PRÉSENTATION DU LOGICIEL DE DATA-MINING WEKA(-PENTAHO) V3.8

26 octobre 2016

Auteurs :

- Eric YABAS
- Manel MERAD
- Marc HOLZWARTH
- Charlemagne ADECHINA

1 – Présentation
Origine

- Logiciel libre et gratuit (Licence GPL)
- Développé en Java
- Créé à l'Université de Waikato en Nouvelle-Zélande
- Le Weka est un oiseau endémique de la Nouvelle-Zélande

1 – Présentation
Historique

1992
- Début du développement de la version originale

1997
- Décision de développer Weka en Java

2005
- Weka reçoit le prix SIGKDD (Data Mining and Knowledge Discovery Service Award)

2006
- Pentaho acquiert une licence exclusive pour utiliser Weka dans sa suite décisionnelle open source (version communautaire et commerciale)

Pentaho est un fort contributeur
http://community.pentaho.com/projects/data-mining/

2 – Installation, bibliothèques et paramétrage

Installation très simple et rapide

1. **Téléchargez la dernière version stable** correspondant à l'environnement de votre poste de travail, à partir de l'un des deux sites suivants :
   - Communauté Pentaho : http://community.pentaho.com/projects/data-mining/
   - Université de Waikato : http://www.cs.waikato.ac.nz/ml/weka/downloading.html

2. **Lancez l'installation et suivez les instructions**.

   Et voilà, c'est prêt !

3. **Optionnel, installer des packages complémentaires (non indispensable pour débuter).**

   Weka comporte un mécanisme permettant d'étendre ses fonctionnalités (algorithmes d'apprentissage, outils…) via un gestionnaire de packages complémentaires. Celui-ci est accessible dans le menu *Tools* de WEKA et permet d'installer l'ensemble des packages publiés dans le dépôt en ligne officiel WEKA (D'autres dépôts peuvent être ajoutés).

