Statistical Inference and Prediction on Wildfire Observations
MATH 561 Fall 2020 Final Project

Justin Nichols[1], Miika Jarvela[2], and John Cox[3]

Colorado School of Mines

Abstract

Introduction

We investigate a spatial database of wildfires that occurred in the United States from 1992 to 2015. The wildfire records were acquired from the reporting systems of federal, state, and local fire organizations. We augment this database using spatial definitions of the Level 1 and Level 3 ecoregions in North America maintained by the US Environmental Protection Agency (EPA). We focus this analysis to the Level 1 ecoregions, which define 12 distinct regions, and incorporate the more specific Level 3 ecoregions only to include average fuel moisture and wind speed information. Previous analysis [1],[3] demonstrate a location's available fuel mass and type of vegetation (related to its ecoregion), are important features to determining the cause of fire ignition. Our work builds on these findings by separating fires by Level 1 ecoregion and completing the modeling tasks for each region individually. We complete three statistical modeling tasks and present the results of these tasks. The combined database consists of 1.71 million observations of wildfires.

Task 1 Overview

We seek to develop models to predict the binary "human" response. This feature is 1 if the cause of the fire can be attributed to humans, and 0 otherwise. We use all other features (which could reasonably be known at the time of discovery) as predictors. Overall, model accuracy is modest at best, with aggregate test accuracy in the 83% region, but we show some intuitive relations, and demonstrate an ensemble classification model.

Task 2 Overview

We further dive into attempting to specifically identify the cause of a fire. Rather than a binary approach, we will melt this into twelve separate sources. Comparing the outcomes of each method to the largest class proportion in the data, we find that the best method improves prediction by 5-30%, while others lose slightly here, they create better interpretation for their conclusions.

Task 3 Overview

We wish to predict the size of a fire and understand which of the predictors are related to the size of the fire. We try various different regression models, but they all had very low predictive accuracy with the highest R^2 value being 0.22 in one ecoregion using a random forest. This likely is because the underlying mechanism which determines the size of the fire is complex and we do not have sufficient predictors to try and describe this mechanism.

Conclusions

The Latitude and Longitude predictors are commonly featured in many models in this work, leading us to believe that spatial distribution of the data remains important even after we separate observations by Level 1 ecoregion. It might be useful to develop a model that is able to take the spatial correlation of the data into account.

In future work, we would consider performing transformations to the discovery day of year, month, Latitude, and Longitude predictors to make them more consistent. Using a spatial model, we would be able to handle latitude and longitude predictors more easily. The discovery day of year and month should be handled by taking into account the natural seasonal variation and should be more cyclical (associating, for instance, day 1 as being right after day 365 rather than being very far apart). Collecting additional predictors would also likely improve modeling capability.

[1]jnichols2@mines.edu
[2]mjarvela@mines.edu
[3]jlcox@mines.edu

COLORADO SCHOOL OF MINES

Contents

List of Figures

List of Tables

1 Introduction

We investigate a spatial database of wildfires that occurred in the United States from 1992 to 2015. The wildfire records were acquired from the reporting systems of federal, state, and local fire organizations. We augment this database using spatial definitions of the Level 1 and Level 3 ecoregions in North America maintained by the US Environmental Protection Agency (EPA). We focus this analysis to the Level 1 ecoregions, which define 12 distinct regions, and incorporate the more specific Level 3 ecoregions only to include average fuel moisture and wind speed information. Previous analysis [1],[3] demonstrate a location's available fuel mass and type of vegetation (related to its ecoregion), are important features to determining the cause of fire ignition. Our work builds on these findings by separating fires by Level 1 ecoregion and completing the modeling tasks for each region individually. We complete three statistical modeling tasks and present the results of these tasks. The combined database consists of 1.71 million observations of wildfires, which we subset in each task for ease of computation.

1.1 Level 1 Ecoregions

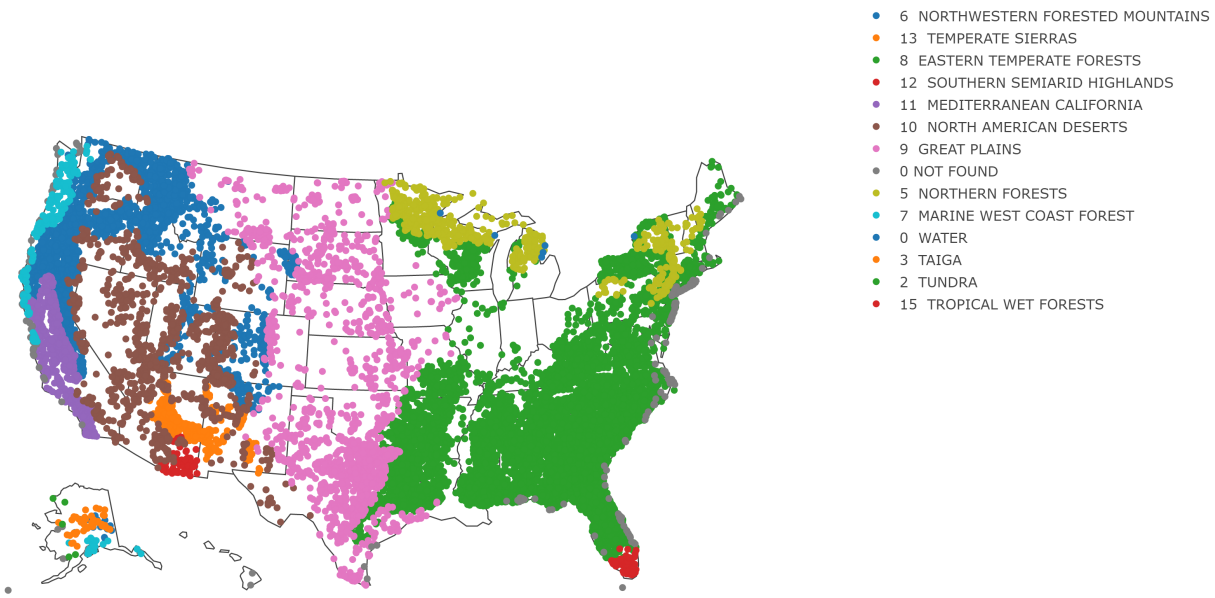


Figure 1: Selection of Observations Showing the Spatial Distribution of Level 1 Ecoregion

We utilize a shape file representation of the Level 1 ecoregions provided by the EPA, consistent with the definition used by Dr. Nagy in her work [2]. We project the latitude and longitude of each fire observation to the equal area projection used by the EPA data and determine the containing egoregion shape. The initial result of this calculation is shown in Figure 1. Note the presence of two troublesome categories: "0 NOT FOUND" and "0 WATER", which are not official ecoregion definitions. We handle these observations before moving on calculate additional features.

Correction of "0 NOT FOUND" Ecoregions
We find that approximately 1.3% of observations are not contained in any ecoregion and are denoted as "0 NOT FOUND". We believe this to be caused by small differences in the reported cartographic projection used in the EPA shape file and our reproduction of this projection, and we therefore desire to remedy this mis-classification. We fit a k-nearest neighbors model to the set of calculated ecoregions to fill in these missing values. We use k=10 and Euclidean distance of the latitude and longitude as a distance metric. The results of this correction are shown in Figure 2, which shows the new ecoregion of the previously "0 NOT FOUND" observations.

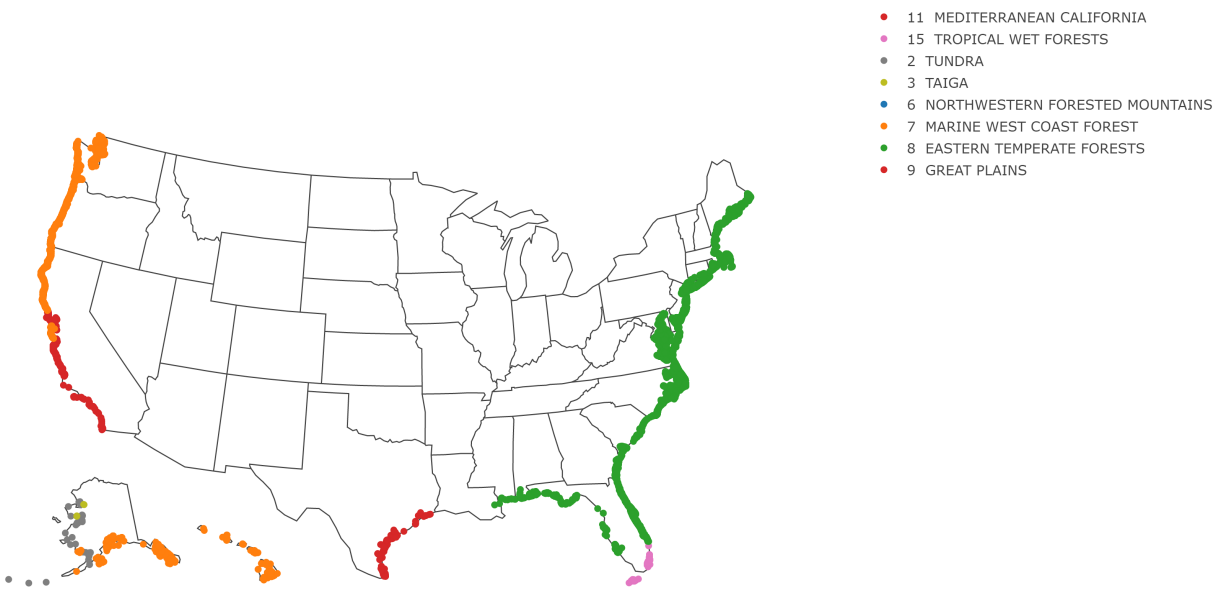


Figure 2: Observations Outside All Ecoregions - Filled by a kNN (k=10) Model

Correction of "0 WATER" Ecoregions
Further, we find that about 0.5% of observation receive an ecoregion of "0 WATER", which is not an official ecoregion but also does not seem to be an issue arising from differences in our projection calculation. These observations are concentrated in the Great Lakes region, and generally consist of very small fires. Rather than calculate a new ecoregion for these observations, we drop them from the database (again because as far as we can tell they are not a caused my a miscalculation on our part and are genuinely located in water). The resulting map of ecoregions follows closely with the actual shape of the Level 1 ecoregions maintained by the EPA on their website. On the advice of Dr. Nagy (that ecoregion is an important feature to understanding the causes of these fires), the Level 1 ecoregion feature calculated here is a main focus in the following work.

1.2 Level 3 Ecoregions

The EPA subdivides the Level 1 ecoregions into Levels 2 through 4, which are hierarchically more specific regions. We utilize a shape file representation of the Level 3 ecoregions provided by the EPA, consistent with the definition used by Dr. Nagy in her work [2]. We project the latitude and longitude of each fire observation to the equal area projection used by the EPA data and determine the containing egoregion shape. We do not plot the results of this projection and bounding shape calculation because there are 105 distinct Level 3 ecoregions, and their visualization is overly cumbersome. We again find that approximately 1.3% of observations are not contained in any ecoregion and are denoted as "0 NOT FOUND". We believe this to be caused by small differences in the reported cartographic projection used in the EPA shape file and our reproduction of this projection, and we therefore desire to remedy this mis-classification. We fit a k-nearest neighbors model to the set of calculated ecoregions to fill in these missing values. We use k=10 and Euclidean distance of the latitude and longitude as a distance metric.
Further, we find that the shape file provided for the Level 3 ecoregions does not include Level 3 definitions for the TUNDRA and TAIGA Level 1 ecoregions. We leave the Level 3 ecoregion as "NOT FOUND" for these observations. Our goal in determining the Level 3 ecoregion was to utilize average fuel moisture and wind speed information, which are provided according to the Level 3 ecoregion. When we merge in this data we fill in fuel moisture and wind speed data with 0 values for any fire occurring in the TUNDRA or TAIGA Level 1 ecoregions.

1.3 Additional Features

We calculate additional features for all observations using the discovery date column in the original database, these are summarized here:

1. **disc_doy**: the day of the year (1-365) the fire was discovered.

2. **disc_dow**: the day of the week (0-6) the fire was discovered (0 corresponds to Monday and 6 corresponds to Sunday).
3. **disc_weekend**: 1 if the day the fire was discovered was Saturday or Sunday, 0 otherwise.
4. **disc_holiday**: 1 if the day the fire was discovered was a US federal holiday, 0 otherwise.
5. **disc_year**: the numeric year the fire was discovered.
6. **disc_month**: the numeric (1-12) month the fire was discovered.
7. **season_fall**: 1 if the month the fire was discovered was 9-11 inclusive, 0 otherwise.
8. **season_spring**: 1 if the month the fire was discovered was 3-5 inclusive, 0 otherwise.
9. **season_summer**: 1 if the month the fire was discovered was 6-8 inclusive, 0 otherwise.
10. **season_winter**: 1 if the month the fire was discovered was 12, 1, or 2, 0 otherwise.

These additional features are inspired by similar features used in Dr. Nagy's work and, where analogous, are calculated similarly [2].

1.4 Observation Subsets

The cleaned database has approximately 1.71 million observations. This is far too many to attempt any but the most simple models on the full data, and would prevent us from conducting much in the way of hyper-parameter tuning. We therefore, desire to subset the data in a reasonable manner to reduce the computational burden. We approach this with Task 1 in mind, which seeks to perform binary classification to predict the **human** response, which is 1 if the cause of the fire was attributable to human activity, and 0 otherwise.

