ILKOM Jurnal Ilmiah Vol. 14, No. 2, August 2022, pp. 142-149
Accredited 2nd by RISTEKBRIN No. 200/M/KPT/2020; E-ISSN 2548-7779 | P-ISSN 2087-1716

Research Article Open Access (CC–BY-SA)

Classification of reading skills of early childhood with C4.5 algorithm

Suherman [a,1,*]; **Agus Nursikuwagus** [b,2]; **Ahmad Sugiyarta** [a,3]; **Indah Komala** [a,4]
[a] *Universitas Serang Raya, Jl Raya Serang-Cilegon Km 5, Serang Banten and 42162, Indonesia*
[b] *Universitas Komputer Indoensia, Jl Dipatiukur No.112-116, Bandung and 40132, Indonesia*
[1]*suherman.unsera@gmail.com;*[2] *agusnursikuwagus@email.unikom.ac.id;*[3]*ahmad.sugiyarta@unsera.ac.id;* [4]*indahhkomalaa@gmail.com*
**Corresponding author*

Article history: Received October 16, 2021; Revised December 15, 2021; Accepted July 09, 2022; Available online August 30, 2022

Abstract

Childhood is a golden age for children, which is important phase for their growth and development. Children need to obtain get a good education according to their development phase to improve their abilities including language. Early reading skill is considered as one type of language skills which is essential at the childhood period. Referencing to teacher's information, one of the problems related to students' reading ability is the absence of determination of the classification of early childhood reading skills. Currently, teachers are challenged in providing the achievement of students' reading skills based on criteria of early reading ability. Consequently, they are not able to improve the learning quality to each student nor prepare material to discuss with students' parents. There is a suggestion given to achieve the goals and resolve it. It is the classification of early childhood reading skills that has been designed where the data processing is carried out using a computer. This study employs a decision tree and Algorithm C4.5 for the method. Algorithm C4.5 is included in the classification based on Decision Tree. Algorithm C4.5 is a method that can handle features with numeric types. The purpose of this study is to perform classification using Algorithm C4.5 as a method that can find out the level of reading skills in early childhood, based on the criteria, so that classification can be done easily and get the right improvement suggestion. The result in this study is that the Algorithm C4.5 method has a good result in solving and calculating for the classification so that the accuracy value of the Algorithm C4.5 calculation that has been carried out is 77,14%.

Keywords: Reading; Early childhood, Data mining, Algorithm C4.5

Introduction

Murdiono proposed early age as a golden age that only occurs once in human development[1]. Children have the right to obtain education that is in accordance with their development in order to maximize their abilities, including language skills. Children's skills can be explored and developed if they are able to read. Sofiana explained that a child's language skills can be observed through receptive language (listening and reading) and expressive language (expressing language both verbally and nonverbally). The language skills needed at this time are initial reading skills, because reading can increase understanding, expand knowledge for children and hone the wording used [2]. Currently, there are still differences of opinion regarding the early reading skills that early childhood needs to posses. According to psychologists, this ability may cause depression to children. On the other hand, experts such as Neuman, Coople, and Bredekamp are of opinion that that learning to read and write is very important for children's success in school and in future life [3]. From birth to eight years of age is the most important time to improve children's literacy. The key phase to improve children's literacy is since their birth up to eight years of age.

In 2013, Sindonews.com revealed the results of Indonesia's national reading comprehension index which was only 0.01, far behind the developed countries of 0.45-0.62 [3]. According to several studies, Indonesia is sadly considered as a book-zero tragedy (tragedy nol buku). In average, Indonesian child reads only 27 pages per year. Finland was ranked 1st with 300 pages in 5 days[3]. Fatimah et al said that there are differences in the ability of each child to recognize letters. This ultimately leads to a low level of early reading ability in children [4]. In the current digital era, the benefits of using gadgets is considered inappropriate for children as they merely get information related to playing and ignore learning understanding such as reading. The low interest in reading is caused by the low awareness of parents who are of opinion that childhood is the time to play with friends. Guidance for children such as early childhood education needs to be taken for granted because it is a place for children to both

http://dx.doi.org/10.33096/ilkom.v14i2.1012.142-149

learn and play. Mukhtar, et al said that because school education only accounts for 20% of the total education received by children, teachers must work hard to maximize it [5].

