The IJICS (International Journal of Informatics and Computer Science)
Vol 3 No 2, September 2019
ISSN 2548-8384 (online)
ISSN 2548-8449 (print)
Page 49 – 58
DOI 10.30865/ijics.v3i2.1382

Implementation of K-Means Methods In Clustering Students Ability Levels in English Language

Cahyo Prianto, Rd. Nuraini, Andi Tenri Wali*

Department of Computer Science, Politeknik Pos Indonesia, Bandung, Indonesia
Department of Computer Science, Politeknik Pos Indonesia, Bandung, Indonesia
Email: [1]cahyoprianto@poltekpos.ac.id, [2*] nuraini@poltekpos.ac.id, [2*] anditenriwali86@gmail.com.

Abstract- Nowadays, English extremely needs to be controlled, especially students, in communicating and reading also understanding literature written in English. In achieving mastery of English, the students, in this case, the students who are not majoring in English are given a common base subject of English. In Politeknik Indonesia, especially majoring in a Bachelor's Degree in Informatics Engineering, teaching English is using the direct method, to find out the results of teaching English within three semesters. Therefore, by doing this research for classifying the level of ability of students into three categories, they are Beginner, intermediate and advanced. The objective of the grouping is to determine how many students who have the capability level is low, medium and high so that the faculty can determine the average level of students' proficiency and the lecturers can intervene to conduct teaching in developing the students' knowledge of English. The classification used the K-Means clustering algorithm, which is one algorithm that classifies the same data on specific groups and different data in the other group. The results of this study by applying the k-means clustering method is the researchers can classify the students based on students' ability levels either they are beginner, intermediate or advanced.

Keywords: English, Level Capabilities, Clustering, Politeknik Pos Indonesia, K-Means method.

1. INTRODUCTION

As we know, English is the most dominant of the international language, so it has become a hope for many people to be able to understand and to communicate in English well. In Indonesia, English is the only foreign language that must be learned from pre-school to university level. Although it has been studied in a long period, there are still many students who have problems in learning English. They had difficulty in speaking, listening, reading and writing in English. This is due to a lack of confidence in speaking English and fears of making mistakes in using grammar, mentioning vocabulary, pronouncing and any others.

Currently, English is very needed to be mastered especially for students, in communicating and reading also understanding literature that written in English. In achieving mastery of English, the students, in this case, the students who are not majoring in English are given a common base subject of English. In Politeknik Pos Indonesia, especially the majoring of Bachelor Degree in Informatics Engineering, for instance, teaching English is conducted within three semesters, namely 1st semester General English 1, 2nd semester General English 2, and 3rd semester General English 3 by using the direct method. To find out the results of teaching English within three semesters, the researchers conducted this research by classifying the level of ability of students into three categories, they are Beginner, intermediate and advanced.

By grouping the students, the student's ability level can be known by seeing their fluency, comprehension, and grammar. The objective of the grouping is to determine how many students have low, medium and high of English level ability. Therefore, the faculty can determine the average level of students' proficiency so that lecturers can intervene in the conduct of teaching in developing the students' knowledge in English. By doing the grouping, the researchers apply the K-means method. K-Means method is needed for students to be able to determine the grouping criteria that could be a reference [1] This k-means clustering algorithm is one of the most widely used in clustering techniques [2] and grouping data into specific figures partition clusters (groups, subsets, or category) [3]. so that the results of this research in applying the k-means clustering method, researchers can classify students based on students' ability levels neither are Beginner, intermediate, and advanced.

2. RESEARCH METHOD

2.1 DSRM Method (Design Science Research Methodology)

A research methodology is a research common approach in taking and implementing research projects. The type of research methodology that is used by the researchers was Design Science Research Methodology (DSRM), this method was used for the framework of the procedures used as well as the understanding of the process of

ISSN 2548-8384 (online)
ISSN 2548-8449 (print)
Page 49 – 58
DOI 10.30865/ijics.v3i2.1382

review to identify and evaluate the results of the research. Design Science Research Methodology (DSRM) consists of seven process methods that researches do [4].

