UPEC
Connaissance - Action

UNIVERSITÉ
PARIS-EST CRÉTEIL
VAL DE MARNE

TP1 : TD ATDN2

Rapport de TP Analyse du Rendement du Maïs

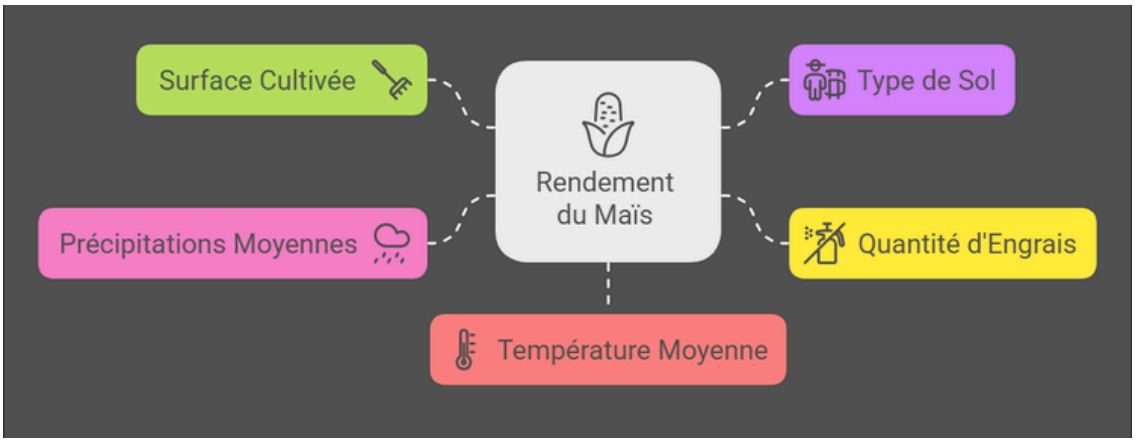
GAKOU YOUSSOUF

26/03/2025

1.Introduction

L'objectif de ce rapport est de comprendre les facteurs qui influencent le rendement du maïs. En étudiant ces variables, la ferme pourra optimiser ses ressources pour améliorer sa production.

Dans ce TP on a un jeu de données contenant plusieurs informations : la surface cultivée, le type de sol, la quantité d'engrais utilisée, la température et le rendement obtenu en tonnes par hectare. L'idée est de voir si certains de ces éléments jouent un role ou pas dans la variation du rendement et voir si il est possible de construire un modèle capable d'optimiser la production.



La problématique est donc la suivante : comment optimiser le rendement du maïs en fonction des variables ? Pour répondre a cette question je vais analyser les différentes variables(SURFACE_HA TYPE_SOL ENGRAIS_KG/HA PRECIPITATIONS_MM TEMPERATURE_C RENDEMENT_T/HA) et appliquer des analyse statistiques dessus ensuite je vais construire un modèle qui va nous permettre d'optimiser la production

2. Compréhension des Données

Pour commencer voici les variables utilisées il y a plusieurs variables explicatives :

- **SURFACE_HA**: Cest la taille du champ en hectares si la surface et grande ca peut augmente la production totale mais pas forcément le rendement par hectare.