Based on the abovementioned studies, it is important for early childhood to be taught about early reading. One of the guidance given is to pronounce letters that are similar or almost the same in shape. This is very useful for the present and also the future in understanding the importance of literacy. Therefore, teachers are required to be able to determine the classification of the level of achievement or early reading ability in children, and also to improve the quality of guidance for each child. It also provides information for discussion with parents of students on the abilities of children.

A fast and precise way is needed to get an accurate classification to avoid subjective judgments as well as make it easier for parents or guardians to understand the results. One of the methods suggested for the classification is the C.45 Algorithm. Saeful Bahri said that one of the reasons for using the C.45 Algorithm is that this algorithm has the highest level of accuracy compared to other algorithms. The advantage of classification using decision trees has the advantage of breaking complex structures into simpler structures so that they are easier to implement [15]. Novandya proposed the accuracy of the C4.5 classification algorithm resulted in a value of 88.89% as evidenced by the resulting program[6]. Judging from previous research, the C.45 Algorithm method is suitable for use in classifying to make it easier for teachers to identify the achievement of sudents' early reading skills.

This study aims to determine the classification of early childhood reading skills by employing the C.45 Algorithm method to simplify and speed up the identification of the classification of early childhood reading skills by implementing 15 criteria, namely, 1. Children are able to pronounce the letters a-z; 2. Children are able to show the letters a-z; 3. Children are able to pronounce vowels; 4. Children are able to show vowels; 5. Children are able to pronounce consonants; 6. Children are able to show consonants; 7. Children are able to pronounce letters that are similar or almost the same in shape/sound; 8. Children are able to show letters that are similar or almost the same in shape/sound; 9. Children are able to read pictures written in simple words; 10. Children are able to mention the letters in simple words; 11. Children are able to read words without pictures; 12. Children are able to connect pictures with words; 13. Children are able to read pictures written in simple sentences; 14. Children are able to read simple sentences; 15. Children are able to read their own names. 160 PAUD students were surveyed in Pandeglang Regency. They are divided into 2, namely 125 training data and 35 test data. The data was processed in the classification of early childhood reading abilities using a rating scale suggested by Sugiono (2017). The rating scale used ranges from a score of 1 (not yet developed), a score of 2 (starting to develop), a score of 3 (developing as expected), and a score of 4 (developing very well)[7].

Method

The algorithms commonly used for data mining are ID3, C4.5, K-Means, Naive Bayes, Support Vector Machines, Apriori algorithms and other data mining algorithms [8]. Classification is the process of grouping a set of objects in the database into classes based on the similarity or classification specified [9]. The method used to perform the classification is the C4.5 Algorithm. The C4.5 algorithm was designed by J. Ross Quinlan. The reason for the naming is because it is derived from the ID3 approach for decision tree stimulation [10]. The C4.5 algorithm uses the gain ratio. To calculate the gain ratio, the bit information value of a collection of objects needs to be calculated first using the concept of entropy [11].

a. Phase of Decision Construction Using C4.5 Algorithm

1. Start with a node that represents the training data sample by creating a tree root node.
2. If all samples belong to the same class, this node is classified as a leaf class. Otherwise, use the gain ratio to select the split attribute. This is the optimal attribute for separating sample data into individual classes.
3. Branches are created for each attribute value and the sample data is subdivided.
4. This algorithm uses a recursive process to form a decision tree for each data partition. If an attribute has been used in a node, it is not reused in child nodes.
5. This process stops when it reaches the condition that all samples in the node are in one class and there are no other attributes that can be used to partition additional samples. In this case, the majority vote will be applied which means turning nodes into leaves and determining the class label that gets the most votes.

b. Entropy Concept

Entropy is used to measure the non-authenticity of the sample (S) [11]. According to Prasetyo [12], entropy can be calculated using the following **Equation 1**:

$$\boldsymbol{E_s = \sum_{i=1}^{m} p(w_i|s) log_2 p(w_i|s)} \tag{1}$$

$p(w_i|s)$is the proportion of class-i in all training data processed at node s. $p(w_i|s)$ is obtained from all rows of data with class label-i divided by the number of rows of all data. While m is the number of different values in the data.

