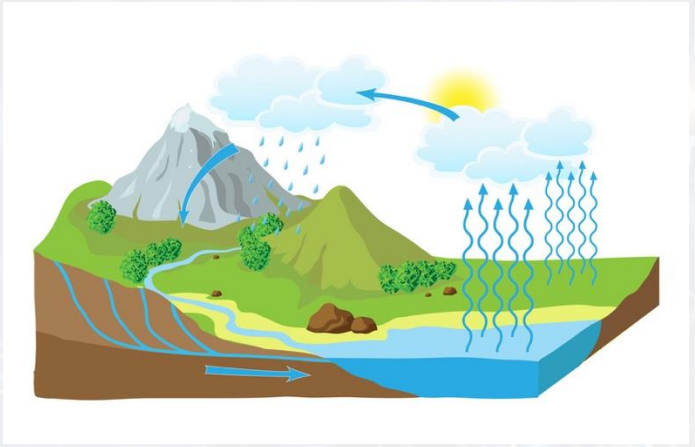
Introduction to Applying XGBoost in Post-Processing Hydrological Models

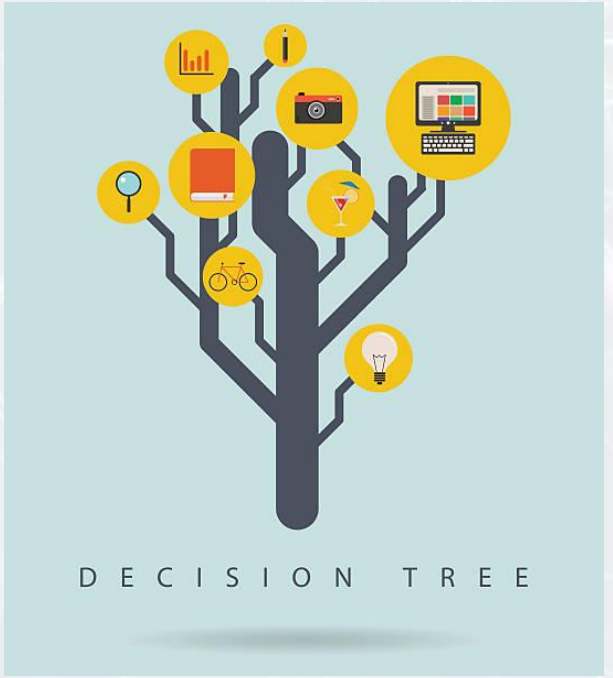
Savalan Naser Neisary[1], Ryan Johnson[2], Dr. Md. Shahabul Alam[2]

[1]Department of Civil, Construction, and Environmental Engineering, The University of Alabama

[2]Alabama Water Institute, The University of Alabama

What are the Workshop Goals?



