
JIB
65e ÉDITION
JOURNÉES DE L'INNOVATION EN BIOLOGIE
LA BIOLOGIE AU SERVICE DU PROGRÈS MÉDICAL
WWW.JIB-INNOVATION.COM
1 - 2 DÉCEMBRE 2022
PALAIS DES CONGRÈS DE PARIS FRANCE

1-2 DÉCEMBRE 2022 | PALAIS DES CONGRÈS DE PARIS FRANCE

JIB | 65e ÉDITION JOURNÉES DE L'INNOVATION EN BIOLOGIE

LA BIOLOGIE AU SERVICE DU PROGRÈS MÉDICAL

Redwood : Un nouveau méta-learner spécifique au gène *FBN1*

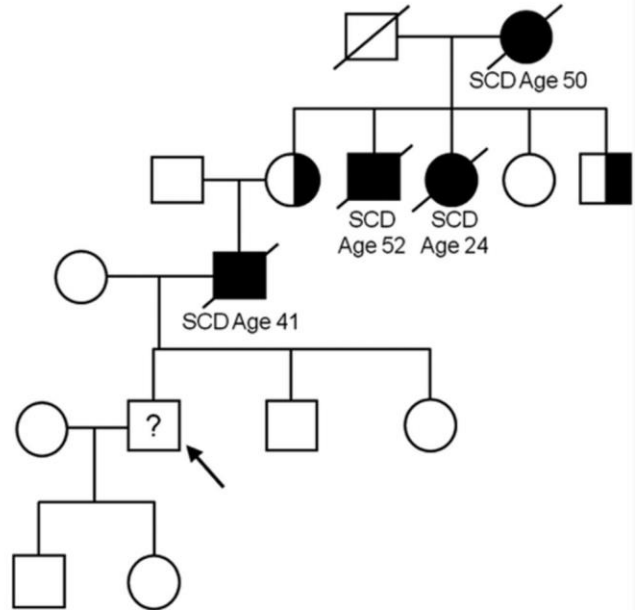
Victor Gravrand, Pauline Arnaud, Caroline Kannengiesser, Catherine Boileau, Nadine Hanna

AP-HP. Nord Université Paris Cité

LNTS UMRS 1148 Laboratory for Vascular Translational Science

FBN1, fibrillin-1 et Syndrome de Marfan

- 65 exons, 60 domaines fonctionnels
- Protéine structurale de la matrice extracellulaire



- **Maladie rare, de forte morbi-mortalité**
- Intérêt du diagnostic génétique : prise en charge
- Majoritairement causée par des variants faux-sens
- **Problématique : mauvaises performances des algorithmes de prédiction**

