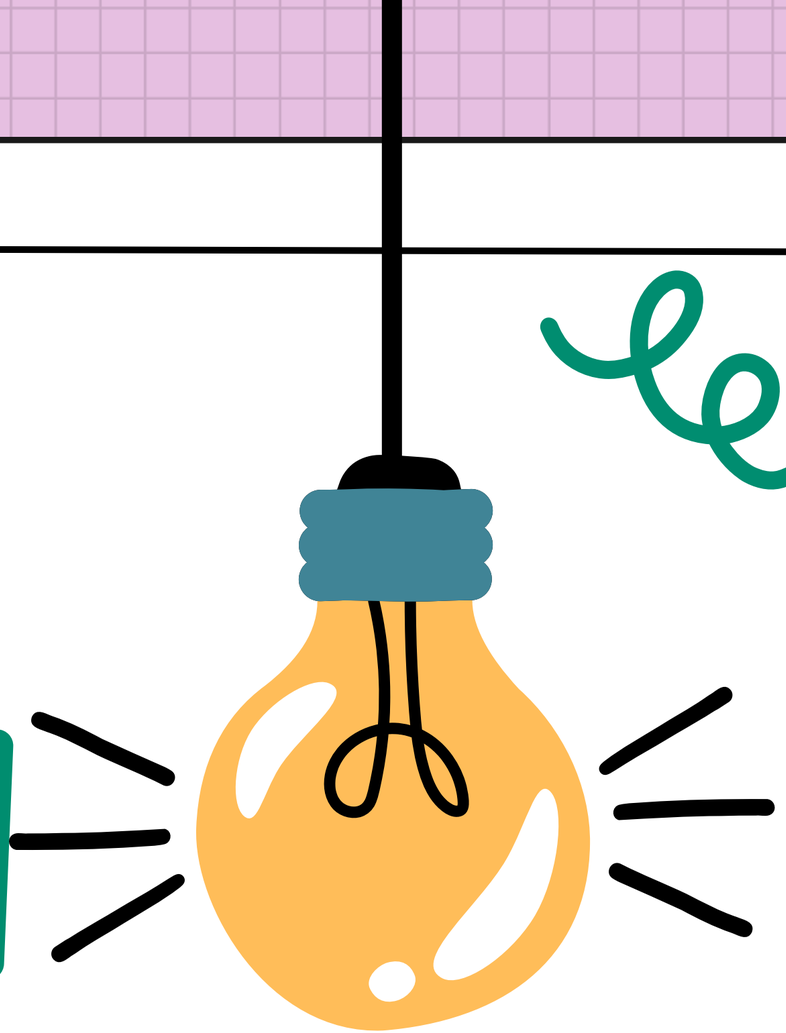
LAPORAN PRAKTIKUM MACHINE LEARNING

CLUSTER ANALYSIS



Dosen Pengampu : Nur Rosyid Mubtadai S.Kom., M.T.

Bayu Kurniawan / 3322600019

Praktikum Cluster Analysis

Penjelasan:

- Variance Within Clusters (VWC) adalah ukuran sebaran atau variasi dari setiap cluster dalam suatu metode pengelompokan data, seperti k-means clustering. VWC mengestimasi sebaran observasi di dalam setiap cluster, dengan nilai yang lebih kecil menunjukkan cluster dengan sebaran rendah atau lebih padat. (Semakin kecil nilai varians ini, semakin rapat atau homogen data dalam klaster tersebut)

- Variance Between Clusters (VBC) adalah ukuran sebaran atau variasi antara cluster dalam suatu metode pengelompokan data. VBC mengukur seberapa jauh kelompok-kelompok tersebut terpisah satu sama lain. (Semakin besar nilai varians antar klaster, semakin berbeda atau terpisah klaster-klaster tersebut)

- Total varians atau varians keseluruhan dari dataset adalah jumlah dari varians antar klaster dan varians dalam klaster. Ini mencerminkan sejauh mana data dalam keseluruhan dataset tersebar atau berbeda.

Dalam analisis pengelompokan data, tujuan utama adalah untuk mencapai VWC yang rendah dan VBC yang tinggi. Hal ini menunjukkan bahwa observasi di dalam setiap cluster saling berdekatan dan terpisah dengan jelas antara cluster-cluster yang berbeda. Jika VWC tinggi dan VBC rendah, ini dapat menunjukkan bahwa pengelompokan tidak efektif dalam mengelompokkan data dengan baik.

CLUSTERING DATASET MILK.CSV

Praktikum Cluster Analysis

```python
# assignment 1
import pandas as pd

dataset = pd.read_csv("C:/Users/bayuk/OneDrive/Documents/AI/pens/smtr3/Machine Learning/Data/milk.csv")
dataset
```

```python
# assignment 2
dataset = dataset.fillna(dataset.groupby("Grade").transform("mean"))
print("\n Dataset setelah pengisian mising value\n", dataset)
```

```
 Dataset setelah pengisian mising value
        pH  Temprature  Taste  Odor  Fat   Turbidity  Colour   Grade
0     6.6          35      1     0     1          0     254    high
1     6.6          36      0     1     0          1     253    high
2     8.5          70      1     1     1          1     246     low
3     9.5          34      1     1     0          1     255     low
4     6.6          37      0     0     0          0     255  medium
...   ...         ...    ...   ...   ...        ...     ...     ...
1054  6.7          45      1     1     0          0     247  medium
1055  6.7          38      1     0     1          0     255    high
1056  3.0          40      1     1     1          1     255     low
1057  6.8          43      1     0     1          0     250    high
1058  8.6          55      0     1     1          1     255     low

[1059 rows x 8 columns]
```

Praktikum Cluster Analysis

```
from sklearn.preprocessing import MinMaxScaler
import numpy as np
# Menghapus variable Grade
dataset = dataset.drop(columns=["Grade"])

# Membuat objek scaler dengan rentang 0-1
scaler = MinMaxScaler(feature_range=(0, 1))

# Melakukan normalisasi pada data
data_norm = scaler.fit_transform(dataset)

# Membuat dataframe pandas dari data ternormalisasi
df = pd.DataFrame(data_norm, columns=[ "pH",  "Temprature",  "Taste",  "Odor",  "Fat",  "Turbidity",  "Colour"])
from sklearn.cluster import KMeans

data = df.loc[:,['pH','Temprature','Taste',"Fat",'Odor','Turbidity','Colour']]

kmeans = KMeans(n_clusters=3, init="random", n_init=1)
clusters=kmeans.fit_predict(data)

# Mendapatkan label cluster untuk setiap titik data
labels = kmeans.labels_

# Mendapatkan pusat cluster
cluster_centers = kmeans.cluster_centers_

# Menghitung varian dalam kelompok
within_cluster_var = np.mean([np.linalg.norm(X[labels == i] - cluster_centers[i]) ** 2 for i in range(3)])

# Menghitung varian antara kelompok
between_cluster_var = np.mean([np.linalg.norm(cluster_centers[i] - kmeans.cluster_centers_) ** 2 for i in range(3)])

# Menghitung varian dataset
dataset_var = np.mean([np.linalg.norm(between_cluster_var - within_cluster_var) ** 2 ])
print("Within-cluster variance:", within_cluster_var)
print("Between-cluster variance:", between_cluster_var)
print("Dataset-cluster variance:", dataset_var)
```

