Coral Bleaching Classification with Machine Learning

What is Coral Bleaching? Why do we care?

- Coral reefs are critical to marine life, housing diverse sea life
- Coral bleaching damages the reefs and biodiversity
  - Occurs in changing water conditions and rising temperatures
- Currently experiencing a global bleaching event[1]
- Bleaching can be reversable if monitored and managed

[1] *Source: "NOAA Confirms 4th Global Coral Bleaching Event." National Oceanic and Atmospheric Administration, 15 Apr. 2024.*

Goals

Predict the presence of bleaching* based on environmental conditions to:

1. **better understand what factors are most critical to coral bleaching, and**
2. **identify areas that are more susceptible to coral bleaching**

*defined as ≥ 5% bleaching observed

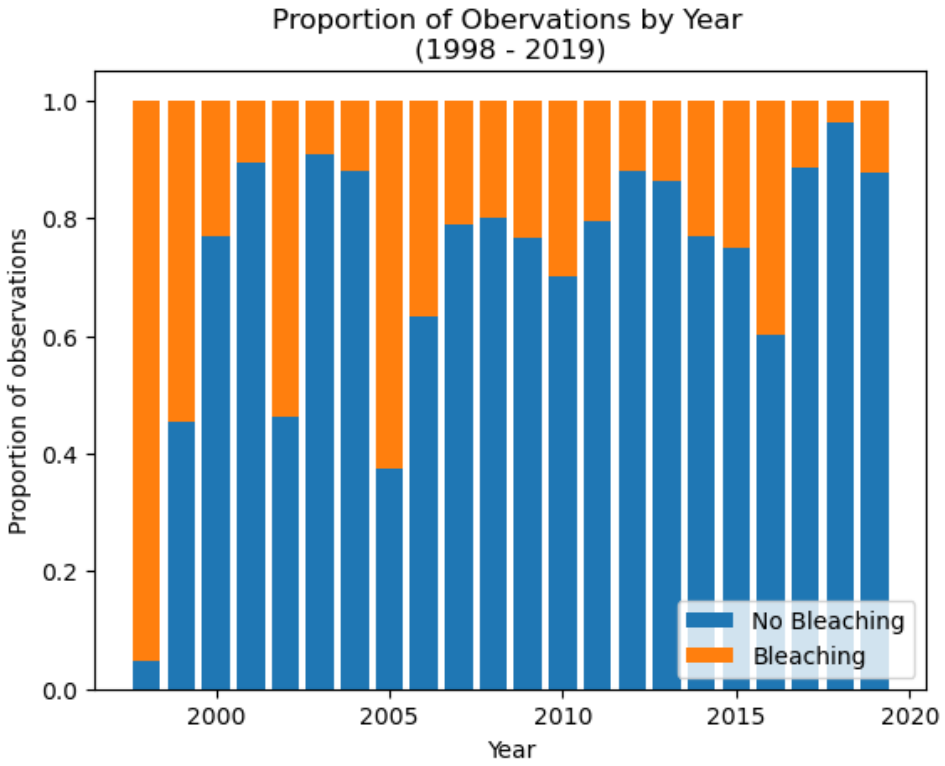
Methodology & Approach

Data Wrangling	EDA	Preprocess	Model	Evaluation & Selection	Future Development
Review and clean the dataset. Understand where missing values occur	Identify patterns and trends that appear in the data.	Prepare for modeling (sampling methods, handle missing data, scaling, and dummy variables)	Compare performance and fine tune hyper-parameters	Determine the best performing model (F1 score)	Consider further analyses that could supplement or improve the model

