STAT306 Group Report

By Bora Guney, Ryan Kwan, Anne Nguyen, Yuchen Zhang

Introduction

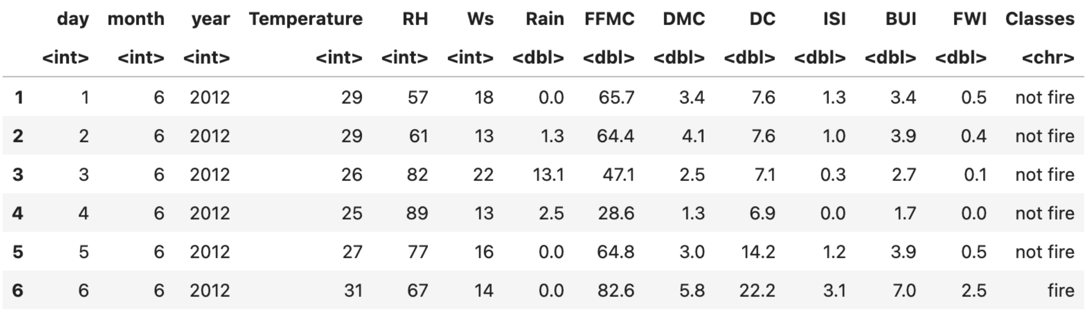
Forest fires are associated with warm and dry conditions and with the changes that have occurred and are occurring to the Earth's climate, these conditions are appearing more frequently (1). As such, we are interested in exploring the relationship between specific weather conditions and forest fires which may allow us to predict future fires. Predictions may allow governments to better prepare for or prevent these fires and allow people to have a wider frame of time to evacuate if the fires cannot be contained. To explore the relationship between weather conditions and forest fires, we examine the Algerian forest fires data set from the UCI Machine Learning database. The data was provided by Abid Faroudja from the Center for Development of Advanced Technologies in Algeria (2). The variables for the data set are described as follows.

Weather condition measurements were recorded each day in two regions in Algeria, the Bejaia region in northeast Algeria and the Sidi Bel-Abbes region in northwest Algeria. Dates cover the months of June to September in the year 2012 and are listed with individual columns for day, month, and year. The temperature is recorded in Celsius on each day at noon for the described dates and ranges from 22 to 42 degrees Celsius. Relative humidity is recorded in percentages and ranges from 21% to 91%. Wind speed is recorded in km/h and ranges from 6 to 29km/h. The total rain each day is recorded in mm and range from 0 to 16.8mm. For each observation, it was noted whether or not there was a fire that day.

In addition to these variables, there are forest fire weather index system components variables that are related to the above variables. Three of these components relate to the moisture content and dead organic matter in the forest while the remaining three relate to fire behaviour (3). The Fine Fuel Moisture Code (FFMC) index is a numeric rating of the moisture content of forest litter and cured fine fuels. FFMC is an indicator of the relative ease of ignition and flammability of the fine fuels and ranges from 28.6 to 92.5 in this data set. The Duff Moisture Code (DMC) index is a numeric rating of the moisture content of loosely compacted organic layers of moderate depth. DMC is an indicator for fuel consumption in moderate duff (decaying organic matter) layers below the litter layer ranges from 1.1 to 65.9. The Drought Code (DC) index is a numeric rating of the average moisture content of deep and compact organic layers. DC is a useful indicator of seasonal drought effects on forest fuels as well as amount of smouldering in deep duff layers and measurements range from 7 to 220.4. The Initial Spread Index (ISI) is a numeric rating of expected rate of fire spread where the calculation of the rating is based on wind speed and FFMC. ISI ranges from 0 to 18.5. The Buildup Index (BUI) is a numeric rating of the total amount of fuel available for combustion and is based on DMC and DC. BUI ranges from 1.1 to 68. The Fire Weather Index (FWI) is a numeric rating of fire intensity in which the calculation of the rating is based on ISI and BUI and ranges from 0 to 31.1.

