Projets Data Mining et Data visualisation

Pr. Najima Daoudi

Les deux projets de data visualisation et Data mining sont intrinsèquement liés. L'objectif du premier est de réaliser une étude descriptive et exploratoire d'un dataset de votre choix qui s'inscrit dans les thématiques de l'année. L'objectif du deuxième projet est de faire une prédiction dans un domaine particulier selon la thématique choisie et ce en s'appuyant sur quatre algorithmes différents (ensemble learning, classification simple ou régression) tout en faisant appel aux compétences acquises durant le cours en termes de features engineering, sélection des features, validation croisée si nécessaires, features importances, évaluation des modèles.

Vous pouvez également intégrer des techniques de NLP ou deep learning. (Les projets ayant intégrés ces techniques auront des bonus).

La thématique est choisie par le professeur selon les thèmes émergents et les préoccupations scientifiques actuelles.

Pour le projet data visualisation :

L'étudiant est appelé dans un premier temps à faire le choix du dataset avec justification de ce choix. Dans un deuxième temps, il doit préciser les objectifs de son étude et choisir les outils techniques nécessaires pour les mettre en place. Enfin, il doit présenter les différentes visualisations en ajoutant son analyse et ses commentaires.

Pour le projet data mining :

L'étudiant est appelé à justifier le choix du dataset par rapport à la problématique de prédiction dans un premier temps. Le jeu de données peut être également créé par les techniques scrapping (les étudiants qui vont créer leur propre jeu de données auront un bonus).

Ensuite, il doit faire le preprocessing du dataset (ceci peut faire partie du projet data vis). Cela étant, il faut justifier le choix des algorithmes et les exécuter après avoir fait appel aux techniques features selection et features engineering. Enfin, il faut procéder à l'évaluation et la validation des modèles et en choisir le meilleur.

Les étudiants qui proposent un environnement de déploiement du meilleur modèle auront un bonus.

1. Organisation, déroulement et livrables :

Le travail sera réalisé par groupe de six étudiants. Chaque étudiant doit exposer une partie qu'il a réalisé lui-même. Le jour de la présentation, le groupe doit livrer un rapport du travail réalisé et une présentation ppt.

2. **Les outils :**

Les outils de visualisation consistent en les technologies utilisées pour réalisées l'étude exploratrice de votre dataset selon vos objectifs prioritaires. Pour cette année, on vous donne le choix entre :

- **Power BI** : https://powerbi.microsoft.com/fr-be/desktop/
- **D3.js :** https://d3js.org/
- **Neo4j**

Pour le projet data mining, il doit être réalisé en python.

3. **Les thématiques de l'année :**

- **Santé**
- **E-learning**
- **Sécurité routière**
- **Environnement**

Remarques importantes :

1. **Trois groupes au maximum choisissent la même thématique.**
2. **Chaque groupe doit travailler sur un jeu de données différent.**
3. **Chaque étudiant aura une note différente à la base de la partie présentée et les réponses aux questions.**