RAPORT DATA CHALLENGE

ENS DATA CHALLENGE — QUBE RESEARCH & TECHNOLOGIES

QRT DATA CHALLENGE
PRÉDICTION DE SURVIE POUR LA LEUCÉMIE MYÉLOÏDE AIGUË

ROBIN GUICHON

03/09/2025 - 15/12/2025

Résumé : *Ce projet développe une pipeline complète pour la prédiction de survie des patients atteints de Leucémie Myéloïde Aiguë (AML) dans le cadre du challenge QRT-Gustave Roussy. À partir d'un jeu de données complexe mêlant variables cliniques tabulaires, mutations génomiques éparses et comptes-rendus cytogénétiques textuels, une stratégie rigoureuse d'ingénierie des fonctionnalités a été mise en œuvre. Celle-ci intègre l'implémentation des règles expertes de l'ELN 2017, le Target Encoding des gènes et l'analyse sémantique (NLP). Pour répondre à la problématique de la censure à droite, des techniques d'expansion temporelle et de correction semi-supervisée ont été appliquées. L'architecture finale repose sur une méthode de Stacking agrégant neuf modèles distincts (dont Histogram Gradient Boosting et MLP). Cette approche a permis d'atteindre un C-Index de 0.7741, surpassant significativement les modèles standards et validant l'apport de l'hybridation entre expertise métier et Machine Learning avancé.*

Keywords : *Stacked Generalization, Histogram Gradient Boosting, Target Encoding, TF-IDF Vectorization, Truncated SVD, Right-Censoring, Discrete-Time Expansion, IPCW C-Index, KNN Imputation, Hyperparameter Optimization*

Bibliographie

[1] Qube Research & Technologies (QRT), *Challenge Data : Prédiction de survie pour la Leucémie Myéloïde Aiguë*, Plateforme Challenge Data ENS, 2025.

[2] D. Micci-Barreca, *A preprocessing scheme for high-cardinality categorical attributes in classification and prediction problems*, ACM SIGKDD Explorations Newsletter, 2001.

[3] G. Salton, C. Buckley, *Term-weighting approaches in automatic text retrieval*, Information Processing Management, 1988.

[4] P. D. Allison, *Discrete-Time Methods for the Analysis of Event Histories*, Sociological Methodology, 1982.

[5] G. Ke et al., *LightGBM: A Highly Efficient Gradient Boosting Decision Tree*, Advances in Neural Information Processing Systems (NIPS), 2017.

[6] Scikit-Learn User Guide/Scikit-learn Developers, *Histogram-Based Gradient Boosting*, Documentation officielle, Section 1.11.

[7] D. H. Wolpert, *Stacked Generalization*, Neural Networks, 1992.

[8] H. Döhner et al., *Diagnosis and management of AML in adults: 2017 ELN recommendations from an international expert panel*, Blood, 2017.