Ml OIVM / TD ATDN

Analyse complète d'un problème de Data Science

Ce travaill vous guide sur tout le cycle d'un projet de Data Science. Les objectifs pédagogiques du TD sont les suivants :

- Appliquer le cycle complet de Data Science sur un jeu de données
- Maîtriser l'analyse exploratoire des données
- Réaliser une analyse statistique descriptive
- Construire et évaluer un modèle prédictif
- Interpréter les résultats et formuler des recommandations

Suppposons que vous travaillez pour une ferme qui cultive du maïs et souhaite **prédire le rendement** (en tonnes par hectare) en fonction de plusieurs facteurs : **type de sol, quantité d'engrais, précipitations, température moyenne, et surface cultivée**. L'objectif est d'optimiser les ressources pour maximiser le rendement. Vous avez à votre disposition un jeu de données rendement_mais.csv avec les colonnes 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

Jeu de données : rendement_mais.csv

SURFACE_HA	TYPE_SOL	ENGRAIS_KG/HA	PRECIPITATIONS_MM	TEMPERATURE_C	RENDEMENT_T/HA
5	Argileux	120	80	22	8.5
3	Sableux	90	65	25	5.2
4	Limoneux	110	75	23	7.3
6	Argileux	130	85	21	9.1
2	Sableux	80	60	26	4.8

Etape 1 : Compréhension du problème

Décrivez les variables disponibles.

Formulez clairement le problème métier.

Identifiez la variable cible et les variables explicatives.

Quelle est la problématique centrale pour la ferme ?

Etape 2 : Analyse statistique descriptive

2.1 Mesures de tendance centrale

Calculez la moyenne, médiane, et mode du rendement.

2.2 Mesures de dispersion

Calculez l'écart-type, variance, et étendue du rendement.

2.3 Visualisation des données

Créez des histogrammes pour le rendement, les précipitations, et la température.

Affichez des boxplots pour identifier d'éventuels outliers.

2.4 Corrélations

Calculez la matrice de corrélation entre les variables numériques.

Affichez une heatmap pour visualiser les corrélations.

Quelles variables semblent avoir le plus d'impact sur le rendement ?

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.

Interprétez la p-value obtenue.

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

Etape 4 : Modélisation

4.1 Séparation des données

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

4.2 Création du modèle

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

4.3 Évaluation du modèle

Calculez les métriques : MAE, RMSE, et R^2 de ces modèles.

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

Etape 5 : Interprétation et recommandations

Analysez l'importance des variables.

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

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

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