

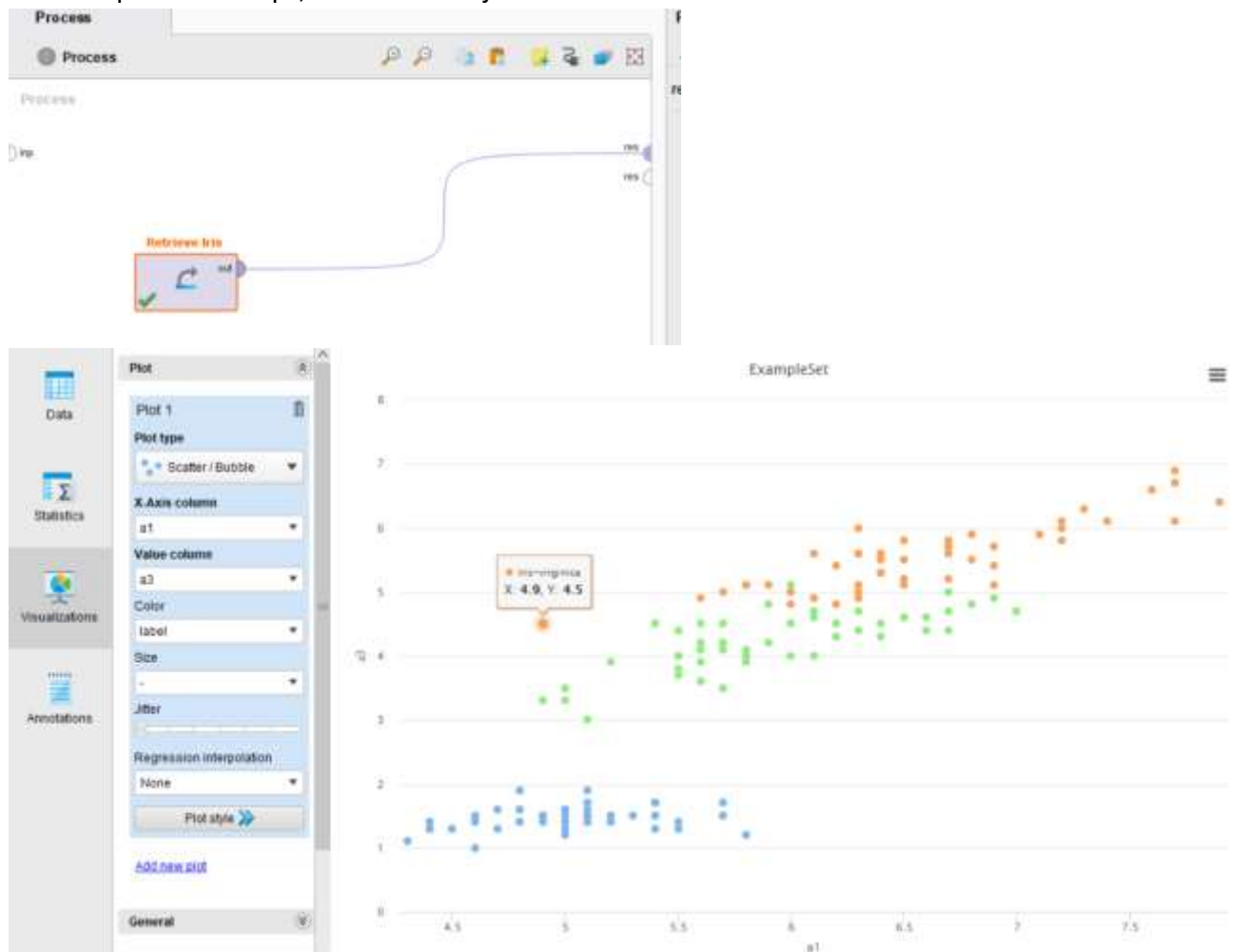
TD5

Regroupement sous RapidMiner

Le jeu de données Iris (dans Samples/Data) contient un ensemble d'enregistrements chacun décrivant une fleur d'Iris. Chaque fleur est décrite par un identifiant Id, le nom de la variété d'iris et quatre valeurs numériques a1, a2, a3 et a4 qui sont des mesures de taille de certains aspects de la fleur (pétale, pistil, ...).

Comme dans ce jeu de données il y a 3 types d'Iris : setosa, versicolor et virginica, nous allons voir dans quelle mesure une technique de regroupement, en l'occurrence K-Means, ne serait-elle pas capable de séparer/identifier les 3 groupes de fleurs.