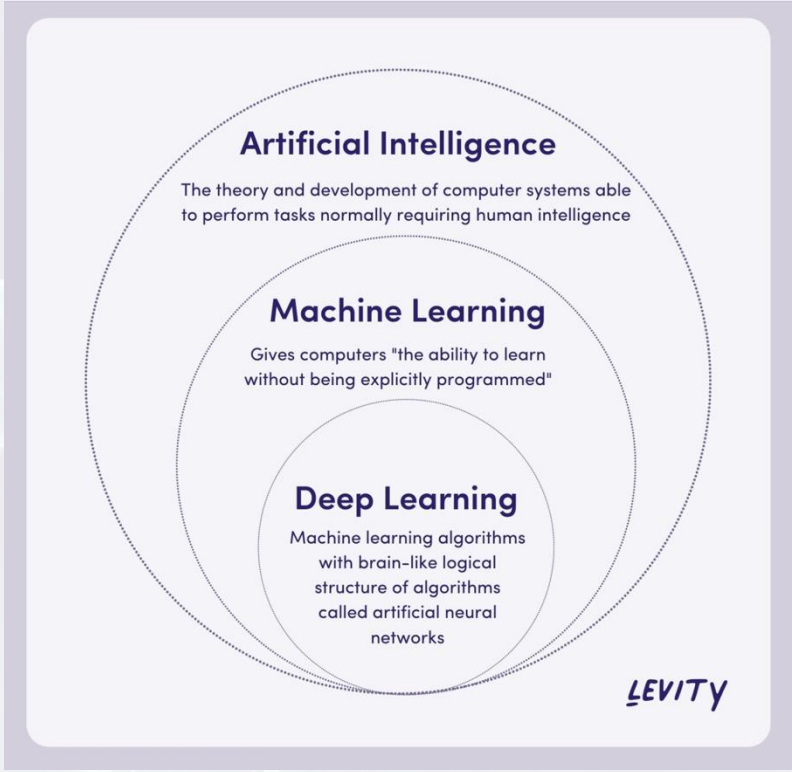


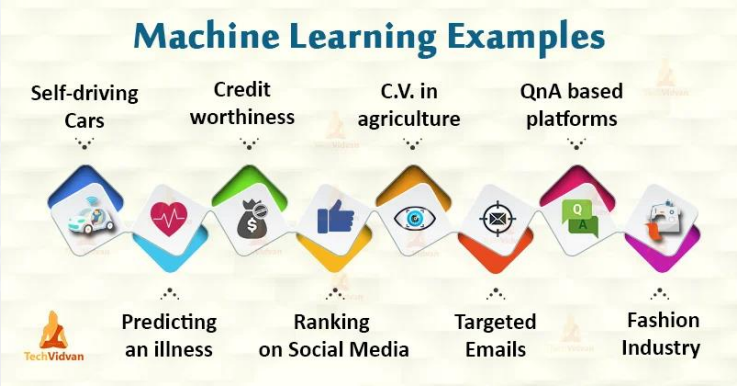
- Understand the basics of machine learning and decision-tree algorithms.
- Learn to apply and train an XGBoost model for hydrological modeling and post-processing.
- Learn how to implement feature selection using the XGBoost algorithm.

What is Machine Learning?

- Study of algorithms that improve their performance for a given task with more experience.

- Spam detection and recommendation systems to medical diagnosis and autonomous driving.

