Nama : Delvanita Sri Wahyuni
NIM : 1301184014
Kelas : IF 42 04

Laporan Tugas Besar Machine Learning

1. **Ringkasan**

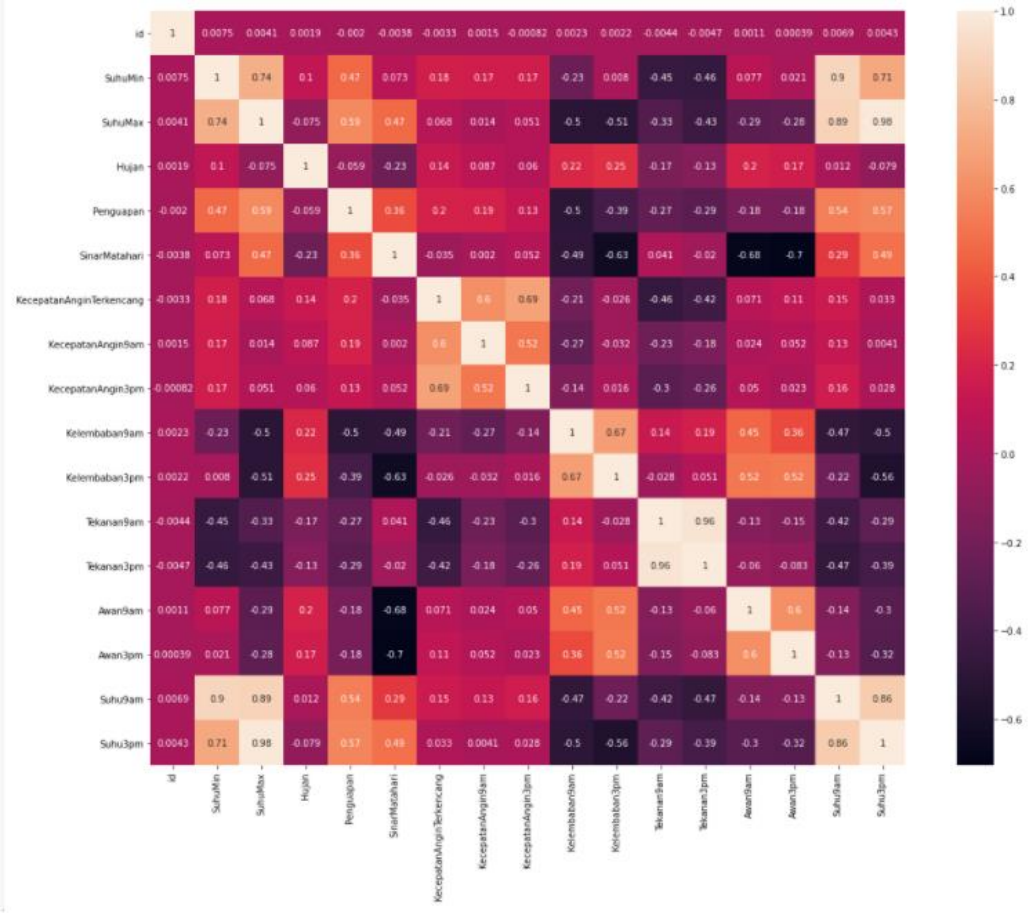
   Pada tugas ini terdapat dataset “salju_train.csv” dengan kolom 24 dan 109095 data. Data tersebut akan dilakukan clustering dan untuk mendapatakan data itu dipilihlah 6 atribut dari 24 atribut yaitu : SuhuMin, SuhuMax, KecepatanAngin9am, KecepatanAngin3pm, Suhu9am dan Suhu3pm.

2. **Data Preparation & Data Exploration**

   Terdapat dataset “salju_train.csv” dengan jumlah baris 109095 dan jumkah kolom 24. Data tersebut akan dibuat clustering dengan dipilih 6 dari 24 artibut yang berpotensi untuk menentukan apakah salju besok turun atau tidaknya. Atribut yang dipakai yaitu : Suhu Min, Suhu Max, Kecepatan Angin 9am, Kecepatan Angin 3pm, Suhu 9am dan Suhu 3am.

