Predicting Forest Fire Behavior in Northeast Portugal

Jackson Dugger, Joshua Hess, Trevor Holt, Hamza Elsiesy, Robert Seiler, Guilermo Felce, Alyssa Guerrero

SDS 322E: Elements of Data Science

Introduction

In this report we will explore how meteorological as well as other data can be used to predict the burned area of forest in the northeast region of Portugal. It is important for forest fire data to be analyzed because the findings can be used in the creation of prevention and control tactics to reduce the serious threat that they pose to our forested landscapes. Forest fires are responsible for eliminating vegetation which is a habitat for many animals and people, releasing harmful pollutants like carbon monoxide and nitrogen oxides into the atmosphere contributing to global warming, and greatly harming many minority communities who have a greater vulnerability to these fires. Through our analysis we predict that we will find drought codes to be the single most significant predictor of fire size, fire size not to be dependent on humidity, and drought to be more significant in predicting fire size than the weather conditions at the time of the ignition.

The Dataset

The dataset used for this analysis is a collection of 517 individual fires from the Montesinho Natural Park, a protected nature reserve in the northeast of Portugal, taking place between January 2000, and December 2003. There are 13 variables which can be split into 5 groups: spatial coordinates, time of fire, FWI system variables, meteorological data, and area. For our purposes, we largely ignored X, Y, and day, and focused primarily on the FWI variables, and the meteorological data.

The Fire Weather Index system, (FWI from hereout), is a group of standardized weather readings, collected at solar noon each day (1). The four FWI measurements used in this dataset are FFMC, DMC, DC, and ISI, which can be seen along with their respective

components in Figure 1 below. It is important to denote the differences between Drought Code (DC), and the Duff Moisture Code (DMC), as they represent different time lags of moisture conditions. DMC has a 15 day lag, while DC is a 53 day lag.

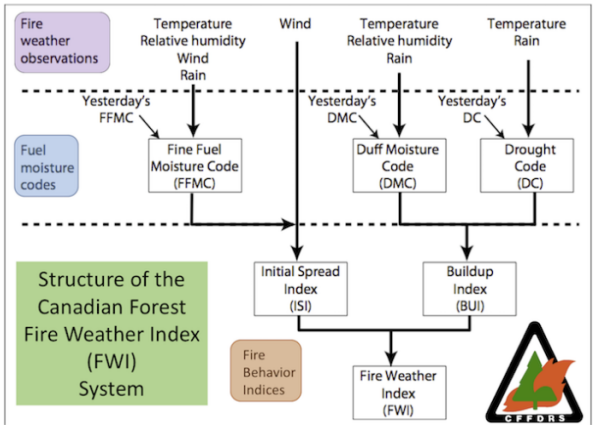


Figure 1

In addition to the FWI variables, the dataset has 4 meteorological measurements: rain, relative humidity (RH), wind, and rain. The rain variable, denotes cumulative precipitation that has fallen 30 minutes prior to fire ignition. The final variable is area, which ranges from 0 hectares, to 1090.84 hectares, acting as our dependent variable. Out of the 517 fires recorded, there were 247 values of “0” in the area column (47.8%), which represent fires smaller than 1/10 hectares, or 100 m^2. In order to increase normalization, and properly visualize the data, we completed a log transformation of the variable “area,” in the form log(1+x) seen in Figure 2 below. In addition to this, the “month” variable was changed to as.factor, so it would be ordered properly in visualizations.

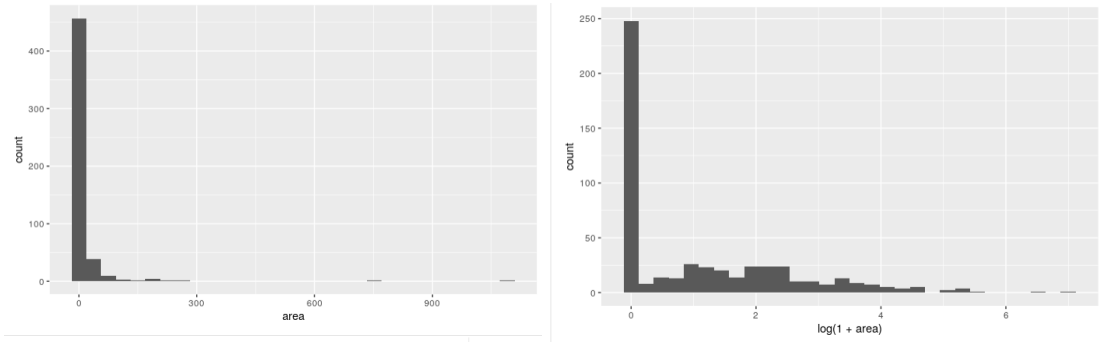


Figure 2:

