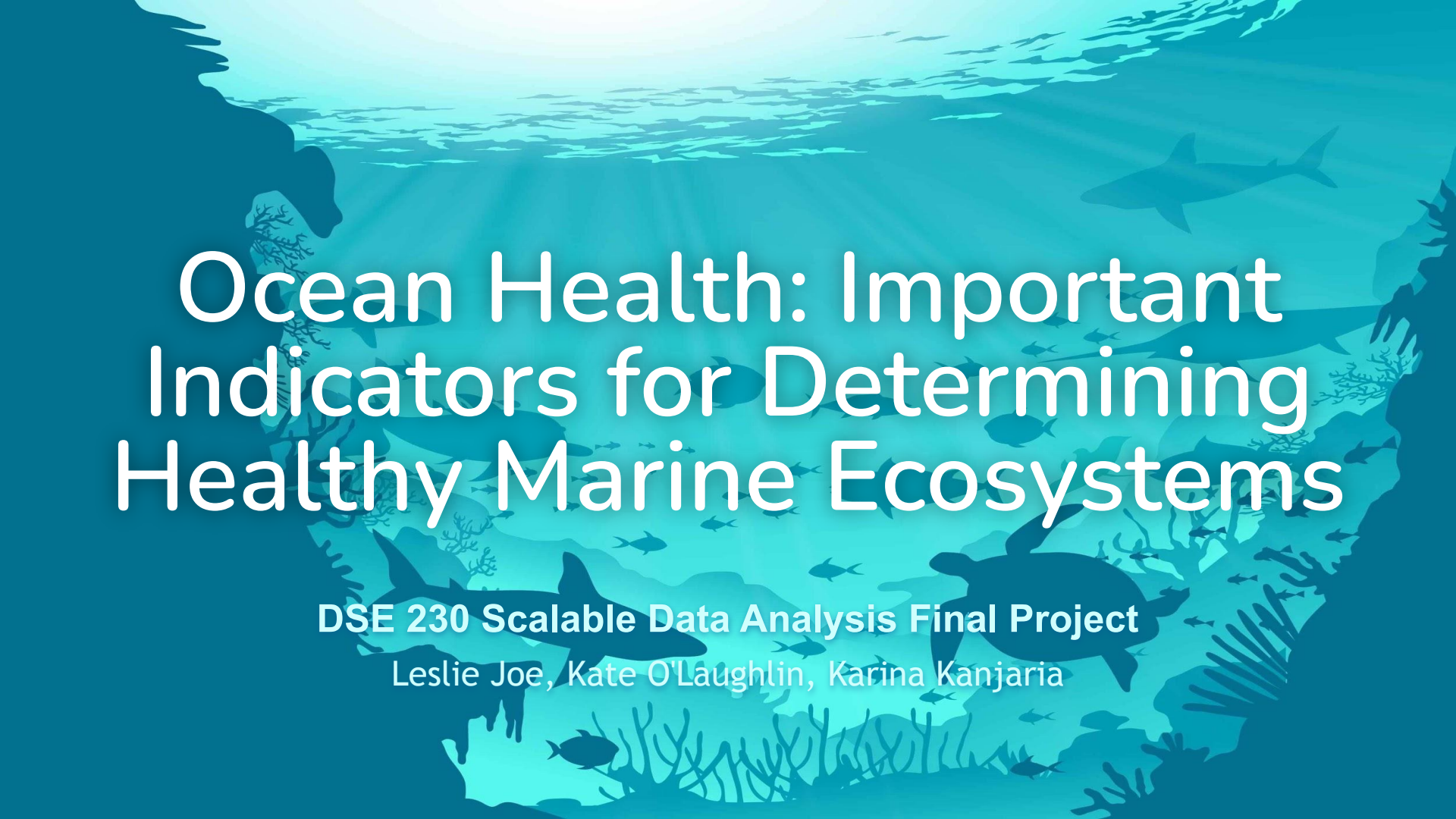


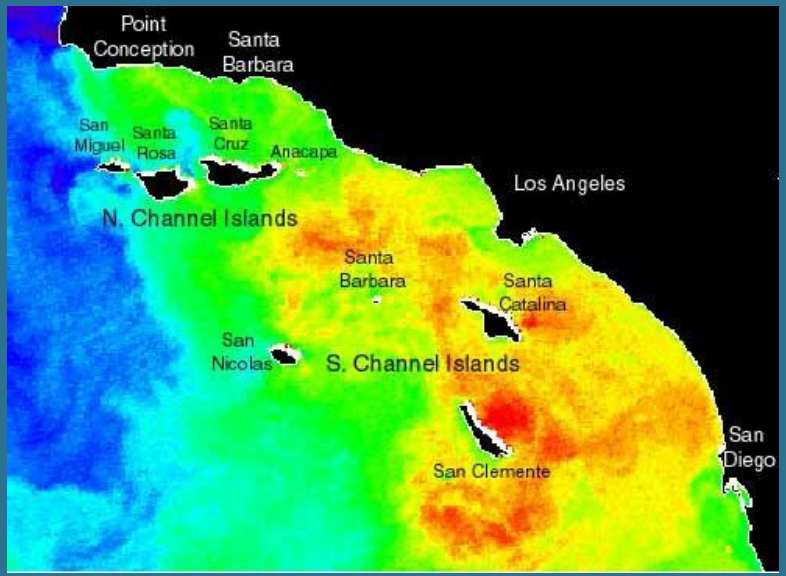
Ocean Health: Important Indicators for Determining Healthy Marine Ecosystems

DSE 230 Scalable Data Analysis Final Project

Leslie Joe, Kate O'Laughlin, Karina Kanjaria

Dataset

- California Cooperative Oceanic Fisheries Investigations (CalCOFI) - Bottle Database:
  - https://www.kaggle.com/datasets/sohier/calcofi
- Dimensions of the dataset
  - 864,850 x 74
- Oceanographic data collected from 1949 to present
- Taken every year on scientific cruises
