Projet 2 de SY09

Anh Tu NGUYEN et Audrick LIBBRAIRE

29 Juin 2018

Table des matières

1 Introduction

Pour ce deuxième rendu de projet de l'UV SY09, nous avons appliqué plusieurs méthodes vues en cours, qui concernent la discrimination et la régression. La discrimination a pour objectif l'apprentissage d'un modèle qui permettra de déterminer la classe à partir des observations des variables explicatives pour de nouveaux individus. La régression nous permet de lier les variables observées à la variable à expliquer pour déduire la valeur d'une variable quantitative. Dans le cadre de ce projet, nous avons effectué des analyses pour évaluer les performances des principales méthodes de discrimination étudiées en cours : analyse discriminante quadratique et linaire, classifieur bayésien naïf, regression logistique et arbre de décision sur différents jeux de données en utilisant les fonctions vues en TDs et celles de la biliothèque **MASS** de R.

2 Discrimination

2.1 Breastcancer

2.1.1 Analyse exploratoire

Tout d'abord, nous avons effectué une analyse exploratoire pour comprendre et s'approcher les données fournies.
Ce jeu de données comporte 30 variables et 569 individus. Nous avons exclu la première colonne car elle correspondait à l'index, qui est inutile dans notre analyse. Cela nous à permis d'obtenir le jeu de données à analyser.
À partir des résultats ci-dessus, on peut voir que toutes les variables sont des variables numériques.

```
> str(breastcancer)
'data.frame':   569 obs. of  30 variables:
 $ X1 : num  18 11.4 12.4 13 15.8 ...
 $ X2 : num  10.4 20.4 15.7 21.8 17.9 ...
 $ X3 : num  122.8 77.6 82.6 87.5 103.6 ...
 $ X4 : num  1001 386 477 520 781 ...
 $ X5 : num  0.1184 0.1425 0.1278 0.1273 0.0971 ...
 $ X6 : num  0.278 0.284 0.17 0.193 0.129 ...
 $ X7 : num  0.3001 0.2414 0.1578 0.1859 0.0995 ...
 $ X8 : num  0.1471 0.1052 0.0809 0.0935 0.0661 ...
 $ X9 : num  0.242 0.26 0.209 0.235 0.184 ...
 $ X10: num  0.0787 0.0974 0.0761 0.0739 0.0608 ...
 $ X11: num  1.095 0.496 0.335 0.306 0.506 ...
 $ X12: num  0.905 1.156 0.89 1.002 0.985 ...
 $ X13: num  8.59 3.44 2.22 2.41 3.56 ...
 $ X14: num  153.4 27.2 27.2 24.3 54.2 ...
 $ X15: num  0.0064 0.00911 0.00751 0.00573 0.00577 ...
 $ X16: num  0.049 0.0746 0.0335 0.035 0.0406 ...
 $ X17: num  0.0537 0.0566 0.0367 0.0355 0.0279 ...
 $ X18: num  0.0159 0.0187 0.0114 0.0123 0.0128 ...
 $ X19: num  0.03 0.0596 0.0216 0.0214 0.0201 ...
 $ X20: num  0.00619 0.00921 0.00508 0.00375 0.00414 ...
 $ X21: num  25.4 14.9 15.5 15.5 20.4 ...
 $ X22: num  17.3 26.5 23.8 30.7 27.3 ...
 $ X23: num  184.6 98.9 103.4 106.2 136.5 ...
 $ X24: num  2019 568 742 739 1299 ...
 $ X25: num  0.162 0.21 0.179 0.17 0.14 ...
 $ X26: num  0.666 0.866 0.525 0.54 0.561 ...
 $ X27: num  0.712 0.687 0.535 0.539 0.397 ...
 $ X28: num  0.265 0.258 0.174 0.206 0.181 ...
 $ X29: num  0.46 0.664 0.399 0.438 0.379 ...
 $ X30: num  0.119 0.173 0.124 0.107 0.105 ...
```

```
> summary(breastcancer)
       X1              X2              X3              X4              X5              X6              X7              X8              X9              X10        
 Min.   : 6.981   Min.   : 9.71   Min.   : 43.79   Min.   : 143.5   Min.   :0.05263   Min.   :0.01938   Min.   :0.00000   Min.   :0.00000   Min.   :0.1060   Min.   :0.04996  
 1st Qu.:11.700   1st Qu.:16.17   1st Qu.: 75.17   1st Qu.: 420.3   1st Qu.:0.08637   1st Qu.:0.06492   1st Qu.:0.02956   1st Qu.:0.02031   1st Qu.:0.1619   1st Qu.:0.05770  
 Median :13.370   Median :18.84   Median : 86.24   Median : 551.1   Median :0.09587   Median :0.09263   Median :0.06154   Median :0.03350   Median :0.1792   Median :0.06154  
 Mean   :14.127   Mean   :19.29   Mean   : 91.97   Mean   : 654.9   Mean   :0.09636   Mean   :0.10434   Mean   :0.08880   Mean   :0.04892   Mean   :0.1812   Mean   :0.06280  
 3rd Qu.:15.780   3rd Qu.:21.80   3rd Qu.:104.10   3rd Qu.: 782.7   3rd Qu.:0.10530   3rd Qu.:0.13040   3rd Qu.:0.13070   3rd Qu.:0.07400   3rd Qu.:0.1957   3rd Qu.:0.06612  
 Max.   :28.110   Max.   :39.28   Max.   :188.50   Max.   :2501.0   Max.   :0.16340   Max.   :0.34540   Max.   :0.42680   Max.   :0.20120   Max.   :0.3040   Max.   :0.09744  
      X11              X12             X13              X14              X15               X16               X17               X18               X19               
 Min.   :0.1115   Min.   :0.3602   Min.   : 0.757   Min.   :  6.802   Min.   :0.001713   Min.   :0.002252   Min.   :0.00000   Min.   :0.000000   Min.   :0.007882  
 1st Qu.:0.2324   1st Qu.:0.8339   1st Qu.: 1.606   1st Qu.: 17.850   1st Qu.:0.005169   1st Qu.:0.013080   1st Qu.:0.01509   1st Qu.:0.007638   1st Qu.:0.015160  
 Median :0.3242   Median :1.1080   Median : 2.287   Median : 24.530   Median :0.006380   Median :0.020450   Median :0.02589   Median :0.010930   Median :0.018730  
 Mean   :0.4052   Mean   :1.2169   Mean   : 2.866   Mean   : 40.337   Mean   :0.007041   Mean   :0.025478   Mean   :0.03189   Mean   :0.011796   Mean   :0.020542  
 3rd Qu.:0.4789   3rd Qu.:1.4740   3rd Qu.: 3.357   3rd Qu.: 45.190   3rd Qu.:0.008146   3rd Qu.:0.032450   3rd Qu.:0.04205   3rd Qu.:0.014710   3rd Qu.:0.023480  
 Max.   :2.8730   Max.   :4.8850   Max.   :21.980   Max.   :542.200   Max.   :0.031130   Max.   :0.135400   Max.   :0.39600   Max.   :0.052790   Max.   :0.078950  
      X20               X21              X22              X23              X24              X25               X26               X27               X28               X29         
 Min.   :0.0008948   Min.   : 7.93   Min.   :12.02   Min.   : 50.41   Min.   : 185.2   Min.   :0.07117   Min.   :0.02729   Min.   :0.0000   Min.   :0.00000   Min.   :0.1565  
 1st Qu.:0.0022480   1st Qu.:13.01   1st Qu.:21.08   1st Qu.: 84.11   1st Qu.: 515.3   1st Qu.:0.11660   1st Qu.:0.14720   1st Qu.:0.1145   1st Qu.:0.06493   1st Qu.:0.2504  
 Median :0.0031870   Median :14.97   Median :25.41   Median : 97.66   Median : 686.5   Median :0.13130   Median :0.21190   Median :0.2267   Median :0.09993   Median :0.2822  
 Mean   :0.0037949   Mean   :16.27   Mean   :25.68   Mean   :107.26   Mean   : 880.6   Mean   :0.13237   Mean   :0.25427   Mean   :0.2722   Mean   :0.11461   Mean   :0.2901  
 3rd Qu.:0.0045580   3rd Qu.:18.79   3rd Qu.:29.72   3rd Qu.:125.40   3rd Qu.:1084.0   3rd Qu.:0.14600   3rd Qu.:0.33910   3rd Qu.:0.3829   3rd Qu.:0.16140   3rd Qu.:0.3179  
 Max.   :0.0298400   Max.   :36.04   Max.   :49.54   Max.   :251.20   Max.   :4254.0   Max.   :0.22260   Max.   :1.05800   Max.   :1.2520   Max.   :0.29100   Max.   :0.6638  
      X30         
 Min.   :0.05504  
 1st Qu.:0.07146  
 Median :0.08004  
 Mean   :0.08395  
 3rd Qu.:0.09208  
 Max.   :0.20750  
```

FIGURE 1 – La structure du jeu de données **breastcancer**

La plupart des variables sont comprises entre 0 et 1. Il y a quelques variables ayant des valeurs qui

sont beaucoup plus grandes que les autres par exemple $X4$ et $X24$. D'autre part, dans le but d'étudier la corrélation entre les variables qui peut produire des influences dans les méthodes de discrimination. Le test de corrélations a été réalisé par la méthode de Pearson avec **p-value** fixée par défaut avec

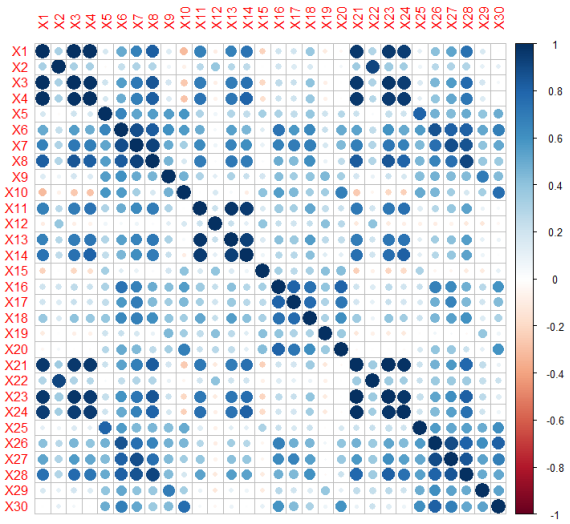

FIGURE 2 – La corrélation entre les variables de **breastcancer**

$\alpha = 0.05$. Grâce au graphique, on peut observer facilement qu'il y a des paires très corrélées comme $(X1, X3)$, $(X1, X4)$, $(X21, X3)$. L'étude de la corrélation est poursuivie par la réalisation d'analyse en composantes principales. Nous avons décidé de réaliser l'analyse en composante principales parce qu'on veut dans un premier temps explorer la géométrie des données et à partir de cela, on peut prévoir le modèle compatible. L'analyse en composantes principales ACP est un outil permettant d'annalyser et visualiser un jeu de données contenant des individus décrits par plusieurs variables. L'ACP nous donne une première vue et une première idée de la géométrie de nos données. Nous avons obtenu d'abord deux graphiques : le cercle de corrélation et l'inertie de valeurs propres des 10 premières valeurs en utilisant la fonction *princomp* et des fonctions du biliothèque *factoextra*. Dans le cadre de

