Projets Réseaux et Graphes (M2 Machine Learning)

Nadia Radouani 21911973 FI & Amira Kouider 21904040 FI

1. Abstract

La cause principale de décès par cancer dans le monde est le cancer de poumon ; ce dernier est à l'origine de plus de décès que les cancers du sein, colorectal et du col de l'utérus réunis. Au moment du diagnostic, 70% des patients présentent une maladie avancée et seulement 15% restent en vie 5 ans après le diagnostic. Plusieurs facteurs tels que l'âge, le tabagisme, la fonction pulmonaire, le stade clinique et pathologique… influencent la survie des patients atteints de ce type de cancer. L'objectif de ce projet est de la reconstruction d'un réseau mettant en jeu tous les paramètres afin de comprendre les relations indirectes entre eux.

2. Introduction

Les poumons font partie de l'appareil respiratoire ; ils sont situés dans le thorax, de chaque côté du cœur. Les cellules du poumon subissent parfois des changements qui rendent leur mode de croissance ou leur comportement anormal. Ces changements peuvent engendrer la formation de tumeurs non cancéreuses (bénignes), comme elles peuvent causer le cancer (de poumon) appelé aussi cancer bronchique ou cancer broncho-pulmonaire.

Les traitements du cancer du poumon comprennent principalement la chimiothérapie (pour les patients atteints d'une maladie avancée ou métastatique) et la résection chirurgicale. Cette dernière n'étant pas sans risques inhérents, il pourrait être bénéfique pour les patients et les médecins d'avoir un aperçu des scénarios risques/bénéfices post-chirurgicaux attendus, dans la mesure où ils peuvent être liés aux antécédents de santé et aux caractéristiques de base du patient au moment du diagnostic et de l'opération. S'il existe un modèle à reconnaître dans les diverses caractéristiques des patients et des maladies, cela aiderait les médecins à prendre une décision plus éclairée sur la question de savoir si un patient donné est un candidat approprié pour une résection chirurgicale. En cas de risque élevé reconnu de décès d'un patient dans l'année suivant l'opération, des traitements alternatifs ou des soins palliatifs peuvent être préférés.

L'objectif principal de ce projet est de reconstruire un réseau permettant d'étudier la relation entre les caractéristiques de base des patients atteints de cancer de poumon et les facteurs influençant leurs survies après la chirurgie. Pour cela, une étude comparative à partir de différents outils de reconstruction ont été effectués : Correlation, Partial Correlation, Hill-climbing, PC approach et Aracne sur la base Thoracic Surgery Data. [Section 3]

3. Data

La base de données utilisée dans le projet est téléchargée depuis le site de l'UC Irvine Machine Learning Repository. Elle fait partie du registre national du cancer du poumon et contient des données compilées au centre de chirurgie thoracique de Wroclaw en Pologne.

Les données ont été collectées auprès de patients ayant subi une opération de résection pulmonaire pour un cancer primaire du poumon entre 2007 et 2011. Elles sont représentées comme suit : 470 lignes représentant les patients et 17 colonne représentant les caractéristiques, l'une de ces caractéristiques indique si le patient donné a vécu ou est décédé dans l'année qui suit l'opération, il s'agit de la variable cible, et les autres caractéristiques comprennent à la fois des variables catégorielles, telles que la toux avant l'opération, et des variables continues telles que la taille de la tumeur d'origine. Le problématique de la base réside dans les données très déséquilibrées : seuls 70 cas parmi 470, soit 14,9 % des données, sont associés à l'étiquette positive (décès dans l'année). Ce problème est résolu avec un algorithme de bootstrapping, SMOTE (Synthetic Minority Over-sampling Technique).

Le deuxième défi réside dans le nombre de variables disponibles : pour effectuer une analyse approfondie des données et des réseaux construits, un certain nombre de nouvelles caractéristiques sont créées pour mieux représenter les relations entre les différentes caractéristiques de l'ensemble de données.

