Nama: Firstly Shures Jeffryson

Npm: 227006516080

Mata Kuliah: Machine Learning R02

Tujuan:

Tujuan dari tugas ini adalah melakukan pengelompokan (clustering) pada dataset yang berisi data numerik tanpa label kelas, untuk menemukan pola atau kelompok yang memiliki karakteristik mirip di dalam data tersebut. Dengan demikian, data dapat dibagi menjadi beberapa cluster berdasarkan kemiripan fitur-fiturnya, yang berguna untuk analisis lebih lanjut atau pengambilan keputusan.

Algoritma yang digunakan:

Algoritma yang digunakan adalah K-Means Clustering.

Alasan penggunaan K-Means:

1. Sederhana dan Efisien: K-Means merupakan algoritma clustering yang paling populer dan mudah diimplementasikan, terutama untuk dataset dengan fitur numerik. Prosesnya cepat dan efisien untuk dataset berukuran sedang hingga besar.
2. Tidak Memerlukan Label: Karena dataset tidak memiliki label (unsupervised learning), K-Means cocok untuk menemukan pola kelompok tanpa perlu data target.
3. Mudah Dipahami dan Diinterpretasikan: Hasil cluster berupa centroid yang merupakan representasi rata-rata dari masing-masing kelompok memudahkan interpretasi hasil.
4. Cocok untuk Data Numerik dan Berbentuk Cluster Bulat: K-Means bekerja baik jika cluster dalam data berbentuk bola/spherical dan jarak Euclidean dapat menggambarkan kemiripan antar data.

Kesimpulan:

Dengan tujuan mengelompokkan data tanpa label dan memudahkan analisis pola, serta memperhatikan sifat dataset, algoritma K-Means menjadi pilihan yang tepat dan efektif untuk tugas ini.

Hasil:

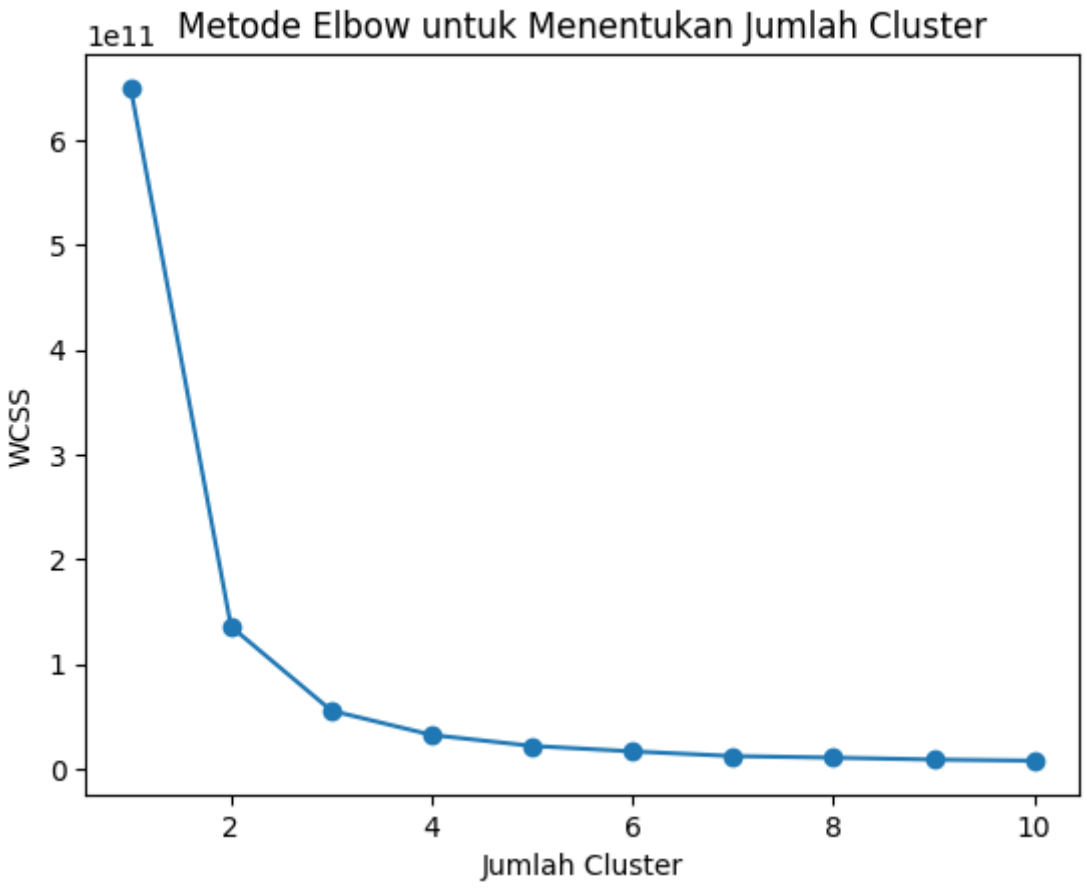
- **Cek data**

```
5 data pertama:
         date             time  epoch  moteid  temperature  humidity  light  \
0  2004-03-31  03:38:15.757551      2     1.0     122.1530  -3.91901  11.04
1  2004-02-28   00:59:16.02785      3     1.0      19.9884  37.09330  45.08
2  2004-02-28   01:03:16.33393     11     1.0      19.3024  38.46290  45.08
3  2004-02-28  01:06:16.013453     17     1.0      19.1652  38.80390  45.08
4  2004-02-28  01:06:46.778088     18     1.0      19.1750  38.83790  45.08

   voltage
0  2.03397
1  2.69964
2  2.68742
3  2.68742
4  2.69964
```