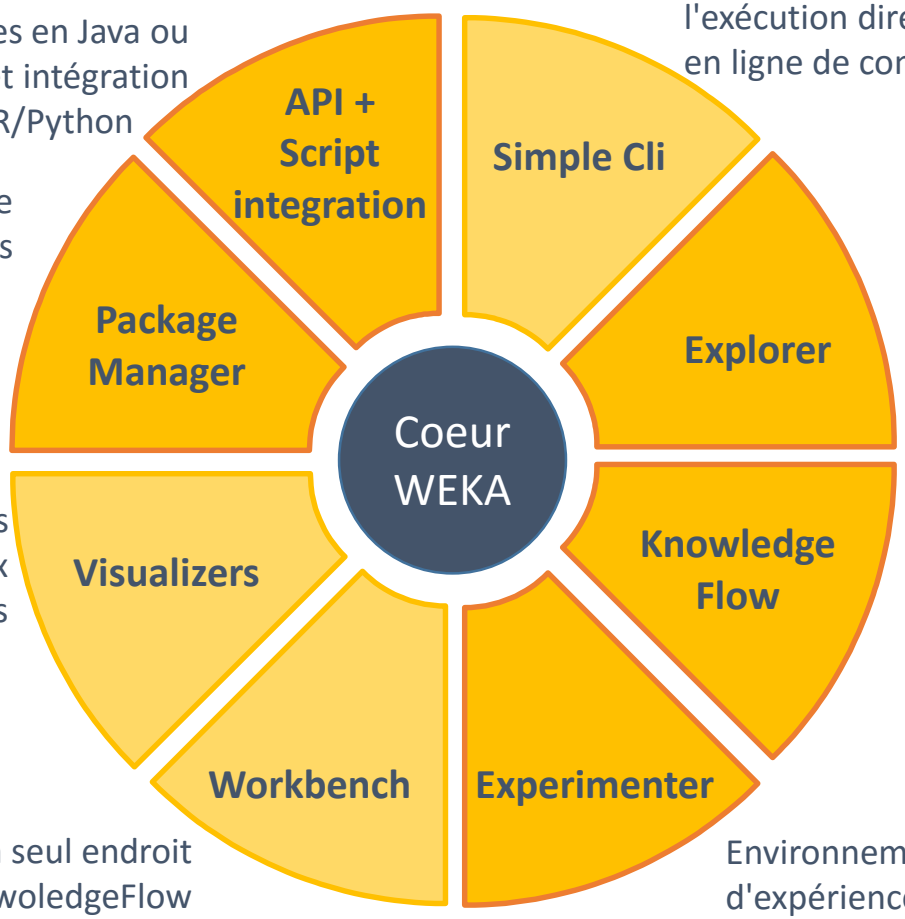
3 – Fonctionnalités et mode opératoire

Interfaces utilisateur Weka

Accès aux fonctions Weka directement depuis des programmes en Java ou d'autres langages, et intégration directe de scripts R/Python

Interface simple (shell) qui permet l'exécution directe des commandes WEKA en ligne de commandes

Gestionnaire de d'extensions



Interface permettant de paramétrer et réaliser une analyse sur un jeu de données

Visualisation des graphiques générés et jeux de données

Interface « drag-and-drop » permettant de créer un processus de workflow complet d'analyse d'un ou plusieurs jeux de données (essentiellement les mêmes fonctions que Explorer.)

Interface regroupant en un seul endroit le SimpleCli, Explorer, KnwoledgeFlow et Experimenter

Environnement pour la réalisation d'expériences de tests et de comparaison de modèles statistiques

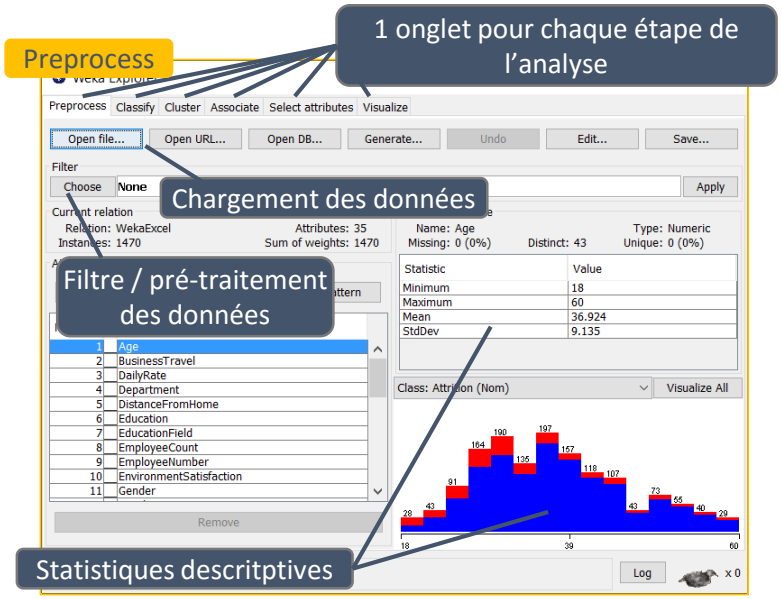
3 – Fonctionnalités et mode opératoire

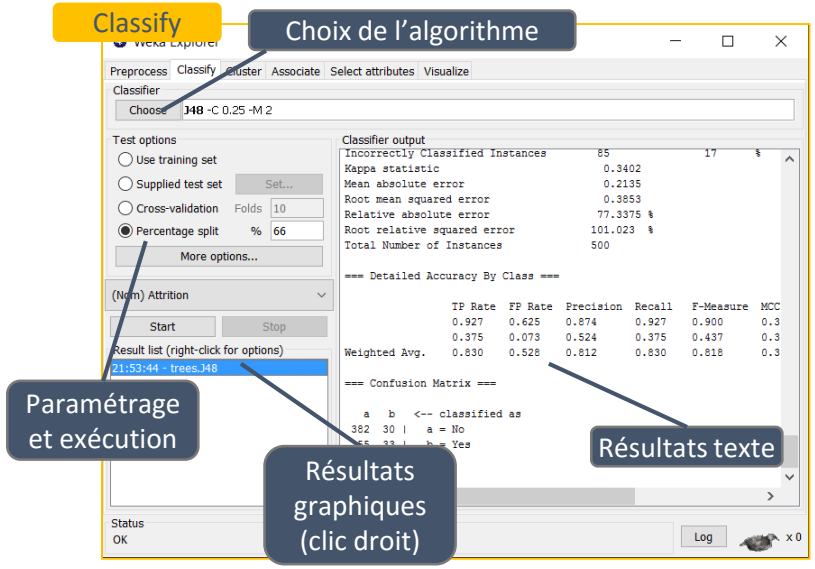
Principales fonctionnalités de traitement des données

