MACQUARIE University
SYDNEY·AUSTRALIA

COMP2200 Assignment

Ngoc Nhat Pham - 47958162

Predicting Coral Reef Bleaching

Introduction

Coral reefs, often referred to as the ***"rainforests of the sea",*** are among the most biodiverse and productive ecosystems on Earth. Occupying less than **0.1%** of the ocean floor, they sustain approximately **25%** of all marine species, providing essential habitats for fish, invertebrates, and countless other organisms. Coral reefs contribute significantly to global ecosystems and human livelihoods by:

- ***Supporting Biodiversity***: They host millions of species, many of which are undiscovered, serving as nurseries and breeding grounds for marine life.
- ***Coastal Protection***: Coral Reefs act as natural barriers, reducing coastal erosion and protecting communities from storm surges and hurricanes.
- ***Economic Value***: Reefs support fisheries, tourism, and coastal economies, generating billions of dollars annually. For example, the Great Barrier Reef contributes over $6 billion to Australia's economy each year through tourism and fishing.

Despite their importance, coral reefs face severe threats, with coral bleaching being a primary concern. Bleaching occurs when corals, stressed by environmental changes such as rising sea temperatures, ocean acidification, or pollution, expel their symbiotic algae (zooxanthellae), leading to whitening and potential mortality due to the loss of these algae's pigments or the algae themselves. ***This phenomenon has escalated in frequency and severity***, with global mass bleaching events recorded in 1998, 2002, 2016, 2017, 2020, and 2022. The loss of coral reefs would disrupt marine ecosystems, coastal

protection, and economic stability, underscoring the urgency of addressing this crisis.

One way to combat coral reef bleaching is to make data-driven decisions, which can be done by Data Science. By analysing comprehensive datasets, such as the one provided, which includes variables like ***Percent_Bleaching, Sea Surface Temperature Anomalies (SSTA), and Thermal Stress Anomaly (TSA),*** we can uncover critical insights. This analysis aims to address key questions:

- ***Which oceans experience the highest occurrence of coral reef bleaching?***
- ***Which ocean exhibits the highest percentage of bleaching?***
- ***How has coral reef bleaching trended over time?***
- ***And most importantly, what environmental factors most significantly drive coral reef bleaching? How can they impact coral reef bleaching and what people can do to address these key factors to reduce coral reef bleaching.***

By leveraging machine learning and statistical models, this study seeks to identify patterns, predict bleaching events, and inform targeted conservation strategies to mitigate environmental stressors and preserve coral reefs for future generations.

Abstract

This study analyzes coral reef bleaching patterns using a comprehensive dataset from the Biological and Chemical Oceanography Data Management Office (BCO-DMO), covering the period from 1980 to 2020 (available at https://www.bco-dmo.org/dataset/773466). The description of the dataset can also be found in the link.

Through thorough data preprocessing and exploratory data analysis, ***the Pacific Ocean*** was identified as the region with the highest occurrence of coral bleaching, ***while the Indian Ocean*** experienced severe bleaching events, notably the 1997–1998 event, which resulted in up to 90% coral cover loss in areas such as the Maldives, Sri Lanka, Kenya, and Tanzania. The target variable, ***Percent_Bleaching***, was categorized into three groups: ***Mild, Moderate, and Severe,*** based on the dataset's ***Bleaching_Comments***.

Multiple Machine Learning models, including ***Logistic Regression, Naïve Bayes, Decision Tree, Random Forest, and Artificial Neural Network (specifically Multilayer Perceptron Classifier),*** were employed to predict bleaching severity.

Evaluation Metric Analysis: The performance of the model was evaluated to determine the most effective algorithm for this application. They are evaluated based on classification accuracy, precision, recall, and F1. ***A classification report*** was used to evaluate the machine-learning algorithms for each dataset. This is one of the performance evaluation metrics used in machine learning, which shows the precision, recall, F1 score, and support of the trained classification model. ***The precision value*** is defined as the ratio of true positives to the sum of true and false positives. On the other hand, ***the recall value*** is the ratio of true positives to the sum of true and false negatives. ***The F1 score*** is the weighted harmonic mean of precision and recall, so it indicates how well the model did. Lastly, the support is the number of actual occurrences of the class in the models. To strengthen the results, more methods were used to evaluate the accuracy of the models. One of these methods is ***comparing the training and testing scores.***

To enhance model accuracy, techniques such as ***statistical imputation for missing values, label encoding, feature selection, feature extraction via Principal Component Analysis (PCA), and Synthetic Minority Oversampling Technique (SMOTE)*** for addressing imbalanced data were applied.

Full Jupyter Notebook can be found HERE:

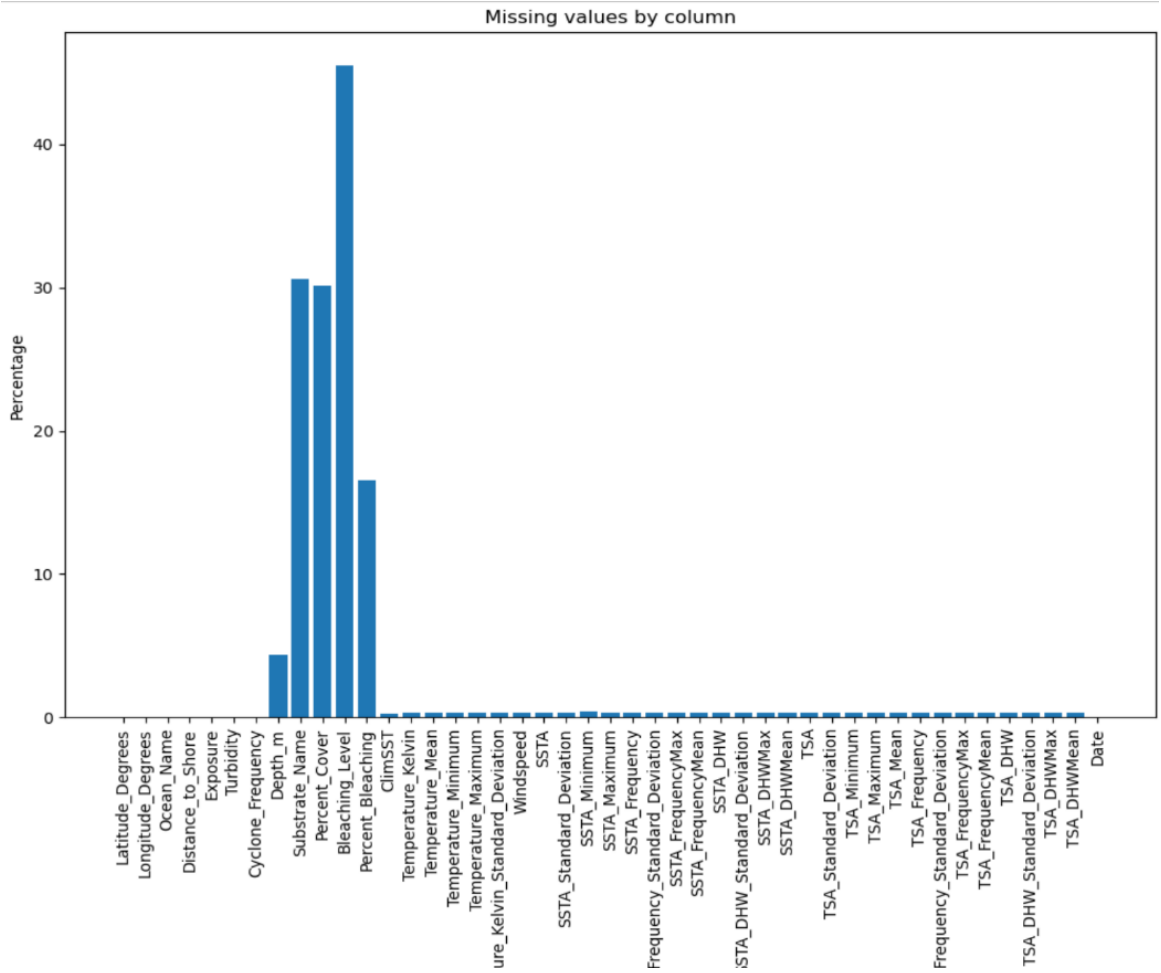
Related Work

Dataset and Preprocessing

The BCO-DMO dataset collects statistical information on the presence and absence of coral reef bleaching from a range of different locations and environmental conditions, from 1998-2020. Measures of depth, on-site exposure, distance to land, turbidity, cyclone frequency, and a range of sea-surface temperature metrics were also included and allowed for comparative analysis and the determination of bleaching thresholds***. There are in total 62 variables with 41361 datapoints***

Prior to the analysis, the dataset underwent a preprocessing step and exploratory data analysis to ensure that it was suitable for the machine learning algorithms used in the study. As the dataset present categorical and numerical variables respectively, different preprocessing methods were applied.

Initially, variables serving as unique identifiers (e.g., Reef_ID, Site_ID), those with excessive unique values (redundant for EDA or modelling), or those derived from the Date variable were removed to streamline the dataset. Next, missing values were addressed:



The ***Bleaching_Level*** variable, intended to include both coral population and colony data, contained only population data, with colony data entirely absent. Similarly, ***Substrate_Name*** had over 30% missing values, justifying the removal of both columns to avoid introducing bias.

Numerical columns such as ***Depth_m, Percent_Cover***, and others exhibited missing values. Dropping these rows would result in significant data loss, so **statistical imputation** was employed. **For normally/close to normally distributed variables** (e.g., SSTA, SSTA_FrequencyMax, TSA, Cyclone_Frequency, Temperature_Maximum), **the mean** was used, while **the median** was applied to **skewed numerical columns**, with distributions determined during the EDA phase.

Also, as I mentioned earlier, The ***Percent_Bleaching*** variable was categorized into three groups: **Mild, Moderate, and Severe** based on insights from the Bleaching_Comments field. Then I used **Ordinal Encoding method**, which is a well-suited feature engineering technique to encode categories with ranking. Similarly, the **Exposure variable (Sheltered, Sometimes, Exposed)** was ordinally encoded

One-Hot Encoding was used to transform categorical data into binary vectors, where each unique category is represented by a column with a value of 1 for presence and 0 for absence, I applied this to ***Ocean_Name.***

Finally, data types were standardized: categorical variables were converted to the category format, the Date variable to datetime format, and continuous variables to float format for consistency and model compatibility.

EDA (Exploratory Data Analysis)

1. Univariate Analysis

A detailed analysis is shown in Jupyter notebook. In conclusion the univariate analysis reveals that :

- ***SSTA and TSA*** distributions are centered near zero with slight skewness, indicating occasional extreme thermal stress events.
- ***Temperature variables*** cluster around typical tropical reef conditions, with maxima indicating potential heat stress.

- ***Depth and Distance_to_Shore*** are right-skewed, showing most reefs are shallow and near-shore.
- ***Percent_Cover*** is low in many areas, and ***ClimSST*** aligns with expected baseline temperatures.
- ***Turbidity*** and Windspeed are right-skewed, indicating variable stressors, while ***Cyclone_Frequency*** suggests occasional high-impact events.

- ***Ocean_Name*** shows the Pacific Ocean with the highest count, followed by the Atlantic, Indian, Arabian Gulf, and Red Sea, suggesting the Pacific is the most studied or affected region.
- ***Exposure*** indicates most sites are Sheltered, with fewer Sometimes or Exposed sites, pointing to a bias toward protected reef environments.
- ***Percent_Bleaching_Category*** reveals most observations fall into the "Mild" category, with fewer "Moderate" and minimal "Severe" cases, indicating that severe bleaching, while less frequent, remains significant

These patterns underscore the environmental factors driving coral bleaching.

