Business Intelligence per i Big Data

Esercitazione di laboratorio N. 4

L'obiettivo dell'esercitazione è:

- **utilizzare il software Rapid Miner per effettuare i preprocessing di dati strutturati (relativi ad una campagna promozionale) e di dati non strutturati (ad esempio, dati testuali relativi ad un argomento specifico) per analisi successive.**
- **applicare i principali algoritmi di clustering disponibili in RapidMiner per segmentare gli utenti della campagna in funzione delle loro caratteristiche anagrafiche e lavorative e le news in base alla similarità dei termini in esse contenuti.**

Dati strutturati

Il dataset denominato UsersSmall (UsersSmall.xls) è disponibile sul sito del corso (http://dbdmg.polito.it/wordpress/teaching/business-intelligence/). Esso 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) Hours per week
(10) Native country
(11) Response.

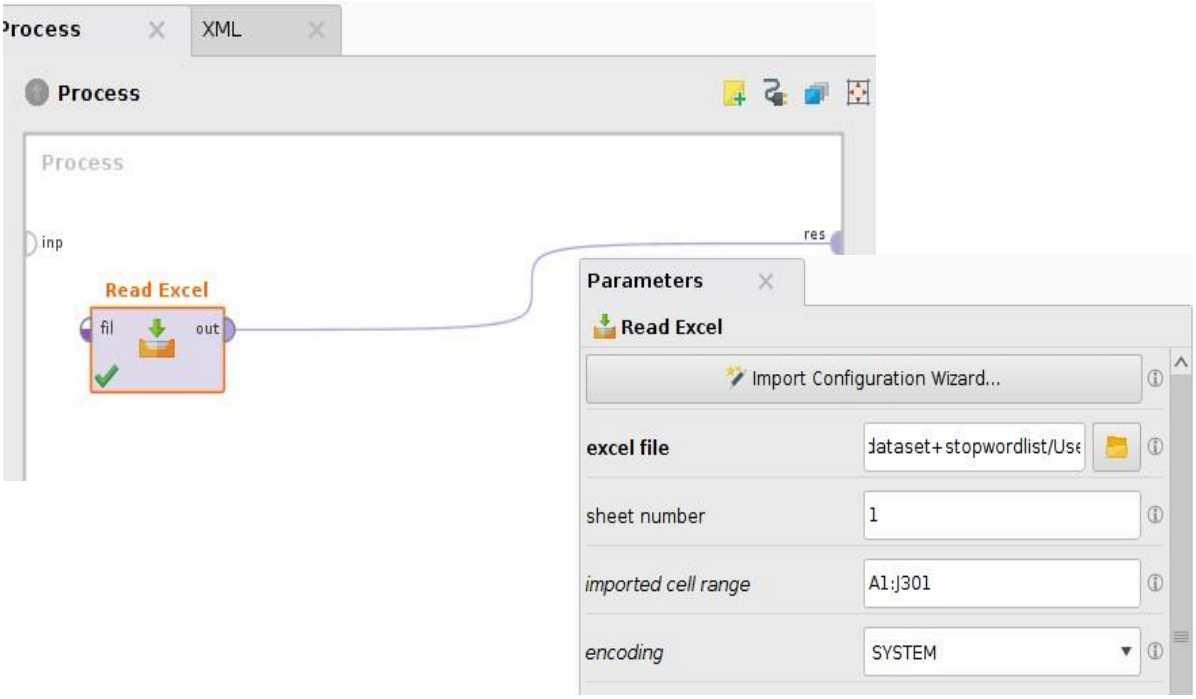
Dati testuali

Il dataset denominato Wikipedia (Wikipedia.zip) è disponibile sul sito del corso (http://dbdmg.polito.it/wordpress/teaching/business-intelligence/). Esso contiene una collezione di 12 articoli di Wikipedia, appartenenti a 3 differenti categorie. In particolare, i documenti appartengono ai seguenti argomenti: matematica, cibo, sport.

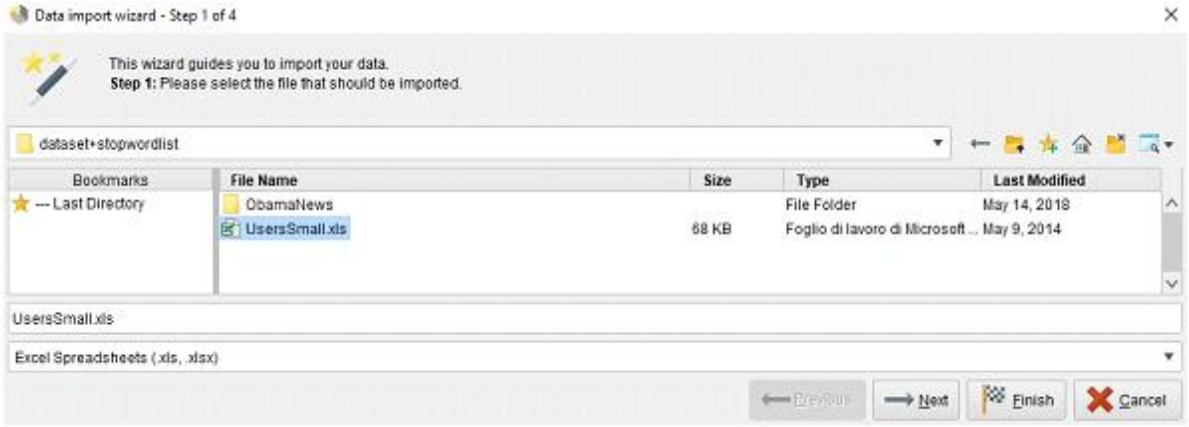
Preparazione dei dati strutturati

Obiettivo 1 – Import dei dati

- Nel pannello **Operators** cercare l'operatore **Read Excel** e trascinarlo nello spazio di lavoro.
- Importare il dataset ***UsersSmall.xls*** utilizzando la procedura guidata ***Import Configuration Wizard***.

