Revue et comparaison de méthodes pour l'identification de facteurs pronostiques sur petits échantillons, Un exemple sur une étude de vie réelle

Journées de Biostatistiques, 17 et 18 Novembre 2022

Lisa Mounier, Cyril Esnault, Julien Dupin, David Pau, Alexandre Civet

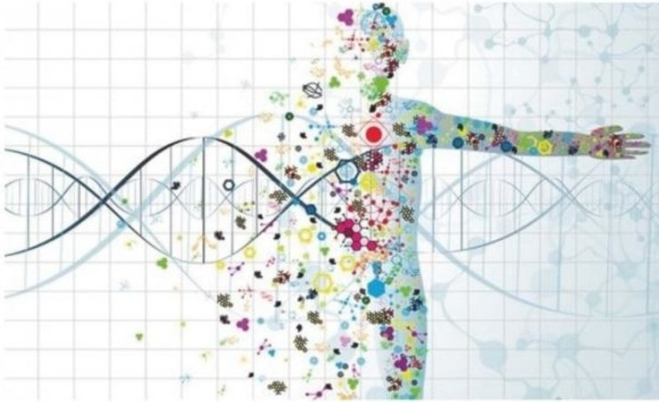
Sommaire

- ❏ Contexte autour de la médecine de précision
- ❏ L'étude de vie réelle REALM
- ❏ Présentation de la revue de littérature sur les méthodes d'identification de facteurs pronostiques combinant des approches de Biostatistiques et de Data Science
- ❏ Présentation des biais et opportunités existantes pour les petits échantillons de données
- ❏ Données et modélisations
- ❏ Comparaisons des méthodes testées
- ❏ Conclusion

Contexte autour de la Médecine personnalisée

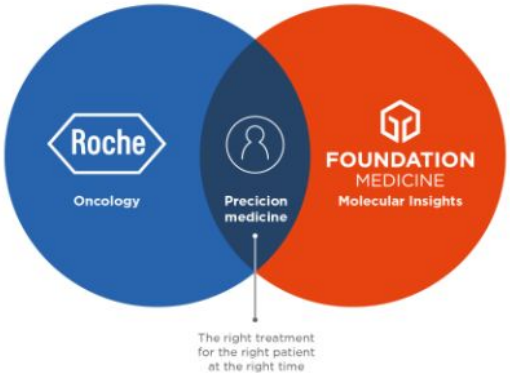
Chaque cancer est unique (les mutations et le nombre de mutations diffèrent d'un individu à un autre)

Besoin en oncologie d'avoir un traitement adapté aux mutations génétiques en fonction du type de cancer.

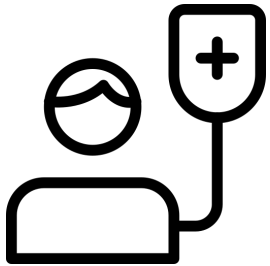


Test FMI commercialisé par Roche :

- ❏ Réalisation d'un profilage génomique contenant les détails sur la tumeur du patient.
- ❏ Les thérapies et essais cliniques à envisager sont proposés par le test FMI, et peuvent être mis en place par les équipes médicales



L'étude REALM



Population

Sous-population provenant d'une base de données pan-tumeurs contenant 416 patients
Etude des 48 patients atteints d'un cancer des voies biliaires et ayant subi un test FMI (sous-population plus homogène)

Outcome

Actionnabilité :
la proposition d'un traitement a été faite aux patients en fonction de l'altération génomique détectée

