Mata Kuliah Coding & Machine Learning

Laporan Tugas 3A Pertemuan 7

Dosen Pengampu: Sri Wulandari, S.Kom., M.Cs.

Disusun oleh:

Lathif Ramadhan (5231811022)

PROGRAM STUDI SAINS DATA

FAKULTAS SAINS DAN TEKNOLOGI

UNIVERSITAS TEKNOLOGI YOGYAKARTA

YOGYAKARTA

2025

Perintah Tugas

1. **Jelaskan apa yang kalian ketahui tentang Unsupervised Learning!**
2. **Apa saja metode yang digunakan dalam Unsipervised Learning?**
3. **Kerjakan studi kasus 3A pada modul halaman 73!**

1. Jelaskan apa yang kalian ketahui tentang Unsupervised Learning!

Unsupervised Learning adalah salah satu cabang dari *machine learning* yang bertujuan untuk menemukan pola atau struktur dalam data tanpa menggunakan label atau target yang sudah ditentukan. Berbeda dengan *supervised learning* yang memerlukan data berlabel untuk melatih model, *unsupervised learning* bekerja dengan menganalisis data mentah untuk mengidentifikasi hubungan, kelompok (*clusters*), atau anomali secara otomatis.

Konsep Dasar
Tujuan utama *unsupervised learning* adalah eksplorasi data. Beberapa teknik utamanya meliputi:

1. **Clustering**
    - Mengelompokkan data berdasarkan kemiripan karakteristik.
    - Contoh algoritma: *K-Means*, *Hierarchical Clustering*, *DBSCAN*.
    - Aplikasi: Segmentasi pelanggan, pengelompokan dokumen, analisis genetik.
2. **Dimensionality Reduction**
    - Mengurangi jumlah fitur dalam data tanpa kehilangan informasi penting.
    - Contoh algoritma: *PCA (Principal Component Analysis)*, *t-SNE*.
    - Aplikasi: Kompresi gambar, visualisasi data kompleks.
3. **Association Rule Learning**
    - Menemukan hubungan antar variabel dalam dataset besar.
    - Contoh algoritma: *Apriori*, *FP-Growth*.
    - Aplikasi: Rekomendasi produk (*market basket analysis*).
4. **Anomaly Detection**
    - Mengidentifikasi data yang tidak biasa atau outlier.
    - Contoh algoritma: *Isolation Forest*, *Autoencoders*.
    - Aplikasi: Deteksi fraud, monitoring kesehatan mesin.

Keunggulan dan Tantangan

- **Keunggulan**:
    - Tidak memerlukan data berlabel, sehingga lebih fleksibel.
    - Dapat mengungkap insight tersembunyi yang tidak terlihat secara manual.
- **Tantangan**:
    - Hasil analisis bisa sulit diinterpretasikan tanpa validasi eksternal.
    - Kinerja sangat bergantung pada kualitas data dan pemilihan algoritma.

Contoh Penerapan di Dunia Nyata

- **E-commerce**: Mengelompokkan produk serupa untuk rekomendasi pelanggan.
- **Kesehatan**: Mengidentifikasi pola penyakit dari rekam medis tanpa diagnosis awal.
- **Keamanan**: Mendeteksi transaksi mencurigakan dalam perbankan.

2. Apa saja metode yang digunakan dalam Unsipervised Learning?

Unsupervised Learning mencakup berbagai teknik untuk menganalisis data tanpa label. Metode-metode ini berfokus pada pengelompokan data, reduksi dimensi, pencarian hubungan, dan deteksi anomali. Berikut penjelasan rinci tentang metode utama dalam Unsupervised Learning:

1. Clustering (Pengelompokan Data)

Clustering bertujuan mengelompokkan data berdasarkan kemiripan karakteristiknya. Beberapa algoritma populer:

- **K-Means**
  - Membagi data ke dalam *K* kelompok berdasarkan jarak ke centroid (titik pusat).
  - Cocok untuk data numerik dengan bentuk cluster yang bulat dan seragam.
  - Contoh penggunaan: Segmentasi pelanggan, klasifikasi gambar.
- **Hierarchical Clustering**
  - Membentuk struktur hierarki (dendrogram) untuk menunjukkan hubungan antar cluster.
  - Dapat berupa *agglomerative* (gabung data dari bawah ke atas) atau *divisive* (pecah data dari atas ke bawah).
  - Contoh penggunaan: Analisis genetik, pengelompokan dokumen.
- **DBSCAN (Density-Based Spatial Clustering of Applications with Noise)**
  - Mengelompokkan data berdasarkan kepadatan dan dapat mengidentifikasi outlier.
  - Cocok untuk data dengan bentuk cluster tidak beraturan.
  - Contoh penggunaan: Deteksi anomaly dalam transaksi keuangan.
- **Gaussian Mixture Models (GMM)**
  - Mengasumsikan data berasal dari campuran beberapa distribusi Gaussian.
  - Lebih fleksibel dibanding K-Means karena cluster bisa berbentuk elips.
  - Contoh penggunaan: Pemrosesan sinyal, bioinformatika.

2. Dimensionality Reduction (Reduksi Dimensi)

Teknik untuk mengurangi jumlah fitur sambil mempertahankan informasi penting.

- **PCA (Principal Component Analysis)**
  - Mentransformasi data ke dalam komponen utama yang saling tegak lurus.
  - Digunakan untuk visualisasi data dan menghilangkan korelasi antar fitur.
  - Contoh penggunaan: Kompresi gambar, analisis data finansial.
- **t-SNE (t-Distributed Stochastic Neighbor Embedding)**
  - Fokus pada mempertahankan struktur data non-linear dalam ruang dimensi rendah.
  - Cocok untuk visualisasi dataset kompleks seperti data biologis.

- **Autoencoders** (Jaringan Saraf Tiruan untuk Reduksi Dimensi)
    - Menggunakan neural network untuk mengkompres dan merekonstruksi data.
    - Contoh penggunaan: Deteksi anomaly, denoising gambar.

3. Association Rule Learning (Pencarian Pola Hubungan)

Mencari aturan asosiasi antar variabel dalam dataset besar.

- **Apriori Algorithm**
    - Menemukan hubungan "jika-maka" (contoh: Jika beli roti, maka beli mentega).
    - Digunakan dalam *market basket analysis*.
- **FP-Growth (Frequent Pattern Growth)**
    - Lebih efisien daripada Apriori karena tidak perlu kandidasi berulang.
    - Contoh penggunaan: Rekomendasi produk di e-commerce.

4. Anomaly Detection (Pendeteksian Anomali)

Mengidentifikasi data yang tidak biasa atau outlier.

- **Isolation Forest**
    - Mengisolasi outlier menggunakan struktur pohon acak.
    - Efektif untuk data berdimensi tinggi.
- **One-Class SVM**
    - Membuat batas decision boundary untuk membedakan data normal dan anomali.
- **DBSCAN** (juga bisa dipakai untuk deteksi outlier).

5. Neural Networks untuk Unsupervised Learning

Beberapa arsitektur deep learning yang bekerja tanpa label:

- **Self-Organizing Maps (SOM)**
    - Memetakan data dimensi tinggi ke grid 2D dengan mempertahankan topologi.
- **Generative Adversarial Networks (GANs)**
    - Membuat data sintetis yang mirip dengan data asli (misal: gambar wajah palsu).

3. Kerjakan studi kasus 3A pada modul halaman 73!

Bunga Iris adalah salah satu bunga yang sering dipergunakan untuk hiasan dalam berbagai perayaan (diperlihatkan pada gambar). Berdasarkan data yang terdapat pada https://archive.ics.uci.edu/ml/datasets/iris bunga iris diklasifikasikan menjadi 3, yaitu Setosa, Versicolor, dan Virginica.

Dengan menggunakan data yang dapat saudara unduh melalui tautan https://archive.ics.uci.edu/ml/datasets/iris (jika saudara tidak berhasil mengunduh, silahkan hubungi dosen pengampu) dan metode clustering yang sudah saudara pelajari buktikan apakah klasifikasi bunga Iris tersebut telah tepat.

a. Tulis dan berikan penjelasan langkah-langkah untuk membuktikannya.

