Projets Réseaux et Graphes (M2 Machine Learning)

Survie post opératoire des patients atteints d'un cancer du poumon

Professeur :

Severine AFFELDT

Étudiant :

Mohamed Abdelhadi Boudjemai

Année universitaire : 2020 – 2021

Table des matières

1. Introduction

À l'échelle internationale, le cancer du poumon reste la principale cause de décès par cancer chez les hommes et les femmes. L'incidence et la mortalité du cancer du poumon sont étroitement liées aux habitudes de tabagisme. Alors que les taux de tabagisme atteignent leur maximum - généralement d'abord chez les hommes, suivis des femmes - l'incidence du cancer du poumon et la mortalité augmentent au cours des décennies suivantes avant de diminuer après le lancement de programmes complets de lutte antitabac. Ces tendances sont apparues plus tôt dans les pays industrialisés que dans le monde en développement.

Le cancer du poumon est une des plus importantes causes de mortalité en France. Il est notamment causé par le tabagisme passif ou actif (9 cas sur 10) et est en général très agressif. L'une des options de traitement est l'ablation chirurgicale de la partie touchée du poumon ; elle permet théoriquement de stopper la propagation du cancer, et d'éviter, si l'intervention a lieu suffisamment tôt, que le cancer ne produise des métastases. Retirer une partie du poumon constitue cependant une prise de risques importante, et le bénéfice en termes de survie engendré par une telle procédure se doit par conséquent d'être avéré.

Ce projet propose d'étudier un jeu de données en rassemblant les observations réalisées sur un nombre de patients ayant un cancer du poumon pour la prédiction, selon les variables définies, l'espérance de vie de ces patients après l'intervention. Cependant, la reconstruction d'un réseau mettant en jeu tous les informations mises en survie en post opératoire est importance.

2. Objectif

L'objectif de ce mini-projet est l'exploitation des différents outils de reconstruction et d'analyse de réseaux afin d'étudier un système complexe à partir de données disponibles en ligne.

Parmi les étapes importantes du projet, il vous faudra notamment :

- Effectuer une étude comparative des résultats obtenus à partir des différents outils de Recontruction (Aracne, constraint-based, search & score...)
- Etudier un jeu de données rassemblant les observations réalisées sur 470 patients ayant subi ce type de chirurgie entre 2007 et 2011. L'enjeu majeur de sa création était la prédiction, selon les variables définies, de l'espérance de vie des patients après l'intervention.

Cependant dans ce projet, étant donné le nombre de variable observées par patient ainsi que leur relative simplicité, la reconstruction d'un réseau mettant en jeu tous ces paramètres peut s'avérer important pour comprendre les relations indirectes entre eux.

3. Description des méthodes utilisées

3.1 Hill climbing

Les méthodes de descente sont assez anciennes et doivent leur succès à leur rapidité et leur simplicité. A chaque pas de la recherche, cette méthode progresse vers une solution voisine de meilleure qualité.
La descente s'arrête quand tous les voisins candidats sont moins bons que la solution courante ; c'est-à-dire lorsqu'un optimum local est atteint.
On distingue différents types de descente en fonction de la stratégie de génération de la solution de départ et du parcours du voisinage : la descente déterministe, la descente stochastique et la descente vers le premier meilleur.

Algorithme 1 Méthode de descente générique

Procédure : φ fonction de coût
Variable locale : S solution courante
Choix d'une solution initiale S_0 ;
Solution courante $S \leftarrow S_0$;
(a.) Génération des candidats par voisinage ;
Choix du meilleur candidat C ;
if $\varphi(C) < \varphi(S)$ **then**
  $S \leftarrow C$;
  Aller en (a.) ;
end if
return S

3.2 Aracne algorithm

Aracne est un algorithme qui utilise l'information mutuelle comme base de départ. Il compare l'information au sens de Shannon apportée par un gène par rapport à un groupe de gènes. Si cette information est nulle il considère que le gène est défini par le groupe.

3.3 Constraint-based :PC algorithm : Probability Collectives Algorithm

L'approche PC présente les caractéristiques clés suivantes qui en font un choix compétitif par rapport aux autres algorithmes d'optimisation des collectifs.
PC est une approche de solution distribuée dans laquelle chaque agent met à jour indépendamment sa distribution de probabilité à tout moment et peut être appliquée à une variables discrètes ou mixtes, etc.,. La distribution de probabilité de l'ensemble de stratégies étant toujours un vecteur de nombres réels quel que soit le type de variable considéré, les techniques conventionnelles d'optimisation des vecteurs euclidiens, comme la descente de gradient, peuvent être exploitées.

