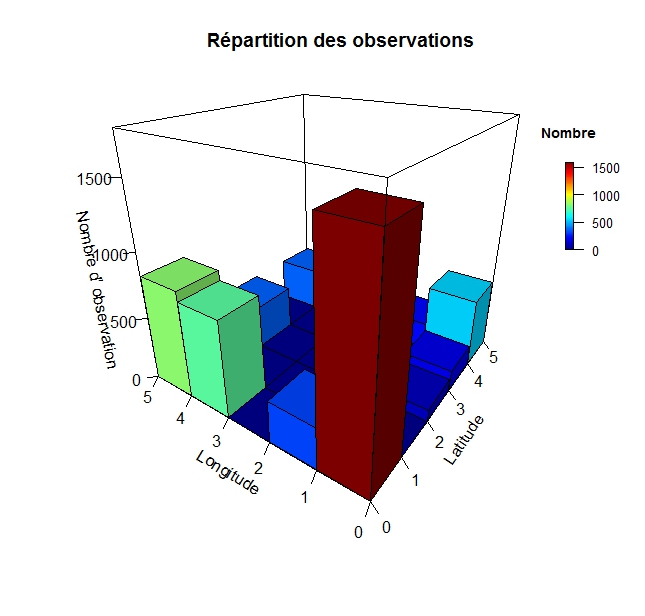
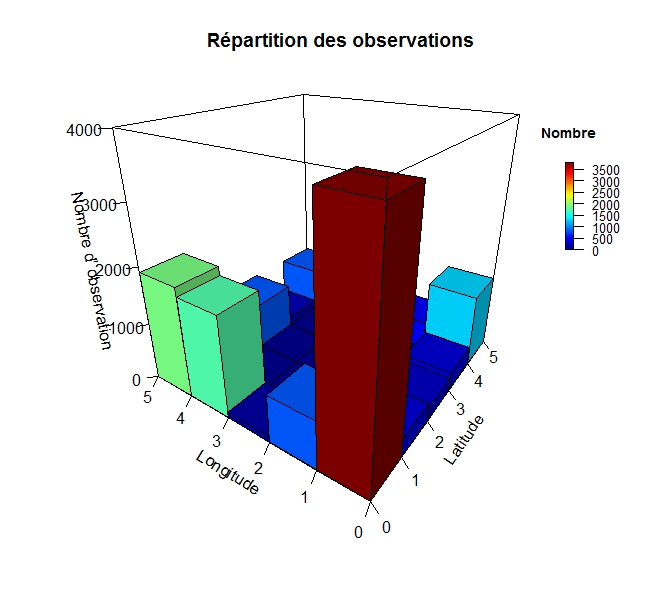
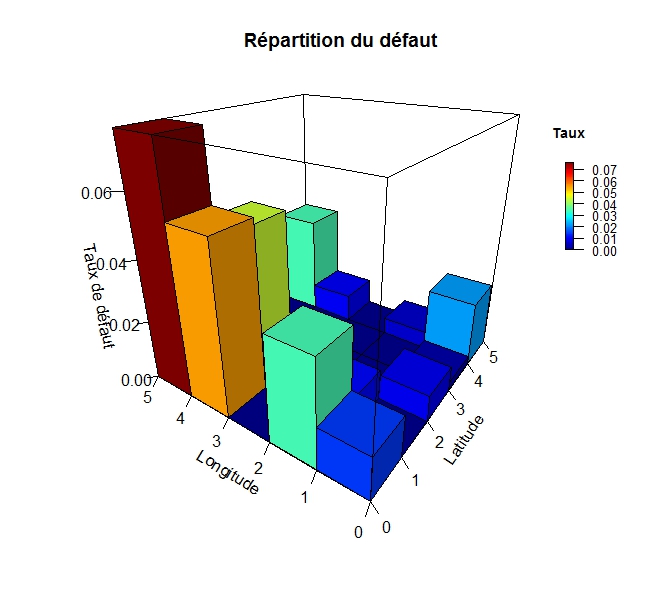
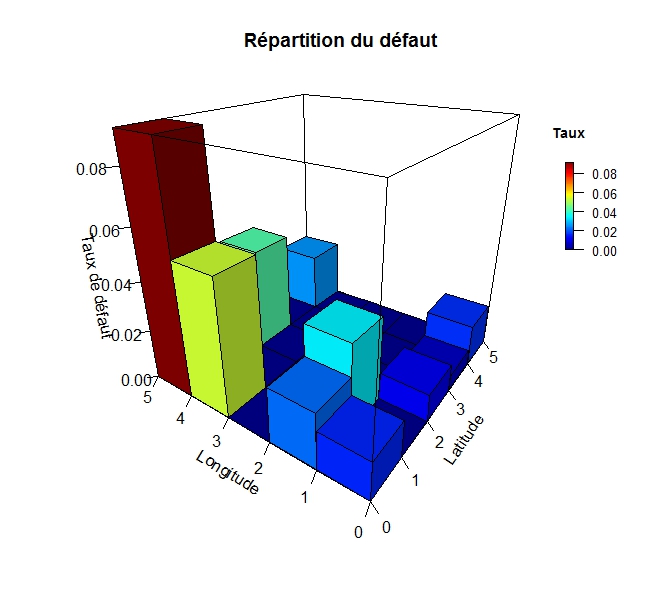
**Partie de SOM :**