Table 1: Count of Two-Class Observations Grouped by Level 1 Ecoregion

	Human	Count	Percent
10 NORTH AMERICAN DESERTS	0	57649	49%
	1	59353	51%
11 MEDITERRANEAN CALIFORNIA	0	3741	**3%**
	1	105680	**97%**
12 SOUTHERN SEMIARID HIGHLANDS	0	2299	21%
	1	8746	79%
13 TEMPERATE SIERRAS	0	28275	66%
	1	14779	34%
15 TROPICAL WET FORESTS	0	2200	29%
	1	5483	71%
2 TUNDRA*	0	584	82%
	1	128	18%
3 TAIGA	0	2489	48%
	1	2696	52%
5 NORTHERN FORESTS	0	2862	**4%**
	1	70323	**96%**
6 NORTHWESTERN FORESTED MOUNTAINS	0	111188	55%
	1	89801	45%
7 MARINE WEST COAST FOREST	0	1618	**6%**
	1	24025	**94%**
8 EASTERN TEMPERATE FORESTS	0	49079	**5%**
	1	899791	**95%**
9 GREAT PLAINS	0	16441	**10%**
	1	153680	**90%**

Table 1 shows the ratio of human to non-human fires after the data has been split by ecoregion. We find several troubling results in this table. First, the "2 TUNDRA" ecoregion is under sampled in the data, having significantly fewer observations than all other ecoregions, further the balance of classes seen in this ecoregion is borderline with human-related fires being a significant minority. We find similar trends (of human-related fires being a significant minority) in several other ecoregions. Considering we desire a balanced allocation among the response classes in our data, we decide to subset according to each task's goal rather than apply an initial sub setting (as Dr. Nagy does with a regional size threshold). This will allow us to ensure a balance of response classes exist in the data we use to train and test our models.

2 Task 1: Binary Classification

In this task we seek to develop models to predict the "human" response. This feature is 1 if the cause of the fire can be attributed to humans, and 0 otherwise. We use all other features (which could reasonably be known at the time of discovery) as predictors. This task is broken into two parts, with the first focused on interpretability, and the second focused on overall accuracy. The features used throughout this task are: 1: LATITUDE, 2: LONGITUDE, 3: disc_doy, 4: disc_weekend, 5: disc_holiday, 6: disc_month, 7: season_fall, 8: season_spring, 9: season_summer, 10: season_winter, 11:fm.mean, 12:Wind.mean. We drop the fm.mean and Wind.mean features for the TUNDRA and TAIGA ecoregions, because we do not have these data for these Level 1 ecoregions.

2.1 Data Sampling

We fit a single model to data from each individual Level 1 ecoregion, to indicate the importance of the ecoregion on the cause of the fires. Yet the number of observations in each ecoregion and the ratio between the response classes vary widely (see Table 1). Therefore we adopt the following random sampling scheme which attempts to 1) balance the response classes among each ecoregion, 2) balance the distribution of fire sizes, 3) reduce the number of observations for ease of computation.

Task 1 Data Sampling Approach

1. Calculate the fraction of data to drop to retain no more than 2000 observations, but not less than 10% of the original data.
2. Calculate the fire size for human and non-human started fires correspnding to this percentile of the data in an ecoregion.
3. Take the minimum of the two thresholds, and drop fires smaller than this threshold.
4. Randomly sample up to 2000 (or the total number remaining) from both response classes.

This method serves to eliminate the smallest fires observed in a way that attempts to balance the range of fire sizes between the two response classes, forces the response classes to be as balanced as possible (in many cases exactly balanced), and reduces the remaining data to a level which allows hyper-parameter tuning. Table 2 shows the ending distribution of data for each ecoregion. We note that the "TUNDRA" and "MARINE WEST COAST FOREST" ecoregions have suspect sampling (reduced count of observations and imbalanced response classes) and we would expect these regions to potentially have issues in modeling. Further, while the minimum fire size between response classes has been equalized there is wide variation in mean and maximum fire size. This is not troubling for task 1 (since this may just be a distinction between human and non-human relation fires), but the very small minimum size in the MARINE WEST COAST forest could be an issue since the cause of very small fires is likely to show higher randomness than larger fires.

Table 2: Data Subset Summary for Task 1 Models

	Human	Count	Percent	Min Size	Mean Size	Max Size
10 NORTH AMERICAN DESERTS	0	2000	50%	23.0	3415.2	558198.3
	1	2000	50%	23.0	947.3	61929.1
11 MEDITERRANEAN CALIFORNIA	0	2000	50%	0.1	196.1	162818.0
	1	2000	50%	0.1	40.5	25156.0
12 SOUTHERN SEMIARID HIGHLANDS	0	2000	50%	0.1	287.4	95000.0
	1	2000	50%	0.1	261.0	222954.0
13 TEMPERATE SIERRAS	0	2000	50%	2.0	824.3	297845.0
	1	2000	50%	2.0	717.4	538049.0
15 TROPICAL WET FORESTS	0	2000	50%	0.1	485.3	68295.0
	1	2000	50%	0.1	269.2	158000.0
2 TUNDRA	0	584	82%	0.1	2667.6	256734.1
	1	128	18%	0.1	542.4	21575.0
3 TAIGA	0	2000	50%	0.1	9799.6	606945.0
	1	2000	50%	0.1	294.7	189688.0
5 NORTHERN FORESTS	0	2000	50%	0.1	98.5	92682.0
	1	2000	50%	0.1	6.8	2377.0
6 NORTHWESTERN FORESTED MOUNTAINS	0	2000	50%	3.0	1539.9	172135.0
	1	2000	50%	3.0	549.8	83323.0
7 MARINE WEST COAST FOREST	0	1618	45%	0.01	277.9	56413.0
	1	2000	55%	0.01	26.9	37336.0
8 EASTERN TEMPERATE FORESTS	0	2000	50%	17.0	190.2	45294.0
	1	2000	50%	17.0	89.9	14626.0
9 GREAT PLAINS	0	2000	50%	40.0	1285.2	158308.0
	1	2000	50%	40.0	589.5	220000.0

2.2 Task 1: Predict Fire Cause - Focus on Interpretability

We fit two types of binary classification models in this portion of Task 1, selected for their ease of interpretability: decision tree classifier and logistic regression models. All models are fit on 70% of the subset data from each ecoregion, with 30% of the data held out as a test set. We use this held-out data to report model accuracy.

2.2.1 Decision Tree Classifier

We fit a decision tree classifier model on data from each Level 1 ecoregion. To reduce the tendency of this model to over-fit the training set, and to improve model interpretability we limit the depth of each tree to 3.

Table 3: Task 1 Part a Decision Tree Classifier Models

	Test Acc	1st Var	2nd Var	3rd Var
8 EASTERN TEMPERATE FORESTS	82.2%	disc_doy	fm.mean	season_summer
6 NORTHWESTERN FORESTED MOUNTAINS	73.4%	LONGITUDE	disc_doy	season_summer
9 GREAT PLAINS	81.8%	disc_doy	LONGITUDE	season_summer
10 NORTH AMERICAN DESERTS	70.6%	LATITUDE	disc_doy	season_summer
11 MEDITERRANEAN CALIFORNIA	74.5%	LONGITUDE	disc_doy	fm.mean
5 NORTHERN FORESTS	76.3%	LATITUDE	LONGITUDE	disc_doy
13 TEMPERATE SIERRAS	76.8%	LONGITUDE	disc_doy	season_summer
7 MARINE WEST COAST FOREST	66.7%	Wind.mean	LONGITUDE	disc_doy
12 SOUTHERN SEMIARID HIGHLANDS	81.1%	disc_doy	season_summer	fm.mean
15 TROPICAL WET FORESTS	81.1%	LATITUDE	disc_doy	season_summer
3 TAIGA*	88.7%	disc_doy	LATITUDE	LONGITUDE
2 TUNDRA*	89.7%	disc_doy	LATITUDE	season_summer

*These ecoregions do not have average fuel moisture and wind speed data.

The test accuracy and top 3 most important features of each model are shown in Table 3. We note that the discovery day of year, latitude, longitude, and season_summer commonly appear in the model's most important features with the fuel moisture and wind speed only appearing in three models. Surprisingly, the model in ecoregion 2 TUNDRA is quite accurate even though this region was significantly under-sampled when compared to the other regions. We plot the resulting decision tree models in Appendix A, but include the model from ecoregion 8 (the ecoregion with the greatest number of total observations) in Figure 3.

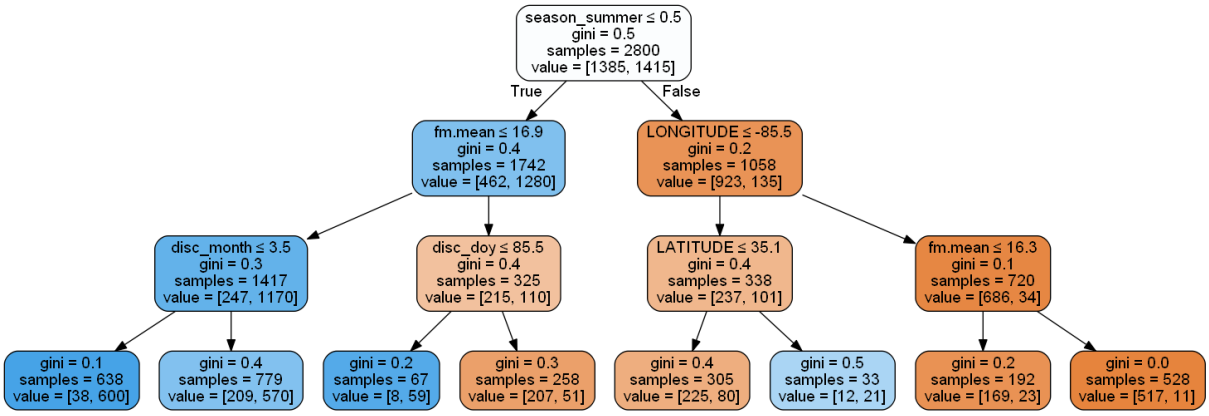


Figure 3: Example Decision Tree Classifier - Ecoregion 8, Blue: human caused, Orange: non-human caused

Note the spatial distribution of observation in ecoregion 8 (the most prevalent ecoregion in the data) in Figure 4. The presence of LATITUDE and LONGITUDE in the decision tree models is consistent with the large area this ecoregion (and several of the Level 1 ecoregions) covers. Inclusion of smaller ecoregions (Level 2 or Level 3) might benefit this analysis to more naturally divide the regions, though we do not include these in this analysis (other than the average fuel moisture and wind speed data) so as not to over-complicate model intrepretation.

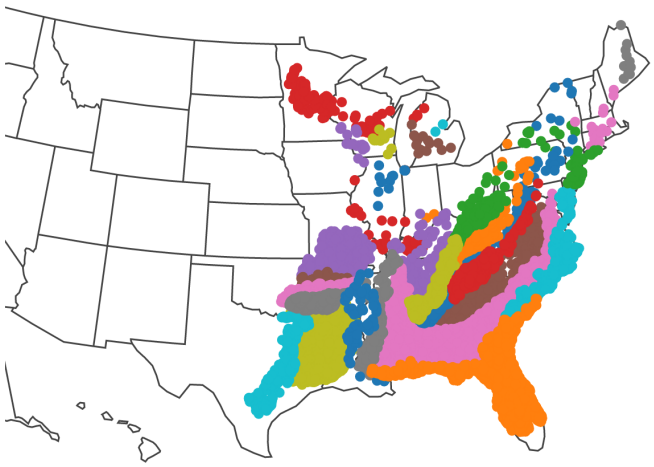


Figure 4: Level 1 Ecoregion 8, Spatial Distribution of Observations Colored by Level 3 Ecoregion

2.2.2 Logistic Regression

We fit a logistic regression model on data from each Level 1 ecoregion. We impose an l2 penalty on the coefficients in the model to discourage model overfitting. We then perform hyper-parameter tuning over 10 values of this l2 penalty, using 10 fold cross validation to determine the best value of the l2 penalty coefficient.

The sign of the relationship between each of the features is shown in Table 4, with the numbered features corresponding to 1: LATITUDE, 2: LONGITUDE, 3: disc_doy, 4: disc_weekend, 5: disc_holiday, 6: disc_month, 7: season_fall, 8: season_spring, 9: season_summer, 10: season_winter, 11:fm.mean, 12:Wind.mean. Interestingly, the sign of the relationship between the features and response is not consistent for any one feature across all ecoregions, with the most consistent features being (positive) disc_weekend, season_winter, disc_holiday, and fm.mean; and (negative) disc_doy, and Wind.speed. This result supports our decision to split data by ecoregion, but the logistic regression models fit have somewhat poor accuracy, and the best regularization coefficient varies widely (six orders of magnitude between the minimum to the maximum). These aspects of the models are somewhat troubling and are not consistent with a good fit to the data.

Table 4: Task 1 Part a Logistic Regression Models

Ecoregion	Test Acc	Best C	1	2	3	4	5	6	7	8	9	10	11	12
8	82.4%	2.78e+00	+	-	+	+	+	-	+	-	-	+	-	-
6	72.8%	2.15e+01	-	-	+	+	+	-	-	-	-	+	-	+
9	78.7%	2.78e+00	+	+	-	-	+	+	+	+	-	+	+	+
10	66.8%	3.59e-01	+	-	-	+	+	-	+	-	-	+	-	-
11	75.5%	1.67e+02	-	-	-	+	+	+	-	-	-	-	+	+
5	73.5%	2.15e+01	-	-	+	+	+	-	+	+	+	+	+	-
13	75.2%	1.67e+02	+	-	-	+	+	-	-	-	-	+	+	-
7	59.6%	4.64e-02	-	-	+	+	+	-	-	+	-	+	+	-
12	81.1%	2.78e+00	-	-	-	+	+	+	+	-	-	+	+	-
15	79.6%	2.78e+00	-	+	-	+	-	+	+	-	-	+	+	+
3*	76.9%	1.00e+04	-	+	-	+	-	+	+	+	+	+		
2*	87.9%	3.59e-01	-	-	-	-	-	+	+	+	-	+		

*These ecoregions do not have average fuel moisture and wind speed data.

