Team Fire Squad
(Jonathan Ortiz, Raya Young, Tiffanie Mac Donald, Jackson Shands, Sanjeet Saikia)
Professor Bolton
IST 707

A Deeper Look at WildFires: An Analysis of Forest Fire Data using Classification Methods in R

Introduction

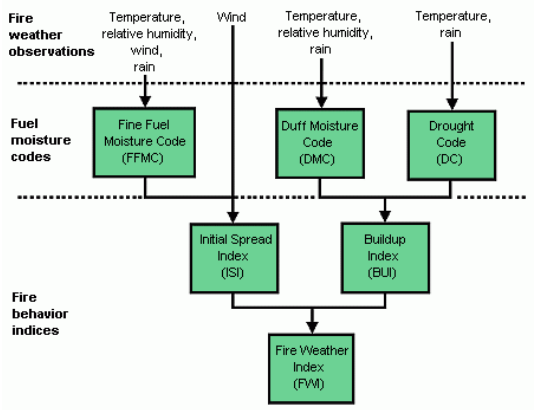
Each year, uncontrollable wildfires cause great devastation to our environment, ecology and economy. They put many that sacrifice to contain the fires at risk and kill many helpless animals. These affected areas take several years to recover from the damage and impact left behind, costing governments billions of dollars to address.

Meteorological data is often collected automatically with instruments placed throughout an area. These instruments can collect data on criteria such as wind speed, humidity, rainfall, and other such environmental characteristics.

Meteorologists were tasked with helping to predict when fires may become prevalent during certain weather conditions. Whenever a wildfire was observed, specific data was collected in hopes that the data could help determine fire watch conditions.

The Fire Weather Index (FWI) was created in 1992 to estimate the risk of a wildfire. The index was computed by Meteo France and the Meteorological Service of Canada. It was first introduced in France and was based upon a Canadian empirical model that has been in place since the 1970's. The meteorological data that was collected for use in the model are temperature, relative humidity, wind, and rainfall amount. Along with the meteorological data, three fuel moisture codes were collected, the Fine Fuel Moisture Code (FFMC), Duff Moisture Code (DMC) and Drought Code (DC). Together with the meteorological data, they yield the Fire Behavior Indices, Initial Spread Index (ISI) and Buildup Index (BUI) which are combined to compute the overall Fire Weather Index (FWI).

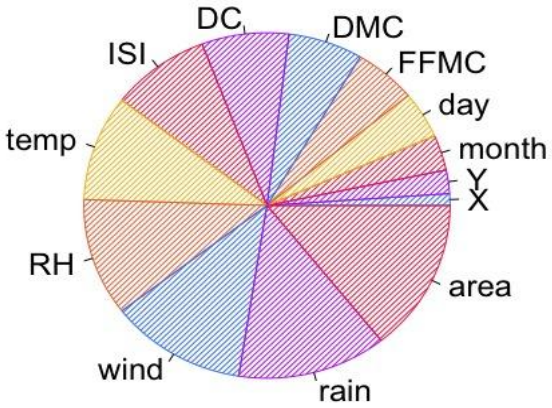
Structure of the FWI System[1]



One particular area of the world which experiences a high frequency of wildfires is the Tras-os-Montes region of northeast Portugal. It is a very well-travelled tourist destination called Montesinho Natural Park. This paper's intent is to use classification methods to determine the likelihood of a significant burn area, by analyzing a combination of meteorological data and environmental data, like soil quality, by utilizing a dataset obtained from the UC Irvine Machine Learning Repository.

About the Data:

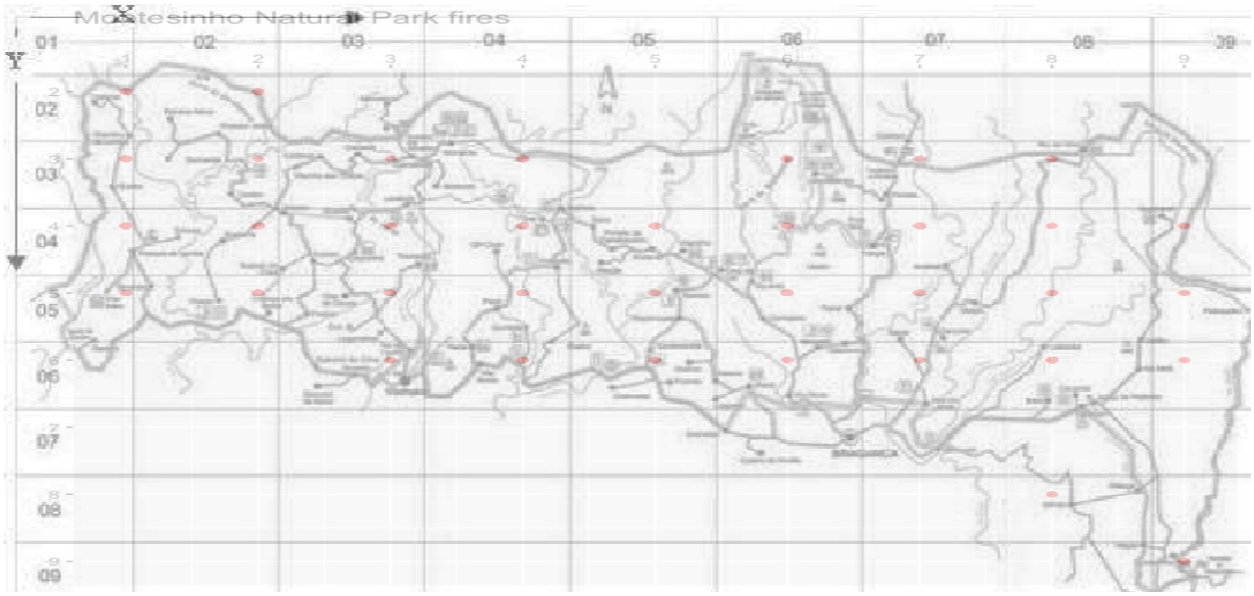
[1] https://cwfis.cfs.nrcan.gc.ca/background/summary/fwi



The data that is utilized in this project was collected by the Department of Information Systems, University of Minho, Portugal[2]. It consists of many variables including the x and y coordinates of the fires, the month, day, FFMC, DMC, DC, ISI, the temperature in celsius, the relative humidity on a scale between 15 and 100 percent, the wind speed, amount of rain, and the area burned.

The Fine Fuel Moisture Code (FFMC) refers to the moisture level and is an indicator of the flammability of fine fuel. The Duff Moisture Code (DMC) also refers to moisture and fuel consumption rate in soil and organic layer composition. The Drought Code (DC) is a moisture rating of deep organic layers and is an indicator of smoldering effects in large logs. The Initial Spread Index (ISI) is a rating of expected fire spread.

[2] https://cwfis.cfs.nrcan.gc.ca/background/summary/fwi



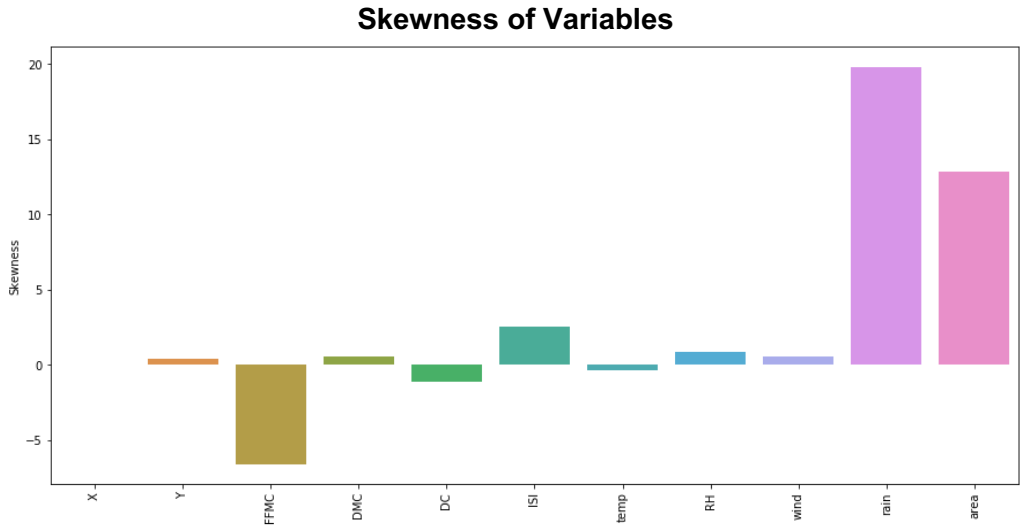
Montesinho Natural Park (Fires marked in red dots)

In the figure above, the points marked in red display the areas within the park where a fire occured. The initial spread is not accounted for above.

