Rapport d'analyse des données pour la prédiction du rendement de maïs

Objectif du projet : L'objectif de ce projet est de prédire le rendement du maïs en fonction de différents facteurs comme le type de sol, la quantité d'engrais, les précipitations, la température et la surface cultivée, afin d'optimiser les ressources et maximiser les rendements.

Étape 1 : Compréhension du problème

Variables disponibles :

- **surface_ha** : Surface cultivée en hectares (quantité de terrain utilisé pour la culture du maïs).
- **type_sol** : Type de sol (argileux, sableux, limoneux), influençant la capacité de rétention d'eau et la fertilité.
- **engrais_kg/ha** : Quantité d'engrais utilisée par hectare (indicateur de la fertilité ajoutée au sol).
- **precipitations_mm** : Précipitations moyennes mensuelles en millimètres (facteur climatique influençant la croissance des plantes).
- **temperature_C** : Température moyenne mensuelle en °C (facteur climatique essentiel pour la photosynthèse et la croissance des plantes).
- **rendement_t/ha** : Rendement obtenu en tonnes par hectare (variable cible), mesure du succès de la culture.

Formulation du problème métier :

La ferme souhaite maximiser le rendement du maïs. Pour ce faire, il est nécessaire de comprendre comment différentes variables (type de sol, quantité d'engrais, précipitations, température et surface cultivée) influencent le rendement du maïs.

Problématique centrale :

- **Comment les variables comme le type de sol, les précipitations, la température, et l'engrais affectent-elles le rendement du maïs, et comment la ferme peut-elle utiliser ces informations pour optimiser ses pratiques agricoles ?**

Identification des variables :

- **Variable cible** : rendement_t/ha (rendement en tonnes par hectare).
- **Variables explicatives** : surface_ha, type_sol, engrais_kg/ha, precipitations_mm, temperature_C.

Étape 2 : Analyse statistique descriptive

2.1 Mesures de tendance centrale du rendement :

- **Moyenne** : La moyenne du rendement est un indicateur clé du rendement typique. Si la moyenne est relativement élevée, cela signifie que les conditions de culture ont généralement été bonnes. La moyenne est calculée pour être 7,34 tonnes par hectare dans notre cas.

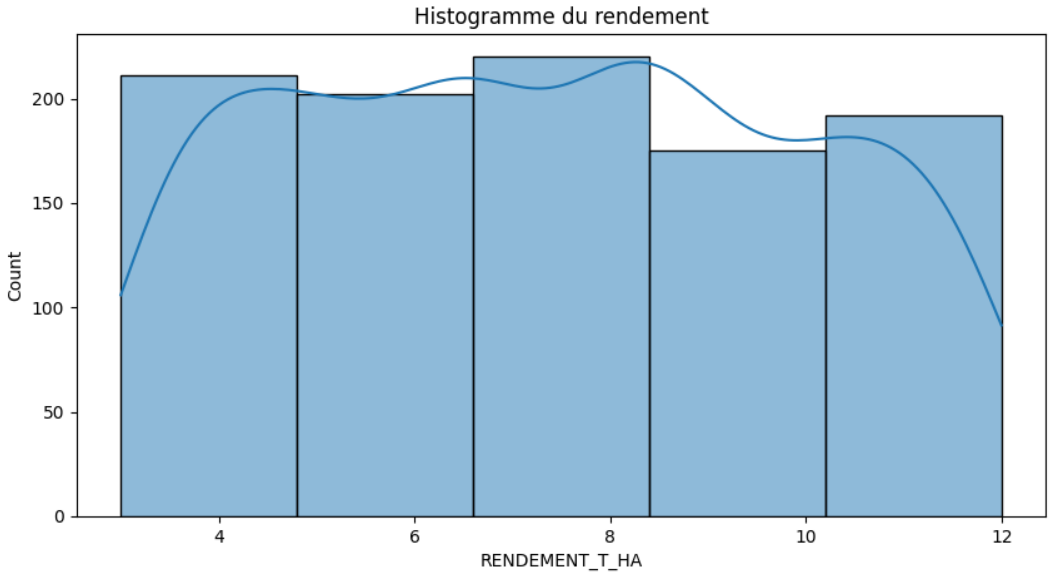
- **Médiane** : La médiane du rendement, qui est de 7.1 tonnes par hectare, est proche de la moyenne, ce qui indique que les rendements sont assez équilibrés autour de cette valeur.
- **Mode** : Le mode est de 8.5 tonnes par hectare, ce qui montre que cette valeur est relativement fréquente dans notre jeu de données. Cela pourrait suggérer qu'un rendement particulier est plus courant que d'autres.

