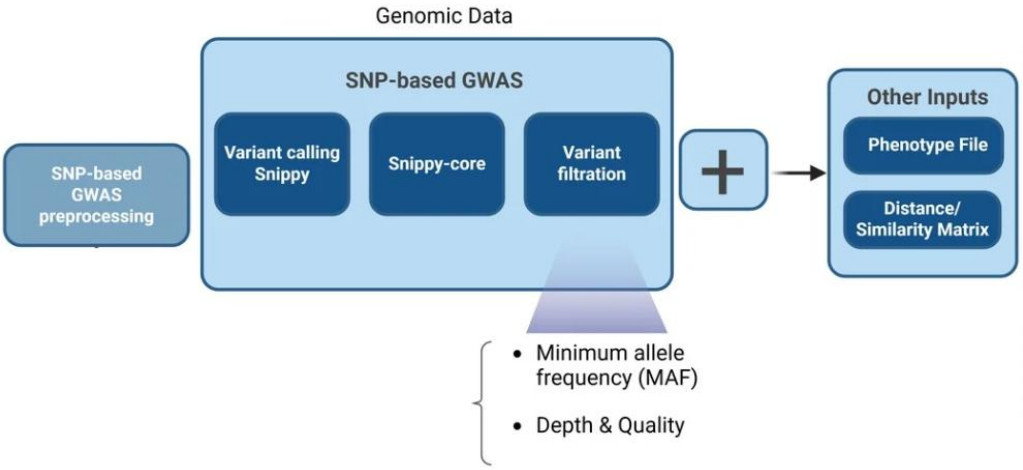
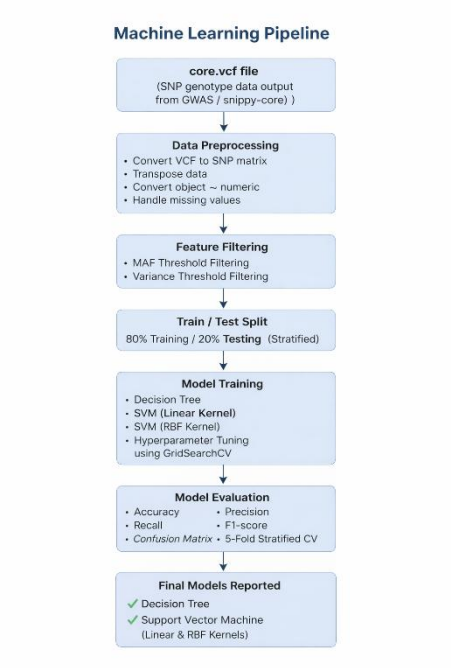
Machine Learning for Biofilm Prediction from Bacterial SNP Datasets

Documentation

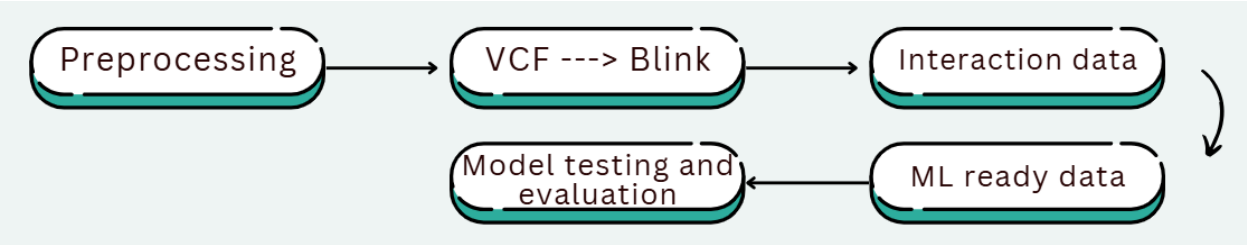
Project Pipeline:

Pipeline for GWAS running and machine learning





As for the pipeline used for Epistatic interactions:



Algorithms used:

GWAS:

- Stand alone; **Snippy**: SNP calling by aligning each isolate's reads to a reference genome, **snippy-core**: merges isolates and extracts shared variable genomic positions → generates
- With Epistasis: use snippy and snippy-core then use Plink epistasis

Machine Learning:

- **Decision Tree classifier** (supervised binary classification)
- **Support Vector Machine (SVM)**:
  - **Linear kernel**
  - **RBF kernel** (reported as best-performing in your results)
- **GridSearchCV** (hyperparameter tuning)
- **Stratified train/test split (80/20)** + **Stratified cross-validation (5-fold)**
- **SMOTE** (optional, for imbalanced classes)
- Evaluation metrics: **Accuracy, Precision, Recall, F1-score, Confusion Matrix**

Package Dependencies:

numpy==2.1.1

pandas==2.2.3

scikit-learn==1.5.2

imbalanced-learn==0.12.4

joblib==1.4.2

matplotlib==3.9.2

cyvcf2==0.31.1

pyyaml==6.0.2

Conda==25.11.1

Snippy==4.3.6

Plink==2

Python==3.14,3.10