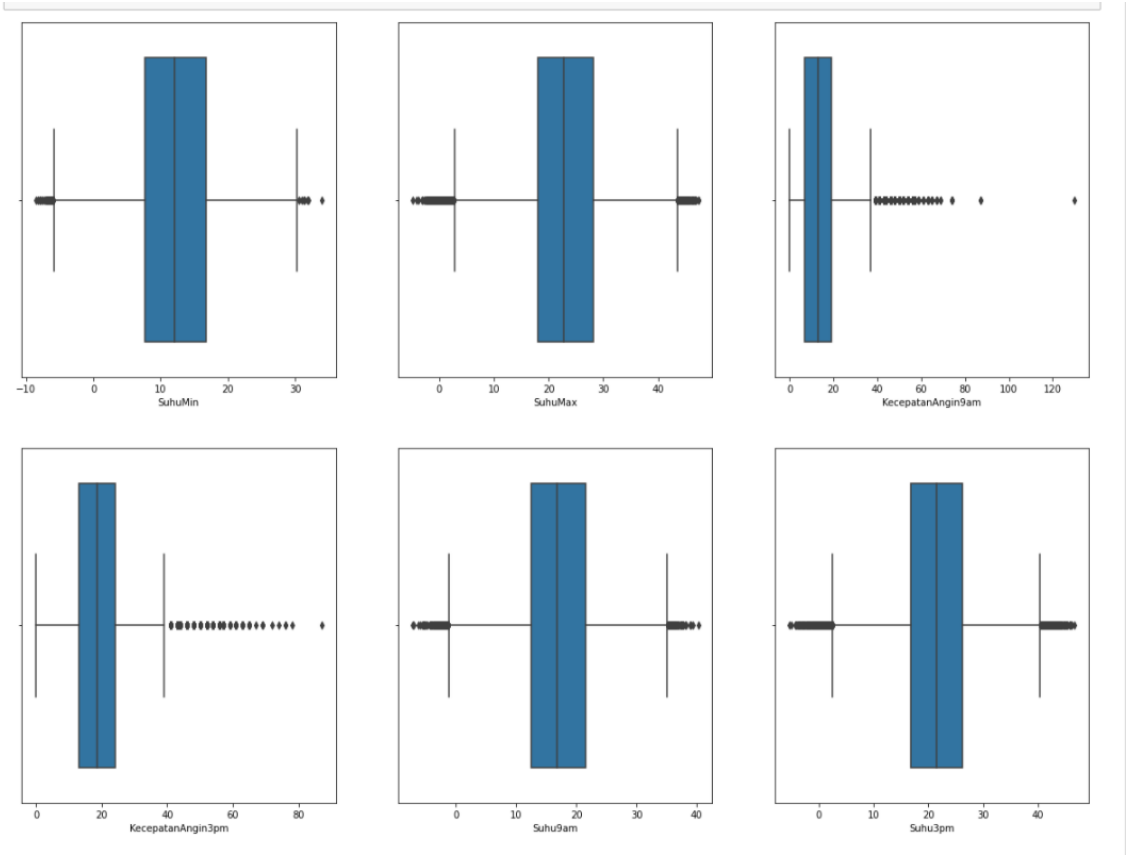
   Teknik yang diperlukan yaitu memasukan dataset agar data itu bisa diolah untuk dibaca jumlah baris dan kolom yang ada, lalu membuat data ringkasan statistik data menggunakan “data2_describe”, penggunaan “dytpes” yaitu untuk menampilkan penggunaan tipe data apa saja yang dipakai oleh statistik data yang telah dimasukan data tersebut akan menampilkan nama kolom dan tipe data yang digunakan. Selanjutnya yaitu teknik “get_numeric_data” yang digunakan sebagai mendapatkan nilai kolom yang bertipe datakan numerik.