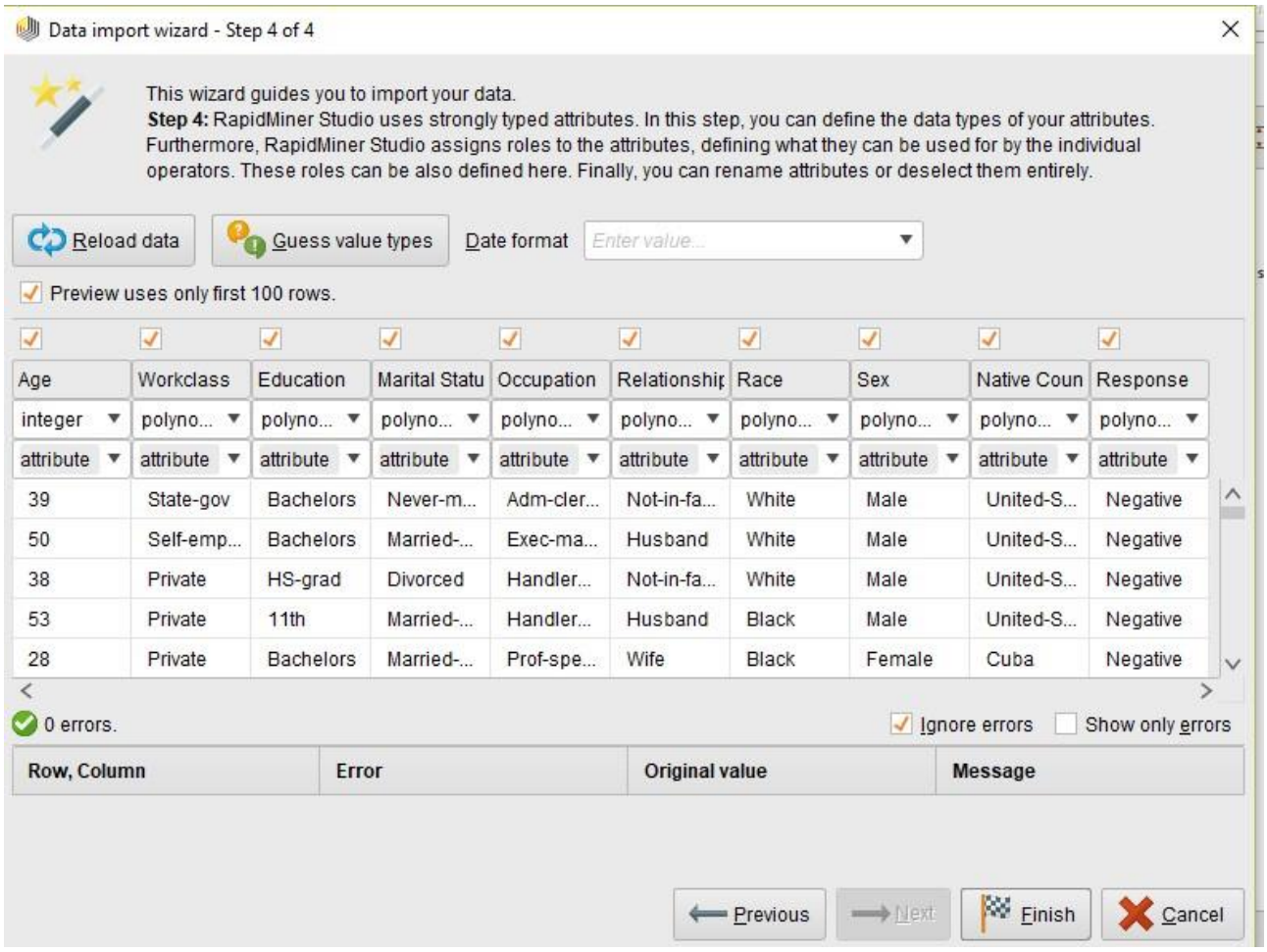


- Selezionare il file all'interno della cartella in cui si sono scompattati i file e selezionare *UsersSmall.xls*.

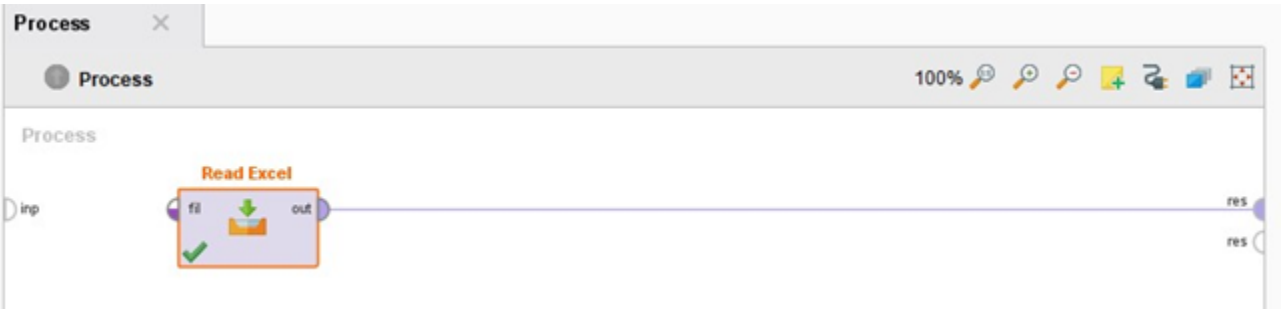


- Cliccare *Next*.
- Controllare che l'intera matrice di dati sia selezionata, in caso contrario selezionare **tutte le colonne**.
- Procedere con *Next*

- Procedere con *Finish*. Si chiuderà il processo guidato.



- Collegare l'uscita dell'operatore Read Excel con res. Usare il tasto destro del mouse.



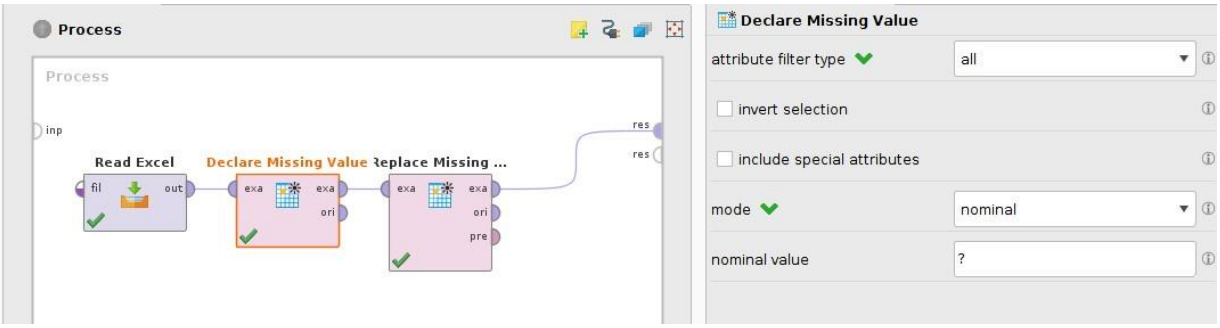
- Per lanciare un processo in RapidMiner usare il triangolo in alto nella barra dei processi.
- Per tornare nel processo principale, cliccare su **design**.

- Analizzare la semantica degli attributi e il loro ruolo a seconda degli obiettivi dell'analisi svolta.

Obiettivo 2 – Gestione dei dati mancanti

Verificare la presenza di eventuali dati mancanti e gestirli con opportuni passi di trasformazione (operatori *Decleare Missing values* e *Replace Missing Values*).

• Dichiarare per tutti gli attributi il '**?**' come valore NULL attraverso l'operatore **Declare Missing Value**.
• Sostituire i valori nulli dichiarati al punto precedente con il valore più frequente usando l'operatore **Replace Missing Values**.



Obiettivo 3 – Outlier detection

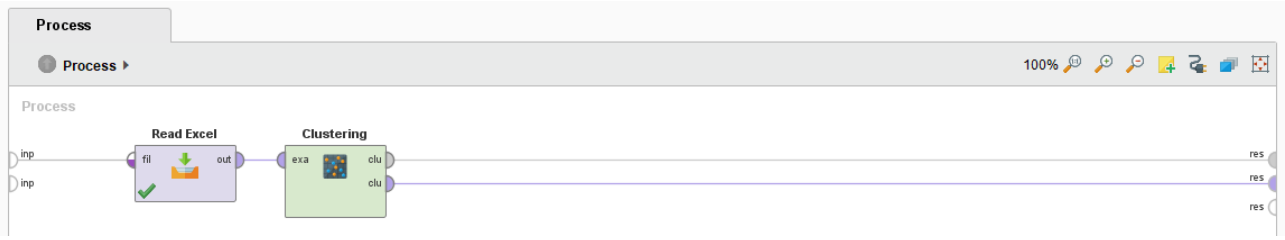
Verificare la presenza di outlier all'interno del dataset, utilizzando una strategia univariata.

- Nel tab R*isultati*, visualizzare le statistiche calcolate per il dataset da analizzare.
- Nel tab R*isultati*, plottare il grafico *Quartiles* per l'attributo *Age*, selezionandolo tra i diversi Charts disponibili.

