**TP1 Initiation à Weka**

**By: Ourmassi Alaa Eddine**

**Specialty: RSSI**

**Group: 02**

**Introduction:**

Weka est une suite logicielle qui propose divers algorithmes d’apprentissage automatique conçus pour les tâches de fouille de données. Ces algorithmes peuvent être appliqués directement à un jeu de données ou intégrés dans des programmes Java. Weka comprend des outils pour le prétraitement des données, la classification, la régression, le regroupement, la découverte de règles d’association et la visualisation. C’est également une plateforme flexible pour le développement et l’expérimentation de nouveaux modèles d’apprentissage automatique.

### **Qu’est-ce que le Data Mining ?**

Le **Data Mining**, ou **fouille de données**, désigne le processus d’extraction de connaissances à partir de grandes quantités de données. Il repose sur des méthodes automatiques ou semi-automatiques permettant d’identifier des relations cachées entre les données et de les transformer en informations exploitables. En d’autres termes, le Data Mining consiste à extraire des pépites de savoir enfouies dans un vaste ensemble de données.

Ce processus est étroitement lié à l’**analyse Big Data**, à l’**intelligence prédictive** et à l’**exploitation des données**, jouant un rôle clé dans la prise de décisions stratégiques basées sur les données.

### **Weka**

Weka est un logiciel libre de **data mining**, développé en **Java** par l’Université de Waikato. Son nom est l’acronyme de **Waikato Environment for Knowledge Analysis**.

Il s’agit d’un ensemble d’outils permettant de **manipuler et analyser des fichiers de données**, en intégrant de nombreux algorithmes de **fouille de données**. Weka offre notamment la possibilité d’expérimenter des techniques de **classification**, de **clustering**, de **génération de règles d’association** avec l’algorithme **Apriori**, ainsi que des **arbres de décision** et des **réseaux de neurones**, comme étudié en cours.

Weka fournit également un répertoire contenant plusieurs jeux de données d’exemple. Ces ensembles peuvent être directement chargés dans l’outil, facilitant ainsi le développement et l’expérimentation des modèles. Les jeux de données sont accessibles via le chemin suivant :  
📂 **C:\Program Files\Weka-3-8\data**  
et sont généralement au format **.arff**.

**Partie explorer:**

**Exploration du Data Mining avec Weka :**

Ce document présente une exploration du **Data Mining** en mettant l’accent sur l’utilisation du logiciel **Weka**. Aucun support visuel descriptif n’est inclus ; l’approche repose uniquement sur des explications textuelles détaillées.

### **Le Data Mining : Définition et Objectifs**

Le **Data Mining** (fouille de données) est un processus visant à extraire des connaissances exploitables à partir de grandes quantités de données. Il est étroitement lié au **Big Data**, à l’**intelligence prédictive** et à l’**exploitation des données**. L’objectif principal est d’identifier des relations entre les données et de les transformer en informations pertinentes.

### **Présentation de Weka**

Weka (**Waikato Environment for Knowledge Analysis**) est un logiciel libre écrit en Java et développé par l’Université de Waikato. Il regroupe plusieurs outils d’analyse et implémente divers algorithmes de **classification**, **clustering**, et **règles d’association**. Weka permet d’expérimenter des méthodes comme **l’algorithme Apriori**, les **arbres de décision**, et les **réseaux de neurones**. Les ensembles de données au format .arff peuvent être directement chargés dans Weka pour l’analyse.

### **Sources de Jeux de Données**

* **Kaggle** : Plateforme collaborative en science des données, hébergeant des jeux de données et organisant des compétitions de Machine Learning.
* **Auto-WEKA** : Outil d’optimisation automatisée des algorithmes d’apprentissage et de leurs hyperparamètres via des techniques bayésiennes.

## **Exploration de Weka**

Weka propose six onglets principaux pour organiser l’analyse :

### **1. Preprocess** (Prétraitement des Données)

Permet de charger un jeu de données et d’examiner ses attributs avant de l’adapter aux méthodes de Data Mining. Voici quelques éléments analysables :

* Nombre et nom des attributs
* Statistiques de distribution des données
* Type des données (binaire, nominal, numérique)
* Nombre total d’instances

### **2. Classify** (Classification)

Module permettant de tester différents **algorithmes de classification**.

