Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 6, Number 1, January 2022
DOI : https://doi.org/10.33395/sinkron.v7i1.112595
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Implementation of C5.0 Algorithm for Prediction of Student Learning Graduation in Computer System Architecture Subjects

Nurfadillah Tanjung[1)*], Deci Irmayani[2)], Volvo Sihombing[3)]
[1)2)3)]Labuhanbatu University, North Sumatra, Indonesia
[1)] nurfadillanjung758@gmail.com, [2)] deacyirmayani@gmail.com, [3)] volvolumbantoruan@gmail.com

Submitted: Jan 15, 2022 | **Accepted** : Jan 28, 2022 | **Published** : Feb 4, 2022

Abstract: Computer system architecture is one of the subjects that must be taken in the informatics engineering study program. In the study program the graduation of each student in the course is one of the important aspects that must be evaluated every semester. Graduation for each student / I in the course is an illustration that the learning process delivered is going well and also the material presented by the lecturer in charge of the course can be digested by students. Graduation of each student in the course can be predicted based on the habit pattern of the students. Data mining is an alternative process that can be done to find out habit patterns based on the data that has been collected. Data mining itself is an extraction process on a collection of data that produces valuable information for companies, agencies or organizations that can be used in the decision-making process. Prediction of graduation with data mining can be solved by classifying the data set. The C5.0 algorithm is an improvement algorithm from the C4.5 algorithm where the process is almost the same, only the C5.0 algorithm has advantages over the previous algorithm. The results of the C5.0 algorithm are in the form of a decision tree or a rule that is formed based on the entropy or gain value. The prediction process is carried out based on the C5.0 algorithm classification using the attributes of Attendance Value, Assignment Value, UTS Value and UAS Value. The final result of the C5.0 algorithm classification process is a decision tree with rules in it.

Keywords: Data Mining; Prediction, Graduation; Computer System Architecture; C5.0 . Algorithm

INTRODUCTION

Computer system architecture is one of the subjects that must be taken in the informatics engineering study program. In the computer system architecture course, the material discussed in the learning process is related to the organization contained in the computer, both the structure and function of each component, hardware and software as well as the work of the Input Output and Processes on the computer. Computer system architecture is a subject that must be completed by every student. If the computer system architecture course is not completed (failed) then it is not permitted for students to take other related courses and the continuation of the computer system architecture course.

In the study program the graduation of each student in the course is one of the important aspects that must be evaluated every semester. Graduation for each student / I in the course is an illustration that the learning process delivered is going well and also the material presented by the lecturer in charge of the course can be digested by students. In addition, student graduation in the course will meet the achievement indicators that have been determined by the study program. Where student graduation in the course will improve the quality or quality of the study program.

Graduation of each student in the course can be predicted based on the habit pattern of the students. Where this pattern can be seen in the data of students who took computer systems architecture courses before. This pattern can be seen based on attendance scores, assignment scores, UTS scores and UAS scores. The application of information technology in education can also produce abundant data regarding student data and the resulting learning value.

Of course, the study program has a lot of data related to student graduation data in the computer system architecture course. The large amount of data that has been collected should be able to process to find information

* Nurfadillah Tanjung

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 6, Number 1, January 2022
DOI : https://doi.org/10.33395/sinkron.v7i1.112595
e-ISSN : 2541-2019
p-ISSN : 2541-044X

stored in the data set. However, the problem is that there is still no extracting this information for study programs or course lecturers to predict student graduation in computer systems architecture courses. This is because there is no process that will be carried out to carry out the data mining process.

Data mining is an alternative process that can be done to solve these problems. By applying the data mining process, it will process all data that has been stored in the study program to get new information(Wu et al., 2014). Data mining itself is an extraction process on a collection of data that produces valuable information for companies, agencies or organizations that can be used in the decision-making process.(Witten et al., 2016)

Prediction of graduation with data mining can be solved by classifying the data set. Classification is part of data mining techniques where the process is carried out based on data grouping(Fricles A Sianturi, Hasugian, Paska Marto, Simangunsong Agustina, 2019). The grouping of data in data mining techniques is based on the relationship of the data to the data sample that has a label or target class. One of the algorithms that can be used in data mining classification is the C5.0 . algorithm(Febrivani & Winanjaya, 2021).