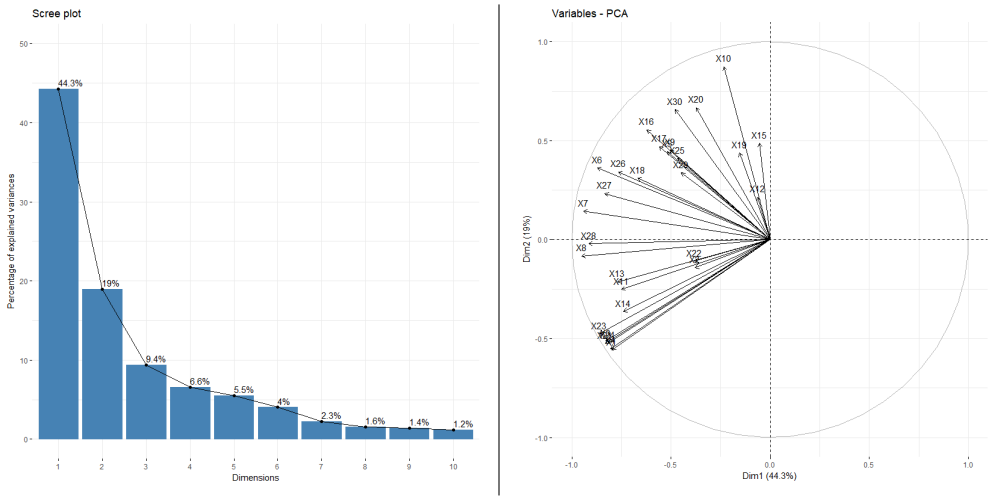

FIGURE 3 – Graphique a : L'inertie de valeurs propres des 10 premières valeurs - Graphique b : Le cercle de corrélation des variables **breastcancer**

notre analyse, nous pouvons voir que les informations se concentrent aux 7 premières composantes avec plus de 90% des informations totales. De plus, le cercle de corrélation des variables expliqué par les deux premières composantes nous montrent la corrélation entre les variables. On peut voir

dans le cercle de corrélation que toutes les variables ont tendance l'inverse au sens de la deuxième composantes et cela est intéressant si on observe la distribution des individus dans le graphique des individus ci-dessous : La distribution des individus dans le graphique est très claire : les individus de

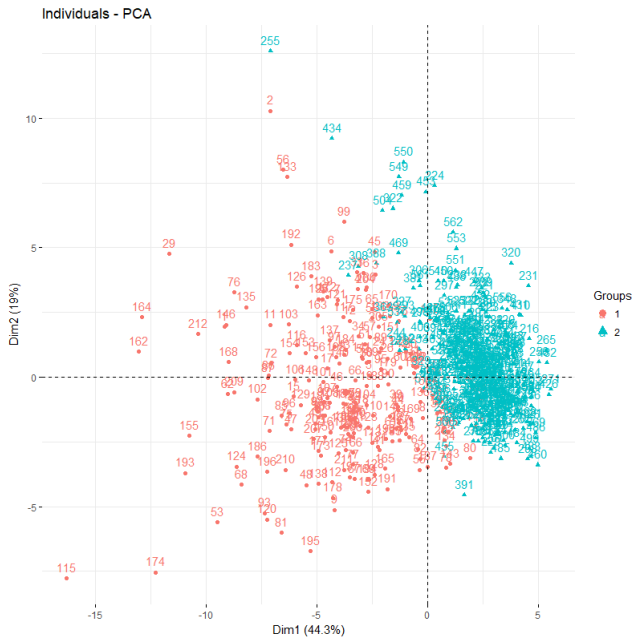

FIGURE 4 – La distribution des individus suivant les 2 premières composantes **breastcancer**

classe 1 sont distribués à gauche de l'axe $y = 0$ et sont distribués selon les variables (les individus se situent très proche des vecteurs des variables) tandis que les individus de classe 2 se situent à droite de l'axe $y = 0$ et ont tendance à être plus concentrés autour de l'axe $x = 0$. Les deux classes sont séparées assez clairement donc nous pouvons prévoir que l'analyse discriminante linéaire peut nous donner une bonne classification. Pour vérifier notre prédiction, nous ferons la réalisation des méthodes de discrimination sur ce jeu de données.

2.1.2 Réalisation des méthode de discrimination

Pour calculer le taux d'erreur ponctuelle $\hat{\epsilon}$ d'une prédiction, il faut simplement compter le nombre de prédictions erronées ϵ_i entre la prédiction faite et les données réelles (d'apprentissage ou de test), puis diviser par le nombre total d'individus n.
On répète cette procédure 100 fois pour obtenir 100 réalisations du taux d'erreur. Pour obtenir l'estimation du taux d'erreur $\bar{\epsilon}$, il ne reste plus qu'à faire la moyenne des 100 réalisations.
L'intervalle de confiance à 95% est obtenu comme-ci : $IC = \left[\mu - 1.96 * \frac{s^*}{\sqrt{N}}; \mu + 1.96 * \frac{s^*}{\sqrt{N}}\right]$ avec :

— μ : La moyenne des 100 réalisation de ϵ

— s^* : L'écart-type corrigé des 100 réalisations de ϵ

— N : Le nombre de réalisation et dans ce cas $N = 100$

Nous avons appliqué les différents modèles d'analyses discriminantes, de régression logistique ainsi que les arbres de décisions. Le résume des résultats sont ci-dessous.

Méthode	**Données**	$\overline{\varepsilon}$	IC
Analyse discriminante quadratique	Apprentissage	0.6104	[0.6092; 0.6118]
	Test	0.6109	[0.6093; 0.6126]
Analyse discriminante linéaire	Apprentissage	0.0800	[0.0785; 0.0816]
	Test	0.0884	[0.0846; 0.0921]
Classifieur bayésien naïf	Apprentissage	0.3702	[0.3699; 0.3704]
	Test	0.3722	[0.3717; 0.3726]
Régression logistique (intercept=1)	Apprentissage	0.0327	[0.0235; 0.0419]
	Test	0.0325	[0.0183; 0.0466]
Régression logistique (intercept=0)	Apprentissage	0.0321	[0.0284; 0.0360]
	Test	0.0426	[0.0337; 0.0515]
Régression logistique quadratique (intercept=0)	Apprentissage	0.0322	[0.0203; 0.0420]
	Test	0.0356	[0.0215; 0.0497]
Régression logistique quadratique (intercept=1)	Apprentissage	0.0282	[0.0232; 0.0331]
	Test	0.0405	[0.0308; 0.0502]
Arbre de décision	Apprentissage	0.0393	[0.0351; 0.0436]
	Test	0.0726	[0.0654; 0.0798]

TABLE 1 – Calcul du taux d'erreur d'apprentissage ε et de test pour le jeu de données breastcancer avec plusieurs méthodes (arrondi à 10^{-4})

On remarque ici que dans le cas de la régression logistique, on ne peut pas réaliser l'analyse sur le jeu de données initial car dans certains cas nous obtenons ce résultat : $det(X^T W_q X) = 0$ donc il n'est pas possible d'inverser cette matrice. On croit que dans ce jeu de données, il y trop de paramètres à estimer, notamment dans le cas de régression logistique quadratique mais on n'a que 569 individus et après avoir séparé en ensemble d'apprentissage et ensemble de test, il n'y a que 379 individus pour l'apprentissage. Nous décidons donc d'utiliser l'ACP, sachant que plus de 90% de l'information se concentre sur les 7 premières composantes.

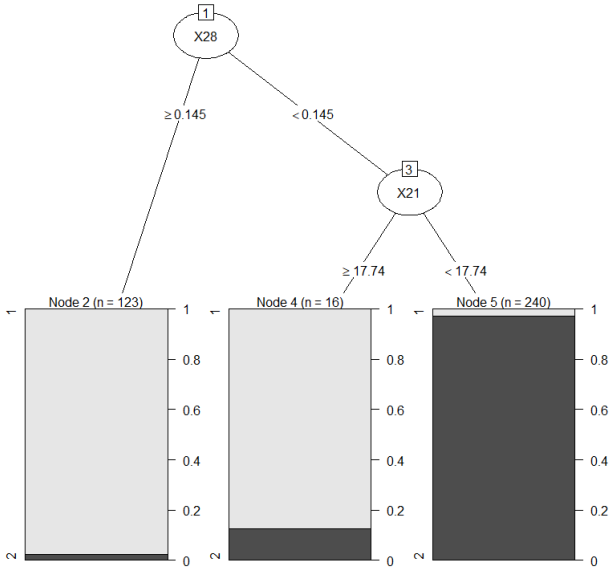

FIGURE 5 – L'arbre de décision **breastcancer**

D'après la table 1, la régression logistique permet d'obtenir de très bons résultats après le traitement de données par l'ACP. Comme nous choisissons les composantes qui conservent la plupart des informations, le nombre de paramètres à estimer dans la régression logistique est inférieur aux autres

méthodes. De plus, la régression logistique est plus robuste et flexible que les autres, notamment dans le cas où les individus ne sont pas distribués normalement. C'est pourquoi on peut obtenir des très bons résultats avec la régression logistique sans perdre beaucoup d'informations en utilisant l'ACP. Le résultat obtenu par l'arbre de décision est remarquable. L'arbre de décision est une méthode interprétable même dans le cas de grandes dimensions. Comme nous l'avons prévu, l'analyse discriminante linéaire peut nous donner un bon résultat. Le classifieur bayésien naïf et l'analyse discriminante quadratique déduisent des mauvais résultats, en particulier pour l'analyse discriminante quadratique. Le classifieur bayésien naïf est trop simple pour exprimer les données **breastcancer** tandis que l'analyse discriminante quadratique est trop compliquée. Nous avons essayé l'analyse discriminante quadratique avec des données traitées par l'ACP comme le cas de la régressions logistique et nous avons obtenu 0.0712 pour le taux d'erreur d'apprentissage et 0.0506 pour le taux d'erreur de test.

2.2 Ionosphere

2.2.1 Analyse exploratoire

Comme le jeu de données précédent, nous avons d'abord effectué une analyse exploratoire pour comprendre et s'approcher les données fournies.
Ce jeu de données comporte 34 variables et 350 individus. On a exclu la première colonne parce que cela correspond à l'index qui est inutile dans notre analyse. Nous pouvons voir que la variable *X2*

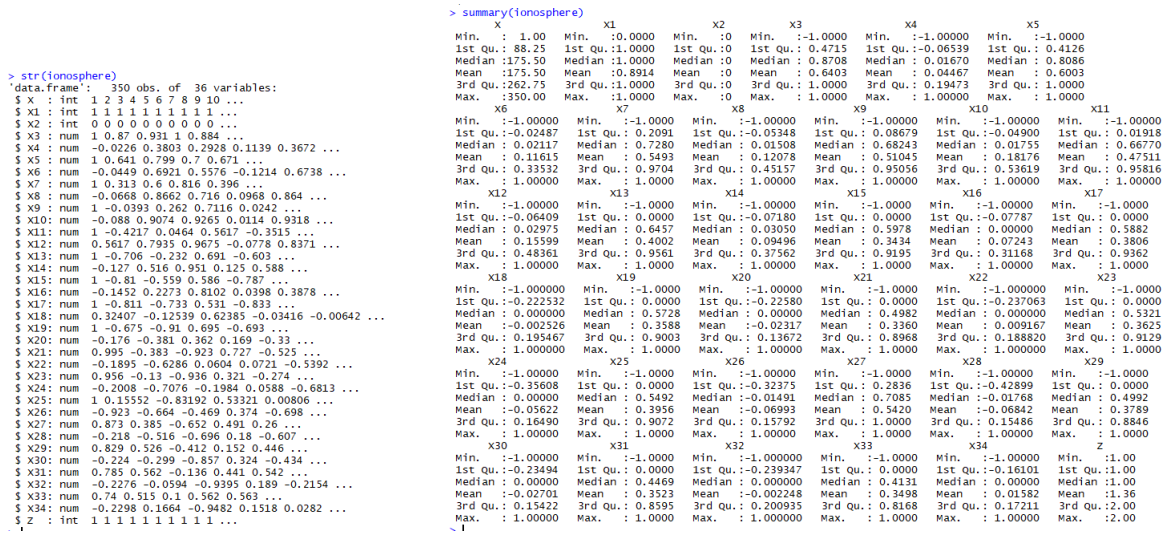

FIGURE 6 – La strucutre du jeu de données **Ionosphere**

est toujours nulle dans ce jeu de données donc nous décidons de la supprimer. A partir des résultats ci-dessus, on peut voir que toutes les variables sont des variables numériques. Toutes les variables sont dans l'intervale [-1,1]. De plus, dans le but d'étudier la corrélation entre les variables qui peut produire des influences dans les méthodes de discrimination.

