Nama : Daffashiddiq Nur Awan
Kelas : IF-42-12
NIM : 1301180311

Laporan Tugas Besar Machine Learning Tahap 1

A. Formulasi Masalah

Memprediksi akankah hari esok bersalju dari dataset salju. Dengan menggunakan metode unsupervised learning menggunakan k-means clustering.

B. Eksplorasi dan Persiapan Data

Data yang digunakan adalah salju_test dan salju_train di gabungkan. Berikut beberapa metode persiapan data yang saya gunakan.

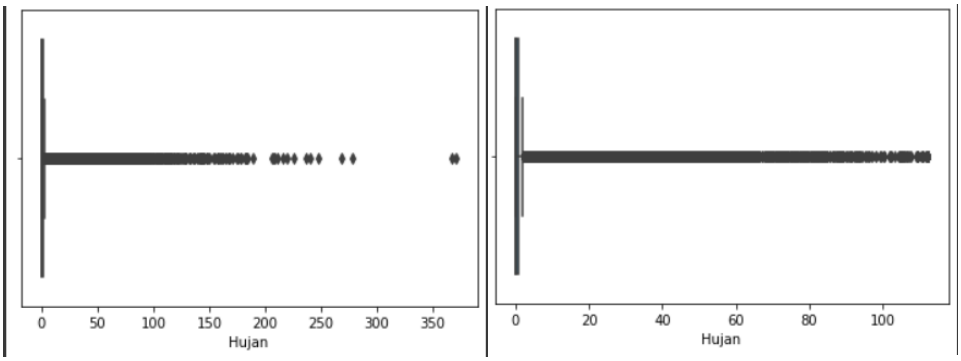
- Handling missing value : dilakukan karena data masih banyak yang bernilai Nan (melebihi 2% dari total dataset).

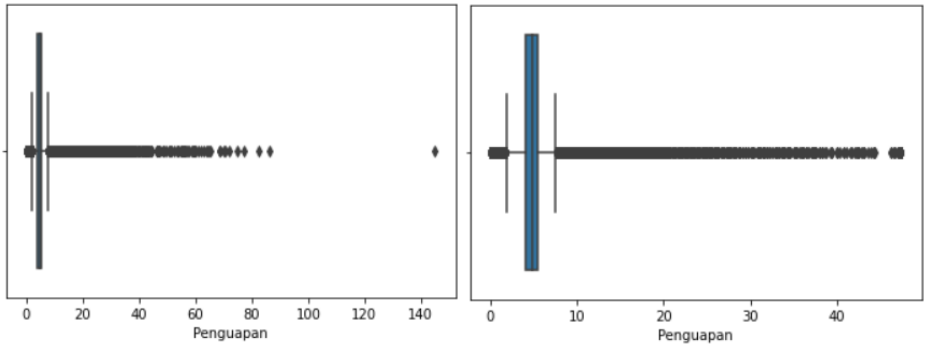
```
[424] dataTrain.isnull().sum()/len(dataTrain)*100

id                          14.285378
Tanggal                      0.000000
KodeLokasi                   0.000000
SuhuMin                      1.011180
SuhuMax                      0.859543
Hujan                        2.214069
Penguapan                   43.118552
SinarMatahari               48.003174
ArahAnginTerkencang          7.090833
KecepatanAnginTerkencang     7.047621
ArahAngin9am                 7.252685
ArahAngin3pm                 2.901545
KecepatanAngin9am            1.218602
KecepatanAngin3pm            2.087573
Kelembaban9am                1.832224
Kelembaban3pm                3.081468
Tekanan9am                  10.364795
Tekanan3pm                  10.340439
Awan9am                     38.409139
Awan3pm                     40.798416
Suhu9am                      1.224887
Suhu3pm                      2.467060
BersaljuHariIni              2.214069
BersaljuBesok                2.239211
dtype: float64
```

