Predicting Forest Fires

Team Fire Squad: Tiffanie M. Mac Donald, Jonathan David Ortiz, Sanjeet Saikia, Jackson F. Shands, Raya Young

Introduction

- Natural (wild) forest fires have a great impact on the natural ecosystem and survival of our species.
- Control of these occurrences are vitally important to the health of the planet.
- Montesinho Natural Park is a naturally forested area that is susceptible to forest fires, has been a focus of study as it is also a popular area for tourists to visit.
- Climate change introduces increased frequency of severe weather events which includes increased risk of forest fires.

Motivation

As phenomenon such as global warming becomes more prevalent, weather event data can be used to make predictions and create more actionable insights to:

- Create solutions to lessen damage
- Prepare for seasonal impact
- Reduce number of those affected, and reduce costs of recovery and restructuring

Problem Statement

A forest fire data set sourced from the UCI Machine Learning Repository, utilizing a combination of meteorological data and spatial coordinates were used to provide insight about forest fires in Montesinho Park, located in the northeast region of Portugal.

Analysis of this dataset will use data mining techniques, such as classification methods including Decision Tree Analysis, Naive Bayes, Support Vector Machines and Random Forest to identify the combination of values that contribute to the likelihood of a significant forest fire.

About the data

X	Y	month	day	FFMC	DMC	DC	ISI	temp	RH	wind	rain	area
7	5	mar	fri	86.2	26.2	94.3	5.1	8.2	51	6.7	0	0
7	4	oct	tue	90.6	35.4	669.1	6.7	18	33	0.9	0	0
7	4	oct	sat	90.6	43.7	686.9	6.7	14.6	33	1.3	0	0
8	6	mar	fri	91.7	33.3	77.5	9	8.3	97	4	0.2	0
8	6	mar	sun	89.3	51.3	102.2	9.6	11.4	99	1.8	0	0
8	6	aug	sun	92.3	85.3	488	14.7	22.2	29	5.4	0	0
8	6	aug	mon	92.3	88.9	495.6	8.5	24.1	27	3.1	0	0
8	6	aug	mon	91.5	145.4	608.2	10.7	8	86	2.2	0	0
8	6	sep	tue	91	129.5	692.6	7	13.1	63	5.4	0	0
7	5	sep	sat	92.5	88	698.6	7.1	22.8	40	4	0	0
7	5	sep	sat	92.5	88	698.6	7.1	17.8	51	7.2	0	0
7	5	sep	sat	92.8	73.2	713	22.6	19.3	38	4	0	0
6	5	aug	fri	63.5	70.8	665.3	0.8	17	72	6.7	0	0
6	5	sep	mon	90.9	126.5	686.5	7	21.3	42	2.2	0	0
6	5	sep	wed	92.9	133.3	699.6	9.2	26.4	21	4.5	0	0
6	5	sep	fri	93.3	141.2	713.9	13.9	22.9	44	5.4	0	0
5	5	mar	sat	91.7	35.8	80.8	7.8	15.1	27	5.4	0	0
8	5	oct	mon	84.9	32.8	664.2	3	16.7	47	4.9	0	0
6	4	mar	wed	89.2	27.9	70.8	6.3	15.9	35	4	0	0
6	4	apr	sat	86.3	27.4	97.1	5.1	9.3	44	4.5	0	0
6	4	sep	tue	91	129.5	692.6	7	18.3	40	2.7	0	0
5	4	sep	mon	91.8	78.5	724.3	9.2	19.1	38	2.7	0	0
7	4	jun	sun	94.3	96.3	200	56.1	21	44	4.5	0	0
7	4	aug	sat	90.2	110.9	537.4	6.2	19.5	43	5.8	0	0
7	4	aug	sat	93.5	139.4	594.2	20.3	23.7	32	5.8	0	0
7	4	aug	sun	91.4	142.4	601.4	10.6	16.3	60	5.4	0	0
7	4	sep	fri	92.4	117.9	668	12.2	19	34	5.8	0	0
7	4	sep	mon	90.9	126.5	686.5	7	19.4	48	1.3	0	0
6	3	sep	sat	93.4	145.4	721.4	8.1	30.2	24	2.7	0	0
6	3	sep	sun	93.5	149.3	728.6	8.1	22.8	39	3.6	0	0
6	3	sep	fri	94.3	85.1	692.3	15.9	25.4	24	3.6	0	0
6	3	sep	mon	88.6	91.8	709.9	7.1	11.2	78	7.6	0	0
6	3	sep	fri	88.6	69.7	706.8	5.8	20.6	37	1.8	0	0
6	3	sep	sun	91.7	75.6	718.3	7.8	17.7	39	3.6	0	0
6	3	sep	mon	91.8	78.5	724.3	9.2	21.2	32	2.7	0	0

Conditional codes

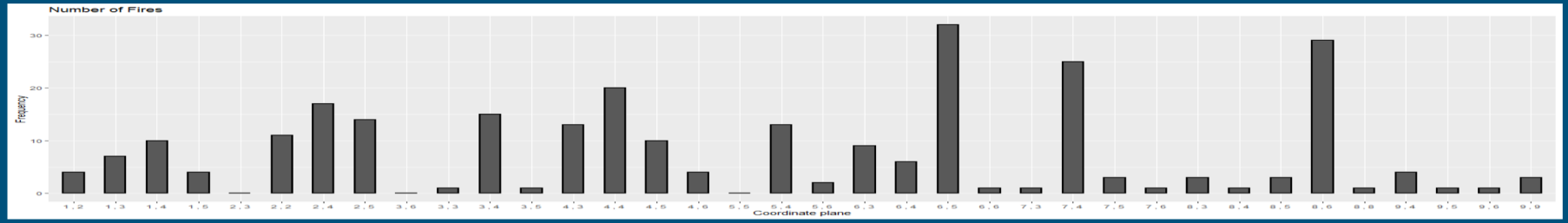
- FFMC - Fine Fuel Moisture Code
- DMC - Duff Moisture Code
- DC - Drought Code