4. Analyse exploratoire

Etant la base de données non équilibrée, Seuls 70 des 470 patients, soit 14,9 % des observations, sont décédés un an après la chirurgie. (Figure 1)

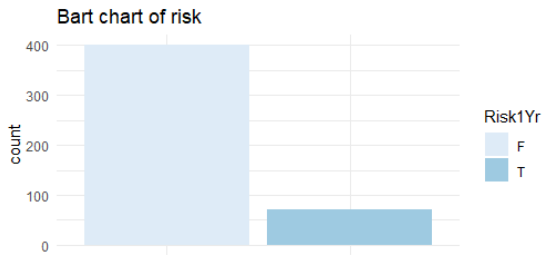

Figure 1 - Distribution des patients qui ont survécu (ou pas) un an après la chirurgie

La variable DGN (recodé en Daignosis) représente le type de diagnostic des patients. D'après les figures 2, on remarque que DGN3 a un grand impact sur les patients qui sont décédés un an après la chirurgie.

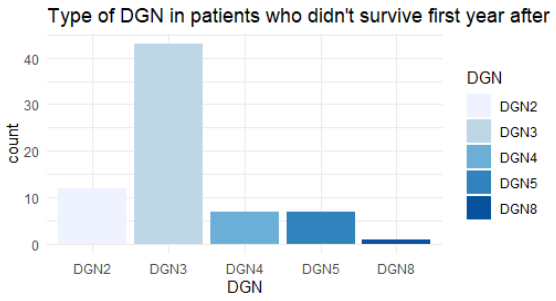

Figure 2 - Le type de DGN avec le plus grand impact sur les patients qui n'ont pas survécu longtemps après la chirurgie

Un déséquilibre est constaté pour toutes les variables observées notamment dans les variables factors binaires (Figure 3).

Distribution of binary variables

	Risk1Yr	PRE7	PRE8	PRE9	PRE10	PRE11	PRE17	PRE19	PRE25	PRE30	PRE32
F	400	439	402	439	147	392	435	468	462	84	468
T	70	31	68	31	323	78	35	2	8	386	2

Figure 3 - Distribution des données binaires

La base ne contient aucune valeur nulle NA mais contient des valeurs abérrantes dans les variables non catégorielles notamment dans les variables Forced_Expiration et Age (Figure 4).

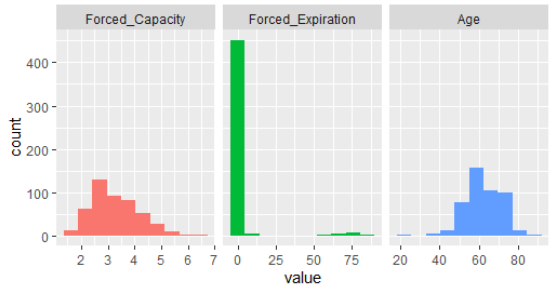

Figure 4 - Distribution des variables contenant des outliers

Cela a même été vérifié en utilisant la méthode Tukey qui confirme avoir 15 outliers dans la variable Forced_Expiration et 1 outlier dans la variable Age. Ce problème à été résolu en supprimant les lignes contenant ces valeurs car elles ne sont pas plausibles dans la réalité et peuvent également avoir un impact négatif sur les performances des modèles appliqués dans la suite. Il reste donc 454 observations dans la base de données, et la distribution des variables sont maintenant légèrement biaisées mais proches de la distribution normale (Figure 5).

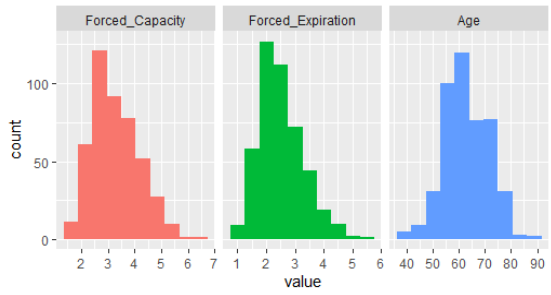

Figure 5 - Distribution des variables sans les outliers

La base contient :

- 3 variables numériques.
- 11 variables binaires catégorielles apparaissant comme « *T/F* » qui ont été recodé en nombre binaires où chaque *F* été converti en 0 et chaque *T* en 1. Ce recodage s'avère nécessaire puisque nombreux algorithmes exigent que les variables d'entrée et de sortie soient numériques.
- 3 variables ordinales/nominales catégorielles : *Zubrod_scale avec* 3 valeurs possibles (PRZ0, PRZ1, PRZ2), *Tumor_size* avec 4 valeurs possibles (OC11, OC12, OC13, OC14) et *Daignosis* avec 7 valeurs possibles (DGN1, DGN2, DGN3, DGN4, DGN5, DGN6, DGN8). Ces 3 variables ont été recodé, en utilisant le one-hot encoding, en plusieurs variables portant le même nom où chacune est représentée par 0 ou 1.

