# Projet d'Apprentissage Statistique
# Sujet 7 : facteurs de risque du cancer du col de l'utérus

Groupe 9

AGUIRRE Paula, KUEHM Timothé, MONTMEAT Florian et NAHAPETYAN Knarik

M1 Mathématiques appliquées

Université Paris Dauphine

8 juin 2018

## Table des matières

# Introduction

*Note au lecteur ou lectrice :*
*Bonjour, nous sommes au courant que le compte rendu de ce projet ne devait -selon les consignes- pas excéder les* 5 *pages. Cependant, nous avons tenu à exposer le plus possible nos résultats, nos commentaires ainsi que nos avis sur les divers résultats obtenus et méthodes appliquées. De plus, nous avons décidé d'utiliser quatre méthodes différentes sur deux échantillons chacune. Nous n'avons pas souhaité en enlever car elles sont, selon nous, toutes intéressantes à traiter. Pour en améliorer la lisibilité et la compréhension, nous avons essayé d'introduire le plus de graphiques et d'explications possibles. De plus, ce projet nous permet de laisser notre plume aller et venir au gré des lignes laissant notre coté littéraire s'exprimer un peu. Nous n'avons donc pas souhaité le limiter et avons écrit légèrement plus de texte que nous aurions dû. Mais hors graphique et autres tableaux, nous obtenons environ sept pages.*
*Nous vous souhaitons désormais tous les quatre une bonne lecture !*

L'objectif de ce projet est de trouver la meilleure méthode pour prédire si une patiente a une grande probabilité d'être atteinte d'un cancer du col de l'utérus et donc de faire les tests de dépistages. Nous cherchons donc à déterminer le profil "type" de la femme présentant un fort risque d'être touchée par le cancer du col de l'utérus.

Pour trouver cette méthode, nous disposons d'une base de données de 858 patientes qui nous donne des informations sur leurs habitudes de vie (par exemple fumeuse, port du stérilet, etc...) ou sur des données médicales (résultats de dépistages ou de diagnostics).

Nous disposons également de résultats de plusieurs stratégies de dépistages du cancer du col de l'utérus, à savoir le test de Hinselmann [1], le test de Schiller [2], la cytologie [3] ou encore la biopsie [4].

Nous avons donc ici affaire à un problème d'apprentissage supervisé et plus précisément un problème de classification supervisée.

Notre objectif consiste à mettre en oeuvre différentes méthodes de prédiction puis les comparer afin de trouver laquelle est la mieux adaptée à notre cas.

1. Le test de Hinselmann consiste à faire un examen colposcopique du col utérin, c'est à dire l'observation des caractéristiques de l'épithélium du col utérin après l'application en trois étapes successives de sérum physiologique, d'acide acétique dilué entre 3 et 5%, et enfin du soluté de Lugol.
2. Le test de Schiller est un test médical où une solution iodée est appliquée sur le col de l'utérus, afin de diagnostiquer un cancer de ce dernier.
3. La cytologie consiste à étudier des cellules isolées, ici celles du col utérin.
4. La biopsie consiste à prélever une très petite partie d'un organe dans le but de les examiner, ici l'utérus.

# 1 Présentation des données

Commençons par présenter la base de données que nous avons. L'étude est menée sur 858 femmes.
Nous avons à disposition 32 variables explicatives que l'on regroupe en deux catégories :

| Entier | Binaire |
|---|---|
| Âge | Fumeuse |
| Nombre de partenaire(s) sexuel(s) | Contraception hormonale |
| Âge du premier rapport sexuel | Port du stérilet |
| Nombre de grossesse(s) | MST |
| Nombre d'années de tabagisme | Diverses MST (nous ne précisons pas ici car peu d'importance) |
| Nombre de paquet(s) de cigarettes par an | Diagnostic de cancer |
| Nombre d'années de contraception hormonale | Diagnostic de CIN |
| Nombre d'années du port du stérilet | Diagnostic de HPV |
| Nombre de MST | Diagnostic "Dx" |
| Nombre de MST diagnostiquées | |
| Nombre d'années depuis le premier diagnostic d'une MST | |
| Nombre d'années depuis le dernier diagnostic d'une MST | |

FIGURE 1 – Variables explicatives

Comme expliqué dans l'introduction, nous disposons des résultats de quatre tests permettant de détecter le cancer de l'utérus qui sont les tests de Hinselmann, de Schiller, la cytologie et la biopsie. Nous avons réfléchi sur comment exploiter ses données et nous avons eu plusieurs idées :

- S'intéresser aux résultats d'un des tests comme variable réponse ?
- S'intéresser uniquement aux personnes dont les quatre tests donnaient le même résultat ?
- Faire une moyenne des résultats des quatre tests et traiter cette nouvelle variable comme variable réponse ?

Nous avons choisi de favoriser la dernière option. En effet, la première idée forçait la comparaison de toutes les méthodes que l'on verra après sur chacun des tests, ce qui la rend extrêmement coûteuse en terme de temps, et surtout, qui n'exploite pas la totalité des données dont nous disposons.

La seconde restreignait beaucoup les données : en effet, il est fréquent que les quatre tests soient négatifs (756 individus sur 858, soit 88.11%), mais il est assez rare que les tests soient tous positifs (6 individus sur 858, soit seulement 0.70% de notre échantillon !). Choisir cette idée aurait donc encore plus accentué le déséquilibre de notre échantillon et l'aurait rendu encore plus difficile à analyser.

Nous avons donc choisi d'utiliser comme variable d'intérêt la variable binaire répondant à la question :

**"La patiente a-t-elle au moins deux tests positifs ?"**.

Notre but va donc être de prédire au mieux cette variable. Évidemment, pour nos tests, nous considérerons que si cette variable vaut 1, la patiente est effectivement atteinte d'un cancer du col de l'utérus car cela signifie qu'au moins deux tests sur quatre sont sortis positifs. Au contraire, nous considérerons qu'une patiente n'a pas le cancer si un seul test est ressorti positif ou si tous les tests se sont avérés négatifs.

**Critiques :** Tout d'abord, rappelons que nous travaillons sur un échantillon relativement petit de données. En effet, bien que nous puissions l'exploiter, un échantillon de seulement 858 patientes reste peu représentatif de la population.

Nous souhaitons dès à présent critiquer le choix de la variable d'intérêt. Comme dit ci-dessus, nous aurions pu la choisir de plusieurs manières différentes. Nous avons considéré ici que deux tests positifs étaient suffisants pour dire qu'une patiente était effectivement atteinte du cancer. Cependant, cela est très largement critiquable. En effet, pouvons-nous vraiment supposer qu'une patiente ayant deux tests positifs et deux tests négatifs a plus de chance d'avoir effectivement un cancer que de ne pas en avoir ? Nous admettrons ici que oui mais nous sommes conscient que cela peut être faux.

## 1.1 Données manquantes

Nous devons également faire quelques remarques sur l'absence de certaines valeurs dans notre base de données. En effet, certaines des patientes n'ont pas répondu à toutes les questions pour des raisons de confidentialité.

R nous apprend que $11,72\%$ des données sont manquantes, et que seules 59 des 858 personnes ont répondu à toutes les questions. Ainsi, éliminer toutes les personnes n'ayant pas répondu à toutes les questions n'est pas envisageable.

Concernant les variables explicatives, les seules avec aucune donnée manquante sont l'âge et le nombre de MST. Inversement, celles avec le plus de données manquantes sont le nombre d'années depuis le premier/dernier diagnostic de MST car, bien évidemment, seules les femmes ayant déjà fait un diagnostic ont pu y répondre ! De même, 117 personnes n'ont pas répondu aux questions concernant le stérilet.

Les autres variables -autres que celles citées ci dessus- ont au plus $13,63\%$ de données manquantes. Nous ne retirerons donc a priori aucune variable. Lorsque cela était nécessaire, nous avons remplacé les valeurs manquantes par la moyenne de la variable en question.

**Critiques :** Nous venons de remplacer les valeurs manquantes par la valeur de la moyenne de la variable. Ceci pourrait être dangereux. En effet, en faisant cela, nous modifions la matrice de variance-covariance et réduisons ainsi artificiellement l'écart-type. Cependant, pour notre étude réduire l'écart-type ne nous pose pas de problème car les classifications que nous allons faire sont indépendantes de ce dernier. De plus, en procédant ainsi, nous ne créons pas de biais supplémentaire, ce qui nous permet de rester fidèles à l'échantillon original.

## 1.2 Inégalité de représentation des classes

Nous avons eu un problème directement lié à l'asymétrie des classes. Nous disposons de deux classes : les individus ayant le cancer et ceux ne l'ayant pas. Nos données présentent effectivement un déséquilibre entre ces deux classes : 92% de l'échantillon n'est pas atteint de cancer et forme une classe majoritaire tandis que seulement 8% en est atteint et forme ainsi une classe minoritaire.