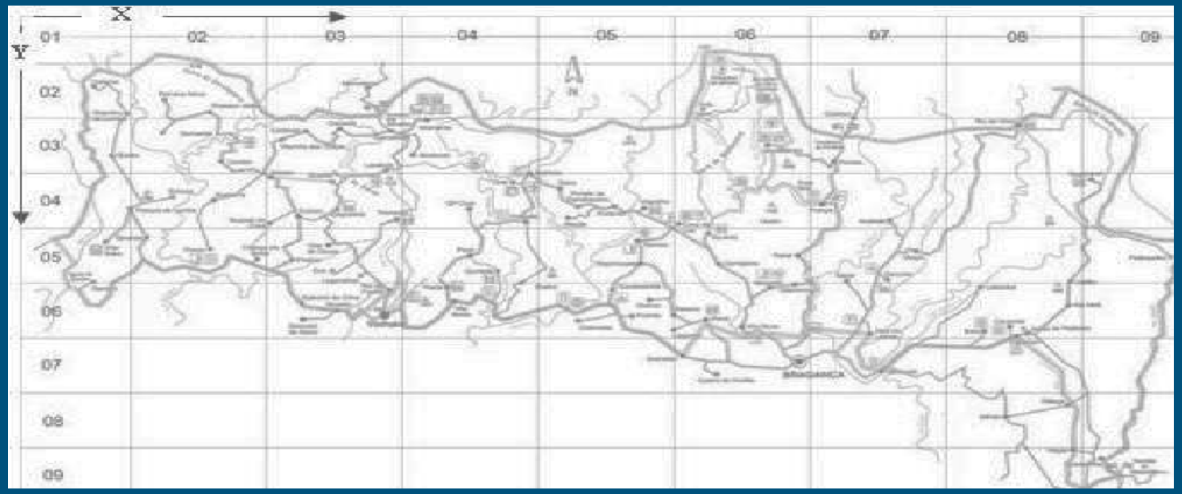
Fire Behavior Indices

- ISI - Initial Spread Index
- RH - Relative Humidity
- AREA - Physical forest area burned

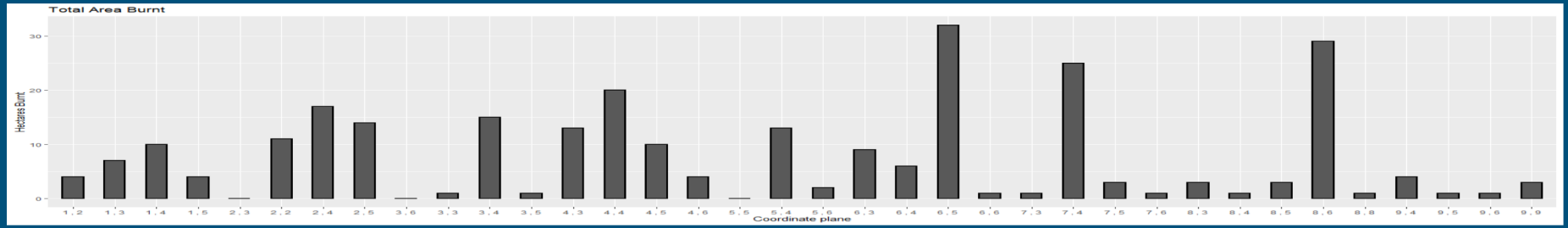
- Other variables:
  - Temp
  - Wind
  - Rain

If an area burned was less than 1ha/100 = 100m2, the dataset rounded down to 0.

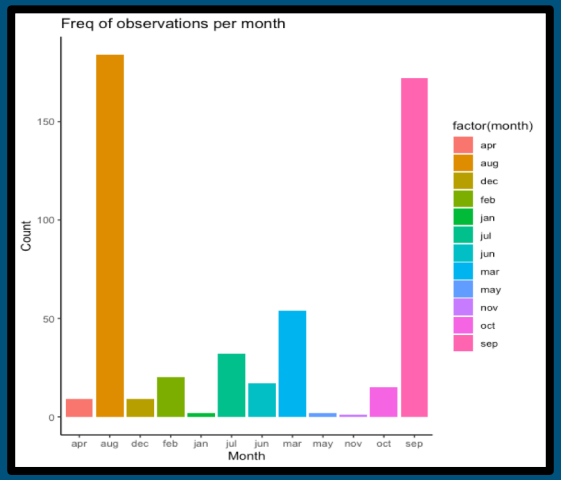




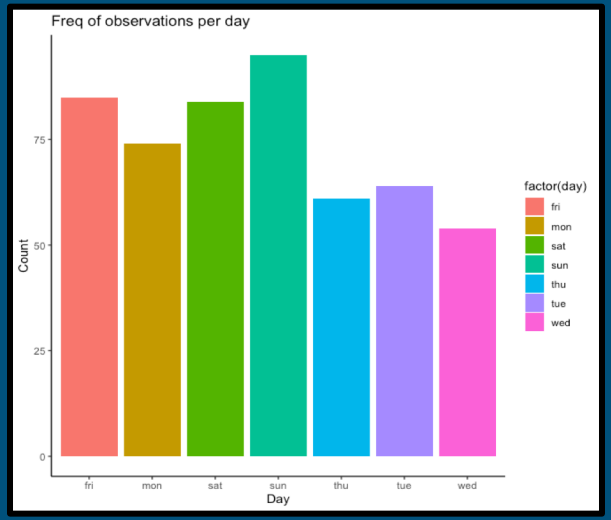
Most burned area and most fires are (“6,5”, “8,6”, “7,4”,”2,4”,”4,4”)
Average area burned : 185 Hectares
Average number of fires : 8 fires over .5 Hectares

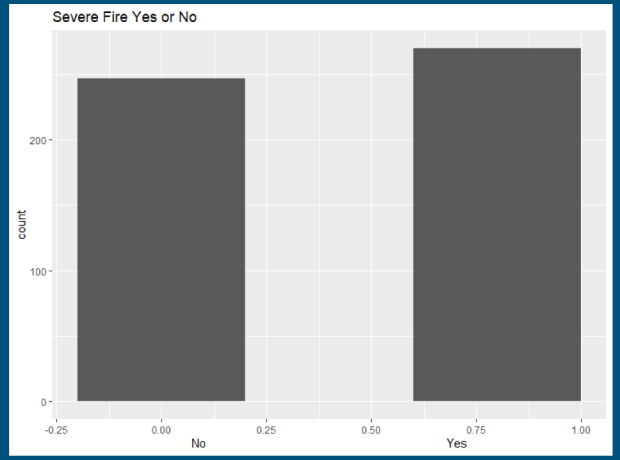


Exploratory Data Analysis



This plot shows the distribution of fires by day of the week. However, information was only recorded when there was a burn, and did not indicate dates or chronology.



