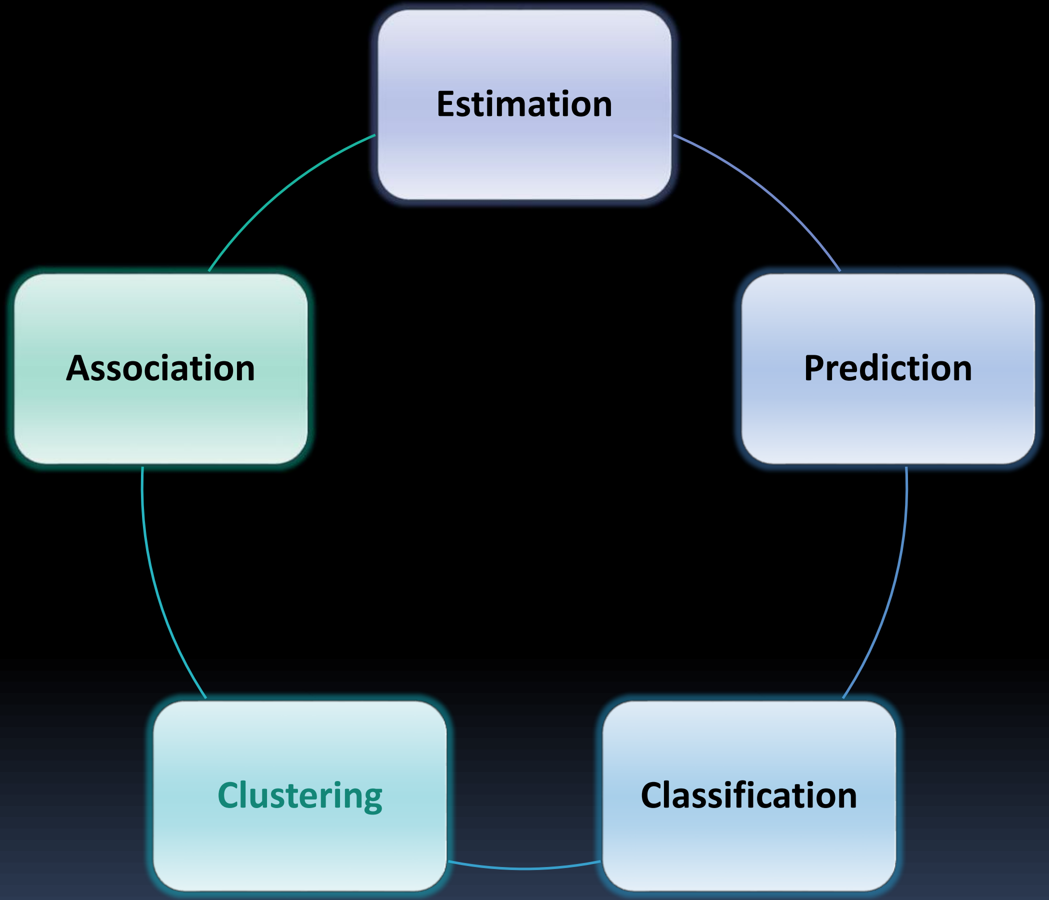
Fajar A. Nugroho, S.Kom, M.CS

CLUSTERING BEST PRACTICE

TEKNIK/METODE Data Mining

1. Estimation
2. Prediction
3. Classification
4. Clustering
5. Association

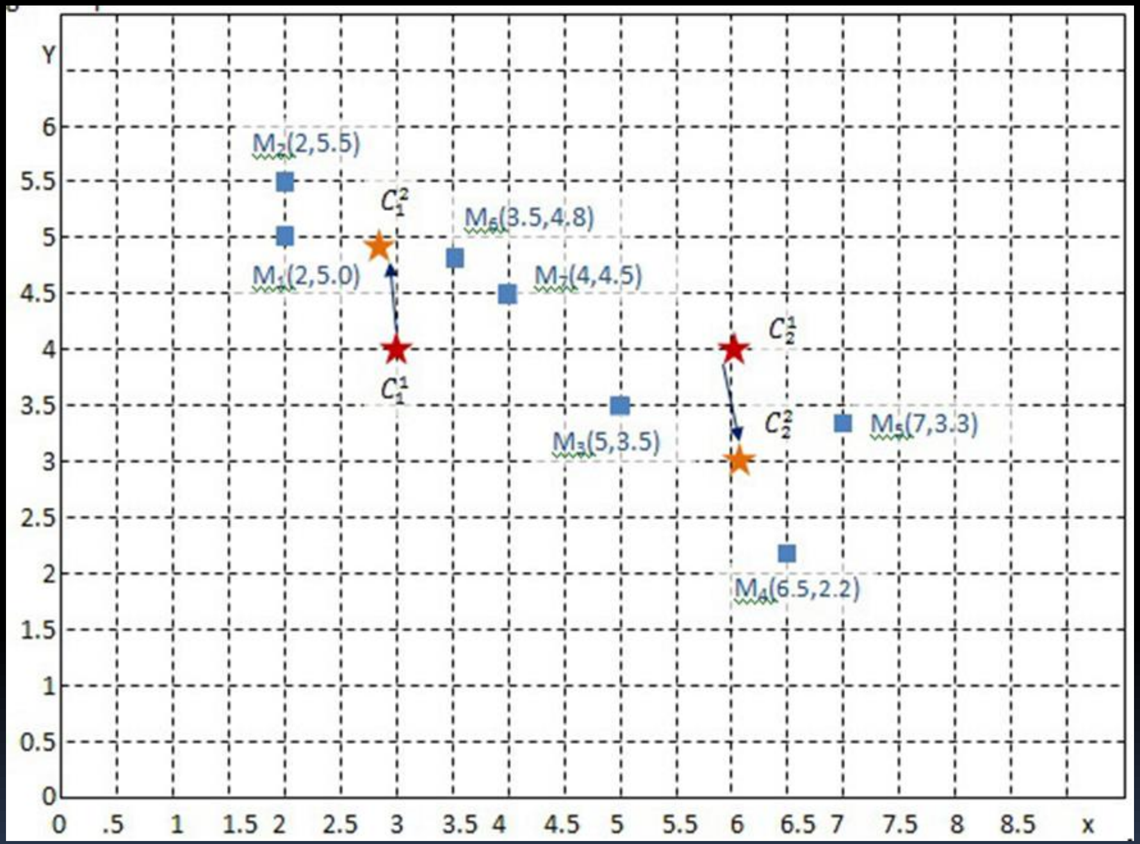


Algoritma Data Mining (DM)

1. **Estimation** (Estimasi):
   - Linear Regression, Neural Network, Support Vector Machine, etc
2. **Prediction/Forecasting** (Prediksi/Peramalan):
   - Linear Regression, Neural Network, Support Vector Machine, etc
3. **Classification** (Klasifikasi):
   - Naive Bayes, K-Nearest Neighbor, C4.5, ID3, CART, Linear Discriminant Analysis, etc
4. **Clustering** (Klastering):
   - K-Means, K-Medoids, Self-Organizing Map (SOM), Fuzzy C-Means, etc
5. **Association** (Asosiasi):
   - FP-Growth, A Priori, etc

Contoh kAsus Algoritma K-Means

- Using K-means algorithm find the best groupings and means of two clusters of the 2D data below. Show all your work, assumptions, and regulations.
- M1 = (2, 5.0),
- M2 = (2, 5.5),
- M3 = (5, 3.5),
- M4 = (6.5, 2.2),
- M5 = (7, 3.3),
- M6 = (3.5, 4.8),
- M7 = (4, 4.5)



Asumsi:
- Semua data akan dikelompokkan ke dalam dua kelas
- Center points of both clusters are $C_1(3,4)$, $C_2(6,4)$

