Algal Bloom Forecast 2024 Erdos Project Executive Summary

Team: Deniz Olgu Devecioglu, Peter Doze, Nathanial Lowry

Github: https://github.com/heineborell/Algae_Bloom_Forecast

Overview:
Harmful algal blooms occur when algae grows out of control and produce harmful toxins. These toxins contaminate the water we drink, and the local ecosystems, and have become a recurrent problem due to climate change. We aim to forecast when algal blooms will occur so that the public can prepare itself.

Stakeholders:
NOAA, State Tourism Board, Fish and Wildlife Department, local fisheries, coastal denizens

KPI:
Predicting/Forecasting an algal bloom accurately, building a model on more reliable measurements, predicting the severity of the bloom (duration and magnitude), predicting severe algal bloom events

Methodology:
We used data collected by NOAA with data ranging from 1950's to 2023. The data covered 4 states Texas, Mississippi, Alabama, and Florida but most of the data came from Florida so we focused on that state. The data has uneven time intervals so we grouped the data with monthly averages for modeling. We used four types of models to predict the time and magnitude of the algae blooms. Only the first two models were used to predict the number of algal blooms.

- **Baseline Forecasting:** Predicts algal blooms based on simpler methods. We used three baseline models: Naive, Average, and Trend baseline.
- **Exponential Smoothing:** Good for forecasting time series data where the underlying structure (trend/seasonal/neither) is strong.
- **ARIMA:** Good for forecasting time series data with a more chaotic underlying structure.
- **Random Forest:** Allows for a multivariate model where we can include the algal blooms dependency on water temperature and salinity.

Results: After using cross-validation to find the optimal parameters for the exponential smoothing, ARIMA, and Random Forest, all of the models had comparable mean square errors but ARIMA performed slightly better than the other models. The double exponential smoothing performed the best on our algal bloom spikes data as there is a clear upward trend.

Future Work: The best quality data was from Florida. We want to investigate the relationships between the Florida blooms and the other states such as Texas or Mississippi. Our EDA has shown that Florida's waters contaminate Mississippi and Alabama. If the blooms in Florida can predict the blooms in other states, we want to see if there is a time lag. This will require us to collect and use wind data which was not readily available for this project.