TP 2024 – Méthodes pour l'analytique

Master MSI - ID

Dans le cadre de ce TP, vous allez développer un modèle de **prédiction du nombre total de médailles que chaque pays (inscrit) pourrait remporter aux Jeux Olympiques d'été 2024 à Paris**. Vous utiliserez le jeu de données "120 years of Olympic history: athletes and results" comme point de départ pour cette tâche.

Afin de vous guider dans vos analyses, la section **Tâches à effectuer** indique les grandes étapes (mais pas toutes les étapes) à ne pas manquer lors de votre analyse. Vous disposez également de données dans le dossier Moodle de ce TP :

- athlete_events.csv et noc_regions.csv : issus de **120 years of Olympic history: athletes and results**, posté par rgriffin (https://www.kaggle.com/datasets/heesoo37/120-years-of-olympic-history-athletes-and-results)
- dictionary.csv : issu de **Olympic Sports and Medals, 1896-2014**, posté par The Guardian (https://www.kaggle.com/datasets/the-guardian/olympic-games)

Vous êtes libres d'utiliser ces données (ou non) et de les compléter par d'autres données trouvées lors de vos recherches.

Tâches à effectuer (ordre suggéré) :

1. Documentation du problème :
   - À 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 pour votre analyse (par exemple : données supplémentaires qu'il faudrait mobiliser lors de votre analyse, *feature* à construire, métriques d'évaluation pertinentes …).
2. Exploration des données :
   - Téléchargez sur Moodle les jeux de données proposés pour répondre à ce problème prédictif, dont le jeu de données "120 years of Olympic history: athletes and results".
   - Analysez les différents champs disponibles dans ce jeu de données, tels que les athlètes, les disciplines, les pays participants et les médailles remportées.
   - Utilisez des méthodes de visualisation pour comprendre les distributions des médailles par pays, les tendances historiques, etc.
   - Calculez des totaux de médailles par années, par pays, par genre des athlètes, etc.
   - Comparez vos résultats à des données trouvées dans le web, telles que https://fr.wikipedia.org/wiki/Tableau_des_m%C3%A9dailles_olympiques_par_nation

- Tirez des conclusions concernant le calcul des médailles par pays (prise en compte des sports collectifs, des évolutions des dénominations des pays).

3. 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 contextualisant les éditions des JO et les pays hôtes).
   - Trouvez des données concernant les JO 2024, qui vous permettront de générer des prédictions de médailles par pays, aux JO 2024 (le recours au Web scraping est possible).
4. Prétraitement des données :
   - Identifiez et traitez les données manquantes ainsi que les valeurs aberrantes dans le jeu de données.
   - Effectuez les transformations nécessaires sur les caractéristiques pour les préparer à être utilisées par les algorithmes de machine learning.
5. *Feature Engineering :*
   - Sélectionnez les *features* les plus pertinents à partir des différentes sources de données pour améliorer la prédiction du modèle.
   - Créez de nouveaux *features* si nécessaire pour enrichir la représentation des données.
6. Modélisation :
   - Choisissez au moins un algorithme de machine learning supervisé pour accomplir vos prédictions.
   - Divisez les données en ensembles d'entraînement et de test.
   - Entraînez vos modèles sur l'ensemble d'entraînement et évaluez leurs performances sur l'ensemble de test.
7. 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.
8. Interprétation des résultats :
   - Analysez et interprétez les résultats obtenus.
   - Identifiez les pays les plus susceptibles de remporter un grand nombre de médailles aux Jeux Olympiques 2024 à Paris en fonction des prédictions de votre modèle.

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. **Bonne exploration et rendez-vous le 11 août 2024 pour connaitre le groupe gagnant de la classe !**

Rendu

Rendre un fichier zippé, au 4 noms de famille de votre groupe (ex. dupont-durant-dupuy-martin.zip) contenant :

- Vos prédictions (fichier csv et/ou affichage du tableau des médailles 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 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 mardi 9 avril 2024, 12h.