CLUSTERING TERHADAP DATASET SALJU MENGGUNAKAN MODEL ALGORITMA K-MEANS

Husnul Aminindya Maheswari (1301174632)

IF-42-02

1. Teknik Eksplorasi dan Pre-Processing Data
    a. Data Eksplorasi

       Pertama, dataset yang akan digunakan untuk pembangunan modelling ini adalah **'salju_train.csv'**. Data eksplorasi menggunakan heatmap untuk melihat korelasi antar atribut. Selain menggunakan heatmap, korelasi dapat dibuktikan dengan fungsi 'correlation' yang dapat menentukan korelasi yang tinggi/kuat antar atribut (Korelasi >0.7). Hasil dari data eksplorasi berikut adalah 4 atribut yaitu Suhu3pm, Suhu9am, SuhuMax, dan Tekanan 3pm, dimana 4 atribut tersebut akan dipakai untuk feature selection dan modelling. Untuk menampung 4 atribut dibuatkan nama file **'dataset_baru.csv'**

    b. Duplicated data

       Setelah menentukan atribut, selanjutnya adalah mengecek apakah dataset mempunyai duplikasi data (duplicated data). Duplikasi data adalah data yang berbagi atau mempunyai nilai yang sama dengan data lainnya. Dalam laporan ini, duplikasi data di drop atau dihapuskan agar model yang akan dibangun memberikan hasil yang lebih optimal.

    c. Missing Values

       Setelah menghapus duplikasi data, saatnya mencari data yang NULL atau tidak ada nilainya (missing values). Di dalam laporan ini, cara mengatasi missing values adalah dengan menggunakan mean. Dataset yang berisi missing values tidak di drop karena di tahap duplicated data, sudah ada beberapa dataset yang di drop, sehingga ingin menahan keoptimalan pembuatan model clustering. Dataset ini juga tidak menggunakan median karena dataset tidak skewed.

    d. Data Outlier

Setelah mengatasi missing values, saatnya mengecek apakah dataset tersebut mempunyai data outlier. Metode deteksi dan penanganan data outliers adalah dengan menggunakan IQR atau InterQuartile Range Rule. Dipakainya IQR karena secara statistik, IQR menghitung distribusi yang adil terhadap setiap data dan penggunaan IQR bisa dibilang sangat sederhana.

e. Feature Scaling

Setelah mengatasi data outlier, dilakukannya feature scaling. Alasannya adalah supaya semua 4 atribut mempunyai rentan nilai yang sama atau mendekati dan menjadi lebih akurat dalam pembuatan modellingnya.

f. PCA

Terakhir di dalam tahap Pre-Processing adalah PCA (Principal Component Analysis). Mengetahui adanya 4 atribut, maka PCA berfungsi untuk mereduksi dan mengkompres data-data di 4 atribut tersebut, sehingga atribut yang akan dipakai untuk modelling ada 2. Untuk menampung hasil PCA, dibuatkan dataset baru bernama **'datatrain_baru.csv'**.

2. Algoritma K-Means

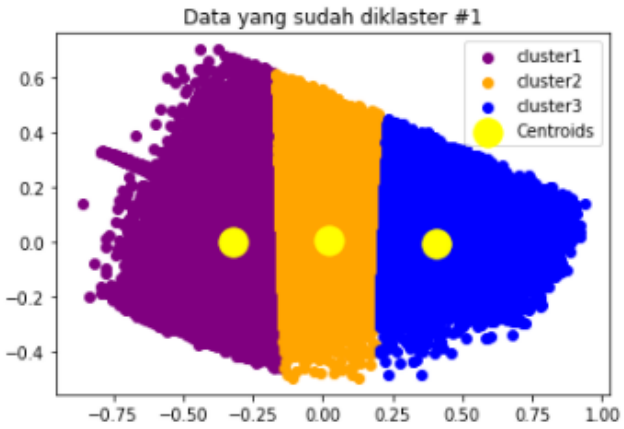
a. Teori

Algoritma K-Means merupakan salah satu metode clustering yang populer karena kesederhanaan dan tingkat akurasi yang tinggi. Cara kerja K-Means iteratif dan sebagai berikut:

1. Membuat centroid dengan letaknya secara acak
2. Menentukan jumlah centroid yang ada
3. Untuk setiap data point, hitung Euclidean Distance antara data point dengan centroid yang terdekat dan meletakkan data point tersebut di centroid tersebut.
4. Menghitung ulang letak centroid sebagai titik tengah cluster
5. Menempatkan data point ke centroid yang baru
6. Jika centroid tidak berubah tempat, hentikan algoritma

3. Hasil dan Analisis

Model machine learning yang dibangun untuk dataset ini adalah dengan KMeans. Dengan menentukan jumlah centroid K sebanyak 3 dan jumlah iterasi sebanyak 50 kali, maka hasil visualisasi clustering adalah sebagai berikut.

