CHAPTER 4

ANALYSIS AND DESIGN

4.1 Analysis

The purpose of this project is to classify and analyze the result of weather classification results for a certain period of time. The data is gained from http://dataonline.bmkg.go.id/ .

1) Data Collection

   The weather data acquired from http://dataonline.bmkg.go.id/ contains 10 attributes which is :

Table 1.1 : Raw Data Attributes

Attributes	**Information**
Tn	Minimum Temperature
Tx	Maximum Temperature
Tavg	Average Temperature
RH_avg	Average Humidity
RR:	Rainfall
ss	Sunshine
ff_x	Maximum Wind Velocity
ddd_x	Maximum Wind Direction
ff_avg	Average wind Velocity
ddd_car	Most wind Direction

For the calculation test, the process only choose 5 attributes which is:

Table 1.2 : Used Data Attributes

Attributes	**Information**
Tavg	Average Temperature
RH_avg	Average Humidity
RR	Rainfall
ss	sunshine
ff_avg	Average Wind Velocity

Table 1.3 : Data

Tanggal	**T_Avg**	**RH_avg**	**RR**	**SS**	**Ff_avg**
1/2/2021	26.8	90	12.7	4.7	4
2/2/2021	27	90	6.7	6.2	4
3/2/2021	28.8	90	5	3.3	6
4/2/2021	26	90	71	2.5	5
5/2/2021	26.7	90	29	3.2	5
6/2/2021	25	90	173.5	4	3
7/2/2021	26.1	94	42.1	0	1
8/2/2021	25.2	96	41.1	0	3
9/2/2021	26.8	90	33.7	0.2	3
10/2/2021	27	94	6.2	6.8	2

2) Processing Data

The next step is to determine the number of cluster, in this example 2 clusters are determined with centroid as follows:

Table 1.4 : Initial Centroid

Cluster	Average Temperature	Average Humidity	Rainfall	Sunshine	Average Wind Velocity
C1	26.8	90	12.7	4.7	4
C2	27	90	6.7	6,2	4

Euclidian Distance is used to calculate distance between each point. The formula is :

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

where

$x_i = First\,Point$

$y_i = Second\,Point$

First, we need to calculate the distance from all datapoint to each centroid using Euclidian distance, For example:

First Iteration

First data point (26.8,90,12.7,4.7,4) to C1:

$$d(1,1)=\sqrt{(26.8-26.8)^2+(90-90)^2+(12.7-12.7)^2+(4.7-4.7)^2+(4-4)^2}=0$$

First data point (26.8,90,12.7,4.7,4) to C2:

$$d(1,2)=\sqrt{(26.8-27)^2+(90-90)^2+(12.7-6.7)^2+(4.7-6.2)^2+(4-4)^2}=6.187891402$$

Second Data Point (27,90,6.7,6.2,4) to C1:

$$d(2,1)=\sqrt{(27-26.8)^2+(90-90)^2+(6.7-12.7)^2+(6.2-4.7)^2+(4-4)^2}=6.187891402$$

Second Data Point (27,90,6.7,6.2,4) to C2:

$$d(2,2)=\sqrt{(27-27)^2+(90-90)^2+(6.7-6.7)^2+(6.2-6.2)^2+(4-4)^2}=0$$

Third Data Point (26.8,90,5,3.3,6) to C1:

$$d(3,1)=\sqrt{(26.8-26.8)^2+(90-90)^2+(5-12.7)^2+(3.3-4.7)^2+(6-4)^2}=8.077747211$$

Third Data Point (26.8,90,5,3.3,6) to C2:

$$d(3,2)=\sqrt{(26.8-27)^2+(90-90)^2+(5-6.7)^2+(3.3-6.2)^2+(6-4)^2}=3.916631206$$

Fourth Data Point (26,90,71,2.5,5) to C1:

$$d(4,1)=\sqrt{(26-26.8)^2+(90-90)^2+(71-12.7)^2+(2.5-4.7)^2+(5-4)^2}=58.35554815$$

