A Comparison of PCA, Lasso and AE for Dimensionality Reduction with SVM

Ashley Babjac
3/16/22

Background

- Using three datasets to compare
    - TARA: Predicting ocean regions from metagenome assembled genomes coming from ocean water
    - Rhizo: Predicting drought tolerance given OTU tables
    - GEM: Predicting cultured/uncultured status based on organism fine/coarse grained abundances
- Each dataset consists of metadata and features

Example of the data

	latitude	longitude	depth	(more metadata)...	organism1_ abundance	organism2_ abundance	...	organismN_ abundance
Site 1	coor1	coor1	15.1	...	0	1	...	5
Site 2	coor2	coor2	1.5	...	4	2	...	3
...	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...
Site N	coor3	coor3	20.6	...	0	0	...	0

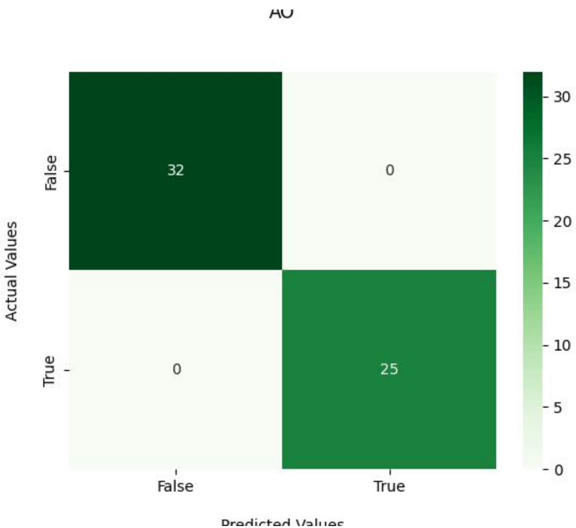
Pipeline

1. Read in/pre-process data (if necessary)
2. Call SMOTE (necessary for TARA, optional for GEM)
3. Run feature selection
   a. Either PCA/Lasso/AE
4. Use the resulting features to call SVM

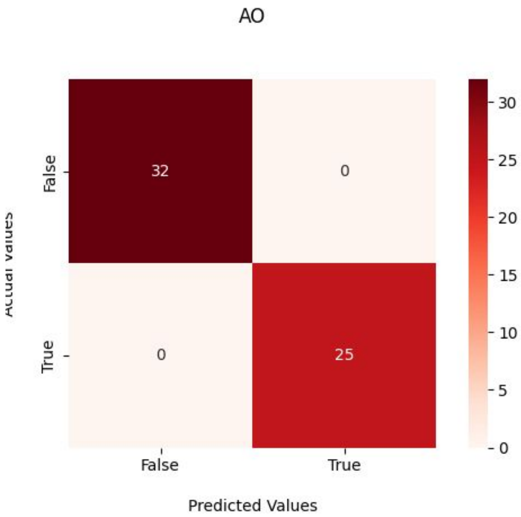
Note: Both feature selection methods and SVM are using 5-fold cross validation with GridSearch of standard parameters.

