**LAPORAN TUGAS BESAR MACHINE LEARNING**

NAMA : Annisa Miranda

NIM : 1301184378

KELAS : IF 42-04

LINK PRESENTASI : <https://youtu.be/Xa2SFQ1bAfs>

**RINGKASAN**

Pada tugas besar machine learning ini saya menggunakan dataset ‘salju’. pada sataset terdapat 109095 data dan 24 atribut. Selanjutnya dilakukan clustering terhadap dataset tersebut, untuk clustering ini saya menggunakan 6 atribut yang menurut saya berpotensi untuk menentukan turun atau tidaknya salju keesokan hari, yaitu hujan, sinar matahari, kecepatan angin 3 pm, kelembaban 3pm, awan 3 pm, dan suhu 3pm.

Saya melakukan percobaan algoritma K-Means untuk melakukan clustering terhadap dataset. Alasan saya menggunakan algoritma K-Means karena mudah diimplementasikan serta memiliki waktu komputasi yang cepat tergantung dari nilai K yang diambil, semakin kecil nilai K, makan semakin cepat juga waktu komputasi yang diperlukan. Percobaan yang saya lakukan antara lain:

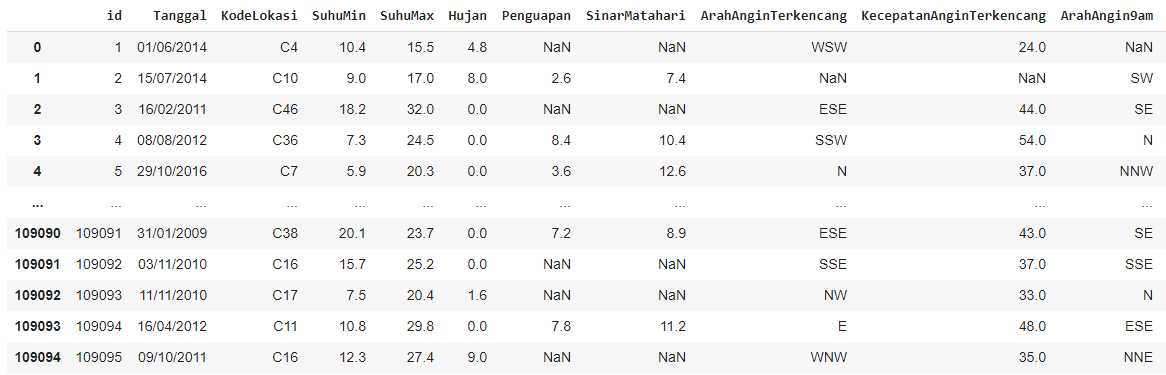
1. K=2 dengan 2 atribut
2. K=2 dengan 3 atribut
3. K=3 dengan 2 atribut
4. K=3 dengan 3 atribut

Saya juga melakukan beberapa teknik eksplorasi dan persiapan data, untuk eksplorasi data saya menggunakan boxplot untuk melihat persebaran data dan membantu dalam melihat ada atau tidaknya pencilan/outlier. Selain itu saya juga menggunakan histogram untuk melihat distribusi data, serta heatmap untuk melihat korelasi pada setiap atribut.

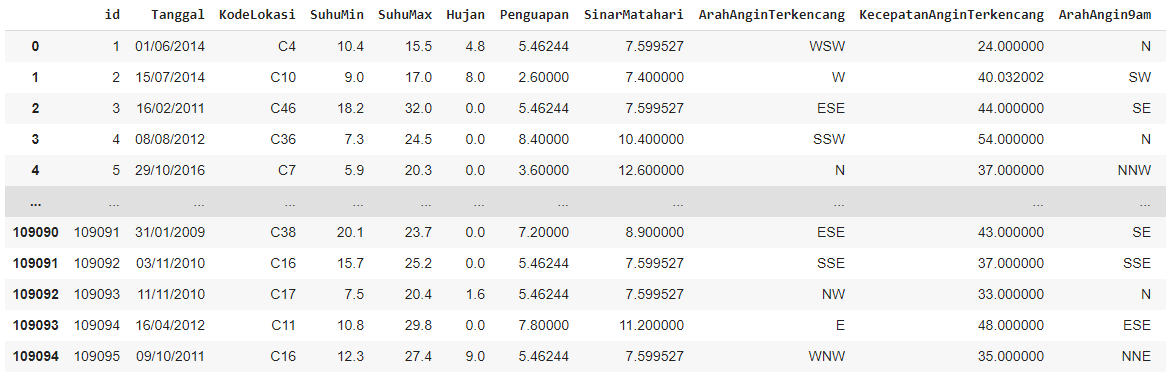
Pada persiapan data, saya melakukan input pada missing value atau atribut dengan nilai null menggunakan nilai mean, serta scalling pada atribut yang sudah saya pilih tadi. Kesimpulan yang diambil berdasarkan dari nilai silhouette score dari hasil clustering. Silhouette score ini dapat mengukur kualitas dari hasil clustering.

