Micro/Nano Plastic Detection via Paper Microfluidic Chips

SIE 533 Group Project

Mervin XuYang Lim, Lexi DeFord, Neo Kai Shuen, Mohammad Wali-ur-Rahman

Agenda

1. Introduction
    a. Micro/Nano Plastics
    b. Data
2. Algorithms
3. Results
4. Discussion

Introduction - Micro/Nano Plastics (MNP's)

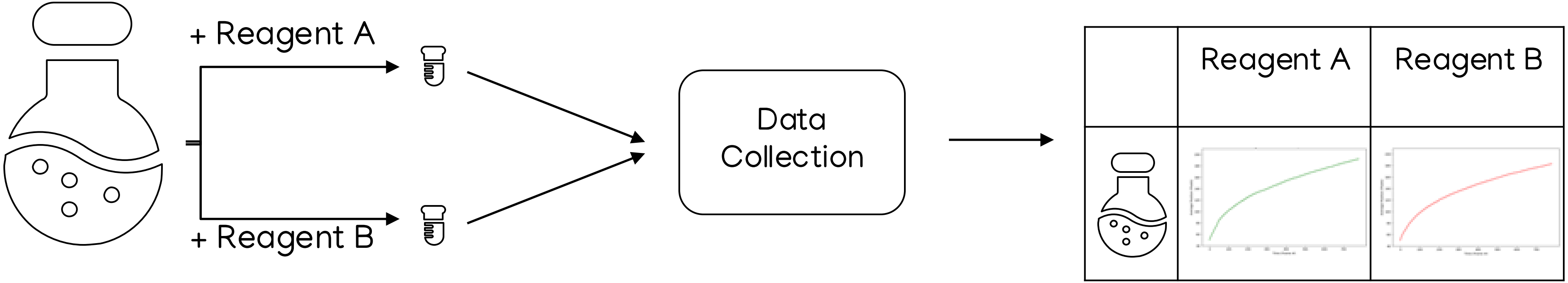
Plastic is everywhere!

Various varieties:

- Type: PMMA, PS, PET, LDPE, PVC…
- Origin: Primary or Secondary
- Size: 1 mm - <1 μm

Goal: Detection & Classification of MNP's in Water (Binary classification: "Plastic" vs "No Plastic")

Introduction - Data Collection Schematic



+ Reagent A

+ Reagent B

Data Collection

	Reagent A	Reagent B

n = 304 x d = 340

Unbalanced Classes: 240 “Plastic” & 64 “No Plastic”

This is a Clean Dataset

- Of the 4C's, we did have a few instances of missing data
    - Averaged the other replicates to complete the dataset
- No converting, creating, or correcting was done
- Classes are unbalanced
    - F1 score
    - Stratified K-fold CV

Hypotheses

Hypothesis 1: Samples are classifiable as “Plastic” or “No Plastic” using various binary algorithms with an F1 score of 0.80 or above.

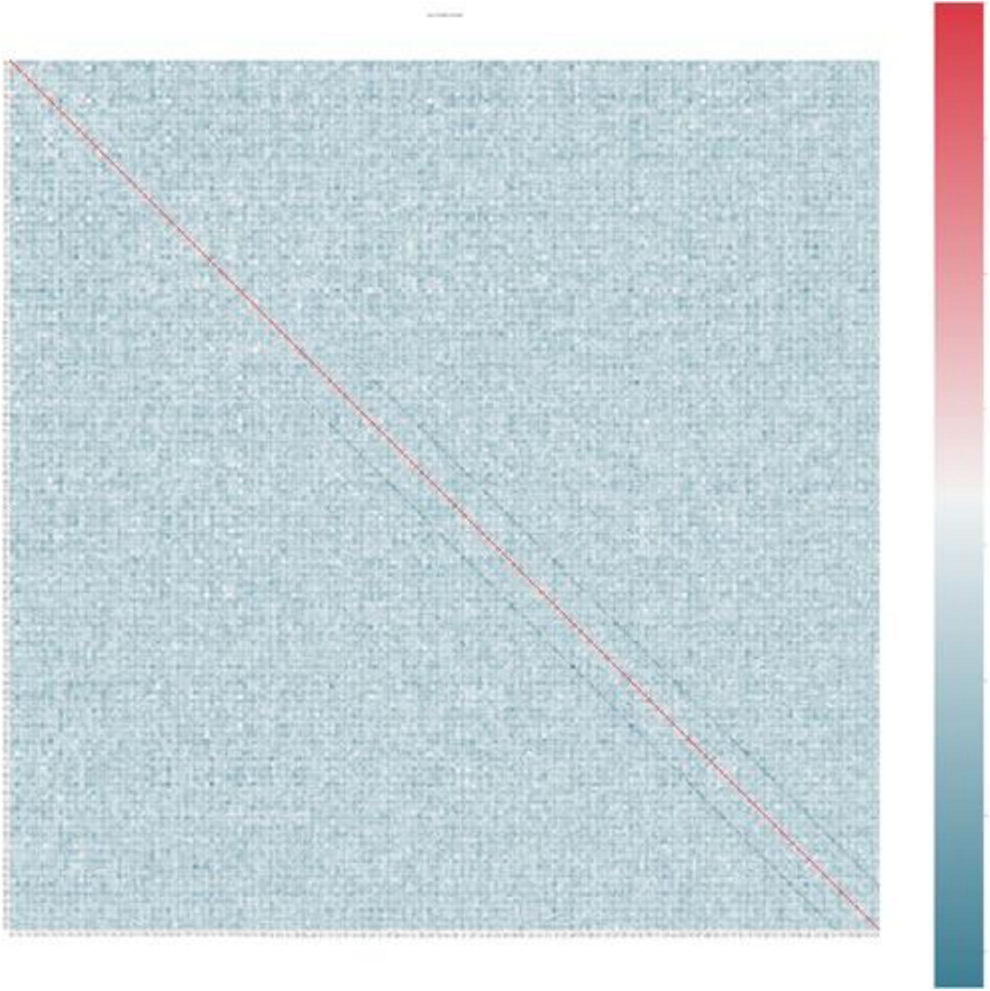
Hypothesis 2: Bovine Serum Albumin (BSA) and Aspartic Acid will be the two most important reagents for this classification across all algorithms.

