Data 6200: Data Manipulation and Visualization
Final Project
Sean Chapman
0959653

Data Set 1: Portugal

Description: The data set contains 517 records acquired from sampling 4 essential weather attributes and calculating 4 forest fire weather system indices using weather data from Montesinho park in Portugal. Each record has an associated spatial coordinate within an 8x9 grid of Montesinho park in Portugal and temporal values for month and day of the week. The area burned in ha is also included. The weather attributes are temperature, relative humidity, wind and rain. The forest fire weather system indices are ISI index (initial spread index), DC index (drought code index), Duff moisture code index and FFMC index (fine fuel moisture code index).

Size: 517 records with 13 attributes; 2 spatial (x, y grid coordinates within the park), 2 temporal (Day, Month), 4 meteorological (temp, relative humidity, wind, rain) and 4 indices (ISI,DC,DMC,FFMC) as well as area burned in ha.

Reference: The data set is stored in the Kaggle repository under the name Forest Fires Data Set Portugal and was compiled by Ishan Dutta. https://www.kaggle.com/ishandutta/forest-fires-data-set-portugal

Data Set 2: Algeria

Description: The data set contains 244 records acquired from sampling 4 essential weather attributes and calculating 4 forest fire weather system indices using weather data from two regions in Algeria. 122 records were sampled from the Bejaia region, while the other 122 were sampled from the Sidi Bel-abbes region. Each record has a temporal component (date DD/MM/YYYY), 4 meteorological components (temperature, wind speed, relative humidity and rain) and 6 forest fire weather indices. Each record is classified as fire or nofire dependant on weather a fire was observed on that day. The The forest fire weather system indices are ISI index (initial spread index), DC index (drought code index), Duff moisture code index, FFMC index (fine fuel moisture code index), BUI (build up index) and FWI (Fire weather index).

Size: 122 records x 2 regions for a total of 244 records. Each record contains a classification of either fire or no fire, along with the 11 attributes.

References: [1] Predicting Forest Fire in Algeria Using Data Mining Techniques: Case Study of the Decision Tree Algorithm, Abid, F., & Izeboudjen, N., International Conference on Advanced Intelligent Systems for Sustainable Development, PP 363-370 Springer, Cham.

Questions:

Question 1: Can we use this data to create a model which accurately predicts forest fires and the size of them? Can we create a model which works in both countries?

Question 2: Are there environmental predictors that are not needed which would still allow an accurate prediction of a fire occurring for a given time and space?

Question 3: What is the busiest time of year for fires? In other words, when should we be most prepared to fight fires, specifically large ones?

Summary Report:

Each data set contains the response variable "Class" which indicates whether a fire had occurred for any given observation. For Portugal, "Class" was computed using the area burned in ha. Any point with a burned area >0 is classified as having had a fire occur.

Table 1: Forest fire indexes included in the data with abbreviation and description.

Forest Fire Index	Abbreviation	Weather Metric Used	Description
Initial Spread Index	ISI	Wind	Represents the ability of a fire to spread in the initial stage of burning
Drought Code Index	DC	Temperature Precipitation	Represents the moisture content of deeper layers of soil and large fuels
Duff Moisture Code Index	DMC	Humidity Temperature Precipitation	Represents the moisture content of duff (organic surface soil) and medium sized fuels
Fine Fuel Moisture Code Index	FFMC	Humidity Temperature Precipitation Wind	Represents the moisture content of fine fuels and organic material on soil (leaves, loose fuel)
Build Up Index	BUI	DC, DMC	Represents the total fuel available to the fire
Fire Weather Index	FWI	BUI, ISI	General rating of fire intensity/fire danger

The predictor variables include relative humidity, wind speed, temperature and precipitation. From these, several forest fire weather indices were calculated. These forest fire weather indices are standard calculations from the weather measurements to represent different aspects of forest fire dynamics which are difficult to measure directly. They are described in table 1.

