DM DE STATISTIQUES EN GRANDE DIMENSION

by

HOUNKONNOU Mehdy

Contents

I wish to thank my teachers.

1. Questions de cours

On considère $\mathcal{L}_n = \{(X_1, Y_1), ..., (X_n, Y_n)\}$ un ensemble d'apprentissage de n couples $(X_i, Y_i) \in \mathbb{R}^p$ x $\mathbb{R}$ i.i.d de même loi q'un couple aléatoires (X, Y) où X représente les p variables explicatives et Y la variable réponse associée.

1.1. On parle de grande dimension lorsque p > n, c'est-à-dire que le nombre de variables explicatives p dépasse le nombre d'observations n.

1.2. Le modèle de regression non-paramétrique :

$$Y = g(X) + \epsilon, \ avec\ g(X) = \mathbb{E}(Y|X)$$

g étant la fonction de régression inconnue représentant le lien entre X et Y.
ϵ étant une variable aéatoire réelle répresentant l'erreur de mesure.
On suppose qu'en moyenne l'erreur de mesure est nulle conditionnellement à X i.e $\mathbb{E}(\epsilon|X) = 0$ presque surement.

1.3. Un predicteur sur $\mathbb{R}^p$ x $\mathbb{R}$ est toute fonction mesurable $f : \mathbb{R}^p \longrightarrow \mathbb{R}$. De plus $f(x)$ est la prédiction de $x \in \mathbb{R}^n$ par le prédicteur f.
L'erreur de prédiction théorique associé est :

$$\mathcal{R}_{P,L}(f) = \mathbb{E}(L(f(X), Y))$$

L étant la fonction de coût et P la loi du couple (X,Y).

1.4. Montrons que la problématique d'estimation de la fonction de régression dans le cadre du modèle non-paramétrique est la même que la problématique de prédiction.
En effet si on considère le coût quadratique on a que pour tout prédicteur f : $\mathbb{R}^p \longrightarrow \mathbb{R}$, l'erreur de prédiction généralisée associé au coût quadratique est donnée par :

$$\mathcal{R}_{P,L}(f) = \mathbb{E}((f(X) - Y)^2)$$

Et en utilisant le modèle non-paramétrique on obtient une réécriture de l'erreur quadratique :

$$\mathcal{R}_{P,L}(f) = \mathbb{E}((f(X) - g(X))^2) - \mathbb{E}(\epsilon^2)$$

Etant donné que pour le problème de prédiction nous cherchons le meilleur prédicteur possible relativiment à la fonction de coût ici le coût quadratique sur $\mathbb{R}^p$ x $\mathbb{R}$ tout entier , on cherche donc le prédicteur qui minimise le plus l'écart entre f et g i.e on cherche donc la fonction la plus proche possible de la fonction de regression. Ce qui nous amène à la conclusion que prédire c'est estimer. C'est ainsi que pour toute fonction de cout, la problématique est la même.

1.5. Dans un arbre CART : En régression, on cherche donc des découpes qui tendent à diminuer la variance des nœuds obtenus et en classification, on cherche à diminuer la fonction de pureté de Gini, et donc à augmenter l'homogénéité des nœuds obtenus, un nœud étant parfaitement homogène s'il ne contient que des observations de la même classe.

1.6. Un arbre maximal est un arbre qui est pleinement developpé, toutes ses feuilles vérifient tous les critères d'arrêt (les conditions d'arret d'un découpage de noeuds) i.e :

1. si le nombre d'observations dans le noeud est inférieur à un seuil fixé
2. si l'impureté du noeud est inférieur à un seuil fixé
3. Si le meilleur découpage du noeud mène à une réduction de l'impureté inférieure à un seuil fixé.

Dans ce cas on obtient un arbre maximal si dans chaque feuille, noeud terminal est pur (totalement homogène).
L'arbre maximal conduit à un problème de sur-apprentissage pour éviter ce problème l'élaguage de cet arbre en un sous-arbre optimal est la stratégie utilisée. L'élaguage d'un arbre en sous-arbre optimal consiste en la recherche du sous-arbre qui minimise l'erreur d'apprentissage pénalisée (pénalité proportionnelle au nombre de feuille) ou le critère de coût-complexité. Cependant cette méthode n'est pas utilisé dans le cadre des forêts aléatoires car les forêts aléatoires réalise un bagging ce qui évite le sur-apprentissage.