Pre processing

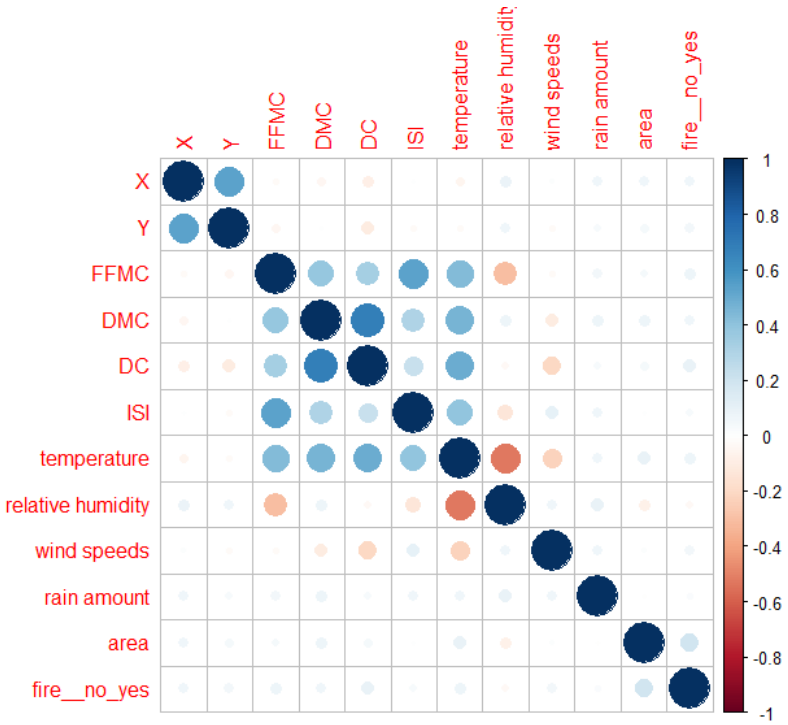
The data is read into R as an excel file, containing 13 different continuous and categorical variables and a total of 517 observations spanning 12 months. In addition to the attributes mentioned above the data includes the month, day of the week, X longitude and Y latitude coordinates of Montesinho Natural Park at the time of a fire. Within each variable no null values which results in the data being fairly clean. At this time, additional preprocessing steps were taken to determine the best format of the data for all model requirements.

The first observation is of the *area* variable containing many zero values. Upon further research, the *zero* value represents the amount of area burned in the park being less than 1 hectare. It should be noted that an additional binary feature is created to represent an "insignificant" fire with a zero. The definition of an insignificant fire is defined as covering less than .1 hectares, and a 1 to indicate a "significant" fire greater than .1 hectares. This variable is used during the classification models.



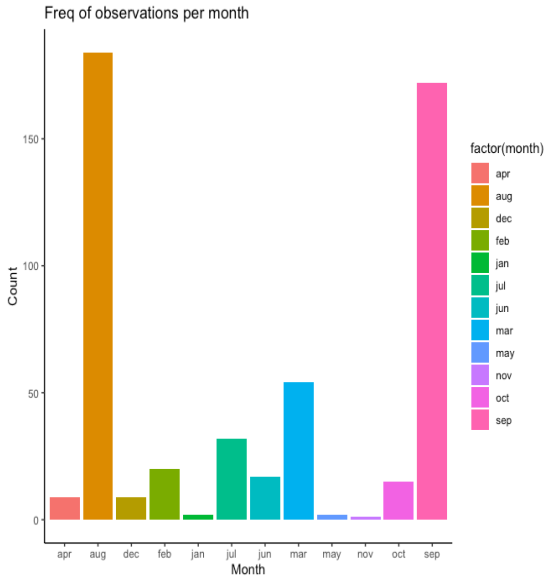
Secondly, the data proposes a challenge as each variable is measured by its individual index. In addition to this the variables FFMC, ISI, rain, and area display a certain amount of skewness. To optimize the performance of the distance based models used the data is normalized by utilizing the min-max scale of the values. Four datasets were created, one is the original dataset with the acquired significant feature, a separate dataset for the scaled data, the last two datasets are created by doing a 67:33 training test split.