Ce déséquilibre pose un problème lors de la minimisation de l'erreur de classification. Un classifieur simple à calculer et d'erreur faible est le classifieur constant égal à 0 -classifieur trivial- puisque son erreur est d'à peine 8% ! Cependant cette prédiction est inutile d'un point de vue pratique car elle n'apporte aucune information.

Nous avons envisagé plusieurs solutions :

- Pondérer différemment l'erreur sur les individus selon leur classe d'appartenance
- Modifier l'échantillon afin de rééquilibrer les classes

Nous avons choisi la deuxième solution car elle nous paraissait plus facile à mettre en oeuvre dans R grâce à la fonction Smote. De plus, réaliser une classification sans rééquilibrer l'échantillon va amener à la prédiction quasi-totale du résultat "non atteint de cancer".

Nous allons donc rééquilibrer notre échantillon. Cependant il faut rester vigilant lors de cette étape. Expliquons brièvement ce qu'est le sur et sous-échantillonage :

- Le sur-échantillonage consiste à ajouter "artificiellement" des individus à la classe minoritaire.
- le sous-échantillonage consiste à supprimer des individus de la classe majoritaire afin d'équilibrer les deux classes. Cette méthode enlève de l'information et rendrait notre échantillon trop petit pour être significatif (environ 140 individus au lieu des 858 du début)

Nous avons donc choisi de corriger la distribution asymétrique des données en réalisant un sur-échantillonage par la méthode Smote. Pour chaque individu de la classe minoritaire, ses $k$ plus proches voisins de la même classe sont calculés, puis un certain nombre d'entre eux sont sélectionnés. Ensuite les individus "synthétiques" sont disséminés aléatoirement le long du segment entre l'individu de la classe minoritaire et ses voisins sélectionnés.

Maintenant que nous avons pris en compte les problèmes apparents de la base de données sur laquelle nous travaillons, nous allons étudier diverses méthodes de prédiction.

# 2 Méthodes de classification

Nous avons essayé plusieurs méthodes. Nous les présenterons dans l'ordre de notre démarche et les commenterons au fur et à mesure.

## 2.1 La régression : une première approche

Notre première approche a été la régression. En effet, bien que n'étant pas de la classification, nous avons décidé de faire ainsi car elle est simple à mettre en oeuvre. Afin de bien prendre en compte le fait que notre variable réponse soit binaire, nous avons utilisé une régression binomiale (Logit).

### 2.1.1 Échantillon non-corrigé

Nous avons donc effectué les régressions Ridge et Lasso sur l'échantillon non-corrigé. Par validation croisée, nous obtenons deux $\lambda$ optimaux différents pour Ridge et Lasso. En effet, la *Figure* 2 ci-dessous représente la déviance moyenne (ou MSE : Minimum Squarred Error) des deux méthodes utilisées en fonction des *log* $\lambda$. A gauche, nous avons représenté celle de l'estimateur Ridge et à droite celle de l'estimateur Lasso.

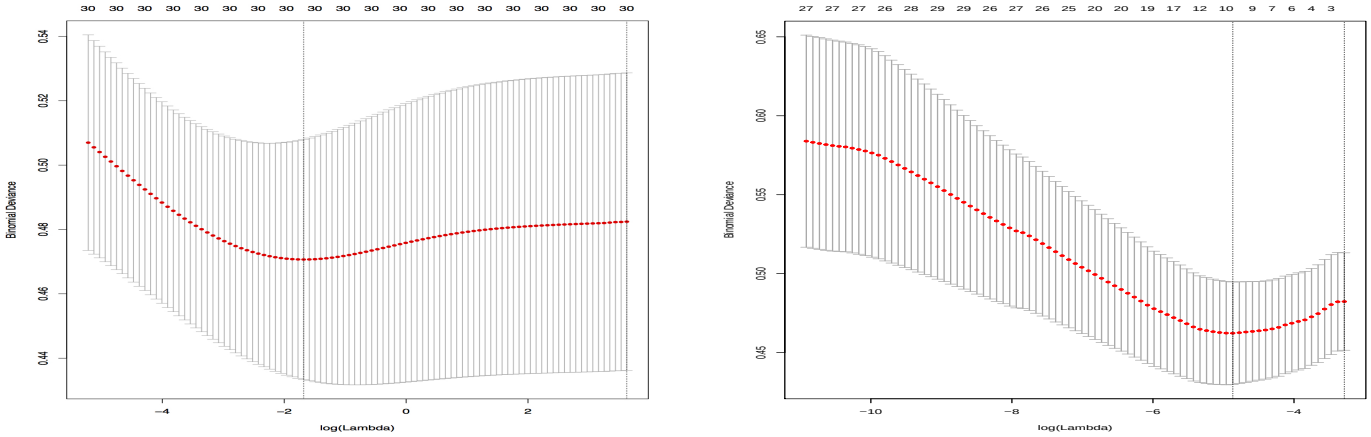


FIGURE 2 – Déviance moyenne par validation croisée : Ridge (à gauche) et Lasso (à droite) - Échantillon non-corrigé

Nous trouvons ainsi des valeurs des $\lambda$ optimaux :

- $\lambda = 0.2111$ pour Ridge
- $\lambda = 0.0087$ pour Lasso

Avec ces valeurs de $\lambda$, nous calculons les prédictions de la variable d'intérêt sur notre échantillon test qui sont presque toujours des "non" (i.e. pas de cancer).

Nous allons à présent introduire une représentation des prédictions que nous allons utiliser tout au long de notre étude. Nous retrouverons dans ces graphiques quatre valeurs différentes ($-1$, 0, 1 et 2), qui indiquent si la patiente a le cancer :

| Dans les faits / Selon notre prédiction | Oui | Non |
|---|---|---|
| Oui | 1 | -1 |
| Non | 2 | 0 |

Ainsi nous trouvons les prédictions suivantes pour Ridge (à gauche) et Lasso (à droite) :

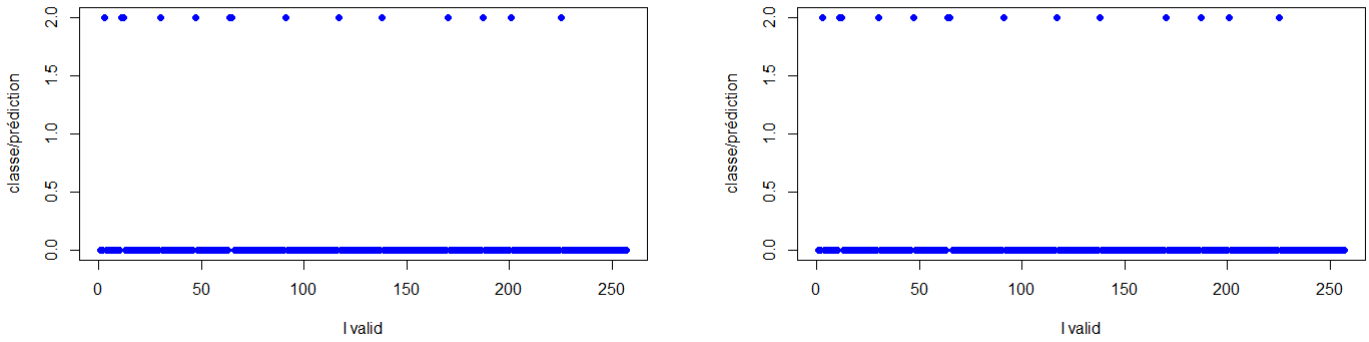


FIGURE 3 – Graphiques de prédiction Ridge(à gauche) et Lasso (à droite) - Échantillon non-corrigé

Nous observons des résultats très similaires pour les deux méthodes. En particulier, nous ne prédisons jamais de cancer, ce qui est inutile ! Nous devons donc modifier notre échantillon pour ne pas à nouveau retrouver le classifieur trivial.

Intéressons-nous aux paramètres du modèle :

- On remarque que la pénalité Lasso est plus efficace que la pénalité Ridge. En effet, Lasso a supprimé 23 variables sur 32, et a donc beaucoup réduit la dimension du problème, ce que Ridge ne fait pas. Cela peut paraître efficace mais dans les faits, nous observons que les résultats des deux estimateurs sont très proches.
- Nous avons observé que les coefficients des estimateurs Ridge et Lasso étaient cohérents entre eux (même ordre de grandeur, même signe).
- Les coefficients sont tous très petits (proche de 0), ce qui explique pourquoi les graphiques de prédictions sont presque uniquement composés de 0.

Nous obtenons cependant des erreurs de classification assez faibles : 9.72% pour Ridge contre 9.72% pour Lasso, mais cela est logique car nous sommes très proche du classifieur trivial.

**Conclusion :** Il faut modifier l'échantillon car nous sommes en train d'utiliser un équivalent du classifieur trivial qui ne nous est d'aucune utilité !

### 2.1.2 Échantillon corrigé

Nous allons désormais faire les mêmes méthodes avec l'échantillon corrigé.

