CSCE 590-1: Course Project
Florida water analysis - Algal bloom detection

By - Fawad Kirmani
Nov 2021

1) Introduction

Algae are plant-like organisms that sustain marine life. They contribute to the food chain and to the oxygen that keeps water bodies healthy. But sometimes when conditions are right—warm water and increased nutrients—certain algae can quickly grow and overpopulate. These foam- or scum-like masses are called blooms, and can be pushed to the shore by winds, waves, tides, and currents. Some blooms release toxins that make ecosystems, animals, and people sick: scientists call these harmful algae blooms or HABs. In Florida, we find HABs along saltwater, freshwater, and brackish water bodies [1]. Florida in recent years have seen wide spread of HABs in its multiple water bodies. In this project we are trying to analyze the water bodies for algal bloom. We have analyzed the water datasets from Florida keeping in mind the water cleaning authorities. For this project, we have built machine leaning models with explainable AI methods.

2) Data Source

We have downloaded from *wateratlas.org*. WaterAtlas is maintained by Water Institute at University of Florida, Tampa. There is total 11 sponsored atlases at *wateratlas.org*:

- Coastal & Heartland National Estuary Partnership (CHNEP)
- Florida Atlas of Lakes
- Hillsborough County
- Lake County
- Manatee County
- Orange County
- Pinellas County
- Polk County
- Sarasota County
- Seminole County
- Tampa Bay Estuary Atlas

We have downloaded three sets of data containing algal bloom information from Pineallas County Wateratlas [2]. We have downloaded three sets of data:

- Historical surface water quality dataset with algal bloom information
- Latest five years of surface water quality dataset with algal bloom information
- Latest two years of surface water quality dataset with algal bloom information

We have downloaded all the parameters available for the dataset. We were not able to work on historical dataset as that data contained issues with multiple columns. We have worked on latest two and five years of data. We started working with two years of data, but that dataset contained very few records of algal bloom. So, we switched to five years of dataset. We have thoroughly analyzed the data with five years of information.

We have added two labels in the datasets:

- **Label 2:** Algal (denoted by 1) or No Algal (denoted by 0)
- **Label 1:** Suspicion of Algal (denoted by 1) or Evidence of Algal (denoted by 2) or Heavy Algal (denoted by 3) or No Algal (denoted by 0)

For this report we have only considered Label 2. We have created Label 1 for future analysis.

3) AI Method

We have:

- Performed data exploration
- Removed every column where all the values are NULL
- Performed correlation between metrics
- Performed imputation of missing values
- Performed feature importance
- Build classification models to classify data into algal bloom and no algal bloom

a) Data Columns and Format

Here is the list of downloaded data columns:

- 'WBodyID',
- 'WaterBodyName',
- 'DataSource',
- 'StationID',
- 'StationName',
- 'Actual_StationID',
- 'Actual_Latitude',
- 'Actual_Longitude',
- 'DEP_WBID',
- 'SampleDate',
- 'ActivityDepth',
- 'DepthUnits',
- 'Parameter',
- 'Characteristic',
- 'Sample_Fraction',
- 'Result_Value',
- 'Result_Unit',
- 'QACode',
- 'Result_Comment',
- 'Original_Result_Value',
- 'Original_Result_Unit'

The dataset is in row format. The *Parameter* column has all metrics information. To convert data format to column format we have used python pandas *pivot_table* function. For indexing, we have used *SampleDate*, *ActivityDepth*, *WBodyID*, and *Result_Comment* columns. We have pivoted *Parameter* column to row format with values as *Result_Value*. After converting the data into column format, the dataset has 25216 records with 148 columns.

