Jurnal Informatika Universitas Pamulang **ISSN: 2541-1004**
Penerbit: Program Studi Teknik Informatika Universitas Pamulang **e-ISSN: 2622-4615**
Vol. 5, No. 3, September 2020 (261-266) **10.32493/informatika.v5i3.5268**

Application of C4.5 Algorithm in Improving English Skills in Students

Kristin D R Sianipar[1], Septri Wanti Siahaan[2], P.P.P.A.N.W Fikrul Ilmi R.H Zer[3], Dedy Hartama[4]

[1,2,3,4]Program Studi Teknik Informatika, STIKOM Tunas Bangsa Pematangsiantar, Jl. Jend. Sudirman Blok A, No. 1,2 dan 3, Kota Pematangsiantar, Sumatera Utara, Indonesia, 21143
e-mail: [1]kristinsianipar7@gmail.com, [2]septriwanti26@gmail.com, [3]fikrulilmizer@gmail.com, [4]dedyhartama@amitunasbangsa.ac.id

Submitted Date: May 31st, 2020 Reviewed Date: June 17th, 2020
Revised Date: September 22nd, 2020 Accepted Date: September 30th, 2020

Abstract

In this world, many languages from other countries can be used as a communication tool. One of them is English. Students who has qualification must know that learning English is very much needed. Because nobody knows what will happen in the next few years. It could be one factor to obtain a position the next few years is our expertise in speaking English. English is a global language used by people to communicate with other people. On this occasion, researchers conducted research to determine what factors can improve students' ability to speak English. To complete this research, researchers resolve by applying the existing algorithm in data mining, namely C4.5 Algorithm. The result of this research can be concluded that the factors that influence to improve students ability in English are hearing from the environment.

Keywords: C4.5 algorithm; English; data mining; improving; ability

1. Introduction

Countries in the world have a language used in communicating with other people. In Indonesia, the language of unity is used, namely Indonesian. However, if you wanted to communicate with foreigners from other countries, it would be difficult to understand what they were saying. This is because humans use the language they have and so do strangers, so there is no common ground in communication. Therefore, English is used as a global language in communication. So, people in the world can communicate and understand what is being said. One technology that can be used to improve learning is computers (Irnanda & Windarto, 2020).

In Indonesia, everyone has learned English since elementary school (SD) or some have learned it from kindergarten. Even in Higher Education, it is applied to provide English courses to students even though the output of the study is not related to English. From this, it can be concluded that the importance of proficiency in foreign languages, especially English is a factor of success in one's academic or to get a good career in work (Megawati, 2016) . However, even though they have been studying English for a long time, there are still many students who find it difficult to learn English.

Students have difficulty learning English. Difficulties that can be experienced, such as difficulties in reading, writing, listening and others. Sometimes there are students who still feel insecure about applying the language in everyday life. The reason is because they are afraid of being ridiculed and considered showing off that they are able to speak English.

There are also problems that are often experienced by students, namely difficulties in reading vocabulary and applying grammar in English. The teaching role is very necessary to be able to improve students' English skills.

With these problems, the authors took the initiative to determine what factors could improve students' skills in English. There are many branches of computer science that can be used to solve complex problems. This branch of computer science is Artificial Intelligence like datamining (Branch, Widyastuti, Gracella, Simanjuntak, & Hartama, 2019; Katrina, Damanik, & Parhusip, 2019; Rofiqo, Windarto, & Hartama, 2018; Sadewo, Windarto, & Wanto, 2018; Sari, Wanto,

& Windarto, 2018; Series, 2019; Swarm, Analysis, Problem, & Prediction, 2018). Based on the above problems, the authors make use of data mining, namely classification in solving problems to improve students' English skills. The algorithm for data mining used is the C4.5 Algorithm (Haryati, Sudarsono, & Suryana, 2015). The purpose of this research is to determine factors that can improve students' English skills.

2. Research Methods

The following are the steps in completing this research, namely:

1. Create the required dataset.
2. Apply the required algorithm to the existing dataset.
3. Perform a manual count.
4. Validate with Rapidminer software.

2.1 Create a Dataset

The dataset is obtained by providing a questionnaire created from Google Form. The questionnaire was given to students to complete the completion of this study.

