K-Means Clustering Algorithm in Web-Based Applications for Grouping Data on Scholarship Selection Results

Anak Agung Ngurah Krisnanda Putra[a], Mohammad Nasucha[a,b,*], Hendi Hermawan[a,b]
[a)]Department of Informatics, [b)]Center for Urban Studies
Universitas Pembangunan Jaya
Tangerang Selatan, Indonesia
*Corresponding: mohammad.nasucha@upj.ac.id

***Abstract*— In our case the selection of scholarship recipients was done previously by a university, by assessing and scoring 7 variables: whether scholarship application form is submitted, whether student study plan is submitted, whether student transcript is submitted, whether student's curriculum vitae is submitted, what is the score of student's scientific article, what is the score of student's presentation on a given topic, and what is the score of student's GPA. This such selection process was done without computation, with the consequence of time consumption and potential human errors. This research aims to overcome that problem, by providing a computed selection process using the same 7 variables and applying a necessary algorithm. In our research the K-means Clustering algorithm is applied although it is understood that other algorithms can be used too. The tests are carried out using black box and white box methods. The result shows that K-means Clustering algorithm is successfully applied to the scholarship selection system, and the K-means Clustering algorithm is successful in grouping students who receive and who do not receive the scholarship.**

Keywords—K-means Clustering, scholarship, euclidean distance, black box testing, white box testing

I. Introduction

Scholarships are assistance in the form of money or other things that are given to students with the aim of helping the continuity of their education. Scholarships can be awarded by government agencies, companies and foundations [1]. Universitas Pembangunan Jaya (UPJ) – a private university in Tangerang Selatan - provides various scholarships, such as Academic, Champion and Special Scholarships to active students, so that students can continue their education. Previously a university staff carried out selection manually, by assessing/scoring seven variables: whether scholarship application form is submitted, whether student study plan is submitted, whether student transcript is submitted, whether student's curriculum vitae is submitted, what is the score of student's scientific article, what is the score of student's presentation on a given topic, and what is the score of student's GPA. This such selection process was done without computation, with the consequence of time consumption and potential human errors. The selection decision then became doubtful.

Previously a number of research activities exercised the K-means Clustering algorithm. Exercise using 4 assessment criteria was done at Nahdlatul Ulama University Blitar [2]. Then, exercise with 5 assessment criteria was done at Makassar State University [3].

This research aims to overcome that problem, by providing a computed selection process using the same seven variables and applying a necessary algorithm. In our case the K-means Clustering algorithm is applied although it is understood that other algorithms can be used too. In this research, K-means Clustering algorithm is applied to classify scholarship applicants based on the seven variables. The classification will distinguish the applications into two groups: granted and ungranted. The remainder of this article is arranged as follows: Section 2 addresses the method, Section 3 reports the results, Section 4 discusses the results, and Section 5 summarizes the content of all sections.

II. METHOD

A. K-means Clustering Algorithm

K-means Clustering is a non-hierarchical clustering method that works by partitioning existing data into one or more groups / clusters [4]. This algorithm performs data partitioning by grouping data that have the same characteristics into one group. The stages are as follows.

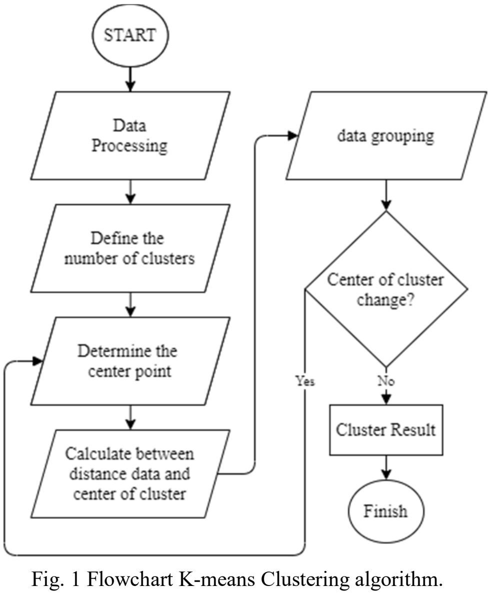

START
Data Processing
Define the number of clusters
Determine the center point
Calculate between distance data and center of cluster
data grouping
Center of cluster change?
Yes
No
Cluster Result
Finish

Fig. 1 Flowchart K-means Clustering algorithm.