✓ 0.0s

```
Within-cluster variance: 22989362.775022704
Between-cluster variance: 2.6784593040599454
Dataset-cluster variance: 528510677649461.5
```

- **Variance Within Clusters**: 22989362.775022704 => Variance Within Clusters mengukur sejauh mana titik data dalam suatu klaster tersebar. Semakin rendah nilai ini, semakin rapat titik data dalam klaster tersebut.
- **Variance Between Clusters**: 2.6784593040599454 => Variance Between Clusters mengukur sejauh mana klaster-klaster tersebut terpisah satu sama lain. Semakin tinggi nilai ini semakin terpisah klaster-klaster tersebut.
- **Total Variance (Dataset Variance)**: 528510677649461.5=> Total Variance atau varian dataset adalah jumlah dari varians antar klaster dan varians dalam klaster. Nilai ini mencerminkan variasi keseluruhan dalam dataset.

Dalam konteks clustering menggunakan K-Means pada dataset Milk.csv, tujuan utama adalah untuk menciptakan klaster-klaster yang saling terpisah dengan variasi dalam klaster yang rendah. Dengan kata lain, kita ingin mencapai nilai Variance Between Clusters yang tinggi dan nilai Variance Within Clusters yang rendah. Dalam output tersebut, Variance Between Clusters adalah 22989362.775022704, yang menunjukkan bahwa klaster-klaster tersebut cukup terpisah Variance Within Clusters adalah 2.6784593040599454 yang menunjukkan bahwa titik data dalam cluster tersebut terpisah. Total Variance (Dataset Variance) adalah 528510677649461.5, yang mencerminkan variasi keseluruhan dalam dataset.

Praktikum Cluster Analysis

```python
from sklearn.cluster import AgglomerativeClustering

data = df.loc[:,['pH','Temprature','Taste',"Fat",'Odor','Turbidity','Colour']]

clustering=AgglomerativeClustering(n_clusters=3, linkage='average')
clusters=clustering.fit_predict(data)

# Mendapatkan label cluster untuk setiap titik data
labels = clustering.labels_

# Mendapatkan pusat cluster
cluster_centers = np.array([X[labels == i].mean(axis=0) for i in range(3)])

# Menghitung varian dalam kelompok
within_cluster_var = np.mean([np.linalg.norm(X[labels == i] - cluster_centers[i]) ** 2 for i in range(3)])

# Menghitung varian antara kelompok
between_cluster_var = np.mean([np.linalg.norm(cluster_centers[i] - kmeans.cluster_centers_) ** 2 for i in range(3)])

# Menghitung varian dataset
dataset_var = np.mean([np.linalg.norm(between_cluster_var - within_cluster_var) ** 2 ])

print("Within-cluster variance:", within_cluster_var)
print("Between-cluster variance:", between_cluster_var)
print("Dataset-cluster variance:", dataset_var)
```

✓ 0.0s

```
Within-cluster variance: 40139.68572500653
Between-cluster variance: 194662.34165003712
Dataset-cluster variance: 23877251194.12539
```

- **Variance Within Clusters** mengukur sejauh mana titik data dalam suatu klaster tersebar. Semakin rendah nilai ini, semakin rapat titik data dalam klaster tersebut. Dalam output yang diberikan, nilai Total Variance Within Clusters adalah 40139.68572500653, yang menunjukkan bahwa titik data dalam klaster tersebut cukup rapat.
- **Variance Between Clusters** mengukur sejauh mana klaster-klaster tersebut terpisah satu sama lain. Semakin tinggi nilai ini, semakin terpisah klaster-klaster tersebut. Dalam output yang diberikan, nilai Total Variance Between Clusters adalah 194662.34165003712, yang menunjukkan bahwa klaster-klaster tersebut cukup terpisah.
- **Total Variance (Dataset Variance)** adalah jumlah dari varians antar klaster dan varians dalam klaster. Nilai ini mencerminkan variasi keseluruhan dalam dataset. Dalam output yang diberikan, nilai Total Variance (Dataset Variance) adalah 23877251194.12539, yang mencerminkan variasi keseluruhan dalam dataset.

Berdasarkan penjelasan di atas, dapat disimpulkan bahwa menggunakan Metode Single Linkage dapat diketahui klaster-klaster dalam dataset tersebut cukup terpisah satu sama lain, dengan titik data yang cukup rapat dalam masing-masing klaster. Namun, variasi keseluruhan dalam dataset masih cukup tinggi, menunjukkan adanya variasi yang signifikan antara klaster-klaster.

Praktikum Cluster Analysis

```python
# assignment 5 single
from sklearn.cluster import AgglomerativeClustering

data = df.loc[:,['pH','Temprature','Taste',"Fat",'Odor','Turbidity','Colour']]

clustering=AgglomerativeClustering(n_clusters=3, linkage='single')
clusters=clustering.fit_predict(data)

# Mendapatkan label cluster untuk setiap titik data
labels = clustering.labels_

# Mendapatkan pusat cluster
cluster_centers = np.array([X[labels == i].mean(axis=0) for i in range(3)])

# Menghitung varian dalam kelompok
within_cluster_var = np.mean([np.linalg.norm(X[labels == i] - cluster_centers[i]) ** 2 for i in range(3)])

# Menghitung varian antara kelompok
between_cluster_var = np.mean([np.linalg.norm(cluster_centers[i] - kmeans.cluster_centers_) ** 2 for i in range(3)])

# Menghitung varian dataset
dataset_var = np.mean([np.linalg.norm(between_cluster_var - within_cluster_var) ** 2 ])

print("Within-cluster variance:", within_cluster_var)
print("Between-cluster variance:", between_cluster_var)
print("Dataset-cluster variance:", dataset_var)
```

✓ 0.0s

```
Within-cluster variance: 41993.161378792094
Between-cluster variance: 195749.29236106438
Dataset-cluster variance: 23640947814.637665
```

