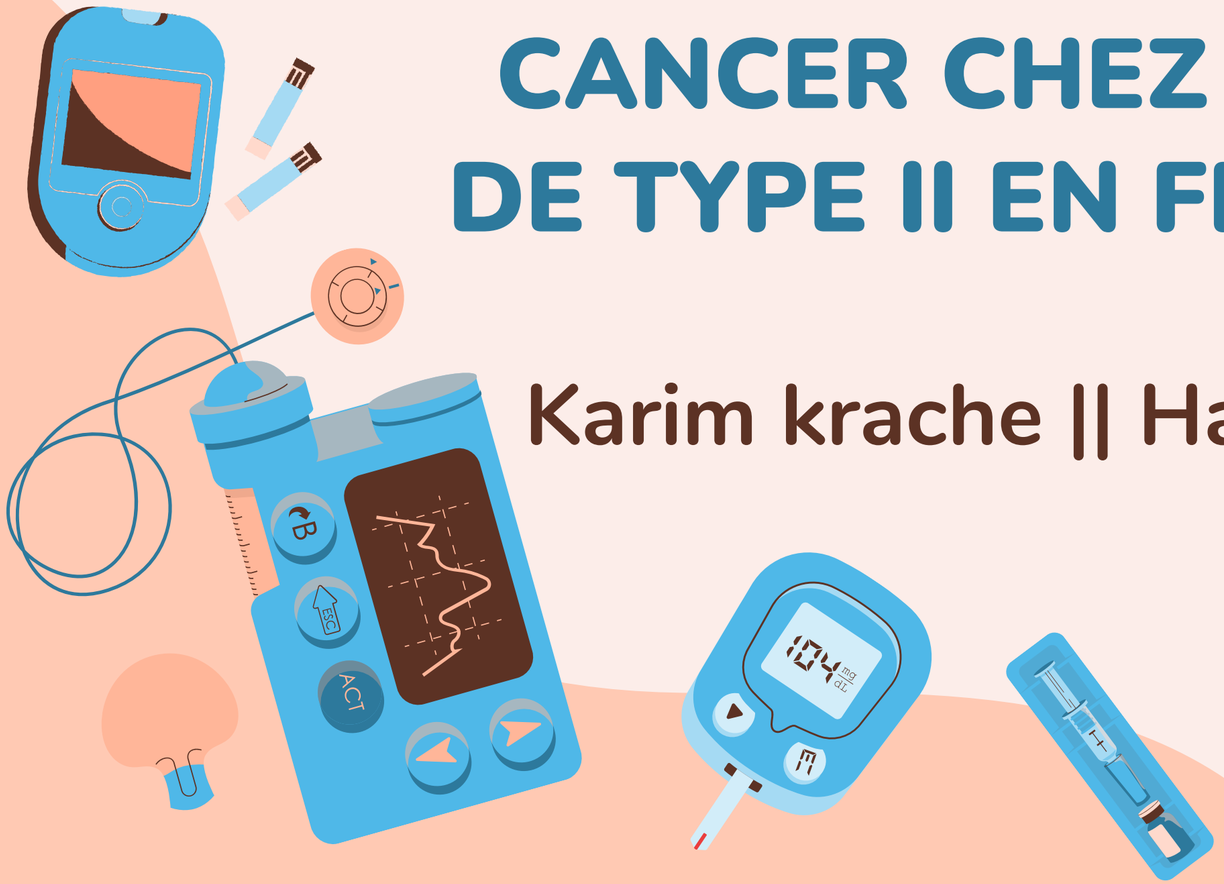
ETUDE DE L'APPARITION DU CANCER CHEZ LES DIABÉTIQUES DE TYPE II EN FRANCE SUR 10 ANS



Karim krache || Hadirou Tamdamba

François Husson,
Professeur, Department Statistics & Computer Science
Institut Agro Rennes-Angers.

CONTEXTE DE L’ÉTUDE

Le cancer hepatocellulaire (CHC ou HCC) c'est...

6000 MORTS PAR AN EN FRANCE

DANS **90%** DES CAS, IL INTERVIENT CHEZ LES PATIENTS ATTEINTS DE CIRRHOSE

12,5 /100 000 CHEZ LES HOMMES ET **2,5** /100 000 CHEZ LES FEMMES.

ENTRE 1980 ET 2018, LA PRÉVALENCE EN FRANCE A AUGMENTÉ DE **12,44** /1000 000

Que nous disent les études ?

Infection par le virus de l'hépatite B (HBV) - .

un risque 100 fois plus élevé de développer un HCC pour les patients infectés par le HBV par rapport aux non infectés

Infection par le virus de l'hépatite C (HCV)

principale cause du CHC en Europe, au Japon, en Amérique latine et aux États-Unis, avec un risque 17 fois plus élevé pour les patients infectés par le HCV par rapport aux non infectés.

Obésité

corrélée à un risque accru de HCC, avec deux fois plus de chances de développer un HCC pour les patients obèses par rapport à ceux qui ne le sont pas.

Consommation d'alcool

la consommation chronique d'alcool, en particulier à des niveaux élevés, est un facteur de risque important pour le HCC.

Diabète sucré

également associé au développement du CHC

sources

Hepatocellular Carcinoma: Risk Factors, Diagnosis and Treatment

Analysis of Risk Factors Associated with the Development of Hepatocellular Carcinoma in Chronic HBV-Infected Chinese: A Meta-Analysis

PROBLÉMATIQUE

QUELS SONT LES FACTEURS CONTRIBUANTS À L'APPARITION DU CHC CHEZ LES PATIENTS ATTEINTS DE DIABÈTE DE TYPE 2 ?

DESCRIPTION DU JEU DE DONNÉES

Taille du jeu de données

Nombre d'individus : **784 308**

Nombre de variables : **325**

Source du jeu de données

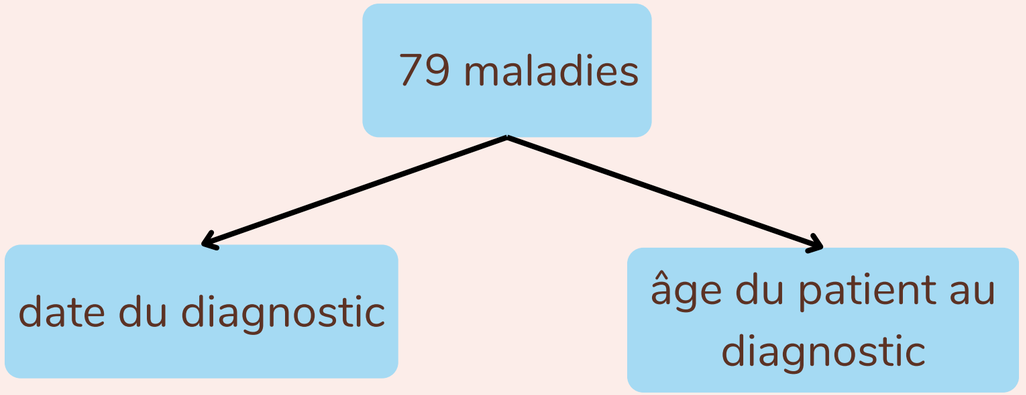
Source des données : Base de données nationale française des hospitalisations

Nature des données : Hospitalisations de patients DT2 de 2011 à 2020.

Contenu : Enregistrements des diagnostics et des actes lors de chaque hospitalisation, incluant l'âge du patient et la date diagnostic.

Description des variables

ID patient	chaine de caractères
sexe	facteur (binaire)
age.min	numérique
age.max	numérique
Death	booléen



Description des variables

325

maladie 1 | maladie 2

anonyme	Age_variable_X1_dp_dr	D_variable_X1_dp_dr	Age_variable_X1_all	D_variable_X1_all	Age_variable_X2_dp_dr	D_variable_X2_dp_dr	Age_variable_X2_all	D_variable_X2_all
######	30	20/07/2010	NA	NA	NA	NA	NA	NA
######	NA	NA	02/02/2012	45	43	02/02/2012	NA	NA
######	NA	NA	NA	NA	NA	NA	NA	NA