(1) Elle est robuste dans le sens que la fonction de coût (objectif global / système) peut être irrégulière ou bruyante, c'est-à-dire qu'il peut accepter des problèmes bruyants et mal modélisés.

(2) L'agent défaillant peut simplement être considéré comme un agent qui ne met pas à jour sa distribution de probabilité, sans affecter les autres agents. D'autre part, il peut gravement nuire à la performance d'autres techniques.

(3) Il fournit des informations de sensibilité sur le problème dans le sens où une variable avec une distribution de pic (ayant la valeur de probabilité la plus élevée) est plus importante dans la solution qu'une variable avec une large distribution ; c'est-à-dire que la distribution de pics fournit le meilleur choix d'action qui peut optimiser l'objectif global.

(4) La formation de la fonction d'homotopie pour chaque agent (variable) aide l'algorithme à sauter hors des minima locaux possibles et à atteindre plus loin les minima globaux.

(5) Il peut avec succès éviter la tragédie des communs, sauter les minima locaux et atteindre davantage les véritables minima globaux.

(6) La charge de calcul et de communication est légèrement moindre et également répartie parmi tous les agents.

(7) Il peut gérer efficacement des problèmes avec un grand nombre de variables.

3.4 miic algorithm (Multivariate Information based Inductive Causation)

MIIC offre la possibilité de tenir compte de la différence entre les interactions dues aux variables latentes et les interactions directes, grâce à l'élargissement de la recherche des contributeurs à toutes les variables du réseau considéré, pas seulement les voisins. Ceci permet de traiter correctement les reconstructions de réseaux complexes comprenant des variables latentes.
Le second apport de miic est la possibilité de calculer un indice de confiance spécifique à chaque lien inféré et d'utiliser cette quantité pour effectuer un filtrage.
MIIC est capable de détecter une autocorrélation à décroissance exponentielle.

4. Etude comparative des Résultats Obtenues

On effectue une étude comparative sur un jeu de données et les résultats obtenus à partir des différents outils de Recontruction (Aracne, constraint-based, search & score...) sont comparés.

Informations sur l'ensemble de données :

Les données ont été collectées rétrospectivement au Wroclaw Thoracic Surgery Center pour les patients ayant subi des résections pulmonaires majeures pour un cancer du poumon primitif dans les années 2007-2011. Le Centre est associé au Département de chirurgie thoracique de l'Université de médecine de Wroclaw et au Centre de Basse-Silésie pour les maladies pulmonaires, Pologne, tandis que la base de données de recherche fait partie du Registre national du cancer du poumon, administré par l'Institut de la tuberculose et des maladies pulmonaires. À Varsovie, Pologne.

Informations d'attribut :

1. DGN: Diagnostic - combinaison spécifique de codes CIM-10 pour les tumeurs primaires et secondaires ainsi que pour les tumeurs multiples le cas échéant (DGN3, DGN2, DGN4, DGN6, DGN5, DGN8, DGN1)
2. PRE4: Capacité vitale forcée - FVC (numérique)
3. PRE5: Volume qui a été expiré à la fin de la première seconde d'expiration forcée - FEV1 (numérique)
4. PRE6: Statut de performance - Échelle de Zubrod (PRZ2, PRZ1, PRZ0)
5. PRE7: Douleur avant la chirurgie (T, F)
6. PRE8: Hémoptysie avant la chirurgie (T, F)
7. PRE9: Dyspnée avant la chirurgie (T, F)
8. PRE10: Toux avant la chirurgie (T, F)
9. PRE11: Faiblesse avant la chirurgie (T, F)
10. PRE14: T dans le TNM clinique - taille de la tumeur d'origine, de OC11 (plus petite) à OC14 (plus grande) (OC11, OC14, OC12, OC13)
11. PRE17: DM de type 2 - diabète sucré (T, F)
12. PRE19: IM jusqu'à 6 mois (T, F)
13. PRE25: PAD - maladies artérielles périphériques (T, F)
14. PRE30: Tabagisme (T , F)
15. PRE32: Asthme (T, F)
16. AGE: Âge à la chirurgie (numérique)
17. Risk1Y: période de survie de 1 an - (T) rue valeur si décédé (T, F)

