



# Hackathon Data Science

## M2 DS4SC

Emmanuel Doumard – [emmanuel.doumard@univ-tours.fr](mailto:emmanuel.doumard@univ-tours.fr)

# Principe du hackathon

- Objectif : résoudre un problème de ML en une journée
- Compétition entre équipes de 3

# Planning

- Jeudi 18/12 :
  - 8h15 – 9h00 : Présentation du hackathon
  - 19h00 : Fermeture du serveur d'évaluation
- Vendredi 19/12 :
  - 8h15 : Présentation des projets

Contexte

# La médecine

- Les progrès de la médecine ont fortement allongé l'espérance de vie à la naissance
- Mais la fin de vie est encore très inégalitaire entre les individus

# La mort

- La plupart des maladies augmentent le risque de mortalité
- Anticiper ces risques permet d'effectuer de la médecine préventive

# Objectif du hackathon

- Prédire le statut de vie d'un patient en 2019
- Prédiction **binaire**
- L'évaluation de la prédiction se fera sur la base du **F1-score**

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

# Le jeu de données

- 59049 patients uniques majeurs
- 1539 variables



Beaucoup de valeurs  
manquantes

- 3 catégories de variables
  - Issues de questionnaires
  - Issues d'examinations
  - Issues de prises de sang (laboratory)



# Le jeu de données

- Les données ont été collectées entre 1999 et 2018
- La mortalité est statuéée en 2019
- Le jeu de données fourni est déjà pré-filtré par un groupe de biologistes

# Les enjeux du hackathon

- Il ne s'agit pas seulement de prédiction, mais aussi d'analyse
- Distributions des patients, prévalence de certaines pathologies, etc.
- Comment prédire efficacement la mortalité ?

# Fichiers disponibles

- `data.csv` : Le jeu de données
- `features_metadata.csv` : Les informations des variables
- `ground_truth_train.csv` : Le statut de vie de 95 % des patients
- `test_indexes.csv` : La liste des individus sur laquelle votre soumission sera évaluée (5 % des patients)

# Soumission des prédictions

- Sur le site <https://hackathon.alexscod.com/>
- Fichier texte avec :
  - Un entête avec l'id de votre groupe
  - Une ligne par prédiction
  - Dans le même ordre que le jeu de données
  - La soumission doit bien faire 59065 lignes (et une colonne) !

364	Groupe 1
365	Groupe 2
366	Groupe 3
367	Groupe 4
368	Groupe 5
369	Groupe 6

# Évaluation

- Baseline /3 : Si vous dépassez le niveau de la baseline
- Compétition /4
- Présentation orale /13 :
  - Méthodologie /4
  - Analyse approfondie du sujet /4
  - Clarté du code présenté /2
  - Travail en équipe /3