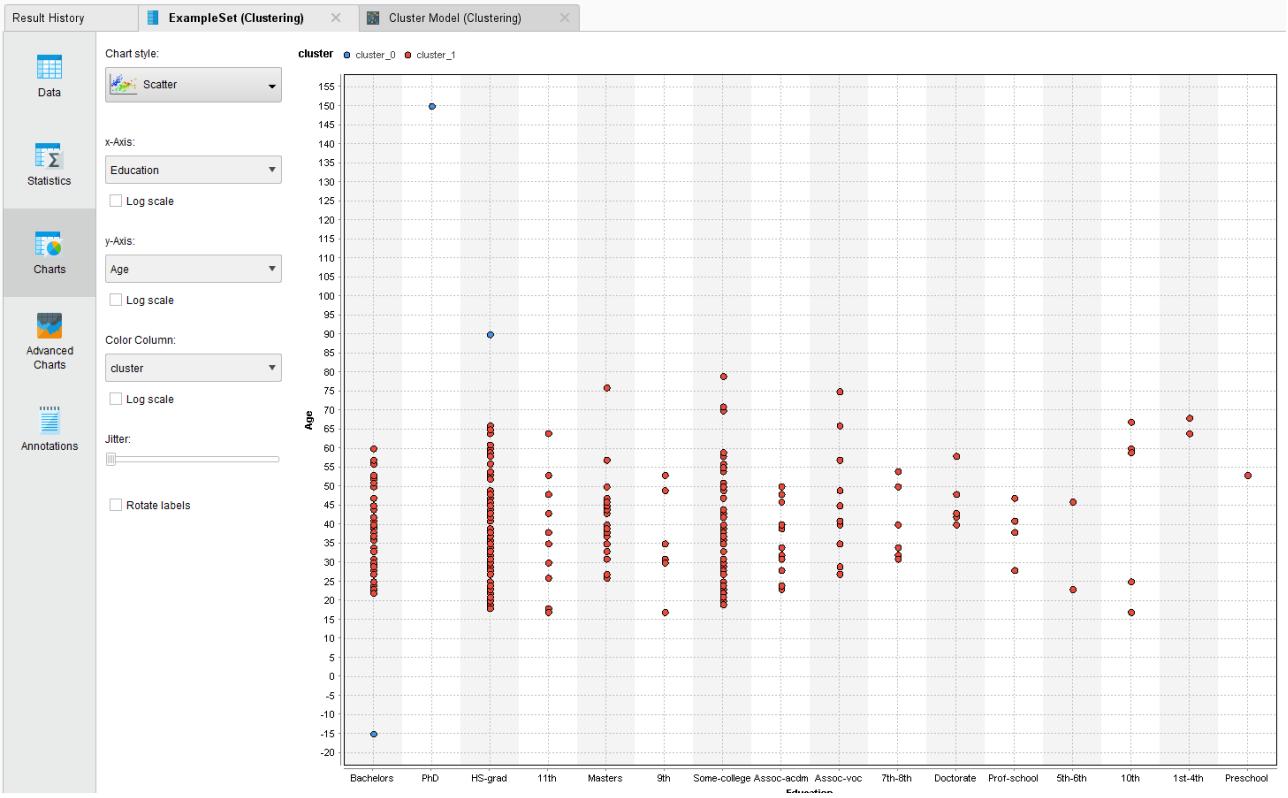
Sono presenti possibili outlier per il dataset in questione? Quali?

Verificare la presenza di outlier all'interno del dataset, utilizzando una strategia multivariata.

- Utilizzare il blocchetto DBScan per eseguire l'analisi multivariata. Settare i seguenti parametri: epsilon = 10, minpoints = 3.

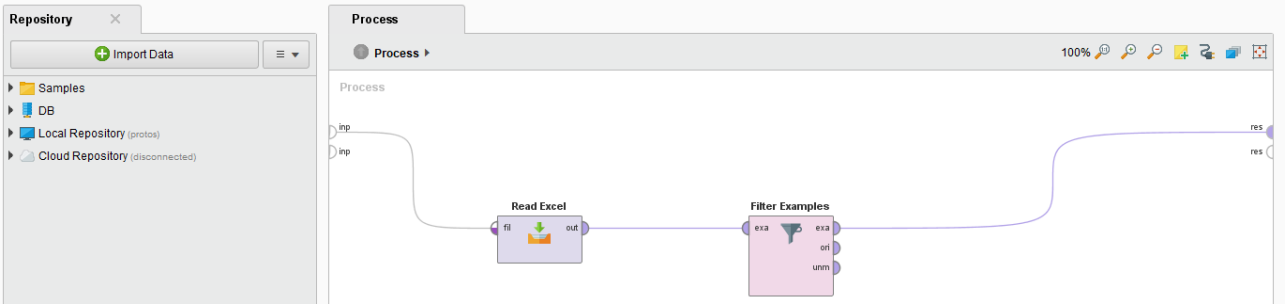


- Collegare entrambe le uscite del DBScan all'uscita.
- Alcuni record sono stati identificati come outlier (Cluster 0)? Visualizzare i sample appartenenti al Cluster 0, sia in forma tabellare che in forma di grafico (Scatter chart, assi x e y uguali a *Age* e *Education* rispettivamente, color column uguale a *cluster*).



Confrontare i risultati dell'analisi univariata e dell'analisi multivariata.

Rimuovere (per tutte le analisi successive) gli outlier identificati, utilizzando l'operatore *Filter Exmaples.*



Obiettivo 4 – Discretizzazione

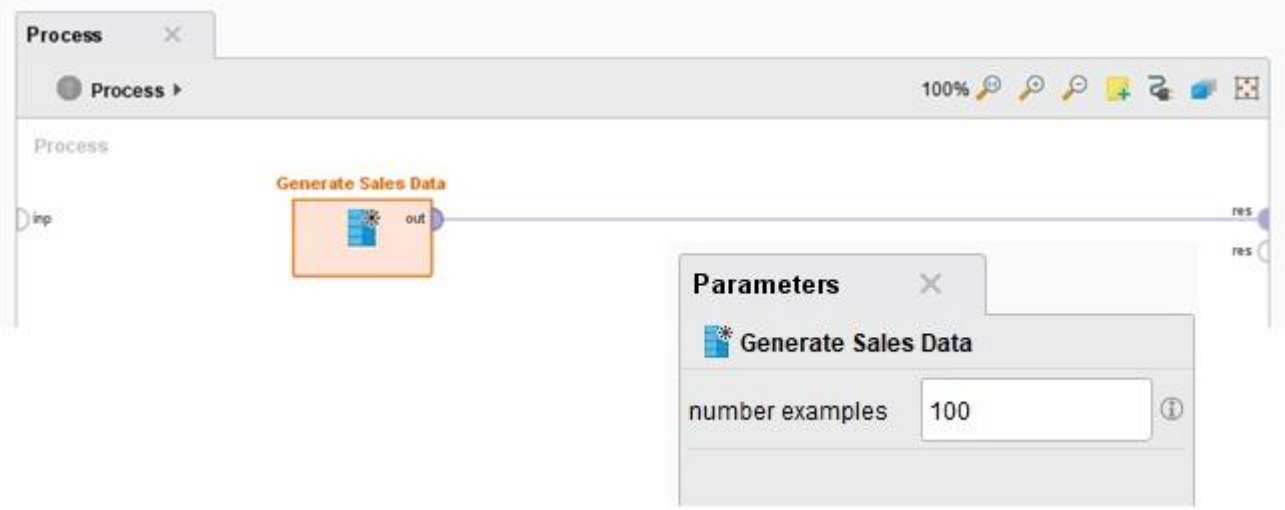
Verificare la presenza di attributi continui nei dati di origine.

- Discutere l'eventuale necessità di applicare un processo preliminare di discretizzazione in funzione degli obiettivi dell'analisi e degli algoritmi di data mining utilizzati.
- Applicare diverse tecniche di discretizzazione (operatori *Discretize by binning*, *Discretize by frequency*, *Discretize by size*) e confrontare i risultati.