- BSA: large, complex structure with hydrophobic and hydrophilic regions
- Aspartic Acid: small, negatively charged region, hydrophilic

Exploratory analysis

Pairwise plot

- Generally uncorrelated
  - We expected the band of correlation around the diagonal due to domain knowledge



Algorithms

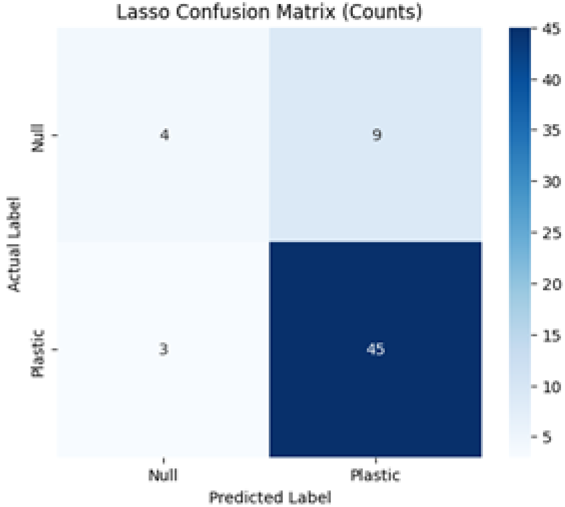
- Linear methods
  - Logistic Regression
    - No regularization
    - LASSO (L1)
    - Ridge (L2)
  - Linear SVM
- Tree methods
  - Random Forest
  - Gradient Boosting
  - Light Boosting
  - XGboost