	DGN	PRE4	PRE5	PRE6	PRE7	PRE8	PRE9	PRE10	PRE11	PRE14	PRE17	PRE19	PRE25	PRE30	PRE32	AGE
1	DGN2	2.88	2.16	PRZ1	F	F	F	T	T	OC14	F	F	F	T	F	60
2	DGN3	3.40	1.88	PRZ0	F	F	F	F	F	OC12	F	F	F	T	F	51
3	DGN3	2.76	2.08	PRZ1	F	F	F	T	F	OC11	F	F	F	T	F	59
4	DGN3	3.68	3.04	PRZ0	F	F	F	F	F	OC11	F	F	F	F	F	54
5	DGN3	2.44	0.96	PRZ2	F	T	F	T	T	OC11	F	F	F	T	F	73
6	DGN3	2.48	1.88	PRZ1	F	F	F	T	F	OC11	F	F	F	F	F	51
7	DGN3	4.36	3.28	PRZ1	F	F	F	T	F	OC12	T	F	F	T	F	59
8	DGN2	3.19	2.50	PRZ1	F	F	F	T	F	OC11	F	F	T	T	F	66
9	DGN3	3.16	2.64	PRZ2	F	F	F	T	T	OC11	F	F	F	T	F	68
10	DGN3	2.32	2.16	PRZ1	F	F	F	T	F	OC11	F	F	F	T	F	54
11	DGN3	2.56	2.32	PRZ0	F	T	F	T	F	OC12	F	F	F	F	F	60
12	DGN3	4.28	4.44	PRZ1	F	F	F	F	F	OC12	F	F	F	T	F	58
13	DGN3	3.00	2.36	PRZ1	F	F	F	T	T	OC11	F	F	F	T	F	68
14	DGN2	3.98	3.06	PRZ2	F	F	F	T	T	OC14	F	F	F	T	F	80
15	DGN3	1.96	1.40	PRZ1	F	F	F	T	F	OC11	F	F	F	T	F	77
16	DGN3	4.68	4.16	PRZ1	F	F	F	T	F	OC12	F	F	F	T	F	62
17	DGN2	2.21	1.88	PRZ0	F	T	F	F	F	OC12	F	F	F	T	F	56
18	DGN2	2.96	1.67	PRZ0	F	F	F	F	F	OC12	F	F	F	T	F	61

Figure1 : le jeu de données

Basic Statistics

Raw Counts

Name	Value
Rows	470
Columns	17
Discrete columns	14
Continuous columns	3
All missing columns	0
Missing observations	0
Complete Rows	470
Total observations	7,990
Memory allocation	48.2 Kb

