# Clustering Dataset Emisi Gas CO dan NOX

Disusun Oleh:
Jaka Pratama Widiyanto (0620101011)
Kresna Bayu Sanyoto (40622110001)

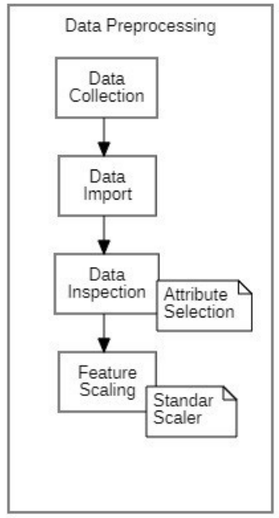
## Deskripsi Masalah

Emisi gas CO dan NOX merupakan masalah dalam lingkungan hidup, dimana pembangkit listrik menghasilkan flue gas, yaitu polutan Karbon Monoksida dan Nitrogen Oksida yang berasal dari pembakaran bahan bakar fosil. Polutan tersebut merupakan salah satu penyebab terbesar pemanasan global.

Pemantauan terhadap emisi pembangkit listrik dilakukan untuk menjaga batas wajar polutan yang dilepaskan ke atmosfir. Hasil penelitian menentukan kewajaran emisi CO dan NOX dapat membantu dalam masalah deteksi anomali pada penelitian lebih lanjut.

Dataset terdiri dari 36733 instance dari 11 pengukuran sensor yang diagregasi selama satu jam. Data dikumpulkan dalam rentang waktu 5 tahun yang disimpan kedalam dataset tersendiri untuk tiap tahunya (terdapat 5 dataset). Dalam kasus ini dataset digunakan adalah dataset tahun 2014. Dataset terdiri dari 7158 instance dan 11 atribut.

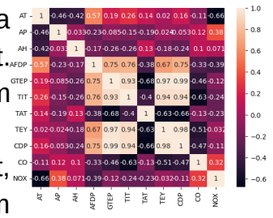
## Data Preprocessing



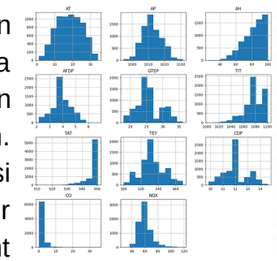

Gambar 1: data processing

Clustering dilakukan pada dataset dengan melakukan beberapa langkah diantaranya: data preprocessing, dimentional reduction, initial centroid, K-Means, dan menampilkan output. Pada data preprocessing dilakukan beberapa langkah untuk menyiapkan dataset sebelum dilakukan clustering.

Langkah pertama yang dilakukan dalam data processing adalah mendapatkan dataset, dataset yang digunakan didapat dari repositori UCI. Lalu, dataset tersebut diimport kedalam program (berbahasa python) menggunakan libraru pandas. Kemudian dicari tahu apakah terdapat missing value atau data null pada dataset, bagaimana persebaran datanya, dan ditentukan atribut apa saja yang akan digunakan. Pada kasus ini setelah diketahui bahwa tidak ada missing value atau data null, dilihat sebaran data menggunakan histogram dan dilihat juga korelasi antara atribut untuk menentukan atribut mana yang akan digunakan. Berdasarkan perhitungan korelasi yang digambarkan pada matriks, atribut yang berkorelasi kuat dengan jumlah gas CO dan NOX adalah: AFDP, GTEP, TIT, TEY, CDP. Terakhir dilakukan normalisasi standar terhadap data yang ditujukan agar suatu independent variable menjadi sama pentingnya dengan independent variable yang lain.

