Equité des algorithmes de l'apprentissage automatique

Encadrante : Nesrine Kaaniche

1 binôme d'étudiants souhaité

Contexte

Les algorithmes s'immiscent de plus en plus dans notre quotidien à l'image des algorithmes d'aide à la décision (algorithme de recommandation ou de *scoring*), ou bien des algorithmes autonomes embarqués dans des machines intelligentes (véhicules autonomes). Déployés dans de nombreux secteurs et industries pour leur efficacité, leurs résultats sont de plus en plus discutés et contestés [1]. En particulier, ils sont accusés d'être des boites noires et de conduire à des pratiques discriminatoires liées au genre ou à l'origine ethnique. L'objectif de ce projet est d'étudier les différentes définitions d'équité [2,3] et d'évaluer les méthodes dites de « pré-processing » pour y remédier. Le projet s'intéresse aux résultats des algorithmes en rapport avec des objectifs d'équité.

Objectifs

L'objectif du projet consiste à :

- Implémenter une méthode de pré-processing, permettant d'évaluer l'algorithmes sur un sous-ensemble d'attributs;
- Evaluer l'impact de la méthode implémentée par rapport aux objectifs d'équité.

Livrables

- Rapport
- Démonstrateur et code source de la méthode implémentée

Références

[1] https://www.kaggle.com/alexisbcook/ai-fairness

[2] Verma, S. and Rubin, J., 2018, May. Fairness definitions explained. In *2018 ieee/acm international workshop on software fairness (fairware)* (pp. 1-7). IEEE.

[3] Bertail, P., Bounie, D., Clémençon, S. and Waelbroeck, P., 2019. Algorithmes: biais, discrimination et équité.