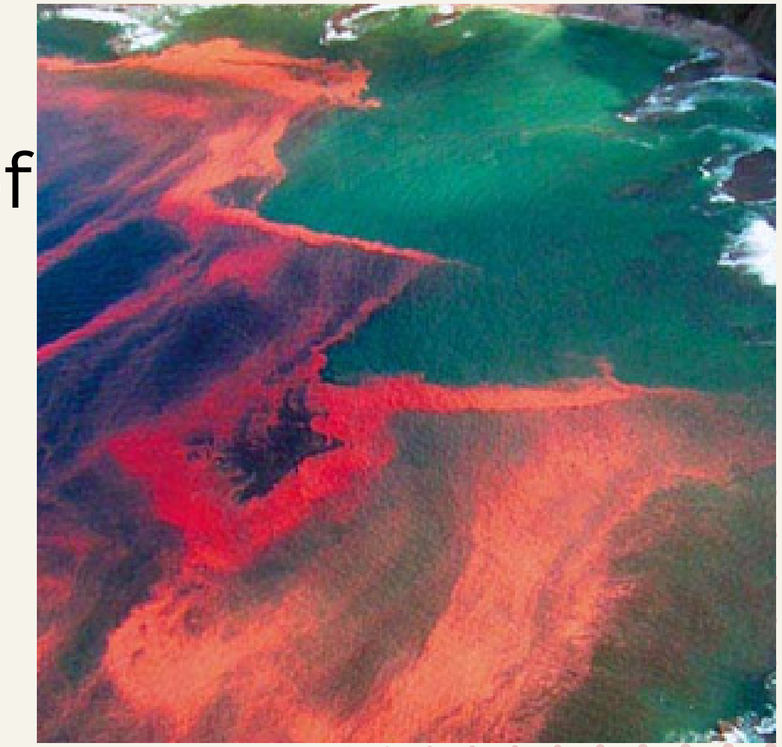
PREDICTING RISKS OF TOXIC ALGAL BLOOMS

Sandre Ann V. Eugenio

Applied Physics 157

BACKGROUND

Harmful Algal Bloom
-grow rapidly and release large amount of toxins



- Public health concern
- Economic risk

MATERIALS AND METHODS

DATA SOURCES

- Southern California Coastal Ocean Observing System
- Pacific Fisheries Environmental Laboratory
- National Oceanic and Atmospheric Administration

2750 samples, 18 variables

	Akashiwo sanguinea (cells/L)	Alexandrium spp. (cells/L)	Ammonia (uM)	Chlorophyll (mg/m3)	Dinophysis spp. (cells/L)	Domoic Acid (ng/mL)	Lingulodinium polyedrum (cells/L)	Nitrate (uM)	Nitrite (uM)	Phosphate (uM)	Prorocentrum spp. (cells/L)	nitzschia delicatissima group (cells/L)	nitzschia seriata group (cells/L)	Sili	cate (uM)	Water Temperature (C)	date	Upwelling Index	ONI
0	0.000000	0.00000	0.470000	8.17	0.000000	0.000	0.000000	7.940000	0.260000	0.750000	0.000000	24550.000000	357041.0	16.25	0000	11.7	2013-05-24	216.0	-0.3
1	0.000000	0.00000	11.545000	1.57	8353.000000	0.000	0.000000	0.975000	0.150000	1.035000	0.000000	232.000000	2320.0	7.19	5000	15.0	2011-11-21	6.5	-1.1
2	937.241026	280.18417	6.191939	3.59	217.504717	0.000	968.184615	3.484584	0.417262	0.703519	4155.671795	8530.082051	2509.0	7.95	8344	15.7	2016-01-04	-5.0	2.5
3	0.000000	0.00000	3.120000	2.79	0.000000	0.000	0.000000	0.510000	0.337276	0.270000	0.000000	46778.000000	33784.0	1.59	0000	16.2	2016-03-28	318.0	1.7
4	400.000000	0.00000	0.230000	11.08	1200.000000	0.044	201200.000000	0.460000	0.040000	0.370000	30400.000000	10400.000000	1200.0	5.90	0000	16.9	2010-05-24	82.0	-0.1