Scatter plots of the transformation of the area variable

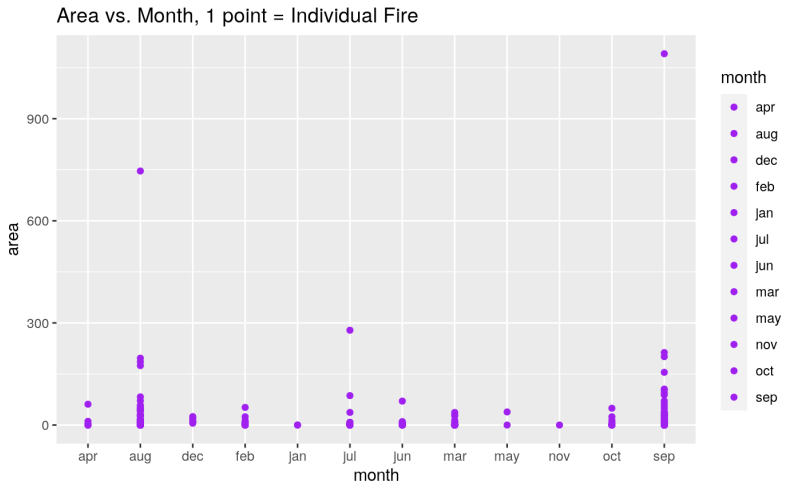


Figure 3

This visualization illustrates how in our dataset, there are a handful of extreme outliers. Most of the data points are concentrated at or around an area of 0, and only 5 are at or above 200 ha. However, one fire burned about 1090 ha, and the next highest burned almost 750.

Exploratory Data Analysis

From this dataset, we were able to construct four hypotheses, and create visualizations with the goal of seeing which variables could be used to predict fire size.

Hypothesis 1

Our first hypothesis is that the dryness of the landscape is a more significant predictor of fire size than the weather conditions at the time of ignition. Portugal's mediterranean climate has a distinct dry season in the summer followed by a cool, wet winter. This hypothesis suggests that the amount of time the landscape has sat in dry and hot conditions prior to ignition is more important than how dry and hot the weather was at the time of ignition. We can see evidence of this by plotting the number of fires by month in a bar chart and comparing it to climate data, seen in Figure 1 below

