Offre de stage de fin d'étude (Bac+5), Santé publique

>LIEU DU STAGE

INSERM (Institut National de la Santé et de la Recherche Médicale)

CépiDc (Centre d'épidémiologie sur les causes médicales de décès)

46, rue Albert 75013 Paris/ Hôpital Paul-Brousse. 12, avenue Paul Vaillant-Couturier. 94804 Villejuif.

>INTITULE DU STAGE

Enrichissement et mise en qualité de la base d'apprentissage des modèles d'intelligence artificielle des causes de décès afin d'améliorer la performance.

>DOMAINE(S) COUVERT(S) PAR LE STAGE

Santé publique, épidémiologie, Datascience, NLP/TAL, qualité

Contexte Le CépiDc de l'Inserm a la charge de traiter la partie médicale des certificats de décès pour leur enregistrement, selon les recommandations de l'Organisation mondiale de la santé (OMS), dans la classification internationale des maladies (CIM). La tâche de codage d'un certificat de décès consiste à attribuer des codes CIM-10 à partir du texte libre renseigné par le médecin et à identifier la cause initiale de décès. Capitalisant sur les millions d'observations annotées par des experts au fil des années suivant de standards internationaux, le CépiDc investit dans le développement de méthodes de traitement automatique des langues pour automatiser l'enregistrement des causes de décès [1]. Un certificat de décès peut être codé par un système expert basé sur la reconnaissance de mots-clés d'un dictionnaire et de règles de décision, par un humain de manière interactive via l'interface du système expert, ou par des modèles d'intelligence artificielle (IA) utilisant des méthodes d'apprentissage profond entraînées sur les données passées. Ces données passées constituent la base d'apprentissage [2,3,4,5]. Le CépiDc souhaite nettoyer sa base d'apprentissage du bruit généré par les soucis de production ou par les variations de codage impactant la décision des modèles IA. L'OMS a publié une nouvelle version de la CIM : la CIM-11 ainsi que des tables de passage non définitives de la CIM-10 vers la CIM-11. L'enjeu est aussi d'anticiper ce passage à la nouvelle nomenclature CIM11.

Objectifs

• Détecter des erreurs de production ou de codage dans la base d'apprentissage ou les changements de règles de codage

• Proposer des corrections à la base d'apprentissage sur les certificats erronés à l'aide de mesures de similarités entre texte de départ et texte de sortie et les mettre en œuvre si elles sont retenues

• Rédiger un document sur l'ensemble des actions réalisées sur le corpus d'apprentissage.
• En fonction de la durée du stage, réaliser le mapping CIM10 vers CIM11 de la base d'apprentissage à l'aide des tables de mapping proposées par l'OMS

Bibliographie

[1] Aurélie Névéol, et al. CLEF eHealth 2018 Multilingual Information Extraction task overview: ICD10 coding of death certificates in French, Hungarian and Italianax http://ceur-ws.org/Vol-2125/invited_paper_18.pdf

[2] Zambetta E, Razakamanana D, Robert A, Clanché F, Martin D, Hebbache Z, et al. Codage des causes de décès 2018 2019 en CIM10 - Approche combinant deep learning, système expert et codage manuel ciblé. *Document de travail CépiDc n2*.

[3] Clanché, Razakamanana, Coudin, Robert, "Les statistiques provisoires sur les causes de décès en 2018 et 2019, une nouvelle méthode de codage faisant appel à l'intelligence artificielle", Drees Méthode n°8

[4] Falissard, Louis, Morgand, Claire,Ghosn, Walid, Imbaud, Claire, Bounebache, Karim and Rey, Grégoire. (2020). Neural translation and automated recognition of ICD-10 medical entities from natural language: Algorithm Development and Validation (Preprint). JMIR Medical Informatics. https://pubmed.ncbi.nlm.nih.gov/35404262/

[5] Zina Hebbache, Aude Robert, François Clanché, Elise Coudin, Diane Martin - Rapport de production - Années de décès 2018-2019 - Données définitives, *Document de travail CépiDc n3*

>CONNAISSANCES ET APTITUDES RECHERCHEES

Compétences :
- connaissances en épidémiologie et méthodes statistiques associées
- rédaction de documents scientifiques
- classification Internationale des Maladies ou PMSI

Aptitudes :
- Logiciels : Python ou R,
- Manipulation de bases de données volumineuses
- Traitement sur données médicales confidentielles
- Anglais lu et écrit courant

>ENVIRONNEMENT DE LA MISSION

Intitulé, activité, compétences statistiques de l'unité d'accueil et du maître de stage :
CépiDc (Centre d'épidémiologie sur les causes médicales de décès). Les missions du CépiDc sont :
- la production de la base nationale des causes médicales de décès,
- la diffusion de cette base pour des objectifs de recherche et de santé publique,
- la production d'analyses statistiques et de recherche sur cette base de données.

Cette dernière mission a donné lieu à l'application des méthodologies statistiques adaptées pour de nombreuses publications dans des revues scientifiques internationales.

Le stage sera co-encadré par :
- Aude Robert, ingénieur au CépiDc spécialisé en traitement automatique des langues,
- Zina Hebbache, responsable du pôle de Codage au CépiDc

Seront aussi associés Daniel Razakamanana datascientist, Elisa Zambetta data engineer, Diane Martin responsable de la production et Elise Coudin directrice du CépiDc.

Ressources mises à la disposition du stagiaire :
Interactions quotidiennes avec les équipes codage et automatisation /datascience du CépiDc.

Gratification : environ 500€ / mois
Durée du stage : 3 mois minimum, date de début négociable.

>PERSONNES A CONTACTER

Aude Robert (aude.robert@inserm.fr)
Zina Hebbache (zina.hebbache@inserm.fr)