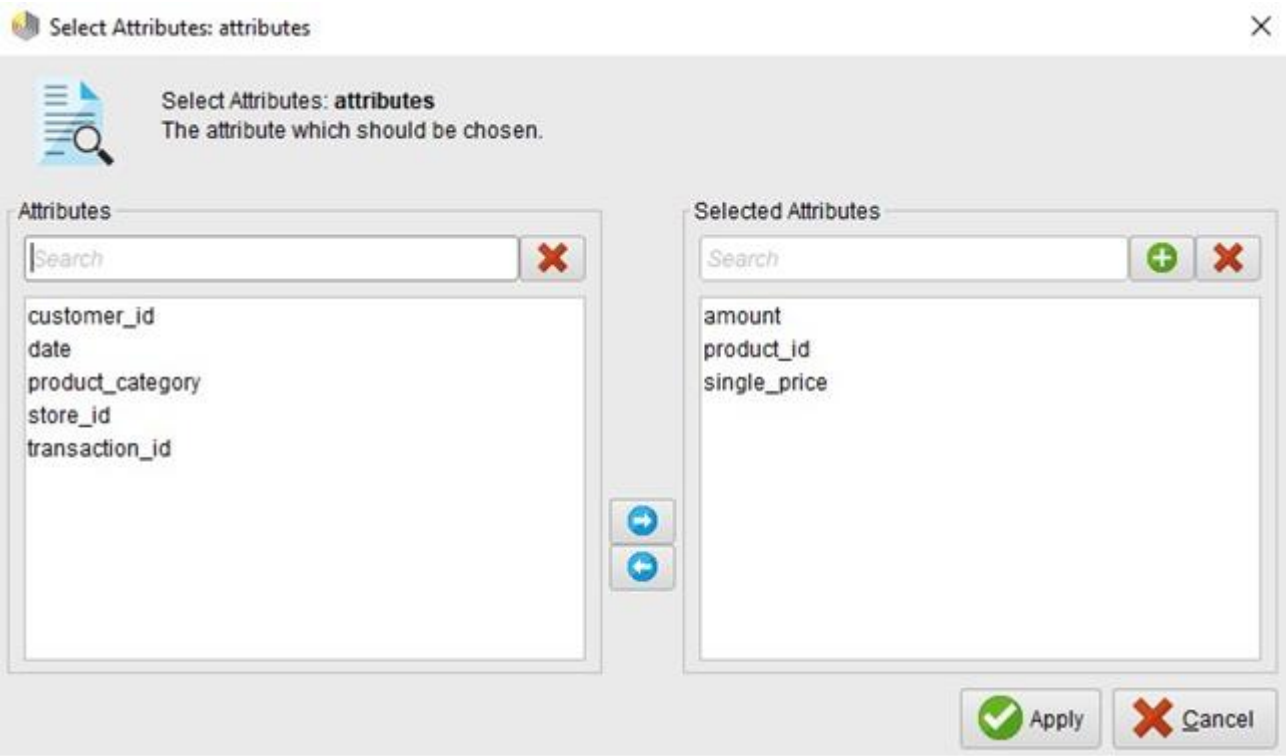
Analizzare i risultati della discretizzazione scelta.

<u>Bonus: dati strutturati con attributi continui</u>

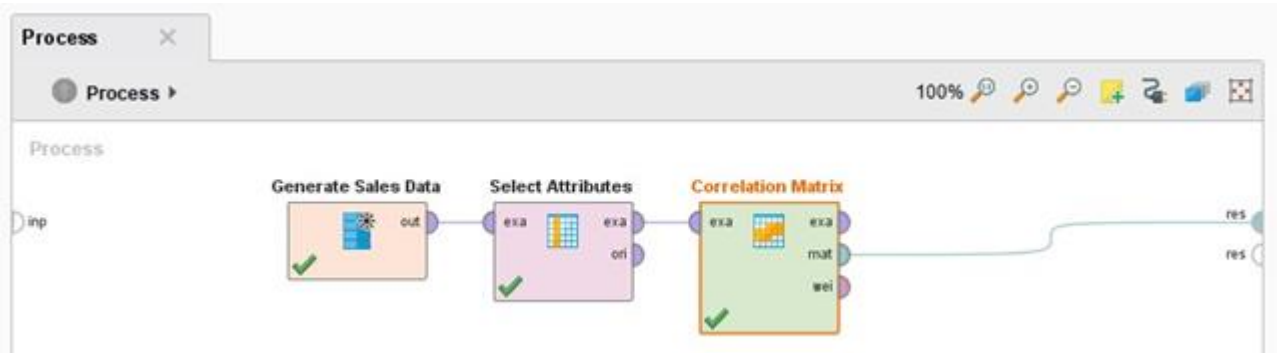
- Utilizzare l'operatore **Generate Sales Data**. Nella barra dei parametri, settare il parametro number examples a 100.



- Selezionare solo gli attributi numerici. Utilizzare l'operatore Select Attribues.

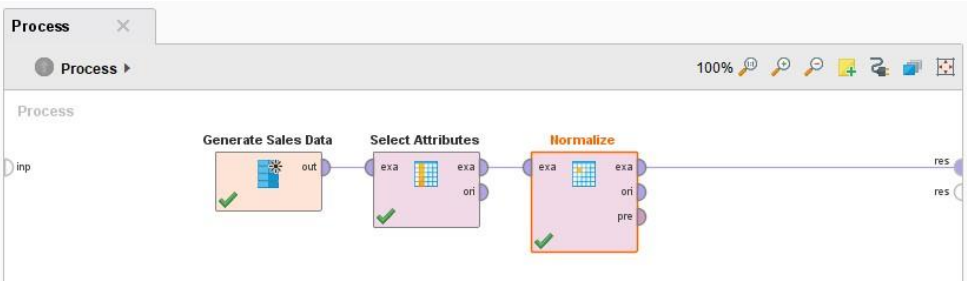


- Analizzare la correlazione tra coppie di attributi (operatore *Correlation Matrix*). Inserire l'operatore "Correlation Matrix" in coda al processo e visualizzare la rispettiva matrice collegando il plug-in del blocco denominato "mat" al plug-in "Result" sulla destra della finestra del processo principale. Il processo così generato sarà analogo al seguente:



- Esiste qualche correlazione elevata?
- Eliminare l'operatore correlation matrix (selezionarlo con il mouse e premere canc).

- Discutere l'eventuale necessità di applicare un processo preliminare di normalizzazione in funzione degli obiettivi dell'analisi e degli algoritmi di data mining utilizzati (operatore *Normalize*).

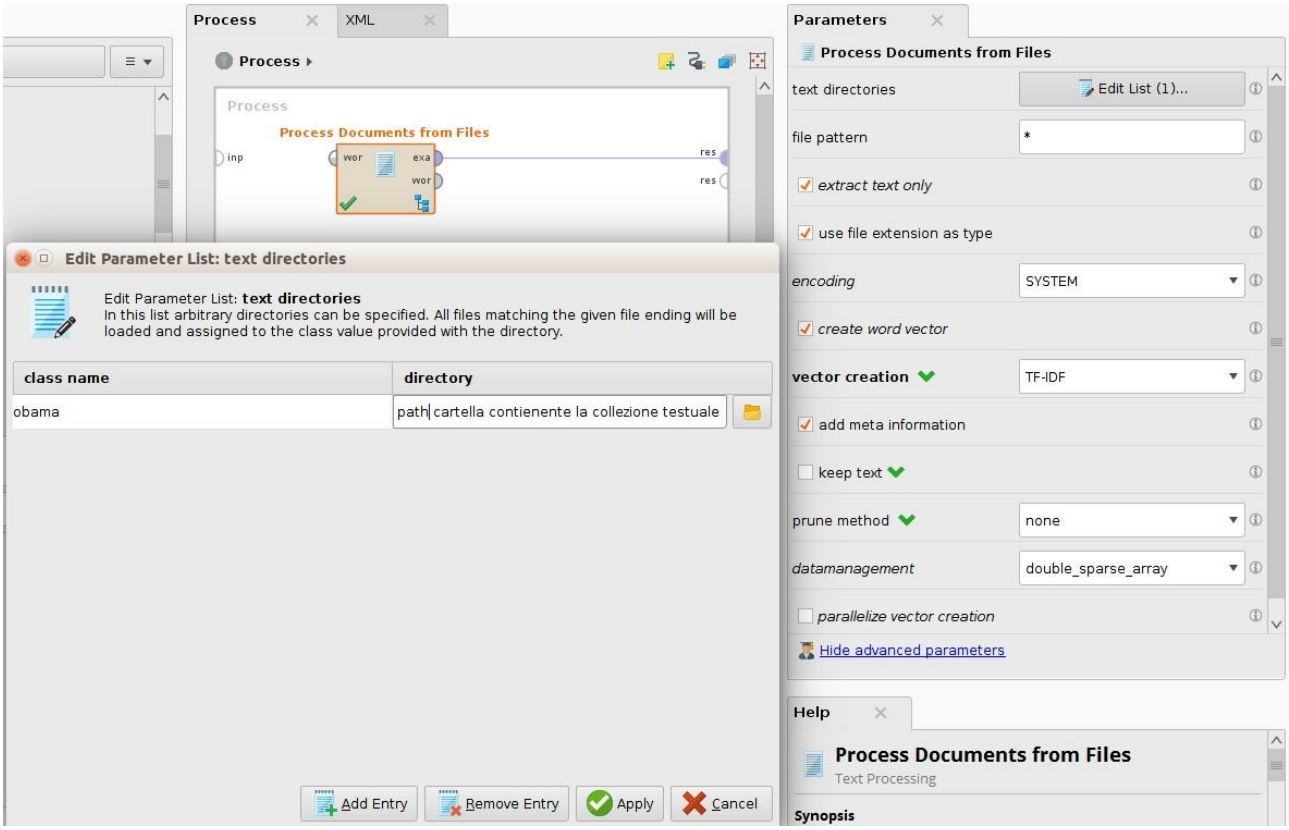


- Quale normalizzazione sceglieresteste? (nel caso non compaia method, cliccare su *Show advanced parameters*)
- Quando è utile normalizzare i dati?