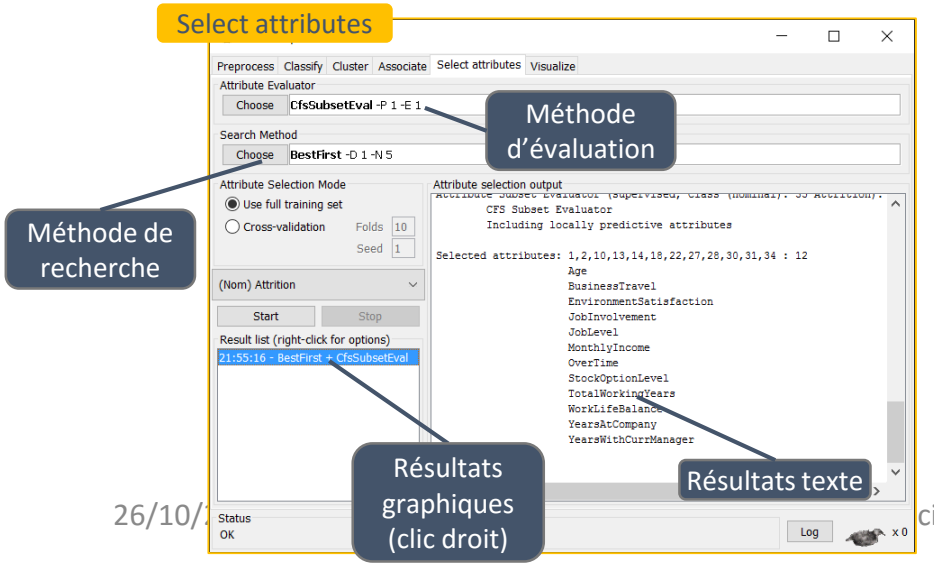
Preprocessing	Import, inspection et préparation/filtre des données
Classification	Mise en œuvre des différents algorithmes de classification
Clustering	Accès aux techniques de clustering comme l'algorithme de k-means
Association	Accès aux apprentissages par règles d'association qui essaient d'identifier toutes les relations importantes entre les variables
Feature (attribute) selection	Choix des variables les plus pertinentes et prometteuses
Visualization	Affichage graphique scatterplot, arbres