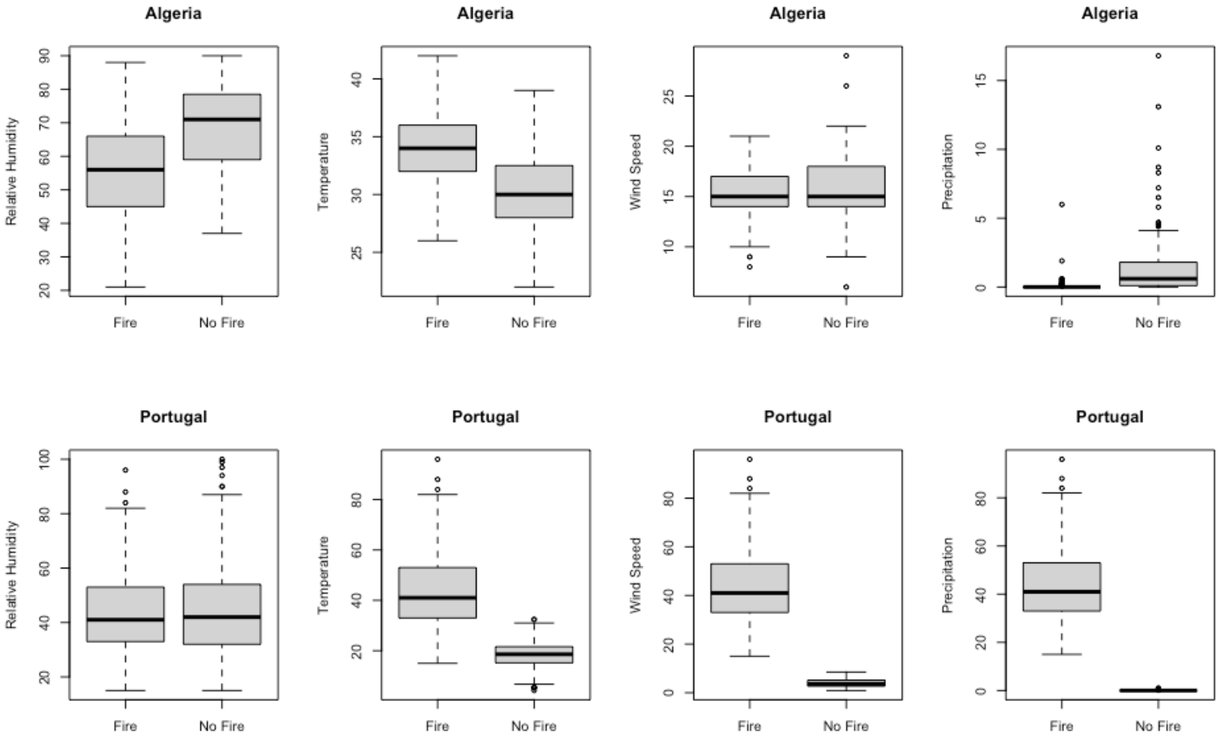

Figure 1: Visualizing the distribution of weather measurements from Algeria and Portugal. Data was separated by the response variable to observe any differences in the weather measurements between responses.

From these boxplots, we can see that most of the distributions of weather measurements are not significantly different between observations of different classes. The weather measurements that are significantly different between observations of fires and no fires are wind speed and precipitation in Portugal. The observed values for these variables are significantly higher in fires than in no fires. This could be due to a sampling bias or some trend in the data. It could be that fires in Portugal occur during storms where wind and precipitation are high, and don't occur during calm weather where there is no wind or rain. The median temperature of fires in Portugal is higher than the upper whisker of the temperature of days with no fire. In general, there is a more distinct difference between the weather patterns of fires and no fires in Portugal compared to Algeria.

In figure 2, we can visualize the weather patterns over time for the Algeria data set. This data set was sampled from 122 days during the summer months from June to September. Each index represents a day, and the data can be divided into the two regions easily. From figure 2, we can see that the daily wind and relative humidity values are randomly distributed. We can also see the trend of temperature rising into the mid-summer days and falling into September. The most interesting trend is that of precipitation. One of the sampled regions does not have many days of high precipitation, and when rain is observed, a fire occurs. The other region has more frequent rainfall and thus has days with high amounts of rainfall and no fire. The first region seems to follow the same trend observed in the boxplots of Portugal. This could mean precipitation will be important in the outcome of the model.