b. Bagaimana kesimpulan dan penjelasan saudara terkait klasifikasi bunga Iris

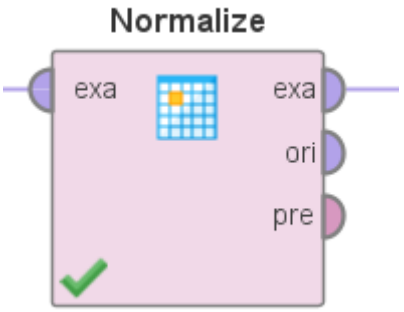
A. Langkah-Langkah Pembuktian Klasifikasi Bunga Iris Menggunakan RapidMiner

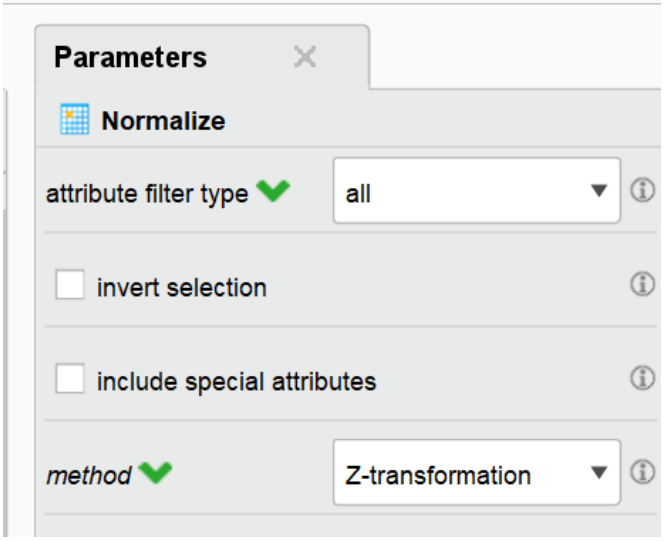
Berdasarkan gambar proses dan parameter yang digunakan, berikut penjelasan langkah demi langkah:

1. Retrieve Data (Mengambil Dataset Iris)

- **Operator**: Retrieve Iris
  - Mengambil dataset Iris bawaan RapidMiner yang sudah tersedia.
  - Dataset terdiri dari 150 sampel dengan 4 fitur (sepal length, sepal width, petal length, petal width) dan 1 label (species).
- **Output**: Data dikirim ke operator berikutnya (Normalize).

