Coral Bleaching Prediction Model

Proposal

Coral reefs are critical to marine life, housing diverse sea life. These reefs are sensitive to water conditions and recently experienced an increase in the frequency and intensity of thermal-stress events that are causing coral bleaching. Coral bleaching can result in the damage and loss of reefs, which in turn reduces marine biodiversity. Some reefs are able to recover from bleaching if given time and appropriate conditions.

This project aims to use environmental factors to predict the presence of bleaching, defined as more than 5% bleaching observed. A model of this sort can indicate what areas are most vulnerable to coral bleaching, and help focus and concentrate conservation efforts to avoid further loss of reefs.

Data

The Global Coral-Bleaching Database (GCBD) encompasses 34,846 coral bleaching records from 14,405 sites in 93 countries, from 1980–2020. Records were obtained and compiled from seven data sources. In addition to the percent of coral bleaching observed, records include site exposure, distance to land, mean turbidity, cyclone frequency, and a suite of sea-surface temperature metrics at the times of survey.

Source: https://www.bco-dmo.org/dataset/773466

Approach

1. **Data Wrangling** – review and clean up the dataset. Understand where missing data is present.
2. **Exploratory Data Analysis** – identify any patterns or trends that appear in the data.
3. **Preprocessing** – prepare the data for modeling, including sampling methods, interpolating remaining missing data, scaling, and dummy variables.
4. **Modeling** – compare performance of various classification models, identifying a few to further tune.
5. **Model Selection** – after hyperparameter tuning, determine the best performing model based on F1 score. Cross validation and train/test score differentials are also used to reduce the risk of overfitting.
6. **Future Development** – consider further analyses that could supplement or improve the model.

1. Data Wrangling

The data was downloaded and saved from BCO-DMO, linked above. The original dataset includes 41,361 observations with 62 columns. This will be a binary classification model, and therefore we must create our target variable called *Bleaching_indicator,* defined as 0 if *Percent_Bleaching* is less than 5%, and 1 if higher.

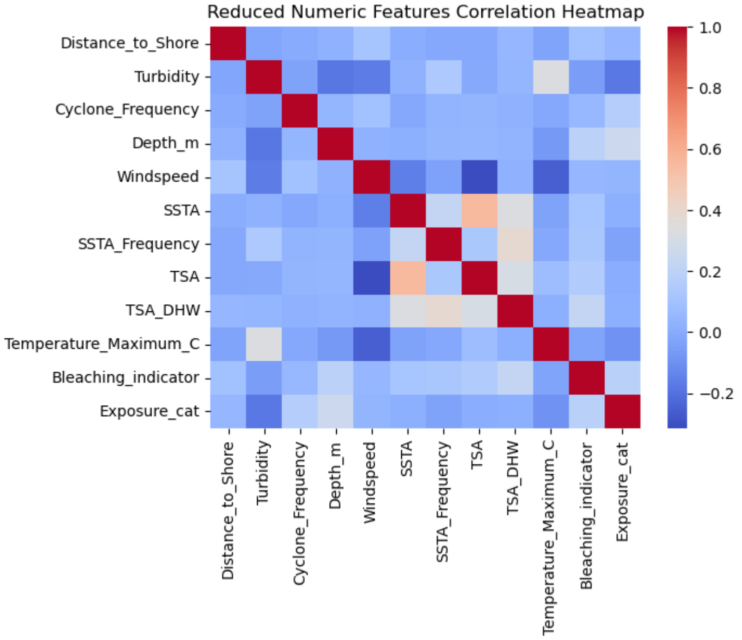
Cleaning the raw data included converting temperature metrics from Kelvin to Celsius for better intuitive interpretability, as well as identifying categorical variables that will be used for modeling. *Exposure* is ordinal and therefore can easily be expressed through numbers, as 0.0, 0.5, and 1.0 for ‘sheltered’, ‘sometimes’, and ‘exposed’ respectively. *Realm_Name* will be converted to a dummy variable in the preprocessing step.

Given the relatively large dataset with over 40,000 observations, we can drop records with missing values without compromising the model. Evaluating missing values revealed that about 16.5% of *Percent_Bleaching* values are null. Since the target variable is derived from this, these records are dropped from the analysis. *Depth_m* is the other notable feature with substantial missing values. It appears missing depth values are concentrated regionally, and would make sense to interpolate missing values based on regional medians.