978-1-6654-4146-9/21/$31.00 ©2021 IEEE

1) Input data. This data is used to determine the average value of data points in a cluster and the distance from each data point to the midpoint of a cluster in the form of value (centroid).

2) Input data into the cluster randomly. This stage is the first time the data points are entered into the cluster randomly without any specific criteria.

3) Calculate the data point centroid in each cluster. The centroid value on K-means Clustering is used as the center of the cluster. By determining the random cluster members in the previous stage, the initial iteration is formed as a random cluster center.

4) Allocation to the nearest centroid, at this stage the results of the centroid from each cluster are known, then the data points are allocated to the nearest centroid based on the distance value of the data point similarity to the centroid. The similarity distance from the data point to the centroid in each cluster is obtained from the calculation of Euclidean Distance. Then the value of the distance of each data point to the centroid cluster is compared and the data point becomes a member of the cluster based on the distance of the data point to the nearest centroid.

5) Convergent, allocating data points to the centroid with the closest distance value, by testing whether the clusters formed have formed convergent clusters or not. Clusters are stated to be convergent if the members of each cluster formed do not change their members. But if there is still a change in cluster members, the stage of counting the centroid of each cluster that is formed will be carried out again, followed by the calculation of the similarity value to the newly formed centroid. This process continues until the cluster results converge [5].

The K-means Clustering algorithm has two main parts: distance and centroid calculation. The distance calculation is to determine the distance among data that is given by the following equation [5].

$$d = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \qquad (1)$$

The formula above is the euclidean distance formula where d is the Euclidean distance, x_i is the data point and, y_i is the centroid. The centroid calculation is to determine the new centroid that is given by the following equation.

$$New\ Centroid = \frac{yield\ value}{a\ lot\ of\ value} \qquad (2)$$

The application of the K-means Clustering algorithm in this case assumes that there are five students who apply for scholarships with the following 7 assessment criteria.

Table I. Scholarship data.

No	Student	Scholarship Form	BRS	CV	Transcript	Paper	Presentation	GPA	Result GPA
1	Budi	100	100	100	100	80	85	3.5	87.5
2	Andi	100	100	100	100	85	75	3.6	90
3	Reza	100	100	100	100	75	90	3.25	81.25
4	Arya	100	100	100	100	90	60	3.4	85
5	Anggi	100	100	100	100	95	70	3.7	92.5

In this case, the value of the Scholarship Form, BRS, CV, and Transcript is 100 for students who have submitted these requirements. If the student does not collect it, they will be given a score of 0. For GPA, multiplication is carried out, namely by multiplying the GPA value by 25 so that the results obtained can be 100. Meanwhile, Paper and Presentations are given an assessment of 0-100. After that, two clusters were made, namely the passing (C1) and the failing (C2) clusters. Furthermore, the data is run randomly on the values that students have and in this example the scores taken are Anggi and Arya as in Table II below.

Table II. First centroid.

Cluster	Student	Scholarship Form	BRS	CV	Transcript	Paper	Presentation	Result GPA
C1	Anggi	100	100	100	100	95	70	92.5
C2	Arya	100	100	100	100	90	60	85

After obtaining the initial cluster from random data collection, the next step is to calculate the distance between centroids using the euclidean distance formula. The results of the distance calculation will be entered in Table III below to group students according to their clusters.

Table III. First calculation.

No	Student	Distance to the Cluster		Result
		C1	*C2*	
1	Budi	21.79	27.04	1
2	Andi	11.46	16.58	1
3	Reza	30.46	33.75	1
4	Arya	13.46	0	2
5	Anggi	0	13.46	1

After getting the results of the first calculation in Table III, it is continued by re-determining the center point of the new cluster to test whether the first calculation is the correct grouping or not using the formula. Initially, the determination of the center point was carried out randomly, but currently calculating using the New Centroid formula as previously explained, so that the results of the new center point are obtained as follows.

