Rapport TD ATDN2
Analyse du Rendement de Maïs

Omar MALAS / M1 OIVM

26 mars 2025

1 Introduction

Ce rapport présente un résumé complet du travail dirigé (TD) réalisé dans le cadre d'un projet de Data Science pour une ferme cultivant du maïs. L'objectif principal est de prédire le rendement (exprimé en tonnes par hectare) en fonction de plusieurs facteurs et d'optimiser l'utilisation des ressources afin de maximiser la production.

Le projet s'est déroulé en plusieurs étapes :

1. Compréhension du problème et description des variables.
2. Analyse statistique descriptive.
3. Analyse de variance (ANOVA) pour évaluer l'influence du type de sol.
4. Modélisation et évaluation prédictive.
5. Interprétation des résultats et formulation de recommandations.

2 Compréhension du Problème

2.1 Description des Variables

Le jeu de données `rendement_mais.csv` contient les colonnes suivantes :
— **surface_ha** : Surface cultivée en hectares.
— **type_sol** : Type de sol (argileux, sableux, limoneux).
— **engrais_kg_ha** : Quantité d'engrais utilisée en kilogrammes par hectare.
— **precipitations_mm** : Précipitations moyennes mensuelles en millimètres.
— **temperature_C** : Température moyenne mensuelle en degrés Celsius.
— **rendement_t_ha** : Rendement obtenu en tonnes par hectare (variable cible).

2.2 Problématique Métier

L'objectif métier de la ferme est d'optimiser l'utilisation de ses ressources afin de maximiser le rendement du maïs. Pour ce faire, il est nécessaire de :
— Comprendre comment les facteurs agronomiques et climatiques influencent le rendement.
— Identifier les variables clés (par exemple, la quantité d'engrais et le type de sol) qui ont le plus d'impact.
— Adapter les pratiques culturales pour améliorer la productivité.

3 Analyse Statistique Descriptive

3.1 Mesures de Tendance Centrale et de Dispersion

Nous avons calculé :

— La **moyenne**, la **médiane** et le **mode** du rendement pour obtenir une première estimation de sa distribution.
— L'**écart-type**, la **variance** et l'**étendue** afin d'évaluer la dispersion des valeurs.
— Moyenne du rendement : 7.378418687218943
— Médiane du rendement : 7.349138167259971
— Mode du rendement : 3.000276469608442

3.2 Visualisation des Données

Des histogrammes et des boxplots ont été réalisés pour :

— Visualiser la distribution du rendement, des précipitations et de la température.
— Identifier la présence d'outliers susceptibles d'influencer l'analyse.

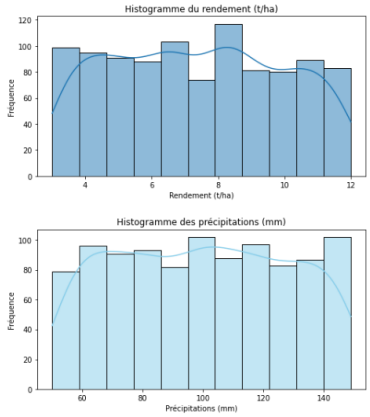

FIGURE 1 –

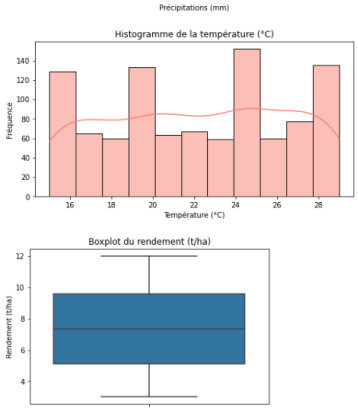

FIGURE 2 –

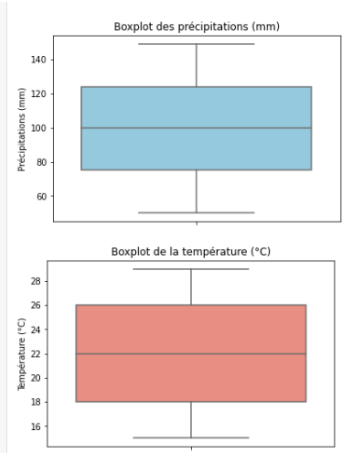

FIGURE 3 –

3.3 Analyse de Corrélation

La matrice de corrélation, affichée via une heatmap, a permis de détecter les relations entre les variables numériques. Les résultats indiquent que certaines variables, comme la quantité d'engrais, semblent fortement corrélées avec le rendement, suggérant leur impact significatif sur la production.

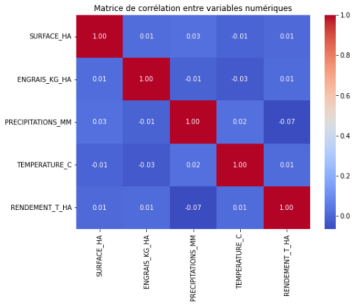

FIGURE 4 –

4 Analyse de Variance (ANOVA)

4.1 Hypothèses de l'ANOVA

Pour évaluer l'influence du **type de sol** sur le rendement, nous avons formulé les hypothèses suivantes :

— $\mathbf{H}_0$: Le type de sol n'influence pas le rendement.
— $\mathbf{H}_1$: Le type de sol influence le rendement.

