Rule Based System to Support Decisions on Determining Employee Status (Lecturers) for Scholarship Student Graduates

Hotma Sadariahta Sipayung
Master of Information Technology
Universitas Teknologi Digital Indonesia
Yogyakarta, Indonesia
email: hotmasipayung@gmail.com

Widyastuti Andriyani[1]
Master of Information Technology
Universitas Teknologi Digital Indonesia
Yogyakarta, Indonesia
email: widya@utdi.ac.id

Bambang P.D.P
Master in Information Technology
Universitas Teknologi Digital Indonesia
Yogyakarta, Indonesia
emial: bpdp@utdi.ac.id

Danny Kriestanto
Informatics
Universitas Teknologi Digital Indonesia,
Yogyakarta, Indonesia
email: danny@utdi.ac.id

ne of the problems that occurred at the Indonesian Digital Technology University (UTDI) is the selection process for prospective Permanent Lecturers of the foundation which is said to be new to be applied to students who receive Masters scholarships at the Masters of Information Technology (MTI) UTDI Yogyakarta. The criteria used in the rules are Semester 1 Achievement Index (IP), Semester 2 IP, Semester 3 IP, Cumulative Achievement Index (IPK), Paper (scientific work), Cooperation, Discipline, Communication, Pre-Thesis, Thesis, C Grade, and Length of Study obtained from MTI UTDI, then will use the C4.5 Algorithm to produce a decision tree that will be used as a rule in the system.
This study uses rules obtained from MTI UTDI by the Head of the Study Program (Kaprodi), namely 41 training data and 8 test data. Using forward chaining as a method in an expert system that seeks solutions through problems, then using the C4.5 Algorithm which is an algorithm used to form a decision tree. The rules formed are then used to predict the eligibility of Masters scholarship graduates to become Permanent Lecturers, Contract Lecturers, or not eligible. The prediction results are then evaluated using a confusion matrix and the accuracy value is 75%, Precision is 77.78% and Recall is 77.78%. So that the C4.5 Algorithm using the RapidMiner application is quite feasible to be used to support decision making in the selection of Masters scholarship students who will be appointed as Permanent Lecturers, Contract Lecturers or those who do not meet the requirements as Lecturers at UTDI Faculty of Information Technology.

KeyWords*: rule-based system, scholarship students, C4.5 algorithm, decision tree, permanent lecturers, contract lecturers*

This Article was:
submited: 11-06-24
accepted: 24-06-24
publish on: 20-07-24

How to Cite:
H. S. Sipayung, et al, "Rule Based System to Support Decisions on Determining Employee Status (Lecturers) for Scholarship Student Graduates", Journal of Intelligent Software Systems, Vol.3, No.1, 2024, pp.36–47, 10.26798/jiss.v3i1.1337

1 Introduction

UTDI has a Master of Information Technology (MTI) study program that provides scholarships to several qualified students at the beginning of admission which will then be re-evaluated for academic results each semester towards the next selection stage to become a Lecturer at UTDI. Not all scholarship recipients can be appointed as lecturers after graduating, but there are criteria that are assessed, including IP, Paper, Grade C, Cooperation, Discipline, Communication, Pre-Thesis, Thesis and Length of Study [1]. Then there are several other requirements that must be met [2]. Using a rule-based system to store and manipulate knowledge to interpret information [3] in a way that is useful for solving problems with rules created based on expert knowledge [4–6] to capture human expertise and decision making using conditions (if) and actions (then) [7]. By using the forward chaining method to find solutions through problems, from existing rules then leading to a conclusion based on these facts [8,9]. Classifying data by forming a decision tree using the C4.5 algorithm [10,11], the results of the analysis of the formed tree diagram are easy to understand, easy to make and require less experimental data, can be implemented with continuous and discrete values [12], the results are easy to understand, the computing time is relatively fast [13] and the accuracy can match other classifications [14]. The data is transformed using the C4.5 algorithm into a decision tree and rules [15], then for problems in classification, the measurements used are precision, recall and accuracy [8].

1.1 Type of Research. This research is included in the type of applied research that uses the action research method which is actually carried out because it has the aim of finding a solution to a problem that is directly faced by the community, or industrial/business organizations [16], researchers are directly involved starting from finding the problem, planning the action, implementing the action and evaluation [16].

1.2 Research Flow Diagram. To achieve the research objectives, the stages of activities to be carried out need to be planned in advance [16]. Starting with formulating the problem, conducting interviews with the MTI team, determining the research objectives, conducting literature studies, collecting rule data, conducting calculations and analysis, conducting trials, then analyzing the research results and making conclusions from the research.

1.3 Pseudocode Algorithm C4.5. The following is pseudocode of the decision tree construction algorithm C4.5 [17].

It is a pseudocode of the C4.5 algorithm that functions to form a decision tree. The calculation starts from calculating the number of

[1]Corresponding Author.

Table 1 Assessment criteria

Kriteria		Nilai
IP	IPS Semester 1 Decimal Score	Decimal Score
	IPS Semester 2 Decimal Score	Decimal Score
	IPS Semester 3 Decimal Score	Decimal Score
	IPK Decimal Score	Decimal Score
Paper	Sinta 1	YES NO
	Sinta 2	YES NO
	Sinta 3	YES NO
	Sinta 4	YES NO
	Sinta 5	YES NO
	Sinta 6	YES NO
	Garuda	YES NO
	Google Schoolar	YES NO
	Scopus	YES NO
Grade C		YES NO
Cooperation		Capable Not Capable
Discipline		Good Not Good
Communication		Good Not Good
Prathesis Preproposal	Prathesis Preproposal	On Time Not On Time
	Proposal	On Time Not On Time
Thesis	Result 1	On Time Not On Time
	Result 2	On Time Not On Time
Duration of Study		Number Value
Expert		Permanent Lecturer Contract Lecturer Does Not Meet

Algorithm 1 FormTree(T)

```
function FormTree(T)
    ComputeClassFrequency(T)
    if OneClass or FewCases then
        return leaf
    end if
    Create a decision node N
    for all Attribute A do
        ComputeGain(A)
    end for
    N.test ← AttributeWithBestGain
    if N.test is continuous then
        find Threshold
    end if
    for all T′ in the splitting of T do
        if T′ is Empty then
            Child of N is a leaf
        else
            Child of N ← FormTree(T′)
        end if
    end for
    ComputeErrors of N
    return N
end function
```

attributes and determining which attribute will be used as the root of the decision tree. Next, the calculation of Entropy, Gain, Split Info and Gain Ratio will be carried out to determine the leaves of the decision tree. After all calculations are complete, the decision tree can be formed based on the calculated Gain Ratio value. The attribute with the highest Gain Ratio value will be located at a higher priority and also have a higher position in the decision tree [18].