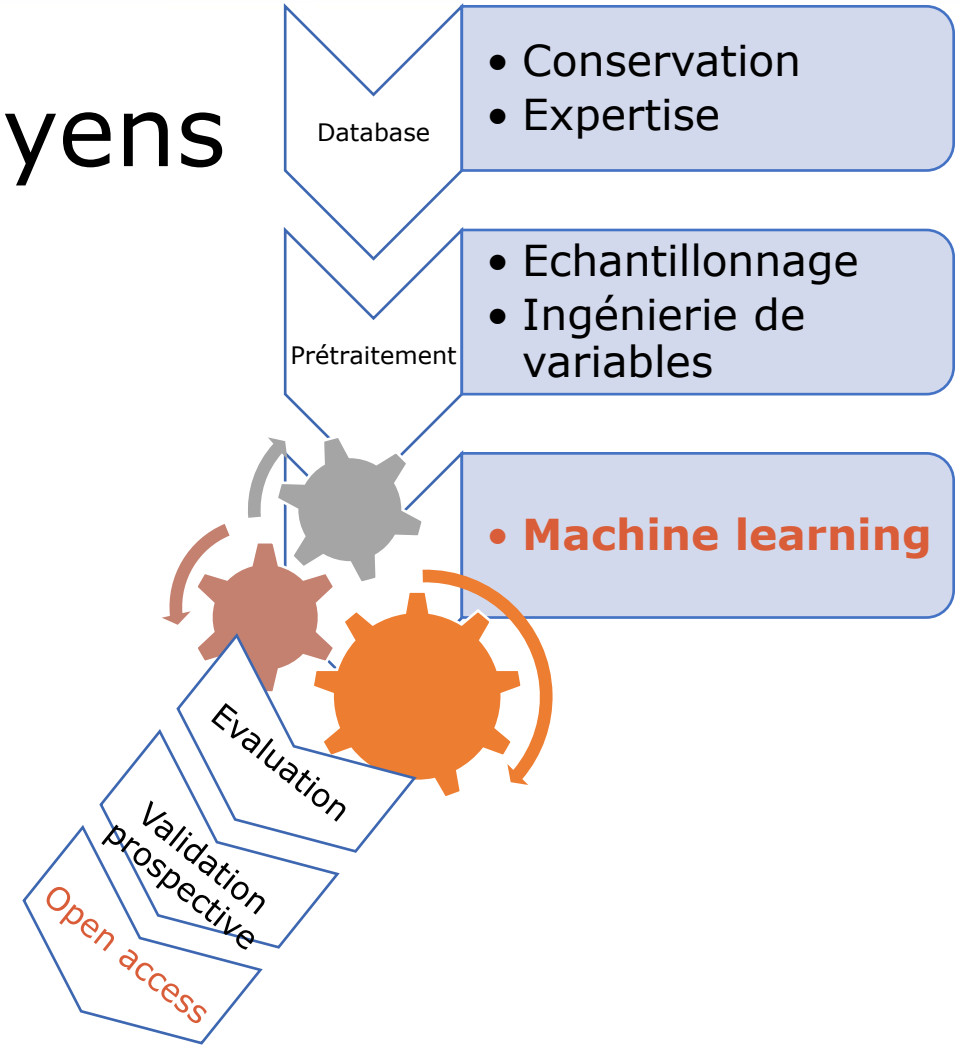
Enjeux du généticien moléculaire

- Variation bénigne ou pathogène ?
- Les critères ACMG :
  - Séparation en critères bénins et pathogéniques
  - Séparation en niveau de preuve (*strong, moderate, supportive,…*)
  - Séparation par type d'arguments (*fréquence allélique dans la pop. générale, localisation dans un hotspot mutationnel, prédictions in silico, …*)

- Exemples :
  - Fréquence allélique > 5% dans la population générale -> Bénin
  - Impact délétère selon les algorithmes prédictifs -> *Plutôt pathogène ?*

Objectifs et moyens

- **Créer notre algorithme prédictif :**
  - Performant aux standards attendus
  - Intégrant notre expertise du gène
  - Cohérent avec une utilisation en routine diagnostique



Création du jeu de données

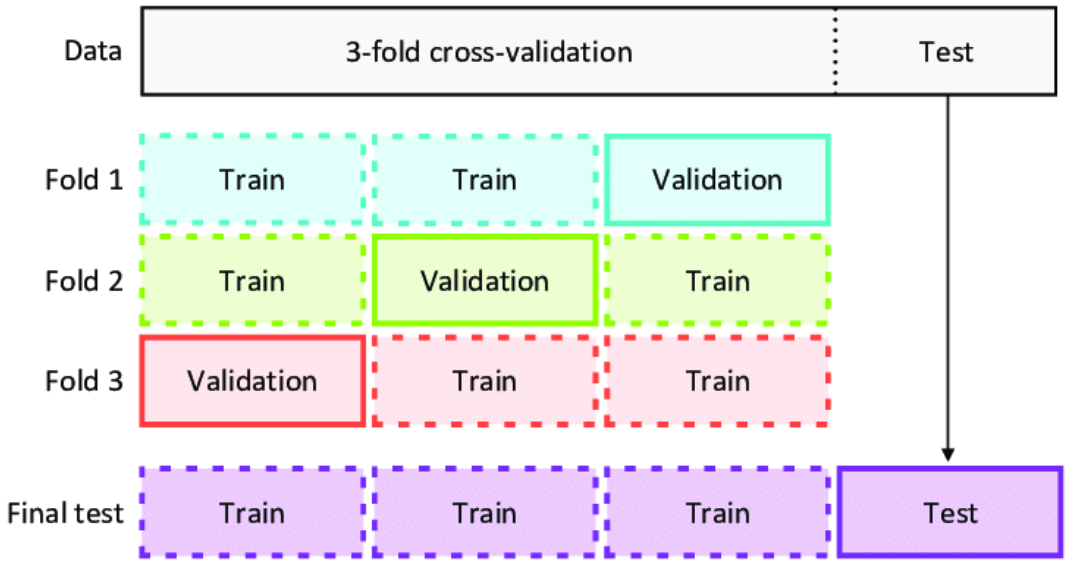
- Pour toutes les variations faux-sens possibles sur *FBN1* (*dbNSFP*)
  - *Récupération des prédictions individuelles des algorithmes majeurs de prédiction*
  - *Récupération des données de conservation*
  - ***Exigence de non-utilisation de l'information de fréquence allélique***

- Labellisation pour l'apprentissage supervisé :
  - Variants pathogènes :
    - Identifiés au LBMR Marfan-Bichat **et** non décrits dans ClinVar (*n = 223*)
  - Variants bénignes :
    - Décrits bénins dans ClinVar **ou**
    - > 5% de fréquence allélique **ou**
    - connus bénins LBMR (*n = 72*)