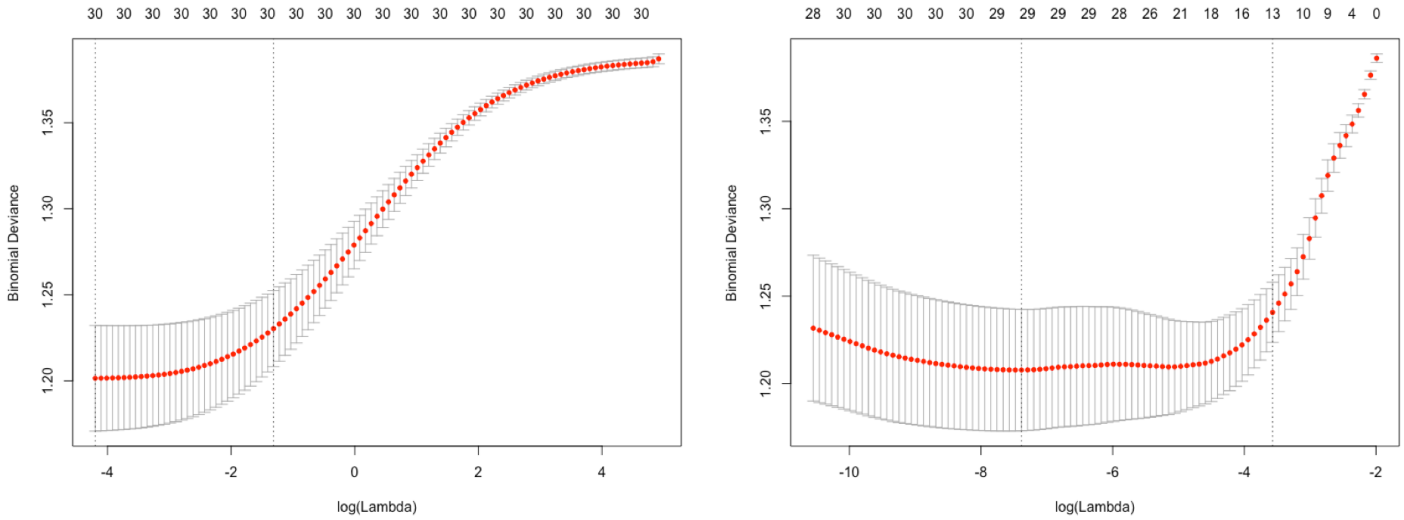


FIGURE 4 – Déviance moyenne en validation croisée - Échantillon corrigé

L'erreur a dans ce cas ci été minimisée pour les valeurs de $\lambda$ optimaux suivantes :

- $\lambda = 0.0499$ pour Ridge
- $\lambda = 0.0100$ pour Lasso

Avec ces valeurs de $\lambda$, nous calculons des prédictions de la variable d'intérêt sur notre échantillon test qui seront presque toujours des "oui" (i.e. a un cancer).

*Rappel des valeurs et de leur signification :*

| Dans les faits / Selon notre prédiction | Oui | Non |
|---|---|---|
| Oui | 1 | -1 |
| Non | 2 | 0 |

Voici à nouveau les graphiques de prédictions obtenus lors des régressions Ridge et Lasso avec le nouvel échantillon :

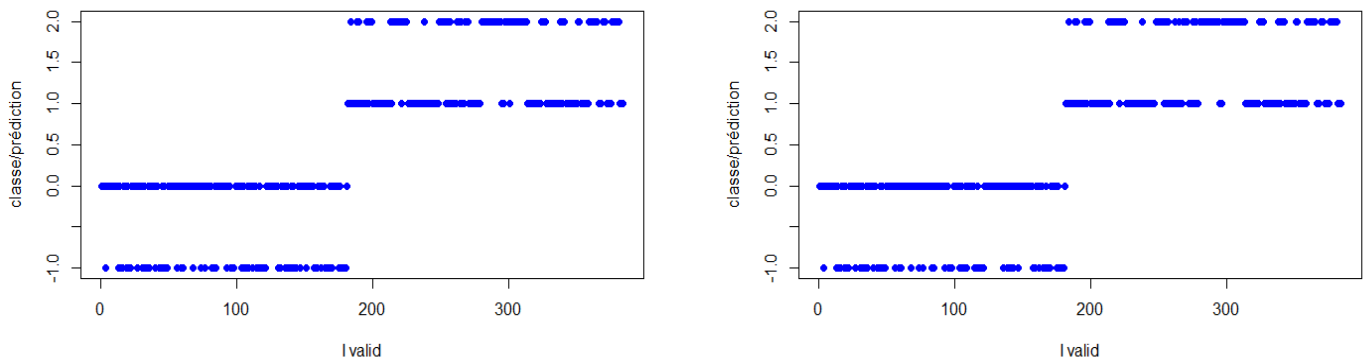


FIGURE 5 – Graphiques de prédiction Ridge et Lasso - Échantillon corrigé

A nouveau, Ridge et Lasso obtiennent des résultats très similaires. On constate une amélioration car toutes les valeurs de prédictions possibles sont prises, mais on observe beaucoup d'erreurs. En effet, toutes les valeurs $-1$ et 2 sont des prédictions erronées. De plus, la valeur 2 (i.e. la patiente n'a pas le cancer selon nous alors qu'elle en a un dans les fait) est très présente. Évidemment, cette erreur de classification est la plus grave d'un point de vue médical et serait la meilleure à minimiser !

Regardons les nouveaux paramètres :

- Cette fois-ci, Lasso a supprimé 7 variables que l'on retrouve également dans les variables supprimées de l'échantillon non-corrigé. Cela signifie que nous estimons qu'il y a plus de variables significatives que dans l'échantillon précédent.
- Les coefficients sont ici aussi de même signe et de même ordre de grandeur pour Ridge et Lasso.
- Nous remarquons que les coefficients prennent des valeurs absolues plus grandes, ce qui explique pourquoi toutes les valeurs sont bien prises cette fois-ci.
- Nous avons observé que le facteur le plus corrélé positivement au cancer est VIH, alors que celui le plus corrélé négativement est HERPES. Dans l'esprit, cela revient à dire qu'avoir de l'herpès implique presque sûrement de ne pas avoir le cancer. Au contraire, avoir le sida donne de grandes chances d'être atteinte du cancer.

Nous obtenons avec cet échantillon des erreurs de classification plutôt grande : $39,07\%$ pour Ridge contre $40,33\%$ pour Lasso, ce qui est loin d'être parfait. Nous devons donc chercher avec d'autres méthodes afin d'améliorer ces résultats !

Les régressions étant faites, il est temps de passer aux classifications !

## 2.2 La méthode des $k$ plus proches voisins ($k$-PPV)

Après avoir essayé diverses régressions, nous avons immédiatement pensé à une méthode de classification les $k$-PPV.

Nous avons tout d'abord réalisé cette méthode sur l'échantillon non-corrigé. Nous obtenons, pour $k = 1$, le graphique de gauche de la figure 6. Ce choix de $k = 1$ provient du fait que, pour tout choix de $k$ assez élevé, nous retrouvons le classifieur trivial. En effet, si on choisissait $k$ plus élevé, on s'intéresserait au vote majoritaire des $k$ plus proches voisins. Or comme notre échantillon est composé majoritairement d'individus n'ayant pas le cancer, nous prédirons -presque- toujours un résultat négatif.

Nous avons ensuite appliqué cette méthode à l'échantillon corrigé. Toujours pour $k = 1$, nous obtenons le graphique de droite de la figure 6.

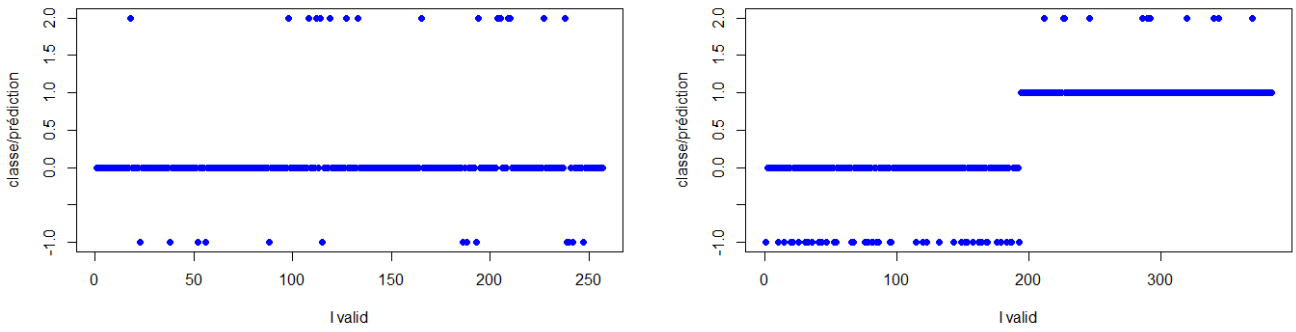


FIGURE 6 – Prédiction des $k$-PPV, avec $k = 1$ - Échantillon non-corrigé (à gauche) et Échantillon corrigé (à droite)

Nous voyons instantanément que cette méthode de classification est mieux adaptée que les régression précédentes quel que soit l'échantillon utilisé.