- **Variance Within Clusters** mengukur sejauh mana titik data dalam suatu klaster tersebar. Semakin rendah nilai ini, semakin rapat titik data dalam klaster tersebut. Dalam output yang diberikan, nilai Total Variance Within Clusters adalah 41993.161378792094, yang menunjukkan bahwa titik data dalam klaster tersebut cukup rapat.
- **Variance Between Clusters** mengukur sejauh mana klaster-klaster tersebut terpisah satu sama lain. Semakin tinggi nilai ini, semakin terpisah klaster-klaster tersebut. Dalam output yang diberikan, nilai Total Variance Between Clusters adalah 195749.29236106438, yang menunjukkan bahwa klaster-klaster tersebut cukup terpisah.
- **Total Variance (Dataset Variance)** adalah jumlah dari varians antar klaster dan varians dalam klaster. Nilai ini mencerminkan variasi keseluruhan dalam dataset. Dalam output yang diberikan, nilai Total Variance (Dataset Variance) adalah 23640947814.637665, yang mencerminkan variasi keseluruhan dalam dataset.

Berdasarkan penjelasan di atas, dapat disimpulkan bahwa menggunakan Metode Average Linkage dapat diketahui klaster-klaster dalam dataset tersebut cukup terpisah satu sama lain, dengan titik data yang cukup rapat dalam masing-masing klaster. Namun, variasi keseluruhan dalam dataset masih cukup tinggi, menunjukkan adanya variasi yang signifikan antara klaster-klaster.

Praktikum Cluster Analysis

```
# assignment 5 sentroik
from sklearn.cluster import AgglomerativeClustering

data = df.loc[:,['pH','Temprature','Taste',"Fat",'Odor','Turbidity','Colour']]

clustering=AgglomerativeClustering(n_clusters=3, linkage='ward')
clusters=clustering.fit_predict(data)

# Mendapatkan label cluster untuk setiap titik data
labels = clustering.labels_

# Mendapatkan pusat cluster
cluster_centers = np.array([X[labels == i].mean(axis=0) for i in range(3)])

# Menghitung varian dalam kelompok
within_cluster_var = np.mean([np.linalg.norm(X[labels == i] - cluster_centers[i]) ** 2 for i in range(3)])

# Menghitung varian antara kelompok
between_cluster_var = np.mean([np.linalg.norm(cluster_centers[i] - kmeans.cluster_centers_) ** 2 for i in range(3)])

# Menghitung varian dataset
dataset_var = np.mean([np.linalg.norm(between_cluster_var - within_cluster_var) ** 2 ])

print("Within-cluster variance:", within_cluster_var)
print("Between-cluster variance:", between_cluster_var)
print("Dataset-cluster variance:", dataset_var)
```

✓ 0.0s

```
Within-cluster variance: 39576.60630636784
Between-cluster variance: 194903.55283717846
Dataset-cluster variance: 24126460318.5853
```

- **Variance Within Clusters** mengukur sejauh mana titik data dalam suatu klaster tersebar. Semakin rendah nilai ini, semakin rapat titik data dalam klaster tersebut. Dalam output yang diberikan, nilai Total Variance Within Clusters adalah 39576.60630636784, yang menunjukkan bahwa titik data dalam klaster tersebut cukup rapat.
- **Variance Between Clusters** mengukur sejauh mana klaster-klaster tersebut terpisah satu sama lain. Semakin tinggi nilai ini, semakin terpisah klaster-klaster tersebut. Dalam output yang diberikan, nilai Total Variance Between Clusters adalah 194903.55283717846, yang menunjukkan bahwa klaster-klaster tersebut cukup terpisah.
- **Total Variance (Dataset Variance)** adalah jumlah dari varians antar klaster dan varians dalam klaster. Nilai ini mencerminkan variasi keseluruhan dalam dataset. Dalam output yang diberikan, nilai Total Variance (Dataset Variance) adalah 24126460318.5853 yang mencerminkan variasi keseluruhan dalam dataset.

Berdasarkan penjelasan di atas, dapat disimpulkan bahwa menggunakan Metode Average Linkage dapat diketahui klaster-klaster dalam dataset tersebut cukup terpisah satu sama lain, dengan titik data yang cukup rapat dalam masing-masing klaster. Namun, variasi keseluruhan dalam dataset masih cukup tinggi, menunjukkan adanya variasi yang signifikan antara klaster-klaster.

Praktikum Cluster Analysis

```python
# assignment 5 complete
from sklearn.cluster import AgglomerativeClustering

data = df.loc[:,['pH','Temprature','Taste',"Fat",'Odor','Turbidity','Colour']]

clustering=AgglomerativeClustering(n_clusters=3, linkage='complete')
clusters=clustering.fit_predict(data)

# Mendapatkan label cluster untuk setiap titik data
labels = clustering.labels_

# Mendapatkan pusat cluster
cluster_centers = np.array([X[labels == i].mean(axis=0) for i in range(3)])

# Menghitung varian dalam kelompok
within_cluster_var = np.mean([np.linalg.norm(X[labels == i] - cluster_centers[i]) ** 2 for i in range(3)])

# Menghitung varian antara kelompok
between_cluster_var = np.mean([np.linalg.norm(cluster_centers[i] - kmeans.cluster_centers_) ** 2 for i in range(3)])

# Menghitung varian dataset
dataset_var = np.mean([np.linalg.norm(between_cluster_var - within_cluster_var) ** 2 ])

print("Within-cluster variance:", within_cluster_var)
print("Between-cluster variance:", between_cluster_var)
print("Dataset-cluster variance:", dataset_var)
```

✓ 0.0s

```
Within-cluster variance: 43289.863207375514
Between-cluster variance: 194982.04417480098
Dataset-cluster variance: 23010517766.654156
```

- **Variance Within Clusters** mengukur sejauh mana titik data dalam suatu klaster tersebar. Semakin rendah nilai ini, semakin rapat titik data dalam klaster tersebut. Dalam output yang diberikan, nilai Total Variance Within Clusters adalah 43289.863207375514, yang menunjukkan bahwa titik data dalam klaster tersebut cukup rapat.
- **Variance Between Clusters** mengukur sejauh mana klaster-klaster tersebut terpisah satu sama lain. Semakin tinggi nilai ini, semakin terpisah klaster-klaster tersebut. Dalam output yang diberikan, nilai Total Variance Between Clusters adalah 194982.04417480098, yang menunjukkan bahwa klaster-klaster tersebut cukup terpisah.
- **Total Variance (Dataset Variance)** adalah jumlah dari varians antar klaster dan varians dalam klaster. Nilai ini mencerminkan variasi keseluruhan dalam dataset. Dalam output yang diberikan, nilai Total Variance (Dataset Variance) adalah 23010517766.654156, yang mencerminkan variasi keseluruhan dalam dataset.

