Dask Exercise

- Data
  - “weather_encoded.csv”
  - Weather data
    - Measurements from weather station on Mt. Woodson, San Diego
    - Processed to remove nulls, one-hot encode categorical features, etc.
- Task
  - Build decision tree classifier
  - Perform hyperparameter tuning using scikit-learn and Dask

Dask Exercise Overview

- Scikit-Learn
  - Load, explore, and prepare data
  - Build decision tree classifier
  - Perform hyperparameter tuning with no parallelism
  - Perform hyperparameter tuning using scikit-learn parallelism
- Dask
  - Build decision tree classifier
  - Perform hyperparameter tuning using scikit-learn with Dask backend

Dask Exercise Steps

- Setup
  - Import libraries
- Load data
  - weather_encoded.csv
- Explore data
  - # rows, # columns, column names, etc.
- Prepare data
  - Use ‘RainTomorrow’ column for labels
  - Use other columns except ‘RainTomorrow’ and ‘RISK_MM’ for features
  - Partition data
- Create & train decision tree
  - Use default parameters
- Evaluate model
  - Calculate accuracy on train and test datasets

Dask Exercise Steps

- Set up grid search
  - Parameters:
    - max_depth: 1 to 10
    - min_samples_split: 2 to 10
    - criterion: gini and entropy
- Tune hyperparameters using scikit-learn
  - Print best set of hyperparameters
  - Calculate accuracy on train and test datasets
- Tune hyperparameters using scikit-learn parallelism
  - Print best set of hyperparameters
  - Calculate accuracy on train and test datasets
- Tune hyperparameters using Dask parallelism
  - Start and connect to local client
  - Use scikit-learn with Dask backend
  - Print best set of hyperparameters
  - Calculate accuracy on train and test datasets
  - Close client connection