**Tugas Besar Tahap Pertama Machine Learning**

**Clustering**

oleh:

Muhammad Hafidh Raditya (NIM 1301184079)

IF-42-03

****

**Program Studi S1 Informatika  
Fakultas Informatika  
Universitas Telkom  
Bandung  
2021**

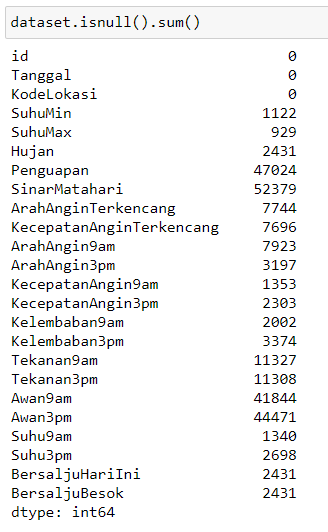
1. **Formulasi Masalah**

Pada tugas besar kali ini, saya mendapatkan dataset nomor 2 yang berisi tentang berbagai macam atribut yang memengaruhi apakah pada suatu hari akan turun salju atau tidak.

Tugas saya adalah untuk mendesain model clustering dari dataset tersebut, kemudian membuat prediksi jika nanti diberikan satu record baru dengan jenis atribut yang sama, apakah hari tersebut akan turun salju atau tidak.

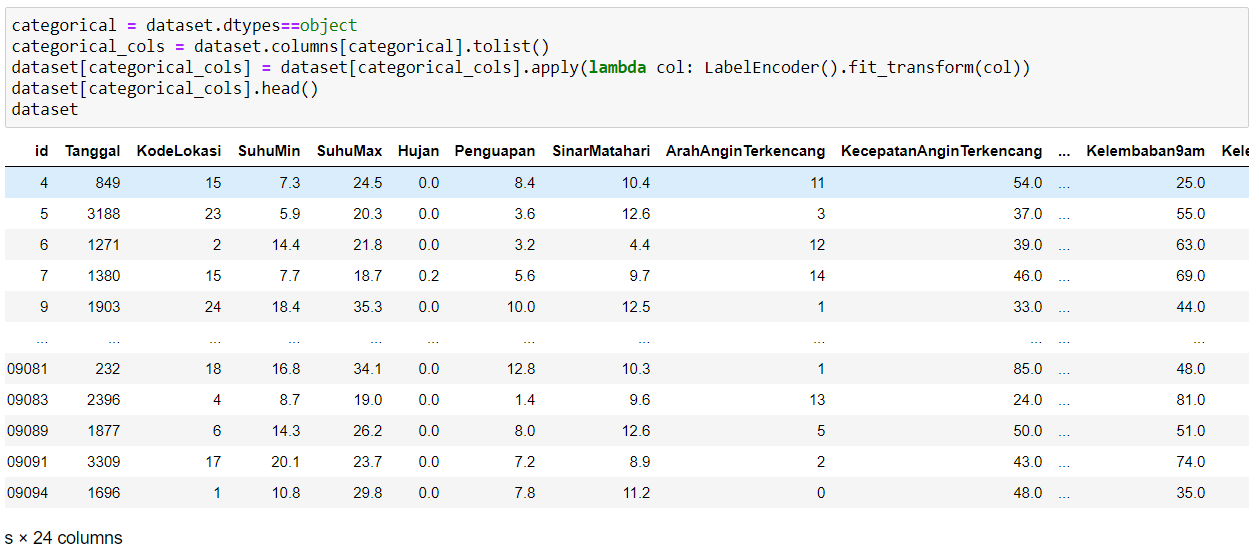
1. **Data Pre-Processing**

Pertama-tama yang saya lakukan adalah mengecek pada dataset, apakah terdapat *record* yang memiliki attribute yang bernilai *null* atau tidak. Jika ada, maka *drop record* tersebut.