Table IV. Second centroid.

Cluster	Student	Scholarship Form	BRS	CV	Transcript	Paper	Presentation	Result GPA
C1	Anggi	100	100	100	100	83.75	80	87.8125
C2	Arya	100	100	100	100	90	60	85

After that, recalculate the data of students who registered for the scholarship selection using a new cluster that had been previously calculated with the Euclidean distance formula and the following results were obtained.

Table V. Second calculation.

No	Student	Distance to the Cluster		Result
		C1	*C2*	
1	Budi	6.26	27.04	1
2	Andi	5.6	16.58	1
3	Reza	14.82	33.75	1
4	Arya	21.14	0	2
5	Anggi	15.76	13.46	2

After comparison between the first calculation in Table III and the second calculation in Table V, it was found that changes in the data for Anggi students were originally from one to two clusters. Because of this change, it must be recalculated until the data does not change again. The recalculation starts by defining a new cluster using the same formula as before, so as to get a new cluster, as below.

Table VI. Third centroid.

Cluster	Student	Scholarship Form	BRS	CV	Transcript	Paper	Presentation	Result GPA
C1	Anggi	100	100	100	100	80	83.33333333	86.25
C2	Arya	100	100	100	100	92.5	65	88.75

After getting the third centroid, calculate data using euclidean distance formula, so that the results are as in Table VII below.

Table VII. Third calculation.

No	Student	Distance to the Cluster		Result
		C1	*C2*	
1	Budi	2.08	23.62	1
2	Andi	10.42	12.56	1
3	Reza	9.72	31.42	1
4	Arya	25.42	6.73	2
5	Anggi	21.02	6.73	2

After comparing again between the third calculation in Table V and the second calculation in Table VII, it turns out that there is no change in clusters or results in the two tables, so it can be concluded that the data is in the appropriate cluster and obtained 3 students received scholarships, namely Budi, Andi and Reza, 2 students did not receive scholarships, namely Arya and Anggi.

B. *Black Box Testing*

Black Box testing is sometimes called behavioral testing or partition testing that focuses on the functional specifications of the software[6]. Black box testing is one of the tests that is carried out only by observing the results of execution through test data and checking the functionality of the software or it can be said as testing that focuses on its functionality without knowing what actually happens in the detailed process [7]. The tester can define a lot of input conditions and perform test on program technical specifications [8]. According to (Mardzotillah & Ridwan, 2020) the function of this black box testing includes several things, namely [9]:

1) Improper or inappropriate functions,
2) Interface error,
3) Errors in data structures and database access,
4) Performance errors, and
5) Initialization and termination errors.