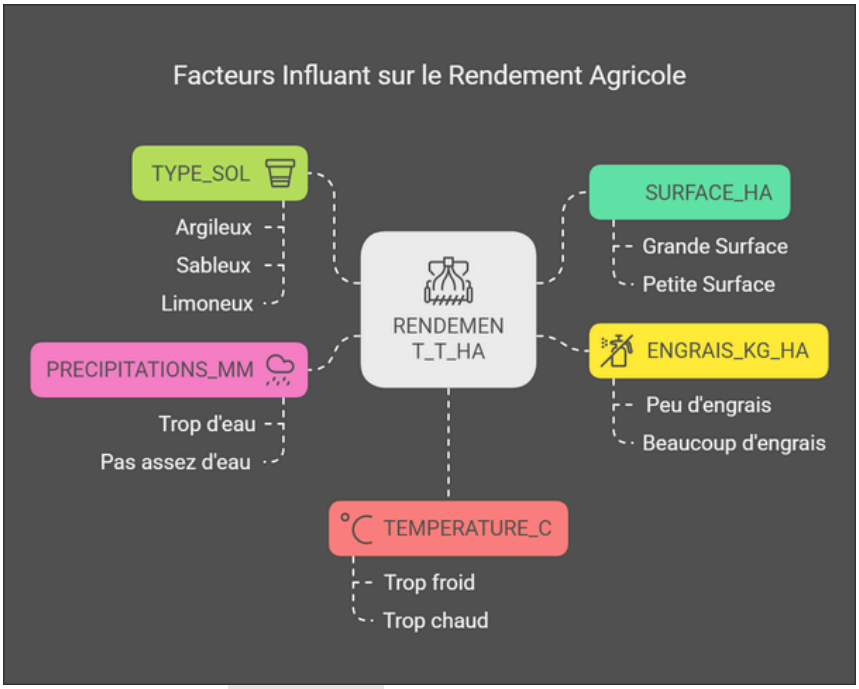
- **TYPE_SOL** : c'est une variable avec trois types : argileux, sableux et limoneux. certains sont plus efficaces que d'autre ca peut donc influencer le rendement.

- **ENGRAIS_KG_HA**: cest une variable qui affecte la croissance des plantes. Plus d'engrais peut augmenter le rendement mais si on en mets bcp ca peut être contre productif.

- **PRECIPITATIONS_MM** : c'est une valeurs numérique ça améliore le rendement jusqu'à un certain point mais après si il ya trop d'eau ca nuit à la production.

- **TEMPERATURE_C**: ca impacte la croissance du maïs si la températures est trop basses ou élevées ca peut nuire au rendement

- **RENDEMENT_T_HA**: cest le rendement en tonnes par hectare cest cque que nous cherchons à expliquer c'est la variable cible.



Je vais examiner l'impact de chaque variable sur le rendement de la production pour déterminer laquelle influence le plus la quantité de maïs produite par hectare.

3.Analyse Statistique Descriptive

Pour commencer j'ai calculé la moyenne, la médiane et le mode de rendement pour avoir une première idée de la distribution des valeurs

```
# Moyenne et médiane
print("Moyenne du rendement:", data["RENDEMENT_T_HA"].mean())
print("Médiane du rendement:", data["RENDEMENT_T_HA"].median())

# Mode
mode_value = data["RENDEMENT_T_HA"].mode()
```

Les résultats montrent que la moyenne du rendement est de 7,38 t/ha ca veut dire que cest la production moyenne dans le jeu de donnees.

```
Moyenne du rendement: 7.378418687218944
Médiane du rendement: 7.349138167259971
Mode du rendement: [ 3.00027647  3.00588052  3.01000885  3.03768699  3.05741139  3.05817922
  3.06190166  3.09256046  3.09729975  3.10172283  3.11607723  3.12314173
  3.12475922  3.14580358  3.14619099  3.14933233  3.17239937  3.18124064
  3.18840706  3.19859534  3.20308061  3.21336978  3.21551639  3.2160454
  3.22474496  3.22679185  3.2312068   3.23192639  3.25067017  3.25787185
  3.27093485  3.27417318  3.29199101  3.29393155  3.31205434  3.31420397
  3.35342473  3.3545265   3.36053207  3.36124973  3.36232203  3.36256967
  3.37421064  3.37723695  3.39659275  3.40497139  3.41026642  3.4352644
  3.43825364  3.44298823  3.44652936  3.45218709  3.45314574  3.45375096
  3.45515793  3.45560225  3.46035185  3.46164464  3.47488732  3.49108867
  3.5004497   3.51012389  3.51193258  3.51324223  3.52708775  3.53628388
```