The Data

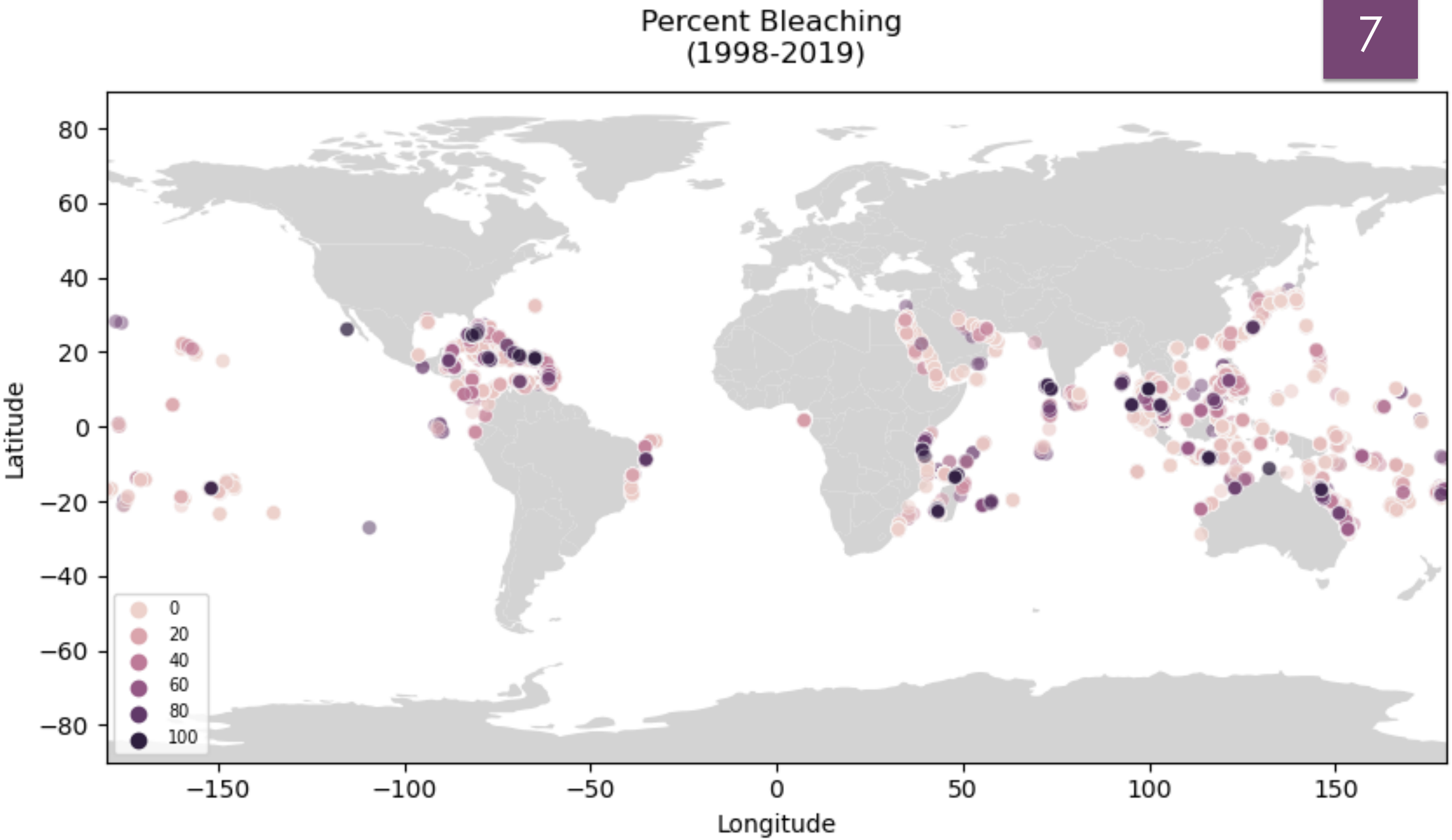
- The Global Coral-Bleaching Database (GCBD) was compiled from seven data sources
  - 34,846 coral bleaching records from 14,405 sites in 93 countries, from 1980–2020
- Observations include:
  - Percent coral bleaching observed
  - Temperature: sea surface temp., frequency and anomaly metrics
  - Environment: site exposure, depth, distance to land, turbidity, windspeeds, cyclone frequency
  - Global regions

When & where bleaching occurred?



