Transaction on Informatics and Data Science
Vol. 1, No. 2, 2024, 85-96
ISSN: 30641772 (online)
DOI: https://doi.org/10.24090/tids.v1i2.12278

Analysis of Tahfidz Program Entrance Examination Using C4.5 Method for Classification of Learning Groups in The Al-Mawaddah Region of Nurul Jadid Islamic Boarding School, Probolinggo, Indonesia

Wahab Sya'roni *[1], Yulia Sri Devi [1], Heriyanto [1], Yanuar Ilmansyah [1]

[1] Informatics Engineering Department, Faculty of Engineering, Nurul Jadid University, Probolinggo, Indonesia, wahab.syaroni@gmail.com

Article Information

Submitted March 12, 2024
Accepted June 26, 2024
Published October 1, 2024

Abstract

The tahfidz program aims to improve understanding and memorization of the Al-Quran in the Al-Mawaddah Region, Probolinggo, Indonesia, including those who organize the *Tahfid* Qur'an program. This research aims to increase the efficiency and accuracy of analyzing the Tahfid program entrance test results using the C4.5 Algorithm for study group classification and visualizing the results through a web-based interactive application using Streamlit. The C4.5 algorithm is used to build a classification model based on the results of the entrance test criteria. The application of Streamlit as a framework for creating interactive web applications makes it easier for users to enter entrance test data and study group classification results, making it easier to make decisions and plan to learn for *Tahfid* program administrators. The research results show that the use of the C4.5 Algorithm method and the streamlit application is practical in analyzing the results of the *Tahfidz* program entrance test and determining study groups, the C4.5 Algorithm succeeded in achieving a high level of accuracy with an accuracy of 94%, then it was implemented using streamlit. With the interactive application, *tahfid* program managers can carry out the selection and grouping process more quickly and precisely and provide a more targeted learning approach according to the student's ability level.

Keywords: Tahfidz program, C4.5, classification, Streamlit.

1. Introduction

In the current era of information and technology, machine learning algorithms and data analytics have become an integral part of various lives, including education [1]. The Al-Mawaddah area is one of the Nurul Jadid Pesantren areas with a *tahfidz* program, which aims to produce *hafidzah* who can memorize the Al-Qur'an well. A critical aspect of the *Tahfidz* program is forming influential study groups. In this context, analyzing entrance test results can provide valuable information to help understand the Al-Qur'an memorization ability of *hafidzah* candidates. This analysis can help pesantren group *santri* to study groups according to ability levels [2].

The division of this study group is essential because, according to the facts in the Quran *Tahfidz* Program, students with fluent memorization abilities become indicators of memorization assessment. Meanwhile, students with less or basic skills will be grouped according to their basic abilities. This study group is divided into four groups: *Khotimat*, Acceleration, *Tahfidz*, and *Tahsin*. So, the division of this study group is based on the assessment results during the entrance test exam so that the ability to do it has a balanced competence and can motivate students to be more active and enthusiastic in learning and memorizing the Al-Qur'an [3]. In remembering, it turns out that it is also like in other places; many methods are used, so it is interesting for researchers to conduct further

* **Author Correspondence**: Wahab Sya'roni: Nurul Jadid University, PP Nurul Jadid road, Karanganyar, Paiton, Probolinggo, East Java 67291,– Indonesia. Email: wahab.syaroni@gmail.com.

