Université de Sherbrooke

Faculté des Sciences

Département d'Informatique

IFT 870 / BIN 710 – Forage de données.

Projet de Session – Livrable 1 :

Prédiction du groupe écologique des oiseaux à partir de leurs mesures osseuses.

Professeur :

Nadia Tahiri

Auteurs :

- Nom et Prénom : Amina Ferraoun (CIP : FERA1401)
- Nom et Prénom : François Soulié (CIP : SOUF2202)
- Nom et Prénom : Paul Tissedre (CIP : TISP1301)

Introduction

Le sujet du projet de session que nous avons choisi vise à explorer la relation entre les mesures osseuses et les groupes écologiques des oiseaux. En utilisant un ensemble de données contenant des mesures osseuses de différentes espèces d'oiseaux, nous chercherons à développer un modèle capable de prédire le groupe écologique d'un oiseau en se basant uniquement sur ses mesures osseuses.

Cette tâche de classification multi classe peut avoir des applications importantes dans la compréhension de l'évolution et de l'adaptation des oiseaux à leur environnement.

Problématique

Peut-on prédire avec précision le groupe écologique d'un oiseau (p. ex. nageant, échassier, terrestre, rapace, grimpeur, chanteur) en se basant uniquement sur les mesures de ses os (longueur et diamètre de l'humérus, de l'ulna, du fémur, du tibiotarse, du tarsométatarsus) ?

Description du projet

Notre sujet de recherche se concentre sur l'exploration et l'analyse des relations entre la morphologie osseuse des oiseaux et leurs groupes écologiques. L'intérêt principal réside dans l'identification des caractéristiques physiques spécifiques qui distinguent les différents groupes écologiques d'oiseaux, en utilisant les mesures de 420 spécimens d'oiseaux représentant une vaste diversité d'espèces.

Ces spécimens sont classifiés en six groupes écologiques principaux, sur la base de leurs habitudes de vie et de leur adaptation à leur environnement. Les données, comprenant 10 mesures morphologiques pour chaque oiseau, offrent une base riche pour l'application de techniques de forage de données et d'analyse statistique afin de découvrir des patterns significatifs.

Notre problématique implique une tâche de classification multi classe (prédire l'une des six classes écologiques) à partir de données numériques (les mesures des os), ce qui pourrait être exploré à l'aide de divers algorithmes d'apprentissage automatique.

Contexte

De multiples espèces d'oiseaux peuplent notre planète, allant des pigeons urbains aux autruches des savanes, en passant par les pingouins des régions polaires. Leurs capacités varient grandement : certains excellents dans l'art du vol, d'autres spécialisés dans la course rapide, ou encore certains maîtres-nageurs sous l'eau, tandis que d'autres préfèrent les eaux peu profondes pour se déplacer.

Ces oiseaux sont répartis en divers groupes écologiques, basés sur leurs modes de vie et habitats. On distingue ainsi huit catégories principales :

- Oiseaux aquatiques
- Oiseaux des zones humides
- Oiseaux de terre
- Oiseaux de proie
- Oiseaux grimpeurs
- Oiseaux chanteurs
- Oiseaux coureurs (absents de notre base de données)
- Oiseaux maritimes (absents de notre base de données)

Notre étude se concentre sur les six premières catégories, qui constituent les groupes principaux présents dans notre base de données.

Les caractéristiques physiques des oiseaux, telles que la robustesse des ailes chez les espèces volantes ou la longueur des jambes chez celles fréquentant les milieux aquatiques peu profonds, témoignent de leur appartenance à un groupe écologique spécifique. En tant que data scientists, notre intérêt pourrait se porter sur l'analyse des liens entre la morphologie osseuse et le groupe écologique de chaque espèce, permettant ainsi d'identifier le groupe écologique d'un oiseau à partir de la structure de ses os.

Description des données

Notre étude se base sur un jeu de données accessible via Kaggle sur ce lien : https://www.kaggle.com/datasets/zhangjuefei/birds-bones-and-living-habits/data.

Il porte sur les oiseaux, leurs os et leurs habitudes de vie. Ce jeu de données recense 420 spécimens d'oiseaux, chacun caractérisé par une série de 10 mesures :

- Longueur et diamètre du humérus
- Longueur et diamètre de l'ulna (cubitus)
- Longueur et diamètre du fémur
- Longueur et diamètre du tibiotarse (tibia)
- Longueur et diamètre du tarsométatarsien

Toutes ces mesures sont des nombres flottants continus exprimés en millimètres. Les squelettes inclus dans ce jeu de données proviennent des collections du Musée d'Histoire Naturelle du Comté de Los Angeles. Ils appartiennent à 21 ordres, 153 genres et 245 espèces différentes.