La mediane est de 7,35 t/ha ce qui signifie que la moitié des observations ont un rendement inférieur a 7,35 et l'autre moitié un rendement supérieur a 7,35 et pour le mode il est pas clairement défini les valure tourne entre 3 et 11 ce qui indique que les valeurs du rendement sont assez dispersées et qu'il n'ya pas de valeur qui revient plus souvent que les autres

j'aiégalement calculé l'écart type la variance et l'étendue du rendement afin de voir si les valeurs sont homogènes ou si elles varient fortement

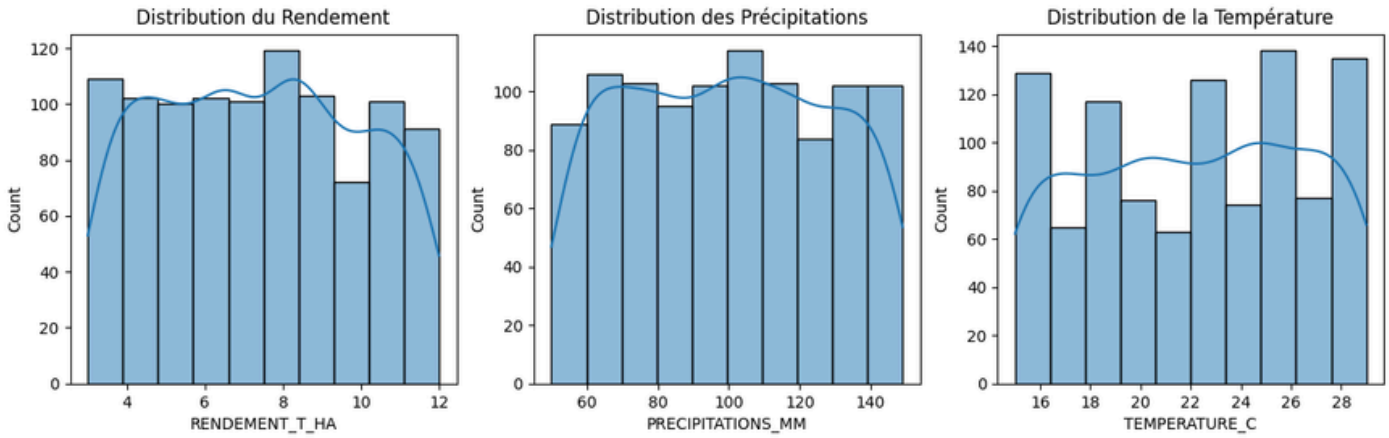
```
print("Écart-type:", data["RENDEMENT_T_HA"].std())
print("Variance:", data["RENDEMENT_T_HA"].var())
print("Étendue:", data["RENDEMENT_T_HA"].max() - data["RENDEMENT_T_HA"].min())
```

```
Écart-type: 2.5699909853267067
Variance: 6.604853664660536
Étendue: 8.995742859645505
```

L' écart type est de 2,57 ca veut dire que les rendements varient significativement autour de la moyenne et la variance est de 6,60 létendue est de 8,99 ca veut dire que que l'écart entre le rendement le plus faible et le plus élevé est tres grand avec ces valeurs je peut en deduire que les rendements sont pas uniformes et dépendent surment des conditions de culture

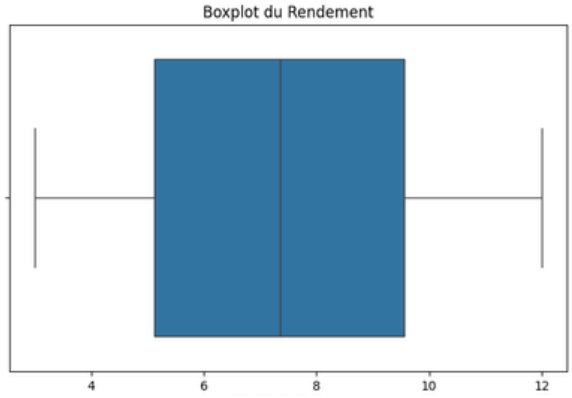
4. Visualisation des données

J'ai tracé des histogrammes pour mieux comprendre la distribution du rendement, des précipitation et de la température.