Dans le graphique de gauche, on remarque que l'on prédit rarement qu'une personne ait le cancer, et que ces prédictions sont toutes fausses. En effet, la valeur 1 n'est jamais atteinte tandis que $-1$ l'est.

Dans celui de droite, toutes les valeurs possibles sont prises. En particulier, nous avons correctement prédit la présence de la maladie pour un grand nombre d'individus. De plus, parmi nos erreurs de prédiction, la moins fréquente est la valeur 2, i.e. la patiente n'a, selon nous, pas le cancer alors qu'elle en a un. Comme dit précédemment, cette erreur de classification est la plus grave d'un point de vue médical, et notre but est de l'éviter au maximum.

Pour finir d'illustrer la différence entre les résultats obtenus sur nos deux échantillons, avant et après correction, comparons leurs risques de classification. Nous obtenons $11,28\%$ pour l'échantillon non-corrigé et $15,17\%$ pour le corrigé. Ce résultat qui peut sembler contradictoire avec nos remarques précédentes nous rappelle que nos deux échantillons ne sont pas comparables.

Dans le premier cas, qu'il y ait erreur de prédiction ou non sur les individus ayant un cancer peut faire varier le risque d'au plus 8% (car les cancéreux représentent 8% de l'échantillon). Cependant, médicalement parlant, nous devrions donner autant d'importance à ces individus qu'aux autres. Ainsi, se tromper sur cette classe devrait faire varier le risque d'au plus 50%. C'est pour cela que le risque du premier prédicteur est faible (environ 10%) alors qu'on se trompe sur toute la classe des individus atteints de cancer.

Dans le second cas, le classifieur trivial égal à 0 enverrait les valeurs $-1$ sur 0 et les valeurs 1 sur 2. Il aurait donc une erreur de 50%, ce qui est bien supérieur au risque trouvé ici de $15,17\%$. Notre classifieur sur cet échantillon est donc meilleur que le classifieur trivial.

Par validation croisée, nous trouvons que le meilleur $k$ à utiliser est $k = 1$.

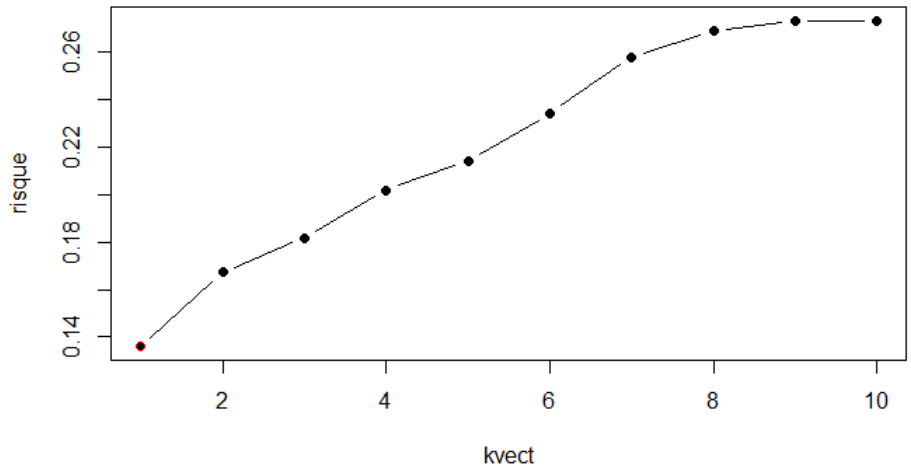


FIGURE 7 – Validation croisée pour l'échantillon corrigé

Cette méthode nous a permis de confirmer ce que nous avions compris avec la régression : il nous faut travailler sur un autre échantillon que celui donné : un échantillon corrigé. Nous sommes de plus convaincus que nous pouvons trouver mieux que les résultats donnés par la régression et les $k$-PPV !

## 2.3 Arbres de décision et forêts aléatoires

Après avoir traité la méthode des $k$ plus proches voisins, nous nous sommes intéressés aux arbres de décisions. Cette méthode a l'avantage principal de donner une solution dont l'interprétation graphique est simple.

### 2.3.1 Échantillon non-corrigé

L'arbre de décision pour cet échantillon est représenté par les *Figures* 8 et 9, obtenus à l'aide de deux échantillons de validation (Ivalid) différents. La partition *Ivalid* / *Iapp* a été faite aléatoirement pour le premier arbre, alors que celle du second a été générée par les 30% premières valeurs de notre base de données pour Ivalid et par les 70% des dernières valeurs pour Iapp. Comme nous pouvons le constater, très peu de "*yes*" sont prédits, d'où la nécessité de rééquilibrer les données.

Pour les deux arbres, nous obtenons des erreurs de classification de $7,32\%$ pour l'arbre 1 contre $7,15\%$ pour l'arbre 2.

