Institut de Cancérologie de l'Ouest

unicancer PAYS DE LA LOIRE

Alternance (Bac +5) - Data Science

Début : à partir de septembre 2024

Service Data Factory & Analytics

Recrutement n°2678

CONTEXTE

L'Institut de Cancérologie de l'Ouest (ICO) est un établissement de santé privé d'intérêt collectif qui assure des missions de prévention, de soin, de recherche et d'enseignement. Afin de développer son activité de recherche sur données de vie réelles, l'ICO développe son propre Entrepôt de Données de Santé (EDS). L'objectif est d'utiliser les différentes sources de données existantes à l'ICO dans le cadre de la recherche ou du soin afin de créer une unique base de données structurées contenant les variables considérées comme les plus importantes pour mener des travaux de recherche sur données observationnelles.

Aujourd'hui l'EDS est alimenté à partir des bases de données structurées disponibles à l'ICO et des travaux sont en cours dans le but d'extraire des données structurées à partir des documents des patients (comptes rendus de consultation, comptes rendus d'anatomopathologie, etc.). Depuis quelques années, de nombreuses études ont montré qu'il est possible d'extraire des données structurées à partir des comptes rendus en utilisant le NLP (Natural Language Processing) mais toutes ces études reposent sur une méthode nécessitant une longue et coûteuse phase d'annotation afin d'entrainer le modèle (1–6). Le stage consistera à utiliser le Large Language Model (LLM) Mixtral 8x7B, afin extraire des données structurées à partir des comptes-rendus médicaux des patients. A noter qu'un travail de stage est en cours sur le sujet et permettra d'avoir une première idée des possibilités offertes par le modèle et des axes de travail.

MISSION

Poste rattaché au Service Data Factory & Analytics (Direction Développement et Innovation).

L'objectif principal est de développer une solution permettant d'automatiser le processus d'extraction d'informations pertinentes à partir de documents médicaux non structurés et d'évaluer les performances de cette solution.

Tâches principales :

- Compréhension des données médicales : familiarisation avec les différents types de comptes rendus médicaux. Analyse des spécificités linguistiques et des structures de ces documents.
- Développement d'un pipeline d'extraction : conception et mise en œuvre d'un pipeline automatisé utilisant Mistral AI pour extraire les variables d'intérêts à partir des documents médicaux, et permettant d'alimenter une base de données structurée.
- Évaluation de la performance de la solution en termes de précision, de rappel et de F1-score en utilisant une base de données manuellement saisie comme Gold Standard.
- Identification des opportunités d'amélioration et itération du modèle pour une extraction plus performante.
- Adaptation du process pour extraire différentes variables.

Cette alternance offre une opportunité unique d'acquérir des compétences pratiques en data science appliquée à la santé, tout en contribuant au développement d'une solution innovante essentielle pour exploiter des données médicales non structurées. L'alternant travaillera en étroite collaboration avec une équipe multidisciplinaire composée de spécialistes en biostatistique et en oncologie.

REFERENCES

1. Schiappa R, Contu S, Culie D, Thamphya B, Chateau Y, Gal J, et al. RUBY: Natural Language Processing of French Electronic Medical Records for Breast Cancer Research. JCO Clin Cancer Inform. 2022 Jul;6:e2100199.

2. Savova GK, Ogren PV, Duffy PH, Buntrock JD, Chute CG. Mayo clinic NLP system for patient smoking status

identification. J Am Med Inform Assoc JAMIA. 2008;15(1):25–8.

3. Holmes B, Chitale D, Loving J, Tran M, Subramanian V, Berry A, et al. Customizable Natural Language Processing Biomarker Extraction Tool. JCO Clin Cancer Inform. 2021 Aug;5:833–41.

4. Hanauer DA, Barnholtz-Sloan JS, Beno MF, Del Fiol G, Durbin EB, Gologorskaya O, et al. Electronic Medical Record Search Engine (EMERSE): An Information Retrieval Tool for Supporting Cancer Research. JCO Clin Cancer Inform. 2020 May;4:454–63.

5. Carrell DS, Halgrim S, Tran DT, Buist DSM, Chubak J, Chapman WW, et al. Using natural language processing to improve efficiency of manual chart abstraction in research: the case of breast cancer recurrence. Am J Epidemiol. 2014 Mar 15;179(6):749–58.

6. Banerjee I, Bozkurt S, Caswell-Jin JL, Kurian AW, Rubin DL. Natural Language Processing Approaches to Detect the Timeline of Metastatic Recurrence of Breast Cancer. JCO Clin Cancer Inform. 2019 Oct;3:1–12.

PROFIL ATTENDU

En prévision de votre dernière année d'études (Bac +5) en Data Science, vous recherchez pour la rentrée prochaine une alternance. Vous devrez disposer de bonnes connaissances des modèles de traitement du langage et du machine learning et être force de proposition. Vous devez être à l'aise avec les langages de programmation Python et/ou R et avoir une appétence pour les applications en santé et l'oncologie. De bonnes capacités de communication, orales et écrites, sont souhaitées.

- Lieu de stage : Institut de Cancérologie de l'Ouest (ICO) - Site de Nantes / Saint-Herblain - Bd Professeur Jacques Monod, 44800 Saint-Herblain
- Encadrant : Florent Le Borgne, Data Analyst – Statisticien
- Date de début : à partir de septembre 2024
- Durée : un an

Merci d'adresser, **au plus tôt**, votre candidature à
La Direction des Ressources Humaines – INSTITUT CANCEROLOGIE DE L'OUEST
par mail : **srh.recrutement@ico.unicancer.fr**
et à Florent LE BORGNE : **florent.leborgne@ico.unicancer.fr**