* **Modes de validation** :
  + Use training set : Test sur les mêmes données d’apprentissage.
  + Supplied test set : Utilisation d’un jeu de test externe.
  + Cross-validation : Division des données en N partitions, avec apprentissage et test successifs.
  + Percentage split : Séparation d’un pourcentage des données pour l’apprentissage et le reste pour le test.
* Les résultats de classification sont affichés sous forme **textuelle** et peuvent être analysés statistiquement.

### **3. Cluster** (Segmentation)

Implémente des algorithmes de **clustering**, comme **K-Means**, pour regrouper les données en clusters distincts. Les résultats peuvent être comparés à des classifications préexistantes.

### **4. Associate** (Règles d’Association)

Utilisation d’algorithmes pour identifier des corrélations et relations significatives entre différentes variables.

### **5. Select attributes** (Sélection d’Attributs)

Détection des attributs les plus pertinents pour améliorer la classification ou le clustering.

### **6. Visualize** (Visualisation)

Offre une projection **2D** des données, affichant les relations entre attributs sous forme de points colorés selon les classes.

## **Mesures d’Évaluation des Modèles**

Pour évaluer la performance des modèles, plusieurs métriques sont utilisées :

* **Précision** : Proportion des prédictions correctes parmi les instances classées.
  + Précision pour la classe Yes = TP / (TP + FP)
  + Précision pour la classe No = TN / (TN + FN)
* **Rappel** : Capacité du modèle à retrouver toutes les instances pertinentes.
  + Rappel pour la classe Yes = TP / (TP + FN)
  + Rappel pour la classe No = TN / (TN + FP)
* **F-mesure** : Moyenne harmonique de la précision et du rappel.
  + F-mesure = (2 × Précision × Rappel) / (Précision + Rappel)
* **Vrai Positif (TP)** : Nombre d’instances correctement classées dans une catégorie donnée.

### **1. Jeu de données d'entraînement : Breast Cancer**

* **Description :** Ce jeu de données contient des informations médicales utilisées pour la classification du cancer du sein.
* **Nombre d'instances :** 286
* **Nombre d'attributs :** 10
* **Objectif :** Servir à l'entraînement d'un modèle d’apprentissage automatique pour classifier les cas de cancer du sein en fonction de caractéristiques diagnostiques.

### **2. Jeu de données de test : Cancer**

* **Description :** Un jeu de données distinct utilisé pour évaluer les performances du modèle.
* **Nombre d'instances :** 89
* **Nombre d'attributs :** 10
* **Objectif :** Fournir une évaluation indépendante de la précision et de la capacité de généralisation du modèle.

**Algorithme 01:**

OneR:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Options | Correctly Classified Instances | Precision | Recall | F-Measure |
| Use training set | 72.7273 % | 0,703 | 0.727 | 0.680 |
| Suplied test set | 76.4045 % | 0,730 | 0,764 | 0,725 |
| Cross validation fold=10 | 65.7343 % | 0.624 | 0.657 | 0.635 |
| Leave one out fold = 60 | 69.2308 % | 0.645 | 0.692 | 0.645 |
| Pourcentage split = 66% | 68.0412 % | 0.656 | 0.680 | 0.644 |

**Discussion des résultats :**

La meilleure méthode d’évaluation pour l’algorithme OneR est **73.3871 %**

**Algorithme 02:**

Naïve Bays:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Options | Correctly Classified Instances | Precision | Recall | F-Measure |
| Use training set | 72.7273 % | 0,703 | 0.727 | 0.680 |
| Suplied test set | 76.4045 % | 0,754 | 0,764 | 0,758 |
| Cross validation fold=10 | 71.6783 % | 0,704 | 0,717 | 0,708 |
| Leave one out fold = 60 | 72.028 % | 0,704 | 0,720 | 0,708 |
| Pourcentage split = 66% | 71.134 % | 0,701 | 0,711 | 0,703 |

**Discussion des résultats :**

La meilleure méthode d’évaluation pour l’algorithme Naive Bayes method avec F-Measure = 0,758