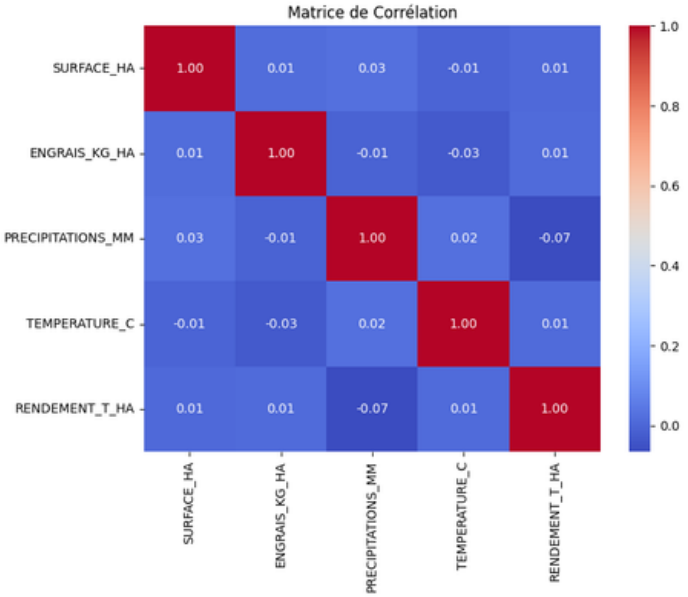
Ces graphiques montrent que le rendement est relativement stable, tandis que les précipitations et la température présentent des variations notables. Cela peut aider à identifier des tendances et des anomalies dans les données.

```python
fig, axes = plt.subplots(1, 3, figsize=(15, 4))

sns.histplot(data["RENDEMENT_T_HA"], bins=10, kde=True, ax=axes[0])
axes[0].set_title("Distribution du Rendement")

sns.histplot(data["PRECIPITATIONS_MM"], bins=10, kde=True, ax=axes[1])
axes[1].set_title("Distribution des Précipitations")

sns.histplot(data["TEMPERATURE_C"], bins=10, kde=True, ax=axes[2])
axes[2].set_title("Distribution de la Température")

plt.show()
```



J'ai également calculé la matrice de corrélation pour voir quelles variables sont les plus liées au rendement.

```python
plt.figure(figsize=(8, 6))


numeric_data = data.select_dtypes(include=[np.number])

# Matrice de corrélation
sns.heatmap(numeric_data.corr(), annot=True, cmap="coolwarm", fmt=".2f")

plt.title("Matrice de Corrélation")
plt.show()
```



Les valeurs de 1 sur la diagonale indiquent une corrélation parfaite de chaque variable avec elle-même. La matrice de corrélation indique que les variables ont des relations très faibles entre elles. Cela signifie qu'aucune des variables n'a un impact significatif sur une autre dans le jeu de données.

5. Test ANOVA (Impact du type de sol sur le rendement)

J'ai utilisé un test ANOVA pour voir si le type de sol a un effet significatif sur le rendement on a donc:
H0 : Le type de sol n'influence pas le rendement.
H1 : Le type de sol influence le rendement

```
# Convertir le type de sol
data["TYPE_SOL"] = data["TYPE_SOL"].astype("category")

# Modèle ANOVA
model = smf.ols("RENDEMENT_T_HA ~ TYPE_SOL", data=data).fit()
anova_table = sm.stats.anova_lm(model, typ=2)

print("Tableau ANOVA :")
print(anova_table)

# Vérifier la p-value
p_value = anova_table["PR(>F)"][0]
if p_value < 0.05:
    print("\nLe type de sol influence significativement le rendement (p =", round(p_value, 4), ")")
else:
    print("\nLe type de sol n'a pas d'effet significatif sur le rendement (p =", round(p_value, 4), ")")
```

```
Tableau ANOVA :
                sum_sq     df         F    PR(>F)
TYPE_SOL     17.900287    2.0  1.356052  0.258151
Residual   6580.348524  997.0       NaN       NaN

Le type de sol n'a pas d'effet significatif sur le rendement (p = 0.2582 )
```