5. Data balancing

Pour l'équilibrage des données, on a utilisé l'algorithme SMOTE. Cet algorithme parcourt toutes les observations de la classe minoritaire, cherche ses k plus proches voisins puis synthétise aléatoirement de nouvelles données entre ces deux points. Notre variable cible est «*Risk1Y*» qui, pour rappel, été renommer en «*Death*» et qui indique si le patient donné a vécu ou est décédé dans l'année qui suit l'opération. Après l'équilibrage, les données sont réparties comme suit : 276 Alive contre 207 Dead. Une dernière vérification sur cette base consiste à supprimer les variables constantes, avant qu'elle soit utilisée pour la reconstruction des réseaux.

6. Correlation Network

Afin d'effectuer un réseau de corrélation, le calcul des corrélations entre les variables est nécessaire (Figure 6).

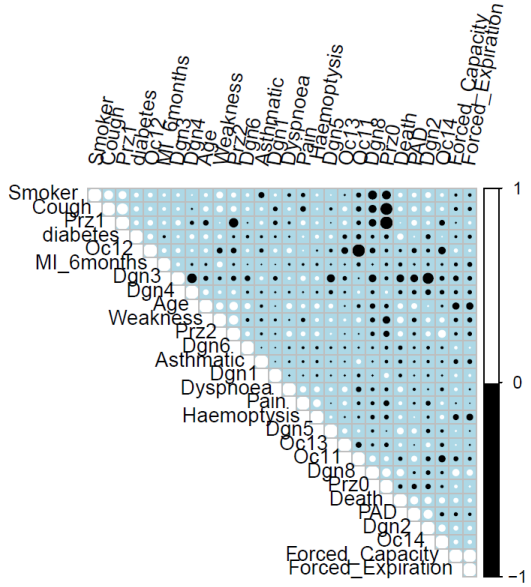

Figure 6 - Corrélogramme

D'après le corrélogramme, on peut repérer certaines corrélations positives fortes ou plus ou moins fortes : entre les variables *Forced_Expiration* (FVC) et *Forced_Capacity* (FEV), *Caugh* (PRE10) et *Smoker* (PRE30), *Prz1* et *Caugh*, *Weakness* et *Smoker*, *Weakness*

et *Caugh*, *Pain* et *Haemoptysis*, *OC14* et *Prz2*, *Dgn8* et *Prz0*.

Et d'autres corrélations négatives fortes ou plus ou moins fortes : entre les variables *Prz0* et respectivement *Caugh*, *Prz1 et Smoker*, *Oc11* et *Oc12*, *Dgn8* et *respectivement Smoker*, *Caugh et Prz1*, *Age* et respectivement *Forced_Expiration et Forced_Capacity.*

Le tableau ci-dessous affiche les corrélations les plus fortes :

	Var1	Var2	Freq
1	Forced_Capacity	Forced_Expiration	0.87
2	Prz0	Cough	-0.74
3	Prz1	Cough	0.60
4	Dgn2	Dgn3	-0.62
5	Prz1	Prz0	-0.82
6	Oc12	Oc11	-0.81