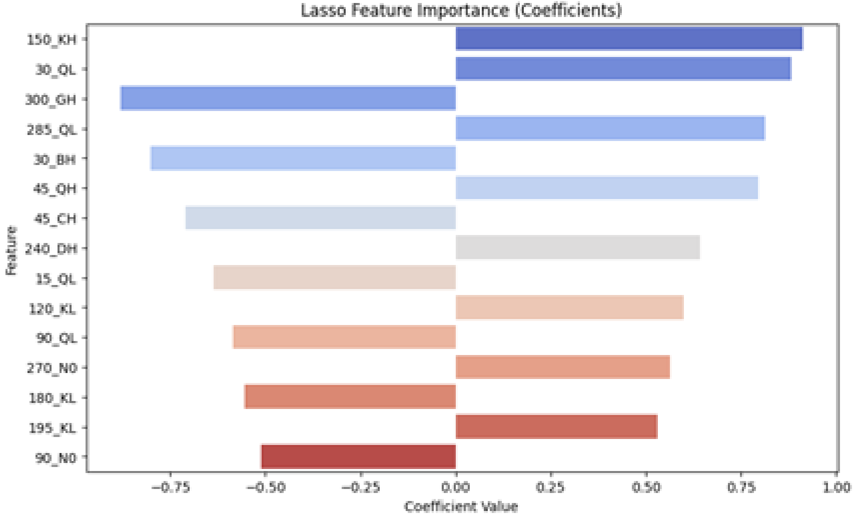
L1 Regularization

(LASSO)
Stratified 10-fold CV.
Optimal penalty: 0.167

Parameters	LASSO
Cross validation	10x
ROC AUC	**0.8830**
Accuracy	0.8033

Params	LASSO
Precision	0.9375
Recall	0.8333
F1-score	0.8824
F2-score	0.9146



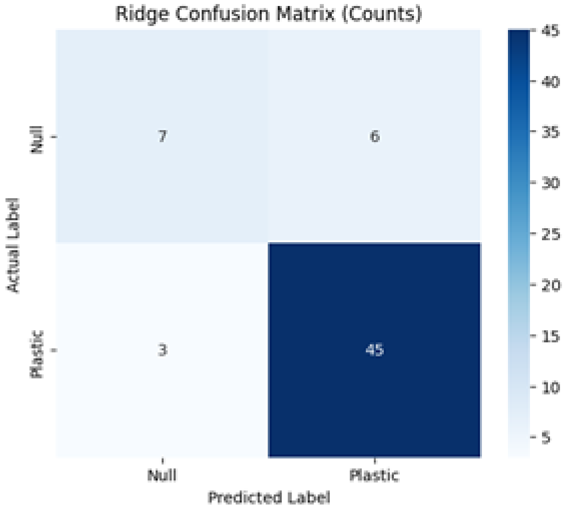


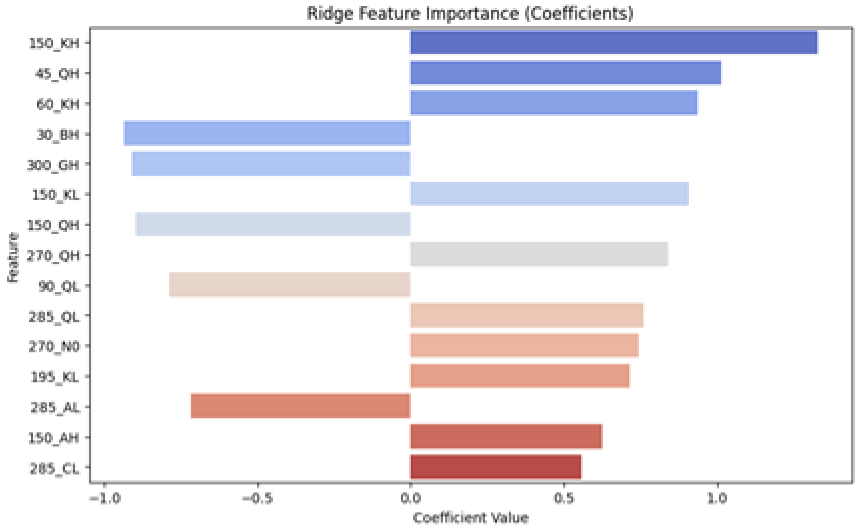
L2 Regularization

(Ridge)
Stratified 10-fold CV.
Optimal penalty: 0.599

Parameters	Ridge
Cross validation	10x
ROC AUC	0.8702
Accuracy	**0.8525**

Params	Ridge
Precision	0.9375
Recall	0.8824
F1-score	0.9091
F2-score	0.9259



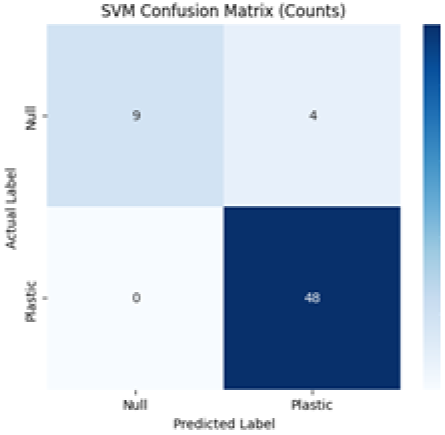


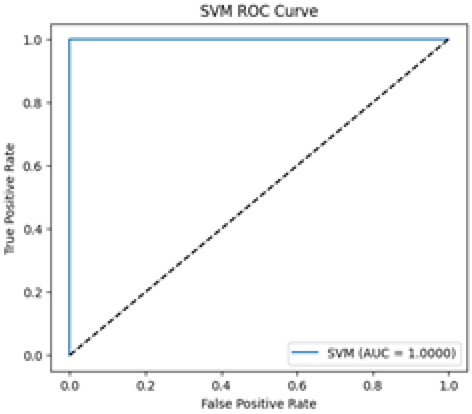
SVM

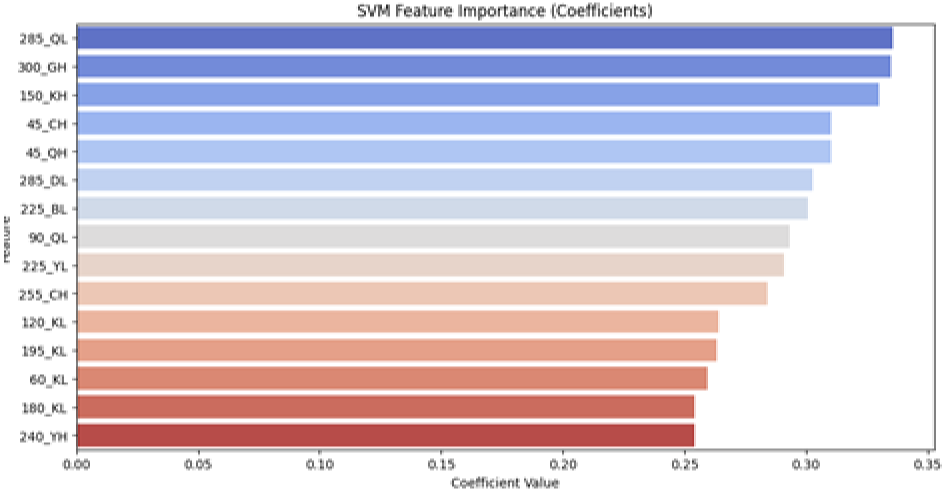
Stratified 10-fold CV.
Kernel: linear (performed better than rbf)
C = 0.01

