Rapport TP1 :

Analyse complète d'un problème de Data Science

Étape 1 : Compréhension du problème

1.1 Variables disponibles

Les variables présentes dans le jeu de données sont les suivantes :

surface_ha : Surface cultivée en hectares

type_sol : Type de sol (argileux, sableux, limoneux)

engrais_kg/ha : Quantité d'engrais utilisée en kg/ha

precipitations_mm : Précipitations moyennes mensuelles en mm

temperature_C : Température moyenne mensuelle en °C

rendement_t/ha : Rendement obtenu en tonnes par hectare (variable cible)

1.2 Formulation du problème métier

Le but est de prédire le rendement du maïs en fonction de plusieurs facteurs (type de sol, quantité d'engrais, précipitations, température, surface cultivée) afin d'optimiser les ressources (engrais, choix du sol, etc.) et maximiser la production de maïs.

1.3 Identification de la variable cible et des variables explicatives

Variable cible : rendement_t/ha (c'est ce qu'on cherche à prédire).

Variables explicatives : surface_ha, type_sol, engrais_kg/ha, precipitations_mm, temperature_C (ce sont les facteurs influençant le rendement).

1.4 La problématique centrale pour la ferme

La ferme souhaite optimiser ses ressources en ajustant l'utilisation d'engrais, le choix du type de sol, et d'autres facteurs pour maximiser le rendement du maïs. L'objectif est d'identifier les principaux facteurs influençant le rendement et d'optimiser les pratiques agricoles.

Etape 2 : Analyse statistique descriptive

2.1 Mesures de tendance centrale

Calcule de la moyenne, de la médiane, et du mode du rendement.

On calcule la moyenne, la médiane, et le mode du rendement pour comprendre la tendance générale du rendement.

Moyenne du rendement : 7.38

Médiane du rendement : 7.35

Mode du rendement : 3.0002

2.2 Mesures de dispersion

Calcule de l'écart-type, de la variance, et l'étendue du rendement.

On calcule l'écart-type, la variance et l'étendue du rendement pour évaluer la dispersion de ces valeurs.