b) Checking Missing Values

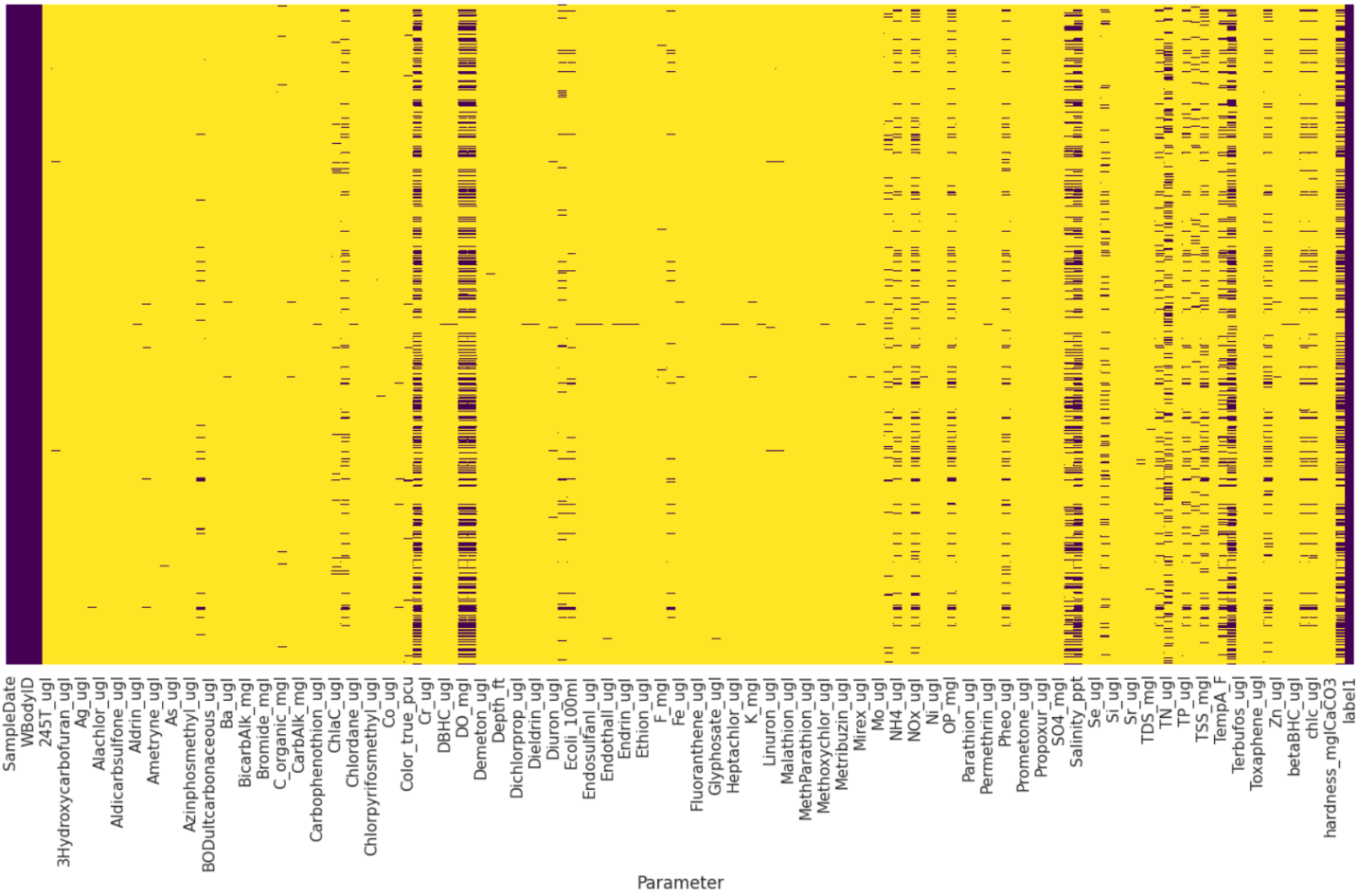


Figure 1: Missing Values Heatmap

Figure 1 is a heatmap showing missing values in five-year dataset from PINEALLAS county wateratlas. Most of the columns have very high number of missing values. We have removed the features where every value was NULL. For classification models, we have imputed the missing values in the remaining features.

c) Correlation Heatmaps

We have performed the correlation between features after creating both Label 1 and Label 2. The correlations graphs below are taken with lot of missing data. Figure 1 shows the correlation heatmap of all the features with Label 1. Figure 2 shows the correlation heatmap of all the features with Label 2. We have performed the correlation without any data imputation for both Figure 1 and Figure 2.

Few observations from Figure 2 and Figure 3. The Salinity (Salinity_PSS) is highly correlated with Specific conductance (Cond_umhocm), Nitrogen (ammonia as N) (NH3_N_ugl), Total Suspended Solids (TSS), Pheophytin-a (Pheo_ugl) and Chlorophyll a (ChlaC_ugl). The temperature TempW_C is highly correlated True color (Color_true_pcu). The pH is also highly correlated with True color (Color_true_pcu). Both Label 1 and Label 2 in Figures 2 and 3 respectively are not highly correlated with any of the features.