Parameters	SVM
Cross validation	10x
ROC AUC	**1.00**
Accuracy	0.9344

Params	SVM
Precision	0.9230
Recall	1.0000
F1-score	0.9600
F2-score	0.9836

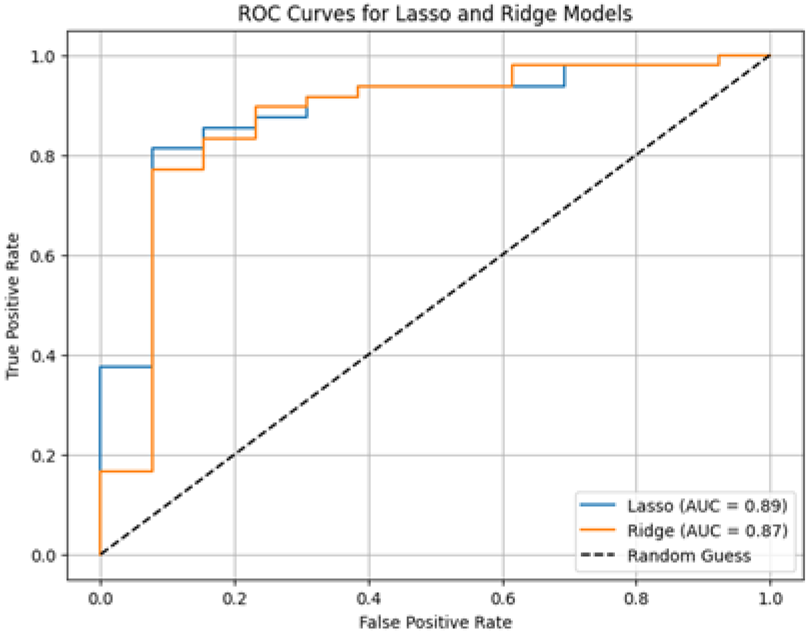






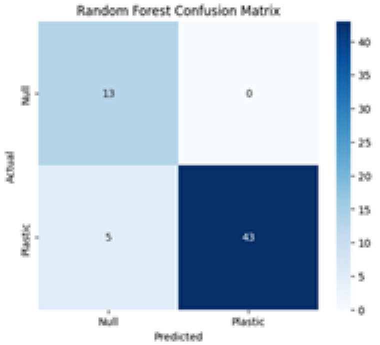
Insights from L1, L2 and SVM

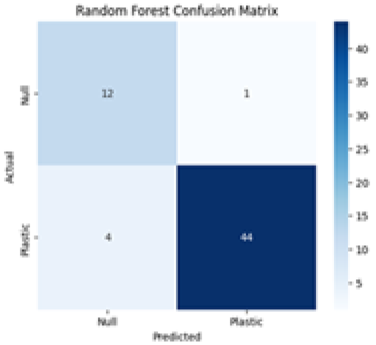
- SVM provided the best performance compared to LASSO and Ridge.
- 15 features provided the optimal results for L1 and L2.
- 150_KH is the most important feature for both regularization, 3rd for SVM.
- 300_GH are within the top 5 feature importance for all 3 models.



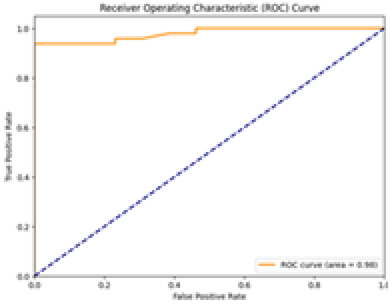
Random Forest Results

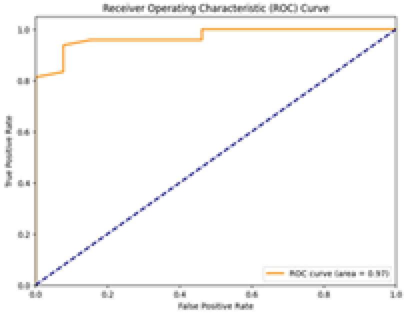
- Stratified 10 Fold Cross Validation
- Grid Search For: 1) Number of estimators, 2) Maximum Tree Depth, 3) Minimum Number of Samples to split an Internal Node.
- Features Selected based on 1) Median Threshold Value and 2) Mean Threshold Value