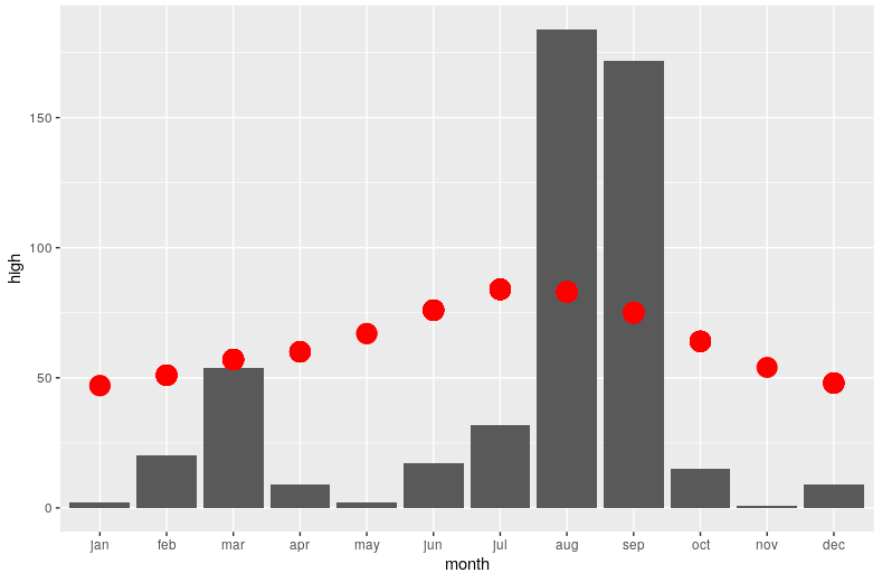
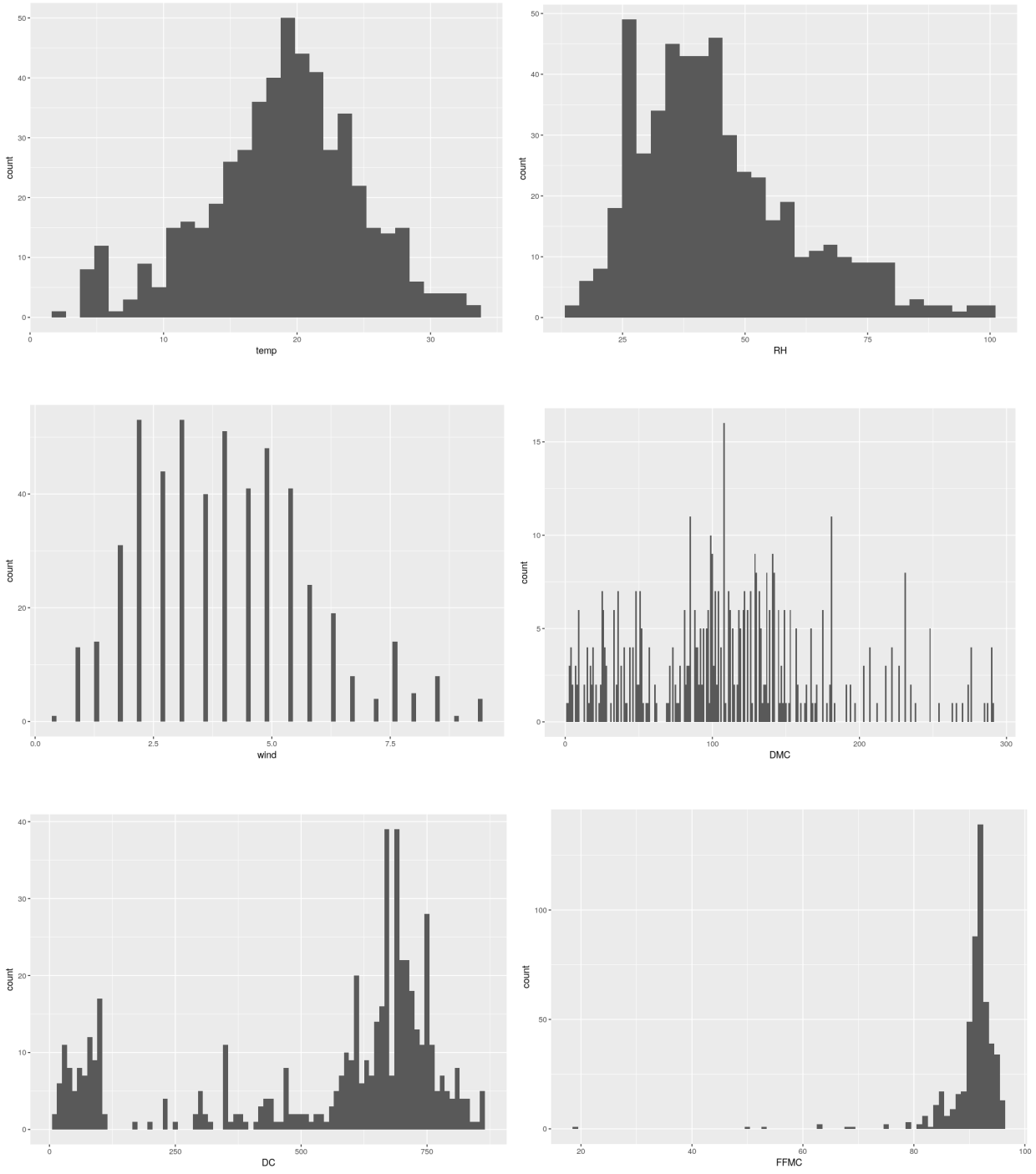


Figure 1

The red dots represent the average daily high temperature by month for this region of Portugal. Average precipitation by month, though not shown, is roughly inverse of temperature. As you can see there is a slight lag between the peak of temperatures and the peak of fire season. September is nearly the worst month for fire activity despite not

being the hottest (nor driest) month. This phenomenon is also observed in parts of the world with similar climates like California.

Under this hypothesis, we would expect drought codes FFMC, DMC, and DC to be the most important predictors of fire size given that they measure the dryness of the landscape. Additional evidence for this hypothesis, albeit perhaps weak, can be seen by plotting histograms of several variables as seen below:



The first row of histograms show temperature and relative humidity. As you can see, the number of fires occurring at different temperatures or humidity levels does not appear to have a significant skew. The second row shows histograms for wind and DMC. Wind

does not appear to be skewed but, surprisingly, neither does DMC. DC and FFMC however, do have a strong rightward skew.