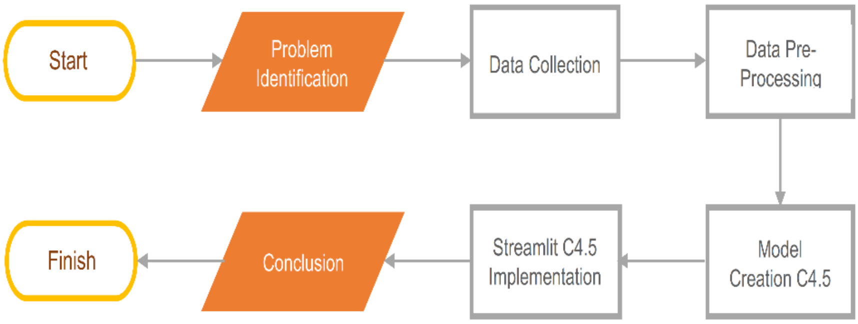
research on memorization methods carried out by students [4]. The requirements for this test include six criteria, including the Number of Memorization, *Munjiyat, Tajweed, Nadzom Hidayatus Sibyan*, Memorization Speed, and *Juz Amma*. They hold all these tests after you register for the *Tahfidz* program. Especially for the *Khotimat* study group, which can be seen from the number of memorizations that have been completed, and for the accelerated study group, specifically for students who are not formal school students. In the process of dividing the study group, the examiner still uses a manual method, namely calculating the value of each test result before the prospective *tahfidz* students enter, so it takes a long time and is less accurate in placing students in each group, resulting in less effective learning because it is not following their abilities [5] [6].

To solve this problem, we need a method to make the group division process faster and more effective. Based on the results of the analysis and reading from several journals, the suitable method is Data Mining Algorithm C4.5, which can classify study groups in the Streamlit-based *tahfidz* program to overcome the above problems so that it can produce accurate data. The C4.5 algorithm is one of the classification algorithms in machine learning and data mining. The purpose of C4.5 is to study the mapping from attribute values to categories that can be used to categorize unknown items into new categories [7], [8]; examiners are expected to use this method to help determine study groups in the *Tahfidz* Program Al-Mawaddah Region.

2. Method

The research stages in this study include problem identification, data collection (literature study, observation, and interviews), data pre-processing (data selection, data cleaning, transformation), modeling the C4.5 algorithm, applying the C4.5 algorithm using Streamlit, and drawing conclusions. Streamlit is an open-source Python framework [9]. The research stage as shown in Figure 1.

Figure 1.
Research stage.



1) Problem Identification

The problem at Al-Mawaddah is that grouping exam test results is still manually using spreadsheets, so it takes a long time to process. Therefore, this research proposes an approach using the C4.5 algorithm, implemented using Streamlit, to overcome the existing problems. The hope is that C4.5 applications built using Streamlit can increase efficiency and accessibility in analyzing and grouping students at Al-Mawaddah.

2) Data Collection

The data collection process was carried out by observation and interviews. Observations were made at the Al-Mawaddah Wilayan Office of Nurul Jadid Islamic Boarding School. At the same time, interviews were conducted with questions and answers with the *Tahfidz* Program Coordinator in the area. The data collected is from program entrance test data from 2019-2023. Some criteria will be used as a dataset for calculation/model building using the C4.5 and streamlit algorithms.

3) Data Pre-Processing

This step aims to sort and normalize the dataset to be used. At this step, there are three processes carried out:

a. Data Selection

Data selection eliminates irrelevant features if features do not significantly contribute to the analysis or classification.

b. Data Cleaning

Data cleaning is used to address missing values contained in the dataset to be used, address outliers or unusual values that may affect the analysis and replace them with appropriate values, and detect and address duplicate data.

c. Transformation

At this step, the process of grouping the test data entered by the program into attribute groups will be applied to data mining analysis. In addition, adding new relevant attributes is possible in the data processing process. This aims to increase the accuracy of the algorithm.

4) Model Building

At this step, the dataset that has been pre-processed is then used to create a classification model using the C4.5 algorithm. At this step, the parameters for C4.5 are also determined to obtain good accuracy results in classifying student study groups. The optimal model obtained in this process will then be used to implement the C4.5 algorithm using the Streamlit framework.

5) Streamlit C4.5 Implementation

At this step, C4.5 is implemented with the previously created model. In this implementation process, the data entered the system will be processed using the previously developed model to produce a decision tree from the C4.5 algorithm classification process.

6) Drawing Conclusions

This step evaluates the results of the classification process carried out using the Confusion Matrix. Confusion Matrix is used to measure the level of accuracy of the C4.5 algorithm in classifying student study groups [10]. Conclusions are drawn to assess the success of implementing the C4.5

algorithm in determining study groups in the *Tahfid* Program. These conclusions can include the level of accuracy that has been achieved, the strengths and weaknesses of the algorithm, and recommendations for further development.

