Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 6, Number 4, October 2022
DOI : https://doi.org/10.33395/sinkron.v7i4.11767
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Data Mining implementation on SMUN Scholarship recipient candidates using the C4.5 algorithm

Rusdiansyah[1)*], Hendra Supendar[2)], Nining Suharyanti[3)], Tuslaela[4)]
[1,2,3)]Universitas Bina Sarana Informatika,Indonesia, [4]Universitas Nusamandiri, Indonesia
[1)]rusdiansyah.rds@bsi.ac.id [2)] hendra.hds@bsi.ac.id, [3)] nining.nni@bsi.ac.id, [4)] tuslaela.tll@nusamandiri.ac.id

Submitted : Sep 3, 2022 | **Accepted** : Sep 26, 2022 | **Published** : Oct 3, 2022

Abstract: This research is motivated by SMUN previously only providing scholarships for underprivileged students, while for students who excel, the method is not yet, therefore, research by implementing data mining with the C4.5 algorithm method so that in the future, if the scholarship program is for outstanding students already exists, then SMUN can immediately apply it. The sample data are 33 scholarship recipients with subjects that have been determined based on the report card scores of prospective scholarship recipients. The selection results for receiving scholarships to outstanding students so far have only been calculated in general terms, the possibility of incorrect results will cause losses for students who have good achievements. As a result, there are students who should get the scholarship rights and do not receive it. The problem of this research is how to make it easier to analyze students who need and are entitled to receive scholarships, how to make it easier to select and determine scholarship recipients. The research uses the C4.5 algorithm to overcome these problems. With the C4.5 algorithm, the percentage accuracy value can be calculated which is used to determine whether the student is entitled to receive a scholarship or not. The data collection technique in this study is to use student report cards. classification value in determining students who get scholarships with the highest accuracy of 81.81%. The accuracy value was obtained by experimenting with the testing process with the number of student data as many as 33 students. It is hoped that it will make it easier to receive scholarships for students who excel academically in school.

Keywords: Algorithm C4.5, Decision Tree, Scholarships, Subjects,

INTRODUCTION

Education is an activity carried out to develop self-potential (Amaliyah, 2021). The institutions responsible for the dissemination of knowledge are educational institutions. Every educational institution, especially high school (SMU) has several work programs that can help students(Basri, 2021), One of them is the scholarship program. Achievement scholarships become learning motivation for students so that they can improve in learning and become an advantage for schools to improve the quality of school education(Sari, 2020).

The school is an institution that organizes academic education for students. Students are input components in the education system which are then processed in the education process so that they become quality human beings in accordance with national education goals(Asmara, 2016). In the learning process at school, within a certain period of time, a large amount of data will be collected which will make it difficult for the school to process the data (Ardiansyah, 2018). This research is based on the observations of researchers, that the target for receiving scholarships is not accurate enough for students who need and are entitled to receive them, because of the difficulties in selecting students and the many criteria. As a result, there are students who should get the right of the scholarship to not receive it. The problem of this research, how to make it easier to analyze students who need and are entitled to receive scholarships(Indrasari, 2018), how to make it easier to select and determine scholarship recipients(Irawan, 2016). The data set will be further processed with data mining to obtain new patterns that can be used to increase effectiveness in the learning process. Data mining is a process that uses statistical, mathematical, artificial intelligence, and machine learning techniques to extract and identify useful information and related knowledge from large databases. SMUN 85 previously only provided scholarships for underprivileged students, while for students who excel had no program, therefore we developed the C4.5 algorithm method so that in the future, if a scholarship program for outstanding students already exists, then

*name of corresponding author

e-ISSN : 2541-2019
p-ISSN : 2541-044X

SMUN 85 has can apply it right away. The sample data used is the value data of prospective scholarship recipients.

LITERATURE REVIEW

Previous research that the author uses as a reference include the following:

1. Research conducted by Erfan Hasmin, et al, 2019. Application of the C4.5 Algorithm. For Determining Student Scholarship Recipients, in this study the problem that usually occurs is that there is no clear formulation used to determine scholarship recipients so that scholarship recipients often occur objectively because they only use one criterion. Because what happens is that candidates who receive scholarships are those who do not need it, because these students are economically able to finance their studies. From the results of the study, it can be concluded that the application of the c.45 algorithm for the recommendation of scholarship recipients already has a fairly good accuracy where the selection already has rules that have been carried out by the classification process with data mining. Thus, this application helps the student body in the selection of prospective scholarship recipients in the following year(Hasmin, 2019).
2. Research conducted by Abdurraghib Segaf Suweleh et al, 2020, Application for Determining Scholarship Recipients Using the C4.5 Algorithm, The results achieved from the test are knowing the accuracy of the implementation of the C4.5 algorithm in the process of determining scholarship recipients reaching 92%,. The conclusion of this study is that the C4.5 algorithm has been successfully implemented in the classification process for scholarship recipients based on an accuracy rate that reaches 92% and the results of valid system testing on each module using the black box method.(Suweleh, 2020).