Results TARA

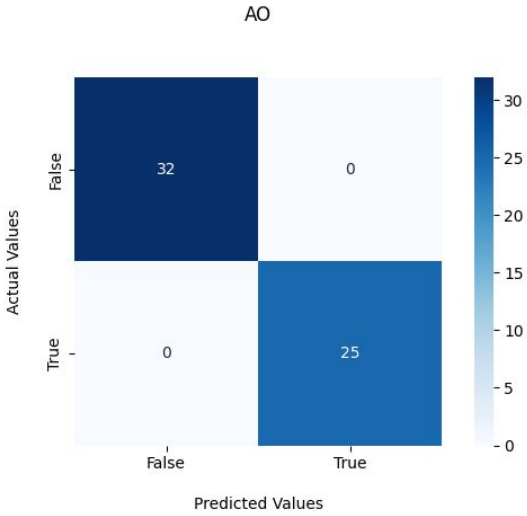
TARA: Arctic Ocean



PCA

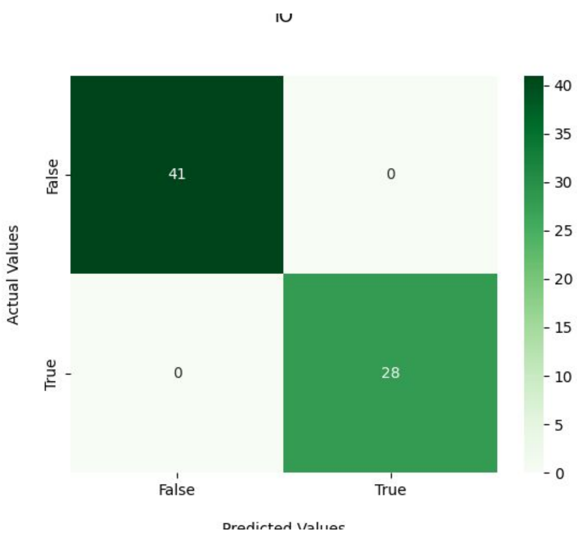


Lasso

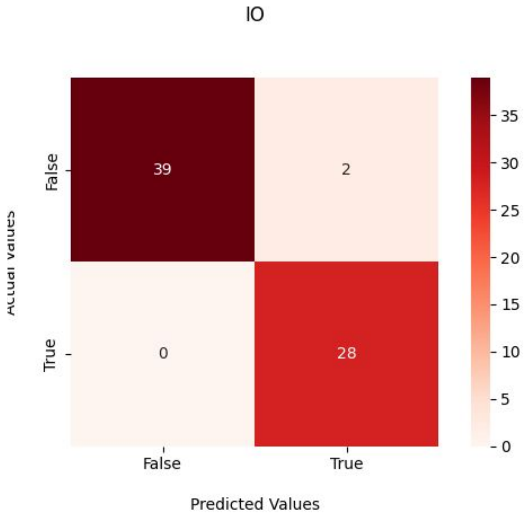


AE

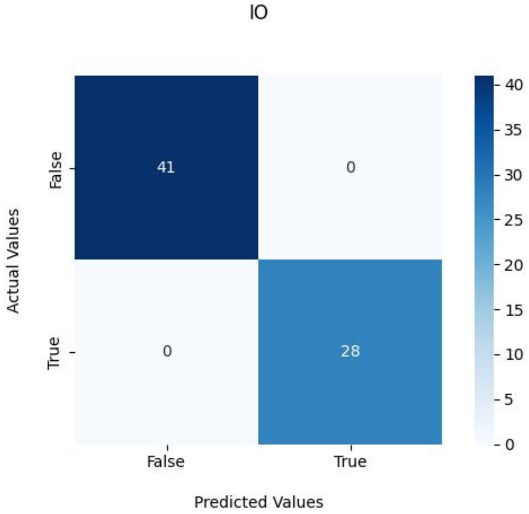
TARA: Indian Ocean



PCA

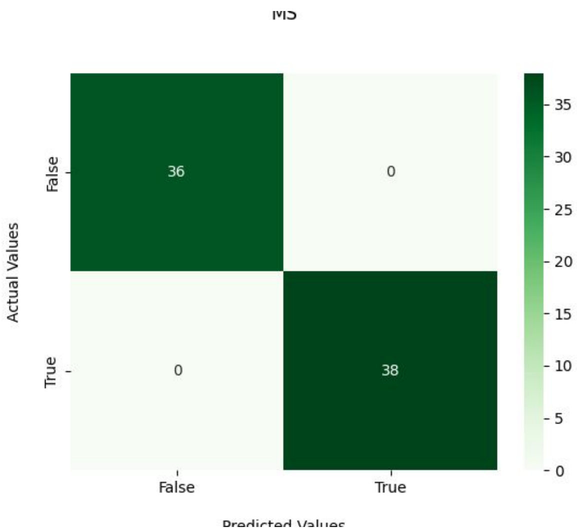


Lasso

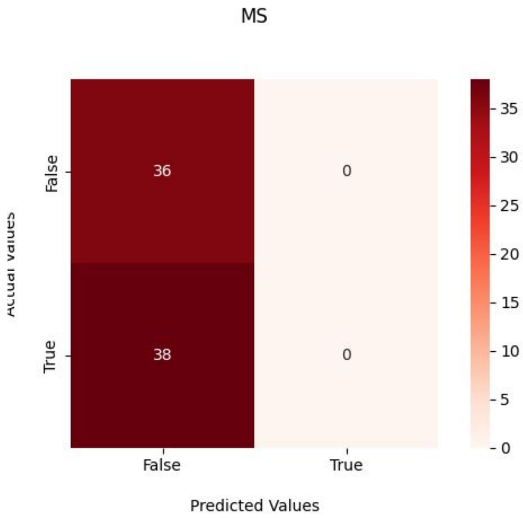


AE

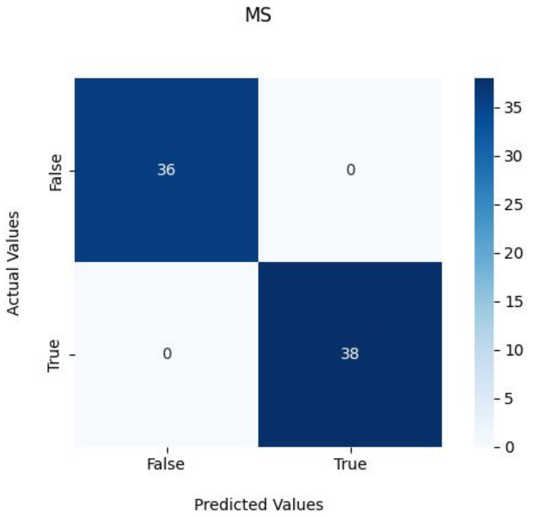
TARA: Mediterranean Sea



PCA

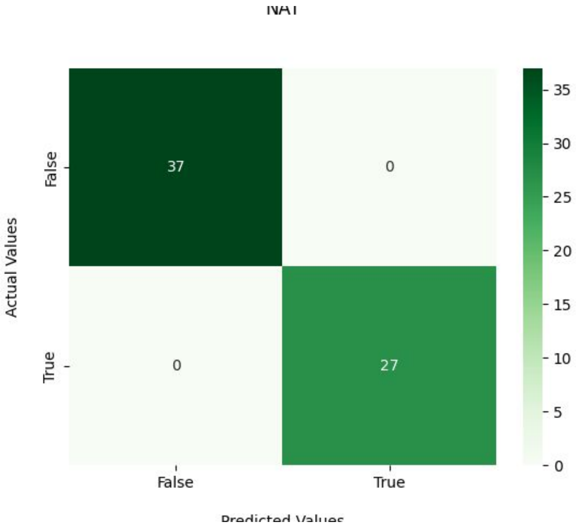


Lasso

