Machine Learning to predict Tissue Type using RNASeq data

Sonali Srijan

Problem Definition

- Using RNASeq data, predict the tissue/organ corresponding to that sample.
- We intend to do this by training an ML model using ortholog genes data from only one species (Arabidopsis).
- Multi-class classification problem with eight tissue classes: Flower, Leaf, Root, Seed, Seedling, Shoot, Silique, Stem

Previous Work

- Two ML models developed: SVM and KNN
- Used log normalized RNASeq data from Arabidopsis th.
- KNN model was found to perform better (5 nearest neighbours used)
- Based on RNASeq profile, the model predicts which tissue the sample belongs to (Multiclass Classification)

Building ML Models with Ortholog Data

- Central Assumption

Most ortholog genes common across the plant species perform similar functions, and hence have similar expression profiles. Orthologs are genes in different species that have evolved through speciation events only.

Thus, training the tissue classifier model using only ortholog gene expression data from Arabidopsis, it is plausible to predict tissue in other plant species as well.

Single Copy Ortholog Genes

- In the Orthogroups.txt file, containing info from multiple plant species, we selected those Orthogroups that only had single-copy orthologs in each of Arabidopsis, Tomato, Maize, Rice. Total number of such orthogroups were found to be 2845 (=2898-53).
- Next, using these orthogroups, the genes corresponding to each orthogroup were extracted for each species.

Building an Ensemble

- Each ensemble consists of SVM, XGB, KNN, Linear models. Two Ensemble models were initially built: Hard-voted and Soft-voted.
- It was found that Soft-voted models performed better for almost all classes. To decide which ML model gets what weight in each class, we tested a range of weights.

Building an Ensemble

SVM:
Accurcay: 0.978
Macro_prec: 0.872

	Flower	Leaf	Root	Seed	Seedling	Shoot	Silique	Stem
Precision	0.8817	0.9879	0.9931	0.9611	0.9737	0.5645	1	0.6101
Recall	0.6721	0.9905	0.9906	0.9611	0.9628	0.7608	0.6	0.9729

XGB:
Accurcay: 0.977
Macro_prec: 0.810

	Flower	Leaf	Root	Seed	Seedling	Shoot	Silique	Stem
Precision	0.8817	0.9875	0.9889	0.9583	0.9834	0.4032	0.6666	0.6101
Recall	0.5774	0.9942	0.9957	0.9663	0.9479	0.8333	0.8	0.9729

KNN:
Accurcay: 0.972
Macro_prec: 0.861

	Flower	Leaf	Root	Seed	Seedling	Shoot	Silique	Stem
Precision	0.9247	0.9825	0.9872	0.95	0.9669	0.5322	1	0.54237
Recall	0.6013	0.9907	0.9864	0.9447	0.9498	0.7173	0.6666	0.9142