SOM non-supervisé :

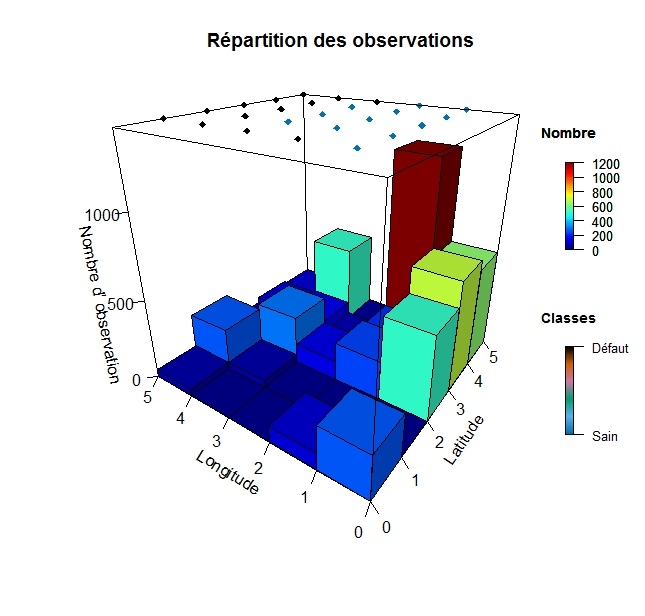
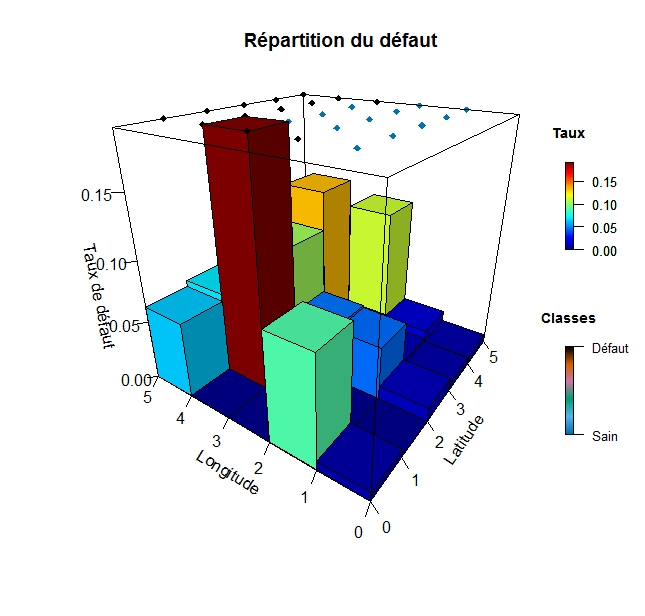