2. Bivariate Analysis

The bivariate analysis highlights that the ***Pacific Ocean*** has the highest bleaching counts, predominantly in the "Mild" category, followed by the Atlantic, with the Indian, Arabian Gulf, and Red Sea showing fewer and less severe events.

Sheltered sites dominate observations with mostly "Mild" bleaching, while Exposed sites show a wider range of severity, suggesting greater stress in open environments. "Sometimes" exposed sites have the lowest counts across categories.

Over time, bleaching severity ***peaks around the mid-1990s and early 2000s***, with a general rise from the 1980s to early 2000s, followed by stabilization and intermittent spikes in the 2010s, reflecting the increasing impact of environmental stressors.

3. Multivariate Analysis

The multivariate analysis reveals that while the Pacific Ocean had the highest bleaching occurrence in univariate analysis, the ***Indian and Arabian Gulf exhibit the highest bleaching levels***, with peaks around 75% in the Indian Ocean and nearly 80% (the highest recorded) in the Arabian Gulf during the mid-1990s and early 2000s, aligning with global mass bleaching events. The Pacific shows moderate peaks around 50%, while the Atlantic and Red Sea have lower, sporadic peaks below 50%. Post-2000, bleaching levels decline across all oceans but show spikes, especially in the Indian Ocean around 2015–2020.

A strong positive correlation (0.85) exists between ***Temperature_Kelvin*** and ***Thermal Stress Anomaly*** (TSA), while SSTA correlates strongly with TSA (0.55) and moderately with Temperature_Kelvin (0.42), underscoring their combined role in bleaching. Other variables show weak or negligible correlations, suggesting limited linear interactions.

Machine Learning Models and Techniques

Feature Selection

Before applying machine learning models, feature selection is conducted as part of data preprocessing to identify the most relevant variables for predicting coral bleaching events. Given the sensitivity of coral reefs and the abundance of variables in the dataset, excessive features can lead to overfitting and model complexity. To address this, feature selection is prioritized.

Here, I specific use ***Recursive Elimination Method***. This is feature selection method that iteratively removes features from a model, evaluating performance at each step to determine the most important features. RFE can reduce the risk of overfitting by choosing the most important features. However, removing important features can also lead to underfitting. To address this issue, I paired with Random Forest to get best features out. ***Random forest*** is a machine-learning method that generally works well with high-dimensional problems and allows for nonlinear relationships between predictors. However, the presence of correlated predictors has been shown to impact its ability to identify strong predictors***. The Random Forest-Recursive Feature Elimination (RF-RFE)*** approach mitigates this issue, ensuring a robust selection of the most influential features for subsequent modelling.

The top 20 best features selected after applying Recursive Feature Elimination (RFE) with Random Forest are:

- ***'Distance_to_Shore'***
- ***'Turbidity'***
- ***'Cyclone_Frequency'***
- ***'Depth_m'***
- ***'Percent_Cover'***
- ***'ClimSST'***
- ***'Temperature_Kelvin'***
- ***'Temperature_Mean'***
- ***'SSTA'***
- ***'SSTA_Standard_Deviation'***
- ***'SSTA_Frequency'***
- ***'SSTA_Frequency_Standard_Deviation'***
- ***'SSTA_DHW'***
- ***'SSTA_DHW_Standard_Deviation'***
- ***'TSA'***
- ***'TSA_Standard_Deviation'***
- ***'TSA_Mean'***
- ***'TSA_DHW'***
- ***'TSA_DHW_Standard_Deviation'***
- ***'TSA_DHWMean'***

Data Splitting, Scaling and Handling Imbalanced

Data Splitting: The data is the splited into ***training (70%) and test (30%)*** sets using train_test_split with stratify=y to maintain the original class distribution in both sets

Scaling: The ***StandardScaler method*** is applied to standardize features by removing the mean and scaling them to unit variance, resulting in a mean of 0 and a standard deviation of 1. This normalization ensures that machine learning algorithms (e.g., Logistic Regression, Neural Networks) perform optimally, as they often assume features are on the same scale. Without scaling, features with larger ranges (e.g., Distance_to_Shore) could disproportionately influence the model compared to those with smaller ranges (e.g., Turbidity).

Handling Imbalanced Data with SMOTE: Given the dataset is imbalanced ***SMOTE (Synthetic Minority Oversampling Technique)*** is also to address this. By resampling the training set (x_train_resampled, y_train_resampled) after scaling, SMOTE helps the model learn from all classes effectively, reducing the risk of poor performance on minority classes. The reason why I applied to SMOTE only on training set is if applied to the test set, the synthetic samples could introduce information from the test set into the training process (e.g, during hyperparameter tuning or model fitting), artificially inflating performance metrics. This violates the principle that the test set should be an independent evaluation of the model's generalization. Since ***Percent_Bleaching_Category*** is likely multi-class and imbalanced (most observations fall into the "Mild" category, with fewer "Moderate", and minimal "Severe" case), SMOTE on the training set ensures the model learns from all categories without altering the test set's natural distribution.

Principle Component Analysis (PCA)

PCA is a dimensionality reduction method that is often used to reduce the dimensionality of large data sets, by transforming a large set of variables into a smaller one that still contains most of the information in the large set. Given the context where there are 20 features after the RF-RFE feature selection step, the dataset is relatively high-dimensional, which can lead to issues like the curse of dimensionality, increased computational complexity, and a higher risk of overfitting in machine learning models. PCA transforms the original correlated features into a smaller set of uncorrelated principal components

(PCs) that capture most of the variance in the data, reducing the feature space while retaining essential information.

Also, the correlation heatmap from the multivariate analysis revealed strong correlations between temperature-related variables (e.g, Temperature_Kelvin and TSA at 0.85, SSTA and TSA at 0.55). PCA mitigates multicollinearity by creating orthogonal components, ensuring that the model does not give redundant information and improving its generalization performance.

Outcome: The analysis indicates that ***PC10 captures 90–95%*** of the variance in the original 20 features, striking a balance between dimensionality reduction and information retention. Selecting PC10 as the lowest component within this range ensures the reduced dataset retains key variability while improving model efficiency and reducing overfitting risk.

