Tab 1

### **Pipeline**

1. **Preprocess Data** (Handle missing values, normalize features).
2. **Feature Selection** (Use relevant numerical features).
3. **Clustering using K-Means** (Determine optimal **K** using the **Elbow Method**).
4. **Assign Cluster Labels** (Check alignment with actual genotypes).
5. **Train a Classifier** (Predict **genotype** using models like **Random Forest, SVM, XGBoost**).
6. **Compare Clustering vs. Classification Results**.

**In Detail:**

#### **1. Data Preprocessing**

* Load the dataset.
* Handle missing values (if any) using imputation techniques.
* Encode categorical variables (e.g., the *Parameters* column) using one-hot encoding or label encoding.
* Normalize or standardize numerical features (e.g., plant height, protein percentage, seed yield) to ensure consistency.
* Split the dataset into **features (X) and target labels (y)** for classification.

#### **2. Clustering (Unsupervised Learning)**

* **Feature Selection**: Choose relevant numerical features for clustering.
* **Elbow Method for Optimal Clusters**:  
  + Use **K-Means clustering** with different values of **K (number of clusters)**.
  + Calculate the **inertia (within-cluster sum of squares)** for each K.
  + Plot the **Elbow Curve** and determine the optimal K (where the curve bends).
* **Cluster Assignment**:  
  + Apply **K-Means** clustering using the optimal K.
  + Visualize clusters using **PCA (Principal Component Analysis) try to plot 3D PCA Projection plot** and **t-SNE**.

#### **3. Classification (Supervised Learning)**

* **Train-Test Split**:  
  + Split the dataset into training (80%) and testing (20%) sets.
* **Model Selection**:  
  + Try different classifiers like **Random Forest, SVM, XGBoost, or Neural Networks**.
  + Tune hyperparameters using **GridSearchCV** or **RandomizedSearchCV**.
* **Model Evaluation**:  
  + Use **accuracy, precision, recall, F1-score** for performance analysis.
  + Generate **confusion matrix** and **ROC-AUC curves** for deeper insights.

#### **4. Comparison of Clustering and Classification Results**

* Compare how clustering results align with actual classes in classification.
* Identify misclassified points and analyze cluster similarities.

**Target Variable:**

Dataset description does not explicitly mention a target variable for classification. Based on the given attributes, possible target variable could be:

1. **Seed Yield** – If the goal is to predict soybean productivity.
2. **Genotype (G)** – If we are classifying different soybean genotypes.
3. **Chlorophyll Content / Protein Percentage** – If predicting plant health or nutritional quality.
4. **Experimental Conditions (Parameters column - G, C, S)** – If classifying based on experimental treatments.

**Recommendation for Classification:**

* **Genotype (G)**: Since the dataset consists of **six different genotypes**, we can classify each sample based on its genotype.
* If genotype is not directly available as a column, we may extract it from the **Parameters** column (which encodes G, C, S).