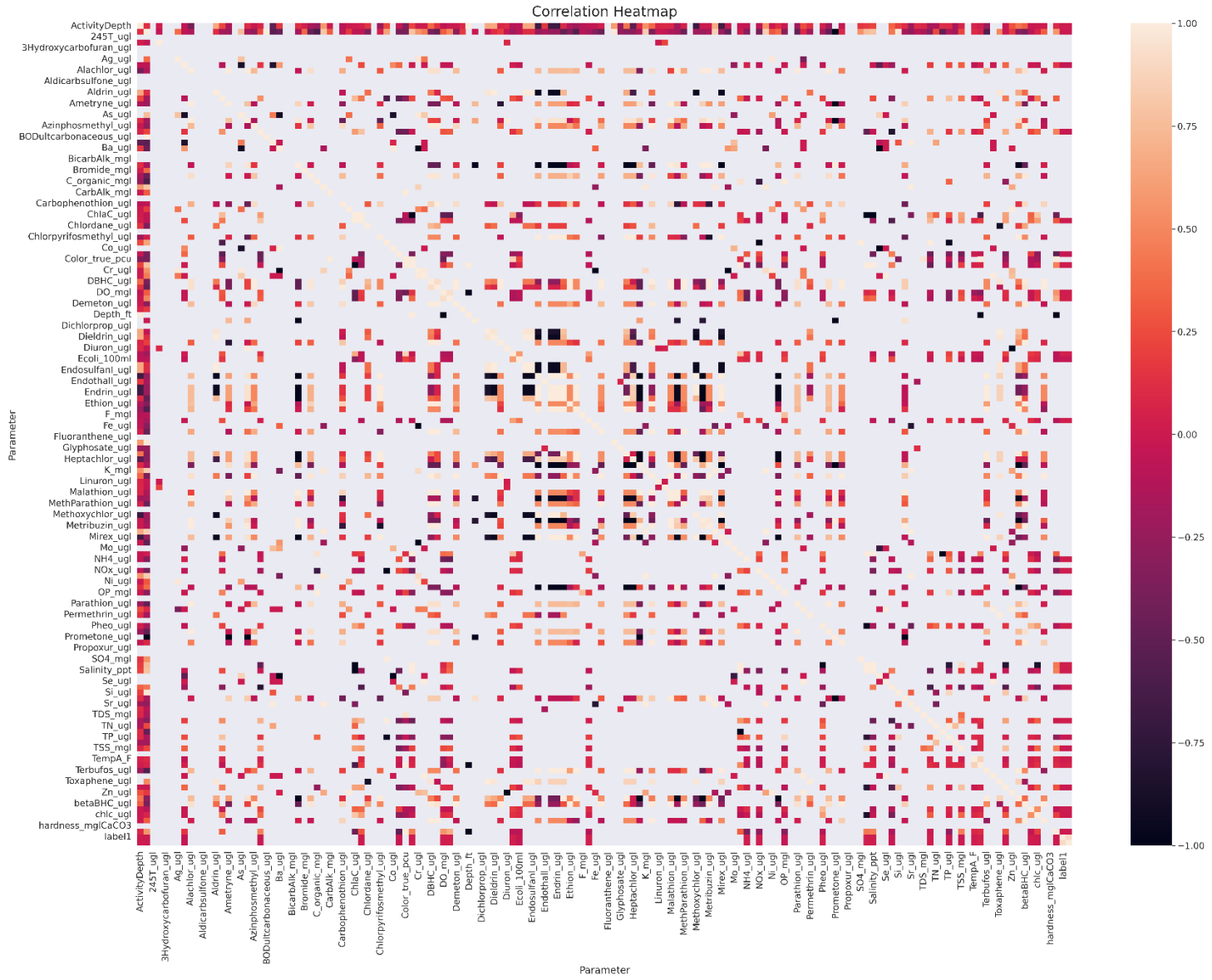


Figure 2: Correlation Heatmap with Label 1

d) Classification

To overcome missing data, we have employed data imputation. We have used *mean* for data imputation. After imputing data, we have built and tested three classification models with all 145 features:

1. XGBoost Classifier
2. Random Forest Classifier
3. Logistic Regression Classifier

We have divided the dataset into 80:20 ratio for training and testing.

I. XGBoost Classifier with all features

We have trained the XGBoost classifier with all the features. To get best hyperparameter for XGBoost classification, we have used Grid Search with 3-fold cross-validation. We achieved ROC-AUC score of 94%. We achieved specificity of 91% and sensitivity of 64%. Figure 4 shows confusion matrix and ROC-AUC curve for the XGBoost classification model trained on all features.

II. Random Forest with all features

We have trained the Random Forest classifier with all the features. To get best hyperparameter for Random Forest classification, we have used Grid Search with 3-fold cross-validation. We achieved ROC-AUC score of 88%. We achieved specificity of 76% and sensitivity of 89%. Figure 5 shows confusion matrix and ROC-AUC curve for the Random Forest classification model trained on all features.

