Prédiction de risque pour la leucémie myéloïde

GUSTAVE ROUSSY
CANCER CAMPUS
GRAND PARIS

ChAllengeData
By MathA

Objectif :

Objectif :

Plan :

1. Analyse du problème
2. Approche choisie
3. Résultats
4. Conclusion

1.1 Définition du risque et métrique associée

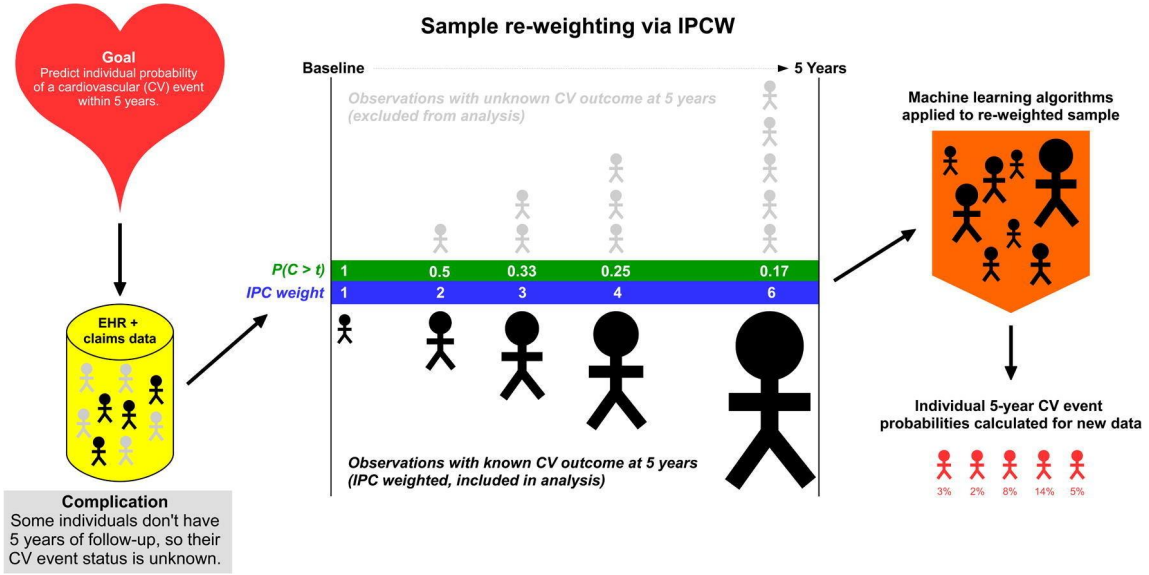
- Risque défini de manière implicite par la métrique
- Métrique : IPCW-C-Index

1.1 Définition du risque et métrique associée

$$\text{C-Index} = \frac{\#\text{ Paires concordantes}}{\#\text{ Paires concordantes} + \#\text{ Paires discordantes}}$$

