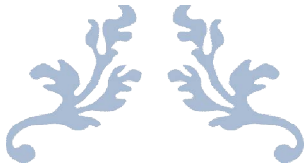

MODULE ENTREPOT ET FOUILLE DE DONNEES

15 JANVIER 2018

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, top10_matrix et data », a été utilisé dans le cadre de ce projet, il renseigne sur certains aspects cliniques de 89 patients atteints du cancer de poumon et l'expression génétique des sous type de cancer.

3. Analyse descriptive des données cliniques

3.1.Interface interactive des données cliniques des patients

L'interface presentée ci-dessous permet de visualiser les données cliniques de chaque patient. Lorsque l'utilisateur clique sur « patient » puis selectionne un patient dans la liste deroulante, il obtient directement un tableau contenant les données cliniques du patient sélectionné. L'interface permet également de visualiser tous les patients et leurs données cliniques.

Figure 1: Interface interactive des données cliniques par patient

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.

3.2.L'analyse descriptive des données cliniques des patients

L'analyse descriptive interactive consiste en premier lieux de sélectionner les variables pertinentes (voir Script) puis réaliser une analyse descriptive de ces variables. L'application réalisée avec Shiny sous R permet de visualiser les graphiques de chaque variable.

Localisation : On distingue que la partie la plus affectée par le cancer est le lobe supérieur droit du poumon chez les patients suivi de la partie supérieure gauche comme le montre la figure suivante.

Figure2 : Représentation graphique de la Localisation de la tumeur

Gender

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

Figure3 : Représentation graphique de la variable Gender

Histologie

Il existe 17 sous types de cancer du poumon, afin de les présenter nous avons eu recours a les classer en 3 sous catégories selon le nombre de patients.

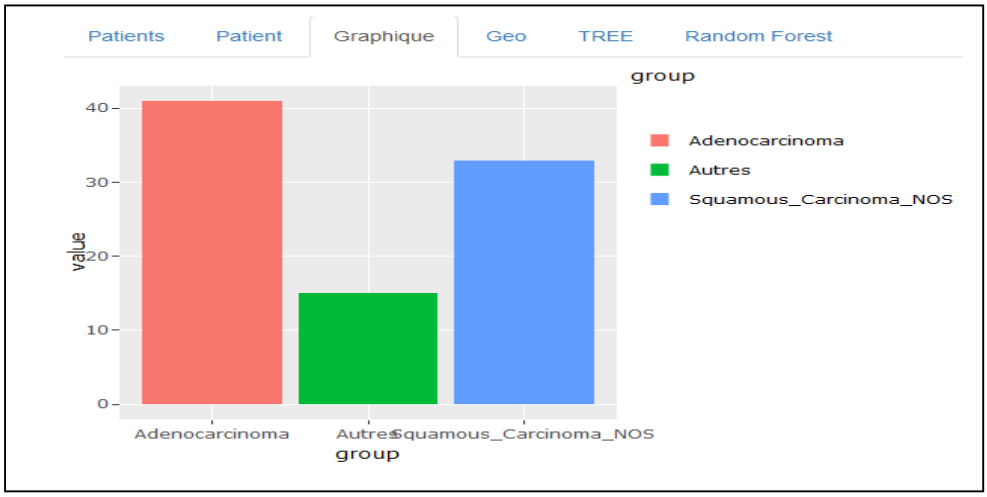

Figure3 : Représentation graphique de l'histologie

Tumor.size(taille de la tumeur)