Hypothesis 2

The evidence and analysis above leads us to our second hypothesis, which is that DC is the single most important predictor of fire size. Additional reasoning for this is that DC has the longest memory of the drought codes. It reaches roughly 53 days into the past versus 13 days for DMC and 16 hours for FFMC. A multiple linear regression of several variables on area shows that DMC and DC are the only two statistically significant variables, though the coefficient on DC is counterintuitive.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  18.23602   69.52751   0.262   0.7932
monthfeb     -9.06937   52.98187  -0.171   0.8642
monthmar    -19.27464   53.68310  -0.359   0.7197
monthapr    -13.00041   56.07272  -0.232   0.8168
monthmay      0.26800   69.50733   0.004   0.9969
monthjun     -6.15303   55.71212  -0.110   0.9121
monthjul     15.89475   57.71091   0.275   0.7831
monthaug     31.00046   60.50089   0.512   0.6086
monthsep     56.79042   63.39012   0.896   0.3707
monthoct     56.86782   65.52153   0.868   0.3859
monthnov    -16.41667   81.72341  -0.201   0.8409
monthdec     31.06318   59.98947   0.518   0.6048
FFMC         -0.08733    0.76562  -0.114   0.9092
DMC           0.20620    0.08570   2.406   0.0165 *
DC           -0.12927    0.05790  -2.233   0.0260 *
ISI          -0.60833    0.82004  -0.742   0.4585
temp          1.32511    1.01060   1.311   0.1904
RH           -0.13566    0.27931  -0.486   0.6274
wind          1.67482    1.75000   0.957   0.3390
rain         -1.87936    9.80125  -0.192   0.8480
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 63.85 on 497 degrees of freedom
Multiple R-squared:  0.0309,    Adjusted R-squared:  -0.006148
F-statistic: 0.834 on 19 and 497 DF,  p-value: 0.6663
```

Hypothesis 3

For the third hypothesis, we wanted to determine if the area could be predicted using the weather variables only. These variables include rain, wind, temperature, and relative

humidity. The area variable was transformed by log(area + 1). It was decided that rain would be left out because there were only 2 data points where rain had occurred when the area variable is greater than 0.

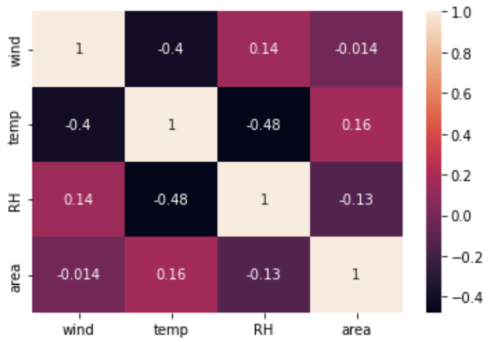


Figure 1

