Business Intelligence per i Big Data

Esercitazione di laboratorio N. 2

Gli obiettivi dell'esercitazione sono:

- **applicare i principali algoritmi di clustering disponibili in RapidMiner per segmentare gli utenti della campagna in funzione delle loro caratteristiche anagrafiche e lavorative e i testi in base alla similarità dei termini che contengono.**

Dati strutturati

Il dataset denominato UsersSmall (UsersSmall.xls) raccoglie dati anagrafici e lavorativi relativi a circa 300 persone contattate da un'azienda per proporgli l'iscrizione ad un loro servizio. Per tali utenti è noto se, dopo essere stati contattati, si sono iscritti al servizio proposto oppure no (valore del campo Response).

La lista completa degli attributi del dataset a disposizione (UsersSmall.xls) è riportata di seguito.

(1) Age
(2) Workclass
(3) Education record
(4) Marital status
(5) Occupation
(6) Relationship
(7) Race
(8) Sex
(9) Native country
(10) Response.

Dati testuali

Il dataset denominato Wikipedia contiene una collezione di 12 articoli di Wikipedia, appartenenti a 3 differenti categorie. In particolare, i documenti appartengono ai seguenti argomenti: matematica, cibo, sport.

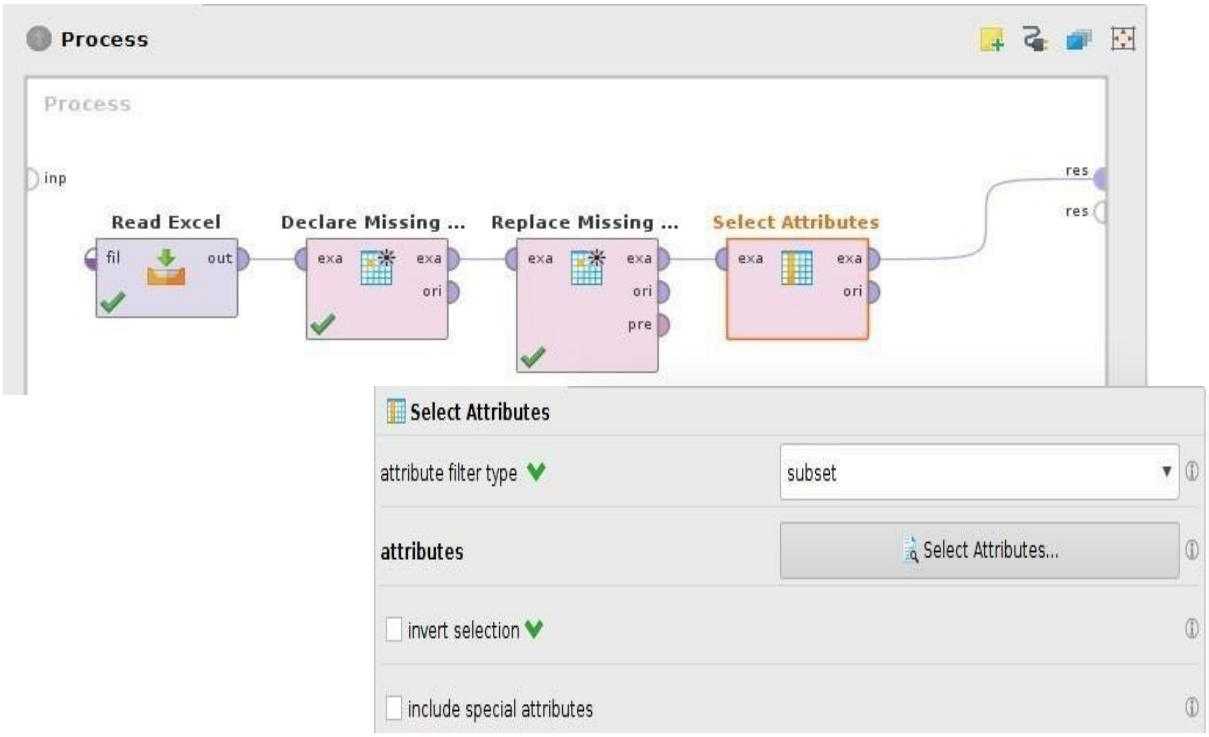
Clustering di dati strutturati

L'obiettivo dell'analisi è raggruppare le persone in gruppi omogenei, tali che persone appartenenti al medesimo gruppo abbiano caratteristiche simili mentre persone appartenenti a gruppi diversi siano dissimili. I gruppi possono rappresentare segmenti di clientela verso cui mirare specifiche promozioni o campagne pubblicitarie.

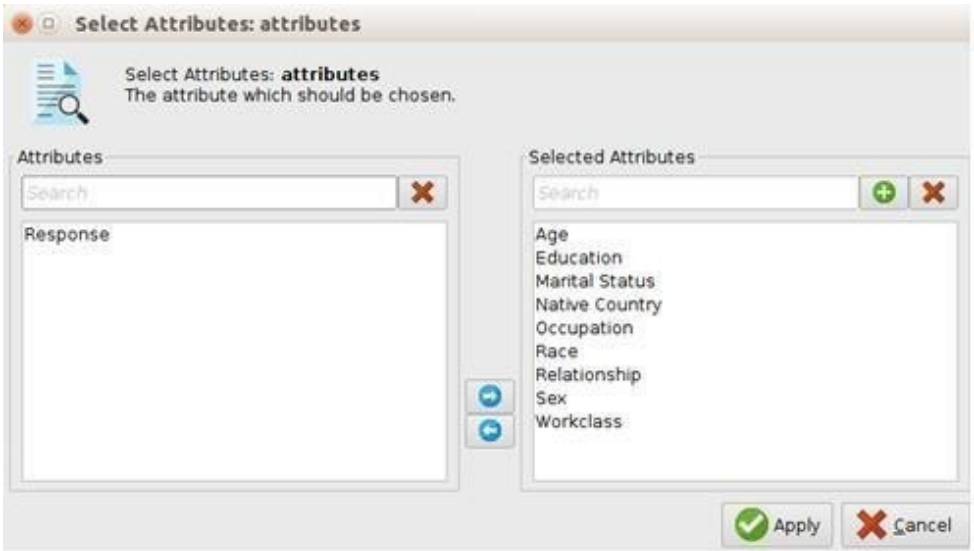
Obiettivo 1 - Import e preprocessing dei dati

Eseguire i diversi step di preprocessing, come imparato nell'esercitazione precedente.

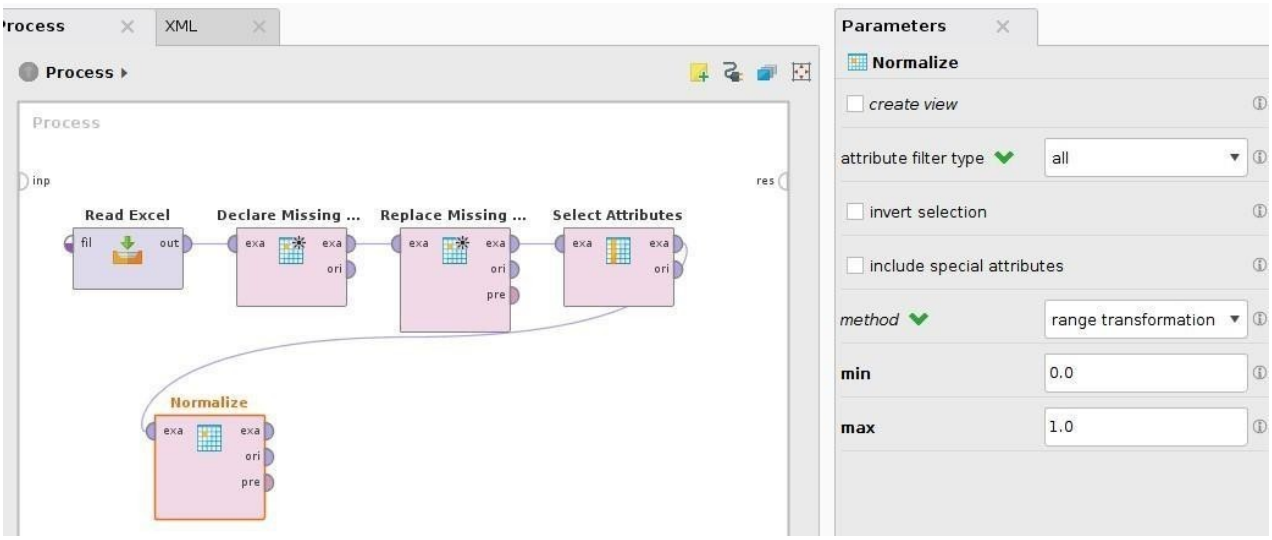
- In particolare, eseguire i seguenti step:
  - Import dati
  - Declare and replace missing values
  - Rimozione degli outliers