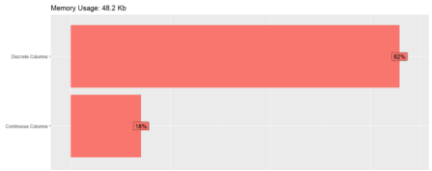


Figure2 : Statistique de Base pour notre jeu de données

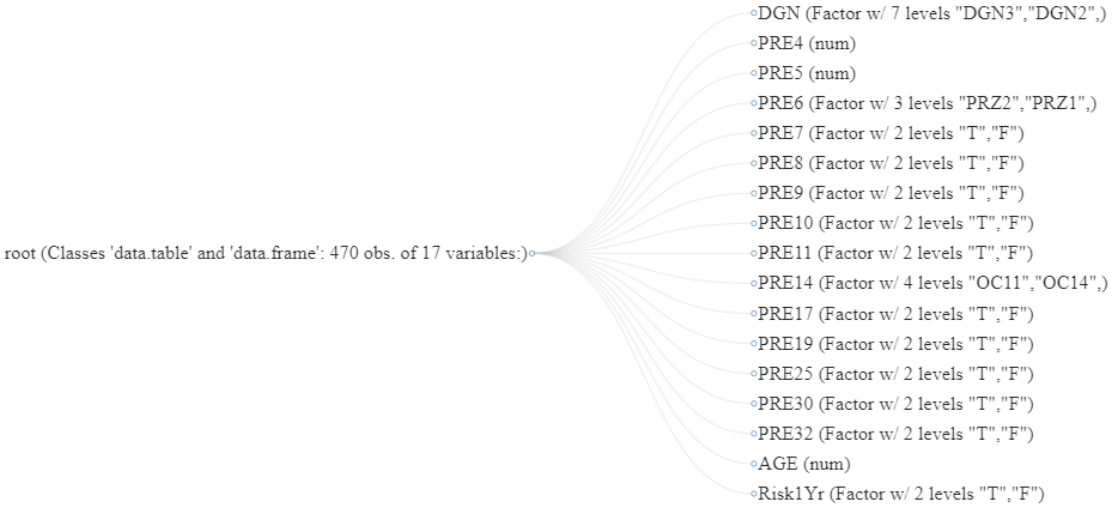


Figure3 : Structure de données

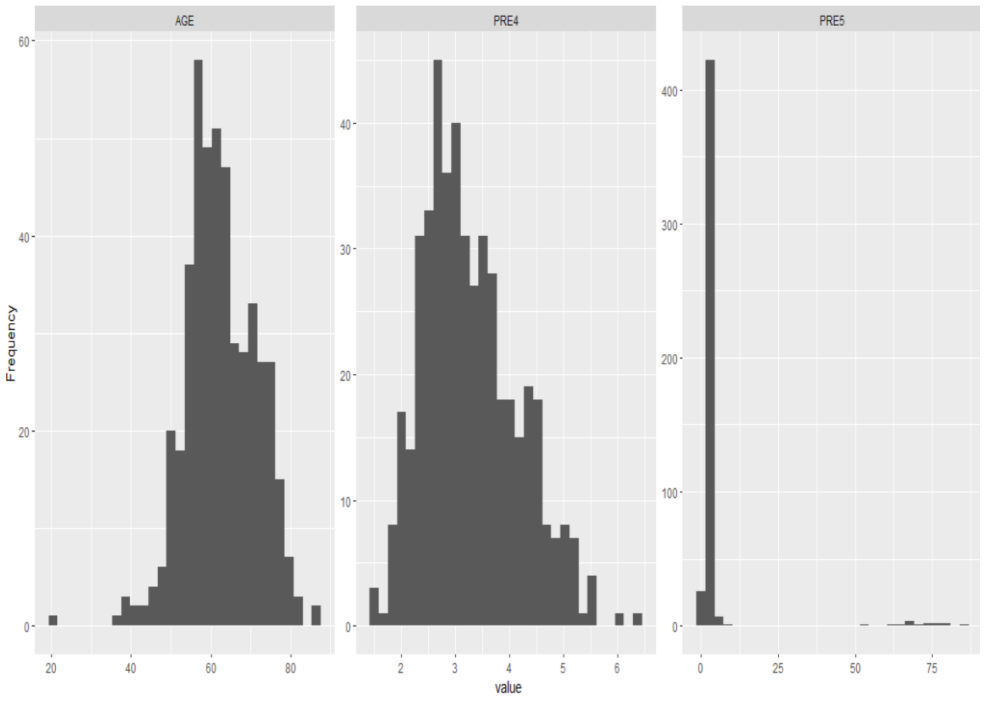


Figure4 : distributions des attributs (PRE5, PRE4, AGE)

