****

**IST 718 – Big Data Analytics**

**M002 | Group 15**

# **Predicting Forest Fires with Spark Machine Learning**

**Submitted by:**

**Zhiwei Wang | Aman Awana | Anjali Chintam**

**Objective**

Every summer wildfires become front-of-mind for thousands of people who live in the west, Pacific Northwest, and Northern Rockies regions of the United States. [Wildland fire suppression costs exceeded $2 billion in 2017](https://www.usda.gov/media/press-releases/2017/09/14/forest-service-wildland-fire-suppression-costs-exceed-2-billion), making it the most expensive year on record for the Forest Service. Fires also tend to explode in size. It is not unusual for fires to grow by 40,000 acres in one day when winds are high, and the terrain is steep. Hence, we are presenting one way in which data science could be applied within the context of streamlining firefighting operations in order to reduce costs and response time. We are attempting to predict the area damaged due to Forest fires and also try to minimize the cost and time required to respond to fires by identifying where firefighting assets should be staged such that they are as close as possible to where fires are likely to occur.

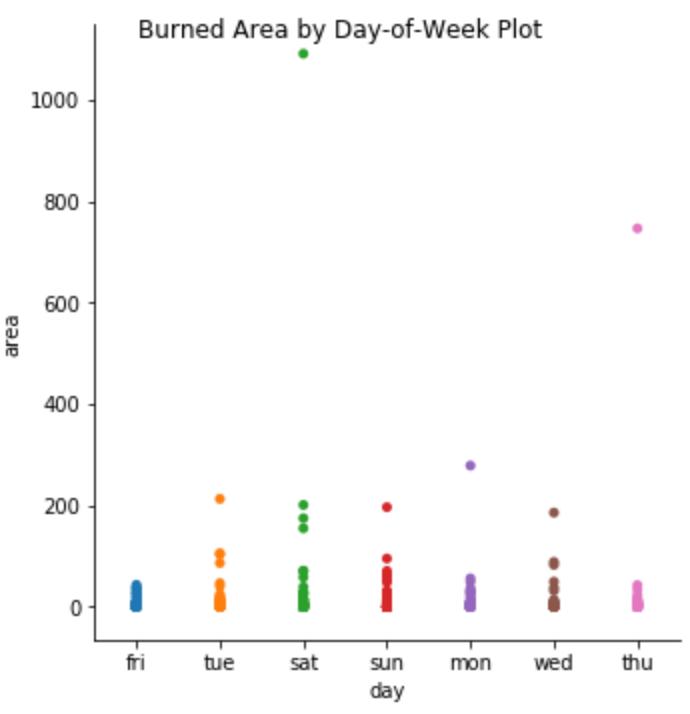
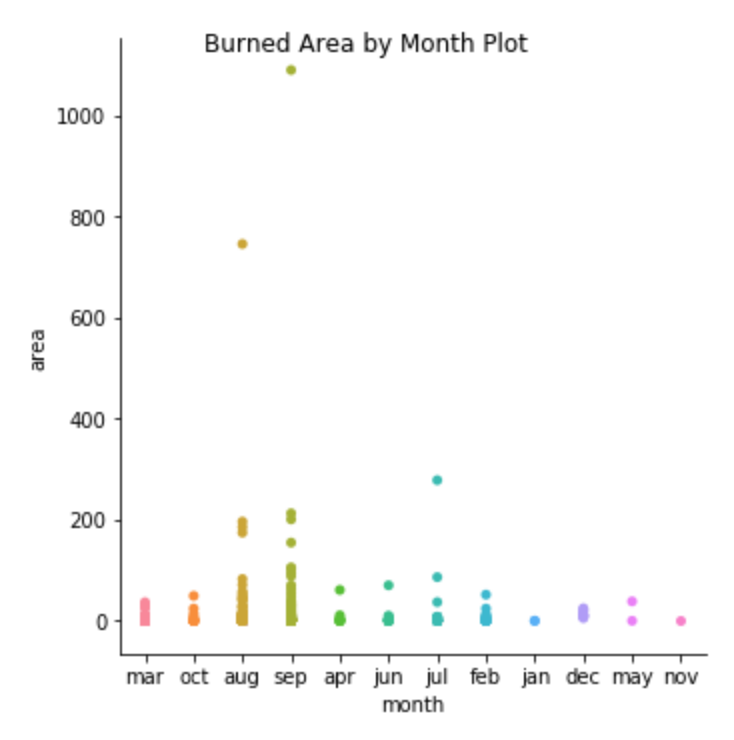
**Data Set Description**

The Forest Fires Data Set contains meteorological and other data for prediction of the burned area of forest fires Portugal. The data set was donated to the UCI Machine Learning Repository in 2008 and it is publicly available for research. The link to the dataset is <http://archive.ics.uci.edu/ml/datasets/Forest+Fires>.

