U4: Entrepôts de données, fouille de données M2 IBM

Sujet : Analyse de données cliniques, des patients atteints d'un cancer du Poumon Sous RStudio

Réalisé Par: CHALAH Diana

Responsable UE4: JANNOT Anne-Sophie

Université PARIS 13 et PARIS DESCARTES

Année Universitaire: 2017/2018

1. Introduction :

L'imagerie médicale est l'un des principaux facteurs qui ont éclairé la science médicale et le traitement et a un grand potentiel pour guider la thérapie, car elle peut fournir une vue plus complète de la tumeur entière et elle peut être utilisée sur une base continue pour surveiller le développement et la progression de la maladie ou sa réponse à la thérapie. En outre, l'imagerie est non invasive et est déjà souvent répétée pendant le traitement dans la pratique de routine, contrairement à la génomique ou la protéomique, qui sont encore difficiles à mettre en œuvre dans la routine clinique.

Un objectif clé de l'imagerie actuelle est la «médecine personnalisée», où le traitement est de plus en plus adapté en fonction des caractéristiques spécifiques du patient et de sa maladie.

On parle dans ce texte d'une analyse radiomics de 440 caractéristiques quantifiant l'intensité, la forme et la texture de l'image tumorale, qui sont extraites de données de tomodensitométrie de 1 019 patients atteints de cancer du poumon ou de la tête et du cou.

Les données cliniques de ce texte vont nous servir à faire notre analyse de données cliniques sur les patients atteints du cancer du poumon.

(Rf :https://www.nature.com/articles/ncomms5006#f2)

2. Matériels et Méthodes :

Notre travail consiste à faire une analyse descriptives des données cliniques des patients atteints d'un cancer du poumon.
On est appelé à faire :
- Une interface interactive pour visualiser ces données cliniques, extraites depuis l'article mentionné au dessus et depuis ce lien :
«*https://wiki.cancerimagingarchive.net/display/Public/NSCLC-Radiomics-*»
- Faire une analyse descriptive .
- Faire une analyse uni-variée interactive permettant de visualiser les relations entre les données d'expression et les sous)types de cancer.
- Implémenter des arbre de décisions interactifs pour prédire les différents sous-types de cancer à partir des données de l'analyse uni-variée.

La travail se fera avec le langage **RStudio** et tous les packages associés pour faire l'analyse, ex de Shiny pour l'interface.
Geo2r pour l'analyse uni-variée (identifier les gènes qui sont exprimés différemment).

3. Résultats :

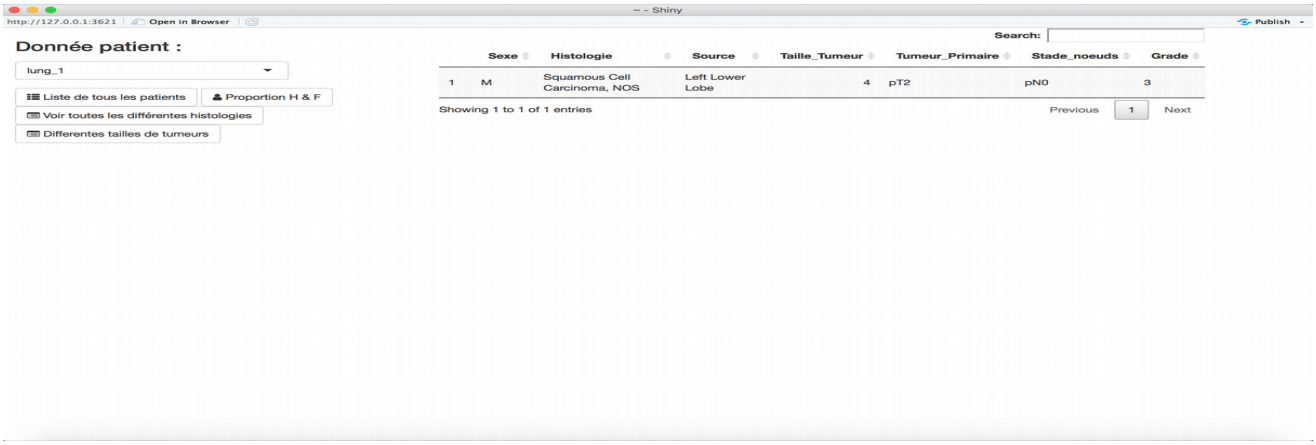
3.1 Réalisation de l'interface interactive pour la visualisation des données cliniques :