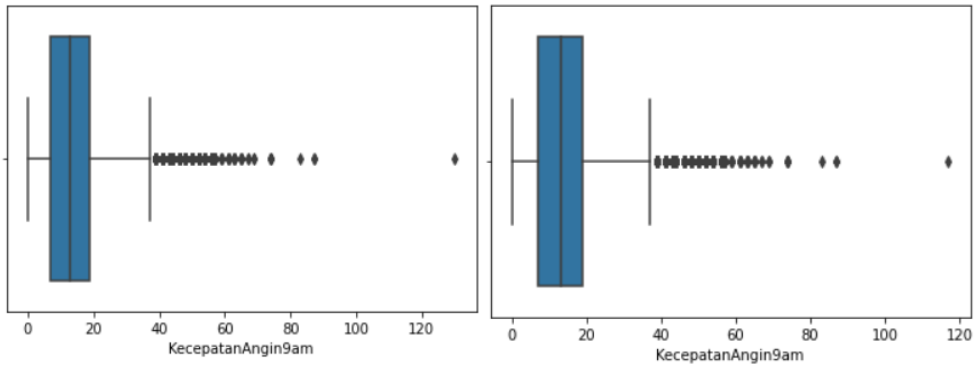
```
dataTrain.isnull().sum()

id                          0
Tanggal                     0
KodeLokasi                  0
SuhuMin                     0
SuhuMax                     0
Hujan                       0
Penguapan                   0
SinarMatahari               0
ArahAnginTerkencang         0
KecepatanAnginTerkencang    0
ArahAngin9am                0
ArahAngin3pm                0
KecepatanAngin9am           0
KecepatanAngin3pm           0
Kelembaban9am               0
Kelembaban3pm               0
Tekanan9am                  0
Tekanan3pm                  0
Awan9am                     0
Awan3pm                     0
Suhu9am                     0
Suhu3pm                     0
BersaljuHariIni             0
BersaljuBesok               0
dtype: int64
```

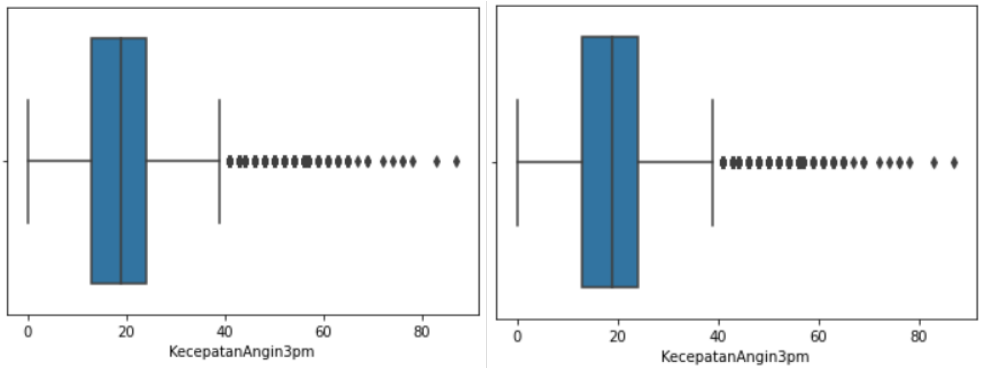
Nama : Daffashiddiq Nur Awan
Kelas : IF-42-12
NIM : 1301180311

- Outlier treatment : dilakukan karena data masih terdapat outlier, jika tidak diberikan tindakan maka data akan berefek dan bias.