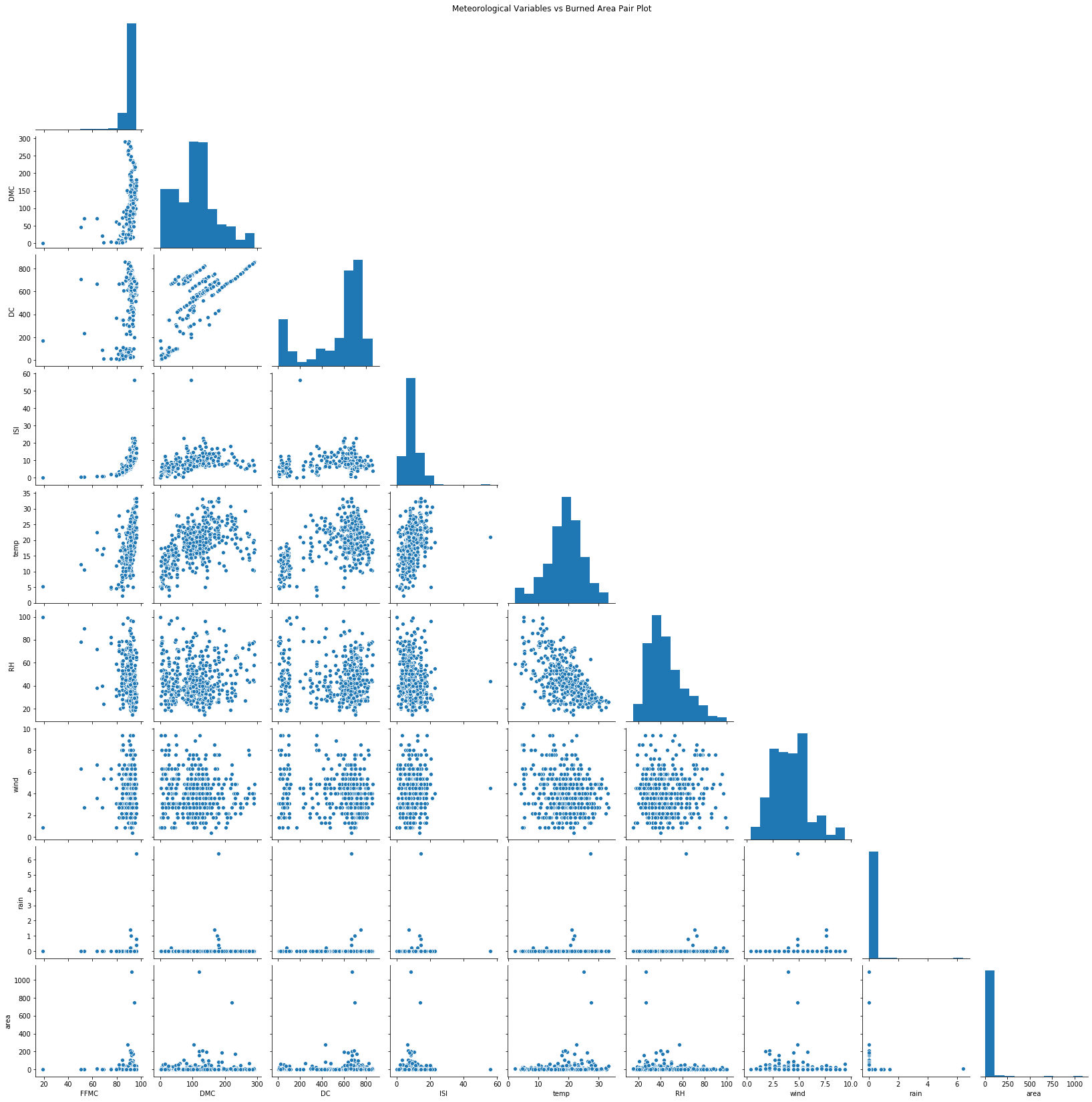
The data set contains 517 rows and 13 columns. The first four variables are about spatial and temporal information: ‘X’ (x-axis coordinate in the map, a numerical variable), ‘Y’ (y-axis coordinate, a numerical variable), ‘month’ (a categorical/string variable from ‘jan’ to ‘dec’), ‘day’ (day of the week, a categorical string variable from 'mon' to 'sun'). The next nine predictors are all numerical variables about meteorological information: ‘FFMC’ (Fine Fuel Moisture Code), ‘DMC’ (Duff Moisture Code), ‘DC’ (Drought Code), ‘ISI’ (Initial Spread Index), ‘temp’ (temperature in Celsius degree), ‘RH’ (Relative humidity), ‘wind’ (wind speed in km/h), and ‘rain’ (outside rain in mm). The last column is the output variable ‘area’ - the burned area of the forest in hectare. It is also worth noting that there is no missing value in the dataset.