**TAHAPAN PROGRAM**

1. Read data ‘salju\_train.csv’, karena beberapa atribut mengalami missing value, maka dilakukan input pada missing value. Pada atribut bertipe numeric, diisi menggunakan nilai mean dari atribut yang berkaitan, serta untuk atribut bertipe string, diisi menggunakan data yang paling sering muncul pada atribut tersebut.



*Figure 1 nilai setiap atribut sebelum diinput*



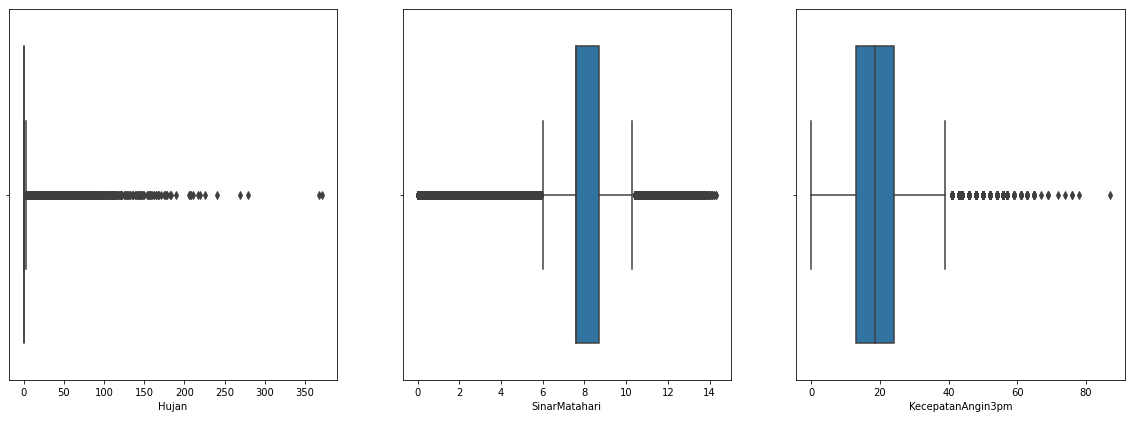
*Figure 2 nilai setiap atribut setelah diinput*

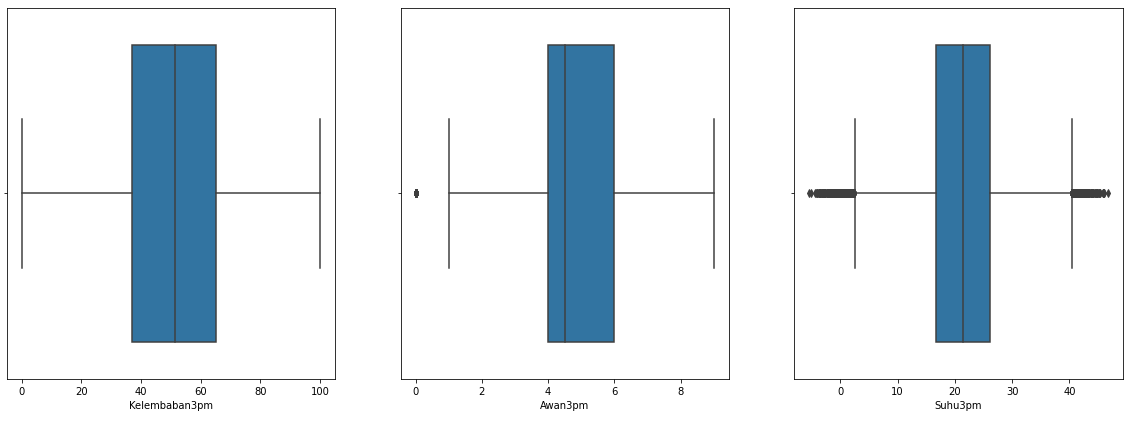
1. Memilih atribut yang berpotensi untuk mengetahui turun atau tidaknya salju keesokan hari ketika melakukan clustering. Atribut yang digunakan yaitu hujan, sinar matahari, kecepatan angin 3pm, kelembaban 3 pm, awan 3 pm, dan suhu 3 pm.