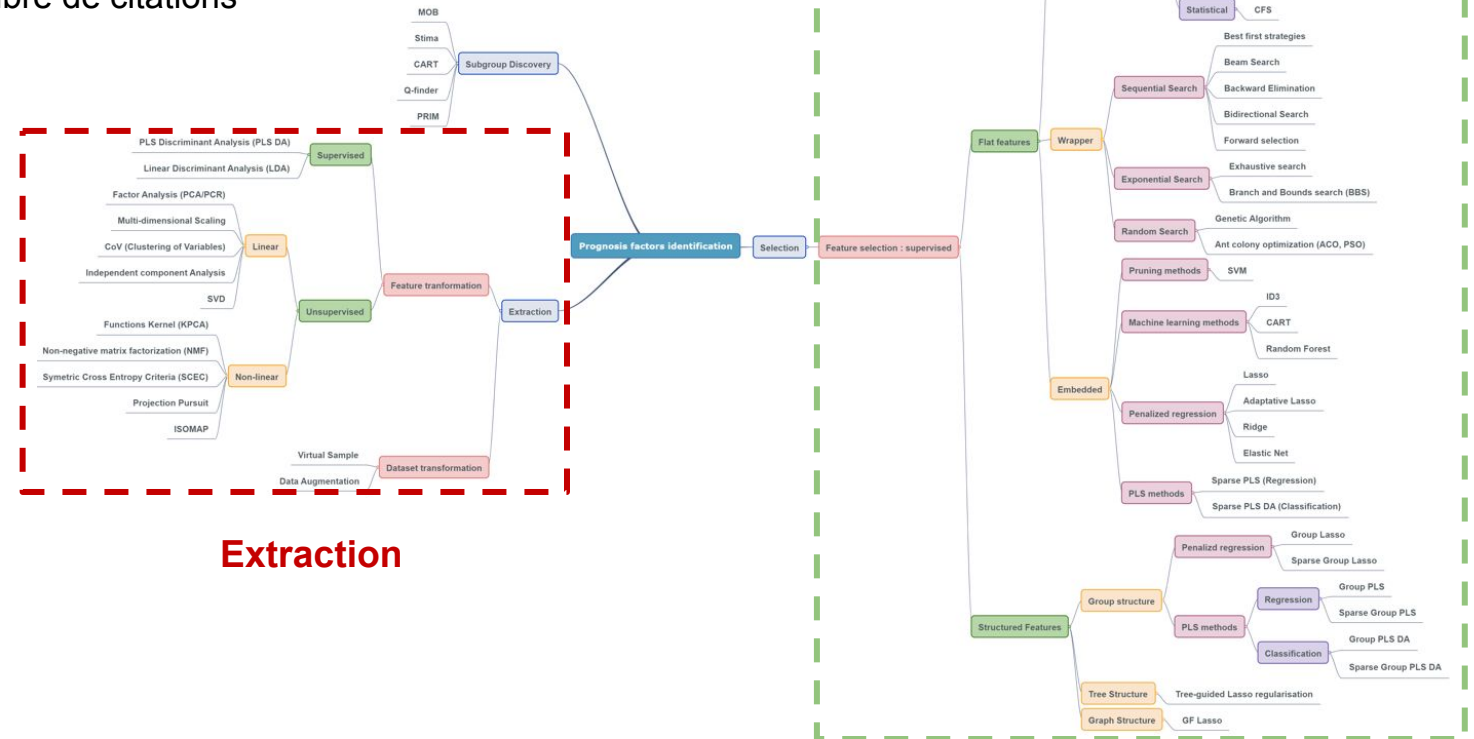
Objectifs

Déterminer les facteurs pronostiques de l'actionnabilité

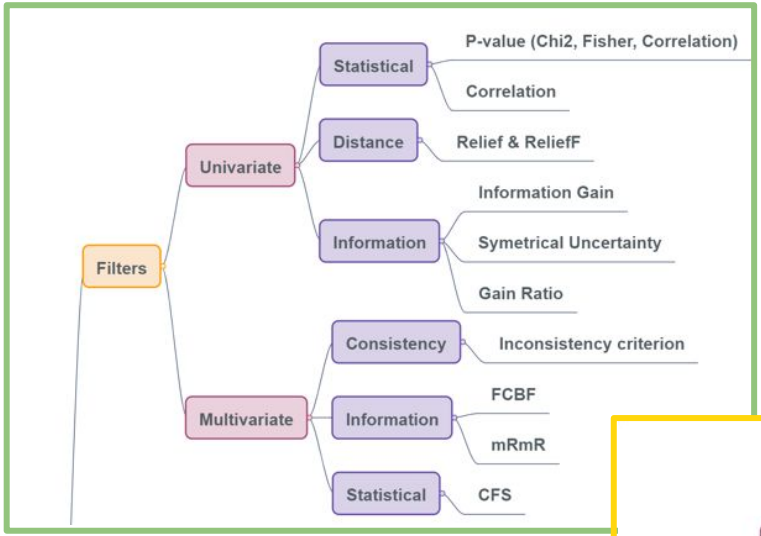
Synthèse des méthodes pour l'identification de facteurs pronostiques