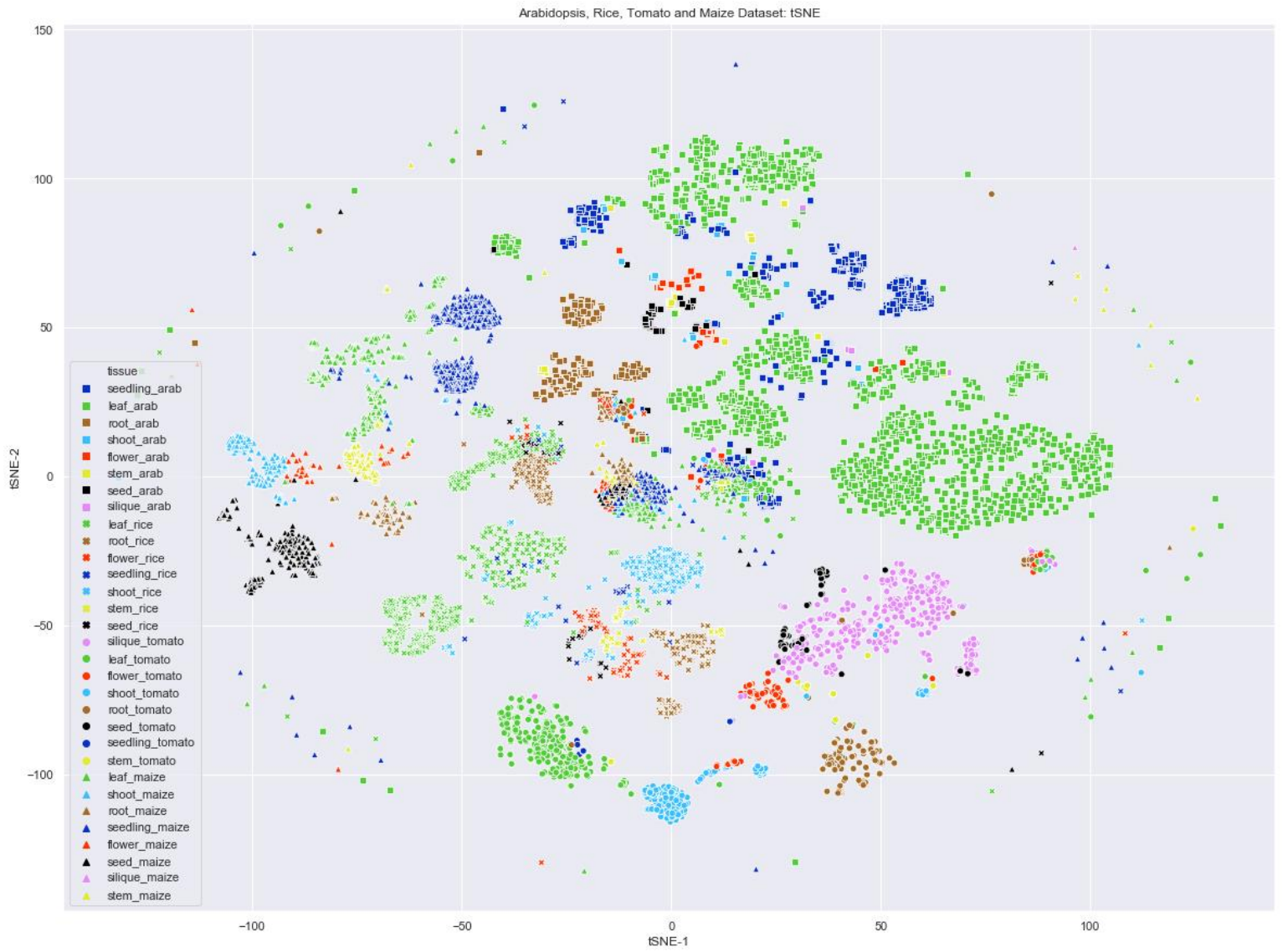
Linear:
Accurcay: 0.953
Macro_prec: 0.817

	Flower	Leaf	Root	Seed	Seedling	Shoot	Silique	Stem
Precision	0.9032	0.9752	0.9353	0.9	0.9512	0.2741	1	0.5932
Recall	0.3206	0.9820	0.9954	0.9473	0.9441	0.6071	1	0.9459

Best Precision models:
Flower: KNN
Leaf: SVM
Root: SVM
Seed: SVM
Seedling: XGB
Shoot: SVM
Silique: Linear
Stem: SVM, XGB

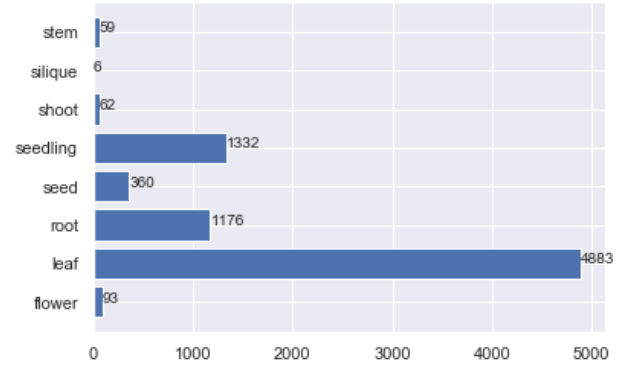
Deciding class-wise weights:
Flower: 0.5*KNN + 0.2*SVM + 0.2*Linear + 0.1*XGB
Leaf: 0.5*SVM + 0.2*XGB + 0.2*KNN + 0.1*Linear
Root: 0.5*SVM + 0.3*XGB + 0.2*KNN
Seed: 0.45*SVM + 0.35*XGB + 0.1*KNN + 0.1*Linear
Seedling: 0.5*XGB + 0.3*SVM + 0.2*KNN
Shoot: 0.4*SVM + 0.4*KNN + 0.1*XGB + 0.1*Linear
Silique: 0.7*Linear + 0.1*SVM + 0.1*KNN + 0.1*XGB
Stem: 0.35*SVM + 0.35*XGB + 0.2*Linear + 0.1*KNN