Écart-type du rendement : 2.57

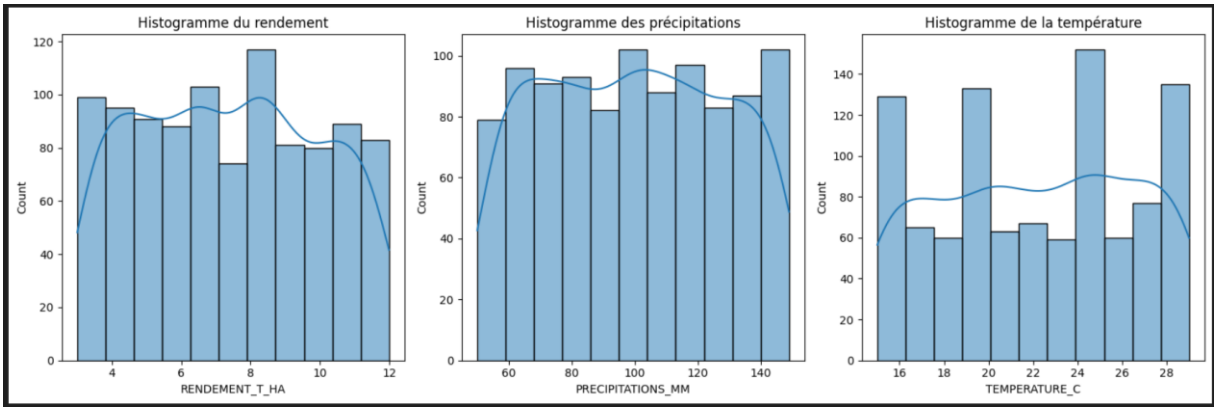
Variance du rendement : 6.60

Étendue du rendement : 8.99

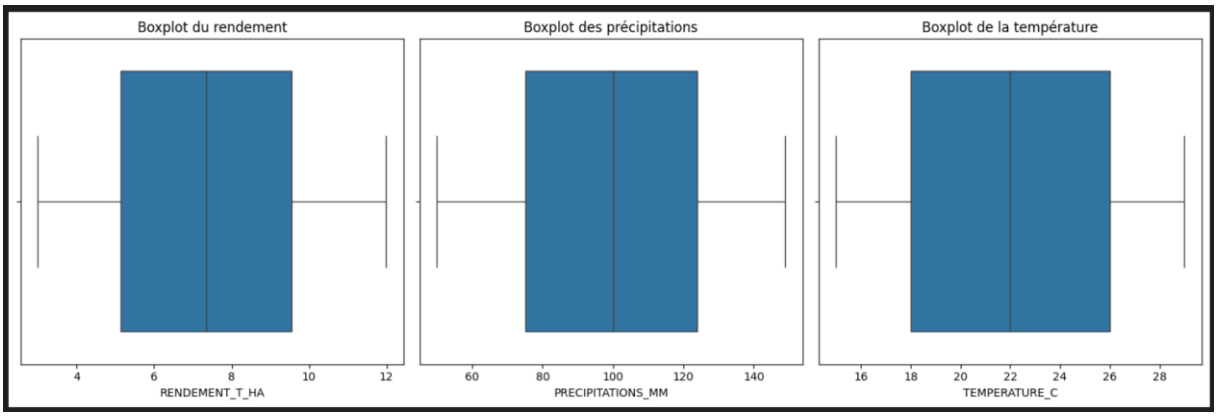
2.3 Visualisation des données

Créez des histogrammes pour le rendement, les précipitations, et la température. Affichez des boxplots pour identifier d'éventuels outliers.

Histogrammes : Affichent la distribution des variables rendement_t/ha, precipitations_mm, et temperature_C pour une première exploration visuelle.



Boxplots : Permettent de détecter d'éventuels outliers dans les variables.



D’après le boxplot :

- Pas de valeurs aberrantes évidentes : aucune valeur ne semble être un outlier extrême.
- Les données sont bien distribuées sans points isolés en dehors des moustaches.

2.4 Corrélations

Calcule de la matrice de corrélation entre les variables numériques. Affichage d’une heatmap pour visualiser les corrélations.

Variables et leur influence sur le rendement :

PRECIPITATIONS_MM (-0.07) :

- Corrélation légèrement négative : C'est la plus forte relation (bien que faible) avec le rendement.

Interprétation : Une augmentation des précipitations pourrait être associée à une légère baisse du rendement, mais le lien est très limité (seulement 0.07).

ENGRAIS_KG_HA (0.01) :

- Corrélation quasi-nulle : Aucun impact significatif détecté.

TEMPERATURE_C (0.01) :

