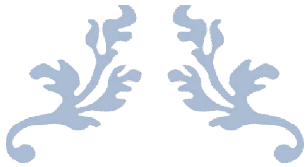


MODULE ENTREPOT ET FOUILLE DE DONNEES

18 DECEMBRE 2017

MASTER 2 INFORMATIQUE BIOMEDICALE

YIMEN KAMGANG MICHELE LOUNICI-ALI SELMA

1. Introduction

Le cancer du poumon représente la première cause de mortalité chez l'homme, ils sont les plus répandus dans le monde avec 900 000 nouveaux cas par an chez l'homme et 330 000 chez la femme, Ainsi le cancer du poumon est l'une des formes les plus fréquentes de cancer avec près de 40000 nouveaux cas par an. Plus de 4 cancers du poumon sur 5 sont liés au tabac. Les hommes sont actuellement plus touchés que les femmes. Néanmoins, cette tendance tend à disparaître compte tenu de l'augmentation du tabagisme féminin.

Le tabagisme représente la principale étiologie du cancer du poumon. Selon les estimations, près de 92 % des décès par cancer des poumons chez l'homme résultent d'une consommation de tabac. Le risque s'accroît en fonction de plusieurs paramètres : dose journalière de tabac, durée du tabagisme… La durée pendant laquelle on fume semble plus importante que la quantité de cigarettes fumées. Néanmoins, d'autres facteurs extérieurs peuvent être impliqués dans la survenue d'un cancer bronchique. C'est le cas, notamment, d'une exposition prolongée à des substances radioactives, à de l'amiante ou à d'autres toxiques (arsenic, nickel, chrome…). On peut également citer parmi les facteurs de risque la consommation de cannabis, les maladies inflammatoires chroniques des bronches ou encore la pollution atmosphérique.

Le cancer du poumon, appelé également cancer bronchique ou cancer broncho-pulmonaire, est une tumeur maligne développée à partir des cellules du poumon. Il existe différentes sortes de cancers pulmonaires, comme le cancer à petites cellules, très agressif, ou les adénocarcinomes, les carcinomes épidermoïdes entre autres...

L'objectif du projet est de faire une analyse descriptive des données cliniques issues d'une cohorte de patients atteints d'un cancer du poumon et d'implémenter des arbres de décision interactifs pour prédire les différents sous-types de cancer à partir des données d'expression génétique en utilisant les packages cart et randomforest.

2. Matériels et méthodes

Cette partie vise à décrire les outils utilisés pour atteindre l'objectif assigné à nos travaux ainsi que les différentes étapes suivies.

2.1.Matériel :

- RStudio version 3.3.2 est un environnement de développement multiplateforme pour R, un langage de programmation utilisé pour le traitement de données et l'analyse statistique. Il est disponible sous la licence libre AGPLv3, ou bien sous une licence commerciale, soumise à un abonnement annuel.

- ❖ GEO2R est utilisé pour comparer deux ou plusieurs groupes d'échantillons afin d'identifier les gènes qui sont différentiellement exprimés dans les conditions expérimentales. Les résultats sont présentés sous la forme d'un tableau de gènes classés par ordre de signification.
- ❖ Jeux de données « Lung3.metadata » a été utilisé dans le cadre de ce projet, il renseigne sur certains aspects cliniques de 89 patients atteints du cancer de poumon.

3. Résultat

3.1 question1

L'interface presenté ci-dessous permet de visualiser les données cliniques de chaque patient. Lorsque l'utilisateur selectionne un patient dans la liste deroulante, il obtient directement un tableau contenant les données cliniques du patient sélectionné, ce tableau est situé dans l'onglet table. Grâce au bouton « voir tous les paients » l'interface permet également de representer tous les patients et leurs données cliniques, de plus elle permet aussi de presenter la repartition et le grade du cancer chez les femmes et chez les hommes à l'aide dun graphe.

