**DS 410 - Mini Project**Big Data Dinosaurs
Natalie Chow, Dylan Crothers, Ryan Dang, Totton Hollenbeak,
Travis Navarro, Yeji Park

Predicting pH After Extreme Weather Events



# **Data Science Question**

How is pH affected by extreme weather events?



### Motivation

- Extreme weather events have become increasingly more common due to climate change
- These have significant and lasting effects on water quality from surface water, groundwater, and other sources
- Consumption of low-quality water post-event poses significant public health and safety risks to to those living in the affected areas
- <u>pH</u> is an important indicator for contamination

## - Modeling and Analysis



# **Dataset Acquisition**

Origin of Dataset: Water Quality Portal

- Local: 81 x 86824

- Cluster: 81 x 2651847

### Data Processing

- remove unnecessary columns
- filter for specific units
- replace NAs with AVG
- feature extraction: making each feature into individual columns
- remove extremities
  - i.e. pH values not in range 0-14 for normal pH scale

### - Challenges

- difference in unit measurements and data types Oxy

Temperature: deg C, deg F Oxygen: mg/l, %, % saturatn

- replacing missing values with a reasonable mean
- rows represent only a single quality measure for one sample

|-- Temperature, water: float (nullable = true)
|-- Specific conductance: float (nullable = true)
|-- Oxygen: float (nullable = true)

|-- HydrologicEvent: float (nullable = false)

|-- pH: float (nullable = true)

# Requirement for ICDS

#### **Large Dataset**

Dataset encompassed historical data across the U.S. spanning 2 years. Gathering averages for each characteristic for each county across the U.S. required lots of power.

**Power Consumption** 

#### **Frequent API Calls**

With the need for 2
API calls for each
row of our dataset,
ICDS allows us to
manage this high
volume of requests
seamlessly.

Leveraging Spark
on ICDS allows us to
use Spark's full
potential for parallel
processing and built
in fault tolerance to
allow us to run
complex code
smoothly and
efficiently.

Spark + ICDS

#### **ICDS Infrastructure**

reliable
environment for
complex code to
run uninterrupted
with minimal
downtime on
powerful hardware,
unlike personal
laptops and PCs.

# Methods

### 1. Multiple Linear Regression

- a method used to evaluate how strong the relationship is between 2 or more independent variables and one dependent variable

pH = 
$$\beta$$
0 +  $\beta$ 1(Temperature)+  $\beta$ 2(Conductance)+  $\beta$ 3(Oxygen)+  $\beta$ 4(Hydrologic event)

### 2. Decision Tree Regression

 a ML technique used for predicting continuous variables which accommodates non-linear relationships and interactions among variables

### 3. Random Forest

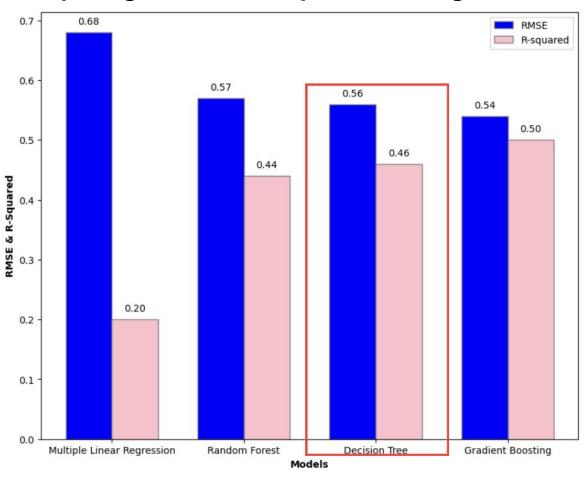
- a ML algorithm that operates by constructing multiple decision trees during training and outputting the average prediction of the individual trees for regression purposes
- combines predictions from multiple trees and mitigates the risk of overfitting and generally yields more robust and accurate predictions compared to individual decision tree

### 4. Gradient Boosting Regression

- a ML technique that constructs trees sequentially, with each tree learning from the mistakes of the previous ones
- is a capable of achieving high accuracy and generalization performance across a wide range of datasets

# Results

## Comparing RMSE & R-squared among the Models



**Inferential Analysis** 

 $\hat{pH} = 0.0268 Temp + .105 Oxygen + 2.67 Conductance - .290 Hydrologic + 6.28$ 

## **ICDS** Utilization

- Changed the number of nodes for each experiment
- Kept memory per node constant16GB
- No correlation found between execution time and the amount of resources

### **Execution Time In Seconds**