1.4 Data Collection. The rule data was taken from the MTI UTDI rule data of 41 training data and 8 test data in batch 1 of 2019. The criteria for the assessment are as follows on Table 1.

Table 2 Target Attributes

Target Attributes	Description
Permanent Lecturer	Appointed as a permanent lecturer at UTDI
Contract Lecturer	Appointed as an intern or contract teaching staff
Not Fulfilling	Considered not to meet the requirements as a permanent or contract teacher

1.5 Target Attributes. The target of this classification is to be able to determine whether a person is included in the Permanent Lecturer/Contract Lecturer class or is not eligible, as in the following Table 2.

1.6 Implementing the Algorithm Used. This study uses the C4.5 algorithm to carry out the data classification process by forming a decision tree [10]. The C4.5 algorithm is described as follows [19,20].

(1) Preparing training data can be seen in Table 1
(2) Calculating the value entropy

$$\text{Entropy } (S) = \sum_{i=1}^{n} -p_i \log_2 p_i \tag{1}$$

Explanation:

S	=	set of cases
A	=	feature
n	=	number of partitions S
i	=	Proportion of to S (Probability obtained from the number (yes) divided by the total cases)

(3) Calculating the Gain Value
In the C4.5 algorithm, the gain value is used to determine which variables are nodes of a decision tree. A variable that has the highest gain will be made a node in the decision tree.

$$\text{Gain}(S, A) = \text{Entropy}(S) - \sum_{i=1}^{n} \frac{S_i}{S} \cdot \text{Entropy}(S_i) \tag{2}$$

No	IPS Sem 1
Rule 11	3,00
Rule 9	3,40
Rule 10	3,50
Rule 7	3,60
Rule 8	3,60
Rule 5	3,70
Rule 6	3,70
Rule 4	3,80
Rule 1	4,00
Rule 2	4,00
Rule 3	4,00

IPS Sem 1	Median
3,00	
	3,2
3,40	
	3,45
3,50	
	3,55
3,60	
	3,65
3,70	
	3,75
3,80	
	3,9
4,00	

Fig. 1 Finding the Median of Numeric Valued Attributes

Explanation:

S = case set
A = attribute
n = A number of partitions of attribute A
$\frac{|S_i|}{|S|}$ = number of cases in attribute partition-i
$|S|$ = number of cases in S

(4) Calculating Split Info Value
Split info is used as a divisor of Gain(A) which will produce Gain Ratio SplitInfoGain Ratio

$$\text{SplitInfo}(S, A) = -\sum_{i=1}^{n} \frac{|S_i|}{|S|} \log_2 \left(\frac{|S_i|}{|S|} \right) \quad (3)$$

Explanation:

S = case set
A = attribute
S_i = number of samples for attribute i

(5) Calculating Gain Ratio Value
The highest Gain Ratio is selected as the test attribute for the node.

$$\text{GainRatio}(A) = \frac{\text{Gain}(S, A)}{\text{SplitInfo}(S, A)} \quad (4)$$

Explanation:

S = case set
A = attribute
Gain (S, A) = info gain on attributeA
Split(S,A) = split info on attributeA

Gain Ratio is another measure used to overcome problems with attributes that have very varied values. The highest Gain Ratio is selected as the test attribute for the node.

(6) Repeat the 2nd process until all attributes are used or meet the above conditions until all branches have the same class

1.7 Manual Calculation. In this study, manual calculations were carried out using existing formulas and were carried out using Microsoft Excel. When performing manual calculations, they were carried out carefully to obtain correct results and in accordance with the Rapidminer software.

1.7.1 Finding the Median of attributes with numeric values(numbers). Changing data into a form that is appropriate so that it can be processed with the C4.5 algorithm calculation [21], Attributes with numeric values are first sorted and then the median is calculated and grouped as shown in Figure 1 below. Performed for all attributes that have numeric values.

1.7.2 Find the values of Node 1 to Node 1.8.B. The highest Gain Ratio value will then be selected as a Node until all branches are fulfilled [22] as in Table 3 and Table 4. Classification is done on MTI UDI rule data. Calculations are done using Microsoft Excel software. Then calculate the data using the formula in the C4.5 algorithm. The calculation results in the 1st iteration can be seen in Table 3.

a. Node: Indicates the node being evaluated. In iteration 1, Node 1 is the starting point for the formation of the decision tree. This is the first node to consider reviewing data based on a particular attribute.
b. Number of Cases (S): The total number of cases or data considered at this node. At Node 1, all data available for analysis is entered.
c. Permanent Lecturer (S1): The number of cases in the node that are classified as permanent lecturers. This is the amount of data that meets the criteria to become permanent lecturers.
d. Contract Lecturer (S2): The number of cases in the node that are classified as contract lecturers. This is the amount of data that meets the criteria to become contract lecturers.
e. Not Fulfilled (S3): The number of cases in the node that are classified as ineligible. This is the amount of data that does not meet the criteria to become permanent or contract lecturers.
f. Entropy: The entropy value for the node. Entropy measures or disrupts data. The lower the entropy, the purer the data division at the node.
g. Gain: The gain value for the attribute used to separate the data at the node. Gain measures how well the attribute reduces brightness.
h. Split Info: Split info is a measure of the information needed to separate data at a node based on a particular attribute.
i. Gain Ratio: Gain Ratio is the ratio of gain to split info. It is used to overcome bias towards attributes with many categories.

Each row in the table details the calculation of entropy, gain, split info, and gain ratio for each data discount condition.

(1) IPS Sem 1: Displays the evaluation attributes of Index Prestasi Semester 1 (IPS Sem 1) to split the data. Values such as entropy, gain, split info, and gain ratio are calculated for various thresholds such as ≤3.10, ≤3.30, etc.
(2) >3.10: Indicates that the data is separated based on whether IPS Semester 1 is greater than 3.10. The number of cases in each category (permanent lecturer, contract lecturer, ineligible) is counted, and the values of entropy, gain, split info, and gain ratio are calculated to enable the effectiveness of such certification.