*Figure 3 tabel atribut yang dipilih*

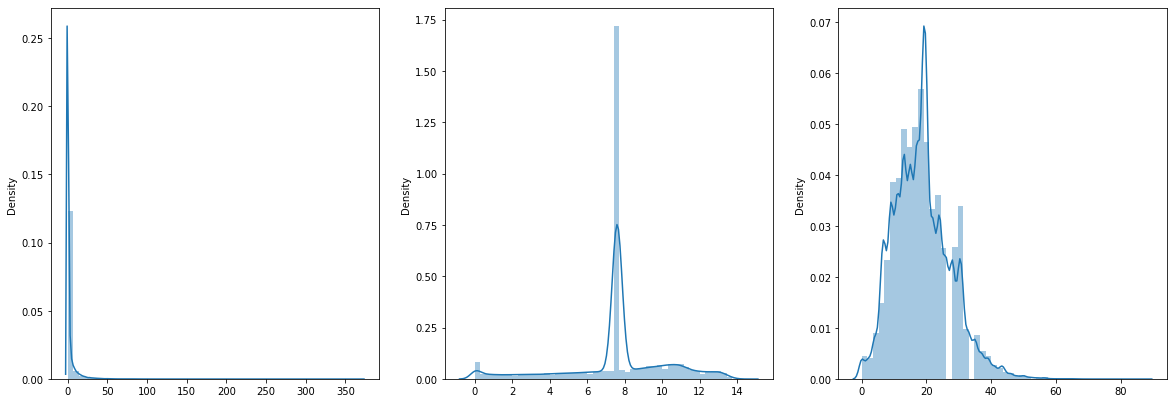
1. Menemukan outlier/pencilan pada setiap atribut yang telah dipilih.