   Untuk mengetahui korelasi antar fitur yang dipilh dari data numerik yaitu mengggunakan plot correlation matrix, hasilnya yaitu :



Heatmap correlation

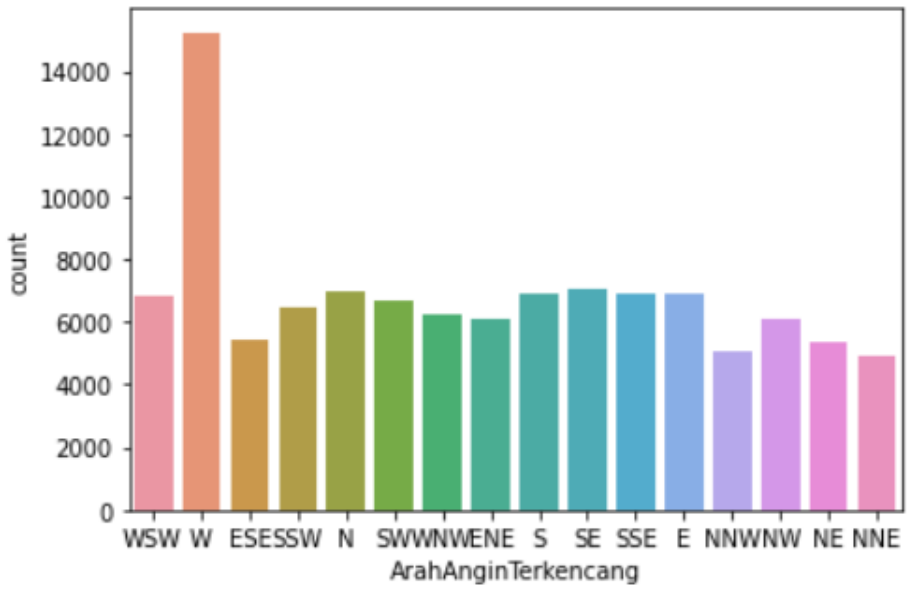
Melakukan missing value dengan nilai null menggunakan nilai mean sebagai pengganti nilai atribut yang bernilai “NaN” menjadi “0.0”. Lalu memlilih atribut yang sekiranya bisa berpotensin untuk memprediksi salju turun besok. Data tersebut dilakukan exploarsi untuk melihat perbedaan dari setiap atribut yang dipilih.



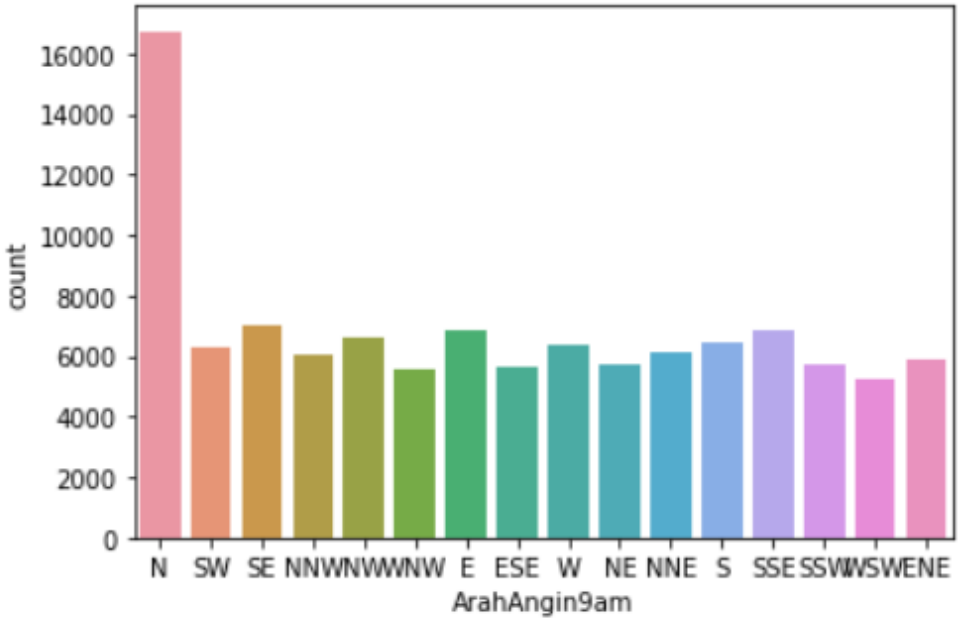
Boxplot

Selain itu, untuk menghitung jumlah kejadian yang sama pada artibut terdapat 5 atribut yang bertipe data object. Baris tersebut dapat dihitung utuk melihat berapa banyak kejadian yang sama pada kolom atribut yang dipilih yaitu :