- Escludere l'attributo ***Response*** dall'analisi usando l'operatore ***Select Attributes***.



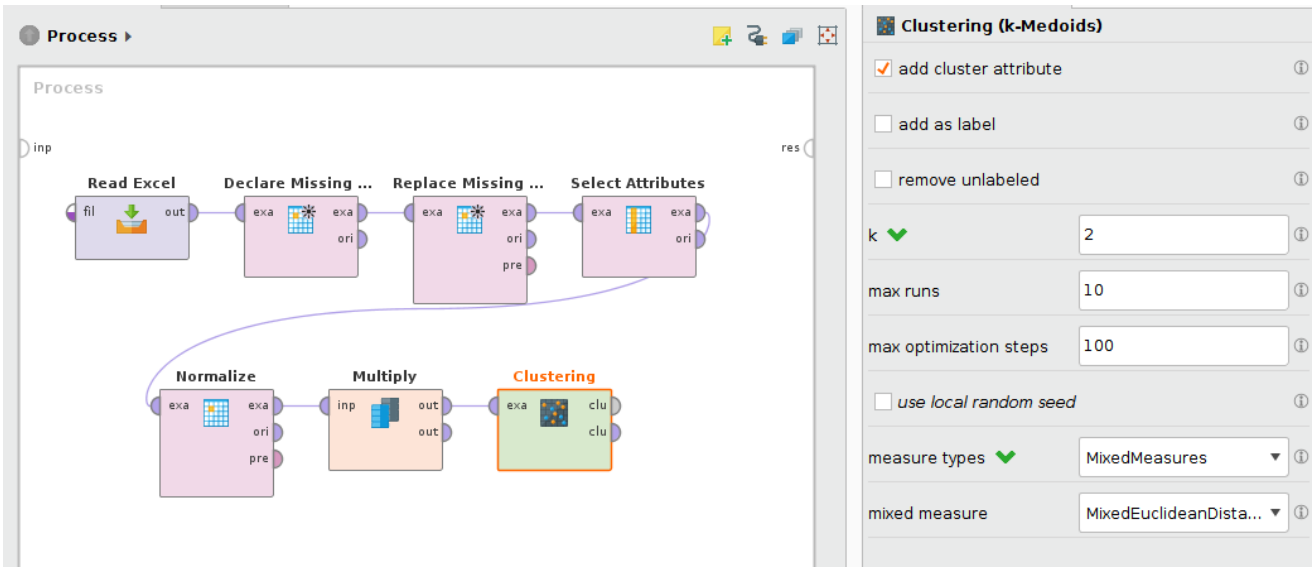
- **Normalizzare** i valori degli attributi numerici indicando come intervallo di valori [0-1] utilizzando l'operatore **Normalize**. L'unico attributo che verrà normalizzato è l'attributo età.