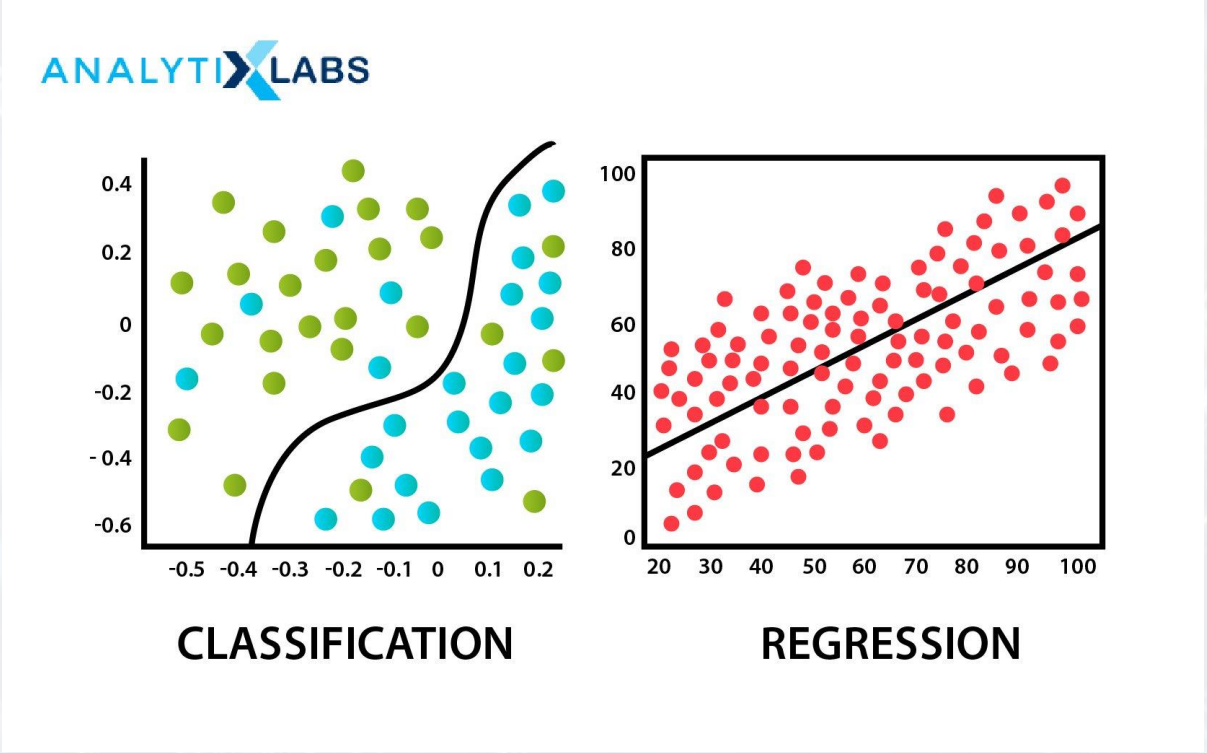




Different Learning Tasks

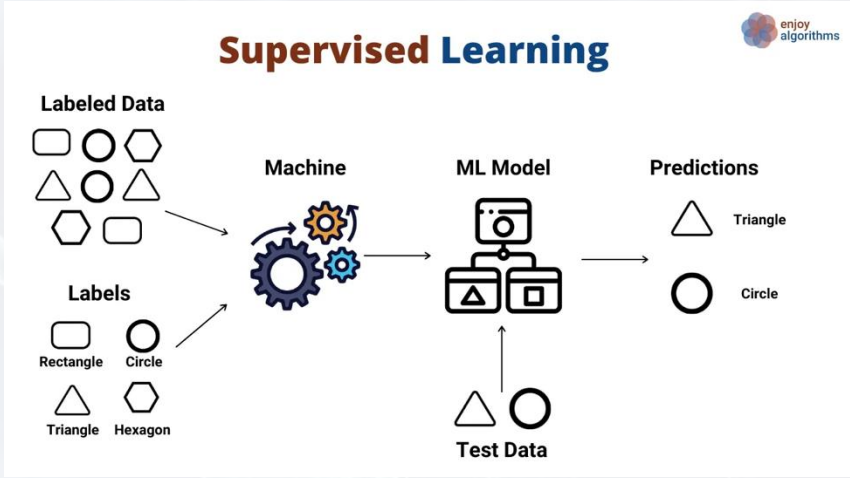
Different Learning Task

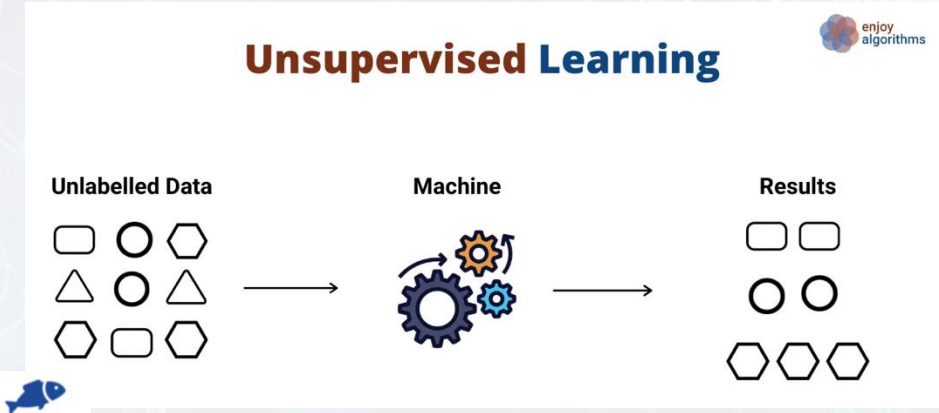
- Classification: Put in categories (classes) based on inputs.
- Regression: Estimate a function/predict a numeric value.