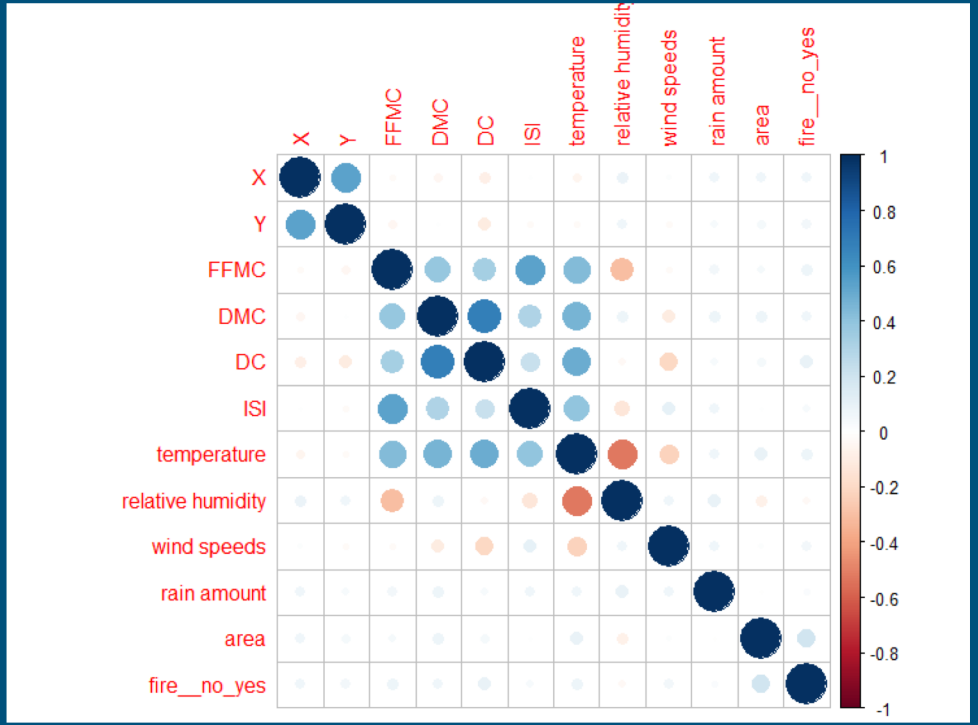


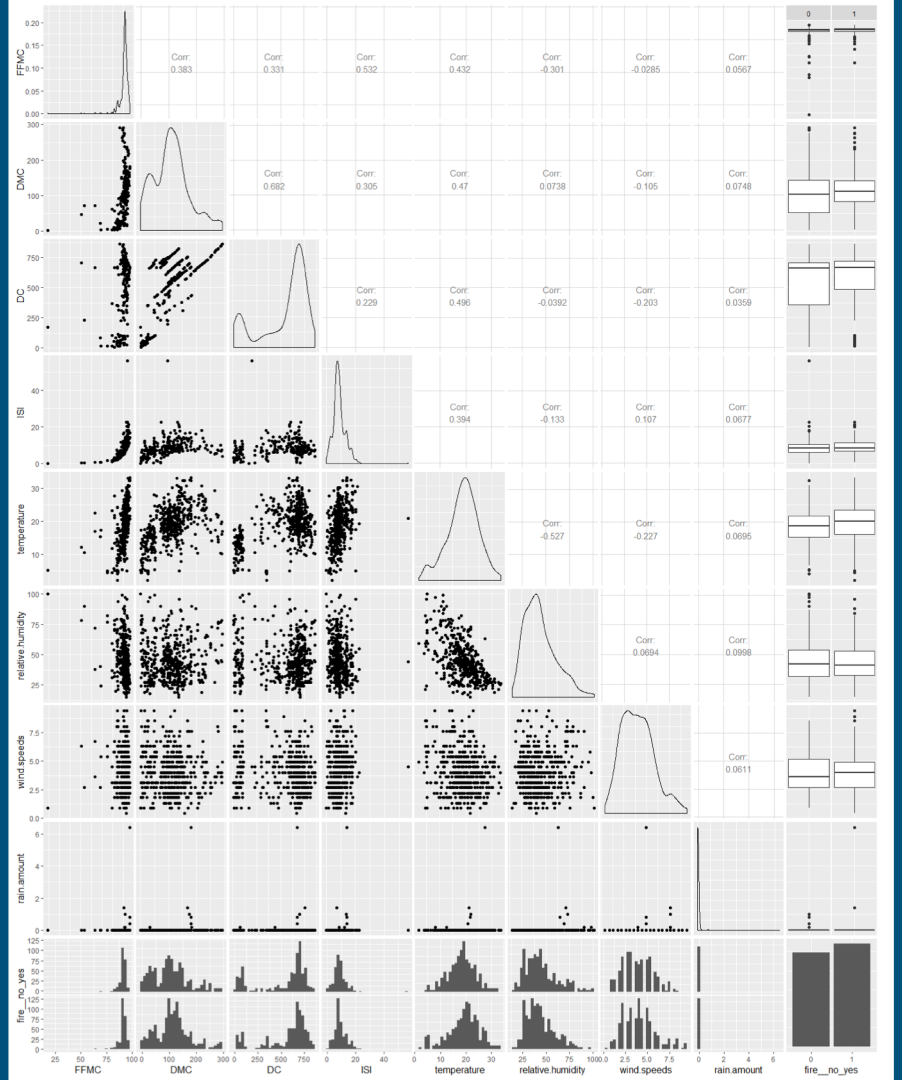
Most significant fires occurred in the months of August and September

Burn area was transformed into a binary variable, indicating if there was a significant fire or not. This plot shows a nearly even distribution of this variable, with 248 ‘No’ and 269 ‘Yes’