Critères utilisés pour la sélection d'articles:

- Mots clés
- Date de publication
- Nombre de citations

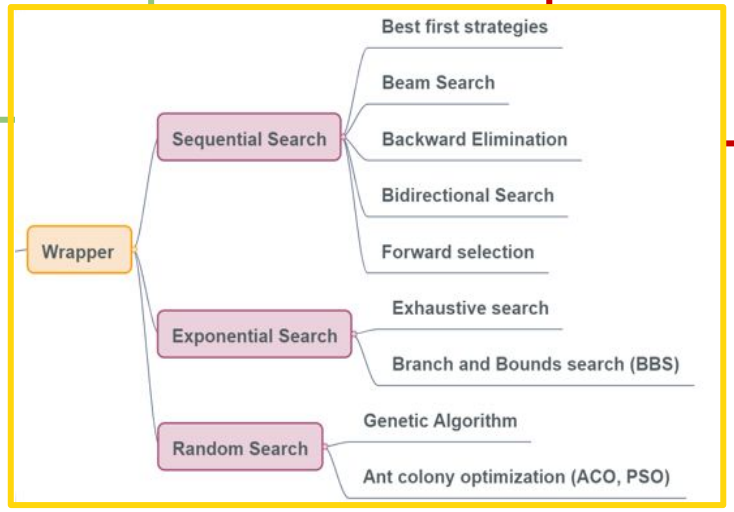


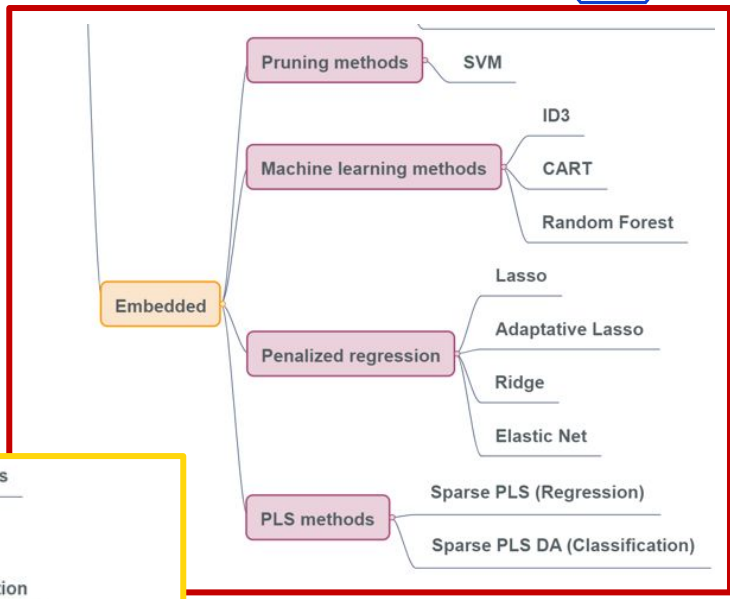
Focus sur les méthodes de Feature Selection



