Présentation des dataset

Dataset 1 : recommendation\_culture.csv

**Contexte :**

L'agriculture de précision est à la mode de nos jours. Il aide les agriculteurs à prendre une décision éclairée sur leur stratégie agricole.  je vous présente un ensemble de données qui permettrait aux utilisateurs de construire un modèle prédictif pour recommander les cultures les plus appropriées à cultiver dans une ferme particulière en fonction de divers paramètres.

Cet ensemble de données a été construite grâce aux précipitations, le climat et les engrais disponibles dans le sol en Inde.

**Champs de données :**

|  |  |  |
| --- | --- | --- |
| Libellé | Description | Type |
| N | Rapport de la teneur en azote dans le sol. | Decimal |
| P | Rapport de la teneur en phosphore dans le sol. | Decimal |
| K | Rapport de la teneur en potassium dans le sol. | Decimal |
| Temperature | Température en degré Celsius. | Reel |
| Humidity | Humidité relative en %. | Reel |
| Ph | Valeur du ph du sol. | Reel |
| Rainfall | Précipitations en mm. | Reel |
| Label ( Variable cible) | Variable cible, indiquant le type de culture correspondant aux autres caractéristiques. | Categoriel  (String) |

Cet ensemble de donné est composé de 2200 lignes et 8 colonnes.

**Objectif :**

L’un des objectifs avec cet ensemble de données est d'appliquer quelques algorithmes de machine learning à ce jeu de données afin de faire des recommandations de culture basées sur les différents paramètres du sol et de l'environnement, Nous évaluerons également les performances de notre modèle et discuterons des implications de nos résultats.

Ci-joint, vous trouverez la dataset, ainsi qu’un code que j’ai réalisé afin d’explorer le dataset en suivant les étapes suivantes :

1. Importation des bibliothèques de manipulation et le dataset pour notre étude.
2. Gestion des valeurs manquantes : Vérifiez s'il y a des valeurs manquantes dans le dataset.
3. Traduction des différentes cultures en francais.
4. Analyse descriptive des données et de la variable cible (Label): J’ai exploré les statistiques descriptives de chaque variable (N, P, K, Temperature, Humidity, Ph, Rainfall, Label), tels que la moyenne, l'écart-type, les valeurs minimales et maximales, les quantiles. Cela a permis d'obtenir un aperçu global des données.
5. Visualisation des données :J’ai utilisé des graphiques (histogrammes, diagrammes des barres.) pour visualiser la distribution des valeurs pour chaque variable. Cela m’a aider à identifier des tendances, des valeurs aberrantes.
6. Corrélation entre les variables : Analyse de la corrélation entre les différentes caractéristiques (N, P, K, Temperature, Humidity, Ph, Rainfall) pour comprendre comment elles sont liées entre elles en utilisant la fonction heatmap.
7. Détection des valeurs aberrantes : Identifions les valeurs aberrantes qui pourraient fausser notre analyse. J’ai utilisé une méthodes statistiques la règles des 1.5\*IQR et des graphiques de boîtes pour repérer les valeurs extrêmes.
8. Exploration des tendances saisonnières : Avez les données sur la température, l'humidité et les précipitations que nous avons, j’ai recherché des tendances saisonnières pour comprendre comment ces facteurs affectent la croissance des cultures.
9. Segmentation des cultures : Exploration des caractéristiques de chaque catégorie pour comprendre les différences et les similitudes entre elles.

Ainsi pour la suite de ce travail, nous pourrions identifier des cluster afin de regrouper les cultures similaires en fonction des caractéristique du sol et du climat, et procéder à la preparation des données et à la création d’un modèle de prédiction en évaluant plusieurs algorithmes et garder la meilleure.