Le test de corrélations a été réalisé par la méthode de Pearson avec **p-value** fixée par défaut à $\alpha = 0.05$. Grâce au graphique, on peut observer facilement qu'il y a des paires corrélées comme $(X13, X5)$, $(X9, X17)$ mais la corrélation n'est pas très significative. L'étude de la corrélation est poursuivie par la réalisation d'une analyse en composantes principales. Comme l'étude précédente, on veut dans un premier temps explorer la géométrie des données et à partir de cela, on peut prévoir le modèle compatible.
Nous avons obtenu d'abord 2 graphiques : le cercle de corrélation et l'inertie de valeurs propres des 10 premières valeurs en utilisant la fonction *princomp* et des fonctions de la biliothèque *factoextra*. Dans le cadre de notre analyse, nous pouvons voir que les informations se concentrent aux 20 premières composantes avec plus de 90% des informations totales. De plus, le cercle de corrélation des variables expliqué par les deux premières composantes nous montre la corrélation entre les variables. On peut

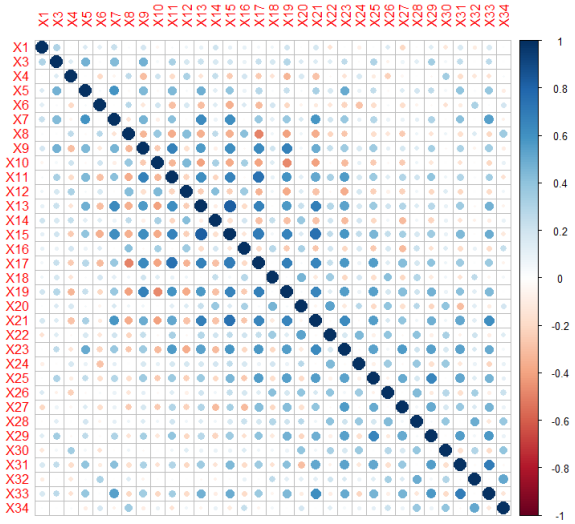

FIGURE 7 – La corrélation entre les variables de **ionosphere**

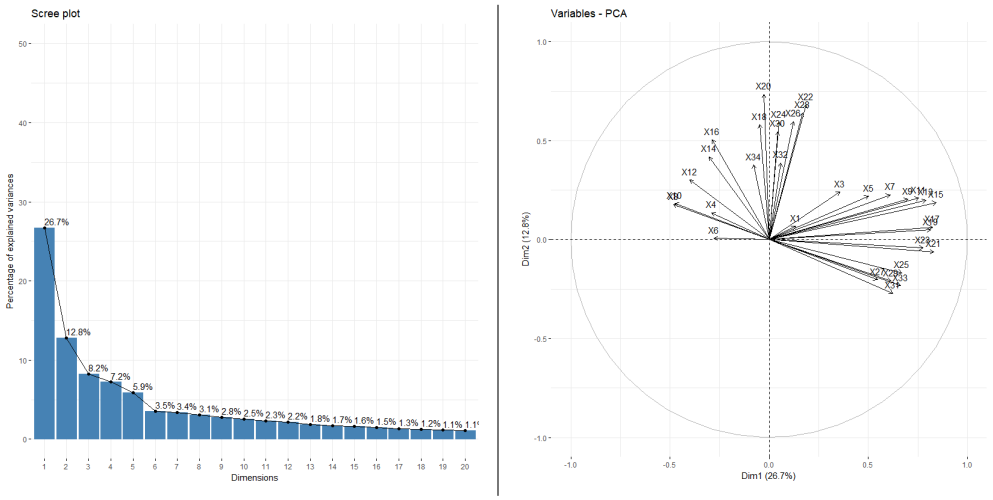

FIGURE 8 – Graphique a : L'inertie de valeurs propres des 10 premières valeurs - Graphique b : Le cercle de corrélation des variables **breastcancer**

voir dans le cercle de corrélation que les variables se séparent en 2 groupes principaux : un groupe a tendance au sens de la première composantes et un groupe a tendance au sens de la deuxmième composantes. On observe la distribution des individus dans le graphique des individus ci-dessous : La distribution des individus dans le graphique dans ce cas n'est plus très claire. Nous pouvons voir que dans la $2^{ème}$ et $3^{ème}$ composante principale, les individus sont plus séparés donc nous décidons de tracer le cercle de corrélation des variables dans ce plan. La distribution des individus est maintenant un peu plus visible : les individus de classe 1 sont distribuées généralement au centre tandis que les individus de classe 2 se situent dans la partie positive de la $3^{ème}$ composante principale. Les deux classes ne sont pas séparées assez clairement pour une frontière linéaire donc nous pouvons prévoir que l'analyse de discriminante quadratique peut nous donner une classification plus précise dans ce cas. Pour la régression logistique, on a plus des variables mais moins d'individus donc il est difficile que la régression logistique nous donne un résultat intéressant avec les données non pré-traitées. Pour vérifier notre prédiction, nous viendrons la réalisation des méthodes de discrimination sur ce jeu de données.

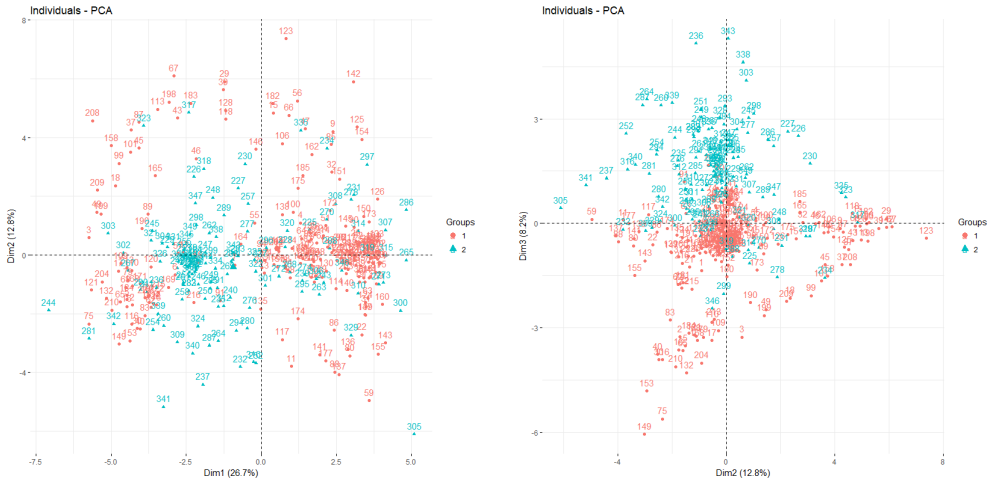

FIGURE 9 – La distribution des individus, graphique a : dans 2 premières composantes - graphique b : dans la 2eme et 3ème composante de **ionosphere**

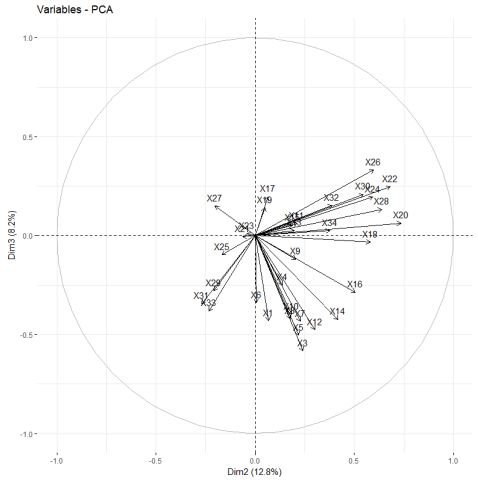

FIGURE 10 – Le cercle de corrélation des variables dans la 2eme et 3ème composante de **ionosphere**

2.2.2 Réalisation des méthode de discrimination

Pour calculer le taux d'erreur ponctuelle $\hat{\epsilon}$ d'une prédiction, il faut simplement compter le nombre de prédiction erronées ϵ_i entre la prédiction faite et les données réelle (d'apprentissage ou de test), puis diviser par le nombre total d'individus n.
On répète cette procédure 100 fois pour obtenir 100 réalisations du taux d'erreur. Pour obtenir l'estimation du taux d'erreur $\bar{\epsilon}$, il ne reste plus qu'à faire la moyenne des 100 réalisations.
L'intervalle de confiance à 95% est obtenu comme-ci : $IC = \left[\mu - 1.96 * \frac{s^*}{\sqrt{N}}; \mu + 1.96 * \frac{s^*}{\sqrt{N}}\right]$ avec :

— μ : La moyenne des 100 réalisation de ϵ

— s^* : L'écart-type corrigé des 100 réalisations de ϵ

— N : Le nombre de réalisation et dans ce cas $N = 100$

Nous avons appliqué des différents modèles d'analyses discriminantes, de régression logistique ainsi que les arbres de décisions. Le résume des résultats sont en table ci-dessous.

Méthode	**Données**	$\overline{\varepsilon}$	IC
Analyse discriminante quadratique	Apprentissage	0.0560	[0.0542; 0.0578]
	Test	0.0285	[0.0258; 0.0312]
Analyse discriminante linéaire	Apprentissage	0.1089	[0.1618; 0.1113]
	Test	0.0891	[0.0850; 0.0933]
Classifieur bayésien naïf	Apprentissage	0.1661	[0.1618; 0.1704]
	Test	0.1578	[0.1517; 0.1638]
Régression logistique (intercept=1)	Apprentissage	0.0578	[0.0549; 0.0606]
	Test	0.1235	[0.1179; 0.1290]
Régression logistique (intercept=0)	Apprentissage	0.0549	[0.0479; 0.0618]
	Test	0.1264	[0.1095; 0.1433]
Arbre de décision	Apprentissage	0.0686	[0.0609; 0.0763]
	Test	0.1239	[0.1064; 0.1414]

TABLE 2 – Calcul du taux d'erreur d'apprentissage ε et de test pour le jeu de données ionosphere avec plusieurs méthodes (arrondi à 10^{-4})

On remarque ici que dans le cas de la régression logistique, on ne peut pas réaliser l'analyse sur le jeu de données initial comme nous avoins prévu car dans certains cas nous obtenons ce résultat : $det(X^T W_q X) = 0$ donc il n'est pas possible d'inverser cette matrice. On croit que dans ce jeu de données, il y trop de paramètres à estimer, notamment dans le cas de régression logistique quadratique mais on n'a que 350 individus. Nous décidons donc d'utiliser l'ACP avec le fait que l'informations se concentrent aux 25 premières composantes avec plus de 90%. La régression logistique fonctionne bien après ce traitement mais la régression logistique quadratique nous donne toujours des erreurs. C'est raisonnable parce qu'avec 25 variables à initial, après la transformation la dimension de notre données est de (350,350). C'est-à-dire qu'on a moins de 350 individus pour l'apprentissage et il y a trop de paramètres. Nous continuons à diminuer le nombre de composantes jusqu'à 6 composantes. Mais avec 6 composantes principales, nous ne conservons pas beaucoup d'informations (50%) donc la régression logistique quadratique n'est pas pratique dans ce cas.