Data mining has attracted a lot of attention in the information systems world and in society as a whole in recent years, due to the widespread availability of large amounts of data and the immediate need to transform that data into useful information and knowledge.(Badrul, 2016). The information and knowledge gained can be used for applications ranging from market analysis, fraud detection, and customer retention, to production control and exploration science. (Prihatmono, 2019). Classification is one of the techniques in data mining. Classification (taxonomy) is the process of placing certain objects or concepts into a set of categories based on the objects used. One of the popular classification techniques used is the decision tree(Pattipeilopy, 2017). Classification itself is divided into two stages, namely classification and learning. At the learning stage, a classification algorithm will build a classification model by analyzing the training data(Iriadi, 2016). The learning stage can also be seen as the stage of forming a function or mapping y=f(x) where y is the predicted class and X is the tuple that the class wants to predict. (Latifah, 2018). The C4.5 algorithm is one of the algorithms that has been widely used, especially in the machine learning area which has several improvements from the previous algorithm, namely ID3(Udariansyah, 2022)(Latifah, 2018). The C4.5 algorithm and the ID3 model are inseparable, because to build a decision tree, a C4.5 algorithm is needed. In the late 1980s, J. Ross Quinlan a researcher in the field of machine learning developed a decision tree model called ID3(Lubis, 2019). There are several stages in making a decision tree in the C4.5 algorithm, namely:

1. Prepare training data. Training data is usually taken from historical data that has happened before and has been grouped into certain classes.
2. Counting tree roots. The root will be taken from the attribute to be selected, by calculating the gain value of each attribute, the highest gain value will be the first root. Before calculating the gain value from the attribute, first calculate the entropy value.

To calculate the entropy value used the formula:

$$Entropy(S) = pi * Log2\ pi \qquad (1)$$

Information:
S : Case Collection
n : Number of Partitions S
pi : Proportion of Si to S

Then after the entropy value for each attribute has been obtained, calculate the gain value using the formula

$$Gain(S, A) = Entropy(S) - * \ Entropy(Si) \qquad (2)$$

Information :
S : Case Collection
A : Features
N : Number of attribute partitions A
|Si|: Number of cases on partition i
|S| : Number of cases in S

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 6, Number 4, October 2022
DOI : https://doi.org/10.33395/sinkron.v7i4.11767
e-ISSN : 2541-2019
p-ISSN : 2541-044X

METHOD

The stages in this research include research steps. The framework in this research is described as follows:

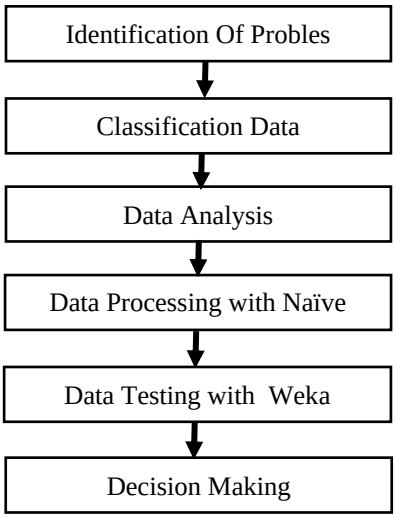

Figure 1: Research Framework

The data used in this study is data based on the criteria used in the calculation, namely the students of SMUN 85 class X which are used for the calculation of the highest alternative to determine students who will receive scholarships. The method proposed for the process as described above is a classification method with the algorithm used is the C4.5 algorithm with the following criteria used:

1. Student Name
2. Indonesian Values
3. English Grades
4. Value of Natural Sciences (IPA)
5. Value of Social Sciences (IPS)
6. Math Grades
7. The value of activity (includes activeness in working on questions and answering questions during tutoring) taking place).
8. Obtaining IQ scores
9. The Value of Tutoring

To determine the conversion of existing data values, the values of subjects and guidance are converted using a range of quality values.

Table 1. Value Conversion

Score	Quality
86-100	A
71-85	B
56-70	C
41-55	D
≤ 40	E

In Table 1. Conversion of subject values into a range of quality values from A, B, C, D, E
In addition to the converted subject and guidance values, IQ scores are also converted using the table below.

Table 2. IQ Conversion

Range	Category	Classification
>=140	Genius	5
120-139	Superior	4
110-119	Diatas rata-rata	3
90-109	Rata-rata	2
<=89	Dibawah Rata-rata	1

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 6, Number 4, October 2022
DOI : https://doi.org/10.33395/sinkron.v7i4.11767
e-ISSN : 2541-2019
p-ISSN : 2541-044X

In Table 2. Convert IQ scores, with categories (genius, superior, above average, average, below average) into a range of values of 5,4,3,2,1.

