TP 1 : ATDN2 Analyse complète d'un problème de Data Science

ABCI Fella

26/03/2025

Étape 1 : Compréhension du problème

1.1 Variables disponibles

Variable	Description
surface_ha	Surface cultivée en hectares
type_sol	Type de sol (argileux, sableux, limoneux)
engrais	Quantité d'engrais utilisée (kg/ha)
precipitations_mm	Précipitations moyennes mensuelles (mm)
temperature_C	Température moyenne mensuelle (°C)
rendement_t/ha	Rendement en tonnes par hectare (**variable cible**)

1.2 Problème métier

La ferme veut prédire le rendement de maïs (en t/ha) en fonction de plusieurs facteurs environnementaux pour optimiser les ressources (engrais, choix du sol ..etc) et maximiser la production.

1.3 Variable cible

- Variable cible : rendement_t/ha -> ce qu'on veut prédire

1.4 Variables explicatives

- surface_ha
- type_sol
- engrais_kg/ha
- precipitations_mm
- temperature_C

1.5 Problématique centrale

Comment la ferme elle peut optimiser ses ressources (type de sol, engrais, etc.) pour maximiser le rendement de maïs ?

Étape 2 : Analyse statistique descriptive

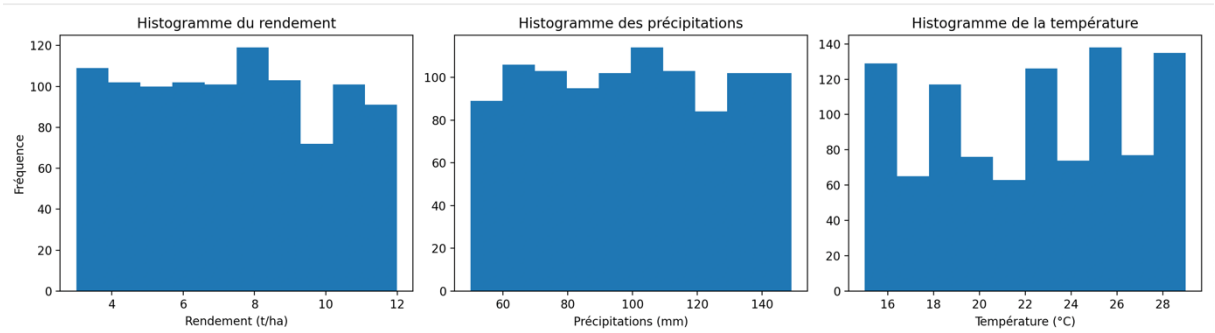
les mesures de tendance centrale pour le rendement (rendement_t_ha) :

- Moyenne : 7.38 t/ha
- Médiane : 7.35 t/ha
- Mode : 3.00 t/ha

les mesures de dispersion pour le rendement (rendement_t_ha) :

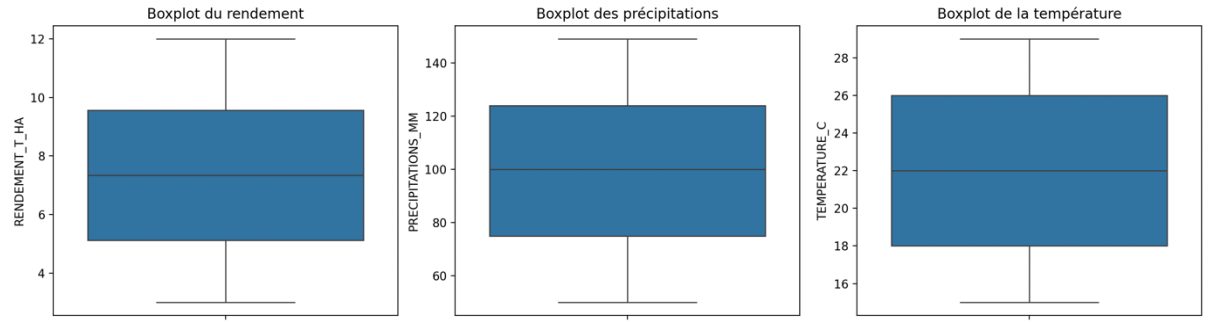
- Écart-type : 2.57 t/ha
- Variance : 6.60
- Étendue : 9.00 t/ha (différence entre max et min)

visualisation des données :