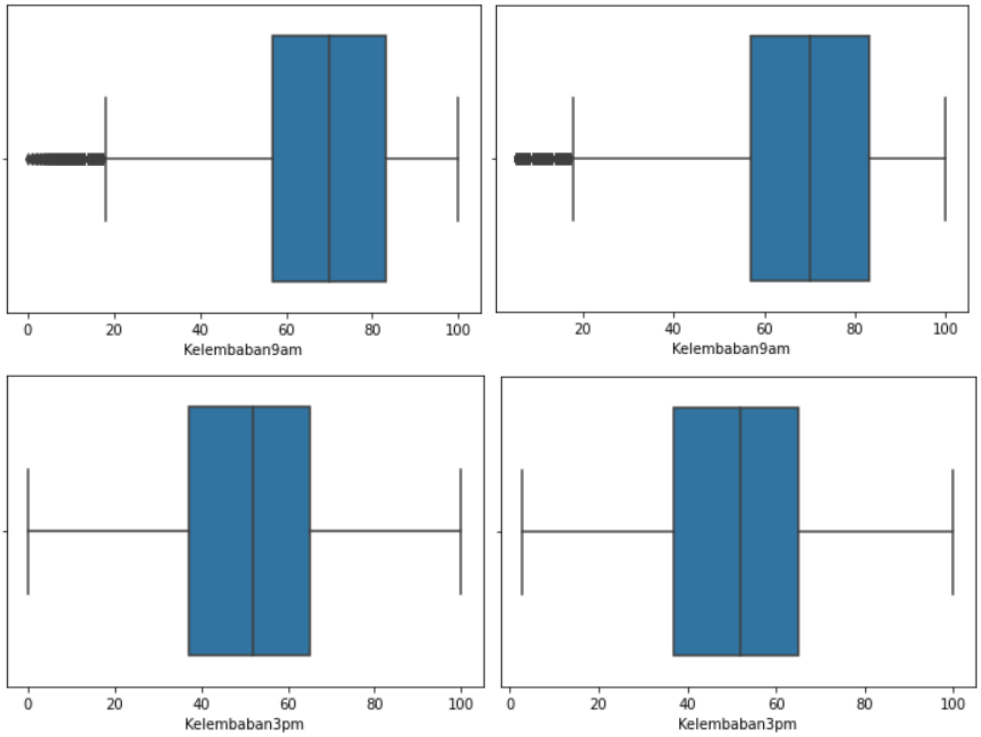






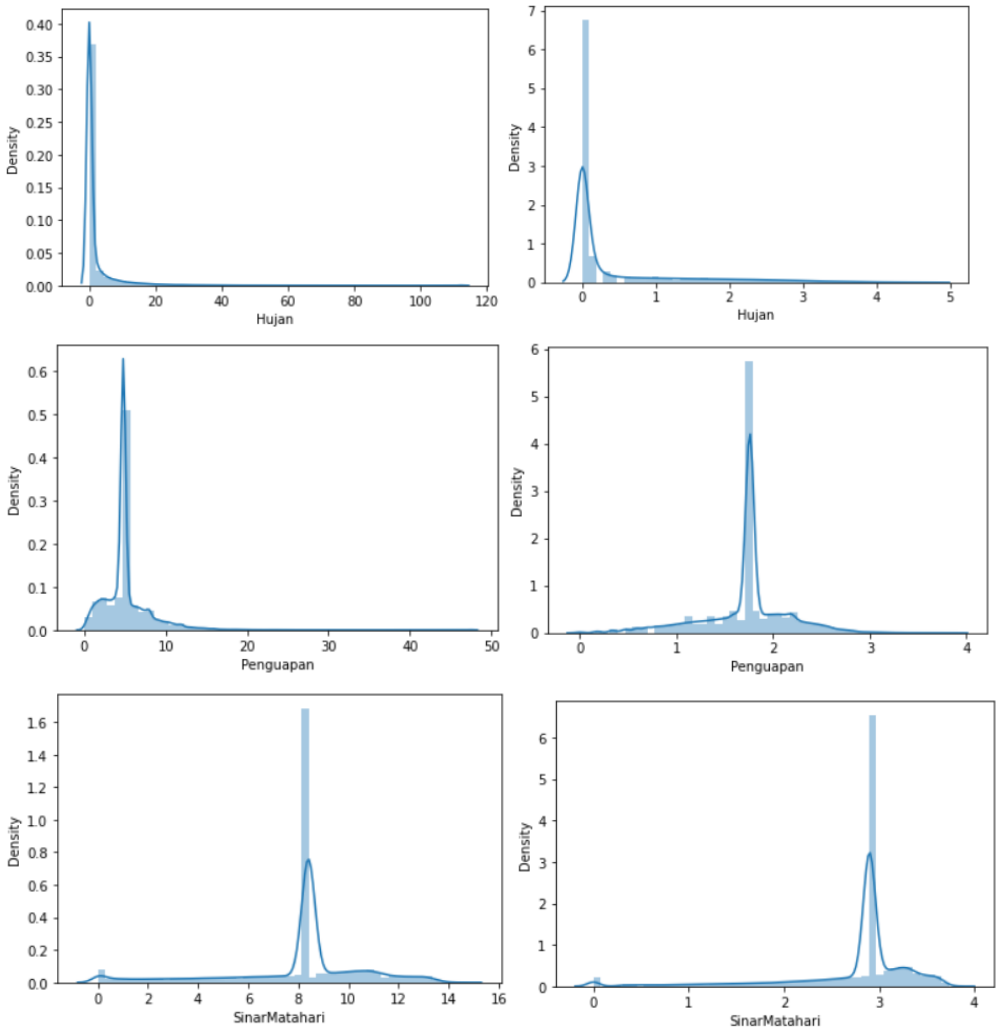


Nama : Daffashiddiq Nur Awan
Kelas : IF-42-12
NIM : 1301180311



Nama : Daffashiddiq Nur Awan
Kelas : IF-42-12
NIM : 1301180311

- Skewed Transformation : dilakukan agar data terdistribusi mendekati distribusi normal.



- Data Transformation : dilakukan agar data memiliki tipe data yang benar, contoh seperti kolom tanggal dibawah memiliki tipe object, maka diganti ke tipe date time menggunakan library pandas.

```
1   Tanggal                 123223 non-null  object
```

```
1   Tanggal                 123223 non-null  datetime64[ns]
```

- Data Deletion : dilakukan agar data yang tidak mempengaruhi model kluster karena data yang disediakan kolom tersebut tidak berpengaruh langsung dengan kluster.

Nama : Daffashiddiq Nur Awan
Kelas : IF-42-12
NIM : 1301180311

- Feature Engineering : dilakukan agar data yang digunakan lebih akurat dari data yang ada.

Tanggal	Bulan
14/11/2012	1
24/10/2015	7
31/08/2010	2
24/08/2013	8
26/10/2011	10

- Normalisasi: dilakukan agar tidak terjadinya ketimpangan antar data

C. Pemodelan

Model yang saya gunakan adalah k means clustering dengan nilai k = 2.

```
import random
kolom = df_norm.columns # menyimpan kolom dari df_norm
# dataframe dijadikan array
X = np.array(df_norm)
#inisialisasi awal centroid
centroid_awal = random.sample(range(0,len(df_norm)),2)
centroids = []
#mencari data dengan index centroid awal
for i in centroid_awal:
  centroids.append(df_norm.loc[i])
#dijadikan array dari list
centroids = np.array(centroids)
#euclidan distance
def euclid_distance(x1,x2):
  return(sum((x1-x2)**2))**0.5
#ic sebagai centroid, x sebagai dataframe
def findClosestCentroids(ic, X):
    assigned_centroid = []
    distElbow=[]
    for i in X:
        distance=[]
        for j in ic:
            distance.append(euclid_distance(i, j))            #jarak tiap centroid ke titik di dataFrame
        assigned_centroid.append(np.argmin(distance))       #jarak minimum dari ke K centroid akan diappe
        distElbow.append(np.min(distance))                  #nilai dari jarak minimum dimasukkan ke list
    return assigned_centroid,distElbow

def calc_centroids(clusters, X):
    new_centroids = []
    new_df = pd.concat([pd.DataFrame(X, columns = kolom), pd.DataFrame(clusters, columns=['cluster'])],
                       axis=1)                                                    #membuat dataFram
    for c in set(new_df['cluster']):
        current_cluster = new_df[new_df['cluster'] == c][new_df.columns[:-1]]          #mengelompokkan d
        cluster_mean = current_cluster.mean(axis=0)       #menghitung mean dari current_cluster
        new_centroids.append(cluster_mean)                #nilai mean dijadikan centroid baru

    return new_centroids,new_df       #nilai centroid baru dan dataframe di return

sum_dist=[]
for i in range(6):
  get_centroids,get_elbow = findClosestCentroids(centroids, X)
  centroids,df = calc_centroids(get_centroids, X)
  get_elbow = np.array(get_elbow)    #merubah tipe data menjadi array
  sum_dist.append(np.sum(get_elbow))  #ditambahkan ke list dari jumlah get_elbow
```

Proses-proses yang terdapat di algoritma tersebut:

- Merubah data frame menjadi np.array agar lebih cepat waktu komputasinya.
- Menentukan 2 titik random dari dataframe.
- Menghitung jarak dari 2 centroid ke tiap baris menggunakan euclidan distance.
- Memasukkan nilai yang minimum ke array.
- Membuat dataframe dengan mengabungkan array dataframe awal dengan array dari kluster.
- Menghitung nilai mean dari tiap kluster.