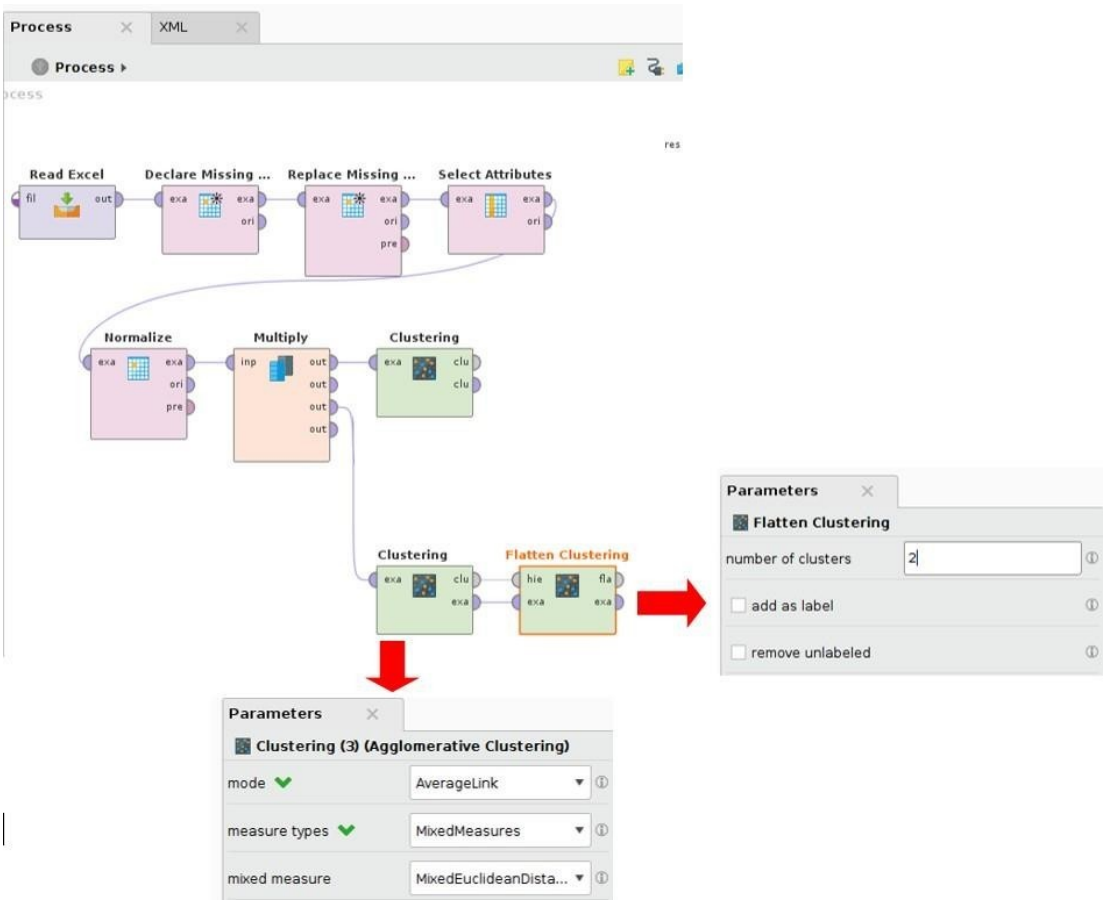
- Quando avete bisogno di utilizzare lo stesso input per diversi algoritmi, utilizzate l'operatore **Multiply**. Nei prossimi step verranno comparati diversi algoritmi di clustering.

Obiettivo 2 - K-Medoids e Agglomerative clustering

- Applicare l'algoritmo di clustering k-Medoids (quali differenze ha rispetto al K-means?) settando a K=2 il numero di cluster. Esegui il processo e analizza i risultati. Come sono distribuiti i due cluster trovati?
- Come sono distribuite le persone con Marital-Status: "*Divorced*" all'interno dei due cluster identificati? Provare a rispondere alla domanda con l'aiuto di un grafico (Bar-column).