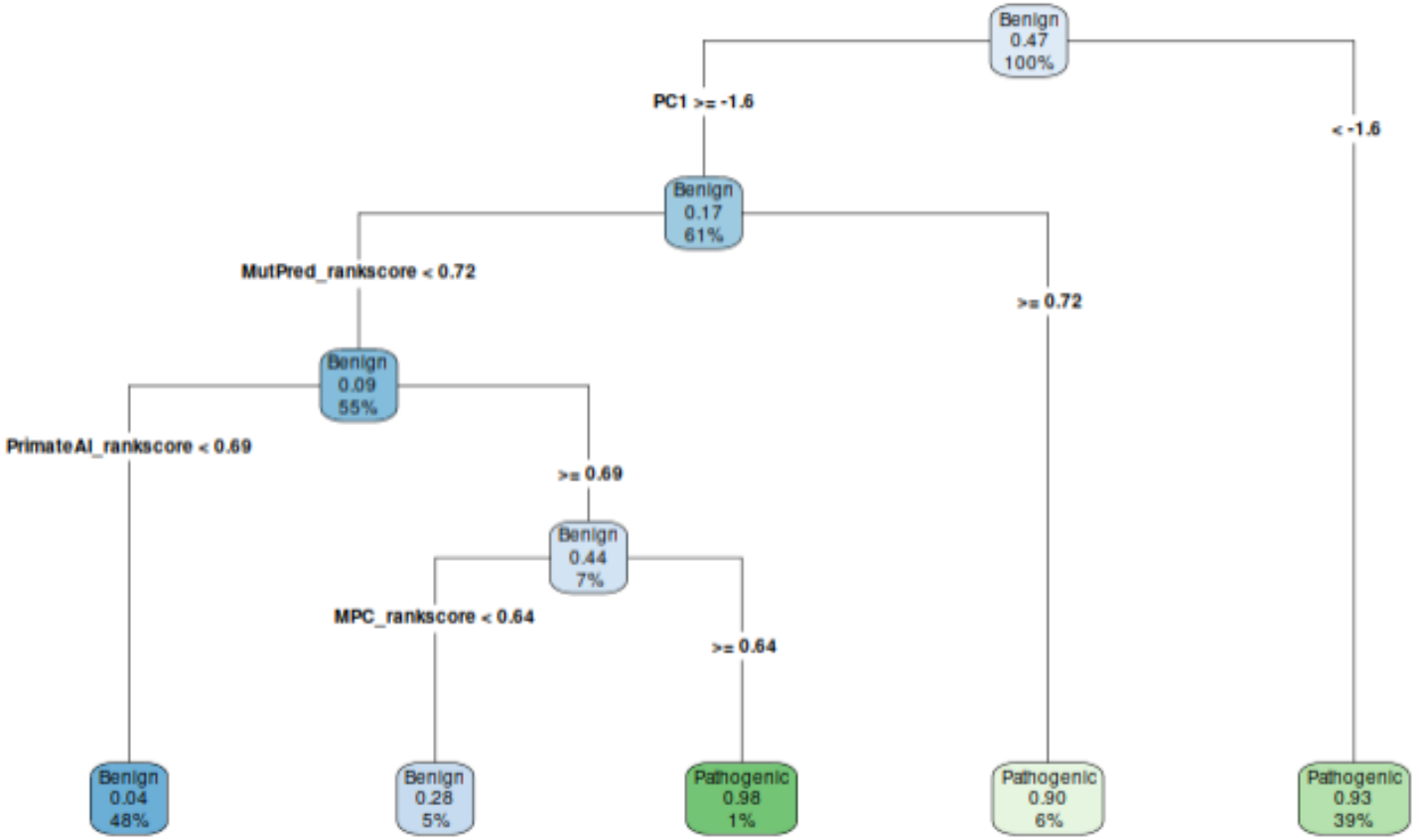
Prétraitement des données

- Allocation du budget de données en deux composantes :
  - 75% dédiés au **jeu d'apprentissage** (*n* = 221)
    - Décomposé en 10 blocs pour validation croisée
  - et 25% à celui du **jeu d'évaluation** (*n* = 74)

- Déséquilibre de classe modulée par suréchantillonnage synthétique de la classe minoritaire (*SMOTE*)
- Création de 12 variables supplémentaires **reflétant la réflexion des biologistes responsables du LBMR**