Fourth Data Point (26,90,71,2.5,5) to C2:

$$d(4,2)=\sqrt{(26-27)^2+(90-90)^2+(71-6.7)^2+(2.5-6.2)^2+(5-4)^2}=64.42189069$$

Fifth Data Point (26.7,90,29,3.2,5) to C1:

$$d(5,1)=\sqrt{(26.7-26.8)^2-(90-90)^2-(29-12.7)^2-(3.2-4.7)^2-(5-4)^2}=16.39969512$$

Fifth Data Point (26.7,90,29,3.2,5) to C2:

$$d(5,2)=\sqrt{(26.7-27)^2+(90-90)^2+(29-6.7)^2+(3.2-6.2)^2+(5-4)^2}=22.52509711$$

Sixth Data Point (25,90,173.5,4,3) to C1:

$$d(6,1)=\sqrt{(25-26.8)^2-(90-90)^2-(173.5-12.7)^2-(4-4.7)^2-(3-4)^2}=160.814707$$

Sixth Data Point (25,90,173.5,4,3) to C2:

$$d(6,2)=\sqrt{(25-27)^2+(90-90)^2+(173.5-6.7)^2+(4-6.2)^2+(3-4)^2}=166.8294938$$

Seventh Data Point (26.1,94,42.1,0,1) to C1:

$$d(7,1)=\sqrt{(26.1-26.8)^2+(94-90)^2+(42.1-12.7)^2+(0-4.7)^2+(1-4)^2}=30.19834433$$

Seventh Data Point (26.1,94,42.1,0,1) to C2:

$$d(7,2)=\sqrt{(26.1-27)^2-(94-90)^2-(42.1-6.7)^2-(0-6.2)^2-(1-4)^2}=36.29614305$$

Eighth Data Point (25.2,96,41.1,0,3) to C1:

$$d(8,1)=\sqrt{(25.2-26.8)^2-(96-90)^2-(41.1-12.7)^2-(0-4.7)^2-(3-4)^2}=29.46540344$$

Eighth Data Point (25.2,96,41.1,0,3) to C2:

$$d(8,2)=\sqrt{(25.2-27)^2+(96-90)^2+(41.1-6.7)^2+(0-6.2)^2+(3-4)^2}=35.52520232$$

Ninth Data Point (26.8,90,33.7,0.2,3) to C1:

$$d(9,1)=\sqrt{(26.8-26.8)^2+(90-90)^2+(33.7-12.7)^2+(0.2-4.7)^2+(3-4)^2}=21.5$$

Ninth Data Point (26.8,90,33.7,0.2,3) to C2:

$$d(9,2)=\sqrt{(26.8-27)^2+(90-90)^2+(33.7-6.7)^2+(0.2-6.2)^2+(3-4)^2}=27.67742763$$

Tenth Data Point (27,94,6.2,6.8,2) to C1:

$$d(10,1)=\sqrt{(27-26.8)^2+(94-90)^2+(6.2-12.7)^2+(6.8-4.7)^2+(2-4)^2}=8.167006796$$

Tenth Data Point (27,94,6.2,6.8,2) to C2:

$$d(10,2)=\sqrt{(27-27)^2+(94-90)^2+(6.2-6.7)^2+(6.8-6.2)^2+(2-4)^2}=4.539823785$$

From the calculation above we can conclude that first datapoint is close to first cluster (0) rather than second cluster (6.187891402). The second datapoint is close to second cluster, and so on. The following is a table of the final results of first iteration.

Table 1.5 : 1st Iteration

Datapoint	C1	C2
1	0	6.187891402
2	6.187891402	0
3	8.077747211	3.916631206
4	58.35554815	64.42189069
5	16.39969512	22.52509711
6	160.814707	166.8294938
7	30.19834433	36.29614305
8	29.46540344	35.52520232
9	21.5	27.67742763
10	8.167006796	4.539823785

From the result above we can conclude that the data points belong to which cluster as follows

Table 1.6 : 1st Iteration Closest Cluster Result

Datapoint	C1	C2
1	1	
2		1
3		1
4	1	
5	1	
6	1	
7	1	
8	1	
9	1	
10		1

The next step is to calculate the point of the new centroid that has been formed.

