



Intitulé de l'offre de stage

Analyse de Séries Temporelles en Réanimation pour la Prédiction de Survie

Stagiaire en	Informatique / Machine Learning
Affectation	CHU Toulouse
Durée	5 à 6 mois
Rémunération	environ 670 € net mensuel
Date de la publication	01/12/2024
Date d'embauche prévue	mars ou avril 2025 suivant disponibilités
Lieu	CHU, 2 rue Charles Viguerie, 31300 Toulouse, France

Le Centre Hospitalier Universitaire de Toulouse

Le Centre Hospitalier Universitaire (CHU) de Toulouse est constitué de plusieurs sites (les principaux étant les sites de Rangueil, Larrey et Purpan, ainsi que celui de l'oncopole en commun avec l'institut Claudius Régaud); il comprend 4 000 médecins et 12 000 personnels hospitaliers. La mission de recherche et d'innovation du CHU fait partie intégrante à la fois de son activité quotidienne et de sa stratégie pour l'avenir. Elle est menée en collaboration étroite avec les facultés et les organismes de recherche que sont notamment l'Institut national de la santé et de la recherche médicale (Inserm) et le Centre national de la recherche scientifique (CNRS).

L'Institut de Recherche en Informatique de Toulouse

L'Institut de Recherche en Informatique de Toulouse (IRIT), une des plus imposantes Unité Mixte de Recherche au niveau national, est l'un des piliers de la recherche en Occitanie avec ses 700 membres, permanents et non-permanents. De par son caractère multi-tutelle (CNRS, Universités toulousaines), son impact scientifique et ses interactions avec les autres domaines, le laboratoire constitue une des forces structurantes du paysage de l'informatique et de ses applications dans le monde du numérique, tant au niveau régional que national.

L'équipe d'accueil

L'informatique, le traitement des données et l'intelligence artificielle sont appelés à prendre une place croissante dans le monde de la recherche médicale. Dans ce cadre, le CHU de Toulouse a créé une équipe de "gestion de la donnée" dont le but est le recueil, l'analyse et le traitement de la donnée médicale, en support aux activités médicales du CHU. Elle collabore activement avec l'IRIT afin d'appliquer les techniques les plus récentes de la science des données et de l'intelligence artificielle dans le domaine de la santé.

Le stagiaire sera également accueilli au sein du pôle d'anesthésie-réanimation du CHU de Toulouse. Le pôle anesthésie-réanimation est composé de 11 unités d'anesthésie réalisant plus de 3000 anesthésies par mois (hors pédiatrie et obstétrique) pour un très large panel d'interventions couvrant l'ensemble des spécialités chirurgicales et interventionnelles. L'encadrement médical sera assuré par le Dr Michaël Poette, anesthésiste-réanimateur sur le site de Rangueil.

Objet du stage

Contexte : La réanimation des patients en soins intensifs s'appuie sur l'analyse continue de séries temporelles issues des données cliniques, de monitoring et biologiques. Ces données, essentielles pour suivre l'évolution des patients, présentent néanmoins des défis liés à leur hétérogénéité, aux données manquantes, et aux anomalies potentielles. L'objectif de ce projet est, d'une part, de traiter ces données pour renforcer la robustesse des prédictions de survie des patients en détectant les données aberrantes et en imputant des données manquantes, puis d'entraîner des modèles de classification supervisée afin de prédire la survie en réanimation.

Objectifs du Stage : Le projet comporte deux grandes étapes :

- > Un pré-traitement des données incluant la détection des anomalies, l'imputation des données manquantes, et la génération de données synthétiques dans un but de rééquilibrage et de diffusion des jeux de données.
- > Une modélisation prédictive pour estimer la survie des patients en réanimation, en s'appuyant sur des méthodologies d'intelligence artificielle.
- > Pré-traitement des Données : Cette première étape vise à préparer les données de façon à garantir la fiabilité des modèles prédictifs en améliorant la qualité et la diversité des données disponibles. Les sous-objectifs incluent :
 - > Détection des données aberrantes : Développement d'algorithmes pour identifier les valeurs aberrantes dans les séries temporelles, qu'elles soient dues à des erreurs de capteurs, des interférences ou des variations cliniques extrêmes.
 - > Imputation des données manquantes : Application de techniques avancées d'imputation, telles que et les modèles attentionnels, pour compléter les valeurs manquantes tout en préservant la dynamique temporelle des données.
 - > Génération de données synthétiques : Utilisation de modèles génératifs comme les cGAN (Conditional GAN) et les modèles de langage (LLM) pour créer des données synthétiques dans le but de rééquilibrer les groupes de patients selon des critères cliniques, et de faciliter la diffusion des jeux de données en préservant la confidentialité.
- > Modélisation et Prédiction de la Survie : Une fois les données prétraitées, le projet passera à la construction de modèles d'apprentissage profond pour prédire la survie des patients :