Based on the calculations in Table 3, the attribute with the highest gain ratio is selected as the attribute for the first node in the decision tree. This means that the attribute provides the most informative data disclosure and reduces the distance in the most effective way. The results of this first iteration are then used to determine further differentiation in subsequent nodes in subsequent iterations. Using this information, the decision tree begins to form with optimal criteria based on the attributes tested, which will aid in more accurate and relevant data classification. The iteration continues until the 13th iteration, namely node 1.8.B as in the Table 4.

Table 4 Results of the 13th Iteration Calculation (Node 1.8 B) contains details about the data division based on different attributes and the calculation of entropy, gain, split info, and gain ratio for each division. These results are used to form nodes in the decision tree using the C4.5 algorithm. The 13th Iteration refers to a process that has involved several previous verification and calculation steps, and Node 1.8 B is one of the nodes in the tree resulting from this iteration.

(1) Node: Indicates a particular node in the decision tree. This node is named after the attribute used to view the data at a particular iteration. Node 1.8 B indicates the 13th iteration and the Division at a particular attribute.

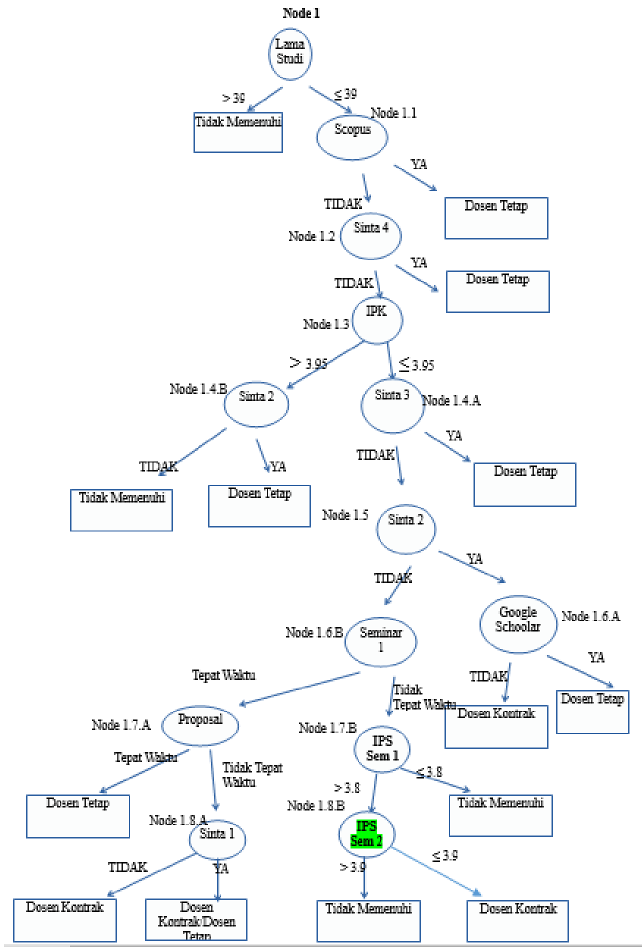

Fig. 2 Decision Tree Result of C4.5 Algorithm Calculation

(2) Number of Cases (S): The total number of cases residing on this node after fragmentation on the considered attribute. It includes all the data corresponding to the condition at that node.
(3) Permanent Lecturer (S1): The number of cases in the node that are classified as permanent lecturers. This category reflects the amount of data that meets the criteria to be a permanent lecturer.
(4) Contract Lecturer (S2): The number of cases in the node that are classified as contract lecturers. This category reflects the amount of data that meets the criteria to be a contract lecturer.
(5) Not Fulfilled (S3): The number of cases in the node that are classified as not fulfilled. This category reflects the amount of data that does not meet the criteria to be a permanent or contract lecturer.
(6) Entropy: The entropy value for the node. Entropy is a measure or disorder in data. The lower the entropy, the purer the data division at that node.
(7) Gain: The gain value for the attribute used for the division at that node. Gain is a measure of information.

1.7.3 Decision Tree from 41 Rules (Training Data). From the calculation using the C4.5 algorithm using 41 training data, a decision tree is formed as in Figure 2.

The decision tree has a simple and easy to implement structure, shaped like an inverted tree, where internal nodes (not leaves) indicate testing on attributes, each branch corresponds to the test results, and each external node (leaf) indicates class prediction [23]. The process in the decision tree is to convert data using the C4.5 algorithm into a decision tree, and into a rule also simplifies the rule [2].

1.8 Data Pre-Processing for Analysis dan Validation .

1.8.1 Data Pre-Processing.

Data Collection. Data collected from the Master of Information Technology (MTI) UTDI which includes 41 training data and 8 test data. The data collected includes various attributes relevant to the study, such as Semester Achievement Index (IP) from several semesters, Cumulative Achievement Index (IPK), scientific publications, collaboration, discipline, communication, pre-thesis, thesis, and length of study.

Data Cleaning. The collected data is evaluated to identify missing or inconsistent values. In the context of this study, the data is checked to ensure all required values are available and ready for further processing. This step is important to ensure data integrity before proceeding with further analysis.

Data Transformation. The raw data that has been collected is transformed into a format that is more suitable for analysis using the C4.5 algorithm. This includes calculating the median value for attributes with numeric values and then grouping them into relevant categories. This transformation helps in handling various types of data (continuous and discrete) so that it can be processed by the C4.5 algorithm to form a decision tree.

Categorical Data Encoding. Categorical data, such as acceptance rates in scientific publications (Sinta 1 to Sinta 6), are encoded into numeric format. This encoding is important to allow the data to be processed by algorithms that usually only accept numeric input.

Data Normalization. Numeric data such as IP and GPA are normalized to ensure that each attribute has a uniform scale. This step is important to prevent attributes with a large range of values from dominating the analysis results.

Numeric Attribute Grouping. Attributes with numeric values such as IP, GPA, and length of study are transformed into groups or categories that are easier to interpret. For example, IP values are divided into groups based on certain thresholds that are relevant to the purpose of the analysis.

Data Separation. The data is divided into training sets and testing sets to build and run the model. In this study, 41 data were used as training data to build the model, and 8 data were used as testing data to broadcast the built model.

Calculation of Gain and Entropy. After the data is ready, the next step is to calculate the entropy and gain values for each attribute. Entropy is used to measure the sharpness in the data, and gain is used to determine which attributes are the most informative to use as nodes in the decision tree.

Calculation of Gain Ratio. Gain ratio is calculated to overcome the problem of attributes that have many categories. Gain ratio is used to select the attribute with the highest gain as the test attribute for nodes in the decision tree.

1.8.2 Validation. The validation process is carried out using RapidMiner software. Validation is carried out using 41 training data obtained from MTI UTDI which are then tested using the C4.5 decision tree.