Figure5 : distributions des attributs

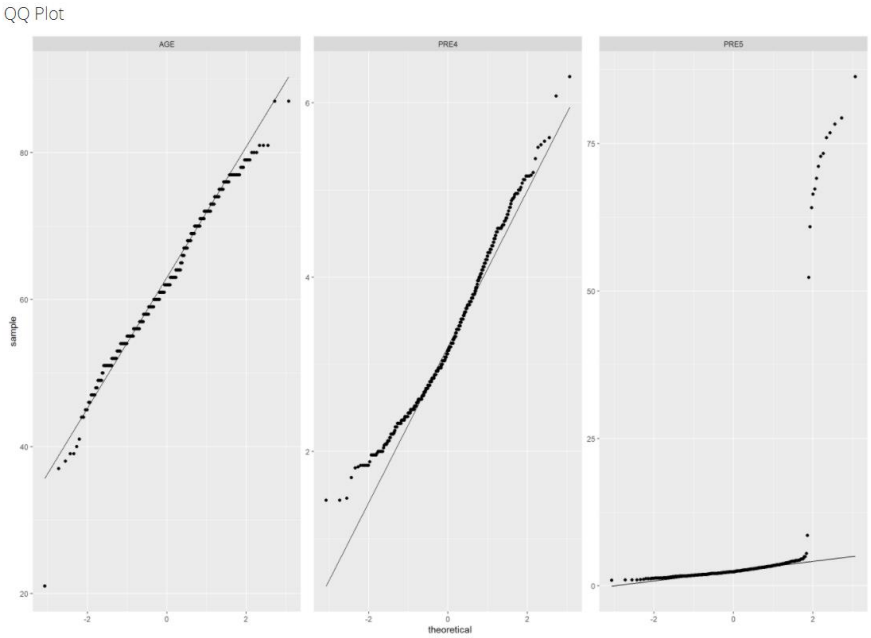


Figure6 : QQ PLOT des attributs (PRE4, PRE5, AGE)

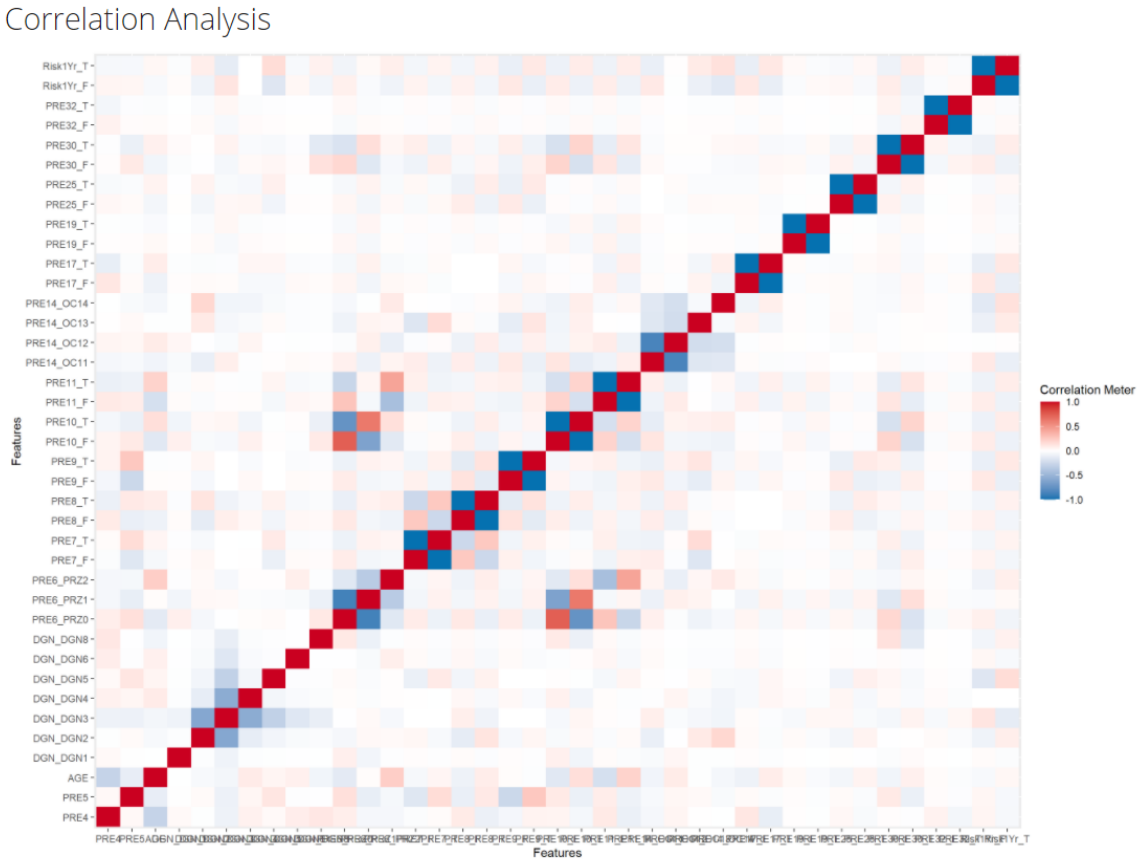


Figure7 : Matrice de corrélations

Hill climbing :

On remarque d'après la figure 8 que les attributs (PRE5, PRE4, PRE17, PRE11, PRE6, PRE10, PRE30, PRE7, PRE8, PRE9) sont en relation entre eux et particulièrement l'attribut PRE5 qui est influencé par (PRE30, PRE6, PRE11, PRE4).