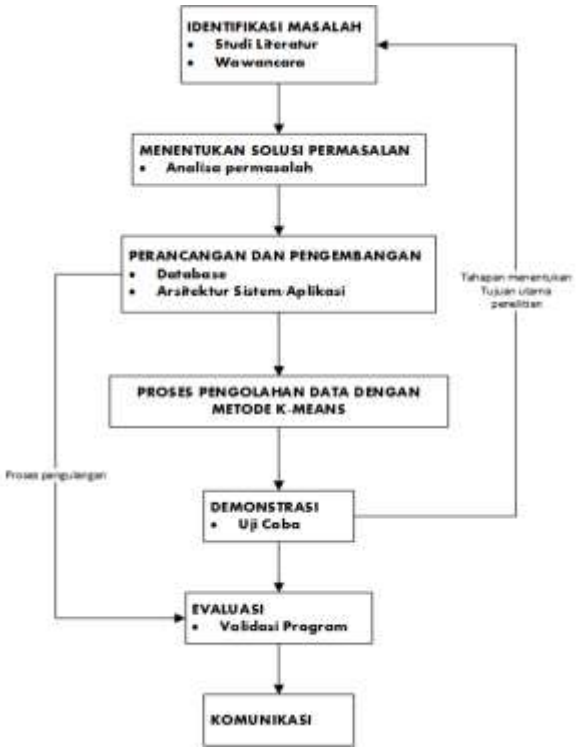

Figure 1. The way of the DSRM Method

2.2.1 Identification of Problems

In defining research problems and finding solutions for the problem. Two things are done to find a solution is to focus on research and client in the study, which will be to find solutions and help to understand the reasons researchers in understanding the problem. The resources required for these activities include knowledge about the state of the problem and the importance of the solutions. The step that is used in identifying a problem is doing data collection, which consists of:

1. The study of literature, searching for references via the internet and journals of national or international.
2. The interview, in this study the researches conducted interviews with English lecturer D4 Informatics Engineering at the Politeknik Pos Indonesia.

2.2.2 Determining the Problem Solution

Summing up the purpose of the existing problems, where the goal is later expected to be better than the present, or how this new artifact can support the settlement of problems that are now handled. The resources required for this phase include knowledge about the state of the current problem.

2.2.3 Design and Development

This stage is making artifacts, namely building models or methods or new properties of technical, social, and or information resources. Conceptually, the design of research artifacts can be in the form of objects that are designed and designed by researchers. This activity includes determining the function of the desired artifact and the architecture of the actual artifact making. At the design stage, the database is carried out, and the design of the application process uses UML.

2.2.4 Processing Data with K-Means

This stage is about how the data is processed using the K-Means, Data clustering using the k-means method is generally done with the basic algorithm.

2.2.5 Demonstration

ISSN 2548-8384 (online)
ISSN 2548-8449 (print)
Page 49 – 58
DOI 10.30865/ijics.v3i2.1382

Demonstrate the use of artifacts to solve one or more of the problems that exist. This could involve the person in charge of the company.

2.2.6 Evaluation

Observe and measure how well the artifacts in solving this problem. This activity involves, comparing both the actual purpose of the results observed in the use of artifacts when demonstrations. This stage requires knowledge of the relevant measure and analysis techniques, depending on the nature of the problem and artifacts, the evaluation can take many forms.

2.2.7 Communication

Communicating the problem and the importance of artifacts between researchers and others interested in the publication of scientific research. Researchers may use the reports or scientific journals such as empirical research process (problem definition, literature review, hypothesis development, data collection, analysis, results, discussion, and conclusions) is a common structure for empirical research papers.

2.3 Clustering

Clustering is a method of grouping data [5]. Data that have similar characteristics to be gathered in the same group or cluster, the data that have different characteristics, will congregate in groups or clusters of different [6] Analysis of the cluster is the method used to divide the data set into groups based similarities predetermined [7] the main objective of the clustering method is a grouping of a number of data/objects into clusters (groups) so that in each cluster will contain the data as closely as possible [8].

2.4 K-Means Methods

K-Means is one method of a non-hierarchical grouping of data (Blocking) which seeks the partitioning of data into the form of two or more groups [9]. This method of partitioning the the data into clusters / groups so that the data having the same characteristics are grouped into the same cluster and the data that have different characteristics are grouped into the other group [10] The data that have the same characteristics are grouped in one cluster/group and the data that have different characteristics grouped by cluster/group to another so the data are in one cluster/group has a small degree of variation [11]. This method is to divide the data into groups with the understanding that any of the data has the same characteristics are grouped into the same group and so Also to any different characteristic properties that the data will be grouped into another group.

