Test technique pour le recrutement de Data Scientist

Objectif : Modèle de classification des URL

Votre tâche consiste à créer un modèle de classification des URLs légitimes et de phishing. Le dataset fourni contient une liste d'URL déjà identifiées comme légitimes ou de phishing.

Instructions :

1. Utiliser une ou des méthodes issues de l'état de l'art de la détection de phishing pour le développement du modèle.
2. Implémenter une méthode de *stacking* avec au moins un modèle de *boosting*.
3. Il est possible d'utiliser les données textuelles ou des caractéristiques numériques créées à partir des URL.
4. Vous êtes autorisé à effectuer une augmentation de données si nécessaire.

Livrables :

1. Notebook contenant :
    - Les étapes de prétraitement des données (exemple : nettoyage, tokenisation).
    - Visualisations aidant à comprendre les données et le *feature engineering*.
    - Le processus de sélection du modèle, y compris la justification des méthodes choisies.
    - Implémentation de la méthode d'ensemble de *stacking* incluant un modèle de *boosting*.
    - Métriques d'évaluation des performances du modèle.
    - Explication des décisions prises dont le choix des données (données textuelles ou caractéristiques numériques créées à partir des URL)
2. Une proposition d'approche pour la maintenance et la mise à jour du modèle à l'avenir. Cela devrait inclure :
    - Les stratégies pour détecter le *drift* ou les changements dans la distribution des données.
    - Les méthodes pour incorporer de nouvelles informations dans le modèle.
    - Suggestions pour surveiller les performances du modèle au fil du temps.

Remarque : Votre capacité à communiquer efficacement votre approche à travers le notebook fourni et la proposition, ainsi que vos compétences techniques en Data Science, sont des éléments importants à mettre en avant.