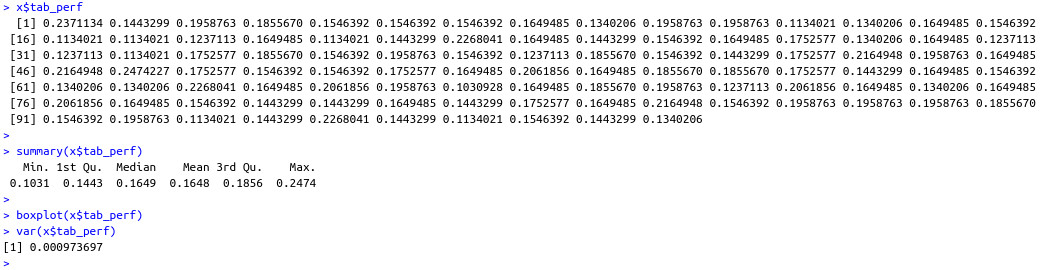
  4-**Classification supervisée :**

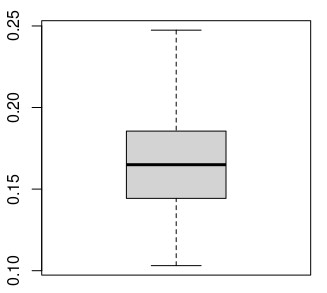
Pour ce qui est de la classification supervisée, nous utiliserons un protocole d'apprentissage et de tests par Bootstrap avec 50% de la population totale pour l’apprentissage et 50% pour les tests. En effet, la population est peu nombreuse (195 individus) et suivant le critère classe, 50% d’individus sains correspond à 24, ce qui est faible mais acceptable - contrairement à 20% habituellement pris et correspondant à 9 -.

Nous voulons noter que sur un test itératif effectué manuellement par nos soins, nous avons une prédominance statistique (17 sur 20 soit 85%) de la variable PPE comme étant la plus prédictive. Pour plus d’efficacité, nous écrivons un script qui va automatiser l’expérience aléatoire et itérer un plus grand nombre de fois.

Ainsi, nous exécutons le script du fichier «bootstrap.R» en choisissant 100 comme nombre d’itérations (Nb\_iter=100).

Sur la distribution des erreurs obtenue, les commandes summary et boxplot donnent :

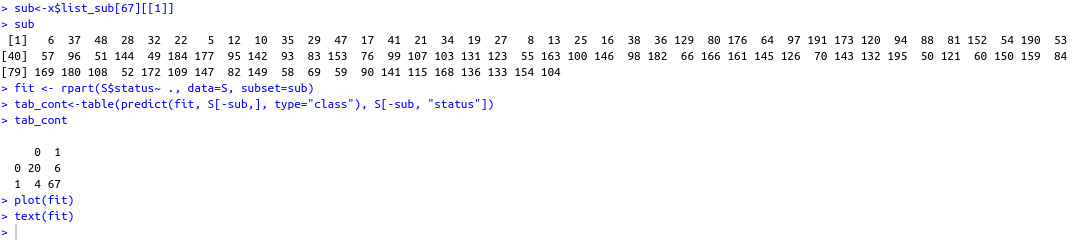


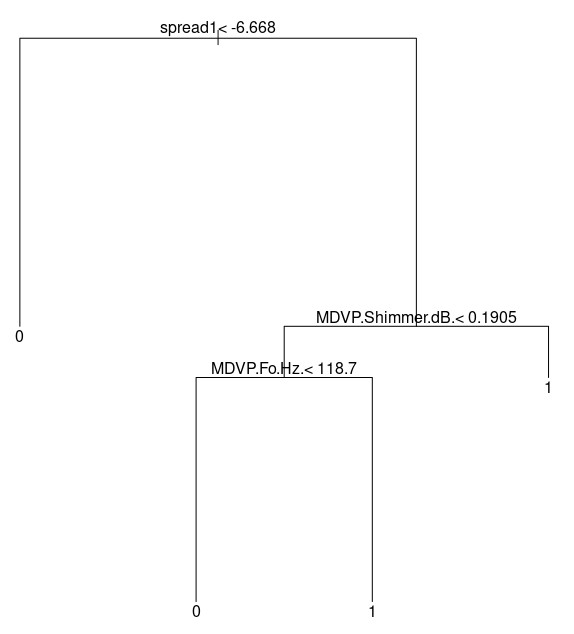


Ainsi le meilleur arbre aurait une erreur de 10,3%; pour une moyenne et une variance de distribution respectivement égales à 16,48% et 0.00097. Ce qui est déjà intéressant comme résultat. Mais dans le souci d’être le plus précis possible, nous choisissons d’enlever cette valeur (10,3%) pour voir comment se comporte la variance. Nous obtenons une erreur de 11,34% pour une moyenne des erreurs égale à 16,55% et une variance de 0.00094. Par conséquent la différence n’est pas significative. Nous restons donc sur les premières valeurs, soit: une performance de 10,3% d’erreur, avec une moyenne de 16,48% et une variance de 0.00097.

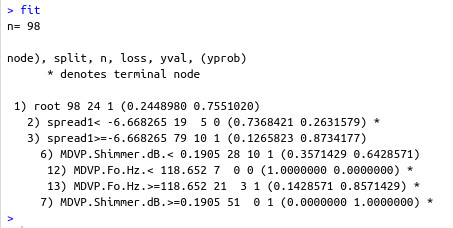
Grâce à la variable ‘list\_sub’ (qui contient la liste des échantillons tirés à chaque itération) retournée par notre script, on retrouve l’indice de l’erreur et l’échantillon associé.

On peut afficher la table de contingence et l’arbre construit :





Par ailleurs la commande ‘fit*’* nous donne le résultat suivant:



Il apparaît ici que **la variable la plus prédictive est ‘spread1’**

Sur 98 individus prélevés la majorité (environ 75,51%) est malade (yval=1) et le reste (24 équivalent à 24,48% environ) est en bonne santé.

Sur les 98 on en a 19 qui vérifient la condition ‘spread1<-6.668’ et la majorité (73,68% environ) est en bonne santé (yval=0)

On peut donc conclure qu’on a une confiance à 73,68% que si un individu a une valeur de ‘spread<-6.668’ alors il est en bonne santé.

De la même manière si ‘spread >=-6.668’ et ‘MDVP.Shimmer.dB>=0.1905’ on a une confiance de 100% que la personne est malade.

Enfin si ‘spread >=-6.668’ et ‘MDVP.Shimmer.dB<0.1905’ et

* si ‘MDVP.Fo.Hz<118.652’ on a une confiance totale (100%) que la personne est en bonne santé.
* si ‘MDVP.Fo.Hz>=118.652’ on une confiance de 85,71% que la personne est en malade.

Cet arbre donne de fortes probabilités pour les étiquettes majoritaires, on peut dire qu’il a une bonne prédictibilité.