RER – Arbre de décision

Contexte :

Se familiariser avec l'algorithme de classification Decision Tree dans le Machine Learning supervisé

Mots clés :

- **Modélisation :** la modélisation est la création d'un modèle à partir des données d'entrainement. Le Data Mining et le Machine Learning sont des formes de modélisation. En entrée nous avons des données et en sortie un modèle, des règles et des connaissances. L'Arbre de décision est un modèle de connaissance (Knowledge).

- **Modèle prédictif :** un modèle qui analyse des données passées afin de pouvoir anticiper et prévoir des comportements futurs. Classifier est un modèle prédictif.

- **Apprentissage automatique supervisé :** type de machine Learning dans lequel on génère un modèle prédictif en se basant sur des données étiquetées (les données disposent d'un attribut cible, target, classes, variable discriminant).
- **Apprentissage automatique non-supervisé :** type de machine Learning dans lequel on génère un modèle prédictif en se basant sur des données non-étiquetées

- **Decision Tree Classifier (ID3) :** est un outil d'aide à la décision représentant un ensemble de choix sous la forme graphique d'un arbre. Les différentes décisions (classes, prédictions) possibles sont situées aux extrémités des branches, et sont atteintes en fonction de décisions prises à chaque étape.

- **Hyperparamètre :** un hyperparamètre est un paramètre de l'algorithme dont la valeur est définie avant la modélisation. Il est utilisé pour influencer la construction du modèle. ***Exemple :*** *pour l'arbre de décision on peut défini le paramètre* ***"max-depth"*** *comme hyperparamètre.*

- **Entropie :** mesure du “désordre”, “impureté”, “chaos”, “hasard”. Valeur calculée qui rend compte de la dispersion d’un ensemble d’éléments entre 2 pôles. Elle se situe toujours entre 0 et 1.

$$\text{Entropy}(S) = - p_+ \log_2 p_+ - p_- \log_2 p_-$$

- **Root(racine) :** premier attribut à considérer pour démarrer le processus de décision.
- **Leaf(feuille) :** les décisions finales, les classes, prédictions.

- **Information :** généralisation de l’entropie. Une entropie lorsqu’on a plusieurs classes.

$$\mathbf{I(A)=-p_1 \log_2 p_1 – p_2 \log_2 p_2 \dots -p_m \log_2 p_m}$$

$$Entropy(S) = -\sum_{c \in C} p(c) \log_2 p(c)$$

- S représente le jeu de données avec lequel l'entropie est calculée, c'est à dire les données de l'attribut à évaluer.
- c représente les classes dans le jeu S.
- $p(c)$ représente la proportion de points de données appartenant à la classe c par rapport au nombre total de points de données du jeu S.

$$I(s_1, s_2, \dots, s_m) = -\sum_{i=1}^{m} \frac{s_i}{s} log_2 \frac{s_i}{s}$$

- **Gain d’information :** paramètre, critère sur lequel va se baser l’algorithme pour choisir les attributs (nœud) afin de construire l’arbre.

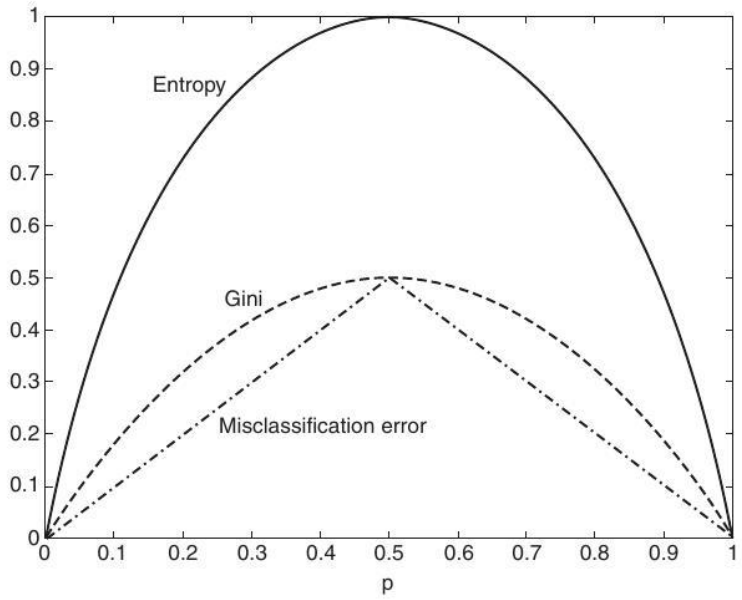
$$Gain(A) = I(s_1, s_2, \dots, s_m) - E(A)$$

- **Régression Logistique :** algorithme de classification basé sur la loi de probabilité logistique. C'est un modèle statistique qui permet d’étudier les relations entre un ensemble de variables d’entrée et une variable de sortie.