784 308

DP_DR
maladie principale qui est diagnostiquée

Description des variables

325

maladie 1 | maladie 2

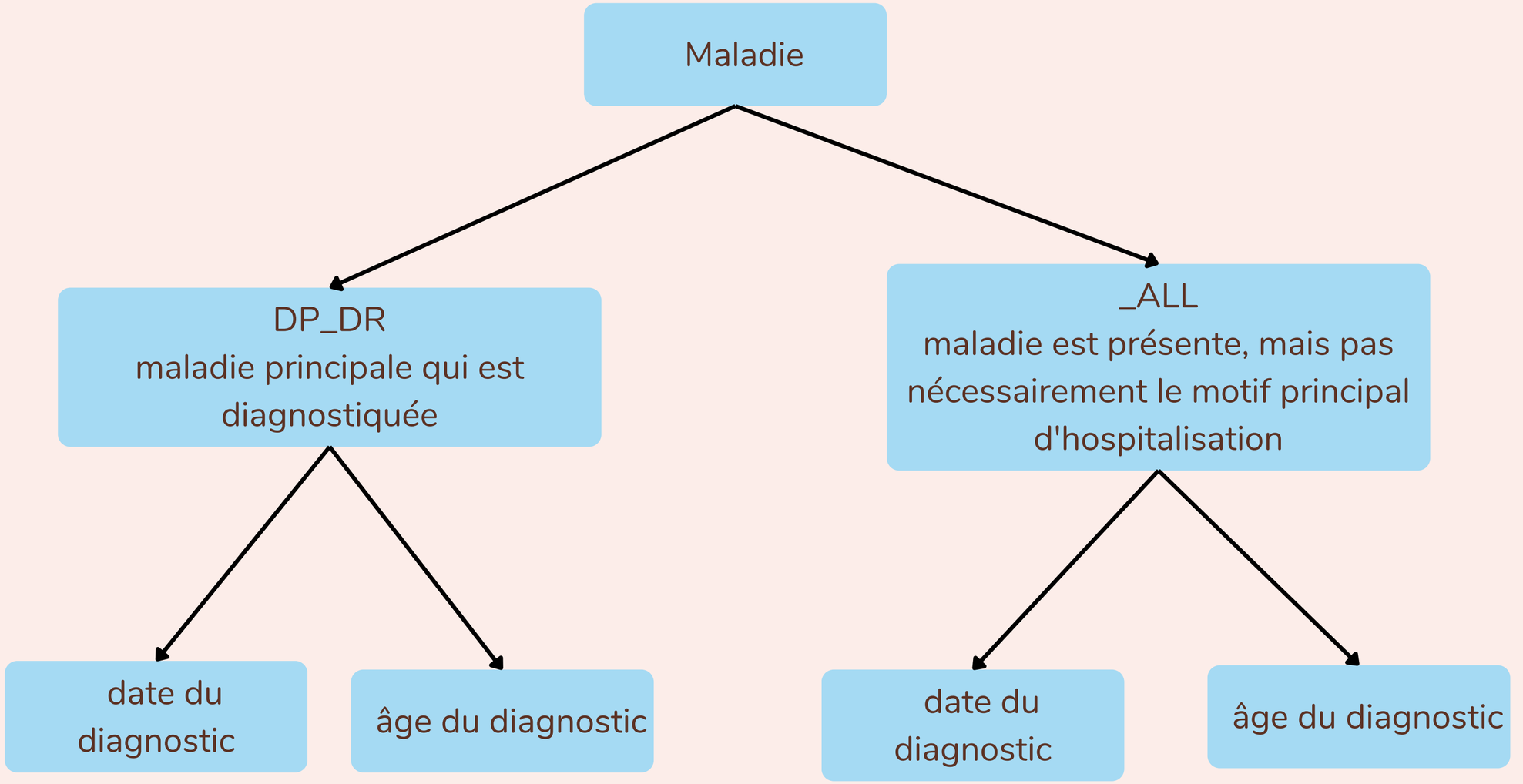
784 308

anonyme	Age_variable_X1_dp_dr	D_variable_X1_dp_dr	Age_variable_X1_all	D_variable_X1_all	Age_variable_X2_dp_dr	D_variable_X2_dp_dr	Age_variable_X2_all	D_variable_X2_all
######	30	20/07/2010	NA	NA	NA	NA	NA	NA
######	NA	NA	02/02/2012	43	43	02/02/2012	NA	NA
######	NA	NA	NA	NA	NA	NA	NA	NA

DP_DR
maladie principale qui est diagnostiquée

_ALL
maladie est présente, mais pas nécessairement le motif principal d'hospitalisation

Description des variables



79

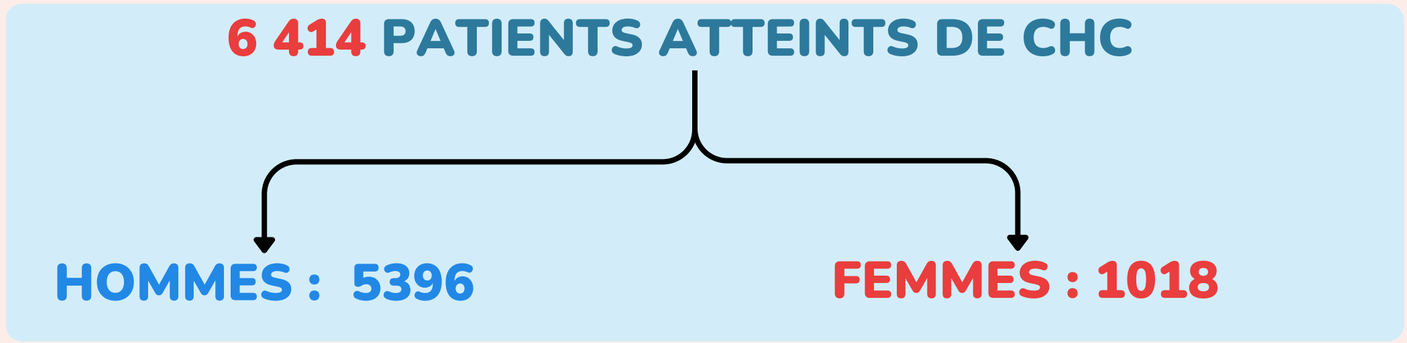
316

ANALYSES DESCRIPTIVES

STATISTIQUES DESCRIPTIVES

L'ÂGE MÉDIAN ÉTAIT DE 62 ANS (52 ANS , 71 ANS)

HOMMES : 54.14 % FEMMES : 45.86%



RÉPARTITION DES TROUBLES DE CONSOMMATION D'ALCOOL ET DU CHC PAR SEXE

HOMMES

TROUBLES DE CONSOMMATION D'ALCOOL	HCC NEGATIIVE	HCC POSITIVE
NON	419 197	4 362
OUI	0	1 034

FEMMES

TROUBLES DE CONSOMMATION D'ALCOOL	HCC NEGATIIVE	HCC POSITIVE
NON	358 697	880
OUI	0	138

STATISTIQUES DESCRIPTIVES

- **RÉPARTITION DE L'OBÉSITÉ ET DU CANCER HCC PAR SEXE**

HOMMES

OBÉSITÉ	HCC NEGATIVE	HCC POSITIVE
NON	419 197	5 320
OUI	0	76

FEMMES

OBÉSITÉ	HCC NEGATIVE	HCC POSITIVE
NON	358 697	987
OUI	0	31

- **RÉPARTITION DES PATIENTS AYANT LA CIRRHOSE ET DU HCC PAR SEXE**

HOMMES

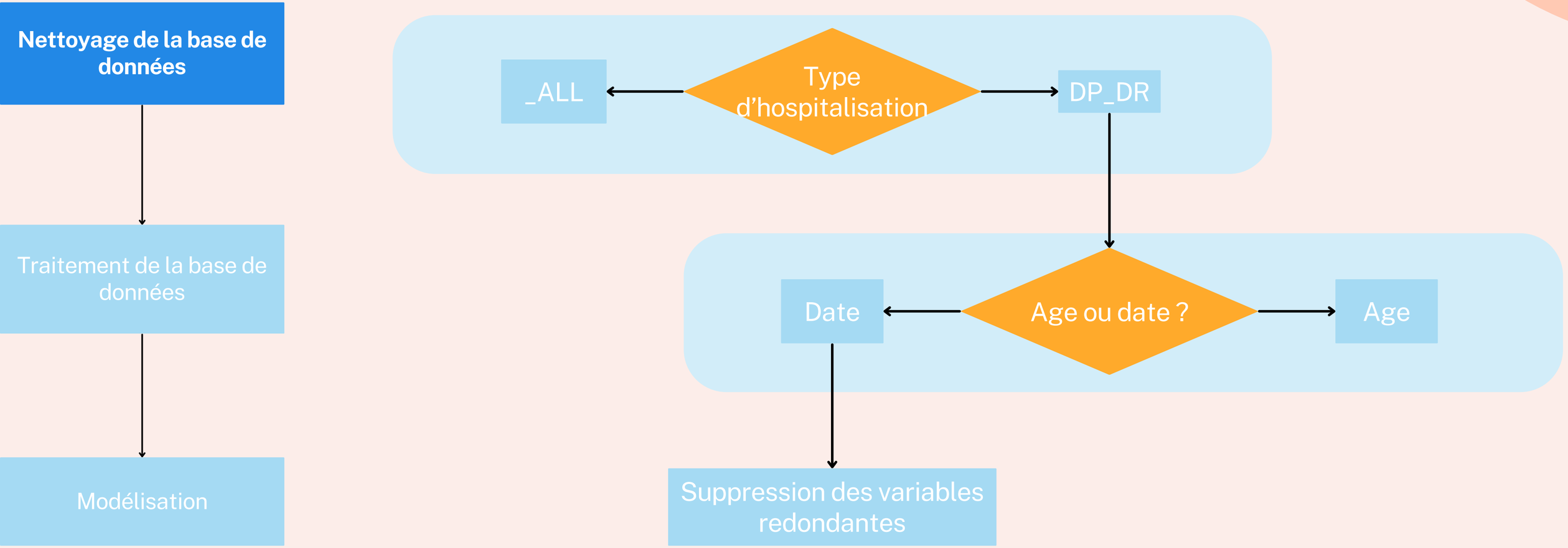
CIRRHOSE	HCC NEGATIVE	HCC POSITIVE
NON	419 197	3 928
OUI	0	1 468

FEMMES

CIRRHOSE	HCC NEGATIVE	HCC POSITIVE
NON	358 697	704
OUI	0	314

DÉMARCHE MÉTHODOLOGIQUE

DÉMARCHE POUR LA MODÉLISATION DES FACTEURS PRÉDICTIFS



DÉMARCHE POUR LA MODÉLISATION DES FACTEURS PRÉDICTIFS

Nettoyage de la base de données

Traitement de la base de données

Modélisation

325

Age_variable_X1_dp_dr	D_variable_X1_dp_dr	Age_variable_X1_all	D_variable_X1_all	Age_hepatocellular_carcinoma_dp_dr	D_hepatocellular_carcinoma_dp_dr	D_hepatocellular_carcinoma_all	D_hepatocellular_carcinoma_all
30	20/07/2010	NA	NA	NA	NA	NA	NA
NA	NA	02/02/2012	45	43	20/09/2010	NA	NA
NA	NA	NA	NA	NA	NA	NA	NA

