ML Classification Pipeline

1. DATA INPUT
- Phyloseq Object (Microbiome Data)
- Data Formatting (Features + Target Variable)

2. DATA PARTITIONING
- Hold-out Split (90% / 10%)
  - 90% → Training Set
  - 10% → External Validation Set

3. NESTED CROSS-VALIDATION
- Outer Loop (K=10 folds) — Outer Fold i (repeated 10×)
- Train Partition (9/10 of data)
- Test Partition (1/10 of data)
- Feature Selection (Boruta Algorithm)
- Selected Features
- Class Imbalance Check
  - Ratio < 0.2 → SMOTE Oversampling
  - Ratio ≥ 0.2 → No Resampling
- Inner Loop (5×5 Repeated CV)
- Random Forest Hyperparameter Tuning (*mtry*)
- Best Model_i
- Evaluate on Test Partition
- Performance Metrics (AUC, MCC, F1)

4. FINAL MODEL SELECTION
- Aggregate CV Results (10 fold performances)
- Select Best Hyperparameters
  - Criterion 1: Mode hyperparameters + Consensus features
  - Criterion 2: Single best fold hyperparameters
  - Criterion 3: Best average performance across outer folds

5. FINAL MODEL TRAINING
- Train Model 1 (Consensus)
- Train Model 2 (Best Fold)
- Train Model 3 (Avg Performance)

6. SHAP FEATURE
- SHAP Analysis Model 1
- SHAP Analysis Model 2
- SHAP Analysis Model 3
- Identify Robust Biomarkers

7. EXTERNAL VALIDATION
- Predict on Hold-out Set
- Final Performance Estimation