Overall, we take these results to mean that the data is not well modeled by the underlying assumptions of logistic regression. Figure 5 shows ROC curves for each of the logistic regression models fit. We note that the model fit to data from ecoregion 7 MARINE WEST COAST FOREST has an especially poor fit, potentially because we have unbalanced response classes in this ecoregion, with the number of non-human fire observations in a slight minority. Additionally observations from this ecoregion include very small fire sizes and the ecoregions the spans an especially diverse geographic area including the US West coast and Alaska. We may be able to improve the accuracy of the model on this region by separating this ecoregion into the Level 2 or 3 ecoregions, which cover a smaller geographic area.

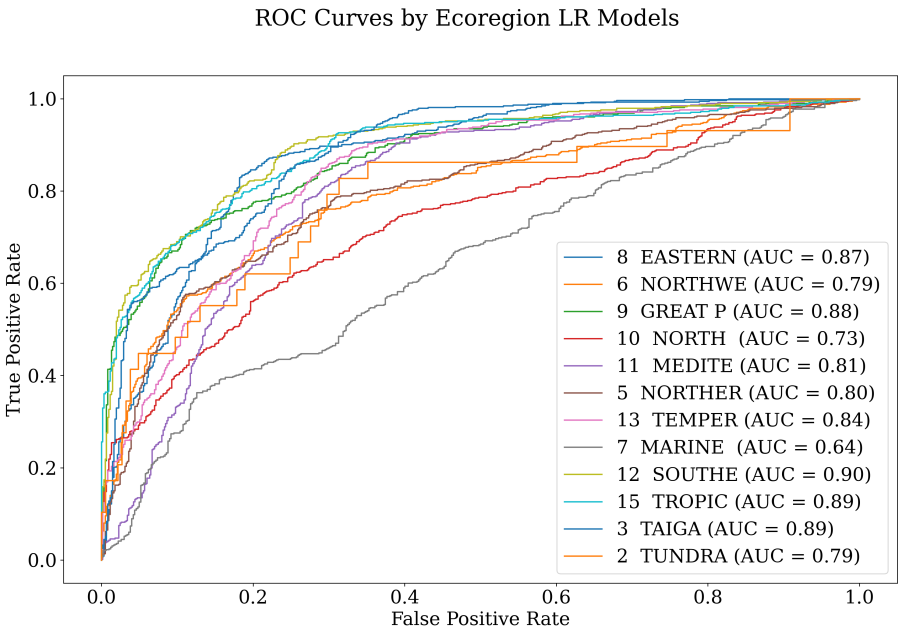


Figure 5: ROC Curves for all Task 1 Logistic Regression Models

2.2.3 Task 1 Interpretability Conclusions

We summarize our conclusions from modeling the binary "human" response using decision tree classifier and logistic regression models here.

1. **Separation of data by Level 1 ecoregion:** The logistic regression models show wide variation in the relation between each feature and the change in the log odds of the response class. We interpret this to mean the relationship between the feature and the response is specific to each Level 1 ecoregion.

2. **Important features:** The decision tree classifier models indicate the LATITUDE, LONGITUDE, disc_doy, and season_summer features are commonly among the most important features for many ecoregions. The presence of the LATITUDE and LONGITUDE features we find somewhat concerning, since it might indicate the need for a more natural spatial separation of observations (such as the inclusion of the Level 2 or Level 3 ecoregions).

3. **Features related to human-caused fires:** The logistic regression models indicate the disc_holiday, disc_weekend, season_winter, and fm.mean features are commonly associated with an increase in the log odds of a human-caused fire.

4. **Features related to non-human-caused fires:** While these same models indicate disc_doy, season_summer, and Wind.speed are are commonly associated with a decrease in the log odds of a human caused fire (and therefore are more associated with a lightning-caused fire).

5. **Overall impression of models:** Generally the models only achieve mediocre accuracy. We take this to indicate that there are likely lurking variables. These may include fuel moisture and wind speed information specific to the time period of each observation (we currently only include this information averaged over the entire time period).

2.3 Task 1: Predict Fire Cause - Focus on Accuracy

In this section we fit more flexible but less interpretable models, to include Linear Discriminant Analysis, Support Vector Machine, Random Forest, and k-Nearest Neighbors. Finally, we extend the data features using all previous models and fit a Multi-layer Perceptron model as a kind of ensemble classification. While there are some opportunities for inference in these models, we instead focus on developing accurate models. We impose the same data subset technique used in the previous portion of task 1, and continue to segregate data by Level 1 ecoregion prior to fitting each model. We provide details of hyper-parameter tuning here:

1. **Linear Discriminant Analysis:** We impose a penalty on the number of components and perform 10-fold cross validation to select best value.

2. **Support Vector Machine:** We iterate over using a linear, radial bias, second and third degree polynomials as well as multiple values of the C and gamma parameters. We use 5-fold cross validation to select the best parameter values for each ecoregion.

3. **Random Forest:** We limit the maximum depth of each tree in the ensemble to 2, and fit 100 individual trees.

4. **K-Nearest Neighbors:** We iterate over a small set of possible k values, and use 5-fold cross validation to select the best value.

5. **Multi-layer Perceptron:** We use a somewhat simplistic network architecture (to limit complexity) and add dropout layers which have an analogous effect to an l1 penalty. We iterate through several hidden layer activation functions, but primarily focus on the rectified linear unit and sigmoid activation functions while tuning the dropout layer rate.

2.3.1 Multi-layer Perceptron Classifier Model Details

We employ the keras Sequential model to train a feed forward neural network on this binary classification task. We merge all predictions from the previous models with the 12 features used in all other models. The input layer of this model has 18 features, developing an ensemble model since this classification model gets the predictions from all other models as input features. The network used has two hidden layers with 32 neurons each. We place a dropout layer after each densely-connected layer to provide regularization and help prevent over-fitting. These layers force a set percentage of connections between layers to "dropout" or have their activation weights set to zero. These layers act somewhat analogously to an l1 penalty in lasso regularization. We use a single neuron output layer with sigmoid activation function for all models constructed (consistent with binary classification). Generally, we find the rectified linear unit activation function and a 10% dropout rate to have the best performance, but we present the best accuracy seen among all models trained during hyper-parameter tuning. Models were fit using the binary cross-entropy loss function (consistent with binary classification), and trained with the "adam" optimizer (which is a computationally efficient stochastic gradient descent method).

Table 5: Task 1 Model Test Accuracy Summary

	DT	LR	LDA	SVM	RF	kNN	MLP**
8 EASTERN TEMPERATE FORESTS	82.2%	82.4%	81.8%	81.7%	82.5%	83.2%	**83.6%**
6 NW FORESTED MOUNTAINS	70.8%	70.2%	70.1%	72.8%	70.1%	**74.3%**	74.2%
9 GREAT PLAINS	81.8%	78.7%	78.0%	84.5%	81.2%	83.7%	**85.0%**
10 NORTH AMERICAN DESERTS	70.6%	66.8%	67.0%	74.1%	71.1%	74.3%	**74.8%**
11 MEDITERRANEAN CALIFORNIA	74.5%	75.5%	73.2%	**80.8%**	74.5%	77.6%	80.6%
5 NORTHERN FORESTS	76.3%	73.5%	72.2%	**77.8%**	71.5%	76.8%	77.7%
13 TEMPERATE SIERRAS	76.8%	75.2%	75.2%	80.7%	73.7%	79.8%	**81.6%**
7 MARINE WEST COAST FOREST	64.2%	62.1%	61.4%	72.4%	60.2%	70.3%	**73.7%**
12 SOUTHERN SEMIARID HIGHLANDS	81.1%	81.1%	80.6%	83.2%	79.9%	83.8%	**84.8%**
15 TROPICAL WET FORESTS	81.1%	79.6%	79.7%	84.6%	82.4%	83.9%	**85.8%**
3 TAIGA*	88.7%	76.9%	74.9%	87.3%	88.5%	88.5%	**89.2%**
2 TUNDRA*	89.7%	87.9%	86.4%	86.4%	88.3%	88.8%	**90.7%**

*These ecoregions do not have average fuel moisture and wind speed data.
**MLP models ensemble all previous models with other features.

2.3.2 Task 1 Accuracy Conclusions

We summarize our conclusions from modeling the binary "human" response using all previous model here.

1. **Benefits of the Ensemble Model:** We see a modest improvement (on average less than 1%) in test accuracy across all ecoregions for the ensemble MLP models, though perhaps not as much improvement as we would hope for this highly flexible model, given six predictions an 12 other features on the binary output. There are several exceptions where the SVM or kNN models perform best. This may be a fluke of the specific train / test split we are imposing or indicate that the underlying assumptions of these models are most consistent with the true realtion between the features and the response.. Generally though, we have validated the concept of using an MLP model as an ensemble predictor, and shown we can successfully construct one using the Keras framework.

2. **Overall impression of models:** Generally the models only achieve mediocre accuracy. We take this to indicate that there are likely lurking variables.

The concept of a ensemble model is predicated on independent predictions on the response, which may not be applicable in this study because we sample the same training and test sets for each ecoregion among all models. But we do at least see a small improvement in test accuracy across most ecoregion models.

3 Task 2: Multi-class Classification

3.1 Data Sampling

In this section, rather than attempting to predict whether a fire is either human or not, we will more specifically aim to identify the cause of the fire. Now, there will be twelve difference classes of a fire for which we will predict. However, as with the previous section, we look at the distribution of the fires within each region by cause, there are clear categories exceeding others. Below gives us an idea of what percent of the fires in a region belong to the twelve different causes of fires. The largest values in each row, will give us a foundation for each region as to what we hope for our model to perform better than.

Table 6: Percentage of Fires by Cause in each Region

	Miscellaneous	Lightning	Debris Burning	Campfire	Equipment Use	Arson	Children	Railroad	Smoking	Powerline	Structure	Fireworks
2 TUNDRA	3.65%	82.02%	4.78%	4.92%	0.98%	0.70%	1.26%	0.00%	1.26%	0.00%	0.28%	0.14%
3 TAIGA	10.13%	48.00%	14.89%	7.27%	2.97%	3.88%	3.90%	0.10%	1.95%	3.86%	2.43%	0.64%
5 NORTHERN FORESTS	24.67%	3.91%	26.06%	6.04%	7.23%	17.15%	7.00%	2.52%	3.10%	1.49%	0.30%	0.53%
6 NORTHWESTERN FORESTED MTNS	10.92%	55.32%	6.80%	12.30%	4.32%	4.71%	1.42%	0.57%	2.42%	0.46%	0.15%	0.62%
7 MARINE WEST COAST FOREST	24.12%	6.31%	25.05%	13.78%	12.86%	5.38%	4.94%	0.46%	4.13%	1.08%	0.83%	1.07%
8 EASTERN TEMPERATE FORESTS	17.07%	5.17%	34.51%	2.69%	7.06%	22.82%	3.65%	2.69%	3.29%	0.62%	0.17%	0.24%
9 GREAT PLAINS	28.72%	9.66%	23.46%	1.63%	12.01%	10.44%	4.14%	1.01%	2.98%	2.39%	0.60%	2.95%
10 NORTH AMERICAN DESERTS	16.72%	49.27%	8.87%	4.13%	7.06%	5.26%	2.35%	1.15%	2.50%	0.83%	0.16%	1.70%
11 MEDITERRANEAN CALIFORNIA	35.09%	3.42%	7.82%	3.62%	29.23%	11.33%	4.69%	0.34%	3.54%	0.83%	0.03%	0.07%
12 SOUTHERN SEMIARID HIGHLANDS	22.74%	20.81%	6.44%	8.85%	4.06%	22.76%	9.27%	0.42%	4.07%	0.06%	0.05%	0.48%
13 TEMPERATE SIERRAS	10.46%	65.67%	3.13%	11.00%	2.32%	2.99%	1.67%	0.16%	2.19%	0.16%	0.10%	0.16%
15 TROPICAL WET FORESTS	13.74%	28.63%	6.72%	2.06%	12.35%	14.89%	4.31%	15.74%	1.11%	0.25%	0.07%	0.14%

Causes greater than 20%

Since this discrepancy exists in the number of fires for each cause, we train over the entire data set for each region. However, we will use a stratified 3–fold cross-validation, which will attempt to distribute the observation so that they're dispersed as equally as possible by these causes. Parameters for each method will be tuned over two trials. Moving forward, due to the extreme imbalance within the TUNDRA region, we will forgo creating a model for it. In order to predict the cause of a fire, we will use the following variables disregarding `fm.mean` and `Wind.mean` in the TAIGA region, due to all values being zero.

Table 7: Features Used

Variable Name	`FIRE_SIZE`	`LATITUDE`	`LONGITUDE`	`season`	`disc_weekend`	`disc_holiday`	`eco3`	`fm.mean`	`Wind.mean`
Variable Type	continuous	continuous	continuous	categorical	binary	binary	categorical	continuous	continuous
Variable Range	$[1e{-}5, 6.17e5]$	$[-90, 90]$	$[-180, 180]$	{F, Sp, S, W}	$\{0, 1\}$	$\{0, 1\}$	45 regions	$[0.00, 18.01]$	$[0.00, 4.95]$

In the following sections we will use these regions and variables in order to predict whether a fires is one of the twelve causes above. We will start by looking into methods that are more interpretable before progressing on to more accurate models. Initially, for each model, we will tune their associated hyperparameters to find which best suit our data. Following this, we will take all of the models and based on accuracy, provide the most visual and accurate models for each region.

3.2 Task 2: Predict Multi-class Fire Cause - Focus on Interpretability

We begin first by looking at two methods, Decision Tree and Logistic Regression, that provide us with an idea of how each region is classifying fires (which cause) as well as which features are influencing each region. We will tune the parameter `C` and look to identify the best `penalty`, L1 or L2, for Logistic Regression. We see the resulting outcome in Table 8a. As for the Decision Tree, we limit the `max_depth` to three, and let cross-validation decide whether to use entropy or a gini index for the `criterion`. These results are found in Table 8b.