We will first tidy up the data then explore the relationships between these variables in the data set. From the analysis, we will determine which variables to use in creating a regression model to predict whether certain conditions lead to forest fires.

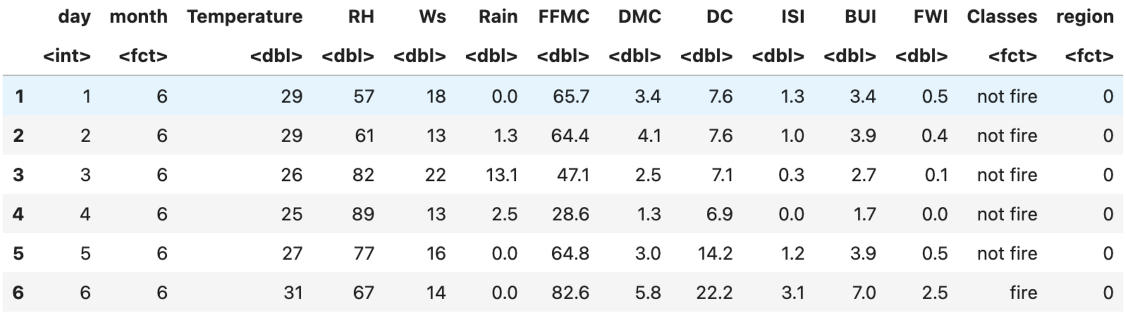
Analysis



	day	month	year	Temperature	RH	Ws	Rain	FFMC	DMC	DC	ISI	BUI	FWI	Classes
	<int>	<int>	<int>	<int>	<int>	<int>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<chr>
1	1	6	2012	29	57	18	0.0	65.7	3.4	7.6	1.3	3.4	0.5	not fire
2	2	6	2012	29	61	13	1.3	64.4	4.1	7.6	1.0	3.9	0.4	not fire
3	3	6	2012	26	82	22	13.1	47.1	2.5	7.1	0.3	2.7	0.1	not fire
4	4	6	2012	25	89	13	2.5	28.6	1.3	6.9	0.0	1.7	0.0	not fire
5	5	6	2012	27	77	16	0.0	64.8	3.0	14.2	1.2	3.9	0.5	not fire
6	6	6	2012	31	67	14	0.0	82.6	5.8	22.2	3.1	7.0	2.5	fire

Table 1. First six observations from the original Algerian forest fire data set for the Bejaia region.

Table 1 shows the original data set we read out via read.csv(). After a series of data cleaning, we removed all values that were not relevant to the analysis and added a column for "region" where 0 represents the Bejaia region and 1 for the Sidi-Bel Abbes region. We obtained the following dataset.



	day	month	Temperature	RH	Ws	Rain	FFMC	DMC	DC	ISI	BUI	FWI	Classes	region
	<int>	<fct>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<fct>	<fct>
1	1	6	29	57	18	0.0	65.7	3.4	7.6	1.3	3.4	0.5	not fire	0
2	2	6	29	61	13	1.3	64.4	4.1	7.6	1.0	3.9	0.4	not fire	0
3	3	6	26	82	22	13.1	47.1	2.5	7.1	0.3	2.7	0.1	not fire	0
4	4	6	25	89	13	2.5	28.6	1.3	6.9	0.0	1.7	0.0	not fire	0
5	5	6	27	77	16	0.0	64.8	3.0	14.2	1.2	3.9	0.5	not fire	0
6	6	6	31	67	14	0.0	82.6	5.8	22.2	3.1	7.0	2.5	fire	0

Table 2. First six observations from the Algerian forest fire data set after cleaning.