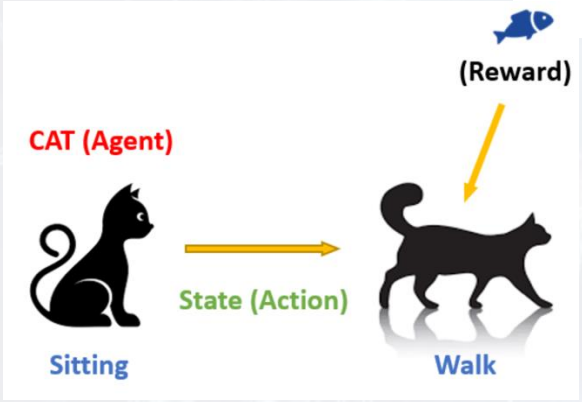


Different Learning Types

- Supervised: 100% expert labeled
- Unsupervised: unlabeled – learn on your own
- Reinforcement learning: The learning system observes the environment and gets rewards based on actions (i.e., training your dog)

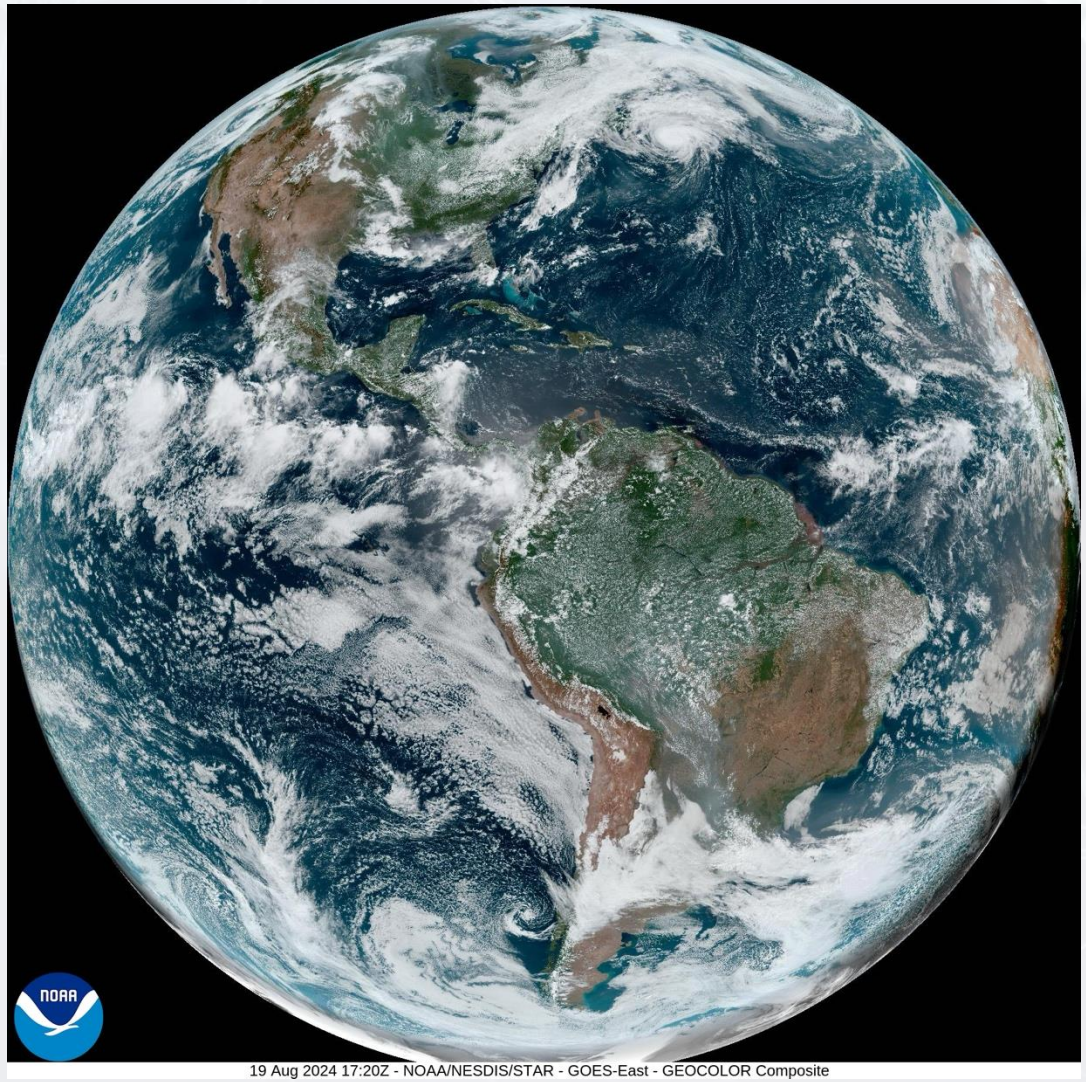






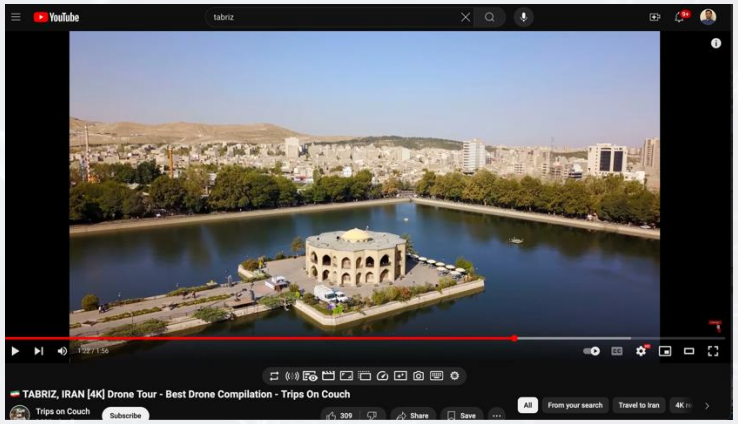
Data Types

Image