784 308

60

D_variable_X1_dp_dr	D_hepatocellular_carcinoma_dp_dr
20/07/2010	NA
NA	20/09/2010
NA	NA

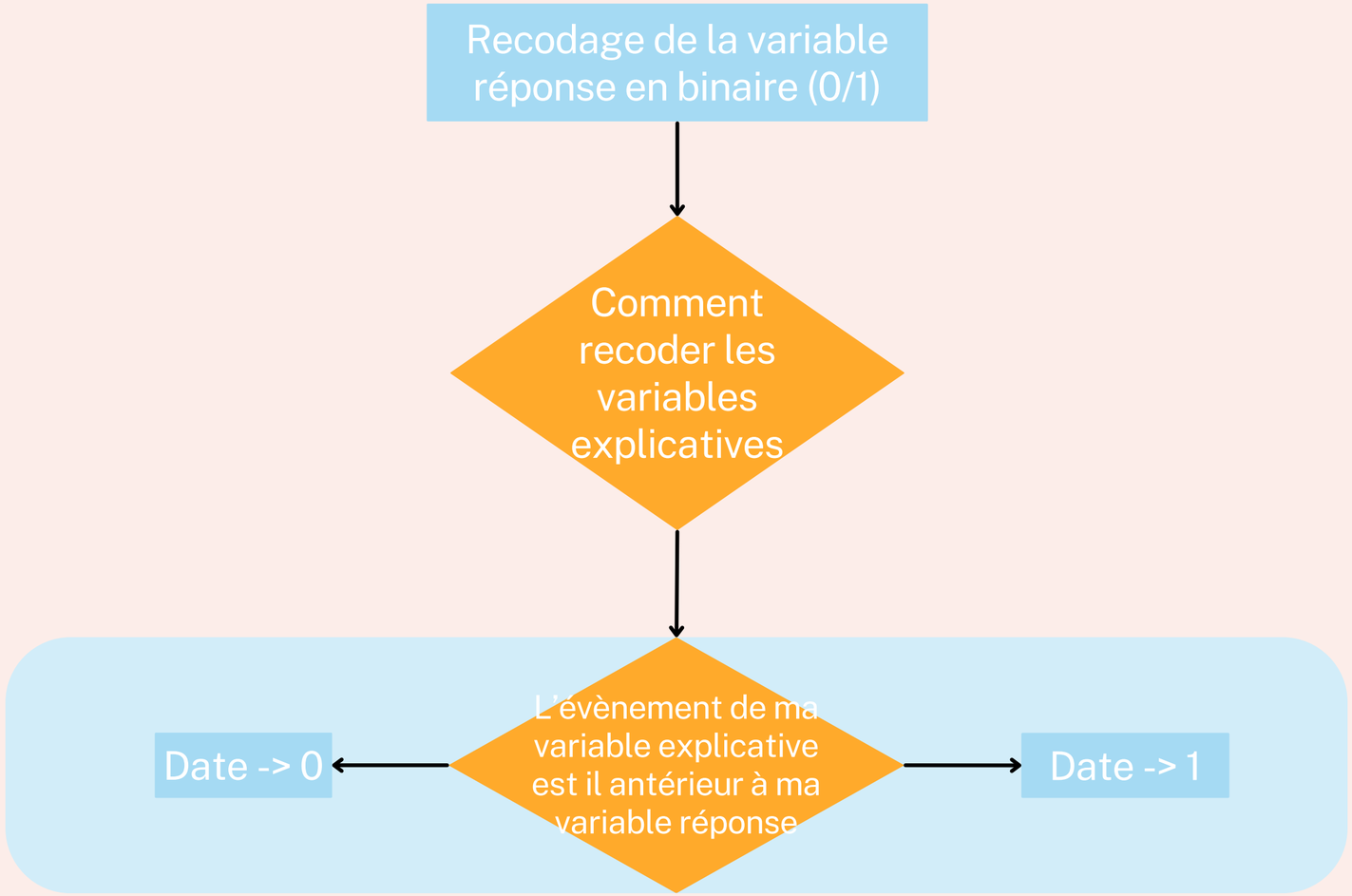
784 308

DÉMARCHE POUR LA MODÉLISATION DES FACTEURS PRÉDICTIFS

Nettoyage de la base de données

Traitement de la base de données

Modélisation



DÉMARCHE POUR LA MODÉLISATION DES FACTEURS PRÉDICTIFS

Variable réponse

Nettoyage de la base de données

Traitement de la base de données

Modélisation

D_variable_X1_dp_dr	D_variable_X2_dp_dr	D_hepatocellular_carcinoma_dp_dr
20/07/2010	NA	NA
25/12/2011	NA	29/01/2012
NA	20/09/2011	24/10/2010
NA	NA	NA

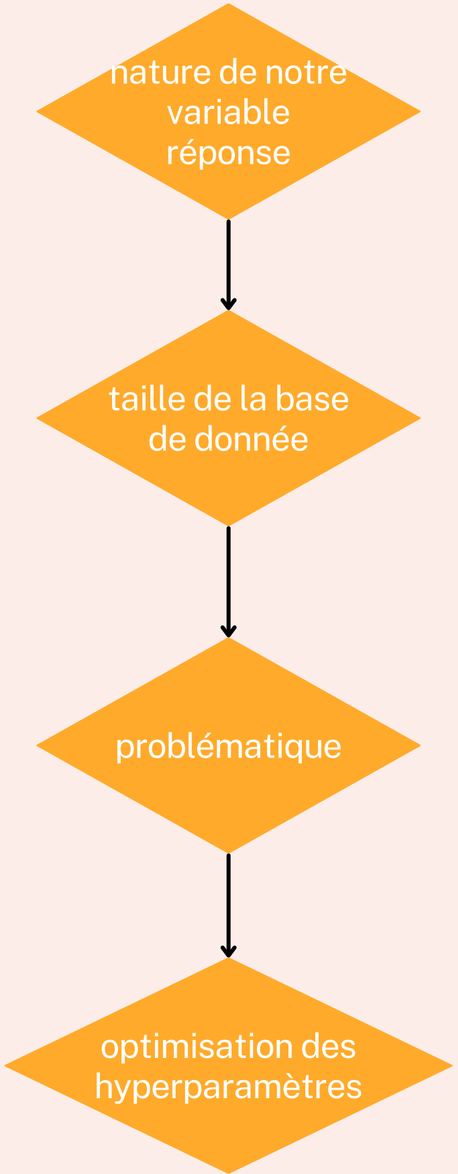
D_variable_X1_dp_dr	D_variable_X2_dp_dr	D_hepatocellular_carcinoma_dp_dr
0	0	0
1	0	1
0	0	1
0	0	0

