

Informations sur le cours

April 11, 2019

1 Introduction au traitement de données

Céline Meillier
meillier@unistra.fr
Bureau C216

1.1 Avant toute chose ...

- Démarrer la machine sur Ubuntu
- Dans chaque nouveau terminal, entrez la commande suivante :

```
export PATH=$PATH:/usr/local/miniconda3/bin
```

- Se rendre sur http://images.icube.unistra.fr/fr/index.php/Page_du_cours
 - → Icube → équipe IMAGEs → Céline Meillier → Enseignement d'ouverture - option ISSD : Introduction à la science des données - Filière généraliste 1A
- Suivre les instructions pour construire l'environnement virtuel python dans lequel on travaillera tout au long du cours.

1.2 Modalité du cours

Prérequis

- Statistiques de lycée
 - Curiosité.
-

Déroulement de l'enseignement : Cours intégré sous la forme de notebook jupyter : notes de cours + exemples en python et applications des notions vues sur un petit projet.

→ **ne pas hésiter à ajouter des commentaires dans votre version du notebook, à modifier le code pour tester les fonctions python**

Projet : Visualiser, traiter et analyser (classification/prédiction) un jeu de données d'apprentissage -> application le jour de la restitution sur un petit jeu de données de test.
→ Par groupe de 2 (éventuellement de 3 ou individuellement)

Organisation :

- 4 séances de CI de 1h45 : 12/04, 07/05, 13/05 matin et après-midi
 - 1 séance de projet 4h en autonomie : 14/05
 - 1 séance de projet encadré/CI 1h45 : 15/05
 - 2 séances de projet 4h en autonomie : 15/05 et 16/05
 - 1 séance de restitution/évaluation 4h : 17/05
-

Evaluation : validation des compétences acquises à travers le projet (cf compétences à valider à l'issue de ce cours) + quelques points de la note finale liée à la performance de l'approche proposée sur un nouveau jeu de données.

1.3 Compétences attendues à la fin du module

- Connaître les outils statistiques pour résumer/représenter un jeu de données afin de déterminer les meilleures approches et analyses pour chaque donnée.
- Connaître les différents moyens de représenter graphiquement des données en fonction de leur nature et de leur taille.
- Savoir utiliser les outils mis à disposition via les packages python : pandas, numpy, seaborn et matplotlib pour représenter graphiquement et numériquement les données.
- Savoir mettre en évidence des relations simples entre les variables et modéliser ces relations en vue de faire de la prédiction ou de la classification.
- Connaître quelques méthodes simples de machine learning pour la classification supervisée et non supervisée.
- Etre en mesure de visualiser, pré-traiter, traiter et analyser un jeu de données à l'aide des différents outils vus en CI.
- Savoir communiquer sur sa démarche et sur les résultats obtenus.
- Savoir critiquer les résultats obtenus.