3. Result and Discussion

The data collection results are the entrance test data of students in the Tahfid Program in the Al-Mawaddah Region of Nurul Jadid Islamic Boarding School. The data is then pre-processed before the classification process using the C4.5 algorithm. The total data obtained is 584 program entrance test data with a period of 2019-2023, with details of the number of students in the study group as shown in Table 1.

Table 1.
Amount of data obtained.

Groups	Amount
Khotimat	7
Acceleration	45
Tahfid	411
Tahsin	121
Total	584

The sample of the dataset as shown in Table 2.

Table 2.
Dataset sample

No	Name	Institution	Amount of memorization	*Munjiyat*	*Tajwid*	*Nadzom*	Speed	*Juz Amma*	Group
1	Luluk Mutiara Maknunah	Formal	0	45S	87	60	70	60	*Tahsin*
2	Irin Nuril Sabila	Formal	0	66	75	55	80	55	*Tahsin*
3	A'isyah Anwar	Formal	0	40	80	55	70	40	*Tahsin*
4	Lyna Auliya Muthi'ah	Formal	0	20	70	60	60	45	*Tahsin*

No	Name	Institution	Amount of memorization	***Munjiyat***	***Tajwid***	***Nadzom***	**Speed**	***Juz Amma***	**Group**
5	Auziah Indah Permatasari	Formal	0	56	60	55	60	60	*Tahsin*
6	Eva Dwi Mariyatul Qivtiyah	Non-formal	0	60	60	65	78	45	Acceleration
7	Sa'daika Mikaela Putri Ramadani	Non-formal	0	45	78	50	60	30	Acceleration
8	Alifia Ilma Sabila	Non-formal	0	55	60	55	85	30	Acceleration
9	Nilna Zahrah Afifah	Non-formal	0	54	85	60	60	25	Acceleration
10	Uswatun Nisa'	Non-formal	0	40	60	60	75	75	Acceleration

Pre-Processing

a. Data Selection

This step involves sorting the data that will be used to implement data mining. Data mining is discovering interesting patterns and knowledge from large amounts of data [11]. The sorting process is carried out after the data has been successfully obtained from the Al-Mawaddah Pesantren Nurul Jadid Regional Tahfid Program Office. The relevant data for this implementation includes several attributes, namely institution data, number of memorizations, *munjiyat* scores, *munjiyat* scores, *tajweed* scores, *nadzom* scores, speed scores, juz 30 scores, and study group information.

b. Data Cleaning

The next step is to clean data by deleting and completing incomplete data, eliminating duplication, and correcting errors. In this dataset, variables that will not be used will be removed as they are irrelevant to the analysis process. The focus will be on the seven variables to be analyzed, namely institution, number of memorizations, *munjiyat* score, *tajweed* score, *Nazem* score, speed score, and juz 30 score, as listed in the table 3.

Table 3.
Exam result according to institution

Institution	**Amount of Memorization**	***Munjiyat***	***Tajwid***	***Nadzom***	**Speed**	***Juz Amma***	**Group**
Formal	0	45	87	60	70	60	*Tahsin*
Formal	0	66	75	55	80	55	*Tahsin*
Formal	0	40	80	55	70	40	*Tahsin*
Formal	0	20	70	60	60	45	*Tahsin*
Formal	0	56	60	55	60	60	*Tahsin*
Non formal	0	60	60	65	78	45	Acceleration
Non formal	0	45	78	50	60	30	Acceleration
Non formal	0	55	60	55	85	30	Acceleration
Non formal	0	54	85	60	60	25	Acceleration
Non formal	0	40	60	60	75	75	Acceleration

c. Data Transform

The next step is to transform the data as follows: First, convert the Institution data by following the following conditions.

Institution	Terms
Formal	1
Nonformal	0

The second transformation converts the value data into an index with a specified range.