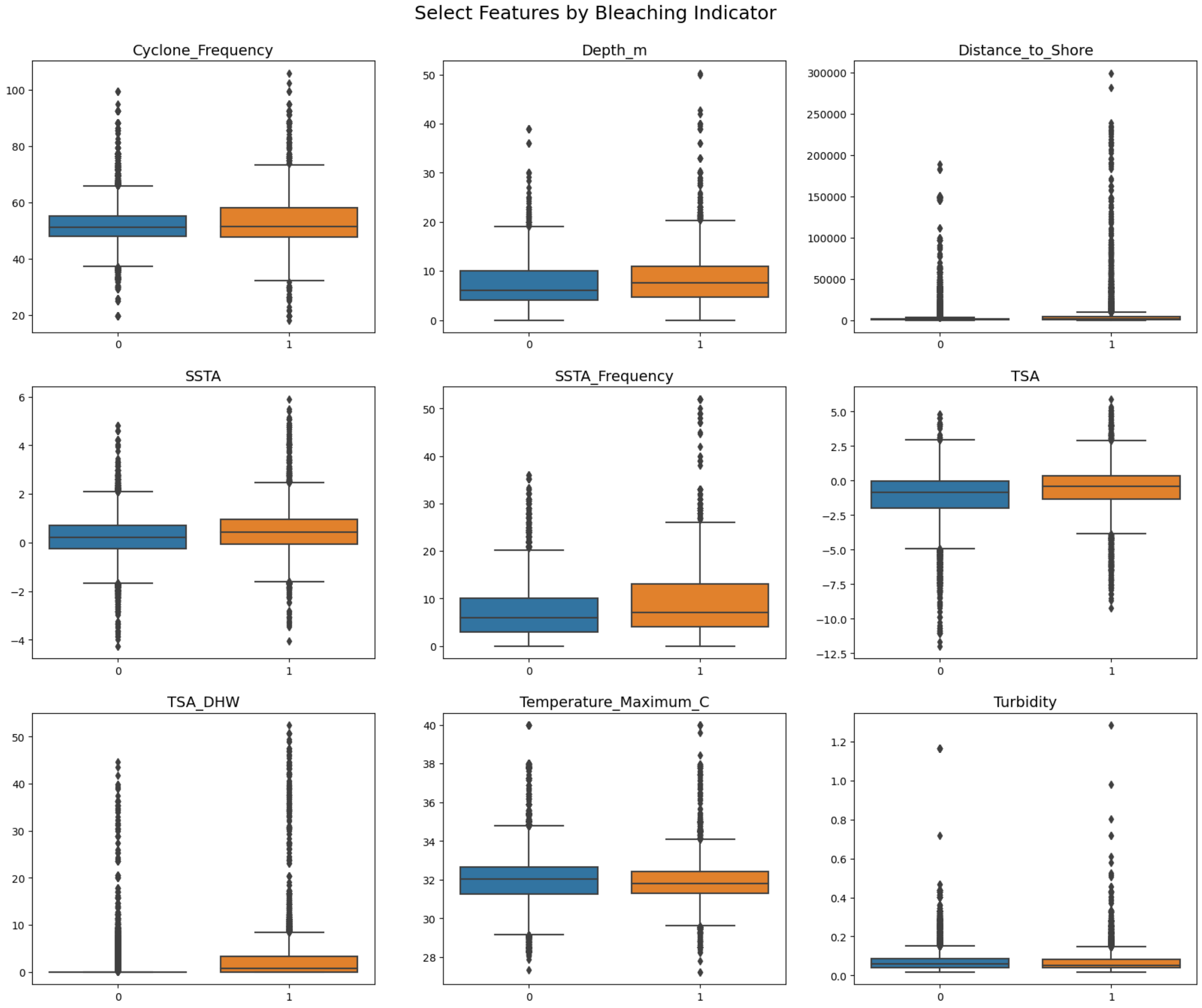
After cleaning and dropping missing values, there are 34,393 observations remaining and the dataset is ready for initial analysis.

2. Exploratory Data Analysis (EDA)

The original dataset comes with over 60 features, most of which will not be used in the final analysis. First, the correlation heatmap of numeric features reveals that there is high correlation amongst many of the temperature metrics. This is not surprising as one would expect higher temperatures results in more thermal stress events, for example. By selectively removing temperature metrics – such as mean, standard deviation, minimum, and maximum – we can see the remaining numeric features have relatively low correlation.



