Given location, crop quality, and weather data, which crops should a farmer plant that would be most resilient to disease?

We are building a model that is predicting which general crop type should be planted. There are four general crop types: fruits, vegetables, legumes and seeds, and grasses. We call this the **general crop model.**

For the general crop model:

* Dataset:
  + Features = percent disease, wellness condition (i.e a combination of percent disease and crop condition), weather conditions, and location.
  + Y = the general crop type.
  + Addressing imbalanced classes: the class legumes and seeds and class grasses are less represented in the dataset. Realistically, we want farmers to plant a variety of crops, so we want to avoid having the model favoring one type of crop over the other. Therefore, we implemented class balance weights.
* Model selection and evaluation
  + We are implementing a multi-class classification model. We will be choosing between XGBoost, random forest, one-vs-rest, logistic regression, k-nearest neighbors, and support vector machines.
  + Criteria for a good model: We also wish for the farmers to plant a variety of crops, so we will choose the model that has the highest accuracy, and also recommend a good mix of crops.
  + We used AUC, precision, recall, F1 score, the confusion matrix, and 5-fold cross validation accuracy score to evaluate each of the models.
* Hyperparameter tuning
* Result of model evaluation
  + XGBoost:
* Final model
  + We decided that XGBoost is the best model, given that it has the highest cross validation score and recommends a more diverse set of crops.

**Clean up documentation**

**Scoring system description (Adam)**

1. What is the scoring system?

Predict best crop on location and region in terms of percent disease and crop condition.

1. Why do we need a scoring system?

Based on the model, we don’t know if the crop is good or not. We are potentially predicting crops that have high percent disease and condition.

Trying to inform the model.

1. Make an example where you use the scoring system (like a mini-tutorial on how to use the scoring system). Scoring system
2. Document the functions.
3. Integrating with the rest of the notebook.

**Model-building description**

1. Provide feedback on (i.e is this an accurate question that reflects the objective of your model): “Given location, crop quality, and weather data, which crops should a farmer plant that would be most resilient to disease?” (**Lauren and Chase**)
2. What is your y\_actual (i.e what exactly are you trying to predict)? (**Lauren and Chase**)
   1. Top 3 specific crops
   2. General crop type
   3. Using k-fold validation
   4. Condition → numeric value; wellness metric = condition \* percent disease (flipped)
   5. Model evaluation: more accurate, less variety due to imbalanced dataset → not realistic for farmers.
3. Analyze confusion matrix (**Lauren**), for general crop type
4. How did we decide to go with the XGBoost (**Lauren**)
   1. XGBoost, random forest, one-vs-rest, logistic regression ← good for multinomial class distribution.
   2. Ensembling method, highest validation accuracy.
5. Different methods we use to tune the hyperparameters of the XGBoost (**Chase**)

Remove, or comment it out.

Use the default parameters.

1. Results of the hyperparameter tuning (**Chase**)
2. How are we evaluating the model (**Lauren and Chase**)
   1. 5-fold cross validation
   2. Precision, recall, f1, accuracy, confusion matrices.
   3. Still gives a variety of results (XGBoost still predicts crops that appear less frequently).
3. Final model: what type, what hyperparameters, and what’s the result? (**Lauren and Chase**)
   1. Class balance weights ← XGBoost focus on underrepresented classes.
   2. XGBoost
   3. Hyperparameters (Chase)

**Interactive packaged product (Eva and Katie)**

Team lead: Elda Pere

Lauren Faulds

Chase Elements

Barnett (Adam) Yang

Kathryn (Kattie) Byers

Eva Sidlo

Kelly Trinh

**Who are we?**

We are Berkeley undergraduates working with ViviendasLeon, a nonprofit dedicated to eliminating rural poverty in Nicaragua and Guatemala.

**Questions to address:**

Given location, soil, and weather data, which crops should a farmer plant that would be most resilient to disease?

**Dataset description:**

Data source: ViviendasLeon

Each row represents a visit from ViviendasLeon to a family. It contains information on crop disease percentage, crop conditions, and any recommendations made.

**Objectives:**

Clean the data to account for missing value, inconsistent names, translate Spanish to English, and scrap weather data to supplement the dataset.

Perform exploratory data analysis to find trends between crop type, effectiveness of recommendations, and geographical area, and disease percentage.

Build a classifier to predict which crop is most disease-resilient given specific weather and geographical information.