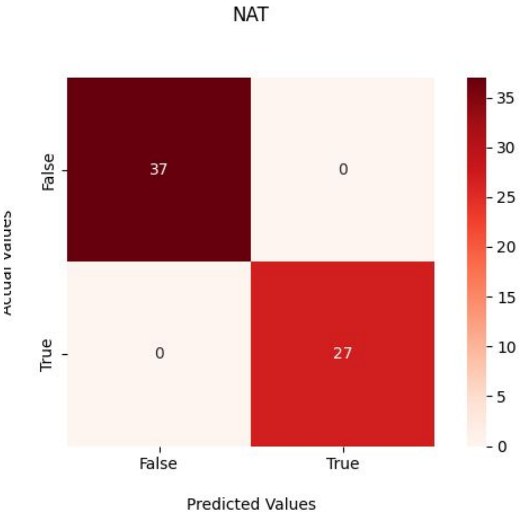


AE

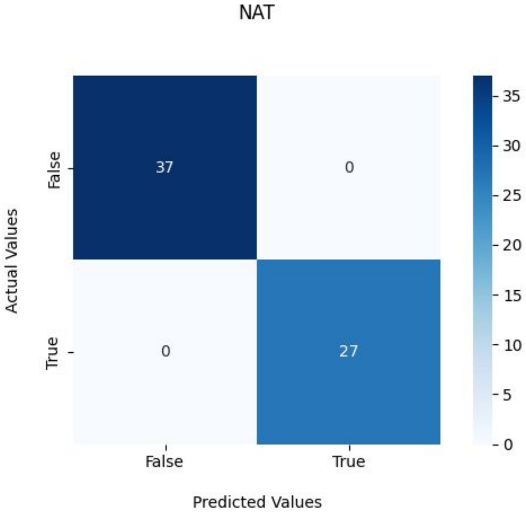
TARA: North Atlantic



PCA

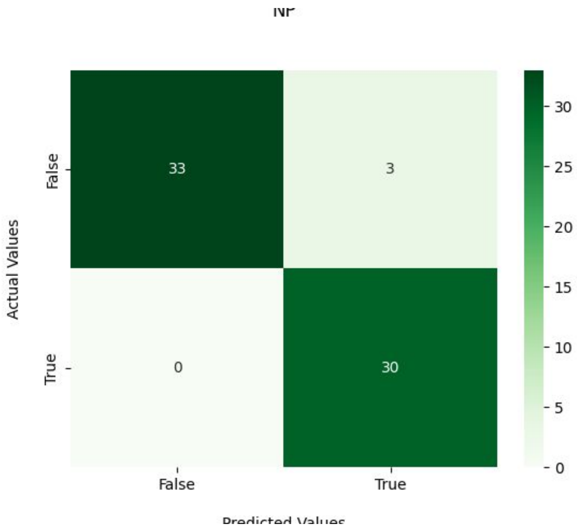


Lasso

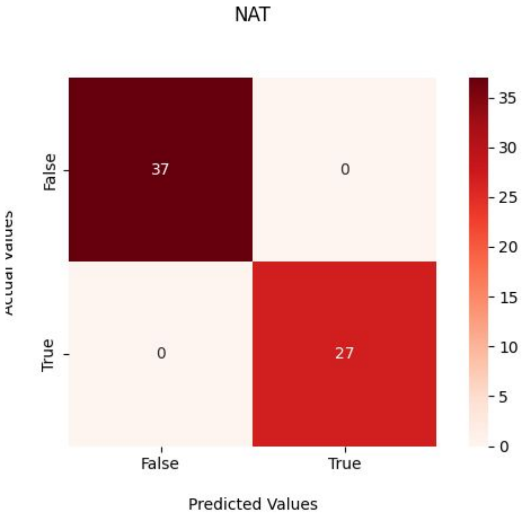


AE

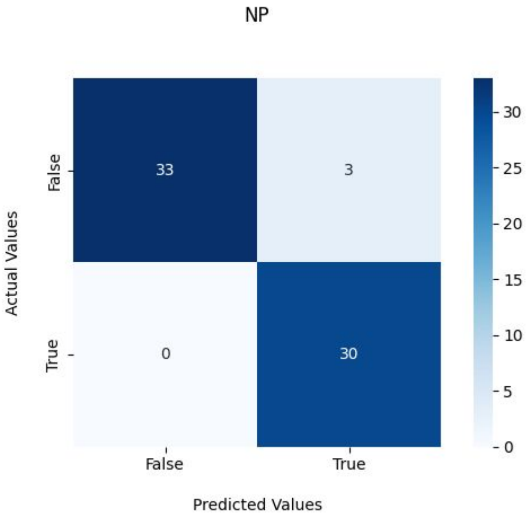
TARA: North Pacific



PCA

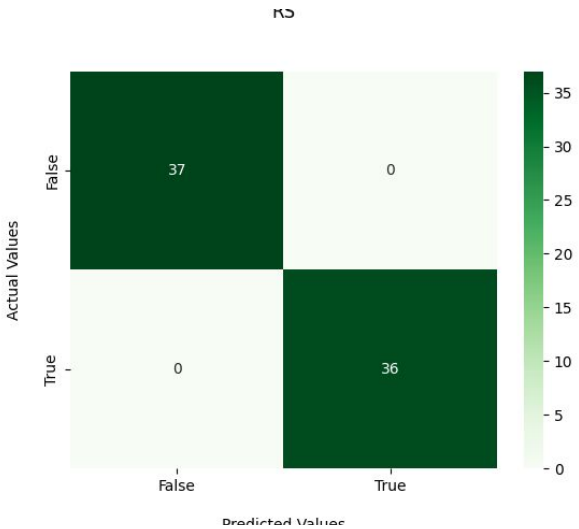


Lasso

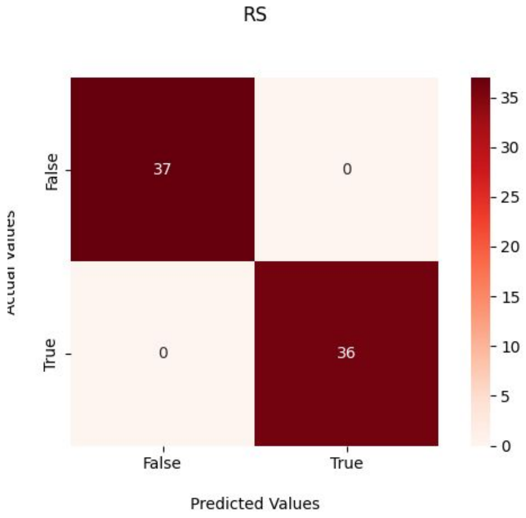


AE

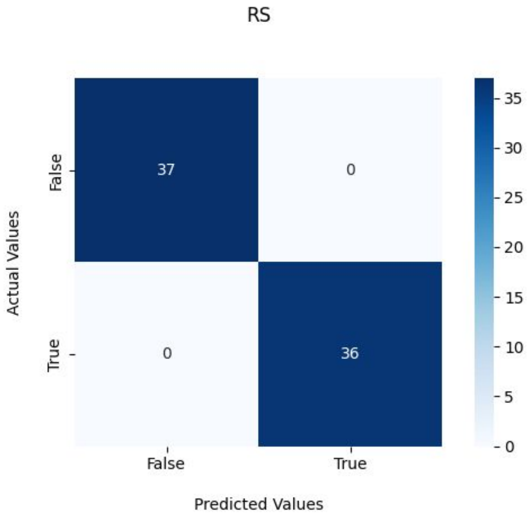
TARA: Red Sea



PCA

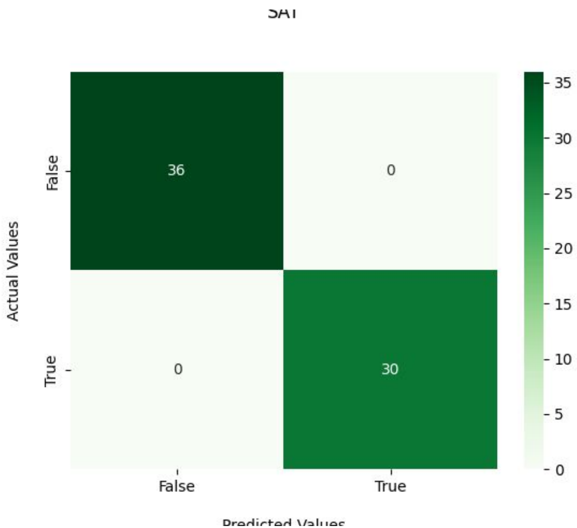