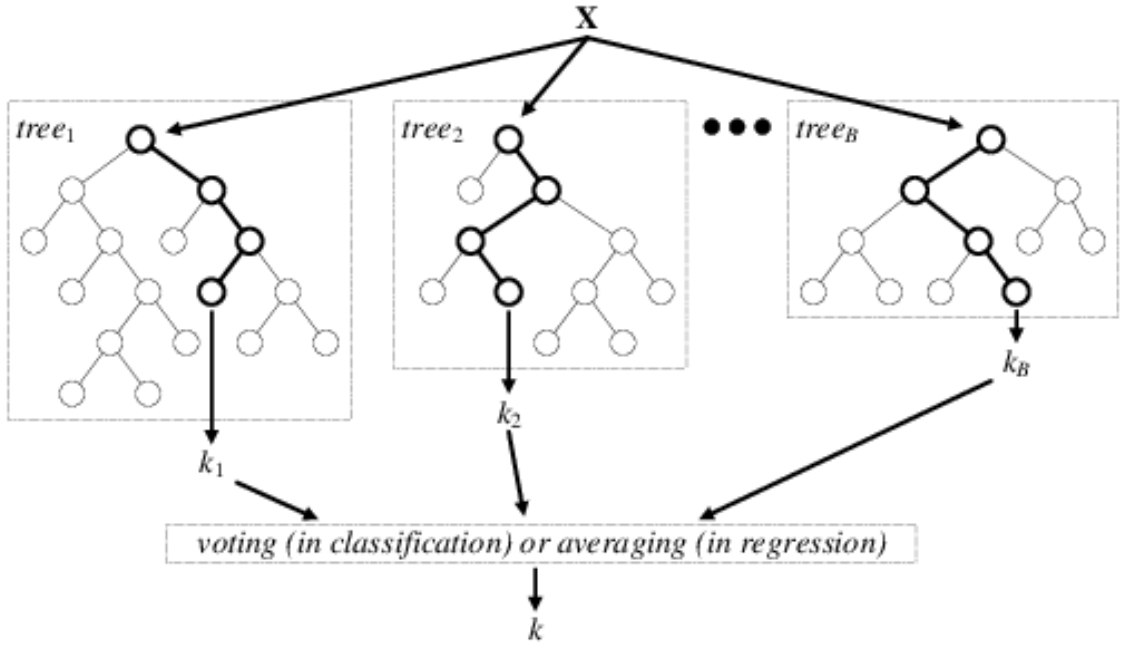


L'arbre archétypal de notre future forêt



Sous le capot de **Redwood**

- Les forêts aléatoires :
  - Plante un arbre décisionnel
  - Celui-ci se voit proposé aléatoirement m variables pour dichotomiser à chaque nœud
  - Sépare l'espace jusqu'à ce que les branches ne soient plus que des feuilles
    - (x < i observations/branche)
  - Répète l'expérience pour 1001 arbres



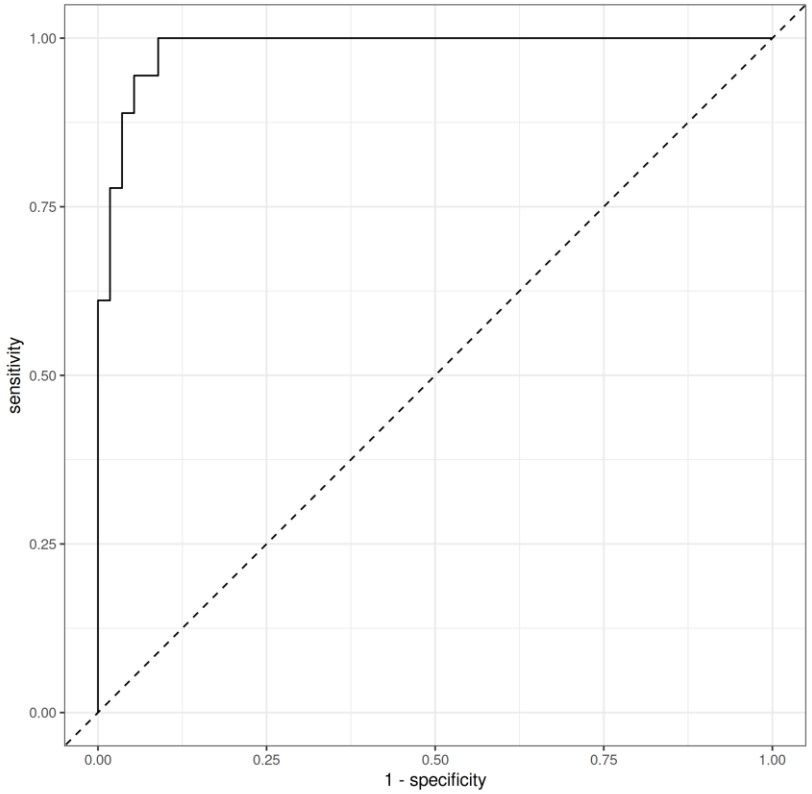
- La forêt vote ensuite sur la classe, **la majorité l'emporte.**
- *L'équivalent en machine-learning d'un staff*