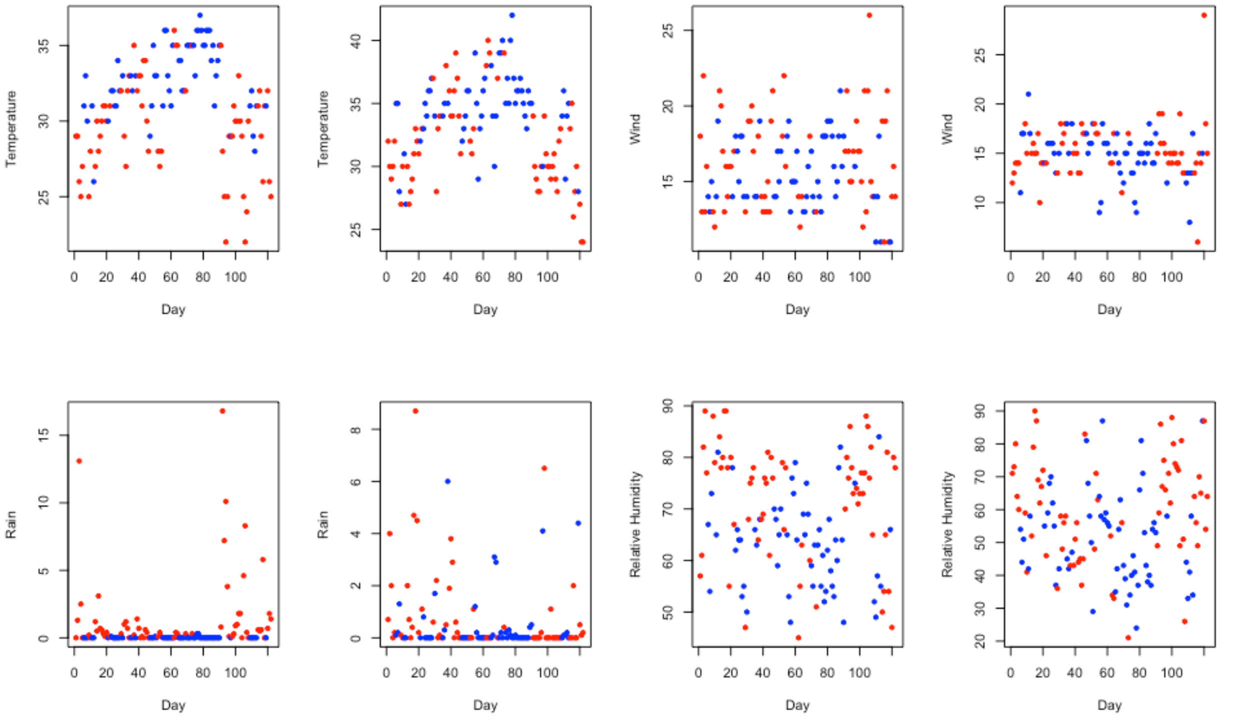

Figure 2: Daily variation in the four weather measurements from the Algeria data set. Red points represent fires while blue points represent no fires. Two graphs were included for each weather measurement, representing each of the regions sampled.

Methods:

Question 1: Can we use this data to create a model which accurately predicts forest fires and the size of them? Can we create a model which works in both countries?

In order to answer this question as best as possible, several different modelling approaches will be attempted. For each modelling approach the data will be split into a training and testing set from a random sample of all observations. The test set will contain 100 of the 768 observations and the remaining 668 observations will be used to train the different models. The trained models will then be used to classify the response of the test set. This classification will be compared to the observed responses of the test set to find the misclassification rate for each model. Once the best model is selected, we need to test if the model can classify the response of one country using the data from the other country. For this test, we will train the ideal model using one of the data sets and use this model to classify the response of the other data set.

The response of the combined data set is binary. The “Class” covariate indicates whether a fire has occurred for an observation. Models which predict the class of an observation need to have a binary response. The models which were built to answer this question included a generalized linear model, a binary classification tree, and a feed forward neural network with one hidden node. Each of the models use all other covariates in the full data set as predictors. These covariates include month, wind, temp, rain, RH, FFMC, DMC, DC and ISI. All covariates are included in order to maximize the predictive capability of the ideal model.

Question 2: Are there environmental predictors that are not needed which would still allow an accurate prediction of a fire occurring for a given time and space?

In order to answer this question, dimensionality reduction analysis needs to be conducted on the models used to answer question 1. In order to reduce the covariates in the generalized linear model, the step() function in R was used. This function iteratively removes one of the covariates from the model in order to reduce the AIC score. The algorithm is complete once the AIC score has reached a minimum and the covariates of this model are shown. In order to minimize the covariates used in the tree, we can tabulate the cross validated error of trees with differing numbers of splits to find the number of splits which minimizes the cross validated error. We can then build this minimized tree using the corresponding cp-score.