Lasso

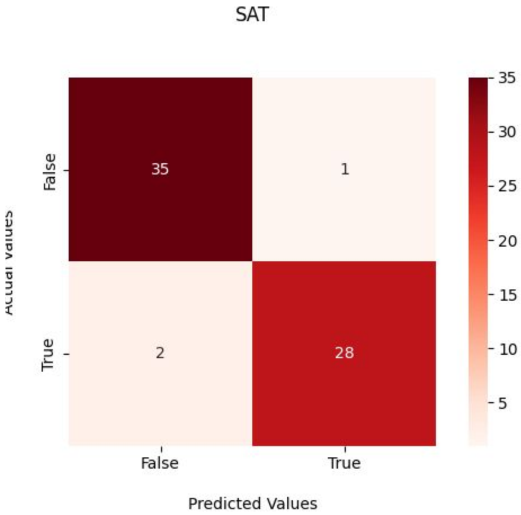


AE

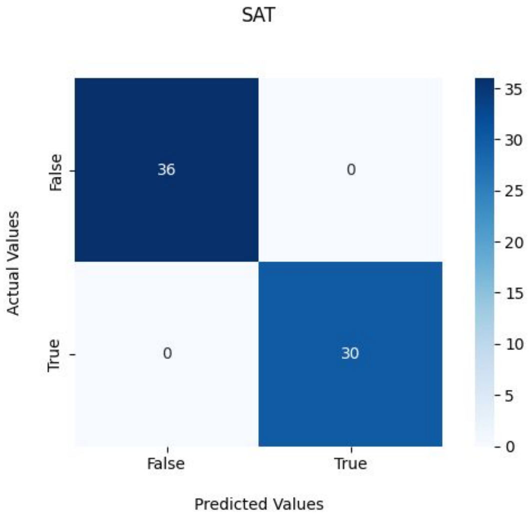
TARA: South Atlantic



PCA

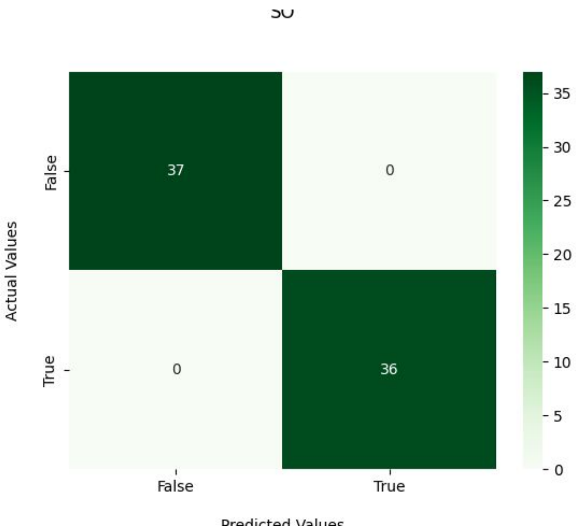


Lasso

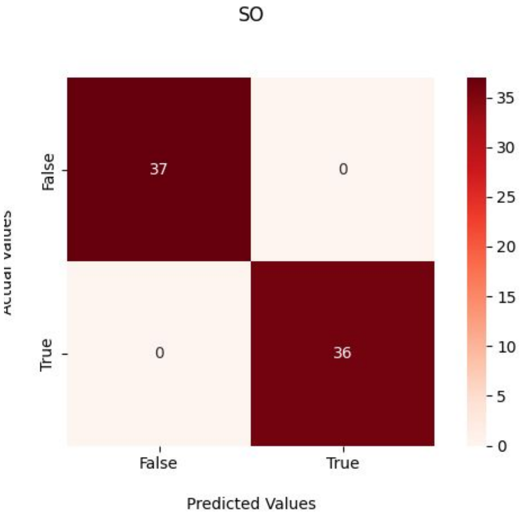


AE

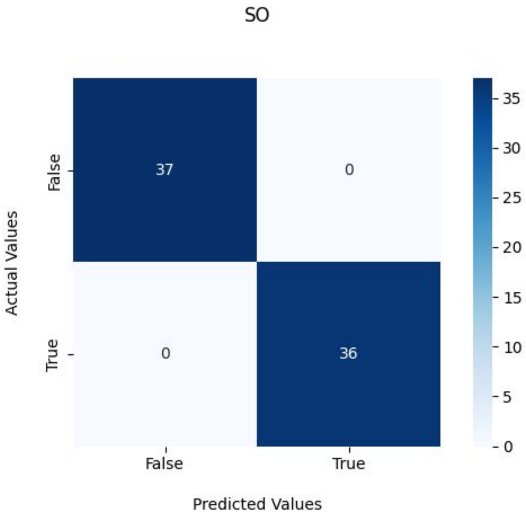
TARA: Southern Ocean



PCA

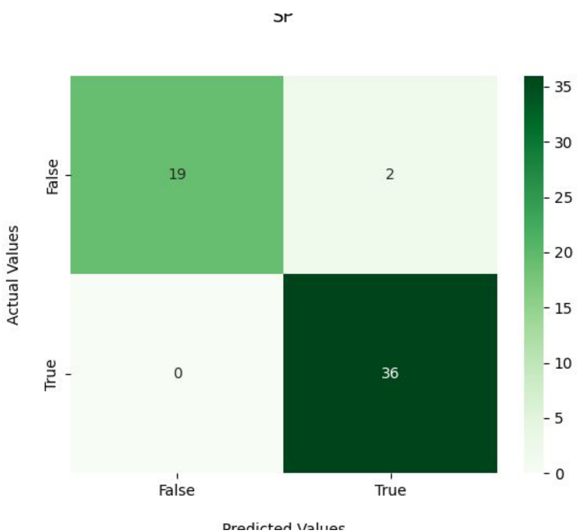


Lasso

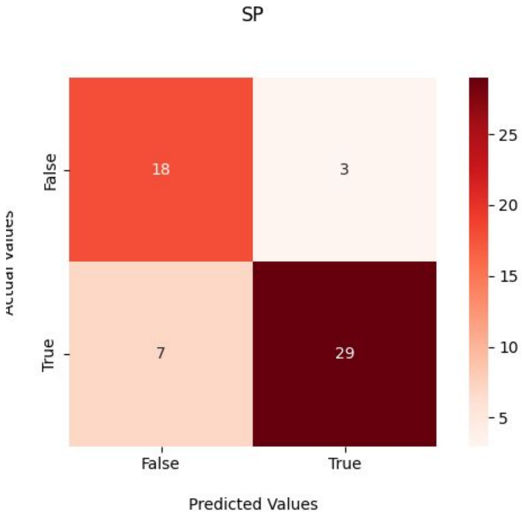


AE

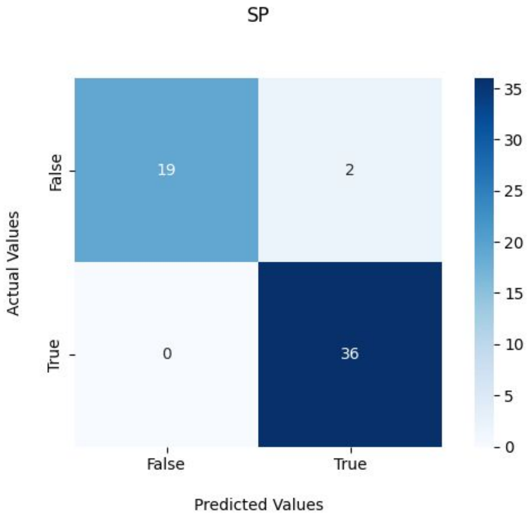
TARA: South Pacific



PCA