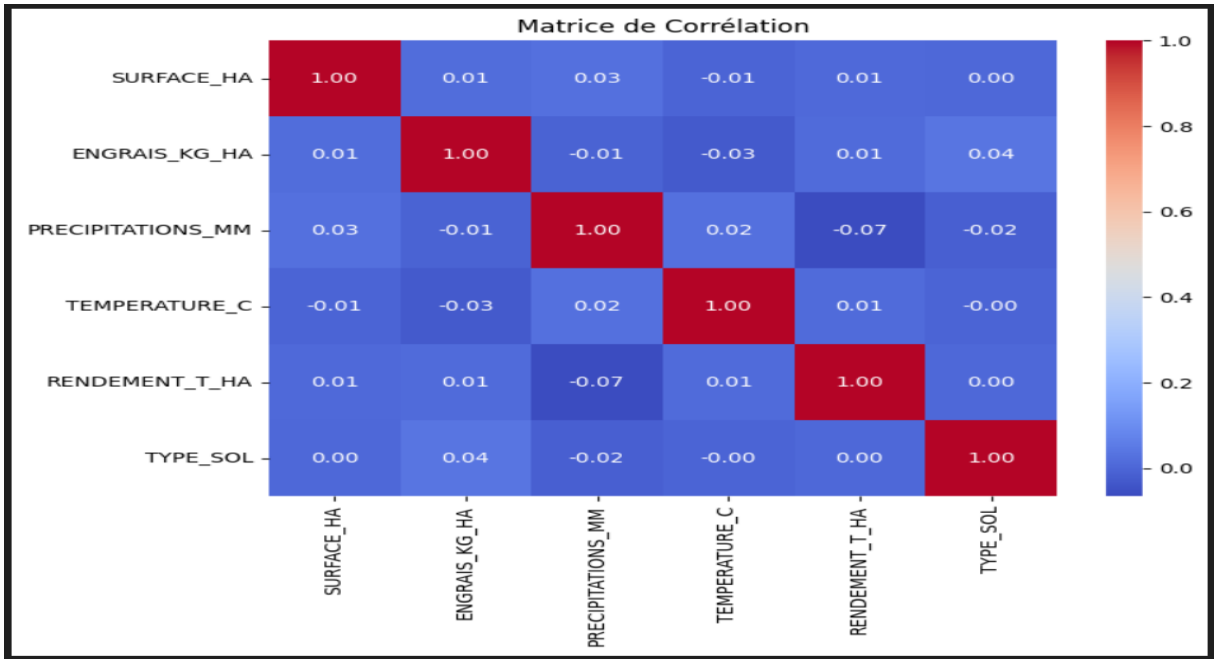
- Aucune corrélation : La température ne semble pas influencer le rendement dans ces données.

SURFACE_HA (0.01) et TYPE_SOL (0.00) :

- Impact nul selon cette analyse.

Aucune variable n'a d'impact significatif sur le rendement dans notre dataset actuel (toutes les corrélations sont proches de 0).

Le seul lien (très faible) est avec les précipitations (-0.07), mais il est trop faible pour être exploitable.



Etape 3 : Analyse de la variance (ANOVA)

3.1 Hypothèses

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

H1 : Le type de sol influence le rendement.

3.2 Test ANOVA

Réalisez une ANOVA sur le type de sol.

- Si la p-value est faible (< 0.05), cela suggère que les rendements sont significativement différents selon le type de sol, ce qui nous amène à rejeter l'hypothèse nulle (H0).
- Si la p-value est élevée (> 0.05), nous ne pouvons pas conclure que le type de sol a un effet significatif sur le rendement.

Interprétation de la p-value obtenue :

Le type de sol a-t-il une influence significative sur le rendement ?

Valeur de la statistique F : 1.36

P-value : 0.26

P-value > 0.05, donc l'hypothèse H0 n'est pas rejetée (Le type de sol n'influence pas le rendement)

Etape 4 : Modélisation

4.1 Séparation des données

Divisez les données en train (80%) et test (20%).

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

4.2 Création du modèle

Entraînement des modèles de votre choix vu précédemment pour prédire le rendement.

4.3 Évaluation du modèle

Calcule des métriques : MAE, RMSE, et R^2 des modèles.

Régression Linéaire : MAE = 2.06, RMSE = 2.42, R^2 = 0.0017

Forêt Aléatoire : MAE = 2.09, RMSE = 2.56, R^2 = -0.11

Gradient Boosting: MAE = 2.11, RMSE = 2.53, R^2 = -0.087

SVR: MAE = 2.24, RMSE = 2.69, R^2 = -0.23

XGBoost: MAE = 2.25, RMSE = 2.71, R^2 = -0.25

Lequel des modèles est-il performant (pourquoi d'après vous) ?

Le modèle le plus performant ici est la Régression Linéaire car il a le meilleur R^2 (0.0017, proche de 0), ce qui signifie qu'il explique légèrement mieux la variance du rendement que les autres modèles. Il a aussi le plus faible RMSE (2.43) et un MAE relativement bas (2.06).

