Modalités du projet

Travail en binôme (ou individuel) avant la date et l'heure indiquées sur Moodle.

Descriptif

Le projet consiste à élaborer un projet de Machine Learning sur le sujet d'apprentissage **supervisé** de votre choix à partir de données ouvertes (par exemple provenant de https://www.data.gouv.fr), tout en explorant et en analysant l'intégration d'outils d'IA générative dans votre travail.

Contraintes pour le jeu de données :

- Le jeu de données ne doit pas être communément utilisé en Machine Learning (notamment, évitez les datasets disponibles sur Kaggle).
- Vous devez vous assurer que le dataset choisi est peu ou pas étudié dans un contexte de Machine Learning (par exemple, en effectuant une recherche avec des mots-clés spécifiques).
- Si votre dataset a déjà été utilisé, mentionnez les travaux existants et décrivez comment vous différenciez votre approche.

Exemple d'un jeu de données data.gouv.fr déjà utilisé qui peut vous inspirer dans votre analyse (mais ne reprenez pas les mêmes données) :
Prédire si un adhérent d'une médiathèque sera emprunteur ou non de document
https://www.data.gouv.fr/fr/datasets/caracteristiques-des-adherents-de-la-mediatheque-la-grand-plage-a-roubaix-en-2017/
https://github.com/olivierviollet/Studies/blob/master/Roubaix_emprunteur.ipynb

Le projet sera noté sur la base d'un fichier Notebook (ipynb) et de la qualité générale de ce support. Une exportation du notebook en fichier HTML est aussi demandée. Ces éléments sont détaillés plus loin. Le barème est donné à titre indicatif et est susceptible de changer

IA générative

Le projet a également pour but de vous faire réfléchir à ce que peuvent apporter des outils d'IA générative (ChatGPT, GPT4All, Bard, Claude, etc.) lors d'un projet de Machine Learning.

Travail demandé

1. Introduction / Sélection des données (2 points)
   - Identifiez au moins un jeu de données pour le projet. Décrivez vos motivations, le processus de recherche.
   - Utilisez l'IA pour explorer des sources de données ouvertes, générer des idées de problématiques ou identifier des angles d'analyse inédits.
   - Analyse critique : L'outil a-t-il proposé des idées pertinentes ou redondantes ?
   - Faites le choix définitif du jeu de données. Expliquez ce choix.

2. Exploration et traitement des données (3 points)
    — Identifiez et décrivez les caractéristiques des données. Effectuez les prétraitements nécessaires à ce stade.
    — Demandez à l'IA de critiquer constructivement votre approche.
    — Proposez une analyse du résultat et, le cas échéant, modifiez votre travail en conséquence.
3. Modélisation et évaluation (9 points)
    — Choisissez et construisez des modèles ou pipelines adaptés au problème.
    — Tester les modèles avec une méthodologie rigoureuse.
    — Évaluer les performances avec des métriques appropriées.
    — Demandez à l'IA de critiquer constructivement vos choix et votre approche.
    — Proposez une analyse du résultat et, le cas échéant, modifiez votre travail en conséquence.
4. Communication des résultats (3 points)
    — Produisez une analyse claire des résultats pour un public non technique.
    — Demandez à l'IA une analyse similaire.
    — Proposez une analyse du résultat et, le cas échéant, modifiez votre travail en conséquence.
5. Retour d'expérience (3 points)
    — L'IA vous a-t-elle été utile ? Quels défis avez-vous rencontrés ? Comment améliorer son intégration dans ce type de projet ? Voyez-vous des considérations éthiques ?

Livrables

Il y a deux livrables. Les dates limites de dépôt sont indiquées sur Moodle.

1. Livrable préliminaire
    — Archive contenant vos données et un notebook présentant les premières manipulations de ces données. A défaut de fichier de données, le notebook peut également contenir un lien vers une version pérenne du jeu de données.
    — Possibilité de changer de données pour le livrable final, à condition de justifier ce choix.
    — Si le livrable préliminaire n'est pas rendu une pénalité de 10% sera décomptée de votre note de projet.
2. Livrable final
    — Archive contenant tous vos fichiers (données incluses ou lien pérenne)
    — Le notebook final doit inclure toutes les analyses, les modèles et votre réflexion sur l'utilisation de l'IA générative.
    — Une version HTML exportée du notebook est requise.

Le travail, et notamment le fichier Notebook, doit être soigné, structuré et clair. Si des références sont utilisées il faut les citer, de même pour du code provenant de sources diverses. Les éléments provenant de ChatGPT (ou autres) doivent être clairement identifiés.

Objectifs pédagogiques

- Développer une méthodologie rigoureuse pour résoudre des problèmes de Machine Learning.
- Exploiter des outils d'IA générative tout en analysant leurs apports et limites.
- Renforcer votre esprit critique face à des outils technologiques en constante évolution.

Version HTML

Afin de pallier à un éventuel souci d'exécution du notebook, il est demandé d'inclure une exportation HTML du notebook (File > Download as *ou* File > Export Notebook As).

C'est essentiellement la méthode qui sera notée et non les résultats spécifiques. Par exemple, il n'est donc pas nécessaire de chercher absolument à avoir une précision de 99% dans vos prédictions. Par contre, si vous choisissez un modèle (pour diverses raisons que vous énoncerez), que vous vous rendez compte que sa précision est de 50% et que vous réussissez à l'améliorer (en expliquant pourquoi et comment) et/ou que vous décidez d'explorer d'autres modèles (pourquoi/comment), cela sera apprécié.

Note : Le cours a principalement évoqué l'évaluation de la classification binaire (deux classes). Il conviendra d'adapter l'évaluation en fonction des besoins.