Classification models

a. Logistic Regression

This is one of the commonly used learning methods that predict the probability of a categorical outcome (Percent_Bleaching_Category with classes 0, 1, 2 representing Mild, Moderate, Severe). In this analysis, it serves as a baseline model due to its simplicity and interpretability, providing a benchmark for evaluating more complex models.

Advantages: Logistic Regression is computationally efficient, easy to implement. In this study, its simplicity makes it less prone to overfitting with the reduced 10 PC feature set.

Disadvantages: It assumes linearity and independence of features, which may not fully capture the complex, nonlinear relationships in coral bleaching data (e.g., interactions between SSTA and Windspeed). It also struggles with imbalanced datasets unless mitigated via SMOTE, and its performance may reduce with high multicollinearity, though PCA may help address this.

Evaluation Metric Analysis:

- ***Training Performance***: The training has a moderate accuracy score (0.56). Macro and weighted averages are also consistent at 0.56 for precision, recall, and F1-score, indicating uniform performance across

classes but moderate accuracy, suggesting the model fits the training data reasonably

- ***Test Performance***: The test accuracy (0.6482) exceeds training (0.5585), suggesting that the test set might be easier or less noisy than the training set. This difference can potentially indicates underfitting model as well. Moreover, the low F1 scores for Classes 1 and 2 (0.28 and 0.30 respectively) highlight difficulty in predicting minority classes. The high precision for Class 0 (0.93) with moderate recall (0.68) indicates that the model is good at identifying Mild bleaching but misses many Moderate and Severe cases

Conclusion: Logistic Regression provides a solid baseline, with test accuracy of **0.65** indicating reasonable predictive power. However, its poor performance on minority classes (Moderate, Severe) suggests the need for advanced models to capture nonlinear patterns and improve class balance handling

b. Naïve Bayes

Naive Bayes is a probabilistic classifier based on Bayes' theorem, assuming independence between features given the class label. It is used here as an additional baseline model following Logistic Regression, using the 10 principal components from PCA.

Advantages: Naive Bayes is computationally efficient and performs well with high-dimensional data, making it a good fit for the reduce 10 PC feature set. It is particularly effective for imbalanced datasets when combined with SMOTE, providing fast training and interpretable probability estimates.

Disadvantages: The independence assumption may not hold for correlated features (e.g., Temperature_Kelvin and TSA at 0.85), potentially underestimating complex interactions in coral bleaching data. It can also be sensitive to feature scaling and may struggle with noisy or overlapping class distributions

Evaluation Metric Analysis:

- ***Training performance***: Similar to Logistic Regression, Naïve Bayes training set has a moderate accuracy score (0.55). The macro and weighted average scores are consistently at 0.55 for precision, recall, and

F1 score, indicating balanced performance across classes but the accuracy is average, indicating consistent training with some limitations.

- ***Testing performance***: The test accuracy of 0.72 is much greater than training, suggesting that there is significant underfitting. The high F1 score for Class 0 (0.85) with strong recall (0.79) shows excellent prediction of Mild bleaching, while Classes 1 and 2 have lower F1 scores (0.28 and 0.32), indicating challenges with minority classes, consistent with the imbalanced test set.

Conclusion: Naive Bayes offers a robust baseline with a test accuracy of 0.72, greater than Logistic Regression (0.65), with improved consistency between training and test sets. However, it also faces the same problem with Logistic Regression: strong performance for majority class but low performance for the minority ones (Moderate and Severe). This suggests the need for advanced models to better capture minority class patterns and nonlinear relationships in the PCA-derived features.

c. Decision Tree

A Decision Tree is a non-linear, tree-based classifier that recursively splits the dataset into branches based on feature values to predict categorical outcomes. It is used here as an advanced baseline model following Logistic Regression and Naive Bayes. In the context of my dataset, its ability to capture complex patterns makes it suitable for exploring the standardized and SMOTE-balanced dataset.

Advantages: Decision Trees can model nonlinear relationships and interactions between features (e.g., SSTA and Temperature_Kelvin), offering flexibility over Logistic Regression. They also handle multiclass problems well and align with the balanced training set.

Disadvantages: They are prone to overfitting, especially with deep trees, and can be sensitive to small data variations. The high dimensionality (even after PCA) may still challenge performance unless tuned, requiring hyperparameter optimization.

Cross-Validation and GridSearchCV: Due to the tendency of Decision Trees to overfit, especially without proper tuning, cross-validation and GridSearchCV are employed. Cross-validation is a model validation technique to assess how

the outcomes of a statistical analysis will generalize to an independent dataset. In this scenario, it splits the training data into 5 parts, training on 4 and validating on the 1 left out, repeating this process to get a reliable average performance. GridSearchCV is a tool from scikit-learn that performs an exhaustive search over a grid of hyperparameters, using cross-validation to evaluate each combination. The tuned hyperparameters are:

- ***max_depth***: Range from 1 to 20, controlling the maximum depth of the tree to prevent overfitting by limiting complexity.
- ***min_samples_split:*** Values [2, 5, 10], defining the minimum samples required to split a node, balancing overfitting and underfitting.
- ***min_samples_leaf***: Values [1, 2, 4], setting the minimum samples in leaf nodes to ensure robust splits and reduce overfitting.