Les autres modèles ont des R^2 négatifs, ce qui indique qu'ils font pire qu'une simple moyenne des valeurs. Cela signifie qu'ils ne parviennent pas à bien capturer les tendances dans les données.

Cependant, les performances globales restent faibles, ce qui suggère que les variables actuelles n'expliquent pas bien le rendement du maïs.

Ceci peut être expliqué par le fait qu'il y a une mauvaise corrélation entre les variables et le rendement comme le montre la matrice de corrélation

- Si les variables explicatives (surface, engrais, précipitations, température) n'ont pas de relation forte avec le rendement, les modèles auront du mal à bien prédire.

Etape 5 : Interprétation et recommandations

5.1 Analyse de l'importance des variables.

D'après les résultats et les performances des modèles, les variables utilisées (SURFACE_HA, ENGRAIS_KG_HA, PRÉCIPITATIONS_MM, TEMPÉRATURE_C) n'expliquent pas suffisamment le rendement du maïs.

Faible impact des variables sur le rendement : Les valeurs négatives du R2 indiquent que les modèles ne parviennent pas à expliquer la variabilité du rendement.

5.2 Recommandations pour augmenter le rendement

Pour optimiser la production de maïs, on peut :

Ajuster l'utilisation des engrais :

- Tester différentes quantités et types d'engrais (azotés, phosphatés, potassiques) pour identifier la meilleure combinaison.
- Appliquer l'engrais au bon moment du cycle de croissance pour maximiser son absorption par les plantes.

Optimiser l'irrigation et la gestion de l'eau :

- Si possible, compléter les précipitations avec un système d'irrigation adapté pour assurer un apport en eau constant.
- Surveiller l'humidité du sol pour éviter le stress hydrique ou l'excès d'eau.

Choisir des variétés de maïs plus adaptées :

- Sélectionner des semences résistantes aux températures élevées et aux conditions climatiques locales.
  Expérimenter différentes variétés hybrides qui pourraient offrir de meilleurs rendements.

Prendre en compte le type de sol :

- Tester la composition du sol et ajuster les nutriments selon ses caractéristiques.

5.3 Limites du modèle et pistes d'amélioration.

Variables explicatives insuffisantes :

- Il manque d'autres facteurs clés comme l'altitude, la densité de plantation, ou encore la qualité des semences.

Méthodes de modélisation à améliorer :

- Tester des modèles plus avancés comme les réseaux de neurones, les modèles de séries temporelles ou les modèles basés sur l'optimisation bayésienne.
- Utiliser des techniques de sélection de variables pour identifier celles qui influencent réellement le rendement.

5.4 Décisions possibles pour optimiser la production

Suivi et analyse des sols : Identifier les besoins en nutriments et ajuster les engrais en conséquence.

Mise en place d'un système d'irrigation : Si l'eau est un facteur limitant, investir dans un arrosage efficace.

Sélection des meilleures variétés de maïs : Tester différentes semences pour voir lesquelles offrent les meilleurs rendements dans les conditions locales.

Collecte de plus de données : Enregistrer plus d'informations sur chaque parcelle pour affiner les modèles et prendre des décisions basées sur les données.

Conclusion

Ce TP montre que les variables utilisées (surface, engrais, précipitations, température) expliquent mal le rendement du maïs, d'où les performances médiocres des modèles. Il y a peut-être d'autres facteurs qui sont probablement plus influents. Ainsi, pour améliorer la prédiction du rendement et optimiser la production agricole, il serait essentiel d'intégrer des données plus diversifiées et de tester des modèles plus avancés. Ce TP met donc en évidence l'importance du choix des variables dans toute modélisation prédictive, car un modèle, aussi sophistiqué soit-il, ne peut être performant si les bonnes variables explicatives ne sont pas prises en compte.