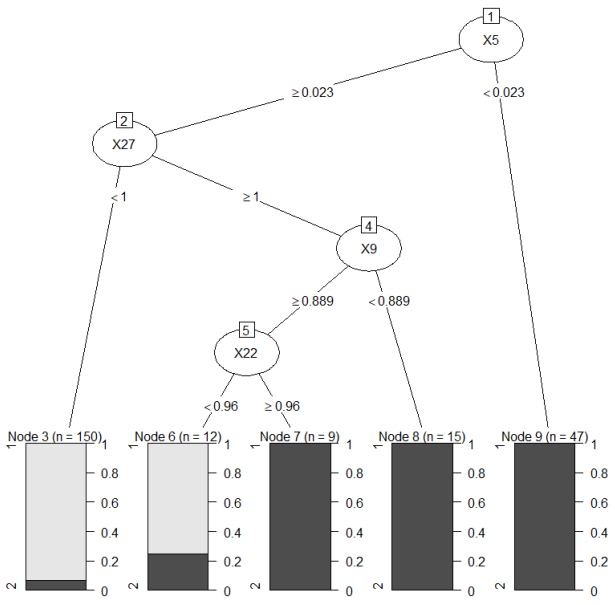

FIGURE 11 – L'arbre de décision **breastcancer**

D'après la table 2, l'analyse discriminante quadratique permet d'obtenir les très bonnes même si avec donnéees initiales. Comme ce que nous avons analysé dans la partie de l'ACP, les individus ne sont

pas séparés très clairement (il y a quelques individus d'un groupe se situent dans un autre groupe) donc nous avons besoin d'un modèle assez compliqué pour bien classifier les individus. Le classifieur bayésien naïf nous donne des taux d'erreurs assez forts car ce modèle est trop simple pour expimer ce jeu de données. L'analyse discriminante linéaire peut nous donner un résultat acceptable car ce modèle est plus robuste que le classifieur bayésien naïf. Les résultats obtenus par la régression logistique et l'arbre de décision sont plus grands que l'analyse discriminante linéaire mais ils sont acceptables. On remarque ici que *overfitting* apparait dans le cas de la régresssion logistique et l'arbre de décision parce qu'on a peu de données dans ce jeu de données. La régression logistique est robuste mais il faut avoir plusieurs de données pour stabiliser ce modèle. On note également ici que l'ajout de l'ordonnée à l'origine n'est pas efficace car il n'y a pas beacoup de valeur nulle dans ce jeu de données (9%) et la valeur maximale de chaque variable est 1. Plus généralement, les classifieurs se basant sur la distribution gaussienne ont de très bons résultats sur ces données car cela correspond effectivement à la vraie distribution des données.

2.3 Sonar

2.3.1 Analyse exploratoire

Comme le jeu de données précédent, nous avons d'abord effectué une analyse exploratoire pour comprendre et s'approcher les données fournies.
Ce jeu de données comporte 60 variables et 208 individus. On a exclu la première colonne parce que cela correspond à l'index qui est inutile dans notre analyse. A partir des résultats ci-dessus, on peut

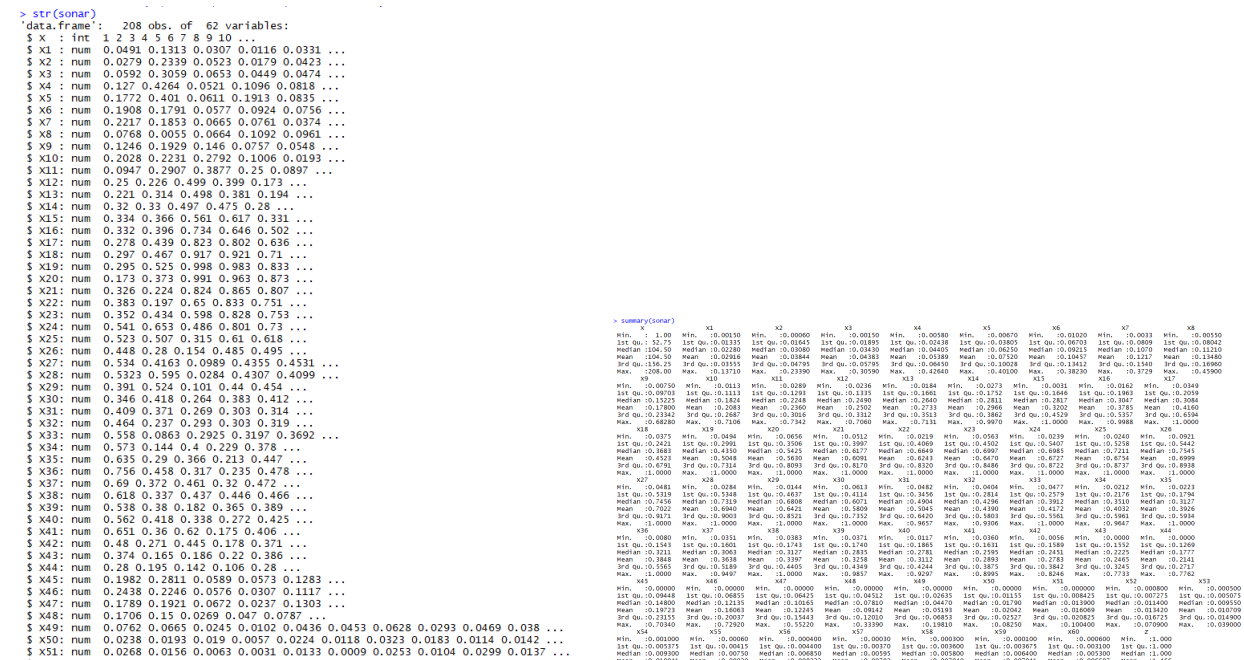

FIGURE 12 – La strucutre du jeu de données **Sonar**

voir que toutes les variables sont des variables numériques. Toutes les variables sont dans l'intervale [0,1]. Nous pouvons voir d'abord que ce jeu de données comporte très peu des individus et beaucoup de variables. Ce sera difficile à appliquer les méthodes qui ont besoin de plusieurs données pour stabiliser le modèle comme la régression ou les méthodes qui doivent estimer beaucoup de paramètres comme l'analyse discriminante quadratique. On croit que **overfitting** apparaîtrait dans plupart des analyses. On puis étudie la corrélation des variables.

Le test de corrélations a été réalisé par la méthode de Pearson avec **p-value** a été fixée par défaut pour $\alpha = 0.05$. Grâce au graphique, on peut observer la corrélation entre des variables n'est pas significative sauf les points autour de la diagonale principale. C'est-à-dire que les variables sont corrélé

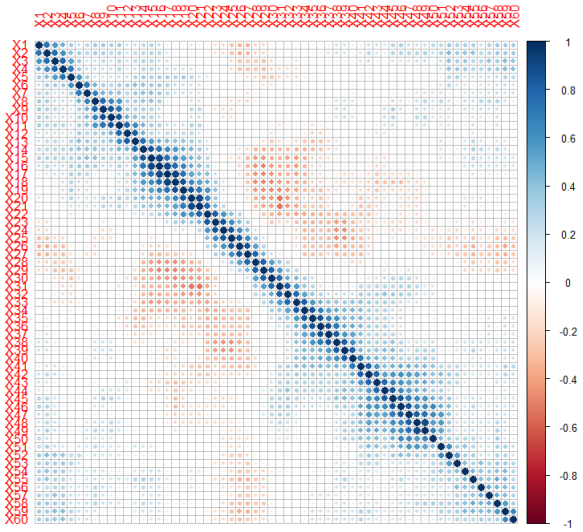

FIGURE 13 – La corrélation entre les variables de **Sonar**

avec les variables "voisines". L'étude de la corrélation est poursuivie par la réalisation d'analyse en composantes principales. Comme l'étude précédent, on veut au premier temps explorer la géométrie des données et à partir de cela, on peut prévoir le modèle compatible.
Nous avons obtenu d'abord 2 graphiques : le cercle de corrélation et l'inertie de valeurs propres de 10 premières valeurs en utilisant la fonction *princomp* et des fonctions du biliothèque *factoextra*. Dans

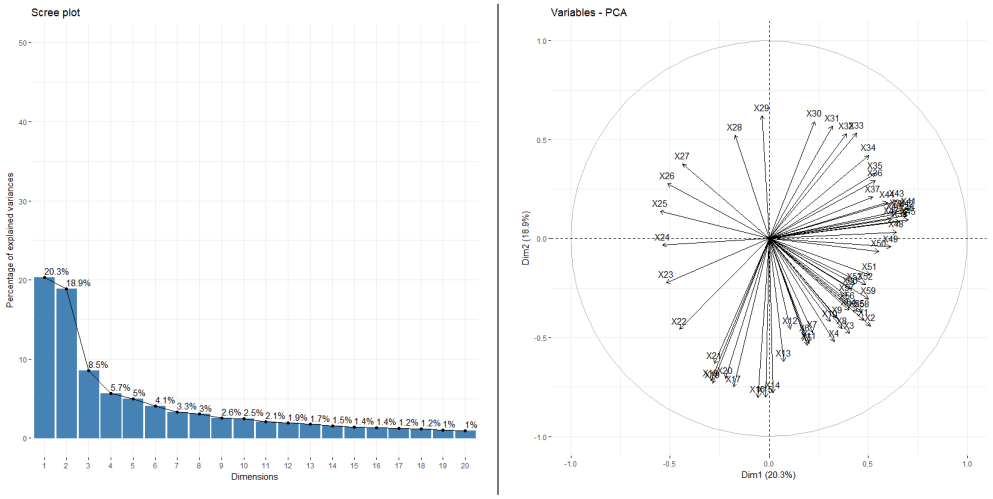

FIGURE 14 – Graphique a : L'inertie de valeurs propres de 10 premières valeurs - Graphique b : Le cercle de corrélation des variables **Sonar**

le cadre de notre analyse, nous pouvons voir que les informations se concentrent aux 25 premières composantes avec plus de 92% des informations totales. De plus, le cercle de corrélation des variables expliqué par les deux premières composantes nous montre la corrélation entre les variables. On observe la distribution des individus dans le graphique des individus ci-dessous.
La distribution des individus dans le graphique dans ce cas n'est plus très claire. Nous pouvons voir que dans la $1^{ère}$ et $3^{ème}$ composante principale, les individus sont plus séparés donc nous décidons de tracer le biplot dans ce plan.
La relation entre les variables et les individus est maintenant un peu plus visible : les individus de classe 1 ont tendance à êtr distribuée généralement selon les variables tandis que les individus de classe 2 ne le sont pas. Les deux classes ne sont pas séparées assez clairement pour une frontière linéaire donc

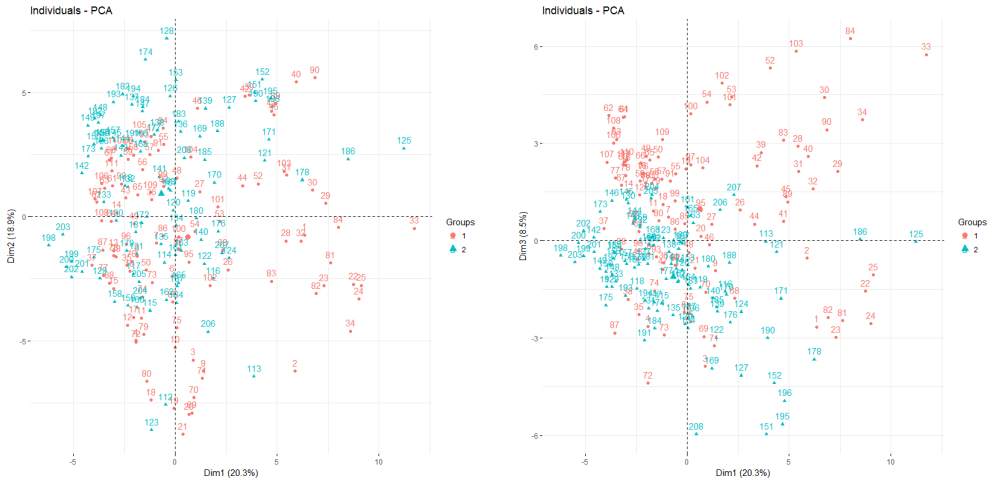

FIGURE 15 – La distribution des individus, graphique a : dans 2 premières composantes - graphique b : dans la 1me et 3ème composante **Sonar**

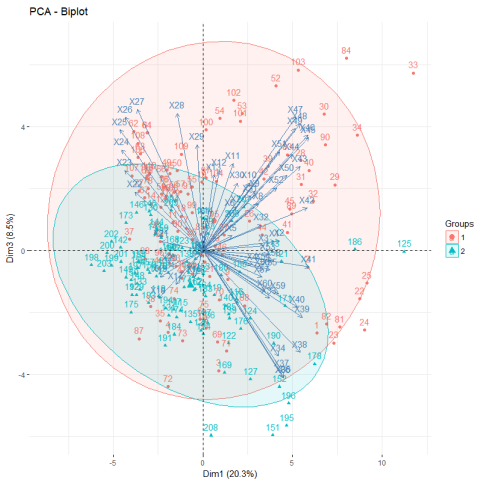

FIGURE 16 – Biplot dans la 1me et 3ème composante **Sonar**

nous pouvons prévoir que l'analyse de discriminante quadratique peut nous donner une classification plus précise mais dans ce cas pour obtenir les résultats sans *overfitting*, il faut pré-traiter ce jeu de données. Avec ce jeu de données peu d'individus comme **Sonar**, le classifieur bayésien naïf peut se fonctionner bien mais dans ce cas, la distribution des individus est compliquée pour que un modèle simple comme le classifieur bayésien naïf puisse nous donner un résultat agréable. Pour vérifier notre prédiction, nous viendrons la réalisation des méthodes de discrimination sur ce jeu de données.