DÉMARCHE POUR LA MODÉLISATION DES FACTEURS PRÉDICTIFS

Nettoyage de la base de données

Traitement de la base de données

Modélisation



Variable binaire

DÉMARCHE POUR LA MODÉLISATION DES FACTEURS PRÉDICTIFS

Nettoyage de la base de données

Traitement de la base de données

Modélisation

nature de notre variable réponse

taille de la base de donnée

problématique

optimisation des hyperparamètres

Variable binaire

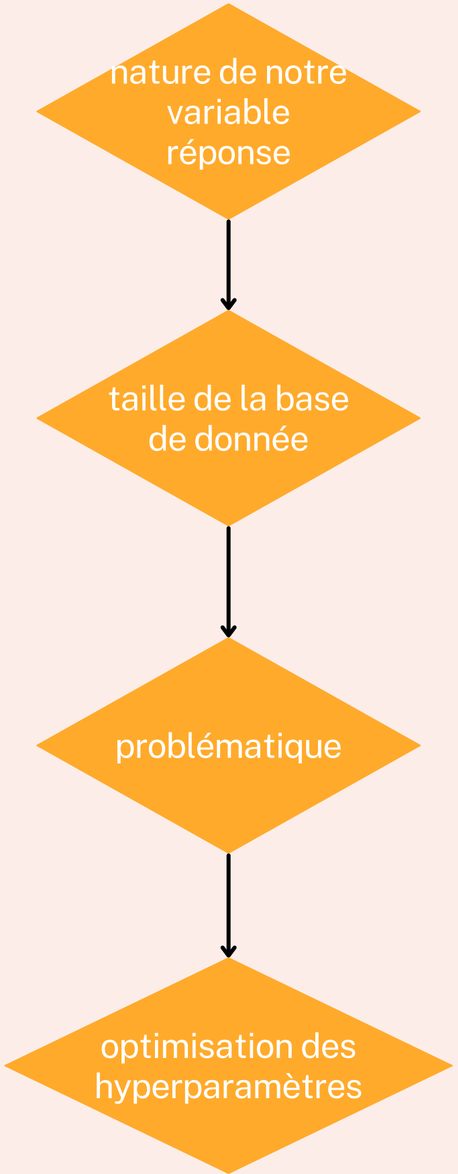
Grande dimension

DÉMARCHE POUR LA MODÉLISATION DES FACTEURS PRÉDICTIFS

Nettoyage de la base de données

Traitement de la base de données

Modélisation



Variable binaire

Grande dimension

Identifier les facteurs d'apparition du CHC chez les DT2

DÉMARCHE POUR LA MODÉLISATION DES FACTEURS PRÉDICTIFS

Nettoyage de la base de données

Traitement de la base de données

Modélisation

nature de notre variable réponse

taille de la base de donnée

problématique

optimisation des hyperparamètres

Variable binaire

Grande dimension

Identifier les facteurs d'apparition du CHC chez les DT2

équilibrage/ajustement du seuil

CHOIX DES ALGORITHMES

Régression log pénalisée par LASSO

- Sélection de variables automatique,
- Bon pour la grande dimensionnalité
- Modèle plus interprétable en identifiant les variables les plus importantes pour la prédiction de la variable réponse

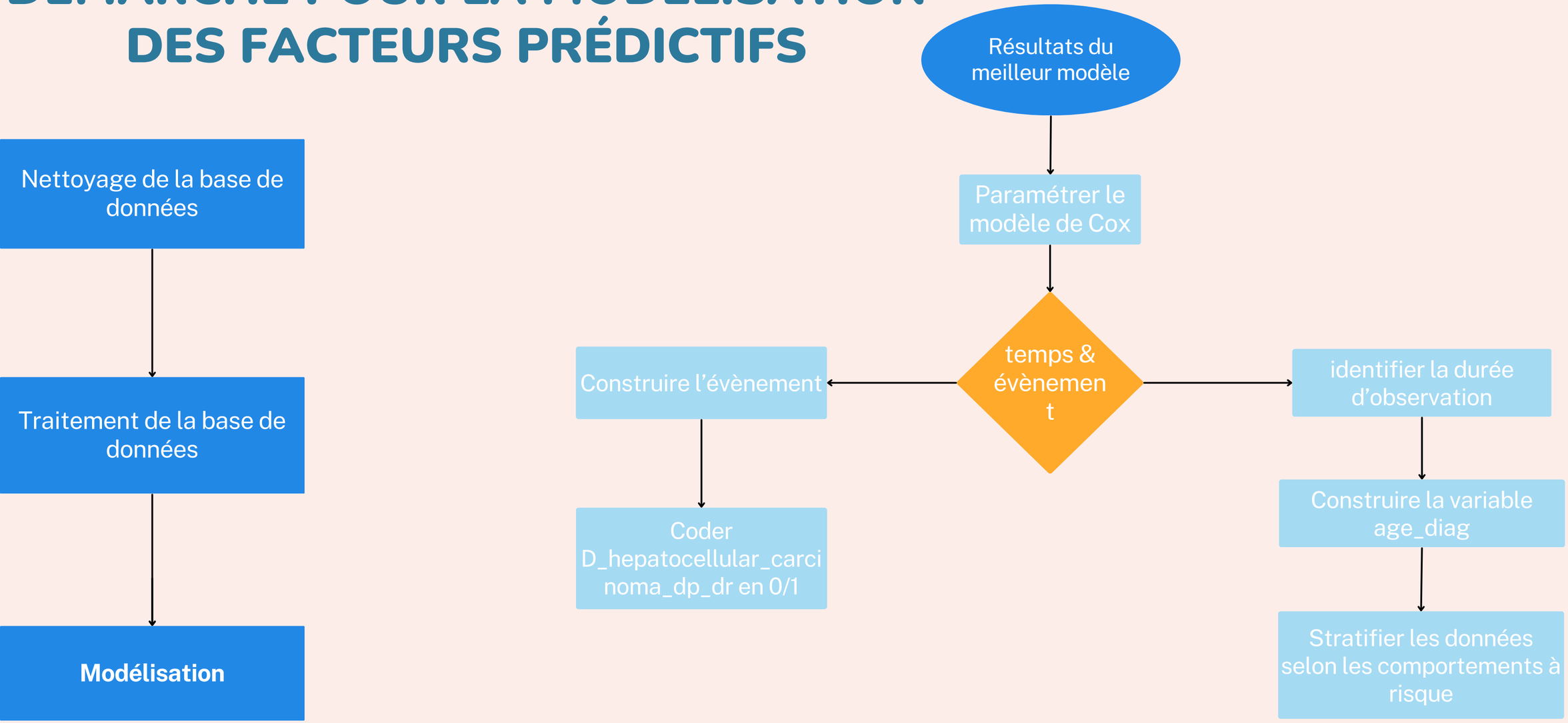
Suppose que les relations entre les variables explicatives et la variable réponse sont linéaires. Si les relations sont non linéaires, le modèle Lasso peut ne pas être le plus approprié.

Random Forest

- Gère les grandes tailles d'échantillons,
- Peut prendre en compte une variété d'équations, y compris des relations non linéaires
- Permet de capturer des modèles plus complexes

Considéré comme une "boîte noire". Difficile d'interpréter les coefficients individuels des variables, car le modèle agrège plusieurs arbres de décision