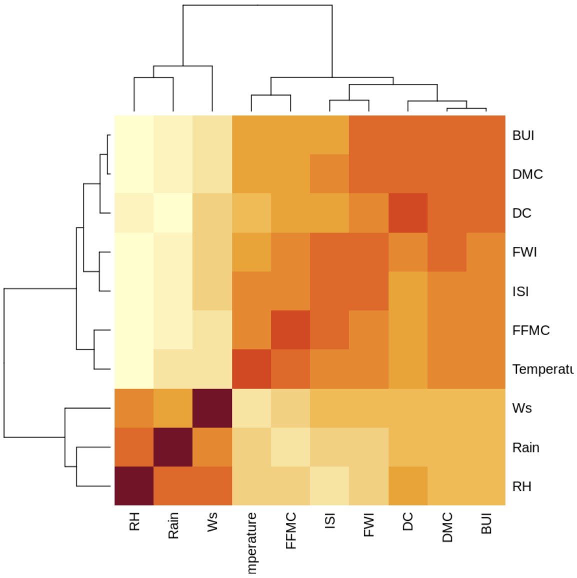


Figure 1. Correlation plot for numeric variables of interest from the data set.

We then made a correlation matrix for a subset of variables from the cleaned data set. The variables include "Temperature", "RH", "Ws", "Rain", "FFMC", "DMC", "DC", "ISI", "BUI", and "FWI". This was done using the cor() function, which calculates the correlation coefficients between pairs of variables. Then we create a correlation plot using the heatmap function, which generates a heatmap visualisation of the correlation matrix.

PCA was then used to help us analyse the contribution of the individual variables to the predicted forest fires, and further analysis using scatterplots was carried out on the variables with relatively high correlation based on the results in the heat map. From PCA, we found about four components covered 90% of the variance.

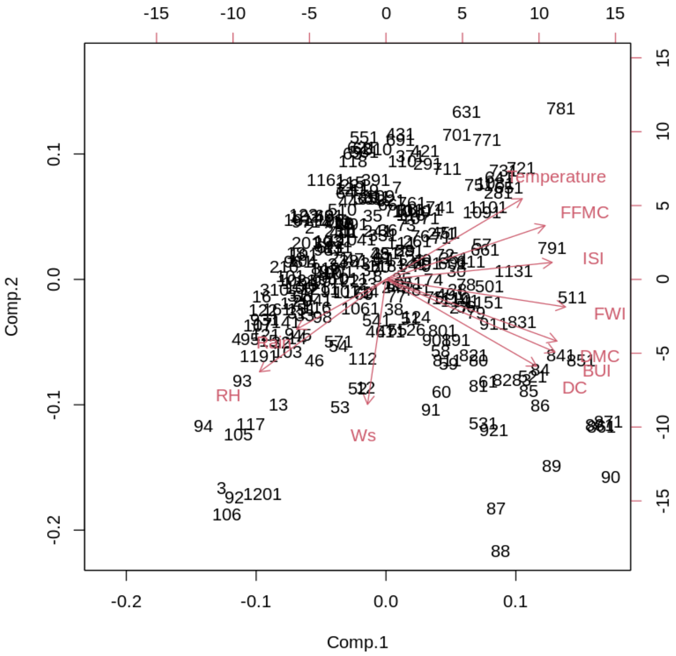


Figure 2. Biplot for the principal component analyses (PCA) on numeric variables of interest.