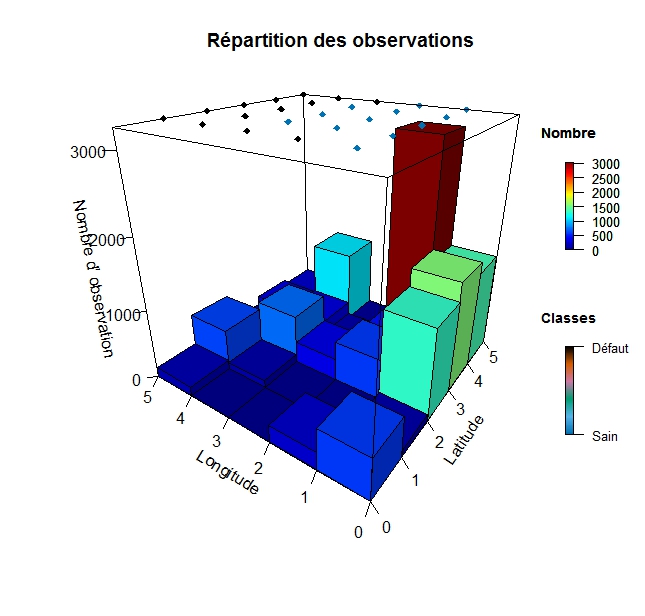
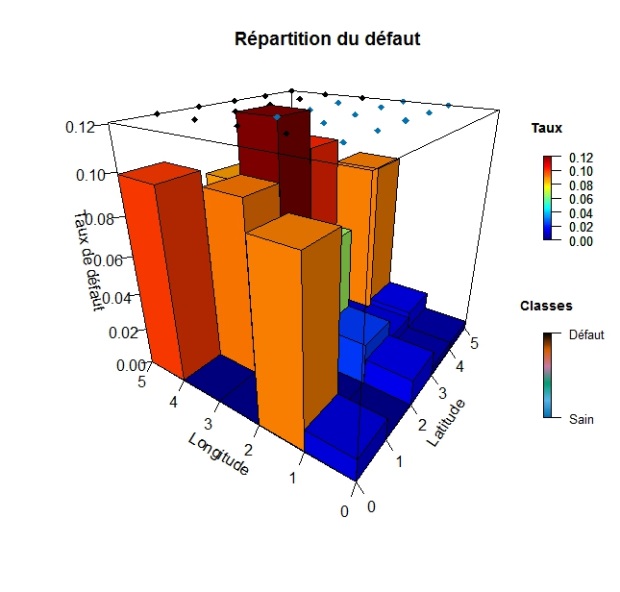




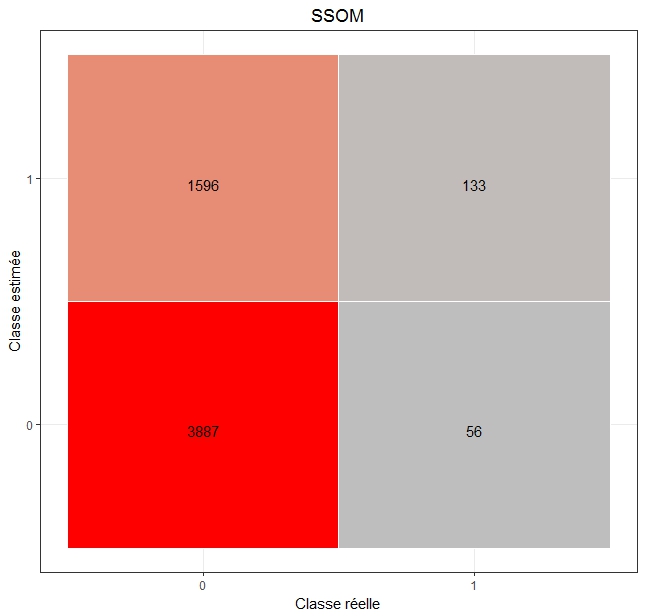
SOM supervisé :