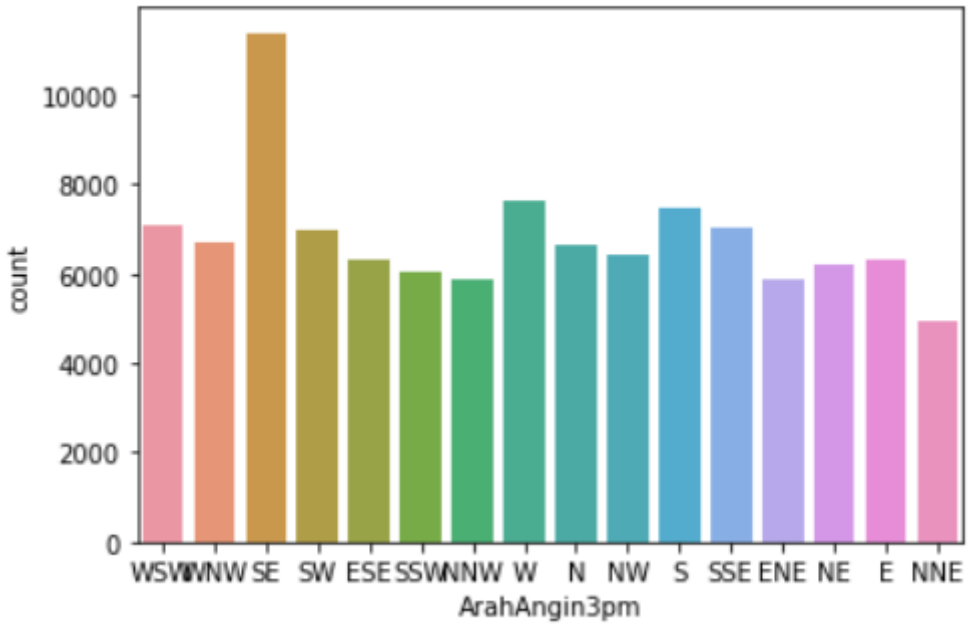
```
: <AxesSubplot:xlabel='ArahAnginTerkencang', ylabel='count'>
```



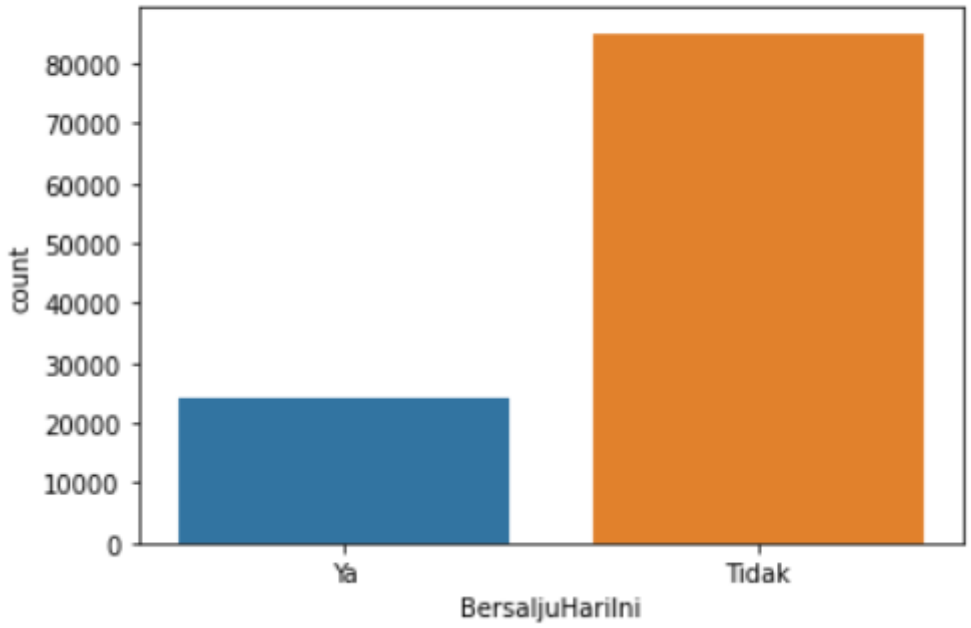
|: <AxesSubplot:xlabel='ArahAngin9am', ylabel='count'>



