École Nationale Supérieure d'Informatique et d'Analyse des Systèmes

Filière : *Génie Logiciel*

DNA Barcoding Using ML & DL

Encadrant :

Pr.TABII Youness

Réalisé par :

ASRI Wassila

2024- 2025

Table des matières

1 Introduction

1.1 Contexte du problème

L'identification des espèces végétales joue un rôle central dans les domaines de la biologie, de l'écologie et de la conservation. Toutefois, les méthodes classiques basées sur des critères morphologiques peuvent s'avérer longues, coûteuses et sujettes à des erreurs, en particulier pour les espèces cryptiques ou les stades précoces de développement. Pour remédier à ces limitations, le **DNA barcoding** a émergé comme une approche rapide et fiable pour l'identification des espèces basée sur des séquences génétiques courtes et standardisées.

1.2 Objectif du projet

Ce projet vise à construire un modèle de classification permettant de **prédire automatiquement l'espèce d'une plante** à partir de sa séquence ADN brute, combinée à sa région génique (*gene region*). L'objectif est de tester et comparer plusieurs approches de machine learning afin de déterminer la plus efficace pour cette tâche de classification multi-classes.

1.3 Intérêt du Machine Learning dans ce contexte

L'utilisation du machine learning dans le DNA barcoding permet :

- D'automatiser le processus d'identification d'espèces sur la base de grandes quantités de séquences.
- D'exploiter efficacement les motifs présents dans les séquences ADN (via des k-mers) et les métadonnées comme la région génique.
- D'améliorer la précision et la rapidité d'identification par rapport aux méthodes manuelles.

Ainsi, l'apprentissage automatique permet d'accélérer la recherche en taxonomie et en biodiversité tout en garantissant une meilleure reproductibilité des résultats.

2 Exploration et Préparation des Données

2.1 Présentation des données

Les données utilisées proviennent de la base **BOLD Systems (Barcode of Life Data Systems)**. Un fichier FASTA a été transformé en CSV et nettoyé pour ne garder que les colonnes pertinentes :

	Gene_Region	Class	Order	Family	Genus	Species	Sequence
0	rbcLa	Magnoliopsida	Gentianales	Apocynaceae	Schizoglossum	Schizoglossum atropurpureum	AGTGTTGGATTCAAAGCCGGTGTTAAAGAGTACAAATTGACTTATT...
1	matK	Magnoliopsida	Gentianales	Apocynaceae	Schizoglossum	Schizoglossum atropurpureum	GATATACTAATACCCTACCCTGTTCATCTGGAAATCTTGGTTCAAA...
3	rbcLa	Magnoliopsida	Gentianales	Apocynaceae	Schizoglossum	Schizoglossum atropurpureum	AGTGTTGGATTCAAAGCCGGTGTTAAAGAGTACAAATTGACTTATT...
4	matK	Magnoliopsida	Gentianales	Apocynaceae	Schizoglossum	Schizoglossum atropurpureum	GATATACTAATACCCTACCCTGTTCATCTGGAAATCTTGGTTCAAA...
11	matK	Magnoliopsida	Gentianales	Apocynaceae	Schizoglossum	Schizoglossum bidens	TCTGGAAATCTTGGTTCAAACCCTTCGCTATTGGGTAAAGGATGCC...

- **Gene_Region** : région du gène séquencée (ex : `rbcL`, `matK`).
- **Class, Order, Family, Genus** : informations taxonomiques hiérarchiques.
- **Species** : étiquette cible à prédire.
- **Sequence** : séquence ADN brute (chaîne de caractères ACGT).

2.2 Prétraitement des données

2.2.1 Nettoyage

Le nettoyage des données a consisté à parser le fichier FASTA brut pour extraire les informations pertinentes : région génique, taxonomie, et séquence ADN. Les colonnes inutiles telles que `ID`, `Country`, `Kingdom`, `Phylum`, `Subfamily` et `Other` ont été supprimées afin de réduire la redondance et se concentrer uniquement sur les variables utiles à la classification.

2.2.2 Traitement des valeurs manquantes

Les lignes contenant des champs critiques vides (`Gene_Region`, `Class`, `Family`, `Genus`, `Order` ou `Sequence`) ont été éliminées. Cette étape garantit que les modèles d'apprentissage ne seront pas biaisés par des données incomplètes.