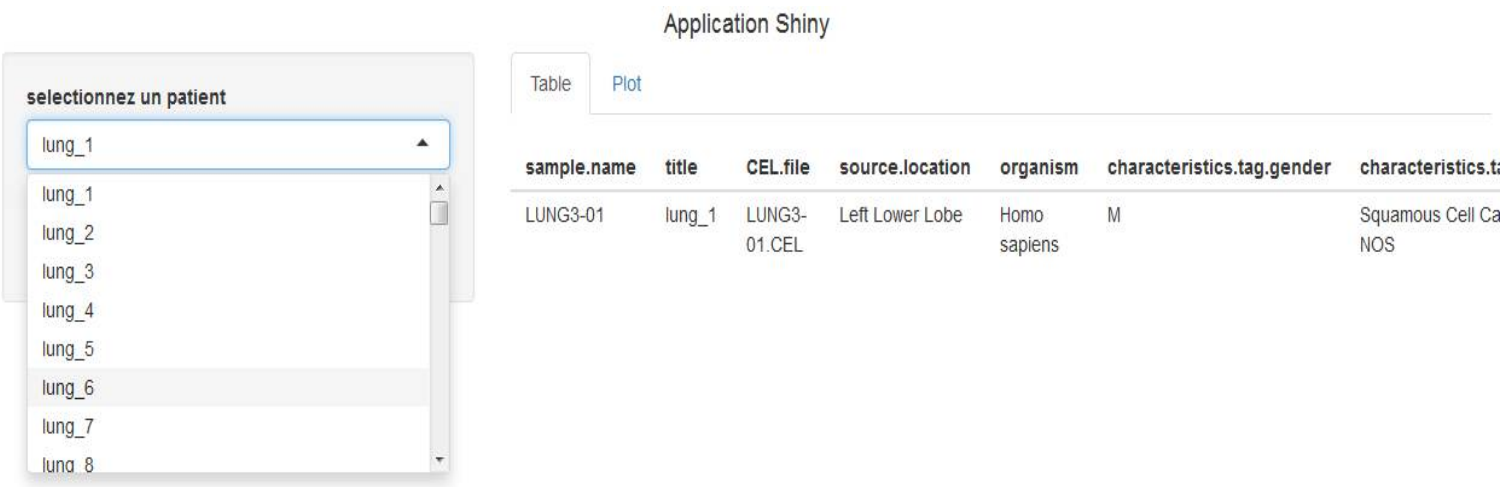


Figure 1: Patient Lung_1 sélectionné à droite données du patient

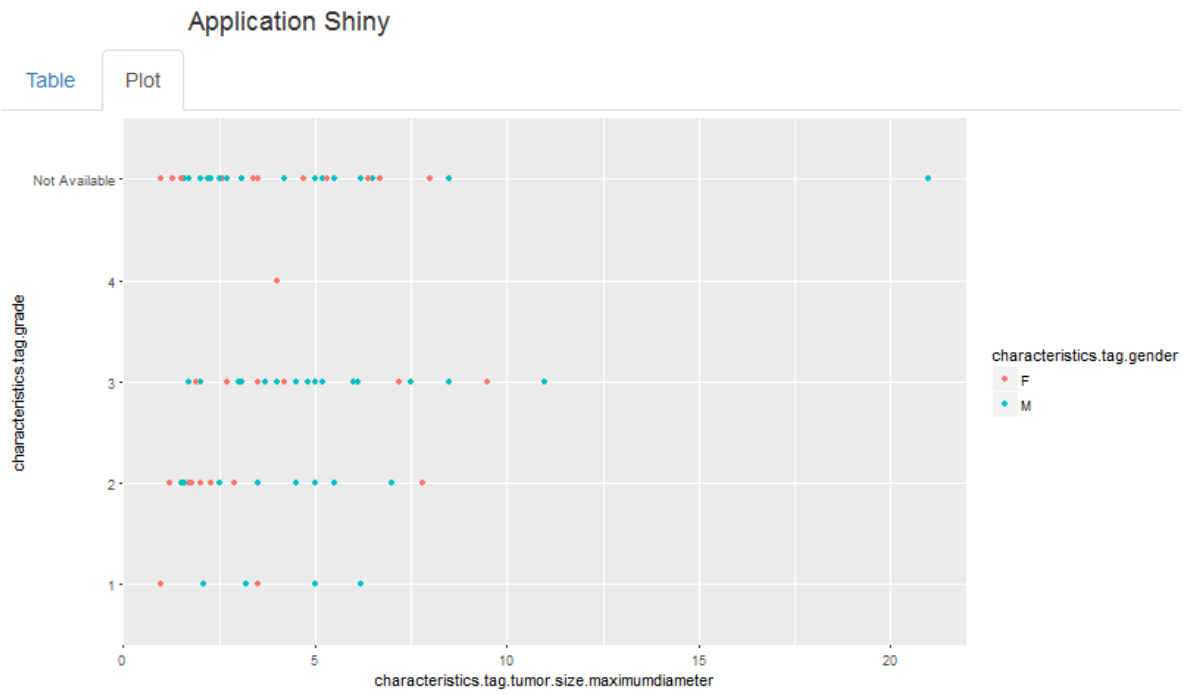
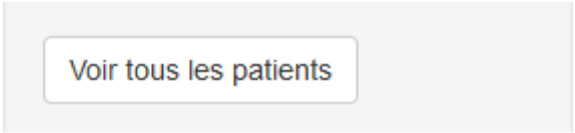


Figure 2:Plot présentant la taille et le grade de la tumeur chez les femmes et les hommes

En cliquant sur le bouton « voir tous les patients » on obtient un tableau avec tous les patients et l'interface donne la possibilité de choisir le nombre de patients qu'on souhaite avoir et donne également la possibilité de faire des recherches sur les patients



w 10 entries Search:

title	organism	CEL.file	characteristics.tag.histology
lung_1	Homo sapiens	LUNG3-01.CEL	Squamous Cell Carcinoma, NOS
lung_2	Homo sapiens	LUNG3-02.CEL	Adenocarcinoma, Papillary, NOS
lung_3	Homo sapiens	LUNG3-03.CEL	Non-Small Cell
lung_4	Homo sapiens	LUNG3-04.CEL	Papillary Type AND Adenocarcinoma, Bronchiolo-alveolar Features
lung_5	Homo sapiens	LUNG3-05.CEL	Squamous Cell Carcinoma, NOS
lung_6	Homo sapiens	LUNG3-06.CEL	Adenocarcinoma, NOS
lung_7	Homo sapiens	LUNG3-07.CEL	Squamous Cell Carcinoma, NOS
lung_8	Homo sapiens	LUNG3-08.CEL	Adenocarcinoma, NOS
