**Laporan Case-Based 2**

**Mata Kuliah Pembelajaran Mesin**

**K-Means**

**Wandi Yusuf Kurniawan – 1301218601 – IFX-45-GAB – IKN**

****

**Program Studi Sarjana Informatika**

**Fakultas Informatika**

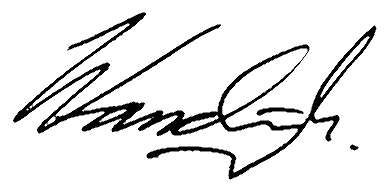
**Universitas Telkom**

**Bandung**

**2022**

**PERNYATAAN KODE ETIK AKADEMIK UNIVERSITAS TELKOM**

Dengan ini saya bersaksi bahwa hasil kerja ini **saya kerjakan sendiri** dan tidak mencontek hasil kerja dari mahasiswa lain dan tidak melakukan kecurangan dalam pengerjaan tugas ini. Jika kesaksian saya ini tidak benar, maka **saya bersedia menerima sanksi diberi minimum nilai E untuk mata kuliah ini dan/atau maksimum untuk semua mata kuliah pada semester ini.**



Wandi Yusuf Kurniawan

Pertama-tama, penulis melakukan import library yang dibutuhkan untuk mengerjakan tugas ini pada Jupyter Notebook menggunakan bahasa Python, diantaranya adalah:

Text

Description automatically generated

**IKHTISAR KUMPULAN DATA YANG DIPILIH**

Kemudian inisialisasi variable data untuk membaca file .data dan .names yang diambil dari <https://archive.ics.uci.edu/ml/datasets/Water+Treatment+Plant> dan lakukan konversi ke format .xlsx dan memanggil .head() untuk memastikan bahwa variable tersebut sudah berisi dataset yang nantinya akan dimasukkan ke dalam algoritma k-means yang dibuat penulis.

Graphical user interface, text, application, email

Description automatically generated

Cek .info() pada data untuk mengetahui informasi kolom, berdasarkan table di bawah terdapat 527 baris dan 38 kolom, dan banyak baris yang bernilai null pada 31 kolom seperti pada barplot di bawah ini.

Chart, histogram

Description automatically generated Table

Description automatically generated

**RINGKASAN PRA-PEMROSESAN DATA YANG DIUSULKAN**

Penulis mengisi data null tersebut dengan metode ‘ffill’ dan ‘bfill’ karena dataset tersebut berupa time-series. Ffill atau forward fill berfungsi untuk mengisi nilai null dengan nilai di row sebelumnya, begitupun untuk bill atau backward fill yang berfungsi untuk mengisi nilai null dengan nilai di row setelahnya..

Graphical user interface, text, application

Description automatically generated

Table

Description automatically generated

Penulis melakukan pengecekan apakah setiap kolom memiliki korelasi yang kuat, Tetapi karena ada 38 kolom, maka sangat sulit untuk mengecek korelasi antar kolom, dan itu akan mempersulit visualisasi algoritma k-means.

A screenshot of a computer

Description automatically generated with medium confidence

Sebagai contoh, penulis memilih dua kolom dengan korelasi terkecil, dalam hal ini adalah ‘SED-E’ dan ‘PH-S’, untuk divisualisasikan dengan dot hitam karena belum ada titik kluster yang mengelompokkan label.

Text

Description automatically generated with medium confidence

Chart, scatter chart

Description automatically generated

**MENERAPKAN ALGORITMA YANG DI PILIH**

Algoritma yang penulis rancang adalah k-means tanpa library sklearn karena dalam aturan pengerjaan tidak diperkenankan. Algoritma ini ditulis menggunakan referensi dari channel Emma Ding (<https://www.youtube.com/watch?v=uLs-EYUpGAw>) dengan sedikit modifikasi.

Text

Description automatically generated

Fungsi di atas membutuhkan dua parameter, yaitu data berisi array dua kolom dan jumlah cluster yang dilambangkan k. Fungsi ini mengembalikan label berupa array yang dihasilkan dari pasangan label, dan centroid berupa array yang berisi titik koordinat pada setiap centroid label.