Figure 1 shows a heatmap with the three feature variables chosen and area. Both RH and wind have a negative correlation with area while temperature has a positive correlation. Therefore we hypothesize that as the temperature increases and RH and wind decrease, the area of the fire will increase. To test this, four linear regressions were fitted. The first three are single linear regressions with one of the three feature variables, and the last is a multiple linear regression with all variables.

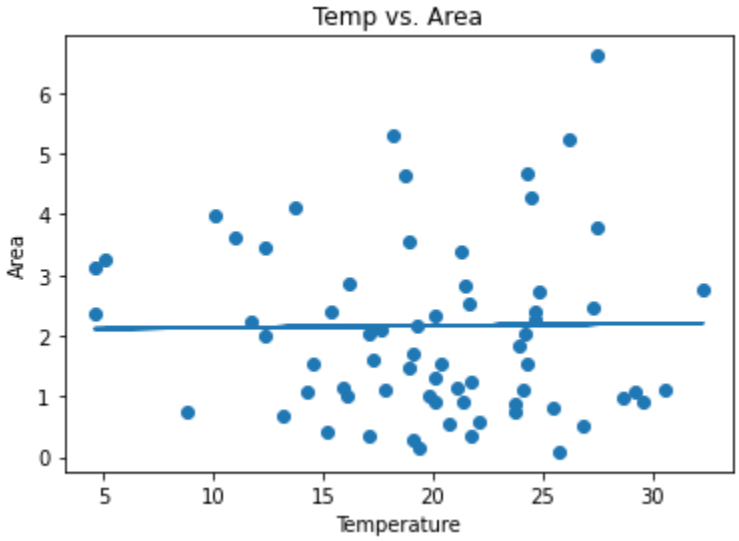


Figure 2

Figure 2 shows the linear regression plot for temperature vs. area. The slope is 0.00335825. A positive slope is to be expected from the heatmap.

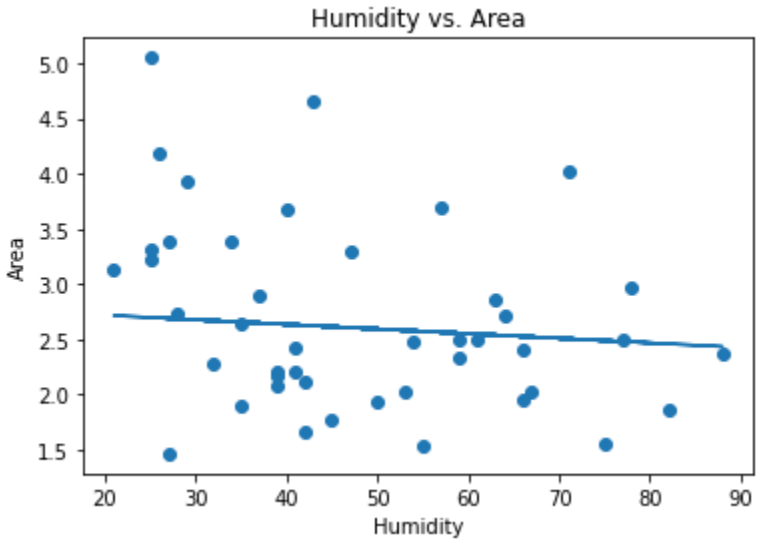


Figure 3

Figure 3 shows the linear regression plot for relative humidity vs. area. This slope is -0.004186. The negative slope is expected from the correlation coefficient in the heatmap.

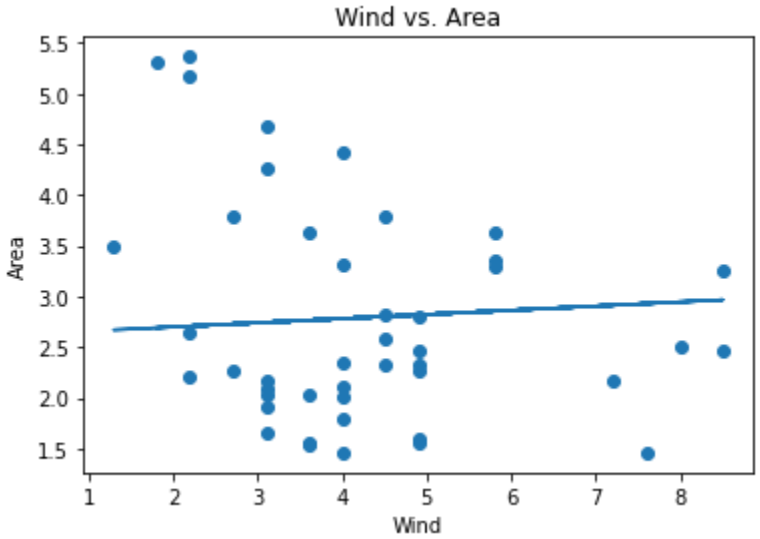


Figure 4

Figure 4 shows the linear regression plot for wind vs. area and the slope has a positive coefficient of 0.04117. This is unexpected because the heatmap shows a negative correlation between wind and area. This could be a consequence of isolating the variables for this regression.

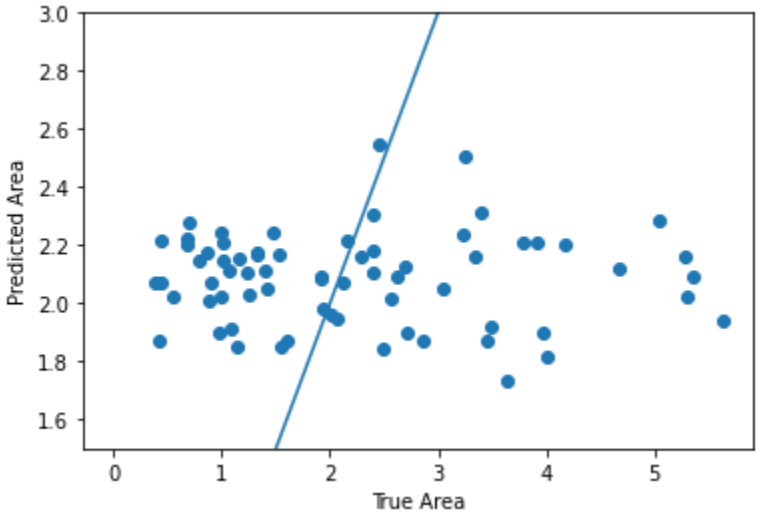


Figure 5

Figure 5 shows the predicted vs. true area for the multiple linear regression plot. The data set was split into train and test sets. The train set was then used to fit the model, while the test set was used to generate a set of predicted area values when given true input values. The predicted area values were then plotted against the real area values for those given inputs. Points on the solid line represent a correct prediction.

	coef	std err	t	P>|t|	[0.025	0.975]
const	12.3520	21.744	0.568	0.571	-30.571	55.275
wind	-0.0045	1.890	-0.002	0.998	-3.736	3.727
temp	0.8550	0.594	1.438	0.152	-0.318	2.028
RH	-0.0330	0.226	-0.146	0.884	-0.479	0.413

Figure 6

Figure 6 shows the coefficients, and p-values for the multiple linear regression. The wind and RH coefficients are both negative while the temperature coefficient is positive which is shown in the correlation heatmap. None of the variables are statistically significant with a p-value of < 0.05. Additionally, the mean squared error for this model is 0.9088 and the r-squared value is 0.0108. These values all indicate that this model is not accurate. Therefore we can conclude that the area cannot accurately be predicted from the wind, temperature, and relative humidity variables alone.

Hypothesis 4

