Travaux pratiques DataMining et Visualisation des données

TP1 – Clustering Knime

Alison PATOU
patou.alison@gmail.com
4/12/2020

Introduction

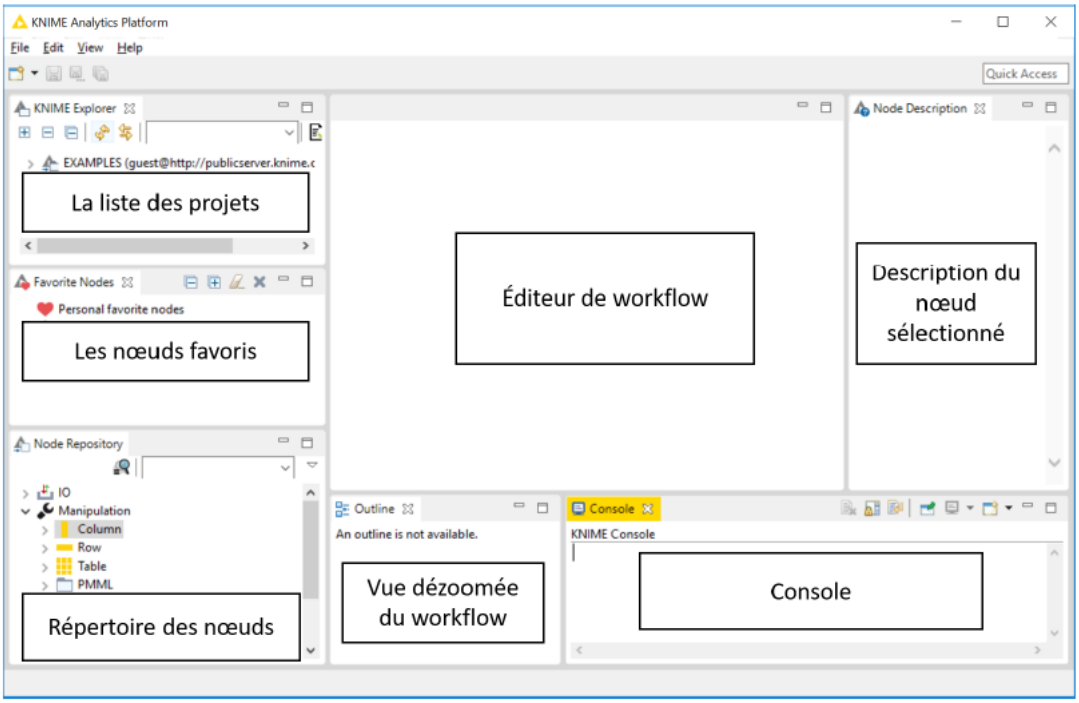
Knime est un outil de préparation de données, permettant également de créer des visualisations de données et de les analyser.
Ce premier TP a pour objectif de vous familiariser avec ce logiciel et notamment avec le concept principal qui sera manipulé tout au long des différents TPs : le workflow et notamment dans notre cas, le workflow Data Science.
Un workflow peut être vu comme une succession de briques (composants élémentaires). Chaque brique possède une fonctionnalité bien précise. En enchainant ces composants, il est alors possible de réaliser des analyses de données plus segmentées, sans avoir besoin de savoir coder et en facilitant déjà, une automatisation du processus.

1. **Télécharger via ce lien Knime** puis installer-le sur votre ordinateur.

2. Présentation de l'interface

L'interface principale par défaut du logiciel est présentée dans l'image ci-dessous et est principalement composée de 6 parties :

- La liste des workflows disponibles (accessibles via Internet ou stockés localement) est disponible dans le coin supérieur gauche de la fenêtre. C'est ici que vous retrouverez vos workflows une fois que vous les aurez enregistré.
- La liste des composants classés thématiquement est disponible dans le coin inférieur gauche
- Une vue générale du workflow est présentée en bas
- A sa droite, il est possible de voir la console, qui permet de voir les traces d'exécution d'un workflow.
- La partie droite de la fenêtre propose une documentation sur le composant sélectionné. Ainsi vous pourrez découvrir tous les composants disponibles sur Knime
- Enfin, la fenêtre principale au milieu de l'écran, permet de créer et composer son workflow et d'interagir avec celui-ci (configurer les composants, les repositionner, les relier, ...)



