Travaux de groupes – École d'automne 2022 - A. Merceron

Exercice 1: Vous démarrez un projet pour prédire l'abandon des études dans l'université dans laquelle vous êtes inscrits actuellement.

- Quelles données pourraient être utiles ? Où sont stockées ces données ?
- Vous devez présenter un concept pour la sécurité des données que vous allez obtenir. Donnez les grands thèmes de ce concept.
- Dans l'esprit "open research" vous aimeriez / devriez publier votre code et vos données. Sous quelles conditions pouvez-vous le faire ?

Exercice 2: Vos données sont similaire aux données montrées sur le transparent 8 du ficher 2022_11_16_Ateliers_EcoleAutmone. Vous pouvez les modifier légèrement pour mieux refléter la réalité de l'université dans laquelle vous êtes inscrits.

- Quels attributs proposez-vous d'utiliser tels quels pour prédire l'abandon des études ?
- Quels attributs, a priori pertinents pour pouvoir prédire l'abandon des études, proposez-vous de calculer à partir de ces données ?
- Donner le type de ces attributs (catégoriels, binaires, numériques) et le domaine de valeurs qu'ils peuvent prendre.

Exercice 3: Utiliser RapidMiner (ou Jupyter Notebook et scikitlearn si vous connaissez déjà) et le fichier MIB-Students_en.csv pour prédire l'abandon des études, voire ce site qui contient aussi plusieurs « process » de RapidMiner : http://gdac.uqam.ca/Workshop@EDM20/#material. Le fichier MiniTutorialRapidMiner peut aussi être utile.

- Lire le fichier et comprendre et visualizer les données.
- Utiliser l'opérateur "Decision Tree" pour prédire l'abandon des études.
- Utiliser l'opérateur cross-validation pour évaluer le modèle.
- Essayer plusieurs algorithmes, en particulier les arbres de décision et les réseaux de neurones et comparer les résultats.
- Si vous avez du temps, inclure l'opérateur d'optimisation des hyper-paramètres.

Exercice 4: Regarder les transparents sur Fairness. Considérer la matrice de confusion pour prédire l'abandon des études pour garçons et filles ci-dessous.

- Est-ce que le "recall" et l'"accuracy" cas Filles et Garçons sont similaires ? Pourquoi ?
- Est-ce que le modèle est "fair" au sens de "equalized-odds" ? de "equal-opportunity"?
- Expliquer avec vos mots à vous ce que ces deux mesures veulent dire.

Garçons

	Pr. abandon	Pr. non abandon	
Ac. abandon	492 (TP)	57 (FN)	(P)
Ac. non abandon	117 (FP)	225 (TN)	(N)

Filles

	Pr. abandon	Pr. non abandon	
Ac. abandon	492 (TP)	117 (FN)	(P)
Ac. non abandon	57 (FP)	225 (TN)	(N)

Exercice 5: Vous avez développé toute une pipe-line pour prédire l'abandon des études dans votre université.

- Quelle "accuracy" minimum doivent avoir vos modèles pour pouvoir être utilisés ? Comment justifier vous cette valeur ?
- Vos modèles pourraient-ils avoir des problèmes de "fairness" ? Lesquels ?
- À qui dans l'université vos modèles pourraient-ils être utiles ? Pourquoi faire ?
- Comment pensez-vous que les étudiants de votre institution répondraient aux questions du transparent 21 ?
- Est-ce que les modèles doivent être explicable (explainable) pour être utilisés ? (Les utilisateurs et utilisatrices comprennent comment la décision est calculée, ou comprennent ce qu'ils / elles pourraient changer pour avoir la prédiction opposée).