**Algorithme 03:**

C4.5:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Options | Correctly Classified Instances | Precision | Recall | F-Measure |
| Use training set | 75.8741 % | 0,760 | 0,759 | 0,716 |
| Suplied test set | 77.5281 % | 0,749 | 0,775 | 0,742 |
| Cross validation fold=10 | 75.5245 % | 0,752 | 0,755 | 0,713 |
| Leave one out fold = 60 | 75.5245 % | 0,752 | 0,755 | 0,713 |
| Pourcentage split = 66% | 68.0412 % | 0,657 | 0,680 | 0,650 |

**Discussion des résultats :**

Discussion des résultats : un paragraphe qui détermine quel est la meilleure méthode d’évaluation pour l’algorithme C45 est F-Measure = 0,742

**Algorithme 04:**

KNN

K=3, avec Euclidean distance

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Options | Correctly Classified Instances | Precision | Recall | F-Measure |
| Use training set | 97.9021 % | 0,979 | 0,979 | 0,979 |
| Suplied test set | 82.0225 % | 0,817 | 0,820 | 0,794 |
| Cross validation fold=10 | 72.3776 % | 0,699 | 0,724 | 0,697 |
| Leave one out fold = 60 | 73.7762 % | 0,717 | 0,738 | 0,714 |
| Pourcentage split = 66% | 75.8621 % | 0,824 | 0,759 | 0,712 |

**Discussion des résultats :**

la meilleure méthode d’évaluation est F-Measure = 0,979.

**Algorithme 05:**

ID3

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Options | Correctly Classified Instances | Precision | Recall | F-Measure |
| Use training set | 78.6713 % | 0,800 | 0,787 | 0,753 |
| Suplied test set | 82.0225 % | 0,817 | 0,820 | 0,794 |
| Cross validation fold=10 | 73.4266 % | 0,712 | 0,734 | 0,699 |
| Leave one out fold = 60 | 76.2238 % | 0,761 | 0,762 | 0,723 |
| Pourcentage split = 66% | 68.9655 % | 0,683 | 0,690 | 0,630 |

**Discussion des résultats :**

la meilleure méthode d’évaluation pour decison table est F-Measure = : 0,794

Conclusion :

Le meilleur algorithme avec le premier dataset Breast-cancer est KNN et avec méthode d’évaluation 97.9021 %

**Deuxieme dataset :**

1. **Jeu de données d'entraînement : Vote**
   * **Description** : Ce jeu de données contient des informations utilisées pour analyser et classifier les votes en fonction de divers attributs.
   * **Nombre d'instances** : 435
   * **Nombre d'attributs** : 17
   * **Objectif** : Servir à l'entraînement d’un modèle d’apprentissage automatique pour la classification des votes en fonction de caractéristiques spécifiques.
2. **Jeu de données de test : Vote2**
   * **Description** : Un jeu de données distinct utilisé pour évaluer les performances du modèle.
   * **Nombre d'instances** : 98
   * **Nombre d'attributs** : 17
   * **Objectif** : Fournir une évaluation indépendante de la précision et de la capacité de généralisation du modèle.

**Algorithme 01:**

OneR:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Options | Correctly Classified Instances | Precision | Recall | F-Measure |
| Use training set | 95.6322 % | 0,958 | 0,956 | 0,957 |
| Suplied test set | 96.9388 % | 0,972 | 0,969 | 0,970 |
| Cross validation fold=10 | 95.6322 % | 0,958 | 0,956 | 0,957 |
| Leave one out fold = 60 | 95.6322 % | 0,958 | 0,956 | 0,957 |
| Pourcentage split = 66% | 90.6977 % | 0,915 | 0,907 | 0,909 |

**Discussion des résultats :**

La meilleure méthode d’évaluation pour l’algorithme OneR est 0,970

**Algorithme 02:**

Naïve Bays:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Options | Correctly Classified Instances | Precision | Recall | F-Measure |
| Use training set | 90.3448 % | 0,907 | 0,903 | 0,904 |
| Suplied test set | 90.8163 % | 0,915 | 0,908 | 0,909 |
| Cross validation fold=10 | 90.1149 % | 0,905 | 0,901 | 0,902 |
| Leave one out fold = 60 | 90.1149 % | 0,905 | 0,901 | 0,902 |
| Pourcentage split = 66% | 81.3953 % | 0,842 | 0,814 | 0,821 |