1. Determine k as the number of clusters you want in the form.
2. Determining the center point (centroid) early in each cluster as many as k.
3. Calculate the distance of each data inputted to each centroid using Euclidean distance formula (Euclidean Distance) that is found within the closest of any data by using centroid. Here is the equation Euclidian Distance:

$$d(Xj, Cj) = \sqrt{\sum_{i=1}^{n}(Xj - Cj)^2} \qquad (1)$$

   Information :
   d = distance
   j = the number of data
   c = centroid
   x = Data
   c = centroid
4. Classify each of data based on its proximity to the centroid (the smallest distance).
5. Renewing the centroid value. New centroid value obtained from the average cluster is concerned with using the formula;

$$Cj = \frac{\sum_{j=1}^{n} X_j}{p} \qquad (2)$$

   Information :
   Cj = Cluster Centers
   p = Number of all members of the cluster
   n = number of items / the number of objects that are members of the cluster
   xj= the object x to i

ISSN 2548-8384 (online)
ISSN 2548-8449 (print)
Page 49 – 58
DOI 10.30865/ijics.v3i2.1382

6. Doing repetition from steps 2 to 5, until the members of each cluster nothing has changed. If step 6 has been fulfilled, then the cluster center value (j) in the last iteration will be used as a parameter to determine the classification of data.

3. RESULT AND DISCUSSION

3.1 Data analysis

The data is used for the data dataset with 51 students in D4 Informatics Engineering 2018 is the data value with the attributes of general English 2. is grammar, fluency, and comprehension. of the three attributes that will be grouped into three groups: beginner with a value of 0-40, 41-70 intermediate value, whereas 71-100 advance.

Table 1. Initial Data Students

No.	Name	GRAMMAR	UNDERSTANDING	FLUENCY
1	students 1	80	85	84
2	student 2	84	84	78
3	students 3	83	85	90
4	students 4	80	80	85
5	students 5	79	83	80
6	students 6	79	83	85
7	students 7	82	80	83
8	students 8	79	85	78
9	students 9	60	60	65
10	students 10	83	84	85
11	students 11	79	85	78
12	students 12	83	93	90
13	students 13	79	83	85
14	students 14	91	84	80
15	students 15	68	87	75
16	students 16	35	40	45
-	-	-	-	-
-	-	-	-	-
49	students 49	86	79	80
50	students 50	81	79	75
51	students 51	92	82	80

3.2 K-Means Clustering Algorithm

Then after the necessary data are ready, the calculation process such data with manual calculation first thing to do is to use clustering techniques, Calculation using K-Means Clustering:

a. Determining K namely Grammar, Fluency and comprehension
b. Determine the number of clusters. In this study, the data will be grouped into three clusters, namely Beginner, intermediate and advanced.
c. Determining the initial center point (centroid), and get the center point.

Table 2. Centroid 1

centroid 1	Grammar	comprehension	Fluency
cahyani	92	82	80
riri Amaliya	60	60	65
m.rosyid Mubarok	35	40	42

a. Calculate the distance of each data inputted to each - each centroid with Euclidean Distance Using the distance formula to find the closest distance of each data by using centroid. The calculation of the distance is as follows:

ISSN 2548-8384 (online)
ISSN 2548-8449 (print)
Page 49 – 58
DOI 10.30865/ijics.v3i2.1382

the initial central point has been determined, and the distance between the first data point cluster center (centroid) The first is:

$$\sqrt{(80-92)^2} + (85-82)^2 + (84-80)^2 = 13$$

The distance between the first data and the cluster center point (centroid) The second is:

$$\sqrt{(80-60)^2} + (85-60)^2 + (84-65)^2 = 37.229$$

The distance between the first data and the cluster center point (centroid) of the three is:

$$\sqrt{(80-35)^2} + (85-40)^2 + (84-42)^2 = 76.249$$

And so the calculation is done for the data and the next attribute as a step above.