Contoh kAsus (iterasi 1) … LANJ

Iterasi 1

a. Menghitung *Euclidean distance* dari semua data ke tiap titik pusat pertama,

Sehingga didapatkan

D11=1.41, D12=1.80,
D13=2.06, D14=3.94,
D15=4.06, D16=0.94,
D17=1.12,

$$D_{11} = \sqrt{(M_{1x} - C_{1x})^2 + (M_{1y} - C_{1y})^2} = \sqrt{(2-3)^2 + (5-4)^2} = \sqrt{2} = 1.41$$

$$D_{12} = \sqrt{(M_{2x} - C_{1x})^2 + (M_{2y} - C_{1y})^2} = \sqrt{(2-3)^2 + (5.5-4)^2} = \sqrt{3.25} = 1.80$$

$$D_{13} = \sqrt{(M_{3x} - C_{1x})^2 + (M_{3y} - C_{1y})^2} = \sqrt{(5-3)^2 + (3.5-4)^2} = \sqrt{4.25} = 2.06$$

$$D_{14} = \sqrt{(M_{4x} - C_{1x})^2 + (M_{1y} - C_{1y})^2} = \sqrt{(6.5-3)^2 + (2.2-4)^2} = \sqrt{2} = 3.94$$

$$D_{15} = \sqrt{(M_{5x} - C_{1x})^2 + (M_{5y} - C_{1y})^2} = \sqrt{(7-3)^2 + (3.3-4)^2} = \sqrt{2} = 4.06$$

$$D_{16} = \sqrt{(M_{6x} - C_{1x})^2 + (M_{6y} - C_{1y})^2} = \sqrt{(3.5-3)^2 + (4.8-4)^2} = \sqrt{2} = 0.94$$

$$D_{17} = \sqrt{(M_{7x} - C_{1x})^2 + (M_{7y} - C_{1y})^2} = \sqrt{(4-3)^2 + (4.5-4)^2} = \sqrt{2} = 1.12$$

Dengan cara yang sama hitung jarak tiap titik ke titik pusat kedua, dan kita akan mendapatkan :

D21 = 4.12, D22 = 4.27, D23 = 1.18, D24= 1.86,

D25 =1.22, D26 = 2.62, D27 = 2.06

Contoh kAsus (iterasi 1) … LANJ

b. Dari penghitungan *Euclidean distance*, kita dapat membandingkan:

		M1	M2	M3	M4	M5	M6	M7
Jarak ke	C1	**1.41**	**1.80**	2.06	3.94	4.06	**0.94**	**1.12**
	C2	4.12	4.27	**1.18**	**1.86**	**1.22**	2.62	2.06

{M1, M2, M6, M7} anggota C1 and {**M3, M4, M5**} anggota C2

c. Hitung titik pusat baru

M1 = (2, 5.0), M2 = (2, 5.5), M3 = (5, 3.5), M4 = (6.5, 2.2), M5 = (7, 3.3), M6 = (3.5, 4.8), M7 = (4, 4.5)

$$C1 = \left(\frac{2+2+3.5+4}{4}, \frac{5+5.5+4.8+4.5}{4}\right) = (2.85, 4.95)$$

$$C2 = \left(\frac{5+6.5+7}{3}, \frac{3.5+2.2+3.3}{3}\right) = (6.17, 3)$$

Contoh kAsus (iterasi 2) … LANJ

ITERASI 2

a) Hitung Euclidean distance dari tiap data ke titik pusat yang baru Dengan cara yang sama dengan iterasi pertama kita akan mendapatkan perbandingan sebagai berikut:

		M_1	M_2	M_3	M_4	M_5	M_6	M_7
Jarak ke	C_1	**0.76**	**0.96**	2.65	4.62	4.54	**0.76**	**1.31**
	C_2	4.62	4.86	**1.27**	**0.86**	**0.88**	3.22	2.63

b) Dari perbandingan tersebut kira tahu bahwa {M_1, M_2, M_6, M_7} anggota C_1 dan {**M_3, M_4, M_5**} anggota C_2

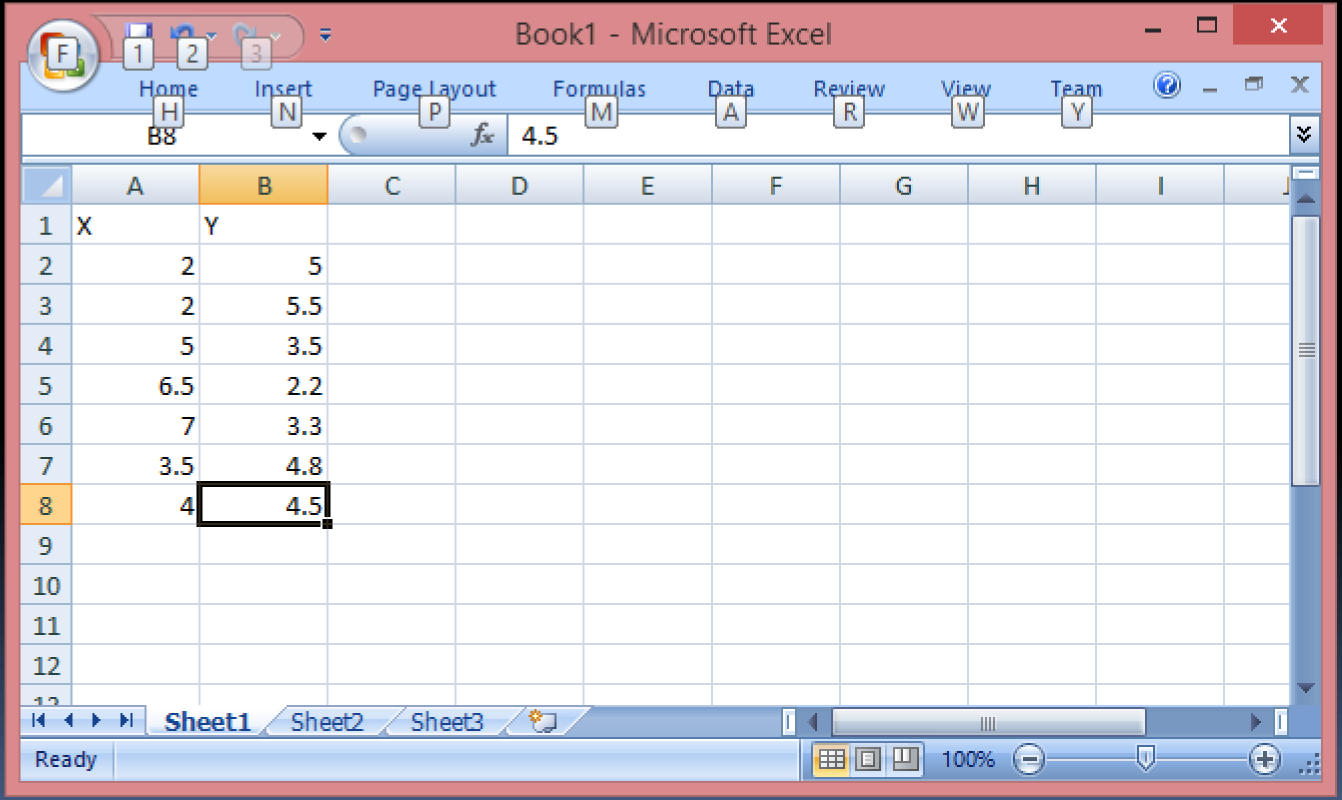
c) Karena anggota kelompok tidak ada yang berubah maka titik pusat pun tidak akan berubah.