On voit aussi que les attributs (PRE7, PRE8, PRE9) ont une relation de causalité entre eux.

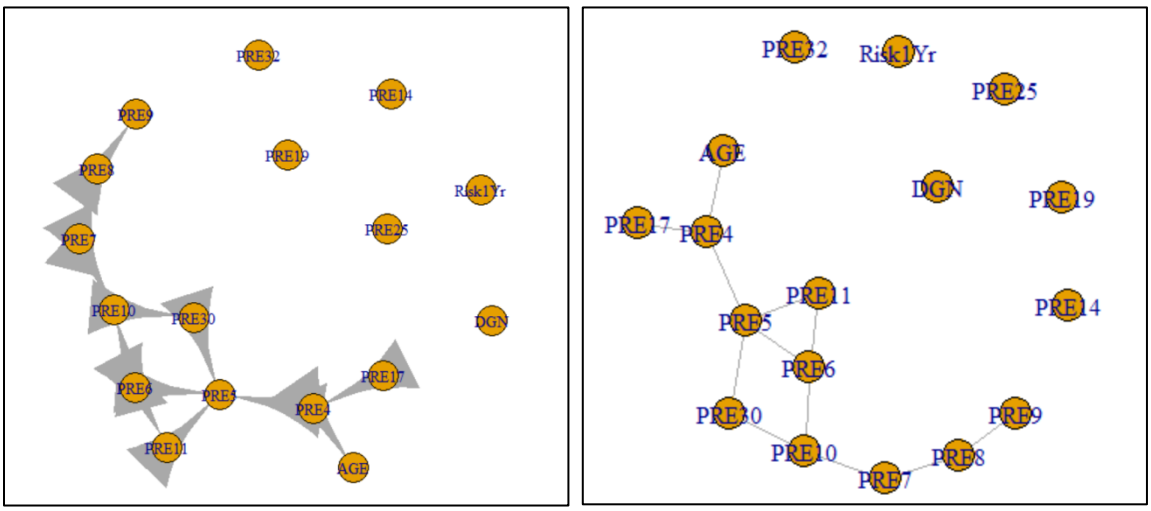
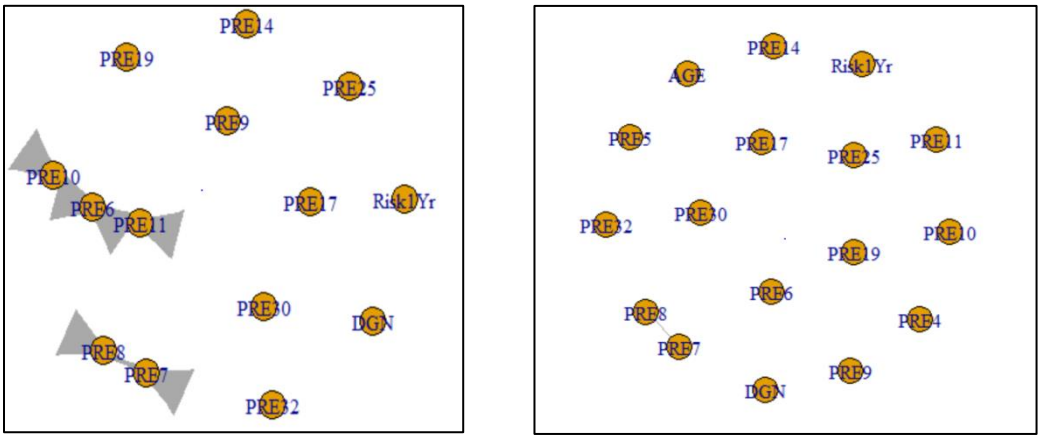


Figure8 : Graphes obtenu avec la méthode HC

PC :

On remarque que la méthode PC ne montre moins de relation de causalité que la méthode HC, ici on constate qu'ils existent des relations de causalités qu'entre les attributs (PRE10, PRE6, PRE11) d'un côté et les attributs (PRER7, PRE8) d'autre coté. Ces mêmes relations on les retrouve dans les résultats de la méthode HC.



