Devoir de projet 1:
Prédicteur de base

Le but de ce devoir est de construire un système de base pour la reconnaissance automatique de l'âge, du genre, et la personnalité des utilisateurs de médias sociaux. Pour implémenter le modèle de base (baseline), il suffit d'utiliser les étiquettes connues (ie, train data). Un jeu de données de formation avec 9500 instances étiquetées est fourni. Votre logiciel doit prédire les informations suivantes sur l'utilisateur en sortie :

- **e genre**, en tant que ¡¡male¿¿ ou ¡¡female¿¿ (qui correspondent à homme et femme respectivement)
- **l'âge**, soit xx-24, 25-34, 35-49 ou 50-xx "35-49", or "50-xx"
- **la personnalité**, sous la forme d'un score compris entre [1,5] pour chacun des cinq traits du modèle de personnalité Big Five, à savoir: ouverture à l'expérience, conscience, extraversion, acceptabilité et stabilité émotionnelle (inversement appelé neuroticisme).

La description de votre projet contient une brève description des caractéristiques de la personnalité du modèle de personnalité Big Five. Prédire le genre est une tâche de classification binaire ou d'apprentissage de concepts. Prédire l'âge est une tâche de classification multi-classe. Prédire les scores de personnalité correspond à la résolution de cinq tâches de régression. Pour ce devoir, vous devez implémenter des prédicteurs de base, soit de base majoritaire, soit de base moyenne pour déduire le genre, l'âge et les traits de personnalité.

- Tâche de classification: Implémenter un modèle de base majoritaire qui prédit l'étiquette la plus commune dans les données de formation.
- Tâche de régression: Implémenter un modèle de base moyenne qui prédit le score qui est le score de valeur moyenne des données de formation.

Les scores de ces algorithmes de base fournissent le point de comparaison requis pour évaluer tous d'autres algorithmes d'apprentissage automatique que vous développerez pour le projet.

Format de soumission requis

La commande de test ift6758 doit prendre en entrée (i) un chemin absolu vers le jeu de données de test (test data) avec de nouvelles instances (sans contenant les étiquettes) et (ii) un chemin absolu vers un répertoire de sortie vide; le format ressemble à:

```
ift6758 -i path/to/test/my-test-data/ -o path/to/output/directory/
```

Le format du jeu de données de test est le même que celui de la formation (train data), avec une structure de répertoire interne identique. Pour chaque utilisateur du jeu de données de test, votre logiciel doit générer un fichier XML correspondant comme le suivant (en anglais):

```
<user
id="8157f43c71fbf53f4580fd3fc808bd29"
age_group="xx-24"
gender="female"
extrovert="2.7"
neurotic="4.55"
```

```
agreeable="3"
conscientious="1.9"
open="2.1"
/>
```

Ce fichier doit être enregistré dans le répertoire de sortie spécifié et porter la valeur id de l'utilisateur comme nom de fichier de base.et "xml" comme extension. Par exemple, le nom du fichier XML dont le contenu est affiché ci-dessus serait:

`8157f43c71fbf53f4580fd3fc808bd29.xml`

Un jeu de données de test public contenant des données de 334 utilisateurs (pas d'étiquettes!) est disponible dans le dossier /data/public-test-data. Vérifiez que votre logiciel fonctionne correctement en exécutant la commande suivante:

```
ift6758 -i /data/public-test-data/ -o ~/results/
```

En utilisant une commande similaire à celle ci-dessus, les capacités prédictives de votre logiciel seront testées sur un jeu de données de test caché de n = 1334 utilisateurs. Ce jeu contient les 334 utilisateurs de le jeu de données de test public, ainsi que 1000 nouveaux utilisateurs qui ne sont ni dans le jeu de données de formation ni dans le jeu de données de test public. Vos solutions pour l'âge et le genre seront évalués en fonction de la précision, c'est-à-dire du nombre d'instances correctement classées, divisé par le nombre total d'instances. Pour l'identification de la personnalité, l'erreur quadratique moyenne (RMSE) sur les cinq traits de personnalité seront utilisés. RMSE est défini comme:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Avec y_i la valeur réelle et $\hat{y}_i$ la valeur prédite.

Échéance: lundi, le 30 septembre

Votre logiciel sera testé entre le mardi 9h et le jeudi 9h. Assurez-vous que le lundi avant 23h59, la commande ift6758 appelle une version stable de votre logiciel sur votre compte (/submissions), et que cela reste disponible au moins jusqu'à mardi 9h. Nous automatiquement récupérons votre logiciel le lundi à minuit.