Table 8: Hyperparameter Tuning (Interpretability)

	C	penalty
3 TAIGA	0.01	L1
5 NORTHERN FORESTS	0.01	L1
6 NORTHWESTERN FORESTED MTNS	0.01	L1
7 MARINE WEST COAST FOREST	0.01	L1
8 EASTERN TEMPERATE FORESTS	1.00	L1
9 GREAT PLAINS	0.01	L1
10 NORTH AMERICAN DESERTS	0.01	L1
11 MEDITERRANEAN CALIFORNIA	0.01	L1
12 SOUTHERN SEMIARID HIGHLANDS	0.10	L1
13 TEMPERATE SIERRAS	0.01	L1
15 TROPICAL WET FORESTS	0.01	L1

(a) Logistic Regression Results

	criterion	max_depth
3 TAIGA	gini	3
5 NORTHERN FORESTS	entropy	3
6 NORTHWESTERN FORESTED MTNS	entropy	3
7 MARINE WEST COAST FOREST	entropy	3
8 EASTERN TEMPERATE FORESTS	entropy	3
9 GREAT PLAINS	gini	3
10 NORTH AMERICAN DESERTS	gini	3
11 MEDITERRANEAN CALIFORNIA	gini	3
12 SOUTHERN SEMIARID HIGHLANDS	gini	3
13 TEMPERATE SIERRAS	entropy	3
15 TROPICAL WET FORESTS	gini	3

(b) Decision Tree Results

We utilize the nine features found in Table 7 for both of these models. However, the Logistic Regression model, using a one-vs-all approach, provides us information through the coefficients as to how each variables impacts a cause. By isolating a class and then comparing it to the other classes we identify the sign of each coefficient in order to determine whether larger values of the variable imply a cause to be more or less (positive or negative) likely. These coefficient signs are found in Appendix B.1. Note that only the sign of these are provided and not the magnitude. Therefore, we only conclude on the how a variable impacts a cause, not how much. Within each table are the causes isolated against the others (each row). Every column represents a variable and within a cell the specific sign. In many of the regions, we tend to see causes of fires to have similar signs associated with particular variables. This implies that no matter which region the fire is in, the variable impacts those causes the same.

Using the Decision Tree for each region we get an idea of how much importance each variable influenced classification. These plots are seen in Appendix B.2. These values are normalized impacts on the `criterion` when a feature is factored in. Therefore, the larger this value, the more important a variable is. Now, in order to get a visualization of how each region decides the cause of a fire, we plot the best estimated decision tree for each region, and color it according to the cause of fire. As we see from Table 8b each region maximized its accuracy by using a depth of three. However, this doesn't account for trimming the tree and thus, possibly redundant branches. The plotted tree for each region is found in Appendix B.3. Not including the Miscellaneous class, the most frequently appearing colors/causes are blue/Lightning, orange/Arson, and green/Debris Burning. This is what we would expect many of the fires to be classified as, due to them making up the majority of fires.

For each internal node of a tree, a splitting variable is listed first (this line isn't included in terminal/leaf nodes), followed by the resulting value for the `criterion`. The number of samples before splitting is listed next, before a list is providing with how many samples are in each class. Then, the last line of the node tells what the observation would be classified if it didn't go further down into the tree. Similarly, and as mentioned before, the color of the node represents what an observation would be classified as if it were to stop at that node. Furthermore, the darkness of the node illustrates how pure the node is for that classification.

These methods help us better visualize what the models are doing when making their decisions. However, by gaining in inference, we sacrifice predictive accuracy. In the next section we look to keen in on this, while losing out on the ability to deduce why the model makes its predictions.

3.3 Task 2: Predict Multi-class Fire Cause - Focus on Accuracy

Beginning with KNN, an easy, fast method, we look to tune the parameter K or `n_neighbors`. For each region, the total number of observations are found, then a maximum of the square root of this total is set as the most neighbors a region uses. This is done to once again accommodate the inconsistencies found in the number of observations within each region. The results for each region are given in Table 9a. For the Random Forest, we relax the restriction on the `max_depth` and restrict the `max_features` used when splitting. Again, we find a suitable scoring `criterion` for each region. Lastly, we identify how many trees or `n_estimators` are sufficient in giving us the best accuracy. This aggregation of many trees is expensive but will decrease the variance we get from our Decision Tree implementation. Since we have already seen that this data has odd behavior, we want a method that will account for this in future use. The outcome for each regions Random Forest is found in Table 9b

Table 9: Hyperparameter Tuning (Accuracy)

	n_neighbors
3 TAIGA	43
5 NORTHERN FORESTS	42
6 NORTHWESTERN FORESTED MTNS	67
7 MARINE WEST COAST FOREST	49
8 EASTERN TEMPERATE FORESTS	142
9 GREAT PLAINS	62
10 NORTH AMERICAN DESERTS	52
11 MEDITERRANEAN CALIFORNIA	50
12 SOUTHERN SEMIARID HIGHLANDS	48
13 TEMPERATE SIERRAS	33
15 TROPICAL WET FORESTS	16

(a) KNN Results

	criterion	max_depth	max_features	n_estimators
3 TAIGA	entropy	10	$\sqrt{n}$	500
5 NORTHERN FORESTS	gini	15	$\log_2(n)$	500
6 NORTHWESTERN FORESTED MTNS	gini	20	$\sqrt{n}$	500
7 MARINE WEST COAST FOREST	entropy	10	$\sqrt{n}$	500
8 EASTERN TEMPERATE FORESTS	gini	20	$\log_2(n)$	500
9 GREAT PLAINS	entropy	15	$\log_2(n)$	500
10 NORTH AMERICAN DESERTS	gini	15	$\sqrt{n}$	300
11 MEDITERRANEAN CALIFORNIA	gini	15	$\sqrt{n}$	300
12 SOUTHERN SEMIARID HIGHLANDS	entropy	10	$\log_2(n)$	300
13 TEMPERATE SIERRAS	entropy	15	$\log_2(n)$	300
15 TROPICAL WET FORESTS	gini	10	$\log_2(n)$	300

(b) Random Forest Results

Comparing the results of the Random Forest to the percent of observations in each region we get a sense that the depth it is choosing correlates to the percent of the fires in each. The Random Forest will reduce the variance we get in the Decision Tree previously thereby avoiding over-fitting. We expect this to provide us with a more accurate model. The reduction of the number of variables we use when splitting will look to isolate out those features we identified previously as most important. This will lead to less of our trees using the same splitting pattern. Being less robust and more flexible seems to be the requirement for this data set.

Since we have more than two classes in this case we now look to fit the data to a Linear and Quadratic Discriminant Analysis as well as a Linear Support Vector Machine. From the first two, we hope to gain information about the variability within the underlying distribution of the data. A regularization parameter is tuned in each case. For LDA and QDA, this improves the estimates of the covariance matrix/matrices while in SVM it acts as the flexibility of the margins. The resulting values for each region are found on the next page.

	shrinkage
3 TAIGA	0.0010
5 NORTHERN FORESTS	0.0127
6 NORTHWESTERN FORESTED MTNS	0.0010
7 MARINE WEST COAST FOREST	0.0089
8 EASTERN TEMPERATE FORESTS	0.0010
9 GREAT PLAINS	1.0000
10 NORTH AMERICAN DESERTS	0.0010
11 MEDITERRANEAN CALIFORNIA	0.0010
12 SOUTHERN SEMIARID HIGHLANDS	0.0014
13 TEMPERATE SIERRAS	0.0010
15 TROPICAL WET FORESTS	0.0010

(c) Linear Discriminant Analysis Results

	reg_param
3 TAIGA	0.950
5 NORTHERN FORESTS	0.575
6 NORTHWESTERN FORESTED MTNS	0.000
7 MARINE WEST COAST FOREST	0.825
8 EASTERN TEMPERATE FORESTS	0.025
9 GREAT PLAINS	0.025
10 NORTH AMERICAN DESERTS	0.000
11 MEDITERRANEAN CALIFORNIA	0.725
12 SOUTHERN SEMIARID HIGHLANDS	0.075
13 TEMPERATE SIERRAS	0.975
15 TROPICAL WET FORESTS	0.975

(d) Quadratic Discriminant Analysis Results

	C
3 TAIGA	0.100
5 NORTHERN FORESTS	0.001
6 NORTHWESTERN FORESTED MTNS	0.001
7 MARINE WEST COAST FOREST	0.001
8 EASTERN TEMPERATE FORESTS	0.001
9 GREAT PLAINS	0.001
10 NORTH AMERICAN DESERTS	0.001
11 MEDITERRANEAN CALIFORNIA	0.001
12 SOUTHERN SEMIARID HIGHLANDS	0.001
13 TEMPERATE SIERRAS	0.001
15 TROPICAL WET FORESTS	0.001

(e) Linear Support Vector Machine Results

3.4 Task 2: Method Results

With parameters tuned for each model, we now determine which provides us with the most accurate results. For each region, a model is selected from each of the previous two sections in order to provide us with the most interpretatino and accuracy. To accomplish this, we set our parameters accordingly and run a 5-fold cross-validation once again on the data. We score each fold and average the resulting accuracy. This is done for each method; the results are recorded below in Table 10.

Table 10: Accuracy Result by Region and Method

	Natural (Basis)	Logistic Regression	Decision Tree	K-Nearest Neighbors	Random Forest	Linear Discriminant Analysis	Quadratic Discriminant Analysis	Support Vector Machine
3 TAIGA	48.00%	48.08%	58.23%	59.36%	61.76%	48.00%	36.27%	40.76%
5 NORTHERN FORESTS	26.06%	26.19%	38.55%	35.89%	43.09%	33.05%	17.02%	33.28%
6 NORTHWESTERN FORESTED MTNS	55.32%	55.32%	57.95%	61.58%	64.13%	55.32%	47.85%	54.92%
7 MARINE WEST COAST FOREST	25.05%	25.22%	33.17%	38.74%	41.91%	29.82%	13.71%	30.96%
8 EASTERN TEMPERATE FORESTS	34.51%	34.63%	40.20%	46.86%	50.30%	37.63%	33.71%	38.93%
9 GREAT PLAINS	28.72%	28.72%	35.86%	43.80%	49.23%	28.78%	27.88%	33.68%
10 NORTH AMERICAN DESERTS	49.27%	49.27%	56.29%	60.96%	64.17%	49.27%	39.59%	52.36%
11 MEDITERRANEAN CALIFORNIA	35.09%	35.09%	37.37%	39.16%	42.32%	35.06%	31.70%	35.95%
12 SOUTHERN SEMIARID HIGHLANDS	22.76%	24.73%	47.85%	47.30%	51.96%	33.57%	25.82%	42.47%
13 TEMPERATE SIERRAS	65.67%	65.67%	67.31%	69.37%	71.07%	65.67%	47.36%	66.30%
15 TROPICAL WET FORESTS	28.63%	28.63%	39.45%	39.54%	48.49%	28.63%	8.92%	30.73%

Most Interpretable Most Accurate

From above, a blue cell represents the choice between Logistic Regression and a Decision Tree. As we see, each region was better predicted using a Decision Tree. This indicates that the division of the data is more complex than a simple hyperplane. As for the orange cell, this indicates the method that performed the best in comparison to the rest of the methods. Once again, a single method, Random Forest, was the leader for all regions. This further supports the prior argument that the data requires more in depth splitting in order to pinpoint a fires cause.

Overall, comparing these values to the largest class probability within each region (basis) we see that each method does perform better for all. The separation isn't large, however, due to their being many classes we could say this is actually a huge improvement. On average, the improvement is between 5-30%. Given that we are also able to interpret what is causing each fire, we could implement intervention with hopes to eliminate future fires.

3.5 Task 2: Conclusion

The results gained from this attempt at predicting the specific cause of a fire were beneficial, however we definitely would hope to improve upon the accuracy. During this process, variables such as `eco3` tended to be identified as unimportant. Other variables that follow this pattern may have hindered the ability for the methods to accurately learn from the data. Therefore, making adjustments for these downfalls we hope would provide a boost to our prediction accuracy.

On the other hand, since there is such high variability in the ways fires occur, we are also concerned the data either doesn't provide enough information, or has given us as much insight as possible. Incorporating other features into the data set would be another additive making identification of a fire less error prone.

4 Task 3: Fire-Size Inference

We wish to perform inference on factors which are related to the size of a fire. We do this by considering various regression approaches in which the size of the fire is the response and all relevant inputs are potential features. We train separate models for each of the various level 1 ecoregions which is important for determining what a large fire is and because we expect the possibility for different effects for various ecoregions.

4.1 Data Sampling

Like above, we primarily focus on estimation for large fires. We follow similar thresholds as in [3] where we only consider the largest 10% of fires for a given level 1 ecoregion. We do this for a couple reasons. Scientifically, large fires are of more interest because they account for a large percentage of the total burned area and they require more effort and resources to suppress. For a statistical inference standpoint, the scale of all fires is very spread out; some fire sizes are on the order of 0.1, whereas the maximum recorded datapoint has a size of 606945. This presents some challenges for statistical modeling due to the presence of outliers. Thus, we only focus on the largest 10% of fires for each level 1 ecoregion.

4.2 Multiple Linear Regression

We begin by trying a multiple linear regression. We first introduce the following notation for ease of model description. We note that we use 3 different auxiliary variables to represent the different possible seasons (with fall as the baseline). These variables are described above but these descriptions are repeated for the new notation. We also note that we try regression of the natural logarithm of the fire size rather than the fire size in order to try and bring down the higher values of fire size. We note that Y over $\log Y$ gives very poor results due to the outliers.

