**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).

**DS/CMPSC 410 Mini-Project Deliverable 2: 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. Principal Component Analysis (PCA) is a technique used to emphasize variation and bring out strong patterns in a dataset. As a first attempt with the creation of the PCA analysis code, we were able to make it run on the 100 MB dataset. After that, the same process was performed in the 5 GB dataset just to make sure it also worked on the official dataset and it effectively ran as expected.
2. Researched some possible ways that we could improve the efficiency of our PCA analysis. Also decided to begin researching and understanding how to correctly use the persist() and cache() methods that we plan to integrate soon in the investigation to potentially increase the efficiency of our code and reduce the amount of run-time.
3. As we were also transitioning from the smaller 200mb dataset to the 5gb version of the dataset we ran into various errors regarding various NaN/null values in a large number of columns for a specific count of 29 rows (in the 5gb), and since we have a large amount of observations we currently believe just removing those rows won't create a signifigant enough of an impact to worry about as we have a very large number of observations to begin with. This decision may change depending on the number of null values that appear in the observations within the large 20+gb dataset. While we could just use dropna() the other code we have can be used for future operations on the dataset as well.
4. After a discussion about what model to implement, we decided to discard using K-means and began to look more into Random Forest Regression modeling on our split dataset.
5. Random Forest regression is a flexible, easy-to-use machine learning algorithm that produces great results. A type of learning model where a group of weak models combine to form a powerful model. In our case PCA analysis is successful, so Random Forest Regression can be used to predict a continuous outcome based on the principal components.
6. There is progress with using map reduce for using the bigger dataset. So far this is a rough draft of the code and will be improved in the future days as we progress with the other parts of the project.
7. Future Steps planned, since we were successful in implementing the PCA analysis we are short-listing some more ways to visualize the results. After visualizing and interpreting the results, we can draw conclusions about our data and start writing the final report document. This will involve identifying key variables that contribute to wildfires and predicting the severity of future wildfires based on certain conditions.

**Link to GitHub:** [**https://github.com/abt5572/DS410-Alpha**](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/>

**DS/CPSC 410 Project Deliverable 1: Project Plan and Preliminary Setup**

**Project Title: Wildfires Big Data Analysis**

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

**Team Composition**

The whole team's plan is to collaborate effectively by constantly communicating our findings, results, and improvements in the project. We will also constantly meet because we all have to know what process to follow for the investigation. Meetings will be done at least once a week, in person preferably.

* Alvaro Tapia: Majoring in Applied Data Science, experience in Python machine learning techniques on developing models and usage of other methods. Also, experience in writing documents and projects. My responsibility in the investigation is to help develop machine learning models and methods such as feature importance, correlation matrix, feature selection, and train-test-validate models in order to find the best model (using PySpark). Also, I will explain the findings in the official report document. Will also support the development of PySpark algorithms such as hyperparameters, MapReduce, data management, and more.
* Chinmay Ghaskadbi: Majoring in Applied Data, my background is predominantly in using machine learning with anonymized medical data. My professional experience is working as a Data Scientist on a project with the American Red Cross as well as an engineering intern with GE Aerospace. My role in this project would be feature engineering as well as implement the machine learning model. Furthermore, I will be heavily involved in writing and development of the project paper.
* Sanidhya Singh: Majoring in Applied Data Science and as such I’ve taken classes working with multiple machine learning models and various datasets. While the majority of the datasets were quite small there were instances in my internships where I was able to learn about classification using larger datasets using models like Xgboost through services like AWS. My primary role will mainly involve hyperparameter testing during the development of this project.
* Jorge R Meza Cabrera: Majoring in Computer Science, I don’t really have experience with data science or machine learning since this is the first data science class I have taken.
* Nakshatra Sharma: In my senior year doing Computer Science. This is my first data science class as I have more experience with computer science and computer engineering classes. My professional experience has ranged from being a part of a management consulting firm McKinsey’s data science team called Quantum Black. Further, I also have experience working with Penn State’s legal team.
* Ahmed Mosaad: Majoring in Computational Data Science. I am a senior, and I have had a mix of classes like computational stats which is the math behind machine learning and other math and stat classes. This is my second machine learning class. I am currently taking another one that introduces the algorithms and basics of machine learning. I have experience with R and making prediction models and such, I also have had experience with using AI or chatbots to boost effectiveness in company workflow.
* Evan Dougherty: I am a senior majoring in computer science. I have taken CMPSC 448, which was about machine learning so I have some experience with that. I also worked with databases for my 2 internships (Veradigm and Peraton), but it was not extensive. Given that I have more experience on the ML side of things, I plan on helping to build the ML model.