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

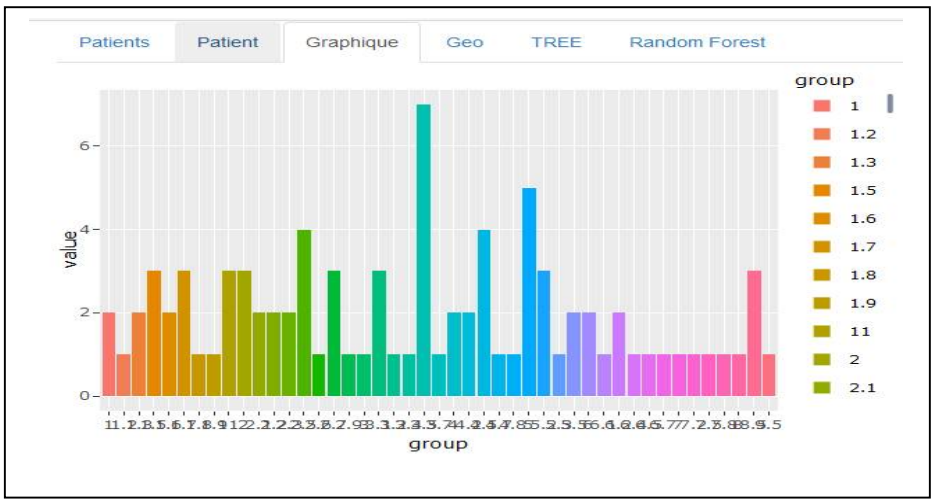

Figure 4 : Représentation graphique de la taille de la tumeur

Stade de la tumeur (T)

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

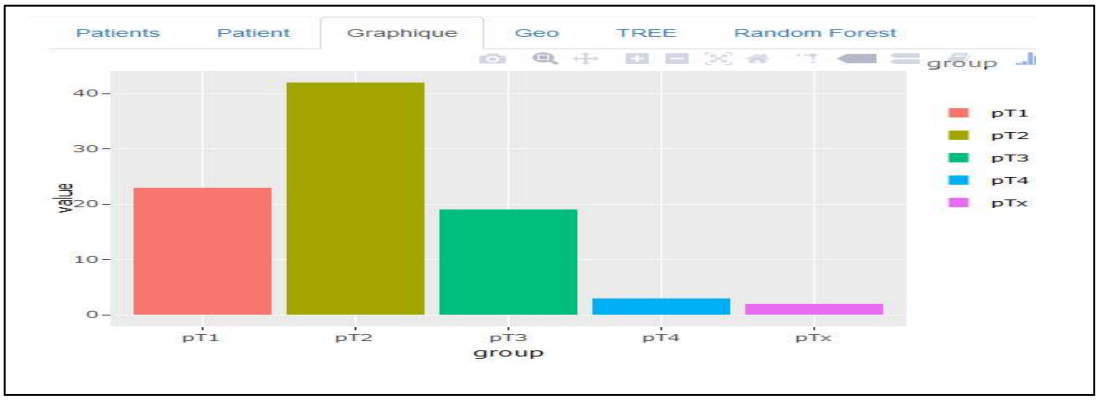

Figure 5 : Représentation graphique du stade de la tumeur

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.

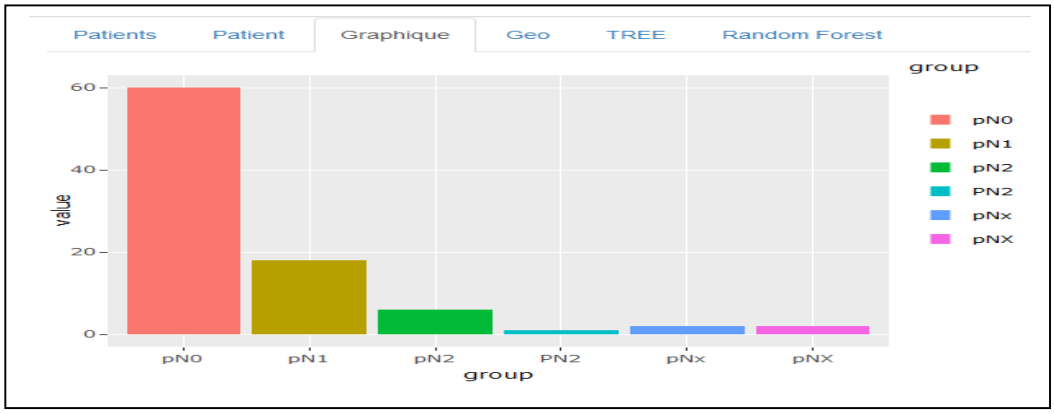

Figure 6 : Stade des ganglions

Stade des métastases :

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.