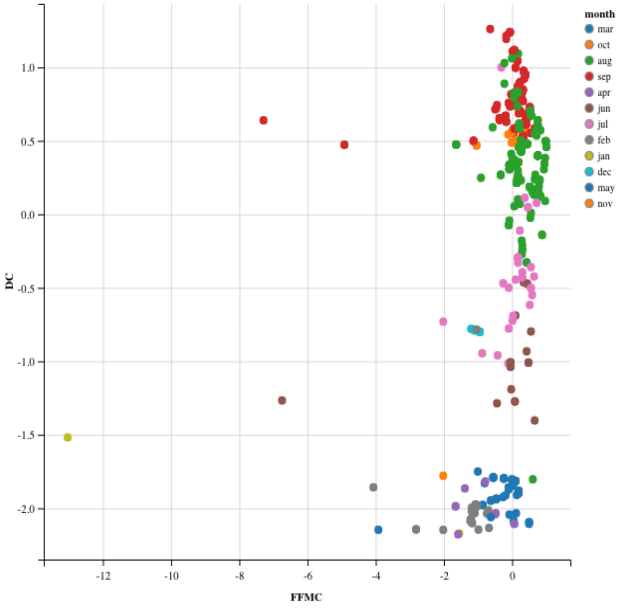
EDA



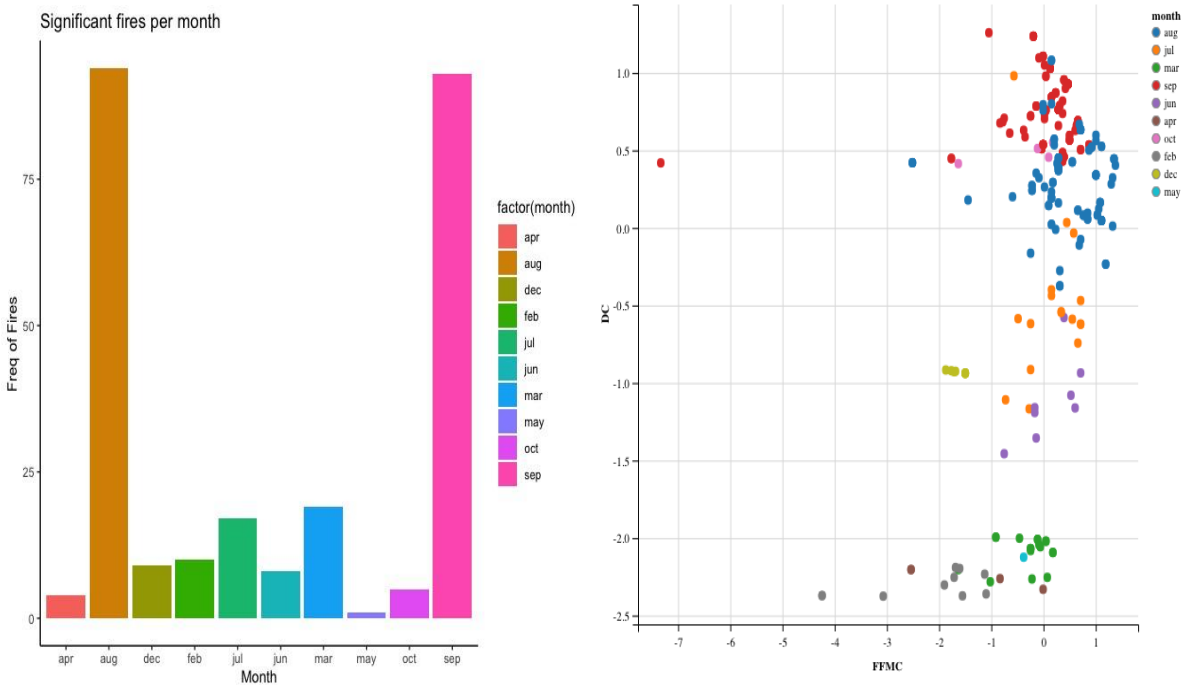
No strong correlations between variables and if there was a fire that day or how much of the area was burned

Scaled DC & FFMC (All data)