Evaluation Metric Analysis:

- Training Performance: The training set has a very high accuracy score of 0.97, with macro and weighted averages are consistent at 0.97 for precision, recall, and F1-score. However, this might indicate near-perfect training fit, suggesting potential overfitting.
- Test Performance: The test accuracy (0.78) is significantly lower than training (0.97), confirming overfitting despite applying hyperparameter tuning. The high F1 score for Class 0 (0.88) with strong recall (0.83) shows good prediction of Mild bleaching, but Classes 1 and 2 still have lower F1 scores (0.42, 0.39), indicating challenges with minority classes

Conclusion: Decision Tree is a strong classification model with a test accuracy of 0.78, surpassing Logistic Regression (0.65) and Naive Bayes (0.55). However, the gap between training (0.97) and test accuracy suggests overfitting issue. Additionally, the poor performance on Classes 1 and 2 indicates a need for ensemble methods to better handle minority classes and reduce overfitting with the PCA-derived features.

d. Random Forest

Random Forest is an ensemble learning method that constructs multiple decision trees and aggregates their predictions to classify outcomes with

classes. Its ability to reduce overfitting through bagging and handle complex, nonlinear relationships makes it suitable for this coral bleaching dataset.

Advantages: Random Forest improves upon single Decision Trees by reducing overfitting via averaging multiple trees, capturing nonlinear interactions), and handling high-dimensional data effectively with the PCA-reduced feature set.

Disadvantages: It can be computationally intensive and less interpretable than simpler models. If not adjusted, its performance may plateau with highly imbalanced test sets

GridSearchCV and Hyperparameters: Similar to Decision Tree, GridSearchCV is also applied to get the best hyperparameters. These hyperparameters are:

- ***max_depth***: Set to None, allowing trees to grow until all leaves are pure or contain fewer samples than min_samples_split
- ***min_samples_leaf***: Set to 1, ensuring each leaf has at least one sample for robust splits.
- ***min_samples_split***: Set to 2, defining the minimum samples required to split a node, balancing overfitting and underfitting.
- ***n_estimators***: Set to 150, specifying the number of trees in the forest, enhancing accuracy.

Evaluation Metric Analysis:

- ***Training Performance***: The model achieves a perfect accuracy of 1.00 on the training set, with macro and weighted averages also at 1.00 for precision, recall, and F1 score. While this indicates the model fits the training data extremely well, it also raises concerns about overfitting, as such perfect performance is uncommon on real-world datasets and may not generalize well to unseen data.
- ***Test Performance***: The test accuracy (0.87) is lower than training (0.9978), confirming overfitting despite being tuned. The high F1-score for Class 0 (0.94) with strong recall (0.93) shows excellent prediction of Mild bleaching, while Classes 1 and 2 have moderate F1-scores (0.53, 0.56), indicating improved but still limited prediction of minority classes

Conclusion: The Random Forest model shows strong performance with a test accuracy of 0.8705, outperforming the other models. However, the significant gap between training and test accuracy indicates overfitting. While the

performance for Class 1 and Class 2 has improved to a moderate level, challenges in accurately predicting these classes still persist.

e. Aritificial Neural Network

An Artificial Neural Network (ANN) is a computational model inspired by the human brain, consisting of layers of interconnected nodes (neurons) that learn complex patterns through backpropagation. In this dataset It is used here to predict classes of coral reef bleaching (Mild, Moderate and Severe)

Advantages: ANNs excel at modeling complex, nonlinear, making them suitable for the intricate patterns in coral bleaching data. They can learn hierarchical feature representations, which can outperform tree-based models on PCA-derived features.

Disadvantages: ANNs are computationally intensive, require careful tuning to avoid overfitting, and lack interpretability compared to simpler models. They also need standardized data and balanced dataset

A Multi-Layer Perceptron classifier is then used in this model with some key hyperparameters:

- ***hidden_layer_sizes***: This specifies the structure of the neural network. Here, the model has two hidden layers: first hidden layer with 60 neurons and second hidden layer with 3 neurons
- The ***learning_rate*** controls how much the model adjusts its weights in response to the loss gradient. Here, 'constant' means the learning rate stays the same throughout training
- ***max_iter*** controls maximum number of training iterations. If the model hasn't converged by 250 iterations, it will stop and possibly raise a warning.

Evaluation Metric Analysis:

- ***Training Performance***: The model has a training accuracy of 0.74, with macro and weighted averages are consistent for precision, recall, and F1-score. This indicates that the model has a moderate performance with balanced class prediction, suggesting a reasonable fit without severe overfitting.

- ***Test performance***: The test accuracy stays the same with the train accuracy (0.74), suggesting no overfitting and effective generalization aided by SMOTE. The high F1-score for Class 0 (0.85) with strong recall (0.77) shows good prediction of Mild bleaching, while Classes 1 and 2 have lower F1-scores (0.36, 0.43), indicating the challenges with minority classes still remain

Conclusion: ANN provides a test accuracy of 0.74, surpassing Logistic Regression (0.65) and Naive Bayes (0.72) but underperforming compared to Decision Tree (0.78) and Random Forest (0.87). Its balanced training and test performance indicates robustness, but the poor performance on Classes 1 and 2 suggests that ensemble methods like Random Forest is more effective for such imbalanced dataset.

Summary

1. Summary of Machine Learning models

	Model	Train Accuracy	Test Accuracy	Mean Precision	Mean Recall	Mean F1-Score
0	Logistic Reg	0.558493	0.648159	0.447892	0.553907	0.455104
1	Naive Bayes	0.548583	0.725132	0.464461	0.550467	0.484247
2	Decision Tree	0.968505	0.781060	0.534411	0.623323	0.564221
3	Random Forest	0.997790	0.870498	0.668131	0.685314	0.676340
4	ANN	0.745014	0.738163	0.518691	0.639083	0.547827

The analysis of the five models, Logistic Regression, Naive Bayes, Decision Tree, Random Forest, and Artificial Neural Network (ANN) reveals distinct performance trends across training and test datasets. ***Random Forest achieves the highest accuracy scores on both training and test data***, alongside the best mean precision (0.67), recall (0.68), and F1 score (0.68). However, the

substantial gap between training and test accuracy strongly suggests overfitting, indicating the model may not generalize well to new data.

In contrast, ANN, having lower precision, recall and F1 score than Random Forest, its small gap between training and test accuracy suggest that the model is less likely to have overfitting risk than Random Forest, indicating robustness and a more balanced performance.

