Data Mining – TP
2020/2021

22 novembre 2020

Résumé

Ce TP se déroule en 2 parties : découverte de Knime via des exemples simples et application d'algorithmes de clustering pour la détection de points d'intérêt ou d'événements géo-localisés dans des médias sociaux. Il est aussi possible de ne pas utiliser Knime et de faire le TP en python. Vous êtes libres d'utiliser Knime ou pas.

Ce TP est à rendre (uniquement la Section 4 **) avant la prochaine séance sur Tomuss : 1 rapport (.pdf) qui décrit les traitements faits ainsi que les résultats, et une archive (ou un lien) contenant le code. Il est possible de le réaliser à plusieurs (3 max), merci de ne faire qu'un seul dépôt sur Tomuss par groupe !**

Si vous avez compris le fonctionnement de Knime, vous pouvez aller directement à la Section 4. N'hésitez pas à consulter des exemples de workflows disponibles (dans KnimeExplorer, clic droit sur EXAMPLES puis se connecter.

1 Installation de Knime

Knime est un logiciel disponible gratuitement. Le lien suivant `https://www.knime.org/downloads/overview` permet le téléchargement de ce logiciel. Cette page demande tout d'abord une inscription. Ensuite, elle fournit les installables pour les différents systèmes d'exploitation.

Knime offre une interface graphique conviviale. La Figure 1 présente les rubriques principales de **Knime**.

Un projet **Knime** est tout simplement un workflow, plus précisément, c'est une suite de nœuds où chacun fait une opération spécifique. La Figure 2 illustre un workflow. Par exemple, Le nœud **«File Reader»** permet de lire un fichier csv contenant les données.

2 Téléchargement de données

Les données sont disponibles à l'adresse suivante : `goo.gl/nu2olb`. Il contient les deux jeux de données suivants :

1. **Iris :** C'est une base de différentes variétés d'iris. Chaque iris (objet) est décrit par 5 attributs :
   — Longueur du sépale.
   — Largeur du sépale.
   — Longueur du pétale.

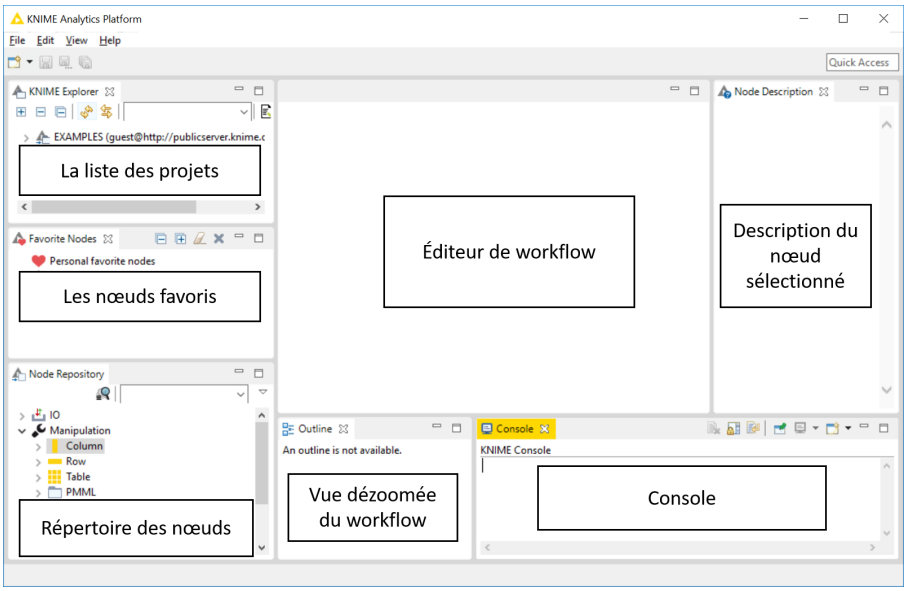

FIGURE 1 – Interface graphique de **Knime**

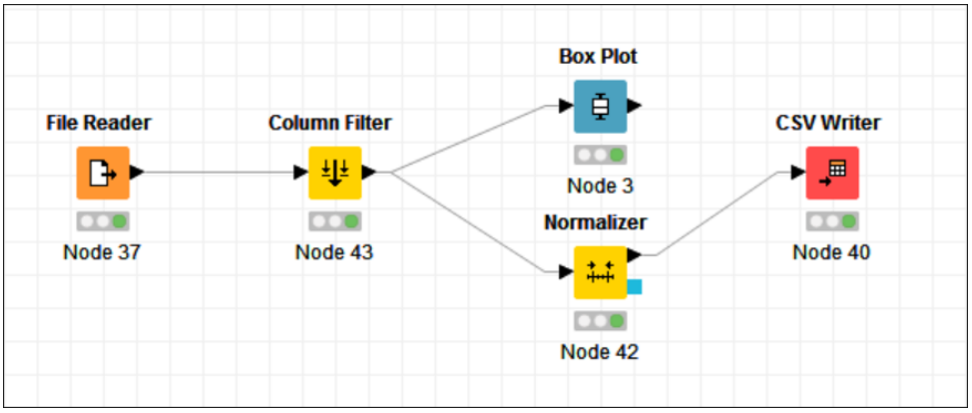

FIGURE 2 – Exemple d'un workflow **Knime**

— Largeur du pétale.
— Variété d'Iris (Classe), il existe 3 classes : Iris Setosa, Iris Versicolour, Iris Virginica.

La base contient 150 instances (50 fleurs pour chaque classe).