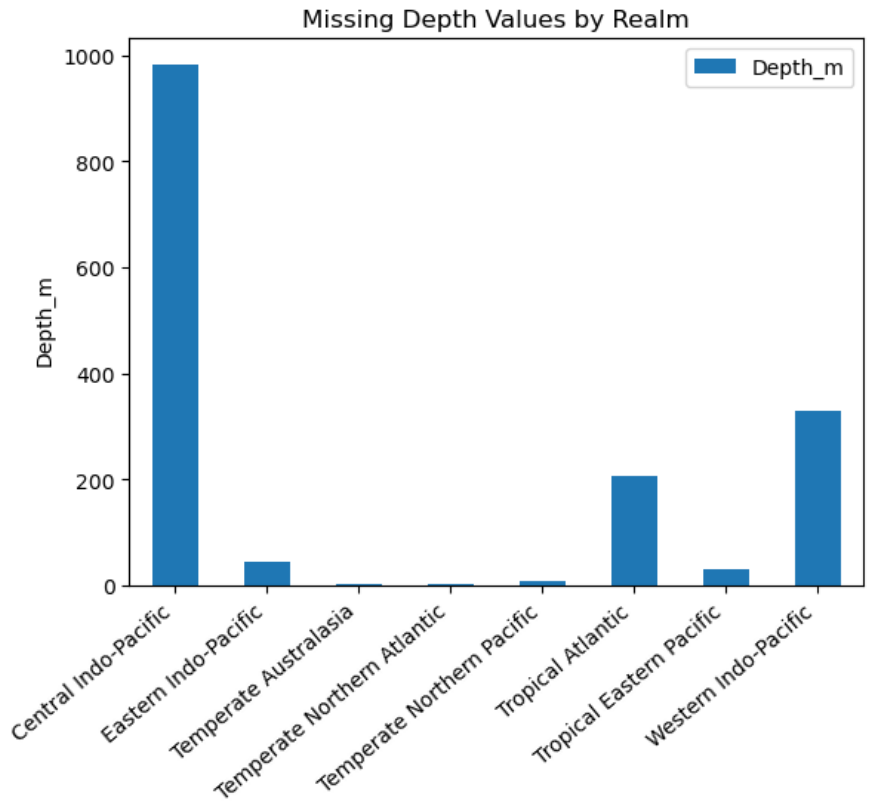
- Global bleaching events:
  - First major event recorded in 1998
  - Subsequent global events in 2010, 2014-17, and most recently 2023-24
- Notable local events:
  - Australia in 2002 and 2006
  - Caribbean in 2005

Source: "Coral bleaching events." Australian Institute of Marine Science.


Percent Bleaching
(1998-2019)
Latitude
Longitude
80
60
40
20
0
−20
−40
−60
−80
−150
−100
−50
0
50
100
150
0
20
40
60
80
100