lung_9	Homo sapiens	LUNG3-09.CEL	Solid Type And Acinar

Figure 3: Tableau de tous les patients

3.2. L'analyse descriptive des données cliniques des patients

A l'aide du logiciel R version 3.3.3 (R est un logiciel libre de traitement des données et d'analyse statistiques mettant en œuvre le langage de programmation S) des statistiques descriptives ont été effectuées sur l'ensemble des variables constituants les données des patients atteints du cancer du poumon. Cela on passant par différentes étapes comme l'importation du fichier « lung3 » la sélection des variables pertinentes en utilisant des commandes propre au langage R (voir Script)

La variable : Localisation

C'est une variable qualitative qui représente la localisation de la tumeur nous avons pu la décrire avec une représentation graphique en camembére (`pie(table(dttp$Localisation))`

Localisation	Patient
Left Lower Lobe	18
Left Upper Lobe	22
Right Lower Lobe	14
Right Middle Lobe	6
Right Upper Lobe	29

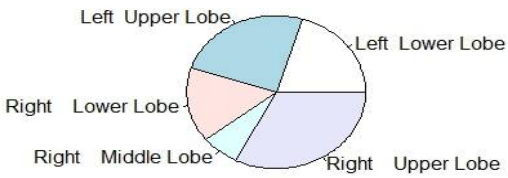


Figure4 : Source.location

La variable Gender

On note que le cancer du poumon est plus fréquent chez les hommes.

