Appendix to Data 793 Practicum Report, *Exploration & Experimentation of Applying Machine Learning Methods to Coral Microbiome Data*

Katherine Cassandra Sperow, MS Data Science Candidate, American University

December 11, 2023

Unsupervised Modeling Settings

K-means

(See also `Coseq_P.Rmd` and `Coseq_S.Rmd` on GitHub)

```
coseq(p_bac_num,
   K= 2:50,
   transformation = “logclr”,
   model = “kmeans”,
   nstart = 100,
   iter.max = 1000
   )
```

Supervised Modeling Settings

(See also `Final_Ridge_Lasso_Modeling_Revised.Rmd` on GitHub)

Ridge Modeling

```
cv.glmnet(X[,-1], # model matrix, take out intercept
       y, # target label 0,1
       alpha = 0, # Ridge, L2 Norm
       type.measure = “class”, # misclassification error
       nfolds = 10, # K-fold cross-validation
       family = “binomial” # logistic
       )
```

Lasso Modeling

```
cv.glmnet(X[,-1], # model matrix, take out intercept
        y, # target label 0.1
        alpha = 1, # Lasso, L1 Norm
        type.measure = “class”, # misclassification error
        nfolds = 10, # K-fold cross-validation
        family = “binomial” # logistic
       )
```

Lambda.min Values with Misclassiffication (Error) Rates

Method	Error Rate 1000 ASVs	Lambda.min	Error Rate 500 ASVs	Lambda.min	Error Rate 200 ASVs	Lambda.min
Ridge	28.0 %	4.464	28.8 %	3.706	15.5 %	0.6629
Lasso	15.1 %	0.05015	15.8 %	0.04569	15.5 %	0.01186

Plots

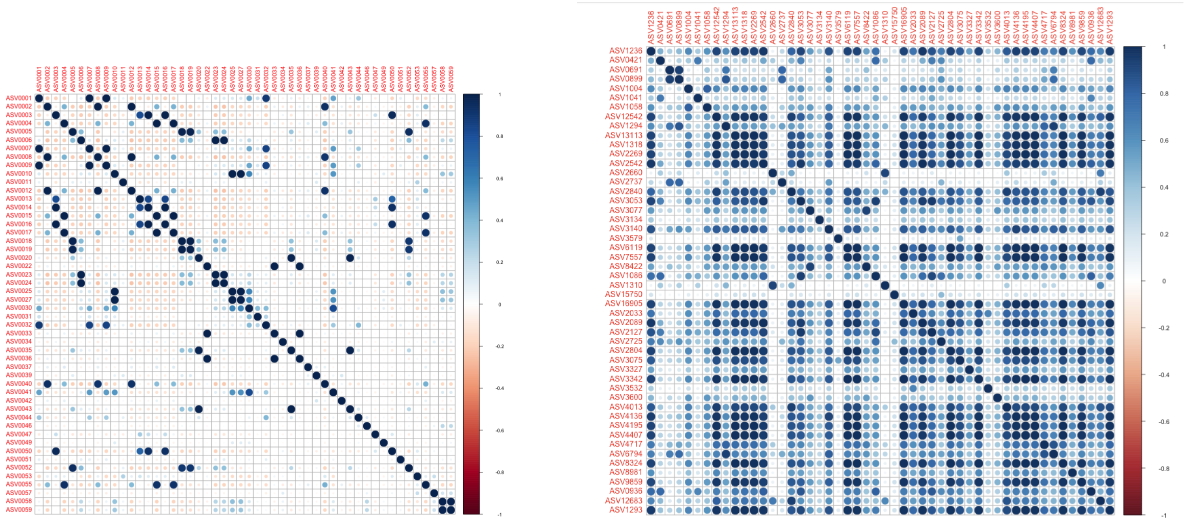

Figure 1: Correlations of the Top 50 ASVs for all observations (left) vs. Clade C (right).

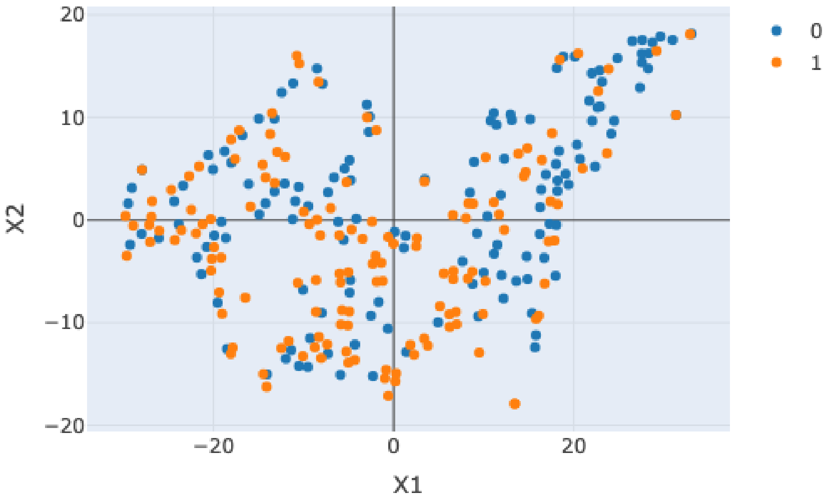

Figure 2: Bacteria data visualized with a t-SNE plot indicating Clade C observations as 1 in orange and non-clade-C as 0 in blue.