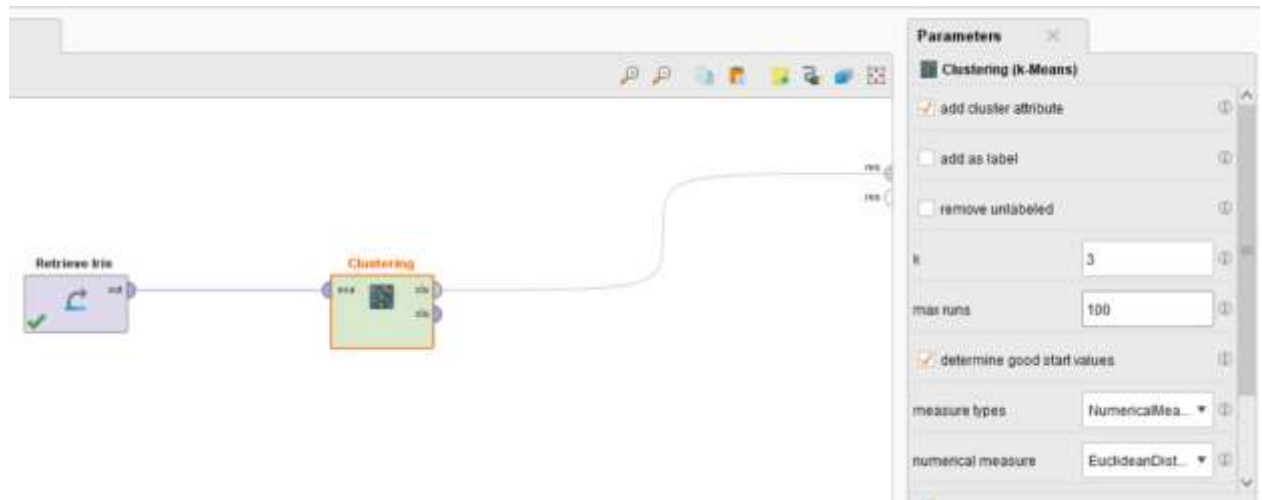
1. Dans un premier temps, visualisons le jeu de données Iris



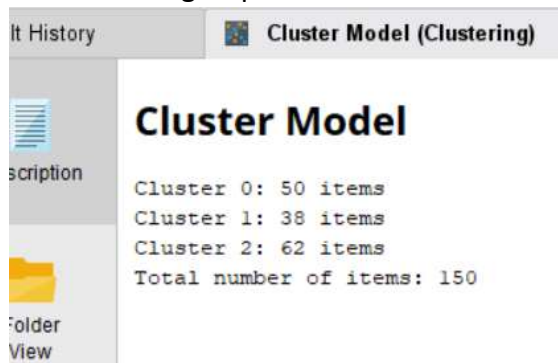
En représentant les enregistrements, on observe que les fleurs de la variété setosa (couleur bleue) sont facilement séparables du reste. On peut s'attendre à ce que tout algorithme de regroupement pourrait « aisément » construire le groupe setosa mais aura plus de mal à séparer les deux autres variétés. Varier les représentations en changeant à chaque fois la paire de variables représentées sachant que sur la figure ci-dessus, ce sont les variables a1 et a3 qui sont utilisées. Vérifier à chaque fois s'il n'y a pas de séparation claire entre les 3 variétés.

2. Utiliser vos connaissances pour afficher le nombre d'enregistrements par variété.

3. On applique K-means à notre jeu de données en (i) fixant le nombre de groupes à 3 et en (ii) choisissant la distance euclidienne comme mesure de distance.



4. On obtient 3 groupes dont voici le contenu en termes d'effectif



5. Clairement, nous n'avons pas réussi à séparer les 3 variétés de fleur (voir résultat obtenu pour la question 2). Afficher le nombre d'enregistrements par variété de fleur ET par numéro de cluster. Est-ce qu'on a réussi à séparer les fleurs setosa du reste comme on pouvait l'espérer ?
6. Remarquer que nous avons appliqué K-means sur un jeu de données que nous n'avons pas normalisées. Ajouter une étape de normalisation (opérateur Normalize avec paramètre Z-transformation pour remplacer chaque valeur par son Z-score). Comparer le résultat obtenu avec le précédent.
7. Est-ce qu'un arbre de décision arrive à séparer les 3 variétés ? Pour le vérifier, il faut (i) construire un arbre de décision à partir du jeu Iris puis (ii) appliquer l'arbre obtenu sur ce même jeu de données Iris en enfin (iii) afficher les performances.