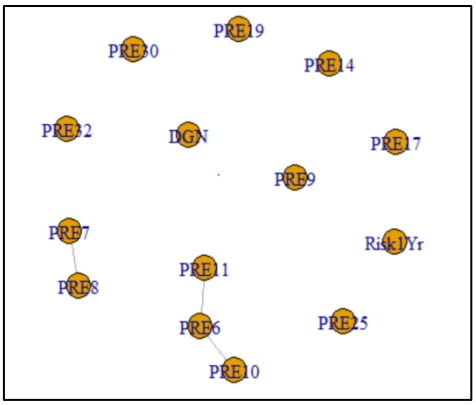


Figure 9 : Graphes obtenue avec la méthode PC

Aracne :

On remarque d'après la figure 10 (Graphes obtenu avec la méthode Aracne) qu'il y'a une relation de causalité mutuelle entre l'attribut PRE4 et les attributs (PRE32, PRE9, PRE30, PRE17, PRE25, AGE, DGN, Risk1Y, PRE19, PRE11, PRE5, PRE6), on constate aussi une relation entre PRE10 et PRE6 ainsi qu'une relation entre PRE5 et (PRE7, PRE8, PRE14).

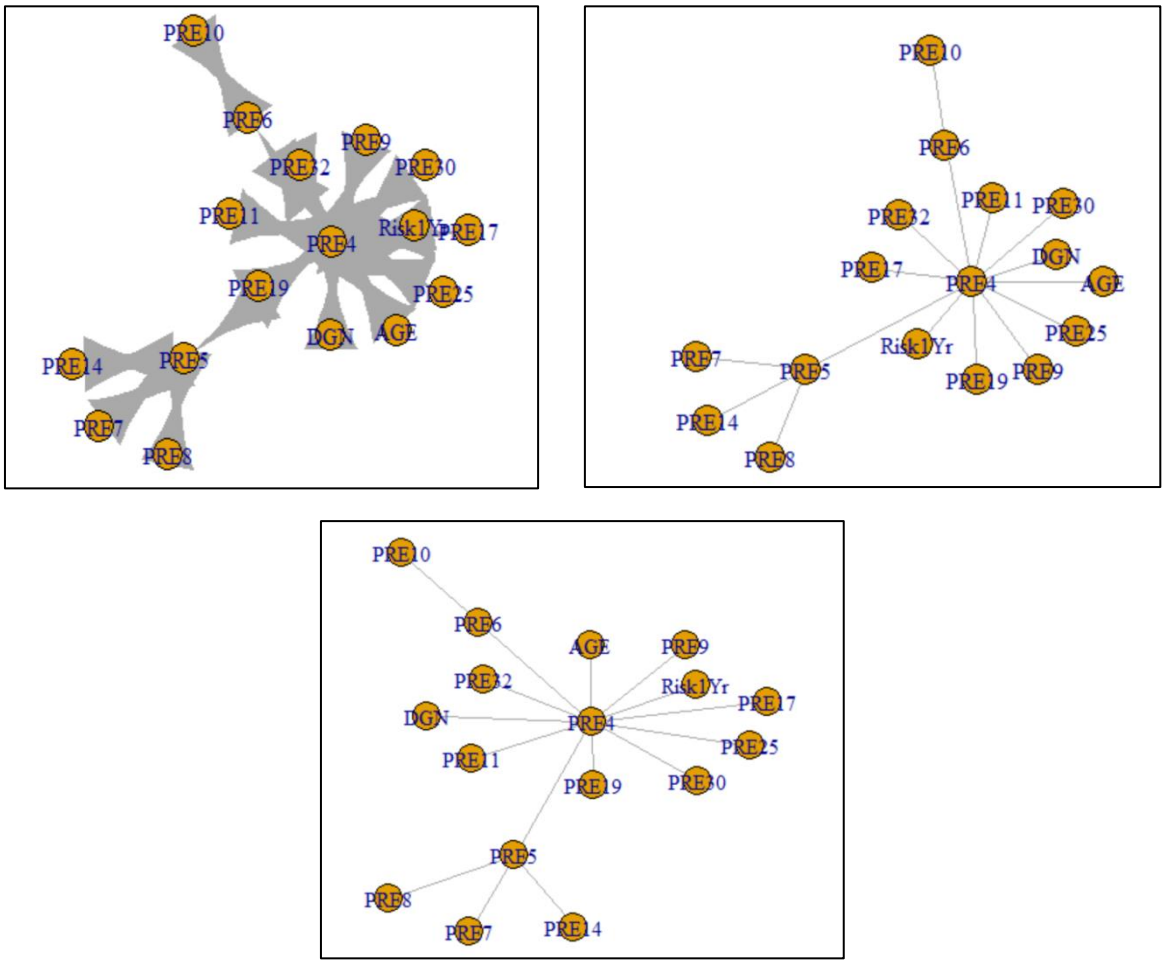


Figure10 : Graphes obtenu avec la méthode Aracne

MIIC :

On remarque d'après la figure 11 (résultats obtenus avec la méthode MIIC) qu'il existe une relation de causalité entre PRE5 et les attributs (PRE4, AGE, PRE9, PRE10), et une relation entre (PRE9, Risk1Y, PRE8), et une autre relation entre (PRE6, PRE10, PRE30).