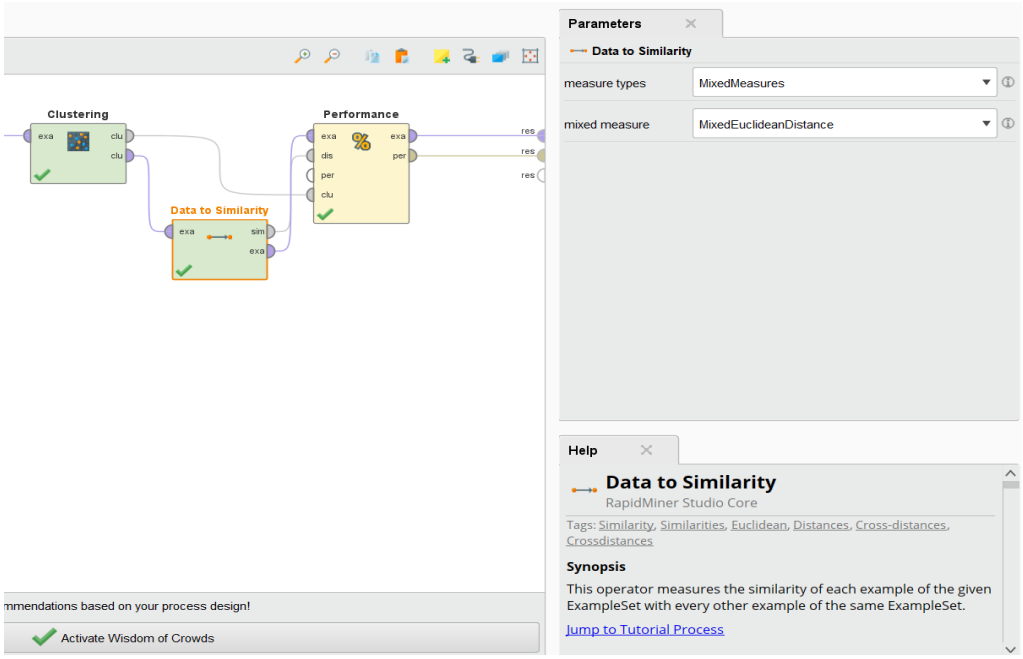


- Applicare l'algoritmo di clustering Agglomerative (Agglomerative Clustering). Selezionare due cluster dal risultato dell'algoritmo di clustering Agglomerative utilizzando l'operatore Flatten Clustering. Esegui il processo e confronta il risultato ottenuto con quello prodotto dall'algoritmo k-Medoids (numero di cluster k=2) svolto precedentemente. Come sono distribuiti gli elementi all'interno dei due clusters?

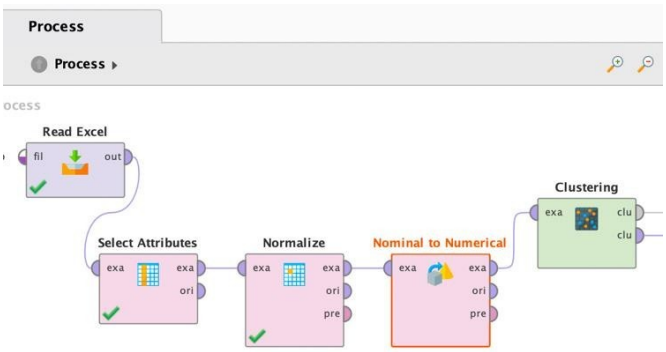


Obiettivo 3 - Valutazione oggettiva dei cluster generati per l'algoritmo K-Medoid

- Calcolare le performance dei cluster generati con l'algoritmo *K-Medoid*. Per il calcolo usare il blocco **Data To Similarity** e il blocco **Performance (Cluster Density Performance)**, che permette di misurare la densità del cluster e quindi quanto questo sia compatto. Utilizzare *MixedMeasures* come *measure types* in **Data To Similarity,** in quanto gli attributi sono sia numerici che nominali.