Berdasarkan data yang ditampilkan, ini adalah cuplikan dari 5 baris pertama sebuah dataset yang berisi informasi dari sensor. Setiap baris mewakili satu pembacaan sensor pada waktu tertentu. Kolom-kolom yang ada meliputi date (tanggal), time (waktu), epoch (jumlah detik sejak 1 Januari 1970), moteid (ID sensor atau mote), temperature (suhu), humidity (kelembapan), light (intensitas cahaya), dan voltage (tegangan). Pembacaan pertama pada tanggal 31 Maret 2004 menunjukkan suhu yang sangat tinggi (122.1530) dan kelembapan negatif (-3.91901), yang mungkin merupakan anomali atau kesalahan pembacaan. Sementara itu, empat pembacaan berikutnya, yang semuanya terjadi pada tanggal 28 Februari 2004, menunjukkan nilai suhu, kelembapan, dan cahaya yang lebih konsisten dan realistis. Data tegangan ditampilkan secara terpisah, di mana setiap barisnya sesuai dengan baris data sensor di atasnya.

- **Tentukan jumlah cluster optimal dengan metode Elbow**



Grafik yang ditampilkan menggunakan **Metode Elbow** untuk menentukan jumlah cluster optimal dalam algoritma *clustering* seperti K-Means.

Penjelasan Metode

Grafik ini memplot **WCSS** (*Within-Cluster Sum of Squares*) terhadap **jumlah cluster**. WCSS mengukur seberapa jauh setiap titik data dari pusat (*centroid*) clusternya. Tujuannya adalah untuk meminimalkan WCSS, yang berarti data dalam setiap cluster sangat berdekatan satu sama lain.

- Sumbu **X** menunjukkan **Jumlah Cluster** yang diuji, dari 1 hingga 10.
- Sumbu **Y** menunjukkan nilai **WCSS** untuk setiap jumlah cluster.

Saat jumlah cluster meningkat, WCSS secara alami akan menurun karena setiap cluster menjadi lebih kecil dan lebih padat. Namun, pada titik tertentu, penambahan cluster baru tidak akan memberikan penurunan WCSS yang signifikan lagi. Titik ini, yang menyerupai **siku** atau **elbow** pada lengan, adalah indikasi jumlah cluster yang optimal.

Interpretasi Grafik

Berdasarkan grafik, kita dapat melihat penurunan WCSS yang sangat curam dari 1 ke 2 cluster. Setelah itu, penurunannya mulai melandai. Siku yang paling jelas terlihat berada di **jumlah cluster 3**. Ini menunjukkan bahwa 3 adalah jumlah cluster yang paling efisien, karena setelah 3, penambahan cluster tidak memberikan banyak manfaat dalam mengurangi WCSS.

