



K-Means Clustering for Determining Quality of Outdoor Temperature Based on BMKG Datasets

M. Rizky Ramadhan⁽¹⁾, Abd. Charis Fauzan⁽²⁾, Nurul Aziz Tri Wahyuni⁽³⁾,
Riska Fitri Nur Alifah⁽⁴⁾, Adi Diantoro⁽⁵⁾

Department of Computer Science, Universitas Nahdlatul Ulama Blitar, Indonesia

E-mail: ⁽¹⁾mohrizky0@gmail.com, ⁽²⁾abdcharis@unublitar.ac.id, ⁽³⁾nurulwahyuni191@gmail.com

Received: 15 December 2019; Revised: 10 March 2020; Accepted: 20 March 2020

Abstrack

The purpose of this study was to determine the quality of outdoor temperatures that are good for conducting activities. One of the factors that cause the body's healthy or not human temperature, either indoors or outdoors. The data used is the BMKG weather dataset. BMKG is Indonesian meteorological and geophysical agency. BMKG weather dataset is a type of quantitative data because it is numeric or nominal and can be calculated. At each day, BMKG also provides information about the temperature according to the state of each area. It is starting from the minimum temperature, maximum, average humidity, to the maximum wind speed. The method used is clustering using the k-means clustering with centroid value. This research resulted in outdoor temperature clustering, which is good, INTERMEDIATE, and bad quality. In the range of values 124.7 indicate good temperature, range values 133.1 indicate moderate temperature, range of values 146.8 indicate bad temperature. Based on research in 32 days produced 28 days of moderate temperature quality, two days of good temperature quality, and two days of poor quality.

Keywords: K-Means Clustering, Outdoor Temperature, BMKG Datasets

Introduction

Temperature change is a phenomenon that is experienced and can be felt directly by the changes and their impact on society. (Kotta, 2008) In conducting outdoor activities, it is necessary to pay attention to the quality of good outdoor temperature for an excellent performance. Likewise, unsuitable temperatures can endanger health, so people must know and understand the importance of good outdoor temperature quality for safe activities. BMKG also urges people to limit outdoor activities by being aware of outdoor temperatures that are too high or too low. Outdoor temperature is essential to understand so that people can determine what activities are suitable for outdoor temperature levels at that time. The public can also be on the lookout for things to look out for with outdoor temperature levels. In determining the outside

temperature group, there needs to be a clusterization of data of minimum temperature, maximum temperature, average humidity, and maximum wind speed. The determination of outdoor temperature quality will be divided into three groups, namely are good, intermediate, and bad. The determination of the quality of outdoor temperature is expected to help the community in carrying out activities according to the appropriate outdoor temperature.

Data clustering can be done through several methods. Two types of data clustering are often used in the process of grouping data, namely Hierarchical and Non-Hierarchical. In this study, we use the non-hierarchical K-Means data clustering method in clustering data. The reason for choosing K-Means compared to other methods is that the K-Means method can group large objects very quickly and thus speed up the group-

ing process. Besides, K-Means also has advantages that are reasons for choosing a method, the time needed to do learning is relatively faster, very flexible, easy to do, is very common to use and use simple principles. (Postrel, n.d.)

Researchers used references from previous research, namely the application of data mining on the eligibility of students in receiving Bidikmisi scholarships according to the specified criteria. (Rahayu, 2019), the grouping of products that sell well and not sell for ease in analyzing the grouping of product sales to self-service (Muningsih & Kiswati, 2015) and grouping levels of drug use for managing drug availability at hospital pharmacies (Taslim & Fajrizal, 2016). So the researchers conducted a study using the K-Means method on the BMKG dataset to determine the quality of the outdoor temperature at the Karangates Geophysical Station, Kab. Malang.

Material and Methods

Data Collection

The type of data used is quantitative data that is the type of data that can be calculated in the form of numbers or nominal. BMKG weather data sets are a type of quantitative data because they are numeric or nominal and can be calculated. More specifically, the data used is matrix data, which is the type of data that has objects and attributes. The data source used in the study is secondary data that is data sources obtained indirectly through an official website "BMKG Database Online Data Center," which can be accessed through http://dataonline.bmkg.go.id/akses_data, which is obtained and recorded by a Non-Departmental Government Institution (LPND), headed by a Head of Agency. The secondary data is basically in the form of published historical evidence or reports, which are then used to support the formation of research. Based on the data source used in this final project research, the data collection method that the author uses is to search, read and collect documents as references, such as articles and literature relating to data mining using the k-means clustering method.