KESIMPULAN

{M_1, M_2, M_6, M_7} anggota C_1 dan {**M_3, M_4, M_5**} anggota C_2

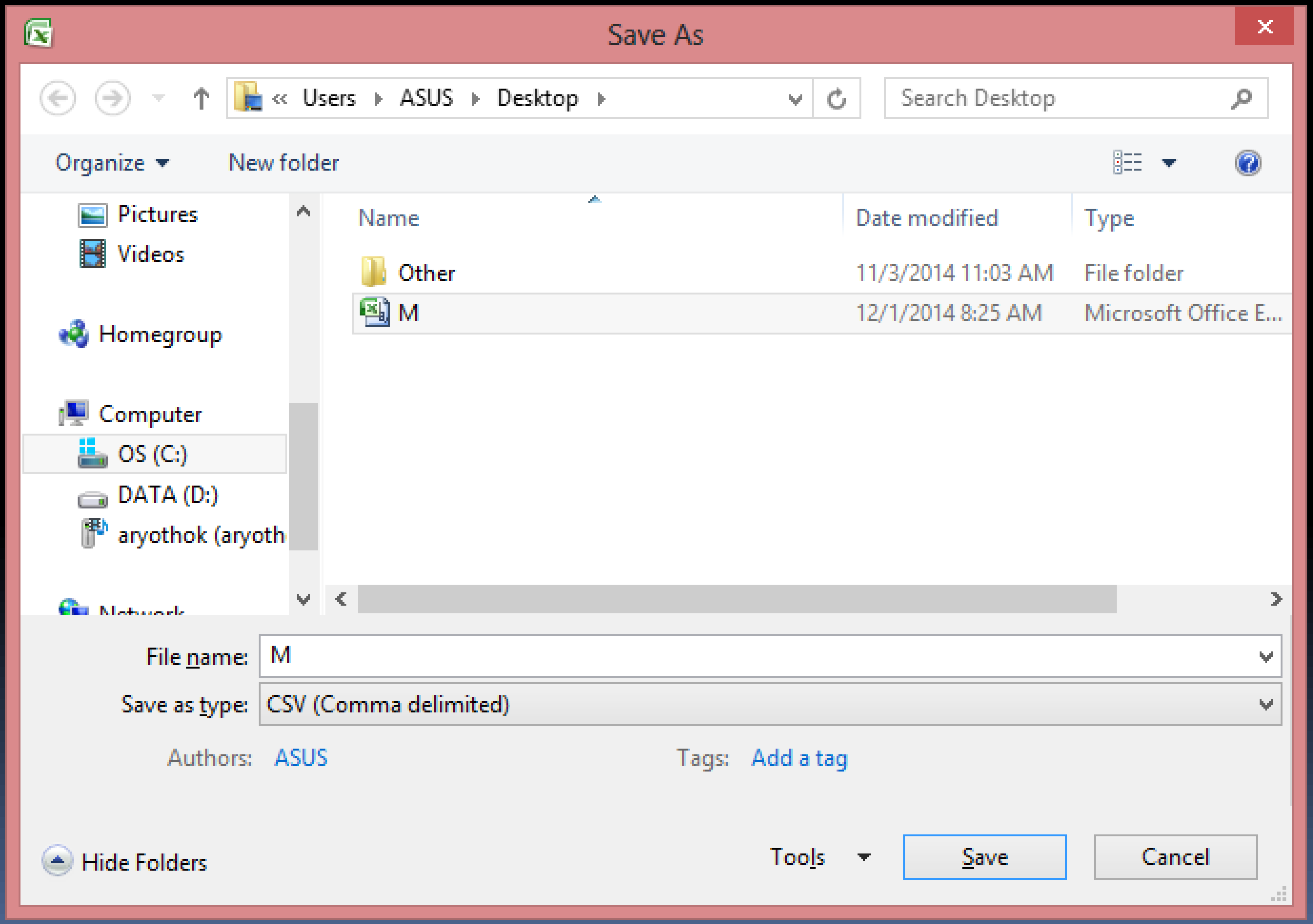
K-Means with WEKA

- Siapkan dataset
- Simpan dgn format CSV



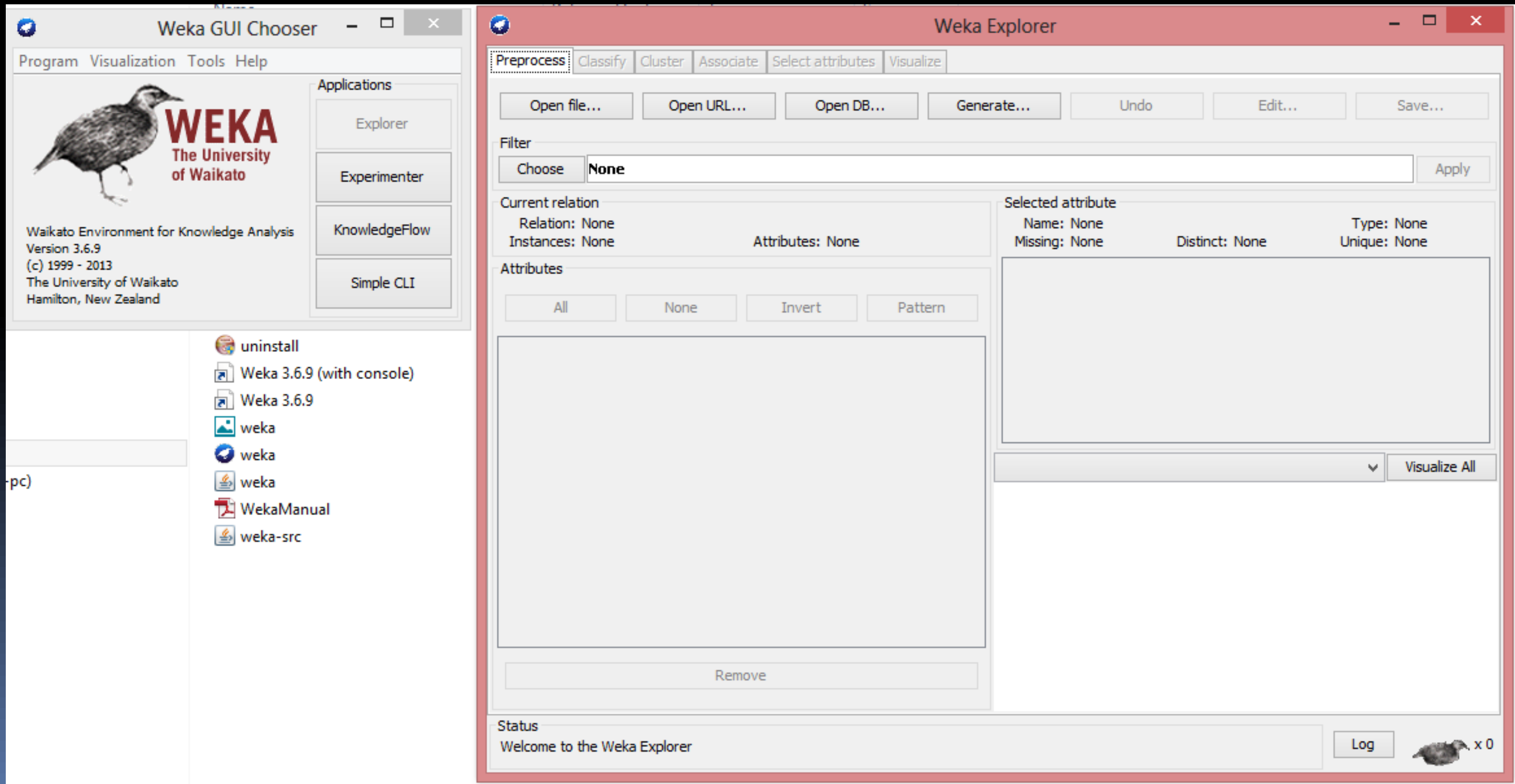
K-Means with WEKA

- Siapkan dataset
- Simpan dgn format CSV