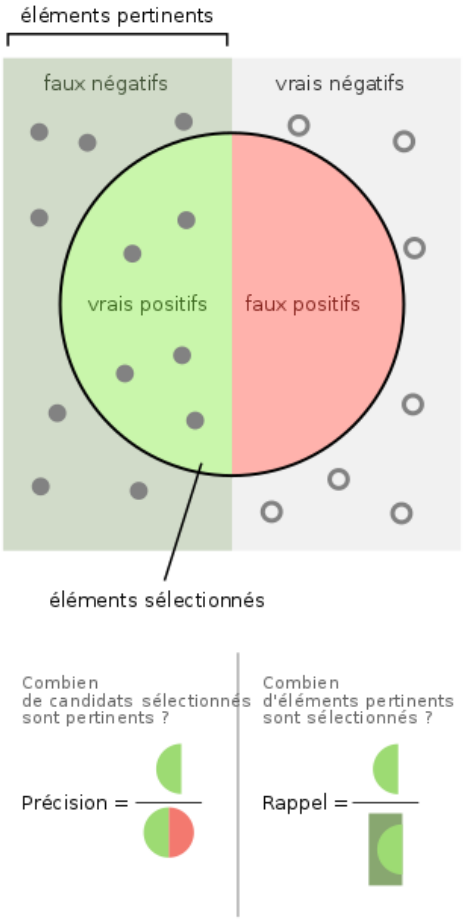
- **CART (Classification And Regression Tree)** : algorithme de classification alternatif à l'ID3 qui utilise l'indice de l'impureté Gini à la place de l'entropie. Il est plus rapide que ID3 parce qu'il utilise le pré-élagage(pruning).

- **Gini :** indice d'impureté utilisé par CART à la place de l'entropie. Il ne calcule pas le logarithme.

$$G = \sum_{i=1}^{C} p(i) * (1 - p(i))$$



- **Métriques :** mesure de performance d'un modèle
  - **Accuracy(exactitude) :** le quotient entre le nombre de prédictions correct et le nombre total de prédictions ; il permet de préciser la qualité générale du modèle. Il indique le % de bonnes prédictions.
  - **Precision:** permet de connaitre le % de prédictions positif ou négatif bien effectué pour chaque classe. La fréquence à laquelle le modèle a été correct pour une classe.
  - **Recall:** ratio de vrais positifs qui permet de répondre à la question de savoir quelle proportion de résultats positifs réels que le modèle a été identifié correctement.



- **F1-score :** c'est un changement de variable dans l'application des mesures de performance en utilisant la précision et le recall. C'est leur moyenne harmonique.

$$\text{F1 Score} = \frac{2 \times (\text{Precision} \times \text{Recall})}{\text{Precision} + \text{Recall}}$$

- **Courbe ROC (Receiver Operating Characteristic) :** c'est un graphique qui montre la performance d'un modèle de classification à tous les seuils de classification. C'est la courbe d'apprentissage.

- **AUC (Area Under the Curve) :** mesure de la totalité de la zone bidimensionnelle située sous la courbe ROC.

- **Matrice de confusion :** tableau permettant une évaluation visuelle du modèle. On y voit le nombre de bonnes et de mauvaises prédictions par classe. On l'obtient à partir de l'évaluation du Test-set.

- **Cross Validation :** la validation croisée est une comparaison entre plusieurs modèles résultant de la division en "k-folds" du Training-set.

- **Overfitting (surapprentissage)** : forte dépendance du modèle prédictif au dataset d'entrainement ce qui ne permet pas la généralisation à d'autres datasets pour l'étape de prédiction.

Problématique(s) :

- En quoi l'Arbre de décision permet-il de réaliser des modèles prédictifs fiables ?
- Comment fonctionne un algorithme d'Arbre de décision ?
- Comment choisir ses paramètres et hyperparamètres pour qu'ils soient les mieux adaptés pour réaliser les prédictions ?

Hypothèses :

1) Les arbres décision de l'IA s'appuient sur la même technique utilisée en science naturelle. **Aude** **VRAI**
2) Les hyperparamètres permettent à l'arbre de décision d'être plus précis (et plus rapides). **Adeline** **VRAI**
3) L'algorithme de l'arbre de décision peut classer des données dispersées, sans tendance. **Etienne** **VRAI**
4) L'algorithme de l'arbre de décision peut atteindre le sur-apprentissage. **Seydou** **VRAI**
5) La validation croisée permet d'augmenter la fiabilité du modèle. **Adrien** **VRAI**
6) Les hyperparamètres sont définis par l'utilisateur avant l'entrainement. **Tetyana** **VRAI**
7) La seule manière d'affiner la proposition de l'arbre de décision est de modifier les hyperparamètres du modèle. **Briand** **VRAI entre autre**
8) Le *Decision Tree* ne permet de prendre que des choix binaires. **Nicolas** **FAUX**
9) Le *Decision Tree* permet de prendre des décisions différentes à chaque branche. **Solenn** **VRAI**
10) Le nombre de paramètres de l'arbre de décision dépend des valeurs du target. **Osman** **VRAI**
11) Il existe une fonction pour calculer toutes les métriques. **Jean Paul** **VRAI**
12) Il est nécessaire de produire plusieurs arbres de décision pour choisir celui qui se rapproche le plus du target. **Briand** **VRAI**

13) L'algorithme de l’arbre de décision est plus rapide dans son exécution que l’algorithme de régression logistique. **Axel** **VRAI**
14) L'entropie est un paramètre de l’algorithme de l’arbre de décision. **Adeline** **VRAI**
15) Grid Search et K-fold sont extrêmement utile pour réduire le coût en temps dans la création des décision-Tree ou Random Forest (ensemble de decision-tree) **Loïc** **VRAI**

Plan d'action :

- Explorer les ressources
- Définir et comprendre les mots clefs
- Répondre à la problématique
- Vérifier les hypothèses
- Faire les Workshop
- RER