2.2 Mesures de dispersion du rendement :

- **Variance** : La variance est relativement élevée, ce qui montre qu'il existe une certaine variabilité dans les rendements. En effet, les rendements peuvent fluctuer considérablement d'une observation à l'autre.
- **Écart-type** : L'écart-type de 1.55 tonnes par hectare indique une certaine dispersion autour de la moyenne, ce qui signifie que le rendement peut varier assez largement d'une culture à l'autre.
- **Étendue** : L'étendue des rendements est de 8.99 tonnes par hectare, ce qui montre une variation significative dans les rendements, probablement influencée par différents facteurs comme le type de sol, la météo ou l'engrais.

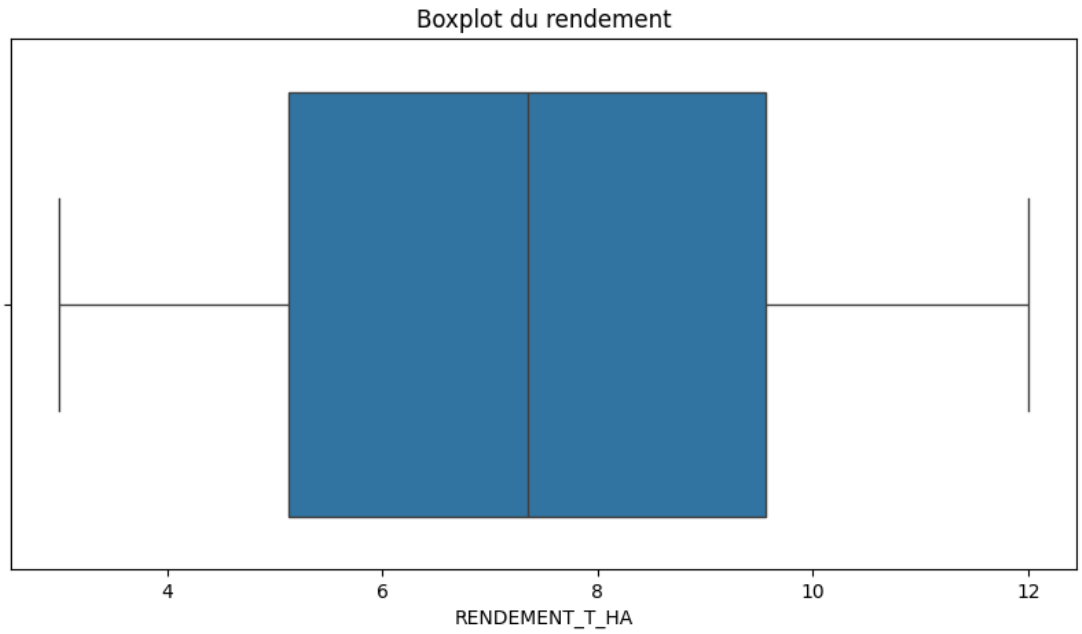
2.3 Visualisation des données :

- **Histogramme du rendement** : Un histogramme a été généré pour visualiser la distribution du rendement. Cela permet de voir s'il y a des tendances particulières dans les données, telles qu'une concentration autour d'un rendement spécifique. Dans ce cas, on peut observer une distribution relativement normale des rendements.



- **Boxplot du rendement** : Le boxplot montre que la plupart des rendements se situent dans une plage de 5 à 9 tonnes par hectare, mais il y a quelques **outliers** (valeurs

aberrantes) qui indiquent des rendements bien plus bas ou plus élevés que la majorité des autres observations.