The variability of natural occurrences and magnitude of the size of the data set lends itself to a large spread of observed values with many outlier values. However, we can distinguish slight differences in the medians and/or interquartile ranges associated with the presence of bleaching in a few features, mainly depth, distance to shore, SSTA frequency, TSA DHW. This is reassuring that the we can see some trends differentiating the presence and absence of coral bleaching.


Select Features by Bleaching Indicator

Cyclone_Frequency

Depth_m

Distance_to_Shore

SSTA

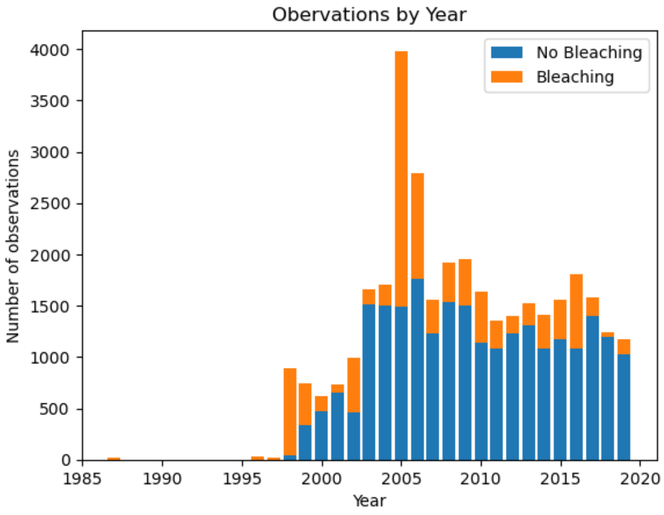
SSTA_Frequency

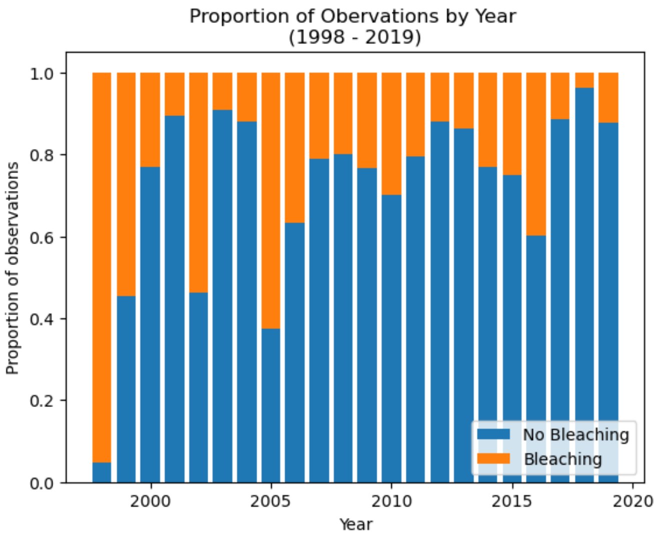
TSA

TSA_DHW

Temperature_Maximum_C

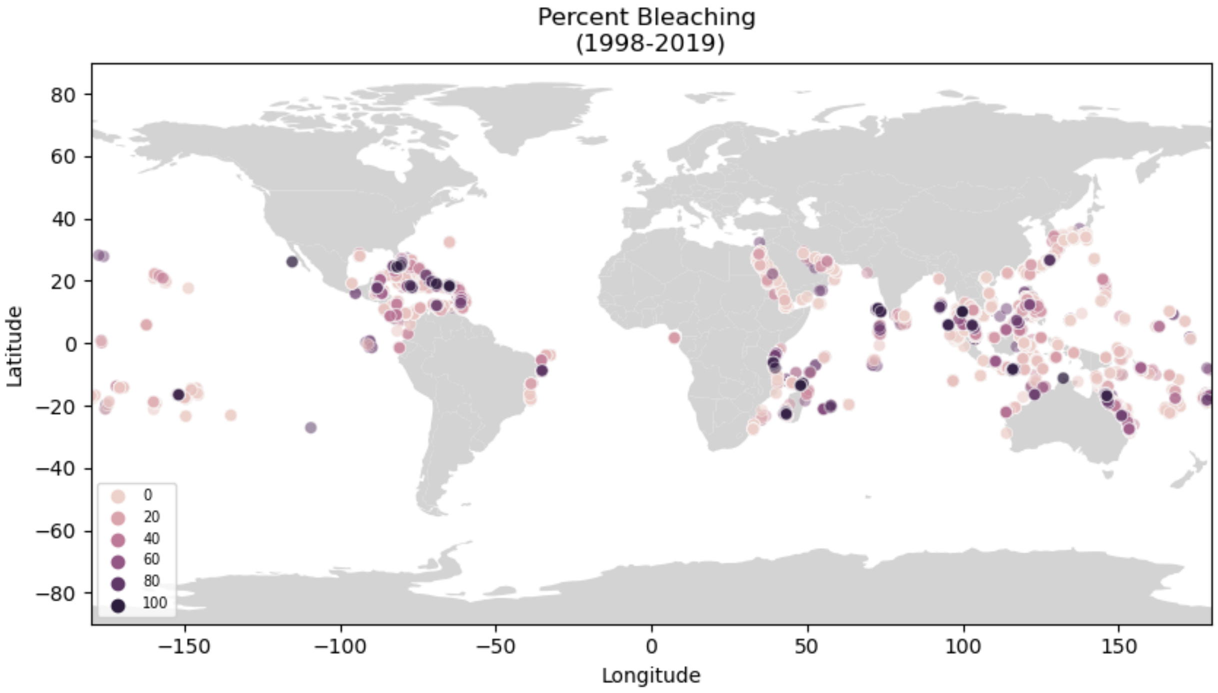
Turbidity





Our data includes complete records from 1995-2019. However, it is not until 1998 when there are consistently over 500 annual observations. This aligns with the first major global bleaching even recorded in 1998, and subsequent global events in 2010 and 2014-2017. In addition to global events, major bleaching events have been recorded locally in the Caribbean and Australian in 2002 and 2005. This is consistent with the above graphs, where most of these years with major bleaching events show an increase in the orange proportion representing observed bleaching.

Demonstrated in the map below, bleaching may be impacted by regional factors, as it appears there is more concentrated bleaching observations in the Caribbean, Eastern Africa, Southeast Asia, and Eastern Australia. This supports the inclusion of a spatial feature, in our case using *Realm_Name* in the form of dummy variables.



3. Preprocessing

The dataset is imbalanced, with only 30% of records with bleaching detected. Both undersampling and oversampling methods were considered to mitigate bias in the model. Undersampling only uses data from the original source, at the cost of reducing the number of observations used. Oversampling (SMOTE) preserves the original size of the dataset, but creates synthetic datapoints to make up for the imbalance.

With both the undersampled and oversampled versions, the following preprocessing steps were taken to prepare the datasets for modeling:

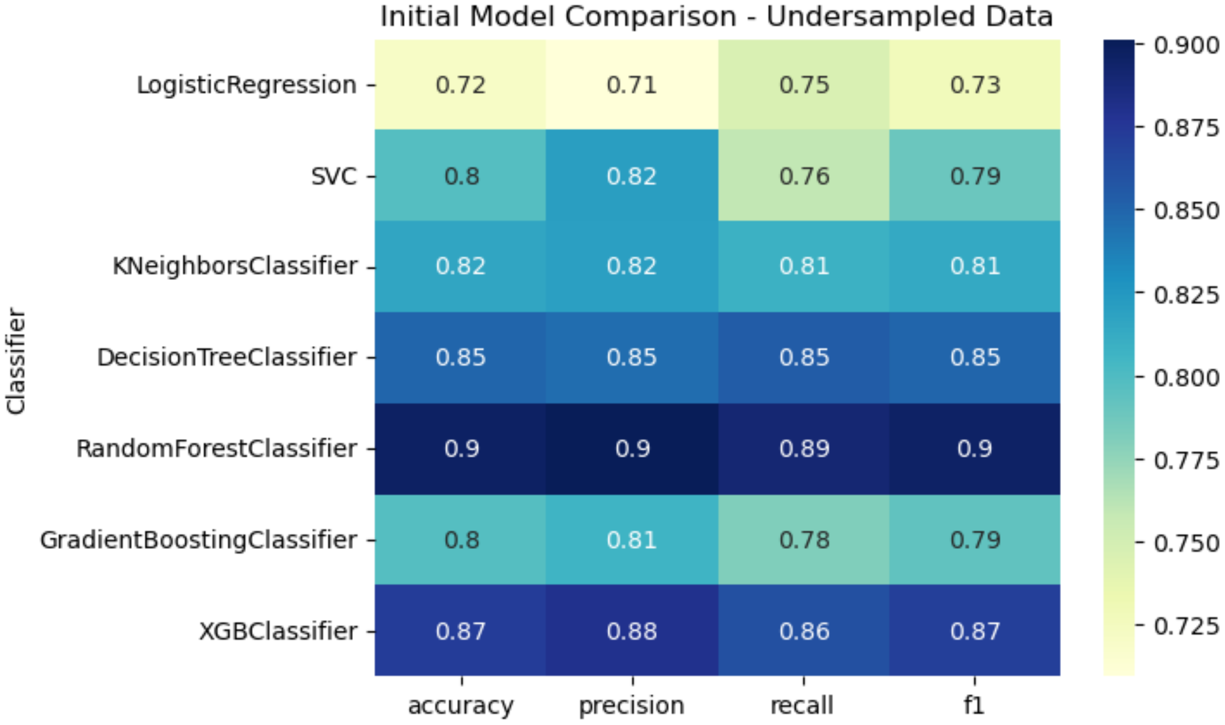
- Splitting data – first splitting into target X (*Bleaching_indicator*) and features y. Then into train and test sets using a 80/20 split. By using train_test_split before the other preprocessing steps, we are preventing data leakage from the unseen test data.
- Impute missing values – *Depth_m* using median by *Realm_Name*. The median is less impacted by outliers.
- Scaling – some of the models we are evaluating are sensitive to the scale of features, such as logistic regression and KNN. Since all features are nonnegative, the StandardScaler is appropriate here and applied to both the training and test data separately.
- Encode categorical variables – *Realm_Name* is still a categorical variable. By encoding this variable, eight new columns will be created for every of the unique realm values to indicate which realm is represented for each observation.

4. Modeling