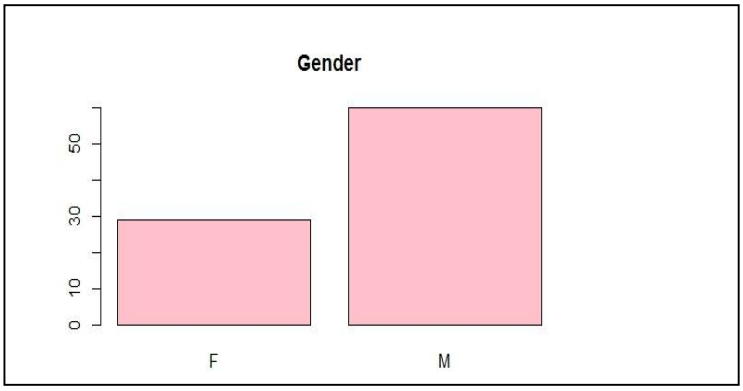
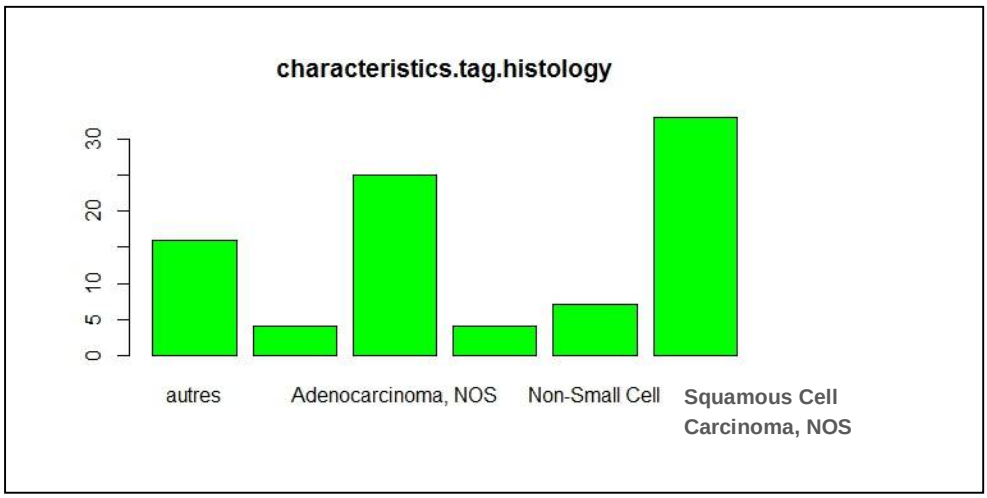


Figure5 : Gender

La variable : caractéristiques histologique

Il existe 17 sous types de cancer du poumon, afin de les présenter nous avons eu recours a les classer dans des sous catégories selon le nombre de patients. Nous avons eu 6 catégories représentées dans la figure suivante :

