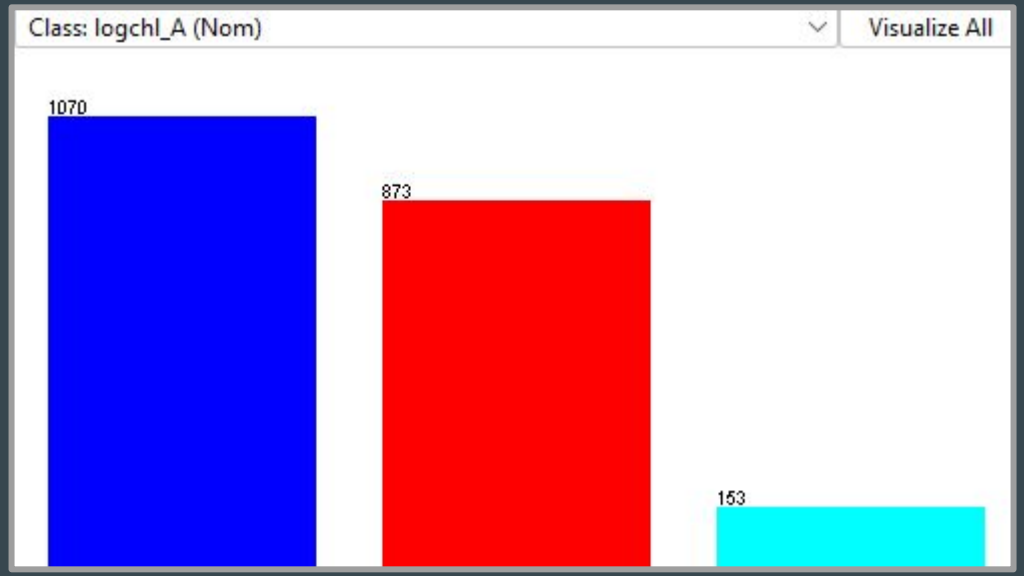
Q1 Project: Chlorophyll-a Predictive Model

Nikhil Alladi, Petr Kisselev, Jacob Dipasupil

Data Information

- 2226 instances (lakes), 67 attributes
- Class: logchl_A
- Heavily right-skewed



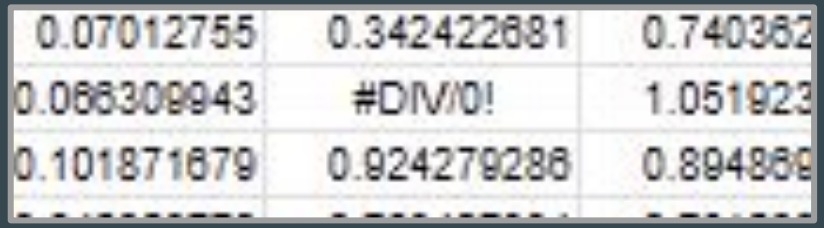
Motivations

- Chlorophyll-a indicates lake health
- High chlorophyll-A concentration leads to algae blooms
- Algae blooms cause:
  - Hypoxia
  - Toxin production
- Early prediction saves health

A harmful algae bloom on Lake Erie.

Preprocessing (I)

- Excel Cleaning (#DIV/0, #NUM, #VALUE)
- Empty class instance removal
- First-pass logic trimming
  - LAKENAME
  - Survey Number
  - SITE_ID