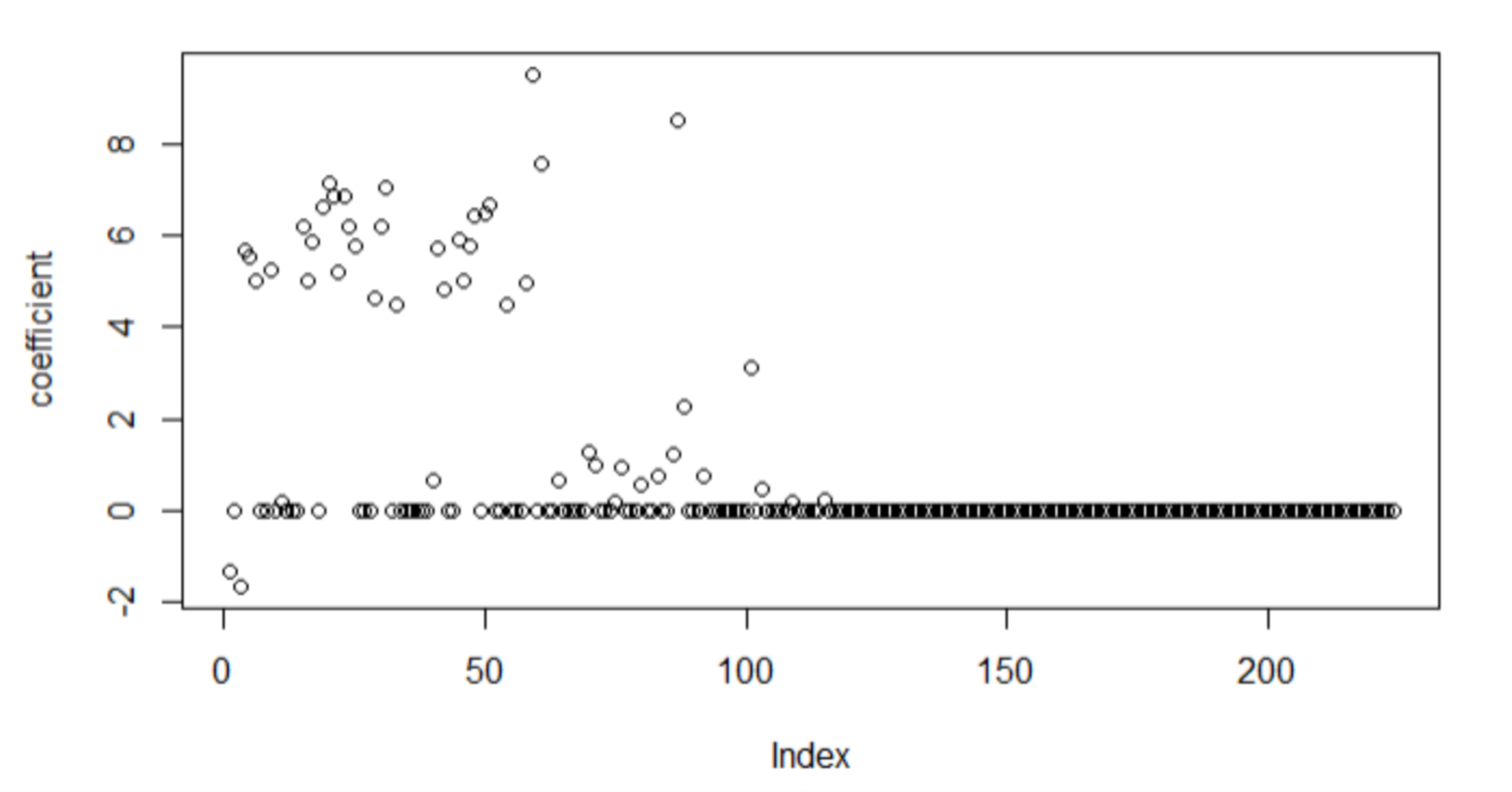
DÉMARCHE POUR LA MODÉLISATION DES FACTEURS PRÉDICTIFS



RÉSULTATS & DISCUSSION

Régression log pénalisée par LASSO

VARIABLES EXPLICATIVES



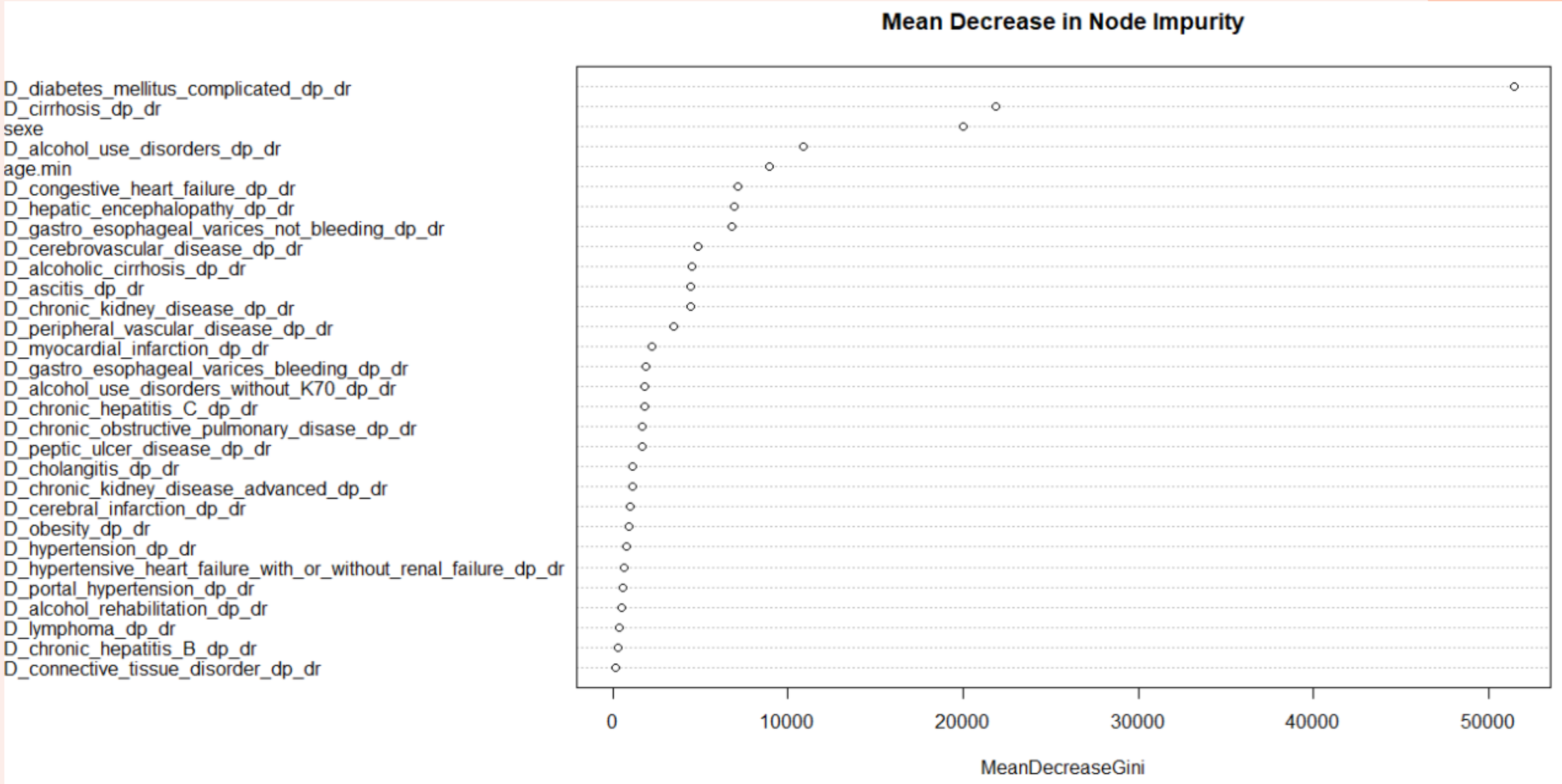
Régression log pénalisée par LASSO

PRINCIPALES VARIABLES EXPLICATIVES

variable explicatives	Coefficient
D_diabetes_mellitus_complicated_dp_dr1	9.04
D_congestive_heart_failure_dp_dr1	7.45
D_decompensated_cirrhosis_dp_dr1	7.32
D_cirrhosis_dp_dr1	7.24
D_chronic_hepatitis_C_dp_dr1	7.16
D_cerebrovascular_disease_dp_dr1	7.15
D_peripheral_vascular_disease_dp_dr1	6.95
D_connective_tissue_disorder_dp_dr1	6.94
D_chronic_obstructive_pulmonary_disase_dp_dr1	6.83
D_peptic_ulcer_disease_dp_dr1	6.78
D_obesity_dp_dr1	6.69
D_myocardial_infarction_dp_dr1	6.68

RANDOM FOREST : VARIABLES D'IMPORTANCE

MEAN DECREASE GINI : IMPACT DE CHAQUE VARIABLE SUR LA RÉDUCTION DE L'IMPURETÉ (GINI IMPURITY) LORS DE LA CONSTRUCTION DES ARBRES DE DÉCISION.



