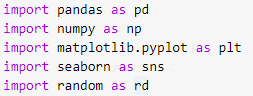
Laporan TUBES 1 Machine Learning

1. Library

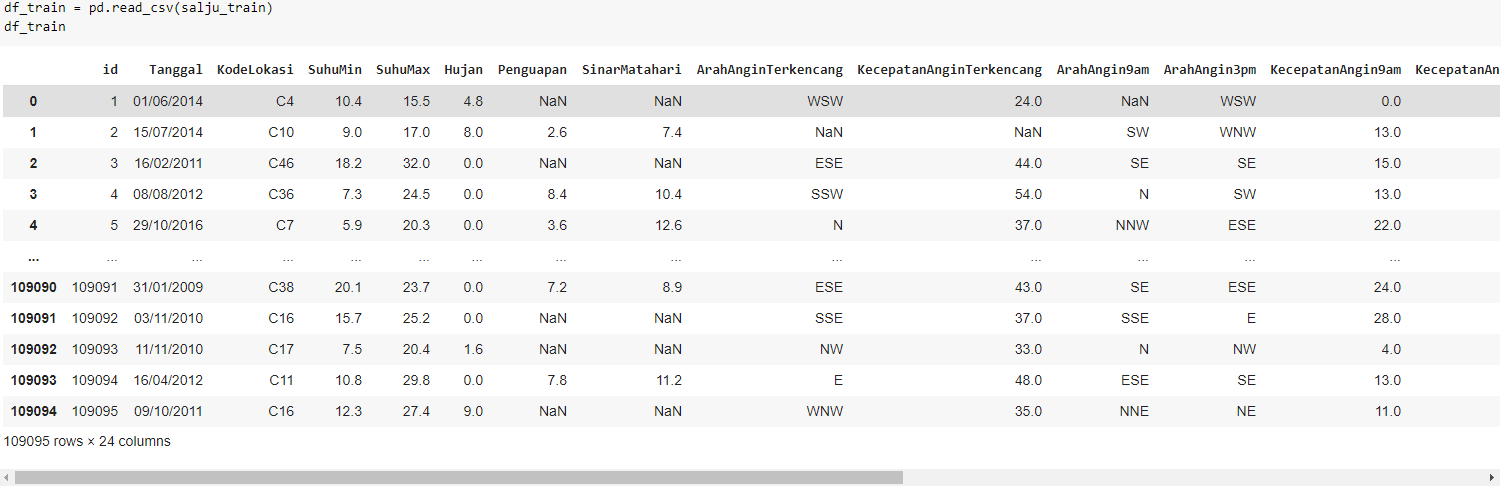


Library yang saya gunakan untuk tubes ini adalah pandas untuk membuat tabel dan mengecek data dan sebagainya, lalu numpy untuk perhitungan, lalu matplotlib untuk membuat plot, lalu seaborn untuk membuat grafik, dan random untuk meng-generate angka random.