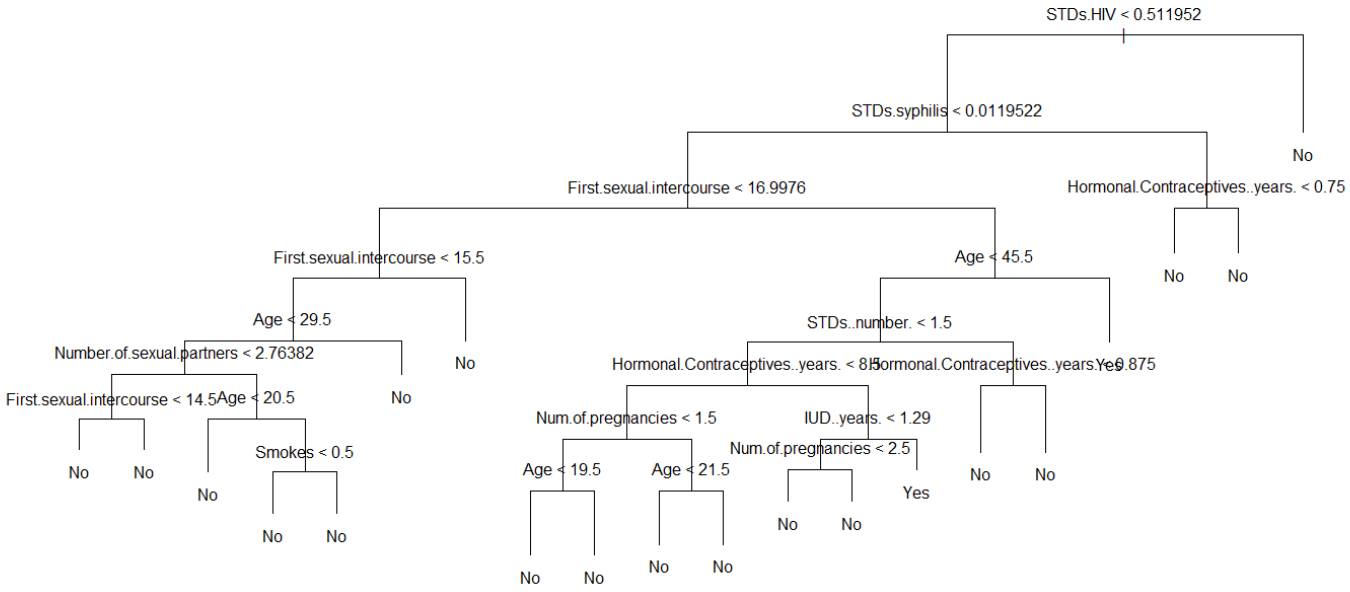


FIGURE 8 – Arbre 1 - Échantillon non-corrigé

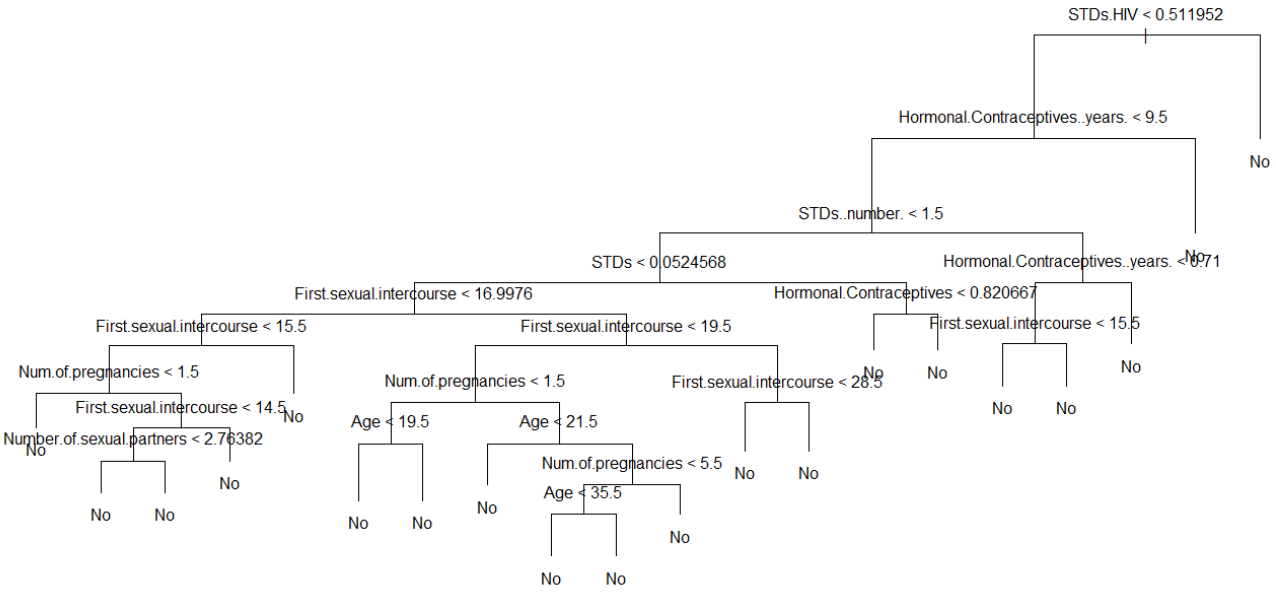


FIGURE 9 – Arbre 2 - Échantillon non-corrigé

On constate que les deux arbres donnent beaucoup -voir dans certains cas uniquement- des "*no*", ce qui revient au classifieur trivial que nous souhaitons éviter. De plus, l'arbre possède 25 noeuds pour répondre "*no*" presque tout le temps, ce qui est redondant.

Représentons maintenant la forêt correspondant à l'échantillon non-corrigé :

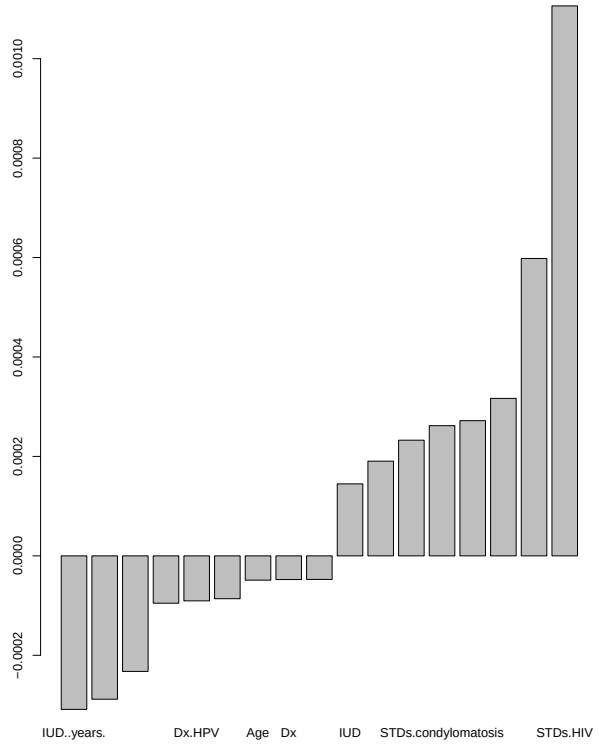


FIGURE 10 – Forêt liée à l'échantillon non-corrigé

L'erreur de classification de cette forêt est de seulement $4,28\%$, mais ceci est normal car nous sommes toujours en train de regarder le classifieur trivial !

### 2.3.2 Échantillon corrigé

Nous avons ensuite appliqué cette méthode aux données équilibrées, dont les arbres de décision sont représentés ci-dessous sur les *Figures* 11 et 12. Contrairement au cas précédent, nous prédisons "*yes*" pour un grand nombre d'individus !

Pour les deux arbres, nous obtenons avec cet échantillon des erreurs de classification de $9,70\%$ pour l'arbre 1 contre $12,49\%$ pour l'arbre 2.

Les erreurs trouvées sont plutôt faibles. En effet, nous obtenons moins de 10% d'erreur pour l'arbre 1.

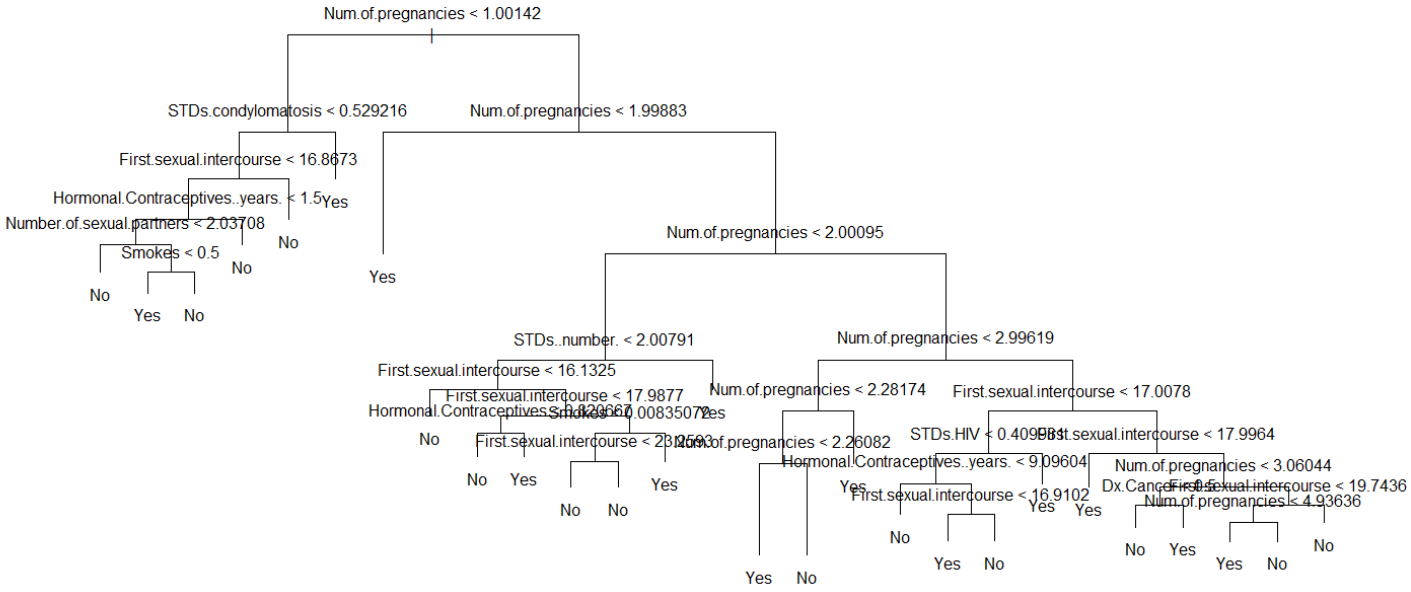


FIGURE 11 – Arbre 1 - Échantillon Corrigé

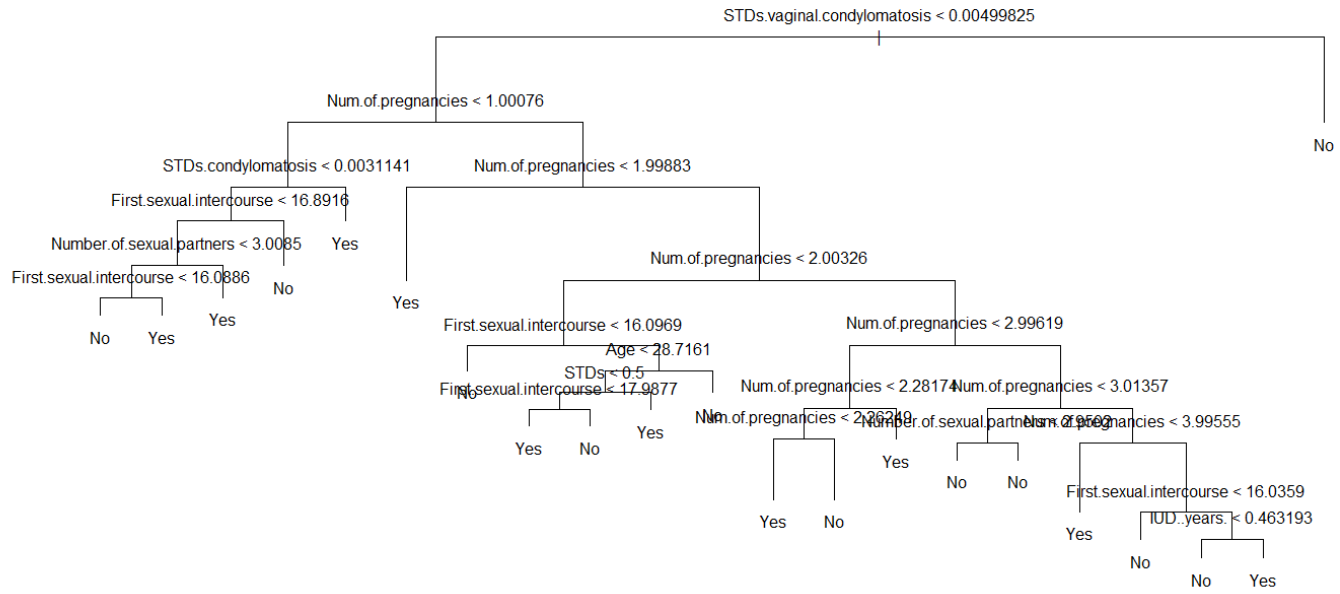


FIGURE 12 – Arbre 2 - Échantillon Corrigé

Intéressons nous maintenant à la forêt correspondant à cet échantillon corrigé :