Range	Terms
>=70	1
<=70	0

The third transformation transforms the study group data under the following conditions:

Group	Terms
Akselerasi	0
Khotimat	1
Tahfid	2
Tahsin	3

The fourth transformation is to change the Memorization Count by following the following conditions.

Amount	Terms
0 Juz	0
1-29 Juz	1
30 Juz	2

The results of the transformation data processing as shown in Table 4.

Table 4.
Transformation of data processing

Institution	**Amount**	***Munjiyat***	***Tajwid***	***Nadzom***	**Speed**	**Juz Amma**	**Group**
1	0	1	0	1	1	1	3
1	0	1	0	1	0	1	3
1	0	1	0	1	1	1	3
1	0	1	1	1	1	1	3
1	0	1	1	1	1	1	3
0	0	1	1	1	0	1	0
0	0	1	0	1	1	1	0
0	0	1	1	1	0	1	0
0	0	1	0	1	1	1	0
0	0	1	1	1	0	0	0

C4.5 Algorithm Modeling

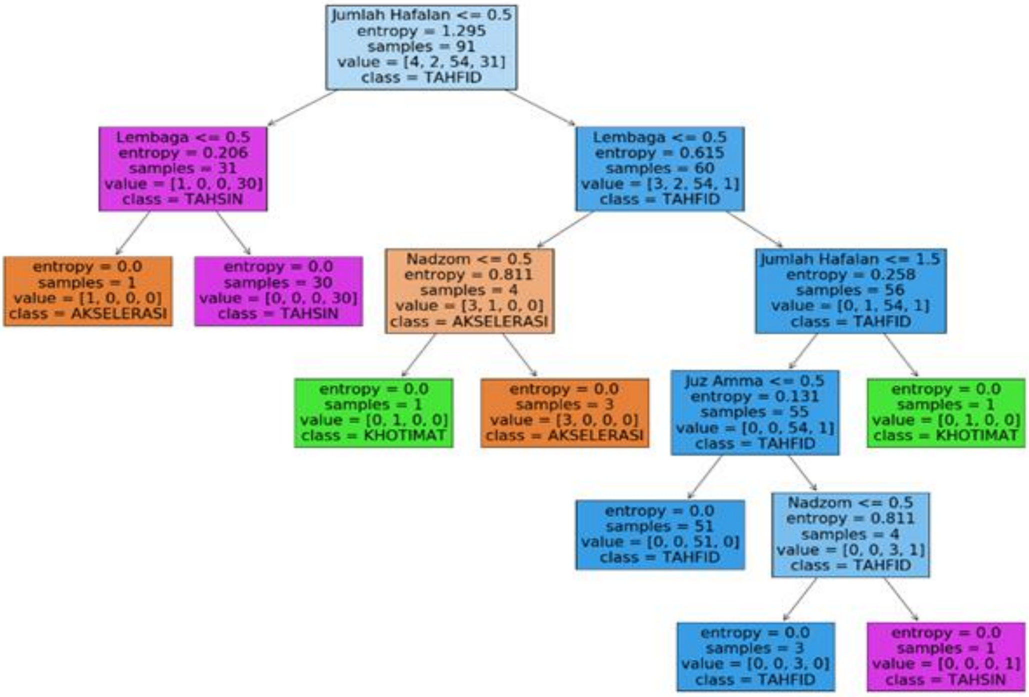
Modeling the C4.5 algorithm on the dataset used in its application using the Python programming language with Visual Studio Code tools [12], [13]. Modeling uses 80% as training data and 20% as testing data [9] and obtains 94% accuracy, 92% precision, and 84% recall and the decision tree [14] as shown in Figure 2.

Streamlit C4.5 Algorithm Model Implementation

Based on the analysis and design results in the previous step, the next step is to test the study group classification application in the tahfid program using the streamlit web-based C4.5 algorithm. The following figure is a view of the main page of the C4.5 algorithm streamlet web.

On the display of determining study groups in Figure 3, there is a feature where users will input alternative data along with criteria that have been weighted so that it will display the results of study group predictions.