1.7. Un surrogate split est un découpage alternatif, de substitution, un suppléant en cas de données manquantes.

1.8. On définit $\mathcal{L}_n$ un n échantillon d'apprentissage, soient $q \in \mathbb{N}$, $m \in \{1, ..., p\}$ et $\mathcal{L}_n^{\Theta_1}, ..., \mathcal{L}_n^{\Theta_q}$, q échantillons de bootstrap associés aux tirages $\Theta_1, ..., \Theta_q$. Soient $\hat{f}_m(\mathcal{L}_n^{\Theta_1}; .), ..., \hat{f}_m(\mathcal{L}_n^{\Theta_q}; .)$ les q arbres randomisés construits sur les échantillons de bootstrap.
Le prédicteur par forêt aléatoire RF-RI est donné par :

$$\hat{RF}_{q,m}(\mathcal{L}_n; .) : \forall x \in \mathbb{R}^p \to \frac{1}{q} \sum_{i=1}^{q} \hat{f}_m(\mathcal{L}_n^{\Theta_1}; x)$$

Le schéma de construction d'une forêt aléatoire est alors le suivant.

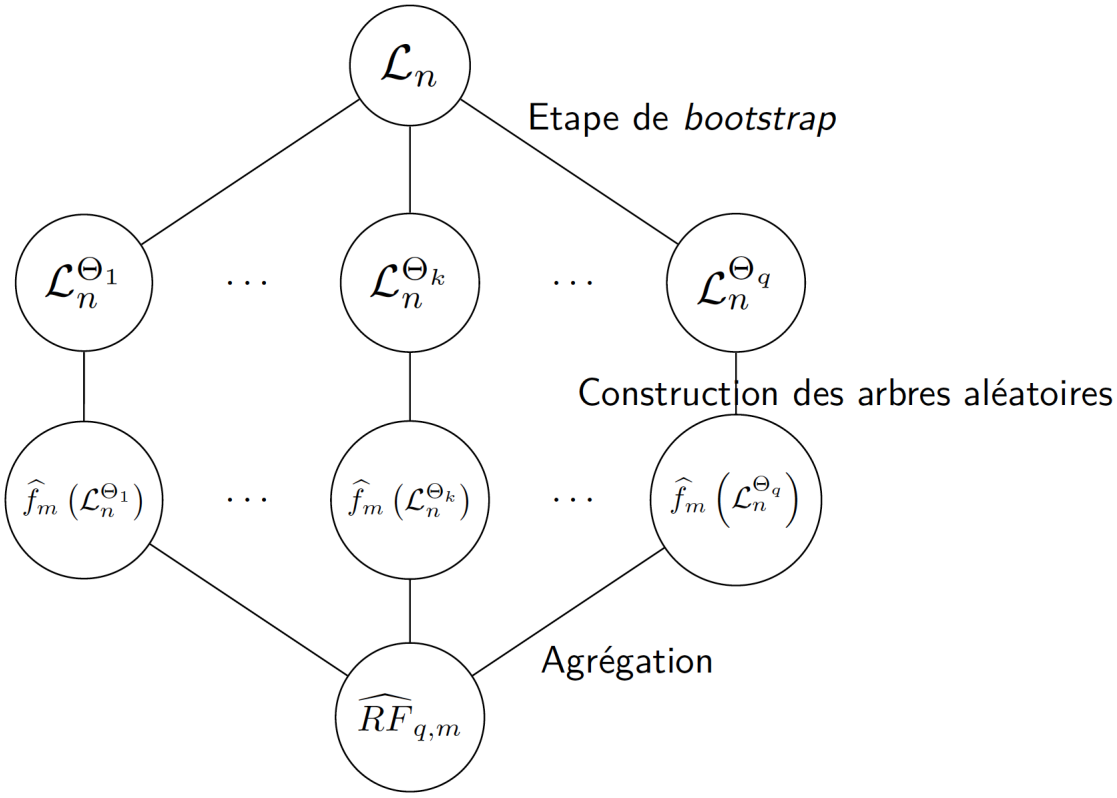

FIGURE 1. Schéma de construction d'une forêt aléatoire

1.9. Les forêts aléatoires ne gèrent pas les données manquantes. En effet la recherche de la scission par substitution nécessite beaucoup de calcul et peut devenir impraticable lors de la croissance d'un grand nombre d'arbres, en particulier pour les arbres complètement saturés utilisés par les forêts. En outre, les divisions de substitution peuvent même ne pas être significatives dans un paradigme de forêt. RF sélectionne des variables de manière aléatoire lors de la division d'un nœud et, de ce fait, les variables d'un nœud peuvent être non corrélées, et une division de substitution raisonnable peut ne pas exister. la division par substitution modifie l'interprétation d'une variable, ce qui affecte des mesures telles que l'Importance de la variable.
En outre l'idée du Bagging, et qu'en appliquant la règle de base sur différents échantillons boots- trap, on en modifie les prédictions, et donc on construit ainsi une collection variée de prédicteurs. L'étape d'agrégation permet alors d'obtenir un prédicteur performant. L'amélioration proposée par les forêts

aléatoires est de baisser la corrélation entre les règles sur les différents échantillons à l'aide d'une étape supplémentaire de randomisation.

1.10. Fixons une observation (X_i, Y_i) de l'échantillon d'apprentissage $\mathcal{L}_n$ et considérons l'ensemble des arbres construits sur les échantillons bootstrap ne contenant pas cette observation, échantillon OOB, c'est-à-dire pour lesquels cette observation est "Out-Of-Bag" on notera

$$O_i = \{j \in 1, ..., q, (X_i, Y_i) \notin \mathcal{L}_n^{\Theta_j}\}$$