De même on trouve une relation entre les attributs (PRE7, PRE9, PRE8) comme dans les résultats de PC et HC.

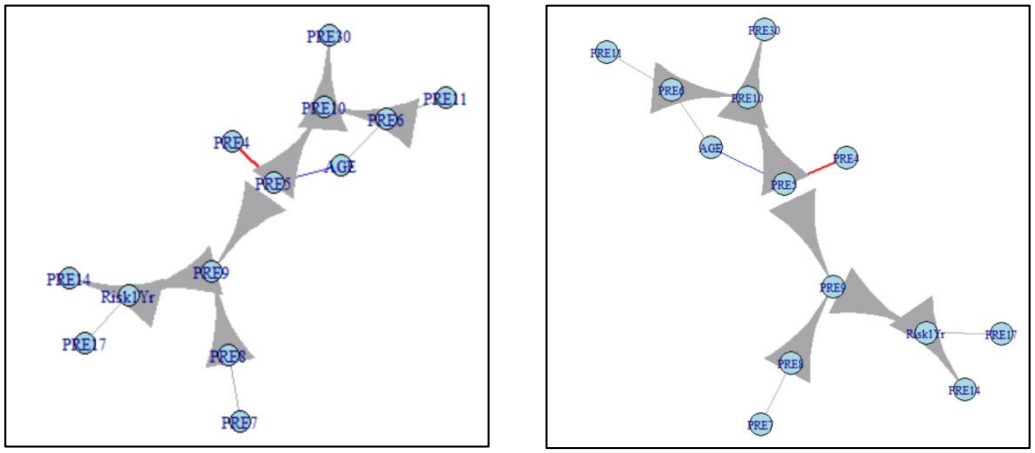


Figure11 : résultats obtenus avec la méthode MIIC

On remarque que la méthode Aracne fait resurgir plus de relations que les autres méthodes.

On remarque aussi que la méthode PC ne montre pas beaucoup de liens entre les attributs par rapport aux autres méthodes.

On constate aussi que la relations entre les attributs (PRE7, PRE8, PRE9) est présente dans les trois méthodes HC PC et MIIC.

5. Conclusion

La prise en charge des patients avec un cancer pulmonaire a connu aux cours des 20 dernières années des évolutions majeures, aussi bien dans le diagnostic qu'en matière de chirurgie, de radiothérapie, dans le domaine des traitements systémiques, de traitements combinés ou du dépistage. Plusieurs études se font pour augmenter l'espérance de vie d'un patient.

L'objectif de ce projet est l'exploitation des différents outils de reconstruction et d'analyse de réseaux afin d'étudier un système complexe à partir de données disponibles en ligne : un jeu de données rassemblant les observations réalisées sur 470 patients ayant subi une chirurgie pulmonaire entre 2007 et 2011.

On a effectué une étude comparative des résultats obtenus à partir des différents outils de Reconstruction (Aracne, constraint-based, search & score...) et comparé les résultats car la reconstruction d'un réseau mettant en jeu tous ces paramètres peut s'avérer important pour comprendre les relations indirectes entre eux.

6. Références

[1] Laetitia Jourdan," Métaheuristiques pour l'extraction de connaissances : Application à la génomique", HAL Id : tel-00007983 https://tel.archives-ouvertes.fr/tel-00007983, 2005.

[2] Louis Verny, "Apprentissage de réseaux causaux avec variables latentes et applications à des contextes génomiques et cliniques", HAL Id : tel-01896778 https://tel.archives-ouvertes.fr/tel-01896778, 2018.

[3] Boosted SVM for extracting rules from imbalanced data in application to prediction of the post-operative life expectancy in the lung cancer patients (Applied Soft Computing 2014, Zieba et al.)