2. Dataset

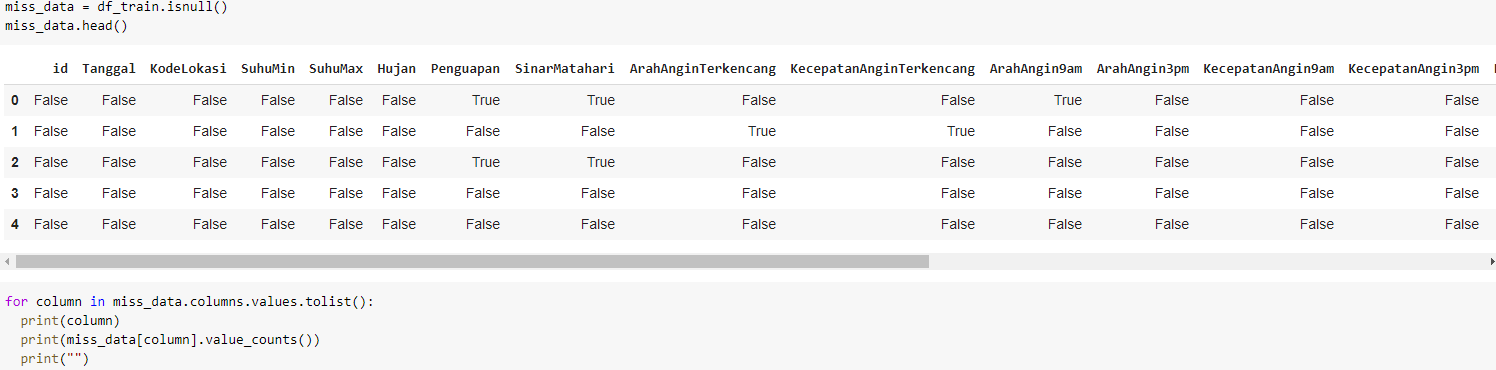


Dataset yang saya gunakan adalah salju\_train, karena memiliki jumlah row yang lebih banyak

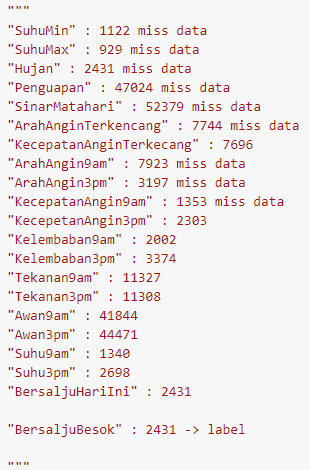


3. Preprocessing data

a. Mengecek nilai yang null

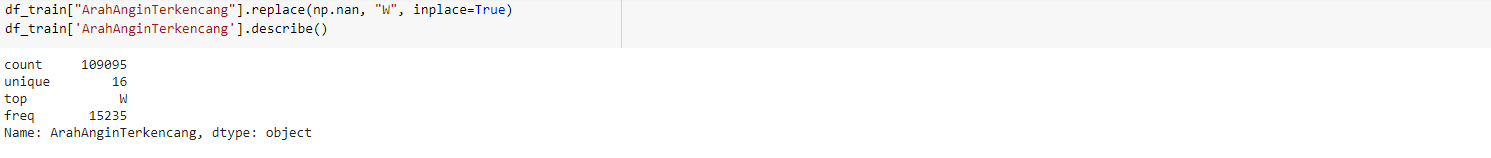


Saya membuat dataframe dengan nama *miss\_data* untuk mengetahui kolom yang memiliki nilai NaN. Kemudian saya menampilkan isi setiap kolom dari *miss\_data* dan dihasilkan jumlah nilai NaN setiap kolom.

Sehingga dihasilkan seperti gambar disamping untuk setiap kolom yang memiliki nilai NaN.

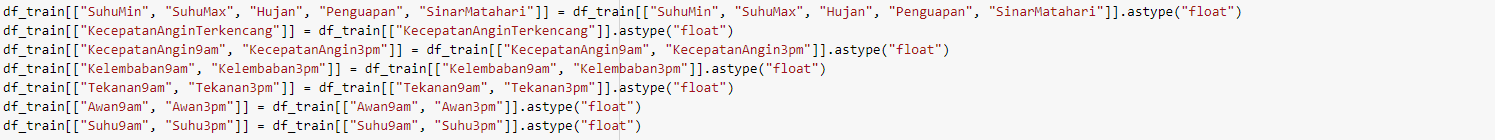
Untuk mengisi nilai NaN saya menggunakan 2 cara, yaitu mengisi setiap nilai NaN pada kolom numerik dengan rata-rata pada setiap kolom dan mengisi nilai NaN pada kolom non-numerik dengan modus dari kolom tersebut







b. Mengubah kolom object numerik menjadi float



Saya merubah kolom numerik dengan tipe data objek menjadi float agar nanti bisa di lakukan normalisasi



c. Normalisasi



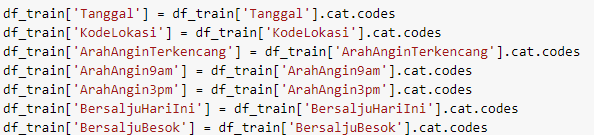
Saya hanya melakukan normalisasi pada kolom *Tekanan9am* dan *Tekanan3pm* saja, karena memiliki nilai ribuan.

