MATHÉMATIQUES
VISION
APPRENTISSAGE

Prédiction de risque pour la leucémie myéloïde

Rapport de challenge de Géraud Ilinca
geraud.ilinca@telecom-paris.fr

1 - Introduction

Dans le cadre du traitement de la leucémie myéloïde, la prédiction précise du profil de risque individuel des patients atteints est essentielle pour adapter les traitements thérapeutiques : les patients à haut risque nécessitent des traitements intensifs, tandis que ceux à faible risque peuvent bénéficier de traitements moins agressifs, limitant ainsi les effets secondaires et améliorant leur qualité de vie. Cependant, cette prédiction reste complexe en raison de la diversité des mécanismes biologiques impliqués, justifiant l'utilisation d'approches de Machine Learning [1].

Dans ce contexte, le challenge proposé par QRT en partenariat avec l'Institut Gustave Roussy revêt une importance particulière. Ce rapport visera d'abord à examiner en détail la problématique abordée, en s'intéressant aussi bien à la métrique utilisée qu'aux données mises à disposition. Nous justifierons par la suite notre choix méthodologique pour la prédiction du risque et procéderons à une analyse des résultats obtenus sur les divers classements.

2 - Analyses préliminaires

2.1 - Métrique d'évaluation

La métrique utilisée pour évaluer la qualité des modèles est l'IPCW C-index (Inverse Probability of Censoring Weighted Concordance Index). Cette mesure étend le C-index classique, qui quantifie la capacité du modèle à ordonner correctement les temps de survie observés. En présence de données censurées à droite, le C-index traditionnel peut être biaisé car certaines régions temporelles contiennent plus ou moins de paires comparables [2]. L'IPCW C-index corrige ce biais en pondérant chaque observation par l'inverse de la probabilité estimée d'être non censurée.

Cependant, cette correction repose sur une hypothèse forte : la non-informativité du mécanisme de censure [2], c'est-à-dire que la probabilité d'être censuré ne doit pas dépendre du risque réel du patient. Dans notre contexte, les modalités de suivi peuvent différer entre les centres médicaux. Ainsi, utiliser ces poids IPCW calculés sur les données d'entraînement pour pondérer nos algorithmes n'est pas forcément une bonne idée. En effet, cette pondération pourrait introduire un sur-apprentissage supplémentaire si la relation reliant le risque du patient et la probabilité de censure varie entre les données d'entraînement et les données de test.

2.2 - Description et spécificités des données utilisées

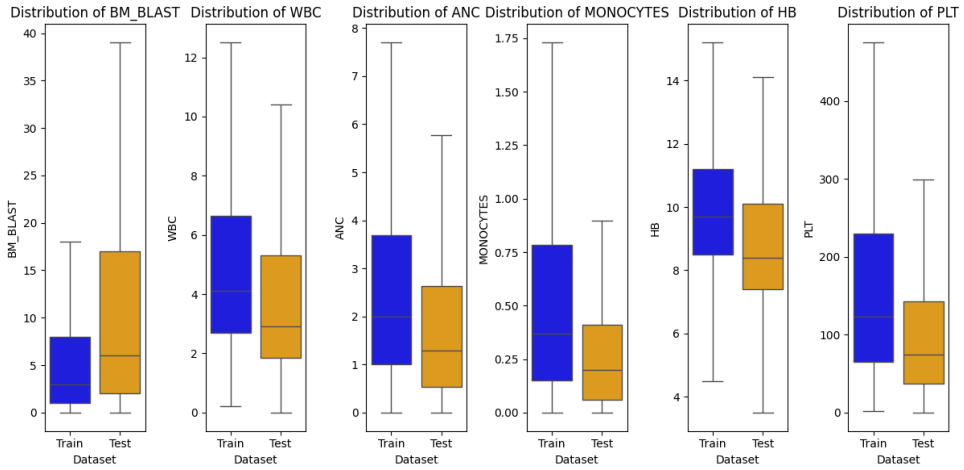
Les données fournies incluent trois catégories principales :

- Variables biologiques continues : concentrations sanguines et autres marqueurs biologiques.
- Données cytogénétiques : mutations chromosomiques décrites selon la nomenclature ISCN.
- Données moléculaires génétiques : mutations ponctuelles avec position chromosomique précise, gène affecté, changement protéique et effet fonctionnel.