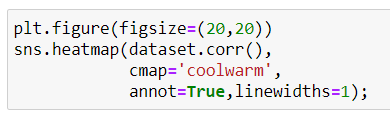




Setelah dataset bersih dari *record* yang bernilai *null*, saya akan mengecek korelasi atau keterkaitan antar attribut dengan menggunakan *heatmap*. Namun dikarenakan dataset tersebut memiliki beberapa atribut kategorikal, maka saya harus mengkonversikan atribut-atribut kategorikal tersebut menjadi atribut numerikal agar dapat dibanding nilainya dan dicek keterkaitannya menggunakan fungsi *corr()* pada bahasa Python.



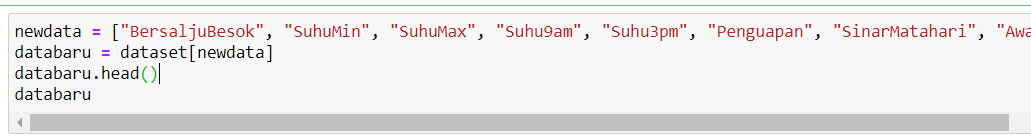
Setelah selesai proses konversi, barulah saya bisa membandingkan keterkaitan antar atribut menggunakan *heatmap*.

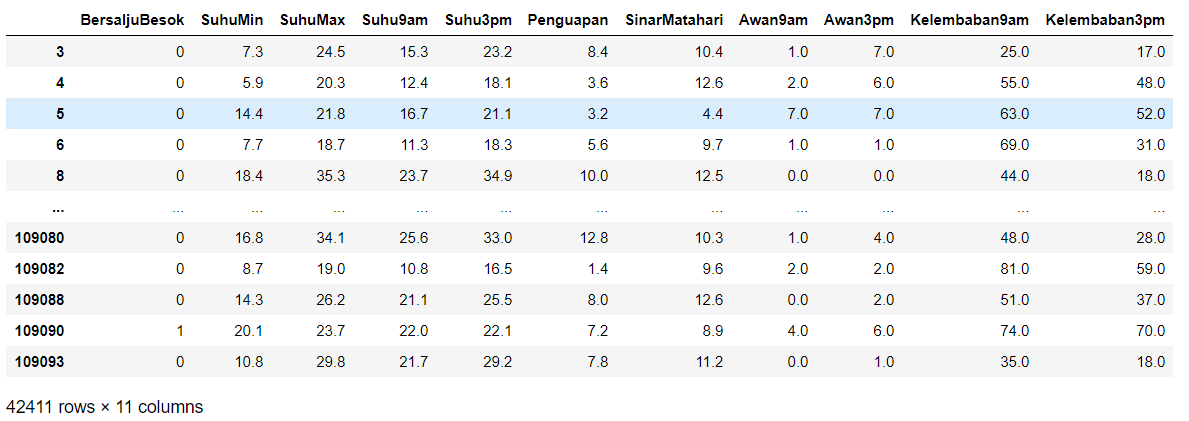


Chart, application, timeline, treemap chart

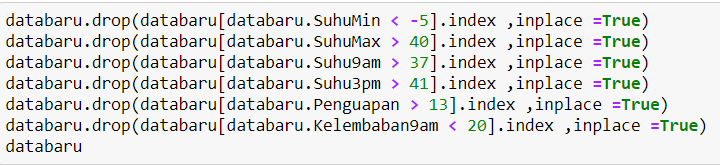
Description automatically generated

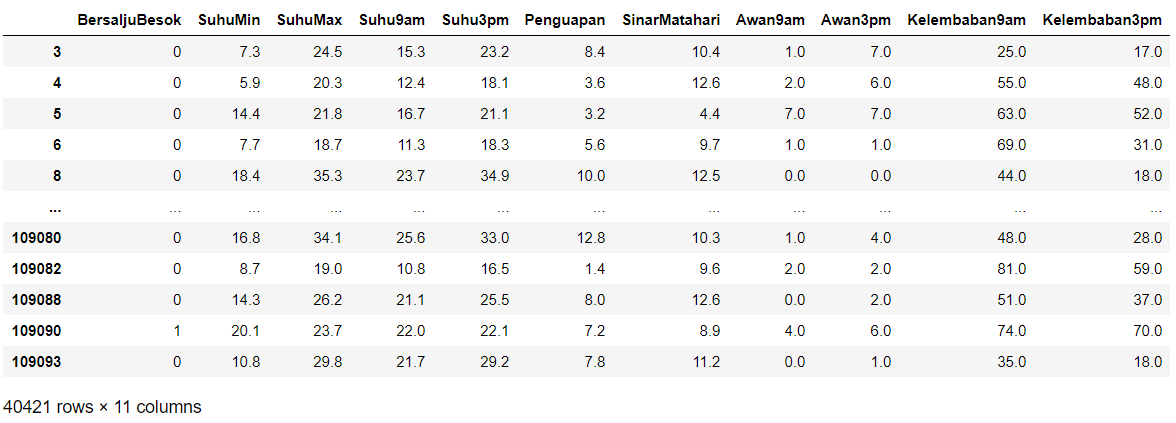
Setelah saya menganalisis hasil dari *heatmap*, saya menyeleksi beberapa atribut yang memiliki nilai korelasi yang tergolong besar yang akan saya pakai selanjutnya. Atribut-atribut