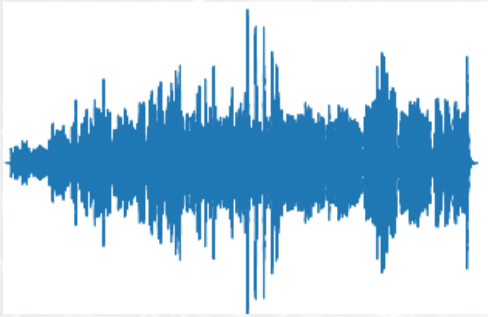


19 Aug 2024 17:20Z - NOAA/NESDIS/STAR - GOES-East - GEOCOLOR Composite

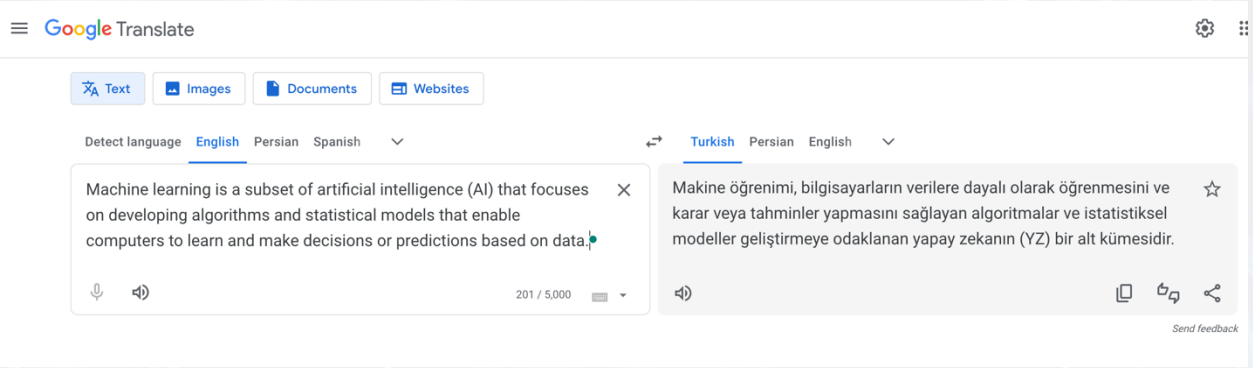
Video



Audio



Text



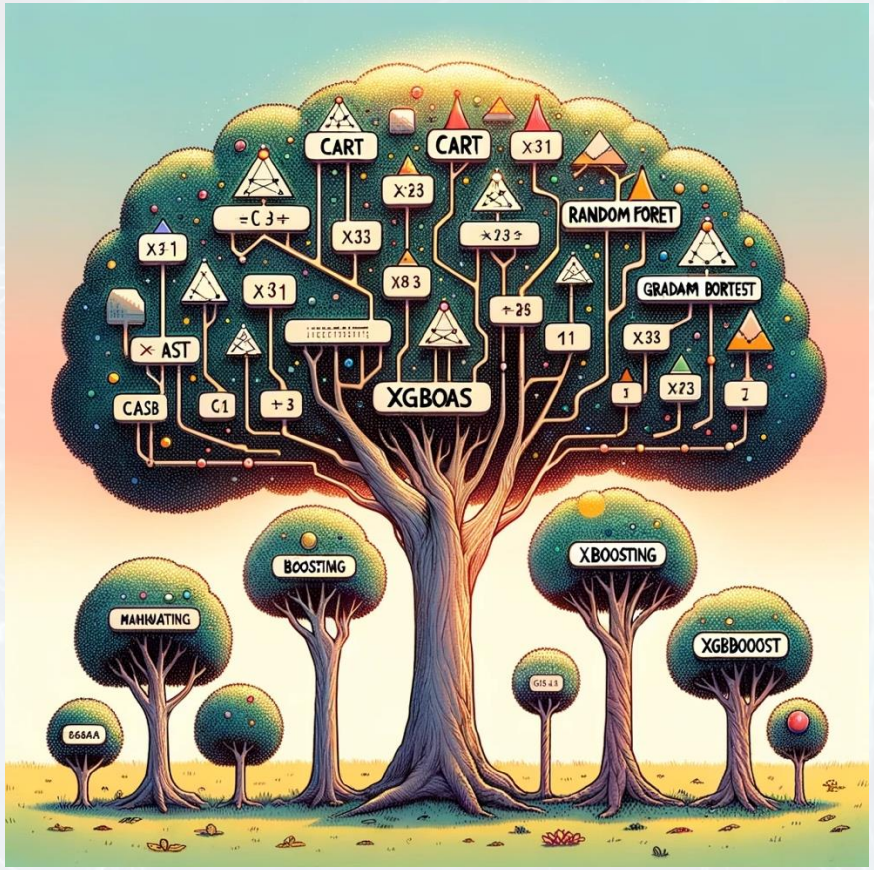
Machine Learning Terminology and Modeling Process

- Train dataset
- Test dataset
- Overfitting
- Underfitting

- Features
- Target
- Hyperparameters

Decision Tree (DT)