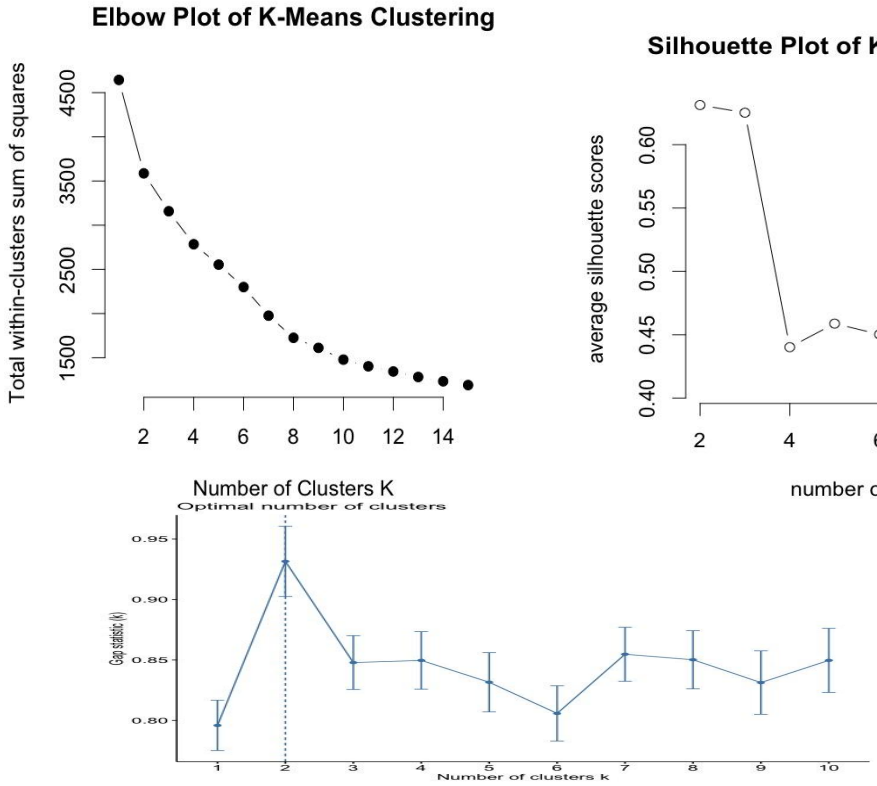
Scaled DC & FFMC (Significant Fires)



The exploratory data analysis revealed that forest fires occur the most in the months of August and September. Furthermore, the drought code and Fine Fuel Moisture Code are above average in these two months for the significant fires occurring. The FFMC and DC are key factors for fire behavior. Last but not least, after February the drought code begins having an exponential growth curve, and the Fine Fuel Moisture Code begins seeing a subtle positive increase as well.

Analysis and Models

K-means Model



To further explore the causes behind the forest fires in Montesinho Natural Park the K-means algorithm is utilized. The euclidean distance is the measure used in the k-means algorithm. The goal for the k-means clustering is to find the homogeneous within clusters and heterogeneous across clusters. In other words, the key operation is the computation of the distance between two cluster centroids. Resulting in the smallest sum of squares being chosen as the most optimal number of clusters.

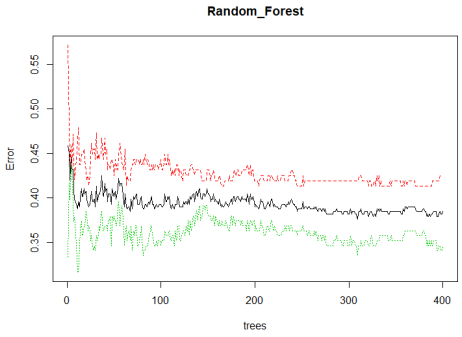
In this exercise, the application of the k-means algorithm is being used to narrow down the transitional points in time when environmental conditions foster more frequent significant fires. The variables used form the dataset for the k-means algorithm are the drought code and the fine fuel moisture code. The drought code is selected because this represents the drying of deep soil and the fine fuel moisture code is chosen due to this factor determining the ease of ignition and flammability. The dataset used is only containing the significant fires with scaled values. The results will show k 2 through 4.