Nama : Daffashiddiq Nur Awan
Kelas : IF-42-12
NIM : 1301180311

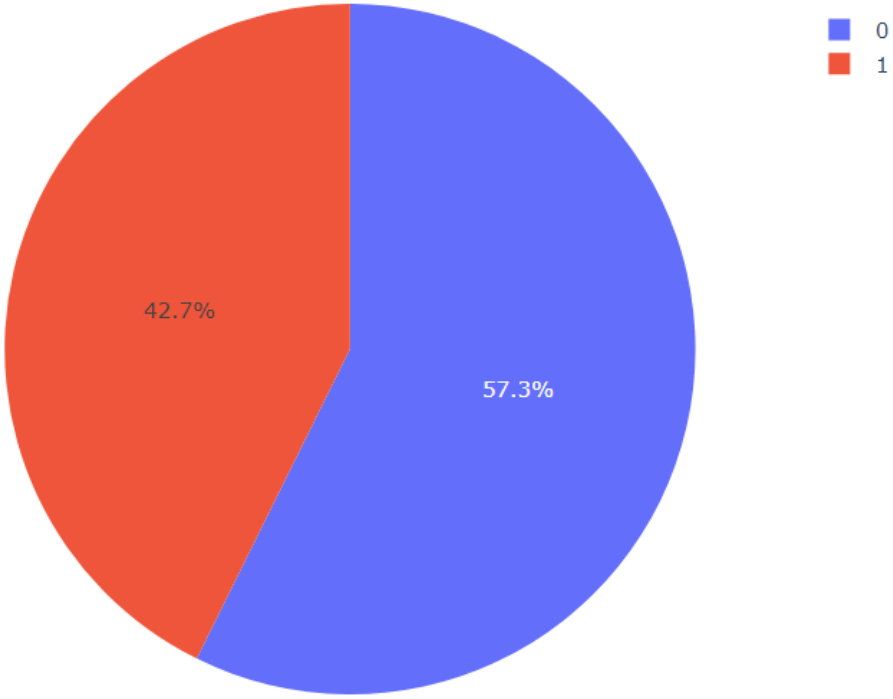
- Menjadikan nilai tersebut sebagai centroid selanjutnya.

Proses nilai Elbow:

- Menambahkan nilai dari jarak centroid dari tiap baris ke array
- Menambahkan nilai dari total jarak centroid ke array
- Ulang tiap nilai