1.9 Limitations and Challenges. The C4.5 algorithm has several limitations and challenges that need to be addressed to provide a more balanced view. One of the main challenges is the tendency to overfit, especially when faced with training data that has many irrelevant attributes or variables. This overfitting causes the model to overfit the training data, resulting in a decrease in model performance on data that has never been seen before. C4.5 requires significant time and computational resources, especially when dealing with large datasets. The process of calculating entropy and gain ratio for each attribute at each stage requires a lot of computation that can affect efficiency. Although this algorithm is able to handle missing values, the way it is handled can affect the final outcome of the decision tree. The presence of irrelevant or noisy data can also affect the structure of the resulting decision tree, making it less stable and more prone to prediction errors. The decision tree produced by C4.5 can be very complex and difficult to interpret if there is no mechanism to prune irrelevant branches. This high tree complexity not only makes the results interpretable, but can also affect the performance of the model in predicting new data. This algorithm can also be biased towards attributes with many categories, as attributes with many categories tend to have higher gain ratios, which causes them to be selected as splitters in the early stages of tree construction, even though they may not provide the most significant information for the final decision. Although decision trees are considered to be easy-to-interpret models, very deep and complex trees can be difficult to interpret and understand for non-technical users. This can be a challenge in communicating the results of the analysis to stakeholders who do not have a technological background. Compared to other machine learning algorithms such as Random Forest or Gradient Boosting, the C4.5 algorithm may not be efficient or accurate in certain cases. These algorithms are able to better handle the problems of overfitting and sensitivity to data noise through various ensemble and regularization techniques. Although the C4.5 algorithm has some advantages in terms of ease of interpretation and handling of complex data, the above challenges indicate that there is a need to consider alternatives or modifications to this algorithm to improve performance and stability in various application contexts.

2 Results and Discussion

Table 3 shows the calculation results in the 1st iteration. The Study Length Criteria gets the highest gain ratio with a value of 0.578907463 and the highest entropy value is "$leq39$" with a value of 1.262260454. From Table 2, because there is the highest gain, it is continued to the 13th iteration. The calculation results in the 13th iteration can be seen in Table 3. Table 4 is the calculation of the 13th iteration and it can be seen that the highest gain ratio is obtained in the IPS Sem 2 attribute with a value of 1.232622907. If entropy has a value of 0 in one of the universes, it indicates that the universe already has leaves.

2.1 Implementasi pada RapidMiner. In this research, using RapidMiner to validate as can be seen in Figure 3 and produce a decision tree as in Figure 4.

Implementation of the C4.5 algorithm in RapidMiner software. RapidMiner is a data analytics platform that is often used for data mining and machine learning processes. In this context, the figure illustrates the process of how data is input and processed using the C4.5 algorithm to produce a decision tree used in research. The first step in the implementation is to import the data into RapidMiner. This data is usually in CSV or Excel format and includes all relevant variables to be analyzed. This data includes attributes such as Semester IP, GPA, publications, collaborations, etc. which are used to classify the status of students into Permanent Lecturers, Contract Lecturers, or Not Fulfilling.

Based on the decision tree above to support decision making for the appointment of MTI Masters scholarship graduates who will become Lecturers at the Faculty of Information Technology in the form of text, is as follows:

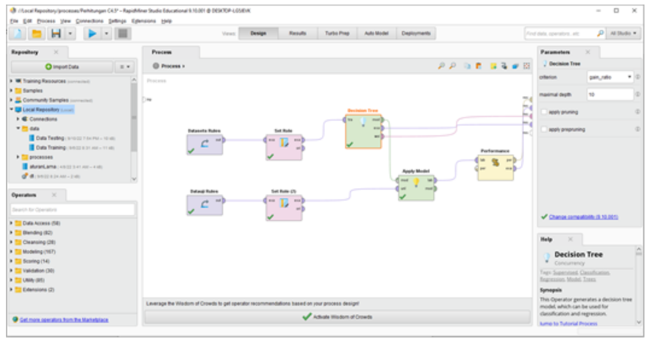

Fig. 3 Implementation on RapidMiner

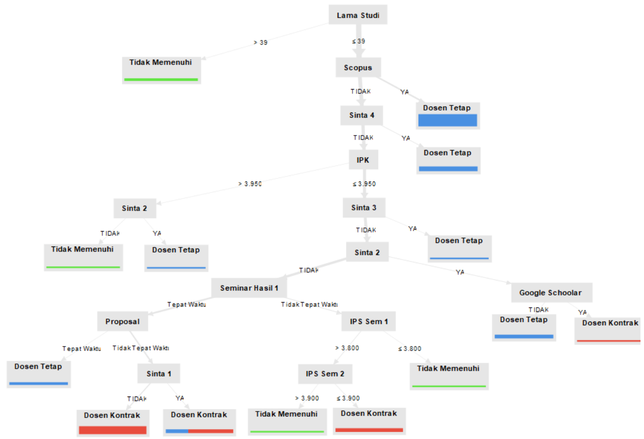

Fig. 4 RapidMiner Results Decision Tree

R1 : IF Length of Study > 39 THEN Not Qualified

R2 : IF Length of Study ≤ 39 AND Scopus = YES THEN Permanent Lecturer

R3 : IF Length of Study ≤ 39 AND Scopus = NO AND Sinta 4 = YES THEN Permanent Lecturer

R4 : IF Length of Study ≤ 39 AND Scopus = NO AND Sinta 4 = NO AND GPA ≤ 3.9 AND Sinta 3 YES THEN Permanent Lecturer

R5 : IF Length of Study ≤ 39 AND Scopus = NO AND Sinta 4 = NO AND GPA > 3.9 AND Sinta 2 YES THEN Permanent Lecturer

R6 : IF Length of Study ≤ 39 AND Scopus = NO AND Sinta 4 = NO AND GPA > 3.9 AND Sinta 2 NO THEN Not Qualified

R7 : IF Length of Study ≤ 39 AND Scopus = NO AND Sinta 4 = NO AND GPA ≤ 3.9 AND Sinta 3 NO AND Sinta 2 = YES AND Google Schoolar = YES THEN Contract Lecturer

R8 : IF Length of Study ≤ 39 AND Scopus = NO AND Sinta 4 = NO AND GPA ≤ 3.9 AND Sinta 3 NO AND Sinta 2 = YES AND Google Schoolar = NO THEN Permanent Lecturer