Decision Trees Model

Three different decision trees were generated to investigate which of the attributes would best predict a significant burn area. The trees were varied according to the set.seed() parameter for reproducibility, and utilized cross-validation. Information Gain was also performed to assess the importance of each variable.

Random Forest Model

The random forest model was used for prediction. The number of trees used was 400. The number of splits the model was allowed to try was 6.



Logistic Regression Model

The logit logistics model was used to predict significant fires using all of the available variables and Checking for statistical significance. A statistical significance level of .095 was used for determining what dependent variables to use.

Support Vector Machine

All four types (Linear, Polynomial, Radial, Sigmoid) of support vector machine kernels were used and compared. The models were ran and compared by accuracy and error type percentage.

Naive Bayes Model

Naive Bayes is a Supervised Machine Learning algorithm based on the Bayes Theorem that is used to solve classification problems by following a probabilistic approach. It is based on the idea that the predictor variables in a machine learning model are independent of each other. Meaning that the outcome of a model depends on a set of independent variables that have nothing to do with each other (which is what makes the model naïve because there will always be some correlations between them).

Our goal with Naïve Bayes is to predict whether there was a fire or not based on the other predicting variables (list variables). With the data prepared and spliced, or split into test and training sets, the preparation is done for analysis. The outcome is a class variable of 0 or 1.

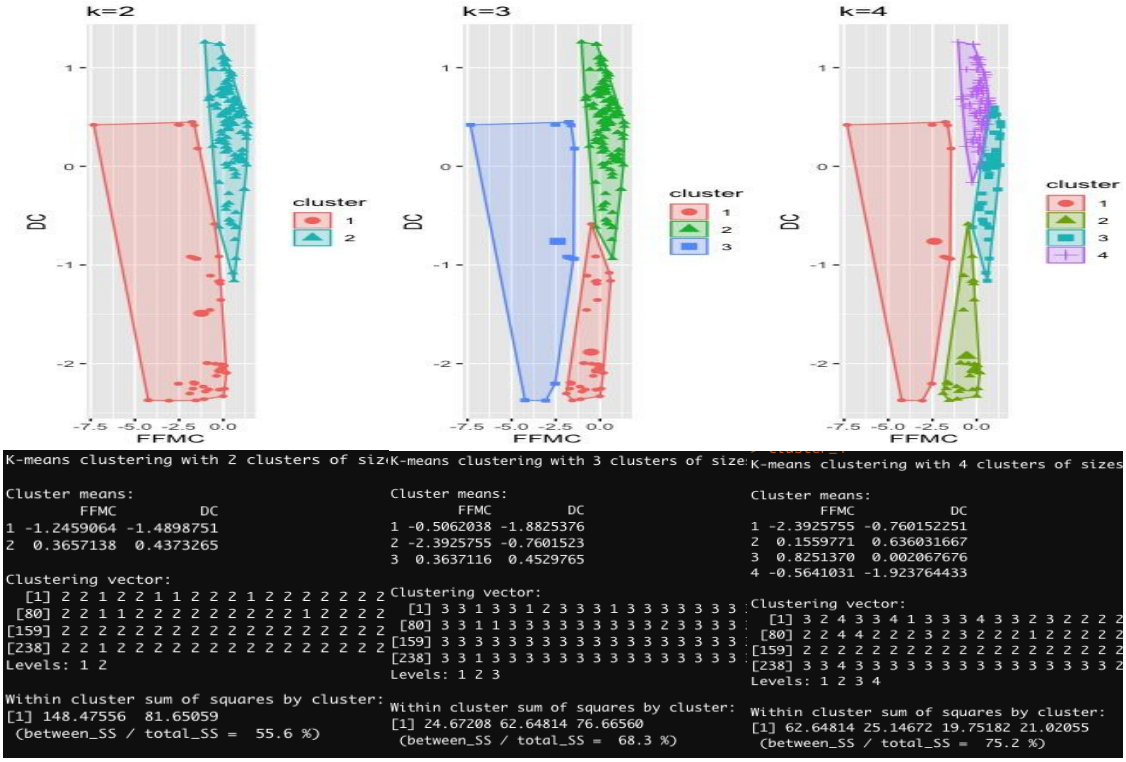
```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 18 18
         1 62 73

               Accuracy : 0.5322
                 95% CI : (0.4545, 0.6087)
    No Information Rate : 0.5322
    P-Value [Acc > NIR] : 0.5312

                  Kappa : 0.0281
 Mcnemar's Test P-Value : 1.528e-06

            Sensitivity : 0.2250
            Specificity : 0.8022
         Pos Pred Value : 0.5000
         Neg Pred Value : 0.5407
             Prevalence : 0.4678
         Detection Rate : 0.1053
   Detection Prevalence : 0.2105
      Balanced Accuracy : 0.5136

       'Positive' Class : 0
```