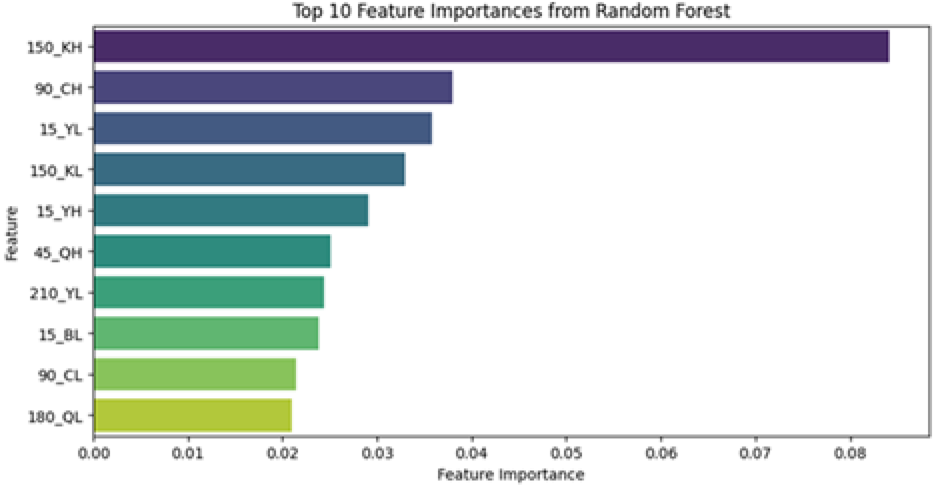




Feature Selection Threshold	Accuracy	Precision	Recall	F1-Score	F2-Score	ROC-AUC	Total Features Selected
Median	0.9180	1.00	0.896	0.945	0.915	0.98	170
Mean	0.9180	0.978	0.917	0.946	0.928	0.97	100

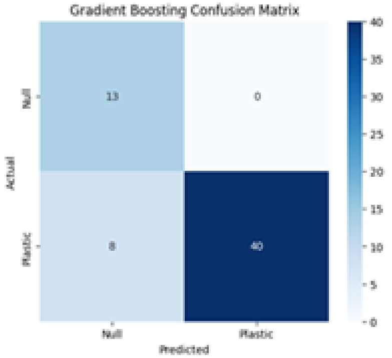


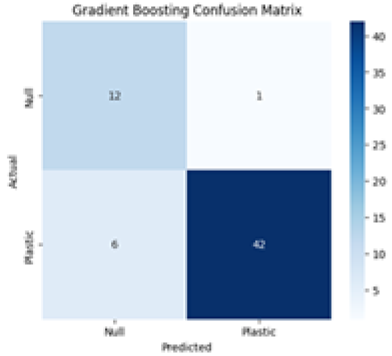


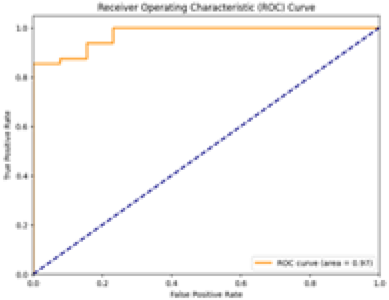


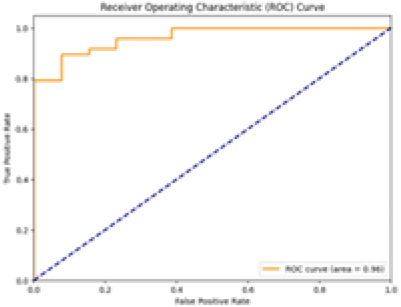
Gradient Boosting Results

- Stratified 10 Fold Cross Validation
- Grid Search For: 1) Number of estimators, 2) Maximum Tree Depth, 3) Learning Rate 3) Minimum Number of Samples to split an Internal Node.
- Features Selected based on 1) Median Threshold Value and 2) Mean Threshold Value