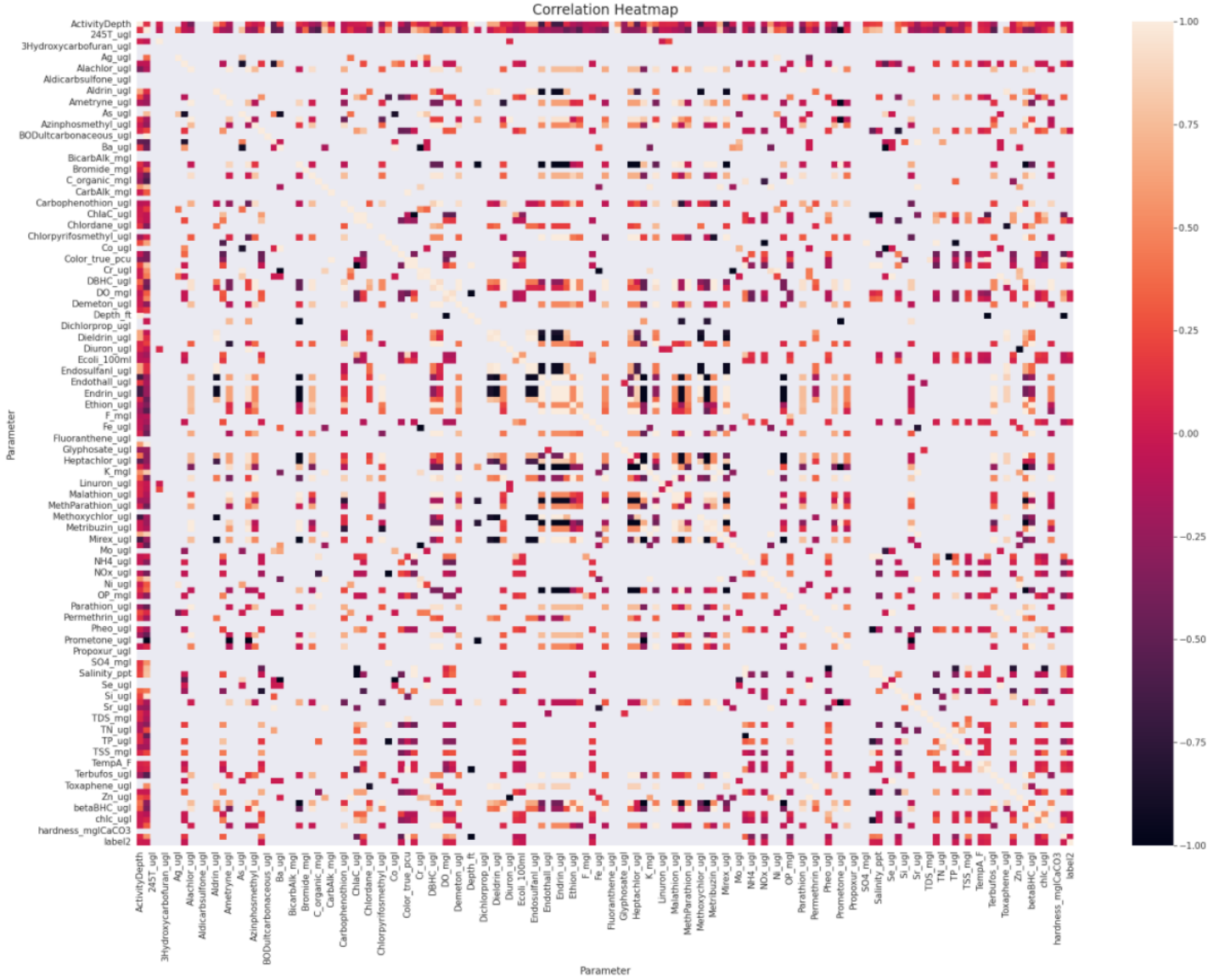


Figure 3: Correlation Heatmap with Label 2

III. Logistic Regression with all features

We have trained the Logistic Regression classifier with all the features. To get best hyperparameter for Logistic Regression classification, we have used Grid Search with 3-fold cross-validation. We achieved ROC-AUC score of 73%. We achieved specificity of 81% and sensitivity of 61%. Figure 6 shows confusion matrix and ROC-AUC curve for the Logistic Regression classification model trained on all features.

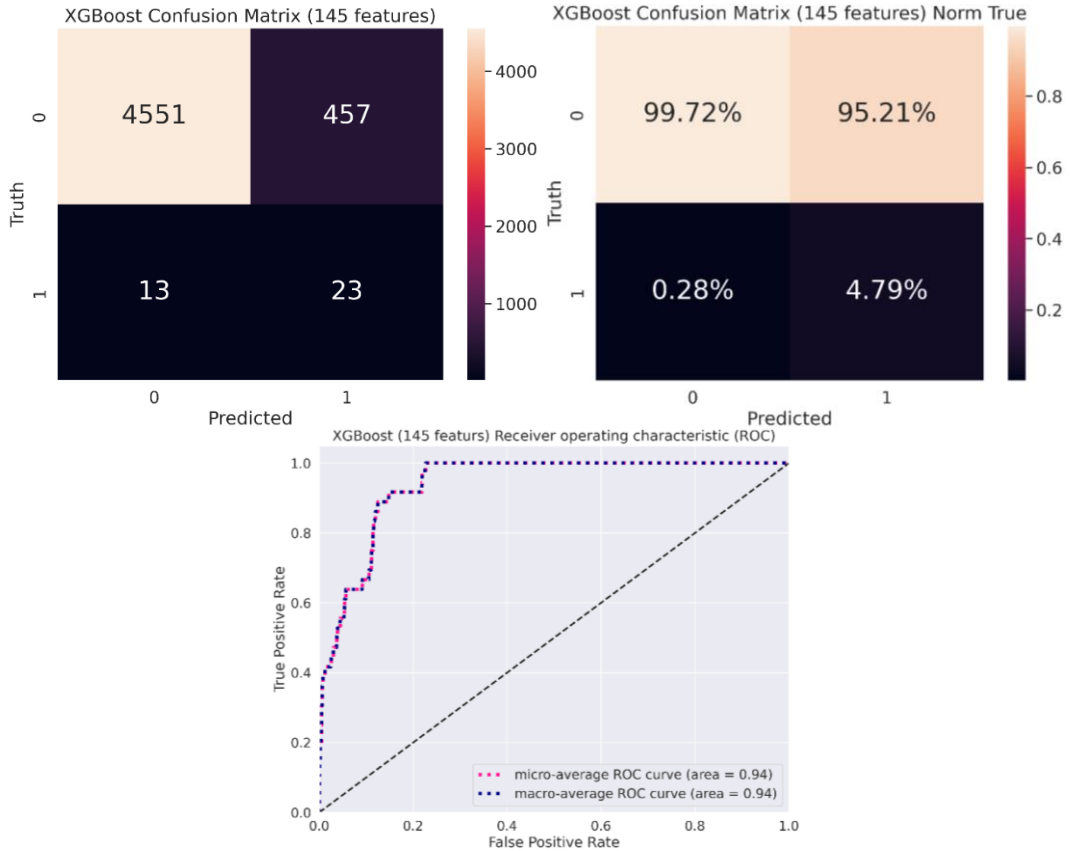


Figure 4: The top left figure is showing the confusion matrix of the XGBoost classification model with all the features. The top right figure is showing normalized confusion matrix by True values of the XGBoost classification model. The bottom figure is showing the ROC-AUC curve of the XGBoost classification model.