Table 11: Linear Regression Model Variable Definitions

Variable name	Description	Potential values
$\log Y$	Natural logarithm of fire size	Continuous in interval $(-\infty, \infty)$
X_1	Cause of fire	1 if the fire was caused by a human and 0 otherwise (lightning)
X_2	Auxiliary variable for spring	1 if the season is spring and 0 otherwise
X_3	Auxiliary variable for summer	1 if the season is summer and 0 otherwise
X_4	Auxiliary variable for winter	1 if the season is winter and 0 otherwise
X_5	Discovery day of week	Values between 0 and 6
X_6	Discovery holiday	1 if the discovery date was a US federal holiday and 0 otherwise
X_7	Discovery year	The year the fire was discovered
X_8	Mean fuel moisture	Continuous in interval $[0, \infty)$
X_9	Mean wind speed	Continuous in interval $[0, \infty)$
X_{10}	Longitude	Continuous in interval $[-180, 180]$
X_{11}	Latitude	Continuous in interval $[-90, 90]$

Thus, we consider the multiple regression model defined by

$$\log Y = \beta_0 + \sum_{i=1}^{11} \beta_i X_i + \varepsilon$$

where ε is a random error term which we assume is independent of X_i and has mean 0. For each level 1 ecoregion, we randomly pick 80% of the remaining large fires as training data and the remaining is used as test data. Note that if a coefficient does not appear (indicated with a dash), it was unable to be estimated due to rank deficiency or insufficient data (for example a factor did not appear in the training data). We fit this model to each of the different data sets for each level 1 ecoregion and report the coefficient values (standard errors in parentheses), number of training observations, the fire size threshold M, test MSE, adjusted R^2, and F-statistic, and p-value for the F-statistic. Any values which are significant at the 5% level are in **bold**.
We also use the following level 1 ecoregion relationship defined in table 12. The number in the first column of each of the following tables refers to the corresponding ecoregion.

Table 12: Ecoregion mapping

2	TUNDRA
3	TAIGA
5	NORTHERN FORESTS
6	NORTHWESTERN FORESTED MOUNTAINS
7	MARINE WEST COAST FOREST
8	EASTERN TEMPERATE FORESTS
9	GREAT PLAINS
10	NORTH AMERICAN DESERTS
11	MEDITERANEAN CALIFORNIA
12	SOUTHERN SEMIARID HIGHLANDS
13	TEMPERATE SIERRAS
15	TROPICAL WET FORESTS

Table 13: Multiple Regression Coefficient Estimates

Ecoregion	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}	β_{11}
2	17.1 (39.0)	**-2.0** (1.0)	-1.0 (1.1)	**-2.5** (1.2)	-0.5 (1.2)	-0.08 (0.06)	-1.7 (1.0)	-0.005 (0.02)	—	—	-0.04 (0.04)	-0.02 (0.04)
3	-8.5 (18.7)	0.3 (0.4)	—	-0.3 (0.2)	—	0.02 (0.03)	**-1.0** (0.4)	0.01	—	—	0.03 (0.01)	-0.02 (0.05)
5	**12.7** (5.6)	**-0.4** (0.09)	-0.1 (0.06)	**-0.2** (0.08)	0.1 (0.2)	0.009 (0.008)	**-0.4** (0.1)	0.003 (0.002)	**-0.5** (0.1)	**-2.2** (0.2)	**-0.03** (0.006)	0.02 (0.02)
6	**-39.1** (5.3)	**-0.6** (0.004)	**-0.4** (0.007)	**0.3** (0.05)	0.01 (0.01)	-0.003 (0.009)	0.0007 (0.1)	**0.02** (0.003)	**-0.2** (0.003)	**0.5** (0.08)	**-0.01** (0.003)	**0.08** (0.006)
7	**-22.2** (10.4)	**-1.3** (0.1)	**-0.3** (0.1)	-0.05 (0.09)	-0.02 (0.2)	0.01 (0.02)	0.2 (0.2)	0.01 (0.01)	0.03 (0.02)	0.06 (0.08)	**-0.05** (0.004)	**-0.03** (0.005)
8	**-10.6** (1.0)	**-0.2** (0.01)	**-0.03** (0.01)	**-0.09** (0.01)	**-0.1** (0.01)	-0.0008 (0.002)	0.02 (0.02)	**0.005** (0.0005)	**0.2** (0.008)	**0.6** (0.02)	0.0007 (0.0007)	**-0.01** (0.001)
9	7.7 (4.3)	**-0.3** (0.03)	0.05 (0.03)	**-0.09** (0.04)	-0.05 (0.04)	0.004 (0.005)	-0.1 (0.08)	-0.003 (0.002)	0.01 (0.01)	**-0.1** (0.05)	**-0.05** (0.005)	**-0.02** (0.002)
10	**16.8** (5.6)	**-0.8** (0.08)	**-0.1** (0.06)	-0.02 (0.04)	**-0.4** (0.1)	-0.001 (0.008)	-0.04 (0.1)	**-0.01** (0.003)	**-0.1** (0.03)	**-0.6** (0.2)	**-0.06** (0.02)	**-0.1** (0.02)
11	**17.0** (5.6)	**-0.8** (0.08)	**-0.1** (0.06)	-0.02 (0.04)	**-0.4** (0.1)	-0.001 (0.008)	-0.04 (0.1)	**-0.006** (0.003)	**-0.1** **(0.03)**	**-0.6** **(0.2)**	**-0.06** (0.02)	**-0.1** (0.02)
12	-10.5 (20.0)	-0.1 (0.1)	**0.8** (0.3)	**0.8** (0.3)	0.5 (0.4)	-0.03 (0.03)	-0.08 (0.5)	**0.02** (0.008)	-0.4 (0.2)	**2.3** (0.7)	**0.2** (0.09)	-0.1 (0.1)
13	**-67.5** (11.6)	-0.2 (0.09)	-0.008 (0.1)	0.09 (0.1)	-0.4 (0.2)	**0.04** (0.02)	-0.4 (0.3)	**0.04** (0.005)	-0.03 (0.4)	0.8 (1.8)	**-0.06** (0.02)	**-0.3** (0.05)
15	-11.0 (20.3)	**-0.5** (0.2)	**0.5** (0.3)	-0.04 (0.3)	0.4 (0.3)	0.009 (0.03)	0.05 (0.4)	0.009 (0.008)	0.7 (0.6)	—	**0.4** (0.08)	**0.9** (0.1)

Table 14: Multiple Regression Results

	n	Threshold M	Test MSE	Adjusted R^2	F-statistic	p-value
2	72	2614.51	2.53	0.12	1.83	0.087
3	519	2961.2	1.51	0.01	1.78	0.089
5	7302	4.8	1.59	0.06	**32.8**	$< 2.2 \times 10^{-16}$
6	19319	3	4.92	0.06	**96.18**	$< 2.2 \times 10^{-16}$
7	2398	2	2.24	0.18	**39.09**	$< 2.2 \times 10^{-16}$
8	93820	18	0.75	0.03	**195.3**	$< 2.2 \times 10^{-16}$
9	15986	50	1.62	0.04	**47.73**	$< 2.2 \times 10^{-16}$
10	11629	30	2.82	0.05	**46.16**	$< 2.2 \times 10^{-16}$
11	10868	7	2.48	0.03	**25.21**	$< 2.2 \times 10^{-16}$
12	1044	20	2.84	0.05	**5.188**	6.35×10^{-8}
13	4075	4	4.56	0.03	**10.19**	$< 2.2 \times 10^{-16}$
15	764	120	1.77	0.09	**7.142**	1.12×10^{-10}

The most important point of note is that these models to not perform that well. There are perhaps a few reasons for this. First, looking at the quantile-quantile plots (which are not included because there are so many), there are large errors committed towards the low and high fire sizes. This is not surprising given the distribution of the data and the presence of outliers towards the higher fire sizes. Moreover, there are not that many predictors available; the only true numerical predictors are X_7 through X_{11}, while all others are categorical in nature. We will discuss this more after examining more models. In general, a negative coefficient means an increase in that variable is associated on average with a decrease in the size of the fire and vice versa for a positive coefficient. With such a low R^2 value, it can be dubious to make too many inferences, as it is clear that the models are not capturing too much variability in the data.

Rather than taking the coefficient values with complete faith, we can more subtly use them (and scientific intuition) to decide how to collect more data. For example, we see in a lot of models that the mean fuel moisture is significant, and this effect is different across different ecoregions. Therefore, we might want to collect more data about the specific type of fuel that is available in the location where the fire started (rather than an average across an ecoregion). The coefficients can also be used to formulate scientific questions. For example, we might be interested in examining the effect of season in different ecoregions, given that some seasons have different signs of the corresponding coefficient. Moreover, one must keep in mind that the test MSE is measured in logarithmic units, so even small test MSEs are in actuality quite large. In general, most of the F-statistics are significant (except in ecoregion 2 and 3, likely due to a low amount of data). This is not too surprising given the large amount of data in most ecoregions.

4.3 Multiple Linear Regression with Interactions

We begin to examine a few more complex models. We only report the test MSE, adjusted R^2 for most models and F-statistic with the p-value for the F-statistic for those models which readily provide one with interpretation. We do not report the coefficient values but they can be obtained through the code. For the multiple linear regression with interactions, we choose to include an interaction between human and fuel moisture (X_1 and X_8), human and wind speed (X_1 and X_9), human and season (X_1 and X_2 through X_4), and human and holiday (X_1 and X_6). We include these interactions because we think they could be relevant. We note the following interaction coefficients which are significant at the 5% level:

- In Northern Forests, the human/wind speed coefficient is significant with a value of -2.88
- In Northwestern Forested Mountains, the human/fuel moisture and human/wind speed coefficients are significant with a value of -0.18 and 0.36
- In the Marine West Coast Forest, the human/holiday coefficient is significant with a value of -2.3

- In the Eastern Temperate Forest, the human/spring and human/fuel moisture coefficients are significant with a value of -0.21 and -0.14
- In the Great Plains, the human/fuel moisture coefficient is significant with a value of -0.12
- In the North American Deserts, the human/spring, human/fuel moisture, and human/wind speed coefficients are significant with a value of -0.53, -0.10, and 0.51
- In Mediterranean California, the human/summer and human/fuel moisture coefficients are significant with a value of -0.70 and -0.37

Table 15: Interaction Terms Multiple Regression Results

	Test MSE	Adjusted R^2	F-statistic	p-value
2	2.53	0.12	1.83	0.087
3	1.51	0.01	1.57	0.13
5	1.59	0.06	**23.71**	$< 2.2 \times 10^{-16}$
6	4.92	0.06	**64.11**	$< 2.2 \times 10^{-16}$
7	2.31	0.19	**27.14**	$< 2.2 \times 10^{-16}$
8	0.75	0.03	**130.6**	$< 2.2 \times 10^{-16}$
9	1.62	0.04	**34.32**	$< 2.2 \times 10^{-16}$
10	2.82	0.05	**31.5**	$< 2.2 \times 10^{-16}$
11	2.48	0.03	**20.31**	$< 2.2 \times 10^{-16}$
12	2.84	0.05	**3.412**	3.87×10^{-6}
13	4.56	0.03	**6.99**	$< 2.2 \times 10^{-16}$
15	1.77	0.09	**5.09**	4.46×10^{-9}

There is not too much of a change in test MSE or the R^2 values. This is not too surprising as the majority of the interaction coefficients were insignificant. This leads to a small decrease in the number of degrees of freedom and thus a small penalty to the adjusted R^2 (which is not really present with only 2 digits). This also leads to a small increase in the p-values from the F-statistics due to there being more parameters and not much additional predictive power. It can be interesting to investigate some of these interaction coefficients and formulate more scientific questions as a result of this, such as trying to determine the differences in the effect of season for human fires versus lightning fires. We now will investigate some more models which are perhaps less interpretable but hopefully will provide an increased prediction accuracy. We do not discuss the interpretation of these models because they are similar to the linear regression model.

4.4 Multiple Linear Regression with Variable Selection

With only 11 available predictors, we are able to do variable selection using best subset selection to try and find some better models with hopefully lower variance and hence better predictive accuracy or R^2 values. We report only the test MSE and adjusted R^2 values.

Table 16: Variable Selection Multiple Regression Results

	Test MSE	Adjusted R^2
2	2.29	0.17
3	1.53	0.02
5	1.59	0.06
6	4.92	0.06
7	2.24	0.18
8	0.75	0.03
9	1.62	0.04
10	2.82	0.05
11	2.48	0.03
12	2.84	0.05
13	4.56	0.03
15	1.77	0.10

These models are very similar. The idea with variable selection is to try and decrease the variance of the models, but it is likely that the linear regression model has a large bias (likely due to misspecification). Therefore, we are not able to receive too much of a benefit from this approach.

4.5 Ridge Regression

We fit a linear model with an L2 penalty on the coefficients. We use 10-fold cross-validation to choose the smoothness parameter λ. We report the test MSE and the deviance ratio (which is the fraction of null deviance explained and is similar to R^2 but an extension for generalized linear models). These are comparable, but perhaps it is more straightforward to compare the test MSE (note that the test set is the same for all models).

Table 17: Ridge Regression Results

	Test MSE	Fraction of null deviance explained	λ
2	2.14	0.07	1.23
3	1.50	0.01	4.41
5	1.59	0.06	0.02
6	4.93	0.06	0.04
7	2.23	0.18	0.05
8	0.75	0.03	0.008
9	1.62	0.04	0.02
10	2.82	0.05	0.11
11	2.48	0.03	0.02
12	2.84	0.05	0.03
13	4.56	0.03	0.17
15	1.80	0.11	0.02

We are able to see a small benefit in some ecoregions, but there are also some ecoregions which due slightly worse. Like we saw for variable selection, the reduction in variance is likely not too important.

4.6 Lasso Regression

We fit a linear model with an L1 penalty on the coefficients. We use 10-fold cross-validation to choose the smoothness parameter λ. We report the test MSE and the deviance ratio (which is the fraction of null deviance explained and is similar to R^2 but an extension for generalized linear models). We also provide the number of non-zero coefficients q.