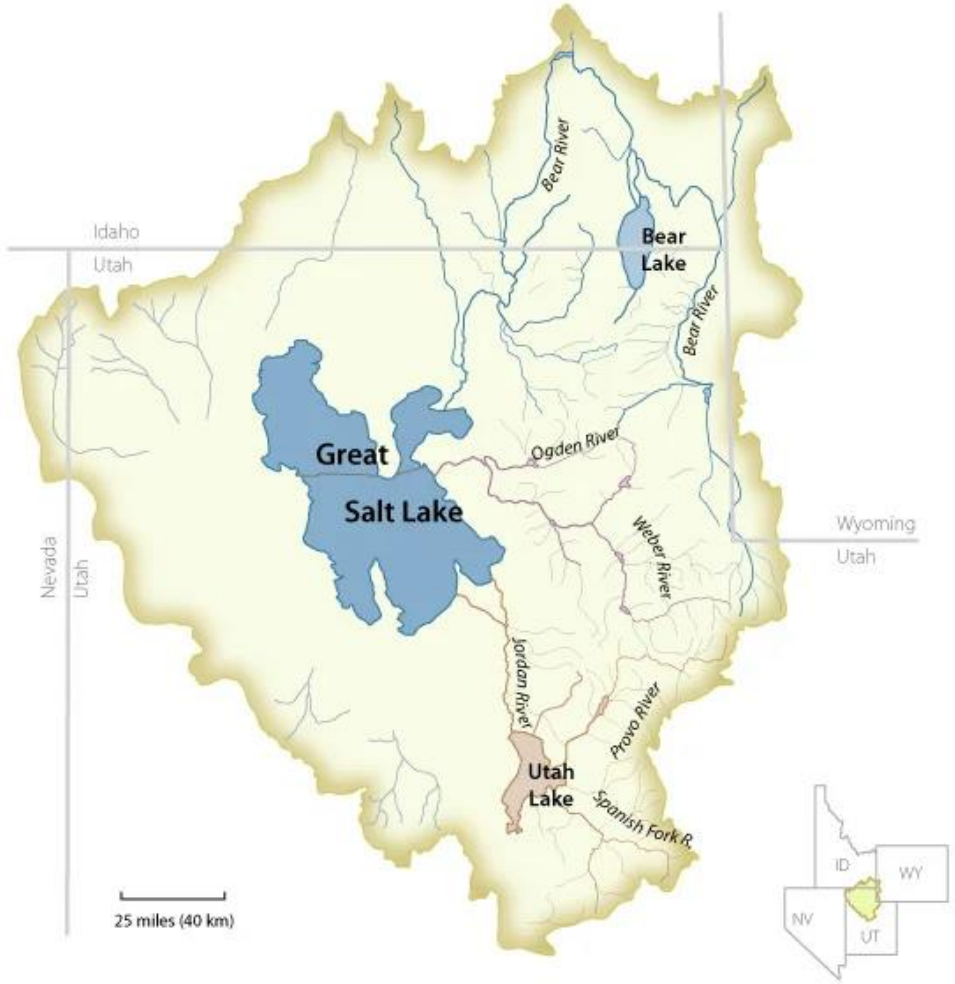
- A decision tree is a non-parametric supervised learning algorithm that performs classification and regression tasks.
- It has a hierarchical tree structure consisting of a root node and branches.
- Conducts a greedy search to identify the optimal split points within a tree.
- It is easy to interpret and requires little data preparation, but it is prone to overfitting.