The C5.0 algorithm is an improvement algorithm from the C4.5 algorithm where the process is almost the same, only the C5.0 algorithm has advantages over the previous algorithm. The C5.0 algorithm is part of the classification technique in data mining that produces information in the form of a decision tree or rule. The formation of a decision tree based on the resulting nodes. The selection of nodes in the decision tree of the algorithm is chosen from the gain value and also entropy(Sowmya & Suneetha, 2017)

The C5.0 algorithm is widely used in research to discuss classification(Kurniawan, 2018). Compared to other classification algorithms, the C5.0 algorithm has better performance, besides that the performance results obtained from the C5.0 algorithm classification process also have a high level of accuracy.(With et al., 2018)

In this study, we will predict student graduation in the computer system architecture course by applying the data mining process and the C5.0 algorithm for the prediction process based on the classification concept that will be carried out.

LITERATURE REVIEW

Data mining is an attempt to extract valuable and useful information in very large databases. Data Mining is also referred to as Knowledge Discovery in Database (KDD) which can be interpreted as the extraction of unknown implicit potential information from a set of data.(Suyanto, 2017). The KDD process involves the results of the data mining process or the process of extracting the tendency of a data pattern, then the results are converted appropriately into information something that is easier to understand.(Sowmya & Suneetha, 2017)

The C5.0 algorithm is(Cynthia & Ismanto, 2018)one of the algorithms which is a refinement of the C4.5 algorithm which uses a tree-shaped representation where each node represents an attribute then the branch represents the value of the attribute and has a leaf where the function is a class. Decision making is based on the largest Gain value from the calculation of all attributes. Here are the steps for using the C5.0 algorithm (Cynthia & Ismanto, 2018)

1. Make a decision system that includes condition attributes and decision attributes. Then describe a decision system consisting of only n objects
2. Counting the number of column data, where the amount of data must be based on certain attribute members whose results are based on certain conditions.
3. Select the attribute that is used as the Node.
4. Create a branch for each Node member.
5. Checks whether the entropy value of each Node member has a value of zero. If the value is 0, then determine the leaf that has been formed. If the entropy value of each Node member is entirely zero, then the process stops.

Node members that have a value greater than zero, then the process is repeated from the beginning with the condition that all members of the Node are zero. Node is an attribute that has the highest gain value of the existing attributes. The process of calculating the gain value of an attribute must use a formula. The following is the formula for calculating the gain value used in the C4.5 . algorithm(Cynthia & Ismanto, 2018):

$$Gain\ (S, A) = Entropy(S) - \sum_{i=1}^{m} \frac{|S_i|}{|S|}\ x\ Entropy\ (S_i)\ (1)$$

In addition, the formula for calculating the entropy value is as follows:

$$Entropy = -\sum_{j=1}^{k} p_j log_2\, p_j\ (2)$$

METHOD

The research methodology describes the stages carried out in the research. The research methodology consists of several stages that are systematically related. This stage is needed to make it easier to conduct research. Before making a methodology, the author first analyzes the topic to be studied. In the research analysis, the author explains

* Nurfadillah Tanjung

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 6, Number 1, January 2022
DOI : https://doi.org/10.33395/sinkron.v7i1.112595
e-ISSN : 2541-2019
p-ISSN : 2541-044X

how the author's process in collecting the data needed for this research. The data collection method is done by using secondary data that is already available based on student scores in the computer system architecture course in the past few years that has been stored in the study program. Overall the research methodology carried out can be seen in the following figure:

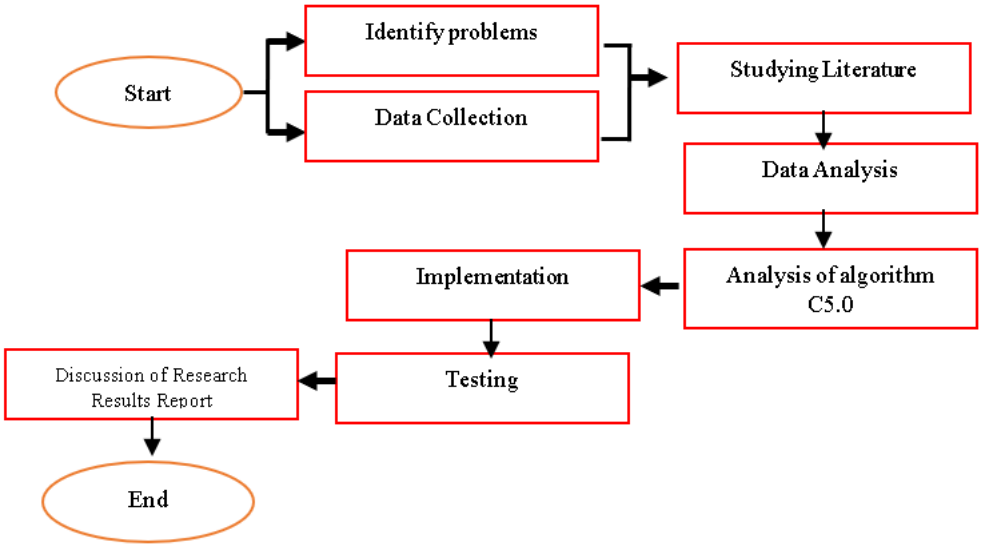

Fig.1 Research Stages

RESULT

The data sample that will be used to predict student graduation in computer architecture courses is based on student score data that is already available in the study program. Predictions will later be grouped into several criteria on their attributes. The following is the criteria data that will be used for the prediction process of student graduation in the computer system architecture course:

Table 1
Data Sample

No	Respondent	Presence	Task	UTS	UAS	Results
1	R1	100	75	80	85	Graduated
2	R2	100	70	85	80	Graduated
3	R3	93	65	75	50	Graduated
4	R4	86	60	70	50	Graduated
5	R5	100	80	70	60	Graduated
6	R6	79	90	65	50	Graduated
7	R7	71	95	60	65	Graduated
8	R8	93	85	70	75	Graduated
9	R9	64	60	60	0	Not pass
10	R10	57	50	50	0	Not pass
11	R11	79	70	50	60	Graduated
12	R12	71	65	50	55	Not pass
13	R13	50	60	60	0	Not pass
14	R14	71	75	80	50	Graduated
15	R15	50	50	60	0	Not pass

In table 1 is the value data from students who take computer systems architecture courses. Then from the data before the classification process using the C5.0 algorithm must be carried out preprocessing the data first. Data preprocessing is done so that the data used in the research is suitable for the process to be carried out on the algorithm. The data that has been collected must be changed with the preprocessing stage. The data results from the preprocessing stage are as follows:

* Nurfadillah Tanjung

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 6, Number 1, January 2022
DOI : https://doi.org/10.33395/sinkron.v7i1.112595
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Table 2
Data Preprocessing

No	Respondent	Presence	Task	UTS	UAS	Results
1	R1	Tall	Tall	Tall	Tall	Graduated
2	R2	Tall	Tall	Tall	Tall	Graduated
3	R3	Tall	Tall	Tall	Low	Graduated
4	R4	Tall	Low	Tall	Low	Graduated
5	R5	Tall	Tall	Tall	Low	Graduated
6	R6	Tall	Tall	Tall	Low	Graduated
7	R7	Tall	Tall	Low	Tall	Graduated
8	R8	Tall	Tall	Tall	Tall	Graduated
9	R9	Low	Low	Low	Low	Not pass
10	R10	Low	Low	Low	Low	Not pass
11	R11	Tall	Tall	Low	Low	Graduated
12	R12	Tall	Tall	Low	Low	Not pass
13	R13	Low	Low	Low	Low	Not pass
14	R14	Tall	Tall	Tall	Low	Graduated
15	R15	Low	Low	Low	Low	Not pass

Table 2 is data that has been preprocessed data. Then from the data in table 2, the classification process can be carried out using the C5.0 algorithm. Then calculate the number of cases, the number of cases for the Pass results, the number of cases for the Disqualified results, and the entropy of all cases and the cases are divided based on the attributes of the Attendance Variable, Assignment Variable, UTS Variable, and UAS Variable. The next step is to calculate the gain for each attribute.

Table 3
Node 1 . Calculation

Node	Attribute	Mark	Number of Cases / Data Set	Graduated	Not pass	Entropy	gain
1			15	10	5	0.9183	
	Presence						0.596
		Tall	11	10	1	0.4395	
		Low	4	0	4	0	
	Task						0.36499
		Tall	10	9	1	0.469	
		Low	5	1	4	0.72193	
	UTS						0.51551
		Tall	8	8	0	0	
		Low	7	2	5	0.86312	
	UAS						0.18934
		Tall	4	4	0	0	
		Low	11	6	5	0.99403	

Based on the results of the calculations above, it can be seen that Presence has a higher gain compared to other gains, namely 0.596. Therefore, Presence is used as the root of the decision tree. In the presence attribute which is the root of the decision tree, there are 2 attribute values, namely High and Low. For the low attribute value, the result of the decision is Not Passed. While on the High attribute there are still 2 decisions, namely Pass and Fail, the process of forming the decision tree is still continued based on the following data.