En utilisant la technique de sous-échantillonnage et ajoutant une couche d’étiquette, la carte Kohonen supervisée s’améliore en le pou discrimination





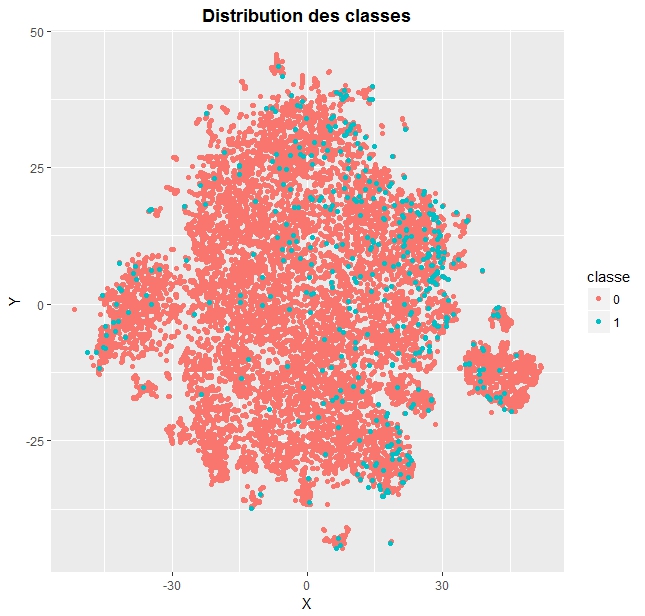
SSOM matrice de confusion sur la base de test : (5672 observations):



Le manque de la distinction du côté de la structure topologique:

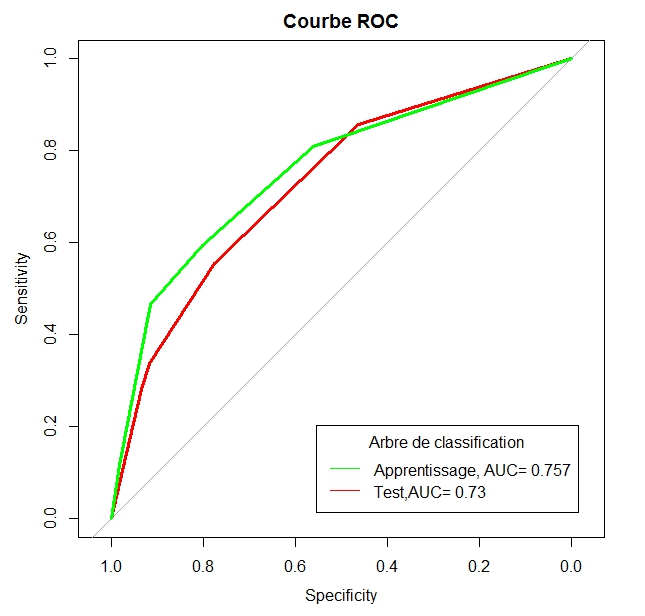
Base de l’iris :

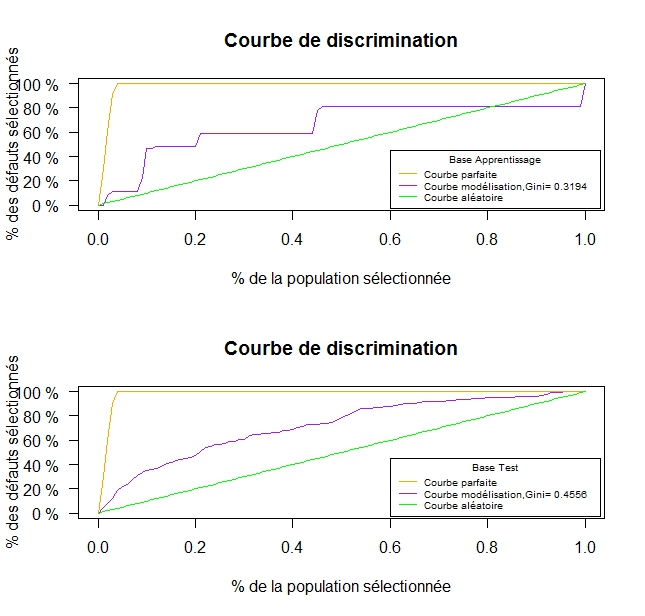
Base du défaut :



