Caractérisation et analyse automatisée des bases de données de santé

Ségolène Combettes, PhD
Directrice R&D - Alia Santé
https://alia-sante.com

Dans le secteur de la santé, l'exploitation et l'analyse des données représente une opportunité sans précédent pour améliorer la recherche et stimuler l'innovation. La quantité de données générées par les recherches cliniques, les dossiers médicaux, les dispositifs connectés de santé et d'autres sources est colossale et ne cesse de croître. Correctement exploitées, ces informations peuvent significativement accélérer l'innovation dans le domaine de la santé, favorisant ainsi l'émergence d'une médecine plus personnalisée, préventive et efficace.

Cependant, l'analyse de ces vastes ensembles de données peut s'avérer complexe, notamment en termes de nettoyage des données, d'analyse statistique, et de visualisation. Les données de santé sont souvent complexes, fragmentées, et contiennent des valeurs manquantes ou erronées qui peuvent biaiser les analyses si elles ne sont pas correctement traitées. La présentation des données de manière intuitive et informative est essentielle pour que les décisions soient prises sur la base des meilleures informations disponibles. Cette première étape de caractérisation et d'analyse des bases de données de santé est indispensable pour transformer les données brutes en données exploitables et valorisables dans un projet d'IA.

Dans ce contexte, le projet proposé par Alia Santé vise à automatiser cette étape de pré-traitement grâce à une pipeline d'analyse de données complète et polyvalente. Cette pipeline pourra être intégrée dans notre produit Alia HDMS, un gestionnaire de base de données. Cet algorithme permettra à n'importe quel utilisateur d'accéder facilement à des analyses fiables et compréhensibles de sa base de données. Cette solution permettra d'aborder et de traiter efficacement n'importe quel dataset, en particulier dans le domaine de la santé, où la précision et la fiabilité des analyses sont cruciales.

Les étudiants seront amenés à créer un algorithme capable de :

- Analyser des datasets en identifiant et en traitant les valeurs manquantes, les doublons et les erreurs potentielles.
- Appliquer des méthodes statistiques pour résumer et comprendre les datasets, incluant la moyenne, la médiane, l'écart-type, et d'autres indicateurs pertinents.
- Mettre en place des techniques de visualisation des données pour faciliter l'interprétation des résultats d'analyses, en utilisant des graphiques et d'autres outils graphiques interactifs.
- Modularité et adaptabilité : Le système doit être conçu pour être facilement adaptable à différents types de données et de besoins d'analyse.