Inserm
La science pour la santé
From science to health

HÔPITAUX UNIVERSITAIRES PARIS OUEST
Corentin-Celton
Hôpital européen Georges-Pompidou
Vaugirard - Gabriel-Pallez

STAGE D'APPLICATION EN STATISTIQUE
3ème ANNEE

FICHE DESCRIPTIVE DU STAGE

L'ENTREPRISE (nom, coordonnées…) :
AP-HP (Assistance Publique-Hôpitaux de Paris)
Hôpital Européen Georges Pompidou
20-40 rue Leblanc
75015 PARIS
N° siret : **267 500 452 00052** **Code APE :** **8610Z**

NOM et Prénom du stagiaire :

INTITULE DU STAGE

Machine Learning en imagerie : Prédiction de la réponse à la chimiothérapie de patients atteints de cancer

DOMAINE(S) COUVERT(S) PAR LE STAGE

Objectif(s) du stage (problématique, missions, méthodologie...) :
(Si vous disposez d'un descriptif détaillé concernant ce travail, veuillez le joindre en annexe)
Grâce aux progrès technologiques et médicaux des 30 dernières années, l'imagerie médicale fonctionnelle occupe une place de plus en plus importante dans le diagnostic puis le suivi de la réponse au traitement de certaines pathologies. Ceci est aujourd'hui pleinement avéré dans le domaine de la cancérologie. Les capacités de stockage augmentent, les techniques d'imagerie se perfectionnent, entrainant un accroissement non-négligeable du nombre des images à traiter et de la taille de chacune d'entre elles.
L'utilisation de méthodes d'intelligence artificielle (IA), comme les réseaux de neurones, les méthodes de plus proches voisins ou les forets aléatoires, par exemple, prend alors tout son sens pour extraire et synthétiser l'information contenue dans ces données volumineuses.
Un certain nombre de packages 'R' ont été mis au point pour implémenter et exploiter ces méthodes d'IA (https://cran.r-project.org/web/views/MachineLearning.html). Mais le choix du package reste aujourd'hui subjectif.
Par ce travail nous souhaitons mener une évaluation plus concrète de différents packages en les comparant sur des jeux de données d'imagerie identiques.
Nous souhaitons pouvoir guider notre équipe et les autres utilisateurs de ces packages en dressant un panorama de leur forces et de leurs faiblesses, en fonction de l'objectif initial et des données dont on dispose (nombre d'images, …) ou d'autres paramètres que nous découvrirons peut-être au cours du projet.

En étroite collaboration avec le tuteur et la radiologue intéressée par le projet, il s'agira de

- S'approprier les outils et programmes de codage et manipulation des images
- S'approprier les différents packages de « R » d'IA correspondant à la question posée par la radiologue (classification supervisée ou non supervisée), au travers de la documentation liée à chaque package mais aussi aux références bibliographiques déjà publiées
- En accord avec la radiologue, choisir un ou plusieurs jeux de données de base et programmer les analyses de classification sur ce(s) jeu(x) de données
- Evaluer les résultats dans ces différentes conditions (nombre d'images, degré de dépendance inter-images, etc)
- Participer à la présentation et discuter les résultats avec les acteurs impliqués dans le projet (tuteur, médecin radiologue, etc)

Si besoin, dans un souci de simplification de problématique trop complexe, des simulations pourront être réalisées puis

analysées.

Domaine scientifique principal : ☒ statistique ☐ économie ☒ informatique

Base de données
Préciser le contenu et la disponibilité :
La radiologue intéressée par le projet met à disposition des images anonymisées de malades venus à l'hôpital pour un diagnostic puis pour le suivi de leur pathologie. La(les) pathologie(s) à étudier plus précisément seront décidées en accord avec la radiologue notamment en fonction du volume d'images à traiter et du nombre de patients atteints.

Résultats attendus :
- Comparatif des différentes méthodes d'IA pour des classifications supervisées ou non-supervisées, sur base de données identiques, par une évaluation rigoureuse.
- Proposer des pistes de recommandation de l'usage de l'un ou de l'autre des packages.