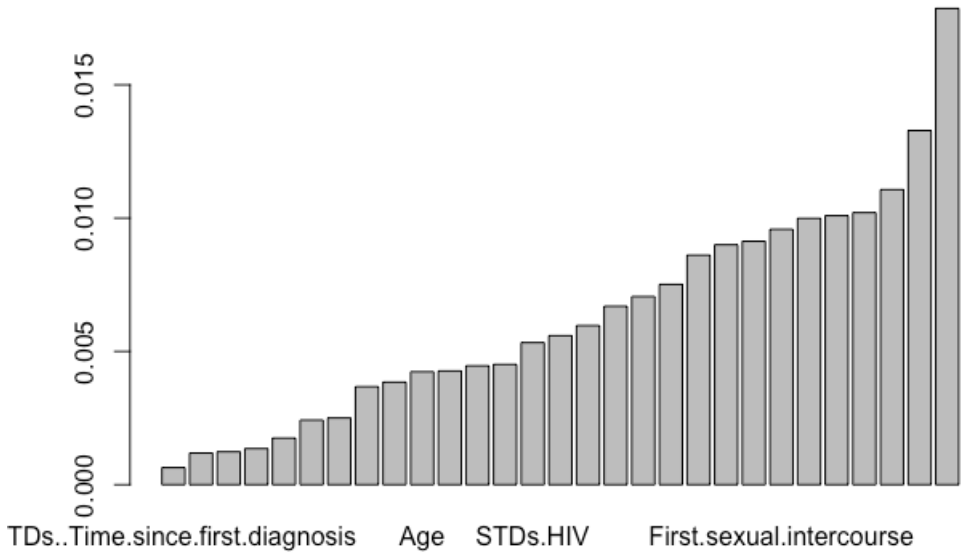


FIGURE 13 – Forêt liée à l'échantillon corrigé

L'erreur de classification de cette nouvelle forêt est de seulement 27.86%.

On remarque que les éléments ayant le plus d'importance dans la classification sont l'âge du premier rapport sexuel ainsi que la présence ou non du SIDA. Pour être un peu plus clair, avoir le VIH augmenterait significativement la probabilité d'avoir un cancer du col de l'utérus. Nous remarquons que c'est bel et bien ce qu'avait déjà prédit la régression.

## 2.4 Les réseaux de neurones

Nous avons ensuite décidé d'utiliser les réseaux de neurones pour voir s'ils étaient plus efficaces que les méthodes précédentes.

### 2.4.1 Échantillon non-corrigé

Comme précédemment, nous commençons notre étude par l'échantillon non-corrigé.

Nous n'utiliserons qu'une seule couche cachée car en utiliser plus d'une devient extrêmement coûteux pour l'échantillon corrigé et qu'il est inutile de comparer deux réseaux de neurones différents.

Traçons les représentations des réseaux de neurones à un (*Figure* 14) ou deux (*Figure* 15) neurones :

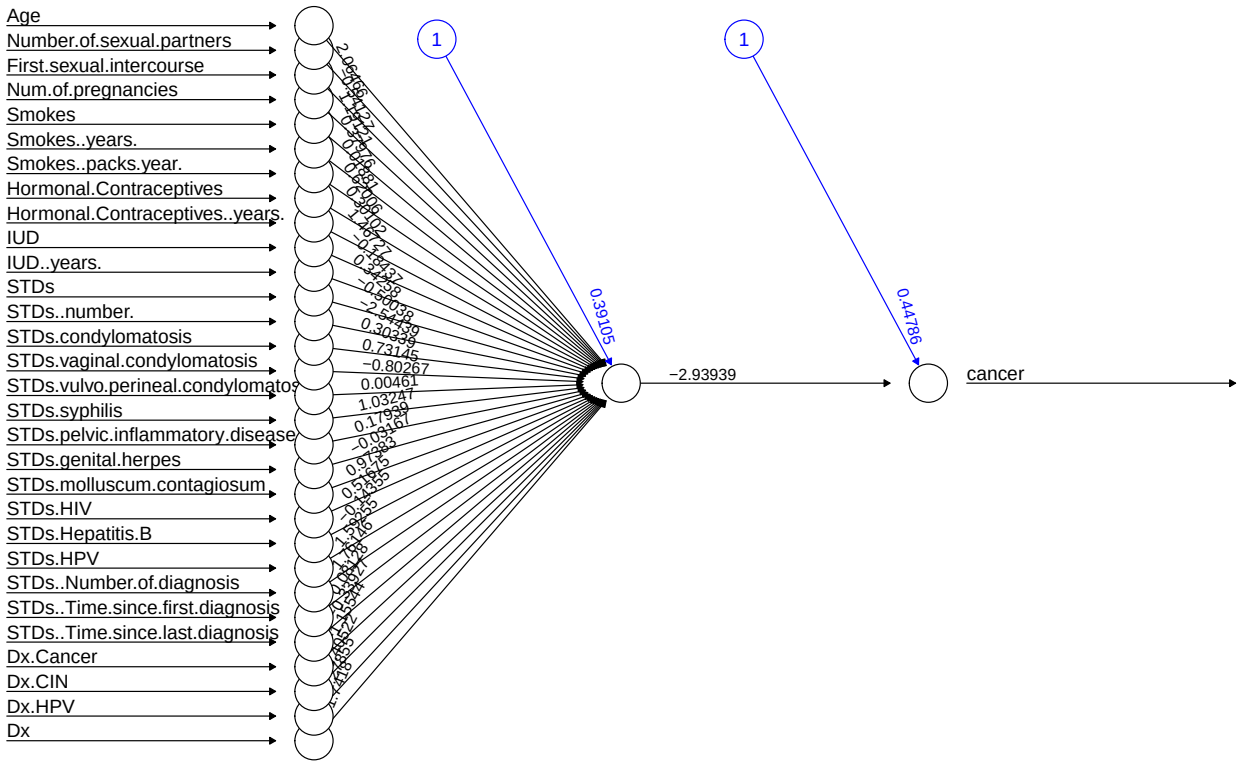


FIGURE 14 – Réseau à un neurone - Echantillon non-corrigé

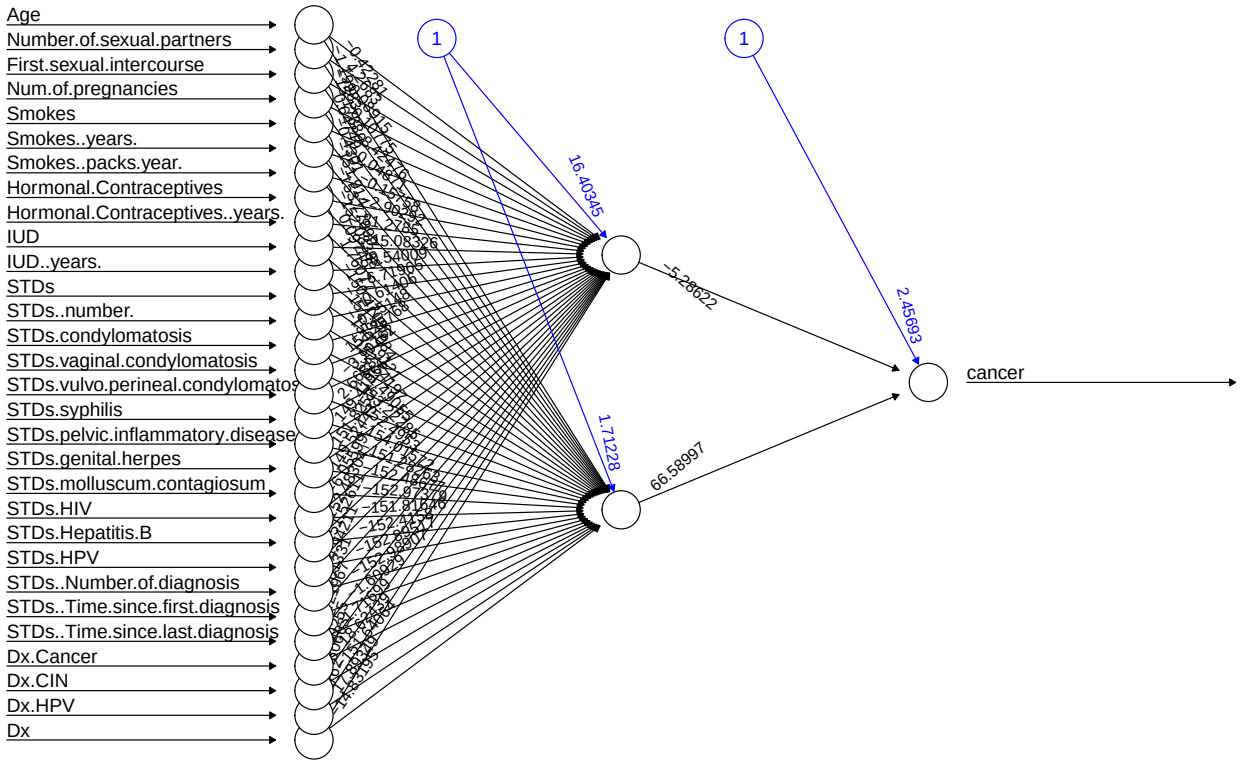


FIGURE 15 – Réseau à deux neurones - Echantillon non-corrigé

*Rappel des valeurs et de leur signification :*

| Selon notre prédiction \ Dans les faits | Oui | Non |
|---|---|---|
| Oui | 1 | -1 |
| Non | 2 | 0 |

