DATA – Année 2025-2026 – Mini-projets

Idée générale : Une séance de TD est consacrée au lancement de ce mini-projet, aux alentours de début décembre. Lors du mini-projet de DATA, chaque binôme doit, en autonomie, trouver un jeu de données sur le sujet de son choix (par exemple sur internet), et le traiter en résolvant une problématique simple, et en mettant en œuvre une ou des méthodes apprises lors des enseignements de l'EC.

Compétences visées : Les mini-projets du cours de DATA de 3GM cherchent à remplir plusieurs objectifs pédagogiques :

- ➔ Construire un programme en autonomie à partir d'une page blanche.
- ➔ Comprendre un jeu de données et en extraire une problématique intéressante.
- ➔ Mettre en œuvre les compétences de traitement de données acquises en cours.

Travail préliminaire : En amont de cette séance, les élèves doivent **impérativement** :

- ➔ Avoir constitué des binômes (si nombre impair dans le groupe, possibilité d'un trinôme ou d'un monôme).
- ➔ Avoir réfléchi à un sujet d'étude.
- ➔ Avoir téléchargé/récupéré un jeu de données de leur choix.
- ➔ Avoir construit une ébauche de problématique à traiter.

Séance de mini-projet de début décembre : L'objet de cette séance est de mener des discussions ouvertes entre les binômes et l'enseignant(e), afin de déterminer :

- ➔ Si le jeu de données choisi est approprié (pas trop simple ni trop complexe, pas trop gros ni trop petit, etc.).
- ➔ Si la problématique choisie est bien dosée et est traitable par les élèves.
- ➔ Quels exemples de stratégie pourraient être appliqués pour traiter cette problématique.

Rendu : le mini-projet fera l'objet d'un rendu, par email à l'enseignant(e), sous la forme d'une archive compressée contenant (i) le jeu de données traité et (ii) un notebook fonctionnel et commenté, respectant le cahier des charges détaillé dans le présent document. Le rendu devra être effectué avant la tenue de la deuxième Interrogation Ecrite, à la fin du mois de janvier.

Evaluation : les critères de notation couvriront les aspects suivants :

- ➔ Pertinence du jeu de données et de la problématique traitée.
- ➔ Respect du cahier des charges et des consignes.
- ➔ Choix et application de la stratégie de traitement et d'analyse.
- ➔ Justesse du code.
- ➔ Qualité formelle du rendu (clarté de la rédaction, commentaires et lisibilité du code, qualités des graphiques, etc.).

Annexe 1 – Cahier des charges

Jeu de données : Il doit s'agir d'un document texte contenant des données quantitatives, idéalement de dimensions 1D ou 2D. Le traitement d'un fichier audio ou d'une image est également possible, mais l'analyse d'une vidéo est déconseillée car trop complexe. Le fichier devrait être suffisamment gros pour permettre une analyse intéressante (au moins quelques centaines de lignes, soit environ **1 ko**), mais suffisamment petit pour être traité facilement (pas plus d'un million de lignes, soit environ **1 Mo**). Le sujet peut être scientifique/technique, mais ce n'est pas une obligation. Les données doivent provenir d'une archive ouverte et être libres de droit. Il peut également s'agir de résultats de simulations issues d'un modèle construit dans une autre EC (auquel cas ce ne sera pas ce modèle qui sera évalué, mais bien le traitement des données qui en seront extraites). Certains jeux de données ont des formats complexes à appréhender (propriétaire, json, etc.) : ils sont déconseillés. **Le premier critère de choix de votre jeu de données doit être : est-il facile à charger sous NumPy en utilisant np.genfromtxt() ?**

Outils de traitement : Le travail doit être effectué dans un Jupyter Notebook, en utilisant uniquement les bibliothèques de base utilisées en cours/TD : NumPy, SciPy, et Matplotlib. Les bibliothèques non-utilisées dans le semestre (scikit_learn, pandas, etc.) sont proscrites. L'utilisation d'une IA génératrice pour l'écriture de code est par ailleurs strictement interdite. Le notebook ne doit pas être trop long. Un bon critère est de l'exporter en PDF, et de vérifier qu'il ne fait pas plus de 4-5 pages A4 (tout compris).

Temps de travail : En plus des deux heures de TD consacrées au lancement du mini-projet, il doit rester modeste. Une durée maximale de 10h de travail personnel semble raisonnable. **Un projet trop long et trop fouillé ne sera pas encouragé par une notation supérieure**.

Techniques mises en œuvre : Le traitement de vos données doit nécessairement s'appuyer sur au moins une méthode vue en cours, en TD, ou en TP, mais ne doit pas utiliser de technique trop complexe non-abordée pendant le semestre. Les méthodes à utiliser sont à choisir dans la liste suivante :

- ➔ Statistiques descriptives (CM2)
- ➔ Estimateurs, lois de probabilité (CM2)
- ➔ Tirages de Monte-Carlo (CM2)
- ➔ Interpolation (CM3)
- ➔ Régression (CM3)
- ➔ Dérivation/intégration (CM3)
- ➔ Filtrage/lissage (CM3)
- ➔ Détection d'extremums (CM3)
- ➔ Transformation de Fourier Discrète (CM4)
- ➔ Analyse spectrale (CM4)
- ➔ Classification supervisée par régression logistique (TP1)
- ➔ Apprentissage non-supervisé par SVD (TP2)

Compte rendu : Tout votre travail doit être détaillé dans un Jupyter Notebook, qui doit contenir **obligatoirement** :

- ➔ Une description du contexte scientifique/technique (ou autre) dans lequel se situe votre travail, et une description de l'origine de vos données (source bibliographique, lien de téléchargement, et méthode d'acquisition : mesures, simulations, enquêtes, relevés, etc.).
- ➔ Une description du jeu de données : type, taille, structure, format, etc.
- ➔ Une description de la problématique que vous souhaitez traiter à partir de ces données, et de la stratégie que vous allez suivre. Dans cette description, vous devez lister quelles techniques, dans la liste ci-dessus, vous avez choisi d'utiliser.
- ➔ Un code en quelques cellules (maximum 4-5 cellules, avec pas plus d'une cinquantaine de lignes de code par cellule), incluant le chargement des données et effectuant la totalité des traitements souhaités, y compris des sorties graphiques. Un graphique est obligatoire, avec une mise en forme correcte. Ce code doit être exécutable par la personne qui corrige le travail, et répondre à la problématique choisie. Il doit être correctement commenté afin d'être lisible et compréhensible.
- ➔ Une conclusion sur la problématique choisie, à la lumière des traitements effectués sur les données.

Format de rendu : Votre travail doit être rendu par email à votre enseignant(e) de TD, dans une archive compressée intitulée « **Nom_Prenom_Groupe.extention** ». Cette archive doit contenir votre jeu de données, et votre notebook intitulé « **Nom_Prenom_Groupe.ipynb** ». Le rendu doit avoir lieu avant la deuxième Interrogation Ecrite. Ces règles formelles, ainsi que l'ensemble de ce cahier des charges, rentrent en compte dans la notation.