- > Entraînement des modèles : Utilisation de modèles de type LSTM (Long Short-Term Memory), CNN (Convolutional Neural Networks) et architectures combinées pour capturer les relations temporelles et structurelles des séries temporelles cliniques.
- > Évaluation des performances : Analyse de la précision prédictive des modèles via des métriques adaptées telles que l'AUROC, la calibration des prédictions, et l'analyse des matrices de confusion. La validation croisée sur différents jeux de données dont des données externes permettra de s'assurer de la généralisabilité des modèles.

Démarche :

- > Préparation des Datasets :
 - > Dataset des patients de réanimation : Utilisation d'un dataset déjà existant de patients issus des réanimations du CHU de Toulouse qui sera enrichi avec des données biologiques et administratives issues de l'Entrepôt de Données de Santé (EDS) du CHU de Toulouse.
 - > Dataset MIMIC-IV : Utilisation après courte formation, afin d'accéder à un ensemble de données complémentaires et diversifiés, qui sera utilisé à travers 2 jeux de données séparés pour limiter le surapprentissage sur nos données locales puis servir de validation externe.
- > Développement des Algorithmes de Pré-traitement et d'Entraînement :
 - > Détection, Imputation et Génération Synthétique : Application des algorithmes pour améliorer la qualité des données et assurer le rééquilibrage des groupes cliniques.
 - > Modélisation prédictive : Implémentation des modèles LSTM et CNN, optimisation des hyperparamètres, et tests de robustesse pour obtenir des prédictions fiables.
- > Validation des Modèles :
 - > Validation externe : Evaluation des modèles sur les deux datasets pour vérifier leur performance dans différents contextes cliniques.
 - > Analyse des erreurs et ajustement : Identification des variables contribuant le plus aux prédictions de survie et ajustement du modèle pour minimiser les erreurs.

Environnement Technique : Des données seront extraites à l'aide de requêtes SQL. Les bibliothèques Python (Pandas, Polars, TensorFlow, PyTorch, Scikit-Learn) seront utilisées pour prétraiter les données, développer les modèles et implémenter les algorithmes de pré-traitement. Les séries temporelles seront traitées avec des frameworks dédiés comme TSFEL pour l'extraction des caractéristiques, ainsi que des architectures LSTM et CNN pour la modélisation.

Impact et Perspectives : Ce projet vise à améliorer l'extraction d'informations issues des séries temporelles en réanimation. Les résultats pourront aider les équipes soignantes à anticiper les évolutions critiques des patients, optimiser la gestion des ressources, et faciliter la diffusion des jeux de données anonymisés pour la recherche en intelligence artificielle appliquée à la santé.

Compétences développées durant le stage : Sur le plan méthodologique, ce stage vise à proposer un projet d'intelligence artificielle de bout en bout depuis l'extraction et le pré-traitement des données jusqu'à la conception de modèles et leur présentation à un panel de professionnels de santé non-experts. Ce stage propose également de mieux comprendre les problématiques liées à l'utilisation des données de santé à travers des visites de services hospitaliers afin de mieux comprendre comment sont renseignées les données, l'explication des enjeux éthiques et la compréhension de la juridiction en vigueur.

Formation

École d'ingénieur, de préférence avec spécialisation en informatique ou mathématiques (en année de césure ou stage long). Master 2 informatique ou mathématiques appliquées.

Compétences attendues

Des connaissances en science des données, apprentissage et réseaux de neurones seront appréciées.

Modalité de candidature

- > CV à envoyer à: anes@stages-medecine-numerique.fr
- > Date limite de candidature : 01/03/2025
- > Encadrants: Dr Michaël Poette (CHU Toulouse), Laure Feuillet (CHU), Sylvain Cussat-Blanc (IRIT), Sandrine Mouysset (IRIT), David Simoncini (IRIT), Daniel Ruiz (IRIT)

Une première sélection sera effectuée sur la base des CVs reçus. Les candidats seront immédiatement informés du résultat, et ceux qui seront retenus à l'issue de la première sélection seront invités à un ou deux entretiens individuels en téléconférence avant sélection définitive.

L'ensemble des offres de stage est disponible sur <http://www.stages-medecine-numerique.fr>.