Results and Discussion

The first stage is the analysis of the calculation of training data of 125 data which is calculated based on 15 criteria divided into 3 parts, namely A, B, and C. In the introduction, part A uses criteria number 1 – 8, part B uses criteria number 9 – 12 , part C uses criteria number 13 – 15. The criteria used are the results of calculations in the first iteration indicated by the following steps:

1. Determine the number of sets of BSB, BSH, MB and BB. In this case the sum of the sets of BSB is 26, BSH is 52, MB is 42 and BB is 5.
2. Calculate the total entropy value. Entropy is used to measure the authenticity of the sample [11]. Entropy can be calculated using the following equation[12].

$$E_{(s)} = -\sum_{i=1}^{m} p(w_i\ |s) log_2 p(w_i|s)$$

$$E_s = -\frac{26}{125}log_2\frac{26}{125} - log_2 - \frac{52}{125}log_2\frac{52}{125} - \frac{42}{125}log_2\frac{42}{125} - \frac{5}{125}log_2\frac{5}{125} = 1{,}712014123$$

3. Calculate the entropy value of each attribute to the total entropy and calculate the information gain value of each attribute.

One of the advantages of the c4.5 Algorithm is that it handles training datasets with missing attribute values. C4.5 allows attribute values to be marked as '?' for lost. The missing attribute values are not used in the gain and entropy calculations [13].

From the calculation analysis above, the results can be shown in **Table 1** below:

Table 1. Entropy of attribute

		Total (s)	**BSB**	**BSH**	**MB**	**BB**	**Entrophy**	**Gain**
total		125	26	52	42	5	1.712014123	
A								1.111999004
	BSB	23	23	0	0	0	0	
	BSH	34	3	31	0	0	0.430551867	
	MB	42	0	21	21	0	1	
	BB	26	0	0	21	5	0.706274089	
B								0.905096926
	BSB	12	12	0	0	0	0	
	BSH	36	14	22	0	0	0.964078765	
	MB	49	0	30	19	0	0.963335546	
	BB	28	0	0	23	5	0.67694187	
C								0.932973824
	BSB	12	12	0	0	0	0	
	BSH	30	14	16	0	0	0.996791632	
	MB	53	0	36	17	0	0.905200297	
	BB	30	0	0	25	5	0.650022422	

4. Determine the root tree

Based on the information gain for each attribute from the results of **Table 1**, the following information gain table is shown in **Table 2**: Based on the information gain for each attribute from **Table 1**, **Table 2** below portraits the information gain:

Table 2. Information Gain

Attribute	**Information Gain**
Attribute_A	1.111999004
Attribute _B	0.905096926

Attribute	Information Gain
Attribute _C	0.932973824

From the results above, the attribute that becomes the root is A with the highest gain of 1.111999004, consisting of 4 values or sub attributes, namely BSB, BSH, MB, BB. Based on the entropy values of the four values, the BSB value has a decision, while the BSH, MB, and BB attributes need to be further calculated for the root node. The decision tree is shown in **Figure 1**. A decision tree is a tree in which each node represents a feature (attribute), each link (branch) represents a decision (rule) and each leaf represents a result (categorical or continuous value)[14].

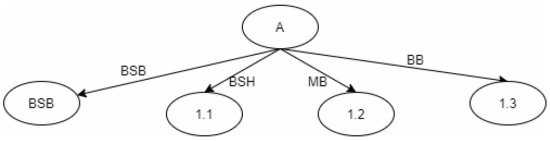

Figure 1. Root of Decision Tree

5. Repeat the step 3 and step 4

After attribute A became the root attribute, the initial training data then shrank after the training data split occurred. The calculations for the other branches is given in **Table 3**.

Table 3. Entropy of Each Node Attribute 1.1

		total (s)	BSB	BSH	MB	BB	Entrophy	Gain
total		34	3	31	0	0	0.430551867	
B								0.171348661
	BSB	0	0	0	0	0	0	
	BSH	10	3	7	0	0	0.881290899	
	MB	24	0	24	0	0	0	
	BB	0	0	0	0	0	0	
C								0.132565165
	BSB	0	0	0	0	0	0	
	BSH	13	3	10	0	0	0.779349837	
	MB	21	0	21	0	0	0	
	BB	0	0	0	0	0	0	