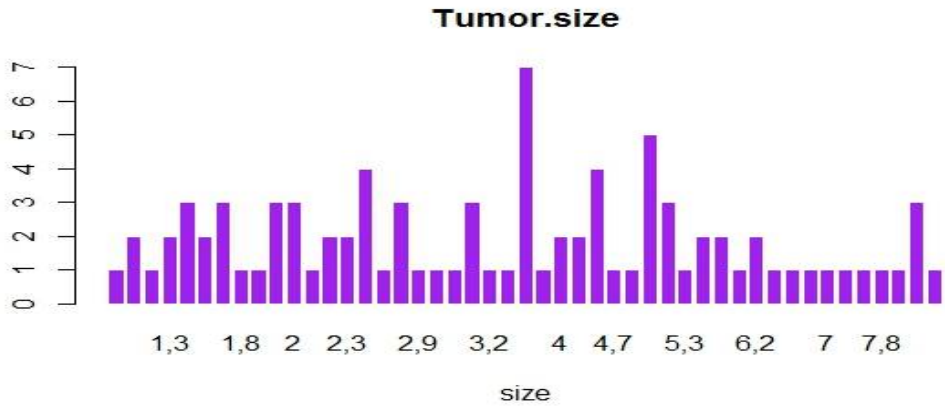


La variable : Tumor size(taille de la tumeur)

C'est une variable quantitative qui correspond à la taille de la tumeur

```
Min.    1st Qu.    Median      Mean       3rd Qu.    Max.    NA's
1.000   2.300       3.500      4.403      5.5         11      1

mean median  var   sd valid.n
Freq 1.89      1 1.62 1.27      47
```



La variable : Stage.primary.tumor

Afin de pouvoir analyser le stade de la tumeur nous avons regroupé les sous catégories pT1a et pT1b dans la catégorie tumeur envahissant le chorion pT1, ainsi que pT2a et pT2b dans pT2 (tumeur envahissant le muscle). Le camembére montre que ces deux tumeurs et pT3(tumeur envahissant le tissu periviscales) sont les plus élevés chez les patients

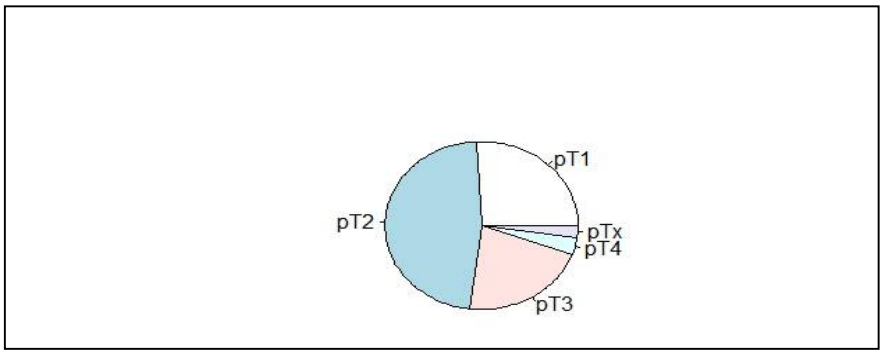


Figure : Stade de la tumeur

La variable: -stage.nodes

Variable qualitative qui représente le stade des ganglions on observe que le stade pN0 est le plus élevé chez les patients, 60 patients qui ne présentent pas de métastases ganglionnaires, 18 patients ont un seul ganglion atteint.

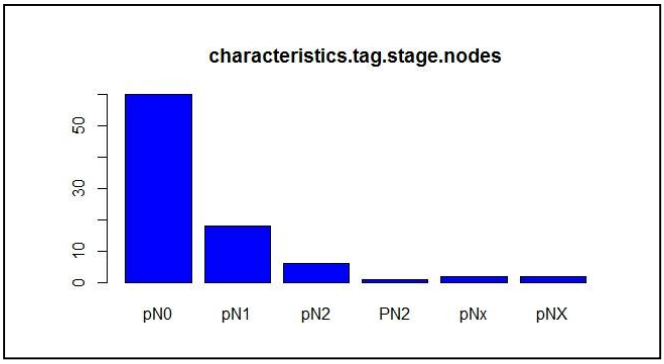


Figure : Stade des gonglion

La variable Stage.mets :

Variable qualitative qui représente le stade de la métastase, le stade pM0 est le plus élevé ce qui révèle que 83 des patient ne présentent pas de métastases à distances. Les métastases qui ne sont pas évalué (pMX), pM1 présence de métastases à distance.

