data or testing data. Please make sure the datasheet we wish to enter is correct.

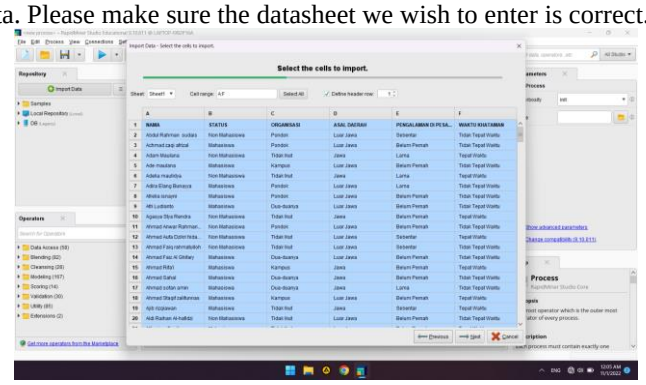

Fig. 1. Entering data into Rapidminer

In this Process, the Decision Tree operator, Apply Model and Performance are entered in the Main Process, then connect the cables. Then click the arrow pointing up. Make sure the wires are connected correctly so no errors occur.

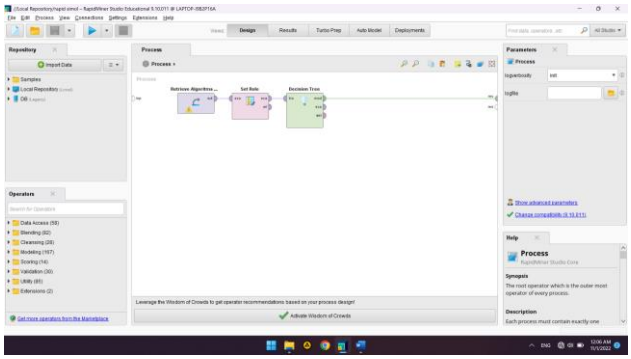

Figure 2. Main Process Decision tree

Determination of Decision Tree Roots. The highest gain is taken to determine the root of the decision tree. From the table above results, the highest gain value is obtained from the status attribute, which is 0.157733836. So the status attribute becomes the root of the decision tree.

Clean Energy and Smart Technology
Volume 01, Number 02, April 2023
DOI : https://doi.org/10.58641/cest.v1i2.39
e-ISSN : 2964-2647

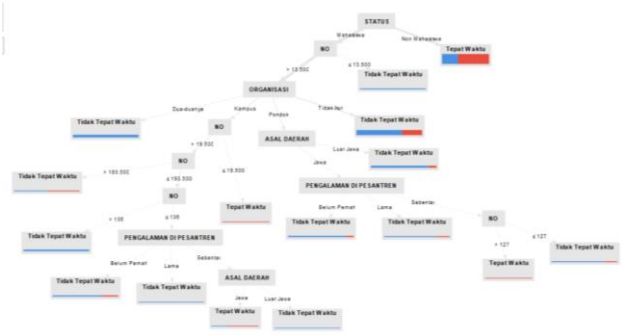

Figure 3. Decision Tree

From the picture above, whether they are students is the most significant influence on timeliness in completing the khataman. They were then followed by the activities of the students themselves, whether they joined the organization or not and the experience of the pesantren and the area of origin of the students. In validating the dataset, the authors use the cross-validation method. The data from the decision tree is then processed, and the results are calculated using variable cross-validation, the results of which are as follows.