a. Having obtained the results for each of the data, and then classifying each data based on its proximity to the centroid (the smallest distance) to the cluster. So from the calculation of the Euclidean distance Distance value has the smallest distance it will be added to the cluster example in the first data is included in the C1 group because it has the smallest distance is 13. And so on grouping for the next data. So that the amount of data for each cluster in the initial calculation is C1 (Advance) is 40 data, C2 (elementary) amounted to 8 data, C3 (beginner) amounted to 3 data.
b. Do the previous process again to get or make sure the cluster value doesn't change. This repetition is called iteration. The initial step of iteration is to update the centroid value. So looking for a new centroid value and not using the initial centroid. To get the new centroid results seen from the first iteration results, here are examples of calculations:

$$\text{C1 Grammar} = \frac{80+84+83+80+ \ldots\ldots\ldots\ldots +92}{40} = 80.325$$
$$\text{C2 Grammar} = \frac{55+56+65+60+ \ldots\ldots\ldots\ldots +30}{8} = 64.25$$
$$\text{C1 Grammar} = \frac{35+35+40}{3} = 36.666$$

And so on is calculated for the next attribute. So we get a new centroid center point like the table below.

Table 3. Centroid 2

centroid 2		
Grammar	comprehension	Fluency
80 325	82 925	80 775
64.25	71 375	76 125
36.66666667	43.33333333	44

The calculation of the distance is as follows:
The distance between the first data and the cluster center point (centroid) The first is:

$$\sqrt{(80-80.325)^2} + (85-82.925)^2 + (84-80.775)^2 = 3.848$$

The distance between the first data and the cluster center point (centroid) The second is:

$$\sqrt{(80-64.25)^2} + (85-71.375)^2 + (84-76.125)^2 = 22.264$$

The distance between the first data and the cluster center point (centroid) of the three is:

$$\sqrt{(80-36.666)^2} + (85-43.333)^2 + (84-44)^2 = 72.207$$

The calculations are done for the data and the next attribute as a step above. And if the value is different from the first iteration centroid beginning of the meal will be done iteration or repetition again and so on until the centroid value equal to the value of the previous centroid. for grouping students, ability levels have occurred 5 times iterations and stop at centroid to 5 because centroid value to 4 and 5 have the same result. Resulting centroid in the table below:

Table 4. Centroid 5

centroid 5		
Grammar	comprehension	Fluency
81.37837838	83	80.8421053
60.6	76.7	78.9
41.25	45	46.75

ISSN 2548-8384 (online)
ISSN 2548-8449 (print)
Page 49 – 58
DOI 10.30865/ijics.v3i2.1382

For grouping, the results for the final centroid or centroid to 5 and the fifth iteration can be seen in the table below.

Table 5. Distance to the nearest and Results Grouping of iterations to 5

Name	Grammar	comprehension	Fluency	C1	C2	C3
students 1	3983	21 708	67 001	1		
student 2	3,993	24 528	65 765	1		
students 3	9513	26 341	72 205	1		
students 4	5,309	20 602	64 727	1		
students 5	2522	19 479	63 044	1		
students 6	4,789	20 382	65 818	1		
students 7	3747	22 037	64 804	1		
students 8	4210	20 205	63 258	1		
students 9	9,925	21,740	67 235	1		
students 10	4,573	24 336	68 754	1		
students 11	4210	20 205	63 258	1		
students 12	13 656	29 843	76 926	1		
students 13	4,789	20 382	65 818	1		
students 14	9710	31 283	71 425	1		
students 15	15 136	13 268	57 250		1	
students 16	9742	31 417	72 378	1		
students 17	72 695	56 137	8192			1
students 18	21 610	33 278	61 559	1		
students 19	49 523	36 269	16 796			1
students 20	74 220	57 998	9307			1
students 21	3,248	21 495	66 450	1		
students 22	3,445	21 025	65 826	1		
students 23	14 402	9047	56 680		1	
students 24	25 749	7,619	53 198		1	
students 25	16 880	5085	53 156		1	
students 26	25 693	7775	42 126		1	
students 27	17 410	12 249	52 589		1	
students 28	35 170	21 736	30 159		1	
students 29	14 802	7,201	54 112		1	
students 30	2,508	20 111	62 534	1		
students 31	18,657	5,124	54 227		1	
students 32	63 920	47 816	5442			1
students 33	9087	15,500	61 066	1		
students 34	6807	16 657	63 041	1		
students 35	7142	15 095	60 055	1		
students 36	2,257	22 979	66 076	1		
students 37	7550	23 834	61 494	1		
students 38	13 741	14 403	63 179	1		
students 39	5,600	18 478	60 163	1		
students 40	4,856	23 538	63 946	1		
students 41	5604	25 703	68 892	1		
students 42	3,204	24 233	65 662	1		
students 43	51 877	31 439	52 498		1	
students 44	4894	25 196	66 971	1		
students 45	4,248	18 693	60 746	1		
students 46	2827	20 036	64 719	1		
students 47	6,062	23 593	63 127	1		
students 48	3,919	18 725	61,283	1		
students 49	6,170	25 527	65 300	1		
students 50	7090	20 896	59 448	1		