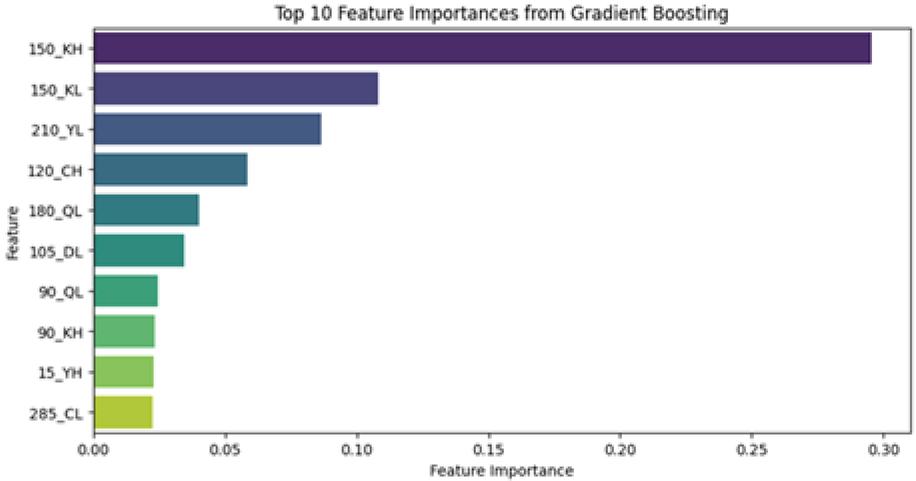






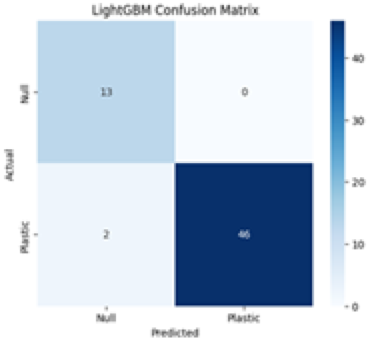


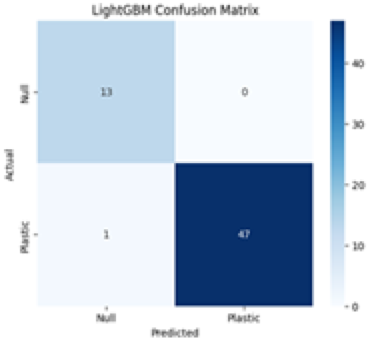
Feature Selection Threshold	Accuracy	Precision	Recall	F1-Score	F2-Score	ROC-AUC	Total Features Selected
Median	0.8689	1.00	0.833	0.909	0.862	0.97	170
Mean	0.8852	0.977	0.875	0.923	0.893	0.96	40



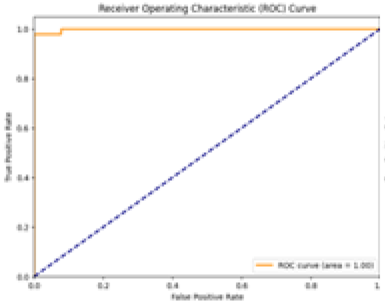
Light Gradient Boosting Results

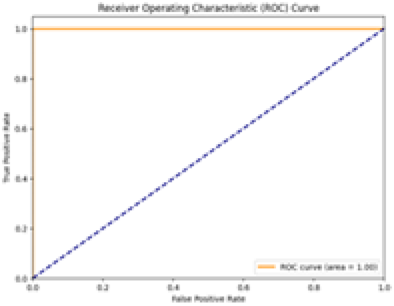
- Stratified 10 Fold Cross Validation
- Grid Search For: 1) Number of estimators, 2) Maximum Tree Depth, 3) Learning Rate, 4) Minimum Number of Samples required in a child node.
- Features Selected based on 1) Median Threshold Value and 2) Mean Threshold Value

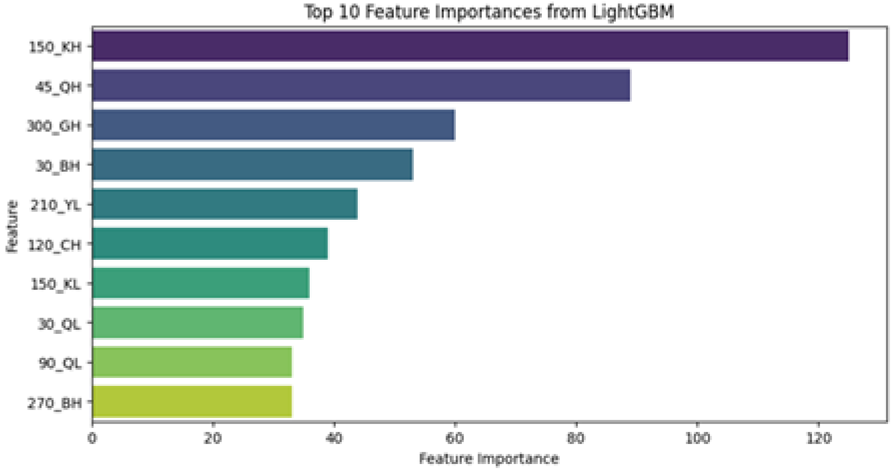




Feature Selection Threshold	Accuracy	Precision	Recall	F1-Score	F2-Score	ROC-AUC	Total Features Selected
Median	0.9672	1.00	0.9583	0.979	0.966	1.00	240
Mean	0.9836	1.00	0.9792	0.9895	0.9833	1.00	71