Table 18: Lasso Regression Results

	Test MSE	Fraction of null deviance explained	λ	q
2	1.96	0	1.23	1
3	1.50	0.005	4.41	2
5	1.59	0.06	0.02	12
6	4.93	0.06	0.04	12
7	2.23	0.18	0.05	9
8	0.75	0.03	0.008	10
9	1.62	0.04	0.02	12
10	2.82	0.05	0.11	9
11	2.48	0.03	0.02	12
12	2.83	0.05	0.03	12
13	4.56	0.03	0.17	11
15	1.79	0.11	0.02	12

For the most part we are able to see very similar results as in ridge regression. Of interesting note are ecoregions 2 and 3, which lead to 1 and 2 non-zero coefficients respectively. In particular, for ecoregion 2, the fraction of null deviance explained is 0, and the test MSE is lower than in previous models. This is because all of the coefficients in this model are 0 except for the intercept–this is an indication that we are likely not able to learn much from these models. This is not surprising given that the variance of the test data is quite large, and this is not easily captured in these models. We should likely reject any conclusions developed from the previous models for this ecoregion considering that the null model performs better. For other ecoregions, the majority of coefficients still remain nonzero after cross-validation and shrinkage.

4.7 Regression Trees

We fit regression trees to the data. We prune these trees using 10-fold cross-validation. For each of the tree methods, we run into a similar issue as for the ridge and lasso models in that the interpretation for R^2 is not so clear. We use the following definition for a pseudo-R^2 value:

$$R^2 = 1 - \frac{\text{MSE}_{\text{test}}}{\text{Var}[Y_{\text{test}}]}$$

It is important to note that this is a measure of how the model compares to the null model (predicted the average for each observation). The range of this R^2 is from $(-\infty, 1]$, where a negative value indicates that the test MSE is higher than the variance of the test data set. The comparison between this value holds between the ridge/lasso model and the standard linear regression methods. For the single regression trees, we report the test MSE, this value for R^2, and the number of nodes in the pruned tree.

Table 19: Regression Tree Results

	Test MSE	R^2	Number of Nodes
2	2.53	-0.23	1
3	1.72	-0.14	6
5	1.36	0.20	6
6	4.96	0.05	4
7	2.25	0.17	6
8	0.75	0.04	3
9	1.64	0.03	2
10	2.87	0.05	3
11	2.48	0.03	3
12	2.69	0.04	4
13	4.62	0.01	3
15	1.74	0.14	7

These models give comparable results for the majority of the ecoregions and also are fairly easy to interpret. Although we do not picture the actual trees, they are fairly simple, with no more than 7 nodes after pruning. Most trees have 4 or less nodes. The specific trees could be useful at a more local scale trying to answer questions specific to a particular ecoregion. While single trees do not give that much of an increase in predictive accuracy over linear regression, we wish to investigate bagging, random forests, and boosting.

4.8 Bagging

Similar to the single regression trees, we report the test MSE and the same definition for R^2. Note that we used 300 trees (and only 100 trees in ecoregion 8 due to computational time). We combined each of the season variables into one variable (for a total of 9 variables). We also report the 3 most important variables.

Table 20: Bagging Results

	Test MSE	R^2	Most Important	2nd Most Important	3rd Most Important
2	2.19	-0.09	season	longitude	discovery holiday
3	1.39	0.08	discovery year	longitude	latitude
5	1.37	0.20	longitude	latitude	discovery year
6	4.75	0.09	longitude	latitude	season
7	2.34	0.14	longitude	latitude	human
8	0.70	0.09	discovery year	longitude	latitude
9	1.50	0.11	longitude	latitude	discovery year
10	2.72	0.10	longitude	human	latitude
11	2.53	0.01	latitude	human	longitude
12	2.72	0.02	longitude	mean windspeed	latitude
13	4.50	0.03	latitude	longitude	discovery year
15	1.87	0.07	longitude	latitude	discovery year

We are able to see a large improvement for most ecoregions by using bagging. We also are able to see some important information from the variable importance. The most important variables in general are latitude, longitude, and discovery year (with a couple others appearing). This is not very surprising, as expect geographical location to be important; we also know that fires have been increasing in size over time.

4.9 Random Forests

Similar to the single regression trees, we report the test MSE and the same definition for R^2. Note that we used 300 trees (and only 100 trees in ecoregion 8 due to computational time). We combined each of the season variables into one variable (for a total of 9 variables), and we only consider up to 3 variables for each tree. We also report the 3 most important variables.

Table 21: Random Forests Results

	Test MSE	R^2	Most Important	2nd Most Important	3rd Most Important
2	2.19	-0.09	season	human	discovery holiday
3	1.39	0.08	discovery year	longitude	discovery day of week
5	1.33	0.22	latitude	longitude	mean windspeed
6	4.59	0.12	longitude	latitude	human
7	2.17	0.20	longitude	human	latitude
8	0.68	0.12	discovery year	season	latitude
9	1.47	0.13	latitude	discovery year	longitude
10	2.65	0.12	longitude	human	latitude
11	2.46	0.04	longitude	human	latitude
12	2.70	0.04	longitude	mean windspeed	latitude
13	4.37	0.06	latitude	longitude	discovery year
15	1.84	0.09	longitude	latitude	discovery year

We see very similar results compared to bagging. However, we are able to get slight improvements in accuracy.

4.10 Boosting

In fitting using boosted trees, we perform 3-fold cross-validation in order to determine the parameter λ, considering values from 10^{-1} to 10^{-4}. We note that for ecoregion 2 we do not perform any cross-validation due to insufficient amount of data and instead use $\lambda = 0.01$. We report the value for λ which was selected from cross-validation, the test MSE, the same definition for R^2 as above, and the 3 most important variables.

Table 22: Boosting Results

	λ	Test MSE	R^2	Most Important	2nd Most Important	3rd Most Important
2	0.01	2.14	-0.04	discovery day of week	discovery year	— (all others 0)
3	0.001	1.43	0.06	discovery year	discovery day of week	season
5	0.003	1.56	0.08	fuel moisture	discovery day of week	discovery year
6	0.01	4.71	0.10	mean windspeed	discovery year	human
7	0.001	2.37	0.12	human	discovery day of week	discovery year
8	0.03	0.72	0.07	mean windspeed	fuel moisture	discovery year
9	0.01	1.55	0.08	discovery year	fuel moisture	discovery day of week
10	0.01	2.74	0.09	discovery year	mean windspeed	fuel moisture
11	0.003	2.50	0.02	discovery day of week	discovery year	fuel moisture
12	0.001	2.72	0.03	discovery day of year	discovery day of week	mean windspeed
13	0.001	4.56	0.02	discovery year	discovery day of week	season
15	0.0003	1.98	0.02	discovery year	discovery day of week	human

We see slightly worse results compared to bagging and random forests. It is interesting to see that the variable importance through boosting gives vastly different results. It is surprising to me to see that discovery day of week is important for a lot of the models.

4.11 Comparison of Models

We summarize all of the R^2 and test MSE results below. The specific definition for the R^2 of each model is defined in each respective section, and each model is evaluated on the same test set for each ecoregion. We **bold** which model is the **best** for each ecoregion (highest R^2 or lowest test MSE).

Table 23: Summary of R^2 for Regression Models (using standard or alternative definition)

	Multiple	Interaction	Variable Selection	Ridge	Lasso	Tree	Bagging	Random Forest	Boosting
2	0.12	0.12	**0.17**	0.07	0	-0.23	-0.09	-0.09	-0.04
3	0.01	0.01	0.02	0.01	0.005	-0.14	**0.08**	**0.08**	0.06
5	0.06	0.06	0.06	0.06	0.06	0.20	0.20	**0.22**	0.08
6	0.06	0.06	0.06	0.06	0.06	0.05	0.09	**0.12**	0.10
7	0.18	0.19	0.18	0.18	0.18	0.05	0.14	**0.20**	0.12
8	0.03	0.03	0.03	0.03	0.03	0.04	0.09	**0.12**	0.07
9	0.04	0.04	0.04	0.04	0.04	0.03	0.11	**0.13**	0.08
10	0.05	0.05	0.05	0.05	0.05	0.05	0.10	**0.12**	0.09
11	0.03	0.03	0.03	0.03	0.03	0.03	0.01	**0.04**	0.02
12	**0.05**	**0.05**	**0.05**	**0.05**	**0.05**	0.04	0.02	0.04	0.03
13	0.03	0.03	0.03	0.03	0.03	0.01	0.03	**0.06**	0.02
15	0.09	0.09	0.10	0.11	0.11	**0.14**	0.07	0.09	0.02

Table 24: Summary of Test MSE for Regression Models

	Multiple	Interaction	Variable Selection	Ridge	Lasso	Tree	Bagging	Random Forest	Boosting
2	2.53	2.53	2.29	2.14	**1.96**	2.53	2.19	2.19	2.14
3	1.51	1.51	1.53	1.50	1.50	1.72	**1.39**	**1.39**	1.43
5	1.59	1.59	1.59	1.59	1.59	1.36	1.37	**1.33**	1.56
6	4.92	4.92	4.92	4.93	4.93	4.96	4.75	**4.59**	4.71
7	2.24	2.31	2.24	2.23	2.23	2.25	2.34	**2.17**	2.37
8	0.75	0.75	0.75	0.75	0.75	0.75	0.70	**0.68**	0.72
9	1.62	1.62	1.62	1.62	1.62	1.64	1.50	**1.47**	1.55
10	2.82	2.82	2.82	2.82	2.82	2.87	2.72	**2.65**	2.74
11	2.48	2.48	2.48	2.48	2.48	2.48	2.53	**2.46**	2.50
12	2.84	2.84	2.84	2.84	2.83	**2.69**	2.72	2.70	2.72
13	4.56	4.56	4.56	4.56	4.56	4.62	4.50	**4.37**	4.56
15	1.77	1.77	1.77	1.80	1.79	**1.74**	1.87	1.84	1.98

For the most part, random forests are able to provide the best predictive accuracy and highest R^2 (lowest test MSE). However, for certain ecoregions such as ecoregion 7, this increase in R^2 is quite small compared to linear regression, so we might choose to use the simpler model in that case. However, in other regions such as ecoregion 3, 5, or 8, random forests are able to provide quite a significant increase in R^2 so these would likely be prefered over the simpler regression models.

4.12 Task 3 Conclusions

Overall, we are not able to find a model that has high predictive accuracy. Therefore, we need to be skeptical about trying to make inference from these models. The best R^2 ranges from 0.04 to 0.22 for the various ecoregions, which is very low. This is not surprising given the low amount of available predictors. Because there are not many quantitative predictors available, we are mainly predicting average fire sizes across the different classes. There is some information that we can extract, such as by using the discovery year and geographical location, but this is not enough to create a good model. In particular, fuel moisture and wind speed seem like they would be helpful predictors, but because they are just averages across the entire level 3 ecoregion, they are not too useful. It would be much more useful to have more specific data, such as the windspeed on the day of the fire discovery and fuel type in the area. Regardless, we are able to develop some potential intuition as to which relationships might be important and hopefully describe which variables might be important to gather for more model creation. I believe models used in this field should be more transparent about model validation and report R^2 and test MSE, for instance. It can be dubious to draw such conclusions when the models are not capturing a reasonable proportion of the variability of the data, but it is also likely that the underlying mechanism is very complex. It would be interesting to instead use classification methods for predicting the fire size class (which is a class between A and G). This might be more successful because there would not be large and small outliers; however, we still would need to worry about the imbalance of classes.

Appendix

A Decision Tree Classifier Models

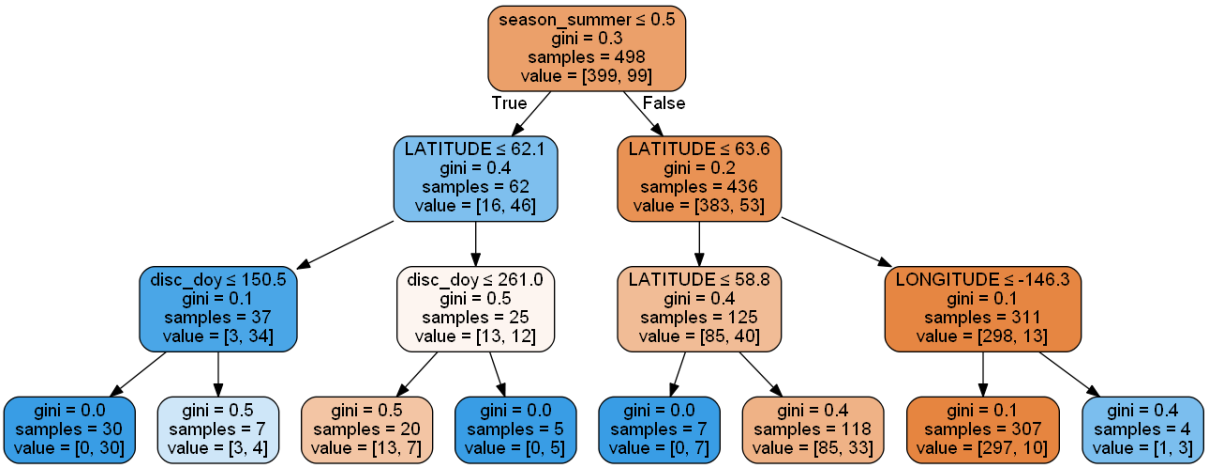


Figure 6: Decision Tree Classifier - Ecoregion 2 TUNDRA (0.04%), Blue: human caused, Orange: non-human caused

