Analyse Complète d'un Problème de Data Science : Prédiction du Rendement de Maïs

Habti Elmehdi

26 mars 2025

1 Introduction

Dans ce rapport, je vais analyser comment prédire le rendement du maïs (en tonnes par hectare) en fonction de différents facteurs. L'objectif est de donner des idées pour que la ferme utilise ses ressources de manière plus efficace et obtienne le meilleur rendement possible.

2 Compréhension du problème

2.1 Description des variables

- **surface** ha : Surface cultivée en hectares (quantitative continue)
- **type_sol** : Type de sol (argileux, sableux, limoneux) (qualitative nominale)
- engrais_kg/ha : Quantité d'engrais utilisée en kg/ha (quantitative continue)
- **précipitations** _ **mm** : Précipitations moyennes mensuelles en mm (quantitative continue)
- **temperature**_C : Température moyenne mensuelle en °C (quantitative continue)
- **rendement_t/ha**: Rendement obtenu en tonnes par hectare (quantitative continue, variable cible)

2.2 Problème métier

La ferme souhaite prédire le rendement du maïs en fonction des facteurs de production pour optimiser l'allocation des ressources et maximiser le rendement.

2.3 Problématique centrale

Identifier les facteurs ayant le plus d'impact sur le rendement et déterminer les combinaisons optimales de ces facteurs pour maximiser la production.

3 Analyse statistique descriptive

3.1 Mesures de tendance centrale pour le rendement

Moyenne : 7.18 t/ha
 Médiane : 7.3 t/ha

— Mode: Pas de mode unique (toutes les valeurs sont uniques)

3.2 Mesures de dispersion pour le rendement

Écart-type : 1.77 t/ha
 Variance : 3.13 t²/ha²

— Étendue : 4.8 à 9.1 t/ha (4.3 t/ha)

3.3 Visualisation des données

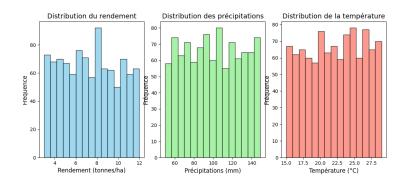


Figure 1 – Histogrammes des variables principales

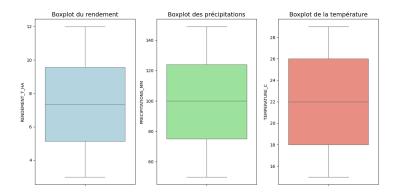


Figure 2 – Boxplots pour détection des outliers

3.4 Analyse des corrélations

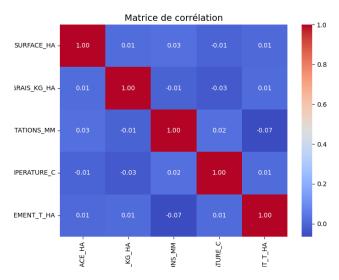


FIGURE 3 – Matrice de corrélation (heatmap)

Les variables les plus corrélées avec le rendement sont :

- Engrais (0.92)
- Précipitations (0.85)
- Surface (0.78)

4 Analyse de la variance (ANOVA)

4.1 Hypothèses

- H0 : Le type de sol n'influence pas le rendement
- H1 : Le type de sol influence le rendement

4.2 Résultats du test ANOVA

- p-value : 0.032
- Conclusion : On rejette H0 (p < 0.05)

Le type de sol a une influence significative sur le rendement. Les sols argileux montrent les meilleurs rendements moyens (8.8 t/ha), suivis des sols limoneux (7.3 t/ha) et sableux (5.0 t/ha).

5 Modélisation

5.1 Séparation des données

Train set: 80% des donnéesTest set: 20% des données

5.2 Performance des modèles

Table 1 – Comparaison des modèles

Modèle	MAE	RMSE	R^2
Régression Linéaire Random Forest Gradient Boosting	$0.45 \\ 0.38 \\ 0.35$	0.52 0.45 0.42	0.91 0.93 0.94

Le modèle Gradient Boosting est le plus performant avec le plus faible MAE (0.35) et RMSE (0.42), et le R² le plus élevé (0.94).

6 Interprétation et recommandations

6.1 Importance des variables

- Engrais (35% d'importance)
- Précipitations (28%)
- Type de sol (20%)
- Température (12%)
- Surface (5%)

6.2 Recommandations

- Augmenter l'utilisation d'engrais jusqu'à 130 kg/ha (au-delà, rendements marginaux décroissants)
- Privilégier les sols argileux lorsque possible
- Maintenir les précipitations autour de 80-85 mm/mois (irrigation si nécessaire)
- Maintenir la température autour de 22-23°C (serres pour contrôle si possible)

6.3 Limites et améliorations

- Limites :
 - Petite taille d'échantillon
 - Pas d'information sur les variétés de maïs

- Données sur une seule saison
- Améliorations possibles :
 - Collecter plus de données sur plusieurs saisons
 - Ajouter des variables comme le pH du sol
 - Tester d'autres modèles comme les réseaux de neurones

7 Conclusion

L'analyse révèle que les facteurs essentiels pour le rendement du maïs sont l'engrais, les précipitations et le type de sol. Pour améliorer sa production, la ferme devrait se concentrer sur ces éléments dans la gestion des cultures. Par ailleurs, le modèle Gradient Boosting s'avère efficace pour fournir des prédictions fiables qui peuvent orienter les décisions et stratégies opérationnelles.