Initial Model Comparison

We first evaluate how a number of binary classification model perform out of the box, with no additional tuning:

- Logistic Regression
- Support Vector Machine (SVM)
- K Nearest Neighbors (KNN)
- Decision Tree
- Random Forest
- Gradient Boosting
- XGBoost



While the results include accuracy, precision, recall, and F1 scores, F1 scores will be used to evaluate the models since it minimizes both false negatives and false positives. When comparing performance of sampling methods, the improvement of using oversampled data was marginal, and therefore we will continue using the undersampled data to maintain the original data. The top three performing models are **Decision Tree, Random Forest,** and **XGBoost**.

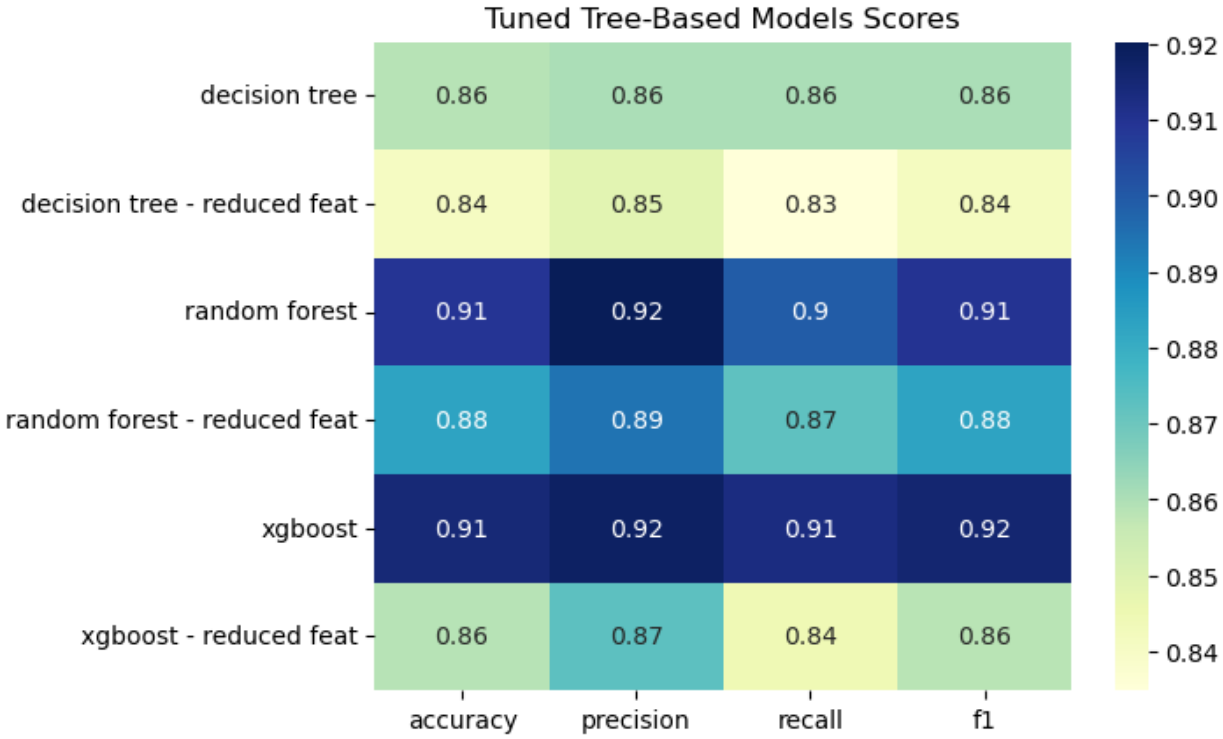
Since the three shortlisted models are tree-based, unscaled data is used moving forward to simplify the interpretability of the results.