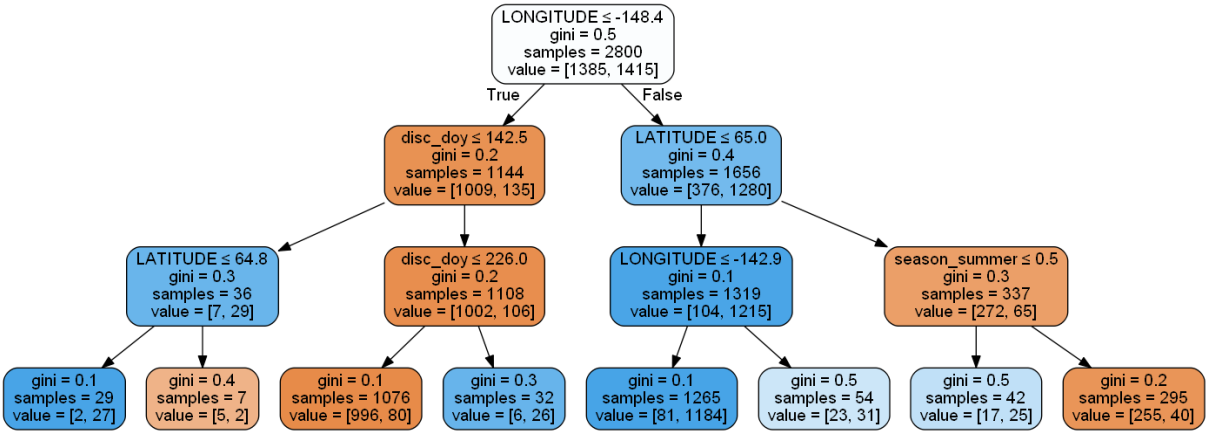


Figure 7: Decision Tree Classifier - Ecoregion 3 TAIGA (0.30%), Blue: human caused, Orange: non-human caused

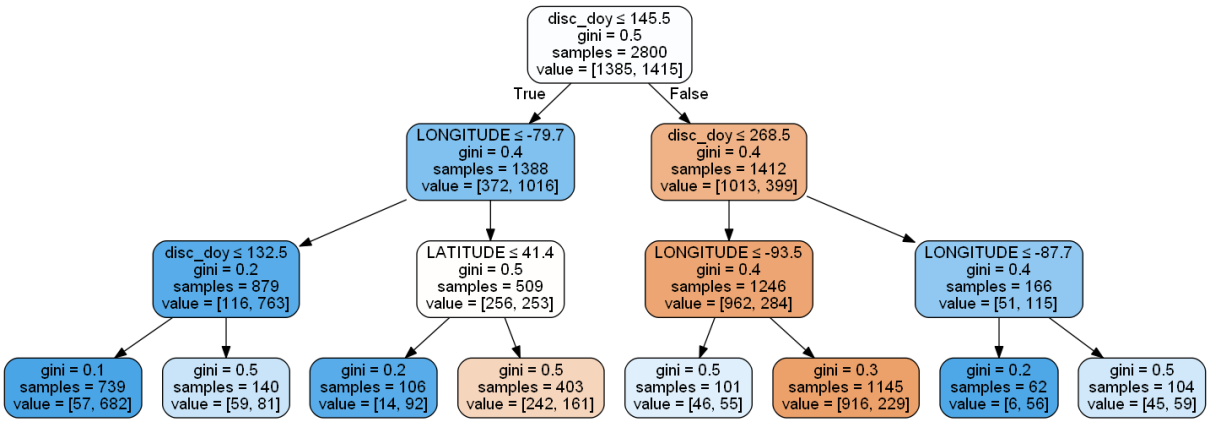


Figure 8: Decision Tree Classifier - Ecoregion 5 NORTHERN FORESTS (4.27%), Blue: human caused, Orange: non-human caused

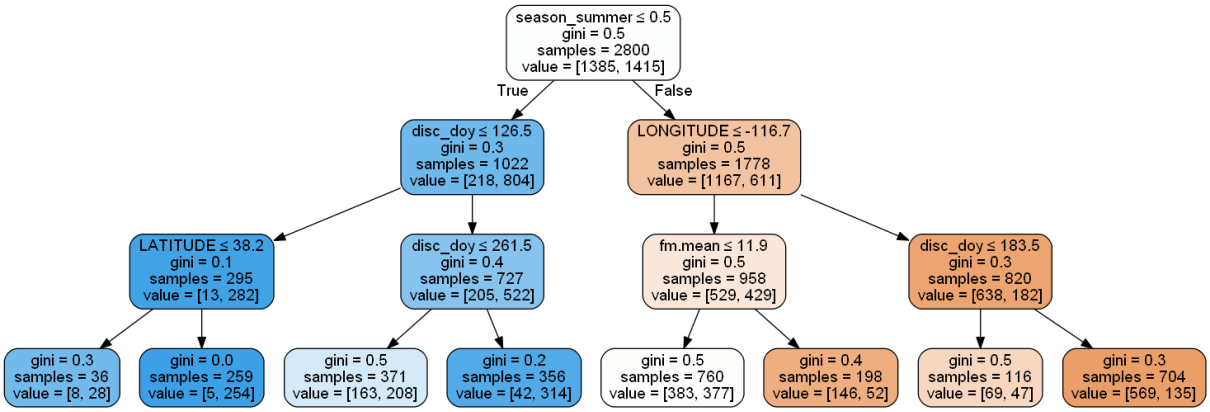


Figure 9: Decision Tree Classifier - Ecoregion 6 NW FORESTED MOUNTAINS (11.73%), Blue: human caused, Orange: non-human caused

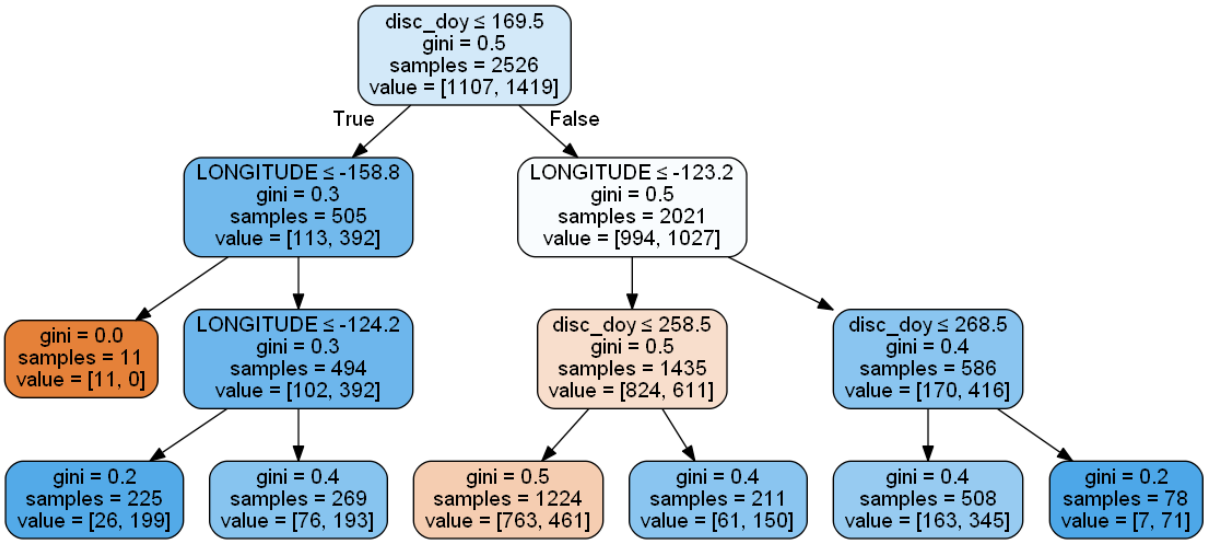


Figure 10: Decision Tree Classifier - Ecoregion 7 MARINE WEST COAST FOREST (1.49%), Blue: human caused, Orange: non-human caused

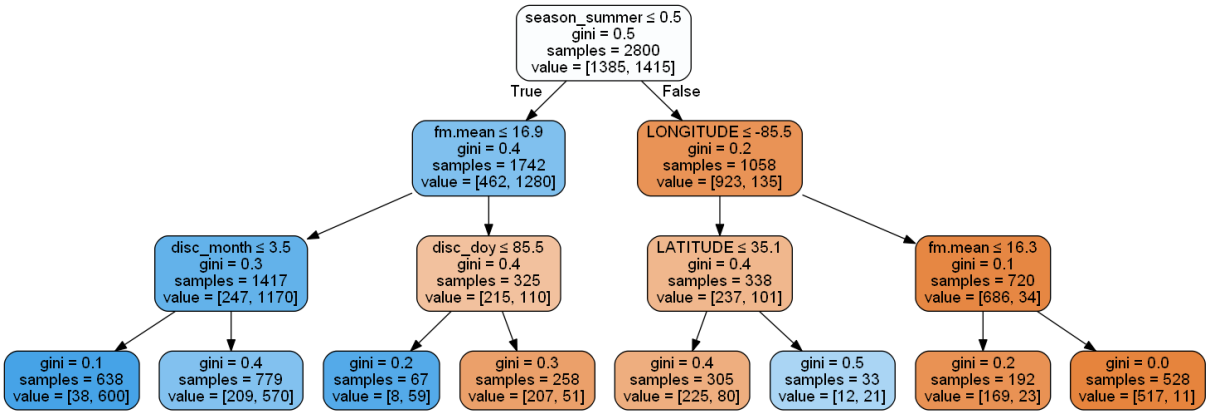


Figure 11: Decision Tree Classifier - Ecoregion 8 E TEMPERATE FOREST (55.40%), Blue: human caused, Orange: non-human caused

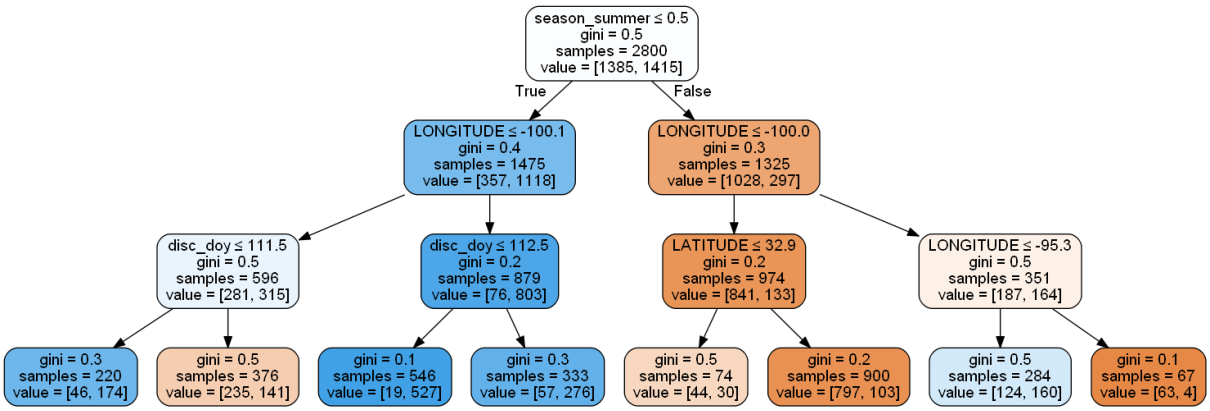


Figure 12: Decision Tree Classifier - Ecoregion 9 GREAT PLAINS (9.93%), Blue: human caused, Orange: non-human caused

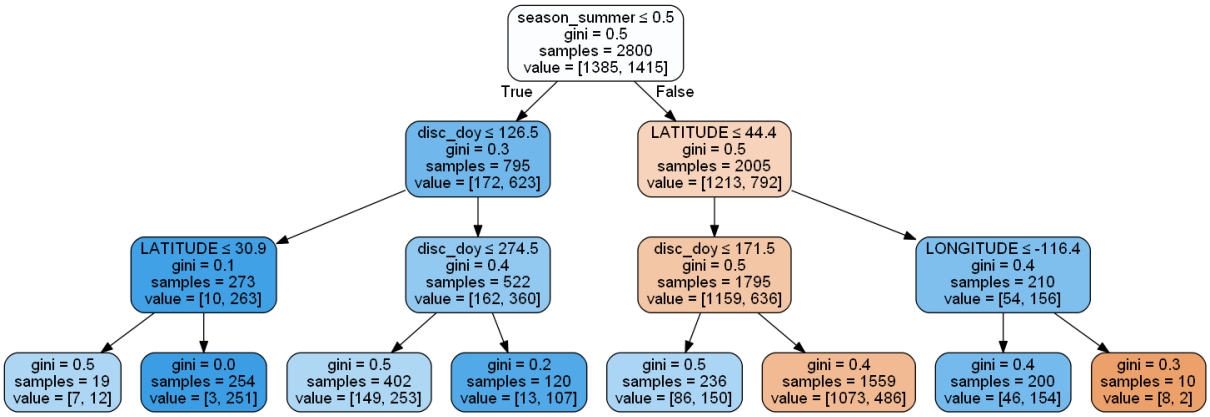


Figure 13: Decision Tree Classifier - Ecoregion 10 N AMERICAN DESERT (6.83%), Blue: human caused, Orange: non-human caused