Figure 2.
C4.5 Algorithm decision tree result



The figure 2 shown illustrates a decision tree generated using the C4.5 method for classifying learning groups in the *Tahfidz* Program at Al-Mawaddah Nurul Jadid Islamic Boarding School. The decision tree categorizes students based on various attributes such as *"Jumlah Hafalan"* (amount of memorization), "*Lembaga*" (institution), "*Nadzom*" (memorization of poetry), and other factors, helping to determine the appropriate group for each student. The target groups in this classification include *Tahsin*, acceleration, *Tahfid*, and *Khotimat*.

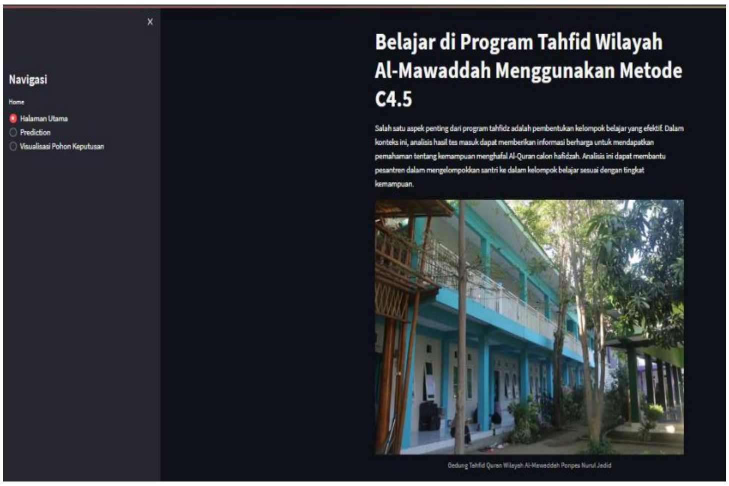
At the top of the tree, the root node is based on the "*Jumlah Hafalan*" (amount of memorized Quran). If the student's memorization is less than or equal to 0.5, the tree splits into two branches based on the value of the "*Lembaga*" attribute. If the "*Lembaga*" value is less than or equal to 0.5, students are typically classified as either Tahsin or Akselerasi, depending on further conditions. If the "*Jumlah Hafalan*" is greater than 0.5, additional attributes such as "*Nadzom*" and "*Juz Amma*" are evaluated to classify the students.

In the first major branch of the tree (students with "*Jumlah Hafalan*" <= 0.5), the decision split for "*Lembaga*" <= 0.5 further categorizes students. For example, if "*Lembaga*" <= 0.5, most of the students are classified into the Tahsin group. However, if specific values of other variables are met (such as "*Nadzom*" or "*Lembaga*"), students may fall into different categories like *Khotimat*.

The second major branch of the tree focuses on students with "*Jumlah Hafalan*" > 0.5. In this branch, the tree splits based on further criteria, such as "*Juz Amma*" and "*Nadzom*." For students who have memorized a larger portion of the Quran (greater than 1.5), they are classified into Khotimat or Tahfid groups. The decision tree uses entropy and information gain to measure the uncertainty at each node, and as the tree progresses, the uncertainty (entropy) decreases until final decisions are made.

Figure 2 demonstrates how the C4.5 algorithm can effectively classify students into learning groups based on their performance in memorizing Quranic verses, institutional background, and other learning metrics. This classification helps the *Tahfidz* Program efficiently allocate students to appropriate learning groups, optimizing their Quranic education pathways.

Figure 3.
Main page.