Table 1. Respondent Data

No	K1	K2	K3	K4	Ability In English
1.	3	4	5	1	Capable
2.	4	3	1	2	Capable
3.	1	5	5	3	Capable
4.	3	5	2	2	Incapable
5.	5	4	4	3	Capable
6.	4	1	2	2	Incapable
7.	5	4	5	5	Capable
8.	4	5	5	3	Capable
9.	3	4	3	1	Incapable
10.	1	3	2	3	Capable
11.	4	2	3	3	Capable
12.	5	5	3	5	Incapable
13.	2	2	3	3	Incapable
14.	4	4	5	5	Capable
15.	5	5	3	5	Capable
16.	1	4	3	2	Incapable
17.	4	1	4	4	Incapable
18.	5	2	2	3	Incapable
19.	2	5	5	1	Incapable
20.	4	4	5	4	Capable
21.	4	3	1	2	Capable
22.	1	5	5	3	Capable
23.	3	5	2	2	Incapable
24.	4	4	5	5	Capable
25.	5	5	3	5	Capable
26.	1	4	3	2	Incapable
27.	4	1	4	4	Incapable
28.	5	2	2	3	Incapable
29.	2	5	5	1	Incapable
30.	1	5	5	3	Capable
31.	3	5	2	2	Incapable
32.	5	4	4	3	Capable
33.	3	4	3	1	Incapable
34.	1	3	2	3	Capable
35.	4	2	3	3	Capable
36.	5	4	1	5	Incapable
37	1	4	3	2	Incapable
38.	4	1	4	4	Incapable
39	4	1	2	2	Incapable
40.	5	4	5	5	Capable
41.	4	4	5	3	Capable
42.	3	4	5	1	Incapable
43.	1	3	2	3	Capable
44.	3	4	5	1	Capable
45.	4	3	1	2	Capable
46.	2	5	5	1	Incapable
47.	5	1	4	5	Incapable
48.	2	2	3	3	Incapable
49.	4	4	5	5	Capable
50.	5	5	3	5	Capable
51.	2	5	5	1	Incapable
52.	4	4	5	4	Capable
53.	4	3	1	2	Capable
54.	2	2	3	3	Incapable
55.	5	4	4	3	Incapable
56.	5	4	5	5	Capable
57.	4	1	2	2	Incapable
58.	5	4	5	5	Capable
59.	4	1	2	2	Incapable
60.	5	4	5	5	Capable
61.	4	3	1	2	Capable
62.	1	5	5	3	Capable
63.	3	5	2	2	Incapable
64.	5	4	4	3	Capable
65.	1	4	3	2	Incapable
66.	4	1	4	4	Incapable
67.	4	1	2	2	Incapable
68.	2	5	5	2	Incapable
69.	4	3	5	4	Capable
70.	4	3	1	2	Capable

This study collected data from 70 respondents by making 4 criteria, namely: Reading References (K1), Hearing from the Environment (K2), Practicing in the Environment (K3), and Utilizing Technology (K4). With sub criteria as follows:

- 5 = Strongly Agree
- 4 = Agree
- 3 = Quite Agree
- 2 = Disagree
- 1 = Very Disagree

ISSN: 2541-1004
e-ISSN: 2622-4615
10.32493/informatika.v5i3.5268

2.2 Applying the Algorithm Used

In this study using C4.5 Algorithm in determining factors that influence students in improving students' English skills. C4.5 algorithm is an algorithm found in the classification technique to solve cases or problems. Decision tree (decision tree) is the basis of C4.5 Algorithm. C.45 algorithm is a decision tree induction algorithm, ID3 (Iterative Dichotomiser 3) (Febriarini & Astuti, n.d.).

The formulas used in C4.5 Algorithm are:

- Calculate entropy

$$Entropi\ (S) = \sum_{j=1}^{k} - p_j\ \log_2\ p_j \quad (1)$$

Explanation:

- S = dataset (case)
- k = number of dataset partitions

- Calculate the gain

$$Gain\ (A) = Entropi\ (S) - \sum_{i=1}^{k} \frac{|S_i|}{|S|} \times Entropi(S_i) \quad (2)$$

Explanation:

- S = dataset (case)
- A = attributes
- N = number of attribute partitions
- | S1 | = large number of i-partition cases
- | S | = number of cases of S

The advantage of using C4.5 Algorithm is that it can make a decision tree so that it is efficient, where the decision tree handles the discrete type and discrete-numeric type attributes, is easy to interpret and has an acceptable level of accuracy (Kamber, n.d.). The weakness of the C4.5 algorithm is that there is scalability in that training data can only be used and stored as a whole at the same time in memory (Luvia, Windarto, Solikhun, & Hartama, 2017).

2.3 Perform Manual Counts

In this study, manual calculations are performed using existing formulas and done using *Microsoft Excel*. When doing manual counting, it must be done carefully in order to obtain correct results and in accordance with the Rapidminer software.