2.3.2 Réalisation des méthode de discrimination

Pour calculer le taux d'erreur ponctuelle $\hat{\epsilon}$ d'une prédiction, il faut simplement compter le nombre de prédiction erronées ϵ_i entre la prédiction faite et les données réelle (d'apprentissage ou de test), puis diviser par le nombre total d'individus n.
On répète cette procédure 100 fois pour obtenir 100 réalisations du taux d'erreur. Pour obtenir l'estimation du taux d'erreur $\bar{\epsilon}$, il ne reste plus qu'à faire la moyenne des 100 réalisations.
L'intervalle de confiance à 95% est obtenu comme-ci : $IC = \left[\mu - 1.96 * \frac{s^*}{\sqrt{N}}; \mu + 1.96 * \frac{s^*}{\sqrt{N}}\right]$ avec :

— μ : La moyenne des 100 réalisation de ϵ

— s^* : L'écart-type corrigé des 100 réalisations de ϵ

— N : Le nombre de réalisation et dans ce cas $N = 100$

Comme nous avons discuté dessus, il est difficile à analyser ce jeu de données sans *overfitting* donc nous d'abord pré-traitons ce jeu de données par la méthode l'ACP comme d'habitude. Nous choisissons 25 premières composantes principales qui nous permettent de conserver plus de 92% de l'informations. Nous avons puis appliqué des différents modèles d'analyses discriminantes, de régression logistique ainsi que les arbres de décisions. Le résume des résultats sont en table ci-dessous.

Méthode	**Données**	$\overline{\varepsilon}$	IC
Analyse discriminante quadratique	Apprentissage	0.0470	[0.0434; 0.0506]
	Test	0.2752	[0.2663; 0.2840]
Analyse discriminante linéaire	Apprentissage	0.1476	[0.1430; 0.1522]
	Test	0.2560	[0.2470; 0.2651]
Classifieur bayésien naïf	Apprentissage	0.1976	[0.1897; 0.2056]
	Test	0.3136	[0.3035; 0.3236]
Régression logistique (intercept=1)	Apprentissage	0.1223	[0.1164; 0.1281]
	Test	0.2444	[0.2361; 0.2527]
Régression logistique (intercept=0)	Apprentissage	0.1349	[0.1297; 0.1401]
	Test	0.2486	[0.2386; 0.2587]
Arbre de décision	Apprentissage	0.1136	[0.0978; 0.1295]
	Test	0.2594	[0.2342; 0.2846]

TABLE 3 – Calcul du taux d'erreur d'apprentissage ε et de test pour le jeu de données sonar avec plusieurs méthodes (arrondi à 10^{-4})

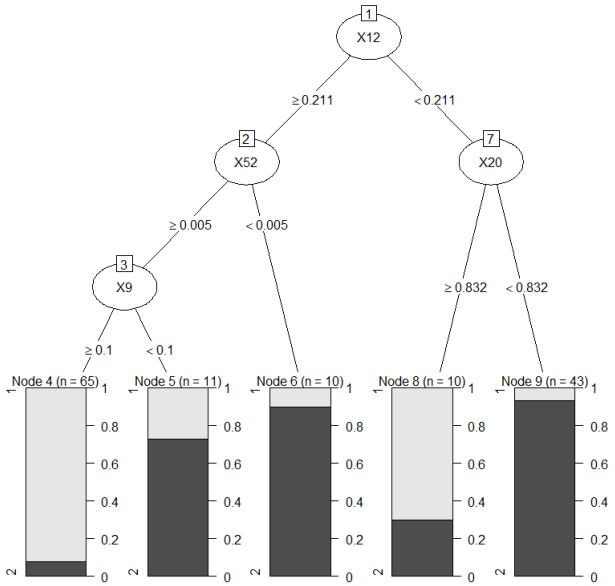

FIGURE 17 – L'arbre de décision **Sonar**

On remarque ici que dans le cas de la régression logistique quadratique, on ne peut pas réaliser l'analyse sur le jeu de données car dans certains cas nous obtenons ce résultat : $det(X^T W_q X) = 0$ donc il n'est pas possible d'inverser cette matrice. C'est reasonable parce que avec 25 variables à initial, après la transformation la dimension de notre données est de (207,350). C'est-à-dire on a moins que 350 individus pour l'apprentissage et c'est évidemment impossible. Nous continuons à diminuer le nombre de composantes jusqu'à 4 composantes. Mais avec 6 composantes principales, nous ne conservons pas beaucoup d'informations (53%) donc la régression logistique quadratique n'est pas pratique dans ce cas.

D'après la table 3, la régression logistique nous donne les meilleurs résultats. Comme ce que nous avons analysé dans la partie de l'ACP, les individus ne sont pas séparés très clairement (il y a beaucoup individus d'un groupe se situent dans un autre groupe) donc nous avons besoin d'un modèle assez flexible pour bien classifier les individus. Le classifieur bayésien naïf nous donne les taux d'erreurs assez forts car ce modèle est trop simple pour expimer ce jeu de données. L'analyse discriminante linéaire peut nous donner un resultat acceptable car ce modèle est plus robuste que le classifieur bayésien naïf. On remarque ici que *overfitting* apparait dans plupart de cas comme nous avons prévu malgé au pré-traitement de données, parce qu'on a peu de données dans ce jeu de données. L'analyse discriminante quadratique est le cas de **overfitting** le plus visble cas le taux d'erreur d'apprentissage est 0.047 mais le taux d'erreur de test est beaucoup plus significatif. L'analyse discriminante quadratique est flexible donc dans certains cas, il peut "apprendre" facilement des bruits.La régression logistique est robuste mais il faut avoir plusieurs de données pour stabiliser ce modèle. Donc le pré-traitement de données par l'ACP ne peut pas résoudre completement le problème de *overfitting*. On note également ici que l'ajout de l'ordonnée à l'origine augmente un peu l'efficace car la précision de la régression logistique dans le cas l'ajout de l'ordonnée à l'origine est un peu plus haute. Mais l'efficacement n'est pas très remarquable. L'arbre de décision n'améliore pas le résultat après le pré-traitement des données cas l'arbre de décision est une méthode non paramétrique donc elle est très sensible. Plus généralement, dans ce jeu de données, nous avons rencontré le problème *overfitting*. Le problème est la conséquence de le fait que on a peu de données pour l'apprentissage. Nous avons essayé d'appliquer l'ACP pour résoudre ce problème. Le *overfitting* diminue après le pré-traitement de données mais les taux d'erreur restent très significatifs.

2.4 Spambase

2.4.1 Analyse exploratoire

Comme le jeu de données précédent, nous avons d'abord effectué une analyse exploratoire pour comprendre et s'approcher les données fournies.
Ce jeu de données comporte 57 variables et 4601 individus. On a exclu la première colonne parce que cela correspond à l'index qui est inutile dans notre analyse. A partir des résultats ci-dessus, on peut

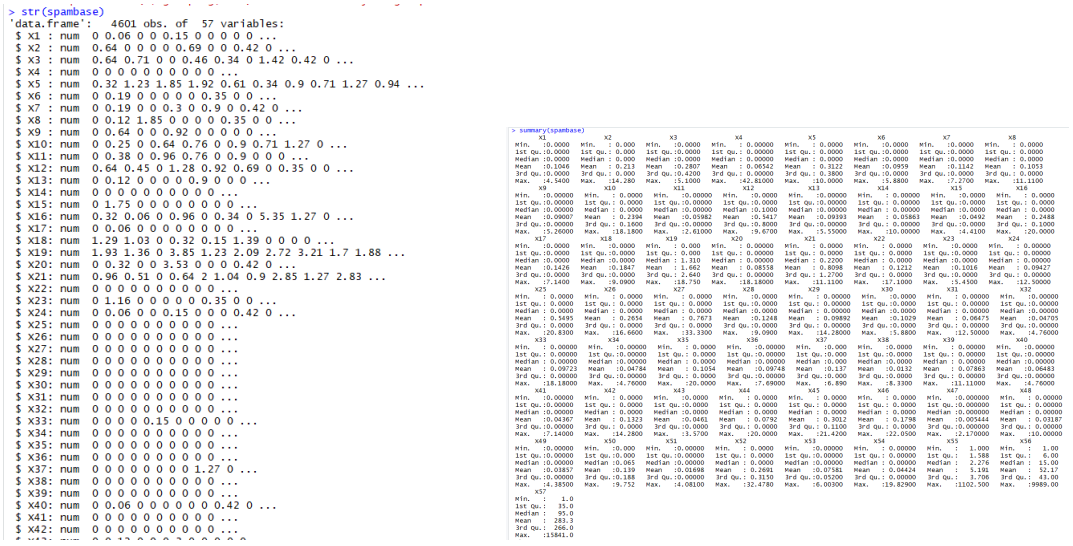

FIGURE 18 – La strucutre du jeu de données **spambase**

voir que toutes les variables sont des variables numériques. On remarque ici qu'il y a quelques variables ayant les valeurs très dominantes que les autres par exemple $X49$, $X55$, $X56$, $X57$. Nous donc voulons savoir si la covariance entre les variables est grande ou non. Après avoir étudié la covariance, on trouve qu'il y a 9 paires de variables dont les covariances sont suppérieures à 1000, 4 paires de variables dont les covariances sont suppérieur à 10000 et une paire dont la covariance est suppérieur à 100000. On

puis étudie la corrélation des variables.
Le test de corrélations a été réalisé par la méthode de Pearson avec **p-value** a été fixée par défaut