R9 : IF Length of Study ≤ 39 AND Scopus = NO AND Sinta 4 = NO AND GPA ≤ 3.9 AND Sinta 3 NO AND Sinta 2 = NO AND Seminar Result 1 = On Time AND Proposal : On Time THEN Permanent Lecturer

R10 : IF Length of Study ≤ 39 AND Scopus = NO AND Sinta 4 = NO AND GPA ≤ 3.9 AND Sinta 3 NO AND Sinta 2 = NO AND Seminar Result 1 = On Time AND Proposal : Not On Time AND Sinta 1 : YES THEN Permanent Lecturer/ Contract Lecturer

IP				Paper													Pratesis		Tesis			
IPS Sem 1	IPS Sem 2	IPS Sem 3	IPK	Sinta 1	Sinta 2	Sinta 3	Sinta 4	Sinta 5	Sinta 6	Garuda	Google Schoolar	Scopus	Nilai C	Kerjasama	Kedisiplinan	Komunikasi	Praproposal	Proposal	Seminar 1	Seminar 2	Lama Studi	Pakar
3.5	4	4	4	TIDAK	YA	TIDAK	TIDAK	TIDAK	TIDAK	TIDAK	TIDAK	TIDAK	Ada	Tidak Mampu	Baik	Baik	Tidak Tepat Waktu	Tidak Tepat Waktu	Tidak Tepat Waktu	Tidak Tepat Waktu	24	Dosen Tetap
3.7	4	3.5	3.8	TIDAK	TIDAK	TIDAK	TIDAK	TIDAK	YA	TIDAK	TIDAK	TIDAK	Ada	Tidak Mampu	Kurang Baik	Kurang Baik	Tidak Tepat Waktu	Tidak Tepat Waktu	Tepat Waktu	Tidak Tepat Waktu	36	Tidak Memenuhi
3.6	3.43	3	3.48	YA	TIDAK	TIDAK	TIDAK	TIDAK	YA	TIDAK	TIDAK	YA	Ada	Tidak Mampu	Baik	Kurang Baik	Tepat Waktu	Tidak Tepat Waktu	Tepat Waktu	Tidak Tepat Waktu	24	Dosen Kontrak
3.8	3.72	4	3.5	TIDAK	TIDAK	TIDAK	TIDAK	TIDAK	TIDAK	TIDAK	YA	TIDAK	Tidak Ada	Mampu	Baik	Kurang Baik	Tepat Waktu	Tidak Tepat Waktu	Tepat Waktu	Tidak Tepat Waktu	18	Dosen Kontrak
4	4	4	4	TIDAK	TIDAK	TIDAK	TIDAK	TIDAK	TIDAK	TIDAK	YA	TIDAK	Tidak Ada	Tidak Mampu	Kurang Baik	Kurang Baik	Tepat Waktu	Tepat Waktu	Tidak Tepat Waktu	Tidak Tepat Waktu	42	Tidak Memenuhi
3.7	3.7	3.2	3.5	YA	TIDAK	TIDAK	TIDAK	TIDAK	YA	TIDAK	TIDAK	YA	Tidak Ada	Tidak Mampu	Baik	Kurang Baik	Tidak Tepat Waktu	Tidak Tepat Waktu	Tepat Waktu	Tepat Waktu	18	Dosen Tetap
3	3.6	3.5	3.4	TIDAK	TIDAK	TIDAK	TIDAK	TIDAK	TIDAK	TIDAK	YA	TIDAK	Ada	Mampu	Baik	Kurang Baik	Tepat Waktu	Tidak Tepat Waktu	Tepat Waktu	Tidak Tepat Waktu	24	Dosen Kontrak
3.6	3.6	4	3.7	TIDAK	TIDAK	TIDAK	TIDAK	TIDAK	TIDAK	TIDAK	YA	TIDAK	Tidak Ada	Mampu	Kurang Baik	Baik	Tidak Tepat Waktu	Tidak Tepat Waktu	Tidak Tepat Waktu	Tidak Tepat Waktu	42	Tidak Memenuhi

Fig. 5 Test Data

R11 : IF Length of Study $\leq$ 39 AND Scopus = NO AND Sinta 4 = NO AND GPA $\leq$ 3.9 AND Sinta 3 NO AND Sinta 2 = NO AND Seminar Result 1 = On Time AND Proposal : Not On Time AND Sinta 1 : NO THEN Contract Lecturer

R12 : IF Length of Study $\leq$ 39 AND Scopus = NO AND Sinta 4 = NO AND GPA $\leq$ 3.9 AND Sinta 3 NO AND Sinta 2 = NO AND Seminar Result 1 = On Time AND Sinta 1 = NO THEN Contract Lecturer

R13 : IF Length of Study $\leq$ 39 AND Scopus = NO AND Sinta 4 = NO AND GPA $\leq$ 3.9 AND Sinta 3 NO AND Sinta 2 = NO AND Seminar Result 1 = Not On Time AND IPS Sem 1 $\leq$ 3.8 THEN Does Not Meet

R14 : IF Length of Study $\leq$ 39 AND Scopus = NO AND Sinta 4 = NO AND GPA $\leq$ 3.9 AND Sinta 3 NO AND Sinta 2 = NO AND Seminar Result 1 = Not On Time AND IPS Sem 1 > 3.8 AND IPS Sem 2 $\leq$ 3.9 THEN Contract Lecturer

R15 : IF Length of Study $\leq$39 AND Scopus = NO AND Sinta 4 = NO AND GPA $\leq$ 3.9 AND Sinta 3 NO AND Sinta 2 = NO AND Seminar Result 1 = Not On Time AND IPS Sem 1 > 3.8 AND IPS Sem 2 $\leq$ 3.9 THEN Contract Lecturer

R16 : IF Length of Study $\leq$ 39 AND Scopus = NO AND Sinta 4 = NO AND GPA $\leq$ 3.9 AND Sinta 3 NO AND Sinta 2 = NO AND Seminar Result 1 = Not On Time AND IPS Sem 1 > 3.8 AND IPS Sem 2 > 3.9 THEN Does Not Meet

2.2 Testing. The following in Figure 5, are 8 (eight) data used to test the decision tree rule results from 41 (forty one) training data in Table 1.

The prediction results using test data are as follows on Figure 6:

2.3 Evaluation. The prediction results using test data are as follows on Figure 6: Based on data processing using RapidMiner software, the system accuracy value is 75.00%, meaning that the rule generated using the rules from UTDI and the C4.5 algorithm in RapidMiner, the level of accuracy reaches 75%. Where the model that has been formed, the level of accuracy is tested by entering test data from 8 (eight) data into the RapidMiner 5.10 application to test the level of accuracy.