3. Au démarrage, le logiciel vous demande de sélectionner votre Workspace. Il s'agit d'un répertoire dans lequel seront stockés les workflows que vous créerez lors des différents TPs.

4. Nous allons créer un workflow vierge qui contiendra le travail de ce TP. Pour ce faire, cliquez sur :
   File > New > New KNIME Workflow.
   Cliquez ensuite sur Finish pour basculer vers la fenêtre principale et disposer de votre workflow prêt à être travaillé.

Données

Nous allons travailler sur les données Adult, comprenant les informations relatives aux salaires annuels de 34 000 personnes (avec seulement l'indication de « supérieur » ou « inférieur » à 50k) ainsi qu'une dizaine d'informations complémentaires sur leurs âges, le nombre d'heures de travail par semaine, etc. Ce dataset est souvent utilisé pour faire de la prédiction. Vous pouvez télécharger l'ensemble des données **sur ce site** et notamment les deux fichiers :

- Adult.Names : documentation sur les données
- Adult.Data : les données

Travail Demandé

1. Import des données

Nous allons dans un premier temps importer les données sous Knime, pour ce faire Glisser-Déposer le composant *File Reader* (dans la sous-catégorie Read) dans votre workflow. Le composant apparaît au milieu de la fenêtre principale. Pour le configurer, double-cliquez dessus.
Renommez l'ensemble des colonnes avec les indications de celles-ci du fichier Adult.Name.

2. Discrétisation de l'Age

Pour cela, trouvez le composant *Auto Binner* (dans la catégorie Data Manipulation>Binning) que vous allez glisser-déposé dans le workflow et ensuite configurer pour discrétiser la colonne Age, avec 5 classes de même taille et dont le nom sera donné par les extrémités de chaque intervalle.

Reliez ces deux composants entre eux à l'aide de la souris. Nous allons maintenant l'exécuter. Pour cela, vous devez repérer en haut de l'interface un bouton vert, avec deux triangles orientés vers la droite. Ce bouton permet de lancer tous les nœuds exécutables présents dans le workflow. Une fois que ce workflow a été exécuté avec succès (l'indicateur d'état est en vert), il est possible de visualiser la sortie de ce composant. Pour cela, faites un clic droit sur ce composant. Vous vous trouvez devant un menu contextuel. Les deux derniers items de ce menu correspondent aux deux sorties du composant.

La sortie qui nous intéresse est **Binned Data**. En faisant un clique-droit sur le composant, vous pouvez visualiser le résultat de la discrétisation.

3. Filtre sur l'âge

Filtrer pour ne conserver que les personnes qui ont un âge compris entre 18 et 60 (inclus). Pour ce faire, vous allez devoir utiliser le composant *Row Filter* (Data Manipulation > Row > Filter).

4. Filtre avec Regex

Excluez les lignes dont la colonne Education matche avec l'expression régulière .*th.*. grâce au même composant utilisé précédemment.
Dans la mesure où on modifie l'ensemble des valeurs possibles pour cet attribut, il est nécessaire de connecter ce filtre à un autre composant : *Domain Calculator*.

5. Suppression colonne

Supprimer la colonne Age à l'aide du composant *Column Filter* dans Data Manipulation > Column > Filter.

6. Modification workflow

Réarrangez votre workflow pour que la discrétisation des âges soit effectuée entre le filtre des âges (de 18 à 60) et la suppression de la colonne Age.

7. Visualisations

L'ensemble des composants se trouves dans la catégorie Views. Si vous souhaitez faire des modifications de couleurs, vous allez devoir utiliser le composant *Color Manager*

A. Regardez la distribution de la colonne Age
B. Regarder la répartition du Salary, du Sex, puis du Salary ET du Sex
C. Visualisez la répartition du niveau d'étude en fonction de l'âge
D. Visualisez la distribution de la colonne Workclass
E. Faire 2 boxplots sur des colonnes qui vous semblent pertinentes.
F. Regarder la distribution de la colonne Hours Per Week selon la colonne Sex avec un boxplot

8. Statistiques descriptives

Vous allez regarder ce que fait le composant Statistics et étudier les tables en sortie.