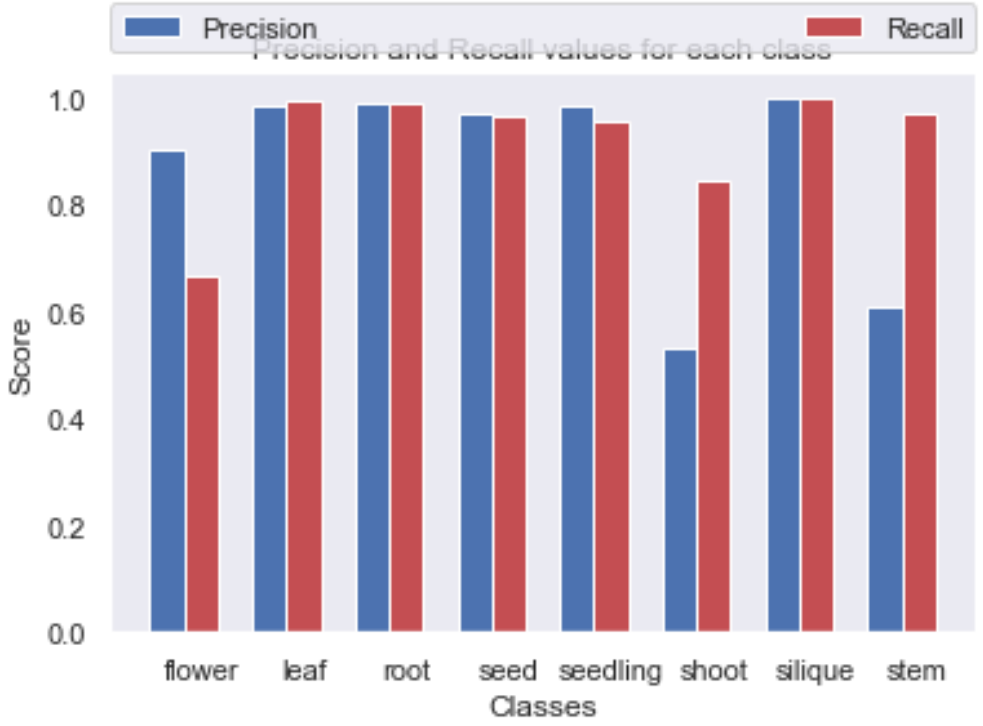
tSNE plots: Visualizing RNASeq Data

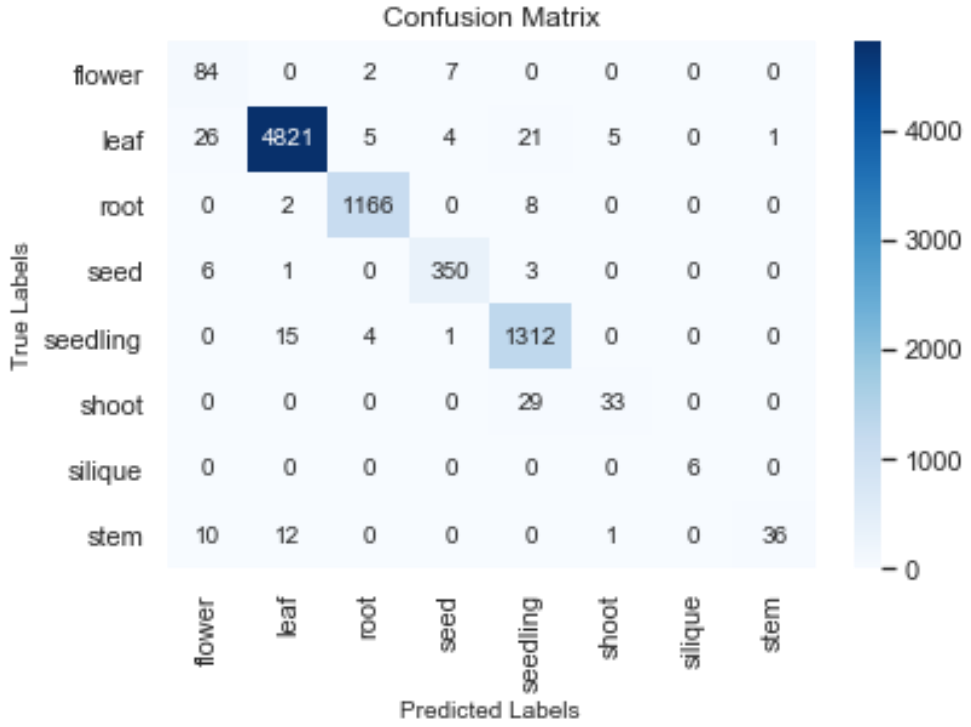


Soft Ensemble Model: Results for Arabidopsis Test Data