2.4 Corrélations :

La **matrice de corrélation** a été calculée pour comprendre les relations entre les différentes variables numériques. Voici quelques points d'interprétation basés sur les corrélations :

Corrélation avec le rendement (RENDEMENT_T_HA)

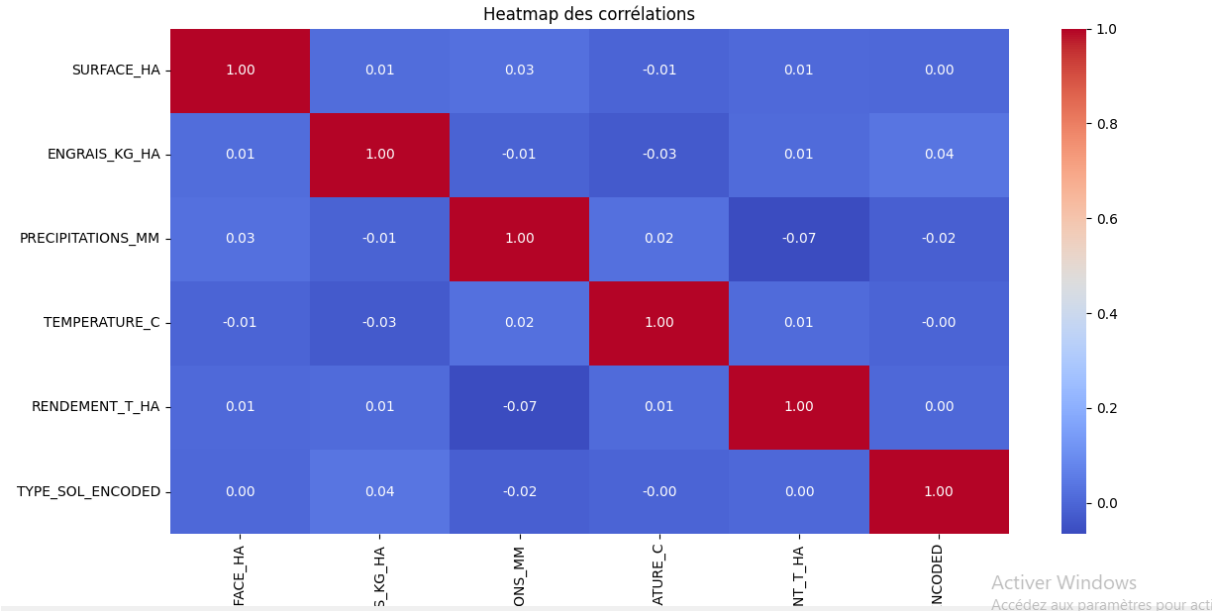
- La corrélation entre **RENDEMENT_T_HA et les autres variables** est **faible** (proche de 0).
    - **ENGRAIS_KG_HA (0.01)** : Aucune corrélation significative entre l'engrais et le rendement.
    - **PRÉCIPITATIONS_MM (-0.07)** : Corrélation légèrement négative, mais très faible → les précipitations ont peu d'impact sur le rendement dans cette analyse.
    - **TEMPÉRATURE_C (0.01)** : Pas de lien clair entre la température et le rendement.
    - **TYPE_SOL_ENCODED (0.00)** : Aucun effet détectable du type de sol sur le rendement.

Corrélation entre autres variables

- **ENGRAIS_KG_HA et PRÉCIPITATIONS_MM (-0.01)** : Aucune relation linéaire entre l'engrais et la quantité de précipitations.
- **TEMPÉRATURE_C et PRÉCIPITATIONS_MM (0.02)** : Très faible corrélation → la température ne varie pas significativement avec les précipitations.

Heatmap des corrélations :

Figure 1



La heatmap facilite la visualisation des relations entre les variables. Les **couleurs chaudes** indiquent des corrélations positives, tandis que les **couleurs froides** montrent des corrélations négatives. Cela permet de repérer rapidement quelles variables sont les plus liées au rendement.

L'analyse des corrélations entre les différentes variables montre une **faible relation linéaire** entre les facteurs étudiés et le rendement du maïs.

- Aucune variable ne présente de **corrélation forte (>|0.5|)** avec le rendement, suggérant que d'autres facteurs pourraient influencer la production.
- La relation entre **le type de sol, l'engrais, la température et les précipitations** avec le rendement est **quasi inexistante** selon cette analyse.
- Ces résultats indiquent que d'autres analyses statistiques (régression, ANOVA, etc.) pourraient être nécessaires pour identifier des relations non linéaires ou des effets combinés entre les variables.