. Nous agrégeons alors uniquement les prédictions de ces arbres pour fabriquer notre prédiction $\hat{Y}_i^{OOB}$ de Yi appélé erreur de prédiction OOB définit comme suit :

$$\hat{Y}_i^{OOB} = \frac{1}{\#O_i} \sum_{j \in O_i} \hat{f}_m(\mathcal{L}_n^{\Theta_i}, X_i)$$

Après avoir fait cette opération pour toutes les données de $\mathcal{L}_n$, nous calculons alors l'erreur commise i.e erreur OOB noté $\hat{\mathcal{R}}^{OOB}$:

1. l'erreur quadratique moyenne en régression :

$$\hat{\mathcal{R}}^{OOB} = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{Y}_i^{OOB})^2$$

2. la proportion d'observations mal classées en classification

$$\hat{\mathcal{R}}^{OOB} = \frac{1}{n} \sum_{i=1}^{n} 1_{\{Y_i \neq \hat{Y}_i^{OOB}\}}$$

Cette quantité est appelée erreur OOB.

1.11. L'importance d'une variable donnée est l'accroissement moyen de l'erreur d'un arbre dans la forêt lorsque les valeurs observées de cette variable sont permutées au hasard dans les échantillons OOB.
Fixons $j \in 1, ..., p$ et détaillons le calcul de l'indice d'importance de la variable X^j. Considérons un échantillon bootstrap $\mathcal{L}_n^{\Theta_l}$ et l'échantillon OOB_l associé, c'est-à-dire l'ensemble des observations qui n'apparaissent pas dans $\mathcal{L}_n^{\Theta_l}$.
Calculons $errOOB_l$, l'erreur commise sur OOB_l par l'arbre construit sur $\mathcal{L}_n^{\Theta_l}$ (erreur quadratique moyenne en régression, proportion de mal classés en classification). Permutons alors aléatoirement les valeurs de la j-ième variable j dans l'échantillon OOB_l . Ceci donne un échantillon perturbé, noté $\hat{OOB}_l^j$. Calculons enfin $err\hat{OOB}_l^j$, l'erreur sur l'échantillon $\hat{OOB}_l^j$. Nous effectuons ces opérations pour tous les échantillons bootstrap. L'importance de la variable X^j, $VI(X^j)$, est définie par la différence entre l'erreur moyenne d'un arbre sur l'échantillon OOB perturbé et celle sur l'échantillon OOB :

$$VI(X^j) = \frac{1}{q}\sum_{l=1}^{q}(errO\hat{O}B_l^j - errOOB_l)$$

Ainsi, plus les permutations aléatoires de la j-ième variable engendrent une forte augmentation de l'erreur, plus la variable est importante. A l'inverse, si les permutations n'ont quasiment aucun effet sur l'erreur (voire même la diminuent ce qui fait que VI peut être légèrement négative), la variable est considérée comme très peu importante.

2. Exercice 1

2.1. Traçons l'arbre correspondant à la figure 1, en précisant pour chaque noeud la coupure et pour chaque feuille la valeur associé.

