Audit Intelligent

1. Introduction

L'objectif de ce projet est de détecter les anomalies dans les transactions comptables en vérifiant si les factures de **Book1** apparaissent également dans **Book2** et **Book3**.
Nous avons utilisé plusieurs algorithmes de **machine learning supervisé** pour classifier chaque transaction comme **présente (1)** ou **non présente (0)**.

Cette démarche permet d'assister les **auditeurs et analystes financiers** dans la détection d'erreurs, de fraudes ou de manquements dans les écritures.

2. Description des données

- **Book1 (Transactions principales)** : Factures identifiées par `NumFacture`.
- **Book2 (Grand Livre)** : Écritures identifiées par `GLDOC`.
- **Book3 (Paiements)** : Transactions de paiement identifiées par `RPDOC`.

Préparation des données :

1. Création de deux labels binaires dans **Book1** :
    a. `in_book2` = 1 si `NumFacture` est trouvé dans `GLDOC`.
    b. `in_book3` = 1 si `NumFacture` est trouvé dans `RPDOC`.
2. Suppression des colonnes non pertinentes (dates, références, taxes, etc.).
3. Conversion de toutes les colonnes en numériques et traitement des valeurs manquantes.

3. Méthodologie

Nous avons testé et comparé **trois modèles de classification** :

- **Gradient Boosting (GB)** : efficace pour des relations complexes.
- **Régression Logistique (LR)** : modèle simple et interprétable.
- **Random Forest (RF)** : ensemble d'arbres de décision, robuste et performant.

Les données ont été séparées en **80% apprentissage** et **20% test**.
Les modèles ont été évalués selon :

- **Accuracy (taux de bonne classification)**
- **Rapport de classification** (Précision, Rappel, F1-score)

4. Résultats

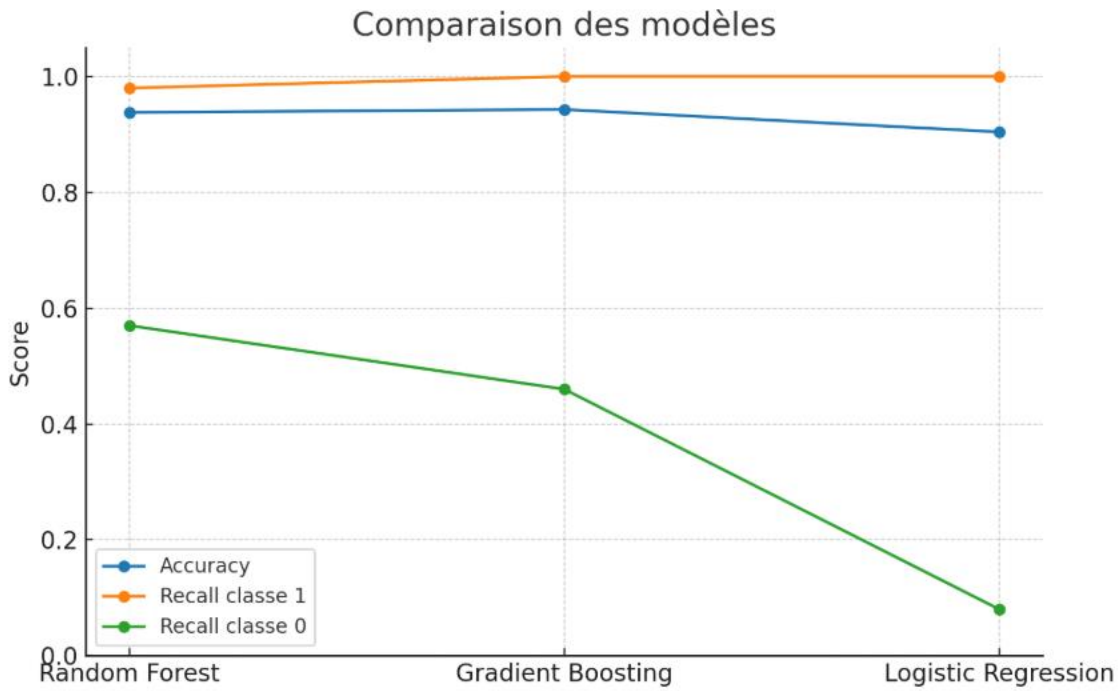
◇ Correspondance Book1 → Book2

- **Gradient Boosting** : (94,3 %)
- **Régression Logistique** : (90,4 %)
- **Random Forest** : (93,8 %)

◇ Correspondance Book1 → Book3

- **Gradient Boosting** : (94,3%)
- **Régression Logistique** : (90,4 %)
- **Random Forest** : (93,8%)

Modèle	Accuracy	Recall classe 1	Recall classe 0	F1 classe 1	F1 classe 0	Commentaire
Random Forest	0.938	0.98	0.57	0.97	0.66	Bon équilibre
Gradient Boosting	0.943	1.00	0.46	0.97	0.62	Très performant sur transactions présentes
Logistic Regression	0.904	1.00	0.08	0.95	0.15	Biaisé vers transactions présentes



5. Export des résultats

Nous avons sauvegardé les prédictions de chaque modèle dans des fichiers CSV pour une analyse transaction par transaction :

Chaque fichier contient :

- L'identifiant de transaction (`NumFacture`)
- Les prédictions du modèle
- Les valeurs réelles observées

6. Conclusion

Le projet a montré l'efficacité de différents modèles de machine learning pour vérifier la cohérence entre plusieurs livres comptables.

- Le **Random Forest** apporte une robustesse et une précision souvent supérieures grâce à son approche par ensembles d'arbres.
- Le **Gradient Boosting** reste performant pour capturer des relations complexes.

- La **Régression Logistique** constitue un modèle de référence simple.

Ce pipeline peut être intégré dans des systèmes d'**audit intelligent**, afin de renforcer la détection d'anomalies et d'automatiser la réconciliation des transactions.