Correlation

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



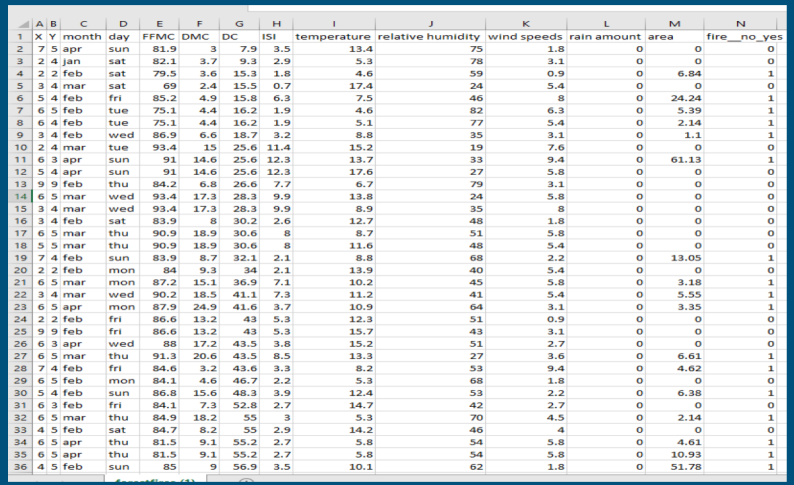

FFMC
DMC
DC
ISI
temperature
relative.humidity
wind.speeds
rain.amount
fire__no_yes
0
1
Corr: 0.383
Corr: 0.331
Corr: 0.532
Corr: 0.432
Corr: -0.301
Corr: -0.0285
Corr: 0.0567
Corr: 0.682
Corr: 0.305
Corr: 0.47
Corr: 0.0738
Corr: -0.105
Corr: 0.0748
Corr: 0.229
Corr: 0.496
Corr: -0.0392
Corr: -0.203
Corr: 0.0359
Corr: 0.394
Corr: -0.133
Corr: 0.107
Corr: 0.0677
Corr: -0.527
Corr: -0.227
Corr: 0.0695
Corr: 0.0694
Corr: 0.0998
Corr: 0.0611

Preparing the data

Cleaning + prep

- Geographic data
- Dates
- Special Indexes

Transformed Area burned into a binary variable and removed area burned



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	X	Y	month	day	FFMC	DMC	DC	ISI	temperature	relative humidity	wind speeds	rain amount	area	fire__no_yes
2	7	5	apr	sun	81.9	3	7.9	3.5	13.4	75	1.8	0	0	0
3	2	4	jan	sat	82.1	3.7	9.3	2.9	5.3	78	3.1	0	0	0
4	2	2	feb	sat	79.5	3.6	15.3	1.8	4.6	59	0.9	0	6.84	1
5	3	4	mar	sat	69	2.4	15.5	0.7	17.4	24	5.4	0	0	0
6	5	4	feb	fri	85.2	4.9	15.8	6.3	7.5	46	8	0	24.24	1
7	6	5	feb	tue	75.1	4.4	16.2	1.9	4.6	82	6.3	0	5.39	1
8	6	4	feb	tue	75.1	4.4	16.2	1.9	5.1	77	5.4	0	2.14	1
9	3	4	feb	wed	86.9	6.6	18.7	3.2	8.8	35	3.1	0	1.1	1
10	2	4	mar	tue	93.4	15	25.6	11.4	15.2	19	7.6	0	0	0
11	6	3	apr	sun	91	14.6	25.6	12.3	13.7	33	9.4	0	61.13	1
12	5	4	apr	sun	91	14.6	25.6	12.3	17.6	27	5.8	0	0	0
13	9	9	feb	thu	84.2	6.8	26.6	7.7	6.7	79	3.1	0	0	0
14	6	5	mar	wed	93.4	17.3	28.3	9.9	13.8	24	5.8	0	0	0
15	3	4	mar	wed	93.4	17.3	28.3	9.9	8.9	35	8	0	0	0
16	3	4	feb	sat	83.9	8	30.2	2.6	12.7	48	1.8	0	0	0
17	6	5	mar	thu	90.9	18.9	30.6	8	8.7	51	5.8	0	0	0
18	5	5	mar	thu	90.9	18.9	30.6	8	11.6	48	5.4	0	0	0
19	7	4	feb	sun	83.9	8.7	32.1	2.1	8.8	68	2.2	0	13.05	1
20	2	2	feb	mon	84	9.3	34	2.1	13.9	40	5.4	0	0	0
21	6	5	mar	mon	87.2	15.1	36.9	7.1	10.2	45	5.8	0	3.18	1
22	3	4	mar	wed	90.2	18.5	41.1	7.3	11.2	41	5.4	0	5.55	1
23	6	5	apr	mon	87.9	24.9	41.6	3.7	10.9	64	3.1	0	3.35	1
24	2	2	feb	fri	86.6	13.2	43	5.3	12.3	51	0.9	0	0	0
25	9	9	feb	fri	86.6	13.2	43	5.3	15.7	43	3.1	0	0	0
26	6	3	apr	wed	88	17.2	43.5	3.8	15.2	51	2.7	0	0	0
27	6	5	mar	thu	91.3	20.6	43.5	8.5	13.3	27	3.6	0	6.61	1
28	7	4	feb	fri	84.6	3.2	43.6	3.3	8.2	53	9.4	0	4.62	1
29	6	5	feb	mon	84.1	4.6	46.7	2.2	5.3	68	1.8	0	0	0
30	5	4	feb	sun	86.8	15.6	48.3	3.9	12.4	53	2.2	0	6.38	1
31	6	3	feb	fri	84.1	7.3	52.8	2.7	14.7	42	2.7	0	0	0
32	6	5	mar	thu	84.9	18.2	55	3	5.3	70	4.5	0	2.14	1
33	4	5	feb	sat	84.7	8.2	55	2.9	14.2	46	4	0	0	0
34	6	5	apr	thu	81.5	9.1	55.2	2.7	5.8	54	5.8	0	4.61	1
35	6	5	apr	thu	81.5	9.1	55.2	2.7	5.8	54	5.8	0	10.93	1
36	4	5	feb	sun	85	9	56.9	3.5	10.1	62	1.8	0	51.78	1