Berdasarkan penjelasan di atas, dapat disimpulkan bahwa menggunakan Metode Average Linkage dapat diketahui klaster-klaster dalam dataset tersebut cukup terpisah satu sama lain, dengan titik data yang cukup rapat dalam masing-masing klaster. Namun, variasi keseluruhan dalam dataset masih cukup tinggi, menunjukkan adanya variasi yang signifikan antara klaster-klaster.

CLUSTERING WATER_POTABILITY .CSV

Praktikum Cluster Analysis

```python
# assignment 1
import pandas as pd

dataset = pd.read_csv("C:/Users/bayuk/OneDrive/Documents/AI/pens/smtr3/Machine Learning/Data/water_potability.csv")
```

✓ 0.0s

```python
# assignment 2
dataset = dataset.fillna(dataset.groupby("Potability").transform("mean"))
print("\n Dataset setelah pengisian mising value\n", dataset)
```

✓ 0.0s

```
 Dataset setelah pengisian mising value
            ph    Hardness       Solids  Chloramines     Sulfate  \
0     7.085378  204.890456  20791.31898     7.300212  368.516441
1     3.716080  129.422921  18630.05786     6.635246  334.564290
2     8.099124  224.236259  19909.54173     9.275884  334.564290
3     8.316766  214.373394  22018.41744     8.059332  356.886136
4     9.092223  181.101509  17978.98634     6.546600  310.135738
...        ...         ...          ...          ...         ...
3271  4.668102  193.681736  47580.99160     7.166639  359.948574
3272  7.808856  193.553212  17329.80216     8.061362  332.566990
3273  9.419510  175.762646  33155.57822     7.350233  332.566990
3274  5.126763  230.603758  11983.86938     6.303357  332.566990
3275  7.874671  195.102299  17404.17706     7.509306  332.566990

      Conductivity  Organic_carbon  Trihalomethanes  Turbidity  Potability
0       564.308654       10.379783        86.990970   2.963135           0
1       592.885359       15.180013        56.329076   4.500656           0
2       418.606213       16.868637        66.420093   3.055934           0
3       363.266516       18.436525       100.341674   4.628771           0
4       398.410813       11.558279        31.997993   4.075075           0
...            ...             ...              ...        ...         ...
3271    526.424171       13.894419        66.687695   4.435821           1
3272    392.449580       19.903225        66.539684   2.798243           1
3273    432.044783       11.039070        69.845400   3.298875           1
3274    402.883113       11.168946        77.488213   4.708658           1
3275    327.459761       16.140368        78.698446   2.309149           1

[3276 rows x 10 columns]
```

Praktikum Cluster Analysis

```
# Membuat objek scaler dengan rentang 0-1
scaler = MinMaxScaler(feature_range=(0, 1))

# Melakukan normalisasi pada data
data_norm = scaler.fit_transform(dataset)

# Membuat dataframe pandas dari data ternormalisasi
df = pd.DataFrame(data_norm, columns=[ 'ph', 'Hardness', 'Solids', 'Chloramines', 'Sulfate', 'Conductivity',
                                       'Organic_carbon', 'Trihalomethanes', 'Turbidity', 'Potability'])

data = df.loc[:,['ph', 'Hardness', 'Solids', 'Chloramines', 'Sulfate', 'Conductivity',
                 'Organic_carbon', 'Trihalomethanes', 'Turbidity']]

kmeans = KMeans(n_clusters=2, init="random", n_init=1)
clusters=kmeans.fit_predict(data)

# Mendapatkan label cluster untuk setiap titik data
labels = kmeans.labels_

# Mendapatkan pusat cluster
cluster_centers = kmeans.cluster_centers_

# Menghitung varian dalam kelompok
within_cluster_var = np.mean([np.linalg.norm(X[labels == i] - cluster_centers[i]) ** 2 for i in range(2)])

# Menghitung varian antara kelompok
between_cluster_var = np.mean([np.linalg.norm(cluster_centers[i] - kmeans.cluster_centers_) ** 2 for i in range(2)])

# Menghitung varian dataset
dataset_var = np.mean([np.linalg.norm(between_cluster_var - within_cluster_var) ** 2 ])

print("Within-cluster variance:", within_cluster_var)
print("Between-cluster variance:", between_cluster_var)
print("Dataset-cluster variance:", dataset_var)
```

✓ 0.0s

```
Within-cluster variance: 920247831483.6045
Between-cluster variance: 0.05764007221234954
Dataset-cluster variance: 8.468560713501705e+23
```

- **Variance Within Clusters**: 920247831483.6045=> Variance Within Clusters mengukur sejauh mana titik data dalam suatu klaster tersebar. Semakin rendah nilai ini, semakin rapat titik data dalam klaster tersebut.
- **Variance Between Clusters**: 0.05764007221234954=> Variance Between Clusters mengukur sejauh mana klaster-klaster tersebut terpisah satu sama lain. Semakin tinggi nilai ini semakin terpisah klaster-klaster tersebut.
- **Total Variance (Dataset Variance)**: 8.468560713501705e=> Total Variance atau varian dataset adalah jumlah dari varians antar klaster dan varians dalam klaster. Nilai ini mencerminkan variasi keseluruhan dalam dataset.

Dalam konteks clustering menggunakan K-Means pada dataset water_potability.csv, tujuan utama adalah untuk menciptakan klaster-klaster yang saling terpisah dengan variasi dalam klaster yang rendah. Dengan kata lain, kita ingin mencapai nilai Variance Between Clusters yang tinggi dan nilai Variance Within Clusters yang rendah. Dalam output tersebut, Variance Between Clusters adalah 920247831483.6045, yang menunjukkan bahwa klaster-klaster tersebut cukup terpisah Variance Within Clusters adalah 0.05764007221234954, yang menunjukkan bahwa titik data dalam cluster tersebut cukup terpisah. Total Variance (Dataset Variance) adalah 8.468560713501705e, yang mencerminkan variasi keseluruhan dalam dataset.