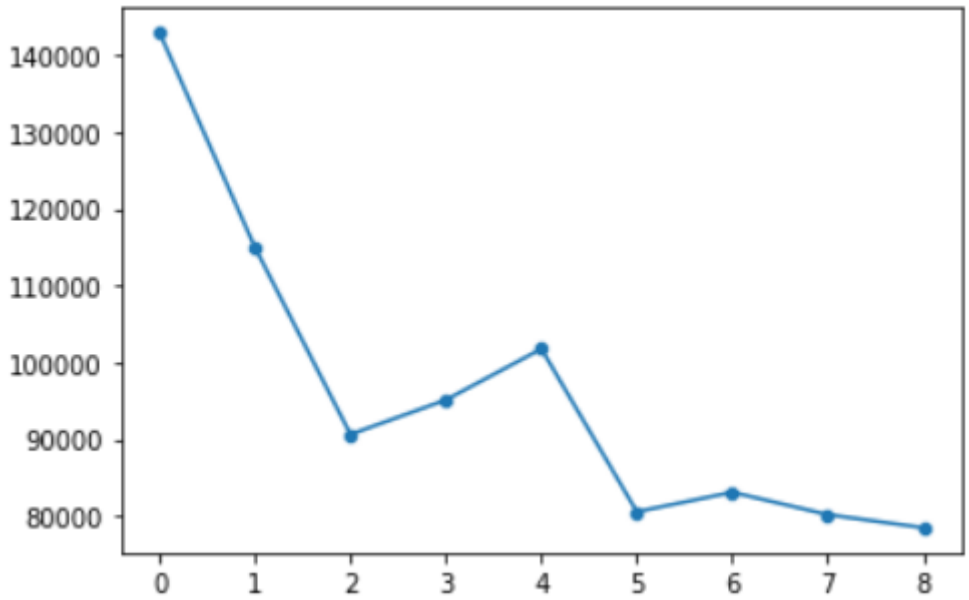
D. Evaluasi

Tahapan setelah memodelkan dataset.

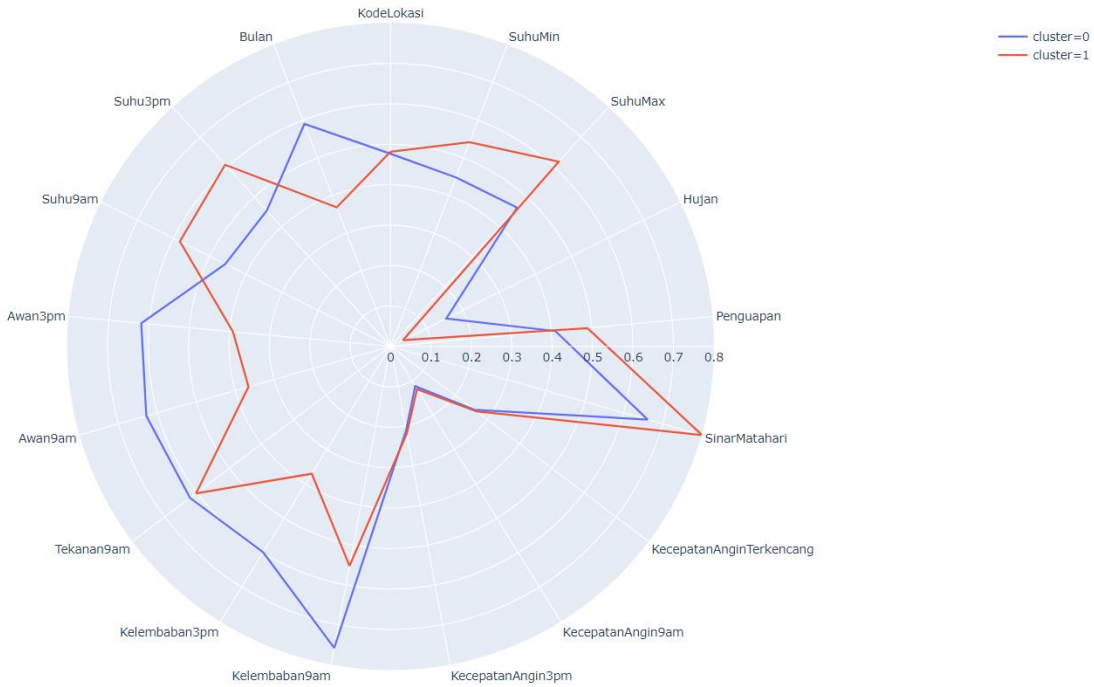


Dapat dilihat dataframe terbagi menjadi 2 kluster dengan kluster 0 (42.7%) dan kluster 1 (57.3%). Dataframe dibagi dengan lumayan rata.

