Titre : Analyse exploratoire et modélisation des crocodiles

Sous-titre : Projet Data Science

: Objectifs du projet

Points clés :

.Identifier quelles espèces de crocodiles existent dans le monde

.Comprendre les différences biologiques entre espèces (taille, poids, âge à maturité)

.Créer des modèles pour prédire certaines caractéristiques des crocodiles

.Fournir des insights pour la conservation et la recherche scientifique

L'objectif n'est pas seulement de collecter des données, mais aussi de les analyser pour aider à protéger les espèces et mieux comprendre leur biologie.

Compréhension du métier

Les crocodiles jouent un rôle clé dans leurs écosystèmes aquatiques

Connaître leur taille, leur poids et leur habitat aide à :

- Protéger les espèces menacées
- Étudier leur comportement et croissance
- Identifier les zones à risque ou vulnérables

Les biologistes et écologistes peuvent utiliser ces données pour des décisions concrètes

Cette diapositive met en contexte l'importance des données et du projet pour la biodiversité et la science appliquée.

* **Description du dataset**

* Identification et classification : Observation ID, Common Name, Scientific Name, Family, Genus
* Caractéristiques physiques : Observed Length (m), Observed Weight (kg)
* Informations démographiques : Age Class, Sex
* Localisation : Country/Region, Habitat Type
* Statut de conservation : Conservation Status
* Notes et observateurs : Observer Name, Notes
* Ces colonnes permettent de décrire chaque crocodile observé et de comprendre ses caractéristiques biologiques et son environnemen

Objectifs de l'exploration des données

Identifier quelles colonnes sont **numériques** (mesures) et **catégorielles** (espèce, sexe, habitat)

Détecter les **valeurs manquantes ou aberrantes** pour assurer la fiabilité de l'analyse

Visualiser la **distribution des tailles, poids et âges**

Chercher des **relations entre les variables** qui peuvent guider la modélisation

Cette étape est cruciale pour nettoyer les données et préparer un modèle fiable.

* **Analyse exploratoire des données**
* **Visualisations possibles :**
* Histogrammes pour observer la répartition des longueurs et poids
* Boxplots par famille ou genre pour identifier les variations
* Countplots pour visualiser le nombre d'observations par espèce ou statut de conservation
* Heatmap pour voir les corrélations entre variables
* On regarde les données sous toutes les angles pour comprendre leur structure avant de construire des modèles.

* **Nettoyage des données**
* Supprimer les doublons pour éviter de biaiser les résultats
* Remplir les valeurs manquantes :
  * Numériques par la **médiane**
  * Catégorielles par la **valeur la plus fréquente**
* Encodage des variables catégorielles pour que les modèles puissent les utiliser
* Normalisation des mesures (taille et poids) pour comparer les différentes espèces
* Un nettoyage minutieux est essentiel pour éviter des erreurs dans les modèles et obtenir des résultats fiables.

* **Préparation pour la modélisation**
* Définir **features** (variables explicatives) et **target** (variable à prédire)
* Séparer les données en **train (80%)** et **test (20%)** pour évaluer le modèle
* Choix des modèles :
  * **Classification** : prédire Sex ou Age Class
  * **Régression** : prédire Observed Weight (kg) ou Observed Length (m)
* On prépare les données pour que les modèles puissent apprendre à faire des prédictions fiables.

* **Modélisation**
* Utilisation des **Random Forest Classifier** et **Random Forest Regressor**
* Optimisation des hyperparamètres avec GridSearchCV ou RandomizedSearchCV
* Évaluation des modèles :
  * Classification **accuracy, precision, recall, F1-score**
  * Régression **RMSE, R^2**
* Ces modèles permettent de prédire les caractéristiques biologiques des crocodiles et de savoir quelles variables sont les plus importantes.

* **Importance des variables**
* Visualisation des variables les plus influentes :
  * Exemple pour la régression du poids : Observed Length (m) et Sex sont les plus déterminants
* Permet de comprendre quels facteurs expliquent le mieux les différences entre individus
* Cela aide les biologistes à savoir quelles caractéristiques sont clés pour la croissance et le poids des crocodiles.

* **Évaluation des modèles**
* **Classification :** Accuracy élevée, F1-score montre la qualité des prédictions de l'âge ou sexe
* **Régression :** RMSE faible et R^2 proche de 1 → modèle fiable pour prédire poids ou taille
* Comparaison des modèles pour choisir celui qui prédit le mieux
* On vérifie si les modèles donnent des résultats réalistes et utilisables pour la recherche.

* **Recommandations**
* Utiliser le modèle pour :
  * Prédire la croissance ou le poids moyen des crocodiles
  * Identifier les espèces vulnérables
  * Orienter la conservation et les études scientifiques
* Créer des dashboards pour visualiser les données par pays ou habitat
* Ces recommandations sont orientées vers des actions concrètes pour la recherche et la protection des crocodiles.

* **Conclusions**
* Les modèles Random Forest donnent de bonnes performances
* Variables clés identifiées : longueur, sexe, âge
* Limites : données incomplètes, biais possibles selon les observateurs
* On résume les points clés pour montrer l'efficacité et les limites du projet.