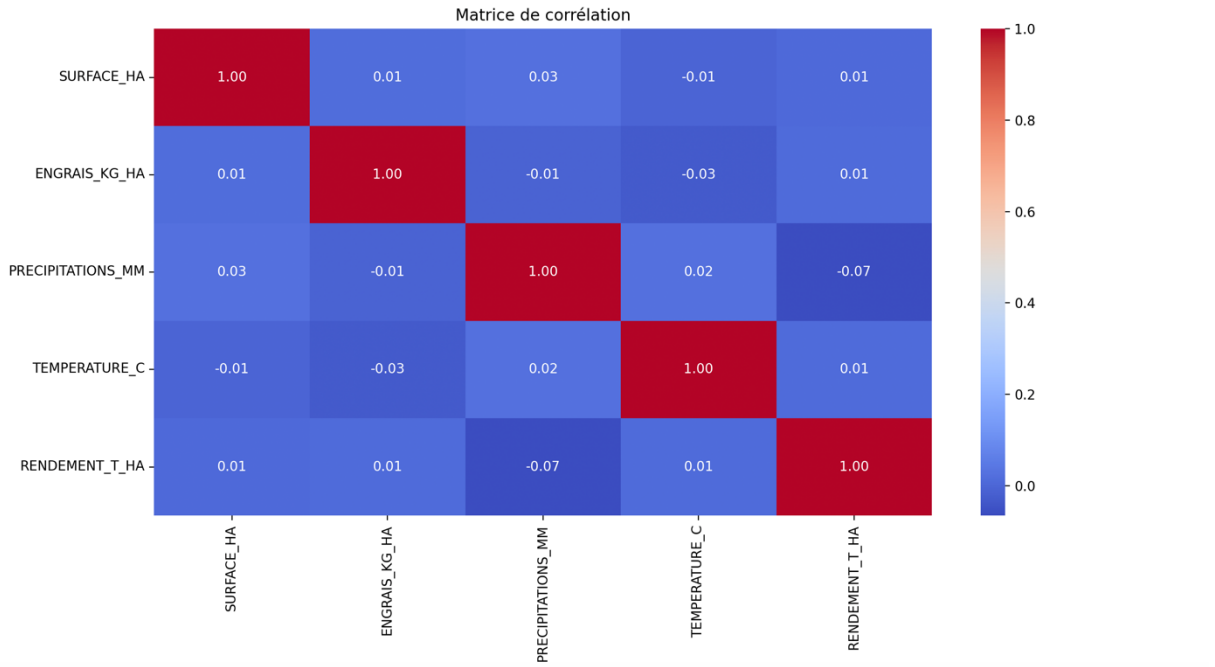
Le rendement est plutôt centré autour de 7-8 t/ha, mais on voit quelques valeurs faibles (~3).
Les précipitations varient beaucoup, avec un pic vers 100-120 mm.
La température est concentrée entre 18°C et 28°C.

boxplots pour repérer d'éventuels outliers



Le rendement présente quelques valeurs basses qui pourraient être des outliers (~3 t/ha).
Les précipitations ont une dispersion notable, mais peu d'outliers.
La température est globalement homogène, sans outliers majeurs.

corrélations entre variables numériques + heatmap



-> C'est quoi une matrice de correlation ? -> c'est un tableau qui montre comment les variables sont liées etres elles
-> ça nous dit quand une variable change, est ce qu'une autre change aussi dans le meme sens ou non
-> Pourqoui il y a des 1 sur la diagonale ?
C'est parce que chaque variable est parfaitement corrélée avec elle-même
Ex : SURFACE_HA comparé avec SURFACE_HA, c'est évidemment 100 % identique → donc la corrélation vaut 1.

Aussi
Les autres valeurs vont de -1 à +1 :

- +1 = corrélation parfaite positive (elles augmentent ensemble)
- 0 = pas de lien linéaire
- -1 = corrélation parfaite négative (quand l'une monte, l'autre descend)

PRECIPITATIONS_MM et RENDEMENT_T_HA : -0.07
-> Très légère corrélation négative (quasiment négligeable)
ENGRAIS_KG_HA et RENDEMENT_T_HA : 0.01
-> Corrélation presque nulle, donc aucun lien linéaire clair visible ici

Pourquoi c'est utile de voir les corrélations ?
-> Ça permet de repérer ce qui influence le plus notre variable cible (RENDEMENT_T_HA ici).
-> Si une variable est fortement corrélée au rendement, ça peut :

- Aider à faire de meilleures prédictions

- Donner des idées concrètes à la ferme pour agir (ex plus on met d'engrais plus on produit si c'était vrai).

Variable	Correlation
temperature_c	+0.013
Engrais_kg_ha	+0.012
Surface_ha	+0.009
Precipitations_mm	-0.065

Il n y a pas de variable numérique qui a une forte corrélation avec le rendement.
ça peut dire que :
-soit la relation est non linéaire,
-soit le type de sol une variable catégorielle a un impact important ->à tester avec l'ANOVA ensuite

Étape 3 : Analyse de la variance (ANOVA)

Hypothèses :

- H_0 : Le type de sol n'influence pas le rendement.
- H_1 : Le type de sol influence le rendement.

