Advanced Clustering and Classification Models for Soybean Agricultural Data

1 Clustering Approaches

Given the presence of categorical variables such as genotype (G), salicylic acid levels (C), and water stress (S), we apply unsupervised clustering techniques to group similar soybean plants based on their characteristics.

1.1 A. K-Means Clustering with Hybrid Distance Measures

Instead of using the conventional Euclidean distance, we employ a custom hybrid distance metric that combines:

- **Gower's Distance**: Ideal for mixed numerical and categorical data.
- **Mahalanobis Distance**: Used to handle correlated numerical features effectively.

To determine the optimal number of clusters, we perform:

- **Silhouette Analysis**,
- **Elbow Method**.

1.2 B. Gaussian Mixture Models (GMM)

We apply soft clustering to model the variability in seed yield and other physiological traits. The model estimates the probabilistic distributions of plant traits, such as seed yield and protein percentage, for each cluster.

1.3 C. Hierarchical Clustering for Genotype Classification

Using Agglomerative Clustering, we analyze the relationships between genotypes and environmental stress factors. A dendrogram is generated to visualize genotype similarities, and clusters can be identified based on specific traits.

2 Classification Approaches

We turn to supervised classification models to predict outcomes such as Soybean Yield or Genotype Category.

2.1 A. Multi-Class Classification with Deep Learning

We train a Neural Network (MLP or CNN for tabular data) to predict the soybean genotype or yield category. To improve generalization and avoid overfitting, we incorporate

- **Dropout Regularization**,
- **Batch Normalization**.

2.2 B. Tree-Based Ensemble Learning

We apply several tree-based methods for classification:

- **Random Forest**: Robust to missing data and offers interpretability.
- **XGBoost**: Handles complex interactions effectively and is optimal for tabular data.
- **LightGBM**: Efficient and scalable for large datasets.

2.3 C. Explainable AI (XAI) for Agricultural Insights

We enhance model interpretability using SHAP (SHapley Additive Explanations), which helps determine the importance of each feature in predicting soybean yield. This allows us to identify the key factors contributing most to yield variability, such as salicylic acid concentration, water stress, and chlorophyll levels.

3 Advanced Statistical and Optimization Techniques

To further enhance, we incorporate advanced techniques:

- **Bayesian Optimization**: Efficient hyperparameter tuning for classification models.
- **Causal Discovery for Yield Prediction**: Use of Bayesian Networks to model how variables such as salicylic acid and water stress affect yield.