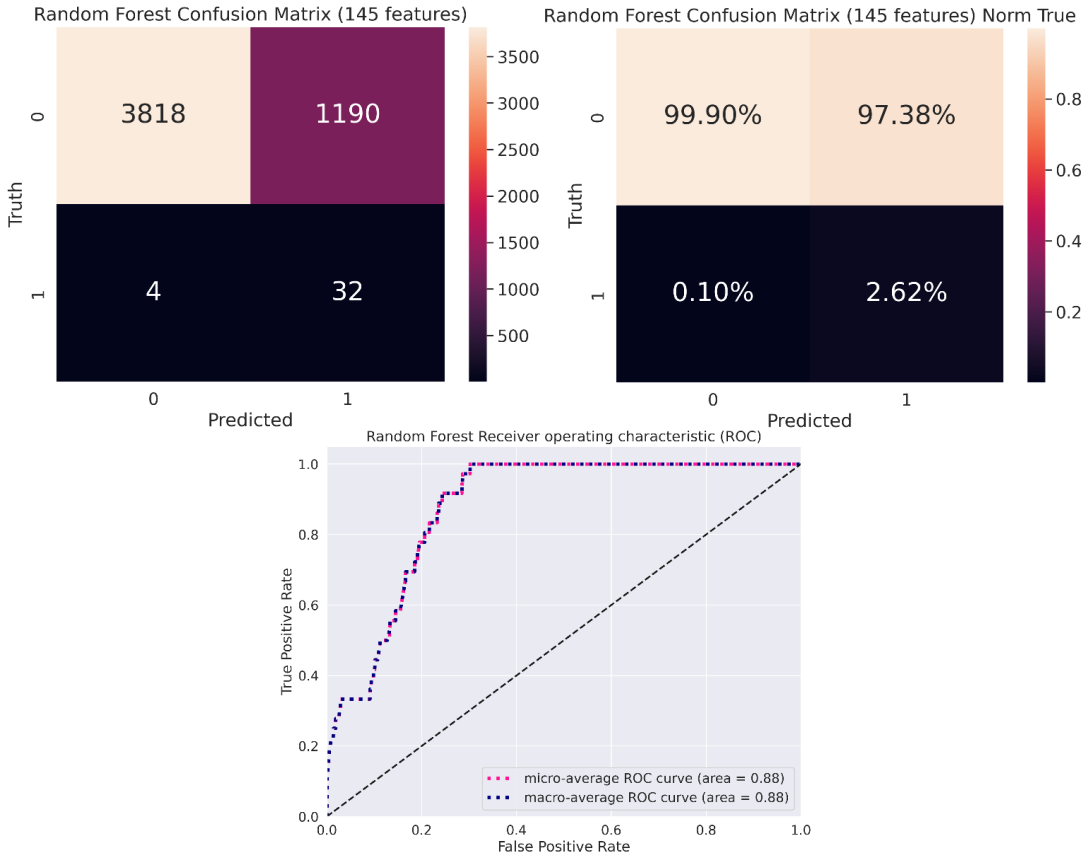


Figure 5: The top left figure is showing the confusion matrix Random Forest classification model with all the features. The top right figure is showing normalized confusion matrix by True values for Random Forest classification. The bottom figure is showing the ROC-AUC curve of the Random Forest classification model.

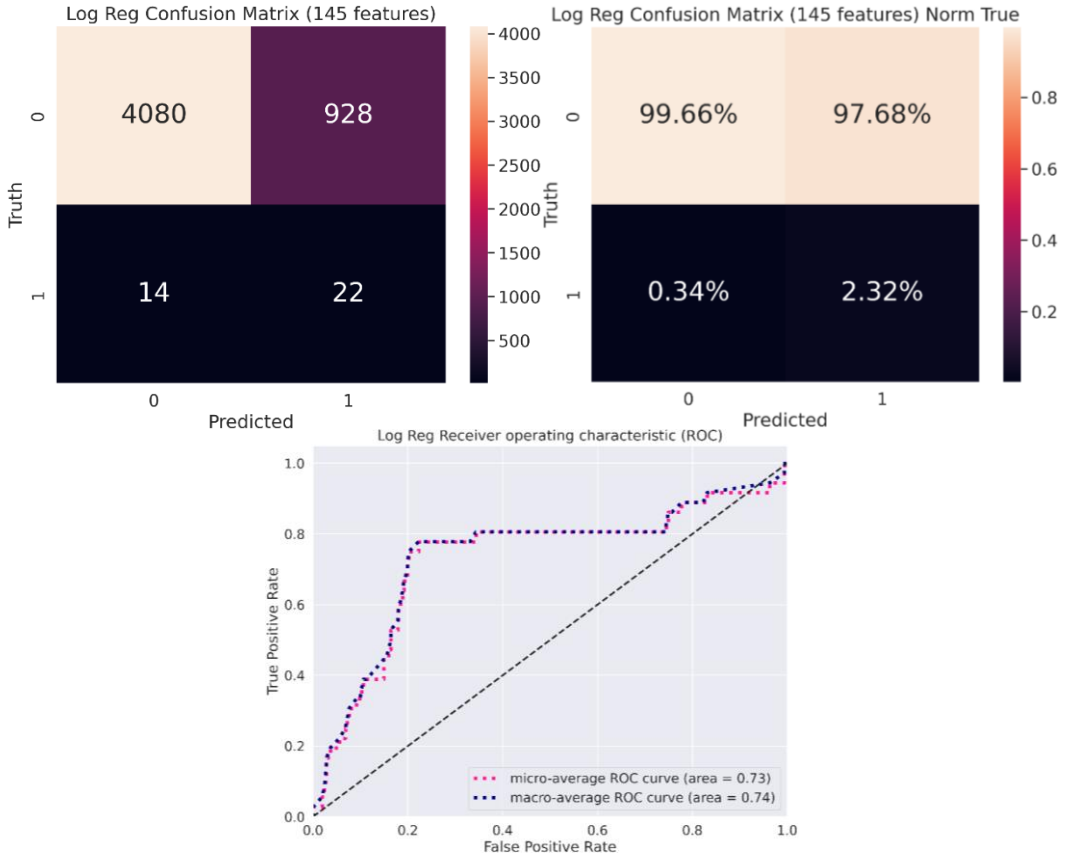


Figure 6: The top left figure is showing the confusion matrix of Logistic Regression model build with all the features. The top right figure is showing normalized confusion matrix by True values of the Logistic Regression. The bottom figure is showing the ROC-AUC curve of the Logistic Regression classification model.

a) XGBoost Feature Importance

To reduce complexity in reading and explaining the model, we have performed XGBoost Feature Importance as XGBoost had the highest AUC score in section 3.d.iii. We have used *weight* as the measure to calculate the XGBoost feature importance. The "weight" is the number of times a feature appears in a tree. The top features from feature importance analysis we got are pH, Dissolved Oxygen percentage, Salinity, Water Temperature, Air Temperature, and Dissolved Oxygen MPG. Figure 8 shows the feature importance score for the top six features. We have also done the correlation analysis these top six features. Figure 9 shows the correlation heatmap and correlation scores for these six XGBoost important features.