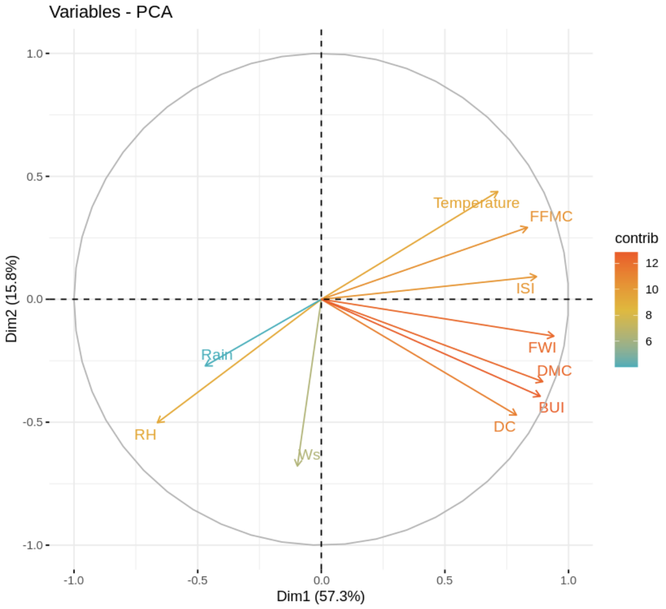


Figure 3. Variable factor map for the variables from the forest fire data set.

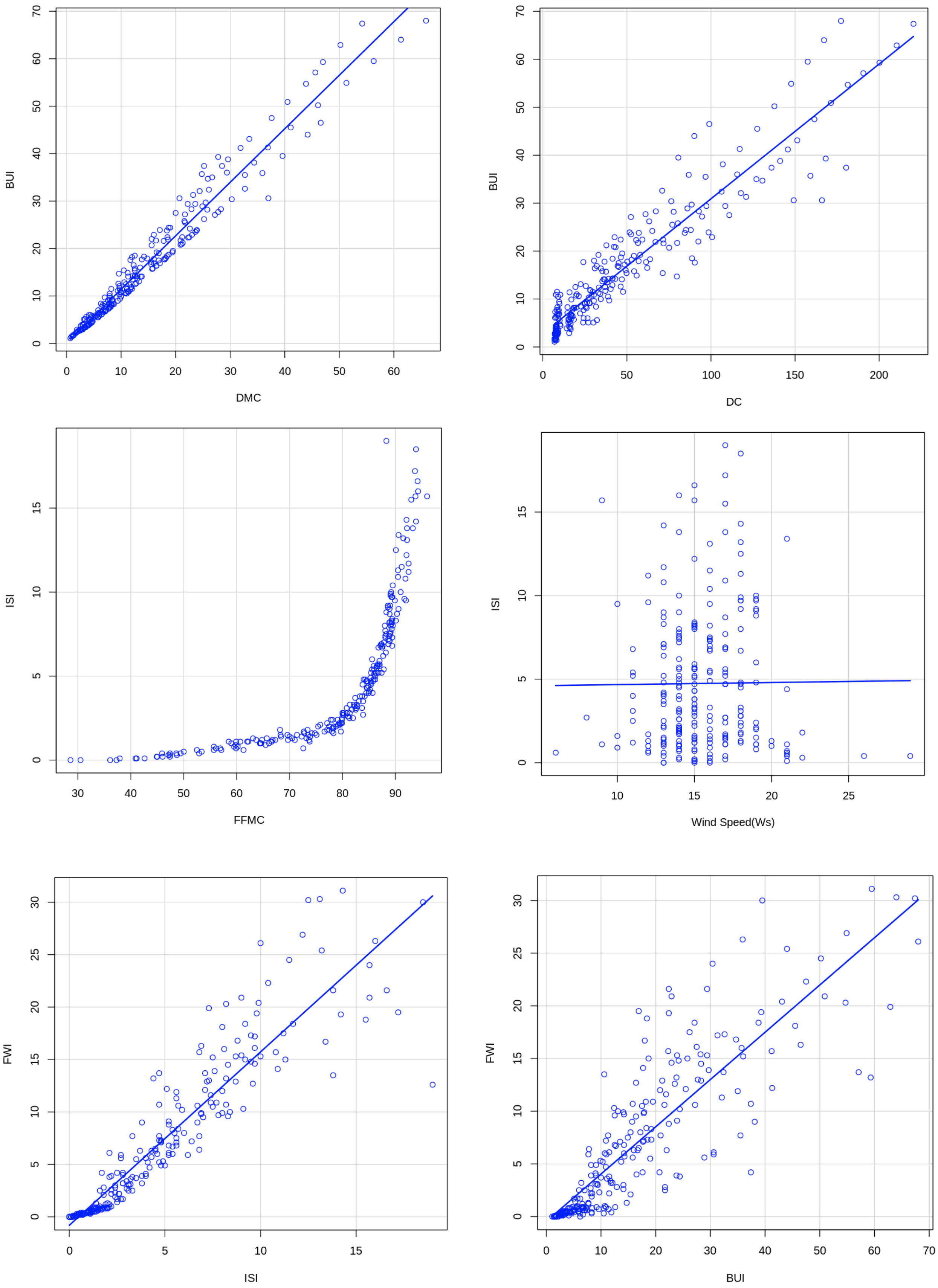
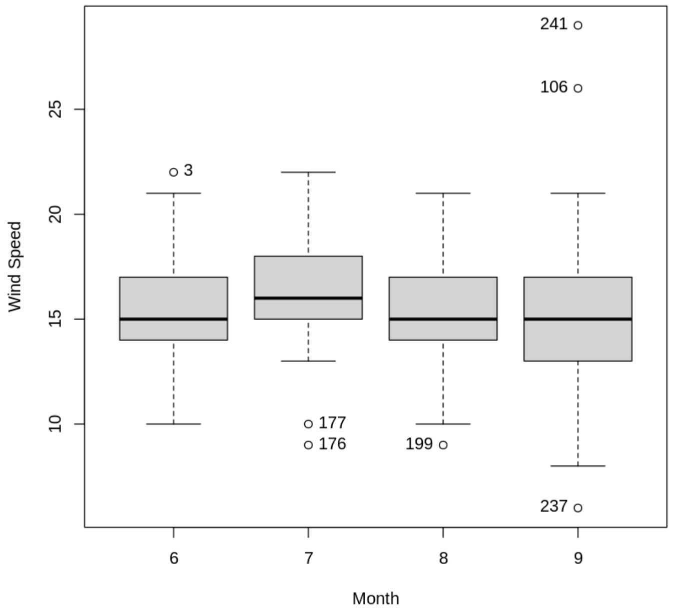