tersebut adalah BersaljuBesok, SuhuMin, SuhuMax, Suhu9am, Suhu3pm, Penguapan, SinarMatahari, Awan9am, Awan3pm, Kelembaban9am, dan Kelembaban3pm.



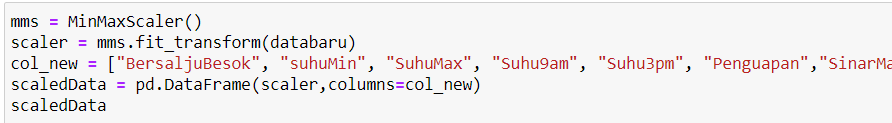


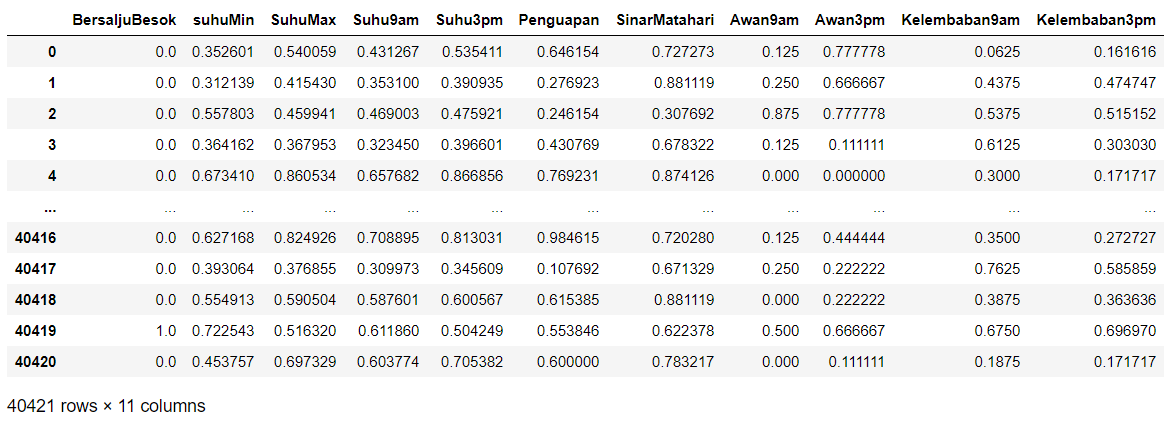
Setelah itu, saya akan mengecek apakah pada setiap atribut memiliki *record* data yang tergolong sebagai *outlier* atau tidak menggunakan boxplot. Setelah saya cek, ternyata ada beberapa atribut yang memiliki outlier. Atribut-atribut tersebut adalah SuhuMin, SuhuMax, Suhu9am, Suhu3pm, Penguapan, dan Kelembaban9am. Maka saya akan *drop* *record* data yang tergolong sebagai *outlier* tersebut.





Setelah dataset bersih dari *null value* dan *outlier*, yang selanjutnya saya lakukan adalah menyeragamkan *scaling* pada setiap atribut. Hal ini dilakukan agar setiap atribut memiliki skala minimum dan maksimum yang sama yang nilainya berkisar antara 0 sampai 1. Tentunya proses ini juga memudahkan nanti dalam proses visualisasi clustering menggunakan *plot*.