Table 4
Node Calculation Data 1.1

No	Respondent	Presence	Task	UTS	UAS	Results
1	R1	Tall	Tall	Tall	Tall	Graduated
2	R2	Tall	Tall	Tall	Tall	Graduated
3	R3	Tall	Tall	Tall	Low	Graduated
4	R4	Tall	Low	Tall	Low	Graduated

* Nurfadillah Tanjung

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 6, Number 1, January 2022
DOI : https://doi.org/10.33395/sinkron.v7i1.112595
e-ISSN : 2541-2019
p-ISSN : 2541-044X

5	R5	Tall	Tall	Tall	Low	Graduated
6	R6	Tall	Tall	Tall	Low	Graduated
7	R7	Tall	Tall	Low	Tall	Graduated
8	R8	Tall	Tall	Tall	Tall	Graduated
11	R11	Tall	Tall	Low	Low	Graduated
12	R12	Tall	Tall	Low	Low	Not Graduated
14	R14	Tall	Tall	Tall	Low	Graduated

Then from the data in table 4 for the calculation of node 1.1, the same process is carried out as the root search (node 1). By finding back the value of entropy and gain. Where later the attribute with the highest gain value will be the next root.

Table 5
Node Calculation 1.1

Attribute Node	Mark	Number of Cases / Data Set	Graduated	Not pass	Entropy	gain
1.1		11	10	1	0.4395	
Task						0.16379
	Tall	10	9	1	0.469	
	Low	5	1	4	0.72193	
UTS						0.66785
	Tall	8	8	0	0	
	Low	3	2	1	0.9183	
UAS						0.54178
	Tall	4	4	0	0	
	Low	7	6	1	0.59167	

Based on the results of the calculations above, it can be seen that UTS has a higher gain value than the other gains, namely 0.66785. Therefore, UTS is used as the root node 1.1 of the decision tree. In the UTS attribute which is the root of the decision tree node 1.1, there are 2 attribute values, namely High and Low. For the High attribute value, the decision result is Passed. While on the Low attribute there are still 2 decisions, namely Pass and Fail, the process of forming the decision tree is still continued based on the following data:

Table 6
Node Calculation Data 1.1.1

No	Respondent	Presence	Task	UTS	UAS	Results
7	R7	Tall	Tall	Low	Tall	Graduated
11	R11	Tall	Tall	Low	Low	Graduated
12	R12	Tall	Tall	Low	Low	Not Graduated

Then from the data in table 6 for the calculation of node 1.1.1, the same process is carried out as the root search (node 1.1). By finding back the value of entropy and gain. Where later the attribute with the highest gain value will be the next root.

Table 7
Node Calculation 1.1.1

Attribute Node	Mark	Number of Cases / Data Set	Graduated	Not pass	Entropy	gain
1.1.1		3	2	1	0.9183	
Task						0
	Tall	3	2	1	0.9183	
	Low	0	0	0	0	
UAS						0.58496

* Nurfadillah Tanjung

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 6, Number 1, January 2022
DOI : https://doi.org/10.33395/sinkron.v7i1.112595
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Tall	1	1	0	0
Low	2	2	1	0.5

DISCUSSIONS

In the node calculation table 1.1.1, the UAS attribute is obtained with the highest gain value of 0.58496. Then you can see the final decision tree in the classification process in the following figure:

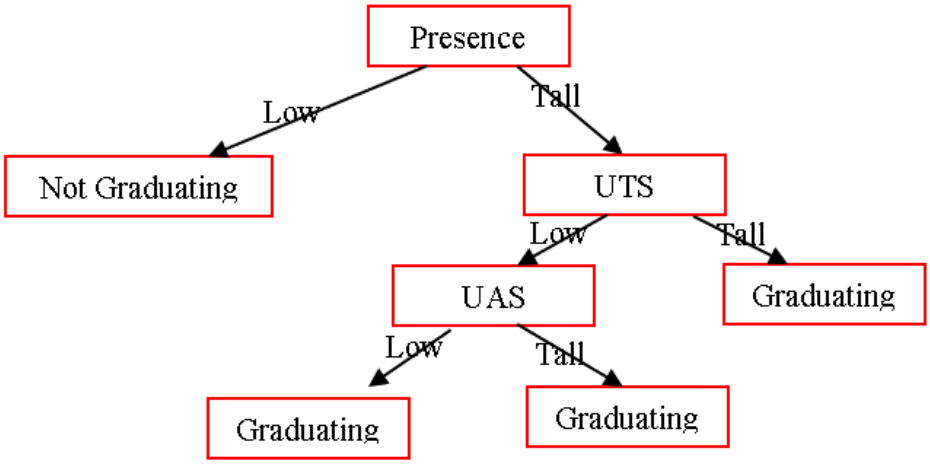

Fig.2 Final Decision Tree

Then after the decision tree is obtained in the classification process, the next step is to test the results obtained to measure the level of accuracy obtained

