DS/CMPSC 410 Mini-Project Deliverable 1: Activity Report

Project Title: Wildfires Big Data Analysis

- Team Members:
    - Alvaro Tapia
    - Sanidhya Singh
    - Chinmay Ghaskadbi
    - Jorge R Meza Cabrera
    - Ahmed Mosaad
    - Nakshatra Sharma
    - Evan Dougherty

Activities performed:

1. The first step for our coding process, aside from loading the data without the kernel crashing, was to preprocess the dataset (5.3GB) and perform feature importance and feature selection. For this purpose, we shrunk the dataset into one that only takes the first 100,000 rows (dataset of around 230MB) for efficiency purposes in order to avoid kernel crashing. We plan to later use the best features selected on the original dataset, the main reason we wanted to first tackle feature selection was due to the large number of columns/features we had available because trying to utilize each one within an ML model would not be entirely feasible. As such we will decide which certain columns/features to prioritize that would have a greater potential impact on the model.

   Note: We do possess a 22 GB version of the dataset. However, for the convenience of uploading and running our initial trials, we are going to be running our code on the smaller versions of the data and once we are confident with our code, we shall run it with the larger dataset.

2. We also initiated our first attempt for applying a machine learning algorithm to the dataset which was lasso regression, as there are many variables we aren't positive on whether or not it is useful to the data/model. The HTML we submitted regarding lasso is currently bare-boned as we primarily wanted to use it as a method to help us utilize the naive method of feature selection.

3. After preprocessing the data and selecting the relevant features, we consider performing MapReduce when handling the large-scale dataset (22GB), so far we have developed code for the MapReduce algorithm but have not yet run it along with all the other code since right now we are using a small version of the dataset and this can be an overkill cause it could introduce unnecessary overhead. This process is very important for the future since MapReduce will help us to handle this large dataset by performing parallel processing and increasing the speed of the analysis by making it scalable.

4. With the help of different libraries and sources available we decided to find potential algorithms to perform data visualization. For this first part, we created a correlation matrix algorithm so we can ensure that the features selected are the correct ones and that will potentially bring the most value to our investigation. In the same way, we developed some code that hasn't been run yet because we are still deciding if it will add relevant value to our investigation the code we did was for the creation of heated bar graphs in order to visualize and identify the impact specific features may have if they are studied separately, and also developed some linear models (lasso regression model) after filtering all features and choosing the most important ones for visualization that will be included in the final document report.

5. We also began to research and understand how to use correctly the persist() and cache() methods that we plan to integrate later in the investigation to increase the efficiency of our code and reduce the amount of run-time.

6. We are also actively looking at various techniques learned in recent previous labs and using any techniques that could be helpful for this project. Such as creating k-means instances and performing visualizations. Applying k-means clustering to group similar data points together. This can help identify patterns or regions where wildfires are more likely to occur based on the features provided. The clusters may represent areas with similar environmental conditions that could be conducive to wildfires.

7. Started implementing PCA to find the principal components (Investigate more about this).

Link to GitHub: https://github.com/abt5572/DS410-Alpha

Link to Dataset: https://drive.google.com/file/d/1B582y8_cPWxNuevpm3ZM-SZf_23HRUAQ/view?usp=share_link

Link to Dataset's GitHub: https://wildfire-modeling.github.io/

Notes:

Following the discussion in class we are considering utilizing the naive method (just dropping certain features) to deal with the very high dimensionality of our dataset. The main problem is that we are not experts on this particular dataset nor wildfires, so we will also dedicate some future time to understanding the dataset on a deeper level. There's also a route involving PCA and potentially the deep learning autoencoder but we believe that's a decision we will solidify following the upcoming class lab session.