Accuracy is: 0.98
Macro precision is: 0.873

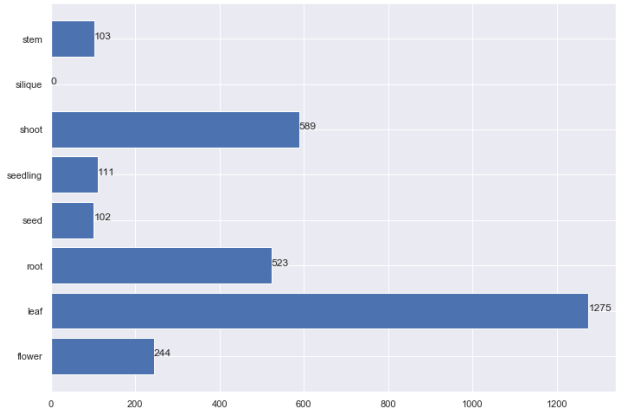


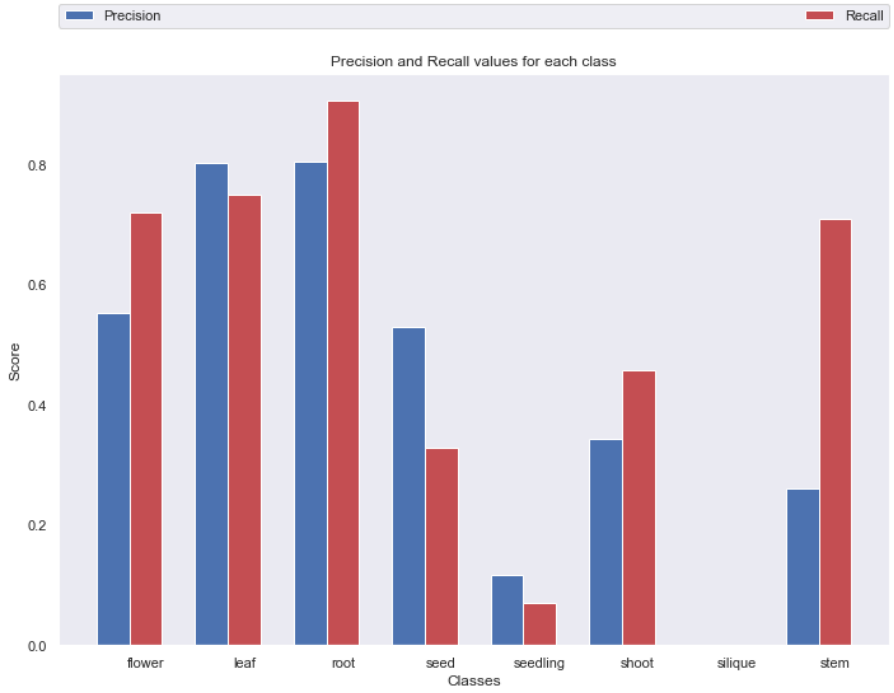


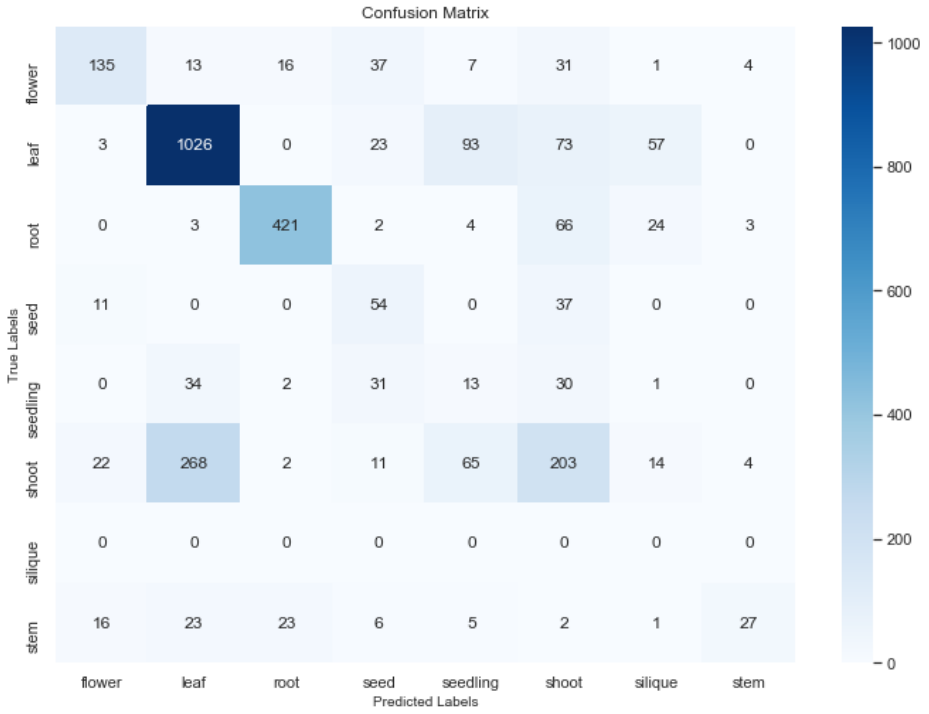