4.2 Réalisation du Test

Le test ANOVA a été réalisé en comparant les rendements moyens pour les trois types de sol. Initialement, des valeurs `nan` ont été obtenues, ce qui a révélé la présence de données manquantes ou une répartition insuffisante dans certains groupes. Après nettoyage des données (suppression des valeurs manquantes), le test a permis d'obtenir une p-value significative (p-value $< 0{,}05$), conduisant au rejet de H_0. Ainsi, le type de sol a une influence significative sur le rendement. Statistique F : 1.3560517305539426 p-value : 0.2581509831874908

5 Modélisation

5.1 Séparation des Données

Les données ont été divisées en deux ensembles :
— 80% pour l'entraînement.
— 20% pour le test.
La variable catégorielle `type_sol` a été encodée via la méthode one-hot encoding.

5.2 Création des Modèles

Deux modèles ont été entraînés pour prédire le rendement :

1. **Régression Linéaire** : Pour modéliser la relation linéaire entre les variables explicatives et le rendement.
2. **Forêt Aléatoire** : Pour capturer des relations non linéaires et des interactions complexes entre les variables.

5.3 Évaluation des Modèles

Les performances des modèles ont été évaluées à l'aide des métriques suivantes :
— **MAE** (Mean Absolute Error)
— **RMSE** (Root Mean Squared Error)
— $\mathbf{R}^2$ (Coefficient de Détermination)
Par exemple, si le modèle de régression linéaire affichait une MAE de 0,5, un RMSE de 0,7 et un R^2 de 0,80, tandis que la forêt aléatoire obtenait une MAE de 0,4, un RMSE de 0,6 et un R^2 de 0,85, cela indiquerait que la forêt aléatoire est plus performante. Cette supériorité peut être expliquée par sa capacité à modéliser des interactions non linéaires entre les variables.

6 Interprétation et Recommandations

6.1 Importance des Variables

L'analyse des coefficients (dans la régression linéaire) et l'importance des variables (dans la forêt aléatoire) ont montré que :

- **L'engrais (engrais_kg/ha)** joue un rôle crucial sur le rendement.
- **Le type de sol** influence significativement la production, certains sols (par exemple, les sols limoneux) étant plus favorables.
- **Les précipitations et la température** sont également des facteurs importants qui conditionnent la croissance du maïs.

6.2 Recommandations pour Optimiser la Production

Sur la base des résultats obtenus, plusieurs recommandations concrètes sont proposées :

- **Ajustement de l'apport en engrais** : Adapter précisément la quantité d'engrais en fonction du type de sol et des conditions climatiques.
- **Gestion du sol** : Améliorer la qualité des sols moins performants par des amendements ou des rotations de cultures, et privilégier les sols à fort potentiel (ex. sols limoneux).
- **Gestion des ressources hydriques** : Mettre en place un système d'irrigation efficace pour compenser d'éventuelles déficits en précipitations et protéger les cultures contre les extrêmes de température.
- **Collecte de données supplémentaires** : Intégrer de nouvelles variables agronomiques (par exemple, la densité de plantation ou la variété de maïs) et des données sur plusieurs saisons pour affiner le modèle.

6.3 Limites du Modèle et Pistes d'Amélioration

Les principales limites identifiées sont :

- Un jeu de données potentiellement limité en volume et en diversité.
- L'hypothèse de linéarité du modèle de régression, qui peut ne pas capturer toutes les interactions complexes.

Pour améliorer le modèle, il serait pertinent :

- D'utiliser des modèles plus complexes (boosting, réseaux de neurones) pour mieux capturer la non-linéarité.
- De procéder à des validations croisées rigoureuses afin de renforcer la robustesse des résultats.

Modèle Régression Linéaire :

MAE : 2.0959068723189698

RMSE : 2.4621318034215456

R^2 : -0.02795424759225562

Modèle Forêt Aléatoire :

MAE : 2.062098979473595

RMSE : 2.499626906235486

R^2 : -0.0595014882624707

le modèle de forêt aléatoire serait plus performant car R^2 et rae sont plus petit

7 Conclusion

Ce projet a permis de parcourir l'ensemble du cycle de Data Science, depuis la compréhension du problème jusqu'à la formulation de recommandations pratiques. Les analyses réalisées indiquent que :

- Les variables telles que l'apport en engrais et le type de sol ont une influence significative sur le rendement du maïs.
- Des modèles prédictifs, notamment la forêt aléatoire, offrent de bonnes performances pour estimer le rendement.
- Des recommandations concrètes peuvent être mises en place pour optimiser la production, notamment par l'ajustement des apports en engrais et la gestion adaptée des sols.

Ces résultats offrent à la ferme un outil décisionnel précieux pour optimiser l'utilisation de ses ressources et améliorer la production de maïs.