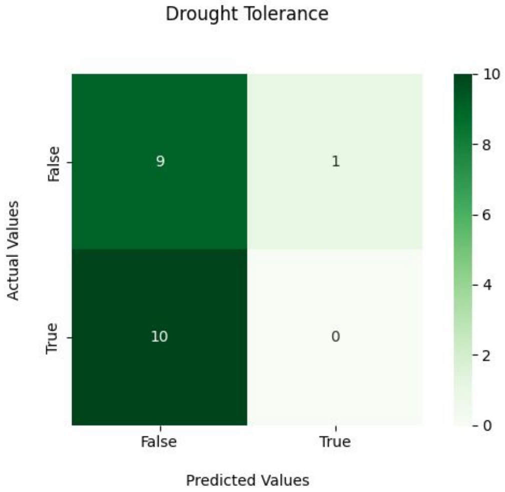
Lasso



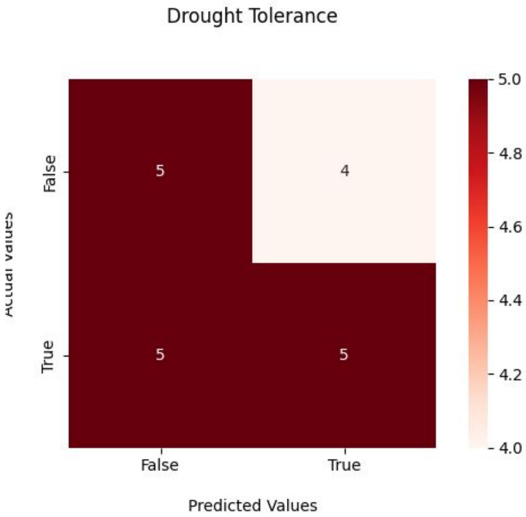
AE

Results Rhizo

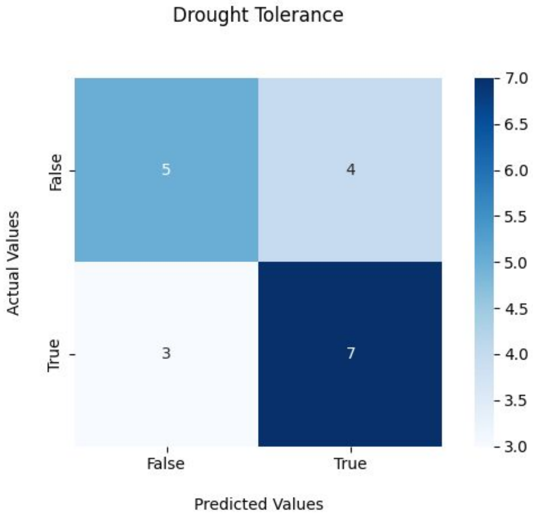
Rhizo: Drought Tolerance



PCA



Lasso



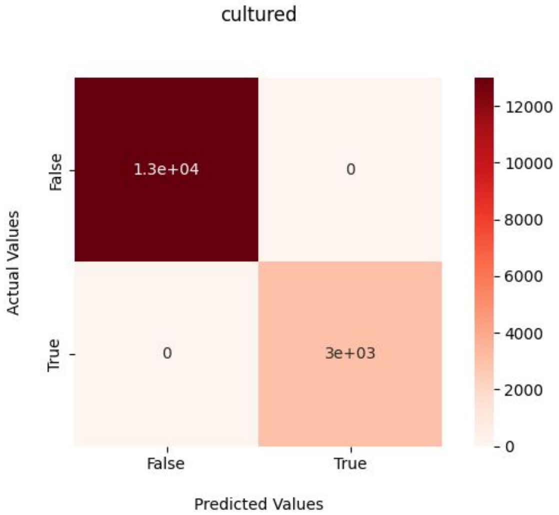
AE

Results GEM

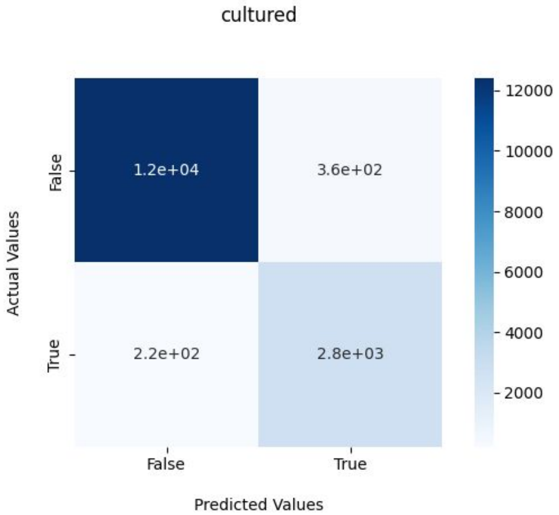
GEM: Cultured/Uncultured

Could not get PCA to run

PCA



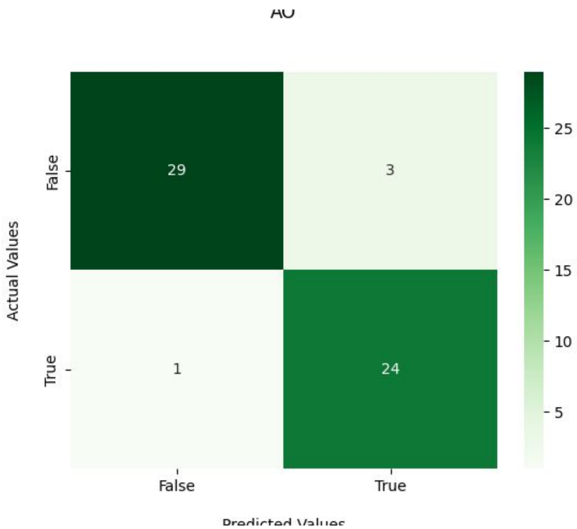
Lasso



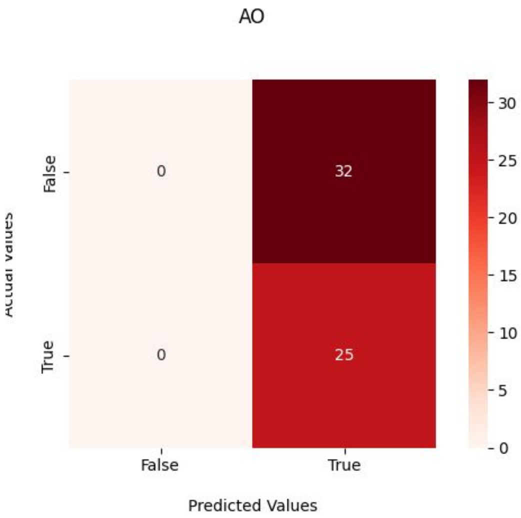
AE

Results TARA (no metadata)