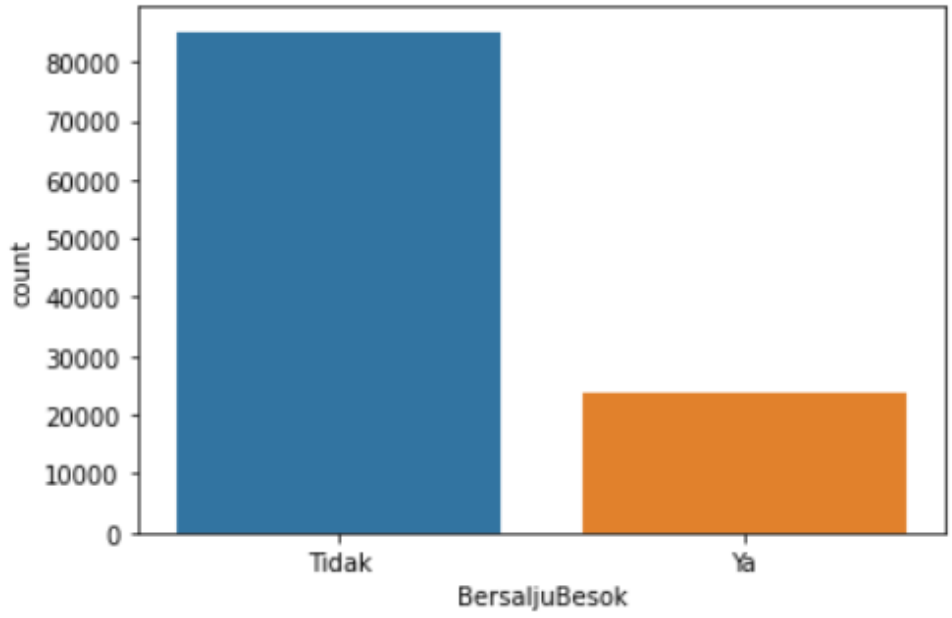
<AxesSubplot:xlabel='ArahAngin3pm', ylabel='count'>



```
<AxesSubplot:xlabel='BersaljuHariIni', ylabel='count'>
```

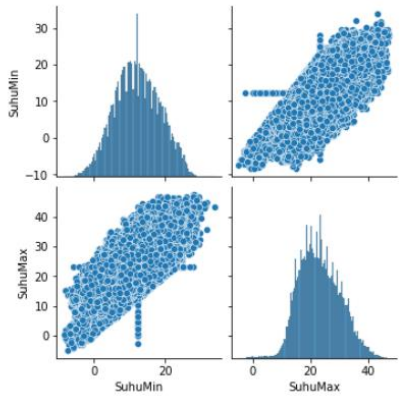



```
: <AxesSubplot:xlabel='BersaljuBesok', ylabel='count'>
```

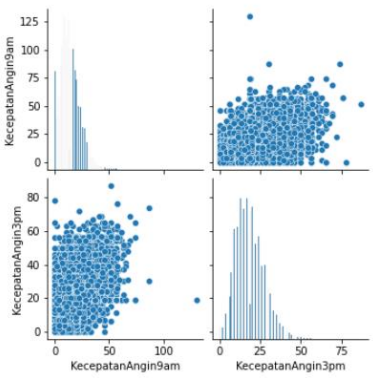


Selanjutnya yang dibutuhkan yaitu melihat perbandingan korelasi dari 6 attribut menggunakan pairplot yang dipilih :

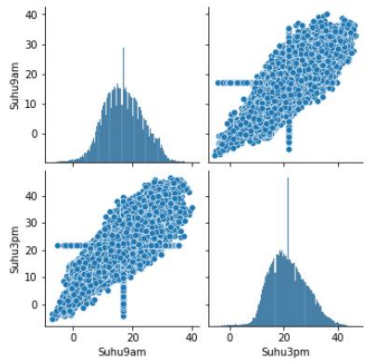
<Figure size 7200x7200 with 0 Axes>



<Figure size 7200x7200 with 0 Axes>



<Figure size 7200x7200 with 0 Axes>



Penggunaan Scaling data juga befungsi sebagai cara untuk membuat datanumerik pada dataset menjadi rentang nilai yang sama. Hasil dari scaling nya yaitu:

:

	SuhuMin	SuhuMax	KecepatanAngin9am	KecepatanAngin3pm	Suhu9am	Suhu3pm
0	0.445755	0.389635	0.000000	0.149425	0.428270	0.391555
1	0.412736	0.418426	0.100000	0.229885	0.402954	0.401152
2	0.629717	0.706334	0.115385	0.298851	0.654008	0.671785
3	0.372642	0.562380	0.100000	0.218391	0.474684	0.548944
4	0.339623	0.481766	0.169231	0.218391	0.413502	0.451056

3. Pemodelan

K-Means Clustering adalah suatu metode penganalisa data atau metode data yang mengelompokkan object dan pencilan object dengan sangat cepat. Dengan pemilihan nilai K yang diinputkan atau mencari nilai K. pada tugas ini K yang

digunakan yaitu bernilai K = [2, 3, 4, 5, 7] yang dapat diprediksi untuk memberikan nilai K yang optimal.

Fungsi yang digunakan yaitu :

```
def euclidian(u, v):
    return sum((p-q)**2 for p, q in zip(u, v))**0.5
```

4. **Evaluasi**

   Pada evaluasi ini yang pakai yaitu menggunakan Silhouette Score. Silhouette Score ini dihitung menggunakan rumus :

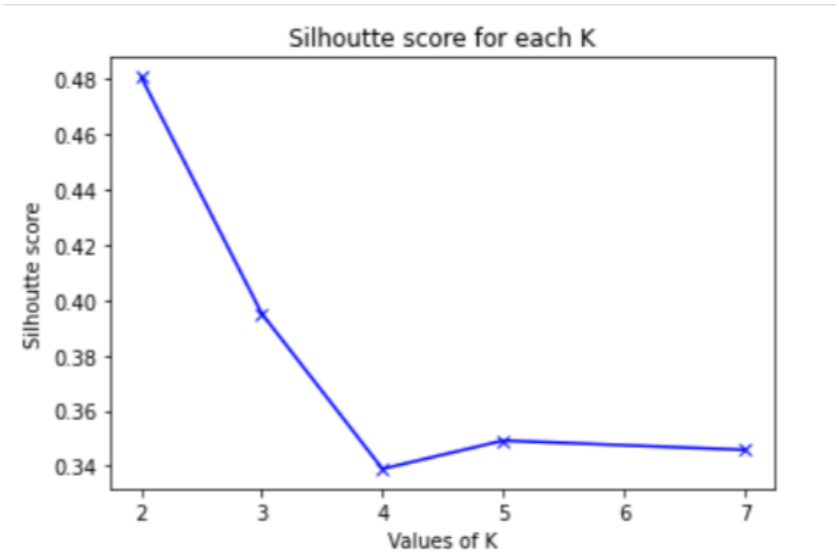
$$S(i) = b(i) – a(i) / \max(a(i), b(i))$$

   a = jarak rata-rata intra - cluster
   b = jarak rata-rata cluster terdekat
   Dengan menggunakan Sillhouette Score maka dapat diketahui seberapa optimalnya setiap tiitk yang dimasukkan. Lalu mencari nilai K terbaik yang dapat digunakan untuk performasi model serta dapat ditingkatkan menggunakan cara eksperimen K-Means yang berulang-ulang dengan nilai K yang berbeda serta hitung nilai Silhouette Scorenya.

