[1] TP No 02 : Arbres de décision & Forets Aléatoires

Exercice 1 (DM-ML-DT)**.** Dans cet exercice, il s'agit d'explorer le processus de la classification supervisée (phase d'apprentissage et phase de classement). L'objectif aussi, est d'expérimenter plusieurs implémentation de l'arbre de décision.

1. A partir du répertoire data de **Weka**, charger le fichier **weather.nominal.arff**. Ensuite, remplir le tableau suivant :

Attribut	Règles	Taux d'erreur	Taux d'erreur total
Outlook	sunny → no overcast → yes rainy → yes		
Temperature	hot → no mild → yes cool → yes		
Humidity	high → no normal → yes		
Windy	false → yes true → yes		

2. Le classifieur **OneR** utilise un seul attribut (celui ayant le plus faible taux d'erreur) pour effectuer la classification. Quelles règles de classification aura-t-on si on utilise **OneR** ?
3. Vérifier la réponse de la question précédente (Onglet *Classify : Classifier/rules/OneR*).
4. Explorer le résultat de la classification
   a. *Detailed Accuracy By Class (TP, FP, Precision, Recall,...)*
   b. *Confusion Matrix*
5. A partir du répertoire en ligne "*Échantillons de données*" de l'*ENT*, choisissez un échantillon de données [2]. Utiliser les différentes implémentations de l'arbre de décision sous Weka à savoir :

1. Ce travail demandé doit être démarré durant la séance de TP, à terminer chez soi pour être remis à votre enseignant avant le 04/12/2020 à 23h55. A remettre dans l'espace de dépôt un dossier numérique contenant le(s) fichier(s)de données générées + un compte-rendu détaillant le travail réalisé et justifiant les résultats obtenus.
2. Le choix de l'échantillon de données doit être validé par votre enseignant

- **HoeffdingTree***
- **J48**
- **LMT**
- **RandomTree**
- **REPTree**
- **SimpleCart**

Pour chaque implémentation :

- Selon le mode opératoire *(Test Options : Use training set)*, visualisez l'arbre généré *(Visualize Tree)*.
- Faites varier les techniques d'évaluation (*Use training Set*, *2 and 10 Cross-validation*, *66% Percentage Split*).
  * Ajouter plus d'options et de critères d'évaluation de classification.
  * Visualiser graphiquement : les erreurs de classification, les courbes de threshold.
- Quelles remarques peut-on faire par rapport aux résultats obtenus ?

6. Dresser un tableau pour comparer les différentes approches. Interpréter les performances de chaque implémentation d'arbre de décision.

Exercice 2 (Challenge-RF)**.** Sous R et en utilisant la base de données "readingSkills" depuis le package "Party", construisez une forêt aléatoire de haute performance qui a pour but de bien classer les observations de la variable "nativeSpeaker" entre "yes" ou "no". Respecter les consignes suivantes :

1. Diviser votre base de données en 70% pour l'apprentissage et 30% pour le test en utilisant le paramètre d'initialisation suivant "set.seed(1234)".
2. Vous êtes libre à choisir les paramètres de votre modèle afin d'optimiser la performance.
3. Vous serez évalué sur les résultats de votre modèle appliqué aux données test. Ainsi, vous êtes invités à mettre, dans le compte rendu, la courbe ROC et l'AUC de votre modèle ainsi que le code sous R.

Exercice 3 (ML-RF)**.** Dans cet exercice, il s'agit d'expérimenter plusieurs implémentation du foret d'arbres de décision.

1. A partir du répertoire en ligne "*Échantillons de données*" de l'*ENT*, considerez le même échantillon de données [3] de l'exercice 1. Utiliser les différentes implémentations du foret d'arbres de décision sous Weka à savoir :
   - **CSForest**
   - **ForestPA**
   - **RandomForest**
   - **SmoothPrivateForest**
   - **SysFor**

   Pour chaque implémentation :

3. Le choix de l'échantillon de données doit être validé par votre enseignant tuteur

- Faites varier les techniques d'évaluation (*Use training Set*, *2 and 10 Cross-validation*, *66% Percentage Split*).
  * Ajouter plus d'options et de critères d'évaluation de classification.
  * Visualiser graphiquement : les erreurs de classification, les courbes de threshold.
- Quelles remarques peut-on faire par rapport aux résultats obtenus ?

2. Dresser un tableau pour comparer les différentes approches. Interpréter les performances de chaque implémentation de forêt aléatoire.