Text

Description automatically generated

Iterasi pada fungsi utama dibantu dengan fungsi should\_stop yang akan terus berjalan sampai jumlah jarak pergerakan centroid lama ke centroid baru dibawah threshold, yaitu 0.001

Text

Description automatically generated

Fungsi initialize\_centroid menentukan posisi centroid secara random dengan rentang nilai minimum dan maksimum pada setiap kolomnya (fungsi random\_sample).

Text

Description automatically generated

Fungsi get\_label menghitung jarak semua titik (kolom) ke centoid dengan rumus euclidan seperti pada fungsi get\_distance. Jika jarak titik ke centroid baru lebih dekat daripada jarak ke centroid lama, maka titik tersebut termasuk ke label baru.

Text

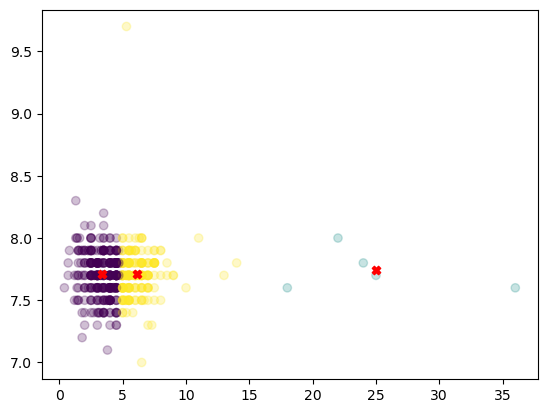
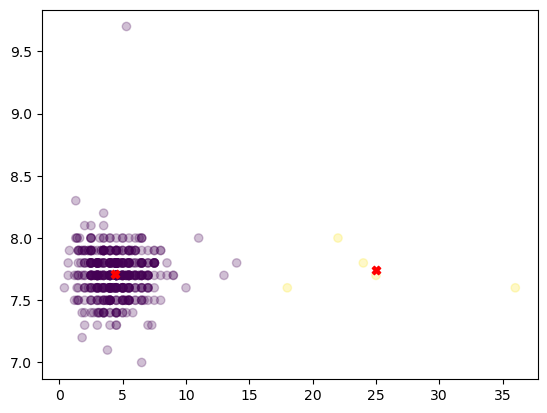
Description automatically generated

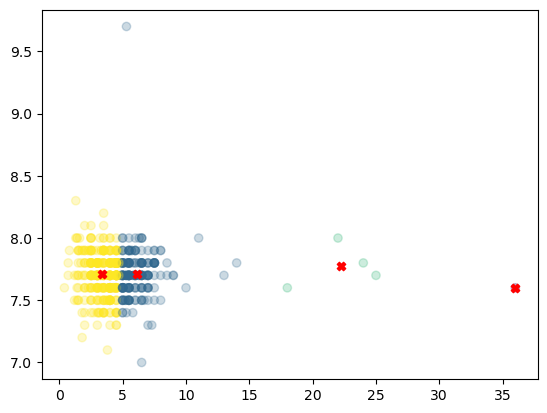
Fungsi update\_centroid menyesuaikan perpindahan centroid berdasarkan rata-rata kelompok titik yang termasuk pada suatu label.

**EVALUASI HASIL**

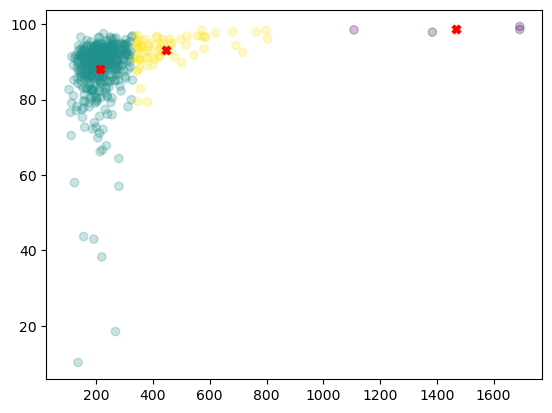
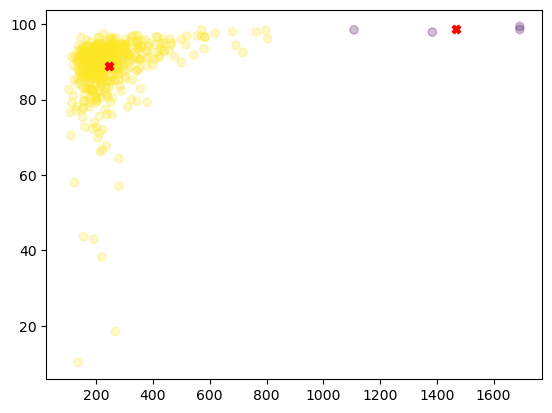
Penulis mencoba tiga pasang kolom dengan korelasi (yang sudah bernilai mutlak) terkecil, mendekati 0.25, dan terbesar mendekati 1 untuk mengecek berapa kluster yang ideal dalam pengelompokkan data.