**Discussion des résultats :**

la meilleure méthode d’évaluation pour l’algorithme Naive bayes est F\_Measure = 0,909

**Algorithme 03:**

C4.5:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Options | Correctly Classified Instances | Precision | Recall | F-Measure |
| Use training set | 97.2414 % | 0,972 | 0,972 | 0,972 |
| Suplied test set | 97.9592 % | 0,980 | 0,980 | 0,980 |
| Cross validation fold=10 | 96.3218 % | 0,963 | 0,963 | 0,963 |
| Leave one out fold = 60 | 96.7816 % | 0,968 | 0,968 | 0,968 |
| Pourcentage split = 66% | 95.3488 % | 0,953 | 0,953 | 0,953 |

**Discussion des résultats :**

Discussion des résultats : la meilleure méthode d’évaluation pour l’algorithme C45 est : 0,980

**Algorithme 04:**

KNN

K=3, avec Euclidean distance

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Options | Correctly Classified Instances | Precision | Recall | F-Measure |
| Use training set | 94.4828 % | 0,946 | 0,945 | 0,945 |
| Suplied test set | 95.9184 % | 0,959 | 0,959 | 0,959 |
| Cross validation fold=10 | 92.6437 % | 0,929 | 0,926 | 0,927 |
| Leave one out fold = 60 | 92.8736 % | 0,931 | 0,929 | 0,929 |
| Pourcentage split = 90% | 86.0465 % | 0,886 | 0,886 | 0,860 |

**Discussion des résultats :**

la meilleure méthode d’évaluation pour l’algorithme Knn est F\_Measure : 0,959

**Algorithme 05:**

ID3

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Options | Correctly Classified Instances | Precision | Recall | F-Measure |
| Use training set | 96.5517 % | 0,965 | 0,966 | 0,965 |
| Suplied test set | 90.8163 % | 0,915 | 0,908 | 0,909 |
| Cross validation fold=10 | 94.9425 % | 0,949 | 0,949 | 0,949 |
| Leave one out fold = 60 | 93.7931 % | 0,938 | 0,938 | 0,938 |
| Pourcentage split = 90% | 90.6977 % | 0,915 | 0,908 | 0,909 |

**Discussion des résultats :**

la meilleure méthode d’évaluation pour l’algorithme decision tree est : 0,965

Conclusion :

le meilleur algorithme avec le deuxième dataset VOTE est C45 avec méthode d’évaluation 97.9592 %

**Conclusion sur la partie explorer :**

**Le meilleur algorithme pour la dataset Breast-cancer est Knn avec la methode F-measer =0,979**

**Et le meilleur algorithme pour la dataset Vote est C45 avec la methode F-measer =0.980**

Partie experimenter

Les résultats d’une exécution de l’interface avec une description des paramètres

Partie knowledgeflow

Knowldgeflow est un processus et l'interface du flux de connaissances vous permet de concevoir graphiquement ce processus et d'exécuter les conceptions que vous créez. Cela comprend le chargement et la transformation des données d'entrée, l'exécution d'algorithmes et la présentation des résultats

- Data source.

- Filters.

- Classifiers.

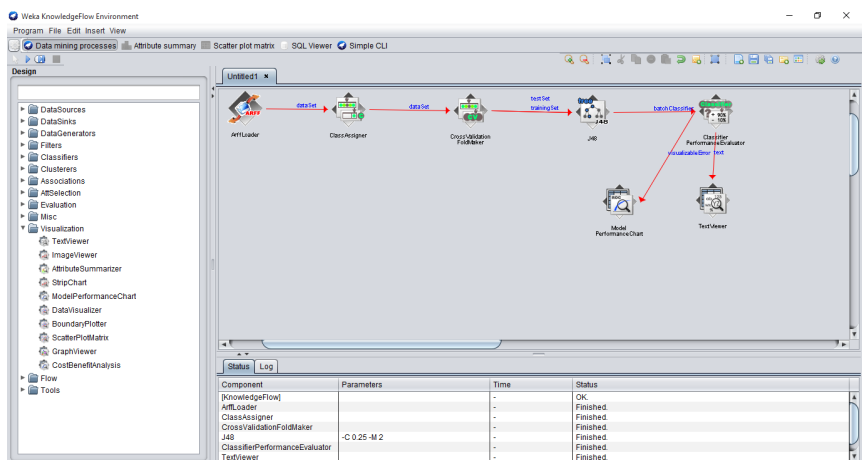
- Evaluation.