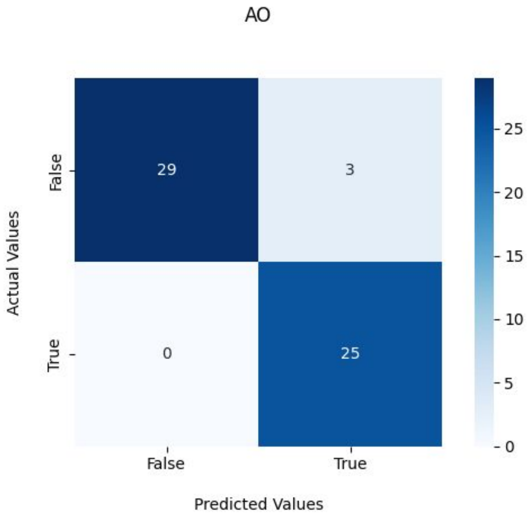
TARA: Arctic Ocean



PCA

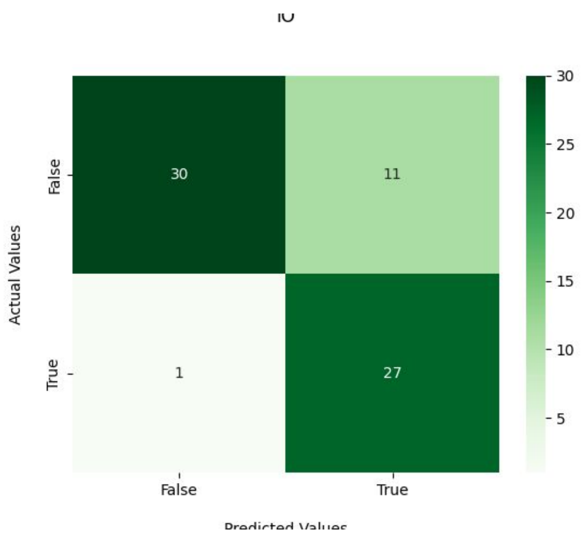


Lasso

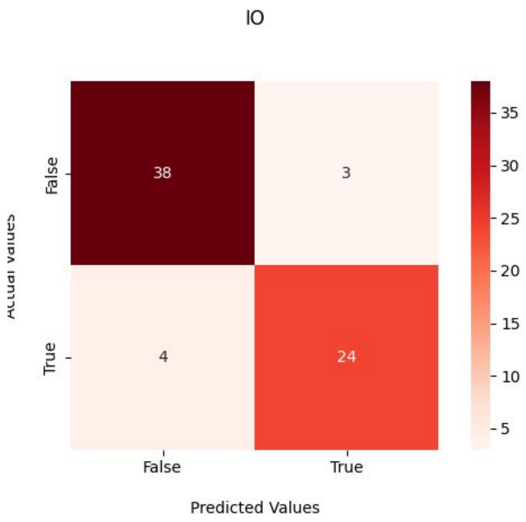


AE

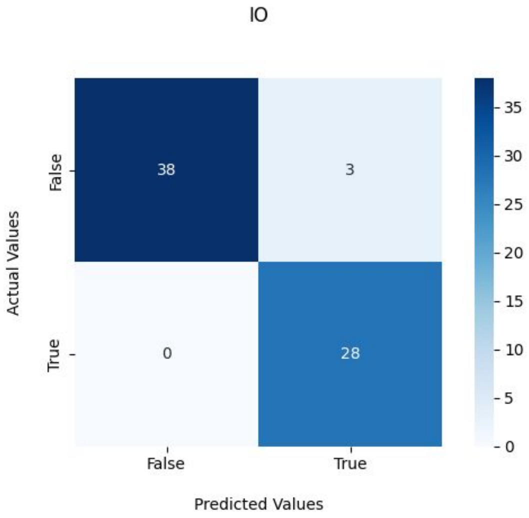
TARA: Indian Ocean



PCA

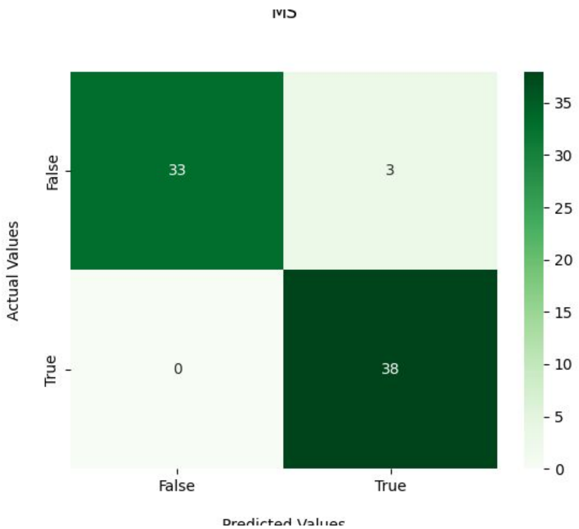


Lasso

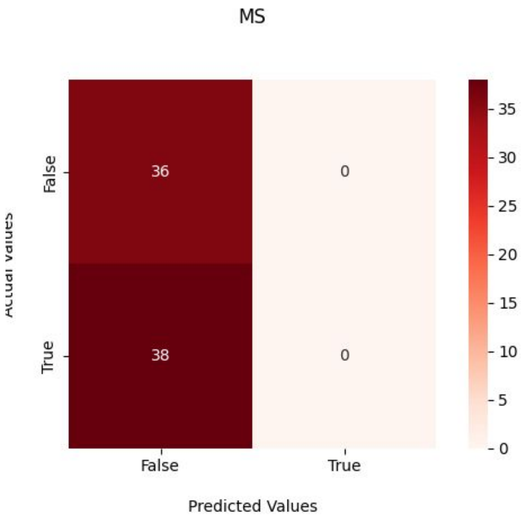


AE

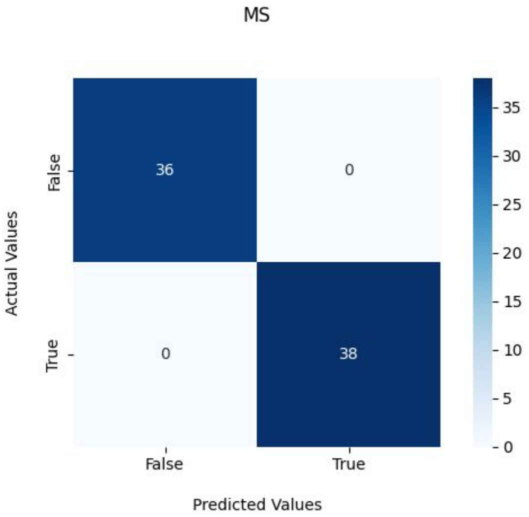
TARA: Mediterranean Sea



PCA

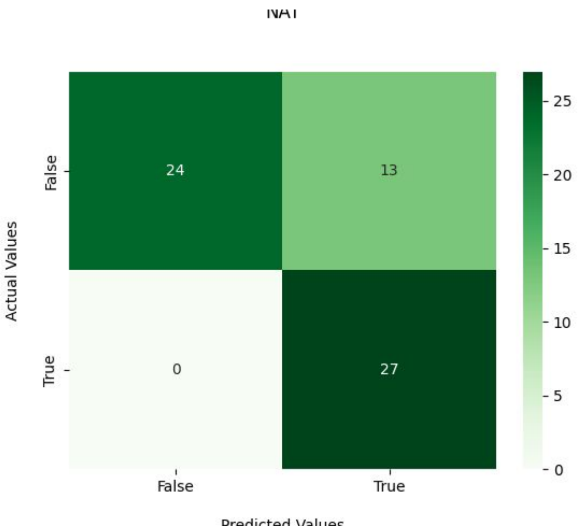


Lasso

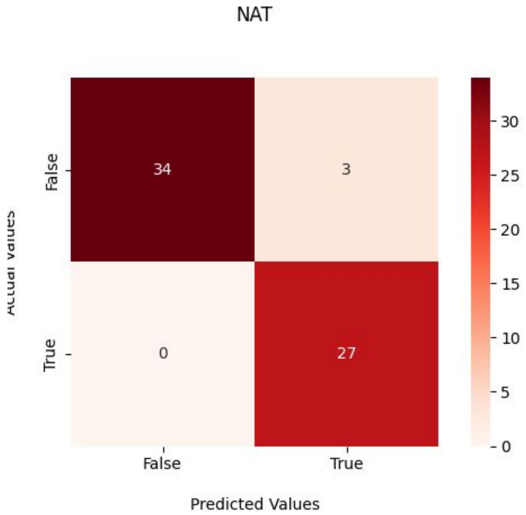


AE

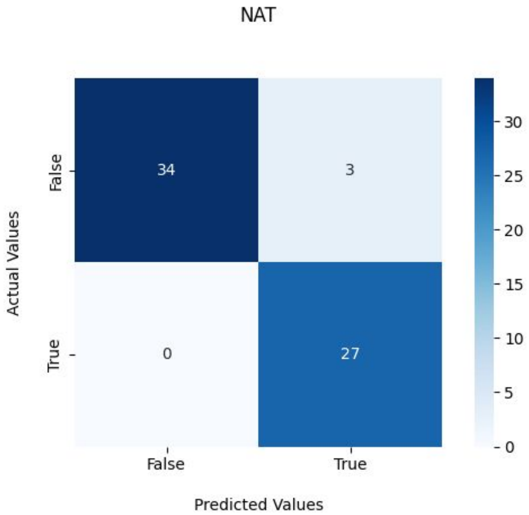
TARA: North Atlantic



PCA

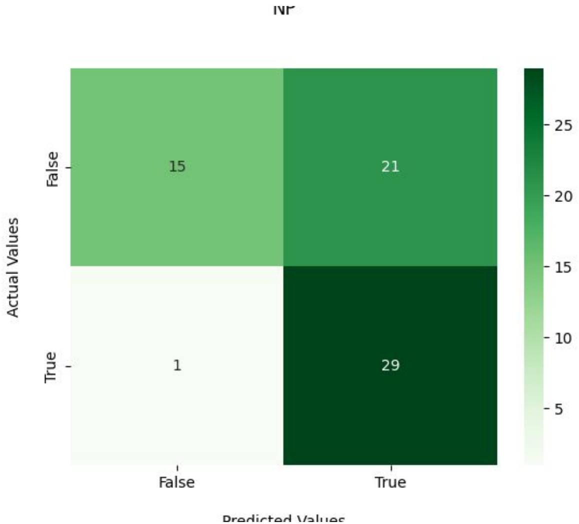


Lasso

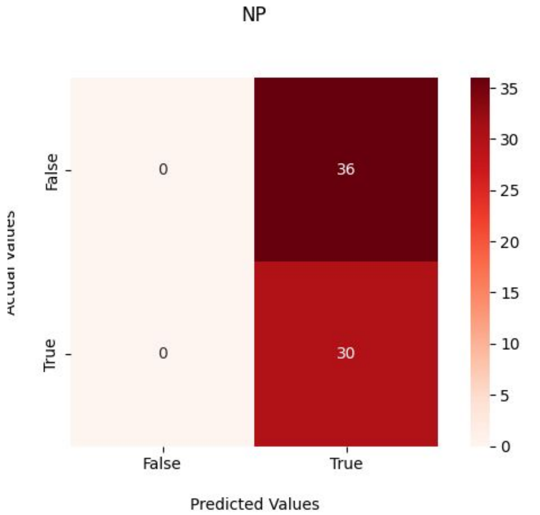


AE

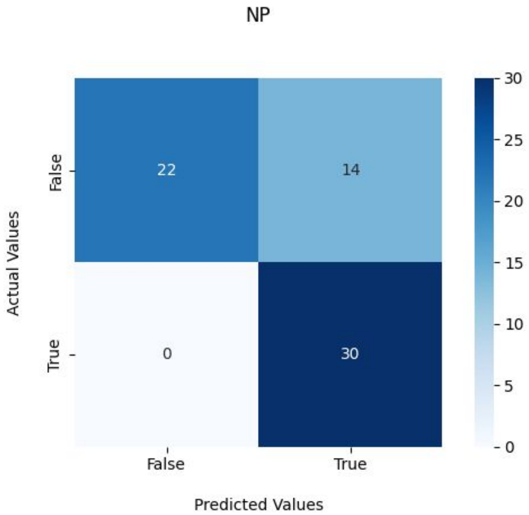
TARA: North Pacific



PCA

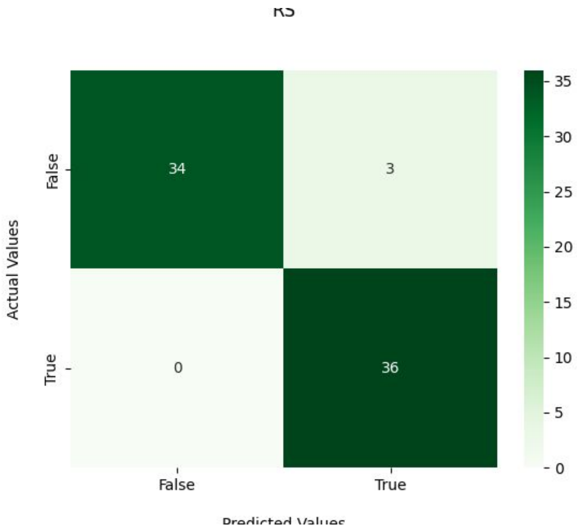


Lasso

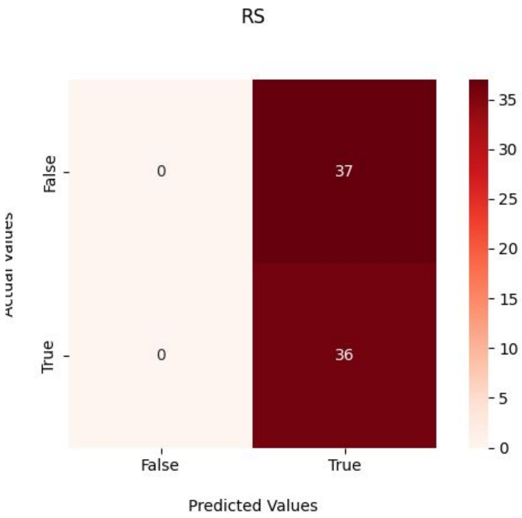


AE

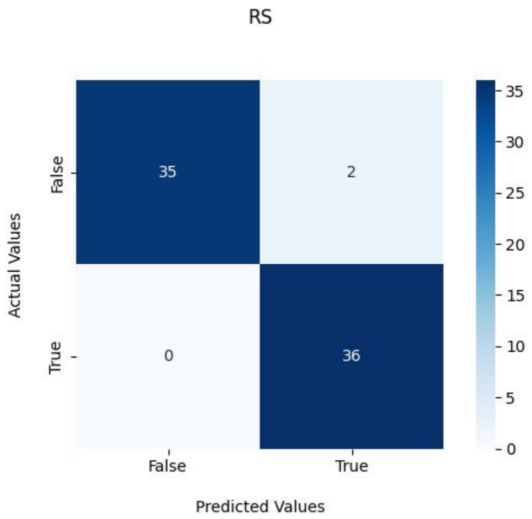
TARA: Red Sea



PCA

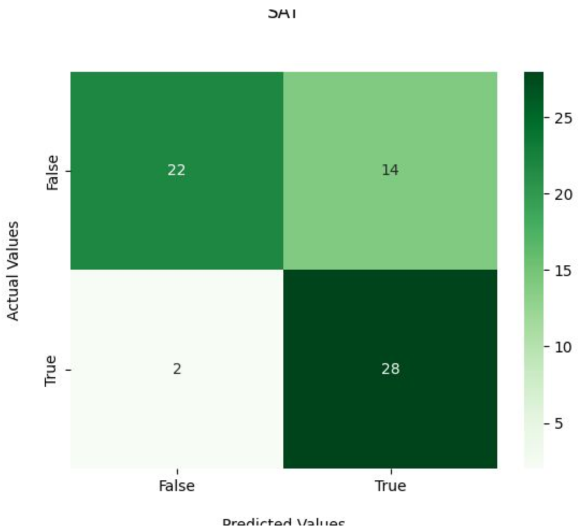


Lasso

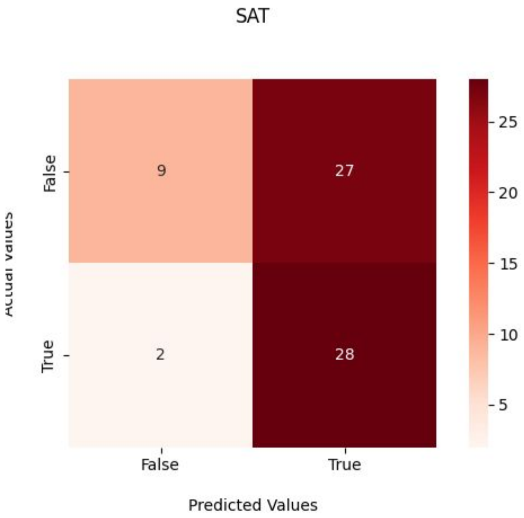


AE

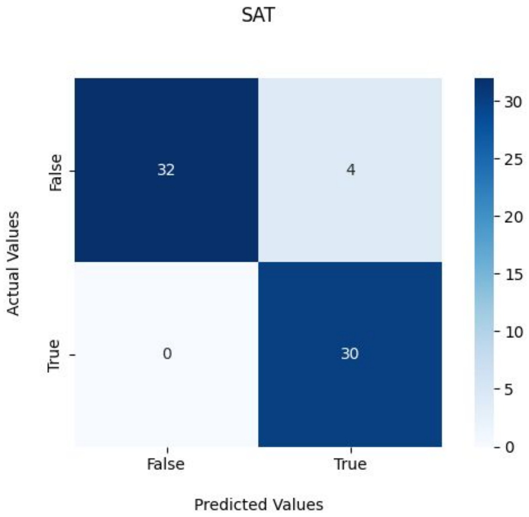