Soft Ensemble Model: Results for Rice

Accuracy is: 0.638
Macro precision is: 0.427

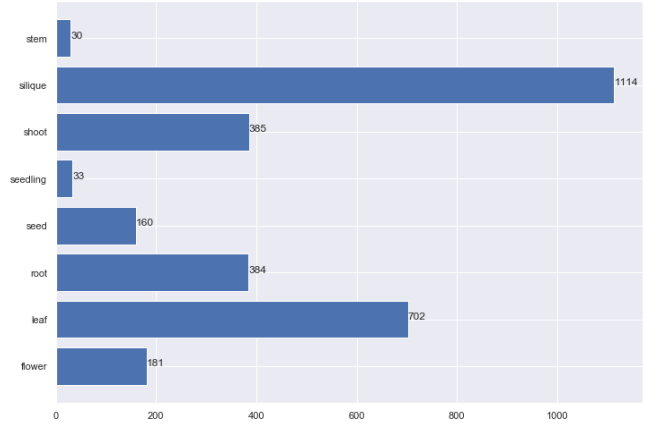


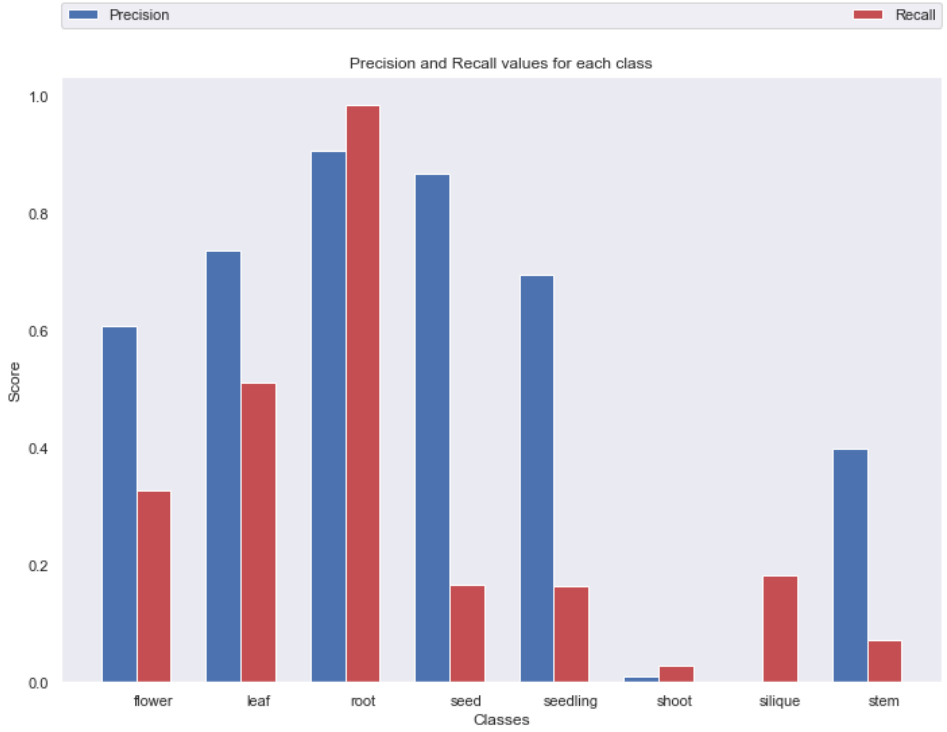