Preparazione dei dati testuali

Obiettivo 1 – Import dei dati

- Importare il dataset Wikipedia in Rapid Miner (operatore *Process Documents From Files*).



- Se volete avere l'informazione del testo all'interno dei risultati, spuntate la voce **Keep Text** nel pannello dei parametri dell'operatore **Process Documents from Files**.

Obiettivo 2 – Generazione dei token, stopwords e stemming

Il **TF-IDF** (*Term Frequency–Inverse Document Frequency*) è una funzione nota nel text mining utilizzata per misurare l'importanza di un termine rispetto ad una collezione di documenti. Il TF-IDF aumenta

proporzionalmente al numero di volte che il termine è contenuto nel documento, ma cresce in maniera **inversamente proporzionale** con la frequenza del termine all'interno della collezione. In questo modo si possono penalizzare le parole molto frequenti che non danno rilevanza alla collezione e dare più importanza ai termini che in generale sono poco frequenti ma più rilevanti per l'analisi.

$$\mathrm{tfidf}_{i,j} = \mathrm{tf}_{i,j} \times \log\left(\frac{\mathrm{N}}{\mathrm{df}_i}\right)$$

$\mathrm{tf}_{i,j}$ = total number of occurences of i in j
df_i = total number of documents (speeches) containing i
N = total number of documents (speeches)

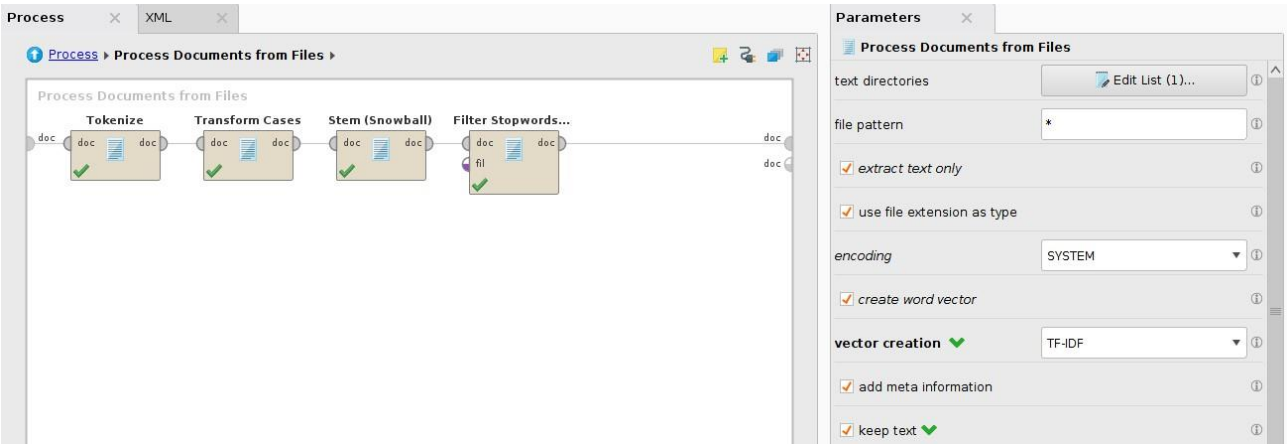
L'operatore *Process Document from Files* ammette un sottoprocesso per poter pulire il dataset e trasformarlo in una tabella chiamata matrice documenti*termini. La tabella avrà una riga per ogni documento della collezione presente nella cartella letta e una colonna per ogni termine presente all'interno della collezione.

- Prima di creare la matrice, guardare l'output. Nel sottoprocesso (per entrare nel sottoprocesso, fare doppio click con il tasto sinistro sull'operatore *Process Document from Files*) collegare le uscite come in figura.

- Per tornare al processo principale, cliccare la freccia blu di fianco a *Process*.
- Collegare l'uscita exa con res ed eseguire il processo.

- Applicare i passi di pre-processing visti nell'esercitazione precedente. Doppio click sull'operatore **Process Documents from Files**. Verrà aperto un sottoprocesso. Utilizzare i seguenti blocchi:
  - Il blocchetto **Tokenize**: splitta ogni documento della collezione Obama in un vettore di parole. L'ordine delle parole non sarà più rispettato. Secondo te ha importanza ai fini dell'analisi? (Settare il parametro non letters).
  - Il blocchetto **Transform Cases**: Trasforma il testo in maiuscolo o minuscolo.
  - Il blocchetto **Stem (Snowball)**: Riduce le parole alla propria radice. La radice è quell'elemento linguistico irriducibile (non ulteriormente suddivisibile) che esprime il significato principale della parola. (Utilizzare la lingua italiana).
  - Il blocchetto **Filter Stopwords (Dictionary)**: Permette di eliminare le parole definite Stopword, parole che non hanno un particolare significato se isolate dal testo e quindi vengono ignorate dai programmi. Sono parole poco significative perché possono essere usate spesso all' interno delle frasi. Ad esempio articoli, congiunzioni e preposizioni non

caratterizzano il significato di un testo, possono essere eliminate a monte di una analisi text mining. Carica il file **stopwordsEnglish.txt** presente sul sito del corso.



- Utilizzare la codifica UTF-8 per il file delle stopword.
- Torna al processo iniziale cliccando sulla freccia blu sotto la voce Process.

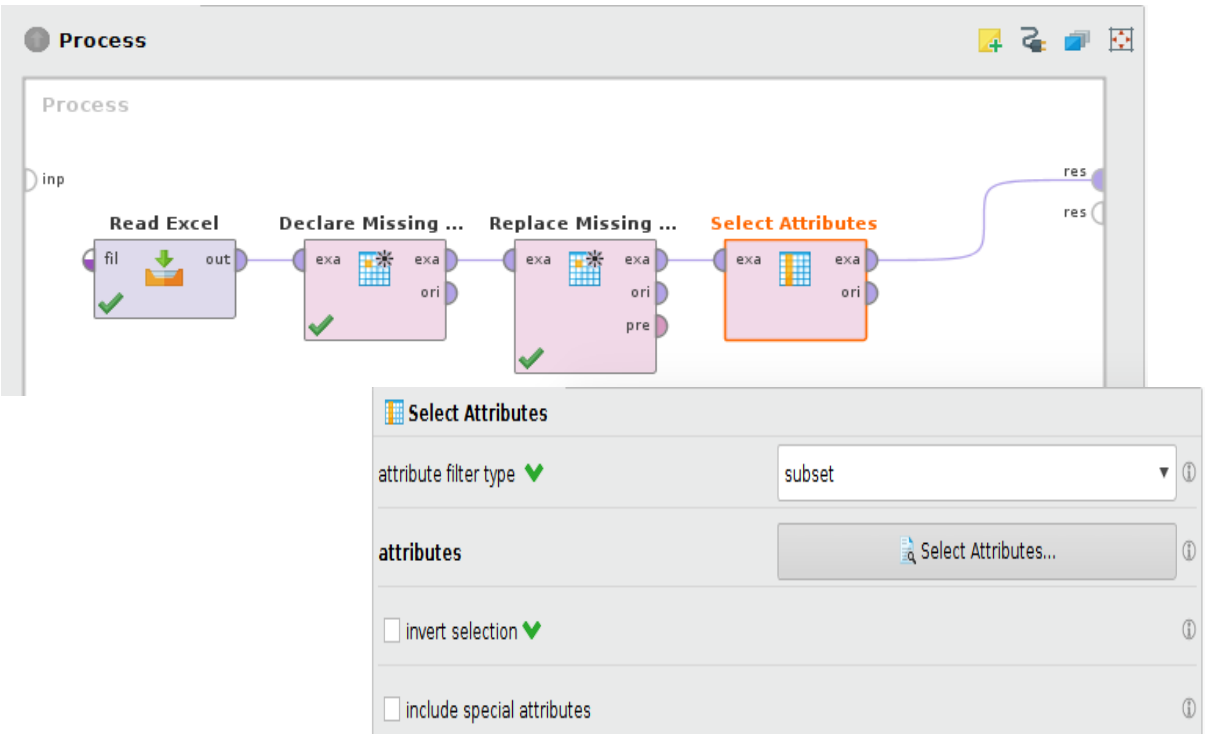
Clustering di dati strutturati

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.