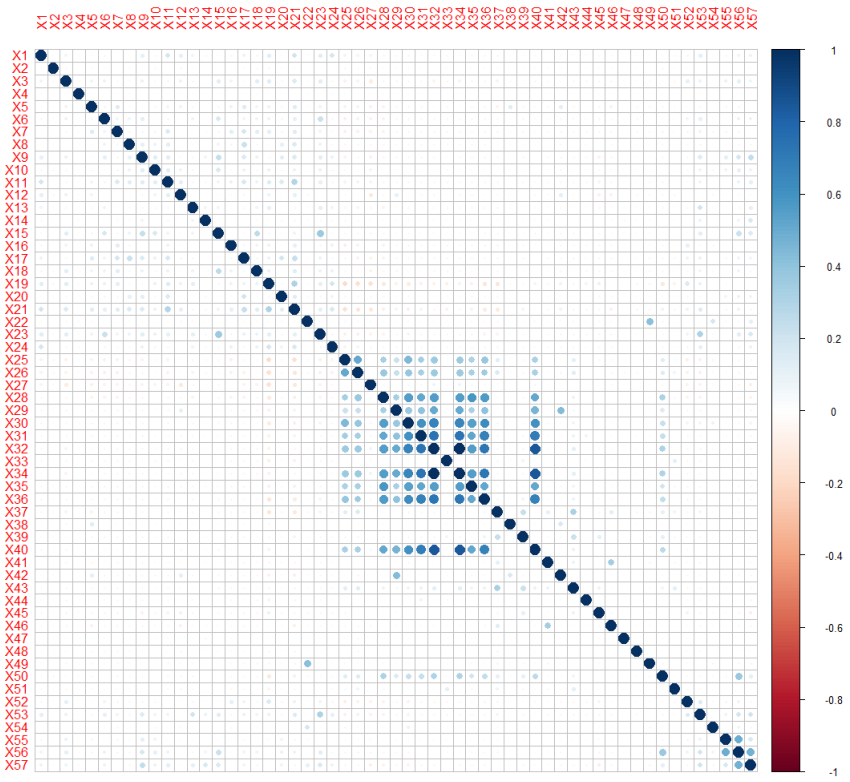

FIGURE 19 – La corrélation entre les variables de **spambase**

pour $\alpha = 0.05$. Grâce au graphique, on peut observer la corrélation entre des variables n'est pas significative sauf $(X32, X34)$, $(X32, X40)$. C'est-à-dire que la plupart de variables sont indépendantes l'une à l'autre. Ce sera peut-être problèmatique pour notre analyse parce que dans nos analyses discriminantes quadratique et linéaire et le classifieur bayésien naïf, nous avons supposé que les vecteurs de caractéristique suit une loi normale multidimensionnelle.

$$f_k(x) = \frac{1}{(2\pi)^{p/2}(det\Sigma_k)^{1/2}} exp\left(-\frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k)\right) \tag{1}$$

A quelques points où la valeur est grande, s'ils sont indépendants l'un à l'autre et ses covariances sont grandes, le terme $-\frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k)$ peut devenir très petite, donc $exp\left(-\frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k)\right)$ sera approximatement égal à 0 et donc f_k sera égale à 0. Si f_k est égal à 0, la densité de mélange devient 0 et la probabilité a posteriori devient indéfinie car on ne peut pas diviser 0 par 0. Avec l'analyse ci-dessus, on croit qu'il y aurais quelque problèmes avec notre analyse discriminante.
L'étude de la corrélation est poursuivie par la réalisation d'analyse en composantes principales. Comme l'étude précédent, on veut au premier temps explorer la géométrie des données et à partir de cela, on peut prévoir le modèle compatible.
Nous avons obtenu d'abord 2 graphiques : le cercle de corrélation et l'inertie de valeurs propres de 10 premières valeurs en utilisant la fonction *princomp* et des fonctions du biliothèque *factoextra*. Dans le cadre de notre analyse, nous pouvons voir que les informations se distribuent uniforement et 95% de données se concentrent aux 45 premères composantes principales. C'est un autre problème de notre analyse car si on veut effectuer une diminuation de dimension par l'ACP, on doit choisir beaucoup de

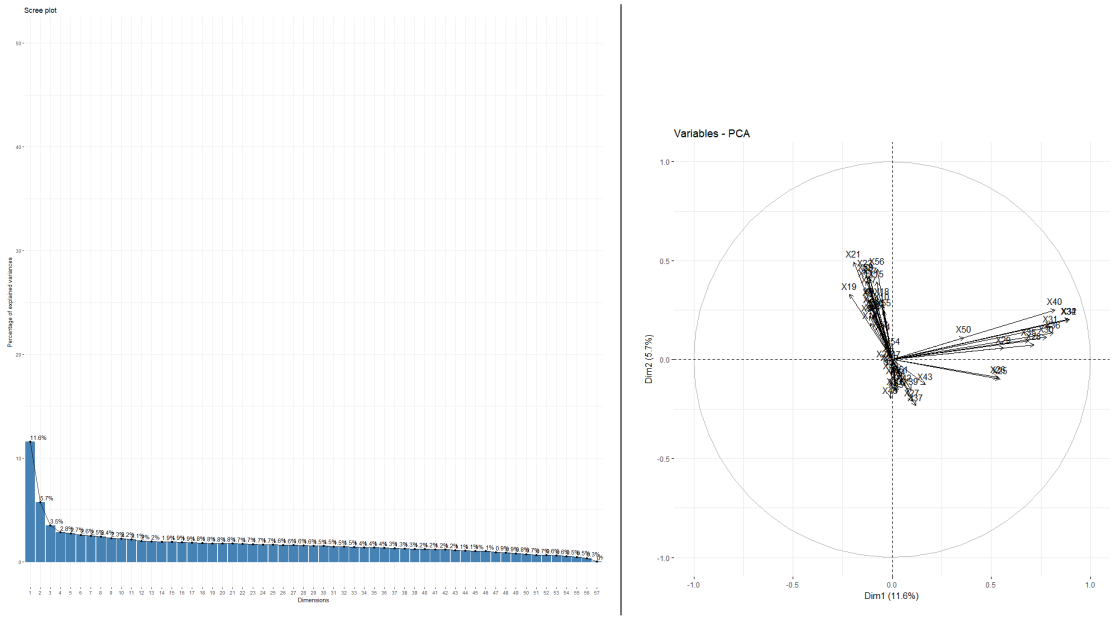

FIGURE 20 – Graphique a : L'inertie de valeurs propres de 10 premières valeurs - Graphique b : Le cercle de corrélation des variables **spambase**

composantes pour conserver plupart de l'informations. Le cercle de corrélation des variables expliqué par les deux premières composantes nous montre la corrélation entre les variables. On observe la distribution des individus dans le graphique des individus ci-dessous.

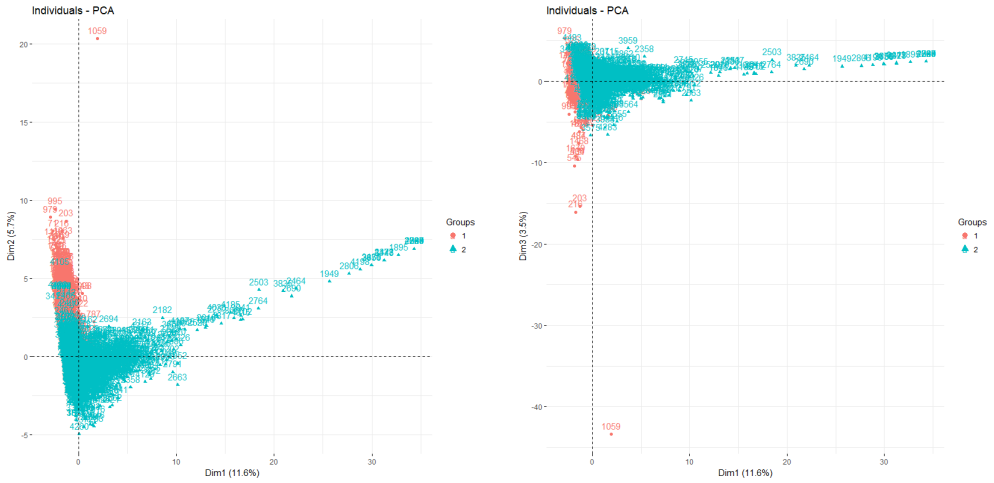

FIGURE 21 – La distribution des individus, graphique a : dans 2 premières composantes - graphique b : dans la 1me et 3ème composante **spambase**

La distribution des individus dans le graphique dans ce cas n'est pas claire. Nous avons visualisé des données dans les autres plans mais nous n'avons obtenu aucun meilleur résultat. Les deux classes ne sont pas séparées assez clairement pour une frontière linéaire donc nous pouvons prévoir que l'analyse de discriminante quadratique peut nous donner une classification plus précise. La régression logistique pourrais nous fournir une classification assez bonne car elle est assez flexible pour un jeu de données comme Spambase. Dans ce cas, la distribution des individus est compliquée pour que un modèle simple comme le classifieur bayésien naïf puisse nous donner un résultat agréable. L'arbre de décision peut performer assez bien dans ce cas parce que cette méthode est interprétable dans une grande dimension. Pour vérifier notre prédiction, nous viendrons la réalisation des méthodes de discrimination sur ce jeu de données.

2.4.2 Réalisation des méthode de discrimination

Pour calculer le taux d'erreur ponctuelle $\hat{\epsilon}$ d'une prédiction, il faut simplement compter le nombre de prédiction erronées ϵ_i entre la prédiction faite et les données réelle (d'apprentissage ou de test), puis diviser par le nombre total d'individus n.
On répète cette procédure 100 fois pour obtenir 100 réalisations du taux d'erreur. Pour obtenir l'estimation du taux d'erreur $\bar{\epsilon}$, il ne reste plus qu'à faire la moyenne des 100 réalisations.
L'intervalle de confiance à 95% est obtenu comme-ci : $IC = \left[\mu - 1.96 * \frac{s^*}{\sqrt{N}}; \mu + 1.96 * \frac{s^*}{\sqrt{N}}\right]$ avec :

— μ : La moyenne des 100 réalisation de ϵ

— s^* : L'écart-type corrigé des 100 réalisations de ϵ

— N : Le nombre de réalisation et dans ce cas $N = 100$

Nous avons d'abord essayé de réaliser les analyses avec le jeu de données initial. Les analyses discriminantes quadratique et linéaire et le classifieur bayésien naïf ne se fonctionnent pas et nous donne toujours des mêmes erreurs. Notre fonction de la régression logistique est bloquée car la dimension de ce jeu de données est très grande. Nous avons essayé des données pré-traitées mais rien ne marche. L'abre de décision est un seul modèle qui se fonctionne bien et donne un bon résultat car comme nous avons indiqué dessus, l'arbre de décision est interprétable même si dans la grande dimension. Nous ensuite cherchons à résoudre les problèmes avec l'analyse discriminante quadratique et linéaire, le classifieur bayésien naïf et avec la régression logistique. D'abord pour la régression logistique, on croit que le problème est la grande dimension du jeu de données et notre fonction doit réaliser trop de calculs, donc la vitesse est ralentie. C'est pourquoi elle ne peut pas retourner le résultat. Nous décidons d'utiliser une fonction *glm* du bibliothèque **ISLR** pour augmenter la vitesse et performance de notre analyse. Pour le problème de l'analyse discriminante quadratique et linéaire et le classifieur bayésien naïf, nous avons d'abord cherché la raison qui produit des erreurs. Nous avons exploré le résultat retourné par la fonction **ad.val** que nous avons implémenté dans TD et nous avons trouvé quelques valeurs *NaN* dans l'ensemble de probabilité et dans l'ensemble de classe prédite. Nous pouvons conclure ici que la risque $f_k = 0$ que nous avons prédite dans la partie d'analyse exploratoire se réalise. Le nombre de valeurs *NaN* apparaissant dans le resultat rétourné par notre fonction dépend de la séparation de données en ensemble d'appentissage et de test mais il n'est jamais trop grand, maximum 5 valeurs *NaN* sur 3068 individus dans l'ensemble d'apprentissage et sur 1533 individus dans l'ensemble de test. Donc on peut négligér ces valeurs *NaN* car son effet n'est pas significatif. Nous avons modifié un peu notre code et toutes les analyses maintenant se fonctionne bien. De plus, comme nous avons discuté dans la partie d'analyse exploratoire, il y a quelques variables dont les valeurs sont beaucoup plus dominantes que les autres. Nous avons donc réalisé un pré-traitement sur ce jeu de données pour obtenir les résultats plus agréable que possible : nous dévisons chaque variable par sa valeur maximale et notre jeu de données devient des valeurs comprises entre 0 et 1. Nous réalisons l'ACP sur le jeu de données pré-traité et nous avons le graphique de individus ci-dessous. Nous pouvons voir que le jeu de données après le pré-traitement est plus visible, les individus sont plus séparés et se distribuent selon 2 groupes de variables.