Pada tahap ini, saya membuat dua macam dataset untuk disimpan, untuk kemudian dibandingkan pada tahap analisis dan kesimpulan. Dataset yang pertama adalah dataset yang sudah dilakukan proses *scaling*. Dataset yang kedua adalah dataset yang belum dilakukan proses *scaling.*

Lalu yang saya lakukan adalah mengecek *heatmap* korelasi dataset yang sudah bersih ini. Hal ini saya lakukan untuk mencari atribut mana yang paling berpengaruh terhadap *value* dari atribut “BersaljuBesok”. Dikarenakan tujuan dari tugas ini adalah untuk memprediksi apakah besok akan turun salju atau tidak, maka saya memilih atribut “BersaljuBesok” sebagai patokan.

Chart, treemap chart

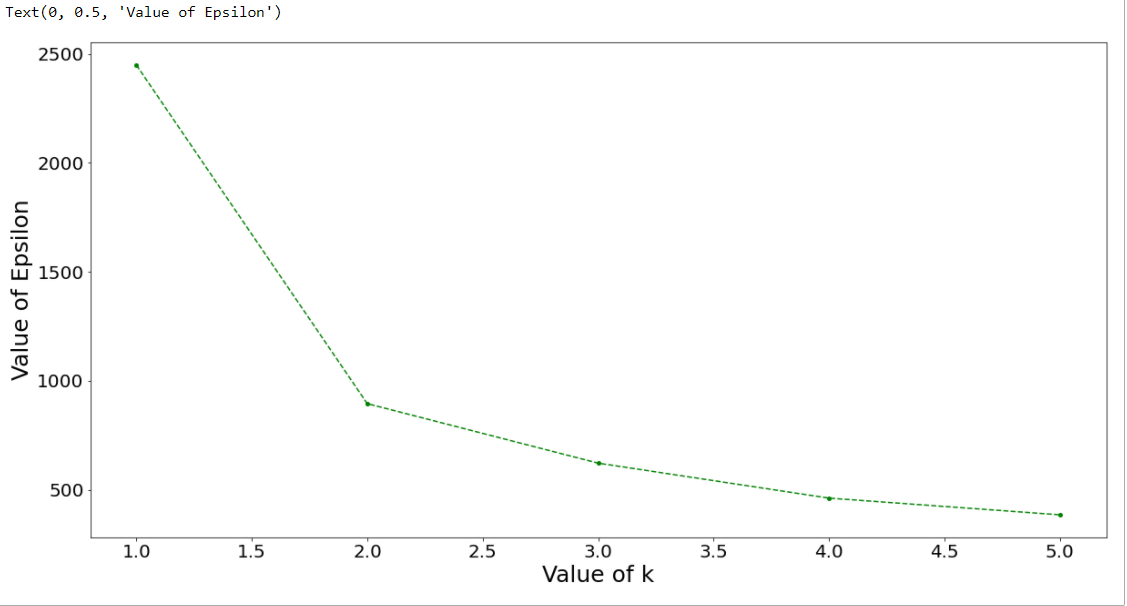
Description automatically generated

Setelah saya analisis, saya mendapatkan kalau atribut Awan3pm dan Kelembaban3pm merupakan atribut yang paling berpengaruh terhadap atribut BersaljuBesok. Maka saya akan menyeleksi kedua atribut tersebut untuk dilakukan proses *clustering*

1. **Proses Clustering**

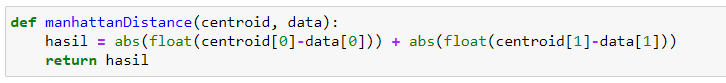
Sebelum saya melakukan proses *clustering* menggunakan metode *KMeans*, saya melakukan proses evaluasi terlebih dahulu menggunakan metode *Sum Square Error* (SSE) dan metode *Elbow* untuk mencari berapa nilai k atau jumlah *cluster* yang optimal untuk digunakan pada dataset ini.