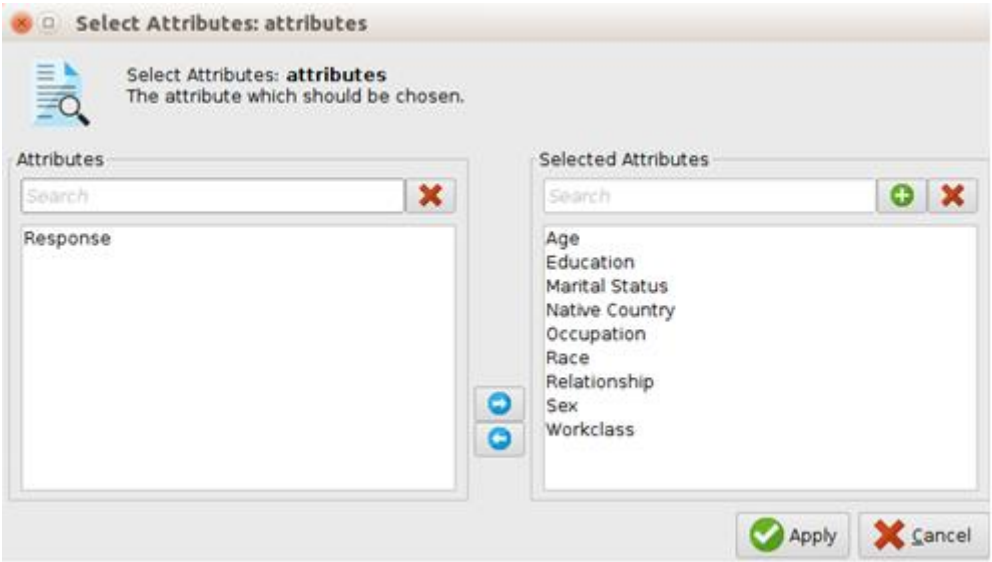
Obiettivo 1 – Import e preprocessing dei dati

Eseguire i diversi step di preprocessing, come imparato nella prima parte dell'esercitazione.

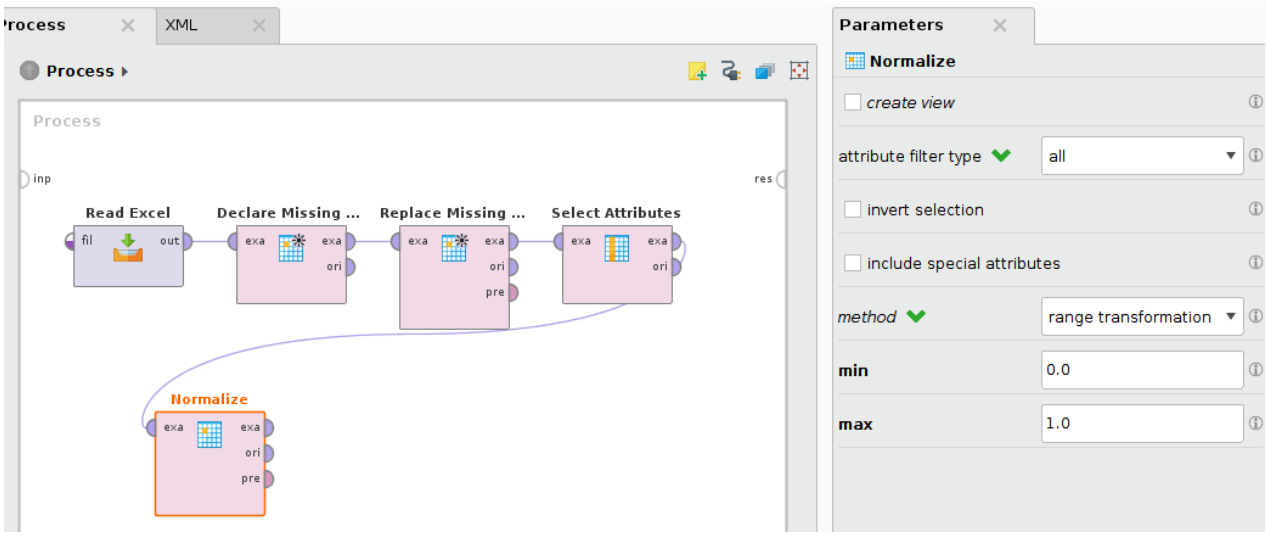
- In particolare, eseguire i gli step:
  - Import dati
  - Declare and replace missing values



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



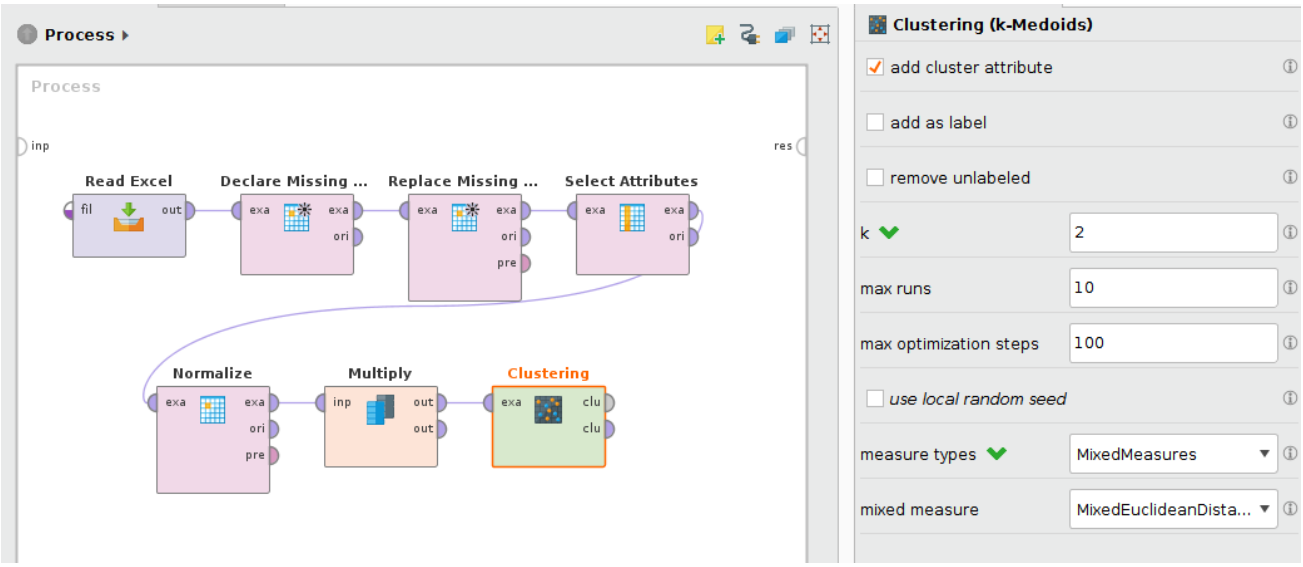
- **Normalizzare** i valori degli attributi numerici indicando come intervallo di valori [0-1] utilizzando l'operatore **Normalize**. L'unico attributo che verrà normalizzato è l'attributo età.



