Group Members:

Seth Jernigan, Ian Stewart, Hamdan Khaled, Daniel Lott

Description of the Problem:

Wildfires pose significant threats to ecosystems, property, and human life. Understanding the cause of wildfires can help inform prevention strategies and resource allocation. Our project aims to classify the cause of wildfire ignition (e.g., lightning, human activity, equipment use) using a supervised machine learning model trained on environmental and geographical data.

Dataset Description:

We are using the 1.88 Million US Wildfires Dataset available on Kaggle. It includes wildfire records from 1992 to 2015 across the United States.

- **Dimensions**: ~1.88 million rows, ~12 key variables
- **Target Variable**: `STAT_CAUSE_DESCR` – describes the cause of the fire (e.g., Lightning, Debris Burning, Equipment Use, etc.)
- **Key Features**:
  - `FIRE_YEAR`: Year of occurrence
  - `STATE`: State in which the fire occurred
  - `DISCOVERY_DATE`: Date the fire was discovered
  - `FIRE_SIZE`: Area burned (in acres)
  - `LATITUDE`, `LONGITUDE`: Geographical coordinates
  - `FIRE_SIZE_CLASS`: Categorized size class of the fire
- **Data Source**: https://www.kaggle.com/datasets/rtatman/188-million-us-wildfires

Supervised or Unsupervised?

Supervised

Regression or Classification?

Classification

Comments and/or Concerns:

The dataset includes class imbalance (e.g., some causes like “Lightning” and “Debris Burning” are far more common than others), which we’ll address using techniques such as resampling or class weighting. We will also need to perform extensive data cleaning (e.g., dealing with missing coordinates or ambiguous cause labels) and may explore reducing the number of cause classes for better model performance.