- **Tampilkan centroid cluster**

```
Centroid cluster:
[[4.06766370e+01 3.15513971e+01 6.54231109e+02 2.50227894e+00
  8.55271409e-12]
 [3.99756558e+01 3.13375312e+01 1.62685420e+03 2.51038239e+00
  1.00000000e+00]
 [3.86786429e+01 3.48146306e+01 1.02550363e+02 2.48351589e+00
  2.00000000e+00]]
```

Berdasarkan gambar, kode tersebut menampilkan hasil dari kmeans.cluster_centers_, yang merupakan atribut dari model K-Means setelah proses clustering selesai. Hasilnya menunjukkan tiga centroid cluster. Setiap baris dalam array tersebut mewakili satu centroid, dan setiap nilai di dalam baris tersebut adalah koordinat centroid untuk setiap fitur atau variabel dalam dataset.

Secara spesifik, output tersebut menampilkan:

- Baris pertama adalah **koordinat centroid** untuk cluster pertama.
- Baris kedua adalah **koordinat centroid** untuk cluster kedua.
- Baris ketiga adalah **koordinat centroid** untuk cluster ketiga.
-

Nilai-nilai ini, seperti 4.06766370e+01 dan 3.15513971e+01, adalah representasi notasi ilmiah dari koordinat, yang menunjukkan titik pusat dari masing-masing cluster dalam ruang multidimensi.

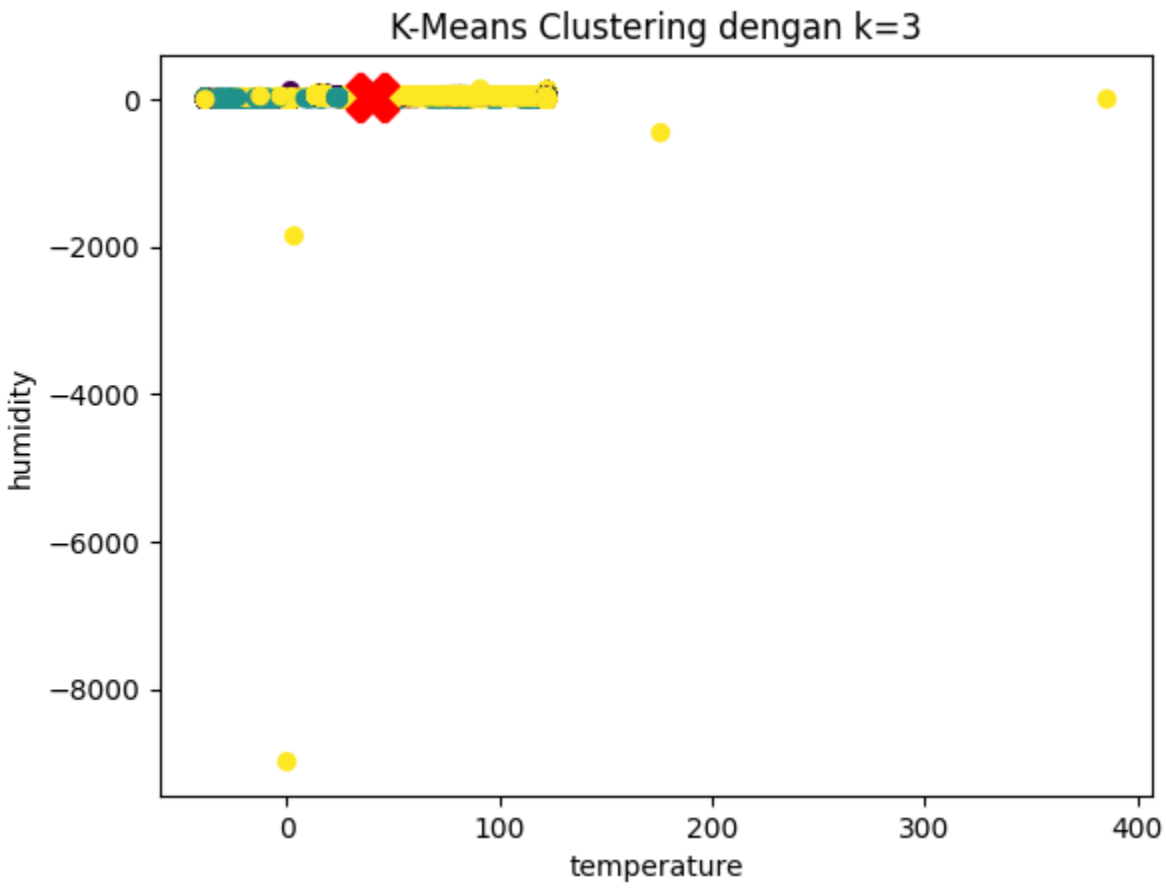
Tampilkan data dengan cluster

```
Data dengan cluster:
         date             time  epoch  moteid  temperature  humidity  light  \
0  2004-03-31  03:38:15.757551      2     1.0     122.1530  -3.91901  11.04
1  2004-02-28   00:59:16.02785      3     1.0      19.9884  37.09330  45.08
2  2004-02-28   01:03:16.33393     11     1.0      19.3024  38.46290  45.08
3  2004-02-28  01:06:16.013453     17     1.0      19.1652  38.80390  45.08
4  2004-02-28  01:06:46.778088     18     1.0      19.1750  38.83790  45.08

   voltage  Cluster_x  Cluster_y  Cluster
0  2.03397        2.0        2.0      2.0
1  2.69964        2.0        2.0      2.0
2  2.68742        2.0        2.0      2.0
3  2.68742        2.0        2.0      2.0
4  2.69964        2.0        2.0      2.0
```