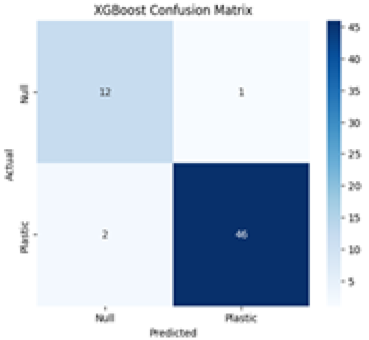


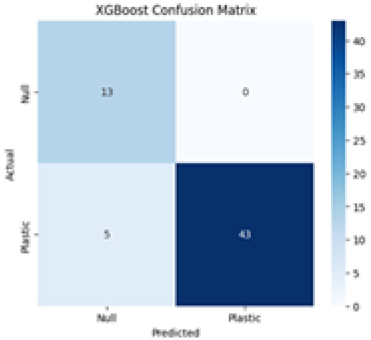




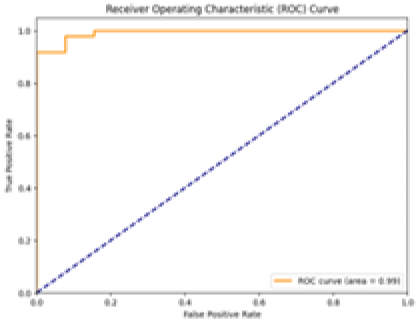
Extreme Gradient Boosting Results

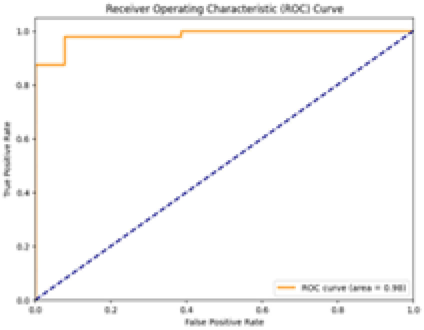
- Stratified 10 Fold Cross Validation
- Grid Search For: 1) Number of estimators, 2) Maximum Tree Depth, 3) Learning Rate 3) Subsample ratio of the training instances
- Features Selected based on 1) Median Threshold Value and 2) Mean Threshold Value

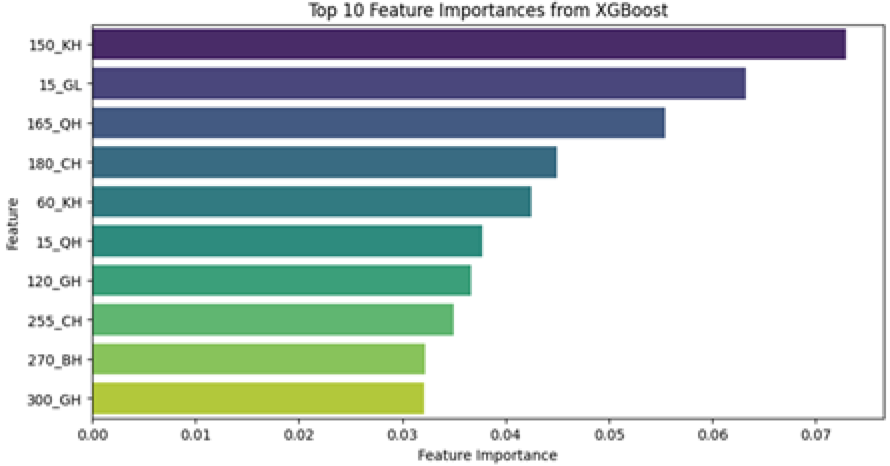




Feature Selection Threshold	Accuracy	Precision	Recall	F1-Score	F2-Score	ROC-AUC	Total Features Selected
Median	0.9180	1.00	0.896	0.945	0.915	0.99	339
Mean	0.9508	0.979	0.958	0.968	0.961	0.98	51



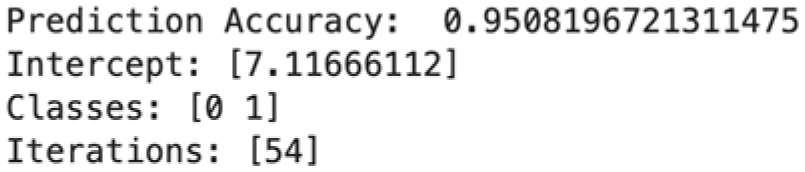




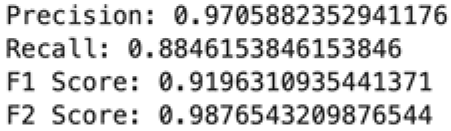
Insights from Decision Tree Based Models

- LightGBM (mean threshold) achieved the highest accuracy (98.36%) and F1-score (0.9895), outperforming all other models.
- The feature “150_KH” was consistently identified as the top predictor across nearly all models and threshold strategies, underscoring its central importance.
- Models using fewer features with only 40 features at mean threshold—still delivered strong F1-scores, highlighting efficient dimensionality reduction.
- High n_estimators across models underscores the importance of a large number of trees for stabilizing predictions and reducing variance.

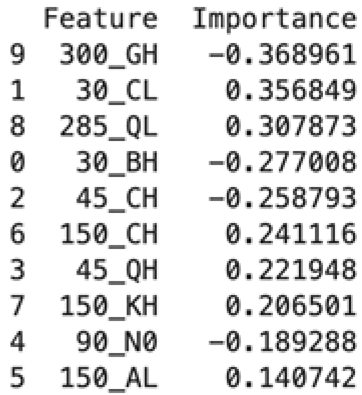
Logistic Regression Results



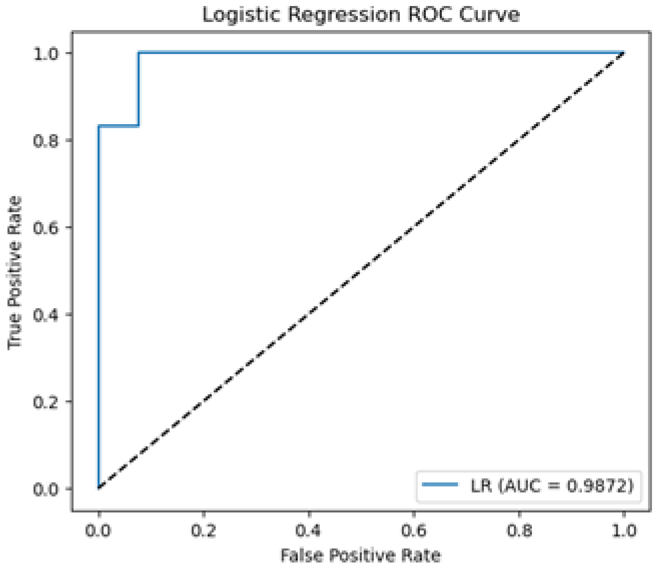
```
Prediction Accuracy:  0.9508196721311475
Intercept: [7.11666112]
Classes: [0 1]
Iterations: [54]
```