Si nous matérialisons à nouveau nos prédictions dans nos graphiques de prédiction pour ces deux réseaux neuronaux, nous obtenons les graphes suivants :

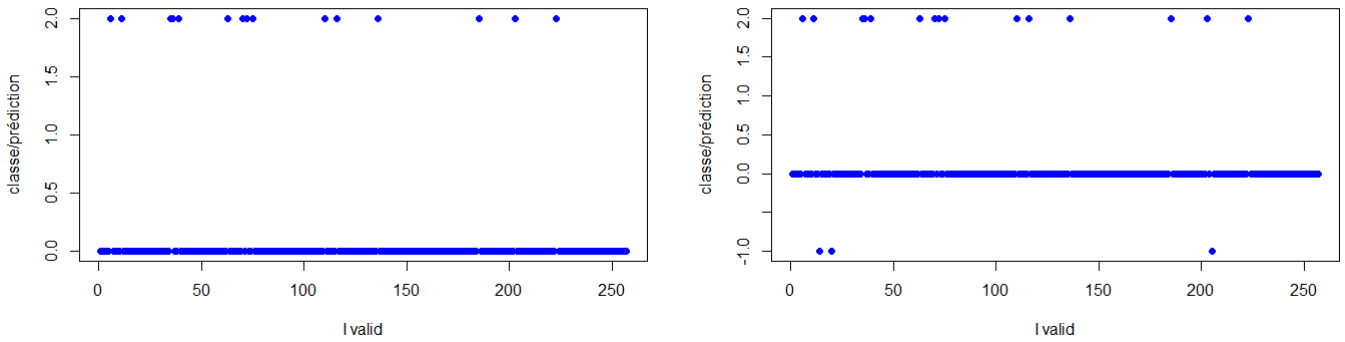


FIGURE 16 – Graphiques de prédiction pour les réseaux à un (à gauche) et deux (à droite) neurones - Échantillon non-corrigé

Nous remarquons que, pour un seul neurone, nous trouvons à nouveau un classifieur proche du classifieur trivial.
Pour deux neurones, nous avons presque à faire au classifieur trivial, à 3 prédictions près. Cependant, les 3 prédictions qui diffèrent valent $-1$. Elles sont donc toutes trois fausses, ce qui n'est pas très convaincant.

Pour l'instant, nos réseaux de neurones ne présentent pas d'excellentes caractéristiques !

Pour un neurone, l'erreur de classification est de $5,84\%$ alors que pour deux neurones, elle est de $7,00\%$

Les erreurs sont très faibles, mais cela s'explique par le fait que nous travaillons sur l'échantillon non-corrigé.

### 2.4.2 Échantillon corrigé

Nous regardons désormais l'échantillon corrigé.
Traçons ici aussi les représentations des réseaux de neurones à un (*Figure* 17) ou deux (*Figure* 18) neurones :

