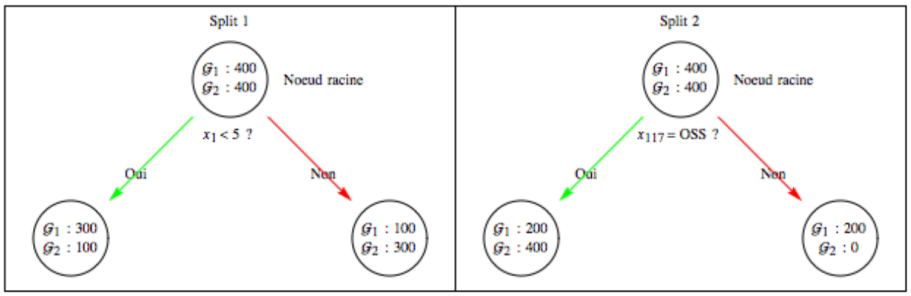
TP1 : Arbre de décision et Foret aléatoire

Indications générales :

1. Le TP se fait impérativement en groupe de 2 à 4 personnes.
2. Le travail doit être démarré durant la séance de TP, à terminer chez soi pour être remis sur Campus avant le 07/11/2023 à 23h55.
3. Un compte rendu obligatoire en format PDF doit être soumis par chaque groupe avant le 07/11/2023 à 23h55.
4. Dans le compte rendu vous présentez le code utilisé pour résoudre chaque partie (dans le cas où le code n'est pas donné) ainsi que les résultats obtenus et l'interprétation détaillée des résultats le cas échéant.
5. L'évaluation est principalement sur votre capacité d'analyser, de critiquer et d'interpréter les résultats. Ainsi, il est essentiel d'expliquer clairement vos conclusions.
6. Les codes sont donnés en R-Studio. Mais si vous êtes plus à l'aise avec un autre langage, n'hésitez pas à l'utiliser.

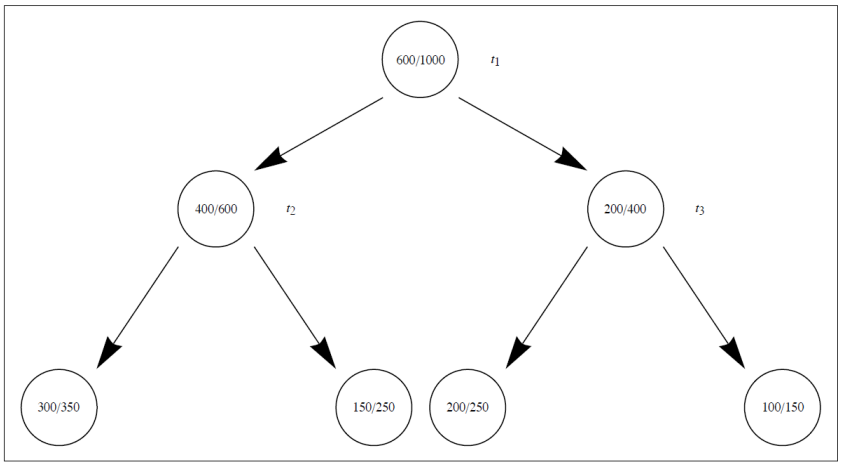
Problème I : Arbres de décision, généralités (Le travail doit être fait à la main c.à.d. pas de code à écrire) :

1. Pour un problème de classification en 2 groupes, sans probabilité a priori ni coût de mauvaise classification, considérer les deux « splits » suivants :



Les nombres qui apparaissent dans chaque cercle (= nœud) sont les nombres d'individus appartenant à chacun des groupes. Déterminer le meilleur d'entre ces deux splits si on utilise pour mesure de l'impureté l'entropie.

2. On considère un problème de classification en 2 groupes. Considérer l'arbre ci dessous où, dans chaque cercle (= nœud), l'expression xxx/yyy indique que, dans ce nœud, il y a xxx individus bien classés sur un total de yyy individus dans le nœud (bien classé selon la règle habituelle du cours).



(a) Calculer la valeur de $R(T)$ et $R_\alpha(T)$ de cet arbre.

(b) En utilisant $R_\alpha(T)$, déterminer les valeurs de α pour lesquelles T_0 est préférable à T_1 = l'arbre constitué de t_1 et des deux feuilles $\{t_2, t_3\}$.