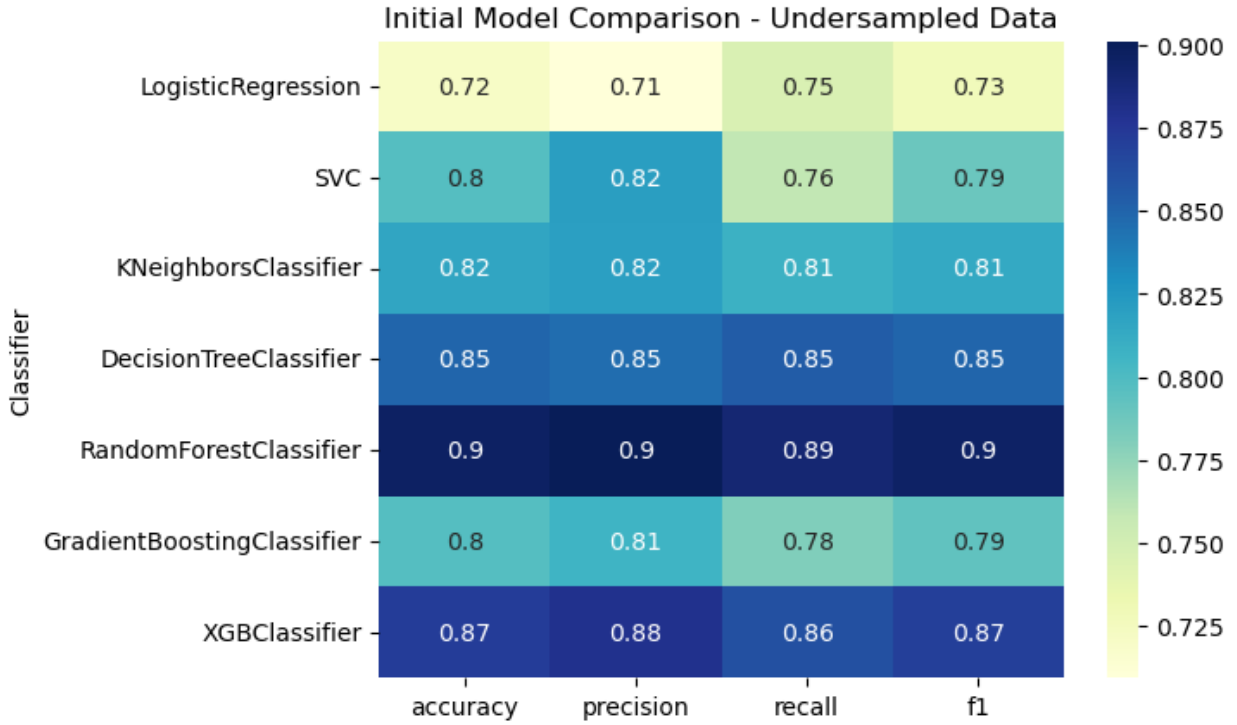
Preparing data for modeling

- Sampling methods to mitigate imbalanced data
  - Random under sampling
  - SMOTE (over sampling)
- Train/Test split (80/20)
- Handling missing data and categorical variables
- Scale for logistic regression