- Data collected from the coast of California
- Took samples at different depths of the ocean in bottles



https://www.csulb.edu/geological-sciences/southern-california-bight-oceanography

Problem

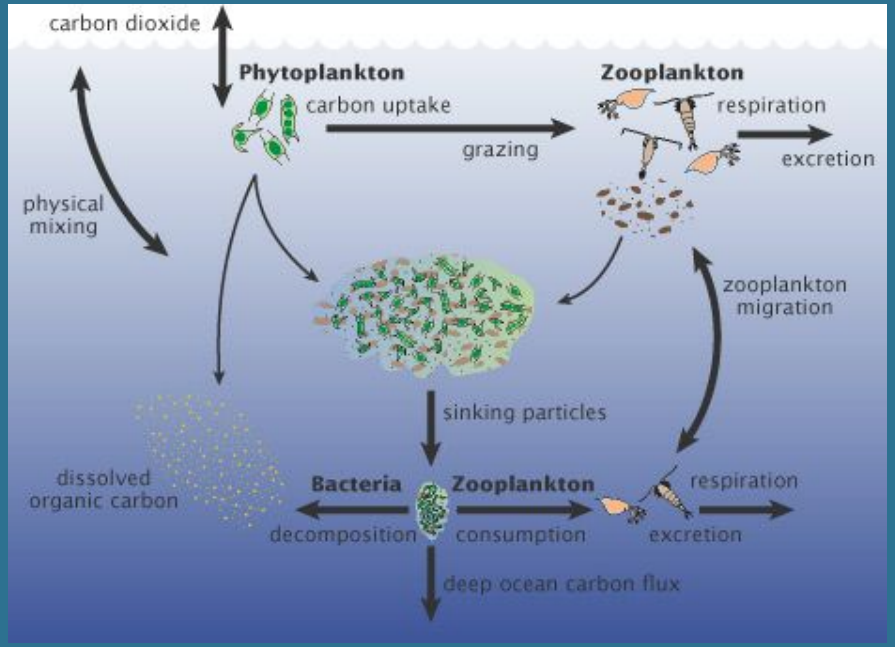
Can we predict chlorophyll levels in a marine ecosystem through measurements of other factors present in the marine environment?

https://mydroll.com/chlorophyll-and-liquid-chlorophyll-benefits-and-risks

If so, which chemicals have the highest influence on the predictions?