ISSN 2548-8384 (online)
ISSN 2548-8449 (print)
Page 49 – 58
DOI 10.30865/ijics.v3i2.1382

students 51	10 702	31 863	71 064	1		

So for the results of grouping the level of student ability in English can be seen in the results of iteration to cluster 1 (Advance) with a total of 37 data and for cluster 2 (intermediate) with a total of 10 data and cluster 3 (beginner) with a total of 4 from 51 data. It can be seen in the graphic image below.

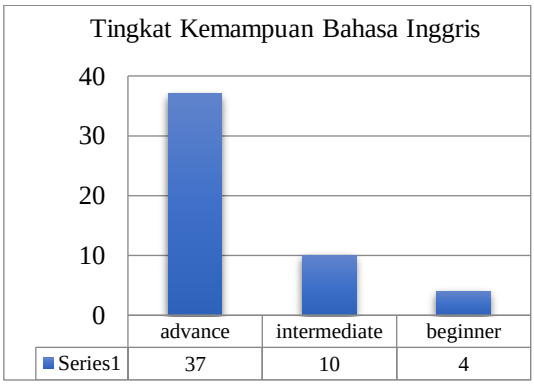

Figure 2. Graph Results Grouping iteration 5

4. IMPLEMENTATION

Following the application of the method of the k-means algorithm on a web-based system.

1. **Analysis System To Be Built.**
2.