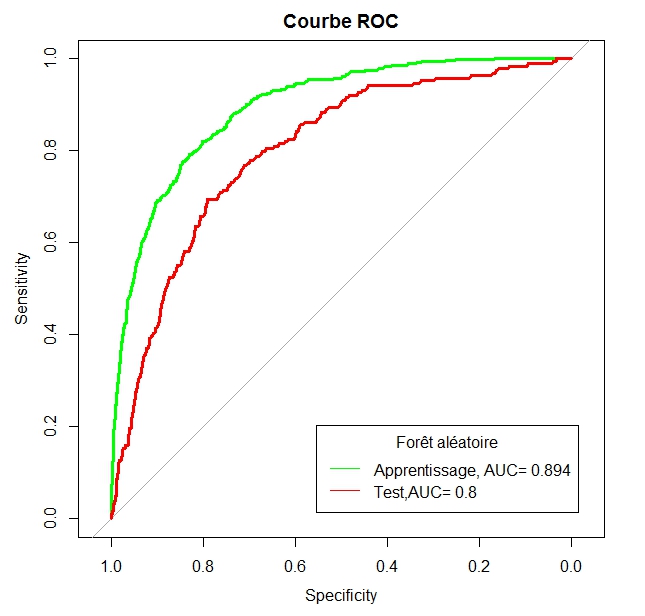
**Partie des méthodes supervisées :**

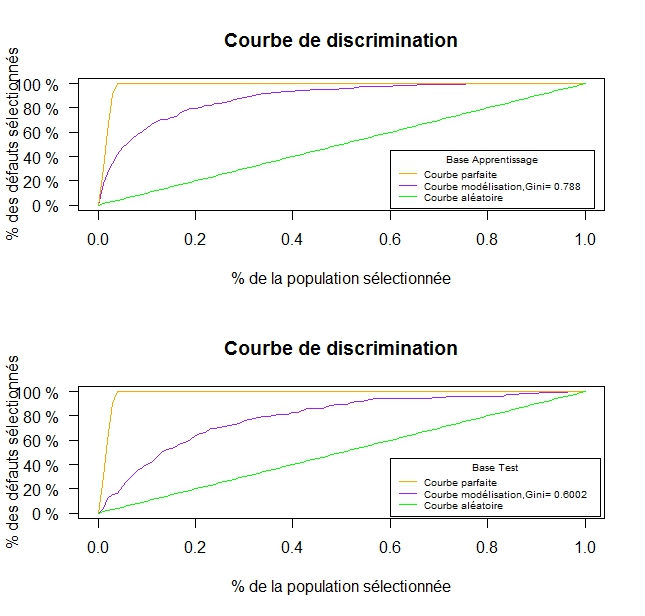
1. Arbre de classification (Modèle référent)