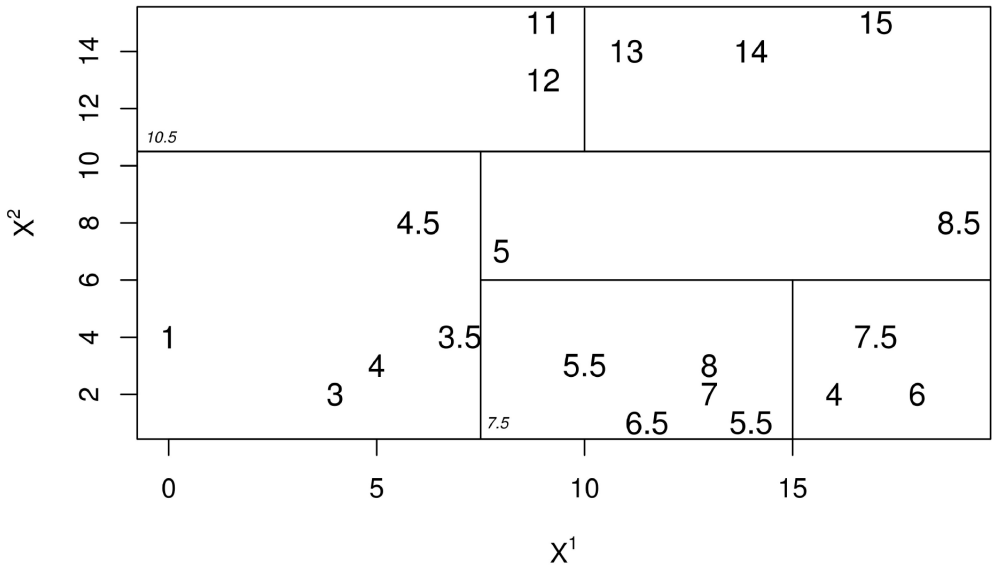

Figure 2

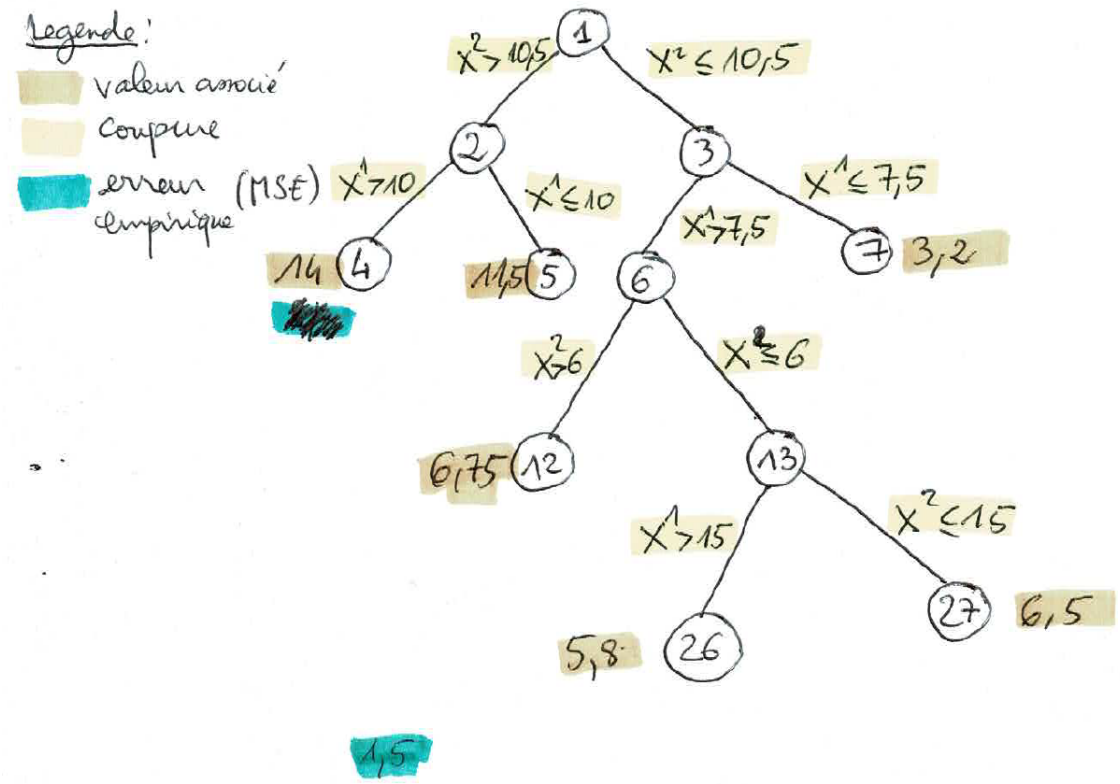

FIGURE 3. Arbre CART associé à la figure 2

2.2. Notre arbre CART possède 6 feuilles et est de profondeur 4.

2.3. Calcul de l'erreur empirique par arbre associé sur les données d'apprentissage :
Ici n = 20 , et le MSE = 1.5.

3. Exercice 2

3.1. Le mot Bagging est la contraction des mots Bootstrap et Aggregating. Étant donné un échantillon d'apprentissage $\mathcal{L}_n$ et une méthode de prédiction (appelée règle de base), qui construit sur $\mathcal{L}_n$ un prédicteur $\hat{h}(.;\mathcal{L}_n)$. Le principe du Bagging est de tirer indépendamment plusieurs échantillons bootstrap $(\mathcal{L}_n^{\Theta_1}, ..., \mathcal{L}_n^{\Theta_q})$, d'appliquer la règle de base sur chacun d'eux pour obtenir une collection de prédicteurs $(\hat{h}(\mathcal{L}_n^{\Theta_1}), ..., \hat{h}(\mathcal{L}_n^{\Theta_q}))$, et enfin d'agréger ces prédicteurs de base. L'idée du Bagging, et qu'en appliquant la règle de base sur

différents échantillons bootstrap, on en modifie les prédictions, et donc on construit ainsi une collection variée de prédicteurs. L'étape d'agrégation permet alors d'obtenir un prédicteur performant.

3.2. Soit $(Z_i)_{1\leq i\leq q}$ des variables aléatoires réelles de même loi. On suppose que la variance est égale à $\sigma^2 \neq \inf$ et que la corrélation pour deux variables aléatoires distinctes est égale à ρ. On montre par la suite :

$$Var(\frac{1}{q}\sum_{l=1}^{q} Z_l) = \rho\sigma^2 + \frac{1-\rho}{q}\sigma^2$$

On a :
