Test technique pour le recrutement de Data Scientist

Objectif : Implémentation d'une approche de détection d'URLs de phishing.

Votre tâche consiste à créer une approche de classification des URLs légitimes (label 0) et de phishing (label 1). Le dataset fourni contient une liste d'URLs annotées en légitimes et phishing.

Instructions :

1. Se baser sur l'état de l'art pour identifier les features textuelles et/ou numériques pertinentes pour la problématique.
2. Utiliser pySpark pour le prétraitement des données (nettoyage, feature engineering, analyse de données tokenisation) cela peut être fait sur votre environnement personnel, sur Google Colab ou sur Databricks Community edition.
3. Développer un modèle en utilisant la méthode de stacking avec au moins un modèle de boosting.
4. Vous êtes autorisé à effectuer une augmentation de données si nécessaire.

Livrables :

Le notebook contenant :

- Les étapes de prétraitement des données réalisées avec Spark.
- Votre approche de feature engineering.
- Le processus de sélection du modèle, y compris la justification des méthodes choisies.
- Métriques d'évaluation des performances du modèle.
- Explication des décisions prises dont le choix des données ainsi qu'une analyse critique des résultats du modèle.