

PROJET DATA SCIENCES

A rendre pour le 16 novembre 2025.

Travail à réaliser en binôme.

Consignes Projet à rendre en **binôme**. Vous devrez mettre en œuvre sur la base de données Breast Cancer Wisconsin

<http://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic>

les différentes méthodes d'apprentissage pertinentes vues en cours. Il est attendu que vous rendiez à la fois un code **lisible** qui permette de reproduire uniquement les résultats pertinents par simple compilation et un document rédigé (LateX ou Notebook) qui détaille votre démarche et vos conclusions en 15 pages maximum.

Description des données La base de données comporte 569 individus pour lesquels 30 variables explicatives sont observées ainsi que la variable binaire **Diagnosis** qui indique si le patient a une tumeur maligne (M) ou bénigne (B) que l'on cherche à classifier.

Objectifs Trouver l'algorithme le plus adapté pour prédire la variable **Diagnosis**. On mettra en œuvre les méthodes vues en cours (lda/qda, régression logistique, forêt aléatoires, k plus proches voisins, ...). On peut suivre le cheminement suivant pour étudier la base de données et rédiger le rapport.

1. **Etude préliminaire.** Donner quelques statistiques descriptives pertinentes et décrire les données (statistiques descriptives univariées, bivariées, les classes sont-elles équilibrées - auquel cas faut-il mettre en œuvre un algorithme pour les rééquilibrer?).
2. **Algorithmes de classifications.** Faire tourner sur les données les algorithmes de classifications vus en cours qui vous semblent pertinents. Dans chaque cas :
 - Justifier en quoi l'usage de l'algorithme est pertinent dans ce contexte.
 - Calibrer les hyper paramètres éventuels.
 - Évaluer le pouvoir prédictif de la méthode (on veillera bien à justifier et à détailler comment l'erreur de classification est évaluée).
3. **Conclusion.** Conclure sur l'algorithme le plus adapté dans ce contexte. Commenter les résultats (quelles variables semblent les plus pertinentes pour expliquer Y ? Comment agissent-elles ? Etc.).