Dans cette partie, 2calsses on été crées sur RStudio :
- Une pour Importer les données de Lung-cancer, faire quelques petite modifications dessus pour rendre la table plus facile à manipuler, ex des colonnes qui semblaient les plus pertinentes à l'étude, renommées, ainsi les différentes données (diagramme et camembert à utiliser dans shiny).Création d'une fonction qu'on va appelé depuis la 2eme classe .

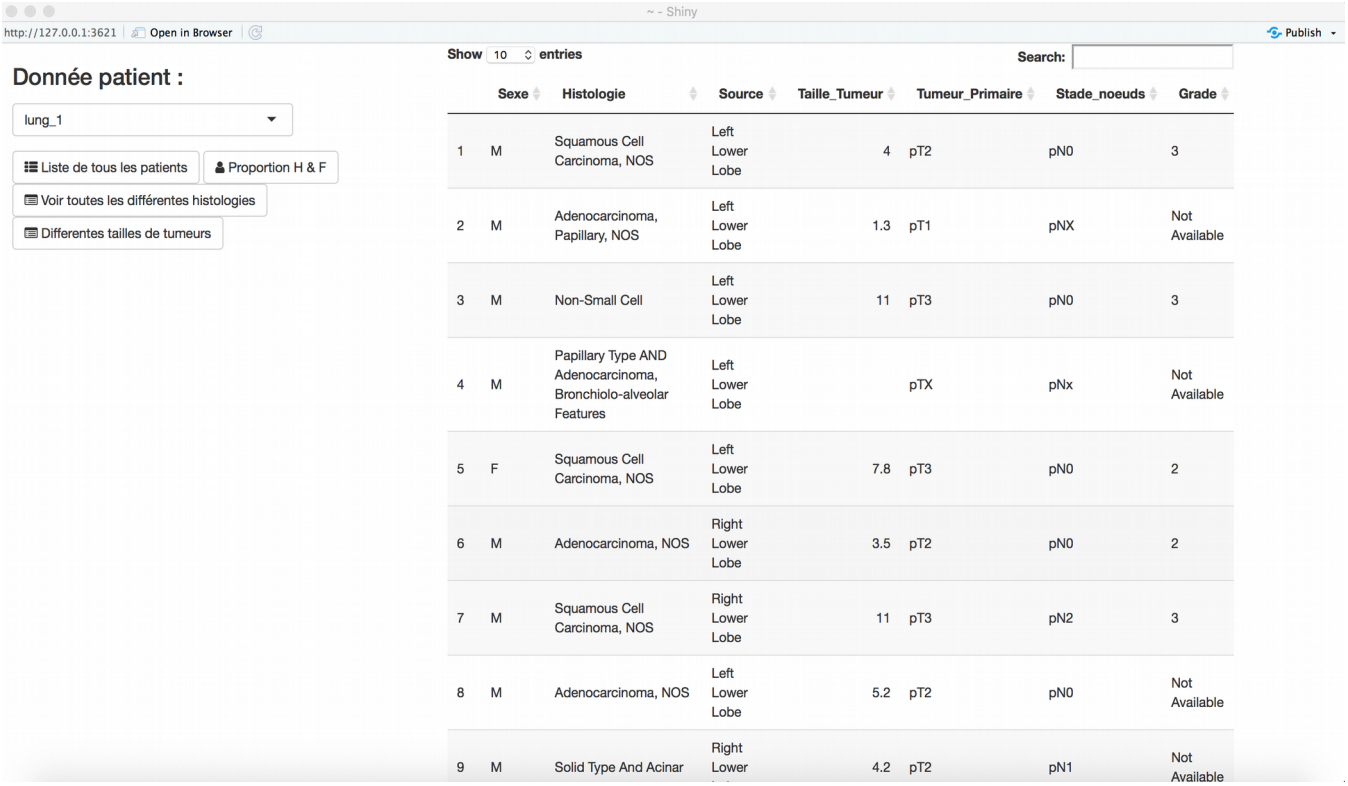
- La seconde va afficher l'interface souhaitée avec le package de Shiny.