TARA: South Atlantic



PCA

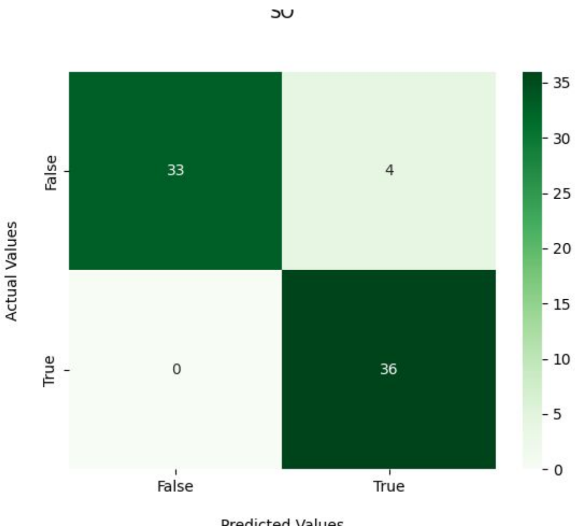


Lasso

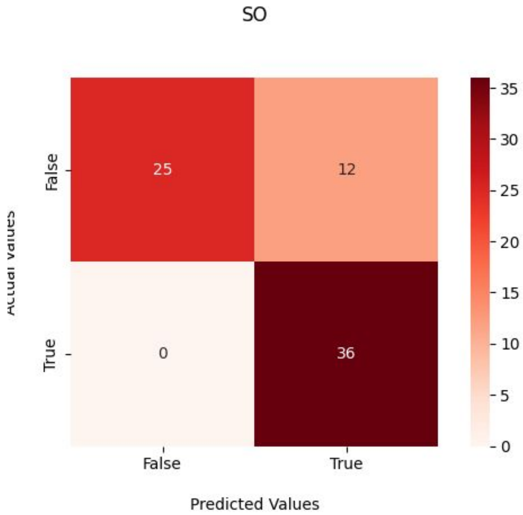


AE

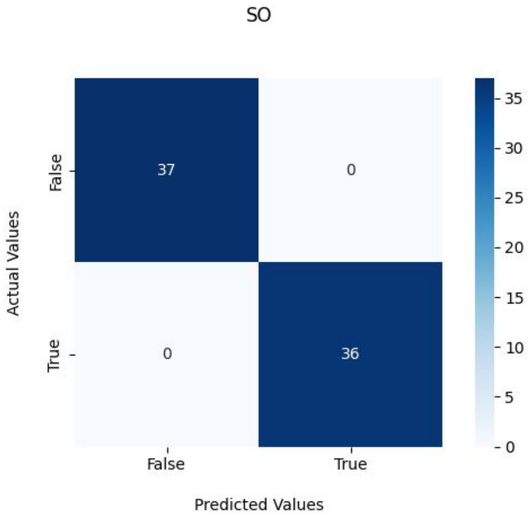
TARA: Southern Ocean



PCA

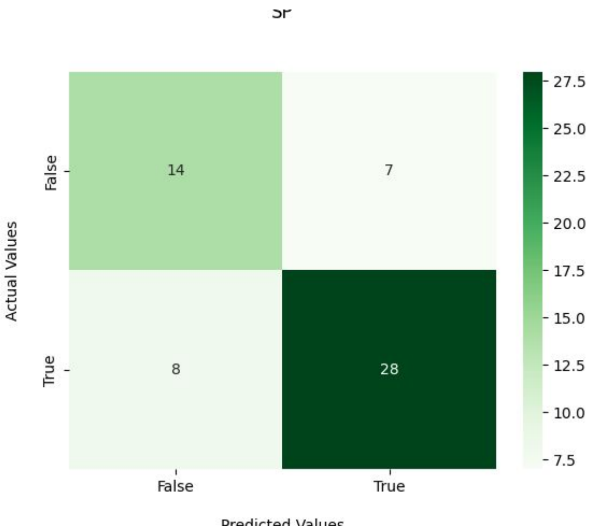


Lasso

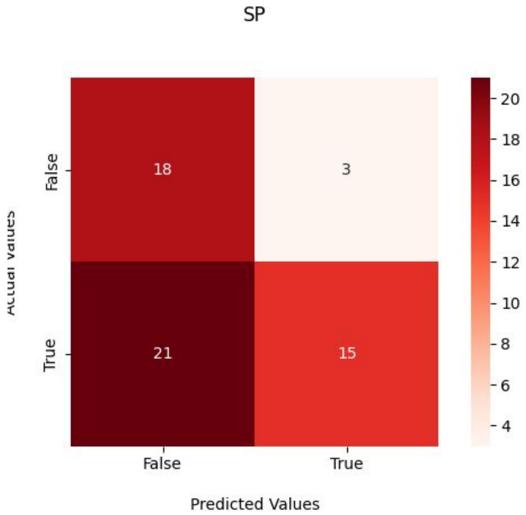


AE

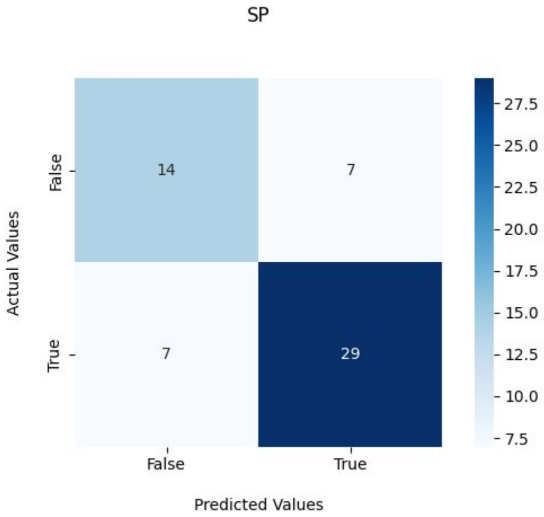
TARA: South Pacific



PCA



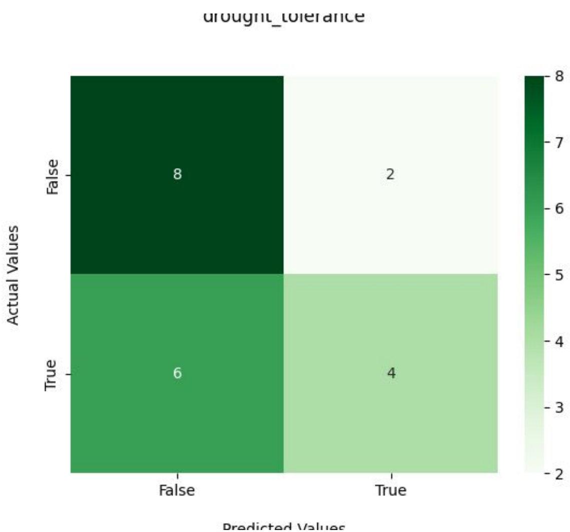
Lasso



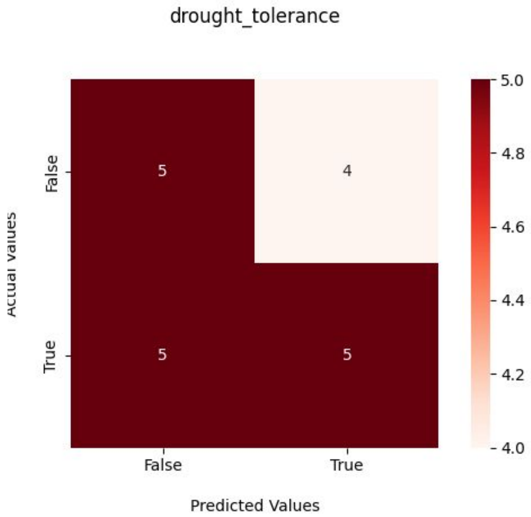
AE

Results Rhizo (no metadata)

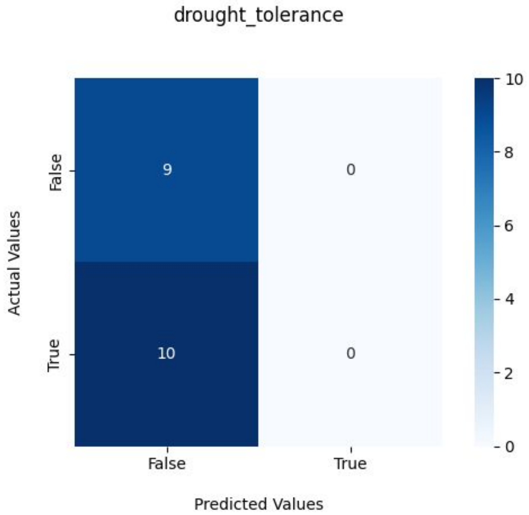
Rhizo: Drought Tolerance



PCA



Lasso



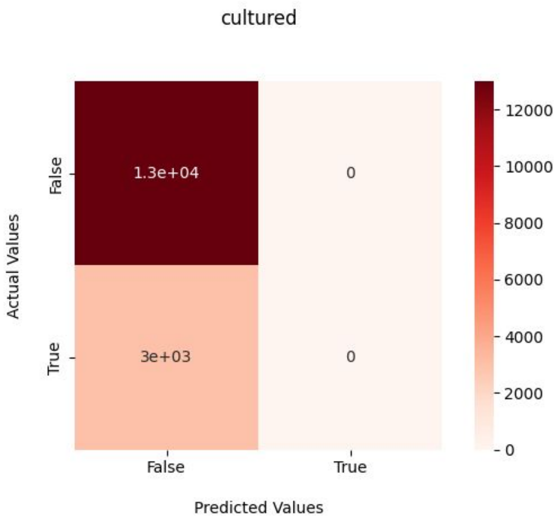
AE

Results GEM (no metadata)

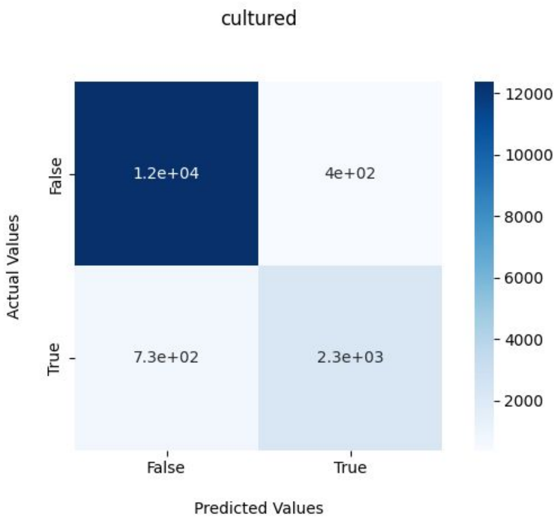
GEM: Cultured/Uncultured

Could not get PCA to run

PCA



Lasso



AE

Conclusions

- PCA, Lasso, and Autoencoders seem relatively comparable if we include the metadata
  - Best model is generally AE, but sometimes Lasso or PCA does slightly better
- Autoencoders consistently do the best when we remove metadata
  - Lasso seems to have trouble picking out “good” features when there is no metadata
  - PCA never finished on GEM but seems inconsistent in terms of whether it produces a good feature selection or not

General insights about runtimes

- Autoencoders run by far the fastest on large datasets
  - AE+SVM: ~3 hours on GEM
  - Lasso+SVM: ~6 hours on GEM
  - PCA+SVM: Ran for >24 hours and never finished for GEM
- Small datasets the runtimes didn’t make much difference but PCA generally ran the fastest

Current problems/Next steps

1. Having an issue with Lasso where all coefficients are being set to 0 (i.e. nothing is deemed important)
   a. Owen suggested trying lower alpha values
   b. Also going to do a test for multicollinearity
2. Cai suggested summarizing in a table with error bars
   a. Going to switch over to AUC instead of accuracy, but this is on my to-do :)
3. Still having an issue with PCA
   a. Appears to just hang in the terminal for certain grid search parameters?
   b. Run it on UTK’s ISAAC?