3. Dans le cours on a vu qu'à chaque nœud, on définit un ensemble de splits possibles : S = {ensemble de questions possibles au nœud t}, puis on choisira le split qui maximise $\Delta Imp(t)$. Chaque split ne va concerner que la valeur d'une des composantes, disons la $j-ième$ X_j, de $X = (X_1, ..., X_p)$. Donner le nombre de split possible dans le cas où (**Nous ne considérons que les splits qui aboutissent à deux nœuds descendants**) :

(a) Cas X_j continue (on pourrait croire qu'il y a une infinité de splits possibles mais en fait, comme il y a N_t individus dans le nœud t, il y aura un nombre fini de splits ayant un impact sur les nœuds descendants de t).

(b) Cas X_j discrète avec des valeurs possibles dans l'ensemble $\{a_1, ..., a_L\}$.

4. On a vu dans le cours le principe des splits suppléants comme étant ceux qui, sur la variable X_m, imitent le mieux l'action de s^* au nœud t. En se basant sur les notes du cours, expliquez **EN DÉTAIL** comment utiliser

le principe des splits suppléants pour :

(a) La gestion de données manquantes,

(b) La mesure de l'importance des variables.

Problème II : Arbres de décision, une application sur des données réelles :

Dans cet exercice, il s'agit d'expérimenter une implémentation de l'arbre de décision pour un problème de classification binaire.

1. A partir du répertoire en ligne "Échantillons de données" de l'ENT, choisissez un échantillon de données (Le choix de l'échantillon de données doit être communiqué à l'enseignant et validé avant de commencer le travail). Commencez par effectuer une analyse exploratoire descriptive de votre base de données (typologie des variables, valeurs manquantes, distributions des observations pour les différentes variables, boxplots, etc.).
2. Utilisez une implémentation de l'arbre de décision sous R (ou autre) pour construire un classifieur ayant pour objectif de prédire la classe de la variable dépendante dans votre base de données après avoir observé les variables indépendantes, en respectant les consignes suivantes :
   (a) Optimisez votre arbre de classification en passant par des techniques vues dans le cours (par exemple, l'optimisation des hyperparamètres par validation croisée, élagage, etc.).
   (b) Selon le mode opératoire, visualisez l'arbre optimal généré.
3. Vous serez évalué sur les résultats de votre modèle optimal **appliqué aux données test** :
   (a) Explorer le résultat de la classification :
      i. Detailed Accuracy By Class (Precision, Recall,...)
      ii. Confusion Matrix
      iii. Etc. (Je vous invite à calculer autant de métriques de performance que possible tout en justifiant les résultats obtenus à chaque fois.)
   (b) Veuillez inclure dans le compte rendu la courbe ROC et l'AUC de votre modèle optimal, ainsi que le code sous R correspondant.
   (c) Donnez une conclusion/interprétation globale par rapport aux résultats obtenus. N'hésitez pas à faire preuve de créativité et à aller au-delà des questions demandées. Toute idée d'analyse à valeur ajoutée sera fortement appréciée. ?

Problème III : Courbe LIFT :

Un vendeur souhaite faire un offre à ses clients propriétaires d'un produit P_1 pour leur proposer un nouveau produit P_2. Il dispose d'une base de données (c'est son $\mathcal{E}$) de $n = 81$ clients parmi lesquels 17 ont, par le passé, acheté P_1, ce qui permet de supposer que ces derniers sont des "Bons" clients ; ainsi $a = 17$.

Il dispose d'un budget maximum de 19 euros pour cette opération et chaque envoi lui coûte 1 euros. S'il choisit au hasard 19 individus de $\mathcal{E}$, il peut espérer toucher $19 \times \frac{17}{81} \simeq 4$ bons clients, soit un coût unitaire de 4.75 euros pour joindre un tel client. Il veut optimiser son envoi de façon à toucher le maximum de bons clients pour le même prix.

Son statisticien a produit comme classifieur l'arbre $\widehat{T}^*_{opt}$. Les données pertinentes à l'évaluation du LIFT apparaissent dans le tableau suivant :

$\hat{T}^*_{opt}$	Score			
	≥ 0.5789	≥ 0.5714	≥0.1428	≥ 0.0
¤ individus	19	26	40	81
VP	11	15	17	17

De ce tableau, on peut constater que 19 clients ont un score ≥ 0.5789. Parmi ceux-ci, $A = 11$ (VP = Vrai Positif du classifieur) sont des bons clients. Son LIFT à 23% ($^{19}/_{81}$) est donc de

$$\frac{^{11}/_{19}}{^{17}/_{81}} = 2.76.$$

Il ne lui en coûtera que $19/11 = 1.73$ euros pour joindre un bon client, soit une jolie économie par rapport à un envoi au hasard.