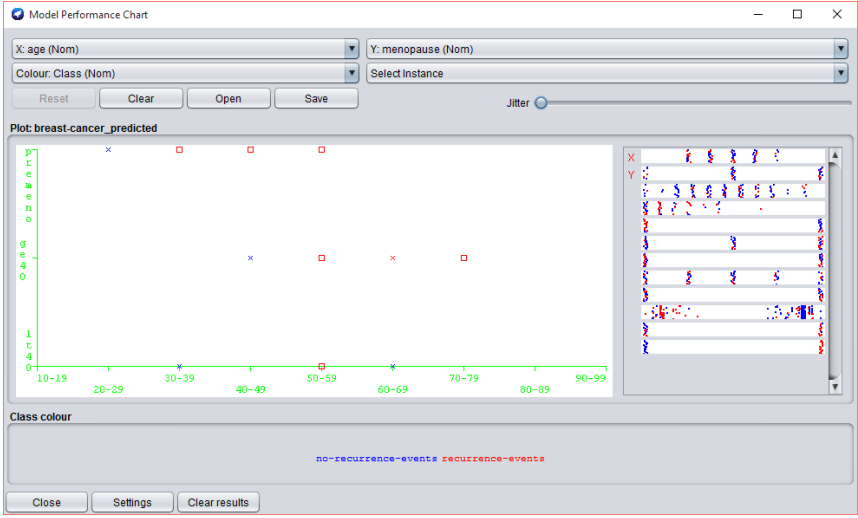
- Visualization.