RESULT

At the classification stage of student data who meet the requirements for applying for a scholarship, it is obtained from the results of the conversion of subject values and IQ scores with the table below.

Table 3. Conversion Results

B.IND	B.ING	IPA	IPS	MTK	ACTIVE	IQ	BIM	RESULT
A	A	A	A	A	A	2	A	Lolos
A	A	A	B	B	K	1	B	Lolos
A	A	A	A	B	K	1	B	Lolos
A	A	A	A	B	A	1	B	Lolos
B	A	A	A	B	A	1	A	Lolos

In Table 3 above, the results of the conversion process of subject values and IQ scores, there are 5 students who are entitled to scholarship applications and will be processed further with the implementation of the C4.5 algorithm.

The calculation of the entropy and gain values for all attributes is carried out to obtain the highest gain value which will be used as the root. The calculation results can be seen in the table below:

Table 4. Calculation results of gain and entropy

NODE	Attribute	Category	Number of Cases (S)	Get away (L)	Did not pass (TL)	Entropy	Gain
1	TOTAL		33	21	12	0,945660305	
	B. Ind	A	20	12	8	0,970950594	0,006405662
		B	13	9	4	0,890491640	
	B. Ing	A	21	18	3	0,591672779	0,274131037
		B	12	3	9	0,811278124	
	IPA	A	21	14	7	0,918295834	0,004974317
		B	12	7	5	1	
	IPS	A	13	13	0	0	0,357205399
		B	20	8	12	0,970950594	
	MTK	A	13	8	5	0,961236605	0,000889487
		B	20	13	7	0,934068055	
	ACTIVE	A	17	10	7	0,977417818	0,007699256
		K	16	11	5	0,896038233	
	IQ	1	32	20	12	0,954434003	0,020148544
		2	1	1	0	0	
	BIM	A	8	4	4	1,000000000	0,018098594
		B	25	17	8	0,904381458	

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 6, Number 4, October 2022
DOI : https://doi.org/10.33395/sinkron.v7i4.11767
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Based on Table 4. the results of the gain calculation show that the social studies subject attribute has the highest gain value, which is 0.357205399 so that the social studies attribute can be used as the root node of the decision tree.

Implementation is done using Rapid Miner software. By generating a decision tree as follows:

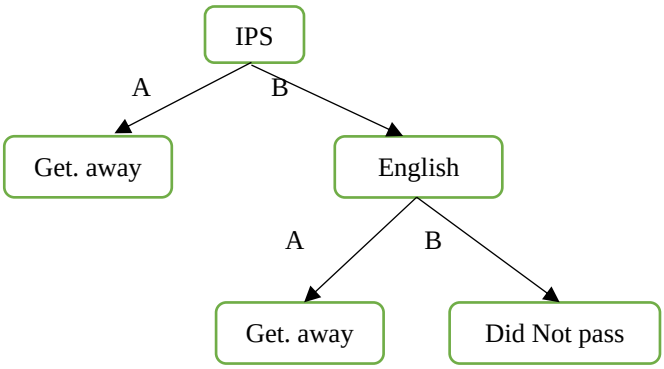

Figure 2. Decision Tree Results

In Figure 2, it can be seen that the root node of the tree is IPS with quality value = A with pass status and English value with A quality value with pass status and B quality value does not pass. The accuracy of all classifications is determined by the number of correct classifications divided by the total number of classification records with a result of 81.81%.

DISCUSSIONS

The results of the Decision tree determination of scholarship recipients resulted in a decision that students who were 'accepted' as scholarship recipients were students who had social studies subjects with a quality of A = genius and English subjects with a value of quality = A. In the research conducted by Erfan Hasmin, the references for articles cannot be used as similarities, because the criteria are economic factors. In research conducted by Abdurraghib, there are similarities in applying for scholarships because the classification is based on the value of the subject with an accuracy rate of 92% for scholarship recipients.

CONCLUSION

Based on the test results on the classification of prospective recipients of the SMUN 85 scholarship, the conclusion is the classification of the selection process for prospective scholarship recipients can classify students in the stages of passing or not in the selection. Of the 33 student data used, it shows an accuracy rate of 81.81%. It can be concluded that the application of the C.4.5 algorithm for the recommendation of scholarship recipients already has a fairly good accuracy the selection already has rules that have been carried out by the classification process with data mining. Thus, this application helps the student body in the selection of prospective scholarship recipients in the following year.

REFERENCES

Amaliyah, A. (2021). Pengembangan Potensi Diri Peserta Didik Melalui Proses Pendidikan. *Journal of Elementary Education*, *5*(1), 28–45.