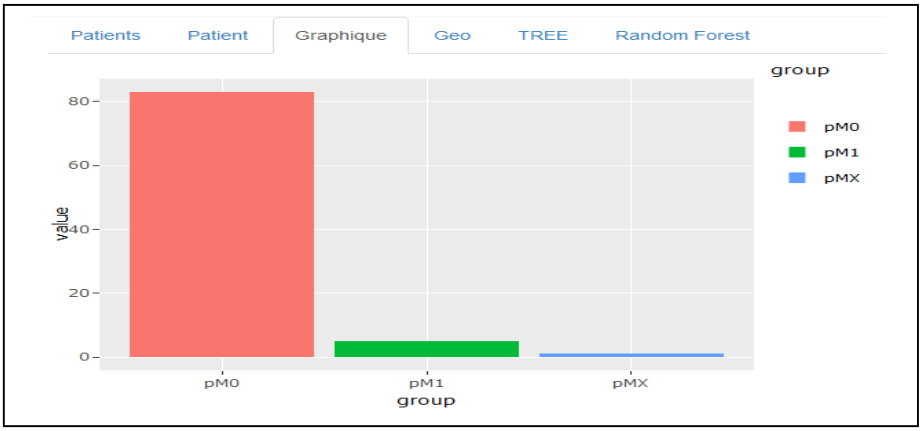

Figure : Stade des métastases

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

Avec GEO2R on a constitué 3 groupes de sous type histologique (Adenocarcinoma et Cell Carcinoma NOS et autres) 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 (par GEO2R et R) :

Patients | Patient | Graphique | Geo | TREE | Random Forest

Show 10 entries Search:

	ID	adj.P.Val	P.Value	F	GB_LIST
1	merck2-NM_030760_at	1.53e-12	2.52e-17	63.7353185	NM_030760 NM_001166215
2	merck2-M76482_at	5.91e-12	1.95e-16	58.5839245	NM_001944
3	merck-NM_001944_a_at	8.18e-12	4.17e-16	56.738089	NM_001944

Figure8: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 Squamous Carcinoma NOS et les autres sous types histologiques.

4. Arbres de décision

L'objectif de cette partie est l'implémentation des arbres de décision interactifs pour prédire les différents sous –types de cancer a partir des données d'expression génétique.

Afin d'établir des arbres de décision sous Shiny en premier lieu nous avons préparé la base de travaille qui est une base de donnée qui contient les 10 premiers gènes les plus expressifs que nous avons pu obtenir a partir de l'analyse GO2R et la matrice de la série GSE58661.

Après la construction de l'échantillon apprentissage et l'échantillon test nous avons pu obtenir l'arbre de décision représenté dans la figure 9.

L'arbre montre 6 classes de sous-type de cancer qui sont caractérisés par la présence des gènes les plus exprimés. Les patients n'ayant pas le gène « merck-NM_001944_a_at » sont atteint du sous-type « Squamous carcinoma Nos »

