Master M2, SITN et Data Science-maths, Université Claude Bernard, Lyon 1
Analyse factorielle, année 2020-2021

SUJET numéro 6,
Projet Partie 1:
Note: la partie 2 sera distribuée au début janvier 2021. Le rapport sera à rendre au plus tard le 27 janvier 2021.

***Pour tous les problèmes**, il faut d'abord faire une **description de l'expérience** qui a permis d'obtenir les données. Ensuite, il faut réaliser une **étude descriptive (avec interprétations)**.*
Si les cours reprennent en présentiel jusqu'au 27 janvier 2021, le rapport (sans le code et les sorties du logiciel) est à rendre sur support papier, sinon, il faut me l'envoyer par mail sous la forme d'un fichier pdf. Le code et les sorties du logiciel sont à envoyer par mail.
Rédaction en anglais acceptée.
Conseil: télécharger les données et leur description le plus tôt possible (pour éviter des problèmes liés à l'indisponibilité du site internet).

Les problèmes utilisent des données du site "Center for Machine Learning and Intelligent Systems at the University of California": *http://cml.ics.uci.edu/* de l'Université de Californie, Etats Unis.
Pour chaque problème, vous trouvez le fichier de données à la rubrique "Data Folder" et la description des données à la rubrique "Data Set description".
En plus de répondre aux questions posées, les résultats obtenus doivent être expliqués et surtout interprétés.
Une documentation supplémentaire sur le sujet pratique étudié sera appréciée.

Problème 1. *(Dermatologie)*
A l'adresse internet:

http://archive.ics.uci.edu/ml/datasets/Dermatology

vous trouvez une courte description du problème concernant des maladies de la peau. Deux types d'attributs sont mesurés: cliniques et histopathologiques. Pour chaque patient ont été mesurées 35 variables (la variable 34 c'est l'âge et la variable 35 représente le type de maladie de la peau). La variable âge ne sera pas considérée dans l'étude.
Etudiez si il y a un lien entre les 11 attributs cliniques et le type de maladie de la peau. Quelles sont les variables et les modalités corrélées? Interprétez les résultats.

Problème 2. *(Espèces Anuran)*
Les données et leur description se trouve à l'adresse:

http://archive.ics.uci.edu/ml/datasets/Anuran+Calls+(MFCCs)

1. Peut-on réduire le nombre de variables numériques?
2. Etudiez si il y a des relations entre les variables numériques. Expliquez ces relations.

3. En utilisant les réponses trouvées plus haut, caractérisez les variables qualitatives.

Problème 3. *(Graines)*
Les données et leur description se trouve à l'adresse:

http://archive.ics.uci.edu/ml/datasets/seeds

On veut décrire et prévoir la variable numéro 8 (qualitative) fonction des autres variables.
1. Vous considérez 80% des données comme base de données d'apprentissage, les 20% restantes serviront comme base de données de test.
2. Quelles sont les variables qui permettent de prévoir la variable numéro 8?
3. Trouvez la meilleure méthode de prévision pour la variable numéro 8. Explications et justifications.
4. Vérifiez la méthode choisie à la question précédente sur les observations de test. Conclusion.