Data => breast\_cancer=>C45

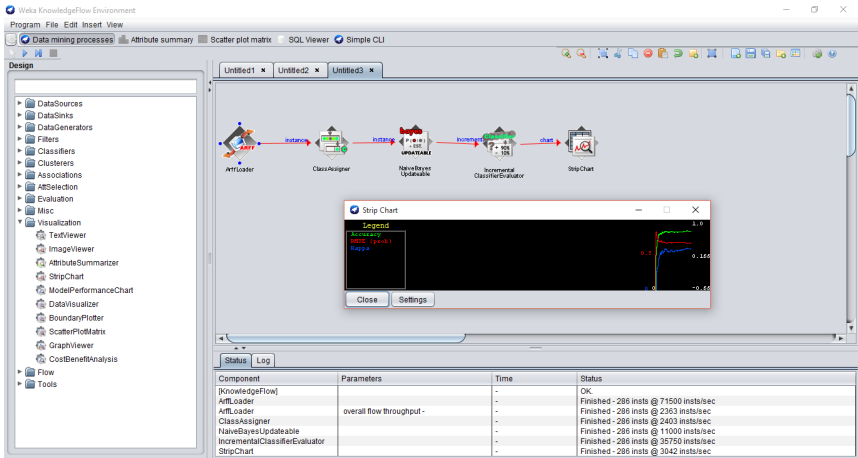
****

evaluat using cross-validation

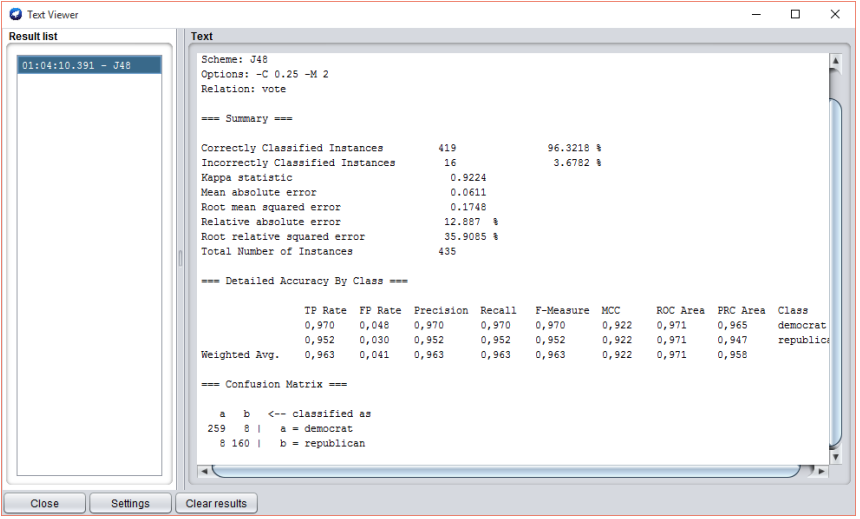
****

****

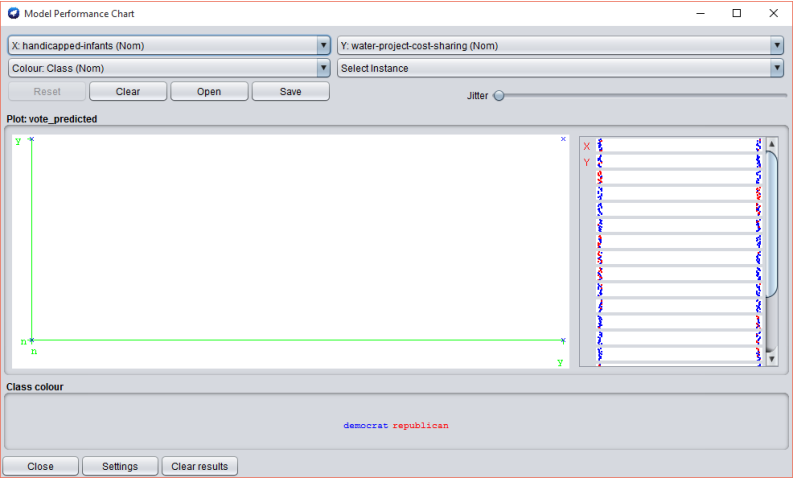
Working with stream data

****

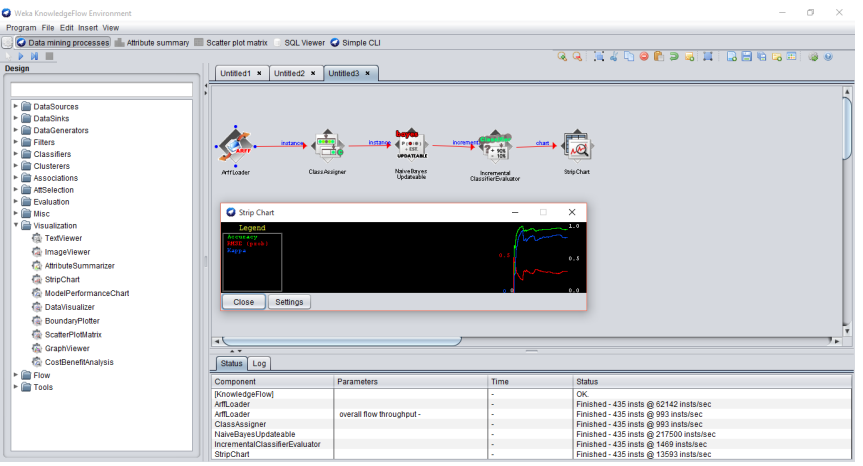
Data=> vote =>C45

****

evaluat using cross-validation

****

Working with stream data

****

**Partie workbench**

Partie SimpleCLI

Les résultats d’une exécution de l’interface avec une description des paramètres

**Conclusion générale**

Weka est une collection d'algorithmes d'apprentissage automatique pour les tâches d'exploration de données. Les algorithmes peuvent être appliqués directement à un ensemble de données ou appelés à partir de votre propre code Java. Weka contient des outils pour le prétraitement des données, la classification, la régression, le clustering, les règles d'association et la visualisation.