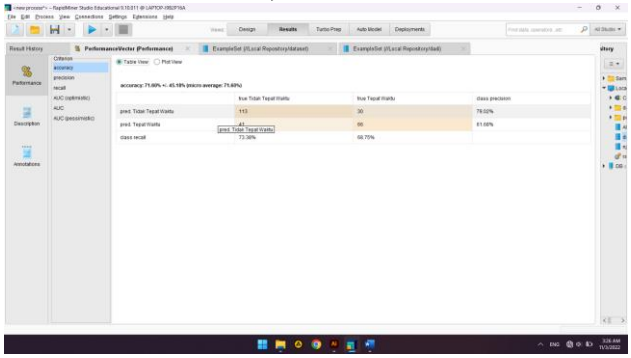

Fig. 4. Cross-validation results

From the cross-validation calculation, it was found that the prediction of the timeliness of finishing the tahfidz students at the PPTQ Al Asy'ariyyah level was 68.75% and 61.68% for those who were on time, while for predictions that were not on time were 73.38% and 79. 02%.

4. CONCLUSION

Based on the results of the evaluation and validation of effects as well as trials in data mining research predicting the timeliness of completing the completion of tahfidz PPTQ Al Asyariyyah students using the C4.5 algorithm, researchers can conclude that the most significant influence for students in achieving their completion time is : non-student status has a greater chance of completing their graduation on time than students, Santri who do not join an organization are more likely to complete their khatam on time than students who take part in an organization, students who come from outside Java, almost all of them are not on time to complete their graduation. then the level of timeliness of PPTQ Al Asy'ariyyah students in completing their khatam is relatively high, but there are still many students who have not been able to complete their khatam on time

5. REFERENCES

Azwanti, Nurul. (2018). *Analisa Algoritma C4.5 Untuk Memprediksi Penjualan Motor Pada PT.Capella Dinamik Nusantara Cabang Muka Kuning.* Batam: Universitas Putera Batam

Firmansyah, (2011). *Penerapan Algoritma Klasifikasi C4.5 Untuk Penentuan Kelayakan Pemberian Kredit Koperasi.* Jakarta.

Harahap, Fitriana. (2015). *Penerapan Data Mining Dalam Memprediksi Pembelian Cat.* Medan: STMIK Potensi Utama.

Clean Energy and Smart Technology
Volume 01, Number 02, April 2023
DOI : https://doi.org/10.58641/cest.v1i2.39
e-ISSN : 2964-2647

Haryati, Sudarsono, dan Suryana. (2015). *Implementasi Data Mining Untuk Memprediksi Masa Strudi Mahasiswa Menggunakan Algoritma C4.5* (Studi Kasus: Univversitas Dehasen Bengkulu), Jurnal Media Infotama Vol. 11 No. 2, September 2015, ISSN 1858 – 2680.

Kusrini. (2007). *Design And Implementation Of Building Decision Tree Using C4.5 Algorithm*. http://elearning.amikom.ac.id/index.php/download/karya/586/a733b5873027a d0abaac6682499a3914 (diunduh pada 20 oktober 2022)

Kusrini, & Luthfi, E. T. (2009). *Algoritma Data Mining*. Yogyakarta: Andi Publishing.

Mardi. (2016). *Data Mining Klasifikasi Menggunakan Algoritma C4.5*, Jurnal Edik Informatika, ISSN: 2407-0491.

Rani, Larissa Navia. (2015). *Klasifikasi Nasaah Menggunakan Algoritma C4.5 Sebagai Dasar Pemberian Kredit.* Padang: Universitas Putera Indonesia YPTK.

Sunge, Aswan Supriyadi. (2018). Prediksi Kompetensi Karyawan Menggunakan Algoritma C4.5 (Studi Kasus: PT. Hankook Tire Indonesia). Bekasi: STT Pelita Bangsa.

Widiarto dan Muchamad Piko Henry. (2011). *Pengambilan Pola Kelulusan Tepat Waktu Pada Mahasiswa Stmik Amikom Yogyakarta Menggunakan Data Mining Algoritma C4.5.* Yogyakarta: Sekolah Tinggi Manajemen Informatika Dan Komputer Amikom. http://repository.amikom.ac.id/index.php/add_downloader/Publikasi_04.22.04 00.pdf/1201 (diunduh pada 25 oktober 2022)