Data Process / Normalization

Data normalization is a technique in the logical design of a relational database that groups the attributes of a relationship to form a good relationship structure (without redundancy). In database science, normalization is used to avoid various data anomalies and data incon-

sistencies. In determining outdoor weather, the data set used is the BMKG weather data set. The reason for using the BMKG weather data set is that a straightforward and flexible data set is needed. Besides that, this data set has a relatively good and reliable level of accuracy because in carrying out its duties and functions, the BMKG is coordinated by the Minister responsible for transportation.

The parameters used in BMKG weather data sets have been through the normalization process before the data testing process is carried out. The parameters used in the official BMKG online data website are minimum temperature, maximum temperature, average humidity, and average wind speed. These parameters are used as the main parameters in determining the class of good, bad, and being in a free place. Other parameters such as maximum wind direction and most wind direction are not included because this parameter does not significantly affect the results in class determination.

K-Means

Some of the simplest and most commonly used clustering techniques are K-means clustering. In detail, this technique uses a measure of dissimilarity to group objects, the degree of dissimilarity is the determinant in the division of data groups. (Metisen & Sari, 2015) The discrepancy can be translated into the concept of distance. Two objects are said to be similar if the distance of the two objects is close. The higher the distance value, the higher the dissimilarity value. K-Means can group large amounts of data with relatively fast and efficient computing time. (Handoko, 2016) However, K-Means has a weakness caused by determining the initial center of the cluster. The results of clusters formed from the K-Means method are highly dependent on the initiation of the initial center value of the cluster provided. (Sibuea & Safta, 2017) The steps of the K-Means method are as follows:

1. Determine the number of clusters to be formed.
2. Determine the centroid (cluster center point).
3. Calculate the distance of each data to each centroid using the correlation formula between two objects (Euclidean Distance) using Equation 1.

$$(a, b) = \sqrt{\sum_{i=1}^n (b_i - a_i)^2} \quad (1)$$

4. Grouping each data based on the closest distance between the data and the centroid.
5. Determine the position of the new centroid (Cb) by calculating the average value of existing data on the same centroid, using Equation 2.

$$C_b = \left(\frac{K_n}{n} \right) \tag{2}$$

Where Kn is the sum of all BMKG weather data set values that are members of the cluster and n is the total number of cluster members.

6. Return to step 3 if the new centroid position with the old centroid is not the same.

Results and Discussion

In Table 1, this research experiment data is based on original data from an online database center data object, the BMKG weather data set. The weather data set was recorded within 32 days, which was used as a reference in class determination. Attributes made as objects consist of the date, minimum temperature (Tn), maximum temperature (Tx), average humidity (RH_avg), and average wind speed (ff_avg).

Table 1. BMKG Weather Dataset

DATE	Tn	Tx	RH_avg	ff_avg
18-10-2019	20,9	35	72	4
19-10-2019	20,4	34,2	72	2
20-10-2019	21,4	33	77	3
21-10-2019	22,9	34,8	62	5
22-10-2019	20,6	33,8	75	3
23-10-2019	20,2	35,4	72	5
24-10-2019	21,6	35,1	74	4
25-10-2019	22,4	34	72	3
26-10-2019	22,8	33,4	73	4
27-10-2019	21,2	34,2	70	4
28-10-2019	22,9	33,7	74	3
29-10-2019	23,2	34,8	70	3
30-10-2019	21	34,2	74	2
31-10-2019	23	34,3	71	4
01-11-2019	23,2	32,4	74	5
02-11-2019	23,4	28,4	91	4
03-11-2019	23	27,4	87	2
04-11-2019	23	32	78	3
05-11-2019	22,5	33,2	67	3
06-11-2019	20,2	33,2	77	2
07-11-2019	22,2	34,2	71	3
08-11-2019	23,9	33,4	74	4
09-11-2019	23,2	33,6	72	3
10-11-2019	23,7	33,8	70	3
11-11-2019	24,2	34	69	3
12-11-2019	24,6	33,8	72	3
13-11-2019	24	34,2	76	2
14-11-2019	24,7	35,5	74	2
15-11-2019	23,3	35	73	2
16-11-2019	21	33,4	72	3
17-11-2019	23,1	35,4	73	2
18-11-2019	24	35,2	78	8