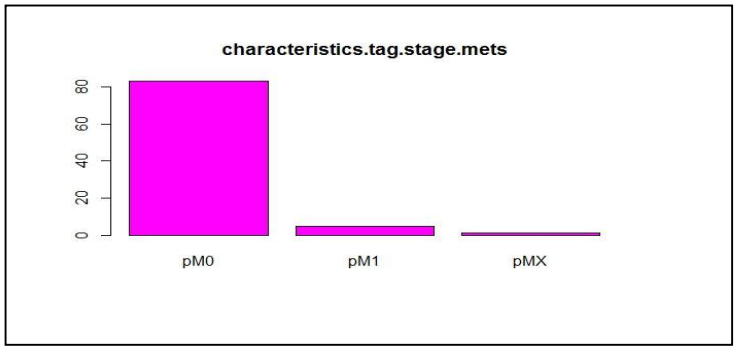


Figure : Stade des métastases

3.2 Analyses univariées entre les données d'expression et les sous-types de cancer

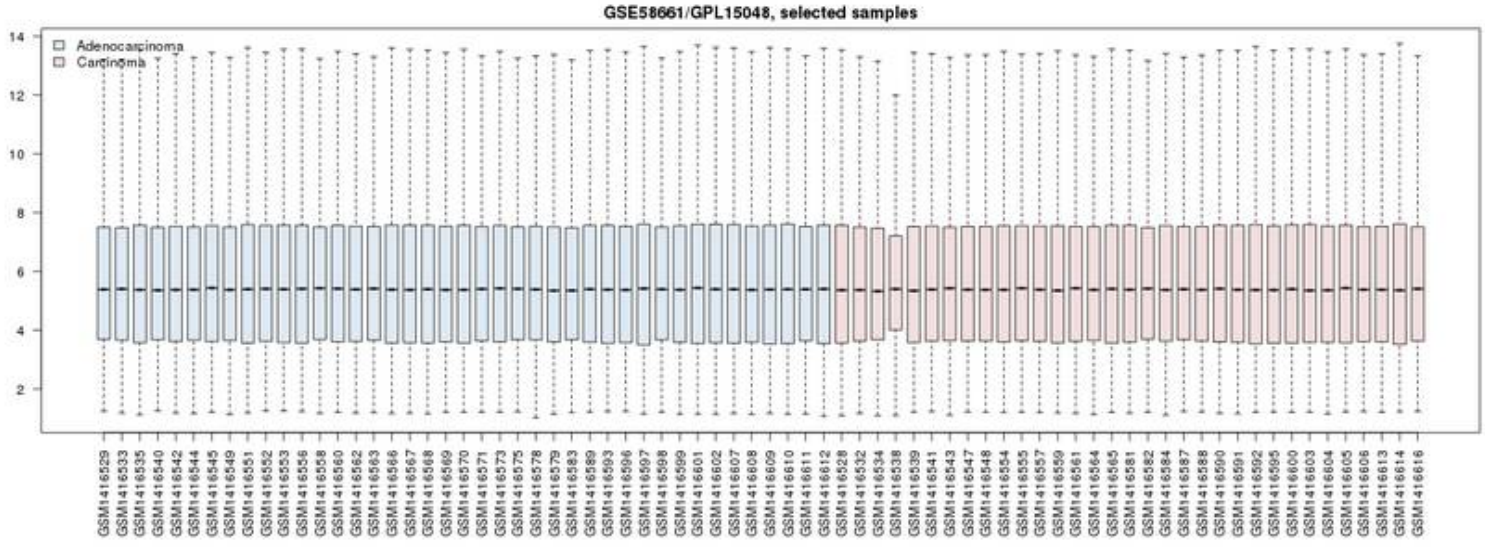
On a utilisé GEO2R pour comparer deux groupes d'échantillons : Adenocarcinoma et Cell Carcinoma afin d'identifier les gènes qui sont différentiellement exprimés dans les conditions expérimentales. Les résultats sont présentés sous la forme d'un tableau de gènes classés par ordre de signification. Les échantillons utilisés proviennent de la série GSE58661.