3 Conclusion

The application of the C4.5 algorithm in this study proved effective in supporting decision-making to determine the status of scholarship recipient students as Permanent Lecturers, Contract Lecturers, or ineligible at Universitas Teknologi Digital Indonesia (UTDI). This approach, using decision trees, offers a structured and easy-to-interpret method to classify candidates based on various criteria such as academic performance, scientific contributions, and other relevant factors. The model's accuracy of 75% highlights its potential reliability in making consistent and objective decisions.

a. Broader Applications and Future Research Opportunities The methodology is not limited to the current context and can be extended to other decision-making scenarios that require similar classification tasks. For example, the C4.5 algorithm can be applied in the hiring process of employees in various sectors, where candidates need to be evaluated based on various criteria such as skills, experience, and cultural fit. It can also be used in educational settings for admission decision-making, where various academic and extracurricular factors need to be considered.
b. Opportunities for Future Research Although the current model shows a reasonable level of accuracy, future research could explore several areas for improvement. One potential direction is to investigate the integration of other techniques to improve the accuracy of the model.

References

[1] Widyastuti, A., 2020, "Kriteria dan Parameter Penilaian Mahasiswa Beasiswa S2 UTDI," Interview.

[2] Rismayana, A. H. and Rosdiana, D., 2019, "Penerapan Algoritma C4. 5 Pada Bidang Pertanian," Jurnal TEDC, **13**(3), pp. 233–238, [Daring]. Tersedia pada: http://ejournal.poltektedc.ac.id/index.php/tedc/article/view/307.

[3] Devarapalli, D., Apparao, A., Kumar, A., and Sridhar, G. R., 2013, "A Novel Analysis of Diabetes Mellitus by Using Expert System Based on Brain Derived Neurotrophic Factor (BDNF) Levels," International Journal, **1**(January), pp. 251–256.

[4] Abu-Nasser, B. and Abu-Naser, S., 2018, "Rule-Based System for Watermelon Diseases and Treatment," International Journal of Academic Information Systems Research, **2**(7), pp. 1–7, [Daring]. Tersedia pada: https://hal.archives-ouvertes.fr/hal-01855441.

[5] Henderi, Aini, Q., Srenggini, A. D., and Khoirunisa, A., 2020, "Rule based expert system for supporting assessment of learning outcomes," International Journal of Advanced Trends in Computer Science and Engineering, **9**(1.2 Special Issue), pp. 266–271.

Row No.	Pakar	prediction(Pakar)	confidence(Dosen Tetap)	confidence(Tidak Memenuhi)	confidence(Dosen Kontrak)
1	Dosen Tetap	Dosen Tetap	1	0	0
2	Tidak Memen...	Dosen Kontrak	0	0	1
3	Dosen Kontrak	Dosen Tetap	1	0	0
4	Dosen Kontrak	Dosen Kontrak	0	0	1
5	Tidak Memen...	Tidak Memenuhi	0	1	0
6	Dosen Tetap	Dosen Tetap	1	0	0
7	Dosen Kontrak	Dosen Kontrak	0	0	1
8	Tidak Memen...	Tidak Memenuhi	0	1	0

Fig. 6 Prediction Results

accuracy: 75.00%

	true Dosen Tetap	true Tidak Memenuhi	true Dosen Kontrak	class precision
pred. Dosen Tetap	2	0	1	66.67%
pred. Tidak Memenuhi	0	2	0	100.00%
pred. Dosen Kontrak	0	1	2	66.67%
class recall	100.00%	66.67%	66.67%	

Fig. 7 Confusion Matrix

[6] Imamoğlu, M. Y. and Çetinkaya, D., 2017, "A rule based decision support system for programming language selection," *2017 2nd International Conference on Knowledge Engineering and Applications, ICKEA 2017*, Vol. 2017-Janua, pp. 71–75, doi: 10.1109/ICKEA.2017.8169904.

[7] Grosan, C. and Abraham, A., 2011, "Rule-Based Expert Systems," *Intelligent Systems Reference Library*, Vol. 17, pp. 149–185.

[8] Idris, M., Mustafid, M., and Suseno, J. E., 2019, "Implementation of C4.5 Algorithm and Forward Chaining Method for Higher Education Performance Analysis," *E3S Web of Conferences*, Vol. 125, Paper No. 2019, pp. 2–6.

[9] Mikulić, I., Lisjak, D., and Štefanić, N., 2021, "A rule-based system for human performance evaluation: A case study," Applied Sciences (Switzerland), **11**(7), pp. 1–19.

[10] Buulolo, E., Medan, K., and Utara, S., 2017, "C4 . 5 Algorithm to Predict the Impact of the Earthquake," International Journal of Engineering Research & Technology (IJERT), **6**(02), pp. 10–15.

[11] Muttaqien, R., Pradana, M. G., and Pramuntadi, A., 2021, "Implementation of Data Mining Using C4.5 Algorithm for Predicting Customer Loyalty of PT. Pegadaian (Persero) Pati Area Office," International Journal of Computer and Information System (IJCIS), **2**(3), pp. 64–68.

[12] Siahaan, H., Mawengkang, H., Efendi, S., Wanto, A., and Windarto, A. P., 2019, "Application of Classification Method C4.5 on Selection of Exemplary Teachers," *Journal of Physics: Conference Series*, Vol. 1235, Paper No. 1.

[13] Rahim, R. et al., 2018, "C4.5 classification data mining for inventory control," International Journal of Engineering and Technology(UAE), **7**, pp. 68–72.

[14] Pah, C. E. A. and Utama, D. N., 2020, "Decision support model for employee recruitment using data mining classification," International Journal of Emerging Trends in Engineering Research, **8**(5), pp. 1511–1516.

[15] Rabcan, J., Vaclavkova, M., and Blasko, R., 2017, "Selection of appropriate candidates for a type position using C4.5 decision tree," *Proceedings of the International Conference on Information and Digital Technologies, IDT 2017*, pp. 332–338, doi: 10.1109/DT.2017.8024318.

[16] Ningse, W. R. S. O., Sumarno, S., and Nasution, Z. M., 2022, "C4.5 Algorithm Classification for Determining Smart Indonesia Program Recipients at MIS Al-Khoirot," JOMLAI: Journal of Machine Learning and Artificial Intelligence, **1**(1), pp. 65–76.