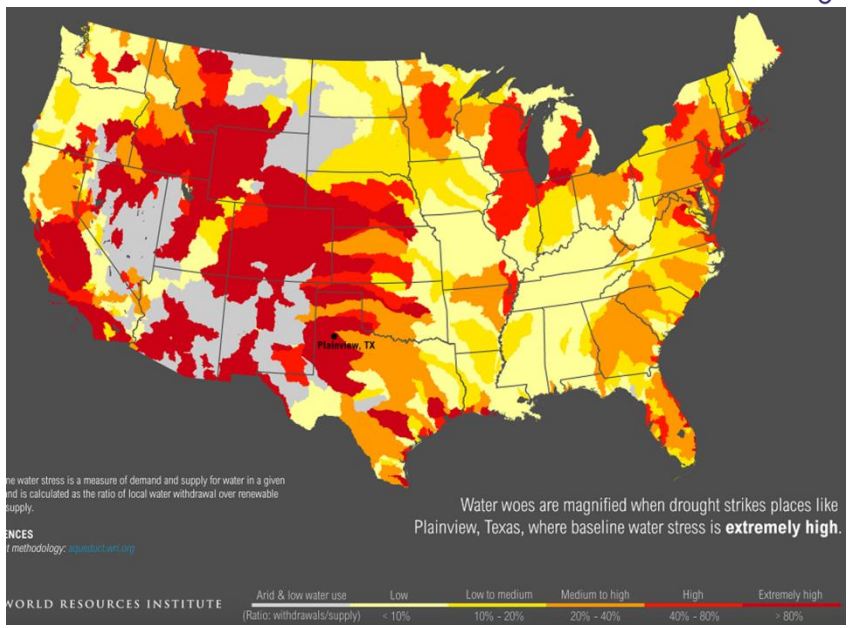


Extreme Gradient Boosting (XGBoost) Algorithm

- XGBoost is one of the algorithms based on the Boosting ensemble method.
- The idea is to train the predictors sequentially, each trying to correct its predecessor.
- Gradient boosting uses a gradient descent algorithm in its core.
- It reduces overfitting but is sensitive to outliers and computationally intensive.