*No null values

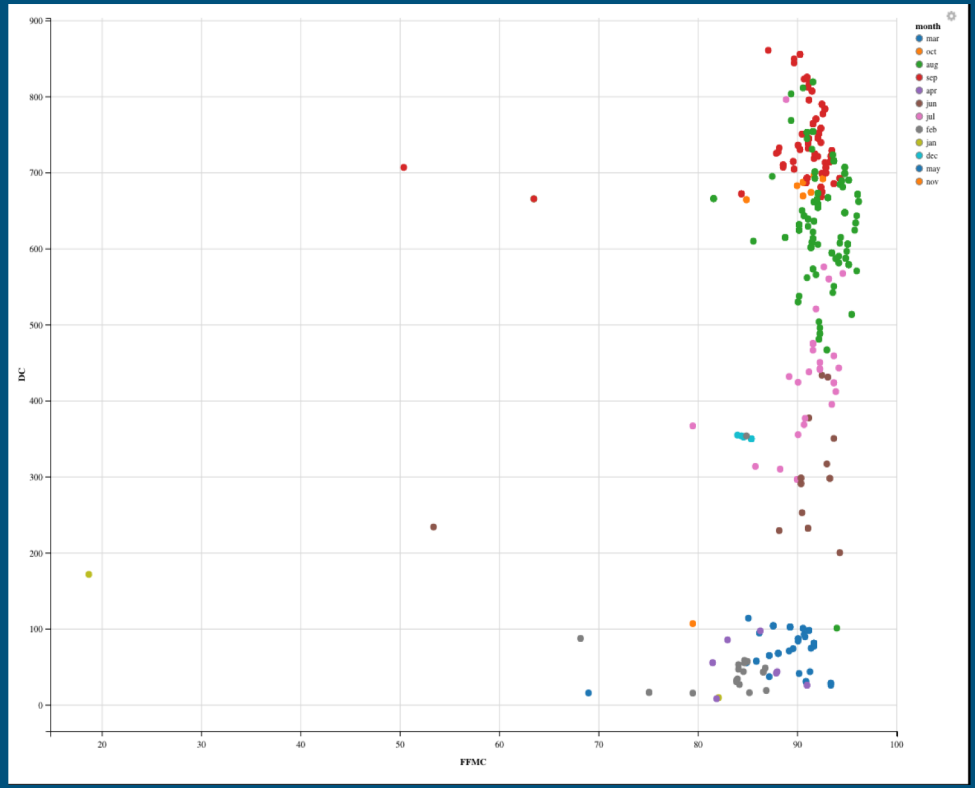
```
Looking at column... X
The num of missing values in column  X is   0

Looking at column... Y
The num of missing values in column  Y is   0

Looking at column... month
The num of missing values in column  month is   0

Looking at column... day
The num of missing values in column  day is   0

Looking at column... FFMC
The num of missing values in column  FFMC is   0

Looking at column... DMC
The num of missing values in column  DMC is   0

Looking at column... DC
The num of missing values in column  DC is   0

Looking at column... ISI
The num of missing values in column  ISI is   0

Looking at column... temp
The num of missing values in column  temp is   0

Looking at column... RH
The num of missing values in column  RH is   0

Looking at column... wind
The num of missing values in column  wind is   0

Looking at column... rain
The num of missing values in column  rain is   0

Looking at column... area
The num of missing values in column  area is   0
```

- Library -> (Dplyr)
- Num & Str converted Factors

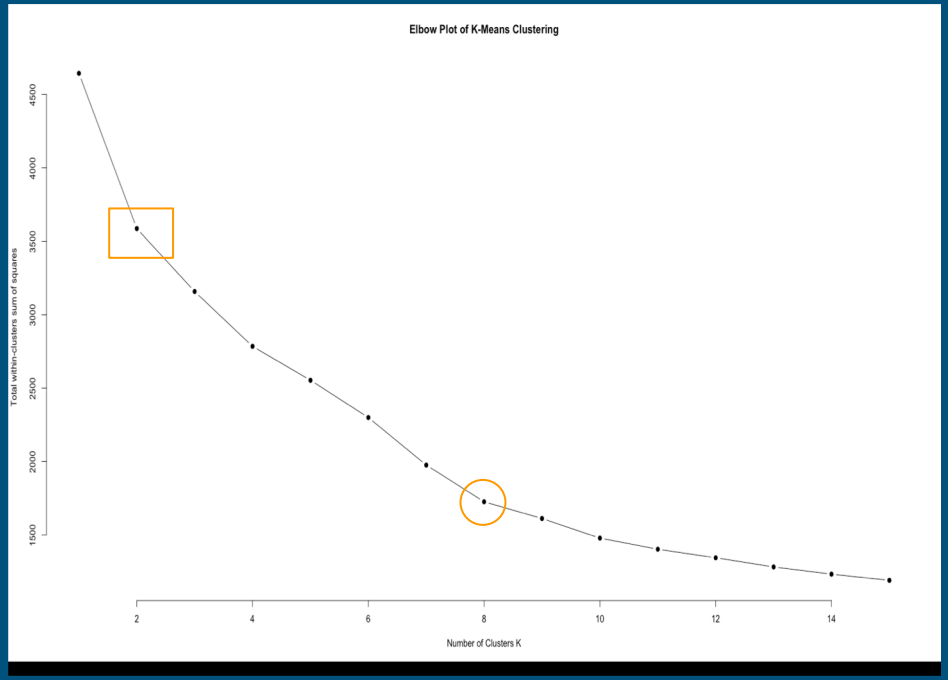
Clustering - kMeans

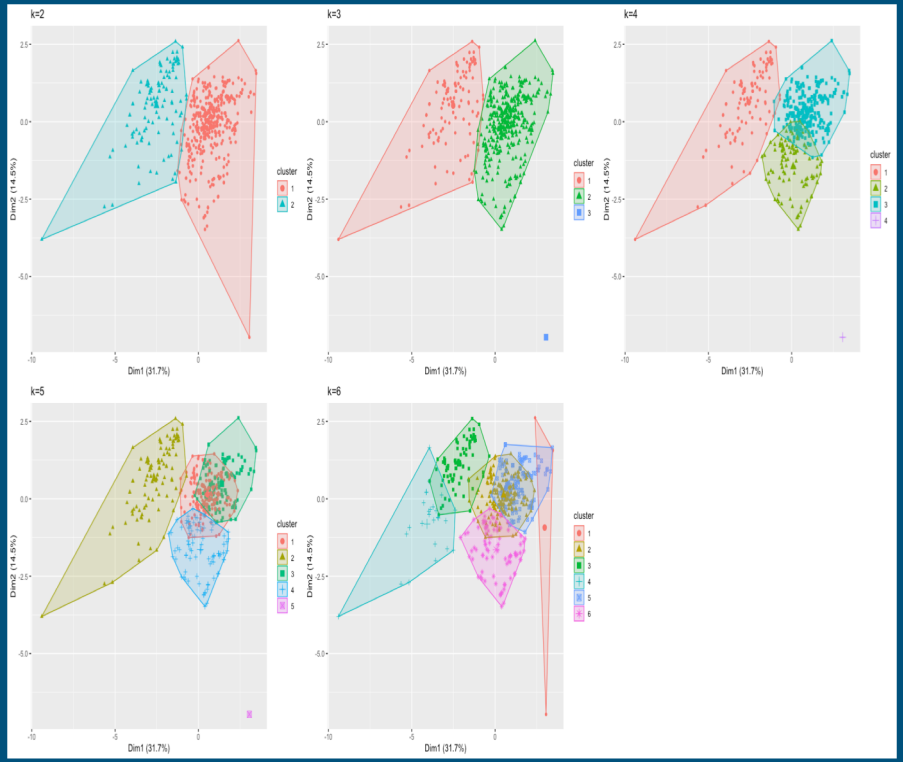
K-Means set up

- Scaled data
  - Dropped (month, day, X,Y)
- Plotted points to find separation of data