On peut voir que le test donne une p-value de 0,258 ce qui est supérieur à 0,05 ca veut dire que le type de sol n'a pas d'effet significatif sur le rendement les différent sol ne semblent pas expliquer les variations de production dans le jeu de données. on ademets donc H0 comme hypothèse

6. Modélisation (Régression Linéaire)

J'ai ensuite construit un modèle de régression linéaire pour tenter de prédire le rendement en fonction des autres variables.

```
# Sélection des variables explicatives
X = data[["SURFACE_HA", "ENGRAIS_KG_HA", "PRECIPITATIONS_MM", "TEMPERATURE_C"]]
y = data["RENDEMENT_T_HA"]

# Séparation en données d'entraînement et de test
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

print("Données d'entraînement :", X_train.shape)
print("Données de test :", X_test.shape)
```

```
Coefficients : [ 0.02125329  0.00104516 -0.00538584  0.01639635]
Intercept : 7.32549529614522
```

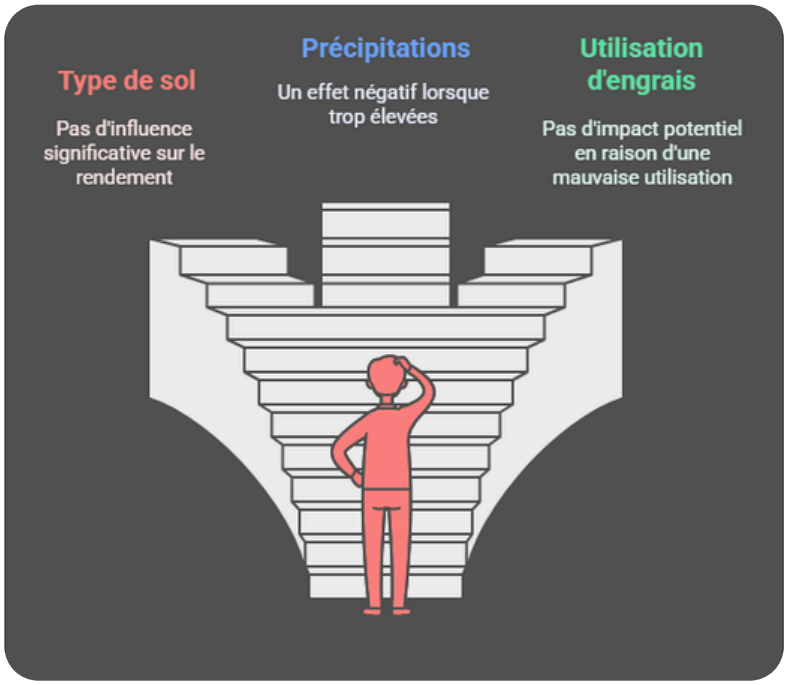
```
# Prédictions
y_pred = model.predict(X_test)

# Calcul des métriques d'évaluation
mae = mean_absolute_error(y_test, y_pred)
rmse = np.sqrt(mean_squared_error(y_test, y_pred))
r2 = r2_score(y_test, y_pred)

print("MAE :", mae)
print("RMSE :", rmse)
print("R² :", r2)
```

```
MAE : 2.0640792435396498
RMSE : 2.426341503870645
R² : 0.0017138549149776638
```

Les résultats montrent que les coefficients des variables sont très faibles , notamment celui de l'engrais, ce qui signifie qu'il n'a presque aucun impact sur le rendement.
L'évaluation du modèle donne un R^2 de 0,0017 ca veut dire que la régression linéaire ne permet pas de bien prédire le rendement.

7. Conclusion et recommandations

Les resultats montrent que le type de sol na pas d'influence significative sur le rendement et que les precipitation pourraient avoir un effet négatif lorsqu'elles sont trop élevées. L'engrais semble ne pas avoir d'impact ce qui peut être du a une mauvaise utilisation.



Pour améliorer le rendement il faudrait peut etre etudier d'autres variables et voir comment mieux gérer l'irrigation pour éviter un excès d'eau et analyser comment l'engrais est utilisé pour voir s'il est bien dosé.