Once these models were reduced, the performance of each was assessed using the same cross validation methods as in question 1. The performances of the reduced models were compared to the full models to directly provide an answer to question 2.

Question 3: What is the busiest time of year for fires? In other words, when should we be most prepared to fight fires, specifically large ones?

In order to answer this question a hidden Markov model was built on the data. This model uses only the month as a covariate to determine whether the probability of a fire depends on the value of the previous data point. The performance of this model was calculated as the percentage of points in the data set which were correctly classified by the model. In order to interpret the results and determine which time of year has the most fires, some visualizations were created which are included in the results section.

Results:

Question 1:

The first model built on the training data was the full GLM. The ideal classification threshold is identified using figure 3 as 0.65. Observations with a predicted probability of a fire higher than this threshold will be classified as having a fire when constructing a confusion matrix. Figure 3 indicates 0.65 as the ideal classification threshold since the proportion of fires and no fires correctly classified is maximized. A lower threshold would classify more of the fires as fires but would also classify more of the no fires as no fires. The receiver operating characteristic (ROC) curve shows that the ideal full GLM has both a high sensitivity and specificity.

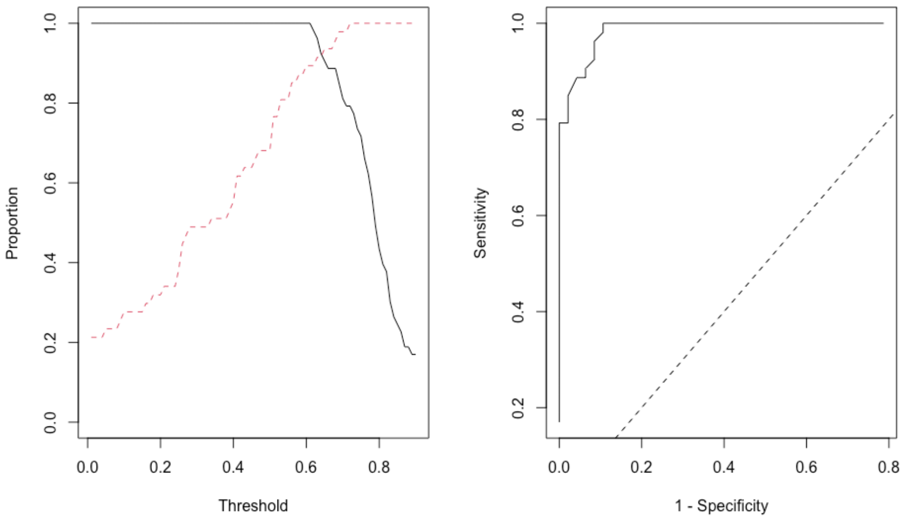

Figure 3: Sensitivity and Specificity of the full GLM predicting fires plotted as a function of the probability function on the left and as an ROC curve on the right. A classification threshold of 0.65 was chosen to maximize both Sensitivity and Specificity of the model.

Table 2: Full models with accuracy and classifications of the test set. Columns 4 through 6 represent the confusion matrix for the model. The first digit represents the true value while the second digit represents the prediction.

Model	Accuracy (%)	1,1	1,0	0,1	0,0
Full GLM	92	47	6	2	45
Algeria GLM	53.3	258	12	229	19
Portugal GLM	63.8	136	1	87	19
Binary Tree	96	53	0	4	43
NN	73	30	23	4	43
Scaled NN	95	51	2	3	44

The Binary Tree model is the most accurate model, having an accuracy of 96% when classifying the results of the test set. The full GLM and Scaled neural network both perform well with 92% and 95% accuracy respectively. The models that were cross validated with each other do not perform with high accuracy. The Algeria GLM, which was trained using the Algeria data set and tested on the Portugal data set, misclassified almost half of the data. The Portugal GLM, which was trained using the Portugal data set and tested on the Algeria data set, performed slightly better, but was still only 63.8% accurate. This suggests that the answer to question 1 is that we can accurately predict whether a forest fire will occur, and that this model works well in both countries. We cannot, however, create a model from one country's data and use it in another country.