2.2.3 Encodage

Les variables catégorielles, telles que la région génique (`Gene_Region`), ont été transformées à l'aide d'un `OneHotEncoder`, permettant ainsi de convertir chaque catégorie en vecteur binaire utilisable par les modèles. Pour la variable cible `Species`, un `LabelEncoder` a été appliqué afin d'obtenir une représentation numérique des différentes espèces.

2.2.4 Vectorisation

La séquence ADN a été vectorisée grâce à deux méthodes distinctes : `CountVectorizer` et `TfidfVectorizer`, en utilisant l'approche des **k-mers**. Chaque séquence est découpée en sous-chaînes de longueur k (ex : $k = 6$) et transformée en vecteur de fréquences d'occurrence ou de poids TF-IDF.

2.2.5 Normalisation

La normalisation n'a pas été appliquée directement car les méthodes de vectorisation textuelle comme `CountVectorizer` et `TF-IDF` génèrent déjà des vecteurs numérisés exploitables par les modèles. Toutefois, pour certains modèles comme les réseaux de neurones, une standardisation ou une mise à l'échelle des vecteurs pourrait être envisagée.

3 Modélisation

3.1 Modèles utilisés

Les modèles suivants ont été entraînés et évalués dans le cadre de ce projet :

- **Random Forest**
- **Naive Bayes**
- **Support Vector Machine (SVM)**
- **XGBoost**
- **Réseaux de neurones (MLP, CNN, RNN)**

3.2 Vectorisation des séquences ADN

Les séquences ont été vectorisées à l'aide de :

- `CountVectorizer` avec k-mers
- `TF-IDF Vectorizer` avec k-mers

3.3 Encodage de la région génique

La région génique a été encodée avec :

- `OneHotEncoder`

3.4 Tuning des hyperparamètres

L'optimisation des hyperparamètres a été réalisée à l'aide de :

- `GridSearchCV`

4 Évaluation des Modèles

4.1 Métriques utilisées

Les performances des modèles ont été évaluées à l'aide des métriques suivantes :

- **Accuracy** : proportion globale des prédictions correctes.
- **F1 Macro** : moyenne des F1-scores de chaque classe (pondérée équitablement).
- **Précision Macro** : proportion de vraies prédictions positives parmi toutes les prédictions positives.
- **Recall Macro** : capacité du modèle à retrouver toutes les instances positives.

4.2 Comparaison des modèles

Le tableau suivant résume les performances des différents algorithmes testés sur le jeu de test :

Modèle	Accuracy	F1 Macro	Précision Macro	Recall Macro
Random Forest	0.87	0.84	0.86	0.85
Naive Bayes	0.58	0.53	0.58	0.55
SVM (SGDClassifier)	0.84	0.80	0.82	0.83
XGBoost	0.83	0.80	0.82	0.81
Réseau de Neurones (MLP)	0.80	0.77	0.78	0.79
CNN	0.34	–	–	–
RNN (LSTM)	0.05	–	–	–

Le **Random Forest** reste le meilleur choix pour ce projet, notamment avec une vectorisation de type **k-mer (k=6)** et un encodage explicite de la **région génique**. Les approches de deep learning n'ont pas surpassé les méthodes classiques ici, ce qui souligne l'importance de l'adéquation entre modèle, volume de données, et qualité de la représentation des séquences.

5 Prédictions sur Données Réelles

5.1 Test sur une nouvelle séquence

on va prédicter l'espèce de la séquence suivante :
sequence="GGTGTTGGATTTCAAGCTGGTGTTAAAGATTATAAATTGACTTACTACACCCCAG
AGTATGAAACTAAGGATACTGATATCTTGGCAGCATTCCGAGTAAGTCCTCAGCCTGGGGTTC
CGCCCGAAGAAGCAGGGGCTGCAGTAGCTGCCGAATCTTCTACTGGTACATGGACAACTGTT
TGGACTGATGGACTTACCAGTCTTGATCGTTACAAAGGACGATGCTATCACATCGAGCCTGTT
GCTGGGGAAGACAACCAATGGATCTGTTATGTAGCTTATCCATTAGACCTATTTGAGGAGGG

TTCCGTTACTAACATGTTTACTTCCATTGTGGGTAACGTATTTGGGTTCAAAGCCCTACGTGCTC
CCCCCTACTTATTCAAAAACTTTCCAAGGCCCGCCTCATGGTATCCAAGTTGAAAGAGATAAG
TTGAACAAGTATGGTCGTCCTTTATTGGGATGTACTATTAAACCAAAATTGGGATTATCCGC
AAAAAATTATGGTAGAGCGTGTTATGAGTGTCTA"
gene_region="rbcLa"

- **RANDOM FOREST** :Espèce prédite : Hordeum jubatum
- **SVM** :Espèce prédite : Hordeum jubatum

les deux modèles ont bien prédictés l'espèce.