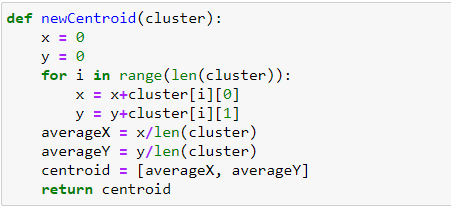


Setelah saya analisis grafik *elbow* yang telah saya buat, pada titik k=2 sudah terbentuk pola siku atau *elbow*. Sehingga dapat disimpulkan kalau jumlah *cluster* yang optimal adalah 2 *cluster*.

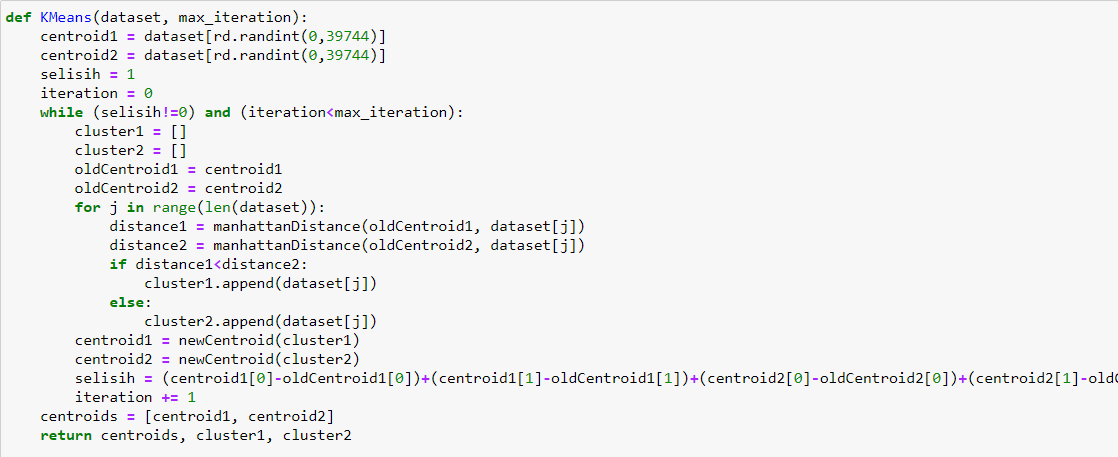
Selanjutnya adalah langkah untuk proses *clustering*. Ada 3 buah fungsi yang saya siapkan untuk mempermudah proses ini.



Yang pertama adalah fungsi manhattanDistance untuk mencari nilai *distance* atau jarak antara dua buah titik.

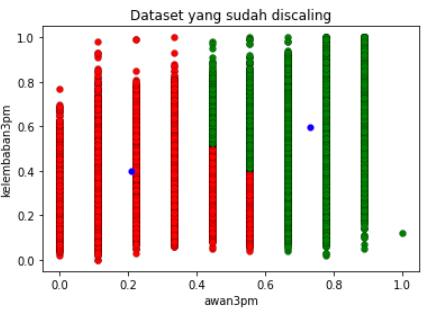


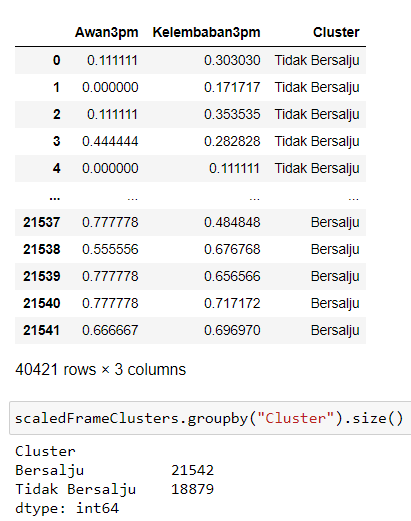
Yang kedua adalah fungsi newCentroid. Fungsi ini digunakan untuk meng-*generate* *centroid* baru pada saat tiap kali iterasi pada proses *clustering*.