Nous avons puis appliqué des différents modèles d'analyses discriminantes, de régression logistique ainsi que les arbres de décisions. Le résume des résultats sont en table ci-dessous.

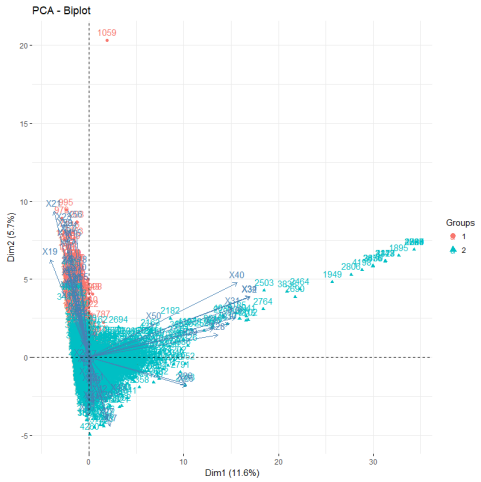

FIGURE 22 – La distribution des individus de **Spambase** après pré-traitement dans le premier plan

Méthode	**Données**	$\overline{\varepsilon}$	IC
Analyse discriminante quadratique	Apprentissage	0.1588	[0.1559; 0.1617]
	Test	0.1236	[0.1106; 0.1366]
Analyse discriminante linéaire	Apprentissage	0.1439	[0.1423; 0.1455]
	Test	0.1412	[0.1384; 0.1441]
Classifieur bayésien naïf	Apprentissage	0.1305	[0.1272; 0.1338]
	Test	0.1311	[0.1280; 0.1342]
Régression logistique	Apprentissage	0.0673	[0.0655; 0.0691]
	Test	0.0743	[0.0719; 0.0768]
Arbre de décision	Apprentissage	0.0972	[0.0958; 0.0987]
	Test	0.1071	[0.1038; 0.1104]

TABLE 4 – Calcul du taux d'erreur d'apprentissage ε et de test pour le jeu de données spambase avec plusieurs méthodes (arrondi à 10^{-4})

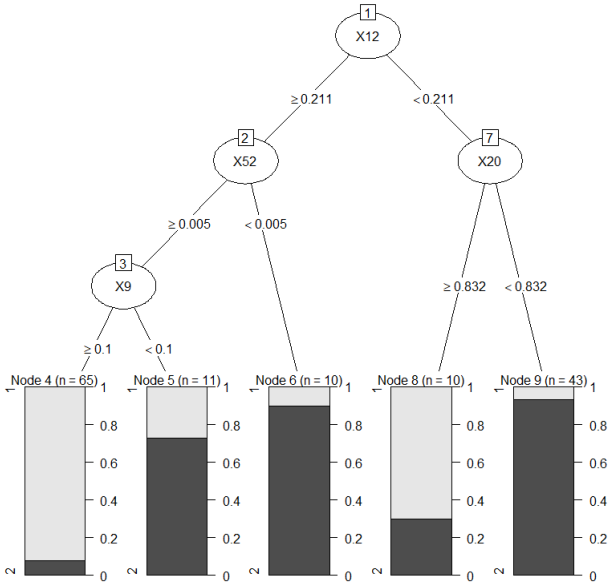

FIGURE 23 – L'arbre de décision **Sonar**

On remarque ici que dans le cas de la régression logistique quadratique, on ne peut pas réaliser l'ana-

lyse sur le jeu de données car dans certains cas nous obtenons ce résultat : $det(X^T W_q X) = 0$ donc il n'est pas possible d'inverser cette matrice.
D'après la table 4, la régression logistique nous donne les meilleurs résultats. Comme ce que nous avons analysé dans la partie de l'ACP, les individus ne sont pas séparés très clairement (il y a beaucoup individus d'un groupe se situent dans un autre groupe) donc nous avons besoin d'un modèle assez flexible pour bien classifier les individus. Le classifieur bayésien naïf nous donne les taux d'erreurs assez forts car ce modèle est trop simple pour expimer ce jeu de données. L'analyse discriminante linéaire nous donne des taux assez grands et un peu plus que le classifieur bayésien naïf. On remarque ici que le jeu de données Spambase est complexe mais le classifieur bayésien naïf nous donne le meilleur resultat que l'analyse discriminante linéaire. L'analyse discriminante quadratique est une méthode aussi flexible mais dans ce cas elle nous donne un résultat qui n'est pas agréable parce que l'analyse discriminante quadratique doit estimer beaucoup de paramétres (dans ce cas le nombre de paramètres est de 3421 !). L'arbre de décision nous donne toujours un bon résultat.

2.5 Spambase 2

2.5.1 Analyse exploratoire

Comme le jeu de données précédent, nous avons d'abord effectué une analyse exploratoire pour comprendre et s'approcher les données fournies.
Ce jeu de données comporte 57 variables et 4601 individus comme le jeu de données précédent. On a exclu la première colonne parce que cela correspond à l'index qui est inutile dans notre analyse. A

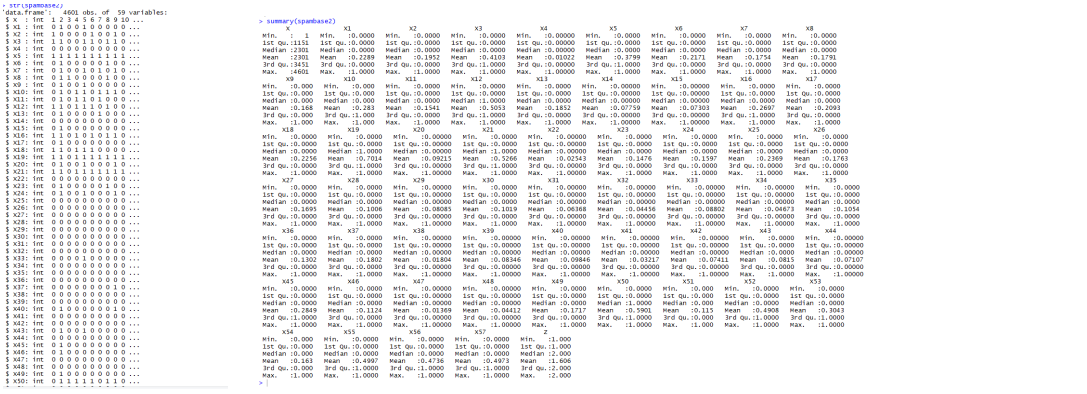

FIGURE 24 – La strucutre du jeu de données **spambase2**

partir des résultats ci-dessus, on peut voir que toutes les variables sont des variables binaire. On puis étudie la corrélation des variables.
Le test de corrélations a été réalisé par la méthode de Pearson avec **p-value** a été fixée par défaut pour $\alpha = 0.05$. Grâce au graphique, on peut observer la corrélation entre des variables n'est pas significative sauf le couple $(X34, X36)$. Comme l'étude précédent, on veut au premier temps explorer la géométrie des données et à partir de cela, on peut prévoir le modèle compatible.
Nous avons obtenu d'abord 2 graphiques : le cercle de corrélation et l'inertie de valeurs propres de 10 premières valeurs en utilisant la fonction *princomp* et des fonctions du biliothèque *factoextra*. Dans le cadre de notre analyse, nous pouvons voir que les informations se concentrent aux 45 premières composantes avec plus de 92% des informations totales. De plus, le cercle de corrélation des variables expliqué par les deux premières composantes nous montre la corrélation entre les variables. Nous pouvons vois que les variables groupent en 2 groupes principaux : un groupe selon la première composante et un autre selon la deuxième composante. On observe la distribution des individus dans le graphique des individus ci-dessous.
La distribution des individus dans le graphique dans ce cas n'est plus très claire. Nous pouvons voir que dans la $1^{ère}$ et $3^{ème}$ composante principale, les individus de groupe 1 ont tendance de se situer

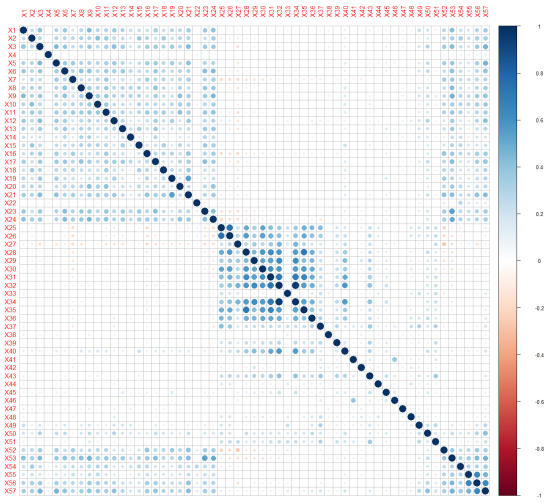

FIGURE 25 – La corrélation entre les variables de **spambase2**

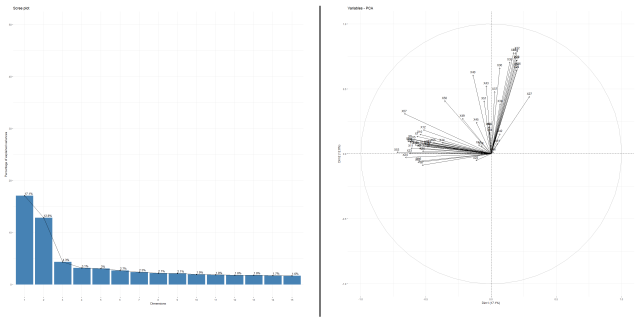

FIGURE 26 – Graphique a : L'inertie de valeurs propres de 10 premières valeurs - Graphique b : Le cercle de corrélation des variables **Sonar**

selon le groupe de variables qui est se distribuent selon la première composante tandis que les individus de groupe 2 ont tendance de se concentrer au centre. Avec ce jeu de données peu d'individus comme **Spambase 2**, le modèle flexible peut nous donner des resultats agréables comme l'analyse discriminante quadratique ou la régression logistique mais l'analyse discriminante quadratique nous donnerais un taux d'erreurs plus significatifs que la régression logistique car il y a beaucoup de paramètres à estimer. De plus, l'arbre de décision peut bien fonctionner dans un jeu de données à grande dimension comme **Spambase 2**. Pour vérifier notre prédiction, nous viendrons la réalisation des méthodes de discrimination sur ce jeu de données.

2.5.2 Réalisation des méthode de discrimination

Pour calculer le taux d'erreur ponctuelle $\hat{\epsilon}$ d'une prédiction, il faut simplement compter le nombre de prédiction erronées ϵ_i entre la prédiction faite et les données réelle (d'apprentissage ou de test), puis diviser par le nombre total d'individus n.
On répète cette procédure 100 fois pour obtenir 100 réalisations du taux d'erreur. Pour obtenir l'estimation du taux d'erreur $\bar{\epsilon}$, il ne reste plus qu'à faire la moyenne des 100 réalisations.
L'intervalle de confiance à 95% est obtenu comme-ci : $IC = \left[\mu - 1.96 * \frac{s^*}{\sqrt{N}}; \mu + 1.96 * \frac{s^*}{\sqrt{N}}\right]$ avec :

— μ : La moyenne des 100 réalisation de ϵ

— s^* : L'écart-type corrigé des 100 réalisations de ϵ

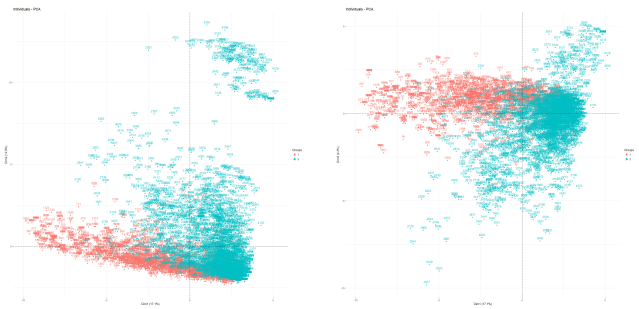

FIGURE 27 – La distribution des individus, graphique a : dans 2 premières composantes - graphique b : dans la 1me et 3ème composante **spambase2**

— N : Le nombre de réalisation et dans ce cas $N = 100$

Nous avons puis appliqué des différents modèles d'analyses discriminantes, de régression logistique ainsi que les arbres de décisions. Le résume des résultats sont en table ci-dessous.