**Definitions**

FRP score: The likelihood of a fire spreading to an adjacent area within 1 day.

**Project Objectives**

Within this project our goal is to predict the FRP score using machine learning to determine the likelihood of a forest fire happening while also utilizing and improving our skills for data cleaning and preparation we acquired in class on a significantly larger dataset.

**Dataset Selection**

For the purpose of this investigation, our group selected the WildFireDB as the data source. This dataset is significant as it utilizes previous historical wildfire occurrences with relevant features extracted from satellite imagery to form an open-source wildfire dataset. This is the first open-source wildfire dataset that contains information about past wildfire occurrences and also contains data about features extracted from satellite imagery. It contains 17 million data points collected from 2012 and 2018. There is a small version of this dataset with a size of 5 GB and also a full version with a size of 30 GB. This dataset contains cumulative information from different websites and sources such as Landfire for topography, Visible Infrared Imaging Radiometer Suite (VIIRS) for thermal anomalies, Metostat for weather data, and some other raw data from the National Oceanic and Atmospheric Administration (NOAA).

**Project Timeline**

* 10/6/2023: First meeting in person to talk about our initial project proposal and start working on the document. Also organizing future meetings and roles that each of us is going to have.
* 10/12/2023: Planned to be the second meeting to develop an approach to tackling the problem and finalizing the dataset as well as finishing the project proposal document.
* 10/18/2023: Meet again either online or in person in order to start implementing feature engineering data algorithms. Also, fix or add any final details to the project proposal.
* 10/27/2023: Start development for the purpose of the investigation, get graphs, tables, images, and more; in order to later start the final report of the project.
* 11/10/2023: Finish code and implement optimization techniques
* 11/17/2023: Implement the findings and results from the code and begin writing the report.
* 11/24/2023: Finish and submit, before this, we should contact the professor to see if everything looks good.
* 11/27/2023 Submission

No issues found yet, it seems that our goal is very clear and the dataset is not hard to manage. We are still testing the small dataset (5GB) and in the future, we will start trying to implement the 30GB dataset.

**Tools and Technologies**

Besides the machine learning methods and model implementation that are going to be used for feature importance, classification, and selection, as well as finding accurate models for prediction, we are going to use the following features and algorithms from the PySpark library:

### Try to find the methods for feature importance selection and ml models to be used in Pyspark.

### CheckPoint to improve the efficiency of iterative processing using Spark

* PCA Analysis to reduce the dimensionality of our dataset and prepare it so that we can better run machine learning on it.

### MLlib to implement machine learning models.

### Hyper-parameters identification for better evaluation of the chosen model with testing data.

### Utilize various Python libraries to establish a form of error evaluation during the evaluation of wildfires.

### Use persists and/or checkpoints to improve performance.

### If useful, use what we learn in the future (Lab8, Lab9, Lab10).

**Project Plan Summary**

The overarching goal of our project is to utilize the WildfiresDB data and clean it so that it is then usable for potential feature analysis and wildfire prediction using machine learning. WildfiresDB is an open-source dataset that links wildfire occurrence with relevant features. These features that were extracted from satellite imagery include pieces relating to weather, vegetation, and the canopy.

The project aims to predict the Forest Fire Probability Rating using Alternating Least Squares. This prediction will help determine the likelihood of a forest fire occurrence. Additionally, the project seeks to enhance skills in data cleaning and preparation on a larger dataset compared to what was covered in class.

In order to complete this project, we will have to improve our skills in some areas:

* Improved data-cleaning knowledge
* Better understanding of how to do PCA on a large data set
* Collaboration
* Project Management
* Communication and Documentation
* Understand the computation behind our models
  + Numerical literacy to understand the dataset

By working on these skills, the team will be well-equipped to clean the WildfiresDB data, perform feature analysis, and build an accurate wildfire prediction model using ALS. This will not only lead to a successful project outcome but also enhance the team's capabilities in data science and machine learning.