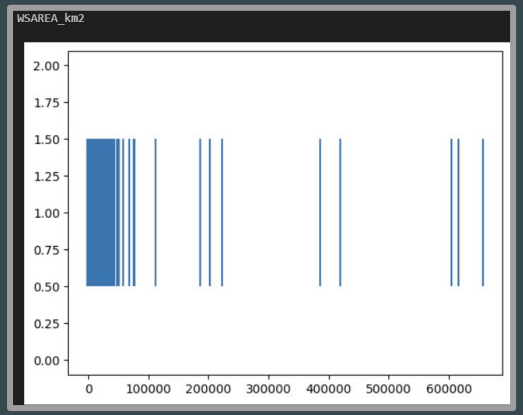


0.07012755	0.342422681	0.740362
0.066309943	#DIV/0!	1.051923
0.101871679	0.924279286	0.894869

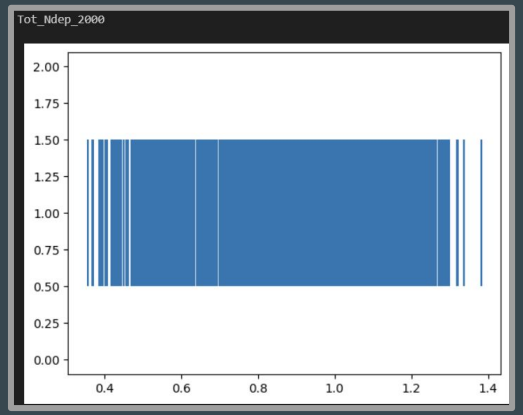
Preprocessing (II)

- Missing values per attribute
  - SNOW

```
0: {'Tmean', 'WSAREA_km2', 'Tot_Sdep_2000', 'Year', 'LST_YrMean', 'Tmean_YrMean', 'OmWs', 'nani', 'SandWs'
1: {"'Total Input'"} -> 0.048%
2: {'logchl_A'} -> 0.095%
32: {'Precip'} -> 1.527%
294: {'wetlands'} -> 14.027%
330: {'SNOW_YrMean'} -> 15.744%
456: {'Human_N_Demand_2007', 'N_Human_Waste_2007', 'N_Fert_Urban_2007'} -> 21.756%
485: {'PctWdWet2011Ws'} -> 23.139%
511: {'AgKffactWs'} -> 24.38%
518: {'P_human_nonfood_demand_kg_2007', 'P_nf_fertilizer_2007', 'P_human_food_demand_kg_2007', 'P_human_wa
615: {'N_Livestock_Food_Demand_2007', 'N_Livestock.Waste_2007', 'PctHbWet2011Ws', 'N_Livestock_N_Content_2
625: {'N_Fert_Farm_2007', 'N_Crop_N_Rem_2007', 'N_CBNF_2007'} -> 29.819%
656: {'P_Accumulated_ag_inputs_2007'} -> 31.298%
659: {'P_livestock_production_2007', 'P_livestock_Waste_2007', 'P_livestock_demand_2007'} -> 31.441%
672: {'P_Crop_removal_2007'} -> 32.061%
713: {'NAPI'} -> 34.017%
718: {'P_f_fertilizer_2007'} -> 34.256%
760: {'Legacy'} -> 36.26%
1080: {'DamDensWs'} -> 51.527%
1925: {'SNOW'} -> 91.842%
```



To be z-score normalized



To be min-max normalized

- Normalization
  - Z-Score vs min-max

Attribute Selection (I)

Learner-Based Information Gain

```
Ranked attributes:
 0.43808     8 ptl
 0.40161     7 ntl
 0.13352     4 lon_dd
 0.12736    59 depth
 0.1254     13 lst_yrmean
 0.10355     9 snow_yrmean
 0.10173    18 tmean
 0.10135    19 tmean_yrmean
 0.10105    57 agkffactws
 0.09842    28 n_fert_farm_2007
 0.0983     12 lst
 0.09745    53 clayws
 0.09506    47 total input
 0.09349    27 n_crop_n_rem_2007
 0.09       36 p_crop_removal_2007
 0.08875    38 p_livestock_demand_2007
 0.08832    14 npp
 0.08732    40 p_livestock_production_2007
 0.08668    39 p_livestock_waste_2007
 0.08393    26 n_cbnf_2007
 0.08101    32 n_livestock_food_demand_2007
 0.08066    34 n_livestock_n_content_2007
```

