TRESCODE Green Organization
JEST TGO Journal of Education, Science and Technology
P-ISSN : 2987 - 4580
E-ISSN : 2987 - 4599
https://ejournal.trescode.org/index.php/jest

Implementation of K-Means Clustering for Analysis Students English Proficiency

Mas'ud Hermansyah[1]*, Nur Andita Prasetyo[2], Yulian Ansori[3], M. Faiz Firdausi[4], Abdul Wahid[5]

Information Systems and Technology, Faculty of Science, Technology and Industry, Mandala Institute of Technology and Science, Jember, Indonesia[1245]

Informatics, Faculty of Engineering, Primagraha University, Serang, Indonesia[3]

Corresponding Author : Mas'ud Hermansyah (masudhermansyah@itsm.ac.id)

Article Info

Received:
March 13, 2023
Revised:
April 22, 2023
Online available:
May 9, 2023

Keyword:
Clustering, K-Means, Davies Bouldin Index

Abstract

English is one of the international languages. Language is a communication tool that is carried out orally or in writing. English proficiency is not only the ability to speak, but also the ability to understand and produce spoken or written texts which are realized in the four language skills namely listening, speaking, reading and writing. With the existence of data mining technology, an analysis of students' English skills can be carried out. This analysis was carried out by grouping students according to ability scores in these empathy skills. In conducting this research, the K-Means clustering method was used to classify students' English skills. With the K-Means clustering technique, it is hoped that the teacher can adjust the learning model according to the students' abilities. Based on the grouping results, the grouping with 3 clusters is the most optimal grouping result with the smallest Davies Bouldin Index (DBI) value, namely 0.365. The application of the K-Means method in grouping student data based on English proficiency scores can produce 3 groups of students who are smart, moderate, and moderate.

Cite this as: Hermansyah, M., Prasetyo, N. A., Ansori, Y., Firdausi, M. F., & Wahid, A. (2023). Implementation of K-Means Clustering for Analysis Students English Proficiency. TGO Journal of Education, Science and Technology, 1(1), 31–35

INTRODUCTION

Indonesia has entered the free market era of the MEA (Asean Economic Community) since early 2015 and the ASEAN economic community or AEC (Asean Economy Community) has agreed that the business language used is English. This is a record for all Indonesian citizens to think hard about how to prepare English language skills, in this case speaking English, so they are able to face the free market (Budiarso, 2017). English is one of the international languages. Language is a communication tool that is carried out orally or in writing. English proficiency is not only the ability to speak, but also the ability to understand and produce spoken or written texts which are realized in the four language skills namely listening, speaking, reading and writing. For high school students, English is a compulsory subject that is taught in developing students' knowledge, language skills, and positive attitudes towards English. So that the English given is presented in an interesting, quality and in accordance with its development. Along with the current development of globalization, English is highly prioritized for high school

students so that these students are able to compete in the field of science and are able to compete with other countries. Students who have the ability to speak English well can communicate their ideas and ideas in the school environment or with foreigners. However, there are still many students at the high school level who still experience difficulties in conveying ideas, thoughts and questions in English using good and correct spoken language (Tambusai & Nasution, 2022).

English is often a problem for students. (Silalahi et al., 2022) in his research explained that the first reason that is most often stated is because English is not the mother tongue so it is difficult to pronounce it. The second reason is feeling lazy to practice listening, speaking, reading and writing so that it makes English even more difficult to understand. This second reason should be a provision for teaching English in class. However, some educators often forget to present the English language needs according to the needs of their students. The aim of this subject is to equip students with active communication skills in English, namely the ability to listen, read and write. Mastery of English is also a means to boost Indonesia's human resources, which according to the Human Development Index are in the lowest category in Asia. Global competition in all English that demands an increase in the quality of human resources, including teaching staff, as the spearhead. The school's output must really be of good quality in order to be competitive and have a high bargaining position. One of the efforts to realize the above is to improve the quality of learning English. Mastery of English will open their horizons to the development of science and technology, including education which is currently easily accessible from various sources.