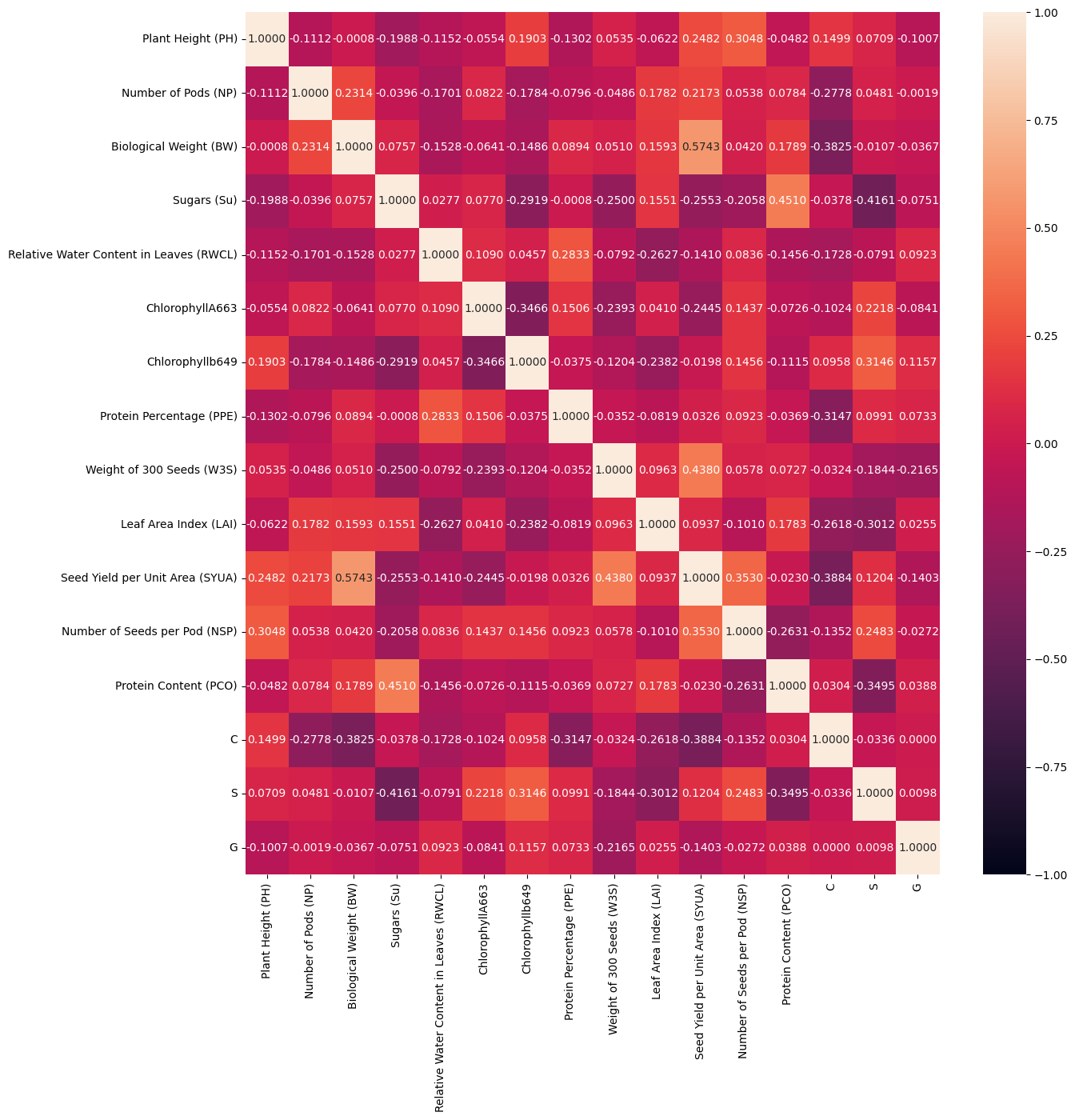
### 

### 

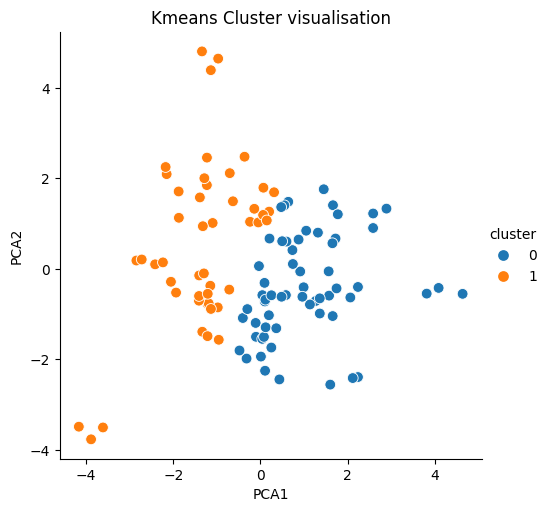
### **Clustering Objective**

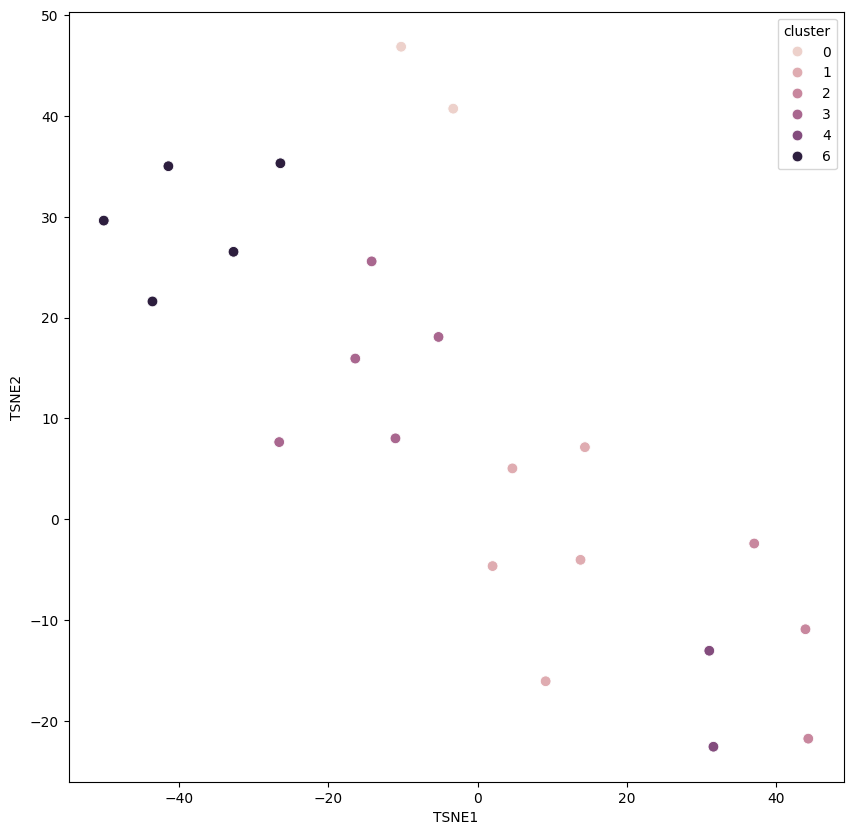
* We will apply **K-Means Clustering** to group similar soybean growth conditions based on features like **plant height, chlorophyll content, protein percentage, seed yield, etc.**
* The goal is to find natural patterns and compare them with the actual genotype classification.

Preprocessing & EDA









Tab 3

<https://scikit-learn.org/stable/modules/clustering.html>

<https://scikit-learn.org/stable/modules/generated/sklearn.manifold.TSNE.html>

<https://medium.com/data-science/t-sne-clearly-explained-d84c537f53a>

<https://scikit-learn.org/stable/modules/generated/sklearn.metrics.davies_bouldin_score.html>

<https://medium.com/@a.cervantes2012/interpreting-and-validating-clustering-results-with-k-means-e98227183a4d>

<https://scikit-learn.org/stable/modules/generated/sklearn.metrics.calinski_harabasz_score.html>

<https://medium.com/@haataa/how-to-measure-clustering-performances-when-there-are-no-ground-truth-db027e9a871c>

<https://medium.com/biased-algorithms/silhouette-score-d85235e7638b>

<https://medium.com/@hazallgultekin/what-is-silhouette-score-f428fb39bf9a>