Filters

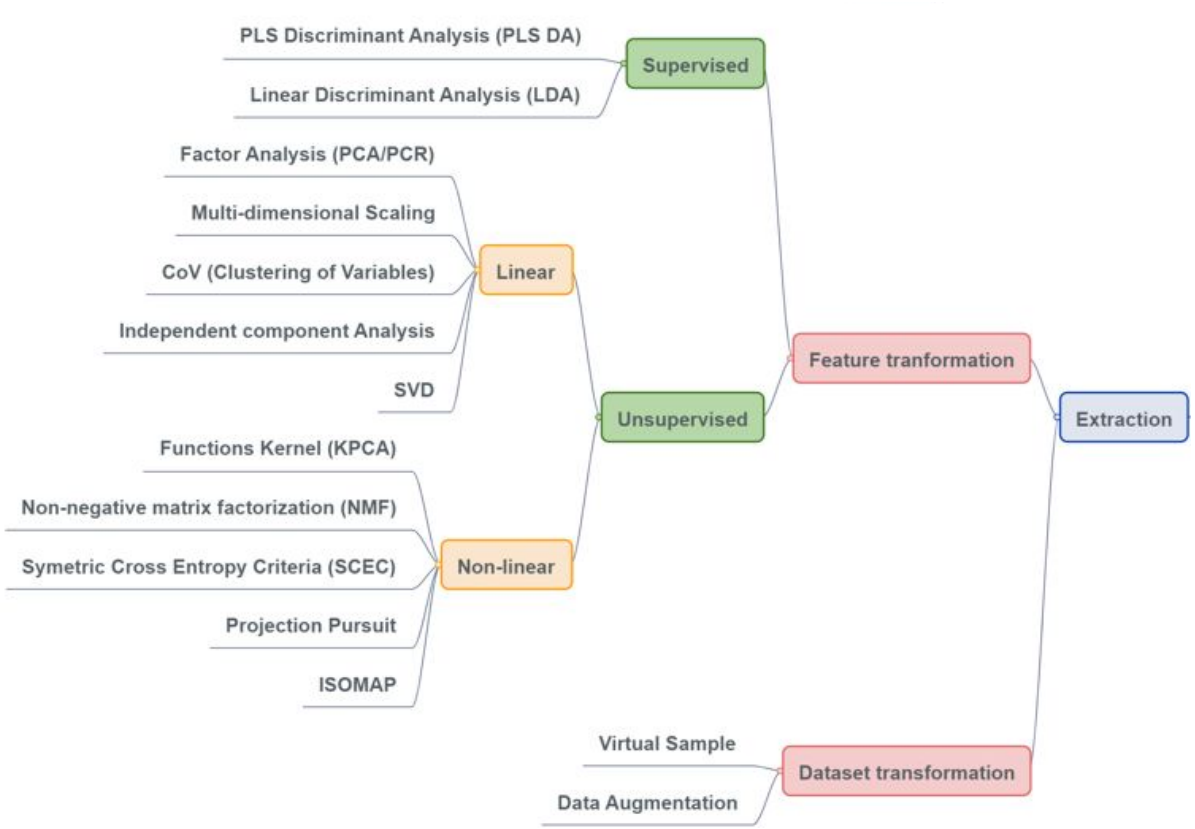
Wrapper

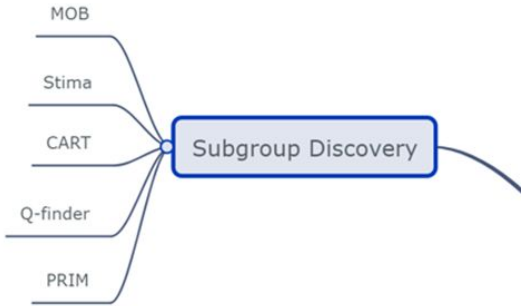




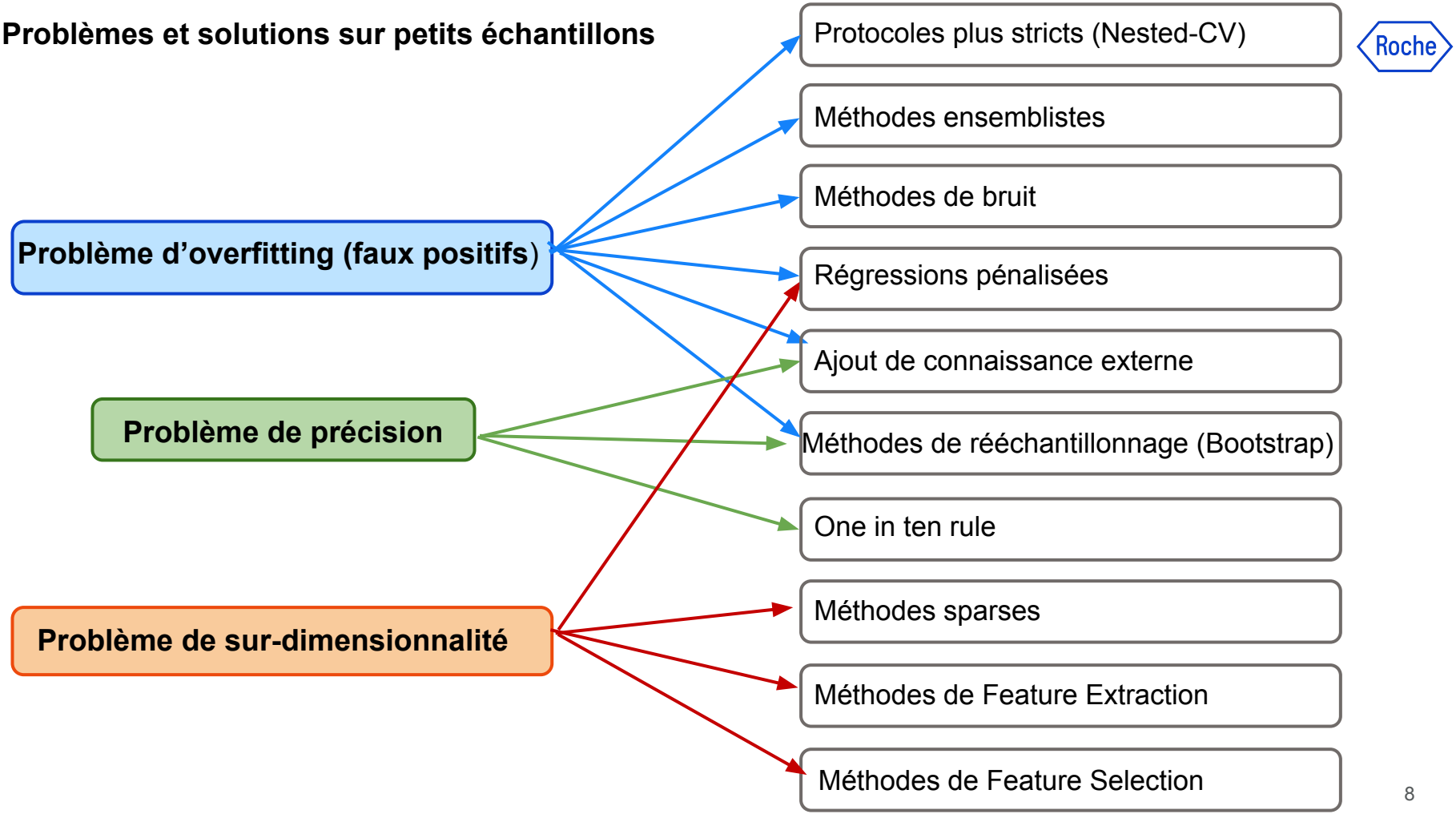
Embedded

Méthodes d’extraction et découverte de sous-groupes





Problèmes et solutions sur petits échantillons



Roche

Données et Modélisation

Variables explicatives

- Sexe
- Classe d'âge
- ECOG
- Classe T
- Classe N
- Statut métastatique
- Temps entre le diagnostic et le test

Variable à expliquer

Actionnabilité

Méthodes de Feature Selection avec comme objectif de challenger la méthode de "référence"