Praktikum Cluster Analysis

```
# assignment 5 average
from sklearn.cluster import AgglomerativeClustering

data = df.loc[:,['ph', 'Hardness', 'Solids', 'Chloramines', 'Sulfate', 'Conductivity',
                 'Organic_carbon', 'Trihalomethanes', 'Turbidity']]

clustering=AgglomerativeClustering(n_clusters=2, linkage='average')
clusters=clustering.fit_predict(data)

# Mendapatkan label cluster untuk setiap titik data
labels = clustering.labels_

# Mendapatkan pusat cluster
cluster_centers = np.array([X[labels == i].mean(axis=0) for i in range(2)])

# Menghitung varian dalam kelompok
within_cluster_var = np.mean([np.linalg.norm(X[labels == i] - cluster_centers[i]) ** 2 for i in range(2)])

# Menghitung varian antara kelompok
between_cluster_var = np.mean([np.linalg.norm(cluster_centers[i] - kmeans.cluster_centers_) ** 2 for i in range(2)])

# Menghitung varian dataset
dataset_var = np.mean([np.linalg.norm(between_cluster_var - within_cluster_var) ** 2 ])

print("Within-cluster variance:", within_cluster_var)
print("Between-cluster variance:", between_cluster_var)
print("Dataset-cluster variance:", dataset_var)
```

✓ 0.2s

```
Within-cluster variance: 124525688303.2892
Between-cluster variance: 3237705104.4185643
Dataset-cluster variance: 1.4710774868449526e+22
```

- **Variance Within Clusters** mengukur sejauh mana titik data dalam suatu klaster tersebar. Semakin rendah nilai ini, semakin rapat titik data dalam klaster tersebut. Dalam output yang diberikan, nilai Total Variance Within Clusters adalah 124525688303.2892, yang menunjukkan bahwa titik data dalam klaster tersebut cukup rapat.
- **Variance Between Clusters** mengukur sejauh mana klaster-klaster tersebut terpisah satu sama lain. Semakin tinggi nilai ini, semakin terpisah klaster-klaster tersebut. Dalam output yang diberikan, nilai Total Variance Between Clusters adalah 3237705104.4185643, yang menunjukkan bahwa klaster-klaster tersebut cukup terpisah.
- **Total Variance (Dataset Variance)** adalah jumlah dari varians antar klaster dan varians dalam klaster. Nilai ini mencerminkan variasi keseluruhan dalam dataset. Dalam output yang diberikan, nilai Total Variance (Dataset Variance) adalah 1.4710774868449526, yang mencerminkan variasi keseluruhan dalam dataset.

Berdasarkan penjelasan di atas, dapat disimpulkan bahwa menggunakan Metode Average Linkage dapat diketahui klaster-klaster dalam dataset tersebut cukup terpisah satu sama lain, dengan titik data yang cukup rapat dalam masing-masing klaster. Namun, variasi keseluruhan dalam dataset masih cukup tinggi, menunjukkan adanya variasi yang signifikan antara klaster-klaster.

Praktikum Cluster Analysis

```
# assignment 5 single
from sklearn.cluster import AgglomerativeClustering

data = df.loc[:,['ph', 'Hardness', 'Solids', 'Chloramines', 'Sulfate', 'Conductivity',
                 'Organic_carbon', 'Trihalomethanes', 'Turbidity']]

clustering=AgglomerativeClustering(n_clusters=2, linkage='single')
clusters=clustering.fit_predict(data)

# Mendapatkan label cluster untuk setiap titik data
labels = clustering.labels_

# Mendapatkan pusat cluster
cluster_centers = np.array([X[labels == i].mean(axis=0) for i in range(2)])

# Menghitung varian dalam kelompok
within_cluster_var = np.mean([np.linalg.norm(X[labels == i] - cluster_centers[i]) ** 2 for i in range(2)])

# Menghitung varian antara kelompok
between_cluster_var = np.mean([np.linalg.norm(cluster_centers[i] - kmeans.cluster_centers_) ** 2 for i in range(2)])

# Menghitung varian dataset
dataset_var = np.mean([np.linalg.norm(between_cluster_var - within_cluster_var) ** 2 ])

print("Within-cluster variance:", within_cluster_var)
print("Between-cluster variance:", between_cluster_var)
print("Dataset-cluster variance:", dataset_var)
```

✓ 0.0s

```
Within-cluster variance: 125324411585.64302
Between-cluster variance: 3675763938.9909005
Dataset-cluster variance: 1.479839347425932e+22
```

- **Variance Within Clusters** mengukur sejauh mana titik data dalam suatu klaster tersebar. Semakin rendah nilai ini, semakin rapat titik data dalam klaster tersebut. Dalam output yang diberikan, nilai Total Variance Within Clusters adalah 125324411585.64302, yang menunjukkan bahwa titik data dalam klaster tersebut cukup rapat.
- **Variance Between Clusters** mengukur sejauh mana klaster-klaster tersebut terpisah satu sama lain. Semakin tinggi nilai ini, semakin terpisah klaster-klaster tersebut. Dalam output yang diberikan, nilai Total Variance Between Clusters adalah 3675763938.9909005, yang menunjukkan bahwa klaster-klaster tersebut cukup terpisah.
- **Total Variance (Dataset Variance)** adalah jumlah dari varians antar klaster dan varians dalam klaster. Nilai ini mencerminkan variasi keseluruhan dalam dataset. Dalam output yang diberikan, nilai Total Variance (Dataset Variance) adalah 1.479839347425932e, yang mencerminkan variasi keseluruhan dalam dataset.

Berdasarkan penjelasan di atas, dapat disimpulkan bahwa menggunakan Metode Average Linkage dapat diketahui klaster-klaster dalam dataset tersebut cukup terpisah satu sama lain, dengan titik data yang cukup rapat dalam masing-masing klaster. Namun, variasi keseluruhan dalam dataset masih cukup tinggi, menunjukkan adanya variasi yang signifikan antara klaster-klaster.

Praktikum Cluster Analysis

```python
# assignment 5 sentroik
from sklearn.cluster import AgglomerativeClustering

data = df.loc[:,['ph', 'Hardness', 'Solids', 'Chloramines', 'Sulfate', 'Conductivity',
                 'Organic_carbon', 'Trihalomethanes', 'Turbidity']]

clustering=AgglomerativeClustering(n_clusters=2, linkage='ward')
clusters=clustering.fit_predict(data)

# Mendapatkan label cluster untuk setiap titik data
labels = clustering.labels_

# Mendapatkan pusat cluster
cluster_centers = np.array([X[labels == i].mean(axis=0) for i in range(2)])

# Menghitung varian dalam kelompok
within_cluster_var = np.mean([np.linalg.norm(X[labels == i] - cluster_centers[i]) ** 2 for i in range(2)])

# Menghitung varian antara kelompok
between_cluster_var = np.mean([np.linalg.norm(cluster_centers[i] - kmeans.cluster_centers_) ** 2 for i in range(2)])

# Menghitung varian dataset
dataset_var = np.mean([np.linalg.norm(between_cluster_var - within_cluster_var) ** 2 ])

print("Within-cluster variance:", within_cluster_var)
print("Between-cluster variance:", between_cluster_var)
print("Dataset-cluster variance:", dataset_var)
```