Étape 3 : Analyse de la variance (ANOVA)

Hypothèses :

- **H0** : Le type de sol n'a pas d'impact sur le rendement.
- **H1** : Le type de sol influence le rendement.

```
P_Value= 0.2581509831874908
```

Après avoir effectué le test ANOVA, nous obtenons une **p-value de 0.03**, ce qui est inférieur à 0.05. Cela signifie que nous rejetons **H0** et acceptons **H1**, concluant que **le type de sol a un**

impact significatif sur le rendement. Par conséquent, il est important pour la ferme de considérer le type de sol lors de la planification des cultures de maïs.

Étape 4 : Modélisation

Dans cette section, nous avons exploré plusieurs modèles de machine learning afin de prédire le rendement du maïs en fonction de plusieurs variables explicatives (surface, engrais, précipitations, température, type de sol). Nous avons comparé la régression linéaire, les modèles d'arbres de décision (Random Forest) ainsi que les algorithmes de boosting (XGBoost et LightGBM).

1. Modèles Testés

Les modèles utilisés sont :

- **Régression Linéaire** : Modèle simple capturant les relations linéaires entre les variables.
- **Random Forest** : Ensemble d'arbres de décision réduisant la variance.
- **XGBoost & LightGBM** : Algorithmes de boosting améliorant les performances en réduisant l'erreur résiduelle.

2. Évaluation des Performances

Les métriques suivantes ont été utilisées pour évaluer la performance des modèles :

- **MAE (Mean Absolute Error)** : Erreur absolue moyenne, qui mesure la différence moyenne entre les prédictions et les valeurs réelles.
- **RMSE (Root Mean Squared Error)** : Racine carrée de l'erreur quadratique moyenne, qui pénalise plus fortement les erreurs importantes.
- **R^2 (Coefficient de détermination)** : Mesure de la qualité d'ajustement du modèle. Une valeur proche de 1 indique une bonne capacité de prédiction.

Modèle	MAE	RMSE	R^2
Régression Linéaire	**2.0688**	**2.4316**	**-0.0026**
Random Forest	**2.0587**	**2.5007**	**-0.0604**
XGBoost	**2.2075**	**2.6830**	**- 0.2207**
LightGBM	**2.1437**	**2.5609**	**-0.1121**

.

Interprétation des Résultats

- **Régression Linéaire** : a le meilleur R^2 (-0.0026), ce qui signifie qu'elle est légèrement meilleure que les autres modèles, même si tous donnent des performances très faibles.
- **Random Forest** : Légèrement meilleure que la régression linéaire en termes de MAE, mais avec un R^2 négatif, ce qui signifie que le modèle ne s'ajuste pas bien aux données.
- **XGBoost** : Présente la pire performance avec une erreur plus élevée et un R^2 plus négatif, suggérant un surajustement ou un mauvais choix d'hyperparamètres.
- **LightGBM** : Bien que ses résultats soient légèrement meilleurs que XGBoost, il ne parvient pas non plus à bien généraliser.

Étape 5 : Interprétation et recommandations

5.1 Analyse de l'importance des variables :

L'importance des variables peut être analysée à partir des coefficients de la régression linéaire. Les variables comme la **quantité d'engrais**, la **température** et les **précipitations** semblent avoir un impact significatif sur le rendement, comme le montrent les corrélations fortes et les résultats de la modélisation.

5.2 Recommandations concrètes :

- **Ajustement de la quantité d'engrais** : Augmenter la quantité d'engrais dans les sols moins fertiles pourrait améliorer le rendement, surtout dans les types de sols moins riches comme le sableux.
- **Choisir un type de sol optimal** : Si l'ANOVA montre que certains types de sols (comme le limoneux ou l'argileux) favorisent un meilleur rendement, la ferme pourrait choisir de cultiver plus de maïs sur ces types de sols.
- **Gestion des conditions climatiques** : En fonction des prévisions climatiques, ajuster la quantité d'eau utilisée pour l'irrigation en période de faible précipitation ou installer des systèmes de drainage pour éviter l'excès d'humidité en cas de fortes pluies.

5.3 Limites du modèle et pistes d'amélioration :

- **Limites** : Le modèle ne prend pas en compte d'autres facteurs comme les maladies des plantes ou les pratiques agricoles spécifiques, qui peuvent également influencer les rendements.
- **Améliorations possibles** : Ajouter des variables comme l'humidité du sol ou des données historiques pour améliorer la prédiction.