Proposition de stage M2 biostatistiques / DataSciences

Imputation de données cliniques manquantes dans la base ESME d'Unicancer – cancer du poumon

Début : Printemps 2025

Durée du stage : 6 mois

Encadrement : Louise Baschet

Contexte :

Le réseau Unicancer coordonne le programme ESMÉ (« Épidémio-Stratégie Médico-Economique »), une initiative académique indépendante qui vise à centraliser des données de vie réelle en cancérologie (données cliniques, administratives et pharmaceutiques), à partir des données préexistantes dans les établissements.

Horiana est une société de conseil spécialisée en biostatistiques et épidémiologie, qui collabore fréquemment avec Unicancer sur l'exploitation et la valorisation de ses bases de données en vie réelle.

Les études réalisées en oncologie doivent prendre en compte de nombreux paramètres individuels afin de pouvoir évaluer les effets individuels spécifiques. Une des variables importantes est le score ECOG (Eastern Cooperative Oncology Group), un score de performance utilisé pour évaluer l'état général des patients et qui est un facteur pronostique connu de l'évolution du patient et de sa réponse aux différents traitements. Dans les études de vie réelle, cette donnée est parfois manquante pour certains patients, car non systématiquement renseignée au cours de la prise en charge en soins courants.

Dans le cadre des « Emulated Target Trials » ou des études épidémiologiques, les critères de sélection sont appliqués à la base de données afin de cibler les patients d'une population d'intérêt. Ces critères de sélection s'appliquent fréquemment sur le score ECOG. Seuls les patients avec un score ECOG renseigné sont aujourd'hui potentiellement éligibles, ce qui peut générer des biais de sélection, réduire la puissance statistique des analyses, et limiter l'extrapolation des résultats.

Dans ce cadre, l'imputation des valeurs manquantes du score ECOG est une piste pour améliorer la qualité et la complétude de la base, pour l'ensemble des études épidémiologiques et cliniques qui peuvent être réalisées lors de la réutilisation de ces données. En complément des méthodes traditionnelles d'imputation, l'utilisation des techniques de machine learning se révèle prometteuse pour prédire ces valeurs manquantes de manière précise.

Objectifs du stage : L'objectif principal du stage est de développer, optimiser et valider un modèle de machine learning capable d'imputer les valeurs manquantes du score ECOG dans la base ESMÉ sur le cancer du poumon, dans le contexte des données de grandes dimensions, et la base poumon d'ESME comprenait 51 067 patients en 2023.

Missions :

1. **Étude de la base de données ESME** :
    - Exploration des données, analyse des variables corrélées à l'ECOG.
    - Analyse des patterns de données manquantes (Missing Completely at Random, Missing at Random, etc.).
2. **Choix des modèles de machine learning pour l'imputation** :
    - Etat de l'art des approches utilisées dans la littérature

- Identifier et implémenter plusieurs algorithmes adaptés à l'imputation de données manquantes (Random Forest, k-NN, modèles bayésiens, réseaux de neurones, etc.), possiblement en prenant en compte l'aspect longitudinal des données, et les modifications de la prise en charge au cours du temps.
- Optimiser les hyperparamètres des modèles choisis via des techniques comme la validation croisée et la recherche sur grille (grid search).

3. **Processus d'optimisation et de validation des modèles** :
    - Partitionner la base de données (train/test) pour évaluer la performance des modèles.
    - Mettre en place un processus de validation rigoureux incluant des mesures comme l'erreur quadratique moyenne (RMSE), les scores de précision (accuracy), ou l'aire sous la courbe (AUC) pour évaluer la qualité de l'imputation.
    - Comparer les performances des différents modèles en termes de biais et variance, y compris avec les approches Multiple Imputation by Chained Equations (MICE) classiques.

Profil recherché :

- Étudiant en M2 Biostatistiques, Data Science ou en dernière année d'école d'ingénieur avec une spécialisation en statistique ou machine learning.
- Compétences solides en programmation (Python ou R), avec une maîtrise des bibliothèques de machine learning (scikit-learn, XGBoost, etc.).
- Bonne connaissance des méthodes statistiques d'imputation et des modèles de machine learning.
- Capacité à travailler avec des bases de données volumineuses et hétérogènes.
- Rigueur, esprit d'analyse et goût pour les problématiques de santé publique.

Lieu :

Bordeaux