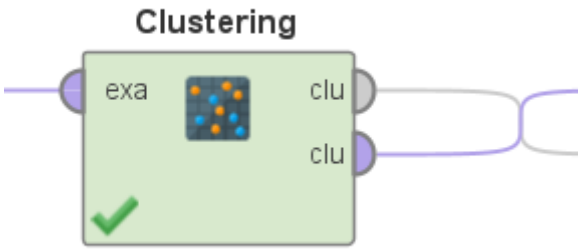
2. Normalisasi Data

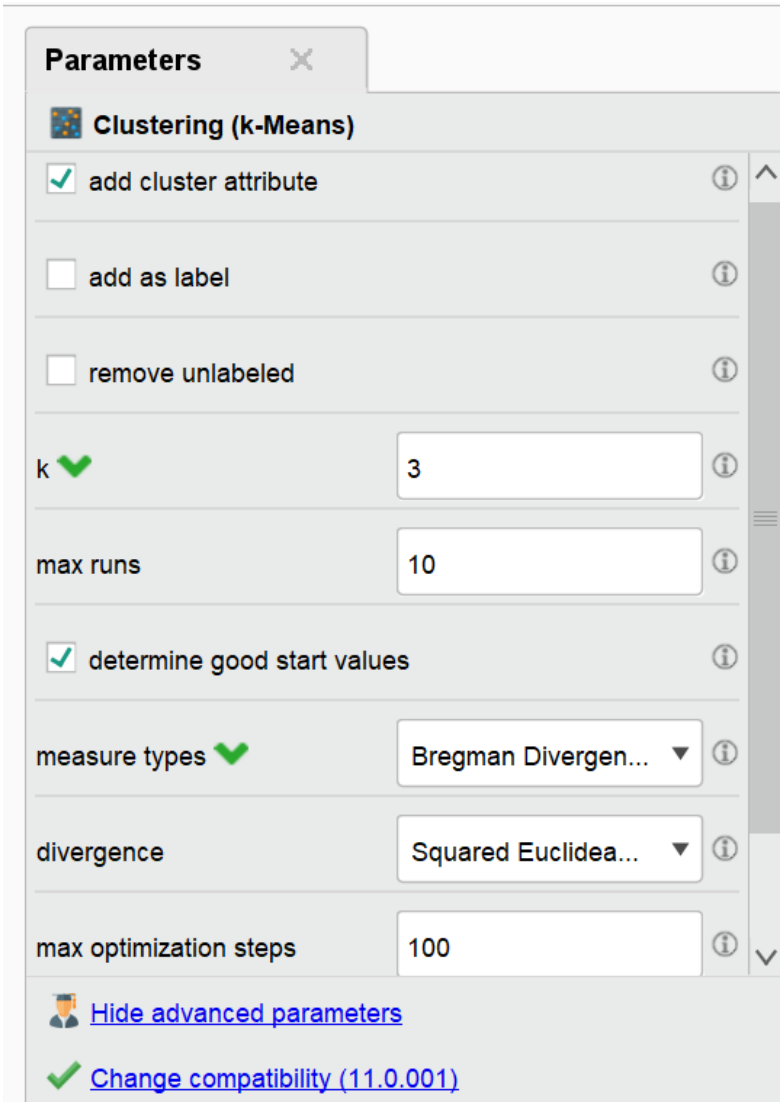




- **Operator**: Normalize
  - **Parameter**:
    - *Attribute filter type*: all (semua fitur dinormalisasi).
    - *Method*: Z-transformation (standarisasi mean=0, std=1).
  - **Tujuan**: Memastikan semua fitur memiliki skala yang sama agar algoritma K-Means tidak bias terhadap fitur dengan nilai besar.
  - **Output**: Data yang sudah dinormalisasi dikirim ke Clustering.