Clustering - K means





Cluster 3

```
Within cluster sum of squares by cluster:
[1]  907.1012  467.9488 1782.9638
 (between_SS / total_SS =  32.0 %)
```

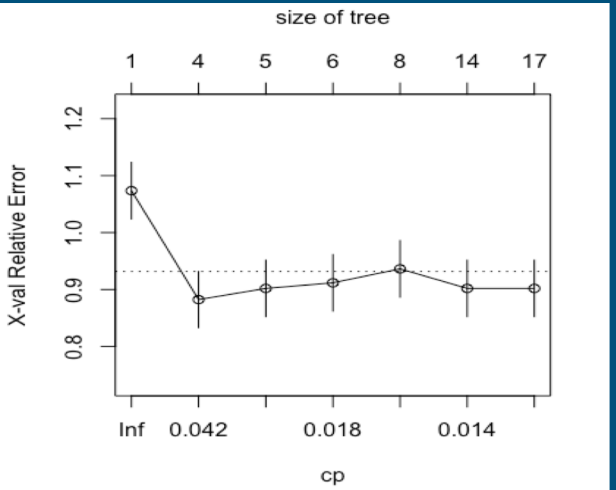
Decision Trees seed(123)

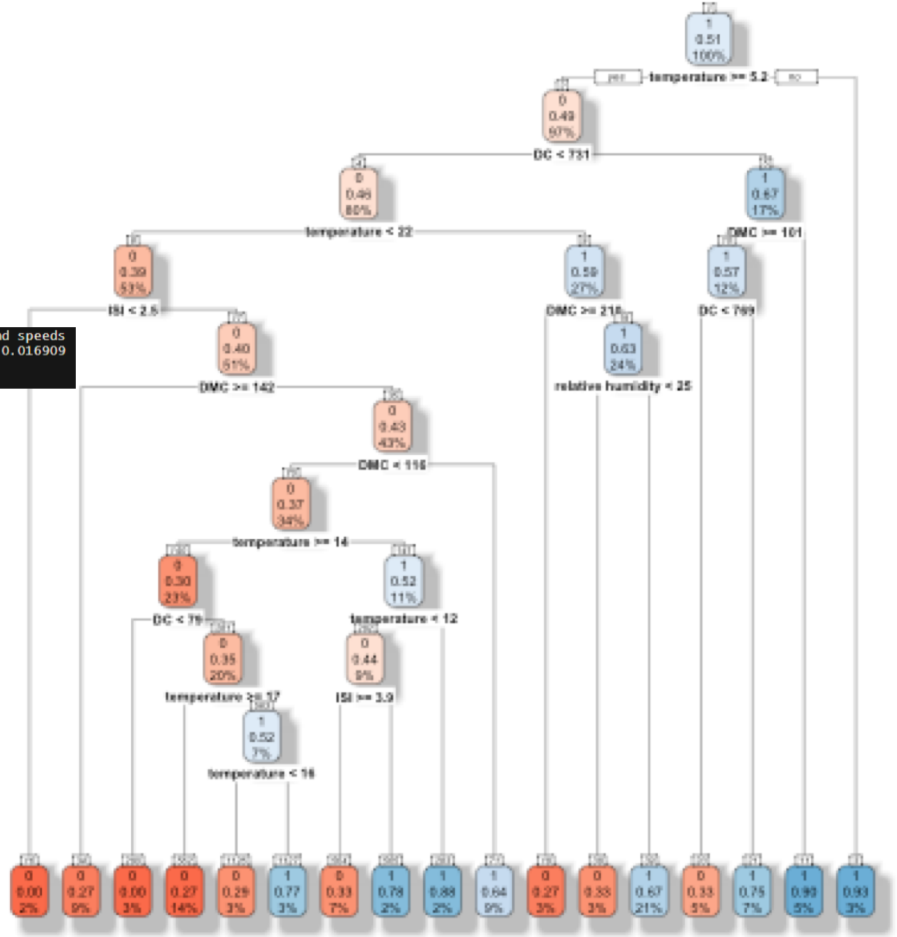
The first tree used a seed of 123 and reflected a 21% chance of a significant burn area when the Relative Humidity less than 25. This tree first split at Temperature, then branched into DC, then Temperature and DMC.

Info Gain :

FFMC	DMC	DC	ISI	temperature	relative humidity	wind speeds	rain amount
0.007438	0.008668	0.014926	0.009935	0.021268	0.008974	0.016909	0.007271

The accuracy for this model 51%.





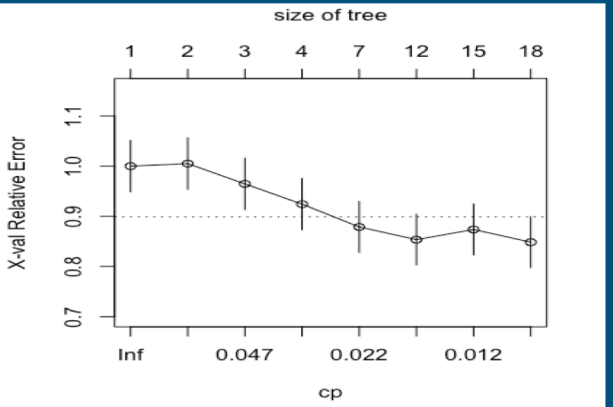
Decision Trees seed(1016)