The final hypothesis we tested was the relationship between relative humidity, and fire size, specifically, as relative humidity increases, fire size decreases. Similar to the other three hypotheses, we used the log transformation of area, (log(1+area)), in order to increase the normalization of the data. The histogram below, Figure 1, shows the distribution of humidity in the dataset, with the majority of values being between 25% and 50%. This is expected, as humidity is the saturation of air, and fires are less likely to happen when there is high humidity.

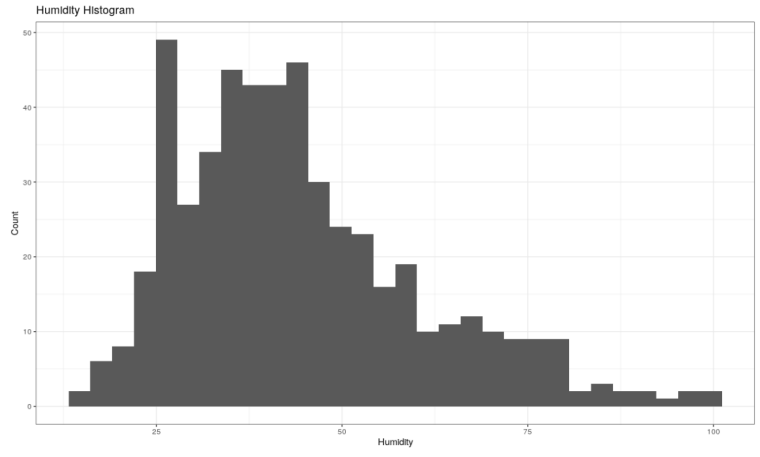
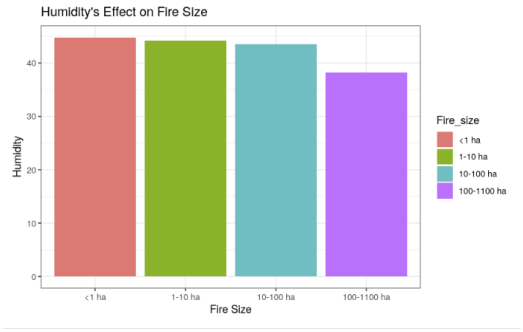


Figure 1

Figure 2 shows the average humidity for four brackets of fire size, with the brackets being fires smaller than 1 hectare, between 1 and 10 hectares, 10-100 hectares, and fires larger than 100 hectares. The graph below, as well as the values in the table show that, although not extreme, as fire size increases, relative humidity decreases.



Fire_size <fctr>	mean_RH <dbl>
<1 ha	44.79197
1-10 ha	44.25676
10-100 ha	43.50000
100-1100 ha	38.18182

Figure 2

When viewing a correlation matrix or heatmap similar to Figure 1 from hypothesis 3, we can see that relative humidity is not an accurate sole predictor of fire size, although there is a slight negative correlation of -0.13. To further prove this view the scatter plot below, Figure 3.

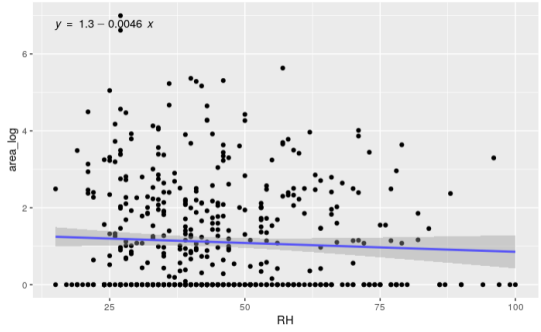


Figure 3

The equation of the regression line is y=-0.0046x+1.3, with an R^2 of 0.00095, according to the coefficients given by the regression model. As stated previously, the negative coefficient is expected, although to a higher magnitude. The p-value is also 0.22, which is far greater than the 0.05 required for this to be statistically significant. This proves that relative humidity is not an accurate sole predictor of fire size, although there is a negative correlation/relationship between the variables.

Model

Random Forest Regressor

We used a random forest regression model to try to predict the area of forest fires. We used a log(area+1) transform on the burned area since these values were heavily skewed towards 0. The model used temp, RH, DMC, and rain as its predictor variables and had 47 trees. This model yielded a mean squared error of about 1.957 and an R-squared of about 0.033. Although this R-squared value wasn't great, it was better than that of the multiple regression and SVR models that we tried.

Random Forest Classifier