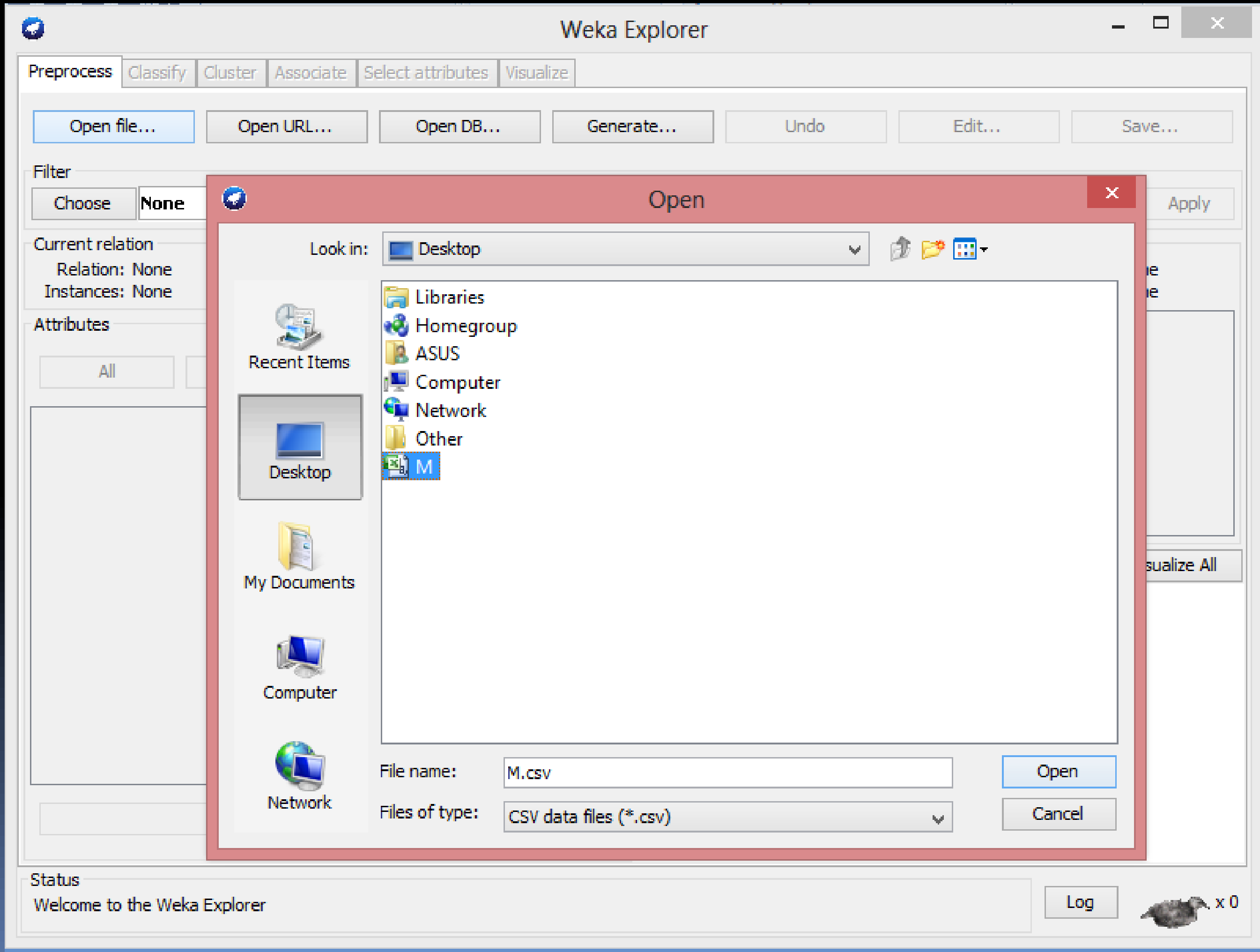
K-Means with WEKA

- Buka WEKA Explorer



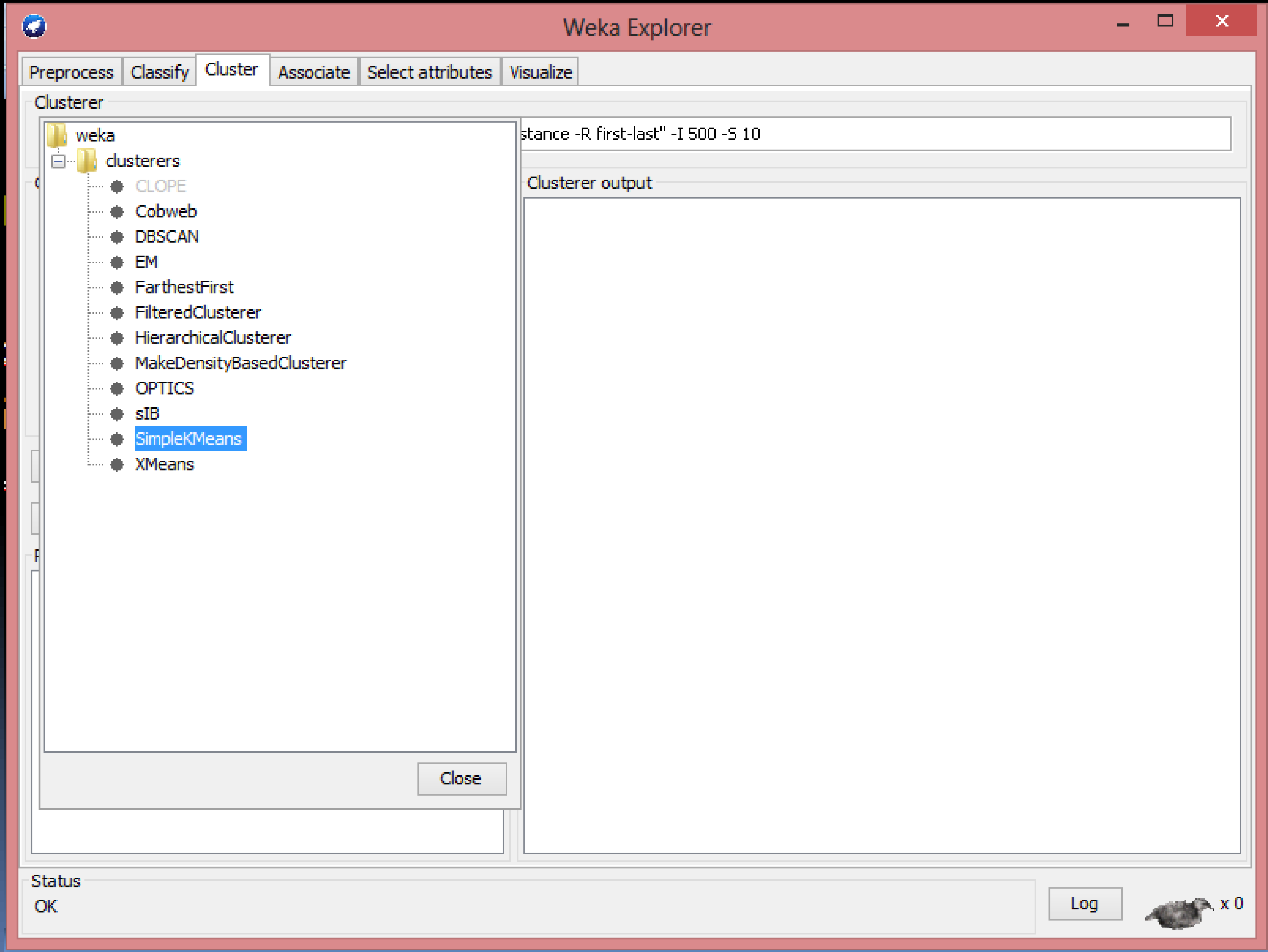
K-Means with WEKA

- Klik tombol “Open file”
- Pilih file CSV yang dibuat sebelumnya



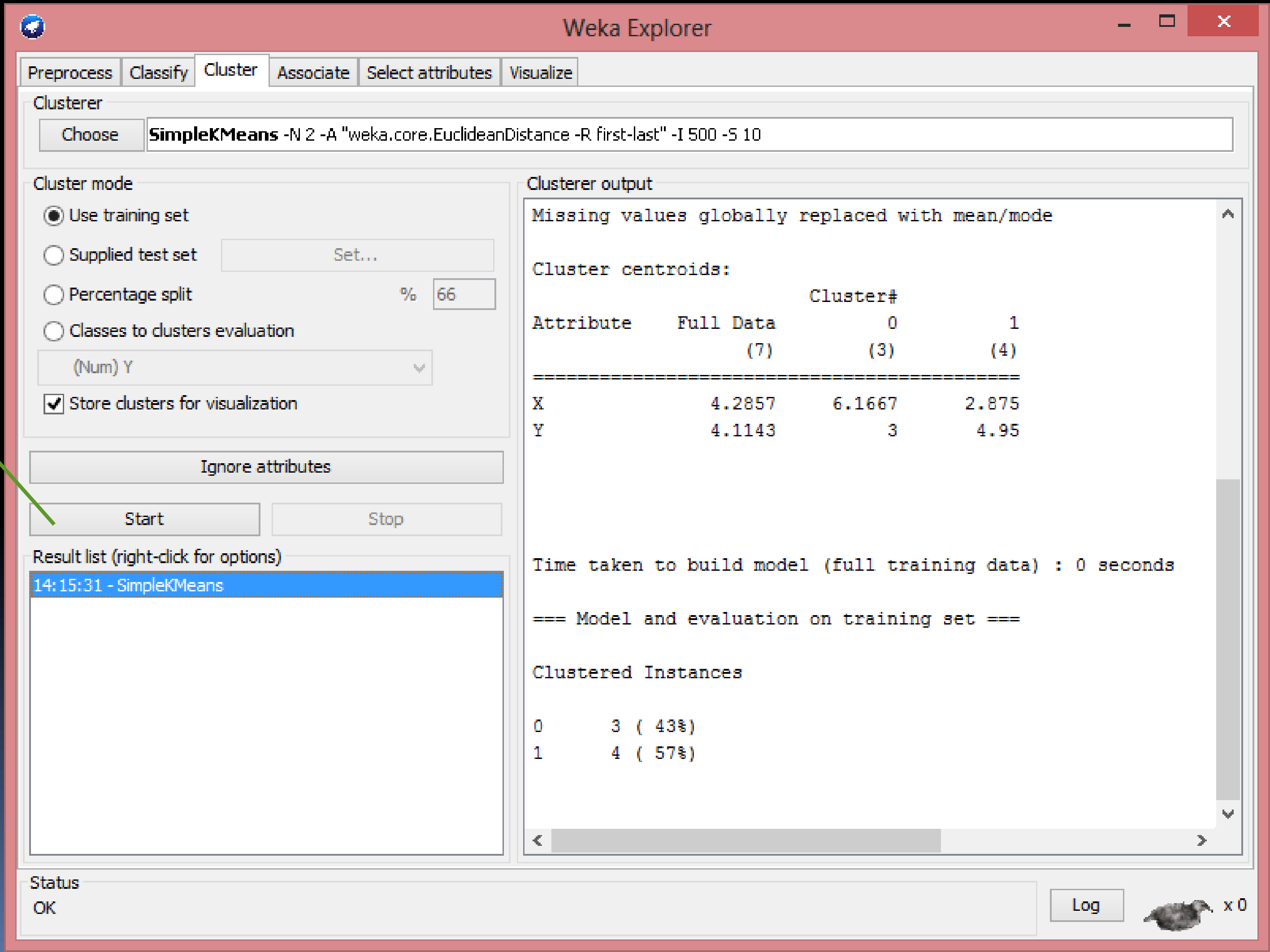
K-Means with WEKA