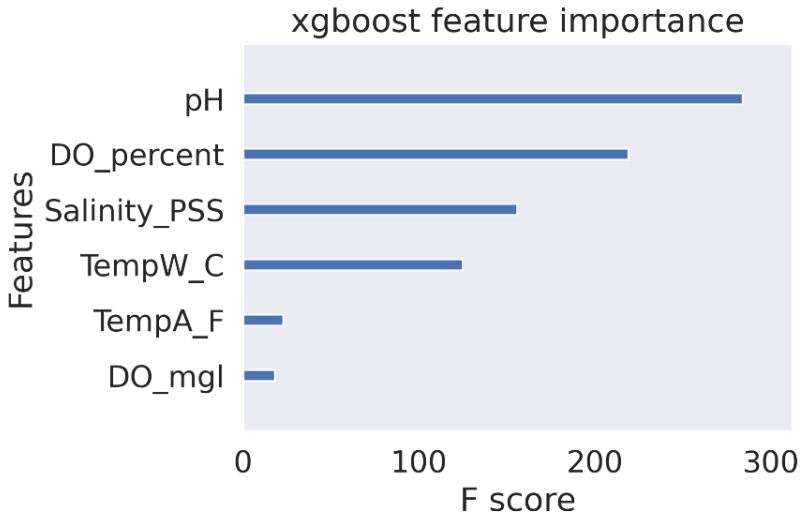
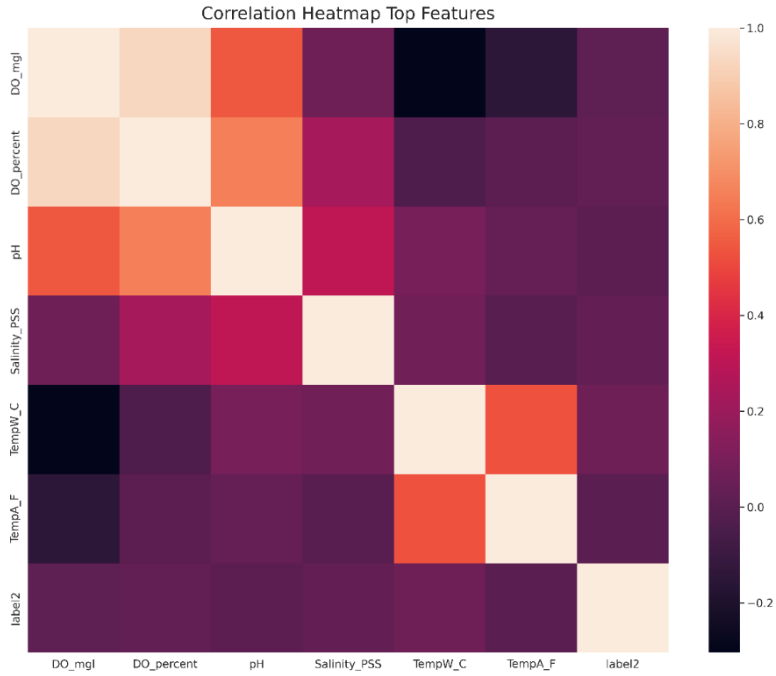


Figure 7: XGBoost Feature Importance



	DO_mgl	DO_percent	pH	Salinity_PSS	TempW_C	TempA_F	label2
DO_mgl	1.000000	0.930895	0.543349	0.060340	-0.303667	-0.148847	0.015778
DO_percent	0.930895	1.000000	0.652375	0.232486	-0.037663	0.010362	0.027018
pH	0.543349	0.652375	1.000000	0.309377	0.091472	0.038972	0.009681
Salinity_PSS	0.060340	0.232486	0.309377	1.000000	0.077212	-0.000005	0.030637
TempW_C	-0.303667	-0.037663	0.091472	0.077212	1.000000	0.530235	0.059903
TempA_F	-0.148847	0.010362	0.038972	-0.000005	0.530235	1.000000	0.005229
label2	0.015778	0.027018	0.009681	0.030637	0.059903	0.005229	1.000000

Figure 8: The top figure shows the Correlation Heatmap for XGBoost Important Features. The bottom figure shows the correlation score of these important features

b) XGBoost Classification model with Top Features

We have trained the XGBoost classifier with all the features. To get best hyperparameter for XGBoost classification, we have used Grid Search with 3-fold cross-validation. We achieved ROC-AUC score of 93%. We achieved specificity of 82% and sensitivity of 83%. This model gave almost similar ACU score of 93% compared to the XGBoost model trained on all features. This model trained on only six featured also gave better sensitivity score of 83% compared to the XGBoost model trained on all features.