[17] Ruggieri, S., 2002, "Efficient C4.5," IEEE Transactions on Knowledge and Data Engineering, **14**(2), pp. 438–444.

[18] Harryanto, F. F. and Hansun, S., 2017, "Penerapan Algoritma C4.5 untuk Memprediksi Penerimaan Calon Pegawai Baru di PT WISE," Maret, **3**(2), p. 95.

[19] Lakshmi, T. M., Martin, A., Begum, R. M., and Venkatesan, V. P., 2013, "An Analysis on Performance of Decision Tree Algorithms using Student's Qualitative Data," International Journal of Modern Education and Computer Science, **5**(5), pp. 18–27.

[20] Setio, P. B. N., Saputro, D. R. S., and Winarno, B., 2020, "PRISMA, Prosiding Seminar Nasional Matematika Klasifikasi dengan Pohon Keputusan Berbasis Algoritme C4.5," *PRISMA, Prosiding Seminar Nasional Matematika*, Vol. 3, pp. 64–71, [Daring]. Tersedia pada: https://journal.unnes.ac.id/sju/index.php/prisma/.

[21] Efendi, A. and Hartanto, A. D., 2020, "Implementation Of The C4.5 Algorithm For Recruitment Of E-Sports Team Members," CCIT Journal, **13**(2), pp. 138–146.

[22] Asidik, I., Kusrini, and Henderi, 2018, "Decision Support System Model of Teacher Recruitment Using Algorithm C4.5 and Fuzzy Tahani," *Journal of Physics: Conference Series*, Vol. 1140, Paper No. 1.

[23] Agarwal, S., 2014, *Data mining: Data mining concepts and techniques*.

[24] LLDIKTI5, 2020, *Buku Standar Pelayanan Publik LLDIKTI Wilayah V*.

[25] Mansurdin and Yurnetti, 1999, *Penelitian Tindakan (Action Research) Dan Aplikasinya Di Lembaga Pendidikan Tenaga Kependidikan (LPTK)*, Vol. 1.

[26] Mulyatiningsih, E., 2011, *Riset Terapan Bidang Pendidikan dan Teknik*.

Table 3 Results of Iteration 1 Calculation (Node 1)

Node			Number of Cases (S)	Permanent Lecturer (S1)	Contract Lecturer (S2)	Not Fulfilled (S3)	Entropy	Gain	Split Info	Gain Ratio
1	TOTAL		41	23	13	5	1.363471961			
	IPS Sem 1							0.037	0.378	0.097
		≤3.10	3	1	2	0	0.918			
		>3.10	38	22	11	5	1.359			
								0.030	0.601	0.050
		≤3.30	6	4	2	0	0.918			
		>3.30	35	19	11	5	1.404			
								0.057	0.801	0.071
		≤3.55	10	7	3	0	0.881			
		>3.55	31	16	10	5	1.444			
								0.028	0.989	0.029
		≤3.65	18	12	4	2	1.224			
		>3.65	23	11	9	3	1.422			
								0.022	1.000	0.022
		≤3.75	20	13	5	2	1.236			
		>3.75	21	10	8	3	1.441			
								0.093	0.901	0.103
		≤3.9	28	19	7	2	1.152			
		>3.9	13	4	6	3	1.526			
	IPS Sem 2							0.021	0.165	0.125
		≤3.10	1	1	0	0	0.000			
		>3.10	40	22	13	5	1.376			
								0.026	0.461	0.056
		≤3.25	4	2	2	0	1.000			
		>3.25	37	21	11	5	1.374			
								0.025	0.535	0.047
		≤3.35	5	3	2	0	0.971			
		>3.35	36	20	11	5	1.389			
								0.008	0.901	0.009
		≤3.55	13	8	4	1	1.239			
		>3.55	28	15	9	4	1.410			
								0.002	0.996	0.002
		≤3.65	19	11	6	2	1.324			
		>3.65	22	12	7	3	1.395			
								0.013	0.989	0.014
		≤3.75	23	14	7	2	1.265			
		>3.75	18	9	6	3	1.459			
								0.038	0.926	0.041
		≤3.85	27	15	10	2	1.280			
		>3.85	14	8	3	3	1.414			
								0.038	0.901	0.043
		≤3.95	28	16	10	2	1.264			
		>3.95	13	7	3	3	1.457			
	IPS Sem 3							0.030	0.601	0.050
		≤3.10	6	4	2	0	0.918			
		>3.10	35	19	11	5	1.404			
								0.024	0.901	0.027
		≤3.25	13	9	3	1	1.140			
		>3.25	28	14	10	4	1.432			
								0.014	0.926	0.015
		≤3.40	14	9	4	1	1.198			
		>3.40	27	14	9	4	1.428			
								0.021	0.979	0.021
		≤3.60	17	10	6	1	1.221			
		>3.60	24	13	7	4	1.428			
								0.047	1.000	0.047
		≤3.75	20	11	8	1	1.219			
		>3.75	21	12	5	4	1.410			
								0.055	0.989	0.055
		≤3.90	23	14	8	1	1.163			
		>3.90	18	9	5	4	1.496			

...continue to next page

...continue Table 3

Node		Number of Cases (S)	Permanent Lecturer (S1)	Contract Lecturer (S2)	Not Fulfilled (S3)	Entropy	Gain	Split Info	Gain Ratio
IPK							0.085	0.281	0.302
	≤3.25	2	0	2	0	0.000			
	>3.25	39	23	11	5	1.344			
							0.037	0.378	0.097
	≤3.35	3	1	2	0	0.918			
	>3.35	38	22	11	5	1.359			
							0.041	0.712	0.058
	≤3.45	8	5	3	0	0.954			
	>3.45	33	18	10	5	1.411			
							0.100	0.901	0.111
	≤3.55	13	10	3	0	0.779			
	>3.55	28	13	10	5	1.488			
							0.105	0.947	0.111
	≤3.65	15	11	4	0	0.837			
	>3.65	26	12	9	5	1.502			
							0.008	0.965	0.008
	≤3.75	25	15	7	3	1.323			
	>3.75	16	8	6	2	1.406			
							0.090	0.601	0.150
	≤3.85	35	22	10	3	1.241			
	>3.85	6	1	3	2	1.459			
							0.114	0.378	0.303
	≤3.95	38	22	13	3	1.275			
	>3.95	3	1	0	2	0.918			
Sinta-1							0.101	0.839	0.120
	YA	11	9	2	0	0.684			
	TIDAK	30	14	11	5	1.475			
Sinta-2							0.037	0.535	0.069
	YA	5	4	1	0	0.722			
	TIDAK	36	19	12	5	1.410			
Sinta-3							0.111	0.535	0.208
	YA	5	5	0	0	0.000			
	TIDAK	36	18	13	5	1.426			
Sinta-4							0.087	0.461	0.189
	YA	4	4	0	0	0.000			
	TIDAK	37	19	13	5	1.414			
Sinta-5							0.087	0.461	0.189
	YA	4	4	0	0	0.000			
	TIDAK	37	19	13	5	1.414			
Sinta-6							0.041	0.712	0.058
	YA	8	5	3	0	0.954			
	TIDAK	33	18	10	5	1.411			
Garuda							0.030	0.601	0.050
	YA	6	4	2	0	0.918			
	TIDAK	35	19	11	5	1.404			
Google Schoolar							0.007	0.872	0.008
	YA	12	6	4	2	1.459			
	TIDAK	29	17	9	3	1.314			
Scopus							0.279	0.839	0.332
	YA	11	11	0	0	0.000			
	TIDAK	30	12	13	5	1.482			
Nilai C							0.007	0.801	0.009
	Ada	10	5	4	1	1.361			
	Tidak Ada	31	18	9	4	1.355			