C. *White Box Testing*

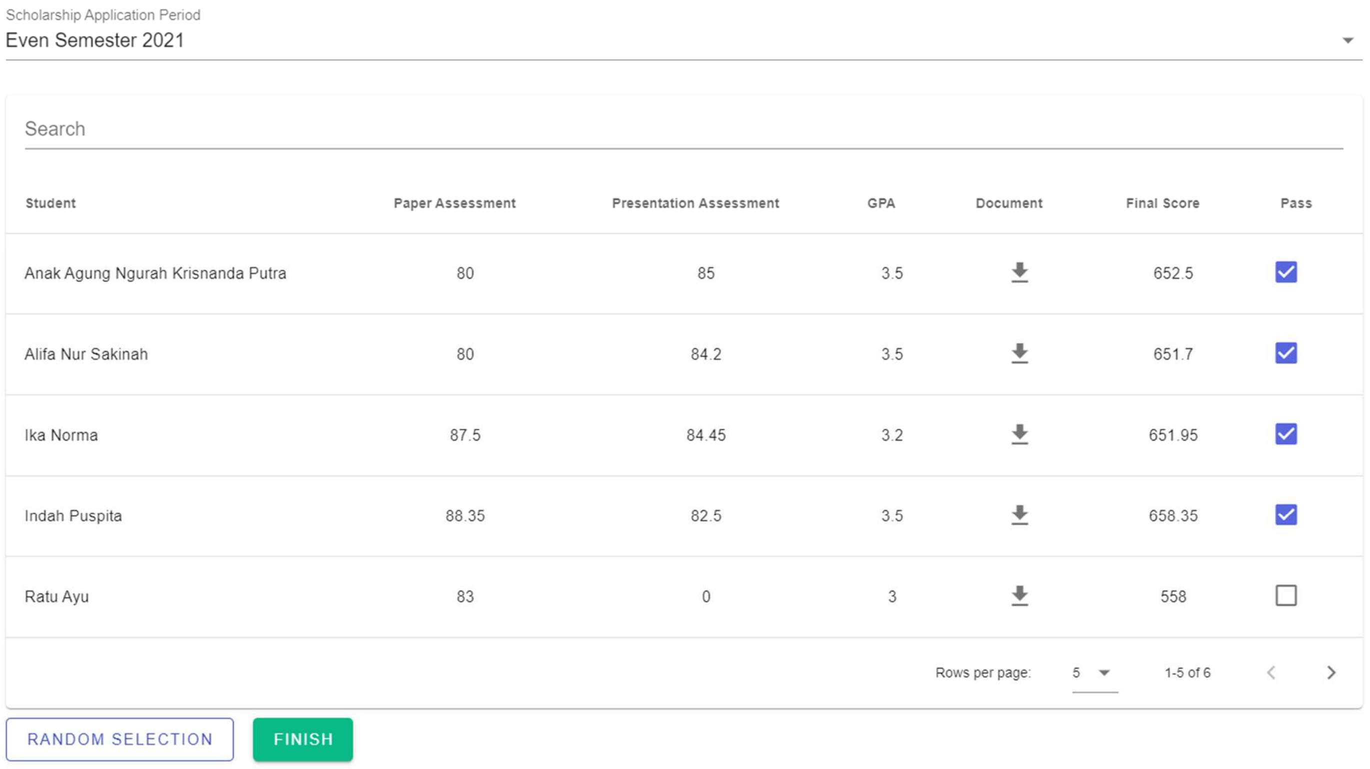
White box testing is one of testing the code of a program and analyzing it to find errors in that section [10]. This test is done without looking at the display, so the test that is carried out only focuses on the program code. White box testing is also used to find out whether the input and output of the program that has been made is as expected or not. White box testing is carried out using flow graphs and cyclomatic complexity with the following formula [10].

$$V(G) = E - N + 2 \qquad (3)$$

Here, V (G) is cyclomatic complexity, E is total number of edges, and N is total number of nodes.

III. RESULT

Results of both simulation and experiment are presented here. In the picture below, it can be seen that there are five students who registered themselves in the scholarship selection at UPJ. When the random selection button is clicked, the K-means Clustering algorithm will classify the student data according to predetermined clusters. So, it was found that four scholarship recipients and one student who did not receive scholarships. On this page there is also information such as paper assessment, presentation assessment, GPA, supporting documents and final score. When the data has been successfully grouped, then click finish and the data will be saved according to the results of the grouping by the K-means Clustering algorithm.



Scholarship Application Period: Even Semester 2021

Search

Student	Paper Assessment	Presentation Assessment	GPA	Document	Final Score	Pass
Anak Agung Ngurah Krisnanda Putra	80	85	3.5	⤓	652.5	☑
Alifa Nur Sakinah	80	84.2	3.5	⤓	651.7	☑
Ika Norma	87.5	84.45	3.2	⤓	651.95	☑
Indah Puspita	88.35	82.5	3.5	⤓	658.35	☑
Ratu Ayu	83	0	3	⤓	558	☐

Rows per page: 5 1-5 of 6

Fig. 2. Implementation of K-means Clustering algorithm.

A. Black Box Testing

Application of the K-means Clustering is depicted by Figure 2. Testing of the K-means Clustering algorithm using the black box testing method is informed by Table VIII.

Table VIII. Black box testing K-means Clustering algorithm.

No	Scenario	Expected Result	Conclusion	Result
C1	Choose the scholarship application period	Displays a list of students for a certain period	Suitable	Success
C2	Click the random selection button	The K-means Clustering algorithm runs and provides data on scholarship recipients and non-scholarship recipients	Suitable	Success

B. White Box Testing

Testing of the K-means Clustering algorithm using the white box method is informed by Table IX.