Ardiansyah, D. (2018). Algoritma c4.5 untuk klasifikasi calon peserta lomba cerdas cermat siswa smp dengan menggunakan aplikasi rapid miner. *Jurnal Inkofar*, *1*(2), 5–12.

Asmara, R. (2016). Dinamika Madrasah Dan Sistem Penyelenggaraan Pendidikan Islam Unggulan. *Revista Brasileira de Ergonomia*, *3*(2), 80–91. Retrieved from https://www.infodesign.org.br/infodesign/article/view/355%0Ahttp://www.abergo.org.br/revista/index.php/ae/article/view/731%0Ahttp://www.abergo.org.br/revista/index.php/ae/article/view/269%0Ahttp://www.abergo.org.br/revista/index.php/ae/article/view/106

Badrul, M. (2016). Algoritma asosiasi dengan algoritma apriori untuk analisa data penjualan. *Pilar Nusa Mandiri*, *XII*(2), 121–129.

Basri, B. (2021). Manajemen Kepala Sekolah dalam Meningkatkan Fungsi Guru di Sekolah Menengah Atas Negeri 4 Merangin. *Jurnal Ilmiah Dikdaya*, *11*(2), 349. https://doi.org/10.33087/dikdaya.v11i2.233

Hasmin, E. (2019). Penerapan Algoritma C4.5 Untuk Penentuan Penerima Beasiswa Mahasiswa. *CogITo Smart Journal*, *5*(2), 308. https://doi.org/10.31154/cogito.v5i2.219.308-320

Indrasari, F. N. (2018). Klasifikasi Penerimaan Beasiswa Menggunakan Algoritma Naive Bayes Classifier.

*name of corresponding author

Universitas Nusantara PGRI Kediri, *3*(2), ثقثقثقق. Retrieved from http://journal.stainkudus.ac.id/index.php/equilibrium/article/view/1268/1127%0Ahttp://publicacoes.cardiol.br/portal/ijcs/portugues/2018/v3103/pdf/3103009.pdf%0Ahttp://www.scielo.org.co/scielo.php?script=sci_arttext&pid=S0121-75772018000200067&lng=en&tlng=

Irawan, M. T. (2016). Penerapan Profile Matching Untuk Pencarian Siswa Smp Penerima Beasiswa Miskin Dan Berprestasi. *JIKO (Jurnal Informatika Dan Komputer)*, *1*(1), 24–29. https://doi.org/10.26798/jiko.2016.v1i1.11

Iriadi, N. (2016). Kajian Penerapan Metode Klasifikasi Data Kelayakan Kredit Pada Bank. *Jurnal Teknik Komputer AMIK BSI*, *II*(1), 132–137.

Latifah, K. (2018). *Analisis Dan Penerapan Algorithma C4.5 Dalam Data Mining Untuk Menunjang Strategi Promosi Prodi Informatika Upgris*. *11*(2).

Lubis, M. R. (2019). Analisa Prediksi Penjualan Produk Dengan Menggunakan Metode C4.5 (Studi Kasus : PT. Kawan Lama Ace Hardware). *Jurnal Riset Komputer*, *6*(5), 545–549. Retrieved from https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom/article/view/1656/1253

Pattipeilopy, W. F. (2017). Pemodelan Dan Prototipe Sistem Informasi Untuk Prediksi Pembaharuan Polis Asuransi Mobil Menggunakan Algoritma C.45. *Prosiding SNATIF*, 791–799.

Prihatmono, M. W. (2019). Implementasi Algoritma C4.5 Menggunakan Python Untuk Klasifikasi Kepuasan Konsumen. *Progres*, 49–55. Retrieved from https://jurnal.stmikprofesional.ac.id/index.php/Progress/article/view/146/22

Sari, N. M. (2020). Pengaruh Kartu Jakarta Pintar (Kjp) Terhadap Motivasi Belajar Siswa Kelas Xii Di Smk Dharma Putra 1 Jakarta. *Research and Development Journal of Education*, *1*(1), 01. https://doi.org/10.30998/rdje.v1i1.6439

Suweleh, A. S. (2020). Aplikasi Penentuan Penerima Beasiswa Menggunakan Algoritma C4 . 5. *Jurnal …*, *2*(1), 12–21. https://doi.org/10.30812/bite.v2i1.798

Udariansyah, D. (2022). Implementasi Algoritma C4.5 Pada Jumlah Penduduk Kota Prabumulih menggunakan Metode Klasifikasi Devi. *Pendidikan Dan Konseling*, *4*. Retrieved from https://core.ac.uk/download/pdf/322599509.pdf

Utami, D. S. (2021). Analisis Sentimen Pinjaman Online di Twitter Menggunakan Algoritma Support Vector Machine (SVM). *SISMATIK (Seminar Nasional Sistem Informasi Dan Manajemen Informatika)*, *1*(1), 299–305.

*name of corresponding author