The final output showed that the Naïve Bayes classifier could predict whether a fire would occur or not with an accuracy of approximately 51.2%. To summarize, the following plot shows how each predictor variable is independently responsible for predicting the outcome.

The figure above is a variable importance evaluation. This evaluation used the model information to see which variables were more closely tied to the model's performance. This evaluation shows the correlation structure between the predictors and the target variable. From the above illustration, it is clear that 'Temperature' is the most significant variable for predicting the outcome for the naive bayes model.

Results

K-means results

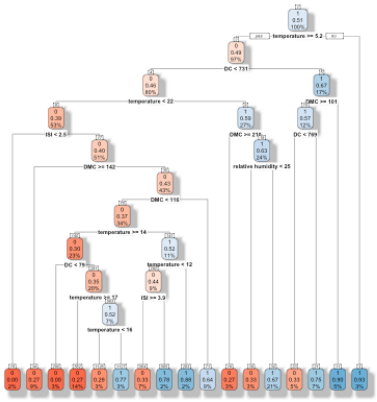


The k-means clustering results suggest that there are well defined points in time where an increase in the frequency of fires are prone to occur more often in provisional situations. The

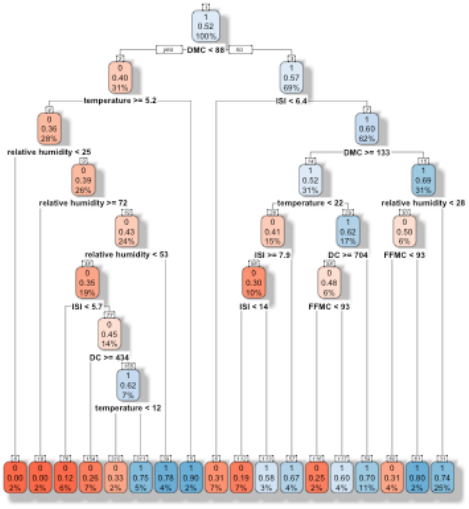
conditions captured in cluster 2 show a group of significant fires with a below average FFMC index and a below average DC index. It is to be noted there are a few outliers and if removed the clusters will likely produce less cloudy results.

Subsequently, the results from cluster three suggest there is a clear transition occurring through cluster one to three. Cluster 1 has the medium amount of results, Cluster 2 has the most frequent occurrences, and Cluster 3 seems to be anomalies which can be important as these can be held as suggestions for other factors being the cause of significant fires occurring like arson or other natural phenoms. Similarly, in the case of cluster 4 a clear departure from one cluster to another is made with the frequency of significant fires occurring in cluster 4, following that is cluster 3. By referring to the plot figured displaying the Month a comparison can be drawn by referencing the correlations between the months.

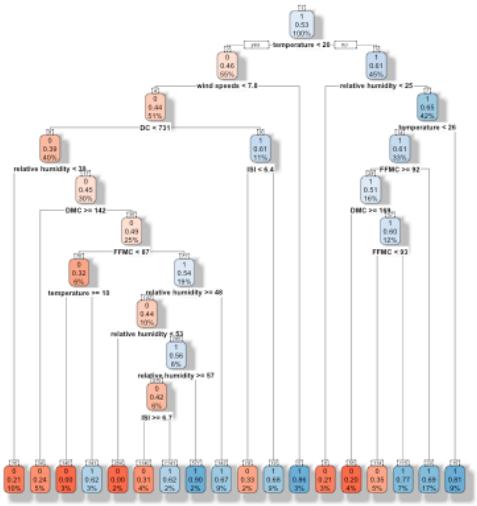
Decision tree results



set.seed(123) had an accuracy of 51% and reflected a 21% chance of a significant burn area when the Relative Humidity less than 25. Information gain showed that Temperature was the most important variable. This tree first split at Temperature, then branched into DC, then Temperature and DMC.



set.seed(1016) had an accuracy of 52% with a 25% chance of a significant burn area when the Relative Humidity was less than 28. Information Gain showed that DMC was the most important variable. This tree first split at DMC, then branched into Temperature and ISI.



set.seed(69) reflected a 17% chance of a significant burn area when the FFMC was greater than or equal to 92. Accuracy was measured at 53%. This tree first split at temperature, then branched into wind speeds and relative humidity. Information gain showed that Temperature was the most important variable. This tree first split at temperature, then branched into wind speeds and relative humidity.