Importance

- Phytoplankton are single-celled organisms that are the base of the ocean's food web
- Chlorophyll are the components that perform photosynthesis to produce energy for the phytoplankton
- Chlorophyll absorb CO2 and sink to the bottom, bringing CO2 out of the atmosphere
- Measure abundance of phytoplankton through measuring chlorophyll



https://earthobservatory.nasa.gov/features/Phytoplankton

Data Preparation

```
deleteList1 = ["Sta_ID","Depth_ID","IncTim","DIC Quality Comment","Cst_Cnt","Btl_Cnt","BtlNum"]
deleteList2 = ["T_qual","S_qual","P_qual","O_qual","O2Satq","Chlqua", "Phaqua","PO4q",
               "SiO3qu","NO2q","NO3q","NH3q","C14A1q","C14A2q","DarkAq","MeanAq"]
deleteList3 = (nullCountsTable.columns if nullCountsTable.select(col)>=threshold)
bottle.dropna(subset="R_CHLA")
deleteList4 = ["Depthm","STheta","SiO3uM","NO2uM","NO3uM","Salnty","O2Sat",
               "O2ml_L","Oxy_µmol/Kg","T_degC","R_POTEMP", "Phaeop"]
deleteList5 = [bottle.columns if lassoModel.coefficients == 0]
```

```
correlation coefficient between
    "Depthm" and "R_Depth":
    0.9999999949168986
Null counts per column:
+------+-------+
|Depthm|R_Depth|
+------+-------+
|     0|      0|
+------+-------+
```

- Data Cleaning
  - Removed unnecessary columns and duplicate columns
  - Removed “bad quality” data points using the Quality Code values
  - Removed Quality Code columns
  - Removed features with too few non-null entries
  - Removed rows were our target value was null
- Perform lasso regression to find relevant features
  - Filled in a temp duplicate dataframe’s remaining nulls with the column means for the Lasso model, and used alpha α = 0.25
  - Removed features with coefficient = 0.0
  - Filled in original dataframe’s null rows with the column means
  - Resulting dataset dimension: 222,293 rows × 11 columns
- Feature Correlation
  - Found 4 additional features with highly correlations
  - Removed 2 features with correlation > 0.9 to reduce redundancy
  - Left with 8 final features for model

```
+-------+-------+-------------------+
|   col1|   col2|        correlation|
+-------+-------+-------------------+
|   R_O2|R_O2Sat|0.9881367985673449|
|R_O2Sat|   R_O2|0.9881367985673449|
| R_SIO3|  R_NO3|0.9754158165712739|
|  R_NO3| R_SIO3|0.9754158165712741|
+-------+-------+-------------------+
```

```
  Null values per column before cleanup:
+------+------+------+------+------+------+-----------+------+------+--
|Depthm|T_degC|Salnty|O2ml_L|STheta| O2Sat|Oxy_µmol/Kg|RecInd|T_qual|S_
+------+------+------+------+------+------+-----------+------+------+--
|     0| 10963| 47354|168662| 52689|203589|     203595|     0|841736| 4
+------+------+------+------+------+------+-----------+------+------+--
Dataset size: 864863 x 74

  Null values per column after cleanup:
+------+------+----------+-------+-------+-----+-----+------+-------+
|S_prec|R_TEMP|R_SALINITY|R_DYNHT|R_O2Sat|R_NO3|R_NO2|R_CHLA|R_PHAEO|
|      |      |          |       |       |     |     |target|       |
+------+------+----------+-------+-------+-----+-----+------+-------+
|     0|     0|         0|      0|      0|    0|    0|     0|      0|
+------+------+----------+-------+-------+-----+-----+------+-------+
Dataset size: 205188 x 9
```

Analysis Approaches

- Categorical amounts of chlorophyll levels are more interpretable than numerical values
- For this reason, we change the data to fit a classification model rather than a regression model
  - Changed target values from chlorophyll measurements to categories
- Logistic Regression is standard for a classification problem
  - Multivariable since we have 3 categorical variables
  - Logistic Regression can also output feature coefficients
    - Helps determine feature importance
- Decision Tree Classification to see whether we can improve classification accuracy
  - Reiterates which features are important