Tableau 1 - les corrélations les plus fortes

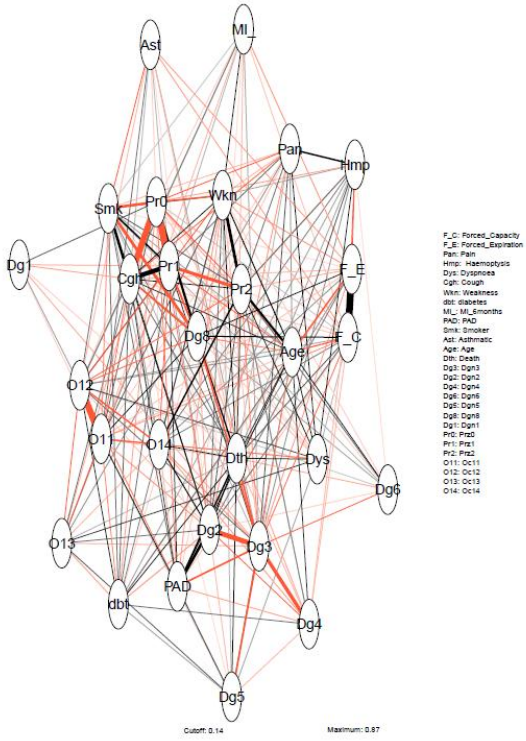

Figure 7 - Network correlation

A partir de ce réseau, on peut repérer des relations entre les variables : une forte relation positive entre les deux variables Forced_Capacity et Forced_Expiration, ce qui nous amène à conclure qu'il y a une relation entre la capacité vitale forcée et le volume qui a été exhalé à la fin de la première seconde d'expiration forcée chez le patient , on remarque aussi qu'il y a une forte relation négative entre l'état des performances à l'échelle Prz0 et la toux avant l'opération (Caugh).

D'autres relations plus ou moins fortes repérées entre la faiblesse du patient avant l'opération et l'état de performance à l'échelle Prz2, entre la présence de la toux et le fait que le patient soit un fumeur ou non, entre la présence des maladies artérielles périphériques et le diagnostique de type DGN2, une relation positive entre celui là et l'espérance de vie du patient après l'opération, et négative entre cette dernière caractéristique et le diagnostique de type DGN3, enfin une relation positive entre l'âge au moment de l'opération et l'état des performances à l'échelle Prz2.

7. Partial Correlation Network

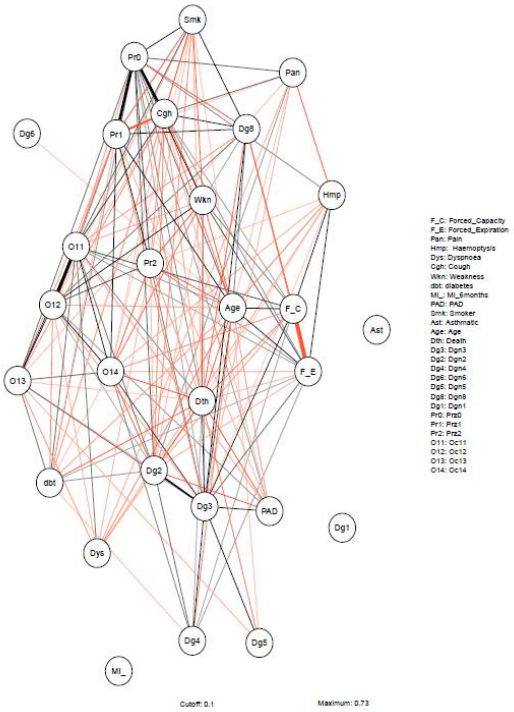

Figure 8 - Partial correlation network

Le réseau construit à partir des corrélations partielles relève les mêmes relations repérées précédemment. On remarque aussi qu'il affiche des caractéristiques n'ayant aucune relation avec les autres, il s'agit de la caractéristique de l'asthme, le MI jusqu'à 6 mois et le diagnostique de type DGN1 chez les patients.

Les réseaux construits à partir des corrélations et les corrélations partielles nous ont indiqués la présence des relations intéressantes entre les caractéristiques. Cependant pour mieux comprendre ces relations, on va

utiliser d'autres outils de reconstruction : Hill Climbing, Aracne et PC approach.

