RER - Classification

# Contexte :

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

# Mots clés :

* Decision Tree Classifier : algorithme de machine learning
* Modèle prédictif : Les modèles prédictifs permettent de générer les informations dont vous avez besoin pour prendre des décisions opérationnelles, de maintenance, de réparation ou de remplacement de composant avisées.
* Apprentissage automatique supervisé : L'apprentissage supervisé (supervised learning en anglais) est une tâche d'apprentissage automatique consistant à apprendre une fonction de prédiction à partir d'exemples annotés, au contraire de l'apprentissage non supervisé. On distingue les problèmes de régression des problèmes de classement1. Ainsi, on considère que les problèmes de prédiction d'une variable quantitative sont des problèmes de régression tandis que les problèmes de prédiction d'une variable qualitative sont des problèmes de classification.
* Fonction Coût : Dans le domaine de l'intelligence artificielle, la fonction de perte ou de coût est la quantification de l'écart entre les prévisions du modèle et les observations réelles du jeu de donnée utilisé pendant l'entraînement.
* Modélisation : La modélisation de l'IA est la création, la formation et le déploiement d'algorithmes d'apprentissage automatique qui émulent la prise de décision logique basée sur les données disponibles.
* Gridsearch :
* Hyperparamètres : Dans l'apprentissage automatique, un hyperparamètre est un paramètre dont la valeur est utilisée pour contrôler le processus d'apprentissage. En revanche, les valeurs des autres paramètres sont obtenues par apprentissage.
  + Coefficient de régression : Le coefficient de régression indique dans quelle mesure la valeur d'une variable dépendante varie avec la variation de la valeur de la variable indépendante (aussi appelée variable explicative).
  + Coefficient de linéarité :
* Entropie :
  + L'entropie mesure la quantité d'information minimum nécessaire pour vous transmettre un message.
  + Augmentation du chaos
* Root(racine) :
* Leaf(feuille) :
* Régression : on cherche à montrer une relation de dépendance entre une variable à expliquer et des variables explicatives. Cette dépendance s’exprime en terme de probabilité d’appartenir à chacune des classes.
* Logistique : la loi de probabilité est modélisée à partir d’une loi logistique.
* Métriques :
  + Accuracy : mesure l’adéquation d’un modèle de classification sous forme de proportion de résultats réels sur le nombre total de cas.
  + Precision : correspond à la proportion de résultats réels sur tous les résultats positifs.
  + Recall : est la fraction de la quantité totale d’instances pertinentes qui ont été réellement récupérées.
  + F1 score : la moyenne harmonique de précision et de rappel.
* Courbe ROC (Receiver Operating Characteristic) : représentation graphique de la relation existante entre la sensibilité et la spécificité d'un test, calculée pour toutes les valeurs seuils possibles.
* AUC (Area Under the Curve) : mesure la zone sous la courbe tracée avec les vrais positifs sur l’axe y et les faux positifs sur l’axe x.
* Cart :
* Matrice de confusion : résumé des résultats de prédiction pour un problème particulier de classification. Elle compare les données réelles pour une variable cible à celles prédites par un modèle.
* Cross Validation : méthode statistique qui permet d'évaluer les performances des modèles d'apprentissage automatique.
* Information :
* Gain d’information :
* Overfitting : Dans le domaine de l'intelligence artificielle, le surapprentissage entraîne un modèle qui correspond trop précisément à une collection particulière de données utilisées pour l'entrainement.

# 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
2. Les hyperparamètres permettent à l’arbre de décision d’être plus précis. Adeline
3. L'algorithme de l’arbre de décision peut classer des données dispersées, sans tendance. Etienne
4. L'algorithme de l’arbre de décision atteint facilement le sur-apprentissage. Seydou
5. La validation croisée permet d’augmenter la fiabilité du modèle. Adrien
6. Les hyperparamètres sont définis par l’utilisateur avant l’entrainement. Tetyana
7. La seule manière d’affiner la proposition de l’arbre de décision est de modifier les hyperparamètres du modèle. Briand
8. Le Decision Tree ne permet de prendre que des choix binaires. Nicolas
9. Le Decision Tree permet de prendre des décisions différentes à chaque branche. Solenn
10. Le nombre de paramètres de l’arbre de décision dépend du target. Osman
11. Il existe une fonction pour calculer toutes les métriques. Jean Paul
12. Il est nécessaire de produire plusieurs arbres de décision pour choisir celui qui se rapproche le plus du target. Briand
13. L'algorithme de l’arbre de décision est plus rapide dans son exécution que l’algorithme de régression logistique. Axel
14. L'entropie est un paramètre de l’algorithme de l’arbre de décision. Adeline

# Plan d'action :

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