Table 8
Classification Performance Accuracy Results

Accuracy : 93.33%

	True Pass	True Not Pass	Precision Class
Pred. Graduated	10	1	90.91%
Pred. Not pass	0	4	100%
Class Recall	100%	80%	

CONCLUSION

Based on the research conducted, the results show that the C5.0 algorithm can be used to predict student graduation in computer systems architecture courses. The prediction process is carried out based on the C5.0 algorithm classification using the attributes of Attendance Value, Assignment Value, UTS Value and UAS Value. The final result of the C5.0 algorithm classification process is a decision tree with rules in it. The performance of the C5.0 algorithm gets a high accuracy rate of 93.33%.

REFERENCES

Cynthia, EP, & Ismanto, E. (2018). Decision Tree Algorithm Method C.45 in Classifying Sales Data for Fast Food Outlets Business. Jurassic (Journal of Information Systems Research and Informatics Engineering), 3(July), 1. https://doi.org/10.30645/jurasik.v3i0.60

With, S., Algorithm, C., & Hidayati, W. (2018). Data Mining Determination of Nurses at Sultan Hospital. 1(2), 1–7.

Febrivani, E., & Winanjaya, R. (2021). Application of Association Data Mining on Drug Inventory. 3(3), 354–365.

Fricles A Sianturi, Hasugian, Paska Marto, Simangunsong Agustina, NB (2019). Data Mining |Weka Theory and Applications. In -: Vol. (Issue).

Kurniawan, YI (2018). Comparison of Naive Bayes Algorithm and C.45 in Data Mining Classification. Journal of Information Technology and Computer Science. https://doi.org/10.25126/jtiik.201854803

* Nurfadillah Tanjung

KP Wirdhaningsih, M. Ratnawati, Dian Eka, UB Malang, D. Mining, and D. Tree, Application of Decision Tree C5.0 Algorithm for Forex Forecasting, 2013th ed. Malang: Brawijaya University Malang, 2013.

LR Haidar, E. Sediyono, and A. Iriani, "Analysis of Drop Out Student Predictions Using the Decision Tree Method with ID3 and C4.5 Algorithms," J. Transform., vol. 17, no. 2, p. 97, 2020, doi:10.26623/transformatika.v17i2.1609.

M. Kamil and W. Cholil, "Comparative Analysis of the C4.5 and Naive Bayes Algorithms on Timely Graduates of Students at Raden Fatah State Islamic University Palembang," J. Inform., vol. 7, no. 2, pp. 97–106, 2020, doi:10.31294/ji.v7i2.7723.

Mardi, Y. (2017). Data Mining :Classification Using the C4.5 Algorithm. Journal of Informatics Education.

Mashlahah, Prediction of Student Graduation Using the Decision Tree Method With the Application of the C4.5 Algorithm, 2013th ed. Malang: Maulana Malik Ibrahim State Islamic University, 2013.

MM Baharuddin, H. Azis, and T. Hasanuddin, "Performance Analysis of K-Nearest Neighbor Method for Identification of Glass Types," Ilk. J. Ilm., vol. 11, no. 3, pp. 269–274, 2019, doi:10.33096/ilkom.v11i3.489.269-274.

Sianturi, FA (2018). Decision Tree Analysis in Student Data Processing. MEANS (Media Information Analysis and Systems), 3(2), 166–172. http://ejournal.ust.ac.id/index.php/Jurnal_Means/

Sianturi, FA, Informatics, T., & Utara, S. (2018). Application of Apriori Algorithm for Level Determination. Penusa Mantik, 2(1), 50–57. http://e-jurnal.pelitanusantara.ac.id/index.php/mantik/article/view/330

Sowmya, R., & Suneetha, KR (2017). Data Mining with Big Data. Proceedings of 2017 11th International Conference on Intelligent Systems and Control, ISCO 2017. https://doi.org/10.1109/ISCO.2017.7855990

Suyanto. (2017). Data mining for data classification and clustering. SpringerReference.

Witten, IH, Frank, E., Hall, MA, & Pal, CJ (2016). Data Mining: Practical Machine Learning Tools and Techniques. In Data Mining: Practical Machine Learning Tools and Techniques. https://doi.org/10.1016/c2009-0-19715-5

Wu, X., Zhu, X., Wu, GQ, & Ding, W. (2014). Data mining with big data. IEEE Transactions on Knowledge and Data Engineering. https://doi.org/101.1109/TKDE.2013.109

* Nurfadillah Tanjung