1 Classification de phytoplancton à partir de mesures spectrales

1.1 Objectif du projet

Ce projet a pour but d'appliquer et de comparer différentes méthodes d'apprentissage supervisé sur un jeu de données simple de mesures spectrales de phytoplancton. Bien qu'il ne s'agisse pas d'un cas appliqué réel avec une problématique scientifique poussée, l'objectif est de manipuler plusieurs algorithmes de Machine Learning et Deep Learning, d'évaluer leurs performances, et d'illustrer leurs avantages et inconvénients sur un jeu de données de petite taille.

1.2 Jeu de données

- **Source** : fichier `phytoplankton_spectra_dataset.csv`
- **Description** : chaque observation correspond à un échantillon de phytoplancton décrit par plusieurs mesures spectrales (intensité lumineuse à différentes longueurs d'onde). La variable cible est la classe du phytoplancton.
- **Prétraitement effectué** :
  - Normalisation des variables spectrales (`StandardScaler`)
  - Encodage de la variable cible (`LabelEncoder`)
  - Découpage en jeu d'entraînement (80%) et jeu de test (20%)

1.3 Méthodes utilisées

Quatre méthodes de classification ont été comparées :

1. **Régression logistique** : classifieur linéaire de base, hyperparamètres : `max_iter=1000`.
2. **Random Forest** : modèle d'ensemble basé sur des arbres de décision, hyperparamètres : `n_estimators=200`, `max_depth=None`.
3. **XGBoost** : méthode de boosting gradienté, hyperparamètres : `n_estimators=200`, `learning_rate=0.1`, `max_depth=6`.
4. **Réseau de neurones (MLP)** : une couche cachée de 64 neurones (ReLU), sortie softmax, optimiseur Adam, `epochs=100`, batch size 16.

Les hyperparamètres ont été trouvés par validation croisé sur l'accuracy.

1.4 Évaluation

- **Métrique principale** : précision (accuracy) sur le jeu de test.
- **Validation croisée** : 5-fold.

Modèle	**Accuracy moyenne (%)**
Régression logistique	85.0
Random Forest	92.3
XGBoost	94.1
Réseau de neurones	91.0

TABLE 1 – Performances moyennes des modèles sur le jeu de test

1.5 Analyse des résultats

La matrice de confusion du meilleur modèle (XGBoost) est :

$$\begin{bmatrix} 12 & 0 & 0 \\ 0 & 14 & 1 \\ 0 & 1 & 13 \end{bmatrix}$$

Les diagonales élevées indiquent que XGBoost discrimine bien les classes. Quelques confusions subsistent entre deux classes proches, probablement en raison de spectres similaires.

1.6 Conclusion

XGBoost est le modèle le plus performant sur ce jeu de données, suivi de près par la Random Forest. La régression logistique fonctionne correctement mais reste limitée par sa nature linéaire. Le réseau de neurones a donné de bons résultats malgré la petite taille du jeu de données, mais n'a pas surpassé les méthodes d'ensemble. Ce projet illustre la diversité des approches en Machine learning / Deep learning et leurs différences de comportement selon la complexité et la taille des données.