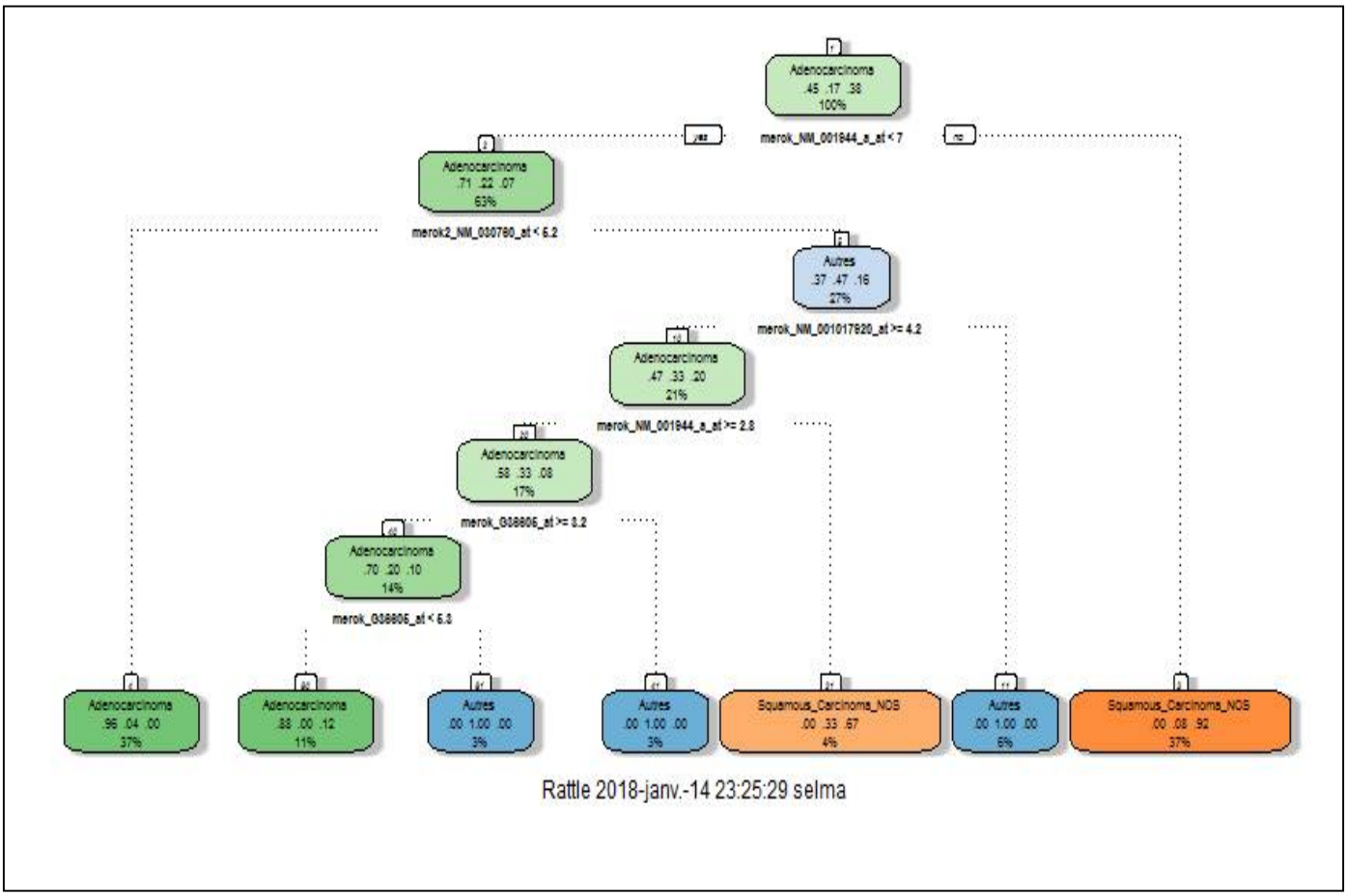

Figure 9 : Arbre de décision pour prédire les différents sous types de cancer

5. Randomforest :

Le randomforest montre que le gène « merck-NM_001944_a_at » a la plus importante contribution dans la prédiction du cancer du poumon.

	MeanDecreaseAccuracy	MeanDecreaseGini
merck_BE563343_s_at	0.04	3.83
merck_BX538327_at	0.04	5.64
merck_G36605_at	0.03	3.25
merck_NM_001010872_at	0.03	3.78
merck_NM_001017920_at	0.02	3.04
merck_NM_001944_a_at	0.09	7.30
merck_NM_005268_at	0.02	2.76
merck_NM_030760_at	0.04	3.24
merck2_M76482_at	0.05	4.96
merck2_NM_030760_at	0.09	5.91

6. Conclusion

L’analyse descriptive des variables montre que les hommes sont plus affectés par le cancer du poumon mais les femmes présentent un grade plus élevé de tumeur de poumon. Ce qui s’explique par leu vulnérabilité 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.

La majorité des patients ne présentent pas de métastases.

L’arbre de décision et le randomforest prédisent que le gène le plus responsable du cancer du poumon chez les 89 patients est « merck-NM_001944_a_at ».

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.