K-MEANS APPLICATION IN CLUSTERING JUNIOR SCHOOLS BASED ON OF NATIONAL EXAM AVERAGE RESULTS IN SURABAYA CITY

EA Widodo [1], C Alvina [2], V C Pranata [3], V Yaphira [4], E S Nugraha [5]
School of Business, Study Program of Actuarial Science, President University, Cikarang, Bekasi 17550, Indonesia
[1]*edward.widodo@sudent.president.ac.id*
[2]*celine.alvina@student.president.ac.id*
[3]*vania.pranata@student.president.ac.id*
[4]*veldelen.yaphira@student.president.ac.id*
[5]*edwin.nugraha@president.ac.id*

ABSTRACT:
In each year each school will submit an individual report about its school to the Education Authorities, which also contains the lowest and highest national exam results, also the average score of all students who took the exam. Because there are so many junior high schools and the equivalent in the city of Surabaya, of course we will find it difficult to find which school gets good national exam scores from the four subjects tested. From the test, the resulting K-Means is one of the basic algorithms for forming clusters and can be applied to the Classification of Junior High Schools (SMP) and the equivalent based on the average national exam results. Clustering using the K-Means algorithm has a fairly good ability in the classification based on the average school exam results. Where in detection, and by input parameters such as threshold. Designing a new system of classifying schools can assist agencies in classifying schools based on average national exam scores. By using K mean clustering, we can classify into several groups based on the existing parameters.

Keywords: *Junior High School, National Examination, K-Means Clustering, Parameters.*

1. Introduction

The role of computers is increasingly in people's lives. Almost all areas of life have used computers as tools. It is hoped that in its development, the benefits of computers can be immediately felt by the community. One of the groups of society that uses it a lot is students or people who are in the sphere of education.
Junior high school education (SMP) and the equivalent are the level of basic education in formal education which has a major role in the continuation of the subsequent educational process. For junior high school students who want to continue their education to a higher level, students must take the National Examination (UN) which is held simultaneously throughout Indonesia. From the results of the national exam, it will determine whether or not a student from junior high school will graduate. Each year each school will submit an individual report about its school to the Education Authorities, which also contains the lowest and highest national exam results and the average score of all students who took the exam. Because of the large number of junior high schools and their equivalents in the city of Surabaya, of course we will find it difficult to find which schools get good national exam scores from the four subjects tested. Therefore, the author will create a program that can classify which schools get good and less good test scores based on the average score of the national exam results.
K-means is an algorithm for classifying or grouping objects based on certain parameters into a number of groups, so that it can run faster than hierarchical clustering (if k is small) with a large number of variables and produce tighter clusters, according to (Abidin, 2009). With this application, we hope to facilitate the process of grouping schools based on the average national exam results.
Therefore, we want to classify the scores of the junior high school national exams in Surabaya with the K-Means clustering method which can classify easily and quickly.

For our problem scope, we have 3 points that the data parameters in this Final Project are Mathematics Subject, Bahasa Indonesia, Bahasa English, and IPA (Natural Sciences), the clustering process is based on early 2019 data, the method used is K-Means.
For our problem formulation, we will use Application of the K-Means Algorithm to school data clustering and how to classify Junior High Schools and their equivalents based on the average National Examination results.
K-means clustering has also been used by some people to group a very large amount of data. For example, (Umran, M., & Abidin, T. F. 2009) document grouping using K-Means and singular value decomposition. From the data using K-Means Clustering, we get a terms-document matrix A measuring 48,512 x 30,000. (Ade Bastian, Harun Sujadi, dan Gigin Febrianto. 2018), Application of k-means clustering analysis algorithm in disease transmitted to humans, the results of the K-means clustering are grouped into 6 groups of diseases. (Baginda Harahap, 2019), Application of K-Means Algorithm to Determine Building Materials Laris (Case Study at UD. Toko Bangunan YD Indarung. The result of this K-Means clustering is to group the names of building materials based on the number of sales.