DATA PROCESSING

- Dataset was shuffled
- Dependent variable: Domoic Acid
  - -Converted into binary variables
  - -Applied threshold (values>0.5: positive case/dangerous level)

```
#Load the data

data = pd.read_csv('clean_data.csv')
data = data.sample(frac=1, random_state=42).reset_index(drop=True) #shuffle
data.head()
```

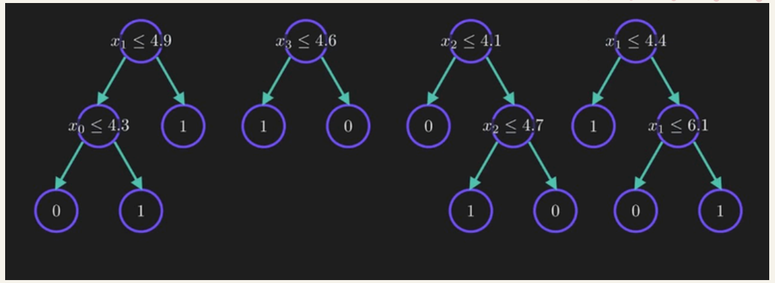
```
#predictor variable
target_data = (data['Domoic Acid (ng/mL)']>0.5) #apply threshold; value>0.5=positive case

#extract for future use
domoic_acid = data['Domoic Acid (ng/mL)']

#remove column from the DataFrame
data = data.drop('Domoic Acid (ng/mL)', axis=1)
```

MACHINE LEARNING

- Data was scaled to normalize.
- Data split into training (75%) and test (25%) sets.
- Random forest classifier is used.
  Hyperparameters were tuned using GridSearchCV.



CRITICAL PARAMETERS AND VALIDATION METHODS

Hyperparameters:

- n_estimators: Number of trees (10, 50, 100, 200)
- max_depth: Maximum depth of the trees (10, 20, None)
- min_samples_split: Minimum number of samples required to split a node (2, 10, 20, 50)
- min_samples_leaf: Minimum number of samples required to form a leaf node (1, 5, 10, 20)

```
model = GridSearchCV(pipeline,params, scoring ='average_precision', n_jobs=-1)
model.fit(xtrain,ytrain)
print('Best Parameters', model.best_params_)
print('Best average precision', model.best_score_)
```

IMPLEMENTATION AND METRIC

Python libraries used:

- pandas
- numpy
- matplotlib
- sklearn
- imblearn

Pipeline :

SMOTE (Synthetic Minority Over-sampling Technique)

RandomForestClassifier

Model Training and Hyperparameter Tuning:

-The model was fitted using the training data, and hyperparameter tuning was performed with grid search.

Evaluation

-The best parameters obtained from the grid search were used to evaluate the model on the test set.

```
pipeline = Pipeline([('sampling', SMOTE(random_state=42)),('rf', RandomForestClassifier(random_state=42))])
```