Sebelum: Sesudah:



d. Mengubah tipe data objek yang tersisa menjadi category



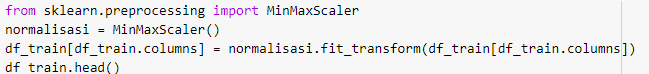


Sisa kolom dengan tipe data objek saya ubah menjadi category kemudian diubah menjadi category berupa numerik agar bisa dilakukan scaling.

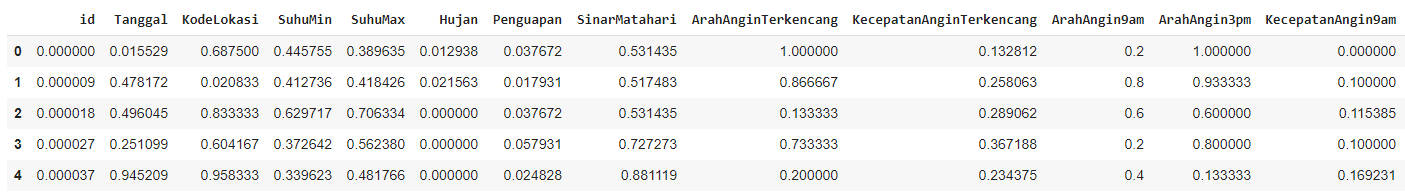




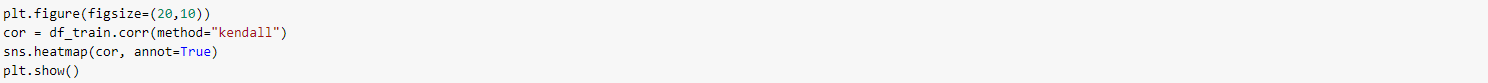
e. Scaling



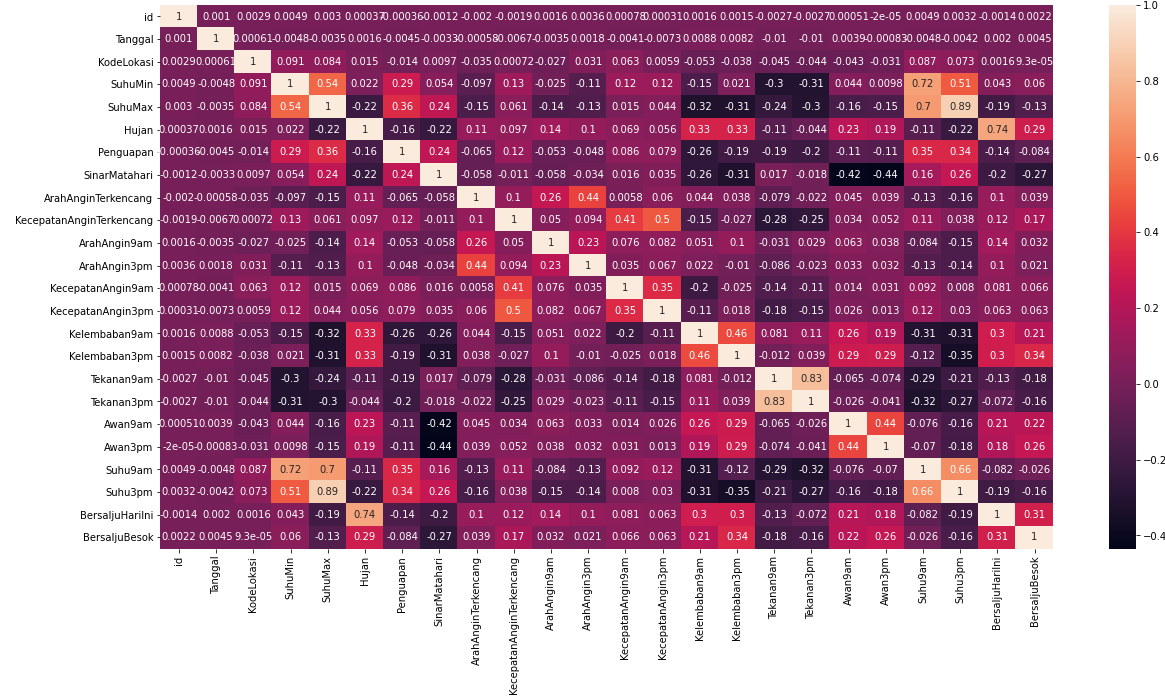
Disini saya menggunakan library preprocessing untuk melakukan min-max-scaler untuk mengubah nilai setiap kolom menjadi rentang 0 sampai 1