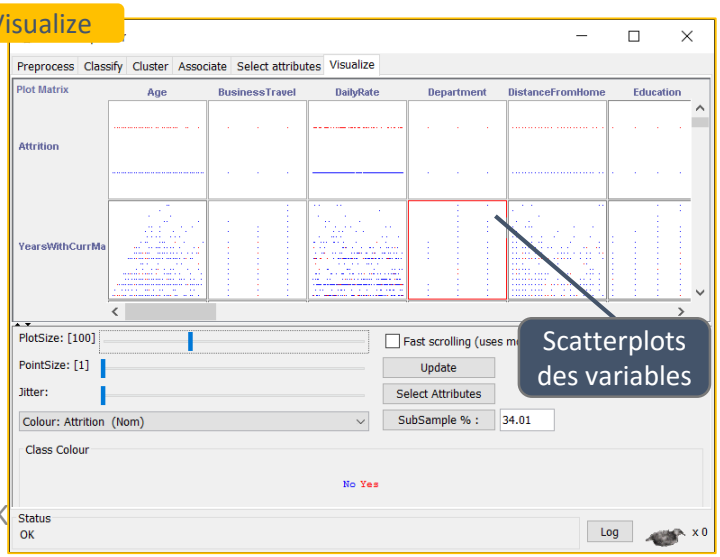
3 – Fonctionnalités et mode opératoire

Focus sur l'interface Explorer



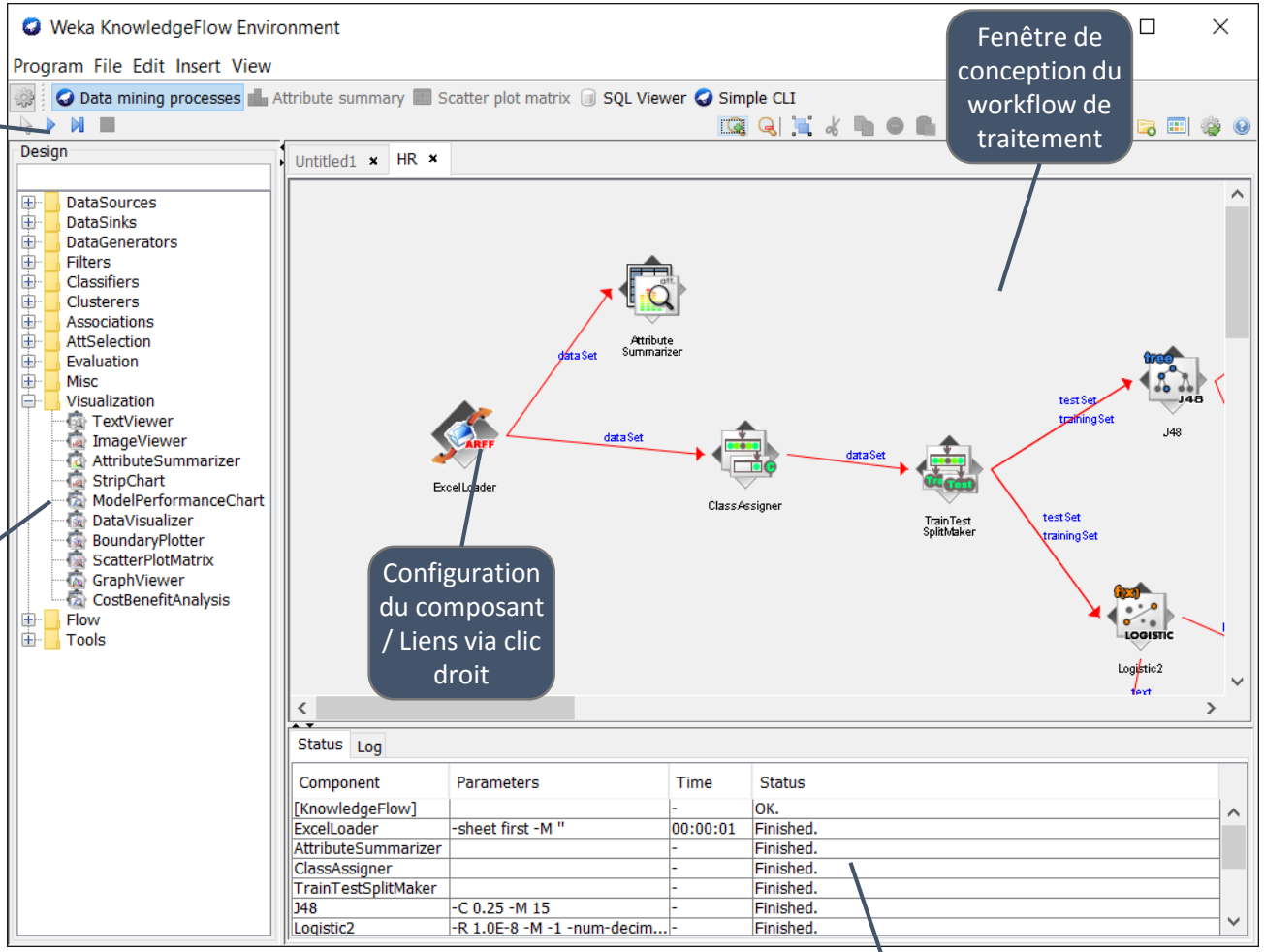






3 – Fonctionnalités et mode opératoire

Focus sur l'interface knowledgeFlow