Principales méthodes statistiques utilisées (exemple : Analyse de données, régression logistique,….) :
- Méthodes de classification.
- Statistiques descriptives

Connaissances et aptitudes recherchées chez le stagiaire :
- Très forte appétence pour l'informatique et la gestion de bases de données éventuellement non relationnelles.
- Maitriser le logiciel R et avoir de solides connaissances dans l'utilisation d'API, si possible
- Etre force de proposition, rigoureux, avec un bon esprit d'initiative. Etre curieux des enjeux et du contexte dans lequel le travail s'inscrit. Savoir s'adapter à différents interlocuteurs.
- Etre capable de comprendre des documents rédigés en anglais, articles scientifiques et documents techniques par exemple.
- Avoir envie de mettre en pratique ses connaissances théoriques en statistique et informatique, dans le domaine de la médecine et de la biologie, en général.

Compétences à acquérir :
- Gérer, nettoyer les bases de données hétérogènes et de grande taille (entrepôts de données)
- Effectuer des analyses de données multidimensionnelles, construire des profils d'individus sur les données
- Ecrire et documenter des applications logicielles adaptées à l'analyse récurrente des données de l'entreprise
- Communiquer et documenter les analyses et les résultats aux commanditaires du projet pour la prise de décision finale

ENVIRONNEMENT DE LA MISSION

Intitulé, organisation, activité, compétences statistiques de l'unité d'accueil et du maître de stage pressenti :
L'étudiant sera accueilli au sein de l'Unité de Recherche Clinique (URC) du Centre d'Investigation Clinique-Epidémiologie Clinique 1418 des HuPO.
Cette unité est située au sein de l'Hôpital Européen Georges Pompidou (HEGP) qui fait partie des 37 centres hospitaliers gérés par l'Assistance Publique des Hôpitaux de Paris.
Le CIC 1418 est une structure de recherche clinique mixte AP-HP-Inserm offrant un soutien aux investigateurs cliniciens et aux chercheurs, allant du support logistique infirmier et médical à une implication totale (aide au montage, financement, réalisation et soutient des projets de recherche reposant sur des patients). Le CIC 1418 participe également à la coordination des activités de recherche du groupe hospitalier. Deux des principales missions du CIC1418 sont de développer des recherches translationnelles, fondées sur des hypothèses originales et de produire de nouvelles connaissances scientifiques et médicales dans le respect des règles éthiques, légales, entre autre.
L'étudiant sera intégré à l'équipe des 5 biostatisticiennes de l'URC, sous la direction du Pr Sandrine KATSAHIAN, médecin hospitalo-universitaire en statistiques et informatique biomédicale. Armelle ARNOUX, qui co-encadrera le stagiaire est docteur en santé publique mention épidémiologie et possède plus de 10 ans d'expérience en Biostatistiques aussi bien dans un milieu privé que publique. Enfin, Laure FOURNIER, radiologue à l'HEGP, s'intéresse à l'imagerie des cancers, afin d'améliorer leur détection précoce et l'évaluation de l'efficacité des thérapies. L'imagerie fonctionnelle permettant de mieux décrire la biologie des cancers, et la recherche sur données (« big data ») sont ses domaines d'expertise.

Ressources mises à la disposition du stagiaire (informatique, bureautiques, logiciels statistiques, matérielles...) :
Un bureau équipé d'un ordinateur sur lequel sont installés les logiciels 'R' et de bureautique (Word, Excel, Powerpoint) ainsi qu'un accès Internet pour effectuer les recherches bibliographiques notamment, sera mis à la disposition du stagiaire.

PERSONNE(S) A CONTACTER

NOM Prénom, Fonction, Service, Mail, Tel :
Pr KATSAHIAN Sandrine
PU-PH Biostatistiques et informatique biomédicale
CIC 1418, Unité de Recherche Clinique
sandrine.katsahian@aphp.fr
01 56 09 55 01

Mme ARNOUX Armelle
Ingénieur de Recherche Sénior, Biostatisticienne
CIC 1418, Unité de Recherche Clinique
armelle.arnoux@aphp.fr
01 56 09 56 37