Machine learning I
TP 2 - Régression

Issam Falih

Instructions : Préparez un rapport incluant le code source et vos résultats, et déposez-le sur Moodle. Il est recommandé d'utliliser streamlit pour afficher vos résultats .Pas plus de 2 personnes par groupe. N'oubliez pas de mettre les 2 noms sur le rendu.

Contexte

Ce TP a pour but d'étudier la relation entre la masse corporelle (`BOW`) et la masse du cerveau (`BRW`) de différentes espèces de chauves-souris, en utilisant la régression linéaire simple. Les données sont fournies dans le fichier `tabBats.txt` sur boostcamp avec les colonnes :

- `Species` : nom de l'espèce observée
- `Diet` : régime alimentaire (1=phytophage ; 2=gleaner ; 3=aerial insectivore ; 4=vampire)
- `BOW` : Body mass (masse corporelle)
- `BRW` : Brain mass (masse du cerveau)
- `AUD`, `MOB`, `HIP` : volumes de différentes structures cérébrales

Partie A :Chargement et exploration des données

1. Chargez le fichier `tabBats.txt` en utilisant la librairie `pandas`.
2. Affichez les premières lignes du tableau, les types de colonnes et les statistiques descriptives.
3. Sélectionnez uniquement les variables pertinentes pour la régression (`BOW` et `BRW`, et éventuellement `Species` pour identifier des outliers).

Partie B : Première régression linéaire simple

1. Tracez le nuage de points (`BRW` en fonction de `BOW`). Décrivez la tendance observée. Y a-t-il des espèces atypiques ?

2. Ajustez un modèle de régression linéaire simple (`BRW` $\sim$ `BOW`) avec `statsmodels`.
3. Affichez le résumé du modèle et commentez :
   - les coefficients estimés,
   - la signification statistique,
   - le coefficient R^2,
   - l'analyse des résidus.
4. Tracez la droite de régression obtenue sur le nuage de points.

Partie C :Analyse avec retrait d'une espèce atypique

1. Créez un second tableau `tab2` en retirant l'espèce *Pteropus vampyrus*.
2. Comparez visuellement les nuages de points (`tab` vs `tab2`).
3. Ajustez un second modèle de régression sur `tab2` et comparez ses résultats avec le premier :
   - coefficients,
   - R^2,
   - qualité des résidus.
4. Superposez les deux droites de régression (avec et sans *Pteropus vampyrus*) et commentez l'effet de cette espèce atypique.