...continue to next page

...continue Table 3

Node			Number of Cases (S)	Permanent Lecturer (S1)	Contract Lecturer (S2)	Not Fulfilled (S3)	Entropy	Gain	Split Info	Gain Ratio
	Kerjasama							0.025	0.926	0.027
		Mampu	27	16	7	4	1.360			
		Tidak Mampu	14	7	6	1	1.296			
	Kedisiplinan							0.219	0.947	0.231
		Baik	26	18	8	0	0.890			
		Kurang Baik	15	5	5	5	1.585			
	Komunikasi							0.065	1.000	0.065
		Baik	20	14	5	1	1.076			
		Kurang Baik	21	9	8	4	1.510			
	Praproposal							0.051	0.996	0.051
		Tepat Waktu	22	14	7	1	1.143			
		Tidak Tepat Waktu	19	9	6	4	1.509			
	Proposal							0.020	0.801	0.025
		Tepat Waktu	10	7	2	1	1.157			
		Tidak Tepat Waktu	31	16	11	4	1.404			
	Ujian SH-1							0.276	0.872	0.317
		Tepat Waktu	29	20	9	0	0.894			
		Tidak Tepat Waktu	12	3	4	5	1.555			
	Ujian SH-2							0.190	0.979	0.194
		Tepat Waktu	17	14	3	0	0.672			
		Tidak Tepat Waktu	24	9	10	5	1.528			
	Lama Studi							0.119	0.926	0.128
		≤21	14	11	3	0	0.750			
		>21	27	12	10	5	1.501			
							0.167	0.712	0.234	
	≤30	33	22	9	2	1.146				
	>30	8	1	4	3	1.406				
							0.163	0.281	0.579	
	≤39	39	23	13	3	1.262	? 1.1.A			
	>39	2	0	0	2	0.000				

Table 4 Iteration 13 Calculation Results (Node 1.8 B)

Node			Number of Cases (S)	Permanent Lecturer (S1)	Contract Lecturer (S2)	Not Fulfilled (S3)	Entropy	Gain	Split Info	Gain Ratio
1.8.B	IPS Sem 1	>3.8	4	0	3	1	0.811			
	IPS Sem 2							0.311	0.811	0.384
		≤3.30	1	0	1	0	0.000			
		>3.30	3	0	2	1	0.918			
								0.311	0.811	0.384
		≤3.5	1	0	1	0	0.000			
		>3.5	3	0	2	1	0.918			
								0.500	1.000	0.500
		≤3.7	2	0	2	0	0.000			
		>3.7	2	0	1	1	1.000			
								1.000	0.811	1.233
		≤3.9	3	0	3	0	0.000			
		>3.9	1	0	0	1	0.000			
	IPS Sem 3							1.000	0.811	1.233
		≤3.60	1	0	0	1	0.000			
		>3.60	3	0	3	0	0.000			
	IPK							0.500	1.000	0.500
		≤3.8	2	0	1	1	1.000			
		>3.8	2	0	2	0	0.000			
	Sinta-1							0.189	0.000	0.000
		YA	0	0	0	0	0.000			
		TIDAK	4	0	3	1	0.811			
	Sinta-5							0.189	0.000	0.000
		YA	0	0	0	0	0.000			
		TIDAK	4	0	3	1	0.811			
	Sinta-6							0.311	0.811	0.384
		YA	1	0	1	0	0.000			
		TIDAK	3	0	2	1	0.918			
	Garuda							0.750	1.000	0.750
		YA	2	0	2	0	0.000			
		TIDAK	2	0	1	2	0.500			
	Google Schoolar							0.189	0.000	0.000
		YA	0	0	0	0	0.000			
		TIDAK	4	0	3	1	0.811			
	Nilai C							0.500	1.000	0.500
		Ada	2	0	1	1	1.000			
		Tidak Ada	2	0	2	0	0.000			
	Kerjasama							0.189	0.000	0.000
		Mampu	4	0	3	1	0.811			
		Tidak Mampu	0	0	0	0	0.000			
	Kedisiplinan							0.311	0.811	0.384
		Baik	1	0	1	0	0.000			
		Kurang Baik	3	0	2	1	0.918			
	Komunikasi							0.311	0.811	0.384
		Baik	1	0	1	0	0.000			
		Kurang Baik	3	0	2	1	0.918			

...continue to next page

...continue Table 4

Node		Number of Cases (S)	Permanent Lecturer (S1)	Contract Lecturer (S2)	Not Fulfilled (S3)	Entropy	Gain	Split Info	Gain Ratio
Praproposal							1.000	0.811	1.233
	Tepat Waktu	1	0	0	0	0.000			
	Tidak Tepat Waktu	3	0	0	0	0.000			
Proposal							0.311	0.811	0.384
	Tepat Waktu	1	0	1	0	0.000			
	Tidak Tepat Waktu	3	0	2	1	0.918			
Ujian SH-2							0.189	0.000	0.000
	Tepat Waktu	0	0	0	0	0.000			
	Tidak Tepat Waktu	4	0	3	1	0.811			
Lama Studi							0.189	0.000	0.000
	≤21	0	0	0	0	0.000			
	>21	4	0	3	1	0.811			
							0.311	0.811	0.384
	≤30	3	0	2	1	0.918			
	>30	1	0	1	0	0.000			