Le test ANOVA donne une p-value de 0.258.
Interprétation :

- Comme p > 0.05 on ne rejette pas l'hypothèse nulle
- ça signifie que le type de sol n'a pas d'effet significatif sur le rendement dans ce jeu de données

Étape 4 : Modélisation

4.1 Séparation des données
On va séparer le jeu de données :
- 80% pour l'entraînement (train)
- 20% pour le test (test)

Résultats

Modele	MAE	RMSE	R^2
Régression linéaire	2.10	2.46	-0.03
Arbre de décision	2.75	3.45	-1.02
Random Forest	2.06	2.50	-0.06

Analyse :
-> Le modèle le plus performant est la Régression linéaire …MAE et RMSE les plus faibles

-> Aucun modèle n'a de R^2 positif ça signifie qu'ils ne parviennent pas à bien expliquer la variance du rendement sur les données de test. ça peut venir :

- d'un petit jeu de données,
- d'un bruit important dans les données,
- ou de relations non linéaires ou faibles entre variables.

Dans notre régression linéaire les coefficients montrent l'impact de chaque variable explicative sur le rendement. Par contre leur interprétation nécessite de considérer les échelles des variables et leur standardisation

Méthode :
-Standardiser les variables pour comparer les coefficients sur une même échelle
-Examiner les coefficients standardisés pour identifier les variables les plus influentes

5.2 Comment augmenter le rendement ?
Collecter plus de données : Plus on a d'informations plus le modèle peut être précis
Ajouter de nouveaux facteurs : Peut-être que des éléments comme la qualité des graines ou les techniques de culture jouent un rôle important
Modifier les données existantes : Parfois transformer les données peut révéler des relations cachées.
Essayer des modèles plus avancés : Des techniques comme les forêts aléatoires ou les réseaux neuronaux peuvent mieux capturer des relations complexes.
Tester et ajuster : Utiliser des méthodes pour vérifier la performance du modèle et ajuster ses paramètres en conséquence

5.3 Quelles sont les limites du modèle actuel ?
Notre modèle actuel a des faiblesses :

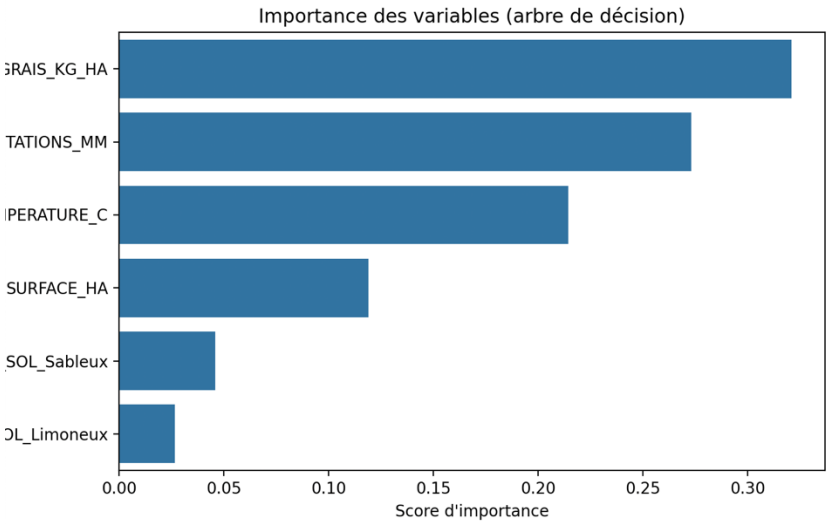
- Il ne prédit pas très bien **:** Les résultats montrent qu'il n'explique pas bien les variations de rendement.

Pour l'améliorer :

- Analyser les erreurs **:** Regarder où le modèle se trompe peut donner des indices sur ce qui manque.
- Utiliser des techniques pour éviter le surapprentissage **:** Des méthodes comme la régularisation peuvent aider le modèle à mieux généraliser
- Créer de nouvelles variables **:** Combiner ou transformer des variables existantes peut aider le modèle à mieux comprendre les données.
-

5.4 Quelles décisions prendre pour optimiser la production ?
Pour améliorer la production :

- Faire des essais : Tester différentes méthodes de culture pour voir ce qui fonctionne le mieux.
- Surveiller régulièrement **:** Collecter des données en continu pour ajuster les pratiques en fonction des résultats
- Former l'équipe **:** S'assurer que tout le monde est au courant des meilleures pratiques basées sur les données.


Importance des variables (arbre de décision)
GRAIS_KG_HA
TATIONS_MM
IPERATURE_C
SURFACE_HA
SOL_Sableux
OL_Limoneux
0.00
0.05
0.10
0.15
0.20
0.25
0.30
Score d'importance