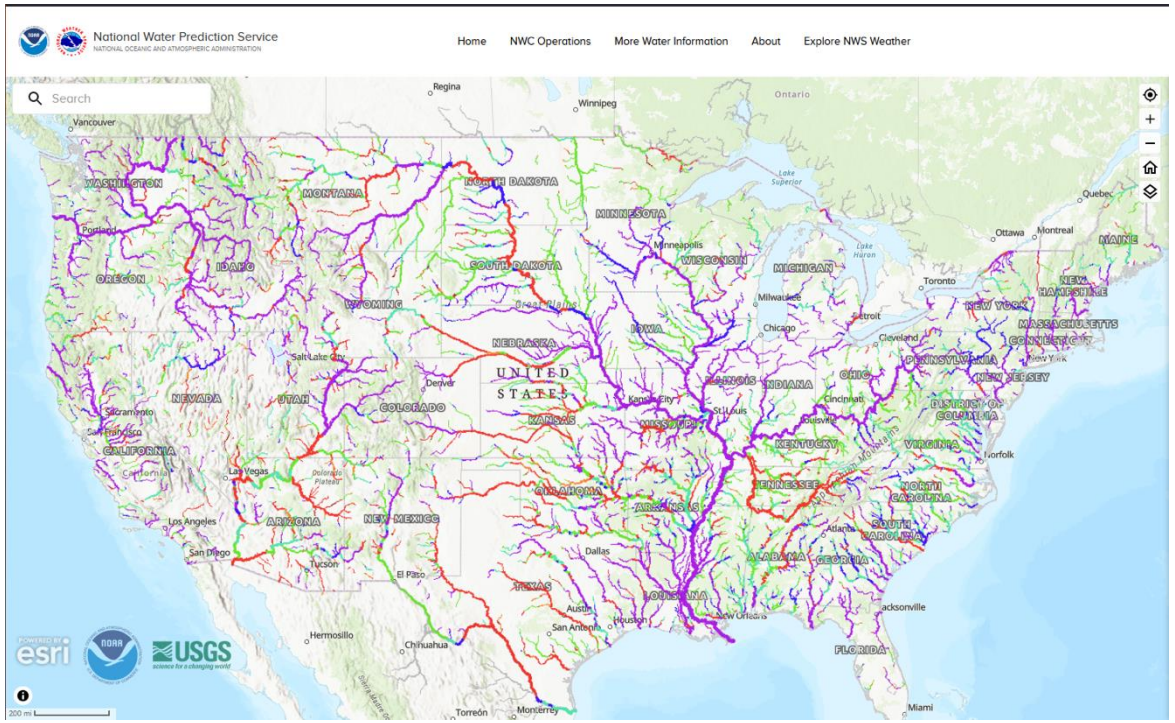
Drought in the Western US





The National Water Model

- Addresses the need for a consistent, large-scale forecast.
- Created by NOAA's Office of Water Prediction.
- Developed based on WRF-Hydro.
- Provides predictions for 2.7 million reaches.
- Our evaluation showed it has low accuracy downstream due to extensive human infrastructure.



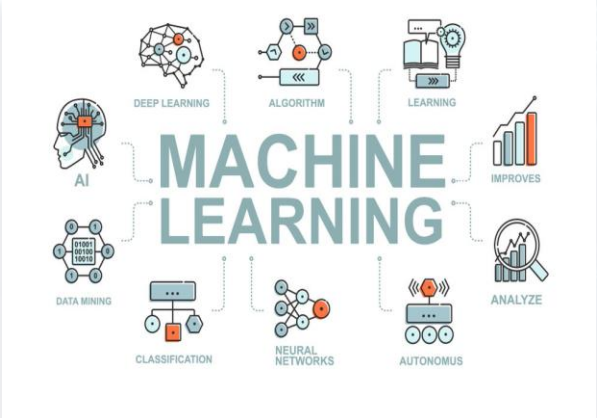
Current River Forecast Points (~3,600)

+

NWM Streamflow Output Points (~2.7 mil)

Post-processing Hydrological Predictions

- There are different ways to improve the predictions, including post-processing.
- Post-processing corrects biases by transforming model outputs based on the relationship between observations and the model.
- ML models proved to be useful in post-processing.



How do we use XGBoost?

Data Set

- 7 NWM reaches collocated with USGS monitoring stations.
- 80% training and 20% test.

Evaluation Metrics/Methods

- Kling Gupta Efficiency (KGE)
- Percent Bias (PBias)

Input Features

- SWE (Snow Water Equivalent)
- Catchment Characteristics
- NWM Flow
- Upstream Storage
- Precipitation and Temperature
- Seasonality Index
- Latitude and Longitude

Tutorial

Book → Tutorial → Decision Tree

Thank You!

Savalan Neisary

snaserneisary@crimson.ua.edu

www.linkedin.com/in/savalann

https://github.com/savalann/

(646)-763-5810