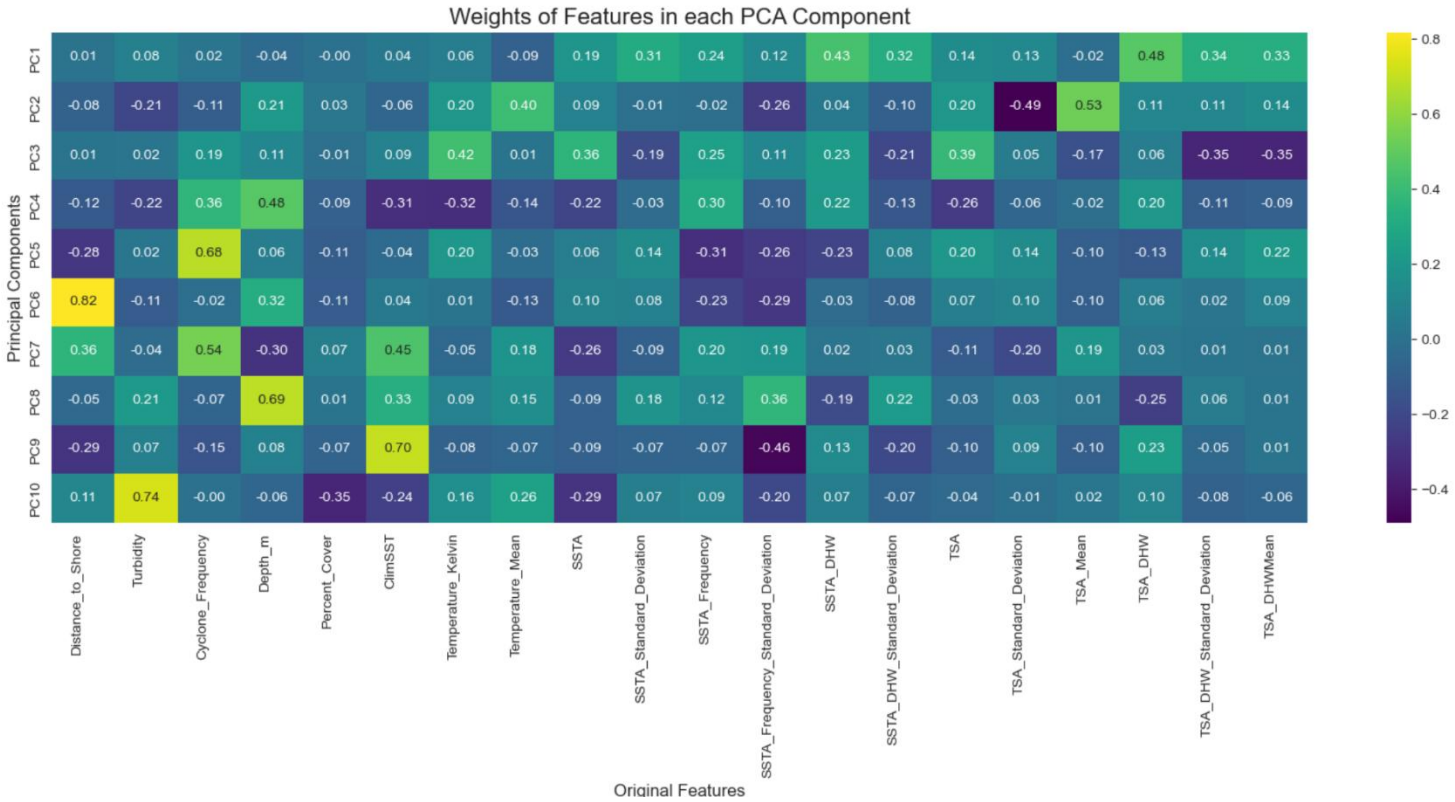
A critical challenge across all models is their limited ability to predict minority classes (Moderate and Severe bleaching), despite I already applied SMOTE to balance the training set. The mean precision, recall, and F1-scores, which hover around or just above average, indicating that while over half of the actual or predicted value are correctly classified, the models struggle with the less frequent classes. This suggests that the current feature set or preprocessing may not fully capture the dynamics of severe bleaching events.

Between Random Forest and ANN, ***neither of them is the best for predicting coral reef bleaching***. Random Forest excels in overall accuracy but risks overfitting, while ANN offers robustness with a more stable performance across datasets but lower predictive power. Both models have notable strengths and limitation. Additional data exploration such as revisiting feature engineering, exploring alternative balancing techniques, or importing more sufficient data, could enhance model performance, particularly for minority class prediction.

2. Feature Importance Analysis for Coral Reef Bleaching

2.1 Feature Importance using PCA

To identify the most influential features driving coral reef bleaching, an additional analysis step was conducted using Principal Component Analysis (PCA) to assess feature importance across the 10 principal components.