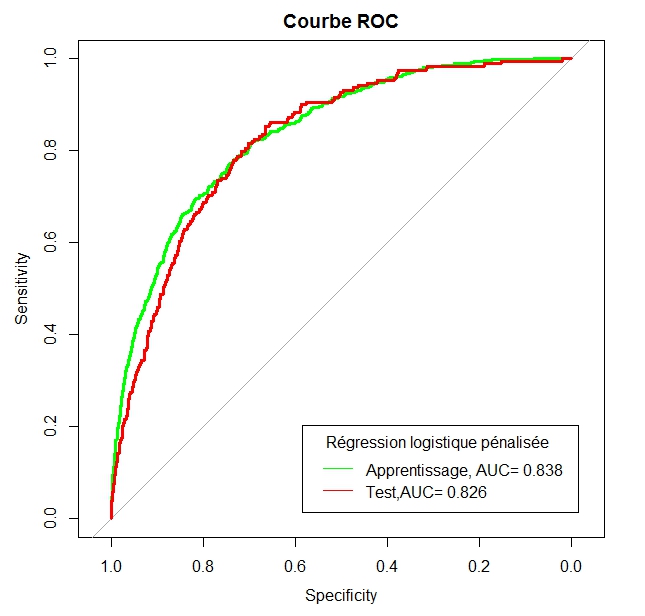


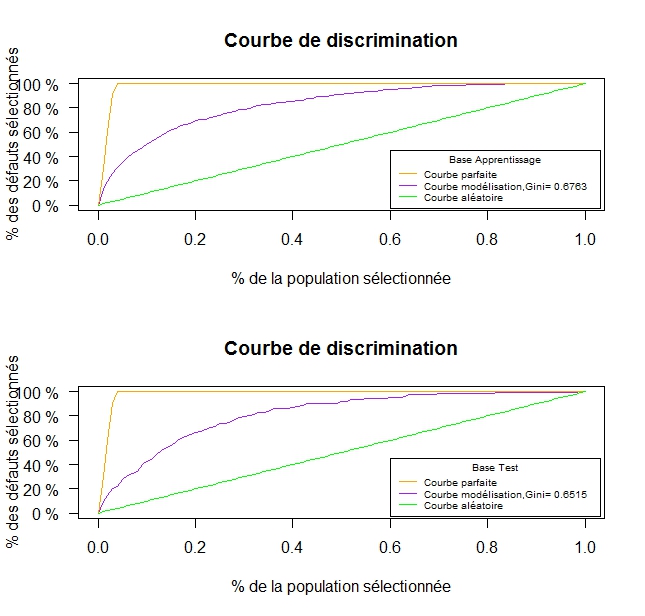
1. Forêt aléatoire



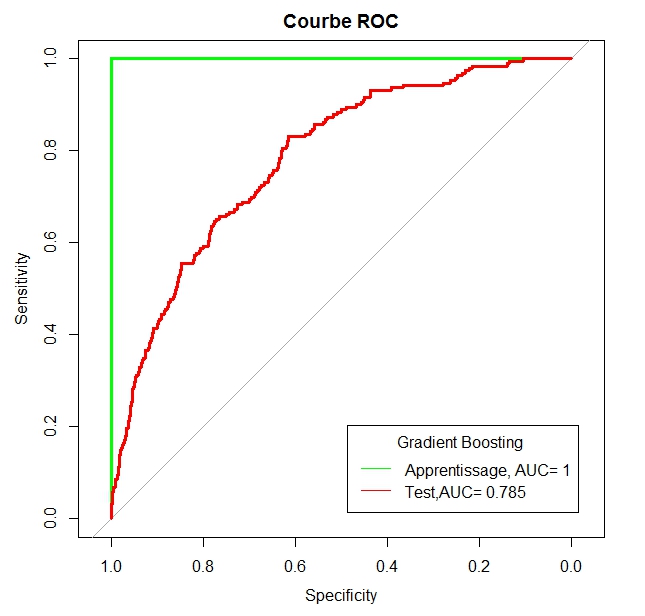


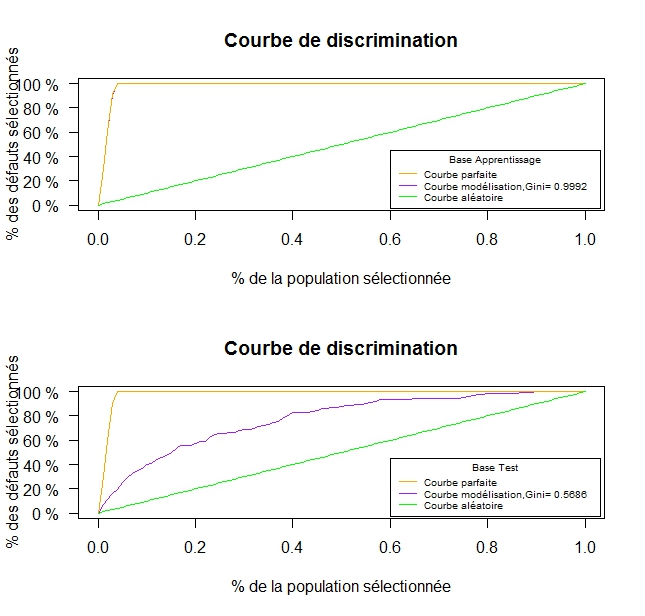
1. Régression logistique pénalisée en utilisant Lasso