Lancement de l'exécution du workflow

Fenêtre de conception du workflow de traitement

Sélection des composants à intégrer dans le workflow

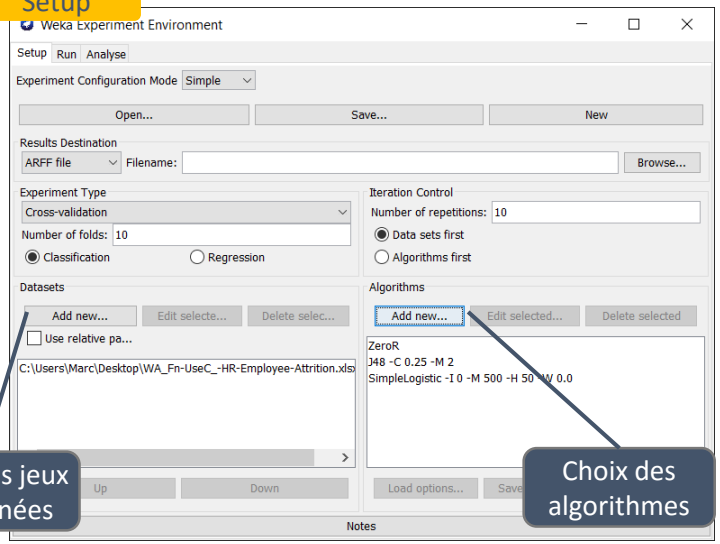
Configuration du composant / Liens via clic droit