Cluster 1 New Centroid

$$C(TempAvg,1)=\frac{26.8+26+26.7+25+26.1+25.2+26.8}{7}=26.08571429$$

$$C(HumiAvg,1)=\frac{90+90+90+90+94+96+90+26.8}{7}=91.42857143$$

$$C(Rain,1)=\frac{12.7+71+29+173.5+42.1+41.1+33.7}{7}=57.58571429$$

$$C(Sun,1)=\frac{4.7+2.5+3.2+4+0+0+0.2}{7}=2.085714286$$

$$C(Wind,1)=\frac{4+5+5+3+1+3+3}{7}=3.428571429$$

Cluster 2 New Centroid

$$C(TempAvg,1)=\frac{27+26.8+27}{3}=26.93333333$$

$$C(HumiAvg,2)=\frac{90+90+94}{3}=91.33333333$$

$$C(Rain,1)=\frac{6.7+5+6.2}{3}=5.96666667$$

$$C(Sun,1)=\frac{6.2+3.3+6.8}{3}=5.43333333$$

$$C(Wind,1)=\frac{4+6+2}{3}=4$$

Table 1.7 : Final Centroid 1st Iteration

Cluster	Average Temperature	Average Humidity	Rainfall	Sunshine	Average Wind Velocity
C1	26.08571429	91.42857143	57.58571429	2.085714286	3.428571429
C2	26.93333333	91.33333333	5.966666667	5.433333333	4

After the calculation Process on first Iteration, the final centroid on first iteration have been concluded on table above. Then the iteration process will continue until the centroid point does not change with the previouse iteration.

Second Iteration

On the second Iteration, the final centroid on first iteration will be used to calculate the distance between data point and centroid. Here are the result from the second iteration process:

Table 1.8 : 2^{nd} Iteration

Datapoint	C1	C2
1	44.99377054	6.904426921
2	51.08313352	1.705220742
3	52.68676401	3.358736535
4	13.58794484	65.12747329
5	28.69272111	23.20246634
6	115.9447687	167.5589084
7	16.02086013	36.76879716
8	17.26253464	35.91175914
9	24.01703477	28.27226753
10	51.6933758	3.61078631

Table 1.9 : 2nd Iteration Closest Cluster Result

Datapoint	C1	C2
1		1
2		1
3		1
4	1	
5		1
6	1	
7	1	
8	1	
9	1	
10		1

Table 1.10 : Final Centroid 2nd Iteration

Cluster	Average Temperature	Average Humidity	Rainfall	Sunshine	Average Wind Velocity
C1	25.82	92	72.8	1.34	3
C2	26.86	90.8	11.92	4.84	4.2

The 2nd Iteration was done and the result can be seen on table above. The centroid on 2nd iteration still differ from 1st iteration, then the iteration process will continue.

Third Iteration

Table 1.11 : Final Centroid 3rd Iteration

Cluster	Average Temperature	Average Humidity	Rainfall	Sunshine	Average Wind Velocity
C1	25.7	91.33333333	95.53333333	2.166666667	3
C2	26.61428571	90.8	11.92	4.84	4.2

The Calculation process on third iteration has been formed and the result can be seen on table above. The centroid on 3rd iteration still differ from 2nd iteration, then the iteration process will continue.

Fourth Iteration

Table 1.12 : Final Centroid 4th Iteration

Cluster	Average Temperature	Average Humidity	Rainfall	Sunshine	Average Wind Velocity
C1	25.5	90	122.5	3.25	4
C2	26.55	91.75	22.0625	3.05	3.5

The Calculation process has been done and the centroid formed on 4th iteration still differ from 3rd iteration, then the iteration process will continue.

Fifth Iteration

Table 1.13 : Final Centroid 5th Iteration

Cluster	Average Temperature	Average Humidity	Rainfall	Sunshine	Average Wind Velocity
C1	25	90	173.5	4	3
C2	26.48888889	91.55555556	27.5	2.988888889	3.666666667

The centroid result from 5th iteration still differ from 4th iteration, then the iteration process will continue.