Table 4. Entropy of Each Node Attribute 1.2

		total (s)	BSB	BSH	MB	BB	Entrophy	gain
total		42	0	21	21	0	1	
B								0.526761738
	BSB	0	0	0	0	0	0	
	BSH	15	0	15	0	0	0	
	MB	25	0	6	19	0	0.795040279	
	BB	2	0	0	2	0	0	
C								0.240243506
	BSB	0	0	0	0	0	0	
	BSH	6	0	6	0	0	0	
	MB	32	0	15	17	0	0.997180399	
	BB	4	0	0	4	0	0	

Table 5. Entropy of Each Node Attribute 1.3

		total (s)	BSB	BSH	MB	BB	Entrophy	gain
total		26	0	0	21	5	0.208538074	
B								0
	BSB	0	0	0	0	0	0	
	BSH	0	0	0	0	0	0	
	MB	0	0	0	0	0	0	
	BB	26	0	0	21	5	0.208538074	

C								0
	BSB	0	0	0	0	0	0	
	BSH	0	0	0	0	0	0	
	MB	0	0	0	0	0	0	
	BB	26	0	0	21	5	0.208538074	

After the calculation results at Node 1.1 shown by **Table 3**, Node 1.2 shown by **Table 4**, and Node 1.3 shown by **Table 5** are obtained, **Figure 2** illustrates the results of a decision tree.

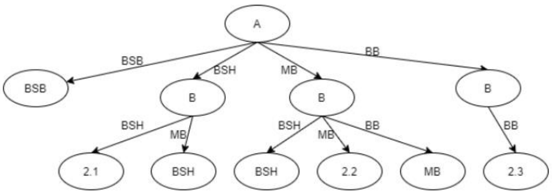

Figure 2. Leaf of Decision Tree

From the decision tree in **Figure 2**, the calculation continues to get results at node 2.1, node 2.2, and node 2.3. The following are the results of the calculation analysis shown by **Table 6**, **Table 7**, and **Table 8**.

Table 6. Entropy of Each Node Attribute 2.1

		total (s)	BSB	BSH	MB	BB	Entrophy	gain
total		10	3	7	0	0	0.881290899	
C								0.881290899
	BSB	0	0	0	0	0	0	
	BSH	3	3	0	0	0	0	
	MB	7	0	7	0	0	0	
	BB	0	0	0	0	0	0	

Table 7. Entropy of Each Node Attribute 2.2

		total (s)	BSB	BSH	MB	BB	Entrophy	gain
total		25	0	6	19	0	0.795040279	
C								0.795040279
	BSB	0	0	0	0	0	0	
	BSH	6	0	6	0	0	0	
	MB	17	0	0	17	0	0	
	BB	2	0	0	2	0	0	

Table 8. Entropy of Each Node Attribute 2.3

		total (s)	BSB	BSH	MB	BB	Entrophy	gain
total		26	0	0	21	5	0.208538074	
C								0
	BSB	0	0	0	0	0	0	
	BSH	0	0	0	0	0	0	
	MB	0	0	0	0	0	0	
	BB	26	0	0	21	5	0.208538074	

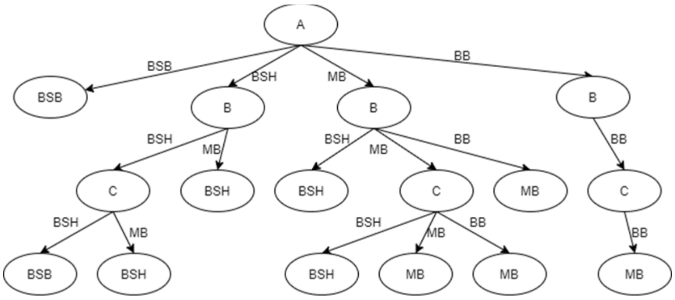

Figure 3. Decision Tree

Based on the above **Figure 3** of decision tree, the rules obtained are shown in **Table 9** as follows.

