**✅ CultSelect Project Roadmap**

**🔹 Dataset Structure:**

* Binary protein presence (1/0) as features
* Disease score (numeric) as target
* 12 datasets (1 per cultivar) — develop for one, replicate for others
* Example cultivar: lancer

**🧪 Step 1: Model Testing**

Train & evaluate the following models:

* Linear Regression
* Lasso Regression
* Decision Tree
* Random Forest
* Gradient Boosting
* XGBoost

Use:

* 70% Training
* 20% Testing
* 10% Final Evaluation
* Cross-validation with K-folds (k ∈ {3, 5, 7, 9})

**🛠 Step 2: Hyperparameter Tuning**

Compare:

* GridSearchCV
* Bayesian Optimization (Optuna)
* Genetic Algorithm (GA from code\_1)

**📉 Step 3: Overfitting/Underfitting Detection**

* Use learning curves
* Generate graphs and/or summary tables

**📊 Step 4: Final Model Evaluation**

* Evaluate final model(s) on **held-out 10%** set
* Report **MAE, MSE, R²**

**🧬 Step 5: Feature Importance**

* For best GA-tuned model (or final selected ones):
  + Rank all features by importance
  + Export as table

**📦 Step 6: CCDM Inference Folder**

* Save top 2–3 models as .pkl
* Organize prediction code for new data
* Base it on CCDM folder structure