We used a random forest classifier to create a model to predict if a fire was going to be a large fire or a small fire. The model used X, Y, month, day, FFMC, DMC, DC, ISI, temp, RH, wind, and rain as the predictor variables. A label was placed on fires with an area greater than $100m^2$.

The model was able to predict if a fire was going to be a large fire with 62% accuracy and produced the following confusion matrix (Figure 1).

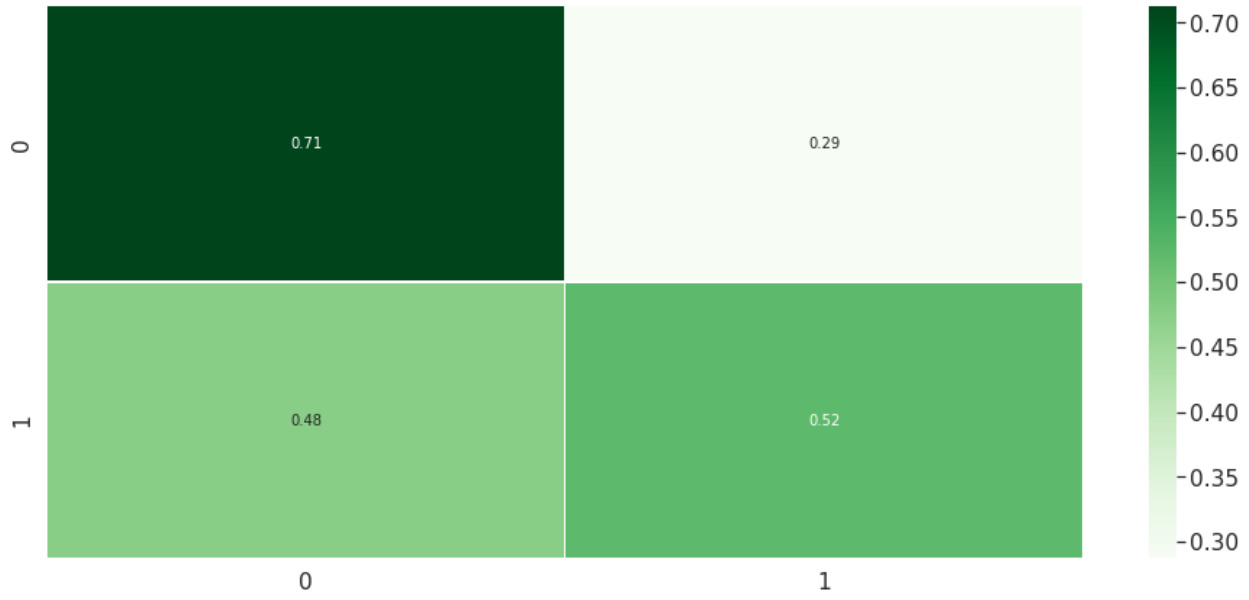


Figure 1

The confusion matrix shows that our model is able to accurately predict a small fire with 71% of the time but is only able to accurately predict a large fire 52% of the time.

Model performance

	Precision	Recall	F1-score
Small Fires (0)	0.61	0.71	0.66
Large Fires (1)	0.64	0.52	0.58

The random forest classifier was able to correctly identify small fires 61% of the time and large fires 64% of the time. The precision of the model combined with the ratio of the correctly identified fires produces the F1 score. The f1 scores were 0.66 and 0.58 for small fires and large fires respectively.

Classifier Model Comparison

Mean Accuracy of Models

	Raw Data	Min Max Scaling	Z Scaling
Random Forest Classifier	0.620	0.579	0.620
AdaBoost Classifier	0.544	0.544	0.544
MLP Classifier	0.591	0.608	0.561

The random forest classifier was tested against the AdaBoost Classifier and the MLP Classifier. The data was also min-max scaled and z-scaled to compare the model with raw data. The random forest classifier had a higher mean accuracy for both the raw data and the z scaled data. For the min max scaled data the MLP classifier was the most accurate at 60.8% but it was still out performed by the random forest classifier with the raw data.

The Z scaled data had a mean accuracy that was equal to the mean accuracy of the raw data. Though the confusion matrix in figure 3 shows that the z-scaled data had a higher accuracy for small fires at 77% (compared to 71%) but a lower accuracy for large fires at 46% (compared to 52%). Since scaling the data showed no significant impact on the model we chose to use the raw data for our model.

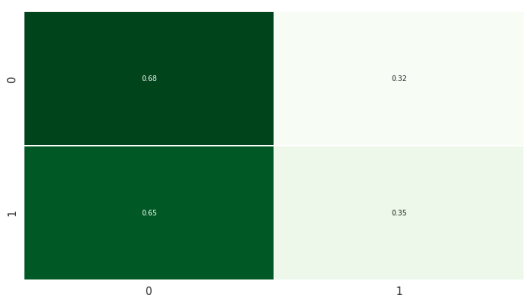


Figure 2 (min max)

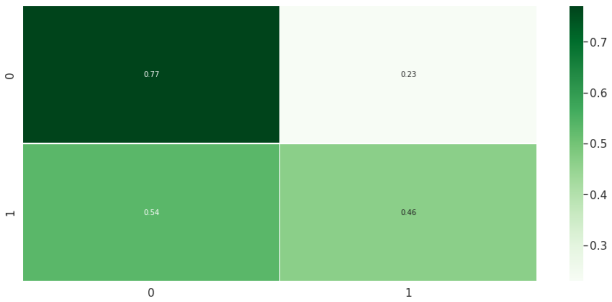


Figure 3 (z scaled)

Results and Limitations