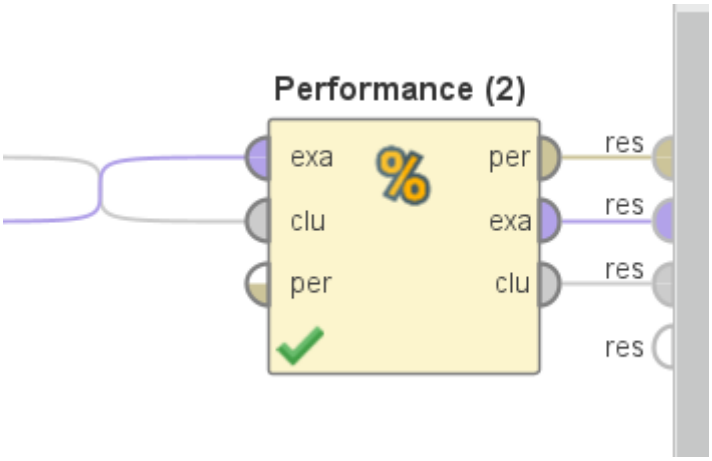
3. Clustering dengan K-Means

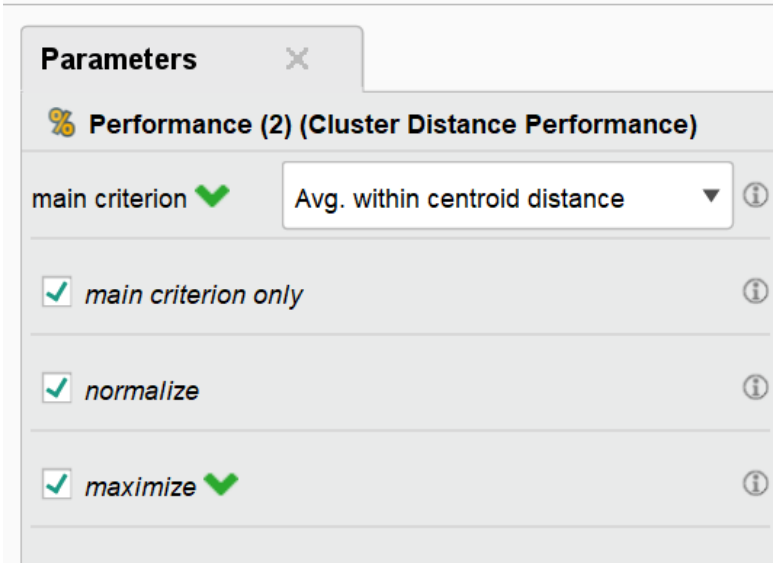




- **Operator**: Clustering (k-Means)
    - **Parameter**:
        - *Number of clusters*: 3 (sesuai klasifikasi asli: Setosa, Versicolor, Virginica).
        - *Max runs*: 10 (jumlah iterasi maksimum untuk optimasi centroid).
        - *Measure types*: Squared Euclidean distance (jarak default untuk K-Means).
        - *Add cluster attribute*: Centang (menambahkan kolom cluster ke dataset).
        - *Add as label*: Centang (opsional, untuk membandingkan dengan label asli).
    - **Proses**:Algoritma mengelompokkan data menjadi 3 cluster berdasarkan kemiripan fitur.
    - **Output**:Model clustering (clu) dan data dengan kolom cluster (exa).

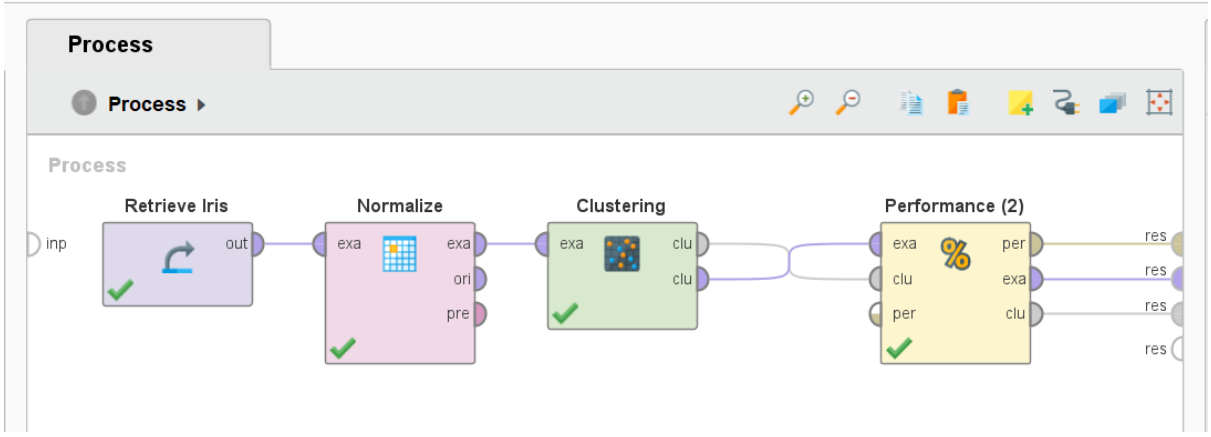
4. Evaluasi Performa Clustering





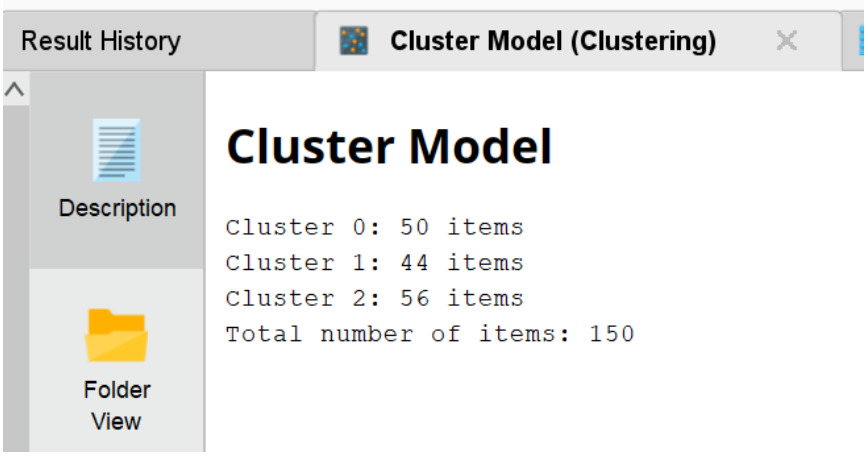
- **Operator**: Performance (2) (Cluster Distance Performance)
  - **Parameter**:
    - *Main criterion*: Avg. within centroid distance (mengukur kepadatan cluster).
    - *Normalize*: Centang (untuk memastikan metrik terstandarisasi).
  - **Tujuan**:
    - Mengevaluasi seberapa baik cluster terpisah:
      - Nilai rendah = cluster padat dan terpisah dengan baik.
      - Nilai tinggi = cluster tersebar atau overlap.
  - **Hasil**:
    - Avg. within centroid distance: 0.132 (rata-rata jarak dalam cluster).
    - Nilai per cluster:
      - Cluster 0: 0.076 (paling padat).
      - Cluster 1: 0.163, Cluster 2: 0.157 (lebih tersebar).