Problème : P(C < t) augmente avec t  IPCW C-Index

1.1 Définition du risque et métrique associée



- **Hypothèse forte** :
Non informativité du temps de survie pour la censure

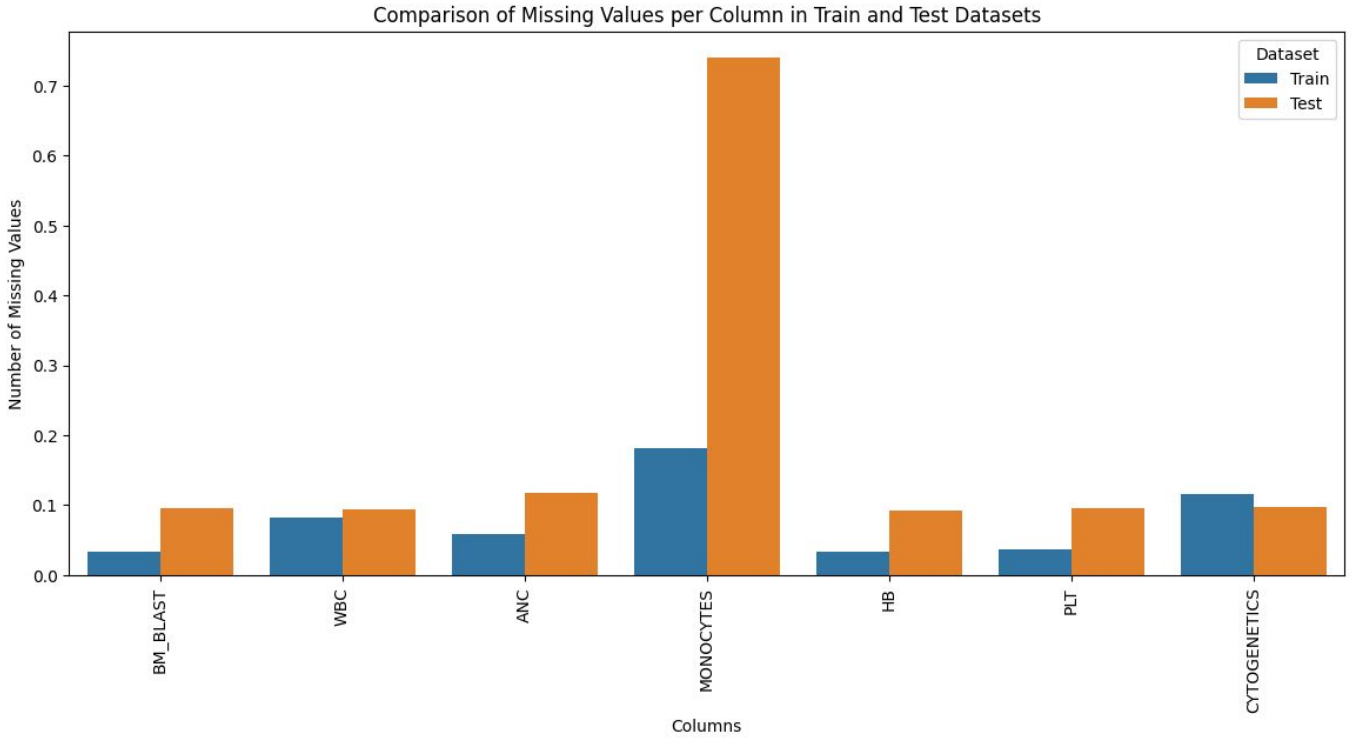
Validité varie d'un centre médical à l'autre

1.2 Données fournies

- Mesures biologiques continues
- Anomalies chromosomiques (encodage ISCN)
- Mutations moléculaires