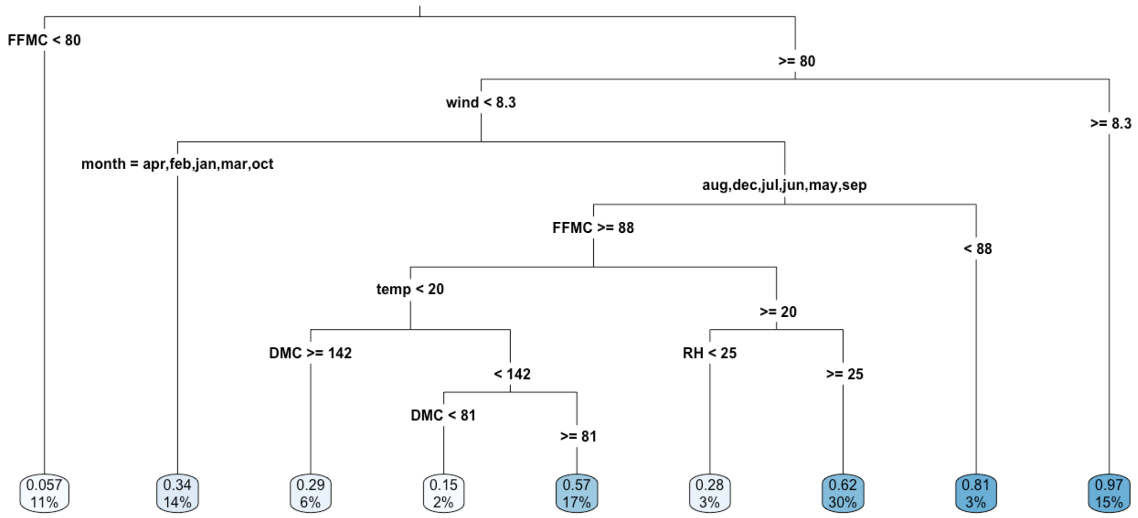

Figure 4: Full binary decision tree model of the data set. Each observation passes through the tree to one of the decision nodes. Each decision node represents the probability that a data point is classified as fire. Darker nodes represent a higher probability of being a fire.

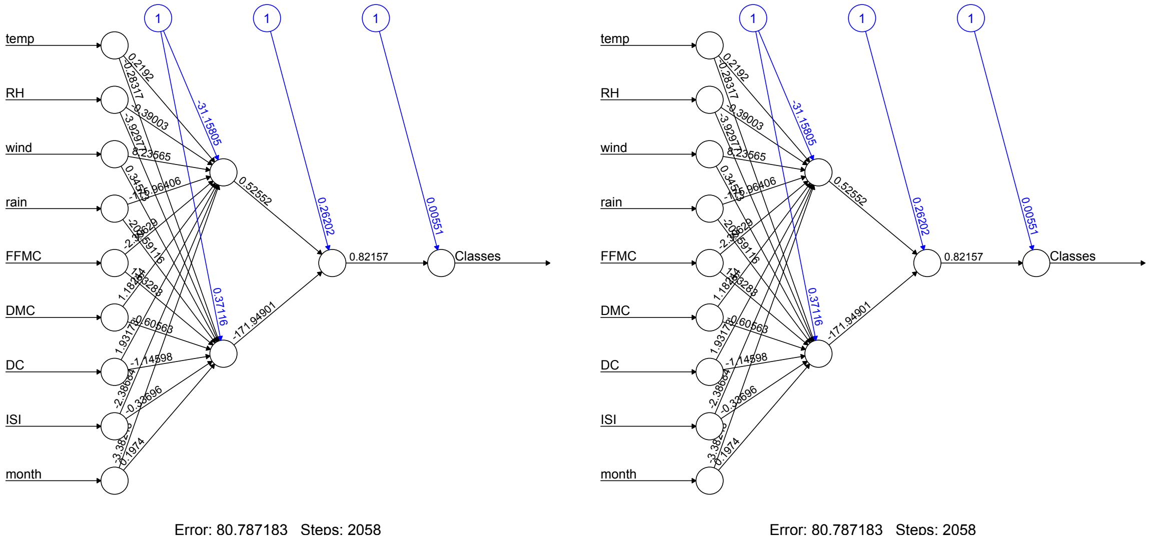

Figure 5: Visualization of the Neural Networks constructed. The network on the left was trained on the raw data while the network shown right was trained on normalized data. The inputs are shown on the left, with each circle representing a node. The networks both contain 10 input nodes, 2 hidden layers with 3 nodes, and a single output node. The weights are shown on each line from node to node.