Performances

	Pathogènes	Bénins
Prédits Pathogènes	54	2
Prédits bénins	2	16

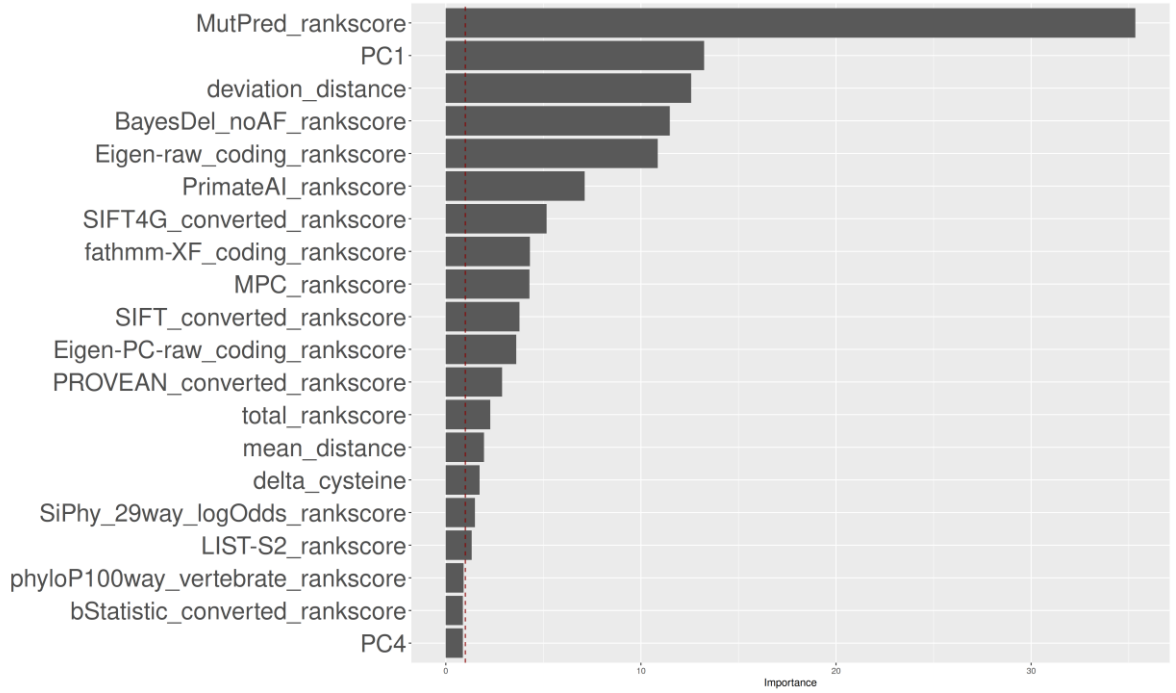
- Actuellement, **Redwood est top 3 sur 11 des 12 métriques testées**
  - **AUROC** 0,985 vs max 0,987 *MetaRNN*
  - **AUPRC** 0,995 vs max 0,996 *MutPred*
  - **F-score** 0,964 vs max 0,973 *BayesDel-AddAF*

- Par des phénomènes de *circularité*, les performances des autres algorithmes sont sur-évaluées.