PERFORMANCE DES ALGORITHMES -
Log pénalisée par LASSO

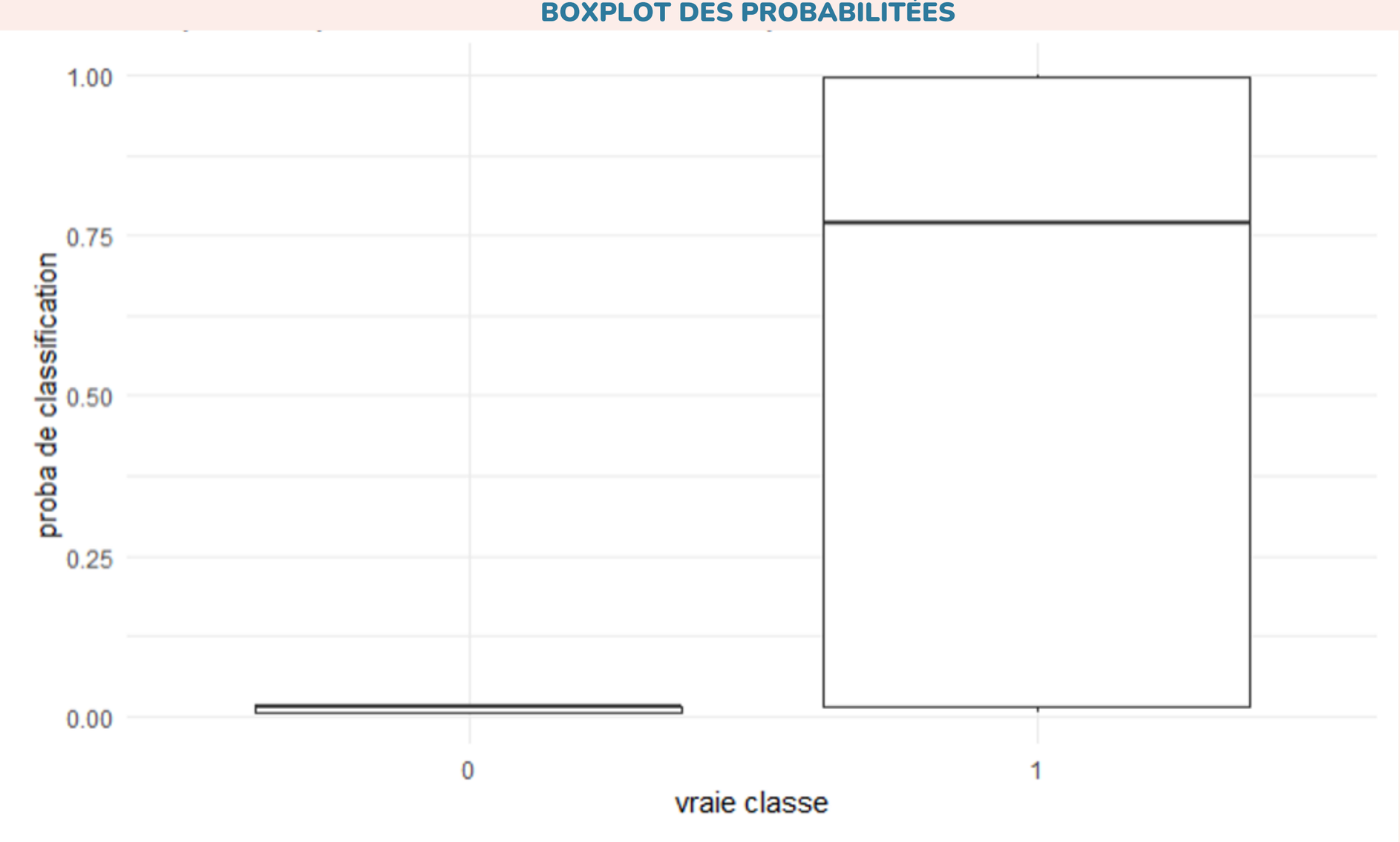
SPÉCIFICITÉ

100%

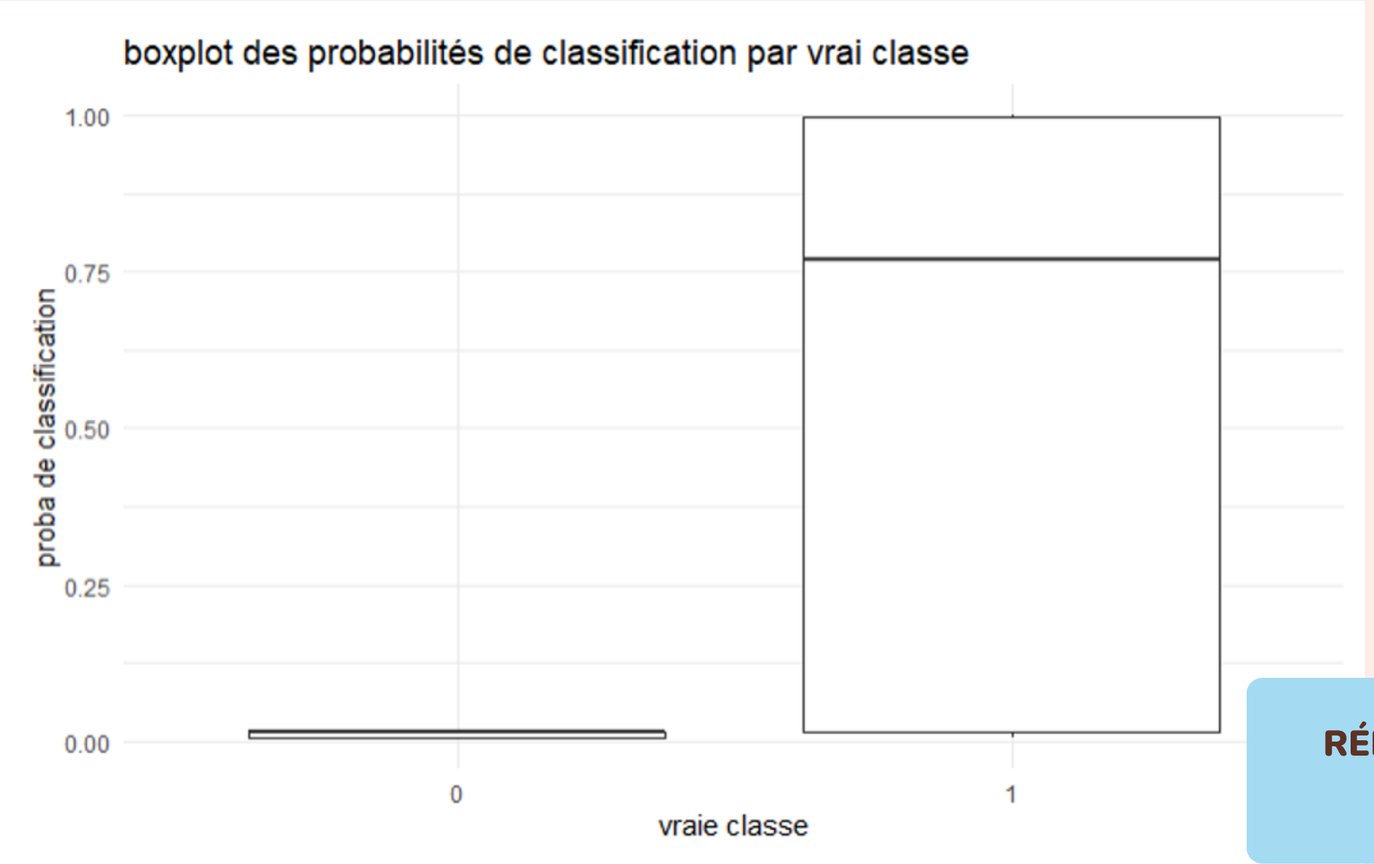
SENSIBILITÉ

53%

PERFORMANCE DES ALGORITHMES - Log pénalisée par LASSO



PERFORMANCE DES ALGORITHMES - Log pénalisée par LASSO



SENSIBILITÉ PERFECTIBLE

Améliorer les performances de prédiction

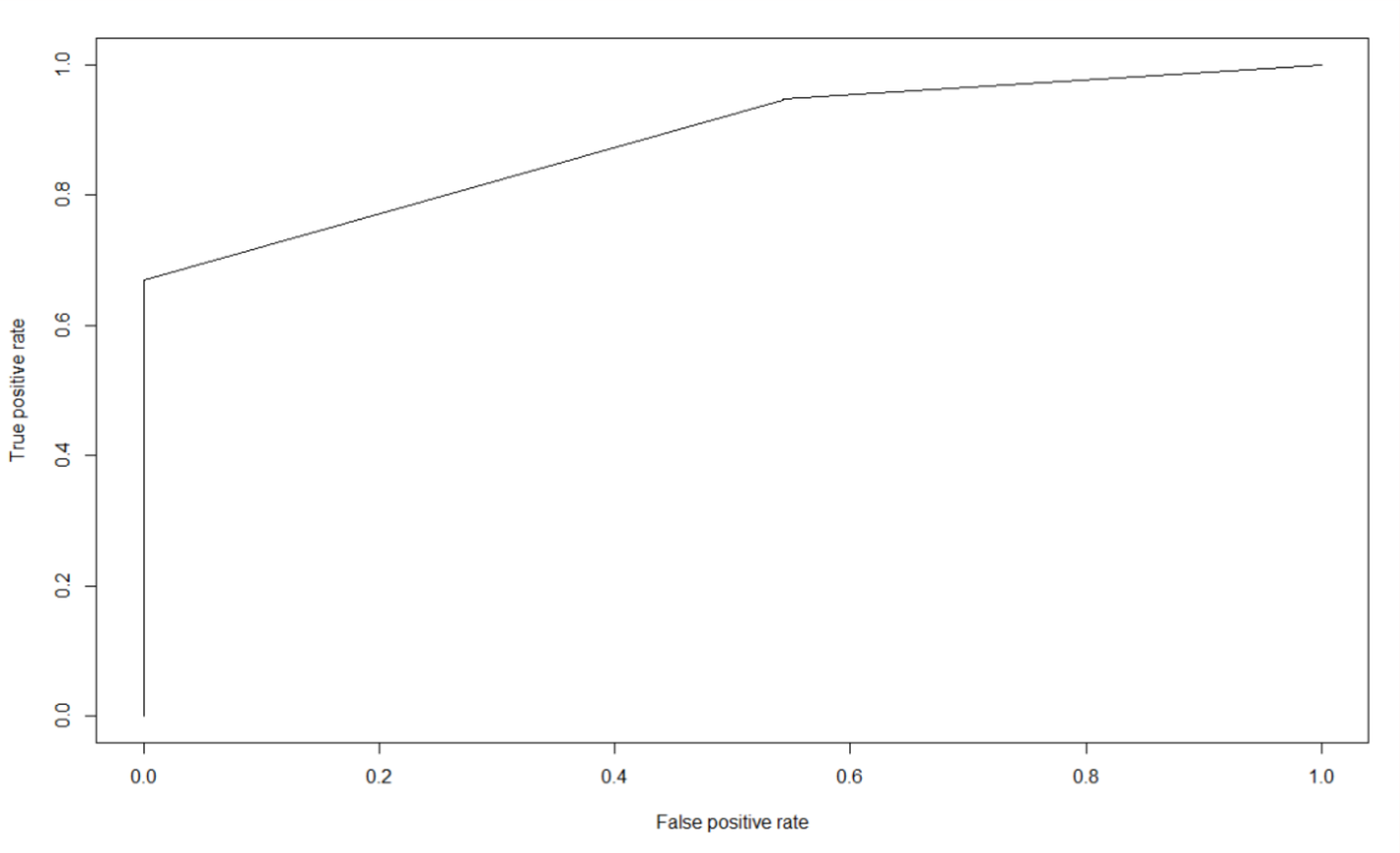
RÉÉQUILIBRER LES CLASSES

MODIFIER LE SEUIL DE CLASSIFICATION

PERFORMANCE DES ALGORITHMES - Log pénalisée par LASSO

CHOIX DU SEUIL OPTIMAL DE CLASSIFICATION

COURBE ROC

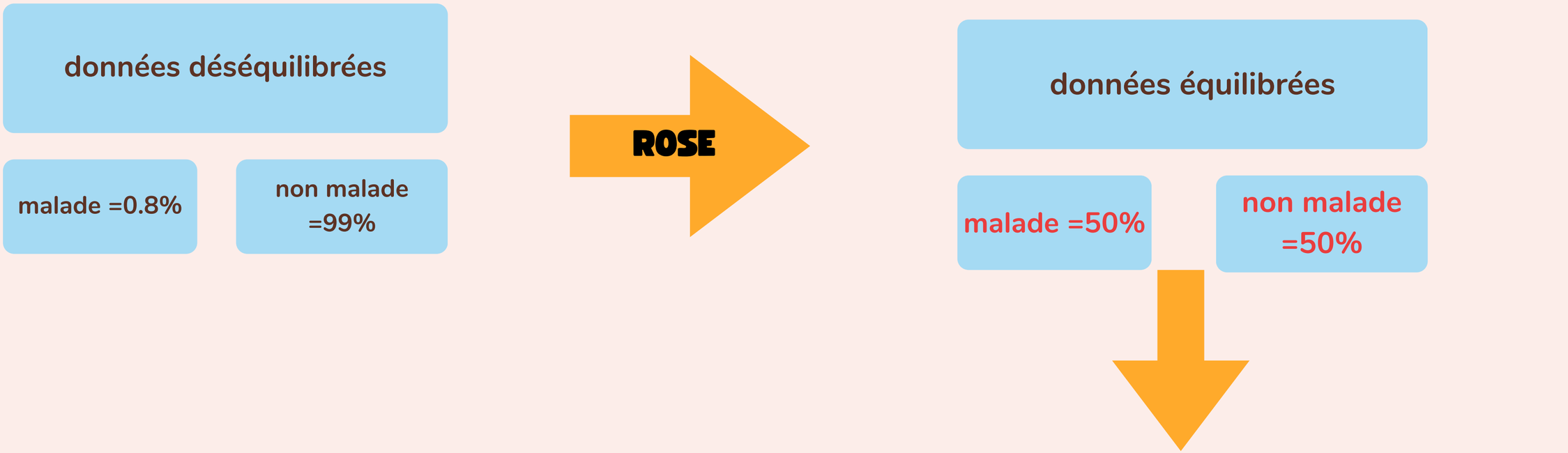


SEUIL OPTIMAL À 0.016

SPÉCIFICITÉ	SENSIBILITÉ
100%	69%

PERFORMANCE DES ALGORITHMES - Log pénalisée par LASSO

RÉÉQUILIBRAGE DES CLASSES



PERFORMANCE DES ALGORITHMES - RANDOM FOREST

SPÉCIFICITÉ

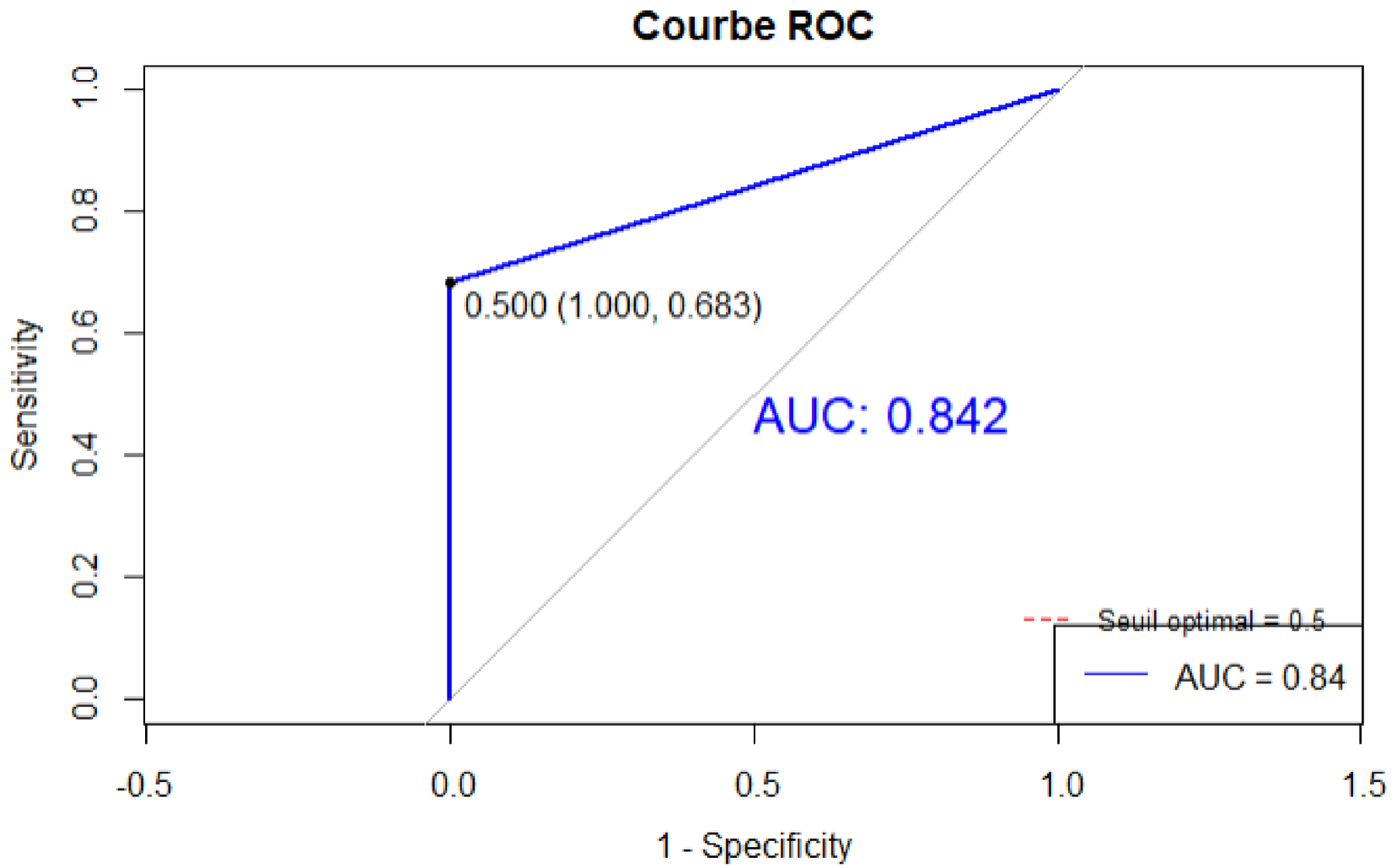
100%

SENSIBILITÉ

68%

Améliorer les performances de prédiction

PERFORMANCE DES ALGORITHMES - RANDOM FOREST



Améliorer les performances de prédiction

PERFORMANCE DES ALGORITHMES - RANDOM FOREST

Rééquilibrage des classes

données déséquilibrées

malade =0.8%

non malade =99%

ROSE

données équilibrées

malade =50%

non malade =50%

SPÉCIFICITÉ

100%

SENSIBILITÉ

68%

PERFORMANCE DES ALGORITHMES