On utilisera le fait que $Cov(Z_i, Z_j) = \sigma_i\sigma_j cor(Z_i, Z_j)$

$$\begin{aligned} Var(\frac{1}{q}\sum_{l=1}^{q}) &= \frac{1}{q^2}\sum_{l=1}^{q} Var(Z_l) + 2\frac{1}{q^2}\sum_{1\leq i<j\leq q} Cov(Z_i, Z_j) \\ &= \frac{1}{q^2}\sigma^2 + 2\frac{1}{q^2}\sum_{1\leq i<j\leq q}\sigma^2\rho \\ &= \frac{1}{q^2}\sigma^2 + 2\frac{q(q-1)}{2q^2}p\sigma^2 \\ &= \rho\sigma^2 + \frac{1-\rho}{q}\sigma^2. \end{aligned}$$

Ensuite on calculera la $\lim_{q\to\infty} Var(\frac{1}{q}\sum_{l=1}^{q} Z_l)$

$$\begin{aligned} \lim_{q\to\infty} Var(\frac{1}{q}\sum_{l=1}^{q} Z_l) &= \lim_{q\to\infty} \rho\sigma^2 + \frac{1-\rho}{q}\sigma^2 \\ &= \rho\sigma^2 \end{aligned}$$

3.3. Les Z_l sont donc les différents prédicteurs, σ^2 leur variance respectifs et ρ la corrélation entre eux. Le fait que la variance de la moyenne des différents prédicteurs soit inférieure à la variance du prédicteur indiviellement ($\rho\sigma^2 + \frac{1-\rho}{q}\sigma^2 \leq \sigma^2$) nous permet de conclure que le Bagging permet de réduire la variance du prédicteur à travers l'aggrégation, la corrélation entre les prédicteurs doit être la plus faible possible.

4. Exercice 3

4.1. Dans la sortie R pour les arbre CART, nous sommes dans un cadre de classification parce que les variables prédictes sont des classes 1,2 et 3 et il ya 3 probabilité à postériori (yval). Il ya 5 variables V1,V2,V3,V4 et V5. Comme dit plus haut il ya 3 classes et pour chaque classe voici le nombre d'observation

dans le jeu d'apprentissage . On utiliser pour cela les probabilité dans le noeud racine.