Analysis Preparation

- Created categories of low, medium, high chlorophyll values
  - Based on research and data distribution
  - https://www.nature.com/scitable/knowledge/library/the-biological-productivity-of-the-ocean-70631104
- Used 8 features to predict chlorophyll values into target categories
  - low, medium, high : 0, 1, 2
- Split data 80, 10, 10: train, test, validation
- Scaled data using Min Max Scalar
- Checked for any data imbalances
  - Each of data split had proportional target counts

Target	Category	Value Range	Counts
0	low chlorophyll values	0 > x > 0.1 µg/L	113,126
1	medium chlorophyll values	0.1 ≥ x ≥ 1 µg/L	17,227
2	high chlorophyll values	x ≥ 1 µg/L	74,835

Analysis Results: Logistic Regression

- Results
  - Accuracy: 0.882
  - False positive rate: 0.119
  - True positive rate: 0.882
  - F-measure: 0.880
  - Precision: 0.884
  - Recall: 0.882
- Important Features
  - Phaeophytin
  - Nitrite
  - Oxygen Saturation
  - Dynamic Height

Model Coefficients			
Feature	**Low**	**Medium**	**High**
S_prec	0.102	0.000	0.000
R_Temp	1.087	0.040	-1.231
R_SALINITY	-1.234	-6.788	1.425
R_DYNHT	8.066	2.345	-14.874
R_O2Sat	-7.570	1.119	10.940
R_NO3	4.859	-2.126	-1.308
R_NO2	-22.395	15.225	15.436
R_PAHEO	-222.394	19.415	137.248

Parameter Tuning (Snippet)		
regParam	**elasticNetParam**	**Validation Accuracy**
0.0001	0.1	0.880
0.0005	0.3	0.880
0.0005	0.4	0.881
0.0005	0.5	0.881
0.0005	0.6	0.881
0.0005	0.7	0.881
⋮	⋮	⋮
0.0005	0.9	0.881
0.0005	1.0	0.881
0.0010	0.8	0.881
0.0010	0.9	0.880
0.0010	1.0	0.881

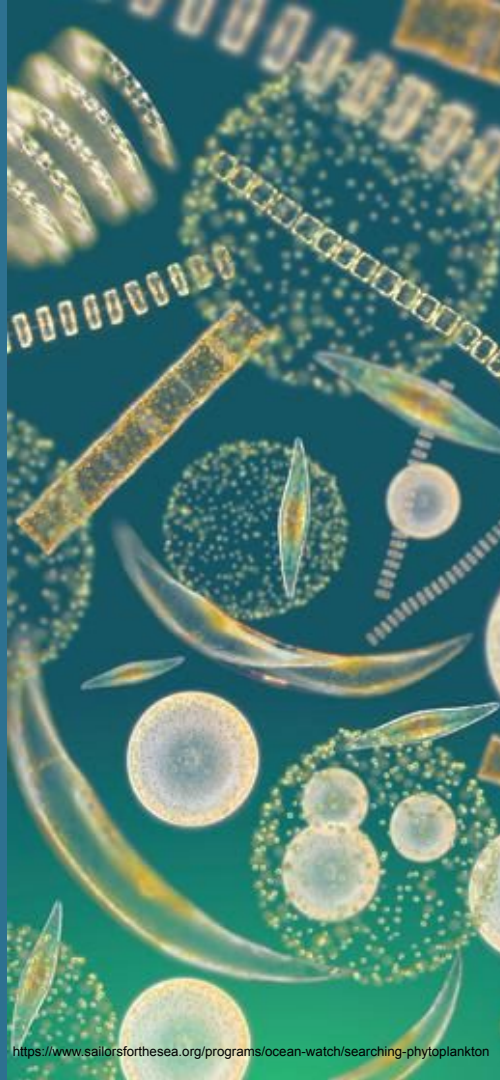
Analysis Results: Decision Tree

- Training Set: 0.882 accuracy
- Validation Set: 0.881 accuracy
- Test Set: 0.878 accuracy
- Important Features Found (in order):
  1. Temperature (°C)
  2. Phaeophytin (µg/L)
  3. Oxygen Saturation (%)
  4. Dynamic Height (m)