1. Method

1.1. The k-means algorithm

of similarity to members of other clusters is very low. The similarity of members to the cluster is measured by the proximity of the object to the mean value of the cluster or can be referred to as the centroid cluster or center of mass (Widyawati, 2010). The following is the distance measurement formula according to (Milde, J. T., 2008):

$$d_{(x,y)} = \|x - y\|^2 = [\sum_{i=1}^{n} \quad (X_i - Y_i)^2]^{1/2} \tag{1}$$

Where d is document point; x is record data, and y is centroid data.
The shortest distance between the centroid and the document determines the position of the document cluster. For example, document A has the shortest distance to centroid 1 compared to the others, then document A enters group 1. Recalculate the position of the new centroid for each centroid (Ci..j) by taking the average of the documents that enter the cluster. early (Gi..j). Iteration is carried out continuously until the position of the group does not change. The iteration formula is defined as follows:

$$C(i) = \frac{x_1 + x_2 + \dots}{\Sigma x} \tag{2}$$

Where x_1 is value of first record data, x_2 is value of second record data, and $\sum x$ is sum of record data.
K-Means is a partitional clustering algorithm that divides a set of data objects into subsets (clusters) that do not overlap, so that each data object is exactly in one cluster. The partitional-clustering strategy that is most often used is based on the square error criterion. In general, the objective of the square error criterion is to obtain a partition (fixed number of clusters) that minimizes the total square error.

1.2. Procedure of K-means algorithm

There are the steps in processing the K-means algorithm: (Larose, 2005):
First, determine the number of k clusters you want. Second, Initialize to determine the center of the cluster. Third, for each row, find the closest cluster center. To calculate the distance between the data and the center of the cluster, the formula is used *Distance Euclidean:*

$$D(X_i, M_k) = [\sum_{i=1}^{p} \quad (X_{ij} - M_{kj})^2 \quad]^{1/2} \tag{3}$$

When Xi is the data, Mk is the centroid of each cluster and p is the number of criteria.
Fourth, determine the group based on the shortest distance. Fifth, For each k cluster, find the centroid (means) of the cluster and update the location of the cluster center into the new centroid values. When Mk is the new mean, Nk is the number of patterns in the k cluster, and Xjk is the number 1 pattern of the k sequence that belongs to the cluster. Sixth, repeat steps 3 to 5 until the iteration value or tolerance value

(the difference between the old and new M allowed to stop the algorithm) is determined or there is still data moving.

1.3. Advantages and Disadvantages of K-Means Method

The advantages of K-Means clustering are easy to implement and run, the time needed to carry out this learning is relatively fast, adaptable, commonly used. Beside that, the disadvantages of K-Means clustering are perfectly matches the initial selection of centroids, it is not clear that how many cluster K is best, only works on numeric attributes.

2. System Design

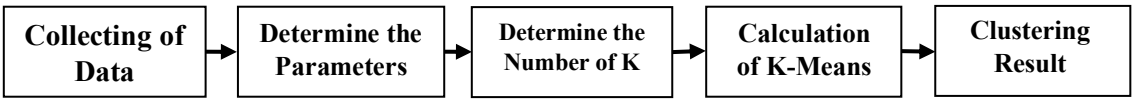

Figure 1. System Design K-Means

In collecting of data, data of average test results are taken from data from each junior high school and its equivalent in the city of Surabaya. In determine the parameters, we determining the parameter is taken from the average subjects tested in the National Examination, there are Mathematics, Natural Sciences, Indonesian, and English. In determine the number of k, we determining the number of k can be determined by the elbow method and the silhouette method to obtain the optimal number of k to cluster this data. In calculation of K-Means, K-Means calculation starts from determining the centroid point for each cluster and calculating the euclidean distance between the centroid points with all data. Then, find the shortest distance between the data and all existing centroids and enter the cluster with the shortest distance. After that, the clustering is repeated until there is no change in the cluster. In clustering result, the results of clustering are the results of class determination for each data, where each data will enter the class with the shortest euclidean distance from the euclidean distance to other centroid points.

3. Result and Discussion

3.1. Data Testing

The data to be processed is the 2019 National Exam data for junior high school students in Surabaya, with a total of 319 data. Data processing was performed using RStudio. The first thing to do is to prepare and present the data (table 1) that we will use, namely "Group2-data.csv". Complete data is listed in the attachment section.

As we can see, to do simple K-Means clustering in R we need some packages such as, "ggplot2" to plot our cluster data, "factoextra" is to run the elbow method and silhouette method, and "kableExtra" is used to present the results. the end becomes a neat table. The "leaflet" package will be used to create a data distribution map, and "dplyr" is activated to use functions such as tables to calculate the amount of data and also the mutate function for adding data columns with the application of formulas. Display data that has been presented into a data-frame in R, see Table 2:

We also do data scaling. This will be of great help in getting good quality results from the K-Means grouping when we have different unit types. The results of the data that have gone through the data scaling process can be seen in the following figure.

