Projet: Analyse Big Data et Data Mining pour la Prédiction des Prix des Logements Airbnb

1 Contexte

Le secteur de la location de courte durée, tel qu'illustré par Airbnb, génère une quantité massive de données exploitables pour comprendre et prédire les tendances des prix. Ce projet vise à analyser les prix des logements Airbnb des grandes villes (Paris, Londres, Rome, Istanbul, Rio, Sydney, New York, Mexico), en appliquant des techniques de Big Data et des modèles de machine learning.
Vous devriez explorer les données fournies, sélectionner les variables les plus pertinentes, tester divers modèles prédictifs et optimiser leurs performances.

1.1 Objectifs

Les objectifs de ce projet sont:

- Construire et entraîner des modèles de machine learning (Random Forest, Gradient Boosting et Régression Linéaire) pour prédire le prix des logements.
- Identifier les variables clés influençant les prix, telles que : localisation, nombre de chambres, notes des utilisateurs, disponibilité, etc.
- Évaluer les performances des modèles en fonction de leur précision et de leur robustesse.
- Effectuer un tuning d'hyperparamètres sur chaque modèle pour améliorer leur performance.
- Identifier et gérer les problématiques d'overfitting et d'underfitting.

1.2 Travail attendu

- **Prétraitement des données :**
  - Nettoyage et normalisation des données si nécessaire.
  - Sélection des variables pertinentes grâce à des techniques statistiques ou algorithmiques (e.g., sélection par importance des variables).
- **Modélisation :**
  - Entraîner au moins quatre modèles : Random Forest, Gradient Boosting et Régression Linéaire.
  - Comparer les performances de chaque modèle sur des métriques comme RMSE, MAE ou F1-score.
  - Justifier le choix des modèles testés en fonction des données disponibles et des objectifs du projet.
- **Tuning des hyperparamètres :**
  - Décrire le processus détaillé d'optimisation des hyperparamètres choisi pour chaque modèle (exemples : profondeur d'arbre, nombre d'itérations, taux d'apprentissage).
  - Effectuer une recherche systématique (grid search).
- **Évaluation et prévention des biais :**

- Évaluer les performances des modèles sur un jeu de test.
- Détecter et corriger les problèmes d'overfitting et d'underfitting en ajustant les paramètres ou les données d'entraînement.

1.3 Format des rendus

- **Rapport écrit au format PDF :**
  - Introduction au projet et problématique étudiée.
  - Description des données et des variables utilisées.
  - Justification des modèles testés et analyse comparative de leurs performances.
  - Description détaillée des étapes de tuning des hyperparamètres.
  - Discussion des résultats obtenus et des défis rencontrés (e.g., gestion des déséquilibres dans les données).
- **Code Python (Script Python ou Notebook) :**
  - Code complet et bien structuré, permettant de reproduire les résultats.