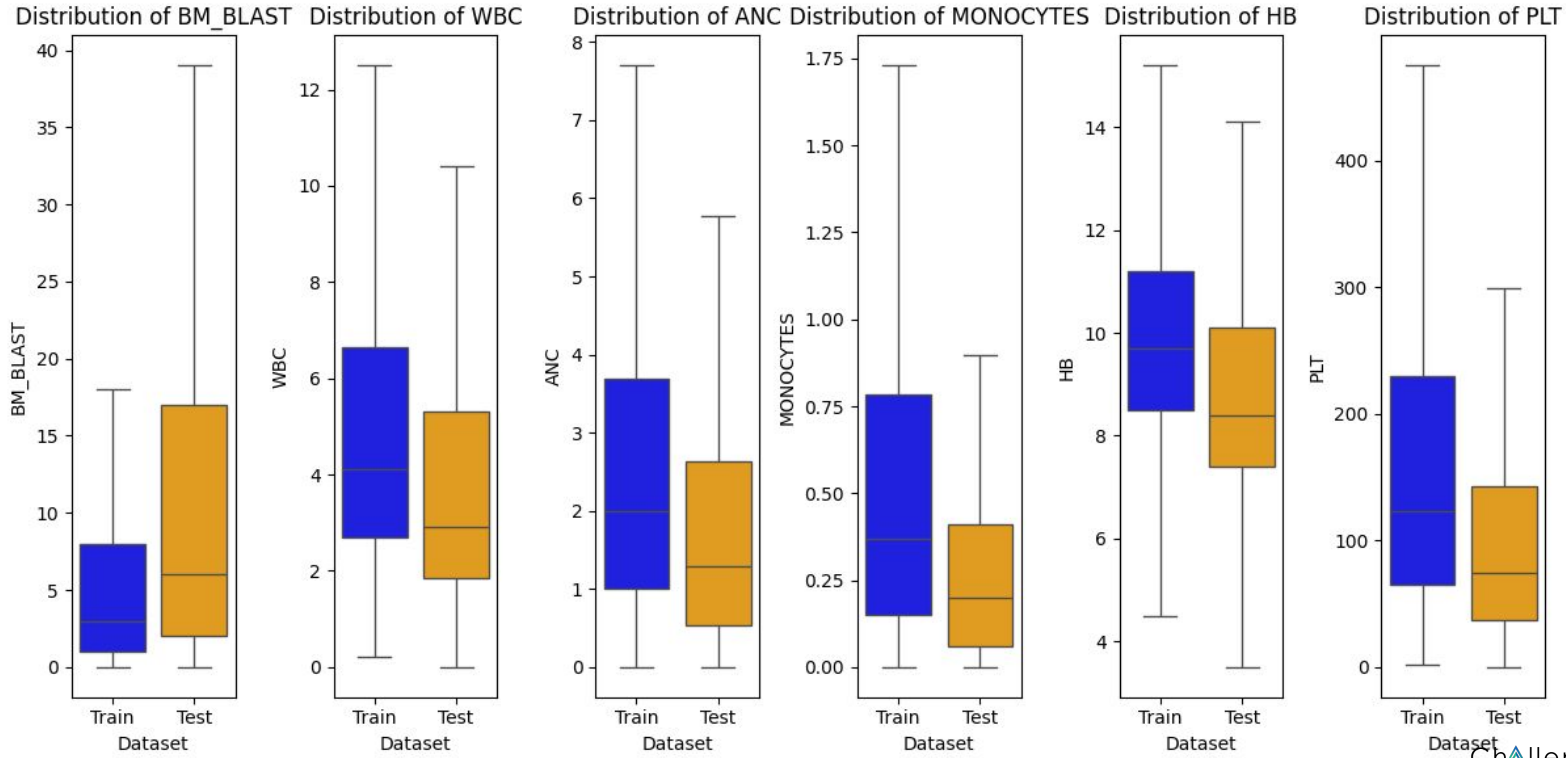
Donnée complexe, et de très haute dimension