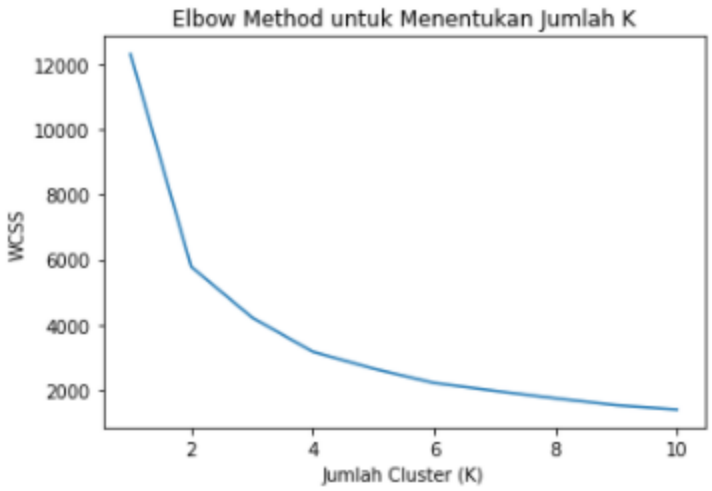


Terdapat 3 cluster dan 3 centroid. Cluster 1 adalah sekumpulan data yang mempunyai nilai PCA di x-axis < -0.15, Cluster 2 adalah sekumpulan data yang mempunyai nilai PCA diantara -0.15 dan 0.20, dan Cluster 3 adalah sekumpulan data yang mempunyai nilai PCA >0.20.

4. Eksperimen

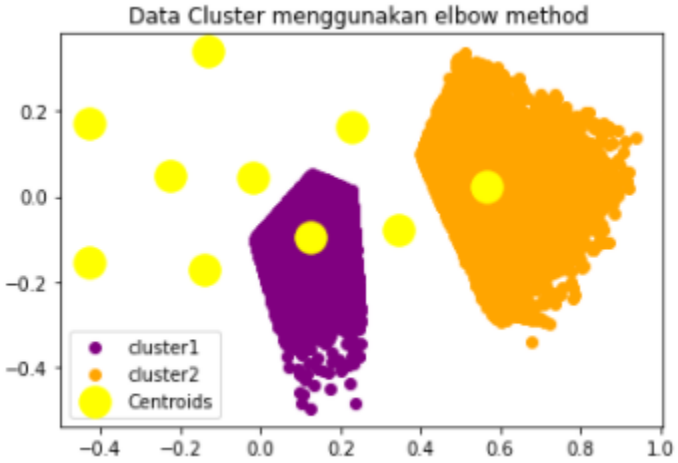
   Di dalam tahap eksperimen, penulis akan membuat sebuah perbanding antara graf yang menggunakan Elbow Method dan yang tidak.

   Sebelum membuat perbandingan, algoritma Elbow Method atau WCSS harus dibangunkan terlebih dahulu. Setelah algoritma tersebut dirun, maka hasil graf adalah sebagai berikut.



Graf di atas menunjukkan Elbow Method untuk menentukan jumlah centroid K. Dapat diamati bahwa siku graf berada di angka 2. Maka dari itu, menurut Elbow Method, jumlah centroid yang paling optimal adalah sebanyak 2.

Jika dibuatkan sebuah scatter plot, dengan K = 2 dan jumlah iterasi sebanyak 50 kali, maka hasil visualisasi graf adalah sebagai berikut.



Dapat diamati bahwa Jika dibandingkan graf dataset yang menggunakan elbow method dengan yang tidak, dapat dilihat bahwa clustering di graf yang menggunakan Elbow Method lebih mudah dicermati dan diklasifikasi.

5. Kesimpulan

Dari hasil pemodelan clustering di atas, dapat dibuktikan bahwa clustering dengan menggunakan WCSS atau The Elbow Method memberikan hasil yang lebih optimal, sehingga mudah untuk mengklasifikasi data test jika ingin membuat prediksi.

LINK VIDEO PRESENTASI: HERE