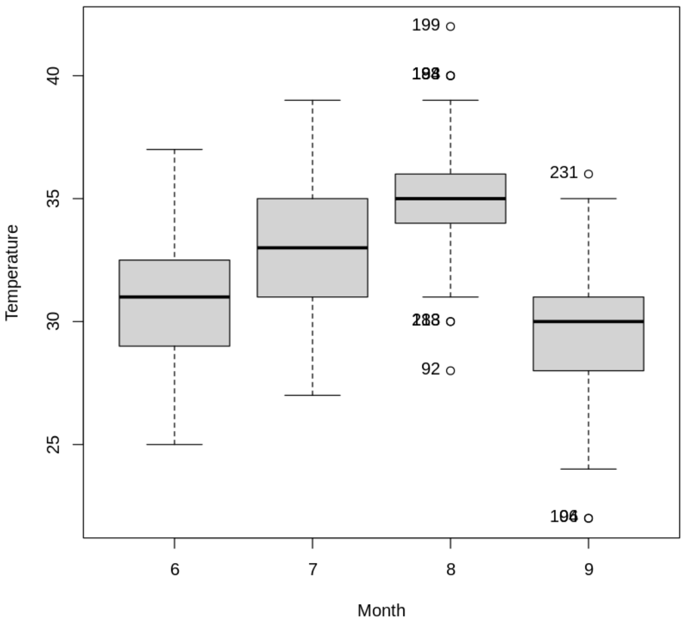
Figure 4. Scatterplots examining the relationships between the variables. From left to right, top to bottom: BUI vs. DMC, BUI vs. DC, ISI vs. FFMC, ISI vs. Wind Speed, FWI vs. ISI, FWI vs BUI.