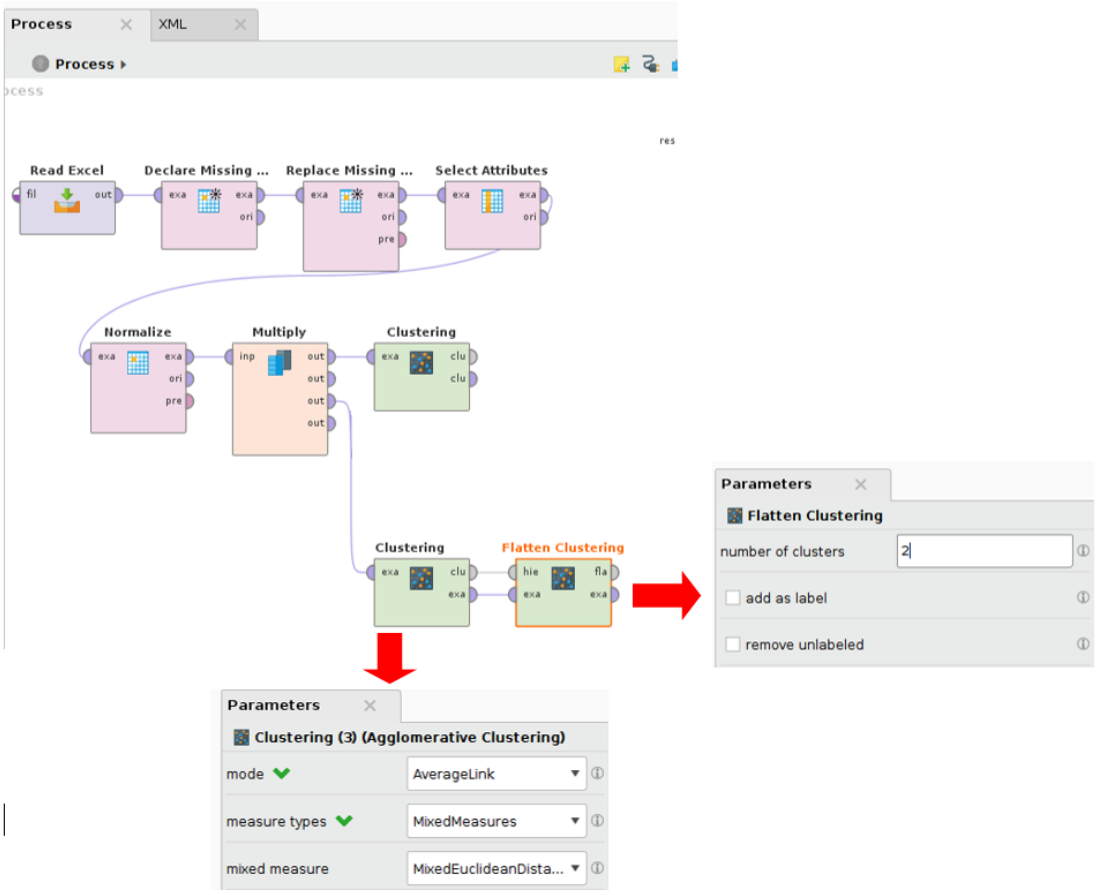
- Quando avete bisogno di utilizzare lo stesso input per diversi algoritmi, utilizzate l'operatore **Multiply**. Nei prossimi step verranno comparati diversi algoritmi di clustering.

Obiettivo 2 – Clustering dei dati

- 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?

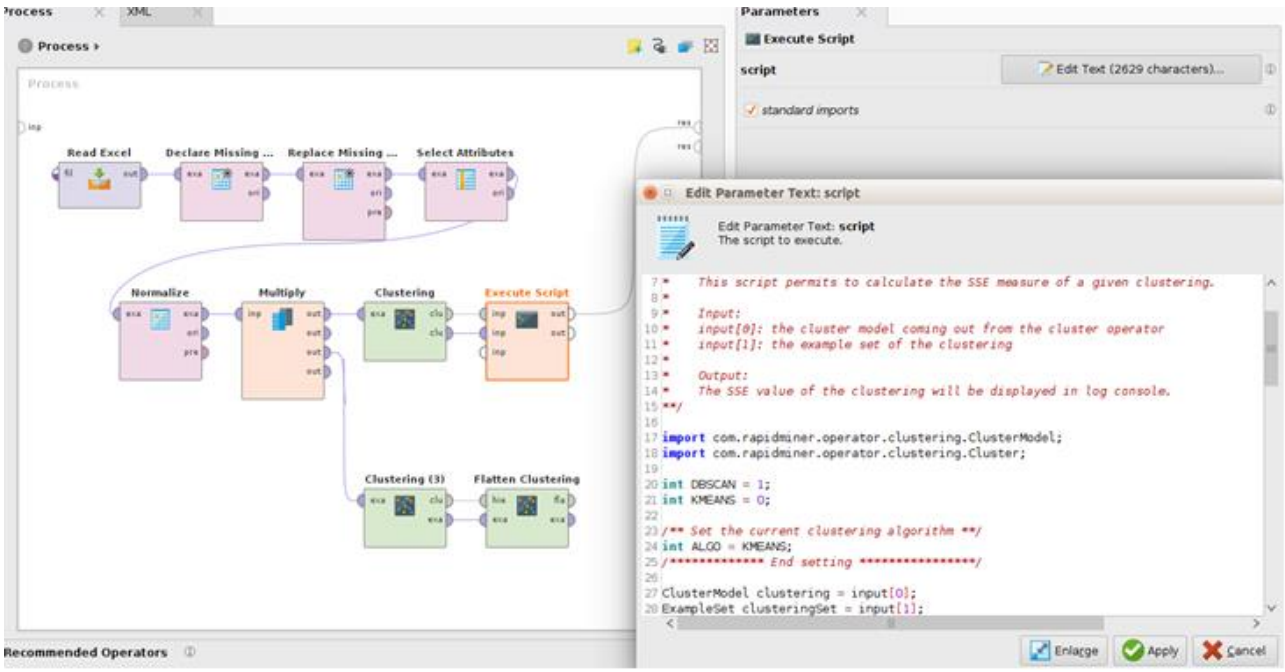


- Applicare l'algoritmo di clustering *Agglomerative (Agglomerative Clustering)*. Selezionare due cluster dal risultato dell'algoritmo di clustering *Agglomerative* utilizzando l'operatore *Flatten Clustering*. Esegui il processo e confrontare il risultato ottenuto con quello prodotto dall'algoritmo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