Question 2:

Each of the models built in question 1 were reduced to construct a model which can predict the class of each observation. The reduced GLM contained the variables month, temp, wind, rain and FFMC. The ideal classification threshold was chosen as 0.65 from figure 4. The reduced tree was created from pruning the full tree using the prune.rpart() function in R. This yielded a tree with three decision nodes, including FFMC, wind and month variables. The accuracy of these reduced models is shown in Table 3.

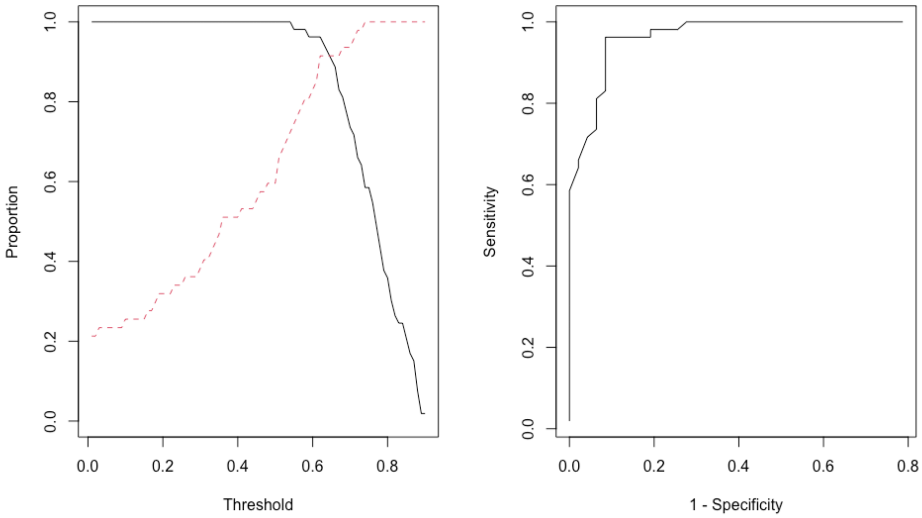

Figure 6: Sensitivity and Specificity of the full GLM predicting fires plotted as a function of the probability function on the left and as an ROC curve on the right. A classification threshold of 0.65 was chosen to maximize both Sensitivity and Specificity of the model.

Table 3: Full models with accuracy and classifications of the test set. Columns 4 through 6 represent the confusion matrix for the model. The first digit represents the true value while the second digit represents the prediction.

Model	Accuracy (%)	1,1	1,0	0,1	0,0
Reduced GLM	91	48	5	4	43
Reduced Binary Tree	96	53	0	4	43

After reducing the models, both the GLM and Binary Tree models perform quite well on the test set. The binary tree maintains 96% accuracy while the GLM drop from 92% to 91% accuracy. The binary tree and the GLM both use the FFMC as one of their variables. This index relies on all four of the weather parameters to calculate. Therefore, while the models can be reduced to include fewer variables and maintain accuracy, all the weather parameters still need to be recorded in order to provide an accurate model of whether a forest fire is expected to occur. The answer to question 2 is that all the environmental parameters are needed to allow an accurate prediction of a fire occurring for a given time and space.

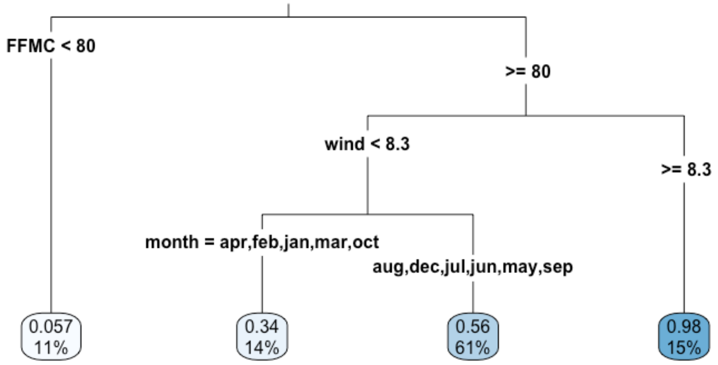

Figure 7: Reduced binary decision tree model of the data set. Each observation passes through the tree to one of the decision nodes. Each decision node represents the probability that a data point is classified as fire. Darker nodes represent a higher probability of being a fire.