Random Forest results

	Actual	Confusion Matrix	
Predicted		0	1
	0	42	39
	1	38	52
		Accuracy:	55%
		False Pos	51%
		False Neg	49%

The random forest results were a 55% accuracy. The false negative error is the lowest of all models at 49%.

Logistic Regression results

The following results are from the logistics regression. None of the variables come close to being significant. DC comes in closest, but is still not significant.

```
Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)        -3.414327   3.291577   -1.04     0.30
FFMC                0.034135   0.036531    0.93     0.35
DMC                 0.000307   0.002534    0.12     0.90
DC                  0.000758   0.000665    1.14     0.25
ISI                -0.011390   0.029489   -0.39     0.70
temperature         0.001910   0.029535    0.06     0.95
`relative humidity` -0.006333   0.009470   -0.67     0.50
`wind speeds`       0.066926   0.066740    1.00     0.32
`rain amount`       0.152089   0.369893    0.41     0.68

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 479.24  on 345  degrees of freedom
Residual deviance: 470.17  on 337  degrees of freedom
AIC: 488.2
```

Support Vector Machine Results

Radial : Below(A) are the results to the Radial kernel SVM model. This model has a 51% accuracy and has a false negative prediction percentage of 55%. This model would be the best kernel to use. Although it does not have the most accurate predictions it has significantly less false negative predictions.

Linear : Below(B) are the results to the Linear kernel SVM model. This model has a 54% accuracy and has a false negative prediction percentage of 71%.

	Actual	Confusion Matrix	
Predicted		0	1
	0	34	38
	1	46	53
		Accuracy:	51%
		False Pos	45%
		False Neg	55%

A

	Actual	Confusion Matrix	
Predicted		0	1
	0	21	22
	1	54	69
		Accuracy:	54%
		False Pos	29%
		False Neg	71%

B

Polynomial : Below(C) are the results to the Polynomial kernel SVM model. This model has a 60% accuracy and has a false negative prediction percentage of 86%.

Sigmoid : Below(D) are the results to the Sigmoid kernel SVM model. This model has a 50% accuracy and has a false negative prediction percentage of 67%.

	Actual	Confusion Matrix	
Predicted		0	1
	0	21	10
	1	59	81
		Accuracy:	60%
		False Pos	14%
		False Neg	86%

C

	Actual	Confusion Matrix	
Predicted		0	1
	0	23	28
	1	57	63
		Accuracy:	50%
		False Pos	33%
		False Neg	67%

D

Conclusion

Based upon the results obtained through the multiple models, predicting which fires would be significant is not entirely accurate with these techniques and data. Random forest had the strongest predictions out of the models tested. There is not one individual model that has outstanding results; however, a few learnings did take place. One learning is that seasonality plays an effect on predictability of significant fires.

Natural phenomenons could occur as outliers in the dataset. Other factors and data which could provide additional insight into forest fire prediction and strengthen the analysis would be data including lightning detection and when populations of campers were increased, leading to the possibility of accidental ignition from camp fires.

In conclusion to better predict, lessen the damage and control the significance of fires going forward better data collection is required. It is suggested to collect points in times when the park is the most populated because having this insight can create better parameters for future research.

Citations:

P. Cortez and A. Morais. A Data Mining Approach to Predict Forest Fires using Meteorological Data.In J. Neves, M. F. Santos and J. Machado Eds., New Trends in Artificial Intelligence,Proceedings of the 13th EPIA 2007 - Portuguese Conference on Artificial Intelligence, December,Guimaraes, Portugal, pp. 512-523, 2007. APPIA, ISBN-13 978-989-95618-0-9.Available at: http://www3.dsi.uminho.pt/pcortez/fires.pdf