Threshold: 0.1

Principal Component Analysis

```
Ranked attributes:
 0.717     1 -0.213total input-0.213nani-0.211p_accumulated_ag_inputs_2007-0.211n_livestock_food
 0.5708    2 -0.241n_human_waste_2007-0.239human_n_demand_2007-0.239p_human_waste_kg_2007-0.237p
 0.475     3 0.307lst+0.271lst_yrmean-0.247npp-0.216omws-0.191lat_dd...
 0.4106    4 -0.252lat_dd+0.224atmo_pdep_2002+0.215atmo_pdep_2007-0.211p_human_food_demand_kg_20
 0.3622    5 0.41 wetlands+0.349pctwdwet2011ws+0.314pcthbwet2011ws-0.29depth+0.225ntl...
 0.3244    6 0.284ntl-0.254depth+0.249ptl-0.223napi+0.218n_rock_2007...
 0.2986    7 0.536wsarea_km2+0.51 lake_area_ha+0.237depth-0.21runoffws-0.208npp_yrmean...
 0.2749    8 0.368lake_area_ha+0.346wsarea_km2-0.292sandws-0.255atmo_pdep_2007+0.255npp...
 0.2523    9 -0.469fire_yrmean-0.42fire-0.339p_f_fertilizer_2007-0.246p2o5ws+0.208wsarea_km2...
 0.2326   10 0.328napi+0.321precip-0.265n_cbnf_2007+0.25 ptl-0.246bfiws...
 0.2141   11 0.524n_rock_2007+0.499damdensws+0.257fire-0.24omws+0.196fire_yrmean...
 0.1973   12 0.626p2o5ws-0.36fire+0.278damdensws-0.278fire_yrmean+0.207wetlands...
 0.181    13 -0.583p2o5ws+0.489damdensws+0.247p_f_fertilizer_2007-0.173fire+0.172depth...
 0.1668   14 0.406damdensws-0.397clayws+0.341p2o5ws-0.301pctwdwet2011ws+0.262sandws...
 0.1535   15 -0.458fire+0.358fire_yrmean+0.342pcthbwet2011ws-0.323damdensws+0.319n_rock_2007...
 0.141    16 0.527fire-0.466fire_yrmean-0.249bfiws+0.242pcthbwet2011ws+0.221n_rock_2007...
 0.1293   17 -0.414omws-0.382fire_yrmean+0.343runoffws+0.284p_f_fertilizer_2007-0.264precip...
 0.1186   18 -0.519n_rock_2007+0.334pcthbwet2011ws-0.294omws-0.275tmean+0.251atmo_pdep_2002...
```

Threshold: 0.2

Attribute Selection (II)

Learner-Based with J48

```
=== Attribute Selection on all input data ===

Search Method:
        Best first.
        Start set: no attributes
        Search direction: forward
        Stale search after 5 node expansions
        Total number of subsets evaluated: 651
        Merit of best subset found:    0.784

Attribute Subset Evaluator (supervised, Class (nominal): 60 logchl_A):
        Wrapper Subset Evaluator
        Learning scheme: weka.classifiers.trees.J48
        Scheme options: -C 0.25 -M 2
        Subset evaluation: classification accuracy
        Number of folds for accuracy estimation: 5

Selected attributes: 4,7,8,20,38,40,50 : 7
                     lon_dd
                     ntl
                     ptl
                     atmo_pdep_2002
                     p_livestock_demand_2007
                     p_livestock_production_2007
                     pctwdwet2011ws
```

OneR Attribute Selection

```
Attribute Evaluator (supervised, Class (nominal): 60 logchl_A):
        OneR feature evaluator.

        Using 10 fold cross validation for evaluating attributes.
        Minimum bucket size for OneR: 6

Ranked attributes:
71.56489     8 ptl
68.2729      7 ntl
59.58969    28 n_fert_farm_2007
59.25573     4 lon_dd
59.16031    45 p_accumulated_ag_inputs_2007
58.77863    57 agkffactws
58.6355     20 atmo_pdep_2002
58.58779    36 p_crop_removal_2007
58.54008    56 bfiws
58.49237    33 n_livestock.waste_2007
58.06298    18 tmean
58.01527    39 p_livestock_waste_2007
57.96756    21 atmo_pdep_2007
57.96756    32 n_livestock_food_demand_2007
57.87214    40 p_livestock_production_2007
57.6813     47 total input
57.58588    27 n_crop_n_rem_2007
57.06107    34 n_livestock_n_content_2007
56.91794    38 p_livestock_demand_2007
56.91794    19 tmean_yrmean
56.82252     9 snow_yrmean
56.82252    12 lst
56.7271     13 lst_yrmean
56.29771    53 clayws
56.25        1 nani
56.10687    22 tot_ndep_2000
```

Threshold: 58.5

Attribute Selection (III)

Self-Picked Attributes:

- lat_dd
- lon_dd
- ntl
- ptl
- atmo_pdep_2002
- n_human_waste_2007
- n_livestock.waste_2007
- p_livestock_waste_200a7
- p_human_waste_kg_2007
- runoffws

- Common appearance in attribute selection
- lon_dd vs lat_dd (added both to be complete)
- Focused on attributes related to Nitrogen and Phosphorus

Classifier Models

What we selected:

Naive Bayes

- Predicts probabilities of a given instance through recursively combining probabilities for parts of it
- Surprisingly effective with small amounts of data

Logistic Regression

- Similar to a linear regression but with the logistic function instead
- Parameters are μ (center) and *s* (scale)

Learner Based / J48

- Java-based open source version of the popular C4.5 algorithm
- Creates decision tree by splitting data in the way to maximize information gain

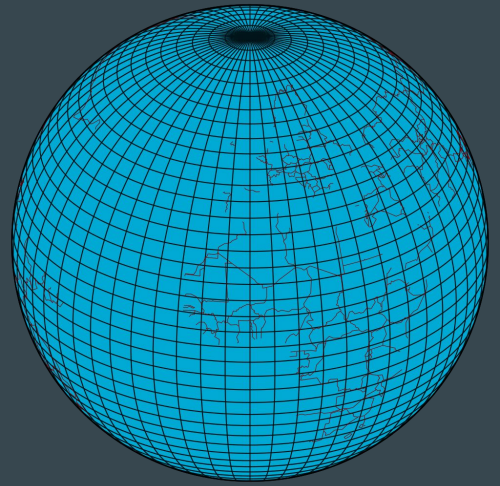
RandomTree

- Constructs a decision tree with *k* branches at each node
- “Random” because at each node the attributes that will be used are randomly selected

Results

Model	Accuracy (%)	TPR High	FPR High	ROC High	TPR Weighted Avg.	FPR Weighted Avg.	ROC Weighted Avg.
InfoGainBayes	72.71%	0.654	0.051	0.935	0.727	0.189	0.852
InfoGainLogistic	78.10%	0.458	0.013	0.948	0.781	0.179	0.879
InfoGainJ48	76.53%	0.641	0.036	0.862	0.765	0.178	0.819
InfoGainTree	68.94%	0.51	0.046	0.732	0.689	0.233	0.728
PCABayes	69.42%	0.209	0.012	0.844	0.694	0.245	0.794
PCALogistic	73.86%	0.307	0.013	0.916	0.739	0.214	0.841
PCAJ48	67.08%	0.275	0.038	0.701	0.671	0.257	0.729
PCATree	61.74%	0.288	0.058	0.615	0.617	0.29	0.664
LearnerBayes	72.47%	0.634	0.058	0.93	0.725	0.198	0.851
LearnerLogistic	76.77%	0.425	0.014	0.945	0.768	0.19	0.869
LearnerJ48	78.01%	0.627	0.022	0.87	0.78	0.174	0.841
LearnerTree	70.66%	0.51	0.046	0.732	0.707	0.22	0.743
OneRBayes	68.70%	0.634	0.057	0.925	0.687	0.236	0.822
OneRLogistic	77.10%	0.438	0.013	0.946	0.771	0.188	0.87
OneRJ48	76.81%	0.588	0.024	0.851	0.768	0.183	0.822
OneRTree	71.37%	0.549	0.043	0.753	0.714	0.215	0.749
HandPickedBayes	72.42%	0.627	0.066	0.914	0.724	0.187	0.846
HandPickedLogistic	77.67%	0.451	0.016	0.951	0.777	0.181	0.882
HandpickedJ48	75.76%	0.588	0.028	0.858	0.758	0.194	0.816
HandPickedTree	71.14%	0.582	0.036	0.773	0.711	0.225	0.743

Conclusion, Limitations, Future Work



- Decently accurate model (78% at best model/attribute pick)
- Use more recent data
- Even class distribution
- Combination of different attributes
    - Many attributes contribute to phosphorus or nitrogen concentrations

Citations

Algae image: https://en.wikipedia.org/wiki/Harmful_algal_bloom#/media/File:Blue-gree_algae_bloom_Lake_Erie.png

Latitude image: By Hellerick - Own work, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=26737079