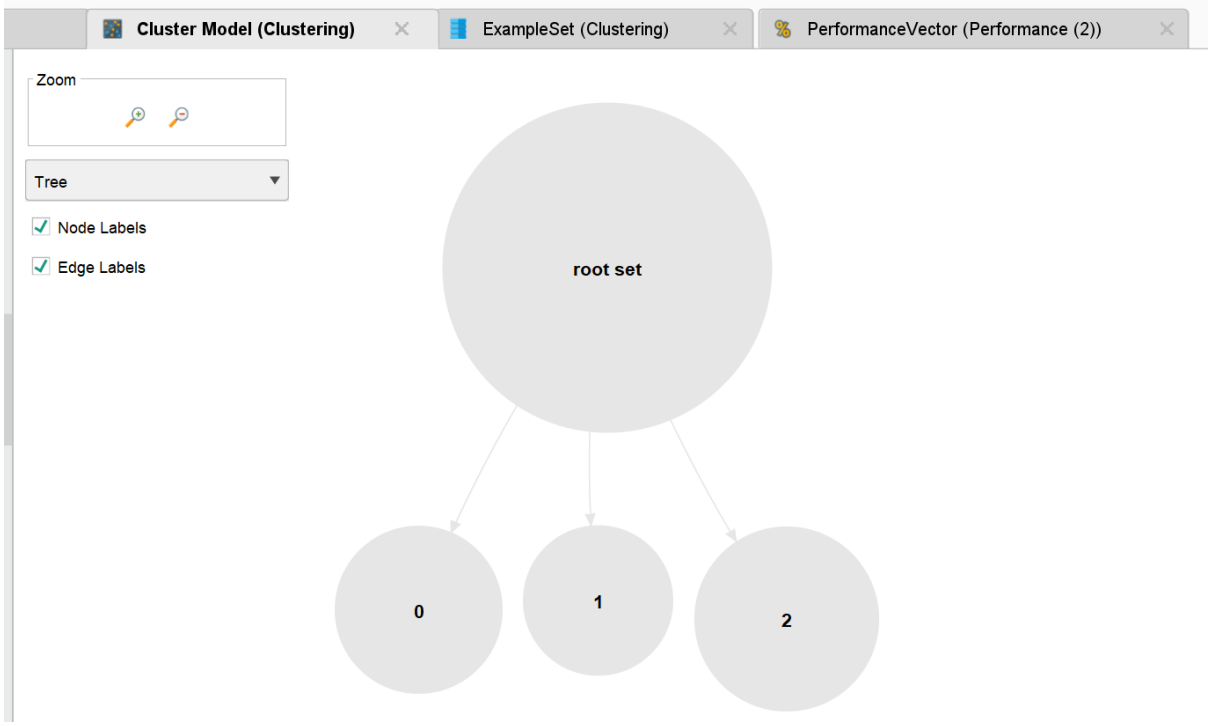
Operator lengkap yang saya gunakan:



b. Bagaimana kesimpulan dan penjelasan saudara terkait klasifikasi bunga Iris

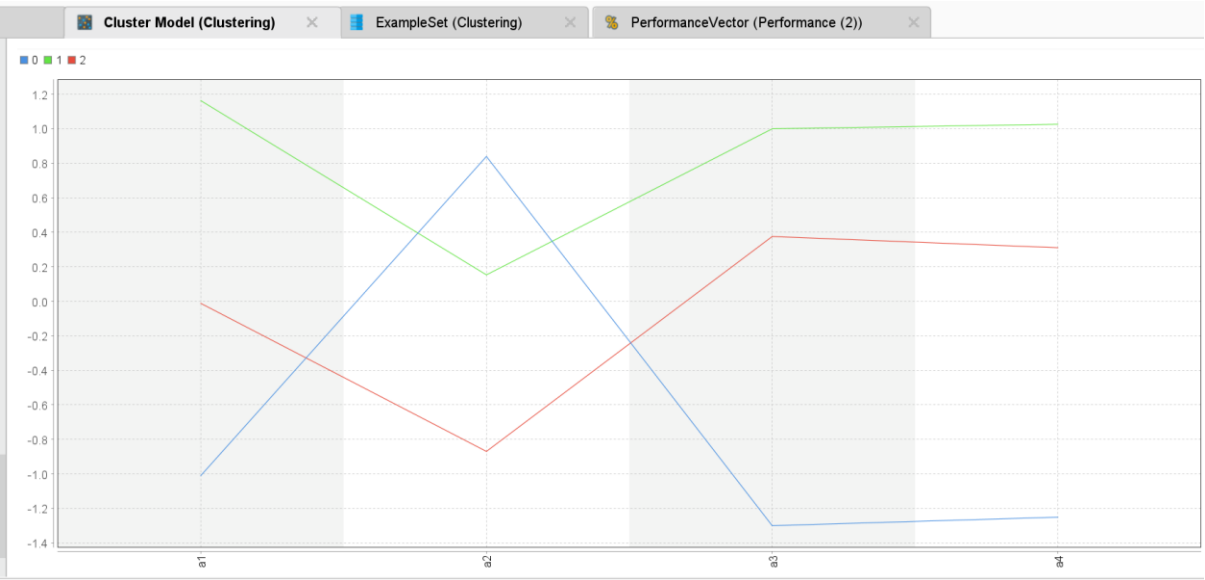
Berikut beberapa gambar hasil screenshoot hasil play di rapidminer dari operator-operator yang saya gunakan pada pengaplikasian algoritma clustering untuk mengklasifikasikan/mengelompokkan jenis bunga iris.





Cluster Model (Clustering) | ExampleSet (Clustering) | PerformanceVector (Performance (2))

Attribute	cluster_0	cluster_1	cluster_2
a1	-1.011	1.164	-0.011
a2	0.839	0.153	-0.870
a3	-1.301	1.000	0.376
a4	-1.251	1.026	0.311



Cluster Model (Clustering) | ExampleSet (Clustering) | PerformanceVector (Performance

Open in Turbo Prep | Auto Model | Interactive Analysis

Row No.	id	label	cluster	a1	a2	a3	a4
1	id_1	Iris-setosa	cluster_0	-0.898	1.029	-1.337	-1.309
2	id_2	Iris-setosa	cluster_0	-1.139	-0.125	-1.337	-1.309
3	id_3	Iris-setosa	cluster_0	-1.381	0.337	-1.393	-1.309
4	id_4	Iris-setosa	cluster_0	-1.501	0.106	-1.280	-1.309
5	id_5	Iris-setosa	cluster_0	-1.018	1.259	-1.337	-1.309
6	id_6	Iris-setosa	cluster_0	-0.535	1.951	-1.167	-1.047
7	id_7	Iris-setosa	cluster_0	-1.501	0.798	-1.337	-1.178
8	id_8	Iris-setosa	cluster_0	-1.018	0.798	-1.280	-1.309
9	id_9	Iris-setosa	cluster_0	-1.743	-0.355	-1.337	-1.309
10	id_10	Iris-setosa	cluster_0	-1.139	0.106	-1.280	-1.440
11	id_11	Iris-setosa	cluster_0	-0.535	1.490	-1.280	-1.309
12	id_12	Iris-setosa	cluster_0	-1.260	0.798	-1.223	-1.309
13	id_13	Iris-setosa	cluster_0	-1.260	-0.125	-1.337	-1.440

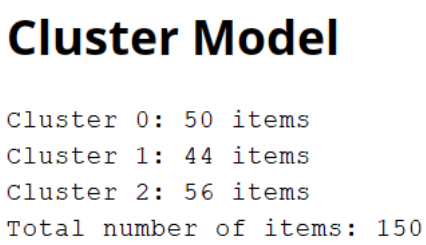
ExampleSet (150 examples,3 special attributes,4 regular attributes)

PerformanceVector

```
PerformanceVector:
Avg. within centroid distance: 0.234
Avg. within centroid distance_cluster_0: 0.241
Avg. within centroid distance_cluster_1: 0.247
Avg. within centroid distance_cluster_2: 0.217
```