From the scatterplots of BUI vs DMC and BUI vs DC, we deduced that BUI is linearly dependent on DMC and DC, which also confirms the definition of BUI. From the ISI vs FFMC scatterplot we determined that there is a quadratic relationship between ISI and FFMC. Yet, we did not detect linear dependency between ISI and Wind Speed (Ws). Since BUI seems to be linearly dependent on DMC & DC and ISI is linearly dependent on FFMC, it is reasonable to drop DMC, DC, and ISI from our model because including them would give misleading results due to collinearity with the other variables. Another derivation we made from the scatterplots is that FWI is also linearly dependent on BUI and ISI. We need to be careful here, since FFMC is linearly dependent on ISI, then FFMC is linearly dependent on FWI as well by induction. Therefore, it makes sense to proceed with just one of FFMC, FWI, BUI, ISI in our analysis. By considering the plot of FFMC vs Classes(“fire” and “not fire”), we decided to choose FFMC for our analysis, where we see a significant correlation between FFMC value and categorization of response as “fire” or “not fire”.

It is important to note that we did not use PCA components in our analysis since we would lose the interpretability of the effect of different variables on whether or not there is a fire. Yet, as one can see from the PCA contributions plot, we explored contributions of different variables on PCA components. To avoid collinearity, we focused on choosing approximately orthogonal variables for our model which also allows us to explain more of the variance. FFMC and Wind Speed(Ws) seemed to be approximately orthogonal with respect to each other. Besides these variables, we also decided to consider Temperature for exploration since the fire category notably has a higher median temperature from the below analysis. This way we have narrowed the focus of our analysis to three variables.

We continued to analyse the important variables summarised from the above plots with regions and months by boxplots. The median FFMC is 34 and is much higher for the fire category while Wind Speed has the same median for both fire and not fire categories.





