Instructions pour le projet informatique de Machine Learning.
12/12/2025

Objectifs

• Trouver un ou deux jeux de données sur internet et s'attacher à prédire les valeurs de deux variables d'intérêt. L'une d'elles doit être catégorielle et donner lieu à un problème de classification, l'autre doit être réelle et donner lieu à un problème de régression. Choisissez des variables pertinentes et expliquez clairement votre démarche et vos objectifs : que voulez-vous prédire ? Pourquoi ? Quels choix avez-vous faits ?

• L'un des deux jeux de données doit être de taille significative (au moins quelques milliers d'individus). Vous pouvez effectuer le problème de classification et celui de régression dans le même jeu de données.

• L'objectif principal du projet est de tester plusieurs méthodes de classification et de régression (pas forcément toutes), de discuter de leur pertinence pour le problème traité et de comparer les différentes approches.

• Le choix du jeu de données peut se faire à partir des exemples proposés pour le mini-projet statistique, ou bien des jeux de données de votre choix.

• Les consignes du mini-projet statistique restent valables : commencer par des statistiques descriptives et des visualisations pertinentes, traiter les données manquantes, détecter et traiter les données aberrantes, étudier la nécessité d'une normalisation, l'encodage des variables catégorielles, etc.

• Étudier la qualité de l'entrainement et de la validation du modèle, justifiez vos choix de métriques d'erreurs. Proposez une analyse critique honnête sur les limites des modèles.

• Commentez le code Python et étudier, si possible, sa robustesse et sa reproductibilité.

• La grille d'évaluation prendra en compte les points suivants, à des poids à peu près équivalents : compréhension et analyse du problème, pré-traitement des données, analyse exploratoire, choix des modèles, entrainement et validation des modèles, interprétation (et interprétabilité), qualité du code, qualité des explications et de la communication des résultats.

Modalités de restitution

Le projet est à m'envoyer par mail au plus tard le 30 janvier 2026, sous la forme d'un compte-rendu au format .pdf, bien rédigé, dans lequel vous expliquerez votre problème, vos choix et vous discuterez des résultats obtenus. Cette partie rédaction est très importante. Un fichier contenant le code (Python ou notebook) accompagnera le compte-rendu afin que je puisse tester les programmes. Si vous me rendez un notebook, le compte-rendu peut y être intégré en utilisant des cellules de texte pour expliquer la démarche et les commentaires.

Si vous utilisez ChatGPT ou d'autres IA pour vous aider (et vous avez parfaitement le droit de le faire) garder à l'esprit le fait que les enseignants passent les rendus de projet dans des applications de détection de plagiat.

Pour ce projet, vous devez travailler en groupe (minimum 3 personnes, maximum 5). Vous expliquerez dans le compte-rendu la répartition du travail de chacun au sein du groupe.

N'hésitez à me contacter si vous avez des questions.