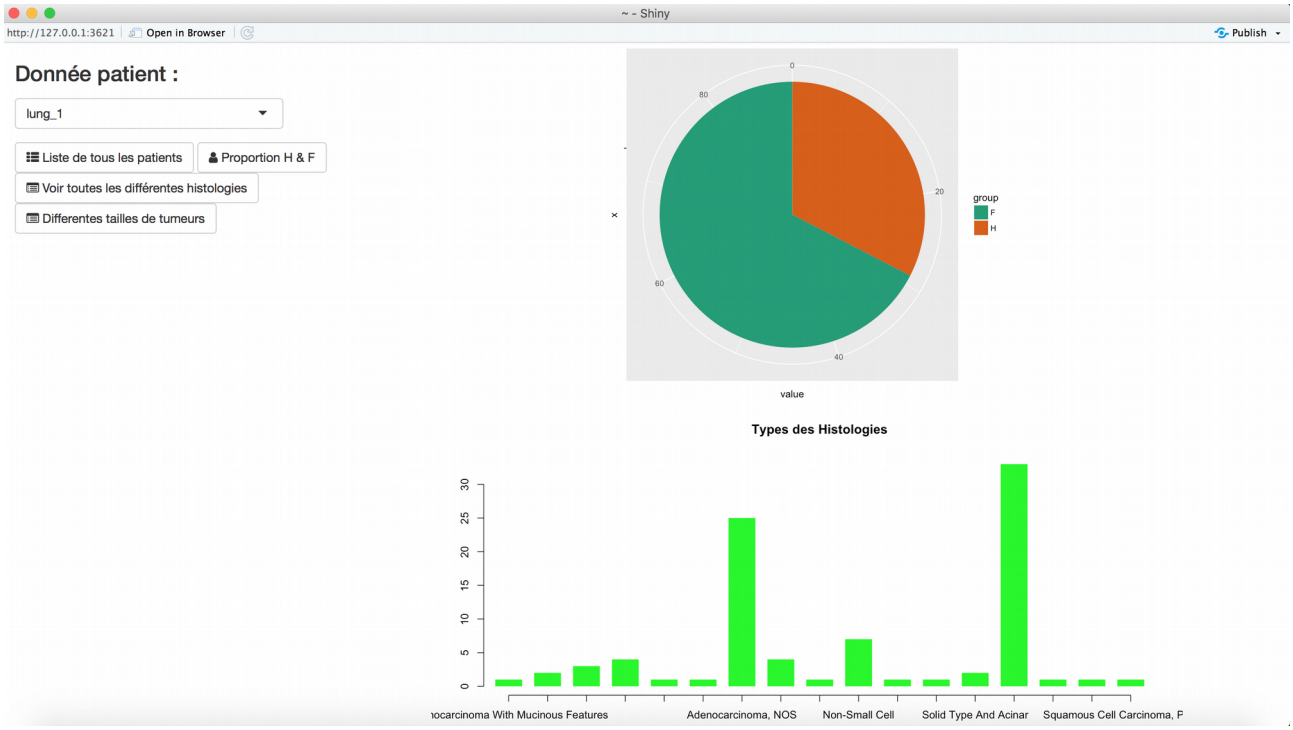
3.2 Analyse descriptive des données cliniques de ces patients :

Choix des données à afficher :
Sexe → Camembert Proportion des H & F dans l'étude
Histologie → Diagramme en battons des différents sous type de cancer trouvés
Taille-Tumeur → Diagramme en battons des différentes tailles de tumeurs

