Instructions pour le mini-projet statistique.
02/10/2026

Objectifs

Il s'agit de trouver des données sur internet (Kaggle, site de l'Insee, data.gov, data.gouv.fr, etc.) et de les étudier. Ce n'est pas utile de prendre une grosse base de données (mais pas interdit non plus). Ensuite il faut choisir une ou des grandeurs d'intérêt, construire un ou deux estimateurs de ces grandeurs, puis tester leurs propriétés sur les données, en effectuant quelques calculs théoriques (si c'est possible) et en visualisant les données et leurs propriétés. Vous pouvez par exemple visualiser la loi forte des grands nombres et/ou le théorème de la limite centrale.

Vous êtes libres d'utiliser les techniques que vous voulez, vues en cours ou non (régressions linéaire ou logistique, méthodes des plus proches voisins, arbres de décision, modèle probabiliste dédié, etc.). Vous pouvez tester plusieurs méthodes sur le même jeu de données.

Dans tous les cas, il faut commencer par une étude de statistique descriptive qui va vous aider à comprendre les déterminants du problème que vous vous posez. Vous devez donc visualiser les données, calculer des moyennes, médianes, variances, etc. Vous devez aussi regarder si la base contient des données manquantes ou pas et le cas échéant, indiquer comment vous les avez traitées.

Exemples de sujets et de bases de données des années précédentes :

- Analyse de la demande dans des statistiques de location de vélo (Kaggle : https://www.kaggle.com/competitions/bike-sharing-demand). On peut essayer d'estimer la demande par une loi de Poisson, une loi binomiale négative ou bien encore une loi de Neyman type A.

- Prédiction du diabète (ou un autre maladie) dans une population donnée. Estimation de la distribution du glucose et/ou de l'IMC par une loi normale ou log-normale (UCI ML Repository : https://archive.ics.uci.edu/ml/datasets/pima+indians+diabetes).

- Distribution des revenus dans un pays et estimation par la loi de Pareto (ou bien une autre loi). Estimation du coefficient de Gini. Plusieurs bases de données sont disponibles :
  .. Luxembourg Income Study (LIS) : https://www.lisdatacenter.org/
  .. World Income Inequality Database (WIID) : https://www.wider.unu.edu/project/wiid-world-income-inequality-database
  .. Eurostat : https://ec.europa.eu/eurostat/web/microdata/european-union-statistics-on-income-and-living-conditions
  .. Adult Income Dataset (Kaggle) : https://www.kaggle.com/datasets/uciml/adult-census-income
  .. Et bien sûr l'Insee ! https://www.insee.fr/fr/statistiques/2414231 ou https://www.insee.fr/fr/statistiques/3560118

- Étude d'une campagne marketing : https://www.kaggle.com/datasets/rodsaldanha/arketing-campaign/data?select=marketing_campaign.csv et de la réponse ou non des clients à cette campagne.

- Prédiction du taux de défaillance d'un client pour l'obtention d'un prêt bancaire.

- Analyse de la distribution des revenus dans les quartiers et politique de la ville. Inégalités de revenus à partir des données Insee

- Loi suivie par l'espérance de vie des français. Comment évolue l'espérance de vie, comment la modéliser ? À partir des données statistiques de l'Insee ou de l'Ined.

- Vous pouvez enfin partir du sujet de TP sur la VVC (CLV), essayer de répondre à une partie des questions théoriques (ce n'est pas obligatoire et les dernières questions sont sans doute trop difficiles) puis essayer de modéliser la VVC et/ou la RFM (Récence–Fréquence–Monétaire) pour la base de données proposée ou bien une autre base de marketing :
  .. CDNow : https://www.brucehardie.com/datasets/
  .. Online Retail Dataset : https://archive.ics.uci.edu/dataset/352/online+retail

.. La librairie « LifeTimes » de Python (https://lifetimes.readthedocs.io/en/latest/Quickstart.html) contient plusieurs jeux de données pour calculer la CLV. Par contre, elle n'est plus maintenue et elle ne fonctionne plus sous Pythonn 3. Si vous réussissez à en faire quelque chose, ça m'intéresse fortement...
.. Kaggle "Customer Personality Analysis" : https://www.kaggle.com/datasets/imakash3011/customer-personality-analysis

Modalités de restitution

Le mini-projet est à m'envoyer par mail avant début décembre 2026, sous la forme d'un compte-rendu au format .pdf, bien rédigé, dans lequel vous expliquerez votre problème, vos choix et vous discuterez des résultats obtenus. Cette partie rédaction est très importante ! Un fichier contenant le code (Python ou notebook) accompagnera le compte-rendu afin que je puisse tester les programmes. Si vous me rendez un notebook, le compte-rendu peut y être intégré en utilisant des cellules de texte pour expliquer la démarche et les commentaires.

Si vous utilisez ChatGPT ou d'autres IA pour vous aider (et vous avez parfaitement le droit de le faire) garder à l'esprit le fait que les enseignants passent les rendus de projet dans des applications de détection de plagiat et que les sorties de ChatGPT (ou autre) en sont truffées.

Vous pouvez travailler par binôme ou trinôme.

Il s'agit d'un « mini » projet. **N'y passer pas trop de temps et restez modestes dans vos ambitions**.

N'hésitez à me contacter si vous avez des questions.