Data Mining & Texte Mining

Travaux pratiques (KNIME)

TP 0 : *Présentation de KNIME*

Master Big Data & Aide à la décision

1ère Année / Semestre 2

ENSA Khouribga

Pr. DARGHAM ABDELMAJID

Année académique : 2018/2019

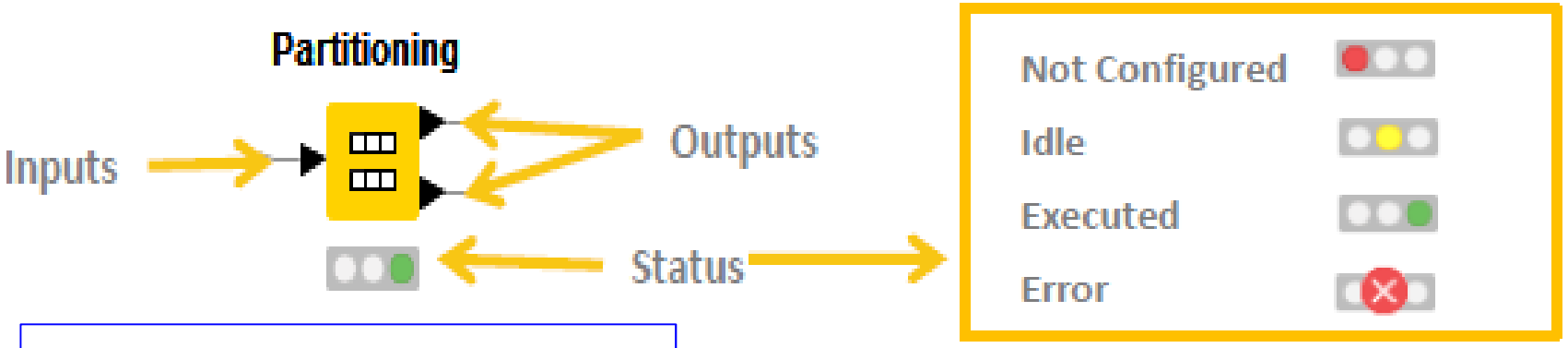
- **KNIME 3.7.1**
  - C'est un logiciel (**open source**) pour **l'analyse**, la **manipulation**, la **visualisation**, et le **reporting** des données.
  - Fondé sur le **paradigme de la programmation graphique**.
  - Fournit plusieurs extensions :
    - **Text Mining**
    - **Network Mining**
    - **Weka machine learning,** etc.

- **Ressources pour KNIME 3.7.1**
  - **KNIME pages (www.knime.org)**
    - **SOLUTIONS pour les exemples de workflows**
    - RESOURCES/**LEARNING HUB www.knime.org/learning-hub**
    - RESOURCES/**NODE GUIDE https://www.knime.org/nodeguide**
  - **KNIME Tech pages (tech.knime.org)**
    - **FORUM pour les questions et réponses**
    - **DOCUMENTATION pour docs, FAQ, changelogs, …**
    - **COMMUNITY CONTRIBUTIONS pour dev/instructions/third party nodes**
  - **KNIME TV sur le YouTube https://www.youtube.com/user/KNIMETV**