Sixth Iteration

Table 1.14 : Final Centroid 6th Iteration

Cluster	Average Temperature	Average Humidity	Rainfall	Sunshine	Average Wind Velocity
C1	25	90	173.5	4	3
C2	26.48888889	91.55555556	27.5	2.988888889	3.666666667

Table 1.15 : Final Result of processing data

Data Point	C1	C2
1		1
2		1
3		1
4		1
5		1
6	1	
7		1
8		1
9		1
10		1

The calculation process has been done on 6th iteration and the centroid of the 6th iteration does not change from 5th iteration. The iteration will not continue and the final centroid has been determined on C1 (25, 90, 173.5, 4, 3) and C2 (26.48888889, 91.55555556, 27.5, 2.988888889, 3.666666667). The cluster result can be seen on table above, C2 have 9 data points, the characteristic result from C2 is the day with light rainfall and moderate rainfall, while in C1 have the characteristic the day with very heavy rainfall, The conclusion from February 2021 for the city of Semarang is that the citizens have to be careful on February 6th, because it has very high rainfall and has the potential for flooding

3) Cluster Quality check on Elbow Method and then with Silhouette Method

Elbow Method

The Elbow Method is used to determine the number of K by determining the percentage of compariosn results between amounts of clusters that will form and elbow at a point (after a point where the result starts to decrease linearly). The Elbow method use the WCSS (WithinCluster Sum Of Squares) on every amounts of clusters. Usually Elbow Method is used to determine the number of k on the K-Means Clustering

Elbow method use WCSS on every amounts of clusters, so we need to calculate the data and classify them into the range of clusters we want, in this case we will try to do WCSS testing from k = 1 to k = 4.

$$WCSS = \sum_{i=1}^{k} (xi - ci)^2$$

where

xi= data distribution on cluster i

ci= centroid of cluster i

K=1

Below is the final centroid on K=1

Table 1.16 : Final Centroid on K=1

Cluster	Average Temperature	Average Humidity	Rainfall	Sunshine	Average Wind Velocity
C1	25	90	173.5	4	3
C2	26.48888889	91.55555556	27.5	2.988888889	3.666666667

Table 1.17 : The amount of WCSS on K=1

Data Point	Distance from Centroid	Class	Distance with Centroid Squared
1	29.48361748	C1	869.2837
2	35.57228837	C1	1265.3877
3	37.20733395	C1	1384.3857
4	28.9757433	C1	839.5937
5	13.25412011	C1	175.6717
6	131.4188103	C1	17270.9037
7	4.808918797	C1	23.1257
8	5.776478166	C1	33.3677
9	9.024616335	C1	81.4437
10	36.22609143	C1	1312.3297
Sum			23255.493
WCSS			23255.493

K=2

Table 1.18 : Final Centroid on K=3

Cluster	Average Temperature	Average Humidity	Rainfall	Sunshine	Average Wind Velocity
C1	25	90	173.5	4	3
C2	26.48888889	91.55555556	27.5	2.988888889	3.666666667

Table 1.19 : The amount of WCSS on K=3

Data Point	Distance from Centroid	Class	Distance with Centroid Squared
1	14.98651245	C2	224.5955556
2	21.11263445	C2	445.7433333
3	22.67835483	C2	514.3077778
4	43.55370886	C2	1896.925556
5	2.556690569	C2	6.536666667
6	0	C1	0
7	15.34050557	C2	235.3311111
8	14.68850647	C2	215.7522222
9	7.012766137	C2	49.17888889
10	21.84556909	C2	477.2288889
Sum			4065.6
WCSS			4065.6

K=3

Table 4.18: Final Centroid on K=3

Cluster	Average Temperature	Average Humidity	Rainfall	Sunshine	Average Wind Velocity
C1	25	90	173.5	4	3
C2	26.16	92	43.38	21.18	3.4
C3	26.9	91	7.65	5.25	4

Table 1.20 : The amount of WCSS on K=3

Data Point	Distance from Centroid	Class	Distance with Centroid Squared
1	5.178320191	C3	26.815
2	1.677796174	C3	2.815
3	3.97932155	C3	15.835
4	27.7703511	C2	771.1924
5	14.75521603	C2	217.7164
6	0	C1	0
7	3.576926055	C2	12.7944
8	4.865429066	C2	23.6724
9	9.961546065	C2	99.2324
10	4.185092592	C3	17.515
Sum			1187.588
WCSS			1187.588