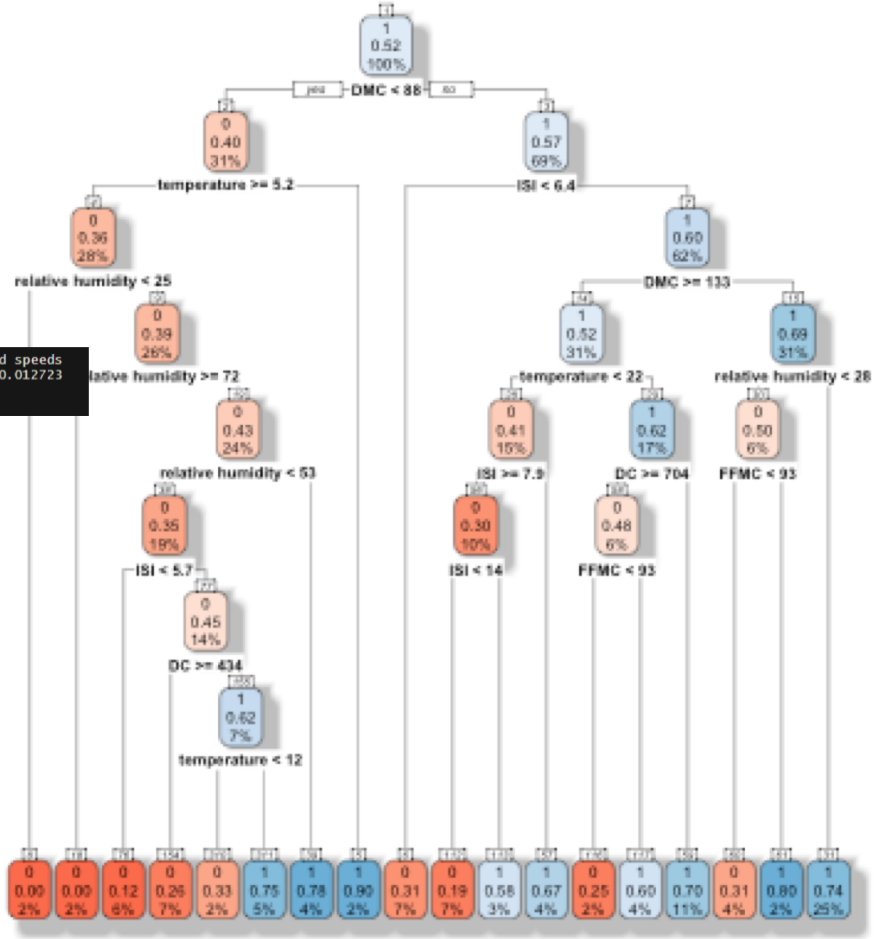
The second model generated used a seed of 1016 and reflected a 25% chance of a significant burn area when the Relative Humidity less than 28. This tree first split at DMC, then branched into Temperature and ISI.

INFO GAIN:

FFMC	DMC	DC	ISI	temperature	relative humidity	wind speeds	rain amount
0.010356	0.018535	0.011259	0.015590	0.016949	0.012125	0.012723	0.005776

The accuracy for this model is 52%





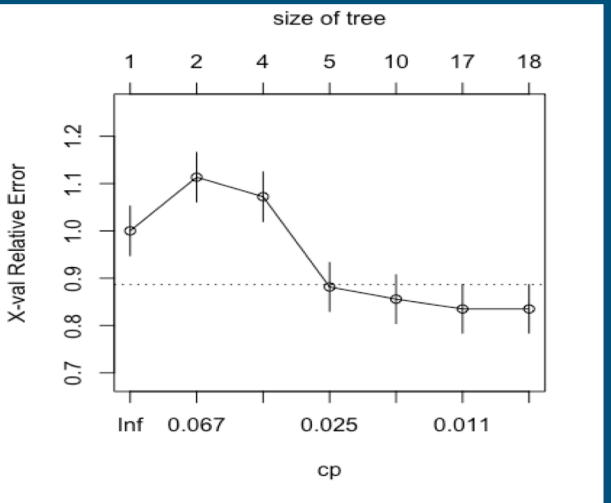
Decision Trees seed(69)

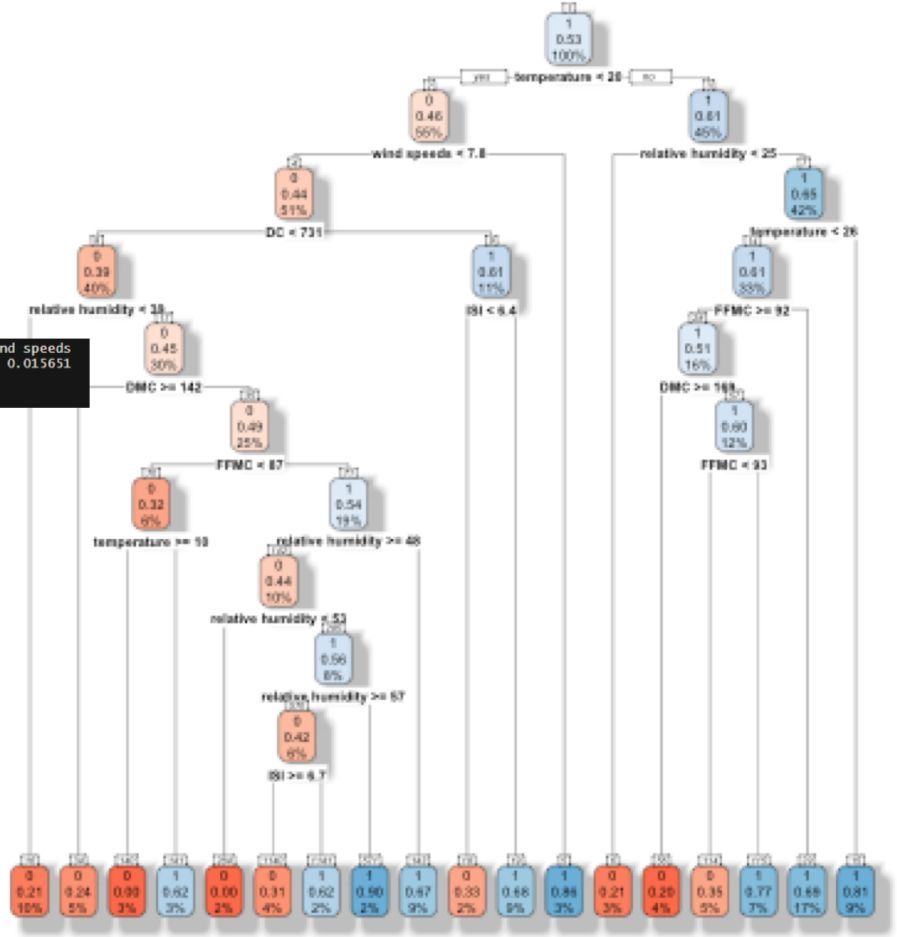
Finally, the third tree used a seed of 69 and reflected a 17% chance of a significant burn area when the FFMC was greater than or equal to 92. This tree first split at temperature, then branched into wind speeds and relative humidity.

