**TD1 ATDN2 : Analyse complète d'un problème de Data Science**

**Etape 1 : Compréhension du problème**

**Variables disponibles :**

* SURFACE\_HA
* TYPE\_SOL
* ENGRAIS\_KG\_HA
* PRECIPITATIONS\_MM
* TEMPERATURE\_C
* RENDEMENT\_T\_HA

**Problème métier :**

Optimiser la production de maïs afin de maximiser le rendement et utiliser les ressources de manière plus efficace et économe.

**Variable cible :** RENDEMENT\_T\_HA.

**Variables explicatives :** SURFACE\_HA, TYPE\_SOL, ENGRAIS\_KG\_HA, PRECIPITATIONS\_MM et TEMPERATURE\_C.

**Etape 2 : Analyse statistique descriptive**

**2.1**

* **Moyenne** ≈ 7.38
* **Médiane** ≈ 7.35

**Interprétation** : Les valeurs proches suggèrent une distribution globalement équilibrer.

* **Mode rendement** ≈ 3.00

**Interprétation** : Ce chiffre est bien inférieur à la moyenne et à la médiane, ce qui indique qu’il existe une fréquence élevée de faibles rendements peut être due à des valeurs aberrante.

**2.2**

* **Ecart-type** ≈ 2.57
* **Variance** ≈ 6.60
* **Etendue** ≈ 9.00

**Interprétation** : Les mesures montrent que la dispersion des données est autour de la moyenne.

**2.3**

|  |
| --- |
|  |
| **Observation :** Sur l’histogramme on observe une pointe entre 8-9, le boxplot le confirme. Ce qui indique qu’environ la moitié des valeurs se situe autour de 7. Pas de valeurs aberrantes de ce cas. |

**2.4**

|  |  |
| --- | --- |
|  | **Observation :**  Faible corrélation entre les variables (pas de valeurs vraiment plus sombres qu’un autres). Pas de variable dominante. |

**Etape 3 : ANOVA**

* **F-value** ≈ 1.36
* **p-value** ≈ 0.258

**Interprétation :** La **p-value** supérieure à 0.05 donc on doit aussi prendre en compte l’hypothèse H0. Ce qui veut dire que dans notre problème le type de sol n’a pas d’influence significative sur le rendement.

**Etape 4 Modélisation et performances prédictives**  
*Modèles RegressionLineaire & SVR :*  
**• Métric Régression linéaire :**

|  |  |  |
| --- | --- | --- |
| **MAE** | **RMSE** | **R^2** |
| 2.06 | 2.43 | 0.0017 |

**Interprétation** : Le R^2 indique que le modèle explique très peu de la variabilité du rendement.

•**Métric SVR :**

|  |  |  |
| --- | --- | --- |
| **MAE** | **RMSE** | **R^2** |
| 2.05 | 2.43 | -0.0025 |

Nous avons des métriques plutôt similaires. Cependant pour le SVR nous avons un R^2 (-0.0025) légèrement négatif, donc inférieure à celle du modèle de RégressionLineaire.

**Modèle performant** : *Régression linéaire.*

**Coefficients du modèle** :

([0.02125, 0.001045, -0.005386, 0.016396] pour [SURFACE\_HA, ENGRAIS\_KG\_HA, PRECIPITATIONS\_MM, TEMPERATURE\_C]).

Les faibles résultats avec un R^2 proche de zéro, peut suggérer que ces variables n’impactent pas la variance du rendement ou du moins pas beaucoup.

**Etape 5**

**Analysez l'importance des variables.**

Avec le coefficient du modèle on observe que SURFACE\_HA et TEMPERATURE\_C sont plus importante que les autres variables avec un coefficient plus élever.

A la différence de PRECIPITATIONS\_MM qui est négatif, donc peut être effet négatif sur le rendement.

**Proposez des recommandations concrètes pour augmenter le rendement (ex : ajuster l’engrais, choisir un type de sol particulier, etc.).**

* Comme vue précédemment on observe une petite importance vis à vis de :
  + **SURFACE\_HA et TEMPERATURE\_C** en termes de rendement. Peut-être approfondir ces paramètres.
* Avec **ANOVA** on a également vue que H0 et H1 sont viable :
  + H0 : Le **type de sol** n’influence pas le rendement.
  + H1 : Le **type de sol** influence le rendement.

Donc peut être selon chaque type de sol tester un type spécifique d’engrais puisque le sol en lui-même n’impacte pas le rendement (ex un engrais que pour les sols sableux).

* **Conditions climatiques :** 
  + On observe également que les conditions climatiques peuvent impacter négativement le rendement. Peut-être faire attentions aux précipitations et/ou prévoir un arrosage adapter.

**Identifiez les limites du modèle et proposez des pistes d’amélioration.**

**Faible capacité prédictive :** Nous avons des R^2 très faible donc une capacité de nos modèles à prédire très peu fiable.

**Modèle Linéaire :** Dans notre cas l’approche linaire n’est peut-être pas forcément la bonne. Cependant elle est la plus simple à mettre en place afin d’avoir un premier avis.

**Quelles décisions la ferme pourrait-elle prendre pour optimiser sa production ?**

**Optimisation de l’engrais :**  
Mener des essais sur le terrain pour déterminer le dosage optimal en engrais selon le sol.

**Adapter l’irrigation :** Adapter l’arrosage selon le temps les saisons. Peut être prévoir des infrastructures pour les plantes (ex serre)