1. Gradient boosting :





**Partie de la méthode hybride :**

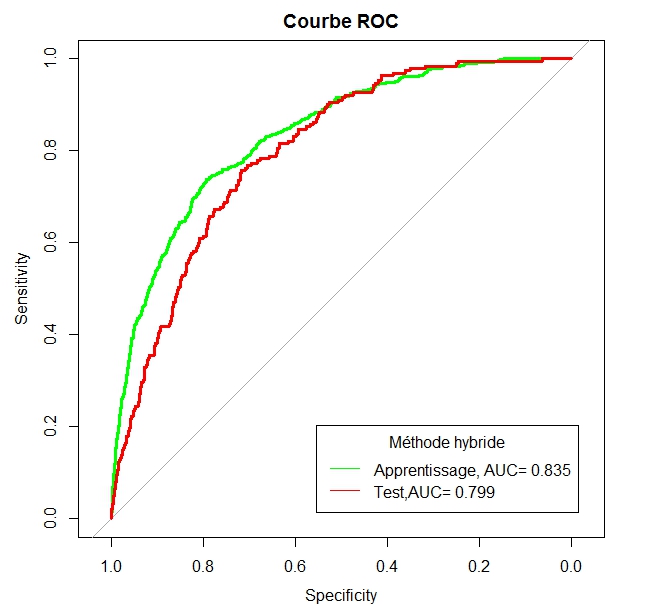
Cette méthode combine la carte Kohonen et les méthodes supervisée en prenant compte de l’hétérogénéité géométrique.

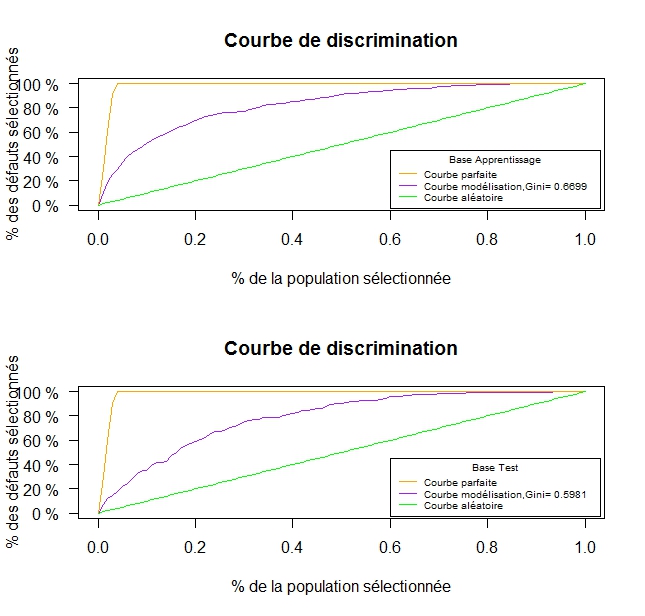
**Hétérogénéité géométrique :**

En raison de la limite de la technique de collecte de données et le renouvellement de la base, il est difficile d’avoir une performance distinguée à partir d’une base d’apprentissage avec l’hétérogénéité.

**La méthode hybride proposée:**

1. Utiliser la carte Kohonen pour diviser les clients sains en plusieurs parties
2. Sur chaque carré de la carte, on construit un classifieur local (ici SVM) pour les clients défaillants et les clients sains partiels
3. Calculer la ’distance signé’ entre les individus et l’hyperplan déterminé par les classifieurs locaux. On projette les individus de l’espace de données à l’espace défini par la ’distance signé’, soit notre variable synthétique.
4. Effectuer une régression logistique lasso sur les variables synthétiques





**Partie de la sélection de variables :**

Nous implémentons six méthodes appartenant à deux catégories : méthodes ‘filter’  et méthodes "embedded".