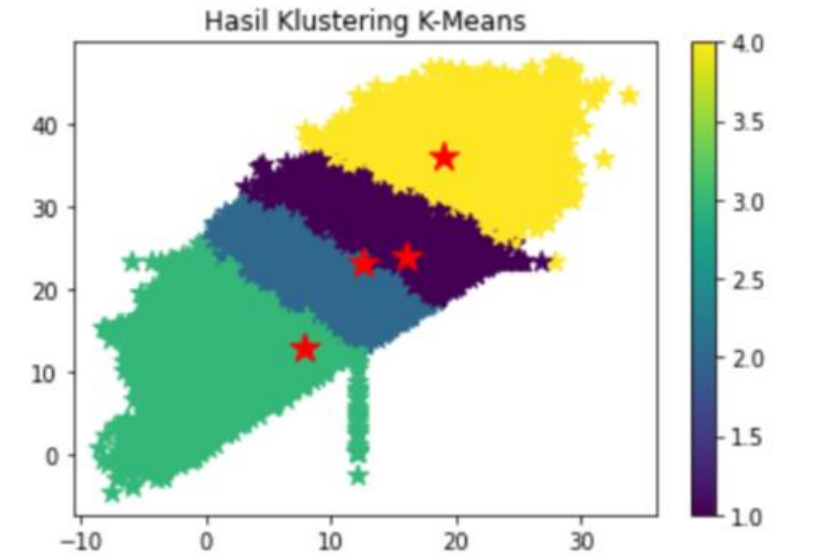
5. **Eksperimen**

   Attribut yang dipilih untuk melakukan eksperimen yaitu Suhu Min, Suhu Max, Kecepatan Angin 9am, Kecepatan Angin 3pm, Suhu 9am , Suhu 9pm. Dilakukan sebanyak 4 kali eksperimen untuk melihat nilai optimal K yang baik. Atribut yang digunakan yaitu :

   - Eksperimen 1 : Suhu Min dan Suhu Max
     Nilai K = 4 , dengan silhoutte score tertinggi yaitu 0.48

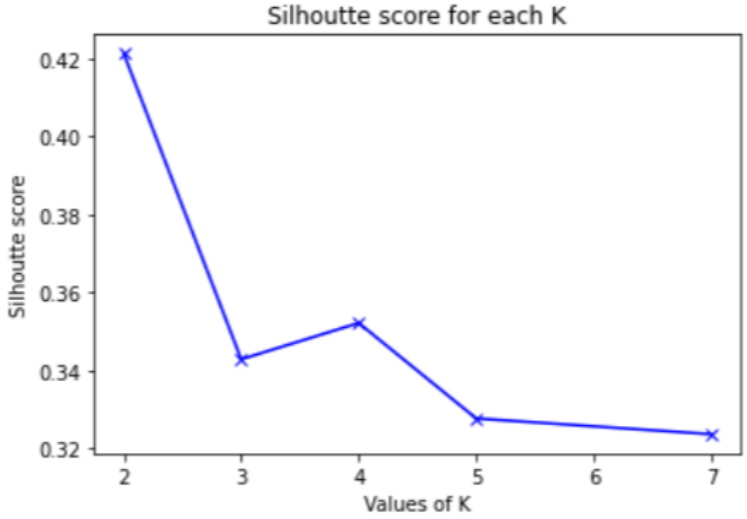


Visualisasi dengan nilai K = 4

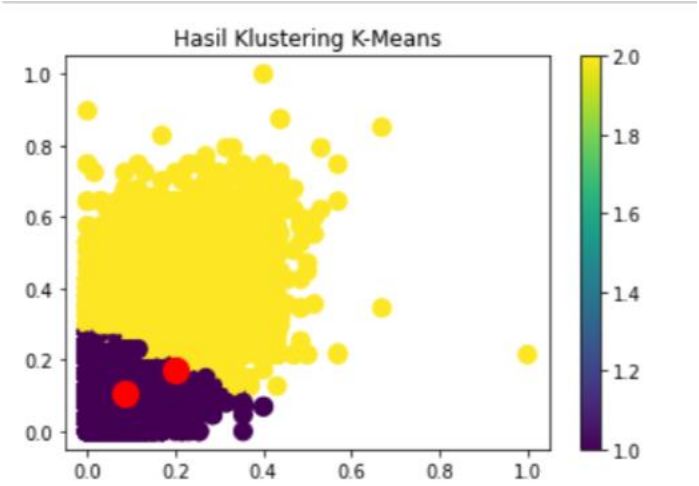


Shilhoutte Score : 0.37176850222109276

- Eksperimen 2 : Kecepatan Angin 3pm dan Kecepatan Angin 3am
  Nilai K : 2, , dengan silhoutte score tertinggi yaitu 0.43