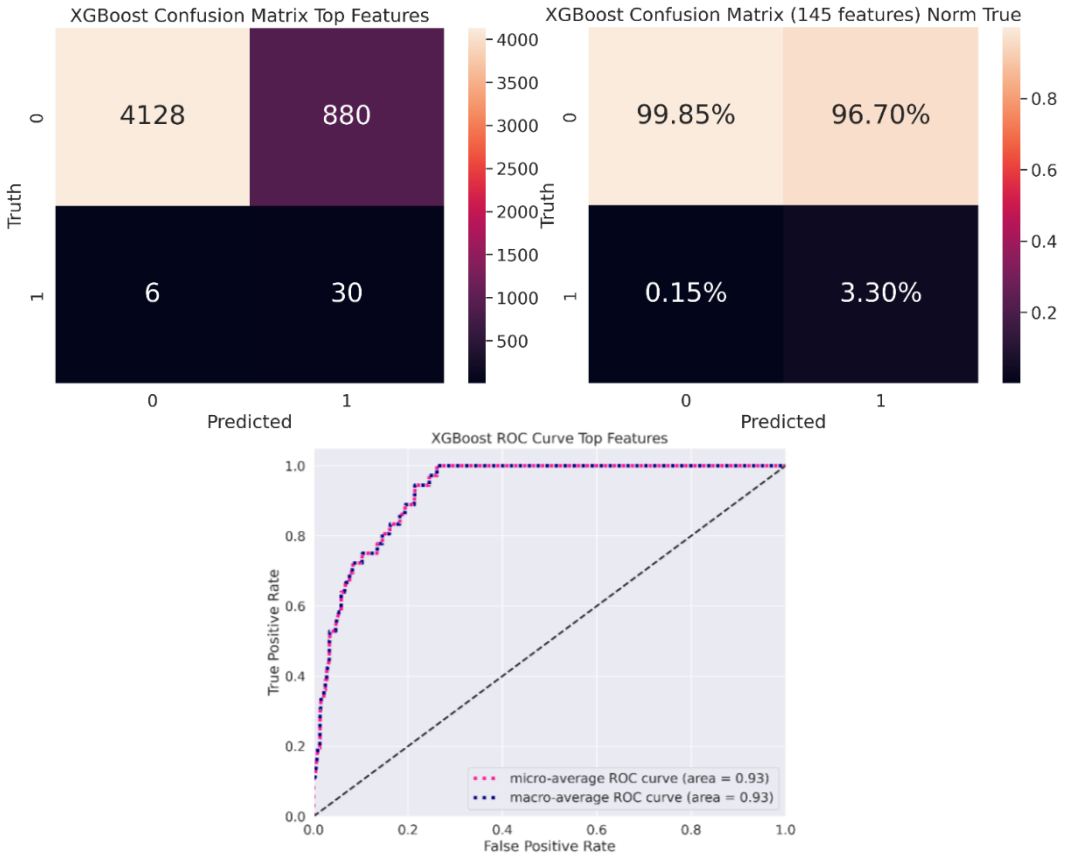


Figure 9: Top left figure is showing the confusion matrix of XGBoost classification with top six features. Top right figure is showing normalized confusion matrix by True values of XGBoost classification with top features. The bottom figure is showing the ROC-AUC curve of the XGBoost classification model with top six features.

Table 1 summarizes the classification scores we achieved with different AI methodologies.

Table 1: Water Bodies Algae Detection Classification Scores

Algorithm	No. of Features	AUC	Sensitivity	Specificity
XGBoost	145	94%	64%	91%
Random Forest	145	88%	89%	76%
Logistic Regression	145	74%	61%	81%
XGBoost	6	93%	83%	82%

4. Human AI

a. LIME

LIME [3] generates a local, linear explanation for any model. It perturbs near the neighborhood of a point of interest, X (Local). Then, it fits a linear function to the model's output (Linear). It then interprets coefficients of the linear function (Explain). Then we can visualize the results through it. LIME can be applied to any classification model.

We have applied LIME to the XGBoost model trained with top six feature only. We have chosen this model because:

- It gave comparable classification scores (AUC, sensitivity, and specificity) compared to the models trained with all the features.
- It is easy explain.
- It is easy to concentrate on top features.

Figure 10 shows the LIME output for a particular record. In the figure, it can be seen that with certain values of water temperature, pH, dissolved oxygen, and salinity for the given row/record, the model is predicting with 93% probability that the water has no algae.

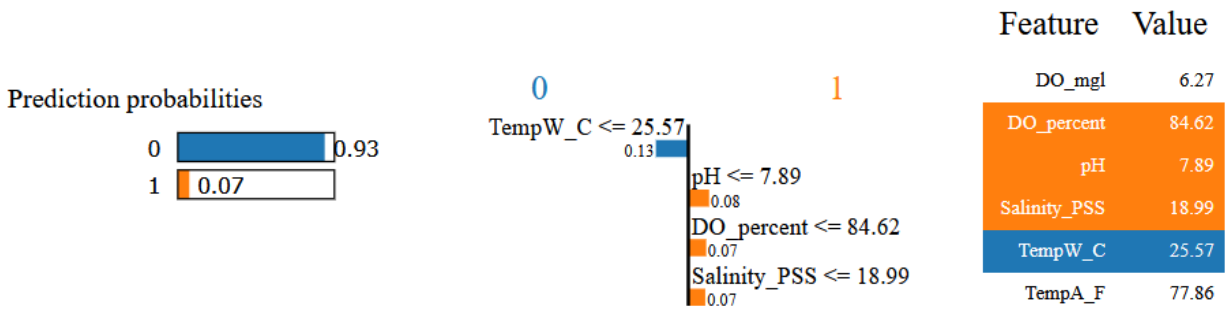


Figure 10: LIME prediction probabilities for a record in test dataset

In Figure 11, we are showing the local linear approximation. By changing the values of each of the top features, we can see how the model prediction will change for that row/record.

```
Decreasing pH to 7.5
P(algal) before: 0.92773473
P(algal) after: 0.884479

Decreasing Salinity_PSS to 17.0
P(algal) before: 0.92773473
P(algal) after: 0.9180464

Decreasing DO_percent to 75.0
P(algal) before: 0.92773473
P(algal) after: 0.78915745

Decreasing TempW_C to 24.0
P(algal) before: 0.92773473
P(algal) after: 0.69907695

Decreasing pH, TempW_C, DO_percent, and Salinity_PSS and TempW_C
P(algal) before: 0.92773473
P(algal) after: 0.2675742
```

Figure 11: Local Linear Approximation of the record in Figure 10

Previous research work supports above findings that specific range of Dissolved Oxygen, pH, Salinity and Temperature are very important factors in harmful algal bloom [5][6][7][8].

b) AIX360:

AIX360 is a toolbox which can be used to explain AI models to different stakeholders to solve problem. We have used Protodash Explainer to explain the model to different authorities working to clean the water bodies in Florida from algae bloom [4].

