**CIBMTR - Equity in post-HCT Survival Predictions Competition**

**Objectif**

L'objectif de cette compétition est d'améliorer la prédiction des taux de survie des patients ayant subi une greffe de cellules hématopoïétiques allogéniques (HCT). L'accent est mis sur l'équité des prédictions en réduisant les disparités liées au statut socio-économique, à l'origine ethnique et à la géographie.

**Évaluation**

* **Indice de Concordance (C-index)** : mesure la capacité du modèle à ordonner correctement les temps de survie en fonction des scores de risque.
* **Indice de Concordance Stratifié** : ajustement du C-index pour s'assurer que la performance du modèle est équitable entre les différents groupes raciaux.

**Approche et Choix Techniques**

**1. Données**

* **Données synthétiques** générées avec **SurvivalGAN**, simulant un grand ensemble de données réelles.
* **59 variables** incluant des caractéristiques démographiques et médicales.
* **Ciblage du modèle** : La variable cible est l’**event-free survival (efs)** et son temps associé **efs\_time**.

**2. Modèles Utilisés**

L'approche repose sur une **modélisation ensembliste** combinant plusieurs modèles :

* **LightGBM** (boosting)
* **Random Forest** (bagging)
* **XGBoost** (boosting)
* **Modèles de Cox** (modélisation de survie avec CatBoost et lifelines)

Chaque modèle est entraîné sur plusieurs cibles dérivées des variables de survie.

**3. Prétraitement des Données**

* **Imputation des valeurs manquantes** :
  + Pour les variables numériques : médiane spécifique à chaque groupe racial.
  + Pour les variables catégoriques : remplissage avec une valeur Unknown.
* **Encodage des variables catégoriques** via LabelEncoder.
* **Standardisation des variables numériques** avec StandardScaler.

**4. Feature Engineering**

Plusieurs nouvelles variables ont été créées, incluant :

* **Score de compatibilité HLA** : basé sur l'appariement des antigènes HLA.
* **Interaction entre comorbidités et score Karnofsky**.
* **Score socio-économique** combinant assurance, éducation et emploi.
* **Score d'accès aux soins** : prenant en compte la distance au centre hospitalier.
* **Interaction entre race et comorbidités**.

**5. Entraînement et Validation**

* **StratifiedGroupKFold (5 folds)** : pour assurer une distribution équilibrée des races dans l’entraînement et la validation.
* **Binisation des cibles continues** pour améliorer la robustesse du split.
* **Stratification selon la race** pour éviter le biais de représentativité.

**6. Optimisation des Poids de l’Ensemble**

* Les prédictions de chaque modèle sont **pondérées** selon un score optimisé via scipy.optimize.minimize(), afin de maximiser le **C-index stratifié**.

**7. Prédiction et Soumission**

* **Finalisation de l’ensemble** en combinant les prédictions de chaque modèle avec des pondérations optimisées.
* **Génération du fichier de soumission** sous le format :

CopierModifier

ID,prediction

28800,0.5

28801,1.2

28802,0.8

* **Évaluation sur l'ensemble de validation** avec le **Stratified C-index**.

**Points Forts**

✅ Approche robuste avec **plusieurs modèles et cibles**  
✅ Prise en compte de **l’équité raciale dans l’évaluation**  
✅ **Feature engineering avancé** pour enrichir les données  
✅ **Optimisation des poids** pour l’ensemble

Ce pipeline assure une approche **précise et équitable** dans la prédiction de survie post-HCT.