Table IX. White box testing K-means Clustering algorithm.

Symbol	Program Code
(1)	public function kmeans(Request $request){
(2)	$periodID = $request->period_id?? null;
(3)	$getSubmissions = $this-> scholarshipSubmissionsModel-> with(['paper_assessments', 'presentation_assessments', 'student.profile', 'period'])->where('period_id', $periodID)->where('next_stage',1)-> where('final_stage', null)->orWhere('final_stage', 0)->get()-> map(function ($value) {
(4)	$countPresentation =count ($value>presentation_assessments);
(5)	$submitScore = !empty($value-> submit_form)&&$value->submit_form ? 100 : 0;
(6)	$brsScore = !empty($value-> brs) && $value->brs ? 100 : 0;
(7)	$raportScore = !empty($value->raport) && $value->raport ? 100 : 0;
(8)	$cvScore = !empty($value-> cv)&&$value->cv? 100 : 0;
(9)	$paperAssessments =$value->paper_assessments? $value->paper_assessments->papers_score : 0;

(10)	$presentationAssessments = $countPresentation ?
	$value->presentation_assessments->
	reduce (function ($carry, $item) {
	return $carry + $item->final_score;
	}) / $countPresentation : 0;

(11)	$ipk = $value->initial_ipk;
(12)	$ipk_final = $ipk * 25;
(13)	return ['id' => $value->id, "student_id" => $value->student_id, "student" => $value->student, "period" => $value->period, "period_id" => $value->period_id, "submit_form" => $value-> submit_form, "submit_score" => $submitScore, "brs" => $value->brs, "brs_score" => $brsScore, "raport" => $value->raport, "raport_score" => $raportScore, "cv" => $value->cv, "cv_score" => $cvScore, "papers" => $value->papers, "other_requirements" => $value->other_requirements!= null ? asset('upload/' . $value->other_requirements): "", "presentation" => $value-> presentation,"next_stage" => $value->next_stage, "final_stage" => $value->final_Stage, "paper_assessments" =>$paperAssessments, "presentation_assessments" => $presentationAssessments, "ipk" => $ipk, "final_score" => $submitScore + $brsScore + $raportScore + $cvScore + $paperAssessments + $presentationAssessments +$ipk_final];});
(14)	return $getSubmissions;}
(15)	kmeans.clusterize(mapData, { k: 2 }, (err, res) => {
(16)	if (err);
(17)	else {
(18)	const clusterOne = res[0].centroid.reduce ((previous, current) => {return previous + current;});

(19)	const clusterTwo =
	res[1].centroid.reduce((previous, current) => {
	return previous + current;});

(20)	selectedCluster = clusterOne>clusterTwo ? 0:1;for (let
	i = 0; i < res[selectedCluster].clusterInd.length; i++) {
	self.beasiswaMahasiswa[res[selectedCluster].
	clusterInd[i]].final_stage = true;}}});

(21)	end

The flow graph and cyclomatic complexity of white box testing as shown in Figure 3.

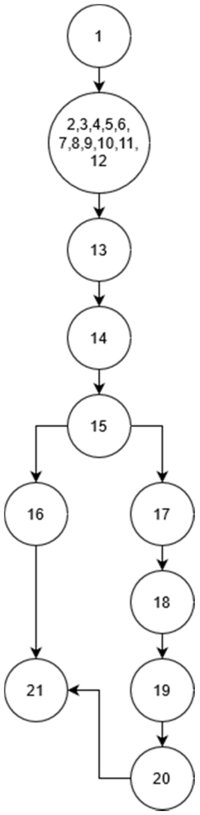


Fig. 3. Flow graph white box testing K-Means Clustering algorithm.