RESULTS

RESULTS

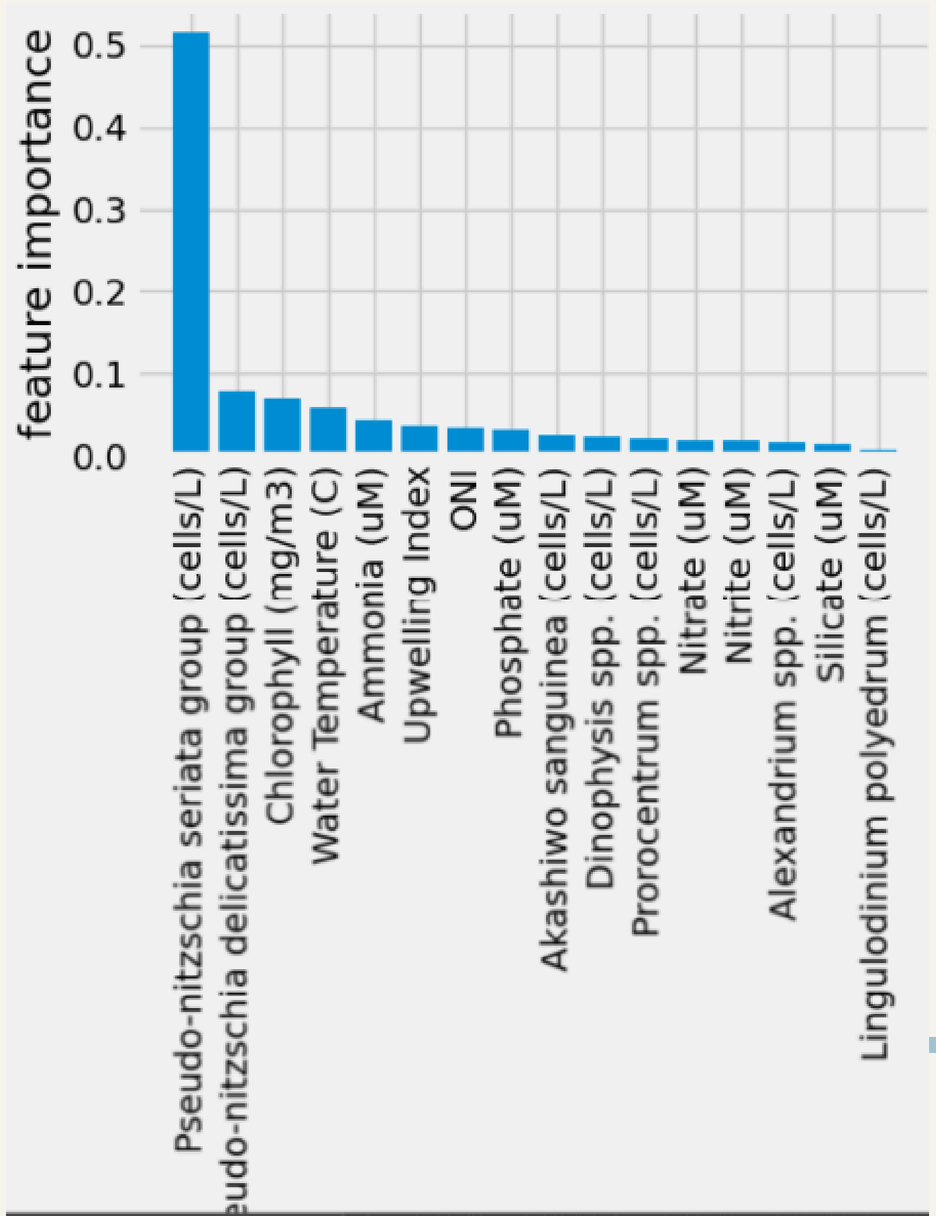
Confusion Matrix and Classification Report

True Negative = 634 False Positive = 25
False Negative = 10 True Positive=19

	precision	recall	f1-score	support
False	0.98	0.96	0.97	659
True	0.43	0.66	0.52	29
accuracy			0.95	688
macro avg	0.71	0.81	0.75	688
weighted avg	0.96	0.95	0.95	688

RESULTS

Feature importance



CONCLUSION

The Random Forest model was successful in predicting toxic algal bloom events with high accuracy and reasonable average precision. The most important features identified were related to the presence of Pseudo-nitzschia seriata, temperature, and chlorophyll concentration.

RECOMMENDATION

- Add more data
- Add more features
- Use other models to compare

REFERENCES

US Department of Commerce, N. O. and A. A. (n.d.). Predicting Harmful Algal Blooms. Oceantoday.noaa.gov. Retrieved May 23, 2024, from https://oceantoday.noaa.gov/predictinghabs/

Liu, Z. Ocean Toxin Prediction, (2018), GitHub repository, https://github.com/zxl124/Ocean-toxin-prediction

Yñiguez, A. T., & Ottong, Z. J. (2020). Predicting fish kills and toxic blooms in an intensive mariculture site in the Philippines using a machine learning model. Science of the Total Environment, 707, 136173. https://doi.org/10.1016/j.scitotenv.2019.136173

Applied Physics 157 | 2024

THANK YOU

Presented By : Sandre Ann V. Eugenio