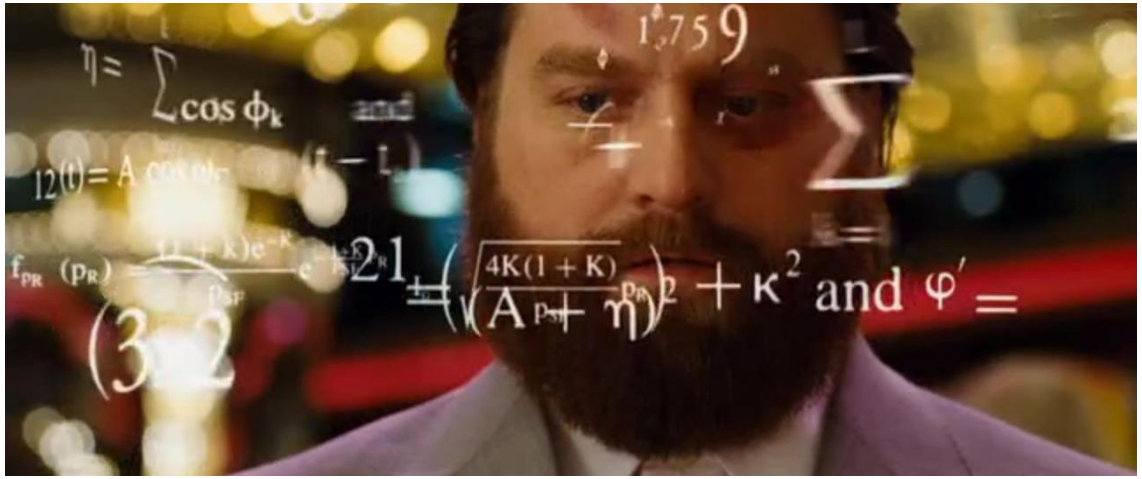


Perspectives

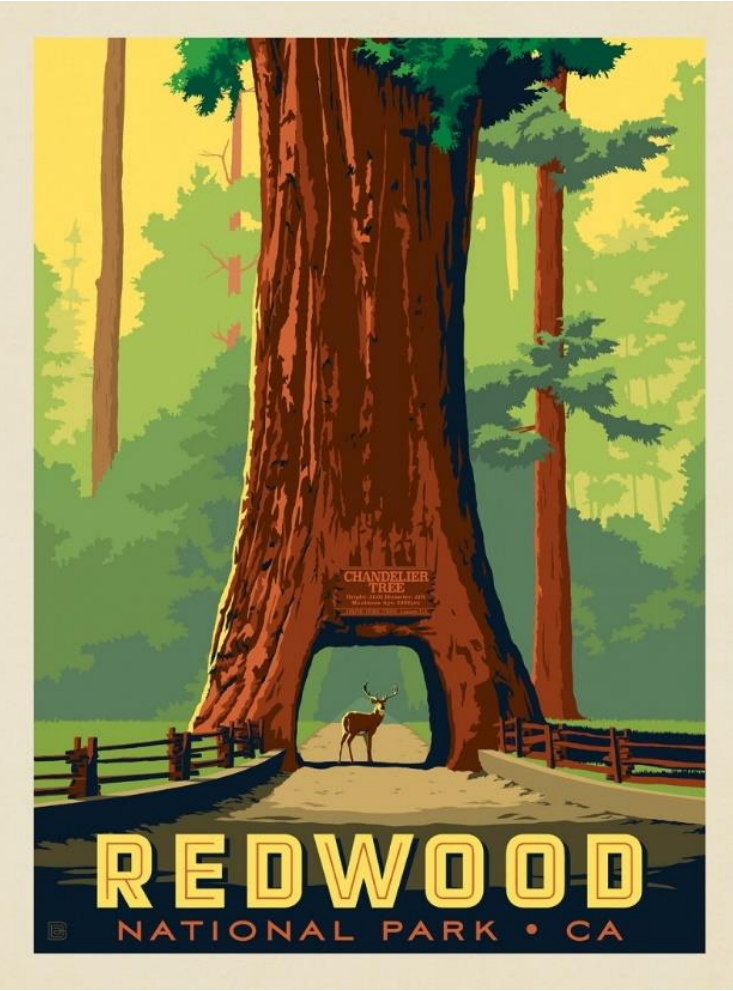
- Optimisation du *SMOTE*
- Evaluation d'autres types de modèles (Lasso, **GBM, NN**, …)
- Travail sur l'interprétabilité de Redwood
  - *C'est à son tour de nous enseigner des choses !*
- **Validation prospective**
- **Extension de la prédiction jusqu'au phénotype du patient**