Table 1. Display of Data Frame in R

NO	CODE	SCHOOL	NPSN	STATUS	STUDENTS	INDONESIA LANGUAGE	INGGRIS	MATHEMATICS	SCIENCE	AVERAGE	LAT	iN
1	5010001	SMP NEGERI 1 SURABAYA	20532613	N	351	91.28	94.34	96.42	92.03	93.52	-7.257286439	112.747683
2	5010002	SMP NEGERI 41 SURABAYA	20532571	N	329	75.3	52.18	54.47	56.16	59.53	-7.241645059	112.750885
3	5010004	SMP NEGERI 2 SURABAYA	20532559	N	331	85.63	77.56	85.57	78.16	81.73	-7.242536703	112.7359835
4	5010006	SMP NEGERI 3 SURABAYA	20532547	N	303	87.78	84.1	90.71	83.84	86.61	-7.256262709	112.736133
5	5010008	SMP WACHID HASYIM 4 SURABAYA	20532592	S	108	58.39	41.96	37.62	39.95	44.48	-7.247804744	112.7368117
6	5010009	SMP MUHAMMADIYAH 7	20532532	S	43	65.53	44.33	39.13	44.53	48.38	-7.24325373	112.7261074
7	5010010	SMP PGRI 29	20532468	S	12	49	37.5	35.63	32.29	38.61	-7.252494674	112.739152

peserta	bind	bing	M	l	rrn
2.010491691	2.226969202	2.153826517	2.884582642	3.1518933790	2.693811299
1.823510102	0.654008971	-0.283604976	0.243638158	0.3558845081	0.187649898
1.840508428	1.670822437	1.183710282	2.201525487	2.0707491626	1.824507300
1.602531861	1.882453507	1.561812795	2.525112010	2.5134960371	2.184321000
-0.054804948	-1.010494002	-0.874462421	-0.817146456	-0.9076589488	-0.922021449
-0.607250551	-0.307681984	-0.737443620	-0.722085045	-0.5506553071	-0.634465418

Figure 2. The Result of Data Scaling Process in R

3.2. Process of Clustering Data

First of all, we will determine the number of clusters using the elbow method and the silhouette method. Here, we can see that it is difficult or ambiguous to determine the elbow, whether k = 2 or k = 3. So, we're going to do the silhouette method and compare the results. The silhouette method suggests using K = 2. So by comparing the 2 results, we decided to use 2 clusters.

The

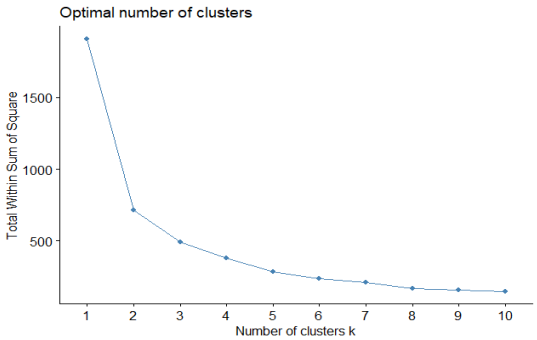

Figure 3. Chart of Elbow Method

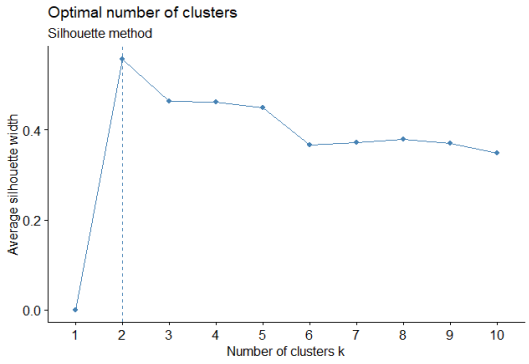

Figure 4. Chart of Silhouette Method

next step is process of clustering. In the K-Means function, "center" is the number of clusters (number of centroids) we want to use in the clustering process, "iter.max" is the maximum iteration in clustering, and "nstart" is the amount of data we will use to select the centroid. The clustering results will be added to our main data to a new column "Cluster". The character of each cluster can be identified by looking at the comparison between centroids. Centroid will type by typing "Data.scaled.clustering" in the console. Display on console:

3.3. The Result of Data Clustering

The next thing to do is plot our data with the cluster results. In the plot we use the number of participants as the x-axis and the mean of the values as the y-axis. The points of the scatter plot are distinguished based on the shape of each group and the color of the dots is differentiated by cluster. "stat_ellipse" helps us plot the area of each cluster. "Alpha" to measure the transparency of the area and "level" to measure the radius of the area, while "geom" in the stat_ellipse is to set the geometric object to be used and displayed. The plot image is as follows.

