Modern Datasets

Balancing and Choosing Dataset

Original Modern Dataset

Lake | Peat | Soil

Downsample soil dataset to n=750

Use SMOTE to upsample Lake and Peat

Lake
Original = 591 +
SMOTE n=159 9 750

Peat
Original n=532 +
SMOTE n=218 = 750

Split datasets into 60:20:20
training (60%), testing (20%), validation (20%)

Test accuracy of the original and SMOTE datasets as classification

Discard low accuracy model

Select highest accuracy model

Tuning and Calibrating Models

Chosen dataset

K-fold test for hypertuning models

Test on unseen data

Calibrate probabilities with a sigmoid or isotonic regression
test with LOGLOSS

Choose best model

Fossil Dataset

Run best model with tuned probabilities on fossil dataset