Nama : Daffashiddiq Nur Awan
Kelas : IF-42-12
NIM : 1301180311

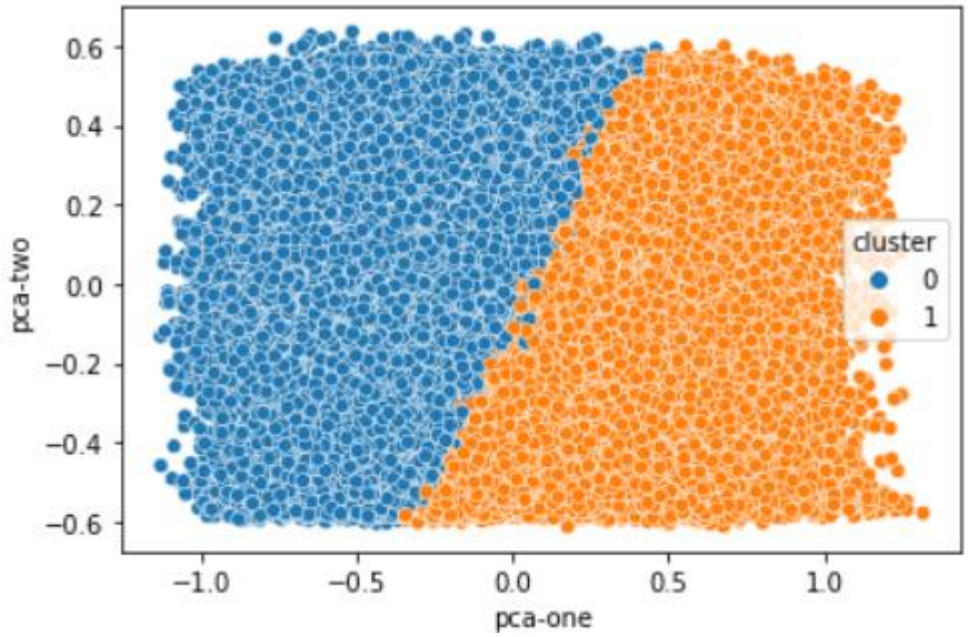


Dari diagram Elbow Method diatas dapat diambil kesimpulan bahwa k = 3 (ditambah 1 karena array dimulai dari 0) adalah nilai yang optimum untuk digunakan. Karena perubahan signifikan terlihat.



Dari diagram polar diatas terlihat beberapa karakteristik kluster kepada fitur yang disediakan.

Nama : Daffashiddiq Nur Awan
Kelas : IF-42-12
NIM : 1301180311

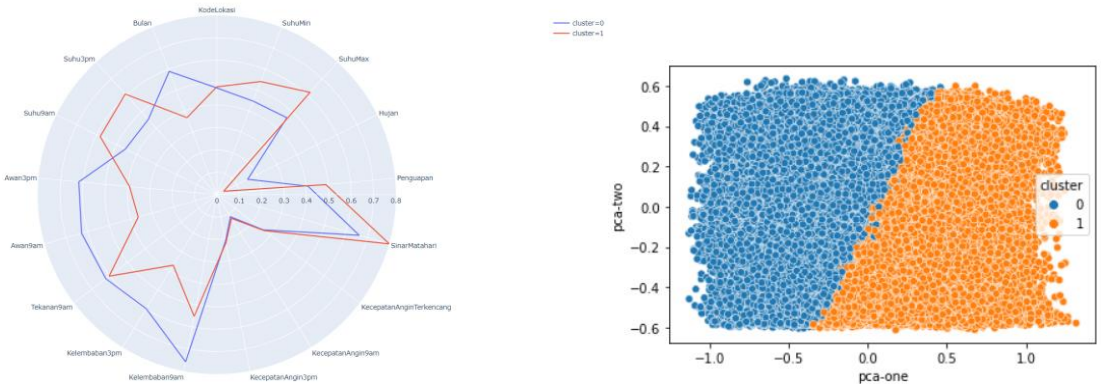


Scatter plot diatas menggunakan metode PCA (*Principal Component Analysis*). Metode tersebut digunakan untuk mereduksi dimensi menjadi 2 saja untuk dapat divisualisasikan dengan 2 dimensi.