- Pindah ke Tab Cluster
- Klik tombol "Choose"
- Pilih "SimpleKMeans"



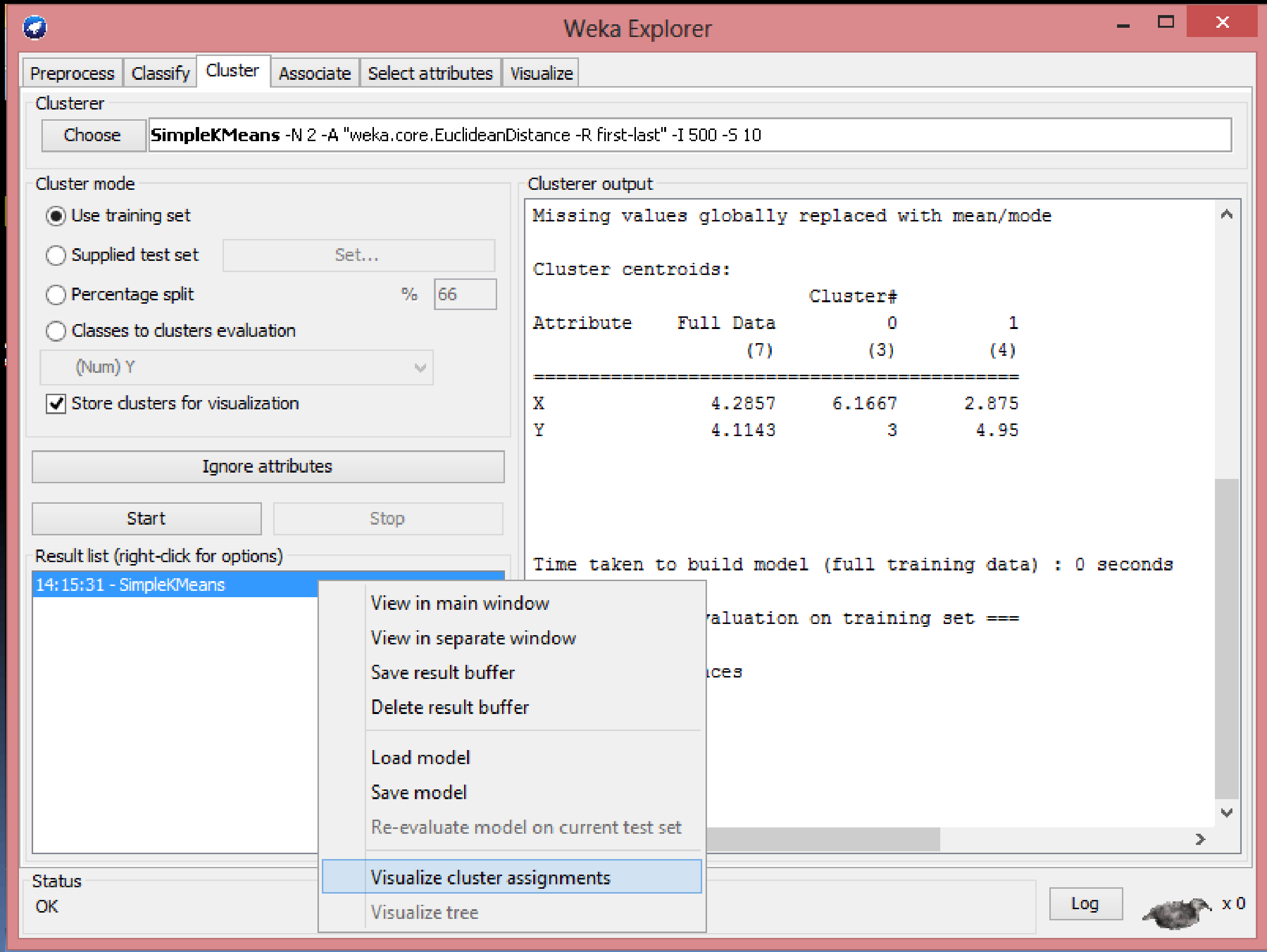
K-Means with WEKA

Klik tombol “Start”



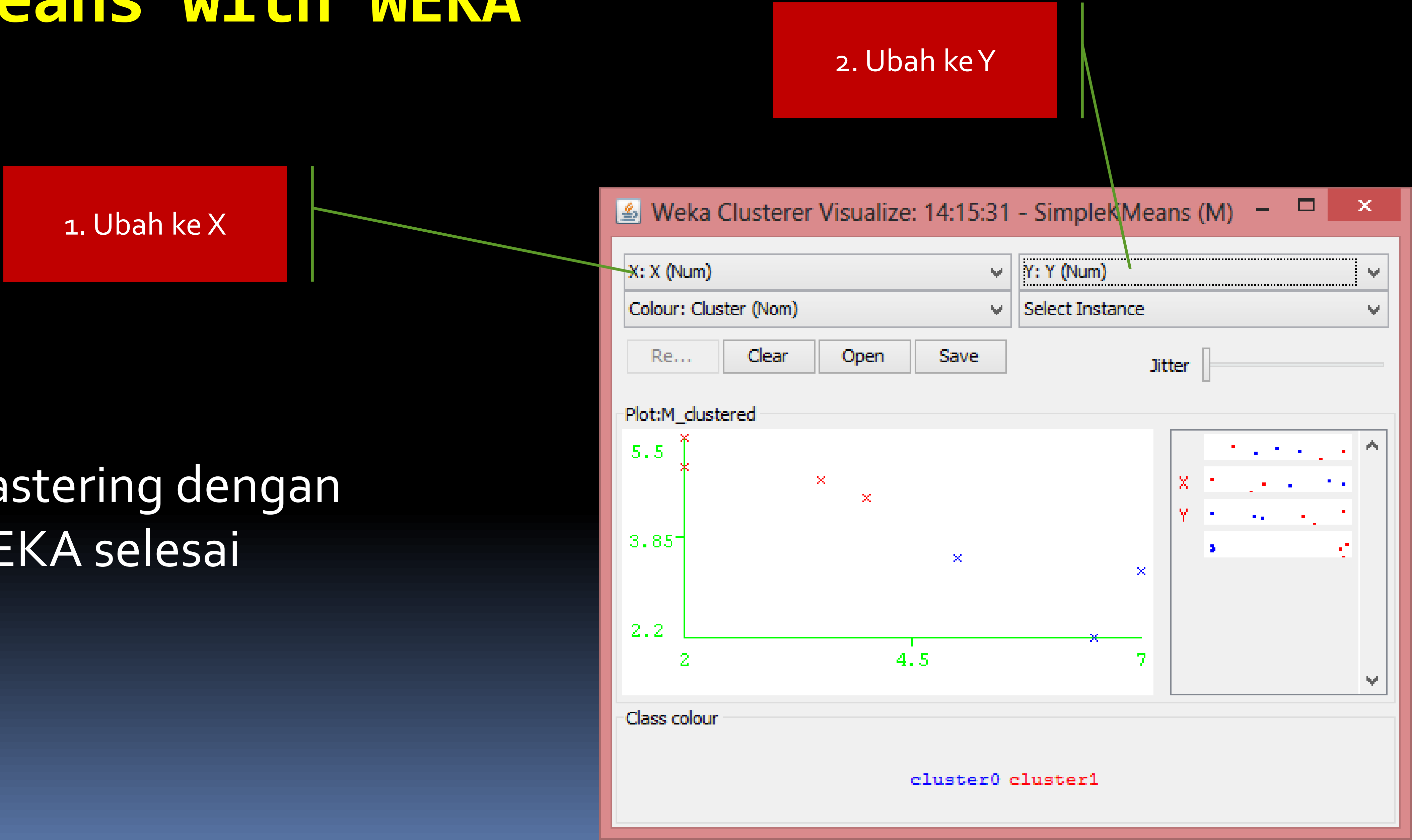
K-Means with WEKA

- Klik kanan pada "Result list"
- Pilih "Visualize cluster"