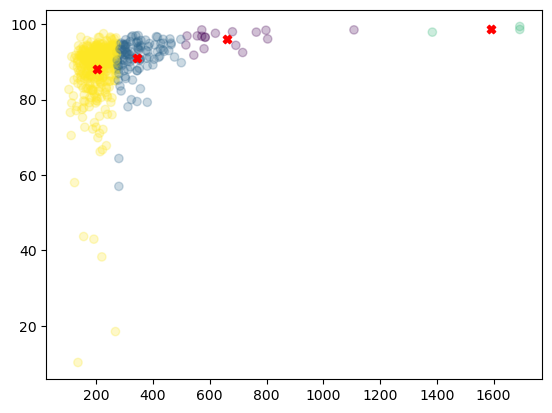
Untuk pasangan kolom yang memiliki korelasi mendekati 0, yaitu kolom ‘SED-E’ dan ‘PH-S’, scatter yang dihasilkan adalah sebagai berikut (untuk k = **2**, 3, 4. **Huruf tebal** menyatakan jumlah kluster terbaik\*):



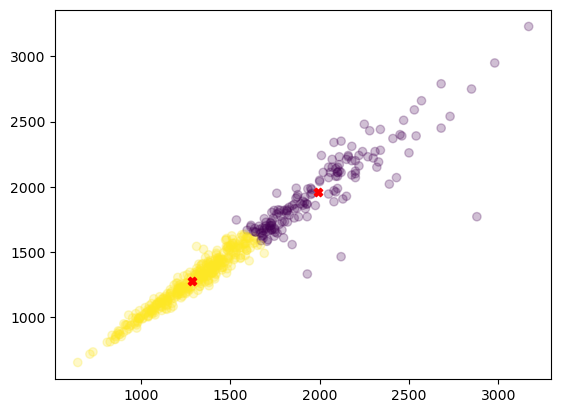


Untuk pasangan kolom yang memiliki korelasi mendekati 0.25, yaitu kolom ‘SS-P’ dan ‘RD-SS-G’, scatter yang dihasilkan adalah sebagai berikut (untuk k = **2**, 3, 4. **Huruf tebal** menyatakan jumlah kluster terbaik\*):





Untuk pasangan kolom yang memiliki korelasi mendekati 1, yaitu kolom ‘COND-P’ dan ‘COND-E’, scatter yang dihasilkan adalah sebagai berikut (untuk k = 2, **3**. **Huruf tebal** menyatakan jumlah kluster terbaik\*):

Chart, scatter chart

Description automatically generated

\*Catatan: Penulis tidak melakukan elbow method karena hanya bisa dilakukan menggunakan library sklearn, yang mana berlawanan dengan aturan pengerjaan yang ditentukan, sehingga penulis hanya bisa mengira-ngira jumlah kluster terbaik berdasarkan distribusi titik. Selain itu, algoritma yang penulis rancang mempunyai kekurangan dalam penentuan jumlah kluster yang lebih banyak, seperti pada keterangan error di bawah yang menyatakan tidak ada label yang bernilai 4).

Graphical user interface, text, application

Description automatically generated

**Link Source Code, Laporan, dan Slide Presentasi :** <https://github.com/onedeetelyu/tugas_machine_learning>

**Link Video Presentasi :** <https://youtu.be/RbKI4LJXrFQ>