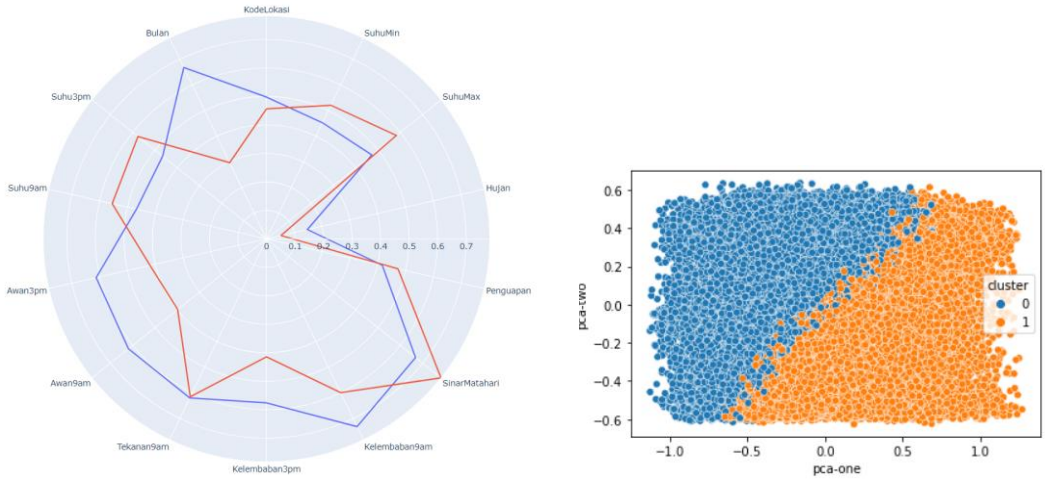
Nama : Daffashiddiq Nur Awan
Kelas : IF-42-12
NIM : 1301180311

E. Eksperimen

1. Masih Original

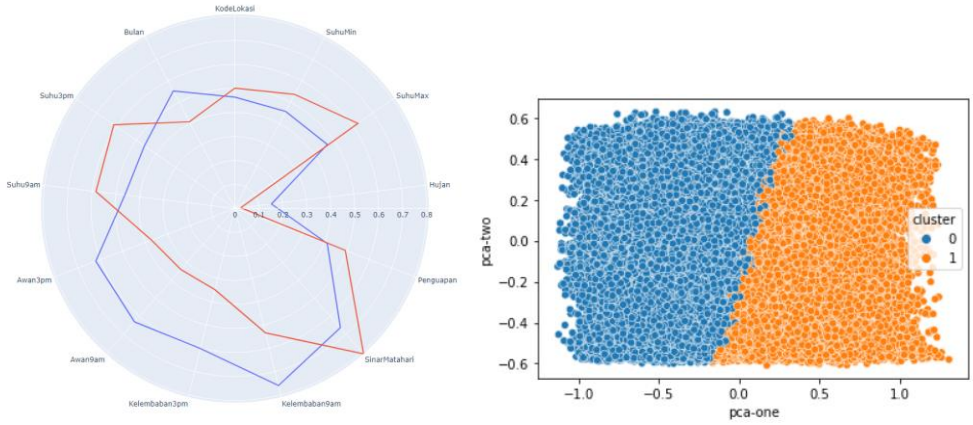


2. Kecepatan Angin 3am, kecepatan angin 9 am, dan kecepatan angin terkencang di drop karena pada diagram polar nilainya tidak terlihat berbeda antara dua kluster.



Terlihat pada diagram polar nilai kluster memiliki karakteristik masing masing pada fitur yang tersedia.

3. Fitur Tekanan 9am dibuang karena terlihat tidak berbeda.



Nama : Daffashiddiq Nur Awan
Kelas : IF-42-12
NIM : 1301180311

F. Kesimpulan

Pada penggunaan model k-means untuk clustering tidak di rekomendasikan karena pada setiap running nilai centroid awal ditentukan secara random. Dapat dilihat di setiap eksperimen pembagian kluster tidak pernah sama. Karakteristik dari kluster 0 yang lebih dominan pada fitur Awan3pm, Awan9am, Kelembaban3pm, Kelembaban9am, Bulan, Hujan. Berikut link Youtube https://www.youtube.com/watch?v=7fdvanmeI1Y dan link data kluster https://drive.google.com/file/d/12M9A3EM1dk2m9zu6fp-Gax04rOfv9Kb2/view?usp=sharing karena melebihi kuota di lms.