2.4 Validation

The validation process is carried out using Rapidminer software. Validation is done by taking a sample of the dataset owned by as many as 70 respondents from questionnaire data given to students and then the data is tested using a decision tree (C4.5).

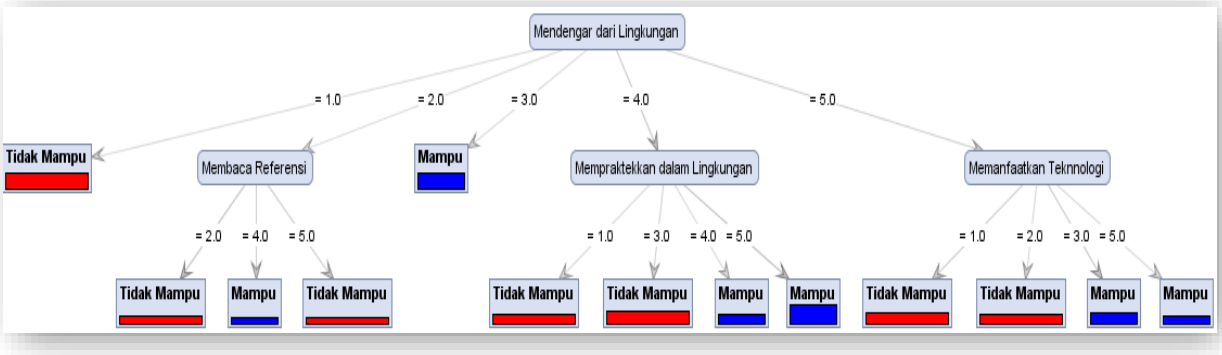

Figure 1. Decision tree from a dataset of 70 respondents

Decision Tree uses performance in the form of a tree (tree) where in each node that is owned to explain the existing attributes, branches interpret the values of existing attributes and leaves present the class (Yulia & Putri, 2019). In the decision tree there is what is known as root. Root is the node at the top of the decision tree. Decision Tree is the best-known data classification technique for use in data mining. Decision tree use relatively fast development. The output of this type of model is made to be easily understood.

3. Result and Discussion

Classification is performed on the respondent's data. The calculation is done using Microsoft Excel software. Then, then perform data calculations using the formula in C4.5 Algorithm. The results of calculations on the 1st iteration can be see in Table 2.

Table 2. Results of the 1st Iteration Calculation

		Jlh Kasus (S)	Mampu (S1)	Tidak Mampu (S2)	Entropy	Gain
Node 1		70	36	34	0,999411	
Membaca Referensi	Sangat Setuju	17	11	6	0,936667	
	Setuju	25	16	9	0,942683	
	Cukup Setuju	9	2	7	0,764205	
	Kurang Setuju	8	0	8	0	
	Sangat Kurang Setuju	11	7	4	0,94566	
						0,188403
Mendengar dari Lingkungan	Sangat Setuju	18	9	9	1	
	Setuju	27	16	11	0,975119	
	Cukup Setuju	9	9	0	0	
	Kurang Setuju	7	2	5	0,863121	
	Sangat Kurang Setuju	9	0	9	0	
						0,279838
Mempraktikkan dari Lingkungan	Sangat Setuju	23	18	5	0,755375	
	Setuju	8	4	4	1	
	Cukup Setuju	15	5	10	0,918296	
	Kurang Setuju	14	3	11	0,749595	
	Sangat Kurang Setuju	10	6	4	0,970951	
						0,151526
Memanfaatkan Teknologi	Sangat Setuju	14	10	4	0,863121	
	Setuju	7	3	4	0,985228	
	Cukup Setuju	20	15	5	0,811278	
	Kurang Setuju	19	6	13	0,899744	
	Sangat Kurang Setuju	10	2	8	0,721928	
						0,149121

Table 3. Results of the 2nd Iteration Calculation

		Jlh Kasus (S)	Mampu (S1)	Tidak Mampu (S2)	Entropy	Gain
Node 2		18	9	9	1	
Membaca Referensi	Sangat Setuju	3	3	0	0	
	Setuju	2	2	0	0	
	Cukup Setuju	4	0	4	0	
	Kurang Setuju	5	0	5	0	
	Sangat Kurang Setuju	4	4	0	0	
						1
Mempraktikkan dari Lingkungan	Sangat Setuju	11	6	5	0,9940302	
	Setuju	0	0	0	0	
	Cukup Setuju	3	3	0	0	
	Kurang Setuju	4	0	4	0	
	Sangat Kurang Setuju	0	0	0	0	
						0,3925371
Memanfaatkan Teknologi	Sangat Setuju	3	3	0	0	
	Setuju	0	0	0	0	
	Cukup Setuju	6	6	0	0	
	Kurang Setuju	4	0	4	0	
	Sangat Kurang Setuju	5	0	5	0	
						1

In Table 2 is the result of the calculation in the 1st iteration. It can be seen that the criteria of Hearing from the Environment get the highest gain with a value of 0.27984 and the highest entopry value is "Strongly Agree" with a value of 1.