Next do the grouping of data using the K means algorithm, the following steps for its completion. Determine the number of clusters, the number of clusters is the number of groups that will be generated. In this study, the number of clusters to be used. It is as many as 3 clusters, namely BAD (C1), INTERMEDIATE (C2), and GOOD (C3). Determine the initial centroid, in determining the initial centroid (the center of the first cluster), is determined by finding the drinking value from the sum of each row of objects with the minimum value obtained is 124.7 as a GOOD cluster (C3), finding the maximum value of the results the sum of each row of objects with the maximum value obtained is 146.8 as a BAD cluster (C1), and looking for a value that approaches the average value of the sum of each row of objects with the average. value obtained is 133.1 as an INTERMEDIATE cluster (C2). The initial centroids obtained are:

$$C1 = (02-11-2019; 23.4; 28.4; 91; 4)$$

$$C2 = (11-17-2019; 23.1; 35.4; 73; 2)$$

$$C3 = (21-21-2019; 22.9; 34.8; 62; 5)$$

Calculate the distance of each existing data to each cluster center, using the Euclidean Distance Space equation:

Equation 3 shows the distance between the data on 18-10-2019 and the center of the BAD cluster (C1).

$$\sqrt{(20,9 - 23,4)^2 + (35 - 28,4)^2 + (72 - 93)^2 + (4 - 4)^2} = 20,26844839 \tag{3}$$

Equation 4 shows the distance between the data on 18-10-2019 and the center of the INTERMEDIATE cluster (C2).

$$\sqrt{(20,9 - 23,1)^2 + (35 - 35,4)^2 + (73 - 93)^2 + (2 - 4)^2} = 3,16227766 \tag{4}$$

Equation 5 shows the distance between 18-10-2019 and the center of the GOOD cluster (C3).

$$\sqrt{(20,9 - 22,9)^2 + (35 - 34,8)^2 + (62 - 93)^2 + (5 - 4)^2} = 10,24890238 \tag{5}$$

With the results of calculations from all data for each first cluster center presented in Table 2.

After calculating the data distance on the centroid, the next step is to group the data. Data grouping is done by finding the minimum value from the results of calculating the distance of each existing data against each cluster center.

Table 2. Results of Calculation of Weather Dataset Distance for Each Centroid.

DATE	BAD (C1)	INTERMEDIATE (C2)	GOOD (C3)
18-10-2019	20,26844839	3,16227766	10,24890238
19-10-2019	20,19009658	3,119294792	10,75220308
20-10-2019	14,90503271	5,064582905	15,31306632
21-10-2019	29,71884924	11,41928194	0
22-10-2019	17,1464282	3,716180835	13,38992158
23-10-2019	20,52413214	4,290687591	10,3754518
24-10-2019	18,36110018	2,709243437	12,11527961
25-10-2019	19,85849944	2,109502311	10,24158191
26-10-2019	18,69117439	2,844292531	11,1341816
27-10-2019	21,89703176	4,248529157	8,261355821
28-10-2019	17,84208508	2,220360331	12,21515452
29-10-2019	21,97726098	3,220248438	8,251666498
30-10-2019	18,23184028	2,617250466	12,5287669
31-10-2019	20,85593441	3,03644529	9,039729375
01-11-2019	17,49399897	4,360045871	12,24132346
02-11-2019	0	19,41880532	29,71884924
03-11-2019	4,6	16,12482558	26,24442798
04-11-2019	13,53218386	6,129437168	16,36612355
05-11-2019	24,51224184	6,496152708	5,632051136
06-11-2019	15,27350847	5,408326913	15,61569723
07-11-2019	20,88252858	2,692582404	9,265527508
08-11-2019	17,7270979	3,104834939	12,16388096
09-11-2019	19,7251109	2,291287847	10,27277957
10-11-2019	21,70829335	3,594440151	8,345058418
11-11-2019	22,737634	4,491102315	7,438413809
12-11-2019	19,81413637	2,60959767	10,38701112
13-11-2019	16,21727474	3,354101966	14,37254327
14-11-2019	18,57686734	1,889444363	12,51918528
15-11-2019	19,27615107	0,447213585	11,41052146
16-11-2019	19,81817348	3,226453161	10,46756896
17-11-2019	19,41880532	0	11,41928194
18-11-2019	15,2184099	7,864477096	16,32084557