Par ailleurs, si on veut toucher 100% des bons clients, alors il faut joindre les 40 individus dont le score est ≥ 0.1428, soit environ la moitié de l'échantillon. Ceci représente une substantielle économie par rapport à un envoi au hasard qui, pour atteindre le même objectif, nécessite alors l'envoi de l'offre à tous les individus de $\mathcal{E}$.

Notre vendeur est relativement satisfait, mais en bon petit capitaliste, les choses ne sauraient s'arrêter là pour lui et ces résultats l'amènent à tenir la réflexion suivante : " Il reste quand même 6 bons clients dans la nature. Combien ça me coûterait pour tenter de tous les joindre, où du moins 3 sur ces 6 clients potentiels ? Pour répondre intelligemment à ce type de question, il faut introduire la notion de "courbe LIFT".

Supposons que le groupe qui nous intéresse soit le groupe $\mathcal{G}_2$ (le groupe des bons clients). Soit $p(2 \mid t)$ le score d'un individu choisi au hasard dans la population quand on le passe dans le classifieur, que l'on suppose un arbre (t est la feuille sur laquelle il a atterri).

Ainsi pour un seuil donné $s \in [0, 1]$,

$$\mathbb{P}[\text{individu} \longmapsto \mathcal{G}_2] = \mathbb{P}[p(2 \mid t) \geq s],$$

estimé par :

$$RPP\,(\text{rate of positive predictions}) = \frac{\#FP + \#VP}{n}.$$

De plus, on rappelle que la sensibilité du classifieur est estimée par :

$$Se = \frac{\#\text{ individus } \in \mathcal{G}_2 \text{ ayant p}(2 \mid t) \geq \text{s}}{\#\mathcal{G}_2} = \frac{\#VP}{\#\mathcal{G}_2}.$$

1. Exprimer l'expression mathématique du LIFT en fonction des RPP et Se.
2. Compléter le tableau suivant qui est adapté au calcul de la courbe LIFT :

$\widehat{T}^*_{opt}$	Score			
	≥ 0.5789	≥ 0.5714	≥ 0.1428	≥ 0.0
# individus	19	26	40	81
RPP				
Se				

3. A partir du tableau de la partie (2), tracer une première version de la courbe LIFT représentant Se en fonction de RPP (i.e. les points de coordonnées (RPP, Se)), pour $s \in [0, 1]$.
4. Sur le graphique de la partie (3), quelle est l'interprétation statistique des points qui se trouvent sur la première bissectrice ?
5. En utilisant une technique d'interpolation linéaire sur le graphique de la partie (3), calculer le $LIFT$ à 10%.
6. Signaler deux différences notables entre les courbes ROC et LIFT, malgré leurs formes comparables.
7. Une autre version de la courbe LIFT est celle normalisée qui représente la courbe des points $(RPP\,(s)\,, {}^{Se(s)}/_{RPP(s)})$. Tracer la courbe LIFT normalisée.
8. On rappelle que pour rejoindre 11 bons clients, il en a coûté 19 euros (avec un LIFT à 23%), soit 1.73 euros par personne. Sachant qu'en se basant sur le graphique de la partie (7), un LIFT de 30 % est de 2,75, combien coûte-t-il en plus pour rejoindre trois personnes de plus au total ? Interpréter le résultat ?
9. Maintenant nous supposons que le vendeur envisage faire son offre par internet (pas de budget limité). Il est intéressé par joindre 80% de sa clientèle sans agacer inutilement trop de "mauvais client" :
   (a) Dans ce contexte, combien de pourcentage de ses clients devrait-il contacter **au hasard** pour espérer toucher 80 % des bons clients ?
   (b) En vous basant sur une lecture de la courbe LIFT de la partie (3), fournissez une solution plus astucieuse avec une réduction substantielle du pourcentage.

Problème IV : Foret aléatoire, une application sur des données réelles :

Dans cet exercice, il s'agit d'expérimenter une implémentation d'une forets aléatoires de décision.

1. A partir du répertoire en ligne "Échantillons de données" de l'ENT, considérez le même échantillon de données du problème II. Utiliser différentes implémentations (au moins deux) du foret d'arbres de décision sous R et pour chaque implémentation Présentez-moi **avec clarté** les étapes de votre raisonnement menant au modèle optimal (Out-of sample estimation, Cross Validation, Élagage, optimisation des hyperparamètres, etc.)
2. Analyser et comparer les erreurs de la classification (Detailed Accuracy By Class (Precision, Recall,...), Confusion Matrix, Courbes ROC, AUC, etc.)
3. Interpréter les performances des différentes implémentations des forets aléatoires des parties (1 et 2) .