From Table 2, because there is the highest root gain, then proceed to the 2nd iteration. The results of the count in the 2nd iteration can be see in Table 3.

In Table 3 is a calculation on the second iteration and can be seen that the same gain has been obtained in the criteria of Reading Reference and Utilizing Technology with the gain value is 1. And the Practicing criteria in the Environment get the lowest gain with a value of 0.3925371.

If the entropy has a value of 0 in one of the Universe, it indicates that it already has leaves.

In this study, we validated the Rapidminer software. We use a decision tree and produce a decision tree as follows:

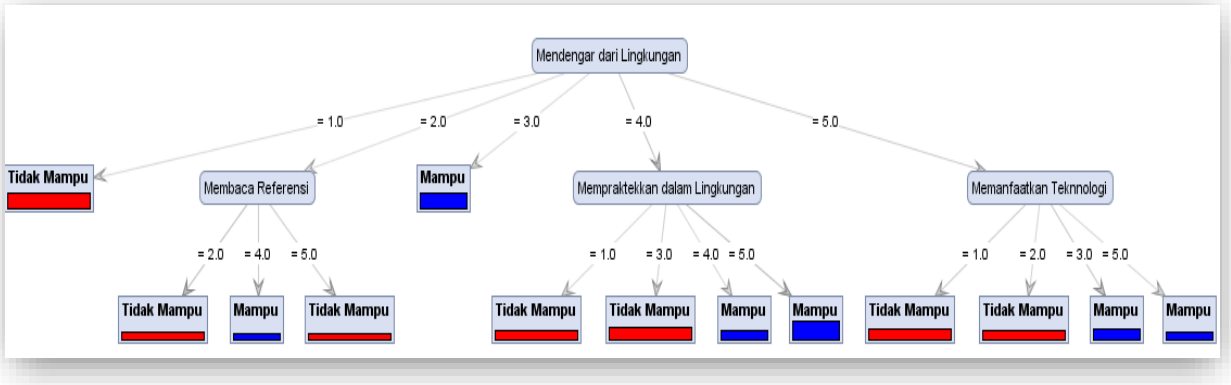

Figure 2. Decision tree in rapidminer

Based on the decision tree above produced a decision tree for increasing students' English ability in text form, as follows:

Mendengar dari Lingkungan = 1.0: Tidak Mampu {Mampu=0, Tidak Mampu=9}
Mendengar dari Lingkungan = 2.0
| Membaca Referensi = 2.0: Tidak Mampu {Mampu=0, Tidak Mampu=3}
| Membaca Referensi = 4.0: Mampu {Mampu=2, Tidak Mampu=0}
| Membaca Referensi = 5.0: Tidak Mampu {Mampu=0, Tidak Mampu=2}
Mendengar dari Lingkungan = 3.0: Mampu {Mampu=9, Tidak Mampu=0}
Mendengar dari Lingkungan = 4.0
| Mempraktekkan dalam Lingkungan = 1.0: Tidak Mampu {Mampu=0, Tidak Mampu=4}
| Mempraktekkan dalam Lingkungan = 3.0: Tidak Mampu {Mampu=0, Tidak Mampu=7}
| Mempraktekkan dalam Lingkungan = 4.0: Mampu {Mampu=4, Tidak Mampu=0}
| Mempraktekkan dalam Lingkungan = 5.0: Mampu {Mampu=12, Tidak Mampu=0}
Mendengar dari Lingkungan = 5.0
| Memanfaatkan Teknnologi = 1.0: Tidak Mampu {Mampu=0, Tidak Mampu=5}
| Memanfaatkan Teknnologi = 2.0: Tidak Mampu {Mampu=0, Tidak Mampu=4}
| Memanfaatkan Teknnologi = 3.0: Mampu {Mampu=6, Tidak Mampu=0}
| Memanfaatkan Teknnologi = 5.0: Mampu {Mampu=3, Tidak Mampu=0}

4. Conclusion

From the results of this study, using the C4.5 Algorithm can make it easier to determine what factors can improve English skills in students. Researchers have completed this research and it can be concluded that the factors that influence to improve students' ability to English are the criteria of "Hearing from the Environment" with the iteration calculation process stopping at the 2nd iteration. With this research, it can provide motivation to others, especially students, namely to improve English language skills so that they can be used in the coming year and become a source of employment.