8. Hill-Climbing

Appelé aussi *score-based method*, *Hill-Climbing est* un algorithme de recherche locale qui se déplace continuellement dans le sens d'une augmentation de la valeur pour trouver la meilleure solution au problème. Il se termine lorsqu'il atteint une valeur de pic où aucun voisin n'a une valeur supérieure. C'est la variante de la méthode *Generate and Test*, qui produit un retour d'information aidant à décider de la direction à prendre dans l'espace de recherche. Il est caractérisé comme étant une approche gourmande, car, l'algorithme se déplace dans la direction qui optimise le coût, et il ne fait pas de retour en arrière dans l'espace de recherche, car il ne se souvient pas des états précédents.

```
nodes:                                  28
arcs:                                   64
  undirected arcs:                      0
  directed arcs:                        64
average markov blanket size:            8.00
average neighbourhood size:             4.57
average branching factor:               2.29

learning algorithm:                     Hill-Climbing
score:                                  BIC (Gauss.)
penalization coefficient:               3.090008
tests used in the learning procedure:   2210
optimized:                              TRUE
```

Figure 9 - Aperçu du modèle HC

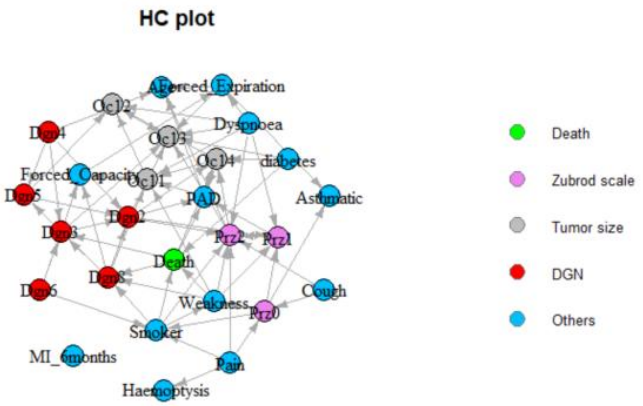

Figure 10 – Réseau obtenu de la méthode HC

On peut voir que la variable *Death* est affectée par les variables *Smoker*, *Weakness* et *Diabetes*. Cela semble relativement évident puisque le diabète et le tabagisme peut causer le décès d'un patient, ainsi, tous les patients atteints de cancer de poumon présentent des symptômes telle que la fatigue. On peut constater une autre relation entre *dyspnoea* et *Asthmatic*, une relation assez évidente puisque un patient souffrant de dyspnée peut souffrir de l'asthme dans l'avenir.

On constate encore des relations entre *Zubrod_scale* et différentes autres variables : Sachant que le score de Zubrod est un système de notation utilisé pour évaluer l'état de santé des patients ; on peut déduire qu'un patient souffrant de dyspnée, de fatigue et ayant mal est considéré comme patient symptomatique.

Dyspnoea	1.0000000
Dgn2	0.9113831
Weakness	0.6593793
diabetes	0.4768666
Prz2	0.4658471

Figure 11 - Top 5 of hubs

On constate que Dyspnoea est le plus grand hub dans le graphe, on peut donc dire que les gênes respiratoires ressenties par le patient est le symptôme ayant le pus de relations avec les autres caractéristique.

Dgn2	37.64286
Oc11	36.18571
Forced_Capacity	34.83333
Dgn8	33.89524
Oc14	25.10000
Prz2	22.91190
Forced_Expiration	21.33333
PAD	18.41667
Prz0	18.23333
Death	17.53333

Figure 12 - top 10 des noeuds en termes de betwenness

En termes de betweenness centrality le type de diagnostique DGN2 se positionne dans le top 10 des nœuds, suivi par la taille de la tumeur d'origine OC11 (la taille la plus petite) et par La capacité vitale forcée. On en déduit que ces trois caractéristiques peuvent avoir une influence considérable au sein du réseau construit en raison de leurs contrôles sur les informations qui passent entre les autres.

	From	To
1	Forced_Capacity	Forced_Expiration
2	Oc11	Forced_Capacity
3	Oc14	Oc11
4	PAD	Oc11
5	Forced_Expiration	Asthmatic
6	Prz0	Dgn8
7	Prz2	Oc14
8	Dgn8	Forced_Capacity
9	Dgn3	Dgn2
10	Dgn2	PAD

Figure 13 - Top 10 des arêtes en termes de betweenness

L'arête **Forced_Capacity -> Forced_Expiration** se positionne dans le top 10 des arêtes en termes de betweenness centrality. Par ailleurs, on remarque que **OC11**, **Forced_Capacity**, **Forced_Expiration** et **Asthmatic** forment une de chaine.

