TP 2025 — Advanced Machine Learning

Master MSI ID

Vous développerez un modèle de prédiction pour les résultats d'une compétition sportive qui aura lieu en 2025. Vous êtes libre de choisir la compétition sportive de votre choix. Pour ce faire, vous devrez identifier des jeux de données des résultats passés sur Internet (par exemple, en allant sur Kaggle ou HuggingFace). Pour vous aider à déterminer le type de données dont vous avez besoin, vous pouvez accéder à un exemple de jeu de données initial pour prédire les médailles aux Jeux olympiques de 2024 (voir le dossier TP sur moodle).

Tâches à effectuer :

1. Documentation du problème :
   - A quel problème souhaite-t-on répondre ? Quel modèle semble adapté pour résoudre ce problème ?
   - Trouver et lire des articles ou sites web traitant de ce problème. Citer ces articles et en tirer des conclusions sur votre analyse (par exemple : données supplémentaires à utiliser, feature à construire, métriques d'évaluation pertinentes, …)

2. Ressources disponibles :
   - Dressez la liste des ressources auxquelles vous avez accès (humaines, matérielles, temporelles).

3. Exploration des données :
   - Trouver un ou plusieurs jeu de données en lien avec la compétition sportive choisie
   - Analysez les différents champs disponibles dans ce jeu de données en produisant un rapport semblable au modèle dans le cours.
   - Utilisez des méthodes de visualisation pour comprendre les distributions des scores, le classement, les tendances historiques, etc.
   - Calculez des métriques supplémentaires pour votre jeu de données : des taux de réussite par équipe, par joueur, etc…

4. Recherche de données complémentaires :
   - Explorez des sources de données externes qui pourraient compléter les informations disponibles dans le jeu de données principal (indicateurs économiques, démographiques, informations conceptualisant la compétition, …) **Cette tâche doit être effectuée à l'aide d'une API, de *web scrapping* ou de la construction documentée du jeu de données à partir de différentes sources d'information.**

5. Prétraitement des données :
   - Identifiez et traitez les données manquantes ainsi que les valeurs aberrantes dans le jeu de données. (N'oubliez pas d'indiquer dans votre rapport les traitements effectués et leur justification).
   - Effectuez les transformations nécessaires sur les caractéristiques pour les préparer à être utilisées par les algorithmes de *machine learning*.

6. *Feature Engineering* :
   - Sélectionnez les *features* les plus pertinentes à partir des différentes sources de données pour améliorer la prédiction du modèle.
   - Créez de nouvelles features si nécessaire pour enrichir la représentation des données.

7. Modélisation :
   - Choisissez au moins un algorithme de machine learning supervisé pour accomplir vos prédictions.
   - Divisez les données en ensembles d'entraînement et de test. Si vous avez besoin de
   - Entraînez vos modèles sur l'ensemble d'entraînement et évaluez leurs performances sur l'ensemble de test.

8. Évaluation et optimisation :
   - Utilisez des mesures d'évaluation appropriées pour évaluer les performances de vos modèles.
   - Optimisez les hyperparamètres des modèles pour améliorer leur précision et leur généralisation.

9. Interprétation des résultats :
   - Analysez et interprétez les résultats obtenus. Un modèle peut avoir un taux de prédiction faible, en fonction des données utilisées (cf. le TP sur la SNCF dans le cours d'introduction au ML). Ainsi, le taux de prédiction n'aura de sens que s'il est lié aux données et au contexte de l'étude.
   - Identifier les possibles vainqueurs de la compétition

Conclusion : Ces travaux pratiques vous offrent une opportunité de mettre en pratique vos connaissances en machine learning sur un problème réel. Veillez à suivre chaque étape avec rigueur et à documenter vos choix et vos résultats pour une analyse approfondie à la fin du projet.

ATTENTION : Ce TP et ses résultats ne sont pas à utiliser dans le cadre de pari sportif.

Rendu

Rendre un fichier zippé, au 4 noms de famille de votre groupe (ex. alaoui-martin-smith-wang.zip) contenant :

- Vos prédictions (fichier csv et/ou affichage du tableau de résultat dans un notebook) ;
- Votre/vos notebooks Jupyter contenant votre/vos modèles ayant conduit à ces prédictions ;
- Un fichier d'analyse et compte rendu contenant votre cheminement au fil des étapes CRISP ;
- Une explication de la répartition des tâches et du travail individuel accompli par les 4 membres du groupe.

N.B. : tous vos fichiers doivent contenir en entête vos 4 noms, la date et l'intitulé de votre master.

Modalités

Le TP est à réaliser par groupe de **4 étudiants**. Il est à rendre, au plus tard, le vendredi 4 avril 2025, 23h59.