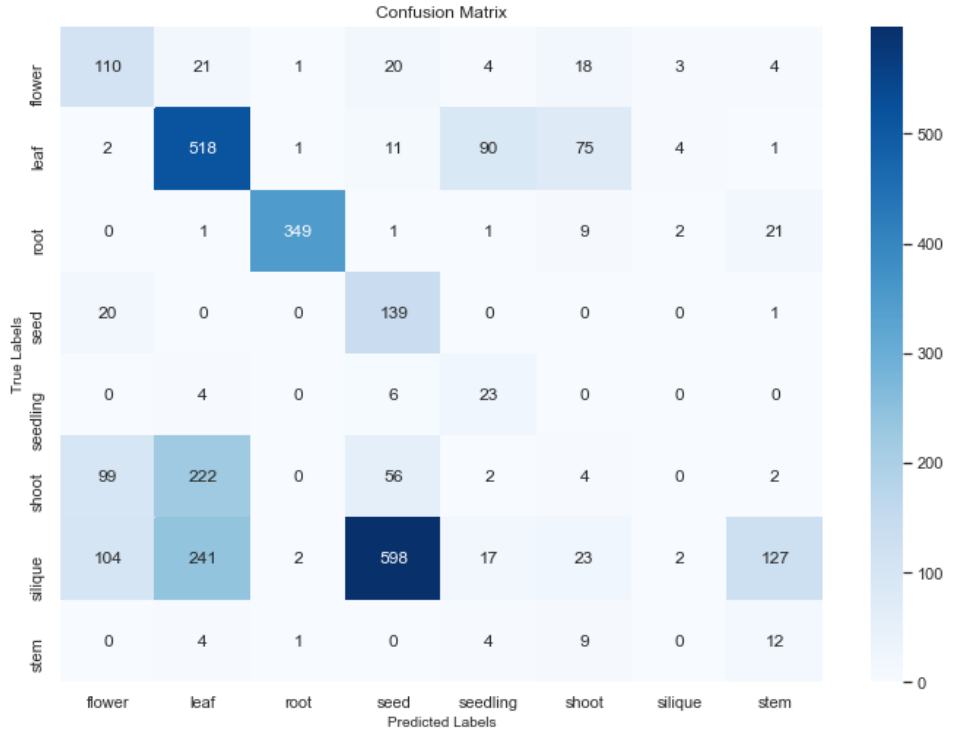


Soft Ensemble Model: Results for Tomato

Accuracy is: 0.387
Macro precision is: 0.529

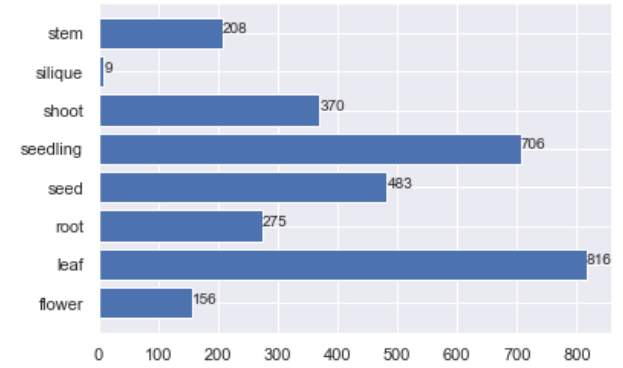