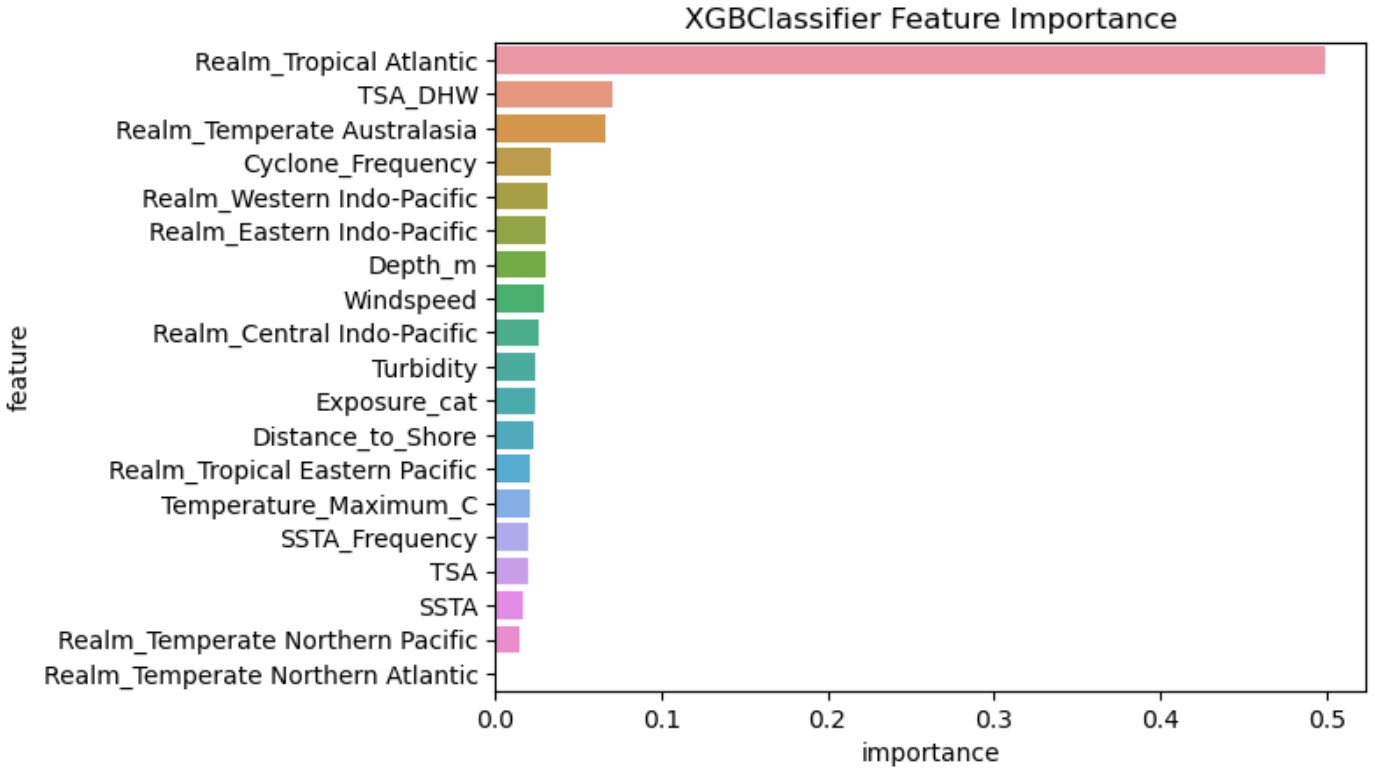


Modeling

- Consider various classification algorithms:
  - Logistic regression
  - Support vector machine
  - K-nearest neighbors
  - Tree-based algorithms



What are the best indicators?



- Regions/Realms
  - Tropical Atlantic (Caribbean)
  - Australasia (Great Barrier Reef)
- Thermal Stress Anomaly Degree Heating Week
  - How frequently a thermal anomaly occurred in the previous 12 weeks

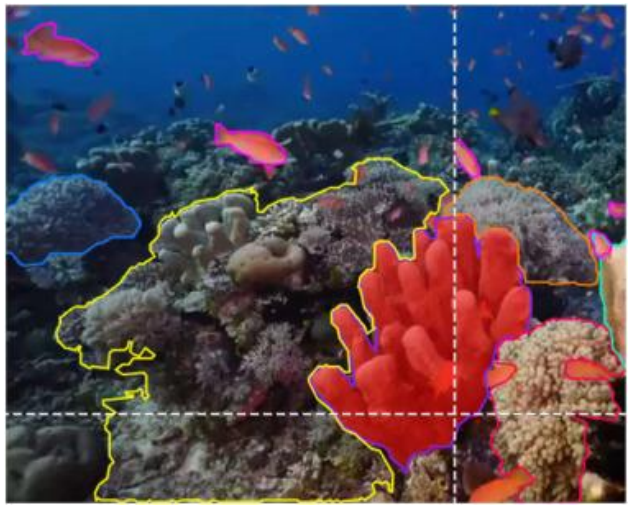
Final Performance

- Extreme Gradient Boost (XGBoost) with full features performed the best overall amongst all metrics, including the above cross validated F1, as well as checking for overfitting

	Cross Validation mean F1-score	Cross Validation STD	Train/Test Score Differential
Decision Tree (full)	0.8472	0.0064	0.1338
Decision Tree (reduced)	0.8257	0.0018	0.1409
Random Forest (full)	0.8978	0.0036	0.0883
Random Forest (reduced)	0.8701	0.0030	0.1149
XGBoost (full)	**0.8987**	**0.0042**	**0.0817**
XGBoost (reduced)	0.8477	0.0091	0.1044

Future Developments

- Improve current model:
  - Feature engineering to reduce the number of features
  - Unsupervised learning – uncover unseen clusters/patterns
- Combine with other coral health data formats (images and audio)



Questions