INFO GAIN :

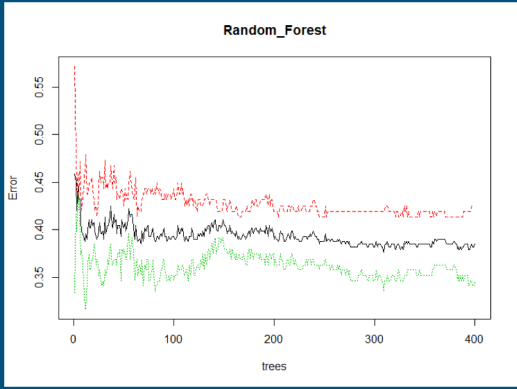
FFMC	DMC	DC	ISI	temperature	relative humidity	wind speeds	rain amount
0.009286	0.012383	0.012410	0.010643	0.016757	0.013328	0.015651	0.003071

The accuracy for this model is 53%.





Random Forest



Number of trees : 400
Accurate : 55%
False Negative : 49 %
Out Of Bag Error : 38%
Number of splits tried : 6

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

Information gain:

FFMC	DMC	DC	ISI	Temp	Rel_Hum	Wind_Speed	Rain_Amt
0.012254	0.026155	0.014761	0.019900	0.022419	0.011685	0.012208	0.006771

Logistic Regression: Logit

```
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
```

	Actual	Confusion Matrix	
Predicted		0	1
	0	32	29
	1	48	62
		Accuracy:	55%
		False Pos	38%
		False Neg	62%

- None of our variables seem to be statistically significant
- Using this in conjunction with our other findings is difficult
    - Variable significance does not coincide with decision tree and naive bayes findings

- Running all possible combinations of our variable the lowest p value for a variable is wind speed at .056 then DC at .12
- With all combinations errors were skewed False Negatives from 62% - 75%

Naïve Bayes

Our goal with Naïve Bayes is to predict whether there was a fire or not based on the other predicting 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.

The final output showed that the Naïve Bayes classifier could predict whether a fire would occur or not with a balanced accuracy of approximately 51.4%.

```
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
```

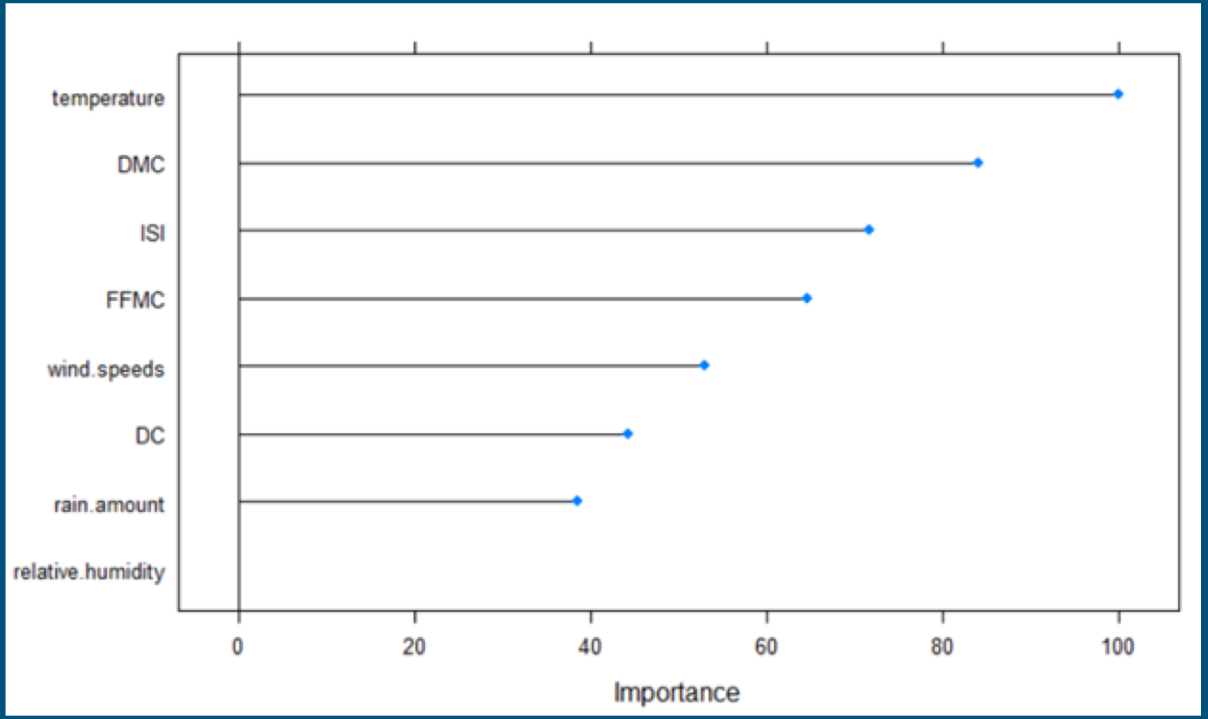
Variable Importance Evaluation

To summarize, the following plot shows how each predictor variable is independently responsible for predicting the outcome.

It is clear that ‘Temperature’ is the most significant variable for predicting the outcome in the Naïve Bayes model.

ROC curve variable importance

	Importance
temperature	100.00
DMC	84.16
ISI	71.70
FFMC	64.72
wind.speeds	53.04
DC	44.30
rain.amount	38.53
relative.humidity	0.00



SVM Models

Radial

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

- We want to avoid false negatives
- Our most accurate model has the highest percentage of false negatives
- Radial SVM although not as accurate, has the lowest false negatives by far and should be used to avoid predicting no fire when there is one

Polynomial

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

Linear

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

Sigmoid

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

Conclusions

- In exploring the data, we found that while some models performed better than others, sometimes underperformance was more due to limitations of the dataset.
    - For example, unable to do time series analysis because days of week and months did not indicate chronological order.
- Overall, the model which performed the best was Random Forest, with an accuracy of 55%, which is still not optimal.
- Finally, from our analysis, we can see that more information is needed, but the detail gleaned from our exploration shows promising direction.
- Other solutions could be determining specific areas of the forest most susceptible to burn, and installing monitoring systems for prevention and recovery such as water canons.

Resources

Citation web link

Dataset web link