9. Aracne

Appelé aussi *Local search method*, *Aracne* commence par un graphe complet, il estime l'information mutuelle entre deux noeuds et la compare à un seuil, ensuite il supprime les liens inférieurs à ce seuil. C'est une approche qui s'applique sur un squelette de graphe, il ne prend pas en considération l'orientation. Ceci ne permet pas de savoir qui contrôle qui, et peut conduire à une suppression de vrais liens.

```
Bayesian network learned via Pairwise Mutual Information methods

model:
  [undirected graph]
nodes:                                28
arcs:                                 47
  undirected arcs:                    47
  directed arcs:                      0
average markov blanket size:          3.36
average neighbourhood size:           3.36
average branching factor:             0.00

learning algorithm:                   ARACNE
mutual information estimator:         Maximum Likelihood (Gauss.)
tests used in the learning procedure: 378
```

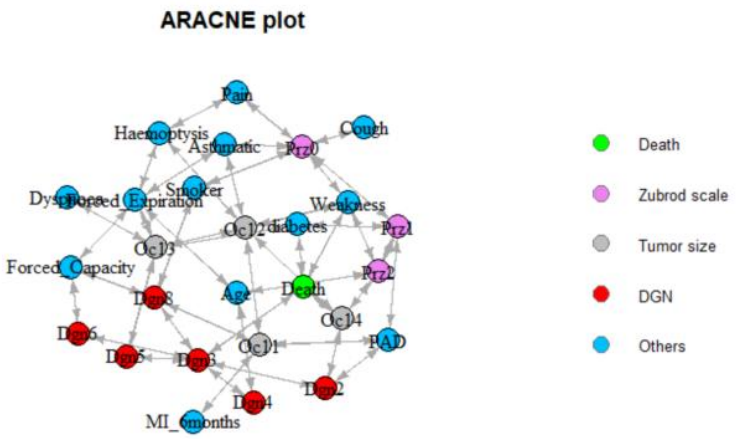

Figure 15 - Réseau obtenu de la méthode ARACNE

La méthode *ARACNE* permet d'avoir un graph avec des arêtes bidirectionnelles. On peut remarquer que la variable *Death* par exemple est affectée par *Age*, *Diabetes*, *Weakness*, *Zubrod_scale*, *Tumor_size*, *DGN*. Ce qui peut être traduit comme suit : un patient peut décéder si il est âgé, souffre de diabète et de fatigue, a un score Zubrod élevé(*Prz2*), a une tumeur de taille grande(*Oc14*) avec un diagnostic *Dgn3*. Un patient avec un score Zubrod bas (*Prz0*) ayant mal, et souffrant de toux est à la base un fumeur.

Un patient avec une tumeur de grande taille est forcément symptomatique et son diagnostic est de niveau 2 (*Dgn2*), score Zubrod élevé, relation entre *Oc14* et *Prz2*.

Un patient avec une tumeur de petite taille (Oc11) peut souffrir d'une crise cardiaque (*MI_6months*) ou de maladie artérielle périphérique (*PAD*) a un diagnostic *Dgn8* tout comme une personne fumeur. Ce patient peut être considéré comme symptomatique mais ambulatoire (*Prz1*).

Oc12	1.0000000
Weakness	0.7930382
Prz0	0.7302697
Oc14	0.7228961
Oc13	0.7013062

Figure 16 - Top 5 of hubs

On constate que OC12 est le plus grand hub dans le graphe, on en déduit qu'une tumeur de taille OC12 chez les patients est la caractéristique ayant le plus de relations avec les autres. (Figure 16)

Cette caractéristique se positionne aussi dans le top 10 des nœuds en termes de betweenness. (Figure 17)

Oc12	121.20274
Prz0	108.22951
Dgn3	105.24921
Oc13	96.76840
Forced_Expiration	85.44235
Oc11	76.35722
Dgn8	66.76768
Prz1	54.89524
Weakness	46.36284
Death	40.65657

Figure 17 - Top 10 des nœuds en termes de betweeness

On déduit que OC12 joue un rôle important et peut avoir une influence considérable au sein du réseau construit.