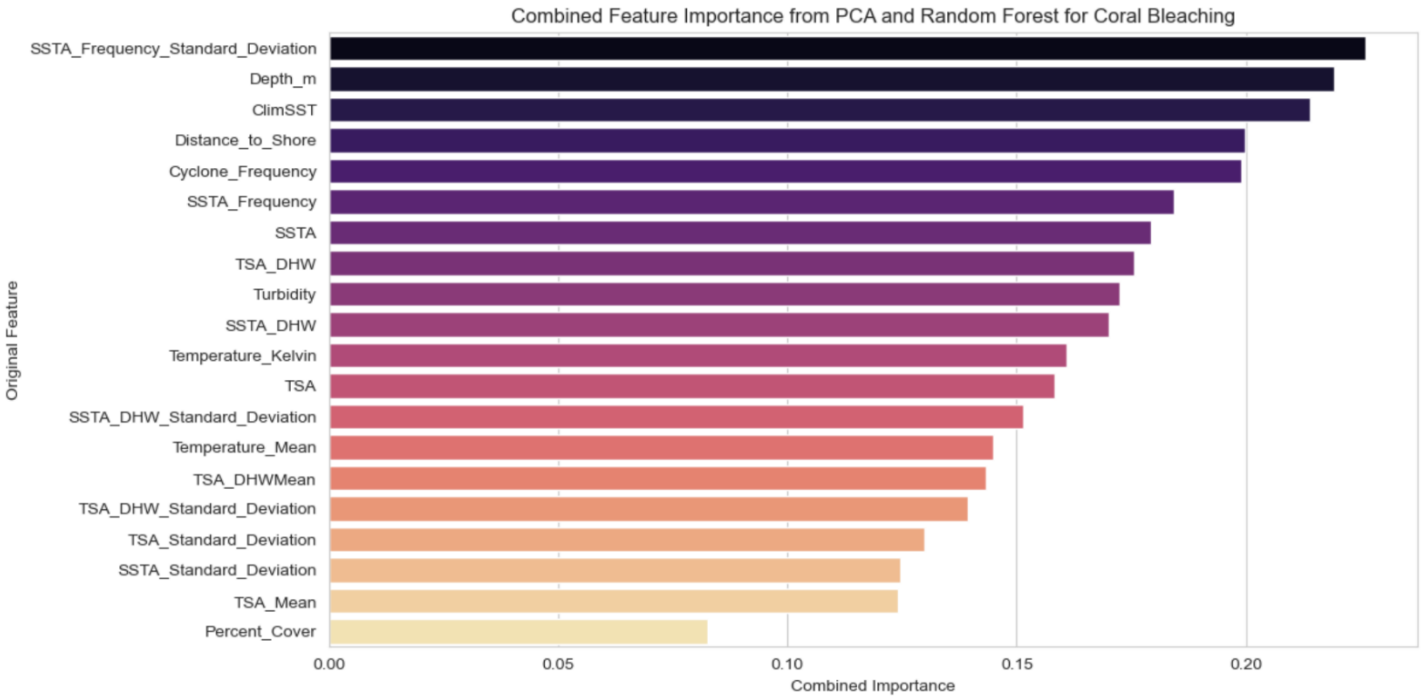


- The heatmap illustrates the weights of the original 20 features within each principal component derived from PCA. ***Notably, in the lower portion of the heatmap (PC5 to PC10), features such as Distance_to_Shore, Turbidity, Cyclone_Frequency, Depth_m, and ClimSST exhibit dominant positive weights***. This suggests that these components are primarily influenced by those specific features(for example, PC6 is largely driven by Distance_to_Shore).

- To enhance this analysis and obtain a more precise estimation of feature importance, PCA was combined with advanced machine learning models, including Random Forest and Artificial Neural Networks (ANN).

2.2 Combined PCA and Random Forest

This is when PCA was combined with an ensembled model Random Forest. I did this by multiplying the PCA component weights with the feature importance weights derived from the Random Forest model, then redistributing the combined importance across the original features

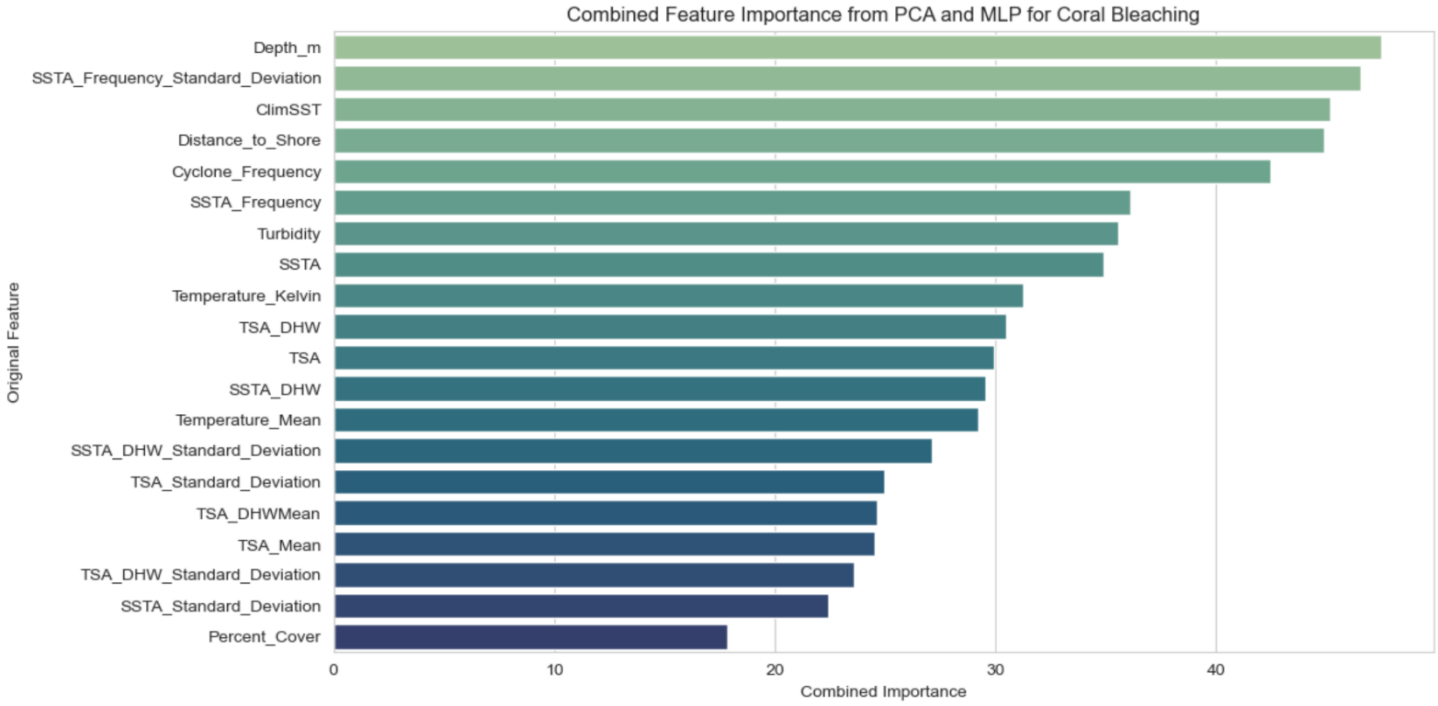


Surprisingly, this approach highlighted ***SSTA_Frequency_Standard_Deviation*** as the most influential feature, followed by ***Depth_m, ClimSST, Distance_to_Shore, and Cyclone_Frequency***.