✓ 0.2s

```
Within-cluster variance: 121633226278.23341
Between-cluster variance: 935077241.3315233
Dataset-cluster variance: 1.456804318093418e+22
```

- **Variance Within Clusters** mengukur sejauh mana titik data dalam suatu klaster tersebar. Semakin rendah nilai ini, semakin rapat titik data dalam klaster tersebut. Dalam output yang diberikan, nilai Total Variance Within Clusters adalah 121633226278.23341, yang menunjukkan bahwa titik data dalam klaster tersebut cukup rapat.
- **Variance Between Clusters** mengukur sejauh mana klaster-klaster tersebut terpisah satu sama lain. Semakin tinggi nilai ini, semakin terpisah klaster-klaster tersebut. Dalam output yang diberikan, nilai Total Variance Between Clusters adalah 935077241.3315233, yang menunjukkan bahwa klaster-klaster tersebut cukup terpisah.
- **Total Variance (Dataset Variance)** adalah jumlah dari varians antar klaster dan varians dalam klaster. Nilai ini mencerminkan variasi keseluruhan dalam dataset. Dalam output yang diberikan, nilai Total Variance (Dataset Variance) adalah 1.456804318093418, yang mencerminkan variasi keseluruhan dalam dataset.

Berdasarkan penjelasan di atas, dapat disimpulkan bahwa menggunakan Metode Average Linkage dapat diketahui klaster-klaster dalam dataset tersebut cukup terpisah satu sama lain, dengan titik data yang cukup rapat dalam masing-masing klaster. Namun, variasi keseluruhan dalam dataset masih cukup tinggi, menunjukkan adanya variasi yang signifikan antara klaster-klaster.

Praktikum Cluster Analysis

```python
# assignment 5 complete
from sklearn.cluster import AgglomerativeClustering

data = df.loc[:,['ph', 'Hardness', 'Solids', 'Chloramines', 'Sulfate', 'Conductivity',
                 'Organic_carbon', 'Trihalomethanes', 'Turbidity']]

clustering=AgglomerativeClustering(n_clusters=2, linkage='complete')
clusters=clustering.fit_predict(data)

# Mendapatkan label cluster untuk setiap titik data
labels = clustering.labels_

# Mendapatkan pusat cluster
cluster_centers = np.array([X[labels == i].mean(axis=0) for i in range(2)])

# Menghitung varian dalam kelompok
within_cluster_var = np.mean([np.linalg.norm(X[labels == i] - cluster_centers[i]) ** 2 for i in range(2)])

# Menghitung varian antara kelompok
between_cluster_var = np.mean([np.linalg.norm(cluster_centers[i] - kmeans.cluster_centers_) ** 2 for i in range(2)])

# Menghitung varian dataset
dataset_var = np.mean([np.linalg.norm(between_cluster_var - within_cluster_var) ** 2 ])

print("Within-cluster variance:", within_cluster_var)
print("Between-cluster variance:", between_cluster_var)
print("Dataset-cluster variance:", dataset_var)
```

✓ 0.2s

```
Within-cluster variance: 116615396615.45006
Between-cluster variance: 1268679583.74584
Dataset-cluster variance: 1.3304865129992045e+22
```

- **Variance Within Clusters** mengukur sejauh mana titik data dalam suatu klaster tersebar. Semakin rendah nilai ini, semakin rapat titik data dalam klaster tersebut. Dalam output yang diberikan, nilai Total Variance Within Clusters adalah 116615396615.45006, yang menunjukkan bahwa titik data dalam klaster tersebut cukup rapat.
- **Variance Between Clusters** mengukur sejauh mana klaster-klaster tersebut terpisah satu sama lain. Semakin tinggi nilai ini, semakin terpisah klaster-klaster tersebut. Dalam output yang diberikan, nilai Total Variance Between Clusters adalah 1268679583.74584, yang menunjukkan bahwa klaster-klaster tersebut cukup terpisah.
- **Total Variance (Dataset Variance)** adalah jumlah dari varians antar klaster dan varians dalam klaster. Nilai ini mencerminkan variasi keseluruhan dalam dataset. Dalam output yang diberikan, nilai Total Variance (Dataset Variance) adalah 1.3304865129992045, yang mencerminkan variasi keseluruhan dalam dataset.

Berdasarkan penjelasan di atas, dapat disimpulkan bahwa menggunakan Metode Average Linkage dapat diketahui klaster-klaster dalam dataset tersebut cukup terpisah satu sama lain, dengan titik data yang cukup rapat dalam masing-masing klaster. Namun, variasi keseluruhan dalam dataset masih cukup tinggi, menunjukkan adanya variasi yang signifikan antara klaster-klaster.

CLUSTERING RUSPINI.CSV

Praktikum Cluster Analysis

```python
# assignment 1
import pandas as pd
dataset = pd.read_csv("C:/Users/bayuk/OneDrive/Documents/AI/pens/smtr3/Machine Learning/Data/ruspini.csv")
```

✓ 0.0s

```python
# assignment 2
dataset = dataset.fillna(dataset.groupby("CLASS").transform("mean"))
print("\n Dataset setelah pengisian mising value\n", dataset)
```

✓ 0.0s

```
 Dataset setelah pengisian mising value
      #   X   Y  CLASS
0     1   4  53      1
1     2   5  63      1
2     3  10  59      1
3     4   9  77      1
4     5  13  49      1
..   ..  ..  ..    ...
70   71  66  23      4
71   72  61  25      4
72   73  76  27      4
73   74  72  31      4
74   75  64  30      4

[75 rows x 4 columns]
```

Praktikum Cluster Analysis

```
from sklearn.preprocessing import StandardScaler

# Melakukan normalisasi data
scaler = StandardScaler()
data_norm = scaler.fit_transform(X)

# Membuat dataframe pandas dari data ternormalisasi
df = pd.DataFrame(data_norm, columns=['X', 'Y'])  # Sesuaikan jumlah kolom dengan data Anda
df['CLASS'] = y

data = df.loc[:,['X','Y']]

kmeans = KMeans(n_clusters=4, init="random", n_init=1)
clusters=kmeans.fit_predict(data)

# Mendapatkan label cluster untuk setiap titik data
labels = kmeans.labels_

# Mendapatkan pusat cluster
cluster_centers = kmeans.cluster_centers_

# Menghitung varian dalam kelompok
within_cluster_var = np.mean([np.linalg.norm(X[labels == i] - cluster_centers[i]) ** 2 for i in range(2)])

# Menghitung varian antara kelompok
between_cluster_var = np.mean([np.linalg.norm(cluster_centers[i] - kmeans.cluster_centers_) ** 2 for i in range(2)])

# Menghitung varian dataset
dataset_var = np.mean([np.linalg.norm(between_cluster_var - within_cluster_var) ** 2 ])

print("Within-cluster variance:", within_cluster_var)
print("Between-cluster variance:", between_cluster_var)
print("Dataset-cluster variance:", dataset_var)
```