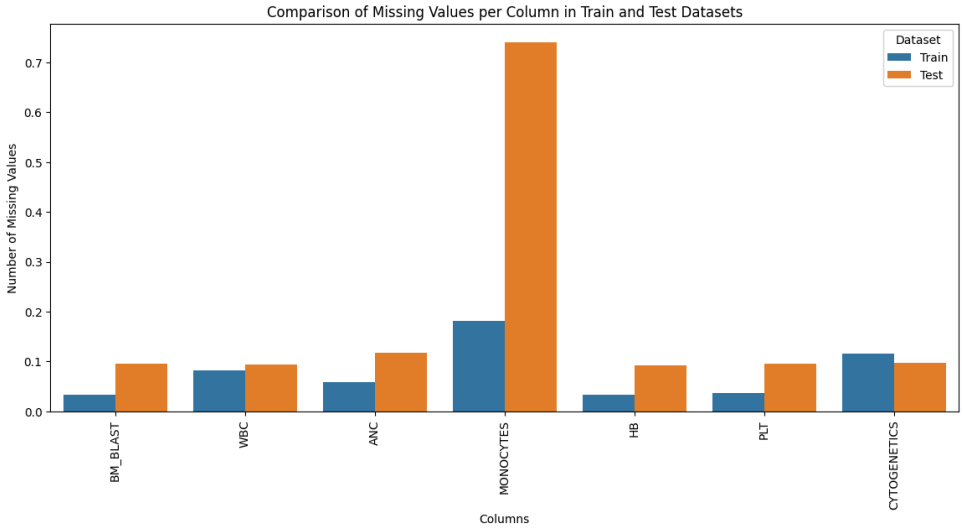
Les données proviennent de 24 centres distincts (23 pour entraînement, 1 différent pour le test). Un point à garder à l'esprit est que d'un centre à l'autre, les distributions conditionnelles $Y \mid X$ (temps

de survie conditionnellement aux caractéristiques), dû aux politiques thérapeutiques variables entre centres, indépendamment de la distribution X.

De plus, on peut observer une différence notable dans les distributions des variables continues des ensembles d'entraînement et de test :



Par ailleurs, on observe une forte variation des proportions de variables manquantes entre les données d'entraînement et de test :



Pour éviter tout data leakage et préserver la validité scientifique des résultats, nous décidons explicitement de ne pas utiliser ces informations spécifiques au dataset de test lors du design ou de l'entraînement des modèles.

Cependant, au vu des fortes différences entre données d'entraînement et de test, on cherchera à privilégier des approches robustes et à intégrer différents biais inductifs.

3 - Approche choisie

3.1 - Architectures

Les Random Survival Forests (RSF) sont un type de Random Forest où, au niveau de chaque noeud cherche, le choix de la feature et du seuil sont effectués de sorte à maximiser une mesure de dissimilarité entre les hazard function des sous-groupes définis par le split. Les RSF constituent une excellente baseline pour plusieurs raisons : elles n'imposent pas l'hypothèse des risques proportionnels, elles obtiennent des résultats compétitifs dans la littérature [1], et elles intègrent une régularisation implicite grâce à leur architecture ensembliste (la fonction de risque est obtenue en moyennant les hazard functions estimées par les arbres). De plus, leur simplicité architecturale facilite les modifications selon les besoins.

Un autre candidat intéressant est le modèle de Cox pénalisé linéaire, qui partage plusieurs qualités avec le RSF : simplicité, robustesse grâce à la régularisation explicite (Lasso ou Ridge), et rapidité d'entraînement. Cependant, ce modèle repose sur l'hypothèse forte des risques proportionnels, qui suppose que l'effet des covariables sur le risque relatif reste constant dans le temps.

Enfin, nous explorons également une approche basée sur le Gradient Boosting via XGBoost, en utilisant deux fonctions de perte adaptées à la survie : la négative log-vraisemblance du modèle de Cox et celle du modèle AFT (Accelerated Failure Time). Ils pourraient surpasser les autres approches si les hypothèses du modèle de Cox ou d'AFT sont bien adaptées au sous-type spécifique de Leucémie Myéloïde étudiée, et si la dépendance aux covariables n'est pas bien modélisée par une relation linéaire.

3.2 - Feature Selection & Feature Engineering

Nous avons construit trois ensembles croissants en complexité :

- Base : Uniquement variables biologiques continues initiales.
- Cytogénétique : Ajout d'informations cytogénétiques binarisées issues des anomalies chromosomiques
- Moléculaire : Ajout supplémentaire des mutations moléculaires encodées par leur Variant Allele Frequency (VAF), binarisation des différentes fonctions possibles pour les mutations

On considère seulement les anomalies chromosomiques et les mutations moléculaires identifiées dans la littérature comme ayant une influence sur le pronostic de la Leucémie Myéloïde [3]. On filtre ces dernières pour avoir un minimum de 4 observations par type d'anomalie/mutation dans le dataset d'entraînement. On ajoute finalement les nombres totaux d'anomalies chromosomiques et de mutations moléculaires.

Par ce biais, on exploite les données cytogénétiques et moléculaires, tout en conservant des nombres de features raisonnables (6 pour base, 27 pour cytogénétique, et 73 pour moléculaire complet).