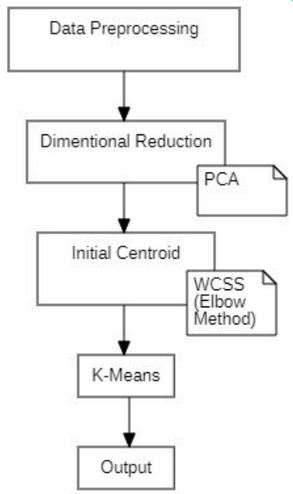



Gambar 2: matriks korelasi

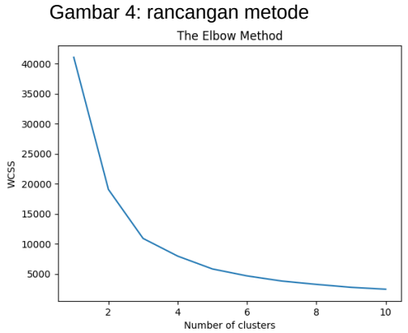


Gambar 3: sebaran data

## Rancangan Metode dan Model




Gambar 4: rancangan metode



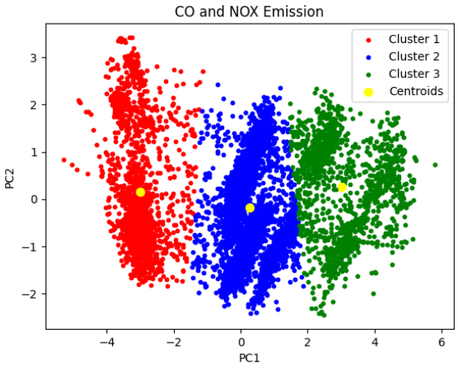

Gambar 5: metode elbow

Setelah dilakukan langkah data preprocessing kemudian dilakukan dimentional reduction menggunakan PCA. PCA adalah teknik pengurangan dimensi linier yang mengubah himpunan fitur yang berkorelasi dalam ruang dimensi tinggi menjadi serangkaian fitur yang tidak berkorelasi dalam ruang dimensi rendah.

Selanjutnya, dicari jumlah centroid optimum untuk menentukan banyaknya cluster. Pada kasus ini jumlah centroid didapat menggunakan metode elbow (WCSS). Dalam metode ini, untuk menentukan nilai k, kita terus menerus mengulang dari k=1 hingga k=n. Untuk setiap nilai k, kita menghitung nilai within-cluster sum of squares (WCSS). Untuk menentukan jumlah cluster (k) terbaik, kita membuat grafik k versus nilai WCSS mereka. Ketika k=1, nilai WCSS paling tinggi, tetapi dengan peningkatan nilai k, nilai WCSS mulai menurun. Kita memilih nilai k dari titik di mana grafik mulai terlihat seperti garis lurus.

Dalam kasus ini clustering dilakukan menggunakan metode K-Means. K-Means adalah algoritma unsupervised learning untuk pemodelan clustering. K-Means digunakan untuk mengetahui bagaimana data dikelompokkan. Algoritma ini bekerja dengan menentukan nilai K atau jumlah cluster. Kemudian setiap titik data dihitung jarak terdekat ke centroid. Titik akan menjadi bagian cluster dan centroid dihitung ulang. K-Means digunakan secara luas karena termasuk algoritma yang cepat dan hasilnya selalu konvergen.

## Hasil Clustering




Gambar 6: hasil clustering

Berdasarkan perhitungan WCSS (Elbow Method), nilai k yang dipilih adalah K=3. Penggunaan nilai k=3 menghasilkan 3 cluster seperti gambar di samping. Hasil K-Means memperlihatkan titik-titik data berkumpul dekat dengan centroid dan tidak ditemukan outlier yang terlalu jauh. Karenanya, dataset tersebut dapat digunakan untuk masalah deteksi anomali.

## Kesimpulan

Dataset yang digunakan adalah data emisi gas CO dan NOX dari pembangkit listrik. Pada kasus ini dataset yang digunakan adalah data emisi gas selama tahun 2014 dengan jumlah data 7158 instance dan 11 atribut. Kemudian dataset tersebut diproses sehingga didapat atribut yang berkorelasi kuat dengan jumlah gas CO dan NOX adalah AFDP, GTEP, TIT, TEY, dan CDP. Dataset yang telah dinormalisasi lalu dikurangi dimensi datanya. Berdasarkan perhitungan WCSS pada tiap iterasi K-Means, jumlah k yang optimum adalah k=3. Hasil K-Means memperlihatkan titik-titik data berkumpul dekat dengan centroid dan tidak ditemukan outlier yang terlalu jauh. Penggunaan metode K-Means dengan perhitungan WCSS dan PCA dapat digunakan sebagai data input untuk masalah anomaly detection.