Data yang ditampilkan merupakan hasil pengelompokan (clustering) dari dataset sensor. Setiap baris data sensor (tanggal, waktu, suhu, kelembapan, cahaya, dan tegangan) telah diberi label cluster. Berdasarkan kolom Cluster, Cluster_x, dan Cluster_y, semua 5 baris data ini ternyata termasuk dalam cluster yang sama, yaitu cluster 2.0. Ini menunjukkan bahwa berdasarkan algoritma clustering yang digunakan, semua data ini memiliki karakteristik yang cukup mirip satu sama lain sehingga dikelompokkan ke dalam satu grup.

- **Visualisasi**

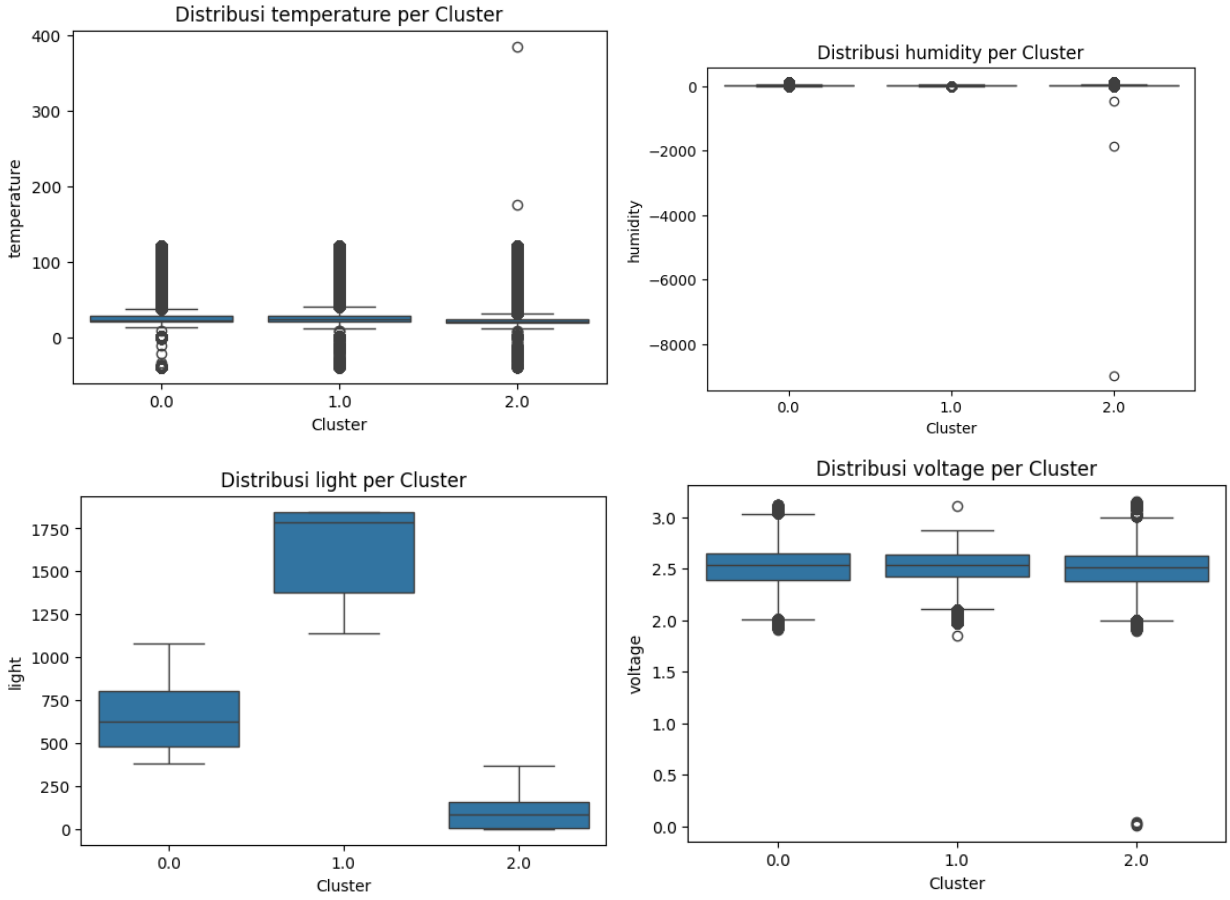


Gambar tersebut menampilkan hasil K-Means Clustering dengan jumlah cluster (k) sebanyak 3. Plot ini menunjukkan hubungan antara temperature (suhu) pada sumbu X dan humidity (kelembapan) pada sumbu Y.

Setiap titik data (lingkaran berwarna) mewakili satu pembacaan sensor yang telah dikelompokkan ke dalam salah satu dari tiga cluster yang berbeda. Tiga cluster tersebut ditandai dengan warna yang berbeda (kuning, ungu, dan hijau).

Simbol tanda silang merah (X) yang besar menunjukkan pusat (centroid) dari masing-masing cluster. Kita bisa melihat ada tiga tanda silang merah, masing-masing berada di tengah-tengah kelompok titik data yang berbeda.

Dari plot ini, terlihat bahwa sebagian besar data berkumpul di sekitar nilai humidity 0. Namun, ada beberapa titik data yang merupakan outlier dengan nilai humidity yang sangat rendah (negatif) dan nilai temperature yang sangat tinggi, yang terisolasi dari kelompok utama.



Berdasarkan keempat diagram boxplot tersebut, kita dapat melihat distribusi masing-masing fitur (light, temperature, humidity, dan voltage) berdasarkan cluster yang telah dibentuk (0, 1, dan 2).