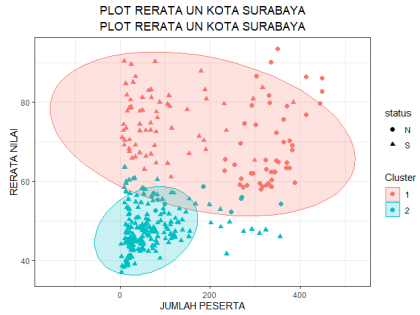

Figure 5. Plot of the average score of the Surabaya City National Examination

Based on the plot results above, it can be seen that cluster 1 has a wider elliptical area, besides that the location is higher than cluster 2. Through this, it can be said that in general the average national examination score for schools included in cluster 1 is higher than the schools included in cluster 2. The wider area of ellipse cluster 1 can also provide the information that schools in cluster 1 have a wider range of test takers than cluster 2.

Next, we will summarize, and display the results of our final grouping. The KableExtra package is here used to create a table for our output. The "kable" function creates a clean and simple table for our output,

while "kable_styling" is for editing our table, adjusting alignment, and formatting the table. The final display of grouping using K-Means in RStudio can be seen in the following figure.

NO	KODE	NAMA SATUAN PENDIDIKAN	NPSN	STATUS	JUMLAH PESERTA	BAHASA INDONESIA	BAHASA INGGRIS	MATEMATIKA	IPA	RATA-RATA NILAI	LATITUDE	LONGITUDE	CLUSTER
1	5010001	SMP NEGERI 1 SURABAYA	20532613	N	351	91.28	94.34	96.42	92.03	93.52	-7.257286	112.7477	1
2	5010002	SMP NEGERI 41 SURABAYA	20532571	N	329	75.30	52.18	54.47	56.16	59.53	-7.241645	112.7509	1
3	5010004	SMP NEGERI 2 SURABAYA	20532559	N	331	85.63	77.56	85.57	78.16	81.73	-7.242537	112.7360	1
4	5010006	SMP NEGERI 3 SURABAYA	20532547	N	303	87.78	84.10	90.71	83.84	86.61	-7.256263	112.7361	1
5	5010008	SMP WACHID HASYIM 4 SURABAYA	20532592	S	108	58.39	41.96	37.62	39.95	44.48	-7.247805	112.7368	2

Figure 6. Table Final Result of Clustering

Then by using the table function, we will calculate the data in each cluster and enter the count results using the for loop into the data frame which will be made into a neat table using kable as in the previous section.
This table below displays the results of calculating the amount of data for each cluster.

CLUSTER	NEGERI	SWASTA	TOTAL
1	49	66	116
2	10	194	206

Figure 7. Table of Results of Calculating the Amount of data per Cluster

Next, we will create a bar plot to compare the average values of national examination in cluster 1 and cluster 2.
Before starting to plot, we need to create a new data frame that contains the average of the average national examination scores for each school in each cluster. To create a bar plot, the ggplot function can be used with the cluster as the x-axis and the average of the UN mean values as the y-axis. If our y-axis is not frequency, then we have to use the stat = "Identity" command in the geom_bar function so that we can assign values to the y-axis.

Figure 8. Bar plot of Average Value of National Examination from each Cluster

To get deeper information from the clustering results, then we will then make a map of the distribution of cluster 1 and cluster 2 data.

We can create maps with available data using the leaflet library. By using the leaflet function, we will enter the final data that has been divided into each cluster. The "addTiles" function is used to create a map base which will later be applied with the data that has been entered. Meanwhile, the "addCircleMarkers" function is used to apply data that has been entered into the map. "Lng" is a part for longitude data and "lat" for latitude data, used to display coordinates on the map. "Color" is used to color each coordinate on the map, "radius" for the number of radius markers of the coordinates, while "popup" is used to display the interface display area that will appear when the coordinates are clicked or pressed. The "addLegend" function is used to add a description or a legend to the map. The final view of the data distribution map is as follows.

