ENSIIE

Projet d'Analyse de Données
2023-2024

Document de travail

1 Présentation

Le projet MAD se déroule en binôme (au sein du même groupe de TD) et compte pour la note du module MAD (% à définir).

Le projet d'analyse de données est l'occasion de mettre en oeuvre les méthodes d'analyse exploratoire et de modélisation sur des données réelles choisies par les étudiants. Pour réaliser ces analyses, il est laissé à l'étudiant la possibilité d'utiliser les logiciels libres tels que R ou MATLAB. Les étudiants qui le souhaitent, pourront travailler en utilisant les langages de programmation tels que Python ou C (C++).

Ce projet constitue aussi l'occasion de développer l'apprentissage de ces logiciels et de leur langage spécifique. Le travail attendu consiste en un rapport d'étude correctement rédigé qui présente les résultats des analyses effectuées. Aucune soutenance n'est prévue.
Le rendu d'un rapport de projet est indispensable pour la bonne compréhension des concepts de l'analyse des données et des statistiques ; l'absence de rendu d'un rapport sera synonyme d'un zéro et sera considérée comme un point négatif en jury d'UE de maths 2A. MAD constitue un pré-requis pour continuer dans diverses options de 3A.

2 Déroulement du projet

2.1 Choix et préparation des données

Le choix de la base de données à analyser est laissé aux étudiants qui ont un libre choix avec cependant les contraintes suivantes :

- La base de données existe déjà sous forme numérique
- Il y a au moins 50 individus et au moins 10 variables

Afin de faciliter la sélection d'une base, il est possible de se rendre sur les sites suivants :

- http ://archive.ics.uci.edu/ml/index.html
- http ://lib.stat.cmu.edu/datasets/
- http ://www.biostat.umn.edu/ lynn/correlated.html
- http ://www.info.univ-angers.fr/pub/gh/Datasets/pbio.htm
- https ://www.kaggle.com/sakshigoyal7/credit-card-customers
- https ://www.kaggle.com/iabhishekofficial/mobile-price-classification
- https ://www.kaggle.com/mcdonalds/nutrition-facts .

Il s'agit de s'assurer de la mise en forme des données sous la forme d'un tableau $n \times p$, avec les individus en ligne et les variables en colonne. Il faut détecter la présence de données manquantes, de valeurs éventuellement aberrantes, s'assurer du type des différentes variables (quantitatives, qualitatives). Au vu de ces données et/ou selon les objectifs pour lesquels la base de données a été constituée, il conviendra

de définir une problématique, ou un ensemble de questions, pour lesquelles des méthodes d'analyse seront proposées.
L'ensemble de cette pré-analyse sera communiquée, sous la forme d'un mini-rapport très concis qui présentera donc les données (leur source notamment), et les objectifs. Celui-ci servira à l'encadrant à évaluer la faisabilité du projet. Le nom du document contiendra les noms des deux membres du binôme. Le mini-rapport pourra apporter 2 points supplémentaires à la note finale.

2.2 Le rapport

Le rapport contiendra une présentation détaillée des données, et il est attendu d'y trouver plusieurs analyses statistiques :

- des analyses descriptives univariées et bivariées
- des analyses factorielles
- de la modélisation / méthodes supervisées (régression)
- de la classification (si le sujet s'y prête, mais c'est souvent le cas)

La date limite de remise des projets sera fixée en fonction de la date de remise des notes par l'enseignant en concertation avec les étudiants et la scolarité. Le format le plus communément adopté pour la remise du rapport est le format pdf. Le rapport pourra alors être envoyé par mail à l'enseignant (qui accusera la réception).

L'évaluation du projet portera sur les points suivants :

- Présentation : qualité de la rédaction, et de la présentation des résultats (utilisation pertinente des tableaux et des graphiques en nombre contrôlé, équilibre avec des annexes).
- Problématique et modélisation : Définition d'une bonne problématique adaptée aux données, bon choix (motivé) des méthodes et modèles proposées.
- Méthodes : évaluation de la bonne compréhension des méthodes utilisées basée sur une bonne description des procédures utilisées et la qualité (justesse) des commentaires des sorties statistiques obtenues par les logiciels.

L'utilisation pertinente et juste de méthodes non-vues directement en cours (par exemple telles que le modèle linéaire généralisée) sera valorisée et donnera lieu à des bonus.