**Preliminary Data Exploration**

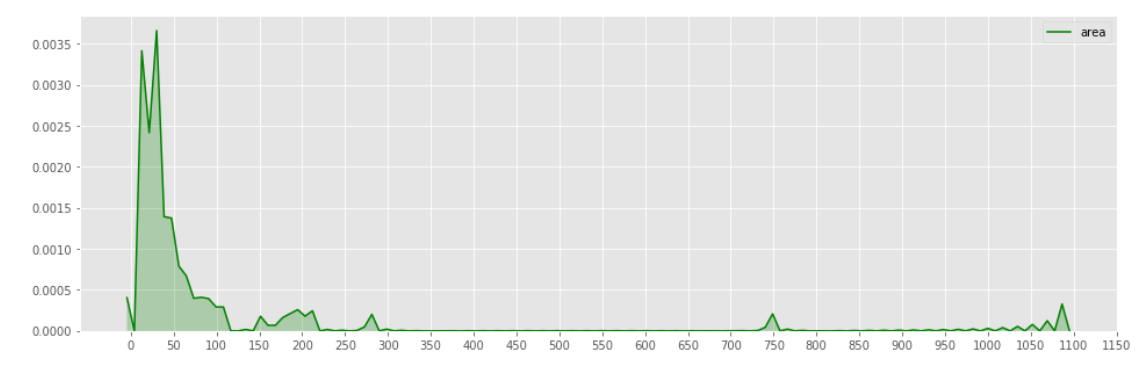
First, we explore the relationship between burned area (output variable) and temporal information. The ‘Burned Area by Month Plot’ suggests that the forest fire resulted in larger damage in July, August and September than the rest of the months. Meanwhile, the ‘Burned Area by Day Plot’ suggests that forest fires burned areas are roughly evenly distributed in a week, while Friday and Thursday seem to have smaller burned areas.



Second, let us take a look at the relationship between meteorological data and burned area.

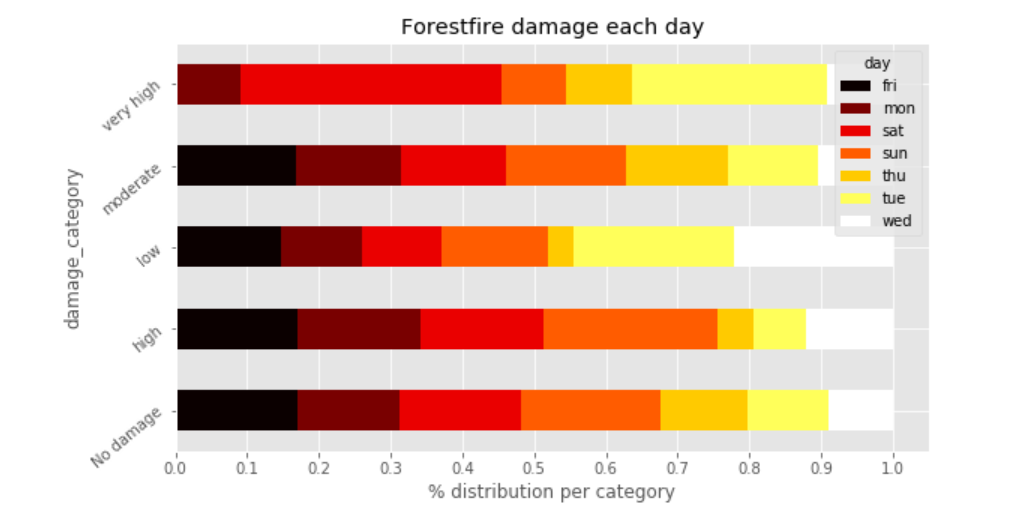


As we can see in the graph, the variable ‘area’, ‘rain’, and ‘ISI’ are very skewed towards 0, while the variable ‘FFMC’ is strongly skewed towards its maximum value (100). Therefore, it is reasonable that we conduct a logarithm transformation of these variables when we build predict models. Meanwhile, except that ‘DMC’ and ‘DC’ seem to be positively correlated, we find no significant patterns among other variables.

I

Skew: 12.846933533934868

Kurtosis: 194.1407210942299

The above plot shows us the skew and kurtosis of the target variable which is the area burned due to the Forest fires.

The above plot gives us the the intensity of forest fire damage caused on different days of week.

**Predictions**

* To predict the area burned based on the intensity of the forest fires.

**Inference**

* To explore how the spatial and temporal information can influence the forest fire level, so the department can allocate their resources accordingly.
* To use clustering techniques, so that the Fire department will be well equipped with their gear at the locations where they can control the increase of fires in those clusters.
* To explore how variables like humidity, wind, and temperature can affect the forest fire level, so the department can set up an alarm based on the meteorological information.

**Non-Spark Packages**

* NumPy
* Pandas
* Seaborn
* Matplotlib