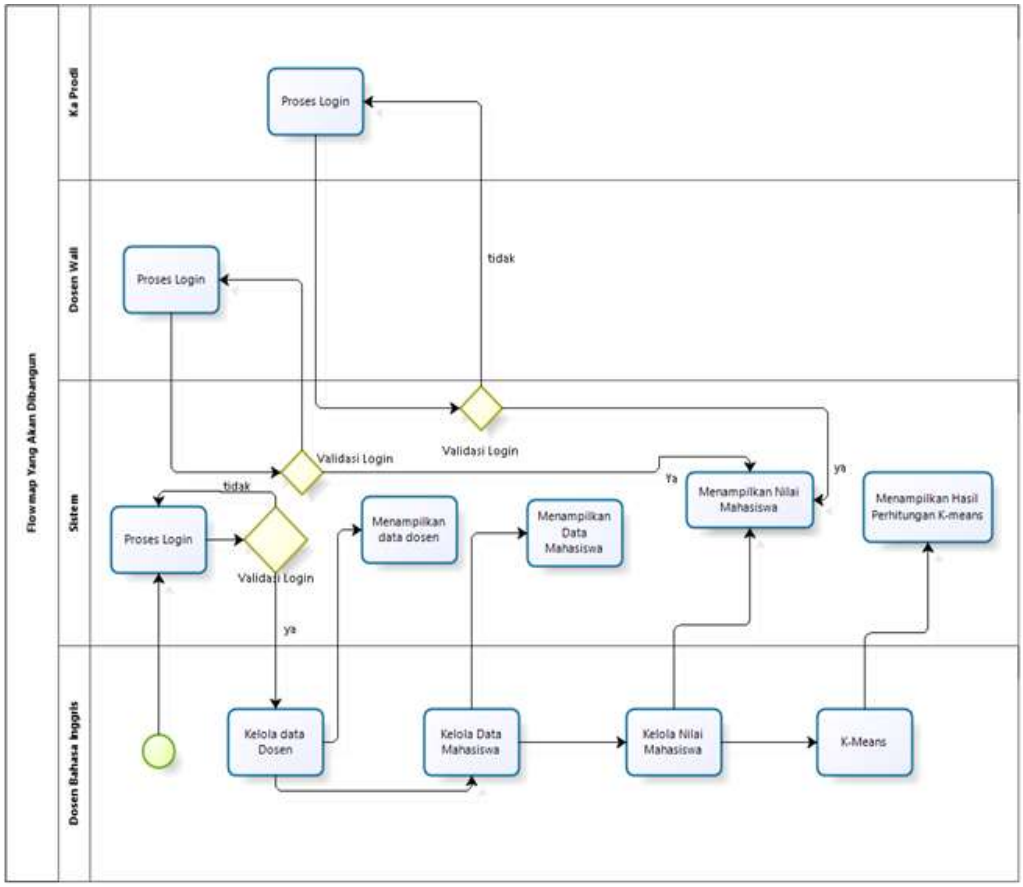

Figure 3. Flowmap Will Be Built

The IJICS (International Journal of Informatics and Computer Science)
Vol 3 No 2, September 2019
ISSN 2548-8384 (online)
ISSN 2548-8449 (print)
Page 49 – 58
DOI 10.30865/ijics.v3i2.1382

3. Database Design

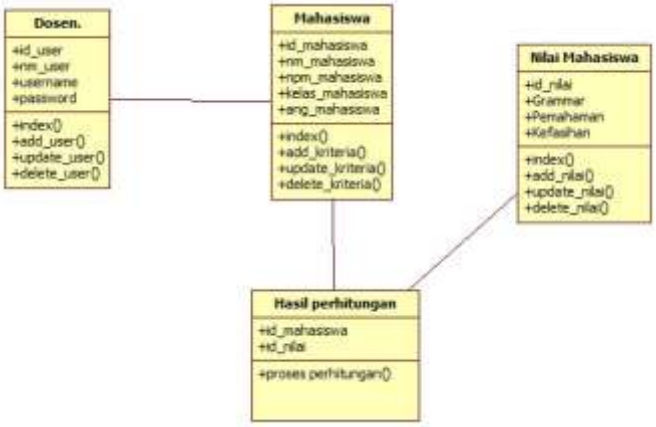

Figure 4. Class Diagram

4. Diagram Use Case

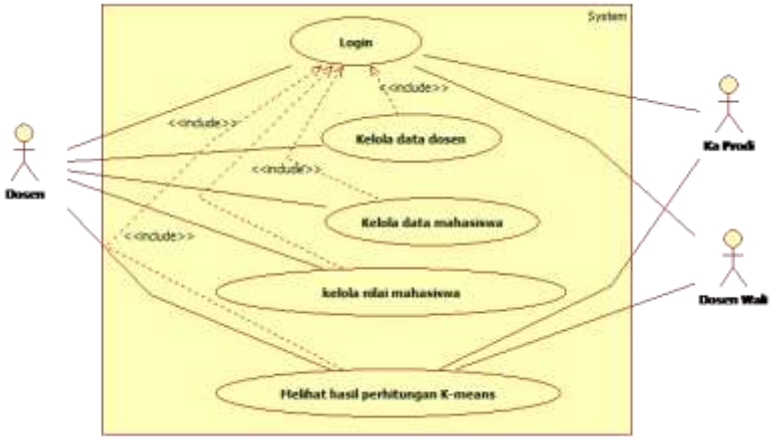

Figure 5. Diagram Use Case

5. User Interface

Page K-means, on this page, can lecturer View graphs and results of the last iteration grouping.