3.3 - Preprocessing

Comme le modèle de Cox pénalisé linéaire et XGBoost font appel à des pénalisations Lasso et Ridge, nous normalisons les données afin d'éviter que l'échelle des variables, liée aux unités physiques ou à l'encodage, n'influence cette pénalisation. Nous appliquons la même approche à la RSF, celle-ci étant insensible aux translations et aux changements d'échelle.

Concernant les données manquantes, nous commençons par les remplacer successivement par la moyenne puis par la médiane. Enfin, une imputation itérative sera expérimentée, compte tenu de la proportion significative de valeurs manquantes.

4 - Résultats

L'évaluation des modèles repose sur une validation croisée à 5 folds définis manuellement. Comme justifié dans les analyses préliminaires, chaque paire de jeux de données (entraînement et validation) provient de centres différents. Chaque fold réserve entre 15 % et 25 % des données à la validation.

Les performances obtenues avec les architectures testées sur les trois jeux de données sont résumées dans le tableau ci-dessous :

	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

Ces résultats mettent en évidence que la principale source de variation provient du choix des caractéristiques utilisées (feature selection et feature engineering). Cette observation est cohérente avec la littérature, où les études comparatives montrent généralement peu de différences entre modèles, mais des écarts notables liés aux méthodes de sélection des variables [1].

Un choix de modèle s'impose cependant : la RSF obtient systématiquement les meilleurs résultats, une tendance confirmée lors de la soumission sur le portail du challenge.

Enfin, on peut remarquer un comportement intéressant. Sur le dataset possédant uniquement 6 features, le modèle linéaire pénalisé de Cox surpasse XGBoost, et cette tendance s'inverse lorsqu'on ajoute les données sur les anomalies chromosomiques et moléculaires. Cela suggère que la relation entre ces données génétiques et le risque est suffisamment non-linéaire pour compenser le surapprentissage accru (réduction du biais suffisante pour compenser l'augmentation de la variance, si on adopte un point de vue compromis biais-variance).

En testant différentes méthodes d'imputation, on obtient les résultats suivants :

	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

Même si ce résultat semble suggérer la supériorité de l'imputation par la médiane, l'évaluation sur le portail officiel montre que la différence observée sur notre cross-validation est probablement fortuite car ce changement n'entraîne pas d'amélioration des résultats.

Concernant l'optimisation des hyperparamètres, une optimisation succincte via Optuna permet une légère amélioration des métriques de validation (entre +0.005 et +0.01). Toutefois, ces gains ne se traduisent pas par une progression notable dans le classement officiel, suggérant un sur-ajustement aux données de la validation croisée.

Sur le portail du challenge, la meilleure performance est obtenue avec la RSF, sans aucune optimisation des hyperparamètres. Les scores sont de **0,7571 (11^{e} place)** sur le dataset public et **0,7123 (5^{e} place)** sur le dataset privé.

La baisse du score sur le dataset privé suggère que cette partie des données appartient à une région moins bien apprise par le modèle, possiblement sous-représentée dans l'ensemble d'entraînement. Cependant, le gain de position dans le classement montre la robustesse de l'approche choisie, bien que des améliorations soient encore envisageables.

5 - Conclusion

Pour la prédiction du risque pour les patients atteints de leucémie myéloïde, ce challenge a mis en évidence que la sélection des variables d'entrée influence davantage la performance des modèles que le choix de l'algorithme. La Random Survival Forest s'est démarquée comme l'approche la plus robuste.

Pour améliorer les résultats, une piste prometteuse consisterait à enrichir l'ensemble des variables disponibles, puis à affiner leur sélection et leur transformation à l'aide de techniques de feature importance. De plus, l'intégration de sources externes, comme la base de données Mitelman [4], pourrait permettre une caractérisation plus précise des anomalies chromosomiques.

Bibliographie

[1] J.-N. Eckardt *et al.*, "Prediction of complete remission and survival in acute myeloid leukemia using supervised machine learning," *Haematologica*, vol. 108, no. 3, pp. 690–704, Jun. 2022, doi: 10.3324/haematol.2021.280027.

[2] G. Dong *et al.*, "The inverse-probability-of-censoring weighting (IPCW) adjusted win ratio statistic: an unbiased estimator in the presence of independent censoring," *Journal of Biopharmaceutical Statistics*, vol. 30, no. 5, pp. 882–899, Sep. 2020, doi: 10.1080/10543406.2020.1757692.

[3] S. Yohe, "Molecular Genetic Markers in Acute Myeloid Leukemia," *Journal of Clinical Medicine*, vol. 4, no. 3, pp. 460–478, Mar. 2015, doi: 10.3390/jcm4030460.

[4] M. Griffith and O. L. Griffith, "Mitelman Database (Chromosome Aberrations and Gene Fusions in Cancer)," *Dictionary of Bioinformatics and Computational Biology*. Wiley, Oct. 15, 2004. doi: 10.1002/9780471650126.dob0996.