Lorsque l'on réalise une analyse d'expression différentielle, on obtient le tableau et l'ensemble de Boxplot ci-dessous (par GEO2R et R) :

ID	P.Value	Gene Symbol
merck2-M76482_at	1.11e-22	DSG3
merck-NM_000458_at	1.82e-22	HNF1B
merck-NM_001944_a_at	1.31e-21	DSG3
merck-BX538327_at	3.97e-21	DSG3
merck-NM_001010872_at	1.13e-20	FAM83B
merck-NM_000424_at	2.93e-20	KRT5
merck2-NM_030760_at	9.88e-20	S1PR5
merck2-NM_012397_at	1.04e-19	SERPINB13
merck2-NM_000424_at	1.06e-19	KRT5
merck2-AK074475_at	1.16e-19	DSC3
merck-NM_012397_s_at	1.63e-19	SERPINB13
merck-BE563343_s_at	2.77e-19	KRT5
merck2-AF297090_at	3.14e-19	DSC3
merck-NM_002855_at	6.38e-19	PVRL1
merck-NM_024423_a_at	6.94e-19	DSC3
merck-NM_001035223_s_at	7.58e-19	RGL3
merck2-NM_006481_at	8.31e-19	HNF1B
merck-NM_198460_s_at	1.72e-18	GBP6
merck-G36605_at	1.90e-18	S1PR5

Figure 4:Tableau de gènes exprimés

On note qu'il y a une différence d'expression entre les patients atteints du type de tumeur Adenocarcinoma et ceux atteints de Carcinoma.



Boxplot de l'analyse univariée

4. Implémenter des arbres de décision interactifs pour prédire les différents sous-types de cancer à partir des données d'expression génétique en utilisant les packages cart et randomforest.

Les arbres de décision sont des outils d'aide à la décision présentés sous la forme visuelle d'un arbre : la base de l'arbre est la racine, contenant une population d'individus à répartir. C'est ce qu'on nomme « base d'apprentissage ». Cette base comprend un ensemble de variables décrivant et de différenciant chacun des individus, ainsi qu'une variable d'intérêt dite « cible ».

Afin d'établir des arbres de décision sous shiny en premier lieu nous avons préparé la base de travaille en la nettoyant et laissant que les variables pertinentes. Puis nous avons construit des échantillons d'apprentissage et des échantillons de test en initialisant le générateur avec la commande (`set.seed(111)`, puis l'extraction des échantillons et cela en établissant le test ratio afin de définir la par du groupe test.

La transformation des variables facteur en variables numériques avec la commande `(dttr=data.frame(matrix(as.integer(as.matrix(dttr[,1:8])),ncol=8,dimnames= dimnames(dttr[,1:8])),Class=data[,8])`

Nous avons eu 71 pour échantillon d'apprentissage et 18 : pour le test.

Nous avons tenté de faire une modélisation avec un arbre de décision avec la commande

```
(fitq.tree=rpart(Class~.,data=datapq,
                 parms=list(split="information"),method="class")
```

Malheureusement nous n'avons pas pu finir le travaille par manqué de temps ainsi que les difficultés que nous avons rencontré dans l'exécution des commandes sur Rstudio (problèmes de packages et de versions)

5. Conclusion

L’analyse descriptive des variables montre que les hommes ont un grade élevé de tumeur de poumon que les femmes. Les femmes sont aussi vulnérables au cancer du poumon qui est la cause de mortalité par cancer chez les femmes, après le cancer du sein. Ce résultat sous-tend une vulnérabilité biologique et/ou génétique du sexe féminin face au tabac.

D’autres études doivent être menées pour déterminer avec exactitude ce qui déclenche ces dommages. Les recherches futures, élargies pour inclure les facteurs environnementaux et génétiques, amélioreront probablement notre compréhension de la pathogenèse de la maladie de cancer de poumon et en fin de compte, conduiront à de nouvelles approches du traitement.