With the existence of data mining technology, an analysis of students' English skills can be carried out. This in data mining that is useful in analyzing data to make it more accurate in solving data grouping problems or dividing a data set into several subsets. The purpose of clustering is to assign data to a group so that the relationship between members in the same cluster becomes strong, while the relationship between members in different clusters becomes weak (Agustina et al., 2012). Objects in a cluster have similar characteristics but have different characteristics from objects in other clusters. Therefore, clustering is very useful in assigning unknown groups or clusters to the data.

In conducting this research, the K-Means clustering method was used to classify students' English skills. With the K-Means clustering technique, it is hoped that schools and parties in the education sector can record students with their respective abilities and can be taught using the right method so as to improve the quality of students' English and student academic achievement. The focus of this research is on grouping student achievement at Vocational High School Lab Business School Tangerang with the K-Means clustering method, where students are grouped according to their level of ability to value listening, speaking, reading and writing.

METHOD

The method used in this study is the data mining clustering method using the K-Means algorithm. Data mining is a series of processes in finding patterns, relationships, extracting added value from large data and information in the form of knowledge with the aim of finding relationships and simplifying data in order to obtain understandable and useful information with the help of statistics and mathematics (Farahdinna et al., 2019).

Clustering is work that separates data or vectors into a number of groups (clusters) according to their respective characteristics. Data with similar characteristics will gather in the same cluster, and data with different characteristics will be separated in different clusters. There is no need for a class label for each data that is processed in a clustering, because later a new label can be assigned when the cluster has been formed. Because there is no target class label for each data, clustering is often called unsupervised learning (Lase & Panggabean, 2019).

TRESCODE Green Organization
JEST TGO Journal of Education, Science and Technology
P-ISSN : 2987 - 4580
E-ISSN : 2987 - 4599
https://ejournal.trescode.org/index.php/jest

K-Means algorithm is an algorithm that works by partitioning data into clusters, so data that is similar is in the same cluster and data that is dissimilar is in another cluster (Rohmawati et al., 2015).
The following are the steps contained in the K-Means algorithm:

1. Determine the number of clusters (k), set the cluster center randomly.
2. Calculate the distance of each data to the center of the cluster
3. Group data into clusters with the shortest distance.
4. Compute the new cluster center.
5. Repeat steps 2 (two) to 4 (four) so that no more data is moved to another cluster.

The clustering process begins with identifying clustered data, using the Euclidean Distance formula as shown in equation (1).

$$d_{ij} = \sqrt{(X_{1i} - X_{1j})^2 + (X_{2i} - X_{2j})^2 + \cdots + (X_{ki} - X_{kj})^2} \quad (1)$$

Information:
D (i,j) = Distance between i to cluster data center j
X ki = data to i on attribute data to k
X kj = center point j on a attribute k

$$C = \frac{\sum m}{n} \quad (2)$$

Equation 2 explains where C is a data centroid, m is a data member belonging to a certain centroid and n is the number of data members belonging to a certain centroid.
Davies bouldin index (DBI) is a method introduced by David L. Davies and Donald W. Bouldin. The Davies Bouldin Index is used to evaluate clusters in general based on the quantity and proximity between cluster members. Calculation of the Davies Bouldin Index value is based on a comparison of the ratio of the i-th cluster and the j-kle cluster. The smaller the Davies Bouldin Index value, the better the resulting cluster (Adhitama et al., 2020). The calculation of the DBI value is presented in equation (3).

$$DBI = \frac{1}{k}\sum_{i=1}^{k} max_{i \neq j}(R_{ij})$$

RESULTS AND DISCUSSION

The data in this study were sourced from grade 11 English grades in the multimedia expertise program at Vocation High School Lab Business School in 2022. The data went through a data pre-processing stage before being used to carry out the clustering process on several attributes. Of the 4 attributes used in the calculation process, namely the ability to value listening, speaking, reading and writing.
Cleansing data is to reduce noise that can affect calculations. In the data cleansing process, data that has vacant values on 4 attributes are not used, so the data used is 72 students out of 76 students. The data that has been processed for data cleansing can be seen in Table 1.