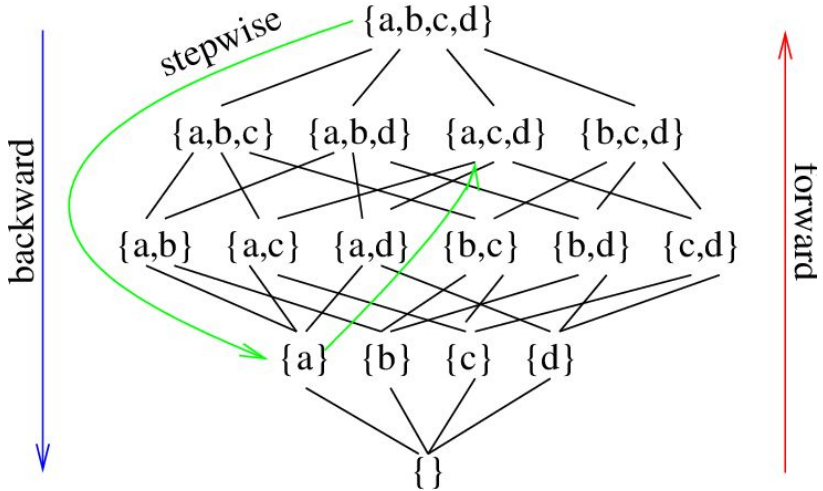
Modélisations effectuées

- Régression logistique avec approche Bidirectionnelle
- Régression logistique avec approche exhaustive
- Régression pénalisée LASSO
- Arbre de décision
- Forêts aléatoires

Analyse de référence (‘Méthode stepwise’)

Méthodologie

- Régression logistique univariée avec p-value < 0.25
- Régression logistique multivariée avec une approche Bidirectionnelle, p-value < 0.10



Forward Selection
+
Backward Elimination

=
Bidirectional Search
(‘stepwise’)

Facteurs pronostiques identifiés

ECOG : p-value < 0.001
Classe d’âge : p-value = 0.010
Statut métastatique : p-value = 0.088

Présentation des méthodologies testées pour challenger la méthode de référence

Approche exhaustive (Wrapper)

- ❏ Régression logistique univariée avec p-value < 0.25
- ❏ Régression logistique multivariée avec une approche exhaustive de sélection de variables

Stratégie n°1
Minimisation de l'AIC

Stratégie n°2
Maximisation de l'AUC
Validation croisée (K = 5)

	Métrique	Variables
Stratégie 1	AIC : 16.03	ECOG, Classe d'âge, Statut métastatique, classe T, Temps entre le diagnostic et le test
Stratégie 2	AUC : 0.94	ECOG, Classe d'âge, Statut métastatique, classe T, Temps entre le diagnostic et le test

Régression Lasso (Embedded)

- ❏ Recherche des coefficients qui minimisent la contrainte suivante :

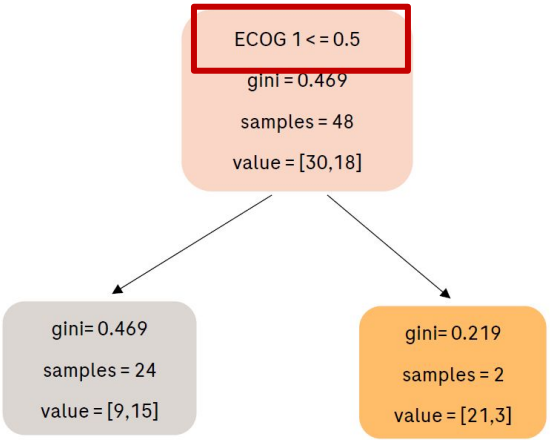
$$L_{lasso}(\hat{\beta}) = \sum_{i=1}^{n}(y_i - x_i^T\hat{\beta})^2 + \lambda\sum_{j=1}^{m}|\hat{\beta}_j|$$

