RÉPUBLIQUE FRANÇAISE
Liberté Égalité Fraternité

Inserm
La science pour la santé
From science to health

Ingénieur.e en apprentissage automatique Data scientist

CDD 3 ans | *Début : dès que possible* | *Villejuif* | *Télétravail partiel* | *Bac +5 ou plus*

L'Inserm est le seul organisme public français entièrement dédié à la recherche biologique, médicale et en santé des populations. Il dispose de laboratoires de recherche sur l'ensemble du territoire, regroupés en 12 Délégations Régionales. Notre institut réunit 15 000 chercheurs, ingénieurs, techniciens et personnels administratifs, avec un objectif commun : améliorer la santé de tous par le progrès des connaissances sur le vivant et sur les maladies, l'innovation dans les traitements et la recherche en santé publique.

Rejoindre l'Inserm, c'est intégrer un institut engagé pour la parité et l'égalité professionnelle, la diversité et l'accompagnement de ses agents en situation de handicap, dès le recrutement et tout au long de la carrière. Afin de préserver le bien-être au travail, l'Inserm mène une politique active en matière de conditions de travail, reposant notamment sur un juste équilibre entre vie personnelle et vie professionnelle.

L'Inserm a reçu en 2016 le label européen HR Excellence in Research et s'est engagé à faire évoluer ses pratiques de recrutement et d'évaluation des chercheurs.

Emploi

Poste ouvert aux candidats

- ☐ Agents fonctionnaires de l'Inserm par voie de mobilité interne
- ☐ Agents fonctionnaires non Inserm par voie de détachement
- ☒ CDD agents contractuels

Catégorie	A
Corps	IR
Emploi-Type	Expert en information statistique E1D44

Structure d'accueil

Unité : CépiDc-US10

A propos de la Structure

Le CépiDc, unité de service de l'Inserm, a pour mission de produire la base de données statistique sur les causes médicales de décès en France, de la diffuser et de réaliser des analyses sur cette base de données. Cette base de données statistique repose sur la collecte et le traitement des volets médicaux des certificats de décès. Ses finalités d'usage sont multiples : la connaissance de l'état de santé de la France et de ses territoires et sa comparaison au niveau international, en vue d'aider au pilotage des politiques de santé publique ; la recherche et les études, les données alimentant le système national des données de santé ; la veille et l'alerte sanitaire, par la production de la donnée la plus pertinente possible dans des délais de quelques jours.
Les principaux traitements réalisés au CépiDc concernent d'abord l'accueil, le contrôle et l'intégration des données collectées, pour une mise à disposition immédiate à des

fins de veille de sanitaire. Puis, le CépiDc assure le dédoublonnage et la correction de non-réponse totale *via* une mise en cohérence avec les décès déclarés à l'état civil et gérés par l'Insee, pour finir par la construction des variables statistiques, avec en particulier le codage des causes médicales de décès.
Concernant ce dernier aspect, il s'agit d'analyser et de coder les textes rédigés par les professionnels de santé lors du constat de décès dans la classification internationale des maladies (CIM). La stratégie de codage combine de façon optimale trois modes : utilisation d'un système-expert de règles en batch et en interactif et utilisation de modèles d'intelligence artificielle prédictifs, construits et entraînés *in-house*. Capitalisant sur les millions d'observations analysées par des experts suivant des standards internationaux et dans le contexte d'une profonde rénovation du processus de production des causes de décès, le CépiDc intègre ces méthodes dans sa chaîne de production pour gagner en temps (afin de respecter les délais réglementaires de diffusion des données) et en qualité, tout en adoptant une démarche statistique rigoureuse et novatrice. L'élaboration de la base de données sur les causes de décès suit les recommandations de l'OMS, et doit satisfaire les normes de qualité d'une statistique officielle et du code des bonnes pratiques en matière de statistique européenne.
Le CépiDc est composé d'une vingtaine d'agents, répartis en deux pôles : pôle production des données et pôle exploitation-diffusion.

Directrice	Hélène Chaput
Adresse	Bâtiment Laplace, Hôpital Paul Brousse, 12 Av. Paul Vaillant Couturier, 94800 Villejuif
Délégation Régionale	Paris Sud