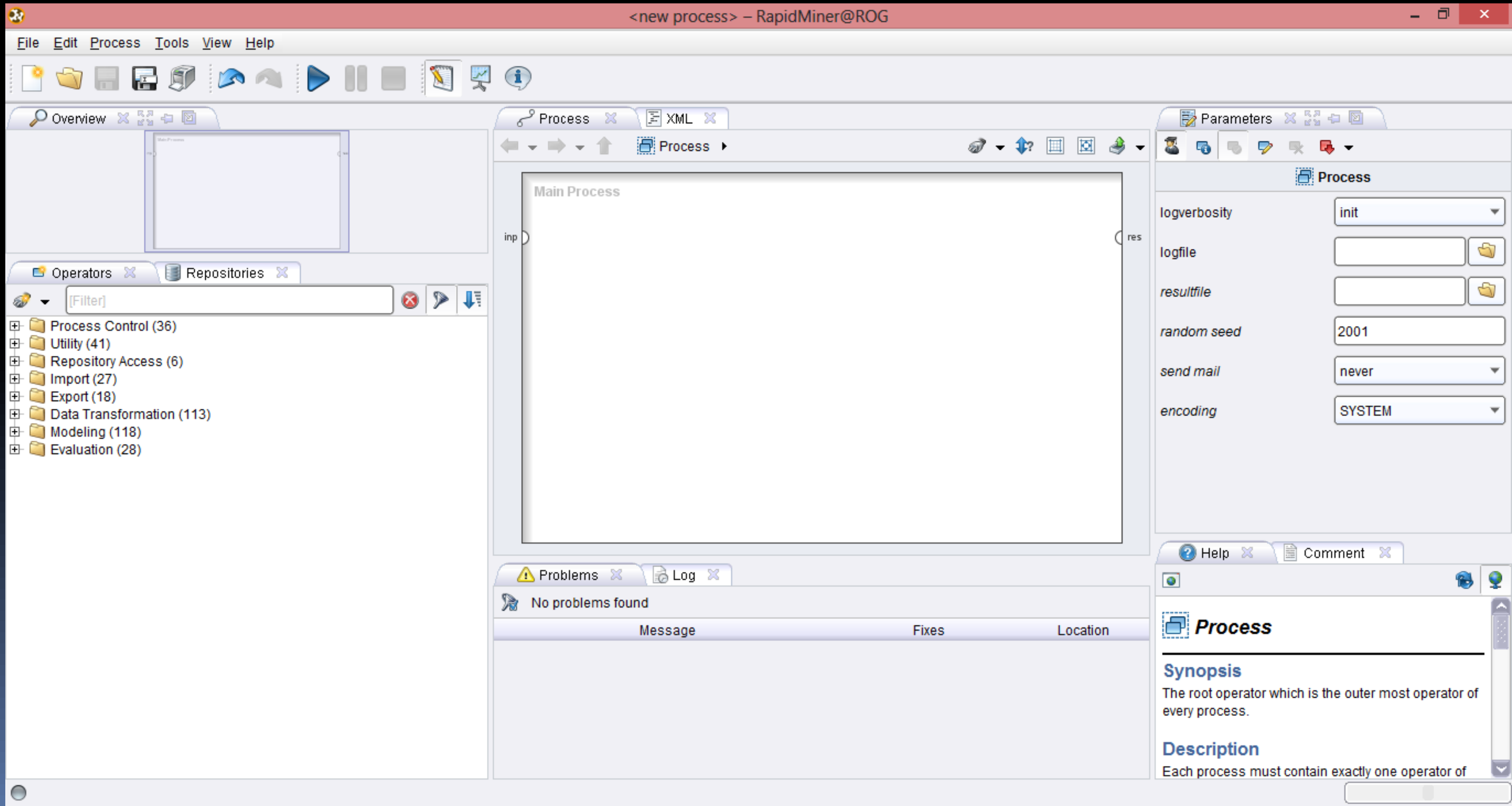
K-Means with WEKA

- Klastering dengan WEKA selesai



K-Means with RAPID MINER

- Buka aplikasi RM
- Klik “New Process”