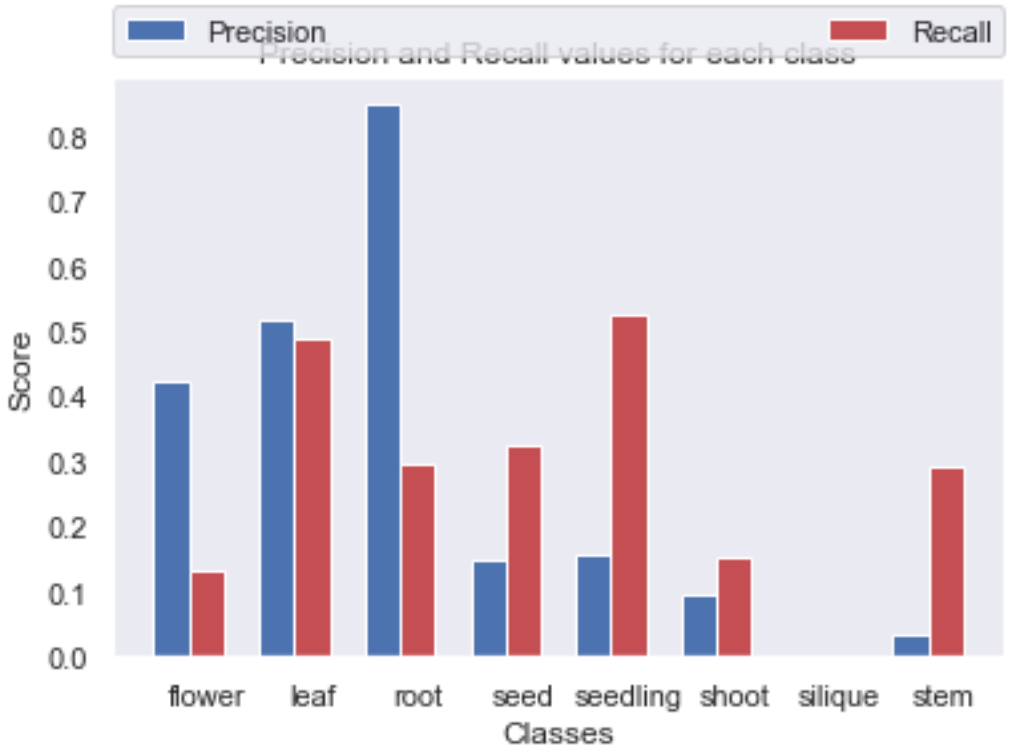


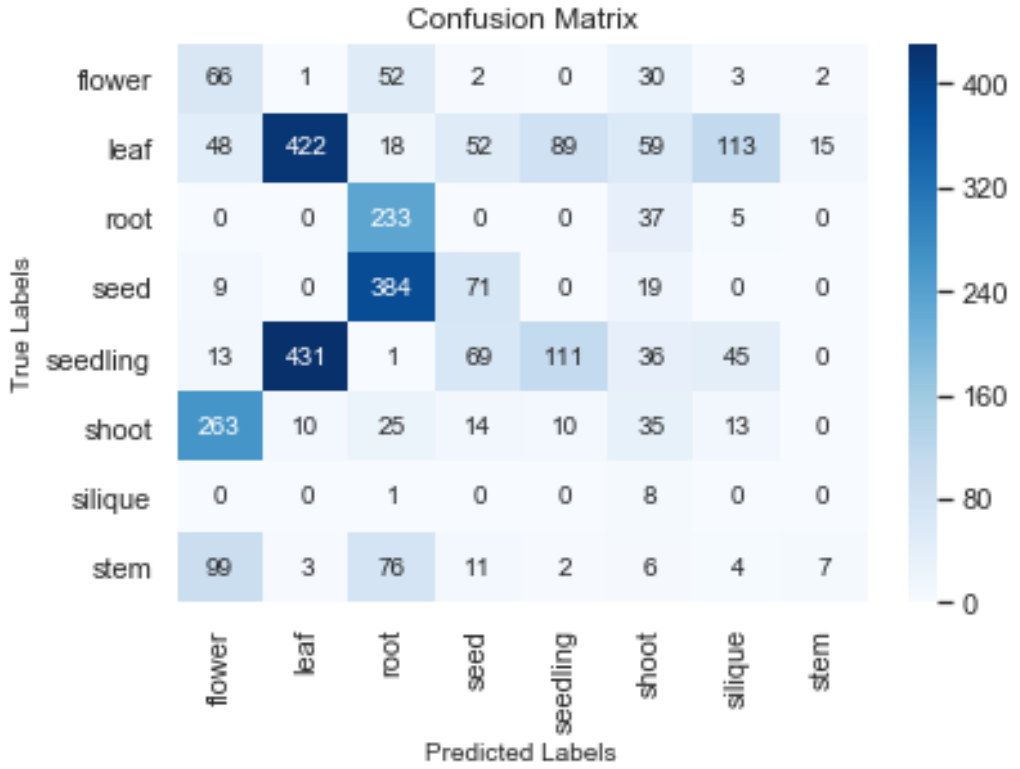


Soft Ensemble Model: Results for Maize

Accuracy is: 0.313
Macro precision is: 0.277







End