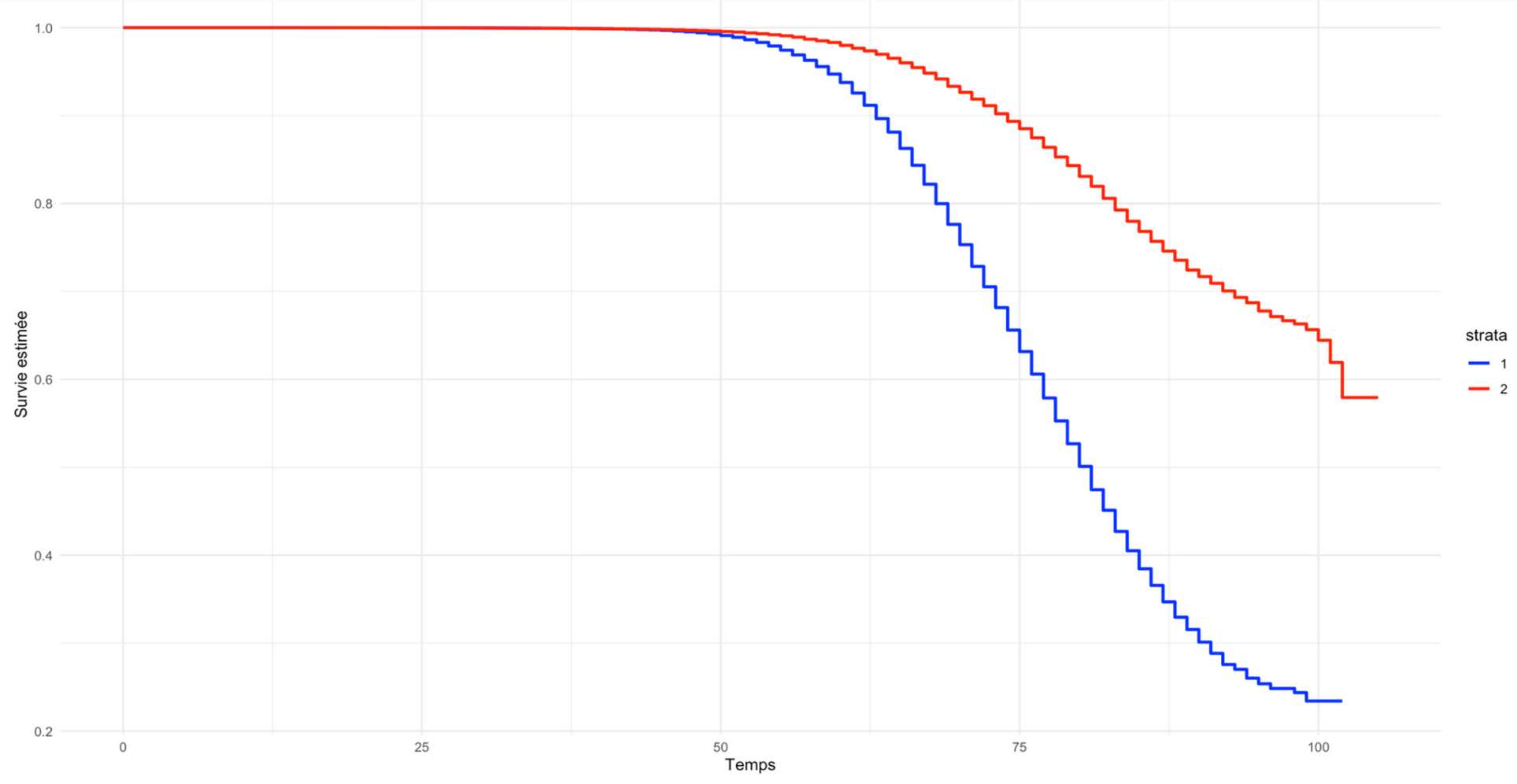
	Régression log pénalisée par LASSO	Random Forest
Sensibilité	70%	68%
Spécificité	100%	100%
Accuracy	83%	99%

MODÈLE DE COX

Variable	Hazard Ratio	% augmentation du risque	Interprétation Narrative
cholangitis	1.07987	7.987%	Risque légèrement augmenté
cirrhosis	3.82754	282.754%	Risque environ quatre fois supérieur
congestive_heart_failure	1.46288	46.288%	Risque environ une fois et demie supérieur
chronic_hepatitis_B	2.67742	167.74%	Risque environ trois fois supérieur
chronic_hepatitis_C	4.79333	379.33%	Risque environ cinq fois supérieur
chronic_hepatitis_D	52.11214	5112.214%	Risque plus de cinquante fois supérieur
decompensated_cirrhosis	1.75823	75.823%	Risque environ une fois et demie supérieur
ascitis	0.80629	-19.371%	Risque légèrement réduit
alcohol_use_disorders	1.76508	76.508%	Risque environ une fois et demie supérieur
alcohol_use_disorders_without_K70	1.69687	69.687%	Risque environ une fois et demie supérieur
cerebrovascular_disease	1.52916	52.916%	Risque environ une fois et demie supérieur
portal_vein_thrombosis	2.11127	111.127%	Risque plus de deux fois supérieur
obesity	3.45222	245.222%	Risque trois fois supérieur
peripheral_vascular_disease	1.28704	28.704%	Risque légèrement augmenté
peptic_ulcer_disease	2.08496	108.496%	Risque plus de deux fois supérieur