Table 9. Decision Tree Rules

No	Rules
1.	IF A=BSB THEN BSB
2.	IF A=BSH AND B=BSH AND C=BSH THEN BSB
3.	IF A=BSH AND B=BSH AND C= MB THEN BSH
4.	IF A=BSH AND B=MB THEN BSH
5.	IF A=MB AND B=BSH THEN BSH
6.	IF A=MB AND B=MB AND C=BSH THEN BSH
7.	IF A=MB AND B=MB AND C=MB THEN MB
8.	IF A=MB AND B=MB AND C=BB THEN MB
9.	IF A=MB AND B=BB THEN MB
10.	IF A=BB AND B=BB AND C=BB THEN MB

After obtaining the rules from the mining calculation of 125 training data, the next step is to test the rules using 35 test data as shown in **Table 10** below:

Table 10. Test of Data Rule Test

NO	NAME	CLASS	A	B	C	INITIAL CLASSIFICATION	RULE	CLASSIFICATION OF RESULTS
1	Aufa Amelia	B	BSH	BSH	MB	BSH	3	BSH
2	Safira	B	BB	BB	BB	MB	10	MB
3	Kapa Husnul Abadiyah	B	MB	MB	BSH	BSH	6	BSH
4	Nayla Jayanti	B	BB	BB	BB	MB	10	MB
5	Anwar	B	MB	MB	MB	MB	7	MB
6	Daffa Maulana Putra	A	MB	MB	MB	MB	7	MB
7	Kayla Salsabila	A	BSH	MB	MB	BSH	4	BSH
8	Ogi Hermawan	A	BSB	BSH	BSH	BSB	1	BSB
9	Inaya Shakila Tsaqib	A	BSB	BSB	BSB	BSB	1	BSB
10	Juan Raitama	A	MB	MB	MB	MB	7	MB
11	Muhamad Rafa Putra	A	MB	BSH	MB	BSH	5	BSH
12	Nyimas Dewi Fatima	A	BSH	MB	MB	BSH	4	BSH
13	Ahdu Rofiah	A	BSH	MB	MB	BSH	4	BSH
14	Arista Adinda Zahranny	A	BB	BB	BB	BB	10	MB
15	Rahmadini	A	BB	BB	BB	MB	10	MB
16	M. Andhika Fardan	A	BSH	MB	BSH	BSH	4	BSH
17	Wasy Mubarak	A	BSH	BSH	BSH	BSB	2	BSB
18	M. Agus Syahputra	A	BSH	MB	MB	BSH	4	BSH
19	Syakila Karisa Putri	A	MB	MB	BB	MB	8	MB
20	Aisyah Ramadini	A	MB	BB	BB	MB	9	MB
21	M.Pratama Mahendra	A	BB	BB	BB	BB	10	MB
22	Khayla Khairunnisa	A	MB	BB	BB	MB	9	MB
23	Zidni Hadya Mulya	A	BSB	BSH	BSH	BSB	1	BSB
24	Raihana Jasmine	A	MB	BB	BB	MB	9	MB
25	Naila Lila Azzahra Somardi	A	BB	BB	BB	BB	9	MB
26	Muhammad Alby Adhyastha	A	MB	BB	BB	MB	9	MB
27	Anita Rahayu	A	BB	BB	BB	BB	10	MB
28	Ratifah Az Zahra	A	BB	BB	BB	BB	10	MB
29	Kianu Rizky Ramadahan	A	BB	BB	BB	MB	10	MB
30	M.Mirza Hazwan	A	BB	BB	BB	MB	10	MB
31	Alesha Nurmaulida Subagja	A	BB	BB	BB	MB	10	MB
32	M.Ghifary Dhanar Prawira	A	BB	BB	BB	BB	10	MB
33	Akhdan Haidar Azhari	A	BB	BB	BB	BB	10	MB
34	Khanza Nabilahat Marani	A	BB	BB	BB	BB	10	MB
35	Raisa	A	BB	BB	BB	MB	10	MB

Next is to calculate the performance of the classification in this study using the contusion matrix method from the results obtained, as follows:

a. Accuracy : $\frac{Jumlah\ data\ yang\ diprediksi\ secara\ benar}{Jumlah\ prediksi\ yang\ dilakukan}$ (2)

Accuracy : $\frac{27}{35} x\ 100 = 77{,}14\%$

b. Error rate : $\frac{Jumlah\ data\ yang\ diprediksi\ secara\ salah}{Jumlah\ prediksi\ yang\ dilakukan}$ (3)

Error rate : $\frac{8}{35} x\ 100 = 22{,}86\%$

Conclusion

Based on the results of the research, the following conclusions can be drawn: Able to understand the procedure for classifying early childhood reading skills with 15 criteria that can be implemented into the classification of early childhood reading skills has been successfully carried out using the C4.5 Algorithm method.The classification that applies the C4.5 Algorithm method produces a good classification in solving and calculating the score for each criterion owned by students who have carried out the testing phase. From the decision tree, accurate results are obtained in the classification process with an accuracy rate of 77.14% with an error rate of 22.86%.With the Classification of Early Childhood Reading Ability, teachers can be more efficient, effective and easier in obtaining the results of the classification of early reading abilities as material for evaluation and discussion with students' guardians.