6 Améliorations Apportées

6.1 Optimisation des paramètres

Une recherche par grille (*Grid Search*) a été réalisée sur le modèle Random Forest afin d'optimiser les hyperparamètres tels que :

- Le nombre d'arbres (`n_estimators`),
- La profondeur maximale (`max_depth`),
- La méthode de sélection des features (`max_features`),
- Le seuil de séparation minimum (`min_samples_split`).

6.2 Recherche du meilleur k

Pour identifier la taille de k-mer optimale lors de la vectorisation des séquences ADN, nous avons testé différentes valeurs de k avec deux méthodes de vectorisation : `CountVectorizer` et `TfidfVectorizer`. Les résultats sont présentés ci-dessous.

Résultats avec `CountVectorizer`

k-mer	Accuracy
3	0.8466
4	0.8571
5	0.8647
6	**0.8707**
7	0.8677
8	0.8677

TABLE 1 – Accuracy obtenue pour différents k avec `CountVectorizer`.

Résultats avec `TfidfVectorizer`

k-mer	Accuracy
3	0.8331
4	0.8316
5	0.8286
6	0.8346
7	**0.8361**
8	0.8316

TABLE 2 – Accuracy obtenue pour différents k avec `TfidfVectorizer`.

6.3 Évaluation croisée

K-Fold	Accuracy moyenne	Écart-type
3	0.8260	0.0069
5	0.8446	0.0047
7	0.8522	0.0123
10	**0.8594**	0.0104

TABLE 3 – Résultats de validation croisée avec K-Fold.

Stratified K-Fold	Accuracy moyenne	Écart-type
3	0.8449	0.0043
5	**0.8642**	0.0108
7	0.8624	0.0098
10	0.8627	0.0078

TABLE 4 – Résultats de validation croisée avec Stratified K-Fold.

6.4 Utilisation des modèles profonds

Des modèles de deep learning ont été testés :

- Un réseau de neurones dense (MLP) qui a obtenu une accuracy de 80%.
- Un modèle CNN (réseau de neurones convolutif) appliqué sur les séquences vectorisées, atteignant une accuracy de 33%.
- Un modèle RNN basé sur LSTM, qui n'a pas dépassé 5% d'accuracy, en raison d'une architecture inadaptée et du manque d'encodage spécialisé.

Ces essais ont permis de valider la supériorité du Random Forest pour ce cas précis, tout en identifiant les limites des approches profondes sur ce jeu de données.

7 Conclusion

7.1 Résumé des résultats

Ce projet a permis d'explorer l'application du Machine Learning à la classification d'espèces végétales à partir de séquences ADN, en utilisant la méthode du DNA barcoding. Plusieurs modèles ont été entraînés, testés et comparés. Le modèle Random Forest, avec un encodage des séquences en k-mers ($k = 6$) via `CountVectorizer` et l'ajout de la région génique en entrée, a donné les meilleurs résultats, atteignant une accuracy de 87%.

7.2 Évaluation globale

Les performances des modèles ont montré que la combinaison de techniques de prétraitement avancées (nettoyage, vectorisation, encodage) et de modèles robustes permet d'obtenir de très bons résultats, même avec un problème multi-classes complexe. Le modèle Random Forest s'est distingué par sa stabilité, sa robustesse, et sa capacité à gérer des données mixtes (catégorielles + séquences).

7.3 Prochaines étapes

Pour approfondir ce travail, plusieurs pistes peuvent être envisagées :

- Prédire la taxonomie d'un espèce à partire de son ADN
- Étendre la base de données à d'autres espèces).
- Intégrer une interface utilisateur pour faciliter l'utilisation du classificateur.

Annexes

Lien vers le dataset :

Les données utilisées dans ce projet proviennent de la base BOLD (Barcode of Life Data System), spécialisée dans les séquences ADN utilisées pour le DNA barcoding. `https://bench.boldsystems.org/index.php/datapackage?id=BOLD_Public.28-Mar-2025`

Lien vers le code source

Le code source complet, incluant le notebook, le traitement des données, l'entraînement des modèles et les visualisations, est disponible à l'adresse suivante :

`https://github.com/Wassila00/DNA-Barcoding.git`