1. classe 1 = 1000*0.286 = 286
2. classe 2 = 1000*0.261 = 261
3. classe 3 = 1000*0.453 = 453

4.2. L'arbre est de profondeur 2 et a 4 feuilles. Les classes pouvant etre prédites sont 1,2,3. L'arbre associé et la classification des observations x_1, x_2, x_3 :

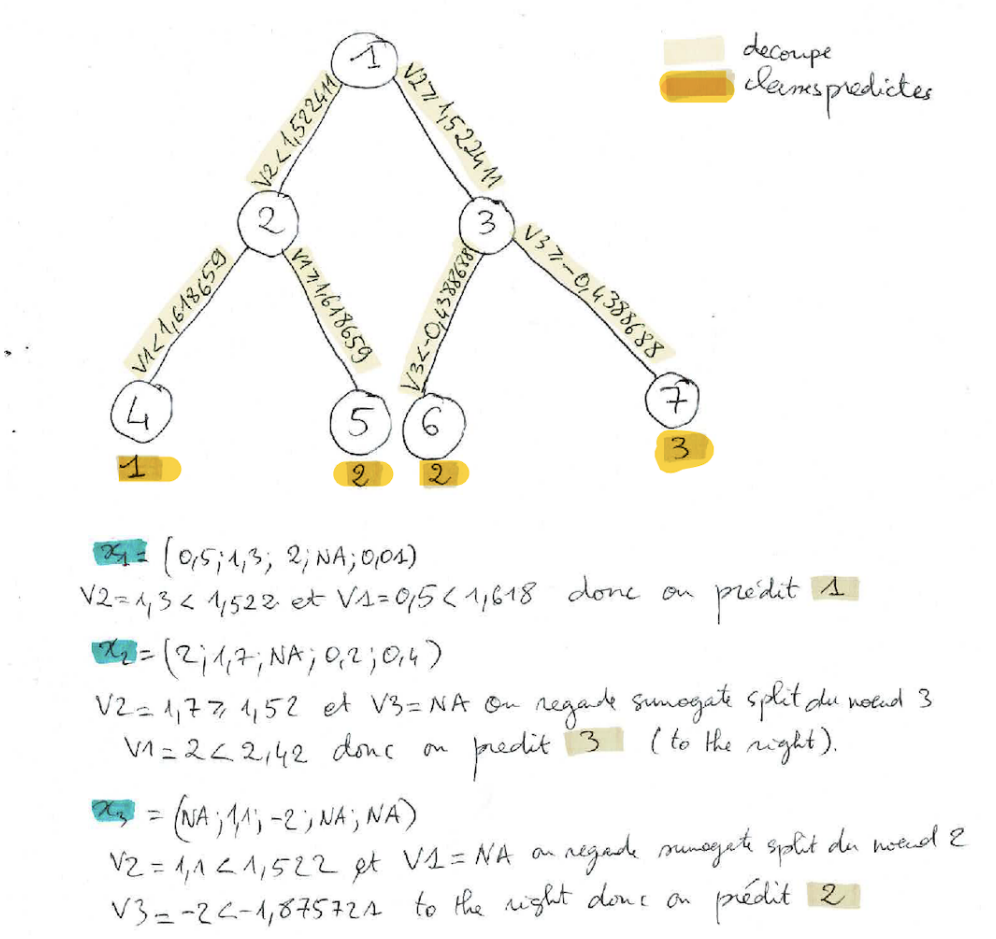

FIGURE 4. Arbre CART associé à la figure 2

4.3. Une sélection de variable sérait d'apres les sorties R : V3,V1 et V2.

4.4. Le paramètre de pénalisation de l'arbre est cp = 0.01.
L'erreur de prédiction est err = (9+3)/1000 = 12/1000 = 0.012, soit 1,2%
La variance associé est var = err - err^2 = 0.012

4.5. Pour la forêt aléatoire le mtry = 2 et le ntrees = 500.
On remarque donc que l'erreur de classification du random forest de 0.5% est plus basse ue celle de l'arbre CART donc le randoms forest est meilleure que l'arbre CART. Avec les forets aléatoires on obtient 5 mal classé tandis que avec l'arbre CART on en obtient 12.

January 2, 2020

H. M., Bordeaux, France • *E-mail :* mehdyhkn@yahoo.com