diabetes_mellitus_complicated	12.62803	1162.803%	Risque plus de douze fois supérieur

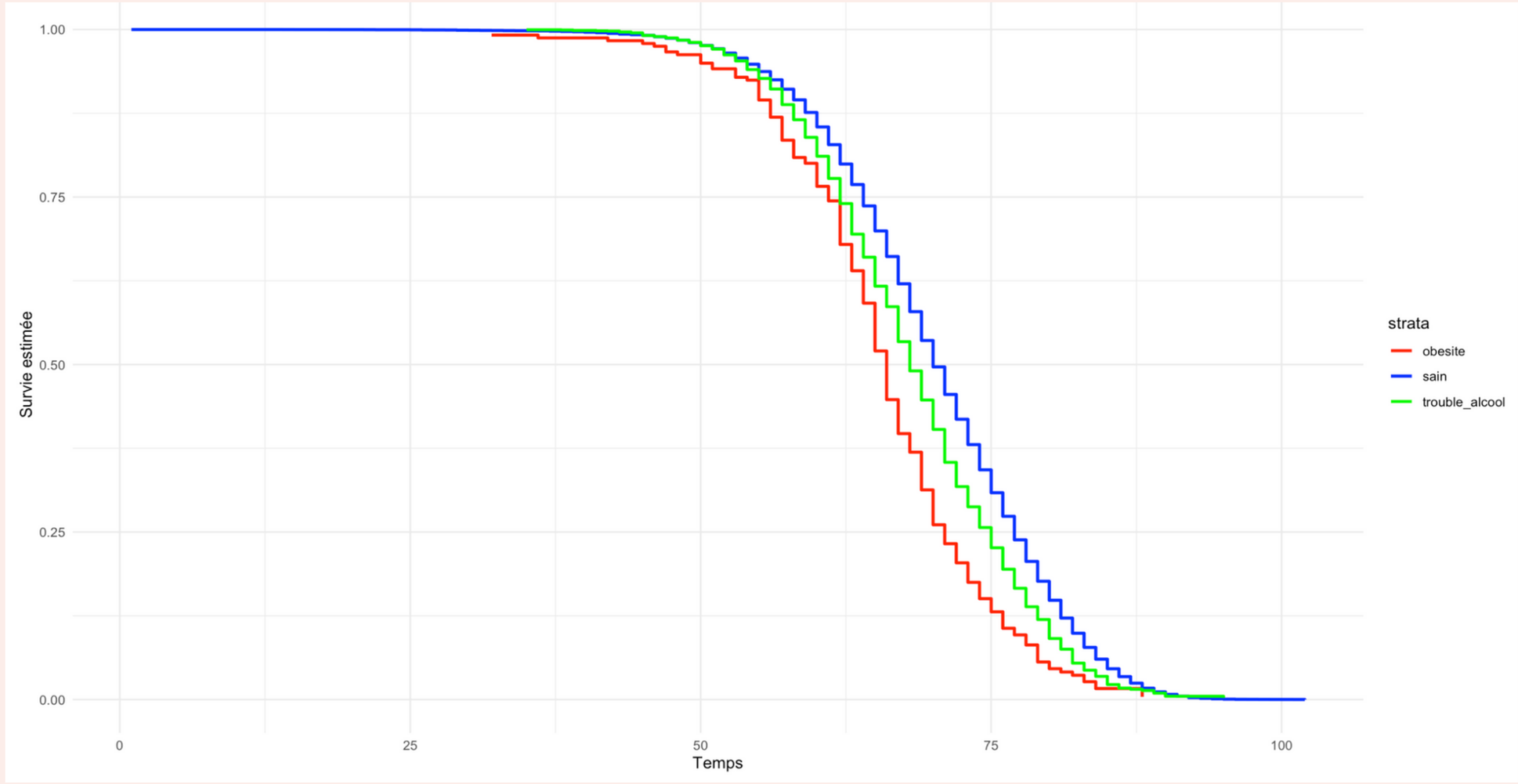
ANALYSE DE SURVIE

PAR SEXE



ANALYSE DE SURVIE

PAR PROFIL



DISCUSSION

Limites & Recommandations

Limites

Défis en termes d'allocation de mémoire lors de l'exécution des algorithmes d'analyse; limitation à la capacité à effectuer une analyse approfondie et exhaustive

Pistes d’amélioration

- optimisation de la gestion de la mémoire : accès à un serveur.
- amélioration de la précision des modèles.
- Validation externe : populations différentes pour évaluer la généralisabilité des conclusions.
- discussion avec un professionnel de santé

CONCLUSION

QUELS SONT LES FACTEURS CONTRIBUANTS À L'APPARITION DU CHC CHEZ LES PATIENTS ATTEINTS DE DIABÈTE DE TYPE 2 ?

Notre modélisation
complication diabète
hépatite C
obésité
cirrhose
hépatite B
obstruction veine porte
ulcère chronique
trouble alcoolique
maladies cerebrovasculaires
maladies cardiovasculaires

VS

Dans la littérature scientifique
hépatite B
hépatite C
obésité
consommation d'alcool
diabète sucré
obstruction veine cave inférieure
obstruction veine hépatique

ASSISTANCE PUBLIQUE HÔPITAUX DE PARIS

Thank you!

Karim Krache Hadirou Tamdamba