Also, ***Turbidity***, which stood out in the initial PCA heatmap, appeared less significant after incorporating the Random Forest model with PCA. To further validate these findings, my next step will involve combining PCA with the Multi-Layer Perceptron (MLP) model for additional insight.

2.3. Combined PCA and Multilayer Perception

In this step, PCA was integrated with a Multilayer Perceptron (MLP) classifier to further evaluate feature importance. The process involved using the input layer weights from the MLP, summing them across the neurons in the first hidden layer, and combining these with the PCA component weights before mapping the importance back to the original features.



The results showed that the top five important features remained consistent with the PCA-Random Forest analysis: ***Depth_m, SSTA_Frequency_Standard_Deviation, ClimSST, Distance_to_Shore***, and ***Cyclone_Frequency***.

However, a notable change occurred where ***Depth_m and SSTA_Frequency_Standard_Deviation swapped positions***, with Depth_m now ranking first, indicating its heightened influence in this combined approach.

Final Conclusion and Key Issues

Insights and Recommendations

After an extensive analysis encompassing data preprocessing, feature engineering, and the application of various machine learning models, this study provides critical insights into coral reef bleaching:

- ***Regional Impact***: The Pacific Ocean, particularly the Great Barrier Reef, has experienced significant coral bleaching events. Conversely, the Indian Ocean and Arabian Gulf exhibit the highest bleaching severity, with peak levels reaching approximately 75% and nearly 80%, respectively.

- ***Environmental Exposure***: Despite many coral reefs being in sheltered or shallow, near-shore environments, they remain highly susceptible to severe bleaching, with 1998 identified as a particularly devastating year
- ***Model Performance***: Random Forest and Artificial Neural Network (ANN) emerged as the most effective models, delivering high accuracy and a balanced gap between training and test datasets. These models also revealed key environmental drivers of bleaching, including Sea Surface Temperature Anomalies (SSTA_Frequency_Standard_Deviation), Depth, Climatological Sea Surface Temperature (ClimSST), Distance to Shore, and Cyclone Frequency

  Below are detailed impacts and actionable recommendations:

  - ***SSTA Frequency Standard Deviation (Sea Surface Temperature Anomalies)***: High variability in sea temperature stresses corals, especially when anomalies are frequent and extreme.
    Actions:

    - ***Reduce Greenhouse Gas Emissions***: SSTA in general are driven by global warming. Individuals and communities can lower carbon footprints by using renewable energy, reducing energy consumption, and supporting policies for emissions reduction.
    - ***Promote Sustainable Practices***: Support sustainable agriculture and transportation (for example, electric vehicles) to reduce CO2, which contributes to ocean warming.

  - ***Depth***: Corals at shallow depths are more exposed to temperature extremes, UV radiation, and wave action.

    Actions:

    - ***Prioritize Deeper Reef Conservation***: Focus protection efforts on deeper zones, which offer greater thermal stability.
    - ***Regulate Shallow Zone Activities***: Limit snorkeling and diving in shallow areas during peak heat seasons to reduce additional stress on corals.

- ***ClimSST (Climatological Sea Surface Temperature)***: Higher average SST is one of the most direct causes of coral bleaching.

  Actions:

  - ***Marine protected areas***: Establish or expand marine protected areas in cooler current-influenced areas to act as thermal refuges.
  - ***Reduce local stressors***: Mitigate overfishing and pollution to enhance coral resilience against heat stress.

- ***Distance to Shore***: Reefs closer to shore are more exposed to human impacts like pollution, sedimentation, nutrient runoff

  Actions:

  - ***Improve coastal zone management***: Enforce buffer zones, reduce urban and agricultural runoff through better wastewater treatment and erosion control.
  - ***Restore mangroves and wetlands***, which filter runoff and protect reefs.

- ***Cyclone Frequency***: Frequent storms physically damage reefs and can worsen bleaching recovery times.

  Actions:

  - ***Support restoration projects*** that plant cyclone-resistant coral species or use structures to stabilize reefs against storm damage

Key Issues in the Study

Although the study shows promising results, with most machine learning models achieving moderate to high accuracy, several important issues need further investigation to improve reliability and real world applicability.

- ***Limited Data Representation***: The dataset includes key factors like Depth and Sea Surface Temperature, but it misses critical real world contributors to coral bleaching, such as **Sunlight Exposure, Pollution,**

Commercial or Industrial Activities, and general Human Impact. Moreover, variables like ***Bleaching_Level, SSTA_Mean*** although are included in the dataset but need to be removed as they were not studied properly. Without these, the models may not fully capture the true drivers of bleaching events, limiting their predictive power.

- ***Overfitting Concerns***: Overfitting is a significant issue, particularly with Random Forest, which achieved a perfect training accuracy of 100%. This likely stems from using too many features, even after applying feature selection and PCA. Additionally, I used SMOTE to balance the training set but kept the test set raw to reflect real-world performance. However, the class imbalance was very marked, most reefs fall under Mild bleaching, with far fewer in Moderate and Severe categories. This imbalance, combined with up sampling, can skew results and contribute to overfitting. As a result, even after combining PCA with Random Forest or MLP, I hesitate to definitively conclude that these are the primary factors driving coral reef bleaching.

- ***Improper Handling of Outliers***: Outliers, which can heavily influence model performance, were not adequately addressed. I chose to retain them because these extreme values reflect real world bleaching peaks, offering insight into how the models perform under such conditions. However, this decision risks skewing the results, as outliers can distort the overall analysis. Future studies should explore better ways to manage outliers while preserving the integrity of real-world data.

THE END