Hyperparameter Tuning

The top three performing models can continue to be refined by adjusting each model's hyperparameters as follows.

Decision Tree:	{'criterion' : ['gini', 'entropy'], 'max_depth' : [5, 10, 15, 25, 30]}
Random Forest:	{'criterion' : ['gini', 'entropy'], 'max_depth' : [5, 10, 15, 25, 30], 'n_estimators' : [50, 100, 150]}
XGBoost:	{'max_depth': [5, 10, 15, 25, 30], 'n_estimators': [50, 100, 150], 'reg_alpha': [0.05, 0.1, 1.0], 'reg_lambda': [0.05, 1.0, 2.0]}

In addition to the full models with all features, each model was tuned with a reduced number of features as an attempt to mitigate overfitting and see how a more simplified model performs.



5. Model Selection

Primary considerations for final model selection are cross validated F1 scores and the difference between training and test F1 scores. Both of these metrics will help find a balance of reducing errors without overfitting to the training data. As is expected, the reduced feature models have lower average cross validated F1 scores. However, the standard deviations of Decision Tree and Random Forest models with reduced features are noticeably lower than that of the other cross validation scores. This implies that these two models are better at generalizing the data and produce more consistent results.

```
Training Cross Validation F1 Scores
Decision Tree (full):     0.8472 mean F1 score with 0.0064 std
Decision Tree (reduced):  0.8257 mean F1 score with 0.0018 std
Random Forest (full):     0.8978 mean F1 score with 0.0036 std
Random Forest (reduced):  0.8701 mean F1 score with 0.0030 std
XGBoost (full):           0.8987 mean F1 score with 0.0042 std
XGBoost (reduced):        0.8477 mean F1 score with 0.0091 std
```

```
Train/Test F1 Score Differential
Decision Tree (full):    0.1338
Decision Tree (reduced): 0.1409
Random Forest (full):    0.0883
Random Forest (reduced): 0.1149
XGBoost (full):          0.0817
XGBoost (reduced):       0.1044
```

After comparing the F1 scores from predictions using training data to the test scores, Random Forest with reduced features has a larger decrease in score compared to Random Forest with all features and both XGBoost models. Therefore, it does not appear to predict the new, unseen data well and implies overfitting despite the results seen in the cross validation scores.

XGBoost with all features has the lowest difference between training and test F1 scores, best performing F1 scores, and middling cross validation standard deviation of F1 scores. **For this initial model, we select XGBoost with all features.**

6. Future Development

There are other approaches to modeling with data that can expand and improve this prediction modeling. Feature engineering to combine the temperature metrics has the potential to reduce the number of features without compromising the accuracy of the predictions. Unsupervised learning could also be used to discover unseen patterns or clusters of bleaching not obvious by the regions defined within the data.

In addition to making further improvements to the existing classification model, the data and model used here can be combined or used in conjunction with other coral health models. For example, image processing has been used to detect bleached reefs, which could be built into a more robust prediction model when combined with environmental conditions. Audio data has also been collected and is being processed to classify reefs as either healthy or damaged.