Méthodes ‘wrapper’  n’est pas prise en compte à cause la grande taille des variables.

Méthodes ‘filter’ : Low variance; Univariate feature selection.

Méthodes ‘embedded’ : L1-Lasso ; Forêt aléatoire ; Gradient Boosting ; Stability selection.

En outre, la moyenne des résultats obtenus est choisie pour construire un nouvel indicateur de l’importance de variable.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | glmImp | rfIMP | xgbImp | stabImp | varImp | chiImp | moyImp |
| COT\_BDF\_ENTREP.Sup | 15,77 | 0,08 | 0,01 | 10,31 | 0,07 | 4,60 | 5,14 |
| TOP\_INCID\_SIREN.0 | 17,54 | 0,08 | 0,02 | 9,79 | 0,04 | 3,02 | 5,08 |
| NOTE\_MOTEUR\_T | 3,55 | 2,66 | 4,38 | 10,50 | 0,22 | 6,92 | 4,70 |
| TOP\_INCD\_12M\_SIREN\_C.0 | 6,53 | 1,87 | 0,01 | 10,61 | 0,93 | 5,99 | 4,32 |
| NBJ\_DEPASS\_MOIS\_SIREN\_12M\_2 | 1,97 | 4,79 | 0,10 | 10,39 | 0,29 | 5,06 | 3,77 |
| NB\_JOUR\_DEBT\_SIREN\_12M | 1,14 | 3,51 | 1,68 | 7,67 | 0,29 | 4,62 | 3,15 |
| ALLUR\_CAV | 0,97 | 1,24 | 0,08 | 7,53 | 0,28 | 4,00 | 2,35 |
| NOTCOF | 2,52 | 0,37 | 0,10 | 6,92 | 0,55 | 2,32 | 2,13 |
| COT\_BDF\_ENTREP.3 | 7,81 | 0,10 | 0,02 | 0,00 | 1,03 | 0,75 | 1,62 |
| COT\_BDF\_ENTREP.5 | 4,01 | 0,06 | 0,01 | 3,31 | 0,81 | 1,36 | 1,59 |
| TOP\_INCD\_12M\_SIREN\_C.0.1 | 0,00 | 2,53 | 0,00 | 0,00 | 0,93 | 5,99 | 1,58 |
| TOP\_INCD\_12M\_SIREN\_C.1.1 | 0,74 | 1,66 | 0,00 | 0,00 | 0,93 | 5,99 | 1,55 |
| TOP\_INCD\_12M\_SIREN\_C.1 | 0,00 | 2,14 | 0,00 | 0,10 | 0,93 | 5,99 | 1,53 |
| TOP\_INCID\_SIREN.1 | 0,00 | 0,07 | 0,00 | 4,34 | 0,04 | 3,02 | 1,25 |
| SLDE\_MOY\_SIREN\_3M | 0,17 | 5,57 | 1,47 | 0,16 | 0,00 | 0,08 | 1,24 |
| MNT\_EE\_CLI\_12M\_SIREN | 0,91 | 2,16 | 0,75 | 2,29 | 0,02 | 0,33 | 1,07 |
| COD\_ACT\_REF.6 | 5,85 | 0,00 | 0,05 | 0,08 | 0,29 | 0,14 | 1,07 |
| NOTMOT | 0,00 | 1,14 | 0,09 | 0,09 | 0,64 | 4,45 | 1,07 |
| SLDE\_MOY\_SIREN\_1M | 0,00 | 4,37 | 1,50 | 0,00 | 0,00 | 0,17 | 1,01 |
| PD\_COFACE | 0,00 | 0,41 | 0,09 | 2,80 | 0,55 | 1,98 | 0,97 |