1.2 Données fournies

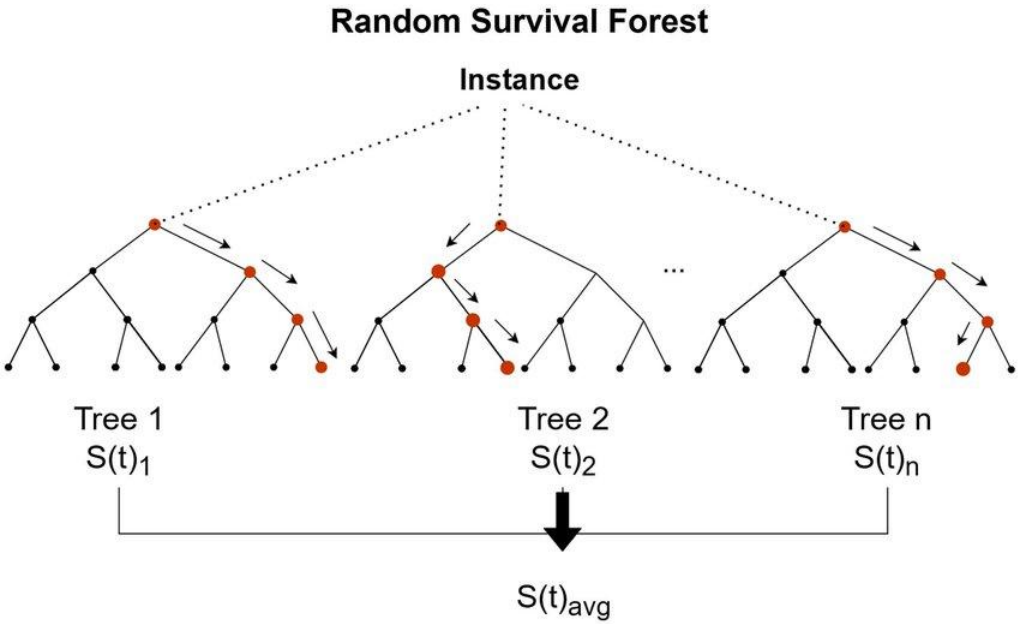


ChAllengeData
By MathA

1.2 Données fournies



2.1 Architectures choisies



- Modèle non-paramétrique et peu d'hypothèses
- Résultats compétitifs
- Mécanique ensembliste qui induit une régularisation naturelle

2.1 Architectures choisies

Modèle de Cox linéaire pénalisé

Risque instantané

$$\lambda\left(t, X_1, \cdots, X_n\right) = \lambda_0\left(t\right) \exp\left(\Sigma_{i=1}^{n} \beta_i X_i\right)$$

Pénalisation

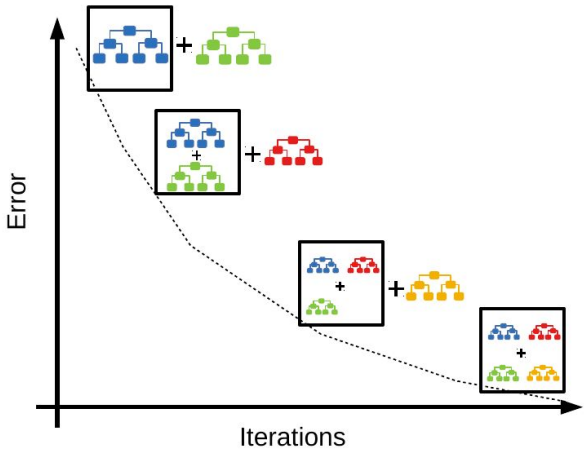
$$\lambda_2 \|\beta\|^2 + \lambda_1 \|\beta\|_1$$

- Hypothèse forte, mais qui est parfois adaptée
- Résultats compétitifs

2.1 Architectures choisies

Modèle de Cox et AFT avec régression XGBoost

$$\lambda(t|\theta) = \lambda_0(t)\theta \qquad \lambda(t|\theta) = \lambda_0(t\theta)\theta$$



- Hypothèses fortes, mais qui sont parfois adaptées
- Résultats compétitifs
- Plus d'hypothèse de linéarité pour la régression

