Offre de stage de fin d'étude (Bac+5), mention Informatique/Intelligence Artificielle

> LIEU DU STAGE

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

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

80, rue du Général Leclerc, 94270 Le Kremlin-Bicêtre

> INTITULE DU STAGE

Evaluation de la qualité de la saisie numérique à l'aide de méthodes d'intelligence artificielle

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

Statistique, Mathématiques, Informatique, Intelligence artificielle, Calcul haute performance

Contexte

Les données sur les causes médicales de décès sont considérées comme les données de santé publiques de référence au niveau national et international. Utilisée par la recherche ainsi que par les multiplies acteurs de la santé publique, l'évaluation de la qualité de ces données est un enjeu primordiale. Le CépiDc est l'unité de service Inserm chargée de produire annuellement de la production de la Base des Causes Médicales de Décès (BCMD), de sa diffusion et du support technico-scientifique relatif à son exploitation. Une problématique inhérente à la production des données est la saisie numérique des certificats manuscrits rédigés par les médecins. Cette opération est actuellement effectuée par un prestataire et représente environ 450 000 documents numérisés et saisis par an soit 85% des décès survenus sur le territoire national.

Ce procédé est une étape fondamentale dans l'identification des causes de décès selon un processus conforme aux recommandations internationales de l'OMS. Dans ce cadre, le CépiDc souhaite mettre en place des méthodes d'apprentissage machine comme l'apprentissage profond afin de reconnaitre et saisir les entités nosologiques manuscrites présentes sur les certificats papiers.

Objectifs

Dans un premier temps, le stagiaire aura pour objectifs de construire un jeu de données adapté à l'exploitation par différents types de méthodes d'apprentissage telles que les réseaux de neurones. Dans un second temps, il sera chargé de mettre au point plusieurs méthodes de type apprentissage profond dédiées à la reconnaissance optique de caractères ainsi que d'évaluer leurs performances.

Méthodes

Source : Plusieurs millions d'images de certificats de décès associée à leur saisie numérique.

Résultats attendus :

- Un jeu de données ayant une granularité assez fine pour permettre l'adaptation d'algorithmes d'apprentissage variés.
- L'implémentation d'algorithmes d'apprentissage machine, dont l'apprentissage profond (réseaux à convolutions, modèles seq2seq à attention ou CTC)
- Des critères d'évaluation de performance et fiabilité de ces derniers.

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

☐ **extrême** ☐ **moyen** ☐ **faible**

Connaissances et aptitudes recherchées chez le stagiaire :

Connaissances des outils suivants :

- *Principes de l'apprentissage statistique et applications*
- *Méthodes de traitement d'image*

Aptitudes :

- *Logiciels : Python, openCV, Tensorflow*
- *Aisance en programmation*
- *Manipulation de bases de données volumineuses*
- *Anglais lu et écrit courant*

> ENVIRONNEMENT DE LA MISSION

Intitulé, activité, compétences statistiques de l'unité d'accueil et du maître de stage :

INSERM-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.

Walid Ghosn, le maître de ce stage, est statisticien-épidémiologiste (diplômé en Ingénierie mathématique de l'université Pierre et Marie Curie) et docteur en épidémiologie-santé publique. Au sein du CépiDc, la mission du stagiaire sera effectuée en collaboration avec l'équipe en charge des recherches et développements qui comprend deux ingénieurs de recherche dont un expert en méthodes d'apprentissage, un statisticien-épidémiologiste ainsi qu'un doctorant spécialisé en apprentissage machine et calcul haute performance.

Ressources mises à la disposition du stagiaire :

Le stagiaire disposera d'un bureau, d'un ordinateur puissant, et du logiciel R et Python (je suppose puisque ça fait partie des compétences requises). Il aura accès à un serveur de calcul haute performance.

La gratification du stage est d'environ 500€ / mois

Durée du stage : 6 mois

> PERSONNE(S) A CONTACTER

M GHOSN Walid,

INSERM-CépiDc

walid.ghosn@inserm.fr

01 49 59 53 37

Code de champ modifié