

La protection de la vie privée dans les bases de données de santé : enjeux et défis de l'apprentissage automatique

Encadrante : Nesrine Kaaniche

2 binômes d'étudiants souhaité

Contexte

L'apprentissage automatique (IA) est une technologie en plein essor qui a le potentiel de révolutionner de nombreux secteurs, dont la santé. En effet, l'IA permet d'analyser de grandes quantités de données de santé pour identifier des tendances, des corrélations et des modèles qui peuvent être utilisés pour améliorer la prévention, le diagnostic et le traitement des maladies.

Cependant, l'utilisation de l'IA dans le domaine de la santé soulève également des enjeux importants en matière de sécurité et de protection de la vie privée. Les données de santé sont en effet des données sensibles qui peuvent être utilisées à des fins malveillantes, comme l'identification de personnes ou le vol d'identité. Ces données sont aussi utilisées pour former des modèles qui alimentent les applications d'apprentissage automatique.

Outre l'intérêt croissant pour les attaques elles-mêmes, il y a un intérêt croissant pour découvrir ce qui cause les fuites de données et dans quelles conditions un modèle est susceptible de différents types d'attaques liées à la vie privée. Il y a plusieurs raisons pour lesquelles les modèles divulguent des informations. Certains d'entre eux sont structurels et ont à voir avec la façon dont les modèles sont construits, tandis que d'autres sont en raison de facteurs tels qu'une mauvaise généralisation ou une mémorisation d'échantillons de données sensibles. Dans cette optique, plusieurs attaques ont été proposées pour évaluer le niveau de protection de la vie privée par les différents algorithmes intelligents [1,2].

Objectifs

L'objectif du projet consiste à :

- Étudier les attaques contre la confidentialité des données et la protection de la vie privée sur la base de données MIMIC ;
- En concertation à l'encadrante, identifier [1] et implémenter les attaques pertinentes pour évaluer le niveau de protection des données en s'appuyant sur ART toolbox [3];
- Evaluer et comparer les résultats

Livrables

- Rapport
- Démonstrateur et code source des outils implémentés

Références

[1] <https://github.com/stratosphereips/awesome-ml-privacy-attacks#privacy-testing-tools>

[2] <https://www.kaggle.com/datasets>

[3] <https://adversarial-robustness-toolbox.readthedocs.io/en/latest/>