4. Feature Selection



Pada feature selection saya membuat figure korelasi dengan heatmap menggunakan metode kendall.

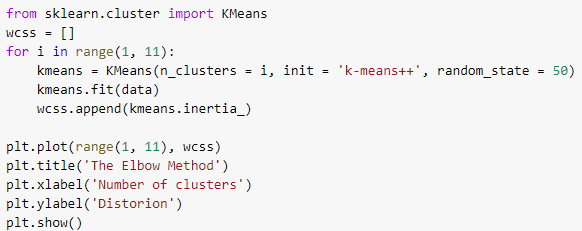


Dari figure diatas dapat didapati bahwa kolom ke-5 yaitu Hujan dan kolom ke-15 yaitu Kelembaban3pm adalah 2 kolom yang sangat berpengaruh pada label

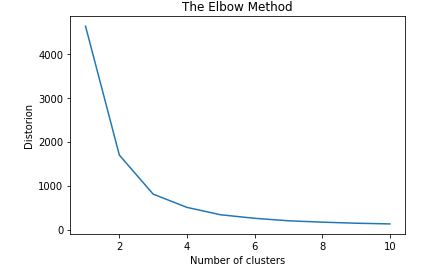
Kemudian saya memasukkan 2 kolom tersebut kedalam dataframe baru dengan nama data.



5. Penentuan nilai k



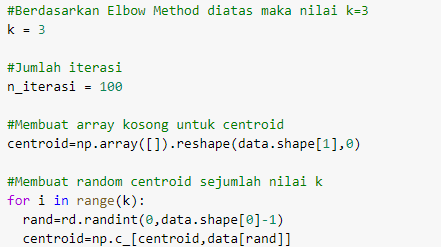
Untuk mendapatkan nilai k yang terbaik saya menggunakan elbow method.



Dari hasil yang didapatkan diatas maka dapat disimpulkan bahwa nilai k yang terbaik adalah 3.

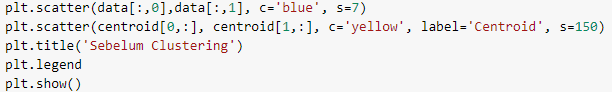
6. Clustering

a.

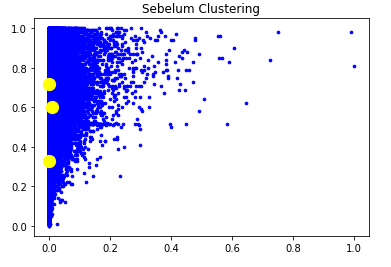


Pertama-taman saya memasukkan nilai k yaitu 3. Kemudian nilai iterasi yaitu 100, karena berdasarkan sumber yang saya pakai 100 sudah cukup untuk mewakilkan. Kemudian saya membuat centroid dengan posisi acak dengan jumlah centroid sama dengan k.