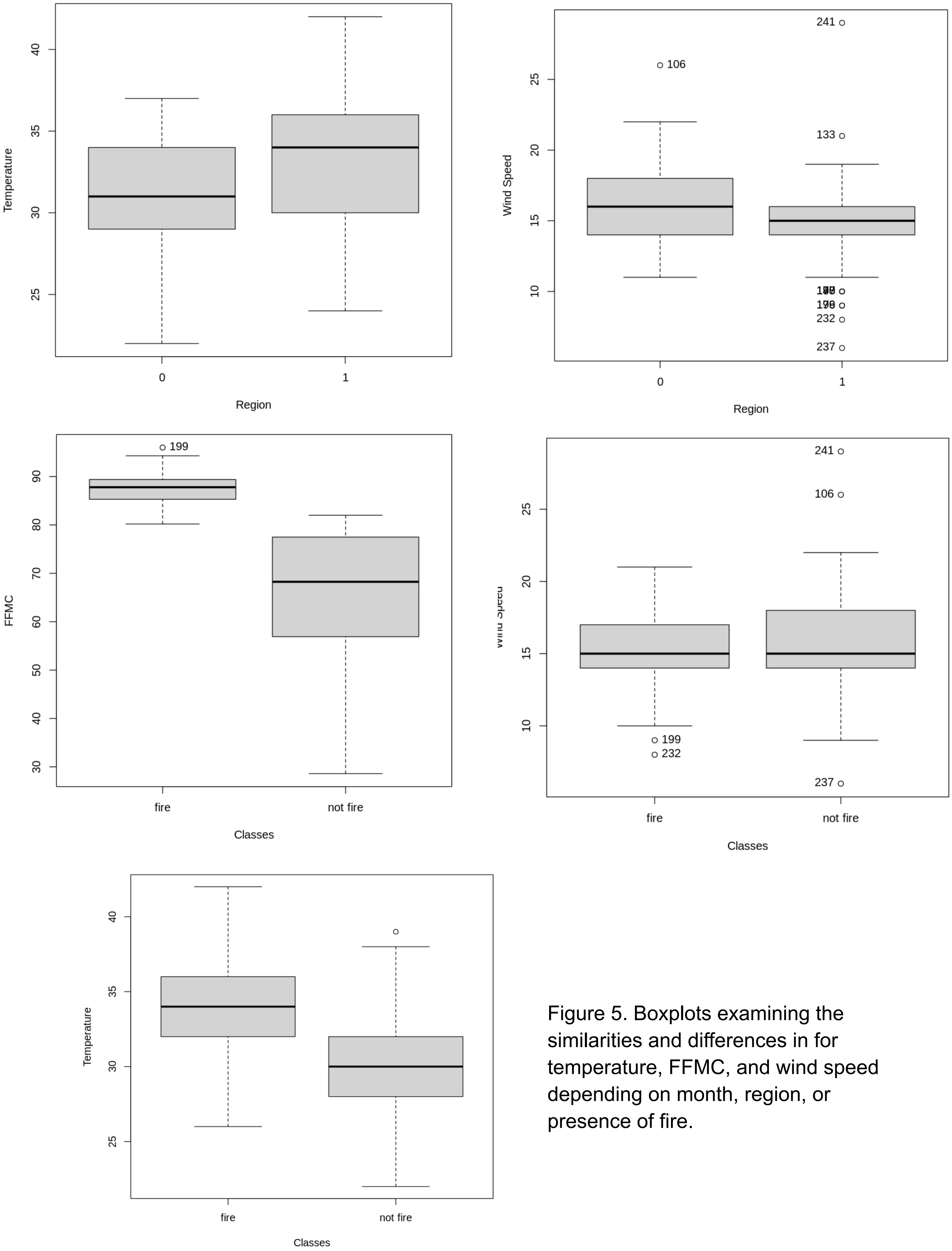


Figure 5. Boxplots examining the similarities and differences in for temperature, FFMC, and wind speed depending on month, region, or presence of fire.

We then created a logistic regression model with whether or not there was a forest fire (Classes) as the binary response variable and temperature, wind speed and FFMC as the explanatory variables (Classes ~ FFMC + Ws + Temperature). The distribution of model is assumed to be binomial. All terms with the exception of temperature were significant at the 5% level. We found the AIC value to be 27.15.

Next we split the data into training and testing sets to examine the prediction accuracy of the model. We sampled 70% of the rows for the training data while the remaining rows constitute the test data set. The logistic regression model is fitted using the training data. Then we perform 10-fold cross-validation on a logistic regression model with the predictors FFMC, Ws, and Temperature to predict the response variable, Classes.

Based on the results of cross-validation, the generalized linear model using the predictors FFMC, Ws, and Temperature has an estimated accuracy of 95.22%, which suggests strong agreement between predicted and observed classes. When tested on the testing data, the model achieved an accuracy of 98.63%. This suggests that the model is performing well and has good predictive power. As well, the ROC curve is almost a perfect right angle.

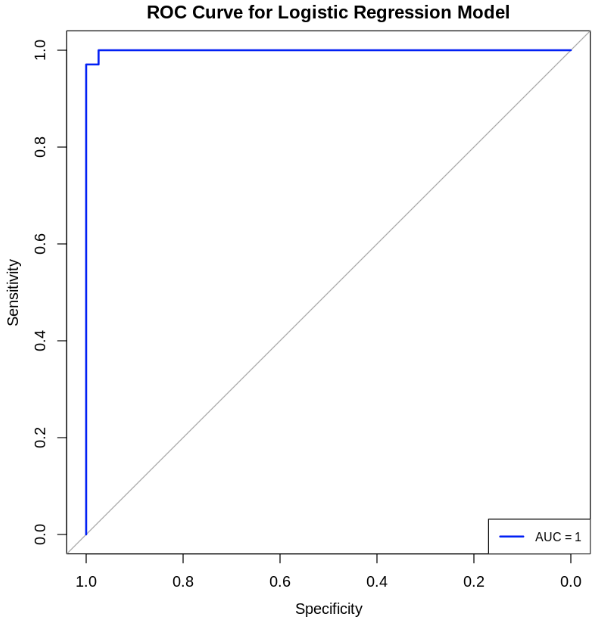


Figure 6. ROC curve for the logistic regression model with variables FFMC, wind speed, and temperature. Since the shape is almost a right angle, it shows the classifier has near perfect performance.