2.2 Preprocessing et Feature Engineering

Variables continues

Données cytogénétiques

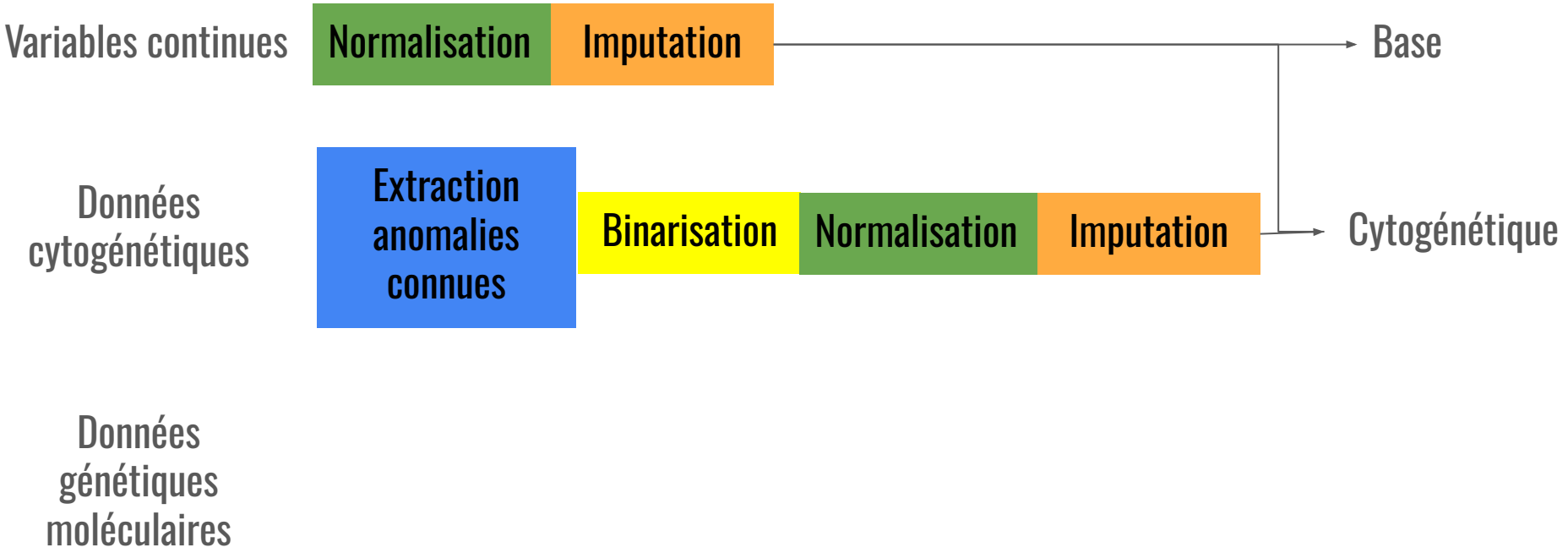
Données génétiques moléculaires

2.2 Preprocessing et Feature Engineering

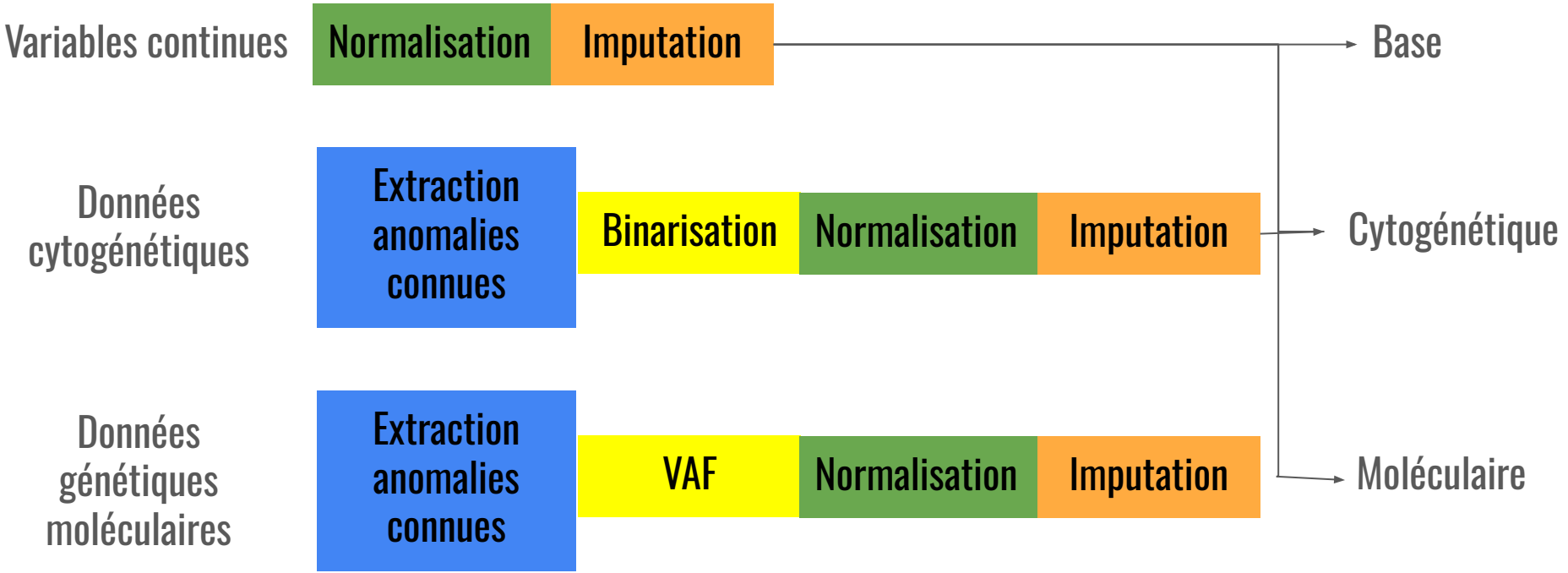
Données cytogénétiques

Données génétiques moléculaires

2.2 Preprocessing et Feature Engineering



2.2 Preprocessing et Feature Engineering



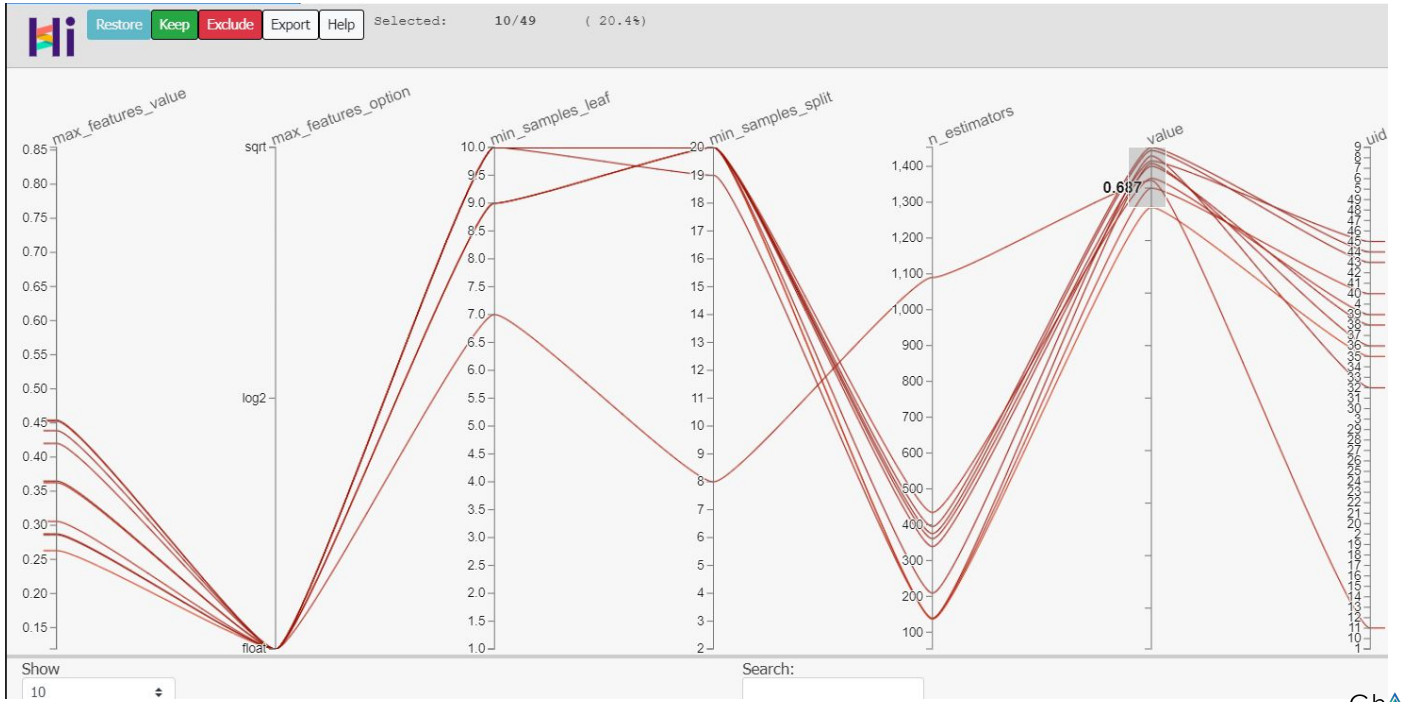
3.1 Résultats - Dépendance aux données

	Base	Cytogénétique	Moléculaire
RSF	**0.664**	**0.673**	**0.690**
Cox linéaire	0.653	0.655	0.663
Cox XGBoost	0.636	0.648	0.676
AFT XGBoost	0.630	0.645	0.673

3.2 Résultats - Méthodes d'imputation

	Moyenne	Médiane	Itérative (Ridge)
RSF	**0.690**	**0.698**	**0.694**
Cox linéaire	0.663	0.672	0.667
Cox XGBoost	0.677	0.676	0.672
AFT XGBoost	0.673	0.664	0.674

3.3 Résultats - HPO ?



