**Projet** **Accompagnement R**
**Octobre 2022** **Durée: 19/10/2022, 18h**

---

**Objectif**: Exploration de données et apprentissage de modèles supervisés.

- Le projet doit être réalisé en binôme ou en trinôme, et **en R**. Merci de préciser les nom et prénom de **chaque membre** du groupe et de mettre en copie votre camarades lors de l'envoi des scripts **commentés** (via Slack, message individuel).
- Les résultats de vos analyses doivent être **commentés** et reprendre les notions vues en cours.
- Votre travail doit s'insérer dans une application Web R Shiny.

---

1. **Données pour ce projet**

   Pour ce mini-projet, vous devez proposer une interface capable de charger **tout type de jeu de données**.

   Il vous faudra donc réfléchir à la prise en compte:
   - du type de chaque variable (ie., qualitative, quantitative)
   - du nombre de catégories en cas de variables qualitative
   - de la présence possible d'outliers
   - des valeurs manquantes
   - de la normalisation
   - de la dummification
   - du déséquilibre des classes

   Pour un jeu de données que vous aurez choisi, vous devrez faire une étude complète à l'aide de l'interface que vous aurez implémentée. La liste des datasets possibles est proposée à la fin de ce document.

2. **Détails des réalisations attendues**

   Votre analyse pourra être mise en ligne *via* votre compte shiny sur `shinyapps.io`. Elle reprendra les différentes explorations vues en cours de *Programmation Web* et des connaissances issues du cours de *Data Science*.

   1. Analyse exploratoire
      (a) Prévoyez un espace d'analyse exploratoire des données.
      (b) Insérer notamment l'analyse unidimensionnelle des variables.
      (c) Insérer également l'analyse bidimensionnelle des variables.
   2. Entraînement de modèles
      (a) Choisissez trois modèles de classification supervisée.
      (b) Entraînez vos modèles sur vos données et faites leur évaluation.
      (c) Affichez les résultats de vos modèles dans l'interface
         i. Precision, Recall, F-score
         ii. courbe ROC, AUC
      (d) Identifiez les features les plus importants.

3. **Jeux de données conseillées**

   Les jeux de données ci-dessous vous sont recommandés. Ils sont mise à votre disposition via le site UCI.

   1. Dermatology: Aim is to determine the type of Eryhemato-Squamous Disease.
   2. Statlog-Heart This dataset is a heart disease database.
   3. Hepatitis This dataset is a hepatitis disease database.
   4. Breast Cancer Wisconsin (Original) Original Wisconsin Breast Cancer Database.
   5. Breast Cancer Wisconsin (Prognostic) Prognostic Wisconsin Breast Cancer Database.
   6. Cardiotocography The dataset consists of measurements of fetal heart rate (FHR) and uterine contraction (UC) features.
   7. Mice Protein Expression Expression levels of 77 proteins measured in the cerebral cortex of 8 classes of control and Down syndrome mice exposed to context fear conditioning, a task used to assess associative learning.
   8. Acute Inflammations The data was created by a medical expert as a data set to test the expert system, which will perform the presumptive diagnosis of two diseases of the urinary system.
   9. Mammographic Mass Discrimination of benign and malignant mammographic masses based on BI-RADS attributes and the patient's age.
   10. Arcene ARCENE's task is to distinguish cancer versus normal patterns from mass-spectrometric data. This is a two-class classification problem with continuous input variables. This dataset is one of 5 datasets of the NIPS 2003 feature selection challenge.

4. **Informations complémentaires**

   - Afin de réduire le problème du déséquilibre des classes, le tutoriel ci-dessous propose des approches de oversampling, qui consiste à augmenter/diminuer le nombre d'instances de la classe minoritaire/majoritaire. Familiarisez-vous avec ces approches et utilisez-les si nécessaire pour réduire le déséquilibre des classes.
     - Resampling strategies for imbalanced datasets