Parameter Tuning		
Min Instances per Node	**Train Accuracy**	**Val Accuracy**
20000	0.790	0.793
10000	0.819	0.822
1000	0.878	0.879
900	0.878	0.879
800	0.880	0.879
700	0.880	0.878
600	0.881	0.878
500	0.881	0.879
400	0.882	0.880
300	0.882	0.881
200	0.882	0.881
100	0.882	0.881
50	0.883	0.881
10	0.882	0.881

Challenges and Solutions

- How to handle Quality Column information
  - Used quality codes and removed rows with the code indicating an “uncertain result”
- How to handle overfitting
  - Reduced number of input features to reduce noise
    - Correlation Analysis
    - Lasso Regression
  - Initially 73 features → reduced to 8 features
  - Decision tree minimum sample split
- Which analysis methods should be implemented?
  - Categorical amounts of chlorophyll levels are more interpretable than concentration amounts
  - Changed targets to categories
  - Used Logistic Regression and Decision Tree Classification models



https://www.sailorsforthesea.org/programs/ocean-watch/searching-phytoplankton

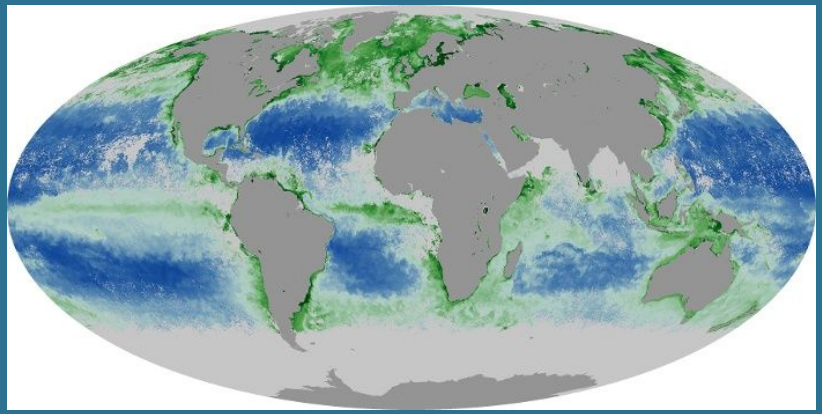
Insights Gained

- Most important features to predict chlorophyll:
  - Phaeophytin (µg/L)
    - a chemical component prevalent in photosynthesis
    - https://link.springer.com/article/10.1023/A:1024990408747
  - Dynamic Height (depth of sample)
    - Most abundant at 200-300m depth
    - https://www.whoi.edu/know-your-ocean/ocean-topics/ocean-life/ocean-plants/ phytoplankton
  - Temperature (°C)
    - Higher temperatures increase phytoplankton growth
    - https://doi.org/10.3389/fmars.2019.00821
  - Oxygen Saturation (%)
    - Phytoplankton produces 20% of all biosphere oxygen production
    - https://oceanservice.noaa.gov/facts/ocean-oxygen.html
  - Nitrite (µg/L)
    - Phytoplankton consume nitrate for fuel
    - https://doi.org/10.4319/lo.1979.24.3.0483
- These features are able to predict chlorophyll levels in a marine ecosystem up to an accuracy of 0.88

https://www.nature.com/scitable/knowledge/library/the-biological-productivity-of-the-ocean-70631104

Future Work

- Can we further reduce the number of features needed?
- Can we further increase the accuracy of our model?
    - Would other classifiers such as KNN or Cluster Analysis produce higher accuracy?
- What is the highest abundance of phytoplankton to increase carbon sequestering, without harming the ecosystem?
- While we know what features can predict chlorophyll concentrations, are there specific weather events or specific latitudes that cause the environment to foster high levels of chlorophyll?



https://earthobservatory.nasa.gov/global-maps/MY1DMM_CHLORA


Thank you
Questions?
https://www.vecteezy.com/vector-art/273920-deep-ocean-background