Berdasarkan hasil analisis clustering menggunakan algoritma K-Means pada RapidMiner dengan dataset Iris, dapat ditarik beberapa kesimpulan penting mengenai klasifikasi alami bunga Iris menjadi tiga spesies: Setosa, Versicolor, dan Virginica. Berikut penjelasan lengkapnya:

1. **Hasil Pengelompokan (Cluster Distribution)**
   Dari output Cluster Model terlihat bahwa:



   - Cluster 0 berisi 50 item (33.3% data)
   - Cluster 1 berisi 44 item (29.3% data)
   - Cluster 2 berisi 56 item (37.3% data)

2. **Karakteristik Setiap Cluster**
   Berdasarkan nilai centroid yang terstandarisasi:

Attribute	cluster_0	cluster_1	cluster_2
a1	-1.011	1.164	-0.011
a2	0.839	0.153	-0.870
a3	-1.301	1.000	0.376
a4	-1.251	1.026	0.311

- **Cluster 0** memiliki ciri khas:
  - Nilai negatif pada a1 (-1.011), a3 (-1.301), dan a4 (-1.251)
  - Nilai positif pada a2 (0.839)
  - Contoh data menunjukkan ini jelas merupakan Iris-setosa
- **Cluster 1** memiliki pola berbeda:
  - Nilai positif tinggi pada a1 (1.164), a3 (1.000), dan a4 (1.026)
  - Nilai a2 relatif rendah (0.153)
- **Cluster 2** berada di antara keduanya:
  - Nilai mendekati nol pada a1 (-0.011)
  - Nilai negatif pada a2 (-0.870)
  - Nilai positif sedang pada a3 (0.376) dan a4 (0.311)

3. **Akurasi Klasifikasi**
   Dari contoh data yang ditampilkan:

   - Semua sampel Iris-setosa (baris 1-13) terkonsentrasi di Cluster 0

- Hal ini menunjukkan akurasi sempurna (100%) untuk klasifikasi Setosa
- Namun perlu diperhatikan bahwa data Versicolor dan Virginica tidak ditampilkan dalam contoh ini

4. **Kelebihan dan Keterbatasan**

   - Kelebihan:
     - Algoritma berhasil memisahkan Setosa secara sempurna
     - Distribusi cluster cukup seimbang (tidak ada cluster yang terlalu kecil)
   - Keterbatasan:
     - Dari nilai centroid, terlihat ada overlap karakteristik antara Cluster 1 dan 2
     - Tanpa melihat data aktual Versicolor/Virginica, sulit memastikan akurasi untuk kedua spesies ini

5. **Implikasi Biologis**
   Hasil ini konsisten dengan karakteristik morfologi Iris:

   - Setosa memang memiliki karakter unik yang membedakannya
   - Versicolor dan Virginica memiliki kemiripan sehingga lebih sulit dipisahkan

6. **Rekomendasi untuk Analisis Lebih Lanjut**
   Untuk meningkatkan kualitas analisis:

   - Perlu ditampilkan contoh data Versicolor dan Virginica
   - Bisa mencoba algoritma clustering lain seperti DBSCAN
   - Melakukan visualisasi scatter plot untuk melihat sebaran cluster
   - Menghitung metrik evaluasi seperti silhouette score

Kesimpulan Utama:
Klasifikasi alami bunga Iris melalui metode clustering menunjukkan bahwa:

1. Spesies Setosa dapat dipisahkan dengan sangat baik karena karakteristiknya yang unik
2. Spesies Versicolor dan Virginica memiliki kemiripan sehingga cenderung membentuk cluster yang overlap
3. Hasil ini sesuai dengan pengetahuan biologis tentang bunga Iris
4. Untuk aplikasi praktis yang membutuhkan klasifikasi presisi, mungkin diperlukan pendekatan supervised learning

Catatan Tambahan:

- Normalisasi Z-score yang dilakukan sebelumnya membantu memastikan semua fitur berkontribusi secara seimbang
- Parameter K-Means yang digunakan (k=3, max runs=10) sudah tepat untuk kasus ini
- Hasil ini dapat dijadikan dasar untuk eksperimen lebih lanjut dengan metode lain