References

[1] L. Nurul Safitri and H. Aziz, “Pengembangan nilai agama dan moral melalui metode bercerita pada anak,” Age Jurnal Ilmiah Tumbuh Kembang Anak Usia Dini, vol. 4, no. 1, pp. 85–96, 2019.

[2] L. Sofiana, “Pengaruh game edukatif mengenal huruf melalui media powerpoint terhadap kemampuan membaca permulaan,” Serang, 2021.

[3] H. Farlina, “Alat permainan edukatif scrabble untuk meningkatkan kemampuan membaca permulaan anak kelompok B,” Jurnal Golden Age, vol. 3, no. 1, pp. 17–29, 2019.

[4] Fatimah, S. Kholijah, and S. Susanti, “Pengaruh media audio visual terhadap kemampuan membaca permulaan pada anak usia dini ra darul isitiqomah 2 desa karang anyar Lampung Timur,” AZZAHRA, vol. 1, no. 1, pp. 39–52, 2019.

[5] M. Qadafi, “kolaborasi guru dan orang tua dalam mengembangkan aspek moral agama anak usia dini,” AWLADY: Jurnal Pendidikan Anak , vol. 5, no. 1, pp. 1–19, 2019, [Online]. Available: www.syekhnurjati.ac.id/jurnal/index.php/awlady

[6] A. Novandya, “Penerapan algoritma klasifikasi data mining C4.5 pada dataset cuaca wilayah Bekasi,” Konferensi Nasional Ilmu Sosial & Teknologi (KNiST), vol. 1, no. 1, pp. 368–372, 2017.

[7] Sugiyono, Statistika Untuk Penelitian. Bandung: Alfabeta, 2017.

[8] N. Rofiqo, A. P. Windarto, and E. Irawan, “Penerapan algoritma C4.5 pada penentuan tingkat pemahaman mahasiswa terhadap matakuliah,” Posiding Seminar Nasional Riset Information Science (SENARIS), pp. 307–317, 2019.

[9] F. Rahman, H.Z. Zahro, and F. Ariwibosono, “Penerapan algoritma C.45 dalam memprediksi asal calon mahasiswa berbasis website (Studi Kasus: Fakultas Hukum Universitas Mataram),” JATI (Jurnal Mahasiswa Teknik Informatika), vol. 4, no. 2, pp. 161–169, 2020.

[10] I. Massulloh and Fitriyani, “Implementasi algoritma C4.5 untuk klasifikasi anak berkebutuhan khusus di ibnu sina stimulasi center,” eProsiding Sistem Informasi (POTENSI), vol. 1, no. 1, pp. 136–144, 2020.

[11] T. Agus Setiawan and A. Ilyas, “Penerapan algoritma C4.5 dalam pemilihan konsentrasi program studi,” IC-TECH, vol. 14, no. 1, p. 4246, 2019, [Online]. Available: http://ejournal.stmik-wp.ac.id

[12] E. Prasetyo, Data Mining-Mengolah Data menjadi Informasi Menggunakan Matlab. Yogyakarta: Penerbit Andi, 2014.

[13] S. Kumar and H. Sharma, “A survey on Decision Tree algorithms of classification in data mining,” Article in International Journal of Science and Research, vol. 5, no. 4, pp. 2094–2097, 2016, [Online]. Available: www.ijsr.net

[14] H. H. Patel and P. Prajapati, “Study and analysis of Decision Tree Based classification algorithms,” International Journal of Computer Sciences and Engineering, vol. 6, no. 10, pp. 74–78, Oct. 2018, doi: 10.26438/ijcse/v6i10.7478.

[15] Saeful Bahri. (2017). Seleksi atribut pada algoritma C4.5 menggunakan Genetik algoritma dan Bagging untuk analisa kelayakan pemberian kredit. Kumpulan jurnaL Ilmu Komputer (KLIK), 4(2), 174-183.