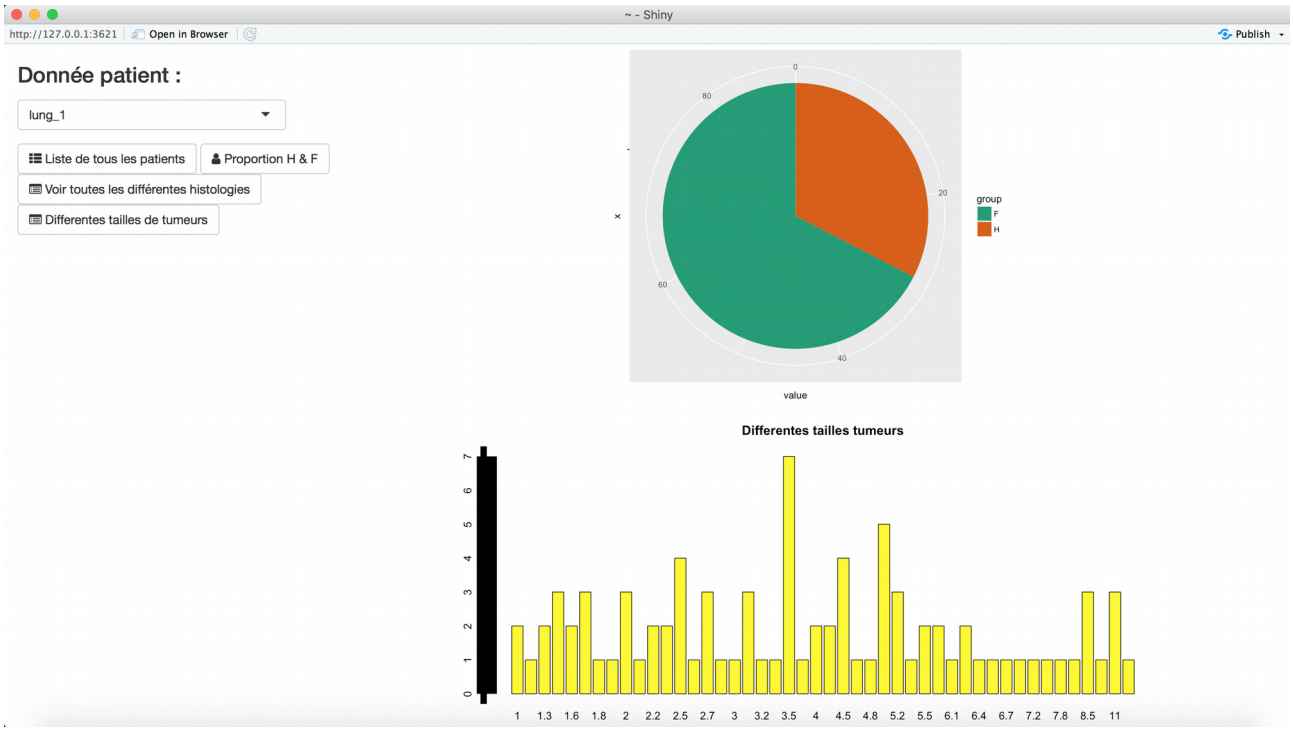


→ Ensemble de l'interface :
- Données patients (Lung) en menu déroulant
- Possibilité d'afficher toutes la liste des patients malades
- Visualisation Du camembert des proportion Sexe
- Visualisation Des Diagrammes (Histologie et Taille-Tumeur)





- Cette fenêtre nous montre le pourcentage des femmes et hommes atteints du cancer dont : environs 30 % sont des hommes et 70 des femmes.
- Le diagramme représente les différents sous types de cancer et à quel pourcentage ils sont représentés dans chez les malades : les tumeurs qui sont les plus fréquentes chez les malades sont les : Adenocarcinoma, NOS de 25 %, NON-Small Cell de moins d 10 % et les Solid Type and Acinar de 30 % environs.



- Diagramme des taille de tumeur : les tailles les plus fréquentes sont celles de : 3.5, 4.5, 5.2, 2.5 etc.