- ❏ Métrique utilisée : AUC
- ❏ Validation croisée avec K=4 pour le paramètre de pénalité

	coef	OR
Intercept	−0.43	0.65
age<60 (ref:age>=60)	0.49	1.63
Metastatic cancer T0 (ref:No)	0.31	1.36
ECOG.1 (ref:ECOG.0)	−1.54	0.21
ECOG.2 (ref:ECOG.0)	−0.15	0.86
T.T2 (ref:Tx)	0.63	1.87
N.N0 (ref:Nx)	0.69	2
N.N2 (ref:Nx)	0.19	1.21

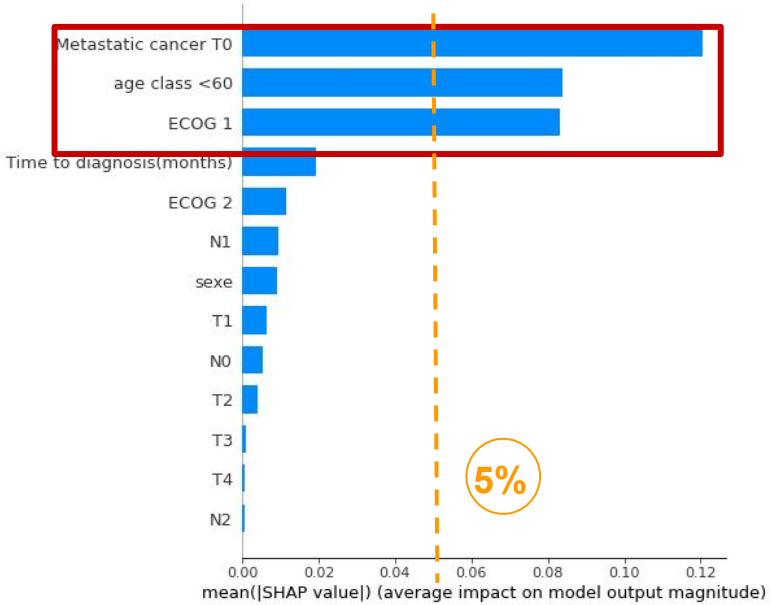
Arbre de décision

- Critère de Gini
- Optimisation des hyperparamètres par grille (GridSearch) avec comme contrainte au moins 4 patients dans les dernières feuilles
- Validation croisée avec K=5
- Métrique AUC (AUC moyen obtenu = 0.74)



RandomForest

- Optimisation des hyperparamètres par grille : 250 arbres, profondeur de 3 niveaux, et un minimum de 4 patients par dernière feuille.
- Validation croisée avec K=5
- Métrique AUC (AUC moyen obtenu = 0.81)

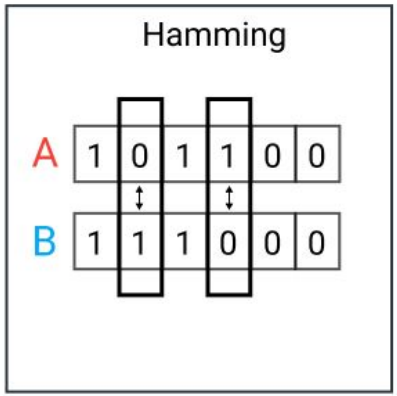


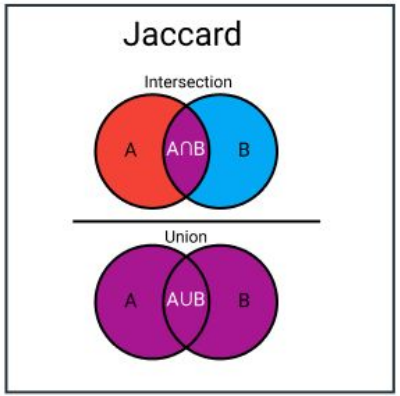
Comparaisons des méthodes

1. Choix méthodologique

Prognostic factors	Stepwise	Exhaustive AIC	Exhaustive AUC	Lasso	Decision tree	Random forest	Frequency
Gender	No	No	No	No	No	No	0.0
Age in class	Yes	Yes	Yes	Yes	No	Yes	0.8
ECOG	Yes	Yes	Yes	Yes	Yes	Yes	1.0
Metastatic cancer	Yes	Yes	Yes	Yes	No	Yes	0.8
Class T	No	Yes	Yes	Yes	No	No	0.6
Class N	No	No	No	Yes	No	No	0.2
Time from diagnosis to test	No	Yes	Yes	No	No	No	0.4