The protodash method can select in a deterministic fashion examples from a dataset, which we term as prototypes that represent the different segments in a dataset. The protodash method is deterministic and not randomized like k-medoids clustering where center is randomly initialized. So, the solutions with protodash are repeatable, and it picks prototypes that are representative as well as diverse, which is important in practical settings where we want to capture all the different segments/modes in the dataset, not missing any of the key behaviors. Another benefit of the protodash method is that, in principle, it can also be applied in non-iid settings such as for time series data since it

performs distribution matching between user/users in question and those available in the dataset. Other approaches which find similar profiles using standard distance measures (viz. euclidean, cosine) do not have this property. Additionally, we can also highlight important features for the different prototypes that made them similar to the user/users in question [3][4].

Below is an example to Protodash explainer on a sample from the test data. Figures 12 and 14 show the similar samples as explanations for a water data predicted as “No Algal (0)” and “Algal (1)” respectively. Figures 13 and 15 show how similar a feature of a prototypical water body is to the chosen water, predicted as "No Algal (0)" and “Algal (1)” respectively.

	0	1	2	3	4
DO_mgl	6.265108	6.265108	5.250000	6.265108	4.19000
DO_percent	84.617505	84.617505	82.600000	84.617505	67.70000
pH	7.892637	8.650000	8.290000	7.892637	8.08000
Salinity_PSS	18.985390	18.985390	32.380000	18.985390	18.98539
TempW_C	25.570452	25.570452	29.800000	25.570452	31.34000
TempA_F	77.864184	77.864184	77.864184	77.864184	88.70000
label2	0.000000	0.000000	0.000000	0.000000	0.00000
Weight	1.000000	0.000000	0.000000	0.000000	0.00000

Figure 12: Similar samples as explanations for a water data predicted as "No Algal (0)"

	0	1	2	3	4
DO_mgl	1.0	1.00	0.29	1.0	0.08
DO_percent	1.0	1.00	0.74	1.0	0.08
pH	1.0	0.07	0.25	1.0	0.52
Salinity_PSS	1.0	1.00	0.08	1.0	1.00
TempW_C	1.0	1.00	0.18	1.0	0.10
TempA_F	1.0	1.00	1.00	1.0	0.08

Figure 13: How similar a feature of a prototypical water body is to the chosen water (predicted as "No Algal (0)")

	0	1	2	3	4
DO_mgl	8.820000	5.620000	7.090000	7.020000	5.42000
DO_percent	130.000000	78.900000	115.500000	103.500000	75.70000
pH	8.530000	7.910000	8.300000	8.250000	7.77000
Salinity_PSS	20.930000	19.780000	29.650000	24.660000	18.98539
TempW_C	29.820000	27.055600	32.770000	28.400000	24.53000
TempA_F	77.864184	77.864184	77.864184	77.864184	79.34000
label2	1.000000	1.000000	1.000000	1.000000	1.00000
Weight	1.000000	0.000000	0.000000	0.000000	0.00000

Figure 13: Similar samples as explanations for a water data predicted as "Algal (1)"

	0	1	2	3	4
DO_mgl	1.0	0.07	0.24	0.23	0.06
DO_percent	1.0	0.09	0.50	0.28	0.07
pH	1.0	0.11	0.43	0.36	0.06
Salinity_PSS	1.0	0.75	0.11	0.39	0.61
TempW_C	1.0	0.37	0.34	0.60	0.15
TempA_F	1.0	1.00	1.00	1.00	0.08

Figure 14: How similar a feature of a prototypical water body is to the chosen water body (predicted as " Algal (1)")

5. Conclusion

In this project we presented an approach to detect algal bloom in water bodies. We presented results from an XGBoost classification model trained only with six features giving AUC score of 93%. We presented the top six features which cause harmful algal bloom from our analysis. Four of the top six features i.e., pH, dissolved oxygen, salinity, and temperature are supported by previous studies as well [5][6][7][8][9]. In this project, we also tried to explain the model results through LIME and AIX360.

6. Future Work

In future, we believe that this project should concentrate our analysis on affected water bodies. We should build a time series model for affected water bodies. We should then apply those learnings to the water bodies which susceptible to algae boom. We should also integrate demographics data for algae bloom affected areas with the water sample data and look for the most affected demographics with algae bloom. We should focus on building explainable machine learning models for affected areas and water bodies

Reference:

1. HABs: Harmful Algae Blooms | Florida Department of Health (floridahealth.gov)
2. Maps / Data - Pinellas.WaterAtlas.org (usf.edu)
3. [1602.04938] "Why Should I Trust You?": Explaining the Predictions of Any Classifier (arxiv.org)
4. [1707.01212] Efficient Data Representation by Selecting Prototypes with Importance Weights (arxiv.org)
5. AIX360/HELOC.ipynb at master · Trusted-AI/AIX360 (github.com)
6. What causes algal blooms? | Center for Earth and Environmental Science (iupui.edu)
7. Impacts of Climate Change on the Occurrence of Harmful Algal Blooms (epa.gov)
8. Researchers Study the Effects of Harmful Algal Blooms | NOAA Fisheries
9. Algal Blooms are Blooming | NEEF (neefusa.org)
10. Low pH levels can eliminate harmful blooms of golden algae, one cause of massive fish kills -- ScienceDaily