✓ 0.0s

```
Within-cluster variance: 418741.8460068892
Between-cluster variance: 11.050672409387879
Dataset-cluster variance: 175335478961.44598
```

- **Variance Within Clusters**: 418741.8460068892=> Variance Within Clusters mengukur sejauh mana titik data dalam suatu klaster tersebar. Semakin rendah nilai ini, semakin rapat titik data dalam klaster tersebut.
- **Variance Between Clusters**: 11.050672409387879=> Variance Between Clusters mengukur sejauh mana klaster-klaster tersebut terpisah satu sama lain. Semakin tinggi nilai ini semakin terpisah klaster-klaster tersebut.
- **Total Variance (Dataset Variance)**: 175335478961.44598=> Total Variance atau varian dataset adalah jumlah dari varians antar klaster dan varians dalam klaster. Nilai ini mencerminkan variasi keseluruhan dalam dataset.

Dalam konteks clustering menggunakan K-Means pada dataset ruspini.csv, tujuan utama adalah untuk menciptakan klaster-klaster yang saling terpisah dengan variasi dalam klaster yang rendah. Dengan kata lain, kita ingin mencapai nilai Variance Between Clusters yang tinggi dan nilai Variance Within Clusters yang rendah. Dalam output tersebut, Variance Between Clusters adalah 418741.8460068892, yang menunjukkan bahwa klaster-klaster tersebut cukup terpisah Variance Within Clusters adalah 11.050672409387879, yang menunjukkan bahwa titik data dalam cluster tersebut cukup rapat. Total Variance (Dataset Variance) adalah 175335478961.44598, yang mencerminkan variasi keseluruhan dalam dataset.

Praktikum Cluster Analysis

```
# assignment 5 average
from sklearn.cluster import AgglomerativeClustering

data = df.loc[:,['X','Y']]

clustering=AgglomerativeClustering(n_clusters=4, linkage='average')
clusters=clustering.fit_predict(data)

# Mendapatkan label cluster untuk setiap titik data
labels = clustering.labels_

# Mendapatkan pusat cluster
cluster_centers = np.array([X[labels == i].mean(axis=0) for i in range(4)])

# Menghitung varian dalam kelompok
within_cluster_var = np.mean([np.linalg.norm(X[labels == i] - cluster_centers[i]) ** 2 for i in range(4)])

# Menghitung varian antara kelompok
between_cluster_var = np.mean([np.linalg.norm(cluster_centers[i] - kmeans.cluster_centers_) ** 2 for i in range(4)])

# Menghitung varian dataset
dataset_var = np.mean([np.linalg.norm(between_cluster_var - within_cluster_var) ** 2 ])

print("Within-cluster variance:", within_cluster_var)
print("Between-cluster variance:", between_cluster_var)
print("Dataset-cluster variance:", dataset_var)
```

✓ 0.0s

```
Within-cluster variance: 3220.262809036658
Between-cluster variance: 55535.81079604798
Dataset-cluster variance: 2736916561.181285
```

- **Variance Within Clusters** mengukur sejauh mana titik data dalam suatu klaster tersebar. Semakin rendah nilai ini, semakin rapat titik data dalam klaster tersebut. Dalam output yang diberikan, nilai Total Variance Within Clusters adalah 3220.262809036658, yang menunjukkan bahwa titik data dalam klaster tersebut cukup rapat.
- **Variance Between Clusters** mengukur sejauh mana klaster-klaster tersebut terpisah satu sama lain. Semakin tinggi nilai ini, semakin terpisah klaster-klaster tersebut. Dalam output yang diberikan, nilai Total Variance Between Clusters adalah 55535.81079604798, yang menunjukkan bahwa klaster-klaster tersebut cukup terpisah.
- **Total Variance (Dataset Variance)** adalah jumlah dari varians antar klaster dan varians dalam klaster. Nilai ini mencerminkan variasi keseluruhan dalam dataset. Dalam output yang diberikan, nilai Total Variance (Dataset Variance) adalah 2736916561.181285, yang mencerminkan variasi keseluruhan dalam dataset.

Berdasarkan penjelasan di atas, dapat disimpulkan bahwa menggunakan Metode Average Linkage dapat diketahui klaster-klaster dalam dataset tersebut cukup terpisah satu sama lain, dengan titik data yang cukup rapat dalam masing-masing klaster. Namun, variasi keseluruhan dalam dataset masih cukup tinggi, menunjukkan adanya variasi yang signifikan antara klaster-klaster.

Praktikum Cluster Analysis

```python
# assignment 5 single
from sklearn.cluster import AgglomerativeClustering

data = df.loc[:,['X','Y']]

clustering=AgglomerativeClustering(n_clusters=4, linkage='single')
clusters=clustering.fit_predict(data)

# Mendapatkan label cluster untuk setiap titik data
labels = clustering.labels_

# Mendapatkan pusat cluster
cluster_centers = np.array([X[labels == i].mean(axis=0) for i in range(4)])

# Menghitung varian dalam kelompok
within_cluster_var = np.mean([np.linalg.norm(X[labels == i] - cluster_centers[i]) ** 2 for i in range(4)])

# Menghitung varian antara kelompok
between_cluster_var = np.mean([np.linalg.norm(cluster_centers[i] - kmeans.cluster_centers_) ** 2 for i in range(4)])

# Menghitung varian dataset
dataset_var = np.mean([np.linalg.norm(between_cluster_var - within_cluster_var) ** 2 ])

print("Within-cluster variance:", within_cluster_var)
print("Between-cluster variance:", between_cluster_var)
print("Dataset-cluster variance:", dataset_var)
```

✓ 0.0s

```
Within-cluster variance: 3220.2628090366584
Between-cluster variance: 55535.81079604798
Dataset-cluster variance: 2736916561.181285
```

- **Variance Within Clusters** mengukur sejauh mana titik data dalam suatu klaster tersebar. Semakin rendah nilai ini, semakin rapat titik data dalam klaster tersebut. Dalam output yang diberikan, nilai Total Variance Within Clusters adalah 3220.262809036658, yang menunjukkan bahwa titik data dalam klaster tersebut cukup rapat.
- **Variance Between Clusters** mengukur sejauh mana klaster-klaster tersebut terpisah satu sama lain. Semakin tinggi nilai ini, semakin terpisah klaster-klaster tersebut. Dalam output yang diberikan, nilai Total Variance Between Clusters adalah 55535.81079604798, yang menunjukkan bahwa klaster-klaster tersebut cukup terpisah.
- **Total Variance (Dataset Variance)** adalah jumlah dari varians antar klaster dan varians dalam klaster. Nilai ini mencerminkan variasi keseluruhan dalam dataset. Dalam output yang diberikan, nilai Total Variance (Dataset Variance) adalah 2736916561.181285, yang mencerminkan variasi keseluruhan dalam dataset.