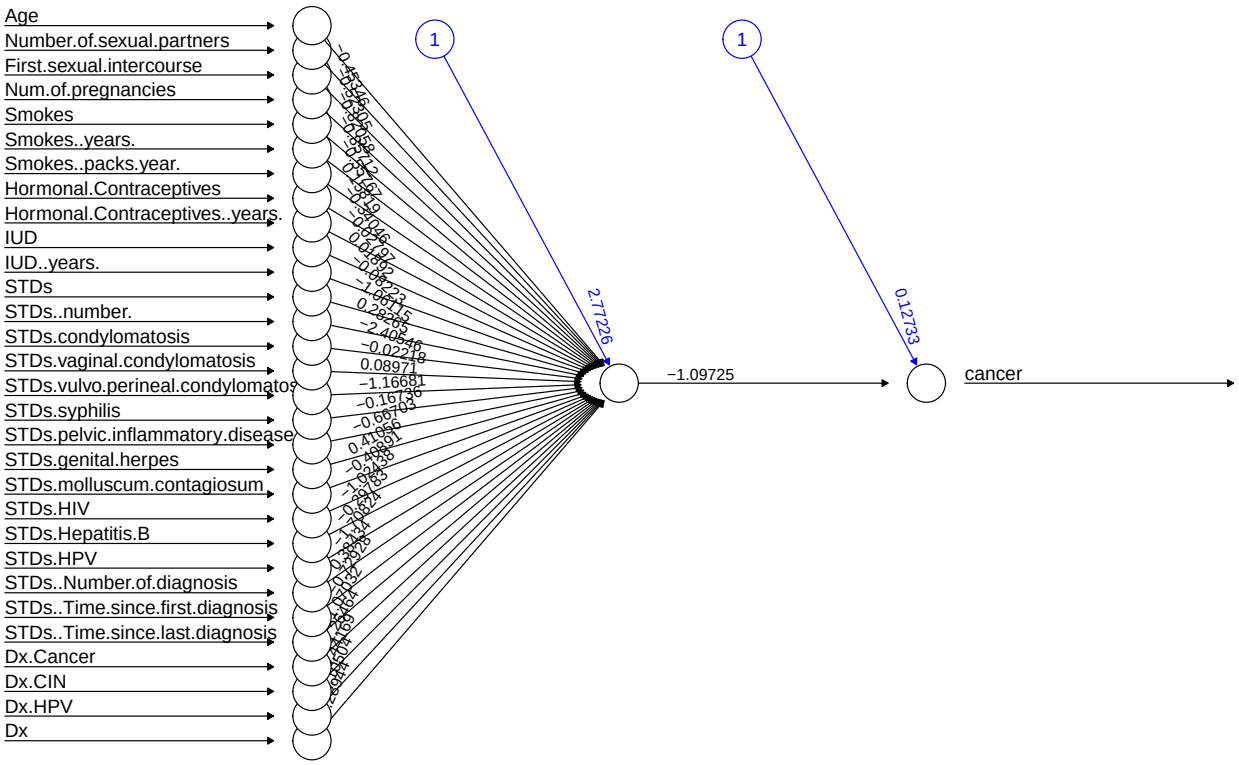


FIGURE 17 – Réseau à un neurones - Échantillon corrigé

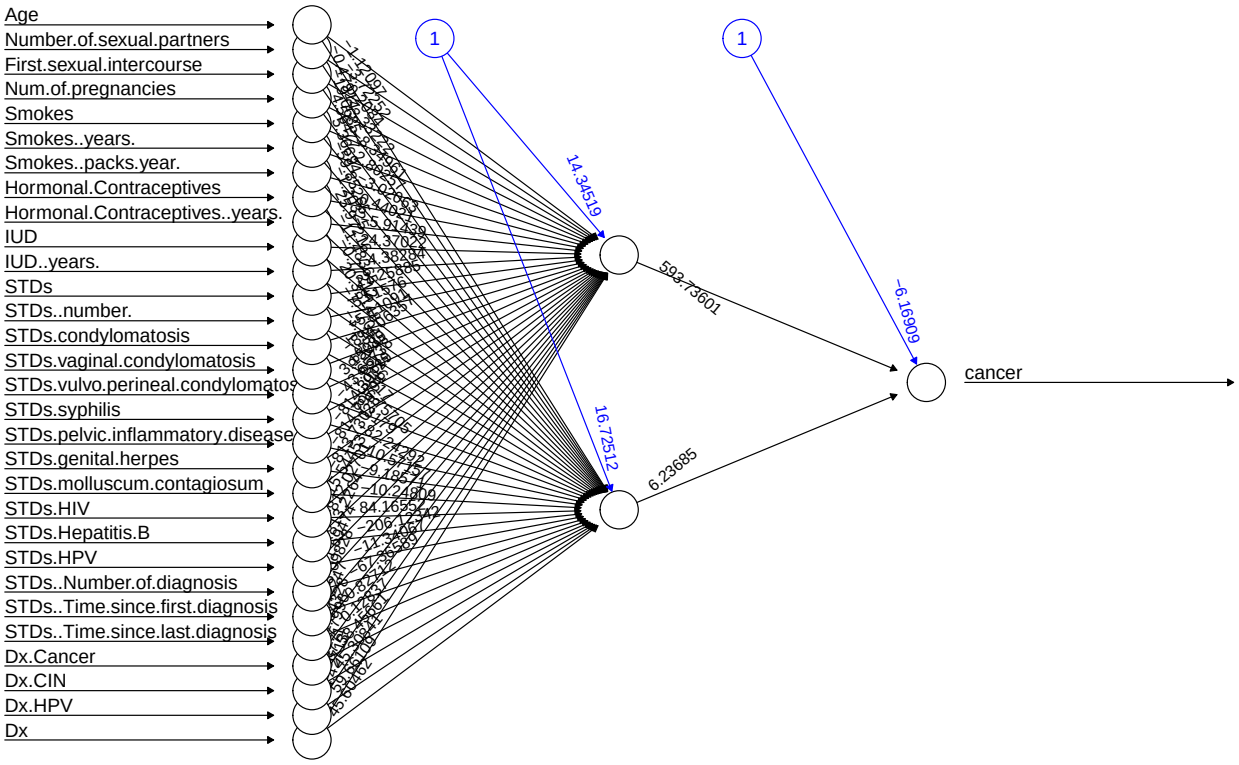


FIGURE 18 – Réseau à deux neurones - Échantillon corrigé

*Rappel des valeurs et de leur signification :*

| Selon notre prédiction \ Dans les faits | Oui | Non |
|---|---|---|
| Oui | 1 | -1 |
| Non | 2 | 0 |

Matérialisons à nouveau nos prédictions dans nos graphiques de prédiction pour ces deux réseaux neuronaux. Nous obtenons les graphes suivants :

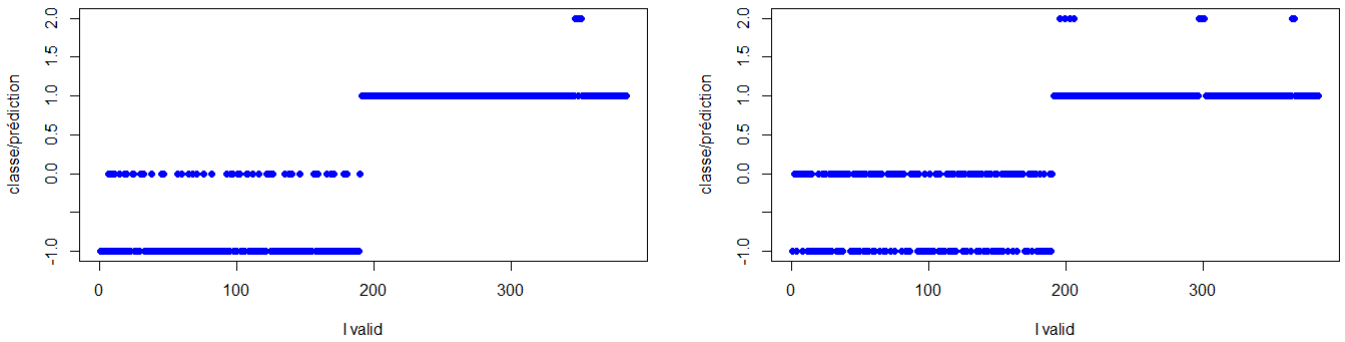


FIGURE 19 – Graphiques de prédiction pour les réseaux à un (à gauche) et deux (à droite) neurones - Échantillon corrigé

Le réseau à un neurone prédit trop de cancers pour les gens qui ne l'ont pas. Cependant, nous obtenons très peu de fois la valeur 2, ce qui était ce que nous cherchions à faire ! Cela paraît donc être un bon compromis car nous avons ainsi peu de chance de ne pas prédire un cancer s'il y en a effectivement un.

Le réseau à deux neurones prédit moins de cancers en général mais prédit également moins de cancers aux personnes en ayant réellement un. Le nombre d'erreur de type 2 reste cependant relativement faible, donc ce réseau de neurones reste intéressant.

Nous pouvons donc conclure que ces réseaux de neurones sur l'échantillon corrigé donnent des résultats plutôt corrects, même s'il faut rester vigilant à la sur et sous-prédiction de cancers.

Pour un neurone, l'erreur de classification est ici de $38,28\%$ alors que pour deux neurones, elle est de $39,58\%$. Les erreurs sont énormes ! Ce n'est donc pas très intéressant, même si la majeure partie des erreurs est liée à une sur-prédiction de cancer (i.e. on prédit un cancer alors qu'il n'y en a pas).

# Conclusion

Nous arrivons au bout de notre étude, il est donc temps de faire le point sur les méthodes essayées.

Tout d'abord, nous avons pu constater que travailler sur l'échantillon non-corrigé était une aberration car nous avons obtenu dans tous les cas le classifieur trivial ou un classifieur très proche de ce dernier, ce qui ne nous est d'aucune utilité du point de vue prédiction. Cela vient bien évidemment du déséquilibre entre les deux classes de notre échantillon : celle atteinte du cancer (8% de notre échantillon total) et celle saine (92%). Nous nous sommes donc plutôt intéressés aux méthodes appliquées à l'échantillon corrigé.

Voici un tableau récapitulatif des erreurs de classification ainsi que des erreurs de type 2 de chaque méthode, car nous avons expliqué que l'erreur de type 2 est l'erreur que nous souhaitons minimiser.

| Méthode | Erreur de classification (en % ) | Erreur de type 2 (Qualitatif) [1] |
|---|---|---|
| Régression Ridge | $39,07$ | $++$ |
| Régression Lasso | $40,33$ | $++$ |
| $k$-PPV ($k=1$) | $15,17$ | $+$ |
| Arbres | 9.70 ou $12,49$ [2] | NA [3] |
| Forêts | $27,86$ | NA [3] |
| Réseau d'un neurone | $38,28$ | $--$ |
| Réseau de deux neurones | $39,58$ | $--$ |

FIGURE 20 – Erreurs des méthodes - Échantillon corrigé

Ainsi nous constatons qu'en fonction de ce que nous souhaitons minimiser comme erreur, nous devons choisir une méthode plutôt qu'une autre.

Dans les problèmes de classification, nous souhaitons toujours par défaut minimiser l'erreur de classification. Mais avec un peu de recul sur notre base de données, les deux types d'erreurs ne sont pas comparables.

Nous avons décidé de privilégier la minimisation de l'erreur de type 2 afin de ne -presque- jamais oublier de prédire un cancer lorsqu'il y en a effectivement un, quitte à en prédire trop. Nous optons donc pour l'utilisation des réseaux de neurones.

Cependant, nous souhaitons améliorer légèrement ce protocole pour prédire moins de cancers. En effet, nous avons expliqué que les réseaux de neurones prédisent trop de cancers en général. Donc nous pourrions envisager d'utiliser pour commencer les réseaux de neurones, puis ensuite d'appliquer une nouvelle méthode ($k$-PPV par exemple) afin de filtrer les individus auxquels nous prédisons deux fois un cancer et ceux qui n'en n'auraient qu'une seule.

Ainsi, seules les personnes ayant eu cette double prédiction de cancer devraient aller faire les tests de dépistages.

Si nous souhaitions approfondir l'étude, nous pourrions chercher à minimiser seulement l'erreur de type 2 - un peu à l'image des tests statistiques qui ne cherchent à minimiser qu'un seul type d'erreur. En effet, un test statistique à deux hypothèses $H_0$ et $H_1$ est dit de niveau $\alpha$ si la probabilité de se tromper sous $H_0$ est inférieure à $\alpha$. On pourrait s'inspirer de ceci en posant les hypothèses :

$H_0$ : **la patiente est atteinte de cancer** contre $H_1$ : **la patiente n'est pas atteinte de cancer**.

Dans ce cadre, on ne s'intéresserait qu'à l'erreur de première espèce qui est la plus grave du point de vue médical.

En effet, il reste préférable de prédire trop de cancer puis d'obtenir des résultats de dépistages négatifs plutôt que de ne pas prédire et de laisser une personne ayant un cancer sans traitement. Cependant, nous ne devons pas oublier que chaque test ou dépistage coûte de l'argent et qu'envoyer trop de personnes passer les tests risque d'être très onéreux.

Nous touchons ici une autre partie de la question, qui relève de l'éthique plus que des mathématiques :

Combien sommes nous prêts à dépenser pour sauver une personne ?

1. $++$ : Beaucoup d'erreur de type 2 ; $--$ : Très peu d'erreur de type 2
2. Nous avons construit deux arbres à partir de deux échantillons d'apprentissages différents, d'où les deux valeurs
3. Nous n'avons malheureusement pas réussi à représenter ces graphiques pour ces deux méthodes.