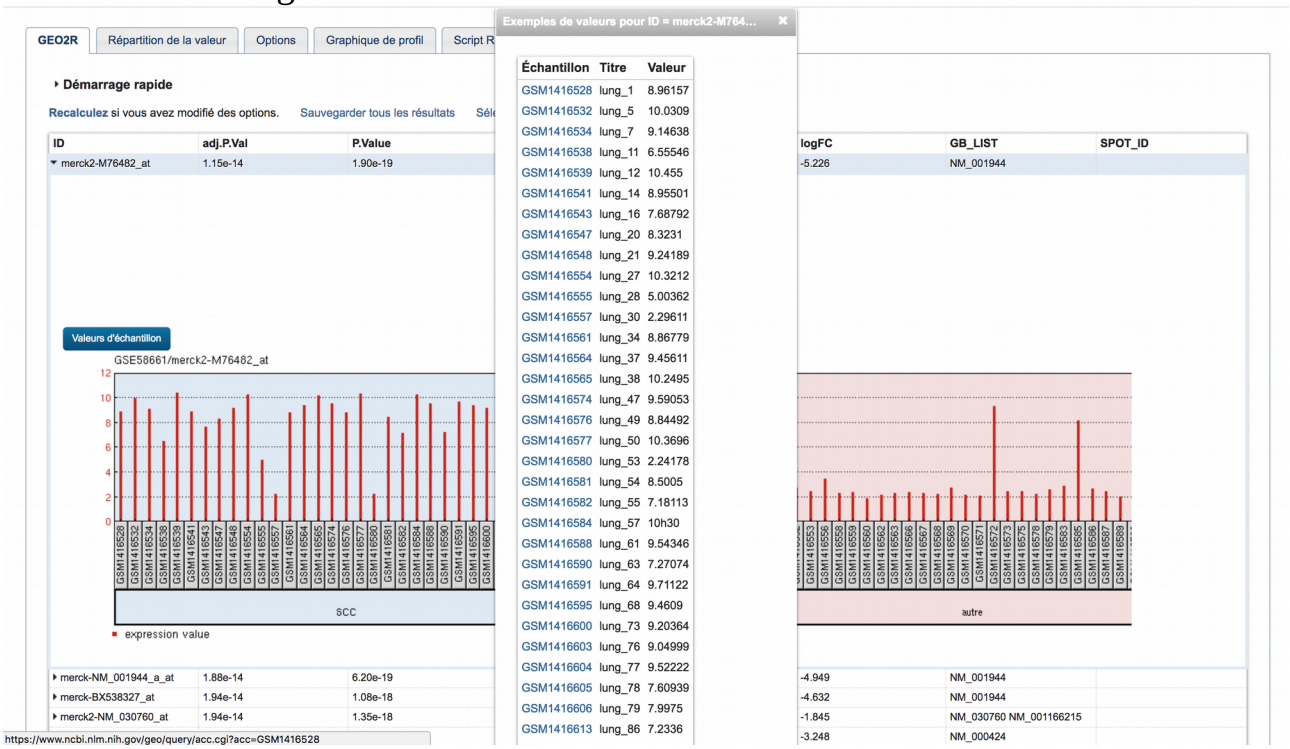
3.3 Analyse uni-variée interactive (relation entre les données d'expression et les sous types de cancer) :

Dans cette partie, j'ai utilisé **Géo2R** qui est un outil web interactif qui permet aux utilisateurs de comparer deux ou plusieurs groupes d'échantillons dans une série GEO afin d'identifier les gènes qui sont exprimés différemment dans les conditions expérimentales. Les résultats sont présentés sous la forme d'un tableau de gènes classés par ordre de signification.
(Rf : https://www.ncbi.nlm.nih.gov/geo/info/geo2r.html#interpret)

Depuis le lien fournit dans le sujet :
https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE58661 → je suis partie sur l'ensemble des données Lung3 des 89 patients, j'ai ensuite fait l'analyse en geo2r, dans ce dernier j'ai d'abord créé 2 groupes d'échantillons d'histologies (SCC et autre) à fin d'effectuer des comparaisons sur ces 2 types de groupes, j'ai par la suite lancer le top 250 pour l'exécution du test.