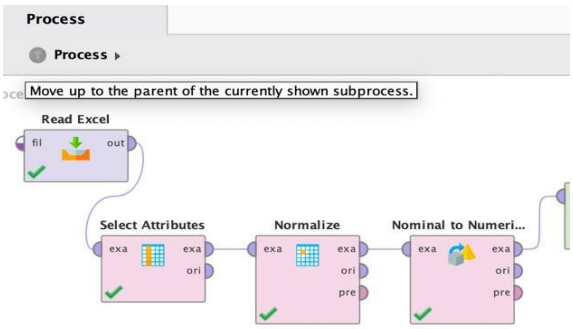
- Provare adesso a trasformare tutti gli attributi nominali in attributi numerici prima di effettuare l'algoritmo di clustering. Utilizzare l'operatore "Nominal to Numerical" con coding_type: **dummy coding.** In che modo vengono trasformati gli attributi nominali con questo operatore? Se tutti gli attributi sono ora numerici è possibile utilizzare altre misure di distanza cambiando *measure types* a *NumericalMeasures* e scegliendone un altra (per ora mantenere *Euclidean Distance*). Come varia il valore (calcolato sempre con K=2)? Il valore ottenuto è migliore o peggiore di quello ottenuto in precedenza?
- Provare a riflettere su vantaggi e svantaggi della tecnica **dummy coding** rispetto al label encoding utilizzato internamente da RapidMiner nell'operatore di clustering. Quando è meglio utilizzare uno e quando l'altro?

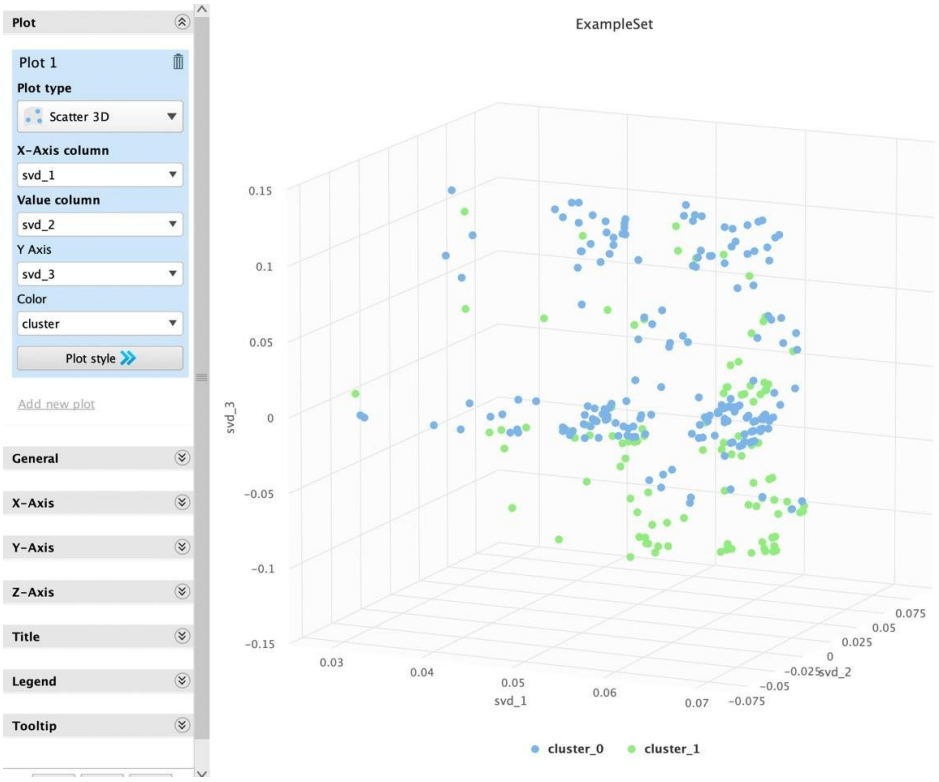


- Rieseguire adesso il processo di valutazione precedente per **differenti valori di K** per l'algoritmo **K-medoids.** Come varia il valore ottenuto all'aumentare di K? Come spieghi questo comportamento?