K=4

Table 1.21 : Final Centroid on K=4

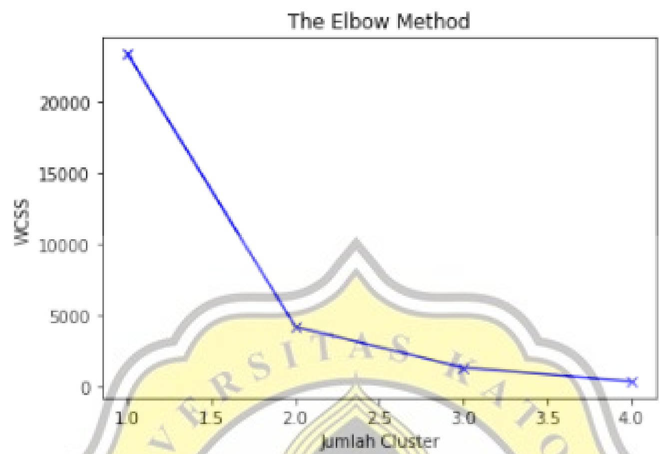
Cluster	Average Temperature	Average Humidity	Rainfall	Sunshine	Average Wind Velocity
C1	26.2	92.5	36.475	0.85	3
C2	26.9	91	7.65	5.25	4
C3	25	90	173.5	4	3
C4	26	90	71	2.5	5

Table 1.22 : The amount of WCSS on K-4

Data Point	Distance from Centroid	Class	Distance with Centroid Squared
1	5.178320191	C2	26.815
2	1.677796174	C2	2.815
3	3.97932155	C2	15.835
4	0	C4	0
5	8.479276207	C1	71.898125
6	0	C3	0
7	6.214750598	C1	38.623125
8	5.946690256	C1	35.363125
9	3.838375307	C1	14.733125
10	4.185092592	C2	17.515
Sum			223.5975
WCSS			223.5975

The calculation process has been done and the amount of WCSS from k=1 to k=4 are 23255.493, 4065.6, 1187.588, 223.5975. and ilustration below will show the graphs of WCSS.

Figure 1.1. Elbow Method



From that ilustration above, the point starts to decrease linearly and form a elbow at the amount of k is equal to 2. so based on Eblow Method the amount of cluster is set to 2

Silhouette Coefficient

Silhouette Coefficient is a method to check the quality amount of each data point to its cluster. Silhouette Coefficient has a range between -1 and 1. The amount of Silhouette Coefficient close to 1 shows that the data point has been classified properly or is right in the middle of the cluster, when is near to 0 indicates that data point is located between clusters, -1 indicates that data point is assigned to the wrong cluster. To check the quality amount of ideal cluster we need to compute the silhouette coefficients for each data point, and then average it all the samples to get the silhouette score.

Below is the formula to compute Silhouette Coefficient to each data points

$$s(i)=\frac{b(i)-a(i)}{max(b(i),a(i))}$$

where

a(i) = The average distance of that data point with all other points on the same cluster

b(i)= The average distance of that data point with all member from closest cluster

Same with Elbow Method, Silhouette Score use every Silhouette Coefficient on each data point on every amounts of clus ters, so we need to calculate the data and classify them into the range of cluster we want, in this case we will try to do Silhouette Score testing from k = 2 to k = 4. Below is the example to calculate silhouette score for first data point on k=2

K=2

we need to calculate the first silhouette coefficient for first data point (26.8, 90, 12.7, 4.7, 4). to find the a(i) for first data point we need to calculate the average distance from data point to all data point in the same cluster . Bellow is the result for classification in K=2

Table 1.23 : Cluster result for K=2

Data Point	C1	C2
1		1
2		1
3		1
4		1
5		1
6	1	
7		1
8		1
9		1
10		1

Find a(i), calculate the average distance from data point to all data point in the same cluster .