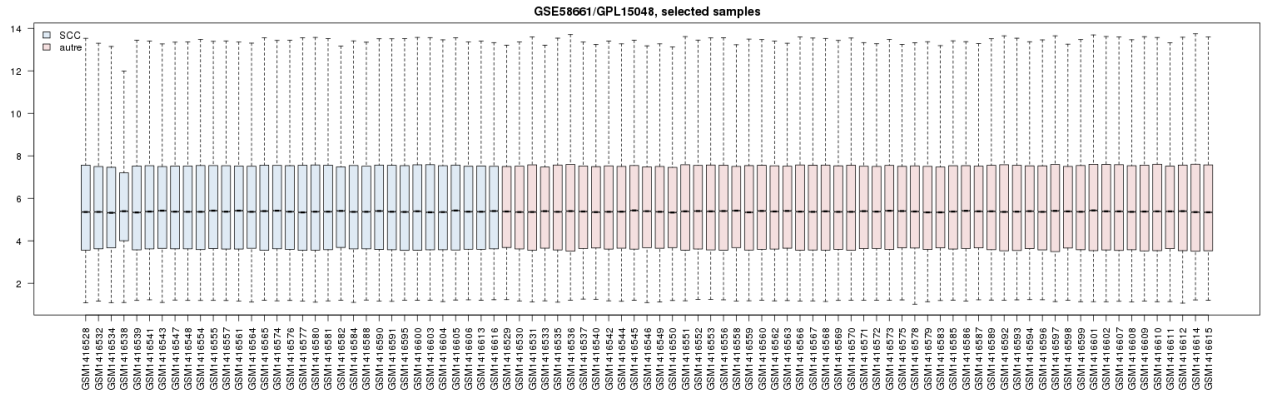
Les résultats sont présentés sous la forme d'un tableau des 250 principaux gènes classés par valeur P. Les gènes ayant la plus petite valeur P sont les plus significatifs.

On peut révéler le graphique du profil d'expression génique pour le gène. Chaque barre rouge dans le graphique représente la mesure d'expression extraite de la colonne de valeur de l'enregistrement d'échantillon.



Il existe en geo2R un bouton sample values, si on clique dessus, un graphe va s'afficher montrant la distribution des valeurs pour les échantillons sélectionnés.

Les distributions de valeurs peuvent être visualisées graphiquement sous forme de boite à moustache également.



Une partie tout aussi interessante, le **scriptR,** Cet onglet imprime le script R utilisé pour effectuer le calcul. Cette information peut être sauvegardée et utilisée comme référence pour la façon dont les résultats ont été calculés.(cette partie nous servira aussi et surtout pour la suite du travail → réalisation de l'arbre de décision).

3.4 Arbre de décision interactif pour prédire les différents sous-types de cancer à partir des données de l'analyse uni-variée :

Pour réaliser les arbres de décision on doit construire 2 types d'échantillons : Apprentissage et Test.
Ici j'ai construit à partir de la table ou matrice obtenu avec le code r de géo2r contenant toutes les données d'expressions génétiques (ex) une autre table (Expression&Tumor) à laquelle j'ai ajouté une dernière ligne (V250) des valeurs des sous types de cancer, dans le but est de définir les différents sous types de cancer à partir des données d'expression génétiques.

Libraire utilisées : rpart et randomforest.

J'ai tout d'abord commencer par charger les données, puis apporter de petite modification comme la suppression de la 1ere colonnes des identifiants.
En suite une petite description de la table avec la commande summary(), donnant la moyenne, la médiane, max, min, 1er et 3eme quantile.

Je suis passée en suite à la construction des échantillons d'apprentissage et des échantillons de test, avec une extraction d'échantillon et le choix d'une part de cet échantillon(validation croisée).

Datapq → représente l'échantillon d'apprentissage
Datestq → représente l'échantillon test

Ces deux derniers vont nous servie à construire l'arbre de décision.

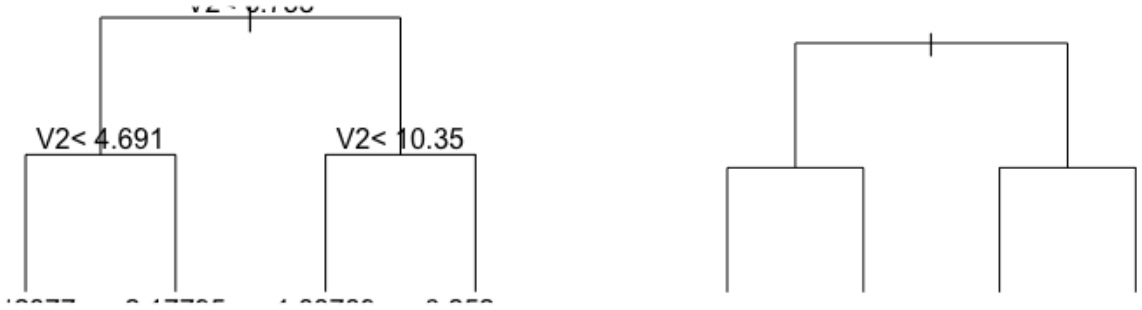
Avec la libraire rpart, je commence à construire mon arbre, on utilisant la commande :
fitq.tree=rpart(V250~.,data=datapq,
parms=list(split="information"),method="class")