Question 3:

The hidden Markov model was not able to accurately predict the fires and had a misclassification rate of 47%. Alternatively, we can interpret the coefficients of the reduced GLM to gain information about the influence of the month on the likelihood of a fire occurring. The reduced GLM contains the month parameter as a factor, setting April as the dummy variable. The coefficients in the model represent the ratio of the log odds of a fire occurring during that month compared to April. Months which are corresponding to a high frequency of fires will have a higher coefficient.

Table 4: coefficients from the reduced GLM corresponding to the months of the year. Each coefficient represents the ratio of log odds

Month	Estimate	Std. Error
Jan	0.315176	0.377021
Feb	0.146305	0.186562
Mar	-0.174369	0.167377
Apr	0	0
May	0.007405	0.362814
Jun	-0.133860	0.175551
Jul	-0.036304	0.173182
Aug	-0.084417	0.163785
Sep	-0.042641	0.161311

Oct	-0.220673	0.197112
Nov	-0.323483	0.489325
Dec	0.585037	0.222424

The months with the highest likelihood of a fire according to the reduced linear model are the months of Dec, Jan and Feb. These months could be more associated with the weather patterns we expect to cause a fire. From the summary report we estimated that data points with high winds and precipitation correspond to fires. These weather patterns could be more common during the winter months. Therefore, it is best to be prepared for a fire during these months. Alternatively, the months of Oct and Nov have the lowest frequency of fires.

Conclusion:

During the analysis, a few problems were encountered. One of these problems was the decision of a classification threshold for the models. For a GLM, there is an algorithm that performs this which I am familiar with using. This algorithm did not work for the Binary Tree or the neural network models. For these models, confusion matrixes were built using 0.5 and 0.65 % classification thresholds and choosing the better of the two. The analysis could be improved by finding an algorithm which creates a figure like figures 3 and 6 for the neural networks and binary trees. Both models performed well in the final analysis so this would improve the performance of the models only moderately.

Another problem which was encountered was the structuring of the data. In order to combine the data, only columns which were common between the two data sets were kept. This lost some information contained in the data sets, including the build-up index and fire weather index from the Algeria data set and the spatial coordinates in the Portugal data set. Additionally, each column in the initial Algeria data set was a factor. In order to perform the analysis on this data set each data point needed to be converted to a numeric, except month, which remained a factor.

Relating to month being a factor, the neuralnet package in R, which was used to build the neural networks used in the analysis was somewhat finicky. The month needed to be converted to a numeric as the network only accepts numeric arguments. Additionally, the data needed to be normalized in order to perform well. The result for the neural net was able to predict the class with a high accuracy so the addition of more hidden nodes would only marginally improve results.

An area for improvement would be to find a model reduction algorithm which could better be applied to question 2. The indices are calculated from the weather covariates, so a model which removes one of the weather covariates should also remove indices calculated from it. This was difficult to do in R. With more time, I could attempt to reduce the model manually to find if there is a weather parameter which could be removed.

References:

Abid, F., & Izeboudjen, N. (2019, July). Predicting Forest Fire in Algeria Using Data Mining Techniques: Case Study of the Decision Tree Algorithm. In *International Conference on Advanced Intelligent Systems for Sustainable Development* (pp. 363-370). Springer, Cham.

Ardianto, R., & Chhetri, P. (2019). Modeling Spatial–Temporal dynamics of urban residential fire risk using a Markov Chain technique. *International Journal of Disaster Risk Science*, *10*(1), 57-73.

De Groot, W. J. (1998, April). Interpreting the Canadian forest fire weather index (FWI) system. In *Proc. of the Fourth Central Region Fire Weather Committee Scientific and Technical Seminar*.

Sakr, Elhajj, I. H., Mitri, G., & Wejinya, U. C. (2010). Artificial intelligence for forest fire prediction. *2010 IEEE/ASME International Conference on Advanced Intelligent Mechatronics*, 1311–1316. https://doi.org/10.1109/AIM.2010.5695809

Natural Resources Canada https://cwfis.cfs.nrcan.gc.ca/background/summary/fwi