Berdasarkan penjelasan di atas, dapat disimpulkan bahwa menggunakan Metode Average Linkage dapat diketahui klaster-klaster dalam dataset tersebut cukup terpisah satu sama lain, dengan titik data yang cukup rapat dalam masing-masing klaster. Namun, variasi keseluruhan dalam dataset masih cukup tinggi, menunjukkan adanya variasi yang signifikan antara klaster-klaster.

Praktikum Cluster Analysis

```python
# assignment 5 sentroik
from sklearn.cluster import AgglomerativeClustering

data = df.loc[:,['X','Y']]

clustering=AgglomerativeClustering(n_clusters=4, linkage='ward')
clusters=clustering.fit_predict(data)

# Mendapatkan label cluster untuk setiap titik data
labels = clustering.labels_

# Mendapatkan pusat cluster
cluster_centers = np.array([X[labels == i].mean(axis=0) for i in range(4)])

# Menghitung varian dalam kelompok
within_cluster_var = np.mean([np.linalg.norm(X[labels == i] - cluster_centers[i]) ** 2 for i in range(4)])

# Menghitung varian antara kelompok
between_cluster_var = np.mean([np.linalg.norm(cluster_centers[i] - kmeans.cluster_centers_) ** 2 for i in range(4)])

# Menghitung varian dataset
dataset_var = np.mean([np.linalg.norm(between_cluster_var - within_cluster_var) ** 2 ])

print("Within-cluster variance:", within_cluster_var)
print("Between-cluster variance:", between_cluster_var)
print("Dataset-cluster variance:", dataset_var)
```

✓ 0.0s

```
Within-cluster variance: 3220.262809036658
Between-cluster variance: 55535.81079604798
Dataset-cluster variance: 2736916561.181285
```

- **Variance Within Clusters** mengukur sejauh mana titik data dalam suatu klaster tersebar. Semakin rendah nilai ini, semakin rapat titik data dalam klaster tersebut. Dalam output yang diberikan, nilai Total Variance Within Clusters adalah 3220.2628090366584, yang menunjukkan bahwa titik data dalam klaster tersebut cukup rapat.
- **Variance Between Clusters** mengukur sejauh mana klaster-klaster tersebut terpisah satu sama lain. Semakin tinggi nilai ini, semakin terpisah klaster-klaster tersebut. Dalam output yang diberikan, nilai Total Variance Between Clusters adalah 55535.81079604798, yang menunjukkan bahwa klaster-klaster tersebut cukup terpisah.
- **Total Variance (Dataset Variance)** adalah jumlah dari varians antar klaster dan varians dalam klaster. Nilai ini mencerminkan variasi keseluruhan dalam dataset. Dalam output yang diberikan, nilai Total Variance (Dataset Variance) adalah 2736916561.181285, yang mencerminkan variasi keseluruhan dalam dataset.

Berdasarkan penjelasan di atas, dapat disimpulkan bahwa menggunakan Metode Average Linkage dapat diketahui klaster-klaster dalam dataset tersebut cukup terpisah satu sama lain, dengan titik data yang cukup rapat dalam masing-masing klaster. Namun, variasi keseluruhan dalam dataset masih cukup tinggi, menunjukkan adanya variasi yang signifikan antara klaster-klaster.

Praktikum Cluster Analysis

```python
# assignment 5 complete
from sklearn.cluster import AgglomerativeClustering

data = df.loc[:,['X','Y']]

clustering=AgglomerativeClustering(n_clusters=4, linkage='complete')
clusters=clustering.fit_predict(data)

# Mendapatkan label cluster untuk setiap titik data
labels = clustering.labels_

# Mendapatkan pusat cluster
cluster_centers = np.array([X[labels == i].mean(axis=0) for i in range(4)])

# Menghitung varian dalam kelompok
within_cluster_var = np.mean([np.linalg.norm(X[labels == i] - cluster_centers[i]) ** 2 for i in range(4)])

# Menghitung varian antara kelompok
between_cluster_var = np.mean([np.linalg.norm(cluster_centers[i] - kmeans.cluster_centers_) ** 2 for i in range(4)])

# Menghitung varian dataset
dataset_var = np.mean([np.linalg.norm(between_cluster_var - within_cluster_var) ** 2 ])

print("Within-cluster variance:", within_cluster_var)
print("Between-cluster variance:", between_cluster_var)
print("Dataset-cluster variance:", dataset_var)
```

✓ 0.0s

```
Within-cluster variance: 3220.2628090366584
Between-cluster variance: 55535.81079604798
Dataset-cluster variance: 2736916561.181285
```

- **Variance Within Clusters** mengukur sejauh mana titik data dalam suatu klaster tersebar. Semakin rendah nilai ini, semakin rapat titik data dalam klaster tersebut. Dalam output yang diberikan, nilai Total Variance Within Clusters adalah 3220.2628090366584, yang menunjukkan bahwa titik data dalam klaster tersebut cukup rapat.
- **Variance Between Clusters** mengukur sejauh mana klaster-klaster tersebut terpisah satu sama lain. Semakin tinggi nilai ini, semakin terpisah klaster-klaster tersebut. Dalam output yang diberikan, nilai Total Variance Between Clusters adalah 55535.81079604798, yang menunjukkan bahwa klaster-klaster tersebut cukup terpisah.
- **Total Variance (Dataset Variance)** adalah jumlah dari varians antar klaster dan varians dalam klaster. Nilai ini mencerminkan variasi keseluruhan dalam dataset. Dalam output yang diberikan, nilai Total Variance (Dataset Variance) adalah 2736916561.181285, yang mencerminkan variasi keseluruhan dalam dataset.

Berdasarkan penjelasan di atas, dapat disimpulkan bahwa menggunakan Metode Average Linkage dapat diketahui klaster-klaster dalam dataset tersebut cukup terpisah satu sama lain, dengan titik data yang cukup rapat dalam masing-masing klaster. Namun, variasi keseluruhan dalam dataset masih cukup tinggi, menunjukkan adanya variasi yang signifikan antara klaster-klaster.

THANK YOU