The calculation is done manually using Microsoft Excel. BAD (C1) indicates cluster areas that have bad outdoor temperatures, INTERMEDIATE (C2) indicates cluster areas that have moderate outdoor temperatures, GOOD (C3) indicates cluster areas that have right outdoor temperatures. After all the data has been successfully grouped into the closest cluster, then begin to recalculate the new cluster center based on the average of members in the same centroid. By adding up the values of all-weather datasets that are members of the cluster then divided by the total number of cluster members: Equation 6 shows the new centroid value in the BAD cluster (C1) in the minimum part temperature value (Tn).

$$Cb = \left(\frac{23,4 + 23}{2} \right) = 23,2 \tag{6}$$

Equation 7 shows the new centroid value in the BAD cluster (C1) in the maximum part temperature value (Tx).

$$Cb = \left(\frac{28,4 + 27,4}{2} \right) = 27,9 \tag{7}$$

Equation 8 shows the new centroid value in the BAD cluster (C1) in the average humidity part value (RH_avg).

Table 3. Results of Grouping Data.

DATE	MIN	CLUSTER
18-10-2019	3,16227766	INTERMEDIATE(C2)
19-10-2019	3,119294792	INTERMEDIATE(C2)
20-10-2019	5,064582905	INTERMEDIATE(C2)
21-10-2019	0	GOOD(C3)
22-10-2019	3,716180835	INTERMEDIATE(C2)
23-10-2019	4,290687591	INTERMEDIATE(C2)
24-10-2019	2,709243437	INTERMEDIATE(C2)
25-10-2019	2,109502311	INTERMEDIATE(C2)
26-10-2019	2,844292531	INTERMEDIATE(C2)
27-10-2019	4,248529157	INTERMEDIATE(C2)
28-10-2019	2,220360331	INTERMEDIATE(C2)
29-10-2019	3,220248438	INTERMEDIATE(C2)
30-10-2019	2,617250466	INTERMEDIATE(C2)
31-10-2019	3,03644529	INTERMEDIATE(C2)
01-11-2019	4,360045871	INTERMEDIATE(C2)
02-11-2019	0	BAD(C1)
03-11-2019	4,6	BAD(C1)
04-11-2019	6,129437168	INTERMEDIATE(C2)
05-11-2019	5,632051136	GOOD(C3)
06-11-2019	5,408326913	INTERMEDIATE(C2)
07-11-2019	2,692582404	INTERMEDIATE(C2)
08-11-2019	3,104834939	INTERMEDIATE(C2)
09-11-2019	2,291287847	INTERMEDIATE(C2)
10-11-2019	3,594440151	INTERMEDIATE(C2)
11-11-2019	4,491102315	INTERMEDIATE(C2)
12-11-2019	2,60959767	INTERMEDIATE(C2)
13-11-2019	3,354101966	INTERMEDIATE(C2)
14-11-2019	1,889444363	INTERMEDIATE(C2)
15-11-2019	0,447213585	INTERMEDIATE(C2)
16-11-2019	3,226453161	INTERMEDIATE(C2)
17-11-2019	0	INTERMEDIATE(C2)
18-11-2019	7,864477096	INTERMEDIATE(C2)

$$Cb = \left(\frac{91 + 87}{2} \right) = 89 \tag{8}$$

Equation 9 shows the new centroid value in the BAD cluster (C1) in the average wind speed section value (ff_avg).

$$Cb = \left(\frac{4 + 2}{2} \right) = 3 \tag{9}$$

The results of the calculation from the whole are presented in Table 4.

Table 4. New Centroids.