ChAllengeData
By MathA

3.4 Résultats - Classement officiel

Académique public

Rang	Date	Participant(s)	Score public
1	15 mars 2025 21:31	l_b & jeremtti	0,7695
2	15 mars 2025 20:44	marcb & tessbreton	0,7656
3	14 février 2025 17:27	SullyCstr & matthieuml	0,7645
4	16 mars 2025 19:53	MoBenyahia & medraki	0,7640
5	15 mars 2025 17:42	gavite & JulienG	0,7634
6	16 mars 2025 19:59	@Sari & mounanaim	0,7630
7	12 mars 2025 14:23	Robenson & Yanis_Kahil	0,7599
8	16 mars 2025 12:44	sachabinder	0,7599
9	16 mars 2025 16:48	pcaucheteux & MANY0427	0,7596
10	7 mars 2025 19:53	flipflop45 & theodore.fougereux	0,7587
11	**15 mars 2025 13:45**	**gilinca**	**0,7571**

Académique privé

Rang	Date	Participant(s)	Score final (dat
1	7 mars 2025 19:53	flipflop45 & theodore.fougereux	0,7156
2	14 mars 2025 16:28	@Sari & mounanaim	0,7138
3	14 février 2025 17:27	SullyCstr & matthieuml	0,7137
4	15 mars 2025 21:31	l_b & jeremtti	0,7136
5	**15 mars 2025 13:45**	**gilinca**	**0,7123**
6	15 mars 2025 17:42	gavite & JulienG	0,7118
7	6 mars 2025 10:14	Werther14 & mathias-grau	0,7114
8	7 mars 2025 03:38	rateddany & bsaadi	0,7100
9	15 mars 2025 20:44	marcb & tessbreton	0,7096
10	13 mars 2025 15:18	MoBenyahia & medraki	0,7087
11	30 janvier 2025 18:19	Ishani	0,7083

4.1 Conclusion - Points clés

- Intégration des données cytogénétiques et moléculaires
- Pas d’overfitting
- Choix du modèle

4.2 Conclusion - Améliorations potentielles ?

- Intégration des données cytogénétiques et moléculaires

  ⟹ Extraire plus de features du dataset, et les enrichir (Mitelman i.e.)

- Pas d'overfitting

  ⟹ Sélection des features par feature importance

- Choix du modèle

  ⟹ Modèles modélisant mieux les intéractions multiples (DL par exemple)

Questions ?