Chaque spécimen d'oiseau dans le jeu de données est également identifié par un label correspondant à son groupe écologique :

- SW : Oiseaux nageurs
- W : Oiseaux vadrouilleurs
- T : Oiseaux terrestres
- R : Rapaces
- P : Oiseaux grimpeurs
- SO : Oiseaux chanteurs

Ces labels facilitent l'analyse comparative et la classification en fonction des habitudes de vie et des habitats des différentes espèces d'oiseaux. C'est ce qu'on cherche donc à classifier. La diversité des mesures fournies permet une exploration détaillée de la morphologie des oiseaux et offre la possibilité d'étudier les relations entre la structure osseuse et les adaptations écologiques.

Importance du sujet et motivations

Dans le contexte du réchauffement climatique et de la disparition de nombreuses espèces d'animaux, être capable de déterminer à quel groupe écologique appartient la dépouille d'un oiseau pourrait nous permettre de prévenir la disparition de ce dernier. Notre projet ne s'inscrirait pas en tant que finalité dans ce procédé mais plutôt comme un outil.

Aussi, en analysant ces données, nous espérons contribuer à la compréhension scientifique des liens entre la morphologie des oiseaux et leur écologie. Cette recherche pourrait fournir des insights précieux sur la manière dont les différentes espèces se sont adaptées à leurs environnements spécifiques au fil de l'évolution.

Finalement, notre curiosité pour la biodiversité et le monde naturel nous as conforter dans notre choix de sujet. Explorer la diversité des oiseaux à travers leurs données morphologiques nous permet de célébrer et de mieux comprendre la complexité de la vie sur Terre.

Évolution des Méthodes : Clés et Études de Cas

Les méthodes actuelles sont principalement axées sur la classification d'images. L'espèce d'un oiseau est alors déterminée en prenant une photographie de ce dernier. Ces méthodes relèvent toujours de l'intelligence artificielle et peuvent utiliser des entraînements supervisés ou non supervisés avec des réseaux de neurones convolutifs, par exemple.

Les méthodes traditionnelles de classification d'images, telles que les k-plus proches voisins, étaient largement utilisées dans les années 2000. L'avènement des réseaux de neurones convolutifs a ensuite révolutionné la classification d'images, y compris la classification des espèces d'oiseaux. De nos jours, l'intégration de données supplémentaires, telles que des enregistrements audios en plus d'images, semble être un domaine de recherche significatif pour améliorer la précision des modèles existant.

Analyse des données

Notre démarche analytique a débuté par l'importation des bibliothèques essentielles et du jeu de données, suivie d'une phase d'analyse exploratoire approfondie. Cette phase initiale nous a permis de comprendre la structure et les caractéristiques des données à notre disposition. Nous avons ensuite procédé à une phase de prétraitement rigoureuse, au cours de laquelle nous avons éliminé l'attribut « id », jugé superflu pour notre analyse, tout en veillant à l'intégrité des données en vérifiant et remplaçant les valeurs manquantes. Cette étape a assuré la complétude de notre dataset. Par ailleurs, la normalisation des données et la conversion des valeurs cibles en nombres à l'aide d'une technique d'encodage ont été mises en œuvre pour préparer le terrain à des analyses plus avancées.

L'examen de la corrélation entre les variables a révélé une forte interdépendance, suggérant la possibilité de réduire la dimensionnalité des données. Nous avons opté pour une Analyse en Composantes Principales (ACP), retenant les trois premiers composants principaux comme les plus informatifs. Cependant, malgré la promesse d'une simplification des données, l'ACP a été mise de côté car elle tendait à diluer l'information essentielle, sans pour autant améliorer significativement la performance de nos modèles de classification.

Pour aborder notre question de recherche, nous avons exploré divers algorithmes de classification, en scindant notre jeu de données en un ensemble d'entraînement (80%) et un ensemble de test (20%), afin d'assurer une représentation fidèle pour l'entraînement de nos modèles. Quatre algorithmes ont été testés, d'abord avec leurs paramètres par défaut, puis en affinant ces paramètres via GridSearch pour optimiser leurs performances :

- K plus proches voisins : Initialement, avec cinq voisins, nous avons atteint une précision de 73%. L'optimisation a poussé ce chiffre à 83%.
- Régression logistique : A produit une précision de 80% avec les réglages par défaut, qui a grimpé à 88% après ajustement des paramètres.
- Centroïde : A démontré une efficacité limitée, avec seulement 48% de précision qui s'est légèrement améliorée à 49% après optimisation.
- Réseau de neurones : S'est révélé être le plus performant avec une précision impressionnante de 90% sans ajustement, bien que l'optimisation ait légèrement réduit cette performance à 86%.

Fondés sur ces observations, nous avons privilégié le réseau de neurones comme notre modèle de choix. Cette décision a été motivée par sa performance supérieure, démontrant une aptitude remarquable à modéliser la complexité de nos données sans nécessiter d'ajustements fastidieux.