CENTROID 2				
BAD (C1)	23,2	27,9	89	3
INTERMEDIATE (C2)	21,72068966	32,90344828	70,65517241	3,172413793
GOOD (C3)	22,7	34	64,5	4

After obtaining a new centroid (CENTROID 2) from each cluster, then begin by recalculating the data with the new cluster center, repeating step 3 until a final pattern is obtained where the cluster is no longer moving. In this study, the data was recalculated until the second iteration, where each cluster did not change again, and no more data moved from one cluster to another. The results can be seen in Table 5.

Table 5. Results and the Latest Pattern of Data Grouping.

DATE	MIN	CLUSTER	MIN	CLUSTER
18-10-2019	3.16227766	INTERMEDIATE(C2)	2.750003783	INTERMEDIATE(C2)
19-10-2019	3.119294792	INTERMEDIATE(C2)	2.570577513	INTERMEDIATE(C2)
20-10-2019	5.064582905	INTERMEDIATE(C2)	6.365999347	INTERMEDIATE(C2)
21-10-2019	0	GOOD(C3)	2.816025668	GOOD(C3)
22-10-2019	3.716180835	INTERMEDIATE(C2)	4.578974074	INTERMEDIATE(C2)
23-10-2019	4.290687591	INTERMEDIATE(C2)	3.700527006	INTERMEDIATE(C2)
24-10-2019	2.709243437	INTERMEDIATE(C2)	4.088052835	INTERMEDIATE(C2)
25-10-2019	2.109502311	INTERMEDIATE(C2)	1.871410158	INTERMEDIATE(C2)
26-10-2019	2.844292531	INTERMEDIATE(C2)	2.755828328	INTERMEDIATE(C2)
27-10-2019	4.248529157	INTERMEDIATE(C2)	1.75108936	INTERMEDIATE(C2)
28-10-2019	2.220360331	INTERMEDIATE(C2)	3.630074832	INTERMEDIATE(C2)
29-10-2019	3.220248438	INTERMEDIATE(C2)	2.498848724	INTERMEDIATE(C2)
30-10-2019	2.617250466	INTERMEDIATE(C2)	3.842247472	INTERMEDIATE(C2)
31-10-2019	3.03644529	INTERMEDIATE(C2)	2.09542279	INTERMEDIATE(C2)
01-11-2019	4.360045871	INTERMEDIATE(C2)	4.11943712	INTERMEDIATE(C2)
02-11-2019	0	BAD(C1)	2.3	BAD(C1)
03-11-2019	4.6	BAD(C1)	2.3	BAD(C1)
04-11-2019	6.129437168	INTERMEDIATE(C2)	7.511928683	INTERMEDIATE(C2)
05-11-2019	5.632051136	GOOD(C3)	2.816025668	GOOD(C3)
06-11-2019	5.408326913	INTERMEDIATE(C2)	6.63564851	INTERMEDIATE(C2)
07-11-2019	2.692582404	INTERMEDIATE(C2)	1.435067023	INTERMEDIATE(C2)
08-11-2019	3.104834939	INTERMEDIATE(C2)	4.107155671	INTERMEDIATE(C2)
09-11-2019	2.291287847	INTERMEDIATE(C2)	2.124107143	INTERMEDIATE(C2)
10-11-2019	3.594440151	INTERMEDIATE(C2)	2.276006997	INTERMEDIATE(C2)
11-11-2019	4.491102315	INTERMEDIATE(C2)	3.180994767	INTERMEDIATE(C2)
12-11-2019	2.6059767	INTERMEDIATE(C2)	3.30843627	INTERMEDIATE(C2)
13-11-2019	3.354101966	INTERMEDIATE(C2)	6.057789915	INTERMEDIATE(C2)
14-11-2019	1.889444363	INTERMEDIATE(C2)	5.308556813	INTERMEDIATE(C2)
15-11-2019	0.447213586	INTERMEDIATE(C2)	3.709787165	INTERMEDIATE(C2)
16-11-2019	3.226453161	INTERMEDIATE(C2)	1.613767315	INTERMEDIATE(C2)
17-11-2019	0	INTERMEDIATE(C2)	3.874020915	INTERMEDIATE(C2)
18-11-2019	7.864477036	INTERMEDIATE(C2)	9.365974927	INTERMEDIATE(C2)

Based on the iteration one and iteration two processes, it appears that the center of the cluster that is being processed remains the same, and no more data is moved from one cluster to another. So the final results of the weather dataset that have been successfully grouped are:

1. In the BAD cluster (C1), there are two days.
2. In the INTERMEDIATE cluster (C2), there are 28 days.
3. In the GOOD cluster (C3), there are two days.