We found that our models were insufficient at predicting fire size. All three models had a relatively low accuracy score, with the random forest classifier, having the highest at 62%. The AdaBoost classifier and MLP classifier scored lower than the random forest classifier, with a 54% accuracy, and a 59% accuracy respectively. Both min-max and z-scaling our data had no effect on our results, as the use of raw data resulted in the highest accuracy. In terms of predicting small fires vs large fires, our random forest classifier was able to predict small fires the best, although the accuracy was only 71%. This could be due to the variables included in the model, or that the information collected in the initial dataset were insufficient in accurately predicting fire size. Overall, our main limitation was the dataset used, and the method in which area was recorded. As mentioned previously, a fire size smaller than 100 m^2, or 1/10 hectare, was recorded as “0.” In the dataset used, there were 247 instances of zero, heavily skewing the distribution leftwards. The dataset also failed to include important variables such as elevation, wind direction, and type of vegetation burned, which could lead to further analysis of the predictors of large/small forest fires that our models could not accurately predict.

Conclusion

In short, we found that our initial analysis suggested that drought is a more important predictor of the size of the fire than the weather conditions at the time of ignition, but further analysis made this unclear. For example, a regression showed that DC was statistically significant, which was expected, but it had a negative coefficient, which is counterintuitive. We proved our hypothesis that there is no dependency on humidity for

fire size by discovering a weak and negative relationship between the variables. We found that the three models we attempted did not perform well at predicting whether a fire would be large or small.

Acknowledgements

Percentage contribution of each team member (doing their part)	
Team member	**Percentage**
Robert Seiler	100%
Trevor Holt	100%
Guillermo Felce	100%
Hamza Elsiesy	100%
Alyssa Guerrero	100%
Jackson Dugger	100%
Joshua Hess	100%

Canadian Fire Weather index System:

https://www.nwcg.gov/publications/pms437/cffdrs/fire-weather-index-system

Climate data for northeast Portugal comes from:

https://weatherspark.com/y/33562/Average-Weather-in-Bragan%C3%A7a-Portugal-Year-Round

Relevant Paper:

[Cortez and Morais, 2007] 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, Guimarães, Portugal, pp. 512-523, 2007. APPIA, ISBN-13 978-989-95618-0-9.