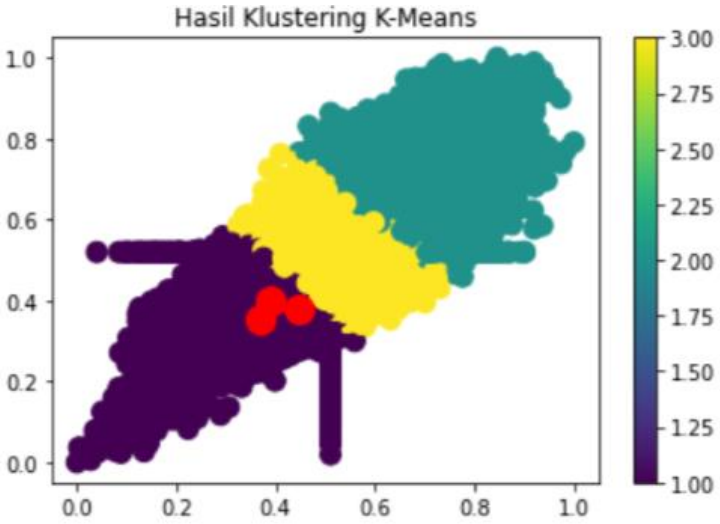
Visualisasi dengan nilai K = 2



Silhoutte Score : 0.4360922334365307

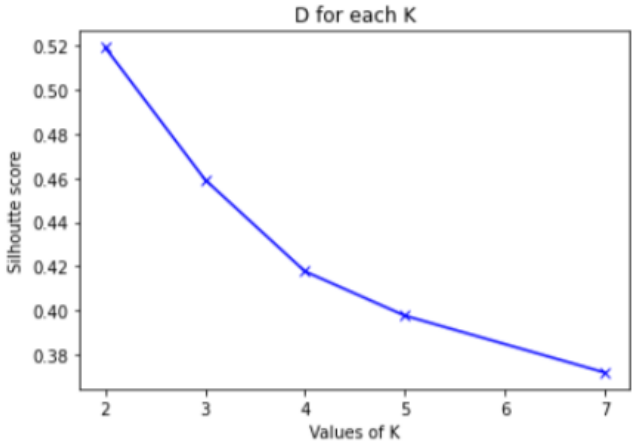
- Eksperimen 3 : Suhu 9am dan Suhu 3pm
  Nilai K : 3, , dengan silhoutte score tertinggi yaitu 0.502

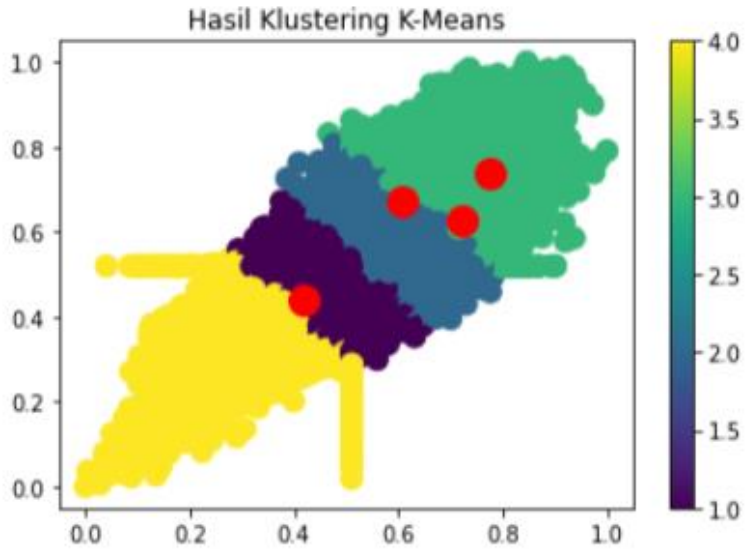
Visualisasi dengan nilai K = 3



Silhoutter Score : 0.4400815882977956

- Eksperimen 4 : Suhu 9am dan SuhuMin
  Nilai K : 4





Silhoutte Score : 0.39250945585456787

6. Kesimpulan

Hasil dari eksperimen diatas menyatakan bahwa pemilihan pengujian menggunakan silhoutter score ini belom efektif karena nilai yang didapatan dari setiap pengujian nilai K tidak ada nilai yang mendekati nilai 1.

Untuk kedepannya disarankan untuk melakukan penyebaran data terlebih dahulu. K-Means menjadi tidak optimal karena data yang tidak teratur sehingga nilai dari silhoutte score pun menjadi tidak bagus.