Offre de stage M1/M2, 2e année/3e année d'école en statistique

>LIEU DU STAGE

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

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

Hôpital Paul-Brousse. 12, avenue Paul Vaillant-Couturier. 94804 Villejuif.

>INTITULE DU STAGE

Détermination de la taille optimale de l'échantillon test pour valider les performances du modèle d'intelligence artificielle utilisé pour déterminer les causes de décès.

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

Contexte La classification automatique des documents médicaux est un domaine scientifique qui connaît un intérêt constant depuis de nombreuses années. 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). Capitalisant sur les millions d'observations annotées par des experts au fil des années, le CépiDc investit dans le développement de méthodes de traitement automatique des langues pour automatiser l'enregistrement des causes de décès [1]. Un système expert basé sur la reconnaissance de mots-clés d'un dictionnaire et des règles de décision permet de traiter automatiquement les deux tiers des données. En outre, les données des années récentes sont en partie prédites par des méthodes d'apprentissage profond entraînées sur les données passées [2,3,4]. Précisément, des modèles seq-to-seq transformers [4,5] et des modèles long-term short- term memory bidirectionnels (BiLSTM, [6]) ont pour tâche de prédire les codes de la CIM version 10 correspondant aux textes rédigés par les médecins constatant les décès et d'identifier la cause initiale du décès, codée dans cette même nomenclature. La stratégie de codage utilisée et les vérifications réalisées sont détaillées dans le rapport de production de la base annuelle de données[7].

Objectifs :

- **Déterminer la taille optimale d'un échantillon test pour le modèle d'intelligence artificielle permettant le codage des causes de décès, en tenant compte de la relation avec la taille de l'échantillon d'apprentissage, qui augmente au fur et à mesure de l'entraînement du modèle.**
- **Déterminer les intervalles de confiance des indicateurs obtenus permettant de valider la qualité des prédictions du modèle d'intelligence artificielle.**
- **De plus pour valider plus globalement la qualité de la production du CépiDc, un échantillon de certificats médicaux dit gold sera codé par des experts. Il faudra estimer la précision des indicateurs estimés à partir de ce gold.**

Bibliographie

[1] Chauvet, Guillaume, 2007. Méthode de Bootstrap en population finie

[2] 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

[3] 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. *Mimeo*. 2023.

[4] 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

[5] 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/

[6] Vaswani, Ashish, Shazeer, Noam, Parmar, Niki, Uszkoreit, Jakob, Jones, Llion, Gomez, Aidan N, Kaiser, Łukasz and Polosukhin, Illia. "Attention is all you need." Paper presented at the meeting of the Advances in Neural Information Processing Systems, 2017. https://arxiv.org/abs/1706.03762?context=cs

[7] Graves, A. and Schmidhuber, J., "Framewise phoneme classification with bidirectional LSTM networks," Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005., Montreal, QC, Canada, 2005, pp. 2047-2052 vol. 4, doi: 10.1109/IJCNN.2005.1556215.

[8] Hebbache Z, Boulet P, Robert A, Zambetta E, Razakamanana N, Coudin E, et al. Rapport de production: année de décès 2022024 Mar. (Document de travail du CépiDc). Report No.: 4. https://www.cepidc.inserm.fr/sites/default/files/2024-03/DT_CEPIDC_N4_Rapport%20de%20production%202021.pdf

Le cas échéant, degré prévisible de confidentialité du rapport de stage : faible

>CONNAISSANCES ET APTITUDES RECHERCHEES

Connaissances des outils suivants :
- Statistique et notamment théorie des sondages
- Bootstrap

Aptitudes :
- Logiciels : Python, R
- Aisance en programmation
- 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
- Fanny Godet, ingénieur statisticienne

Seront aussi associés Daniel Razakamanana datascientist, Elisa Zambetta data engineer et Elise Coudin directrice du CépiDc.

Ressources mises à la disposition du stagiaire :
Données nationales d'enregistrement des causes de décès (plus de 3 millions d'enregistrements annotés)
Plateforme de calcul du CépiDc (sur base de 3 GPU).
Intéractions quotidiennes avec l'équipe automatisation /datascience du CépiDc.

Gratification : environ 500€ / mois
Durée du stage : 4 mois (négociable,3 mois minimum), date de début négociable.

>PERSONNES A CONTACTER

Aude Robert (aude.robert@inserm.fr)
Fanny Godet (fanny.godet@inserm.fr)