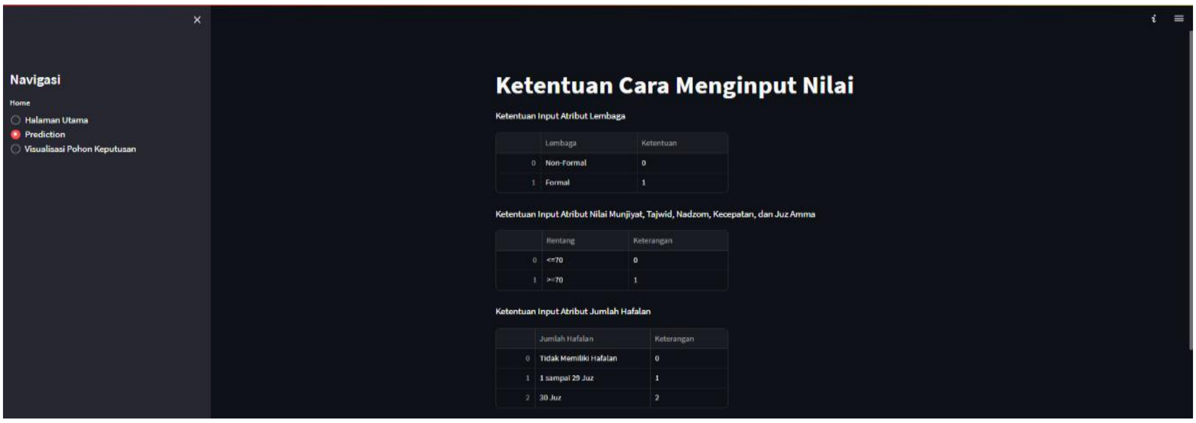


Belajar di Program Tahfid Wilayah Al-Mawaddah Menggunakan Metode C4.5

Salah satu aspek penting dari program tahfidz adalah pembentukan kelompok belajar yang efektif. Dalam konteks ini, analisis hasil tes masuk dapat memberikan informasi berharga untuk mendapatkan pemahaman tentang kemampuan menghafal Al-Quran calon hafidzah. Analisis ini dapat membantu pesantren dalam mengelompokkan santri ke dalam kelompok belajar sesuai dengan tingkat kemampuan.

Gedung Tahfid Quran Wilayah Al-Mawaddah Ponpes Nurul Jadid

Figure 4.
Study group determination page.



Ketentuan Cara Menginput Nilai

Ketentuan Input Atribut Lembaga

	Lembaga	Ketentuan
0	Non-Formal	0
1	Formal	1

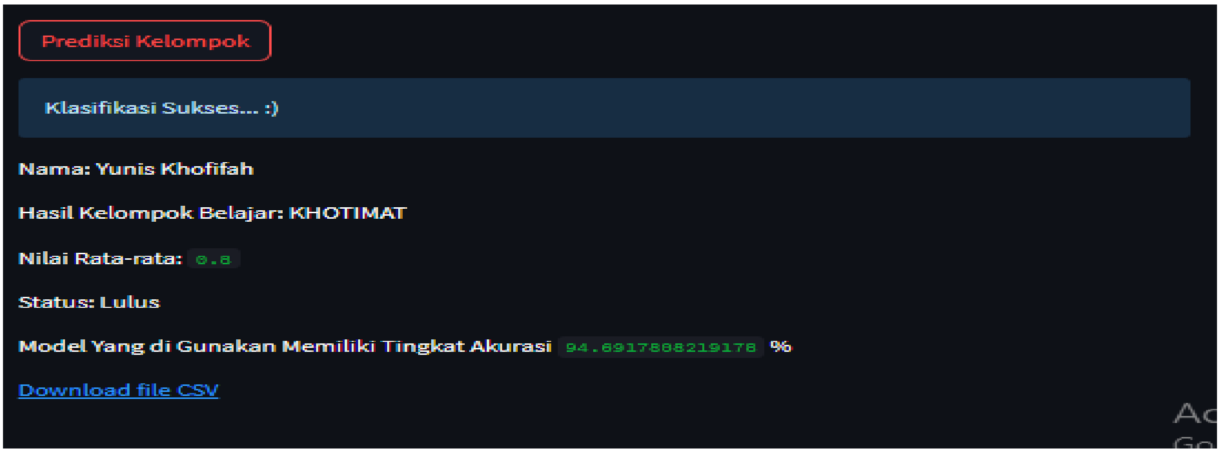
Ketentuan Input Atribut Nilai Munjiyat, Tajwid, Nadzom, Kecepatan, dan Juz Amma

	Rentang	Keterangan
0	<=70	0
1	>=70	1

Ketentuan Input Atribut Jumlah Hafalan

	Jumlah Hafalan	Keterangan
0	Tidak Memiliki Hafalan	0
1	1 sampai 29 Juz	1
2	30 Juz	2

Figure 5.
Accuracy result display.