Obiettivo 4 - Visualizzazione/validazione del risultato di un processo di clustering tramite l'uso di tecniche di riduzione delle dimensioni dei dati - Principal Component Analysis attraverso SVD (Singular Value Decomposition)

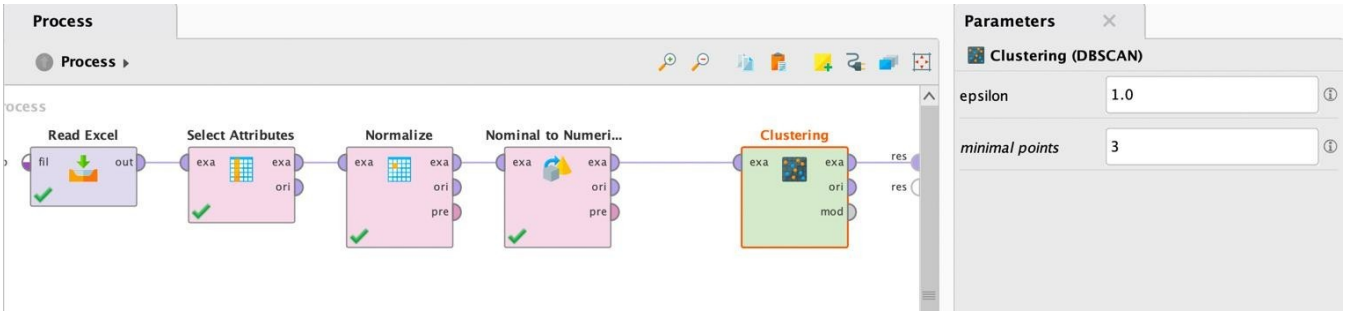
- Analizzare la qualità del clustering generato mediante una tecnica di riduzione della dimensionalità dei dati, nota come **Principal Component Analysis**. La più nota tecnica algebrica per effettuare la PCA è attraverso la decomposizione matriciale ***Singular Value Decomposition*** (SVD). La SVD trasforma i dati linearmente. Scegliendo le prime K componenti (dimensioni) ottenute dalla SVD, si si possono proiettare i dati, originalmente a N-dimensioni in uno spazio a K-dimensioni, con K scelto dall'utente e minore di N.
- Applicate l'operatore *SVD (Singular Value Decomposition)* sul dataset generato dal processo di clustering realizzato al passo precedente. Eseguire il processo impostando K=3 e visualizzare su un grafico di tipo scatter i dati rispetto alle tre dimensioni individuate dall'operatore SVD. Usare l'attributo **cluster** come attributo per la **scelta dei colori dei punti**. I cluster sono ben definiti?


Process
Process
Move up to the parent of the currently shown subprocess.
Read Excel
fil
out
Select Attributes
Normalize
Nominal to Numeri...
exa
ori
pre


Plot
Plot 1
Plot type
Scatter 3D
X-Axis column
svd_1
Value column
svd_2
Y Axis
svd_3
Color
cluster
Plot style
Add new plot
General
X-Axis
Y-Axis
Z-Axis
Title
Legend
Tooltip
ExampleSet
svd_3
0.15
0.1
0.05
0
-0.05
-0.1
-0.15
0.03
0.04
0.05
0.06
0.07
svd_1
0.075
0.05
0.025
0
-0.025
-0.05
-0.075
svd_2
cluster_0
cluster_1

Obiettivo 3 – DBSCAN

- Applicare l'algoritmo di clustering DBScan settando min points = 3. Esegui il processo e analizza i risultati.