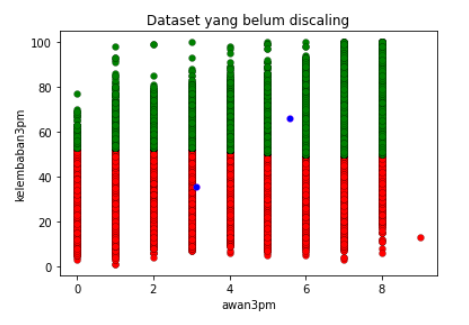
Dan yang terakhir adalah fungsi Kmeans yaitu fungsi utama dari proses *KMeans clustering*. Jika dijelaskan menggunakan bahasa natural, maka algoritma KMeans memiliki langkah-langkah sebagai berikut:

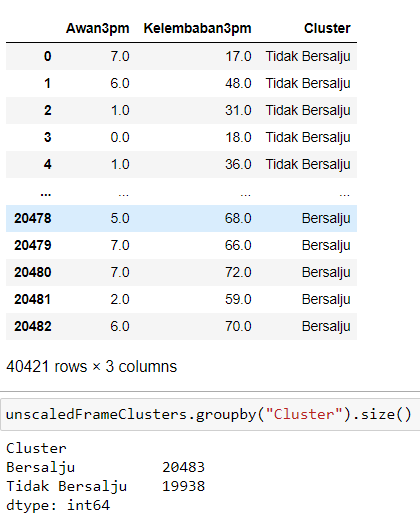
1. Buat 2 centroid baru secara random.
2. Hitung jarak antara setiap data terhadap kedua centroid yang sudah dibuat.
3. Jika jarak terhadap centroid 1 lebih kecil daripada jarak terhadap centroid 2, maka masukkan data tersebut ke dalam cluster 1. Begitu juga sebaliknya.
4. Buat 2 buah centroid baru dari nilai rata-rata pada masing-masing cluster.
5. Ulangi langkah 2-4 sampai menemui kondisi dimana centroid pada iterasi ke-n sama dengan centroid pada iterasi ke n+1, atau pada saat jumlah iterasi menyentuh nilai max\_iteration yang sudah ditentukan.
6. **Analisis**





Sudah discaling





Belum discaling

Setelah saya lakukan proses labeling pada cluster 1 dan cluster 2 pada kedua jenis dataset tersebut, secara garis besar, data pada cluster 1 dan cluster 2 antara kedua dataset yang sudah dilakukan scaling dan belum dilakukan scaling memiliki jumlah yang tidak begitu berbeda jauh, namun memiliki persebaran data pada cluster yang sangat berbeda. Namun jika dilihat visualisasi atau peta persebaran clusternya, asumsi saya adalah cluster yang dibuat menggunakan dataset yang sudah dilakukan scaling lebih baik daripada cluster yang dibuat menggunakan dataset yang belum dilakukan scaling. Alasannya adalah jika dilihat pada visualisasi datanya, pada dataset yang sudah dilakukan scaling, terdapat peralihan antara cluster 1 dan cluster 2 yang cukup jelas. Dapat dilihat pada saat interval atribut awan3pm berkisar antara 0.4-0.6, terdapat dua jenis cluster pada interval tersebut yang menunjukkan bahwa terdapat peralihan yang jelas antara cluster 1 dan cluster 2. Namun analisis ini masih berupa hipotesis saya saja, dikarenakan untuk melihat bagus atau tidaknya proses clustering yang sudah dibuat, harus dilakukan proses klasifikasi untuk mengecek akurasi pada proses clustering yang sudah dibuat, yang artinya saya harus melakukan proses klasifikasi pada tubes tahap 2 terlebih dahulu untuk mengetahui jenis dataset mana, yang sudah discaling atau belum discaling, yang lebih baik akurasinya.

1. **Kesimpulan**

Kesimpulan yang saya dapat dari tugas kali ini adalah, ternyata jika hasil clustering pada dua jenis dataset yang sudah discaling dan belum discaling memiliki persebaran data yang sangat berbeda, terbukti pada saat dilihat dari visualisasi datanya. Namun untuk mengetahui jenis dataset mana yang lebih baik, harus dilakukan proses klasifikasi terlebih dahulu untuk mengetahui akurasi clusteringnya.