Klasifikasi Sukses... :)

Nama: Yunis Khofifah

Hasil Kelompok Belajar: KHOTIMAT

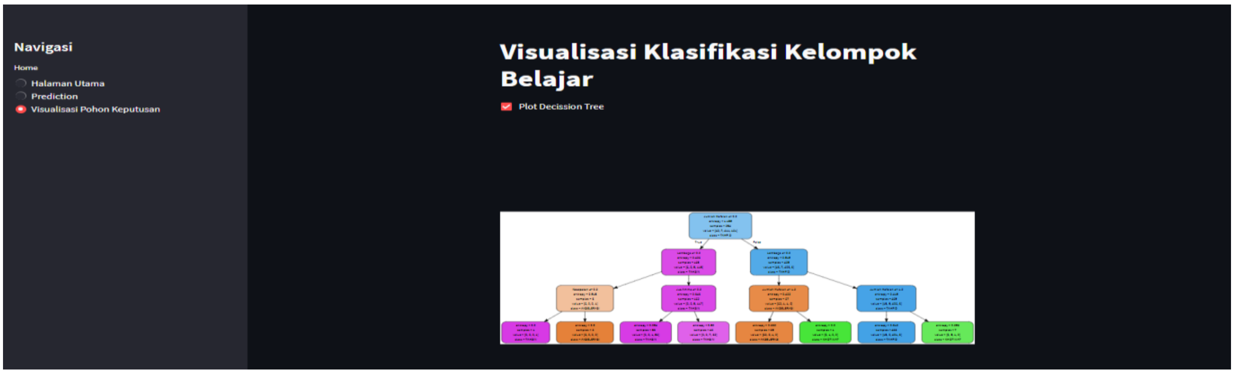
Nilai Rata-rata: 0.8

Status: Lulus

Model Yang di Gunakan Memiliki Tingkat Akurasi 94.8917808219178 %

Download file CSV

Figure 6.
C4.5 algorithm decision tree view in Streamlit.



Visualisasi Klasifikasi Kelompok Belajar

☑ Plot Decission Tree

4. Conclusion

Based on this research, it can be concluded that the classification of study groups in the *Tahfid* Program using the C4.5 Algorithm method was successfully implemented. The results of applying this classification technique can provide an overview and prediction of the results of the classification of *Tahfid Santri* study groups using variables related to the classification of study groups. In the modeling trial using the C4.5 algorithm, the accuracy result was 94%, the recall was 92%, and the precision value was 84%, with 584 data used and 117 data testing data. Therefore, the use of the C4.5 algorithm in the classification of study groups at the Al-Mawaddah Regional **Tahfid** Program can be categorized as a reasonable classification.

5. References

[1] A. P. Sudaryanto and S. Hanny, "Manajemen Sumber Daya Manusia Sektor Publik Menghadapi Kemajuan Kecerdasan Buatan (Artificial Intelligence)," *Musamus J. Public Adm.*, vol. 6, no. 1, pp. 513–521, Jul. 2023, doi: 10.35724/mjpa.v6i1.5402.

[2] M. M. El Iq Bali and M. A. A. Fatah, "Pengelolaan Program Tahfidz Dalam Meningkatkan Kemampuan Membaca dan Menghafal Al Qur'an," *J. Educ. FKIP UNMA*, vol. 9, no. 2, pp. 534–540, May 2023, doi: 10.31949/education.v9i2.4835.

[3] S. Dewi and S. Fatimah, "Implementasi Metode Clustering Algoritma K-Means untuk Menentukan Kelompok Tahfidz dan Tahsin di Pesantren Siswa Al-Ma'soem," *J. Account. Inf. Syst.*, vol. 6, no. 1, pp. 10–18, Mar. 2023, doi: 10.32627/aims.v6i1.701.