Méthode	**Données**	$\overline{\varepsilon}$	IC
Analyse discriminante quadratique	Apprentissage	0.0680	[0.0674; 0.0686]
	Test	0.0762	[0.0750; 0.0773]
Analyse discriminante linéaire	Apprentissage	0.1057	[0.1050; 0.1063]
	Test	0.1094	[0.1079; 0.1108]
Classifieur bayésien naïf	Apprentissage	0.1695	[0.1680; 0.1709]
	Test	0.1669	[0.1629; 0.1709]
Régression logistique	Apprentissage	0.0559	[0.0548; 0.0570]
	Test	0.0621	[0.0597; 0.0644]
Arbre de décision	Apprentissage	0.1126	[0.1113; 0.1139]
	Test	0.1190	[0.1174; 0.1207]

TABLE 5 – Calcul du taux d'erreur d'apprentissage ε et de test pour le jeu de données spambase 2 avec plusieurs méthodes (arrondi à 10^{-4})

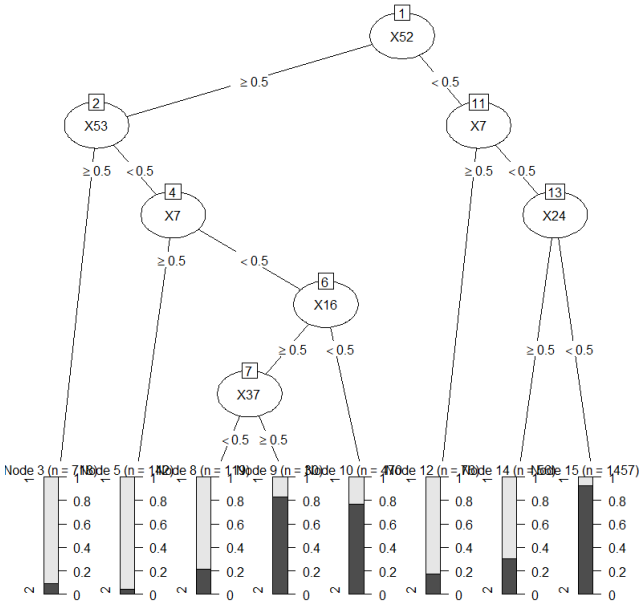

FIGURE 28 – L'arbre de décision **Spamebase 2**

On remarque ici que dans le cas de la régression logistique, on ne peut pas réaliser l'analyse sur le jeu de données par notre fonction le jeu de données comporte trop de variables et observations donc notre fonction doit réaliser beaucoup de calculs. C'est pourquoi la vitesse est ralentie et elle ne peut pas retourner le resultat. Donc nous avons utilisé la fonction *glm* du bibliothèque **ISLR**.
D'après la table 5, la régression logistique nous donne les meilleurs résultats. Comme ce que nous avons analysé dans la partie de l'ACP, les individus ne sont pas séparés très clairement (il y a beaucoup individus d'un groupe se situent dans un autre groupe) donc nous avons besoin d'un modèle assez flexible pour bien classifier les individus. L'analyse discriminante quadratique nous donne également un très bon resultat. L'analyse discriminante quadratique et la régression logistique sont assez complexe pour exprimer des données. Le classifieur bayésien naïf nous donne les taux d'erreurs assez forts car ce modèle est trop simple pour expimer ce jeu de données. L'analyse discriminante linéaire peut nous donner un resultat acceptable car ce modèle est plus robuste que le classifieur bayésien naïf mais il n'est pas assez flexible pour ce jeu de données. Malgré à la grande dimension, l'arbre de décision nous donne aussi un resultat stable.

3 Modèle

1. On a :

$$\mathbb{P}(X^j = 1|Z = \omega_k) = p_{kj}$$

D'où

$$\mathbb{P}(X^j = 0|Z = \omega_k) = 1 - p_{kj}$$

Cela correspond donc à une distribution de Bernouilli de paramètre p_{kj}. C'est à dire $X^j \sim_{w_k} B(p_{kj})$. Nous pouvons donc en déduire :

$$\mathbb{P}(X^j = x_j|Z = \omega_k) = p_{kj}^{x_j}(1 - p_{kj})^{1-x_j}$$

2. On suppose que les variables $X^1, ..., X^p$ sont indépendantes

$$\begin{aligned}
\mathbb{P}(\mathbf{X} = \mathbf{x}|Z = \omega_k) &= \mathbb{P}(X^1 = x_1, ..., X^p = x_p|Z = \omega_k) \\
&= \mathbb{P}(X^1 = x_1|Z = \omega_k, ..., X^p = x_p|Z = \omega_k) \\
&= \prod_{j=1}^{p} \mathbb{P}(X^j = x_j|Z = \omega_k) \\
&= \prod_{j=1}^{p} p_{kj}^{x_j}(1 - p_{kj})^{1-x_j}
\end{aligned}$$

3. En considérant que le i^{eme}exemple d'apprentissage consiste en un couple (x_i, z_i), écrire la probabilité jointe $\mathbb{P}(X = x_i, Z = z_i)$:

$$\begin{aligned}
\mathbb{P}(X = x_i, Z = z_i) &= \prod_{k=1}^{g} \mathbb{P}(X = x_i, Z = \omega_k)^{z_{ik}} \\
&= \prod_{k=1}^{g} (\pi_k \times \prod_{j=1}^{p} p_{kj}^{x_j}(1 - p_{kj})^{1-x_j})^{z_{ik}}
\end{aligned}$$

4. On en déduit que la vraissemblance jointe des paramètres du modèle est :

$$\begin{aligned} L((x_1,...,x_n);(p_{kj},\pi_k)) &= \prod_{i=1}^{n} \mathbb{P}(X=x_i, Z=z_i) \\ &= \prod_{i=1}^{n}\prod_{k=1}^{g} \left[\pi_k \times \prod_{j=1}^{p} p_{kj}^{x_{ij}}(1-p_{kj})^{1-x_{ij}}\right]^{z_{ik}} \end{aligned}$$

5. On cherche à calculer l'EMV de chaque paramètre p_{kj} (k, j) $\in [\![1,g]\!] \times [\![1,p]\!]$.

$$\begin{aligned} lnL((x_1,...,x_n);(p_{kj},\pi_k)) &= \sum_{i=1}^{n}\sum_{k=1}^{g} ln\Big([\pi_k \prod_{j=1}^{p} p_{kj}^{x_{ij}}(1-p_{kj})^{1-x_{ij}}]^{z_{ik}}\Big) \\ &= \sum_{i=1}^{n}\sum_{k=1}^{g} z_{ik} \times ln\Big(\pi_k \prod_{j=1}^{p} p_{kj}^{x_{ij}}(1-p_{kj})^{1-x_{ij}}\Big) \\ &= \sum_{i=1}^{n}\sum_{k=1}^{g} z_{ik} \times \Big[ln(\pi_k) + \sum_{j=1}^{p} x_{ij} ln(p_{kj}) + (1-x_{ij}) ln(1-p_{kj})\Big] \end{aligned}$$

En dérivant par rapport à p_{kj}, on obtient :

$$\frac{\partial lnL((x_1,...,x_n);(p_{kj},\pi_k))}{\partial p_{kj}} = \sum_{i=1}^{n} z_{ik}(x_{ij} \times \frac{1}{p_{kj}} - (1-x_{ij}) \times \frac{1}{1-p_{kj}})$$

On cherche alors à annuler cette dérivée afin d'obtenir l'EMV de p_{kj} :

$$\begin{aligned} \sum_{i=1}^{n} z_{ik}(x_{ij} \times \frac{1}{p_{kj}} - (1-x_{ij}) \times \frac{1}{1-p_{kj}}) &= 0 \\ \sum_{i=1}^{n} z_{ik}(\frac{x_{ij}}{p_{kj}(1-p_{kj})} - \frac{p_{kj}}{p_{kj}(1-p_{kj})}) &= 0 \\ \sum_{i=1}^{n} x_{ij} &= \sum_{i=1}^{n} p_{kj} \end{aligned}$$

Or $\sum_{i=1}^{n} p_{kj} = n$
On en conclut que :

$$\boxed{\frac{1}{n}\sum_{i=1}^{n} x_{ij} = \hat{p_{kj}}} \tag{2}$$

De même, on peut calculer l'EMV de chaque probabilité à priori grâce au lnL calculé précédemment. On a la contrainte suivante :

$$\sum_{k=1}^{g} \pi_k = 1 \tag{3}$$

La formulation lagrangienne de ce problème est la suivante :

$$\mathcal{L}(L((x_1,...,x_n);(p_{kj},\pi_k)),\lambda) = lnL((x_1,...,x_n);(p_{kj},\pi_k)) - \lambda(\sum_{k=1}^{g} \pi_k - 1) \tag{4}$$

avec λ étant le multiplicateur de Lagrange associé à la contrainte.

On applique donc les conditions d'optimalité à ce Lagrangien :

$$\frac{\partial \mathcal{L}}{\partial \pi_k} = 0 \Leftrightarrow \frac{1}{\pi_k} \sum_{i=1}^{n} z_{ik} = \lambda \Leftrightarrow \frac{1}{\lambda} \sum_{i=1}^{n} = \pi_k$$
$$\frac{\partial \mathcal{L}}{\partial \lambda} = 0 \Leftrightarrow \sum_{k=1}^{g} \pi_k = 1$$

on en déduit :

$$\sum_{k=1}^{g} \pi_k = 1 \Leftrightarrow \sum_{k=1}^{g} \frac{1}{\lambda} \sum_{i=1}^{n} z_{ik} = 1 \Leftrightarrow \lambda = \sum_{k=1}^{g} \sum_{i=1}^{n} z_{ik} \Leftrightarrow \lambda = n$$

Finalement,

$$\boxed{\hat{\pi_k} = \frac{1}{n} \sum_{i=1}^{n} z_{ik}} \tag{5}$$

4 Programmation

Voici la fonction **binaryNBCfit**

```
1  binaryNBCfit <- function(X, z)
2  {
3    n <- nrow(X)
4    p <- ncol(X)
5    g <- max(unique(z))
6    param <- NULL
7    pkj <- array(0, c(g,p))
8    pik <- rep(0, g)
9    for (k in 1:g)
10   {
11     indk <- which(z==k)
12     Xk <- X[indk,]
13     nk <- length(indk)
14     pkj[k,] <- apply(Xk, 2, mean)
15     pik[k] <- nk / n
16   }
17   param$pkj <- pkj
18   param$pik <- pik
19   return(param)
20 }
```

Voici la fonction **binaryNBCval**

```
binaryNBCval <- function(param, Xtst)
{
  n <- nrow(Xtst)
  p <- ncol(Xtst)
  g <- length(param$pik)
  res <- list()
  prob <- matrix(0, nrow=n, ncol=g)
  for(k in 1:g)
  {
    pik <- param$pik[k]
    for(i in 1:n)
    {
      prob.ik <- 1
      for(j in 1:p)
      {
        pkj <- param$pkj[k,j]
        prob.ij <- pkj ** Xtst[i,j] * (1-pkj) ** (1 - Xtst[i,j])
        prob.ik <- prob.ik * prob.ij
      }
      prob[i,k] <- prob.ik * pik
    }
  }
  prob <- prob / apply(prob,1,sum)
  pred <- max.col(prob)
  res$prob <- prob
  res$pred <- pred
  return(res)
}
```