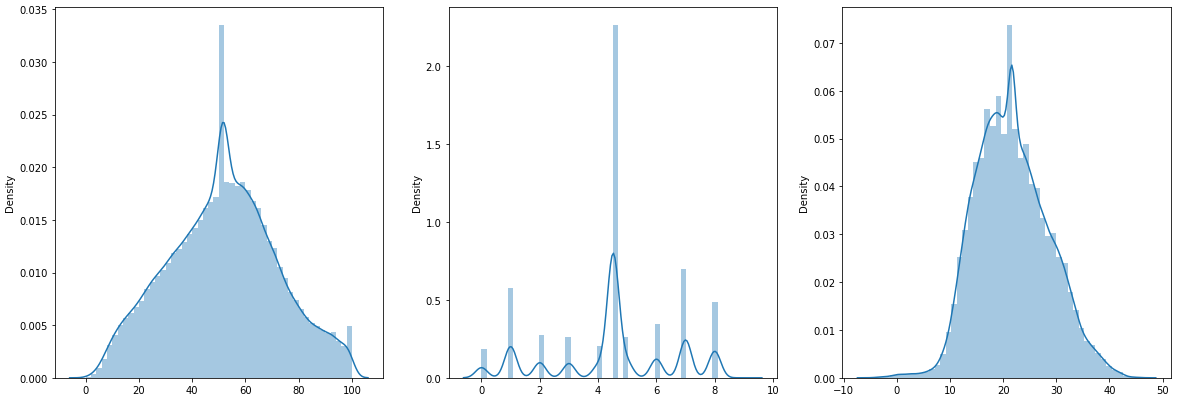




*Figure 4 boxplot untuk melihat outlier*

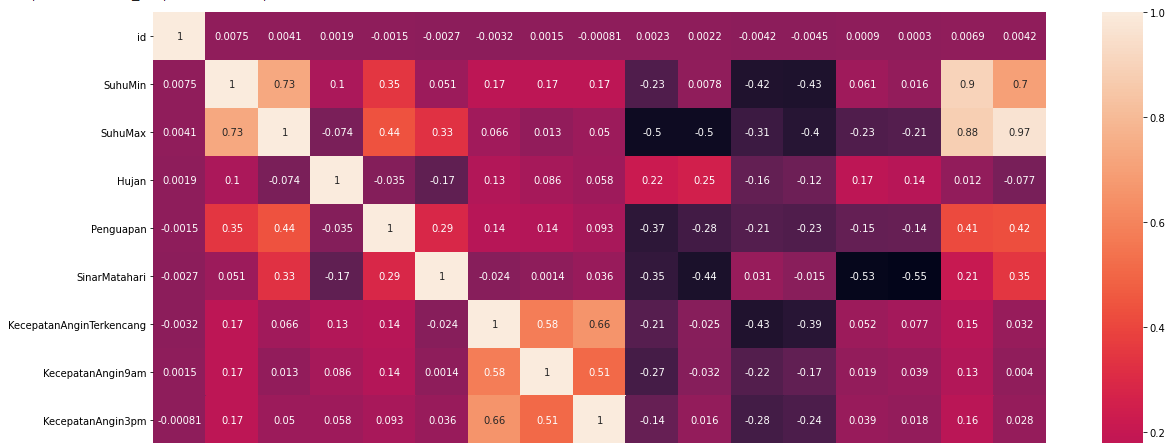
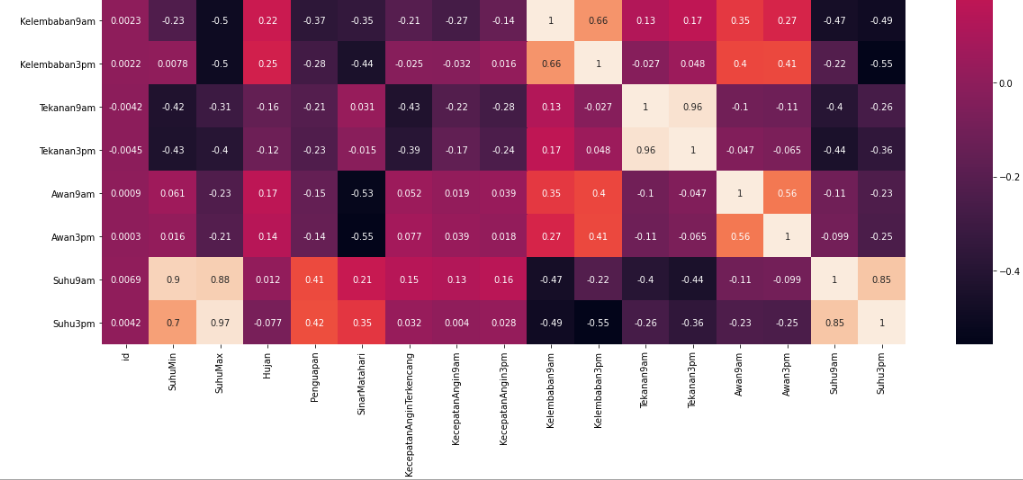
1. Menampilkan distibusi/persebaran data dari atribut yang dipilih menggunakan histogram.





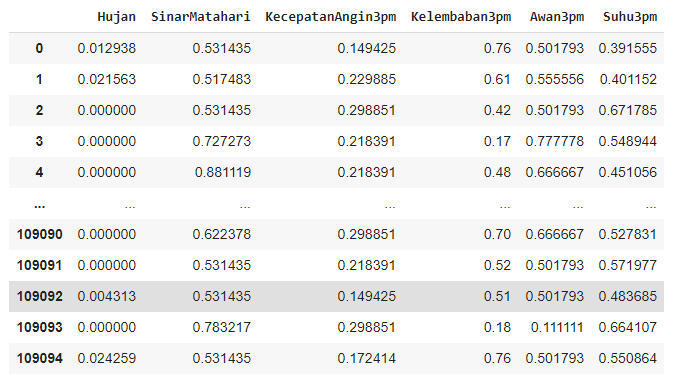
*Figure 5 histogram distribusi atribut yang dipilih*

1. Menampilkan korelasi dari setiap atribut menggunakan heatmap.



*Figure 6 heatmap untuk menampilkan korelasi antar atribut*

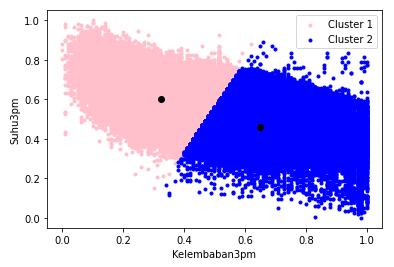
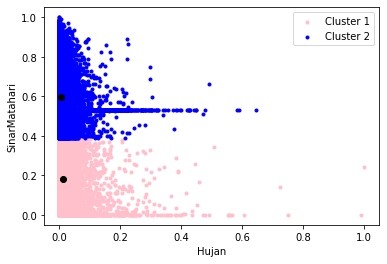
1. Melakukan scalling terhadap atribut yang sudah dipilih agar memiliki rentang nilai yang sama, agar menghasilkan nilai Euclidean distance yang bagus untuk clustering.