[4] A. Arlina, M. S. Bagus, M. I. Mazid, A. Limbong, and E. A. Elsil, "Metode Menghafal Al-Qur'an di Yayasan Tahfidz Qur'an Al-Husna Sei Kepayang," *J. Educ.*, vol. 5, no. 2, pp. 3184–3192, Jan. 2023, doi: 10.31004/joe.v5i2.984.

[5] B. Arifin and S. Setiawati, "Gambaran Strategi Pembelajaran Tahfidz Al-Quran," *J. Pendidik. Tambusai*, vol. 5, no. 2, pp. 4886–4894, 2021, doi: https://doi.org/10.31004/jptam.v5i2.1709.

[6] N. Nurhasanah, "Pengembangan Tes Untuk Mengukur Kemampuan Penalaran Mahasiswa Mata Kuliah Geometri," *Pepatudzu Media Pendidik. dan Sos. Kemasyarakatan*, vol. 14, no. 1, p. 62, May 2018, doi: 10.35329/fkip.v14i1.186.

[7] A. Purwanto and H. W. Nugroho, "Analisa Perbandingan Kinerja Algoritma C4.5 dan Algoritma K-Nearest Neighbors untuk Klasifikasi Penerima Beasiswa," *J. Teknoinfo*, vol. 17, no. 1, p. 236, Jan. 2023, doi: 10.33365/jti.v17i1.2370.

[8] S. Febriani and H. Sulistiani, "Analisis Data Hasil Diagnosa Untuk Klasifikasi Gangguan Kepribadian Menggunakan Algoritma C4.5," *J. Teknol. dan Sist. Inf.*, vol. 2, no. 4, pp. 89–95, 2021, doi: https://doi.org/10.33365/jtsi.v2i4.1373.

[9] M. F. Nur Syahbani and N. G. Ramadhan, "Klasifikasi Gerakan Yoga dengan Model Convolutional Neural Network Menggunakan Framework Streamlit," *J. MEDIA Inform. BUDIDARMA*, vol. 7, no. 1, p. 509, Jan. 2023, doi: 10.30865/mib.v7i1.5520.

[10] F. M. A. Sofyan, A. P. Riyandoro, D. F. Maulana, and J. H. Jaman, "Penerapan Data Mining dengan Algoritma C5.0 Untuk Prediksi Penyakit Stroke," *J-SISKO TECH (Jurnal Teknol. Sist. Inf. dan Sist. Komput. TGD)*, vol. 6, no. 2, p. 619, Jul. 2023, doi: 10.53513/jsk.v6i2.8578.

[11] Z. Nabila, A. R. Isnain, and Z. Abidin, "Analisis Data Mining Untuk Clustering Kasus Covid-19 di Provinsi Lampung dengan Algoritma K-Means," *J. Teknol. dan Sist. Inf.*, vol. 2, no. 2, p. 100, 2021, doi: https://doi.org/10.33365/jtsi.v2i2.868.

[12] S. Junaidi, M. Devegi, and H. Kurniawan, "Pelatihan Pengolahan dan Visualisasi Data Penduduk menggunakan Python," *ADMA J. Pengabdi. dan Pemberdaya. Masy.*, vol. 4, no. 1, pp. 151–162, Jul. 2023, doi: 10.30812/adma.v4i1.2963.

[13] A. Harper and T. Monks, "A Framework to Share Healthcare Simulations on the Web Using Free and Open Source Tools and Python," in *Proceedings of SW21 The OR Society Simulation Workshop*, Operational Research Society, Mar. 2023, pp. 250–260. doi: 10.36819/SW23.030.

[14] S. S. Shabrilianti, A. Triayudi, and D. A. Lantana, "Analisis Klasifikasi Perfomance KPI Salesman Menggunakan Metode Decision Tree Dan Naïve Bayes," *J. Ris. Komput.*, vol. 10, no. 1, pp. 182–191, 2023, doi: http://dx.doi.org/10.30865/jurikom.v10i1.5628.

This page is intentionally left blank