10. PC Approach

L'algorithme vérifie les noeuds deux à deux, et refait le test en conditionnant sur une variable tierce et en grandissant, en plus en plus, son ensemble du conditionnement. Il a l'avantage de pouvoir donner des orientations exprimant les causalités mais son inconvénient réside dans le fait que le test d'indépendance est sensible au bruit des données.

Pour appliquer la méthode PC, un recodage des variables continues : Age, Forced_Capacity et Forced_Expiration en variables discrètes est nécessaires. On a recodé de telle sorte à avoir des classes équilibrées : 4 classes pour chacune de ces trois variables, où chaque classe contient à peu près 113 observations.

```
> pc_model
Object of class 'pcAlgo', from Call:
pc(suffStat = suffStat, indepTest = disCItest, alpha = 0.5, labels = colnames(dataPC))
Number of undirected edges:  3
Number of directed edges:    15
Total number of edges:       18
```

Figure 18 - Aperçu du modèle PC

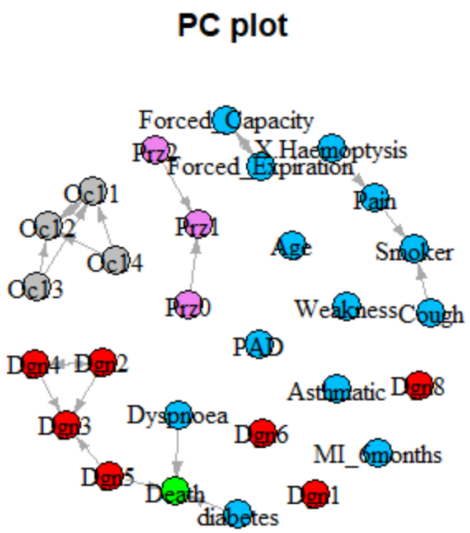

Figure 19 - Réseau obtenu de la méthode PC

On remarque que l'espérance de vie du patient après l'intervention chirurgicale est affectée par 3 caractéristiques : le type de diagnostique DGN5, la présence de la dyspnée avant l'intervention et le fait que le patient soit diabétique ou non. On peut donc dire que ces trois caractéristiques jouent un rôle important dans la mort ou la survie d'un patient un an après l'intervention.

La présence de la toux et de la douleur chez le patient affecte la caractéristique du fumeur: un patient fumeur est plus susceptible d'avoir la toux et les douleurs.

La caractéristique Pain est affectée par Haemoptysis : un patient souffrant de l'hémoptysie et qui crache du sang va forcement avoir plus de douleurs.

Les deux caractéristiques Forced_Capacity et Forced_Expiration s'affectent mutuellement, ce qui nous amène à conclure qu'il y a une relation mutuelle entre la capacité vitale forcée et le volume qui a été exhalé à la fin de la première seconde d'expiration forcée chez le patient.

Par ailleurs, l'approche PC a supprimé plusieurs relations qu'on a pu avoir précédemment en utilisant les approches HC et Aracne.

	From	To
1	Prz0	Prz1
2	Weakness	Prz0
3	Prz1	Prz2
4	Forced_Capacity	Forced_Expiration
5	Forced_Expiration	Forced_Capacity
6	Pain	X.Haemoptysis
7	X.Haemoptysis	Pain
8	Weakness	Prz2
9	Dgn2	Dgn3
10	Dgn4	Dgn3

Figure 20 - Top 10 des arêtes en termes de betweenness

En analysant le top 10 des arêtes en termes de betweenness centrality, on constate la présence d'arêtes qui peuvent nous intéresser dans ce classement, il s'agit de **Weakness -> Prz0**, et **Forced_Capacity -> Forced_Expiration.**

11. Conclusion

En utilisant différentes méthodes de reconstruction de réseaux, on a pu repérer des relations intéressantes entre les variables, comprenant ainsi les différents facteurs qui entre en jeu en termes d'espérance de vie post opératoire des patients atteints d'un cancer de poumon.

12. Références

[1] Boosted SVM for extracting rules from imbalanced data inapplication to prediction of the post-operative life expectancyin the lung cancer patientsMaciej Zieba, Jakub M. Tomczak, Marek Lubicz, Jerzy Swiatek

[2] An effective structure learning method for constructing gene networks. Xue-wen Chen, Gopalakrishna Anantha, Xinkun Wang .