Valutazione oggettiva dei cluster generati per l'algoritmo DBSCAN

- Quanti cluster vengono identificati dall'algoritmo DBSCAN? Qual è il cluster che contiene il numero maggiore di elementi?
- Provare a ripetere l'esperimento cambiando il parametro *minimal points* (provare 2 e 4 come valori). Come cambia il numero di cluster individuati al variare di questo parametro? Questo comportamento è in linea con quanto ti aspettavi?
- Settare nuovamente il parametro *minimal points=3*. Come bisogna cambiare il parametro *eps* per ottenere meno punti etichettati come rumore? Perché? Prova a ripetere l'algoritmo con diversi valori di *eps* e guarda come cambia la cardinalità del cluster contenente punti rumorosi

Clustering di dati testuali

La seconda parte di questa esercitazione prevede l'analisi attraverso l'algoritmo K-Means della collezione di articoli denominata *Wikipedia*. Scompatta la cartella presente sul sito del corso ed esegui i passi seguenti.

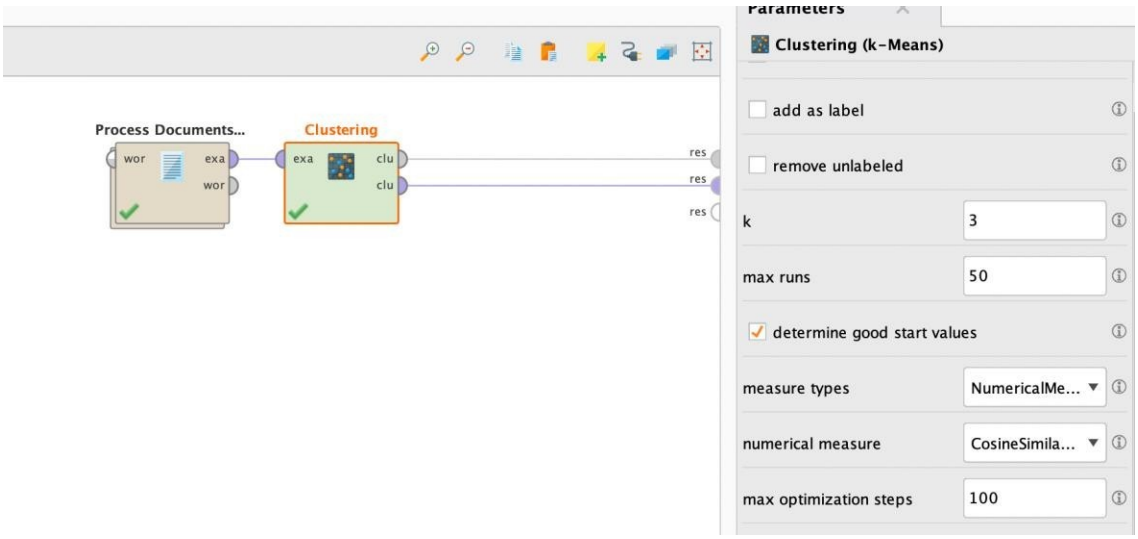
Obiettivo 1 - Import e preprocessing dei dati

- Trasforma la collezione di documenti nella matrice document*term. Per eseguire questa trasformazione, eseguire i diversi step di preprocessing, come imparato nella prima esercitazione.

Obiettivo 2 - Clustering dei dati

- Utilizzare l'algoritmo di **K-Means** per dividere la collezione in gruppi omogenei di documenti che parlino di uno stesso topic. Per i dati testuali la misura per calcolare la distanza tra punti (in questo caso tra due documenti) è la **CosineSimilarity.**

- Impostare K=3 e max_runs=50: come vengono suddivisi i 12 testi iniziali tra i 3 clusters? Selezionare l'opzione "keep_text" all'interno dell'operatore "Process Documents from Files" per tenere traccia del testo originale.



- Identificare all'interno di ciascun cluster le 3 parole che hanno un'importanza maggiore. (SUGGERIMENTO: andare nella sezione "Centroid Table")
- Provare infine a visualizzare i cluster identificati attraverso la tecnica SVD (3 dimensioni). I 3 clusters identificati sono ben distinti nello spazio tridimensionale?