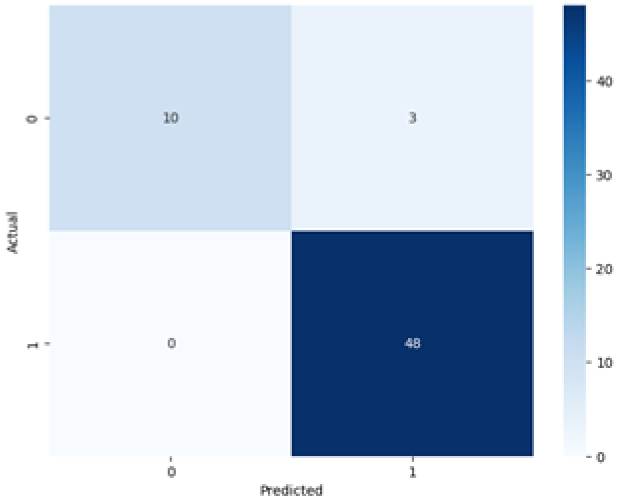


```
Precision: 0.9705882352941176
Recall: 0.8846153846153846
F1 Score: 0.9196310935441371
F2 Score: 0.9876543209876544
```



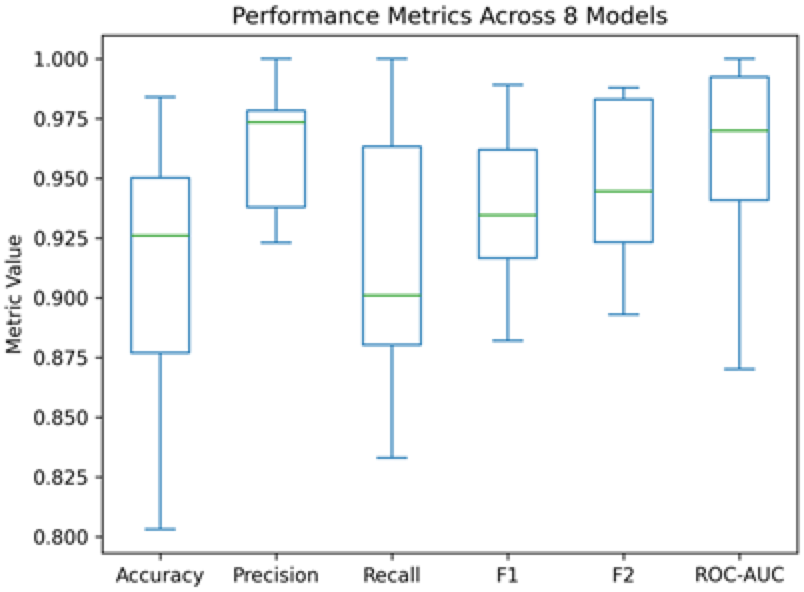
```
  Feature  Importance
9  300_GH   -0.368961
1   30_CL    0.356849
8  285_QL    0.307873
0   30_BH   -0.277008
2   45_CH   -0.258793
6  150_CH    0.241116
3   45_QH    0.221948
7  150_KH    0.206501
4   90_N0   -0.189288
5  150_AL    0.140742
```





Results Summary

Model	Accuracy	Precision	Recall	F1	F2	ROC-AUC
L1	0.803	0.938	0.833	0.882	0.915	0.883
L2	0.853	0.9375	0.882	0.909	0.926	0.870
SVM	0.934	0.923	1.000	0.960	0.983	1.00
RF	0.918	0.978	0.917	0.946	0.928	0.96
GB	0.885	0.977	0.875	0.923	0.893	0.96
LGB	0.984	1.00	0.979	0.989	0.983	1.00
XGB	0.950	0.979	0.958	0.968	0.961	0.98
LR	0.951	0.970	0.885	0.919	0.988	0.99



Top Features Discussion

- Lysine
  - Polar, easily dissolves in water, positive charge
- Glycine
  - Nonpolar, easily dissolves in water, negative charge
- BSA
  - *Various different regions around this large molecule*
- Glutamine
  - Polar, easily dissolves in water, neutral charge
- Cysteine
  - Polar, complex water interaction, neutral charge

Chemical *variety* is important for classifying this dataset

Discussion & Conclusion

Hypothesis 1 was supported; yes, data were classifiable via binary algorithms as “plastic” or “no plastic” above an F1 score of 0.80. Model F1 scores were between 0.882-0.989 with a median of 0.935.

Hypothesis 2 was not supported; BSA and Aspartic Acid were not always the top reagents for classification.