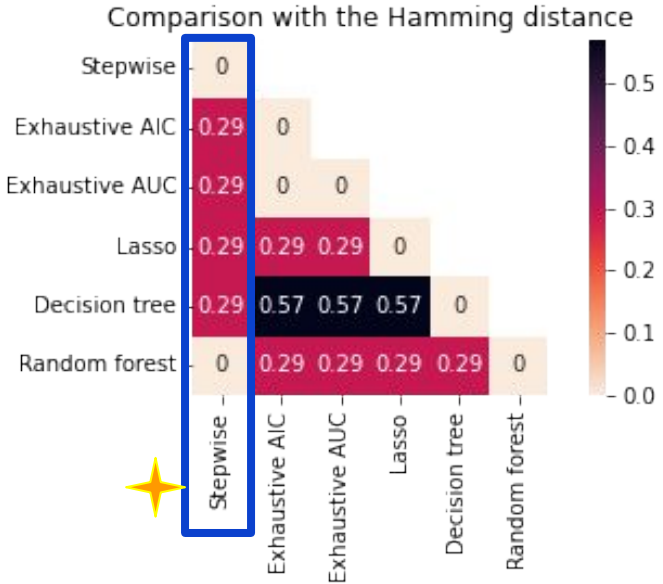
Quelle mesure choisir ?

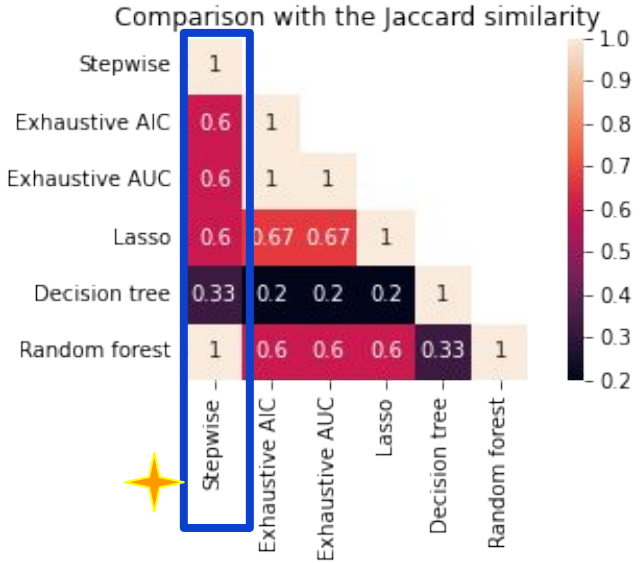




Comparaisons des méthodes

2. Résultats





3. Limites

- Le set de variables utilisé est restreint, laissant moins de place à la variabilité entre les méthodes
- Face à la diversité des méthodes, que donnerait une approche bayésienne ? une méthode d'échantillonnage ?

Conclusion

Roche

Sur l'étude REALM

Les facteurs pronostiques de l'actionnabilité les plus identifiés :

- ❏ Le statut métastatique
- ❏ L'ECOG
- ❏ La classe d'âge

Sur le plan méthodologique

- ❏ Certains facteurs sont identifiés que par certaines méthodes, montrant que le set de facteurs pronostiques identifié est dépendant de la méthodologie choisie.

- ❏ Tester différentes approches dans le cas des petits échantillons est nécessaire pour accroître la confiance en les résultats et pour ne pas être méthodologie-dépendant ! Attention, certaines approches semblent plus appropriées que d'autres…

Merci pour votre attention !

Mes remerciements à Alexandre Civet, Cyril Esnault, Julien Dupin et David Pau pour leurs accompagnements.