References

Branch, P., Widyastuti, M., Gracella, A., Simanjuntak, F., & Hartama, D. (2019). *Classification Model C . 45 on Determining the Quality of Custumer Service in Bank*

BTN Classification Model C . 45 on Determining the Quality of Custumer Service in Bank BTN Pematangsiantar Branch. https://doi.org/10.1088/1742-6596/1255/1/012002

Febriarini, A. S., & Astuti, E. Z. (n.d.). *Penerapan Algoritma C4 . 5 untuk Prediksi Kepuasan Penumpang Bus Rapid Transit (BRT) Trans Semarang*. 95–103. https://doi.org/10.30864/eksplora.v8i2.156

Haryati, S., Sudarsono, A., & Suryana, E. (2015). *Implementasi Data Mining untuk Memprediksi Masa Studi Mahasiswa Menggunakan Algoritma C4.5 (Studi Kasus: Universitas Dehasen Bengkulu)*. *11*(2), 130–138. https://doi.org/10.37676/jmi.v11i2.260

Irnanda, K. F., & Windarto, A. P. (2020). Penerapan Klasifikasi C4.5 Dalam Meningkatkan Kecakapan Berbahasa Inggris dalam Masyarakat. *Seminar Nasional Teknologi Komputer & Sains (SAINTEKS)*, 304–308. Medan: STMIK Budi Darma.

Kamber, M. (n.d.). *Data Mining : Concepts and Techniques Third Edition*.

Katrina, W., Damanik, H. J., & Parhusip, F. (2019). *C . 45 Classification Rules Model for Determining Students Level of Understanding of the Subject C . 45 Classification Rules Model for Determining Students Level of Understanding of the Subject*. https://doi.org/10.1088/1742-6596/1255/1/012005

Luvia, Y. S., Windarto, A. P., Solikhun, S., & Hartama, D. (2017). Penerapan Algoritma C4.5 untuk Klasifikasi Predikat Keberhasilan Mahasiswa di AMIK Tunas Bangsa. *Jurasik (Jurnal Riset Sistem Informasi Dan Teknik Informatika)*, *1*(1), 75–79. https://doi.org/10.30645/jurasik.v1i1.12

Megawati, F. (2016). Kesulitan Mahasiswa dalam Mencapai Pembelajaran Bahasa Inggris Secara Efektif. *PEDAGOGIA: Jurnal Pendidikan*, *5*(2), 147–156. https://doi.org/10.21070/pedagogia.v5i2.246

Rofiqo, N., Windarto, A. P., & Hartama, D. (2018). Penerapan Clustering pada Penduduk yang Mempunyai Keluhan Kesehatan dengan Datamining K-Means. *KOMIK (Konferensi Nasional Teknologi Informasi Dan Komputer)*, *2*(1), 216–223. https://doi.org/10.30865/komik.v2i1.929

Sadewo, M. G., Windarto, A. P., & Wanto, A. (2018). Penerapan Algoritma Clustering dalam Mengelompokkan Banyaknya Desa/Kelurahan Menurut Upaya Antisipasi/Mitigasi Bencana Alam Menurut Provinsi dengan K-Means. *KOMIK (Konferensi Nasional Teknologi Informasi Dan Komputer)*, *2*(1), 311–319. https://doi.org/10.30865/komik.v2i1.943

Sari, R. W., Wanto, A., & Windarto, A. P. (2018). Implementasi Rapidminer dengan Metode K-Means (Study Kasus: Imunisasi Campak pada Balita Berdasarkan Provinsi). *KOMIK (Konferensi Nasional Teknologi Informasi Dan Komputer)*, *2*(1), 224–230. https://doi.org/10.30865/komik.v2i1.930

Series, C. (2019). *The Application of Data Mining in Determining Patterns of Interest of High School Graduates The Application of Data Mining in Determining Patterns of Interest of High School Graduates*. https://doi.org/10.1088/1742-6596/1339/1/012042

Swarm, P., Analysis, P. C., Problem, O. D., & Prediction, E. P. (2018). *Data mining tools | rapidminer : K-means method on clustering of rice crops by province as efforts to stabilize food crops in Indonesia*. https://doi.org/10.1088/1757-899X/420/1/012089

Yulia, Y., & Putri, A. D. (2019). Data Mining Menggunakan Algoritma C4.5 untuk Memprediksi Kepuasan Mahasiswa Terhadap Kinerja Dosen di Kota Batam. *Computer Based Information System Journal*, *7*(2), 56–66. https://doi.org/10.33884/cbis.v7i2.1373