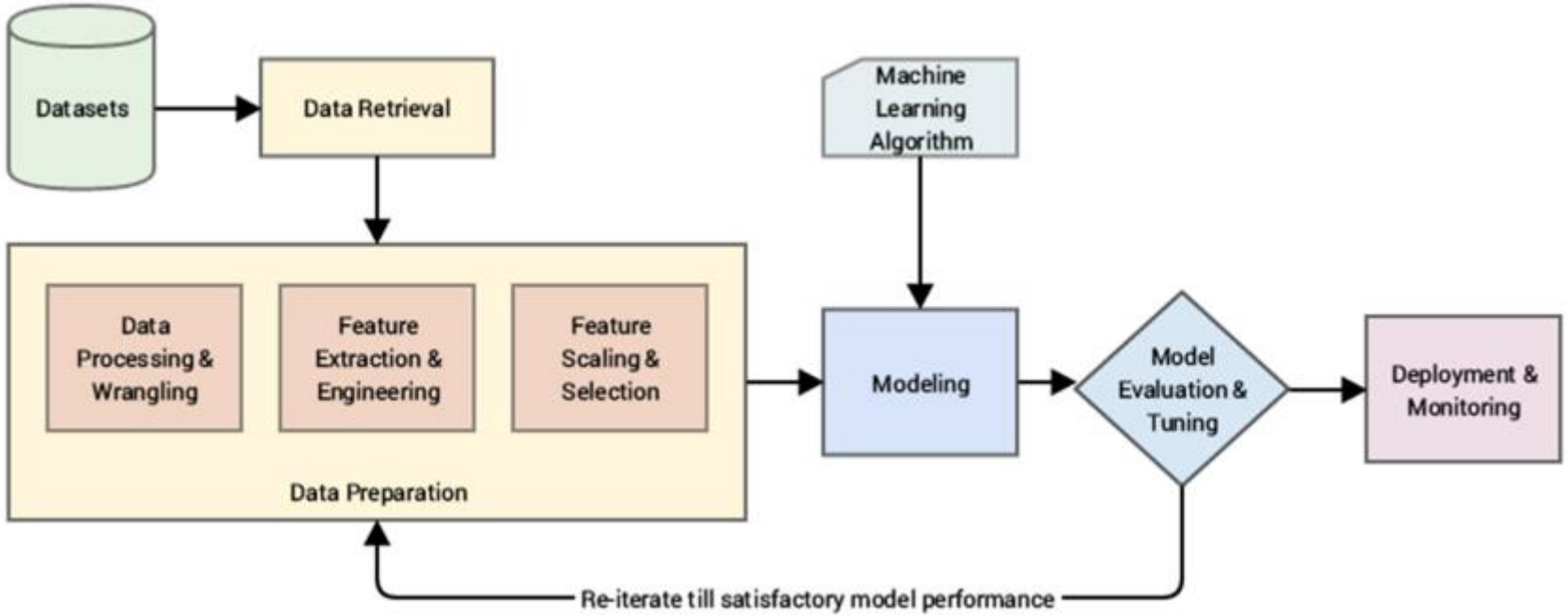
Discussion et remerciements

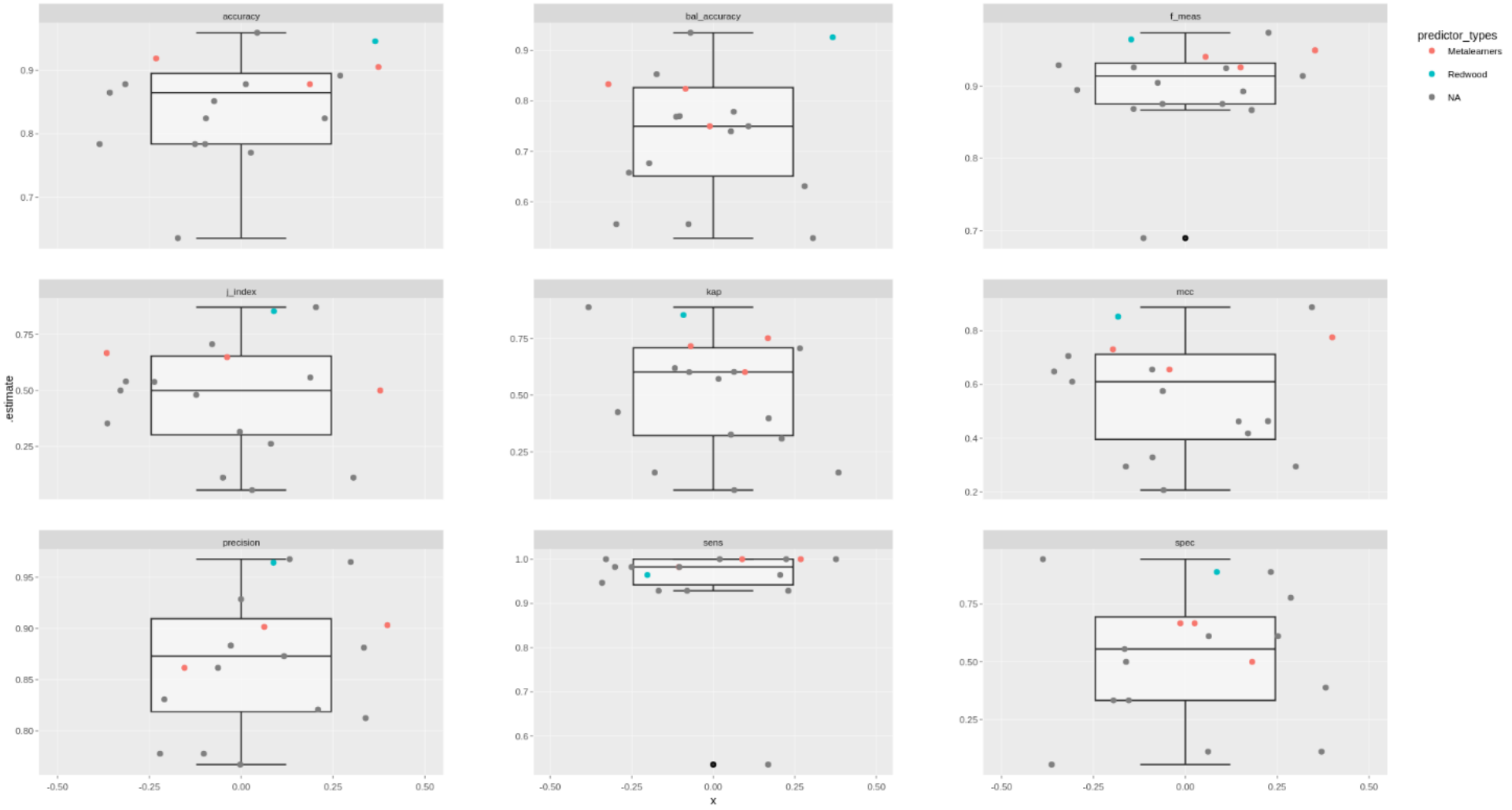


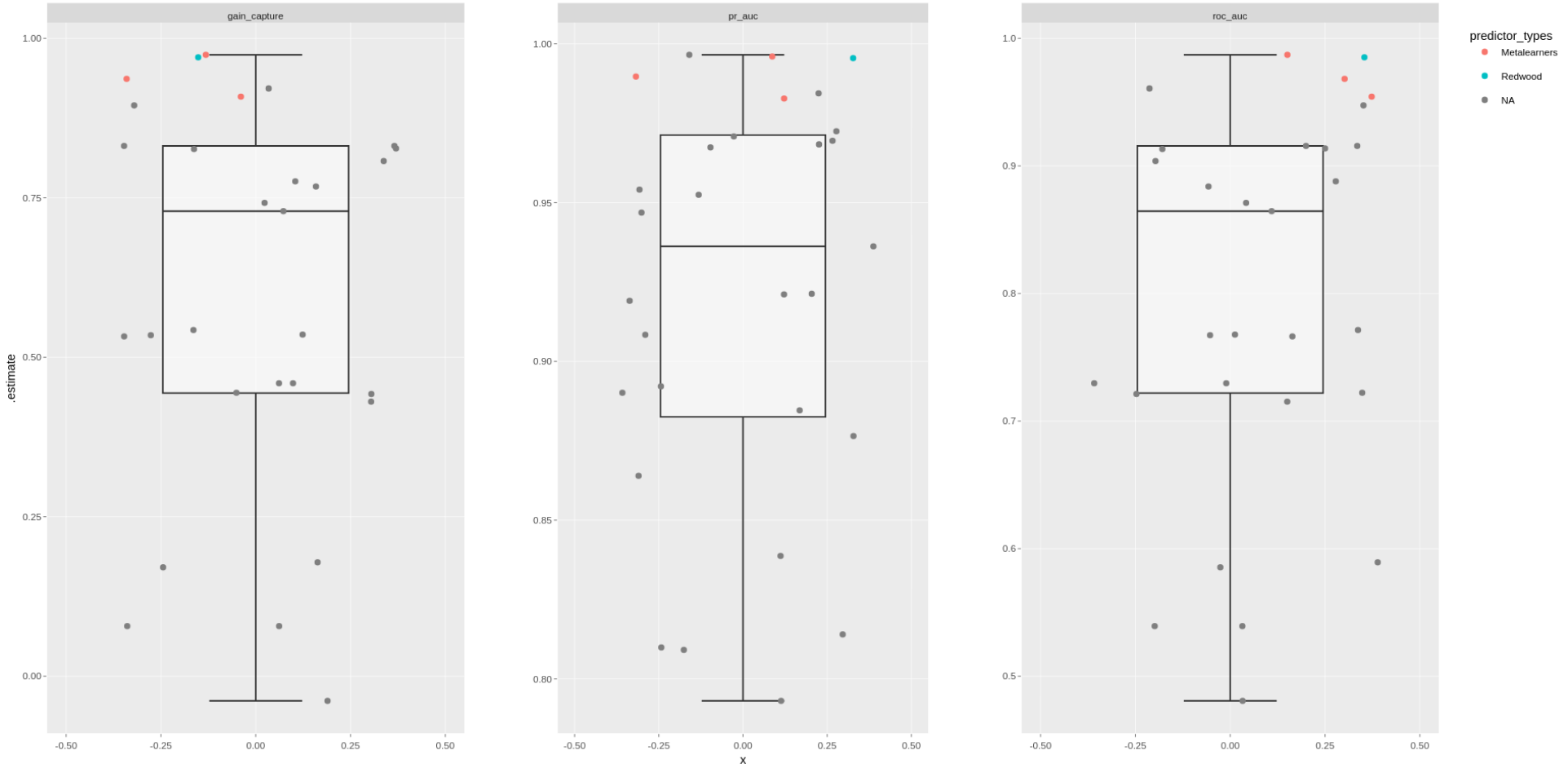
Je remercie tous l'équipe du labo de génétique de Bichat de m'avoir supporté un semestre.
Et prie Cochin de faire de même !



Pipeline d'apprentissage supervisée




accuracy
bal_accuracy
f_meas
j_index
kap
mcc
precision
sens
spec
.estimate
x
predictor_types
Metalearners
Redwood
NA


gain_capture
pr_auc
roc_auc
.estimate
x
predictor_types
Metalearners
Redwood
NA