Figure 1. Centroid range.

A result is obtained where the quality of the outdoor temperature in the INTERMEDIATE cluster (C2) is more dominant than other clusters. This can occur because the INTERMEDIATE(C2) centroid is at the midpoint centroid and has two ranges of distance, namely the distance between the GOOD(C3) centroid and the INTERMEDIATE(C2) centroid and the sparse range between the BAD(C1) centroid and the INTERMEDIATE(C2) centroid as in Figure 1.

Conclusion

Based on the results of the study, the clusterization of room temperature based on repeti-

tion stops at iteration two which shows that the cluster area that has an outdoor temperature that is INTERMEDIATE (C2) is more dominant than the other clusters, from 32 days produces 28 days of moderate temperature quality, two days right quality temperature, and two days of poor quality. This means that the Karang Kates Geophysics location in Malang Regency shows the state of the quality of the outdoor temperature being moderate or the condition of relatively average temperature to carry out activities, both indoors and outdoors.

Suggestion

In this study, which has reached the results concluded, then this study can be used as a reference for parties related to the location of the dataset to make decisions in anticipating weather based on conditions and conditions. It is recommended that further research be carried out using methods and adding dataset attributes, as well as a broader range of regions so that in future studies, it will be more accurate and more thorough, and also the better data sets taken are the latest data.

Reference

A.E. Rahayu, K. Hikmah, N. Yustia, A.C. Fauzan. (2019). Penerapan K-Means Clustering Untuk Penentuan Klasterisasi Beasiswa Bidikmisi Mahasiswa. *ILKOMNIKA: Journal of Computer Science and Applied Informatics* 1 (2), 82-86

Handoko, K. (2016). Penerapan Data Mining Dalam Meningkatkan Mutu Pembelajaran Pada Instansi Perguruan Tinggi Menggunakan Metode K-Means Clustering (Studi Kasus Di Program Studi Tkj Akademi Komunitas Solok Selatan). *Jurnal Teknologi Dan Sistem Informasi*, 02(03), 31-40.

Kotta, M. H. (2008). *SUHU NETRAL DAN RENTANG SUHU NYAMAN MANUSIA INDONESIA (Studi Kasus Penelitian Pada Bangunan Kantor Di Makassar)*. 6, 23-29.

Metisen, B. M., & Sari, H. L. (2015). Analisis clustering menggunakan metode K-Means dalam pengelompokan penjualan produk pada Swalayan Fadhila. *Jurnal Media Infotama*, 11(2), 110-118.

Muningsih, E., & Kiswati, S. (2015). Penerapan Metode K-Means untuk Clustering Produk Online Shop dalam Penentuan Stok Barang. *Jurnal Bianglala Informatika*, 3(1), 10-17.

Postrel, V. (n.d.). *10 Clustering*. (1).

Ramadhani, R. D. (2014). Data Mining Menggunakan Algoritma K-Means Cluster-

- ing Untuk Menentukan Strategi Promosi Universitas Dian Nuswantoro. *Industrial Marketing Management*, 1(1), 1–9. <https://doi.org/10.1016/j.indmarman.2016.05.016>
- Sibuea, M. L., & Safta, A. (2017). Pemetaan Siswa Berprestasi Menggunakan Metode K-Means Clustering. *Jurteks*, 4(1), 85–92. <https://doi.org/10.33330/jurteks.v4i1.28>
- Taslim, T., & Fajrizal, F. (2016). Penerapan algoritma k-mean untuk clustering data obat pada puskesmas rumbai. *Digital Zone: Jurnal Teknologi Informasi Dan Komunikasi*, 7(2), 108–114. <https://doi.org/10.31849/digitalzone.v7i2.602>
- Marzuki, I. (2013). *Temu Kembali Informasi Big Data Menggunakan K-Means Clustering*. (June).
- Rizky, A., & Amiq, F. (2013). *Penerapan Metode Clustering Dengan Algoritma K-Means Untuk Rekomendasi Pemilihan Jalur Peminatan Sesuai Kemampuan Pada Program Studi*.