Cluster Assignments

- Note: Predictor rows are doubled to show the cluster assignment results per species for the first forty coefficients. Complete data frame of joined predictors, cluster assignments and taxonomy in `Final_Modeling_Clustering_Final_Analysis` in `analysis` folder on GitHub.

coef	predictor	seed23__run2	seed105__run3	seed12__run4	coral__species
0.947	ASV1321	3	6	1	P
0.947	ASV1321	10	3	6	S
-0.888	ASV2269	3	6	1	P
-0.888	ASV2269	12	5	4	S
-0.657	ASV4059	3	6	1	P
-0.657	ASV4059	10	3	6	S
-0.605	ASV4407	3	6	1	P
-0.605	ASV4407	10	3	6	S
-0.442	ASV7680	3	6	1	P
-0.442	ASV7680	10	3	6	S
-0.386	ASV6267	3	6	1	P
-0.386	ASV6267	10	3	6	S
-0.290	ASV3532	3	6	1	P
-0.290	ASV3532	10	3	6	S
0.276	ASV1445	11	3	7	P
0.276	ASV1445	10	3	6	S
-0.275	ASV9936	3	6	1	P
-0.275	ASV9936	10	3	6	S
-0.273	ASV5987	3	6	1	P
-0.273	ASV5987	10	3	6	S
0.268	ASV3327	3	6	1	P
0.268	ASV3327	10	3	6	S
-0.232	ASV2822	3	6	1	P
-0.232	ASV2822	10	3	6	S
0.231	ASV9424	3	6	1	P
0.231	ASV9424	10	3	6	S
-0.218	ASV11359	3	6	1	P
-0.218	ASV11359	10	3	6	S
0.217	ASV1365	3	6	1	P
0.217	ASV1365	2	5	4	S
0.213	ASV3075	3	6	1	P
0.213	ASV3075	10	3	6	S
0.201	ASV2132	3	6	1	P
0.201	ASV2132	10	3	6	S
0.197	ASV3342	3	6	1	P
0.197	ASV3342	10	3	6	S
0.194	ASV6984	3	6	1	P
0.194	ASV6984	10	3	6	S
0.194	ASV1315	3	6	13	P
0.194	ASV1315	15	3	6	S

First 10 Ridge 200 Model Predictors (in order of coefficient absolute value)

Please see `Final_Modeling_Clustering_Final_Analysis` in `analysis` folder for complete list on GitHub.

coef	predictor	Kingdom	Phylum	Class	Order	Family	Genus	Species
0.947	ASV1321	Bacteria	Proteobacteria	Gammaproteobacteria	Alteromonadales	Alteromonadaceae	Aestuariibacter	NA
-0.888	ASV2269	Bacteria	Proteobacteria	Gammaproteobacteria	Alteromonadales	Marinobacteraceae	Marinobacter	NA
-0.657	ASV4059	Bacteria	Proteobacteria	Gammaproteobacteria	Pseudomonadales	Moraxellaceae	Acinetobacter	NA
-0.605	ASV4407	Bacteria	Proteobacteria	Gammaproteobacteria	Pseudomonadales	Moraxellaceae	Acinetobacter	NA
-0.442	ASV7680	Bacteria	Proteobacteria	Alphaproteobacteria	Puniceispirillales	SAR116_clade	NA	NA
-0.386	ASV6267	Bacteria	Proteobacteria	Gammaproteobacteria	Alteromonadales	Pseudoalteromonadaceae	Pseudoalteromonas	NA
-0.290	ASV3532	Bacteria	Bacteroidota	Bacteroidia	Cytophagales	Flammeovirgaceae	Flammeovirga	NA
0.276	ASV1445	Bacteria	Proteobacteria	Gammaproteobacteria	Oceanospirillales	Marinomonadaceae	Marinomonas	NA
-0.275	ASV9936	Bacteria	Bacteroidota	Bacteroidia	Bacteroidales	NA	NA	NA
-0.273	ASV5987	Bacteria	Proteobacteria	Gammaproteobacteria	Pseudomonadales	Moraxellaceae	Acinetobacter	NA