Description du poste

Mission principale

Vous participez à la mise en œuvre en production courante des outils d'intelligence artificielle pour le codage des causes de décès. Ces outils fondés sur de l'apprentissage profond et du traitement automatique des langues améliorent la qualité et la rapidité de codage automatique, de façon à satisfaire les délais réglementaires de diffusion de la base. Vous êtes responsable de leur adaptation pour tenir compte du prochain changement de nomenclature (passage de la CIM 10 à la CIM 11) et vous êtes partie prenante de l'évolution du système d'information du CépiDc qui en découle. Vous bénéficiez d'un accès à des ressources de calcul (GPU) permettant de concevoir, entraîner et tester des modèles et de réaliser des prédictions.
Au sein du pôle production des données du CépiDc, vous travaillez dans l'équipe automatisation, sous la responsabilité de la cheffe d'équipe, et en étroite collaboration avec le *data scientist senior*. Vous collaborez également avec le reste de l'équipe pluridisciplinaire (codeurs, nosologistes, responsables de production, statisticiens) et vous êtes partie prenante de l'écosystème formé avec les partenaires de recherche et développement (médecins spécialisés en informatique médicale et datascientists, de l'AP-HP, LISN-CNRS, Insee, Santé publique France, Inserm).

Activités principales

- Mettre en production, maintenir, monitorer et valider une chaîne de traitements de données textuelles comprenant des prédicteurs de type réseaux de neurones (*transformers*) pour aider/automatiser le codage du texte libre des certificats de décès dans la CIM (annotation, training/fine-tuning, monitoring).
- Mettre en production le ciblage des certificats à allouer aux différentes modalités de codage (IA, manuel), évaluer l'amélioration continue du codage automatique (en taux de codage et en qualité) en vue d'une boucle d'apprentissage continue (on line) à partir de la validation/correction des codeurs des propositions de l'algorithme.
- Adapter l'architecture du modèle et le *feature engineering* en vue d'améliorer la classification des causes, en adéquation avec la finalité statistique du traitement et les bonnes pratiques.
- Participer à l'internationalisation de ces méthodes en lien avec les instances représentatives

	françaises à l'OMS et au sein de l'Europe. • Assurer une veille scientifique sur les modèles et les algorithmes à l'état de l'art dans le domaine. • Participer activement à des groupes d'échanges de bonnes pratiques existants ou à construire regroupant datascientists, statisticiens et chercheurs en épidémiologie et informatique (Insee, DREES, Inserm, Inria,...) autour de l'usage de l'IA/TAL sur ces thématiques.
Spécificité(s) et environnement du poste	• Confidentialité des données • Contraintes de production.
Connaissances	• Apprentissage automatique, traitement automatique des langues, *deep learning*, sciences des données • Maîtrise de l'ensemble des étapes allant du développement à la mise en production • Maîtrise des environnements de production • De bonnes bases statistiques • Des connaissances en biostatistique et un intérêt pour l'épidémiologie sont des plus
Savoir-faire	• Très bonne maîtrise de Python et des librairies de *deep learning* (Tensorflow, Pytorch) en particulier celles appliquées au traitement automatique des langues. • Entraînement et monitoring d'algorithmes de *deep learning* • Mise en production d'algorithmes de *machine learning,* MLops • Git, outil de versioning • Design et maintien de pipeline de *machine learning*, ces expériences sont des plus, de même que l'utilisation de Docker, MLFlow, et de technologies cloud
Aptitudes	• Proactivité, force de proposition • Aisance relationnelle, sens de la communication et de la pédagogie • Capacités d'organisation, de planification et de rigueur • Discrétion et confidentialité • Savoir s'insérer et interagir avec des équipes multidisciplinaires : pôle de production, experts métiers chargés de production, statisticiens, stagiaires, chercheurs • Savoir se maintenir à l'état de l'art des connaissances
Expérience(s) souhaité(s)	• Ce poste convient à un sortant d'école motivé, formé à l'usage de Python et des librairies d'apprentissage profond.
Niveau de diplôme et formation(s)	• Diplôme d'ingénieur de grandes écoles, Master en data science ou équivalence professionnelle
Informations Générales	
Date de prise de fonction	Dès que possible
Durée (CDD et détachements)	36 mois Renouvelable : ☒ OUI ☐ NON
Temps de travail	• Temps plein • Nombre d'heures hebdomadaires : 38h30 • 45 jours congés annuels et RTT
Activités télétravaillables	☒ OUI, en partie ☐ NON * Préciser les modalités de télétravail possible.
Rémunération	• **Selon l'expérience et le profil de candidature**

- Prise en charge d'une partie de la mutuelle.

Modalités de candidature	
Date limite de candidature	15/02/2026
Contact	recrutement.cepidc@inserm.fr ; aude.robert@inserm.fr
Contractuels	• Envoyer CV et lettre de motivation à recrutement.cepidc@inserm.fr • Précisez vos prétentions salariales.
Pour en savoir +	• Sur l'Inserm : **https://www.inserm.fr/** ; site RH : **https://rh.inserm.fr/Pages/default.aspx** • Sur la politique handicap de l'Inserm et sur la mise en place d'aménagements de poste de travail, contactez la Mission Handicap : **emploi.handicap@inserm.fr**