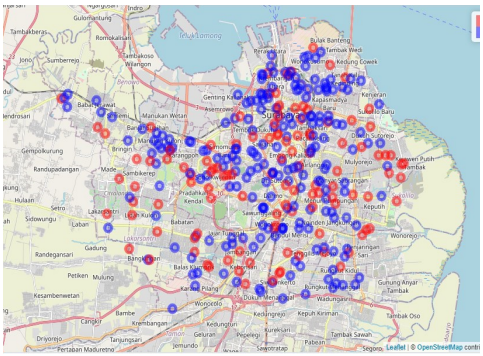

Figure 9. Map of the Distribution of Surabaya City Junior High Schools According to Clusters

Based on the results of the distribution of data on the map above, the distribution of school locations including cluster 1 and cluster 2 does not appear to be specifically collected in an area, only at a few points but is quite evenly distributed throughout the city of Surabaya. However, if you pay attention to the number of clusters and the position of their distribution in areas far from the capital, cluster 2 has more numbers than cluster 1.

4. Conclusion

After clustering, analysing, and testing data using the K-Means method, the following conclusions are obtained:

K-Means is one of the basic algorithms for grouping data that can be used and applied in grouping data on the results of the National Examination, especially for the data that has been used, namely the National Exam data for junior high school students in Surabaya in 2019.

The K-Means method has a fairly good ability in terms of data grouping or data classification based on the characteristics of the centroid (cluster centre) which is influenced by input parameters.

Data grouping using the K-Means method can draw a lot of information from a data set, for example, as in the data used previously, it can be seen that from the results of the National Examination for junior high school students in Surabaya, we can find out which school groups with average National Exam scores are high and low, as well as knowing the distribution of data in the area.

K-Means can help agencies in the school classification process based on the average National Examination results.

References

GIS4DEV. (2019, October 1). Google Maps: How to get longitude and latitude from google maps [EN]. Retrieved from https://www.youtube.com/watch?v=2KdqdzT98A4

Godfrey, K., Godfrey, K., K., Romero, J., L., Fit, P., . . . Roth, G. (2018, October 21). Determining The Optimal Number Of Clusters: 3 Must Know Methods. Retrieved from https://www.datanovia.com/en/lessons/determining-the-optimal-number-of-clusters-3-must-know-methods/

Kementerian Pendidikan dan Kebudayaan. (n.d.). LAPORAN HASIL UJIAN NASIONAL | KEMENTERIAN PENDIDIKAN DAN KEBUDAYAAN. Retrieved from

https://hasilun.puspendik.kemdikbud.go.id/#2019!smp!capaian!05&01&999!T&T&1&T&1&unbk!3!&

Leaflet for R - Introduction. (n.d.). Retrieved from http://rstudio.github.io/leaflet/

Sekolah menengah pertama. (n.d.). In *Wikipedia bahasa Indonesia, ensiklopedia bebas*. Retrieved April 30, 2021, from https://id.wikipedia.org/wiki/Sekolah_menengah_pertama

Umran, M., & Abidin, T. F. (2009). Pengelompokan Dokumen Menggunakan *K-Means* dan Singular Value Decomposition. *Studi Kasus Menggunakan Data Blog*, 2. Retrieved from http://www.informatika.unsyiah.ac.id/tfa/pdf/papers/SESINDO-2009.pdf

Xie, Y. (2015, June 24). Leaflet: Interactive web maps with R. Retrieved from https://blog.rstudio.com/2015/06/24/leaflet-interactive-web-maps-with-r/

Yobero, C. (2018, January 2). RPubs - *K-Means Clustering Tutorial*. Retrieved from https://rpubs.com/cyobero/k-means

Ade Bastian, Harun Sujadi, dan Gigin Febrianto. (2018). Penerapan algoritma k-means *clustering* analysis pada penyakit menular manusia (studi kasus kabupaten majalengka). Retrievered from https://www.neliti.com/id/publications/238400/penerapan-algoritma-k-means-clustering-analysis-pada-penyakit-menular-manusia-st

Baginda Harahap, (2019). Penerapan Algoritma K-Means Untuk Menentukan Bahan Bangunan Laris (Studi Kasus Pada UD. Toko Bangunan YD Indarung) Retrieved from https://ptki.ac.id/jurnal/index.php/readystar/article/view/82/pdf