# Contrôle Science des données 2

L3 MIASHS

10 décembre 2023

1h30

Notes de cours et TP autorisées, documentation concernant python et ses bibliothèques autorisées.

Tout recours à des outils tels que (mais pas seulement) ceux connus sous le nom d’ « IA génératives » est évidemment strictement interdit (et totalement inutile).

Il est bien évident qu’il s’agit d’un travail personnel réalisé par la seule personne qui écrit son nom ci-dessous.

Prénom :

Nom :

Ci-dessous vous indiquez la réponse aux questions. Toutes les commandes python doivent m’être envoyées à philippe.preux@univ-lille.fr dans un fichier Prénom-Nom.cc-sd2.py avec le compte-rendu en Prénom-Nom.cc-sd2.**pdf**. Le fichier .py contient le python qi permet de répondre aux questions ; le fichier pdf contient les réponses aux questions. Vous remplacez Prénom et Nom dans les noms de fichier par votre prénom et votre nom.

Le fichier python que vous m’envoyer doit s’exécuter sans erreur.

Remarque : dans ce document, le style des paragraphes où vous indiquez les réponses aux questions se nomme réponse. Merci d’utiliser ce style : il affiche le texte en bleu ; si votre réponse n’apparaît pas en bleu, c’est que vous n’utilisez pas le bon style.

**Conseil : réfléchissez avant d’agir !**

# Première partie

On utilise le jeu de données disponible à l’url

<https://philippe-preux.github.io/ensg/miashs/l3-sd2/tp/cc2/exo-cc2.csv>.

Ce jeu de données est artificiel : je l’ai créé pour ce contrôle. Il contient des données, chacune étant décrite par des attributs et une classe (attribut dénommé classe).

Vous répondez aux questions suivantes. Votre fichier Python doit contenir les commandes permettant de répondre aux questions. Indiquez clairement à quelle question vous répondez dans le fichier Python.

1. Dans le fichier Python, chargez le fichier dans un objet dénommé exocc2.
2. Quel est le nombre de données (individus) contenues dans ce jeu de données ?
3. Quel est le nombre d'attributs décrivant les données ?
4. Quelle est la liste des attributs ?

Ce jeu de données contient des exemples de différentes classes, chacune identifiée par une couleur.

1. Combien y a-t-il de classes et combien y a-t-il d’exemples de chaque classe ?
2. On réalise une exploration visuelle du jeu de données en vue de déterminer si la prédiction de la classe semble faisable par un arbre de décision. Ci-dessous, vous expliquez comment vous procédez et ce que vous constatez ; n’hésitez pas à insérer des figures qui vous paraîtraient utiles pour justifier votre démarche. Un arbre de décision semble-t-il pouvoir déterminer parfaitement la classe ? Dans votre fichier Python, vous indiquez les instructions que vous exécutez afin de répondre à cette question. N’hésitez pas à y mettre des commentaires pour que l’on comprenne ce que vous faites.
3. En vous appuyant sur votre réponse à la question précédente, comment faites-vous pour construire un bon (le meilleur possible) arbre de décision ? Vous expliquez ci-dessous votre démarche. Dans le fichier python, vous mettez le code qui permet de faire tout cela.
4. Insérez ci-dessous une représentation graphique de l’arbre obtenu. Faites-en sorte qu’il soit joli et lisible.
5. Quelle est l’erreur de test de cet arbre ? Vous mesurez cette erreur en induisant 100 arbres de décision dont vous mesurez le taux de succès et vous en indiquez la moyenne et l’écart-type.
6. Quelle est l’erreur de test de cet arbre pour chacune des classes ?

# Seconde partie

On utilise le jeu de données disponible à l’url

<https://philippe-preux.github.io/ensg/miashs/l3-sd2/tp/cc2/sleep.csv>[.](https://philippe-preux.github.io/ensg/miashs/l3-sd2/tp/cc1/australian-crabs.csv)

Ce jeu de données est composé d'un ensemble de lignes. Chaque ligne correspond à un mammifère. Il a été collecté pour étudier le sommeil de différentes espèces de mammifères et révéler des relations entre la durée du sommeil, en distinguant sommeil paradoxal (pendant lequel les animaux rêvent) et sommeil non paradoxal (pendant lequel les animaux ne rêvent pas) et certaines caractéristiques de ces animaux. Les attributs sont les suivants :

* Species est le nom de l'espère animale.
* Body.Weight est la masse moyenne d'un animal de cette espèce (en kg).
* Brain.Weight est la masse du cerveau (en g).
* Non.Dreaming est la durée du sommeil non paradoxal (en heures/jour).
* Dreaming est la durée du sommeil paradoxal (en heures/jour).
* Sleep est la durée totale de sommeil (en heures/jour).
* Lifespan est la durée de vie (en années).
* Gestation est la durée de gestation (en jours).
* Predation est un indice de prédation entre 1 et 5 : de 1 pour les animaux les plus prédateurs (les moins proies), à 5 pour les animaux les moins prédateurs (les plus proies).
* Sleep.exposure est un indice d'exposition au danger pendant le sommeil de 1 à 5 : de 1 pour les animaux qui sont les moins exposés au danger pendant leur sommeil, à 5 pour les animaux les plus en danger pendant leur sommeil.
* Endangered est un indice synthétisant le fait qu'un animal est en danger des autres animaux ou pas : 1 pour les moins en danger, 5 pour les plus en danger.

Les attributs Species, Gestation, Predation, Sleep.exposure, Endangered sont des attributs catégoriques. Des valeurs d’attribut sont manquantes.

Avec ce jeu de données, on se pose deux questions : y a-t-il des relations entre les attributs liés à la durée du sommeil et les autres ? Et peut-on utiliser ces relations pour donner une valeur à la place de certaines valeurs inconnues ?

Vous répondez aux questions suivantes. Votre fichier Python doit contenir les commandes permettant de répondre aux questions. Indiquez clairement à quelle question vous répondez dans le fichier Python.

1. Dans le fichier Python, chargez le fichier dans un objet dénommé sommeil. Faites ce qu’il convient pour les attributs catégoriques.
2. Quel est le nombre de données contenues dans ce jeu de données et quel est le nombre d’attributs par donnée ?
3. Par exploration visuelle, déterminer des corrélations qui existeraient entre l’un des attributs liés à la durée du sommeil (Non.Dreaming, Dreaming, Sleep) et un autre attribut.
4. Déterminer des modèles linéaires liant l’un de ces attributs à un autre (appuyez-vous sur ce que vous avez constaté visuellement à la question précédente). Expliquez comment vous faites ci-dessous.
5. Estimez les valeurs manquantes des attributs Non.Dreaming, Dreaming, Sleep. Pouvez-vous le faire pour toutes les espèces où il manque l’une de ces données ?