Table 1. Data Cleansing Results

No. Absen	Listening	Speaking	Reading	Writing
1.	86	85	88	82
2.	79	72	76	74
3.	68	72	72	78
4.	84	83	85	86
5.	80	73	80	76
…	…	…	…	…

TRESCODE Green Organization
JEST TGO Journal of Education, Science and Technology
P-ISSN : 2987 - 4580
E-ISSN : 2987 - 4599
https://ejournal.trescode.org/index.php/jest

…	…	…	…	…
71.	74	78	74	79
72.	70	75	80	68

Source: student scores

This study aims to find the best cluster distribution by measuring the DBI value to classify students' abilities in English proficiency. Furthermore, the mining process is carried out with the aim of finding information or patterns for clustering using the K-Means algorithm. The implementation of the K-Means algorithm in this study uses the Rapidminer software. Based on research (Hermansyah et al., 2020) K-Means clustering modeling with Rapidminer can be seen in Figure 1

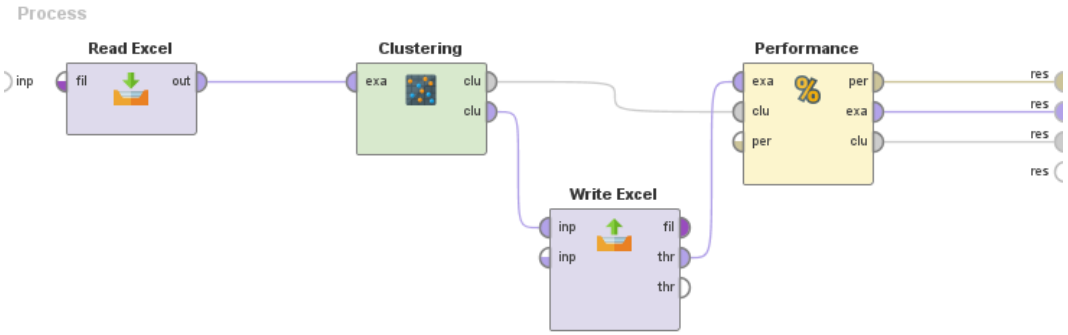

Figure 1. K-Means Clustering Modeling with Rapidminer

Source: data processing results

1. Read Excel is an operator to read ExampleSet from the specified Excel file.
2. Clustering is an operator that performs a grouping process using the K-Means algorithm.
3. Write Excel is an operator for creating clustering results reports with file type.xlsx.
4. Performance is the operator used to evaluate the performance of the clustering method based on the centroid.

In this step, 4 experiments were carried out with the clustering process with the number of clusters starting from 2, 3, 4, and 5 clusters.

Tabel 2. Results of K-Means Clustering Modeling with Rapidminer

Number of Clusters	Name of Cluster	Number of Members
2 cluster	Cluster 1	57
	Cluster 2	15
3 cluster	Cluster 1	24
	Cluster 2	35
	Cluster 3	13
4 cluster	Cluster 1	13
	Cluster 2	16
	Cluster 3	19
	Cluster 4	24
5 cluster	Cluster 1	5
	Cluster 2	9
	Cluster 3	24
	Cluster 4	8
	Cluster 5	16

Source: data processing results

The final step is to explain the meaning of each cluster formed. This step is done by dividing the centroid of the last iteration by the number of members in each cluster. So that it can be concluded the description of each cluster. In addition, this step also

evaluates clusters using the Davies Bouldin Index method to calculate the average value of each point in the data set.

Several experiments were carried out on the model that was built. The first model is clustering the dataset without normalizing it. The number of clusters is determined from 2, 3, 4, and 5. The segmentation results formed will be evaluated using the Davies Bouldin Index. Davies Bouldin Index is a cluster validation method from clustering results.