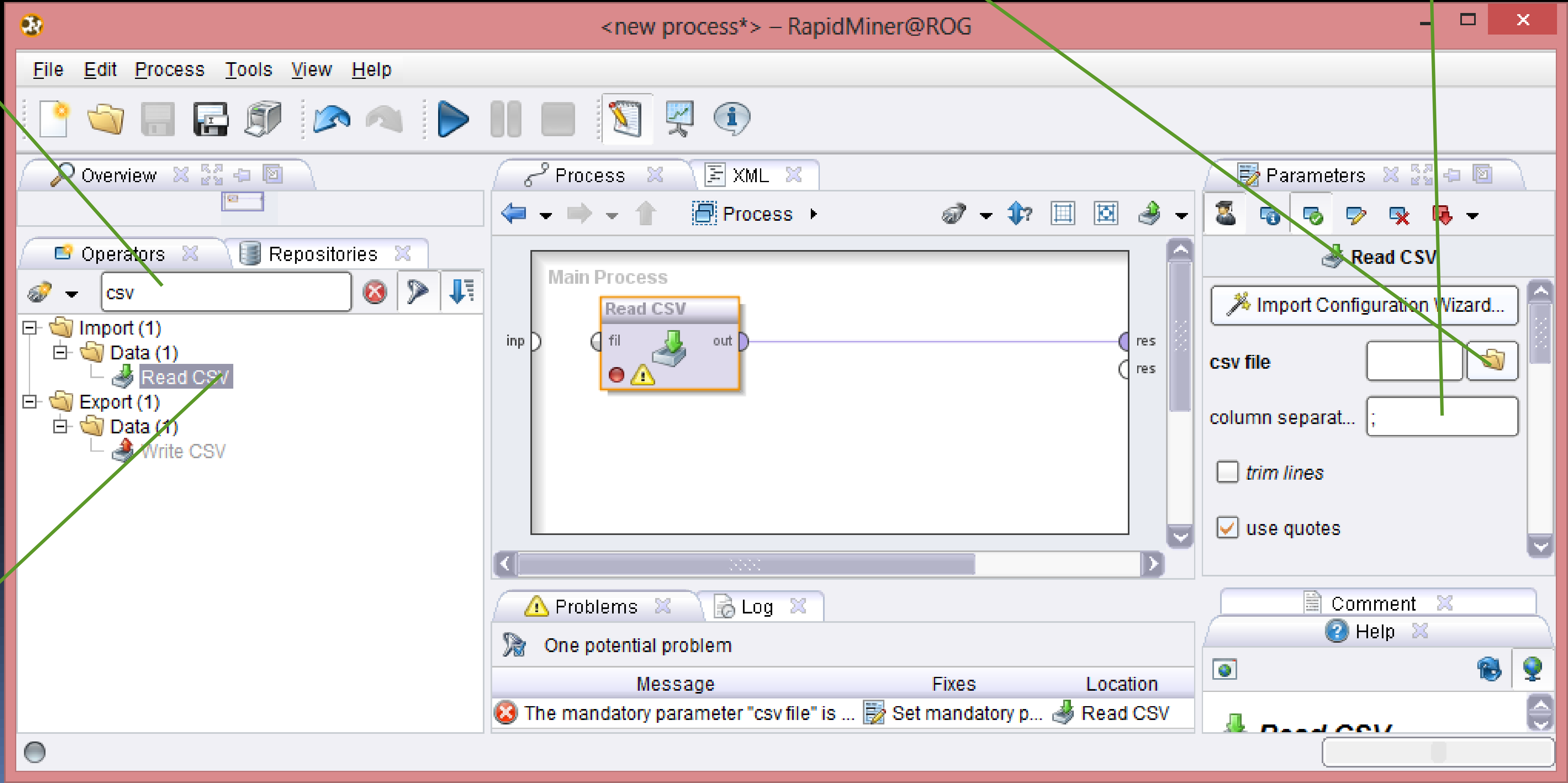
K-Means with RAPID MINER

3. Klik csv file, cari file data (*.csv)

4. Gunakan koma ","

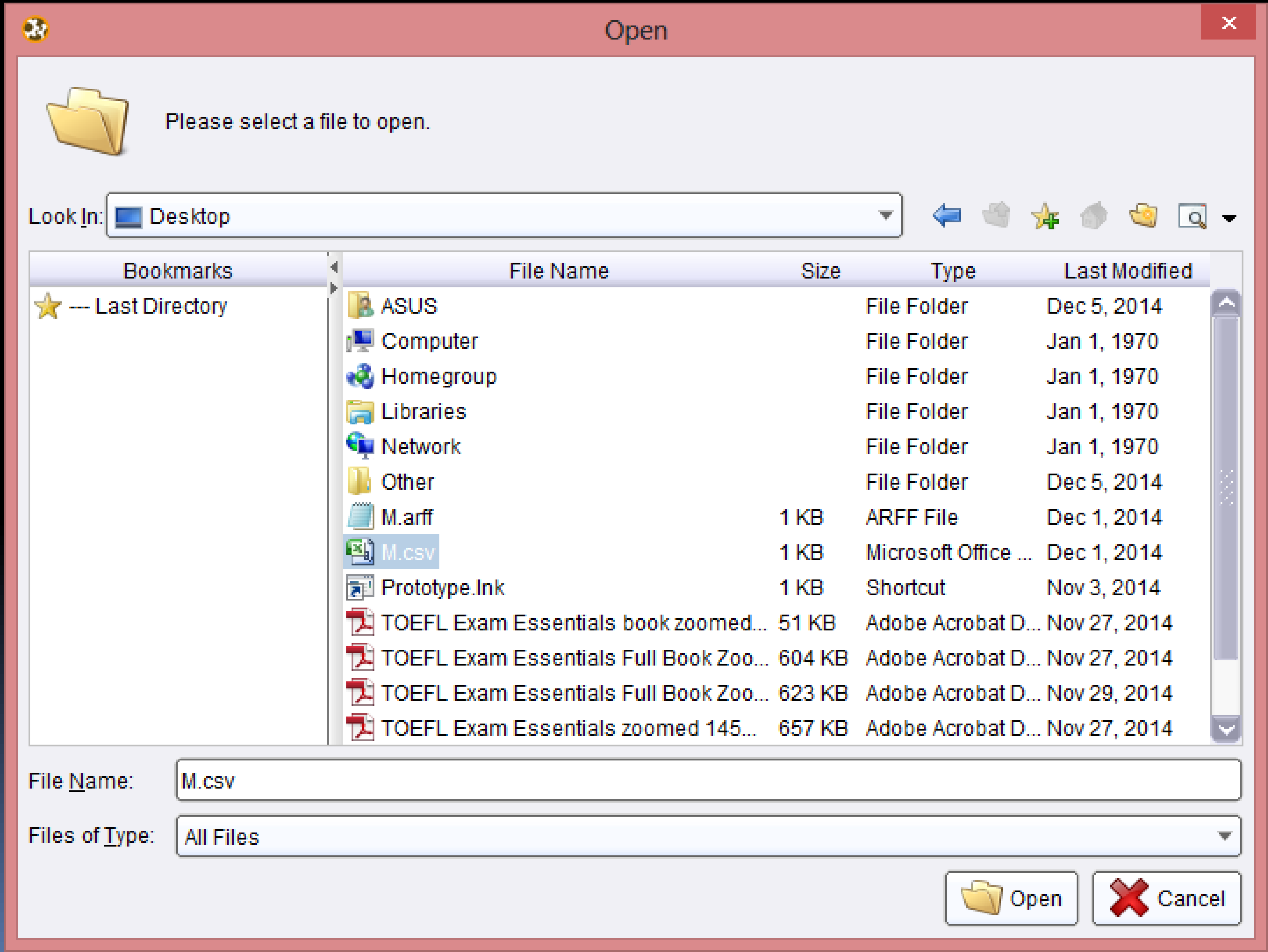
1. Ketik "csv" pada pencarian operator

2. Klik 2 kali



K-Means with RAPID MINER

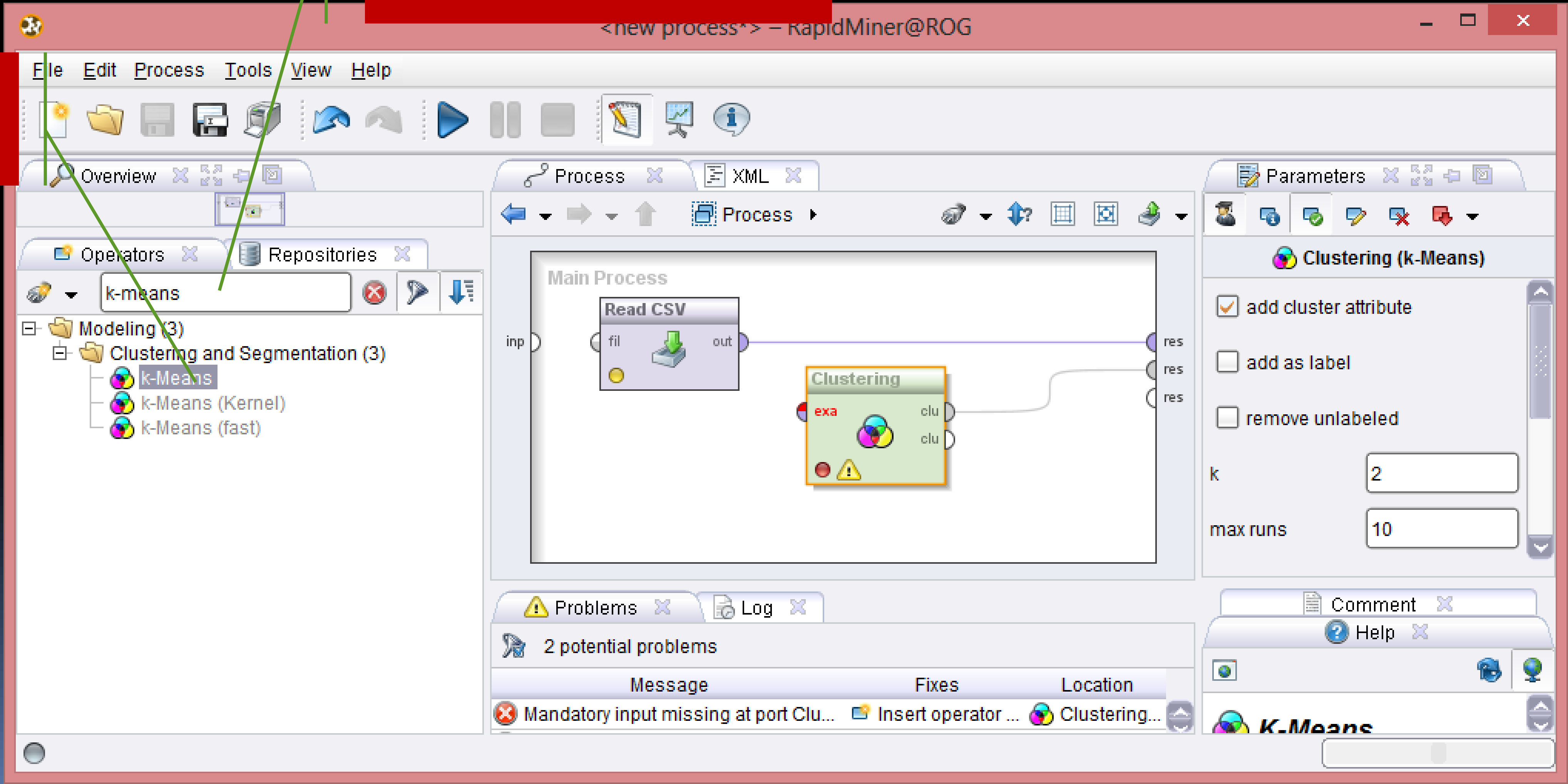
- Cari file csv yang sudah dibuat
- Klik Open