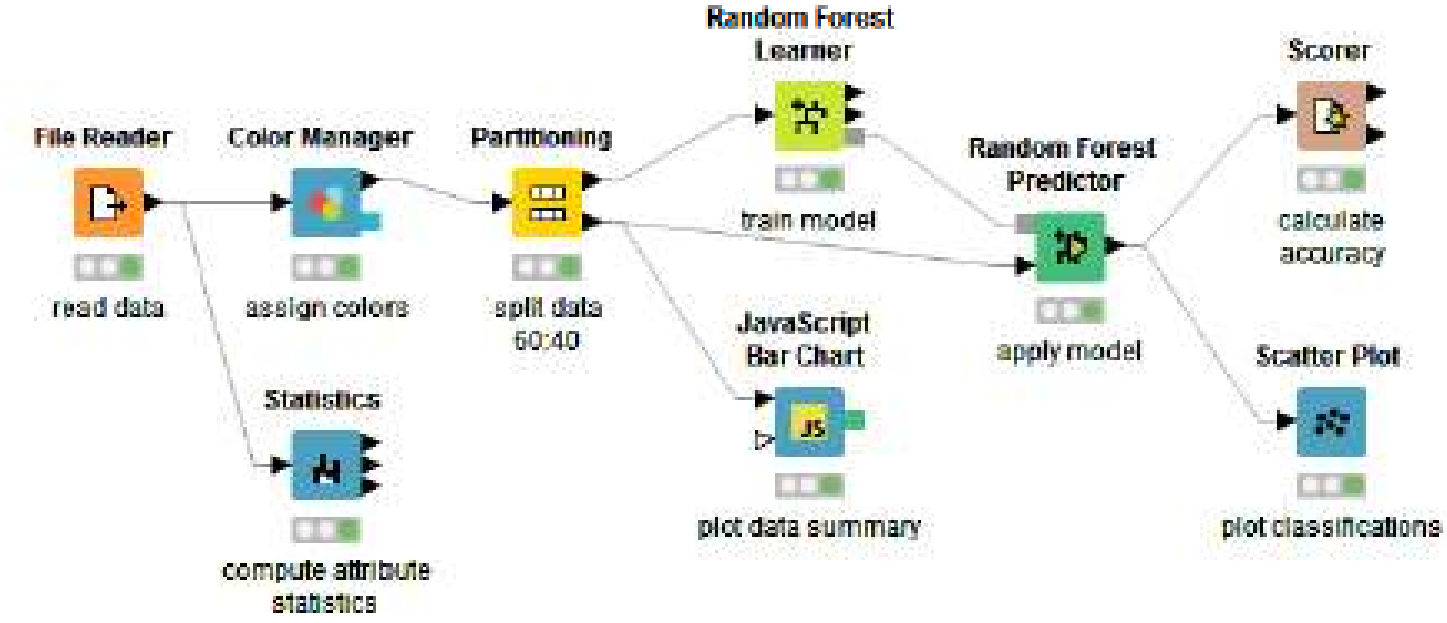
- **Atouts de KNIME**
  - **Open source**
  - **Facilité d'utilisation**
  - **Interface graphique simple**
  - **Il peut lire de très nombreux formats de données (Texte, CSV, Exel, Acces, XML, etc)**
  - **Il comporte de très nombreuses solutions pour prétraiter, analyser et visualiser des données et des résultats d'analyses**
  - **Extensible**

- **Notions clés de KNIME**
  - Les **nœuds : exécutent des tâches** sur des données.
  - Chaque **nœud** porte un **nom spécifique**, effectue une **tâche spécifique**, possède des **entrées** (**inputs**) et des **sorties** (**outputs**) et se trouve dans un **état** (**status**).
  - Le **workflow** : un **réseau de nœuds** ➔ les **nœuds** sont combinées pour former un **workflow** (c'est donc un **ensemble de travaux**).

TP 0 : Présentation de KNIME



Composants d'un nœud



Composants d'un workflow

- **Accès aux données**
  - **Bases de données**
    - MySQL, PostgreSQL
    - JDBC (Oracle, DB2, MS SQL Server)
  - **Fichiers**
    - CSV, TXT
    - Excel, Word, PDF
    - SAS, SPSS
    - XML
    - Images, Textes, networks

- **Accès aux données**
  - **Web, Cloud**
    - REST, Web services
    - Twitter, Google
- **Big Data**
  - **Spark,**
  - **HDFS support,**
  - **Hive,**
  - **Impala,**
  - **HP Vertica, In-database processing**

- **Transformation de données**
  - **Préprocessing**
    - **Row, column, matrix based**
  - **Data blending**
    - **Join, concatenate, append**
  - **Aggregation**
    - **Grouping, pivoting, binning**
  - **Feature Creation and Selection**

TP 0 : Présentation de KNIME

- **Analyse / Data Mining**
  - **Régression : linéaire / logistique**
  - **Classification : arbre de décision, ensembles, SVM, MLP, Naïve Bayes**
  - **Clustering : K-Means, DBSCAN, hiérarchique**
  - **Validation : cross-validation, scoring, ROC**
  - **Misc : PCA, MDS, Item Set Mining**
  - **External : R, Weka**

- **Visualisation de données**
  - **Interactive :**
    - **Scatter plot, histogram, pie charts, box plot**
    - **Highlighting (brushing)**
  - **JFreeChart**
  - **JavaScript**
  - **Misc :**
    - **Tag cloud, open street map, networks, molecules**
  - **External : R, Weka**