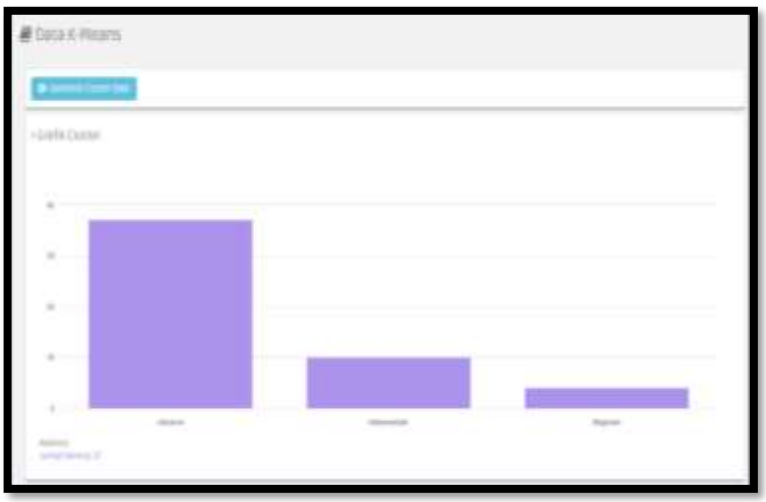

Figure 6. Graphic and Results Grouping Iteration 5

ISSN 2548-8384 (online)
ISSN 2548-8449 (print)
Page 49 – 58
DOI 10.30865/ijics.v3i2.1382

6. Page detail iterations, on this page teachers, can see the results of groupings and calculations of each iteration.

Figure 7. Detail Iteration

Figure 8. Calculation Results and Grouping Iteration

5. CONCLUSION

1. Conclusion The problem results of this study have shown that the problem in knowing the level of students' ability to speak English is based on the criteria used in the calculation and can be applied in order to facilitate a lecturer in knowing the average the level of students' knowledge so that it can be done in the handling of the learning process
2. Conclusion Methods By applying k-means clustering method then of student data 51 are 37 students were included in the Advance group, 10 students included in a group of 4 students Intermediate and Beginner seen including groups of criteria specified. Its requirements, namely grammar, fluency, and comprehension.
3. Conclusions Testing System-level students' ability to speak English can classify students by level of ability in the English language using the k-means clustering.

REFERENCES

[1] Suprawoto, T. (2016). Classification of student data using the k-means to support the marketing strategy selection, 1 (1), 12-18.
[2] Gan, G., & Ng, MK (2017). PT US CR. Pattern Recognition Letters. https://doi.org/10.1016/j.patrec.2017.03.008
[3] Xu, R., Member, S., & Ii, DW (2005). Survey of Clustering Algorithms, (May).

ISSN 2548-8384 (online)
ISSN 2548-8449 (print)
Page 49 – 58
DOI 10.30865/ijics.v3i2.1382

[4] Kao, H., Yu, M. Masud, M., Wu, W., Chen, L., & Wu, YJ (2016). Computers in Human Behavior Design and evaluation of hospital-based business intelligence system (HBIS): A foundation for design science research methodology. Computers in Human Behavior, 62, 495-505. https://doi.org/10.1016/j.chb.2016.04.021
[5] DATA MINING: RapidMiner APPLICATION BY K-MEANS CLUSTER AREA IN contracted Dengue Hemorrhagic Fever (DHF) BY PROVINCE. (2018), 3 (2), 173-178.
[6] DATA clustering UMROH congregation of Saints TOUR & TRAVEL ON USING K-MEANS CLUSTERING. (2019), V (2), 97-104.
[7] Gunawan, S. (2019). IMPLEMENTATION OF K-MEANS, SUFFIX TREE AND DEWEY Decimal Classification LIBRARY BOOKS FOR SHELVING Implementation of the K-Means, Suffix Tree and Dewey Decimal Classification for Book Library Shelving, (1), 121-129.
[8] Windarto, AP, Study, P., Information, S., & Mining, D. (2017). Application of Data Mining In Fruit Exports by Country of Destination Using K-Means Clustering, 16 (4), 348-357.
[9] Sadewo, MG, Eriza, A., Windarto, AP, & Hartama, D. (2019). In the K-Means algorithm Grouping / District Village According Presence User friendly Electrical and Lighting Source Main Street by Province, 754-761.
[10] Widiyaningtyas, T., Indra, M., Prabowo, W., & Pratt, MAM (2017). Implementation of the K-Means Clustering Method to Distribution of High School Teachers (September), 19-21.
[11] Bastian, A., Sujadi, H., Febrianto, G., Studies, P., Informatics, T., Majalengka, U., ... No, M. (nd). No Title, (1), 26-32
[12] Agustin, FEM, Fitria, A., & S, AH (nd). (CASE STUDY: Junior STATE 101 JAKARTA) Information Engineering Program, Faculty of Science and Technology State Islamic University Syarif Hidayatullah, 73-78.