a(1)= 6.187891402 + 8.077747211 + 58.35554815+ 16.39969512 + 30.19834433 + 29.46540344 + 21.5 + 8.167006796 / 8 = 22.29395456

a(2)= 6.187891402 + 3.916631206 + 64.42189069 + 22.52509711 + 36.29614305 + 35.52520232 + 27.67742763 + 4.539823785 / 8 = 25.1362634

and so on, below is the result for a(i) on each data point on K=2

Table 1.24 : a(i) result for K=2

Datapoint	**a(i)**
1	22.29395456
2	25.1362634
3	26.56471335
4	49.20128116
5	20.40512803
6	0
7	24.71544535
8	24.39186691
9	21.16802814
10	26.13087076

After determining the a(i), the next step is to calculate b(i), b(i) is the average distance from data point to all data point in the closest cluster. For example:

To calculate b(i) for first data point, we need to check the closest cluster from cluster that data points belong to other cluster, for first data point (b1) the closest cluster are C1, so we need to calculate the distance from data point to all members on C1

Table 1.25 : The distance from C2 to all other clusters on K=2

Cluster	Distance based on Centroid	Closest cluster
C2	146.0209003	C1

On C1 in the first cluster there is only one data point (25, 90, 173.5, 4, 3), then calculate the distance between data point to (25, 90, 173.5, 4, 3)

$$d(1,6)=\sqrt{(26.8-25)2+(90-90)2+(12.7-173.5)2+(4.7-4)2+(4-3)2}=160.814707$$

caculating the b(i)

$$b(1)=\frac{131.5415144}{1}=160.814707$$

below are the result of b(i) from K=2

Table 1.26 : b(i) result on k=2

Datapoint	b(i)
1	160.814707
2	166.8294938
3	168.5377702
4	160.814707
5	146.0209003
6	146.0700742
7	131.5415144
8	132.59638
9	131.5415144
10	167.3861703

Now we have the result of a(i) and b(i) on each data point on K=2, the next step is to calculate the silhouette coefficient on each cluster

$$s(1)=\frac{160.814707-22.29395456}{160.814707}=0.861368683$$

$$s(2)=\frac{166.8294938-25.1362634}{166.8294938}=0.849329619$$

and so on, bellow is the result of s(i) on each datapoint on K=2

Table 1.27 : S(i) result on K=2

Datapoint	a(i)	b(i)	s(i)
1	22.29395456	160.814707	0.861368683
2	25.1362634	166.8294938	0.849329619
3	26.56471335	168.5377702	0.842381246
4	49.20128116	160.814707	0.694049866
5	20.40512803	146.0209003	0.858813497
6	0	146.0700742	1
7	24.71544535	131.5415144	0.812109162
8	24.39186691	132.59638	0.816044247
9	21.16802814	131.5415144	0.839077205
10	26.13087076	167.3861703	0.843888711

The result of s(i) on each data point have been concluded, the next step is to count the average of all s(i) on each data point

$$Silhouette\ Score = \frac{0.861368683+0.849329619+0.842381246+0.694049866+0.858813497+1+0.812109162+0.816044247+0.839077205+0.843888711}{10} = 0.841706224$$

The silhouette score from k=2 is 0.841706224. The next step is to calculate the silhouette score for K=3 and K=4

K=3

Table 1.28 : s(i) result on K=3

Datapoint	**a(i)**	**b(i)**	**s(i)**
1	7.477548469	160.814707	0.760210465
2	4.881448797	37.28915216	0.849329619
3	6.252258147	38.75218647	0.838660506
4	34.91828937	63.48427296	0.449969453
5	19.16882405	21.641432	0.114253435
6	0	130.2125052	1
7	14.21666017	35.21423053	0.596280823
8	14.39291856	34.39081527	0.581489463
9	15.64095114	26.69510515	0.414089173
10	6.489742202	37.91554789	0.828836914

K=4

Table 1.29 : s(i) result on K=4

Datapoint	a(i)	b(i)	s(i)
1	7.477548469	24.39086072	0.693428266
2	4.881448797	30.50596753	0.839983807
3	6.252258147	31.93591546	0.804224865
4	0	137.1317916	1
5	11.55454372	21.641432	0.466091536
6	0	102.5353598	1
7	9.104038271	35.21423053	0.741467068
8	8.965860315	34.39081527	0.739294918
9	8.376977097	26.69510515	0.686198011
10	6.489742202	31.10883923	0.79138591