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

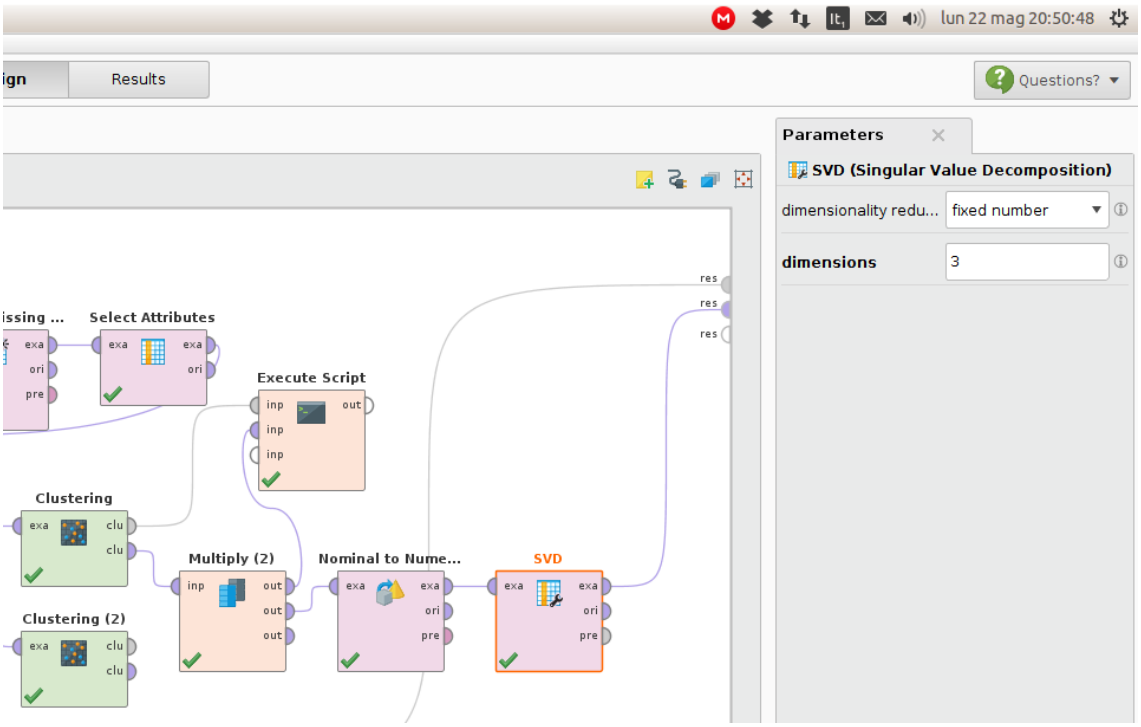
- Calcolare l'**SSE** (Sum of Squared Errors) dei cluster generate con l'algoritmo *K-Means*. Per il calcolo usare lo script **sse.script** disponibile sul sito del corso; copia e incolla il codice dello script all'interno dell'opportuna textbox dell'operatore ***Execute Script***. *(NB. Settare all'interno del codice il corretto algortitmo (KMEANS o DBSCAN)).*



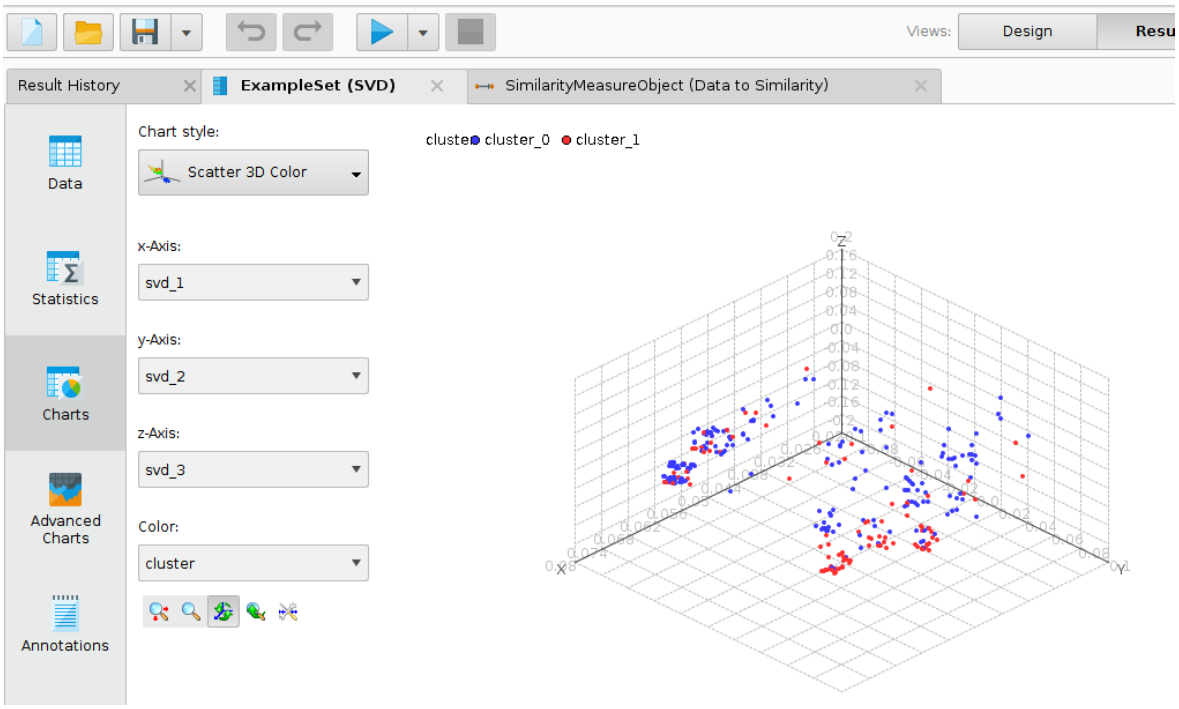
- Eseguite il programma. Il valore di SSE viene riportato nel tab **Log** della pagina dei risultati.
- Rieseguire il processo di valutazione precedente per **differenti valori di K** per l'algoritmo **K-medoids**. Come scegliere il valore ottimale di K? (Suggerimento: utilizza il metodo del ginocchio o del gomito visto durante la spiegazione del clustering in classe).
- **Eliminare** il ramo del Clustering agglomerativo (o disabilitarlo con il tasto destro).

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

- Analizzare la qualità del clustering generato mediante una tecnica di riduzione della dimensionalità dei dati, nota come ***Singular Value Decomposition*** (SVD). SVD permette di proiettare dei dati 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?



- Moltiplica l'uscita del clustering K-medoids e utilizza l'operatore **Nominal to Numerical**. Collegalo all'operatore **SVD** e esegui il programma. Salva il processo sul tuo computer.



Clustering di dati testuali

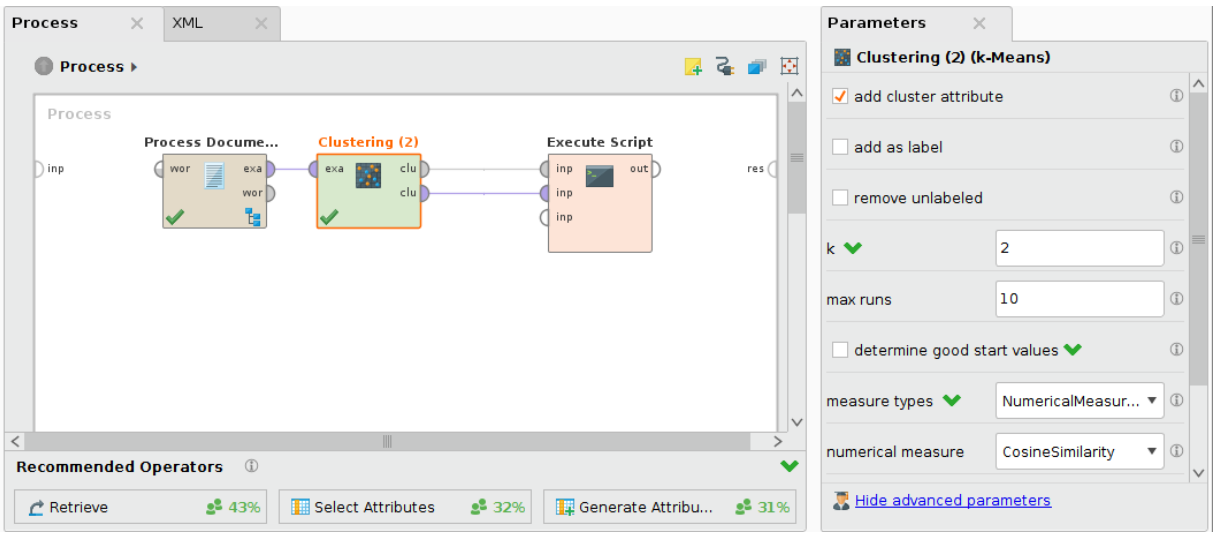
La seconda parte di questa esercitazione prevede l'analisi attraverso l'algoritmo K-Means della collezione di articoli denominata *Wikipedia*. Scompatta la cartella presente sul sito del corso ed esegui i passi seguenti.

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 parte dell'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 **Cosine Similarity**. Come viene calcolata? Rispetto a cosa si differenzia dalla norma Euclidea?

- Riesegui il K-means diverse volte per cercare il parametro K migliore. Utilizza l'SSE come metrica di valutazione (guarda la prima parte dell'esercitazione).
- Visualizza i cluster trovati attraverso la tecnica SVD.
- Ha senso calcolare la correlazione prima di eseguire il K-means? Perché?