*Figure 7 hasil scalling dari atribut yang dipilih*

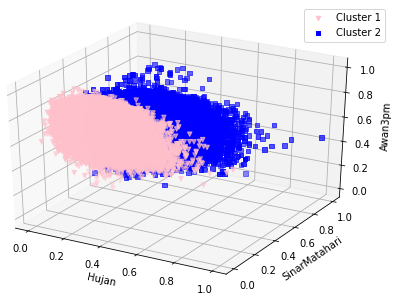
1. Clustering

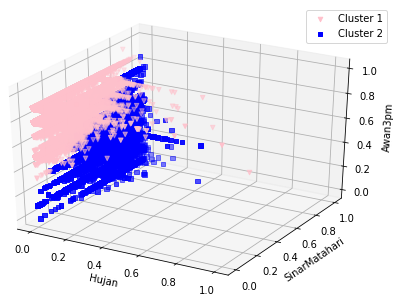
* Percobaan menggunakan 2 atribut yaitu Hujan-SinarMatahari dan Kelembaban3pm-Suhu3pm dengan nilai K = 2. Hasil yang didapatkan adalah sebagai berikut.

*Figure 8 hasil clustering menggunakan 2 atribut dan K=2*

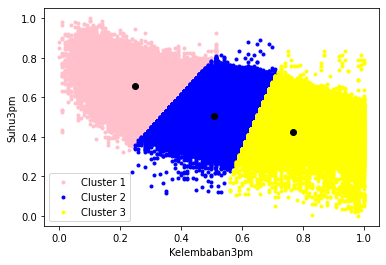
* Percobaan menggunakan 3 atribut yaitu Hujan-SinarMatahari-Awan3pm dan –KecepatanAngin3pm-Kelembaban3pm-Suhu3pm dengan nilai K = 2. Hasil yang didapatkan adalah sebagai berikut.

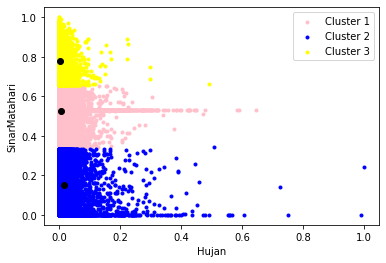




*Figure 9 hasil clustering menggunakan 3 atribut dan K=2*

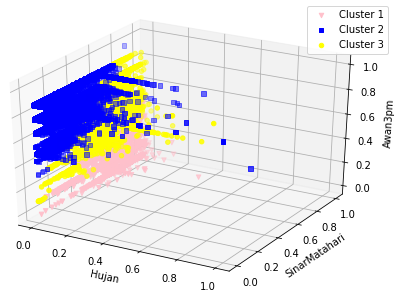
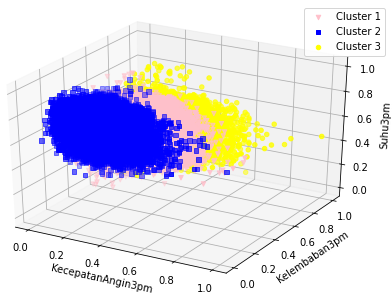
* Percobaan menggunakan 2 atribut yaitu Hujan-SinarMatahari dan Kelembaban3pm-Suhu3pm dengan nilai K = 3. Hasil yang didapatkan adalah sebagai berikut.





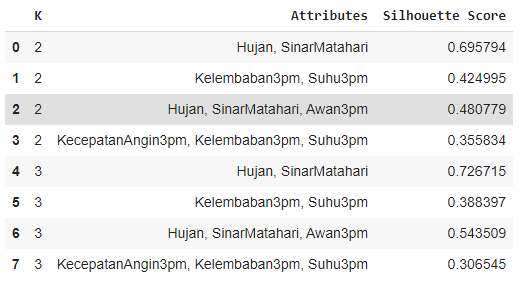
*Figure 10 hasil clustering menggunkan 2 atribut dan K=3*

* Percobaan menggunakan 3 atribut yaitu Hujan-SinarMatahari-Awan3pm dan –KecepatanAngin3pm-Kelembaban3pm-Suhu3pm dengan nilai K = 3. Hasil yang didapatkan adalah sebagai berikut.



*Figure 11 hasil clustering menggunakan 3 atribut dan K=3*

1. Evaluasi hasil clustering berdasarkan silhouette score.



*Figure 12 tabel perbandingan silhouette score*

**KESIMPULAN**

Dari hasil percobaan dapat disimpulkan bahwa silhouette score tertinggi dimiliki oleh hasil clustering atribut Hujan dan SinarMatahari dengann nilai K=3. Menurut saya nilai K yang digunakan dalam percobaan serta atribut yang dipilih sangat berpengaruh dalam proses clustering, alangkah lebih baik jika sebelum melakukan clustering kita menentukan terlebih dahulu nilai K yang optimal untuk percobaan yang akan dilakukan, serta lebih memprhatikan atribut yang berpotensi untuk dipilih.