And now after the calculation process from K=2 to K=4 we now have all the silhouette coefficient score as follows:

Table 1.30 : Cluster quality result with Silhouette Score

Amount of Clusters	Silhouette Coefficient
2	0.841706224
3	0.645288222
4	0.776207438

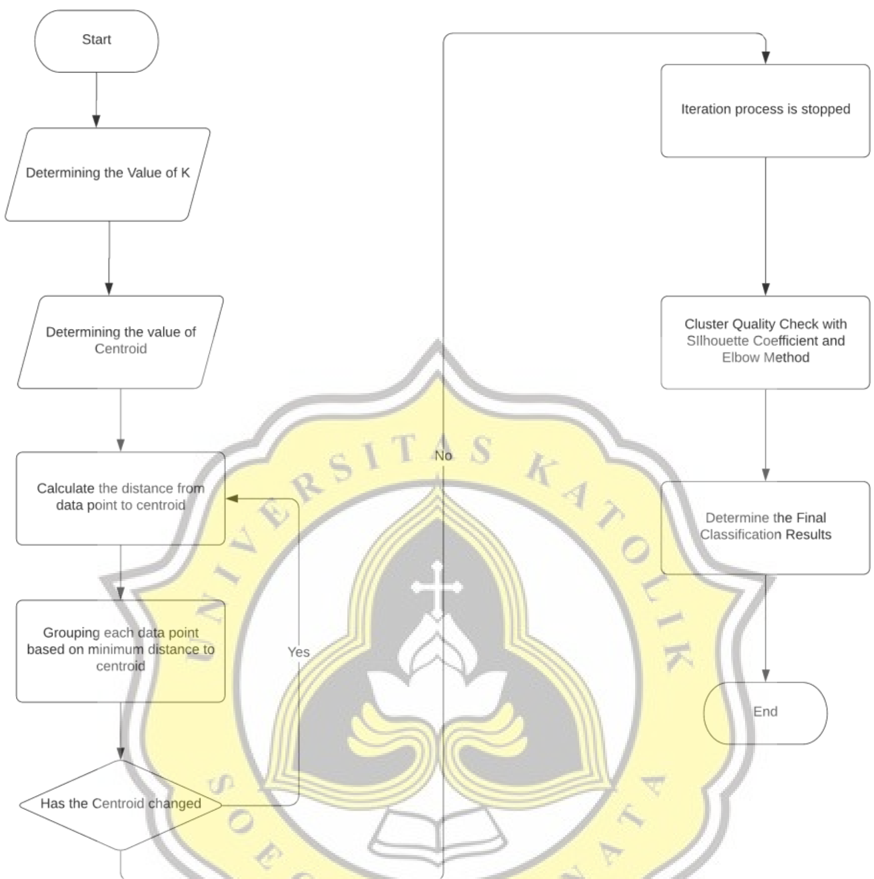
From the table above the amount of 2 cluster have the best score for silhouette coefficient, this makes the k=2 is an optimal clusters

4) Final Analysis

From the calculation and classification process above, we can conclude that the Elbow Method has the same value for k with Silhouette Coefficient. The amount of silhouette coefficient on k = 2 is 0.841706224, on k= 3 is 0.645288222 and k = 4 0.776207438, the highest amount of silhouette coeffficient is 0.841706224, so the optimal number of cluster is 2. C2 have 9 data points, the characteristic result from C2 is the day with light rainfall and moderate rainfall, while in C1 have the characteristic the day with very heavy rainfall, The conclusion from February 2021 for the city of Semarang is that the citizens have to be careful on February 6th, because it has very high rainfall and has the potential for flooding

4.2 Design

Figure 1.2. Flowchart Process



The work process of this project can be seen on the Flowchart above, the first step is determining the value of K and centroid by user, then the classification process is started, On each iteration the value of newly formed centroid will be compared with the previous iteration, if the centroid changes then the iteration process will continue