2. **Bears :** C'est une base où chaque objet décrit un ours à l'aides des attributs suivants :
   — Age : L'age en mois.
   — Month : le mois où les mesures sur cet ours ont été prises.
   — Sex : le sexe de l'ours.
   — Headlen : taille de la tête en pouce.
   — Headwth : la largeur de la tête en pouce.
   — Neck : Le tour du cou en pouce.
   — Length : La taille du corps en pouce.
   — Chest : le tour de la poitrine en pouce.
   — Weight : le poids de l'ours en livres.

   La base contient 54 enregistrements (ours).

3 Premiers pas - Jeu de données Iris

L'objectif de cet exercice est l'exploration et la fouille du jeu de données Iris. Tout d'abord, on crée un nouveau projet **(File → New → New Knime Workflow)**.

Chargement de données : Pour charger les données, on sélectionne le nœud **«File Reader»**. Ensuite, on ouvre la fenêtre de configuration du nœud (clique droit sur le nœud, choisir **«configure»**). Sur cette fenêtre, on spécifie le fichier de données à charger.

Exploration de données : Dans cet étape, on fait l'exploration et la visualisation des données.

- Pour identifier facilement les trois classes de fleurs dans les visualisations, il est intéressant de colorer les données (attribuer une couleur à chaque classe). Pour ce faire, on peut utiliser le module **«Color Manager»**.
- Knime regroupe les nœuds de visualisation dans la rubrique **«Views»** située dans le répertoire des nœuds. Par exemple, on peut utiliser le nœud **«Box Plot»**, **«Conditional Box Plot»**, et **«Scatter Plot»**. On peut voir la corrélation entre les attributs en utilisant le nœud **«Linear Correlation»**. La Figure 3 montre un exemple de workflow pour cette étape.

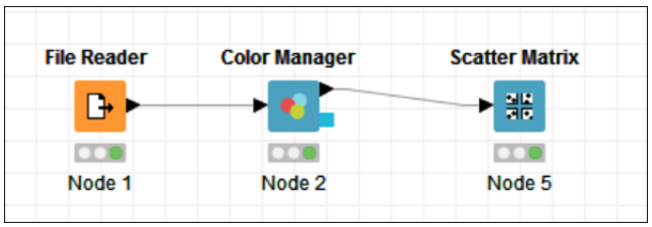

FIGURE 3 – Exploration de données Iris

Clustering : En utilisant les 4 attributs numériques des fleurs, on vise à partitionner les données.

1. Pour équilibrer l'impact des attributs sur le clustering, on normalise ces attributs. On peut utiliser le nœud **«Normalizer»**. Ce nœud offre plusieurs types de normalisation, par exemple : la normalisation min-max.
2. Pour faire le clustering, on peut choisir une des méthodes offertes dans la rubrique **(Analytics → Mining → Clustering)**.
3. Enfin, on peut visualiser le résultat pour voir la relation entre les clusters trouvés et les différents attributs.

La Figure 4 montre un exemple de workflow pour cette étape.

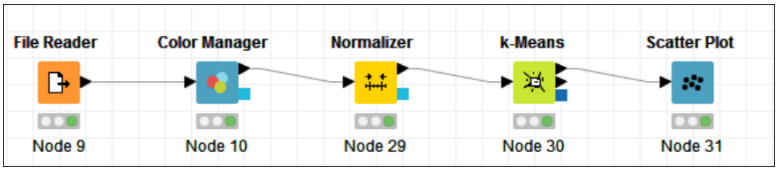

FIGURE 4 – Clustering de données Iris

4 Détection de points d'intérêt

L'objectif de cet exercice est de trouver de manière automatique des points d'intérêts intéressants dans la ville de Lyon, définis par une activité forte de prise de photos. Pour cela, on veillera à détailler chaque étape du processus de KDD (à l'aide du logiciel Knime).

- Compréhension, nettoyage des données, visualisation et statistiques. Il faudra par exemple : vérifier la cohérence des données (dates, positions GPS) ; supprimer les doublons, afficher les points sur une carte monde, ... On utilisera entre autres les noeuds File Reader, GroupBy, Row Filter, Geo- Coordinate Row Filter, OSM Map View, Missing Value.
- Sélection des attributs intéressants pour l'analyse courante (Column Filter).
- Fouille de données avec du clustering : comparer, discuter k-means et DBSCAN. Quel est l'algorithme le plus adapté pour détecter des points d'intérêt ?
- Une fois des points d'intérêt détecté, les caractériser à en considérant les légendes et tags associées aux photos de chaque cluster. On pourra utiliser des règles d'association [1] . On pourra aussi afficher des nuages de mots associés à chaque cluster.
- Évaluation, interprétation, visualisation (sur une carte), discussion des résultats. Comment votre analyse peut-elle aider le Grand Lyon ? Quelles connaissances lui apporte-t-elle ? La dernière étape est souvent négligée, mais elle est capitale. Un résultat de fouille de données ne sert à rien s'il n'est pas actionnable : il doit servir à quelque chose, et le mode d'emploi doit être donné.

Les données sont disponibles sur le lien suivant :
`http://liris.cnrs.fr/~mplantev/ENS/DMTP/flickr_data.csv`

1. Consulter le tutoriel proposé par Knime sur la fouille de texte, construire une table document/terme binaire. On peut alors y chercher chercher des motifs fréquents de termes pour chaque cluster, ou encore des règles d'association qui concluent sur des numéro de cluster