Distribusi light per Cluster

Diagram boxplot untuk light menunjukkan perbedaan yang jelas antar cluster. Cluster 1 memiliki nilai cahaya tertinggi dengan median di atas 1500, diikuti oleh Cluster 0 dengan median sekitar 600, dan Cluster 2 memiliki nilai cahaya terendah dengan median di bawah 100.

Distribusi temperature per Cluster

Untuk temperature, ketiga cluster memiliki distribusi yang sangat mirip dengan rentang interkuartil (IQR) yang berdekatan. Median ketiga cluster juga hampir sama, yaitu sekitar 20. Namun, terdapat banyak data outlier (pencilan) pada ketiga cluster, terutama pada Cluster 2, yang memiliki nilai temperature mendekati 400.

Distribusi humidity per Cluster

Mirip dengan temperature, distribusi humidity di ketiga cluster juga terlihat sangat mirip dengan median di sekitar 0. Namun, terdapat beberapa outlier ekstrem pada Cluster 2 dengan nilai kelembapan negatif yang sangat rendah, bahkan ada yang mendekati -9000, yang kemungkinan besar adalah data anomali atau kesalahan pembacaan.

Distribusi voltage per Cluster

Distribusi voltage di ketiga cluster juga serupa, dengan median berada di sekitar 2.5 hingga 2.6. Namun, terdapat beberapa outlier yang signifikan pada Cluster 2, dengan satu pencilan berada di nilai yang sangat rendah, mendekati 0.

Secara keseluruhan, perbedaan paling mencolok antar cluster terletak pada fitur light. Sementara fitur temperature, humidity, dan voltage menunjukkan distribusi yang relatif seragam di antara ketiga cluster, meskipun terdapat outlier yang signifikan, terutama pada Cluster 2. Hal ini mengindikasikan bahwa fitur light adalah variabel yang paling dominan dalam proses pengelompokan data ini.