Résultats de l'exécution de chaque composant

3 – Fonctionnalités et mode opératoire

Focus sur l'interface Experimenter

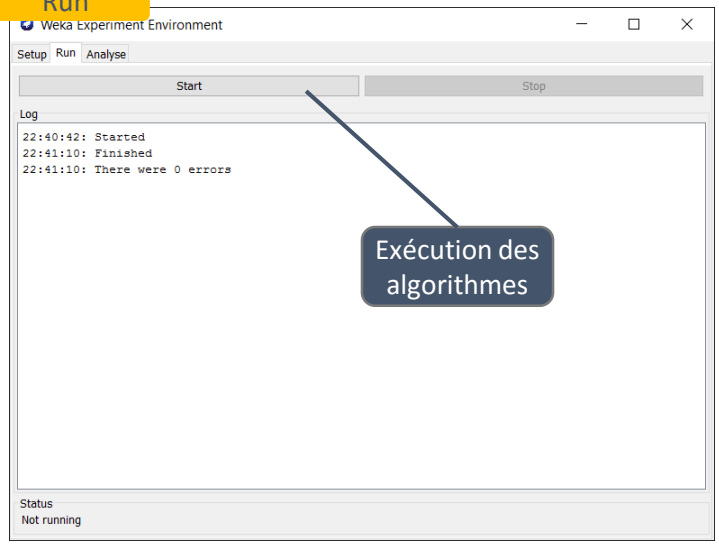
Setup



Choix des jeux de données

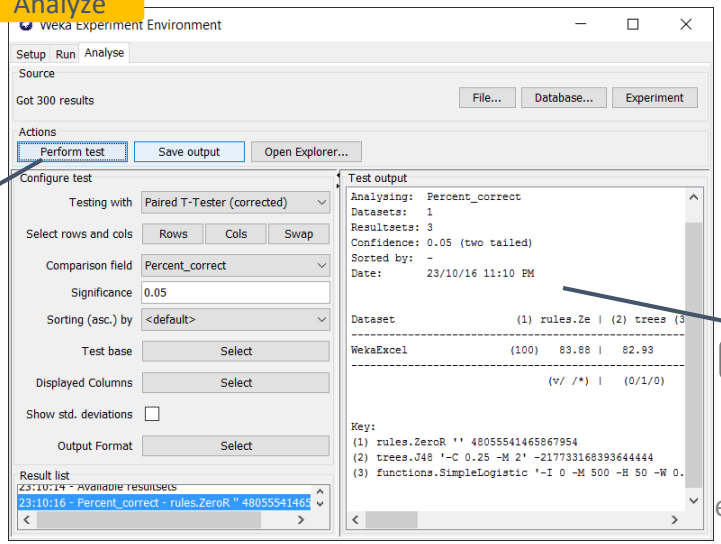
Choix des algorithmes

Run



Exécution des algorithmes

Analyze



Test des différents modèles

Résultats

4 – Méthodes de datamining proposées

	Nombre d'algorithmes	Exemples
Classification	134	Méthodes Bayésiennes, arbres de décision, régressions, réseau de neurones, séparateurs à vaste marge, boosting, bagging…
Clustering	12	CLOPE, Cobweb, DBScan, EM (maximisation de l'espérance), FarthestFirst, FilteredClusterer, HierarchicalClusterer (classification ascendante hiérarchique), MakeDensityBasedClusterer, OPTICS, sIB, SimpleKMeans, XMeans
Association	7	Apriori, FilteredAssociator, FPGrowth, GeneralizedSequentialPatterns, HotSpot, PredictiveApriori, Tertius

La liste exhaustive est disponible sur le wiki de la communauté Pentaho.
http://wiki.pentaho.com/display/DATAMINING/Data+Mining+Algorithms+and+Tools+in+Weka

5 – Gestion de la volumétrie et rapidité

Un outil prêt pour la gestion du Big Data

Les larges volumes de données liés au Big data entraîne rapidement des problèmes de saturation mémoire lors de l'utilisation des logiciels de data-mining. Weka met en œuvre un ensemble de techniques et d'architecture permettant de contourner ces limites et de réussir à gérer ces problématiques Big Data :

1. **Sparse data**
   La compression, en amont du traitement, des fichiers de données contenant beaucoup de zéros permet de réduire l'empreinte mémoire nécessaire aux traitements de données.
2. **Incremental / anytime algorithms**
   L'utilisation d'algorithmes incrémentiels permet de construire un modèle par un traitement des données "ligne après ligne", ne nécessitant ainsi pas le chargement des données complètes en mémoire. La limite mémoire devient celle de la taille du modèle.
3. **Reservoir sampling algorithms**
   La génération d'échantillons aléatoires permet de générer un modèle en minimisant la mémoire nécessaire au traitement, sans toutefois dégrader de beaucoup la qualité du modèle.
4. **Ensemble classifiers**
   La division des données en sous-ensembles permet de générer des sous-modèles en s'affranchissant des limites mémoire du volume original. Le modèle final est ensuite généré à partir de l'ensemble de ces modèles.
5. **Data stream mining**
   La gestion de la fouille des flots de données permet de traiter les données arrivant en flux continue, en adaptant automatiquement les paramètres du modèle en fonctions des nouvelles données reçues.
6. **Distributed architecture**
   La gestion des traitements sur une architecture distribuée, fonctionnant avec Hadoop ou Spark, permet ainsi virtuellement de n'avoir plus aucune limite dans la taille des données.

6 – Points forts et points faibles

Avertissement : les points forts et faibles décrits ci-dessous sont limités à notre compréhension et nos constatations issues de nos travaux pour ce projet.

Points forts

+ **Gratuité**
+ **Richesse des algorithmes**
+ **Gestion du big data**
+ **Outil très "extensible"** via la gestion de packages supplémentaires et la possibilité d'intégrer des scripts R ou Python
+ Interface de **comparaison des performances des modèles**

Points faibles

– **Ergonomie et lisibilité pas toujours évidente** pour une prise en main par les débutants

– **Erreurs non parlantes pour un non développeur** (souvent des exceptions java sans message clair à l'utilisateur).