K-Means with RAPID MINER

1. Ketik “k-means” pada pencarian operator

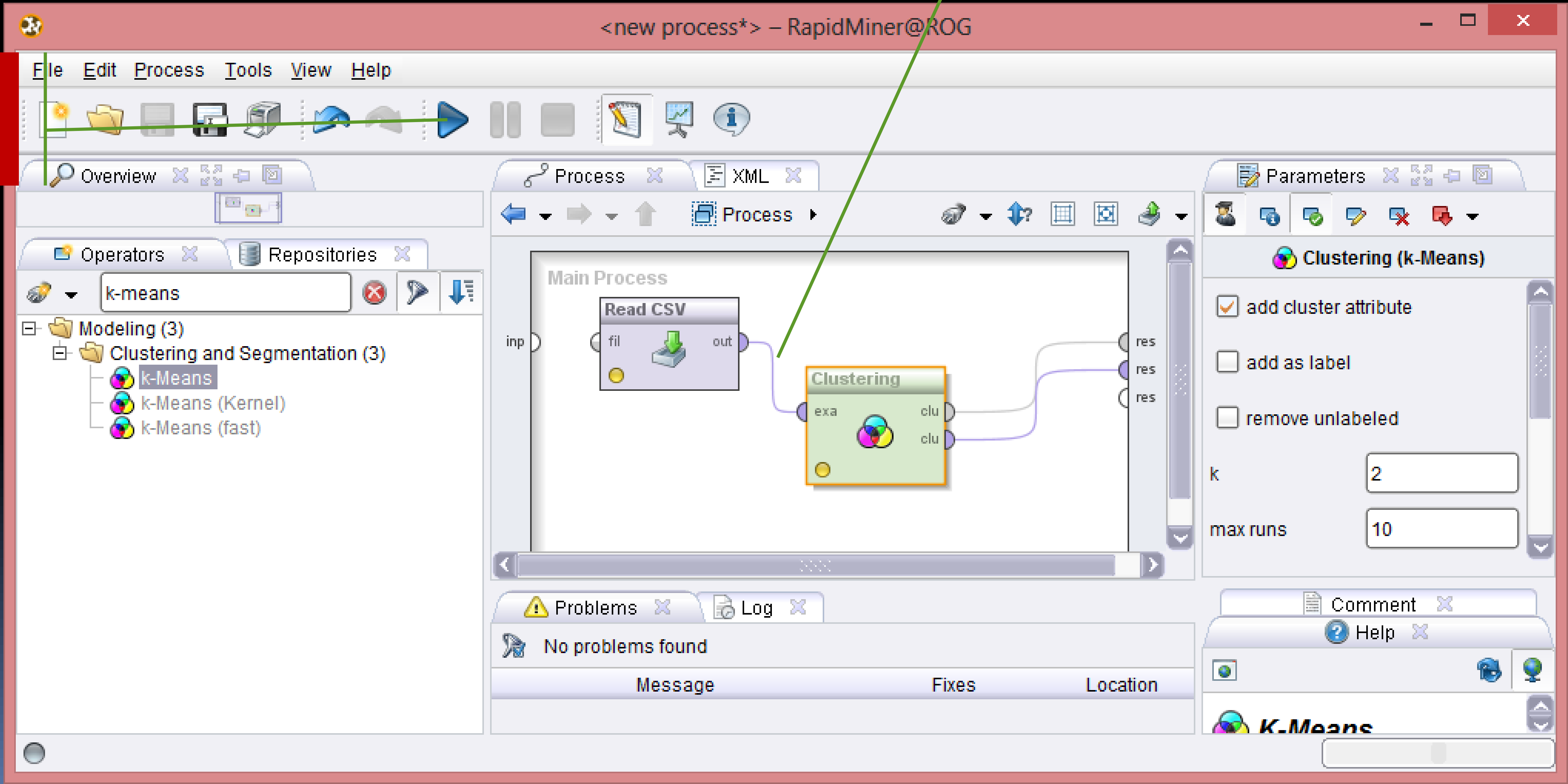
2. Klik 2 kali



K-Means with RAPID MINER

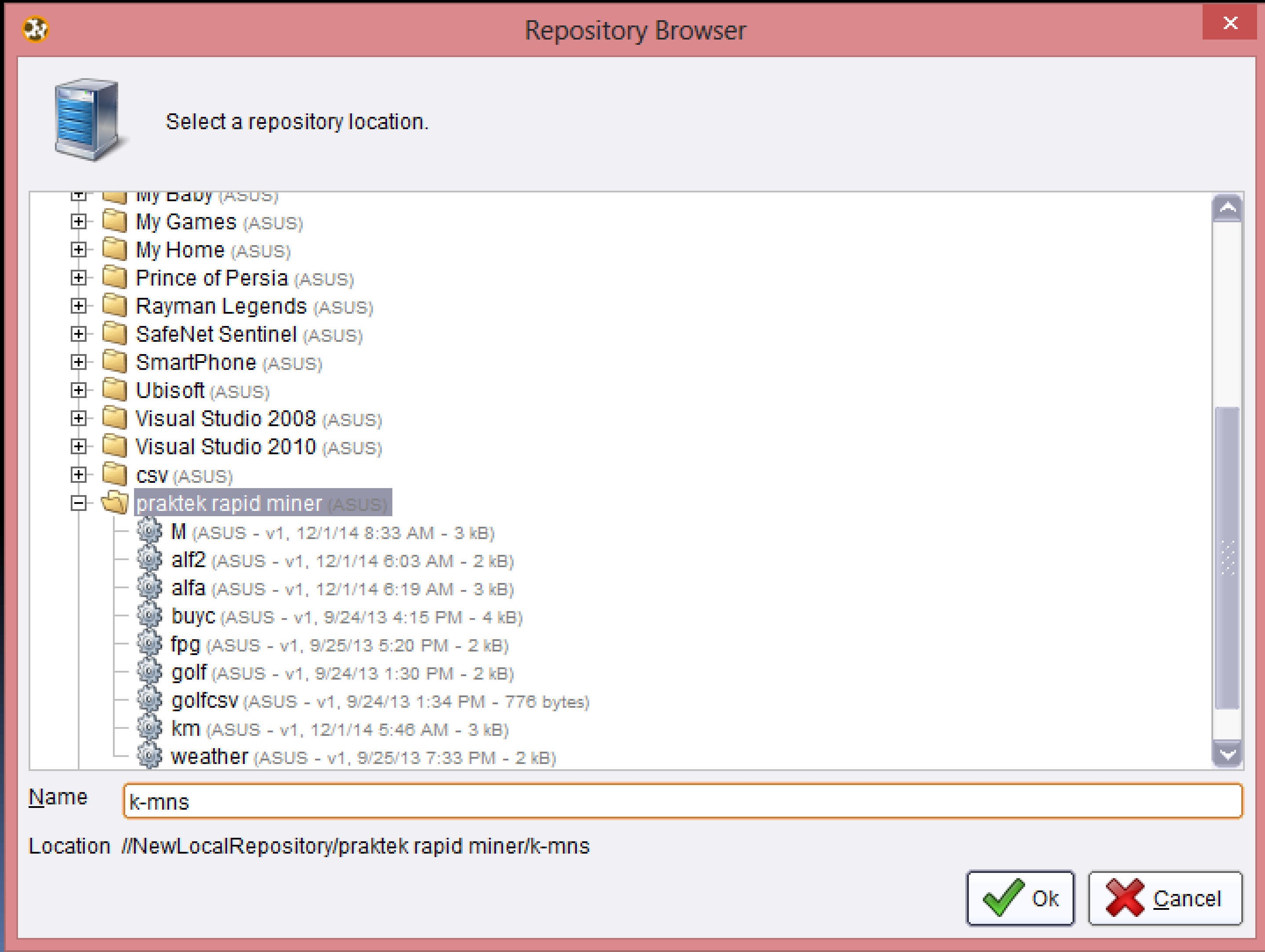
1. Atur koneksi seperti contoh

2. Klik run



K-Means with RAPID MINER

- Simpan proses



K-Means with RAPID MINER

1. Pilih “Plot View”

2. Pilih “X”

3. Pilih “Y”

4. Pilih “cluster”