Tabel 3. Davies Bouldin Index Value of Experiment Results

Number of Clusters	DBI value
2 cluster	0,404
3 cluster	0,365
4 cluster	0,454
5 cluster	0,599

Source: data processing results

From the results of the experiments conducted, the K-Means algorithm with Rapidminer, the number of 3 clusters produces better quality clusters compared to the number of clusters 2, 4, and 5. The results of the cluster evaluation show that the K-Means algorithm with the number of clusters 3 is more optimal with a value The smallest DBI, at 0.365.

Based on table 2, the results of this comparison get an analysis that the score of English skills with the 3 best clusters classifies the smart cluster as many as 24 students, the medium cluster as many as 35 students, and as many as 13 students as sufficient. The results of the cluster grouping are used as a reference for grouping study groups.

CONCLUSION

The application of the K-Means method in grouping student data based on English proficiency scores can produce groups of smart, moderate, and moderate students. The results of clustering can be a reference for English teachers to set learning strategies for each study group according to the cluster. Based on the grouping results, the grouping with 3 clusters is the most optimal grouping result with the smallest DBI value, namely 0.365.

REFERENCES

Adhitama, R., Burhanuddin, A., & Ananda, R. (2020). Penentuan Jumlah Cluster Ideal Smk Di Jawa Tengah Dengan Metode X-Means Clustering Dan K-Means Clusterin. *JIKO (Jurnal Informatika Dan Komputer)*, *3*(1), 1–5. https://doi.org/10.33387/jiko.v3i1.1635

Agustina, S., Yhudo, D., Santoso, H., Marnasusanto, N., Tirtana, A., & Khusnu, F. (2012). Clustering Kualitas Beras Berdasarkan Ciri Fisik Menggunakan Metode K-Means. *Clustering K-Means*, 1–7.

Budiarso, I. (2017). Analisis Kemampuan Keterampilan Berbicara Bahasa Inggris terhadap Kinerja Karyawan PT Berrys Internasional Jakarta. *JABE (Journal of Applied Business and Economic)*, *3*(1), 1. https://doi.org/10.30998/jabe.v3i1.1752

Farahdinna, F., Nurdiansyah, I., Suryani, A., & Wibowo, A. (2019). Perbandingan Algoritma K-Means Dan K-Medoids Dalam Klasterisasi Produk Asuransi Perusahaan Nasional. *Jurnal Ilmiah FIFO*, *11*(2), 208. https://doi.org/10.22441/fifo.2019.v11i2.010

Hermansyah, M., Hamdan, R. A., Sidik, F., & Wibowo, A. (2020). Klasterisasi Data Travel Umroh di Marketplace Umroh.com Menggunakan Metode K-Means. *Jurnal Ilmu Komputer*, *13*(2), 8. https://doi.org/10.24843/jik.2020.v13.i02.p06

Lase, Y., & Panggabean, E. (2019). Implementasi Metode K-Means Clustering Dalam Sistem Pemilihan Jurusan Di SMK Swasta Harapan Baru. *JUTIKOMP : Jurnal Teknologi Dan Ilmu Komputer Prima*, *2*(2), 375–379.

Rohmawati, N., Defiyanti, S., & Jajuli, M. (2015). Implementasi Algoritma K-Means Dalam Pengklasteran Mahasiswa Pelamar Beasiswa. *Jitter : Jurnal Ilmiah Teknologi Informasi Terapan*, *I*(2), 62–68.

Silalahi, M., Purba, A., Benarita, B., Matondang, M. K. ., Sipayung, R. W., Silalahi, T. F., Saragih, N., Girsang, S. E., Damanik, I. J., & Sibuea, B. (2022). Analisis Kesulitan Belajar Bahasa Inggris Siswa Sma Negeri 1 Narumonda Kabupaten Tobasa. *Community Development Journal : Jurnal Pengabdian Masyarakat*, *3*(2), 728–732. https://doi.org/10.31004/cdj.v3i2.4686

Tambusai, A., & Nasution, K. (2022). Tingkat Pemahaman Bahasa Inggris Bagi Siswa Sekolah Menegah Atas (SMA). *Jurnal Pema Tarbiyah*, *1*(1), 44–53.