– **Gestion CSV par défaut calamiteux**

7 – Positionnement par rapport aux autres outils

	Weka	RapidMiner	SQL Server Data-Mining add-ins	Orange	Rattle-Gui	Dataiku
Licence utilisation	Open source / Commerciale	Open source / Commerciale	Commerciale (version gratuite)	Open source	Open source	Commerciale (versions gratuite et payante)
Machine/OS Supporté	Windows / Linux / Mac OS	Windows / Linux / Mac OS	Windows / Mac OS	Windows / Linux / Mac OS	Windows / Linux / Mac OS	Linux
Gestion des séries temporelles	Oui	Oui	Oui	Non	Oui	Non
Gestion graphique des workFlows	Oui	Oui	Non	Oui	Non	Oui
Gestion Big Data	Oui	Non	Non	Non	Oui	Oui

7 – Conclusion

Weka est un logiciel très puissant mais dont la prise en main n'est pas évidente pour un débutant (ergonomie/lisibilité limitée).

En revanche, une fois la prise en main effectuée, la génération et le test d'un modèle sont réalisables très rapidement.

Le choix de Weka comme logiciel de datamining dépendra probablement de la complexité de la problématique à traiter (big data, algorithmes spécifiques, intégration avec d'autres logiciels de business analytics), de la maîtrise statistique/informatique de l'utilisateur, et du type de support souhaité (commercial ou non).

ANNEXES

A1 – Documentation WEKA

Manuel Weka

- http://prdownloads.sourceforge.net/weka/WekaManual-3-8-0.pdf?download

MOOC Weka

- https://weka.waikato.ac.nz/explorer
- (Lien direct vers la chaîne Youtube https://www.youtube.com/user/WekaMOOC)

Documentation communauté Pentaho

- http://wiki.pentaho.com/display/DATAMINING/Pentaho+Data+Mining+Community+Documentation

Ouvrage « Practical Machine Learning Tools and Techniques", 2011

- http://www.cs.waikato.ac.nz/ml/weka/book.html

A2 – Bibliographie

Fonctionnalités et mode opératoire

- WEKA, un logiciel libre d'apprentissage et de data mining http://docplayer.fr/4841246-Weka-un-logiciel-libre-d-apprentissage-et-de-data-mining.html
- How to run your first classifier in Weka http://machinelearningmastery.com/how-to-run-your-first-classifier-in-weka/
- Page Wikipedia sur WEKA https://fr.wikipedia.org/wiki/Weka_(informatique)

Gestion de la volumétrie

- Mining Big Data using Weka 3 http://www.cs.waikato.ac.nz/ml/weka/bigdata.html
- More Data Mining with Weka (5.4: Meta-learners for performance optimization) https://www.youtube.com/watch?v=dfUZdxXl_kU
- Handling Large Data Sets with Weka http://wiki.pentaho.com/display/DATAMINING/Handling+Large+Data+Sets+with+Weka
- Handling Large Data Sets with Weka: A Look at Hadoop and Predictive Models https://www.youtube.com/watch?v=04por7YgCfA
- Advanced Data Mining with Weka (4.1: What is distributed Weka?) https://www.youtube.com/watch?v=OjguI77sYoU
- WEKA API 17/19: Combining Models .. Boosting, Bagging, Stacking and Voting https://www.youtube.com/watch?v=062w-dGDRr0
- https://fr.wikipedia.org/wiki/Fouille_de_flots_de_donn%C3%A9es

Positionnement

- A Comparison of Open Source Tools for Data Science, 2015 http://proc.conisar.org/2015/pdf/3651.pdf
- Comparative Study of Data Mining Tool, 2014 https://www.ijarcsse.com/docs/papers/Volume_4/6_June2014/V4I6-0145.pdf