Since p-value for temperature is not significant, we decided to make a logistic regression model with whether or not there was a forest fire (Classes) as the response variable and wind speed and FFMC as the explanatory variables.

The output shows that both "FFMC" and "Ws" are significant predictors of "Classes" at the 5% level of significance. The intercept is also significant. The residual deviance is relatively small, indicating that the model fits the data well. The AIC value of 25.157 suggests that this model is a good fit for the data.

We then perform the same 10-fold cross validation on the second model. The resulting accuracy score shows that the model correctly predicted the occurrence or non-occurrence of forest fires in the testing set with 100% accuracy. The resulting ROC curve for this model had a perfect right angle. This result may be due to overfitting, and there may also be other issues with the model, such as data bias. If there is data bias, the model may simply be dividing the data into categories without truly learning the features of the data. We will further discuss it in the summary section.

Conclusion

After cleaning and parsing the data, we performed an analysis of all the variables by using a correlation matrix with a heatmap to help visualise correlations between the variables from the data. Afterwards, we created a PCA of the data and create a 2D model. This model helped us determine which variables are linearly dependent on each other. We found that BUI was linearly dependent on DMC and DC and ISIS was linearly dependent on FFMC. Due to this, we dropped DMC,DC, and ISI because of collinearity.

An exploration was done on Wind Speed and Temperature indexed by month which gave a negative correlation. We found that through box plots we discovered a high median for FFMC and temperature for the fire category while wind speed had similar medians for both fire and non-fire's category. As a result, we fitted a logarithmic model for FFMC, Wind Speed, and Temperature which gives an AIC of 27.15 while a model with just FFMC and wind speed gives an AIC of 25.157.

The generalised linear model with the predictors FFMC, wind speed, and temperature gave an estimated accuracy of 97.02% which suggests that there is a strong agreement between predicted and the observed classes. Testing the following model on the testing data, we achieved an accuracy of 98.63% indicating that the model is good and has a good predictive power.

We found that if the model only includes wind speed and FFMC, it results in an accuracy of 100% regardless of how the training data and testing data are processed. Adding temperature to the model increased the AIC by 2 but the model appears more reasonable during cross-validation. The sample size of 243 might be too low which may explain how the model with FFMC and wind speed resulted in an accuracy of 100%.

Overall, we decided to select the model with wind speed and FFMC as predictors due to higher accuracy and lower AIC. Our model having an accuracy of 100% means that it is statistically significant to conclude that there is a high correlation between FFMC and wind speed with forest fires.

We decided not to use the PCA components but the contributions of variables to the variance which helps us derive the relevant variables in our prediction. Therefore, for future Algerian forest fires, it is possible to make predictions with these specific conditions.

References

1. Halofsky, J.E., Peterson, D.L. & Harvey, B.J. Changing wildfire, changing forests: the effects of climate change on fire regimes and vegetation in the Pacific Northwest, USA. fire ecol 16, 4 (2020). https://doi.org/10.1186/s42408-019-0062-8

2. Faroudja, A et al. Predicting Forest Fire in Algeria using Data Mining Techniques: Case Study of the Decision Tree Algorithm. International Conference on Advanced Intelligent Systems for Sustainable Development (AI2SD 2019). 08-11 July, 2019. Marrakech, Morocco.

3. Natural Resources Canada. (n.d.). Canadian Forest Fire Weather Index (FWI) System. Government of Canada. https://cwfis.cfs.nrcan.gc.ca/background/summary/fwi