📄 Rapport de Projet : Prédiction et Analyse du Rendement du Maïs en Afrique

1. Introduction et Objectifs du Projet

Ce projet vise à analyser et prédire le rendement agricole du maïs sur le continent africain. En s'appuyant sur un jeu de données de 203126 observations avant nettoyage puis **16 873 observations après.** L'étude combine une analyse exploratoire approfondie (EDA), une modélisation prédictive par apprentissage automatique (Machine Learning) et une solution de déploiement opérationnelle.

L'objectif principal est de fournir des outils et des insights permettant d'améliorer la sécurité alimentaire en identifiant les facteurs clés de productivité et en offrant des prédictions fiables pour les acteurs du secteur agricole.

---

2. Analyse Exploratoire des Données (EDA)

L'analyse initiale a permis de dresser un état des lieux de la production de maïs en Afrique.

2.1 Statistiques Descriptives

Le rendement moyen observé est de **1,234 t/ha**, avec une forte variabilité (écart-type de 0,834 t/ha). Cette disparité souligne des différences majeures entre les exploitations et les régions.

Variable	Moyenne	Médiane	Min	Max
Surface (ha)	7 797,22	4 000,00	0,50	49 931,00
Production (t)	9 462,28	4 217,85	1,00	49 979,97
Rendement (t/ha)	**1,234**	1,028	0,10	8,00

2.2 Principaux Enseignements

- **Corrélation Structurelle** : Une corrélation parfaite existe entre la production et la surface, ce qui est attendu par définition (Production = Rendement \times Surface).

- **Analyse en Composantes Principales (ACP)** : L'ACP a révélé que 77,7 % de la variance est expliquée par deux dimensions : la **taille de production** (surface/production) et la **productivité intrinsèque** (rendement).
- **Facteurs Influents** : Le facteur "Pays" est statistiquement significatif (ANOVA, $p < 0,05$), confirmant que le contexte géographique et les politiques nationales jouent un rôle prépondérant dans la performance agricole.

---

3. Modélisation Prédictive

Trois familles d'algorithmes ont été testées pour prédire le rendement (variable cible continue).

3.1 Algorithmes Sélectionnés

1. **Régression Ridge** : Modèle linéaire régularisé servant de référence (baseline).
2. **Random Forest** : Méthode d'ensemble (Bagging) excellente pour capturer les non-linéarités.
3. **Gradient Boosting** : Méthode séquentielle (Boosting) visant à minimiser les erreurs résiduelles.

3.2 Performance des Modèles

Les modèles non-linéaires surpassent nettement la régression linéaire, confirmant la complexité des interactions agricoles.

Modèle	MAE (t/ha)	RMSE (t/ha)	R^2 Score	Statut
Ridge Regression	0,4737	N/A	0,3440	Baseline
Random Forest	0,4261	**0,6344**	**0,4425**	**Meilleur Modèle**
Gradient Boosting	**0,4215**	0,6422	0,4287	Challenger

> **Note sur le R^2** : Un score de ~0,44 indique que 44 % de la variance est expliquée. Le reste (56 %) est probablement lié à des facteurs non inclus dans le dataset, tels que la pluviométrie précise, la qualité des sols ou l'accès aux intrants.

---

4. Déploiement et Automatisation

La solution a été industrialisée pour permettre une utilisation en temps réel et une maintenance continue.

4.1 Architecture Technique

L'infrastructure repose sur des technologies modernes garantissant scalabilité et portabilité :

- **API Web** : Développée avec **FastAPI** pour des performances optimales.
- **Conteneurisation** : Utilisation de **Docker** (multi-stage build) pour réduire la taille de l'image (~500 MB).
- **Hébergement** : Déploiement sur la plateforme Cloud **Render** avec intégration continue (CI/CD) via GitHub.

4.2 Pipeline de Réentraînement Automatique

Pour maintenir la précision du modèle face aux nouvelles données, un système automatisé a été mis en place :

1. **Détection de changement** : Utilisation d'un hash MD5 pour vérifier si les données source ont évolué.
2. **Compétition de modèles** : À chaque cycle, les trois modèles (Ridge, RF, Gradient Boosting) sont réentraînés.
3. **Sélection automatique** : Le modèle affichant le meilleur score R^2 est automatiquement déployé en production.

---

5. Conclusion et Recommandations

5.1 Synthèse

Le projet a permis de passer d'une donnée brute à une application fonctionnelle capable de prédire le rendement du maïs avec une erreur moyenne de **0,42 t/ha**. Le modèle **Random Forest** a été retenu pour sa robustesse et sa précision globale.

5.2 Recommandations

- **Enrichissement des données** : L'intégration de variables climatiques (précipitations, températures) et pédologiques (qualité du sol) est indispensable pour dépasser le plafond de performance actuel.

- **Interventions ciblées** : Les disparités régionales suggèrent que les politiques de soutien devraient être adaptées spécifiquement aux contextes nationaux identifiés comme moins performants.
- **Utilisation de l'API** : L'interface déployée permet aux décideurs d'effectuer des simulations rapides pour anticiper les récoltes selon les scénarios de plantation.