The flow graph above illustrates the logical control flow of the K-means Clustering algorithm to classify scholarship applicants. The Calculation of cyclomatic complexity based on the flow graph is as follows.

$$V(G) = 11\ edges - 11\ nodes + 2 = 2 \qquad (4)$$

Based on the calculation of the cyclomatic complexity, there are 2 independent paths where the results are in accordance with the cluster. The description is as follows.

path 1 = 1-2-3-4-5-6-7-8-9-10-11-12-13-14-15-16-21

path 2 = 1-2-3-4-5-6-7-8-9-10-11-12-13-14-15-17-18-19-20-21

IV. CONCLUSION

The conclusion that can be drawn from research on the use of the K-means Clustering algorithm in the scholarship selection system is that the K-means Clustering algorithm has been successfully applied to the scholarship selection system. K-means Clustering algorithm also works in grouping scholarship recipients with 7 assessment criterias and the data will be stored according to the results of the grouping by the K-means Clustering algorithm.

REFERENCES

[1] E. W. Saputra, “Optimasi Fungsi Keanggotaan Fuzzy Mamdani Menggunakan Algoritma Genetika Untuk Penentuan Penerima Beasiswa,” vol. 02, no. 02, pp. 160–175, 2019.

[2] A. E. Rahayu, K. Hikmah, N. Y. Ningsih, and A. C. Fauzan, “Penerapan K-Means Clustering Untuk Penentuan Klasterisasi Beasiswa Bidikmisi Mahasiswa,” *Ilk. J. Comput. Sci. Appl. Informatics*, vol. 1, no. 2, pp. 82–86, 2019.

[3] R. Rizki, “Sistem Pendukung Keputusan Seleksi Penerima Beasiswa Bidik Misi Bagi Mahasiswa Baru Universitas Negeri Makassar Menggunakan Algoritma K-Means Clustering,” *Indones. J. Fundam. Sci.*, vol. 5, no. 1, pp. 33–46, 2019.

[4] A. K. Turkhamun, B. Panjaitan, and R. Guntara, “Implementasi Data Mining Clustering Data Mahasiswa Teknik Informatika Menggunakan Algoritma K-Means,” *Semin. Nas. Cendekiawan*, vol. 1, no. ISSN (P) : 2460-8696 ISSN (E) : 2540-7589, pp. 805–810, 2018.

[5] A. E. Wicaksono, “Implementasi Data Mining Dalam Pengelompokan Peserta Didik di Sekolah untuk Memprediksi Calon Penerima Beasiswa Dengan Menggunakan Algoritma K-Means (Studi Kasus SMA N 6 Bekasi),” *Jur. Tek. Inform. Univ. Gunadarma*, vol. 21, no. 3, pp. 206–216, 2016, [Online]. Available: https://ejournal.gunadarma.ac.id/index.php/tekno/article/view/1599/1358.

[6] Zamtinah, E. Supriyadi, and Soeharto, “Functional test of the online Recognition of Work Experience and Learning Outcome System using black box testing,” *J. Phys. Conf. Ser.*, vol. 1446, no. 1, 2020, doi: 10.1088/1742-6596/1446/1/012060.

[7] P. Astuti, “Penggunaan Metode Black Box Testing (Boundary Value Analysis) Pada Sistem Akademik (Sma/Smk),” *Fakt. Exacta*, vol. 11, no. 2, p. 186, 2018, doi: 10.30998/faktorexacta.v11i2.2510.

[8] S. Sutiah and S. Supriyono, “Software Testing on The Learning of Islamic Education Media Based on Information Communication Technology Using Blackbox Testing,” *... (International J. Inf. Syst. ...*, vol. 3, no. 36, pp. 254–260, 2020, [Online]. Available: http://ijistech.org/ijistech/index.php/ijistech/article/view/57.

[9] Q. Mardzotillah and M. Ridwan, “Sistem Tracer Study Dan Persebaran Alumni Berbasis Web Di Universitas Islam Syekh-Yusuf Tangerang,” *Jutis (Jurnal Tek. Inform.*, vol. 8, no. 1, pp. 90–106, 2020, [Online]. Available: http://ejournal.unis.ac.id/index.php/jutis/article/view/705.

[10] E. sita Eriana, “Pengujian Sistem Informasi Aplikasi Perpustakaan Berbasis Web Dengan White Box Testing,” *J. Teknol. Inf. ESIT*, vol. XV, no. 10, pp. 28–33, 2020.