En faisant un summary, il donne les variables les plus importantes, hors les données d'expression les plus significatifs :

Variable importance
V2 V117 V32 V101 V204 V25 V133 V17 V19 V217 V9 V125 V126 V208 V210 V99
20 8 8 8 8 8 4 4 4 4 4 4 4 4 3 3

et 6 nœuds et les informations qui vont avec (predict class etc..).

j'ai pu obtenir cet arbre de décision :



L'arbre nous donne comme Racine la variable V2 qui représente l'expression génétique ou le gène qui définit le plus les sous type de cancer avec les valeur < 4,6 et < 10.3.

Elagage :
Il sert à obtenir l'arbre le moins complexe donnant l'erreur par validation croisée la plus fiable.
Calcul de cp= coefficient de pénalisation de la complexité de l'arbre.

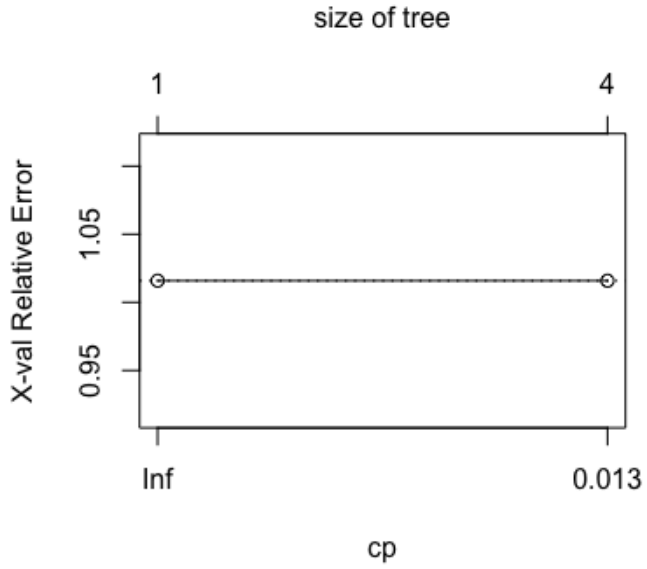
Résultat :
boucle allant de i a 0 :
cp= 0.49
boucle allant de i a 1 :
cp= 0.343
boucle allant de i 0:1 :
cp= 0.2401
cp= 0.16807

Variables actually used in tree construction:
[1] V2

Root node error: 63/64 = 0.98438

n= 64

	CP	nsplit	rel error	xerror	xstd
1	0.015873	0	1.00000	1.0159	0
2	0.010000	3	0.95238	1.0159	0



Random Forest :
Amélioration du bagging dans le cas spécifique de l'algorithme CART.
L'objectif est de rendre le modèle arbres construits plus indépendants entre eux.

L'algorithme des forêts d'arbres décisionnels effectue un apprentissage sur de multiples arbres de décision entraînés sur des sous-ensembles de données légèrement différents.

```
 fit=randomForest(V250~.,data=datapq,do.trace=20,
+                  importance=TRUE,norm.vote=FALSE)


> print(fit)

Call:
 randomForest(formula = V250 ~ ., data = datapq, do.trace = 20,       importance =
TRUE, norm.vote = FALSE)
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 83

          Mean of squared residuals: 0.2965728
                    % Var explained: 95.96
```

4.Discussion :

Par faute de temps j'ai pas pu avancer dans mon travail, j'ai eu beaucoup de difficulté à construire mes arbres .
Package rattle ne fonctionne pas.

Reference :
https://github.com/Diana1192/Projet_R_UE4/blob/master/Arbre%20de%20d%C3%A9cision

https://github.com/Diana1192/Projet_R_UE4/blob/master/Arbre%20de%20d%C3%A9cision

https://www.dropbox.com/sh/bmv5xqpgb5bekci/AABZqCaUHGmBclWhLqKdienVa/UE4/Cours%20du%2011%20d%C3%A9cembre/Machine%20learning-M2-2017.pptx?dl=0