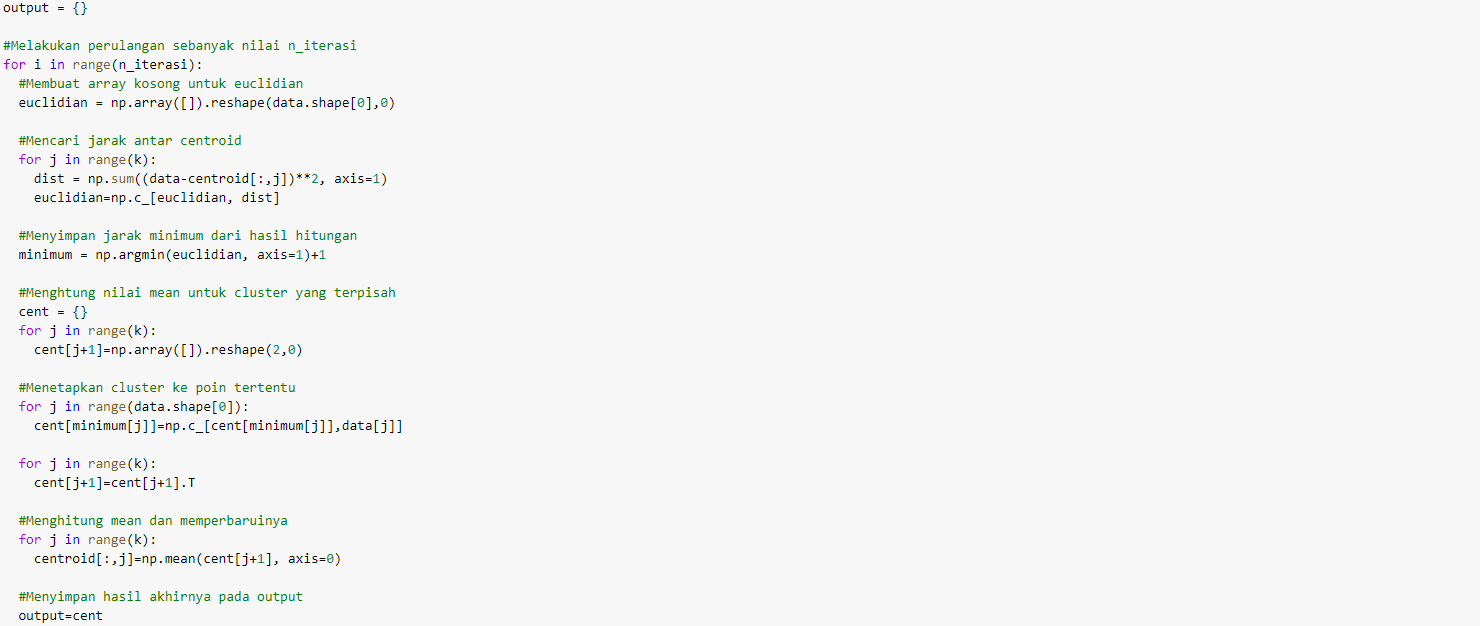
b.



Lalu saya membuat plot untuk setiap data dan centroid yang telah dibuat sebelumnya (plot belum dilakukan clustering)



c.

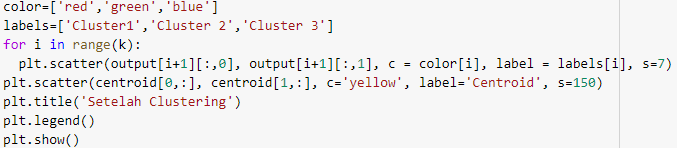


Kemudian saya mencari jarak antara poin-poin dari setiap centroid dan disimpan pada variabel euclidian. Disini saya menggunakan Euclidian distance karena merupakan metode yang paling sering digunakan. Disini juga saya menyimpan jarak minimum pada variable minimum.

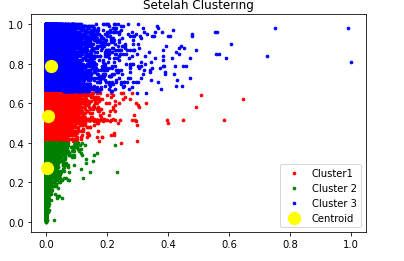
Kemudian setiap poin data dikelompokkan berdasarkan nilai minimum dan disimpan pada output. Kemudian juga dilakukan perhitungan mean untuk setiap cluster acak dan ditentukan sebagai centroid baru. Kemudia cent digunakan untuk menyimpan solusi iterasi tertentu.

Tahapan tersebut dilakukan terus-menerus selama iterasi yang telah ditentukan

d.



Setelah dilakukan clustering maka tampilan nya akan menjadi seperti berikut.



Link youtube : https://youtu.be/6Ad2IsaQzc8