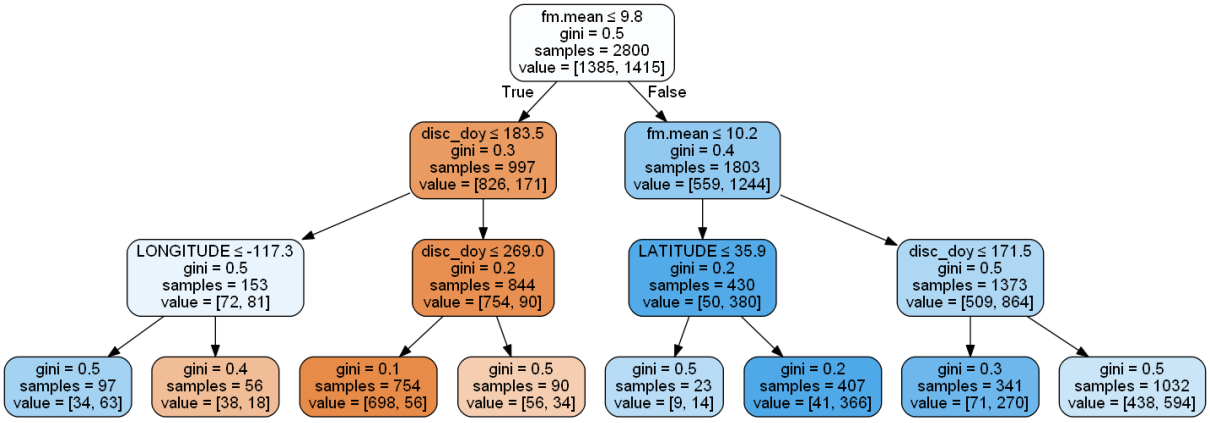


Figure 14: Decision Tree Classifier - Ecoregion 11 MED CALIFORNIA (6.38%), Blue: human caused, Orange: non-human caused

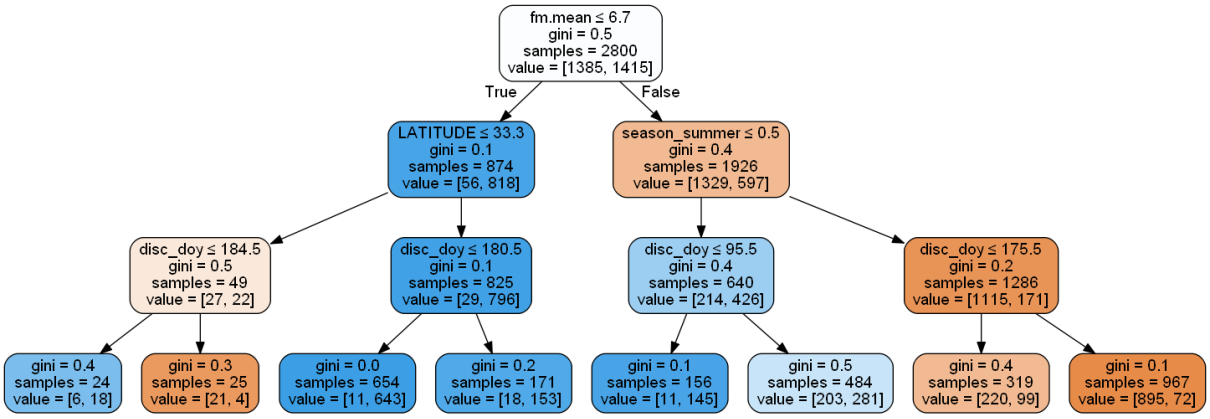


Figure 15: Decision Tree Classifier - Ecoregion 12 S SEMIARID HIGHLANDS (0.64%), Blue: human caused, Orange: non-human caused

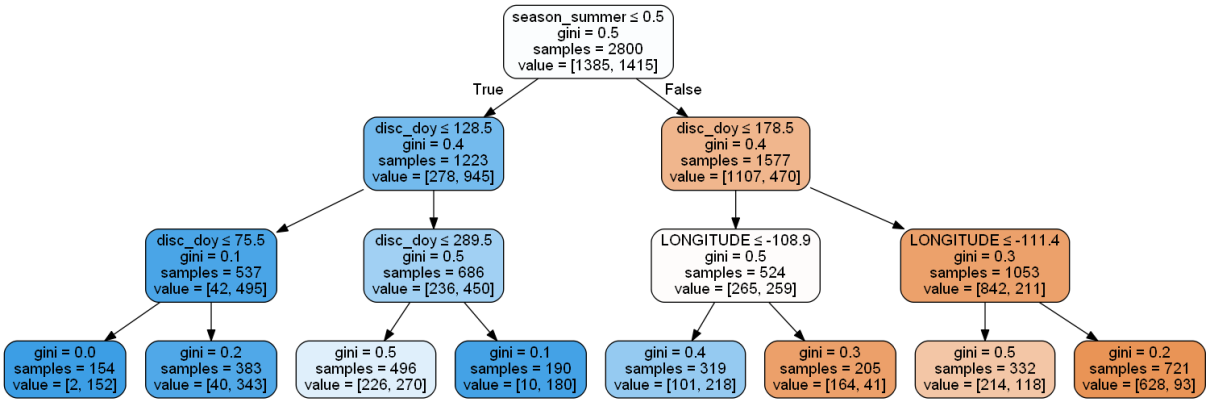


Figure 16: Decision Tree Classifier - Ecoregion 13 TEMPERATE SIERRAS (2.51%), Blue: human caused, Orange: non-human caused

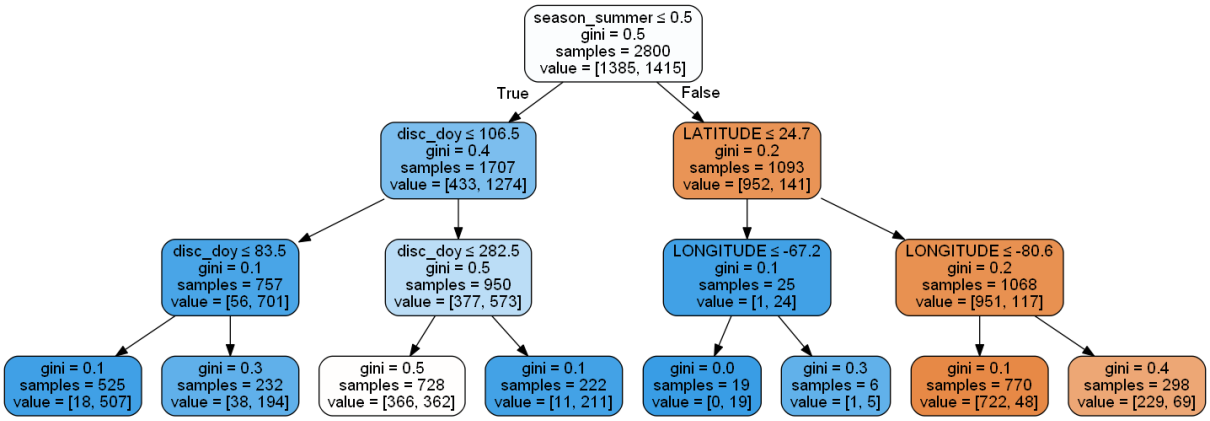


Figure 17: Decision Tree Classifier - Ecoregion 15 TROPICAL WET FOREST (0.45%), Blue: human caused, Orange: non-human caused

B Task 2

B.1 Logistic Regression Coefficient Signs

3 TAIGA	FIRE_SIZE	LATITUDE	LONGITUDE	season	disc_weekend	disc_holiday	eco3	fm.mean	Wind.mean
Arson	+	−	+	−	0	0	0	NA	NA
Campfire	−	−	+	−	0	0	0	NA	NA
Children	−	−	+	−	0	0	0	NA	NA
Debris Burning	+	+	−	+	+	0	0	NA	NA
Equipment Use	+	−	+	−	0	0	0	NA	NA
Fireworks	−	−	+	−	−	0	0	NA	NA
Lightning	+	+	−	+	+	0	0	NA	NA
Miscellaneous	+	+	−	0	0	0	0	NA	NA
Powerline	−	−	+	−	0	0	0	NA	NA
Railroad	−	−	+	−	0	0	0	NA	NA
Smoking	−	−	+	−	0	0	0	NA	NA
Structure	−	−	+	−	0	0	0	NA	NA

5 NORTHERN FORESTS	FIRE_SIZE	LATITUDE	LONGITUDE	season	disc_weekend	disc_holiday	eco3	fm.mean	Wind.mean
Arson	+	+	−	+	+	0	+	+	+
Campfire	+	−	+	+	+	0	−	−	−
Children	−	+	−	−	+	0	−	+	+
Debris Burning	+	+	−	+	+	0	+	+	+
Equipment Use	+	+	−	+	−	0	+	+	+
Fireworks	−	−	+	−	−	0	−	−	−
Lightning	+	−	+	−	−	0	−	−	−
Miscellaneous	+	+	−	+	+	+	+	+	+
Powerline	−	−	+	−	−	0	−	−	−
Railroad	−	−	+	−	−	0	−	−	−
Smoking	−	−	+	−	−	0	−	−	−
Structure	−	−	+	−	−	0	−	−	−

6 NORTHWESTERN FORESTED MTNS	FIRE_SIZE	LATITUDE	LONGITUDE	season	disc_weekend	disc_holiday	eco3	fm.mean	Wind.mean
Arson	+	−	+	−	−	0	−	−	−
Campfire	+	+	−	+	+	+	+	+	+
Children	−	−	+	−	−	−	−	−	−
Debris Burning	−	−	+	−	−	0	−	−	−
Equipment Use	+	−	+	−	−	−	−	−	−
Fireworks	−	−	+	−	−	0	−	−	−
Lightning	+	−	+	−	−	0	−	−	−
Miscellaneous	+	+	−	+	+	+	+	+	+
Powerline	−	−	+	−	−	0	−	−	−
Railroad	−	−	+	−	−	−	−	−	−
Smoking	−	−	+	−	−	−	−	−	−
Structure	−	−	+	−	−	−	−	−	−

7 MARINE WEST COAST FOREST	FIRE_SIZE	LATITUDE	LONGITUDE	season	disc_weekend	disc_holiday	eco3	fm.mean	Wind.mean
Arson	−	−	+	+	0	0	−	−	−
Campfire	+	+	−	+	+	0	+	+	+
Children	−	−	+	+	0	0	−	−	−
Debris Burning	−	+	−	+	+	0	+	+	+
Equipment Use	−	+	−	+	0	0	+	+	+
Fireworks	+	−	+	−	−	0	−	−	−
Lightning	+	+	−	+	0	0	+	+	+
Miscellaneous	+	+	−	+	+	0	+	+	+
Powerline	+	−	+	−	−	0	−	−	−
Railroad	−	−	+	−	−	0	−	−	−
Smoking	+	−	+	−	0	0	−	−	−
Structure	−	−	+	−	−	0	−	−	−

8 EASTERN TEMPERATE FORESTS	FIRE_SIZE	LATITUDE	LONGITUDE	season	disc_weekend	disc_holiday	eco3	fm.mean	Wind.mean
Arson	+	+	−	+	+	+	+	+	+
Campfire	−	−	+	−	−	−	−	−	−
Children	−	−	+	−	−	+	−	−	−
Debris Burning	−	+	−	+	+	+	+	+	+
Equipment Use	+	+	−	+	−	−	+	+	+
Fireworks	−	−	+	−	−	−	−	−	−
Lightning	+	−	+	+	−	−	+	−	−
Miscellaneous	+	+	−	+	+	+	+	+	+
Powerline	−	−	+	−	−	−	−	−	−
Railroad	−	−	+	−	−	−	−	−	−
Smoking	−	−	+	−	−	−	−	−	−
Structure	−	−	+	−	−	−	−	−	−

9 GREAT PLAINS	FIRE_SIZE	LATITUDE	LONGITUDE	season	disc_weekend	disc_holiday	eco3	fm.mean	Wind.mean
Arson	+	+	−	+	+	+	+	+	+
Campfire	−	−	+	−	−	−	−	−	−
Children	−	−	+	−	−	−	−	−	−
Debris Burning	+	+	−	+	+	+	+	+	+
Equipment Use	+	+	−	+	+	+	+	+	+
Fireworks	−	−	+	−	−	+	−	−	−
Lightning	+	+	−	+	+	−	+	+	+
Miscellaneous	+	+	−	+	+	+	+	+	+
Powerline	+	−	+	−	−	−	−	−	−
Railroad	−	−	+	−	−	−	−	−	−
Smoking	−	−	+	−	−	−	−	−	−
Structure	−	−	+	−	−	−	−	−	−

10 NORTH AMERICAN DESERTS	FIRE_SIZE	LATITUDE	LONGITUDE	season	disc_weekend	disc_holiday	eco3	fm.mean	Wind.mean
Arson	+	−	+	−	−	0	−	−	−
Campfire	−	−	+	−	−	0	−	−	−
Children	−	−	+	−	−	−	−	−	−
Debris Burning	+	+	−	−	0	0	+	+	+
Equipment Use	+	−	+	−	−	0	−	−	−
Fireworks	−	−	+	−	−	+	−	−	−
Lightning	+	+	−	+	+	+	+	+	+
Miscellaneous	+	+	−	+	+	+	+	+	+
Powerline	−	−	+	−	−	−	−	−	−
Railroad	−	−	+	−	−	−	−	−	−
Smoking	−	−	+	−	−	−	−	−	−
Structure	−	−	+	−	−	−	−	−	−

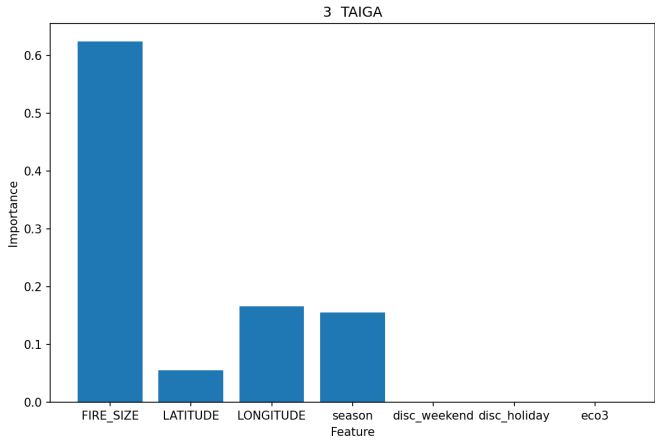
11 MEDITERRANEAN CALIFORNIA	FIRE_SIZE	LATITUDE	LONGITUDE	season	disc_weekend	disc_holiday	eco3	fm_mean	Wind_mean
Arson	+	+	−	+	+	+	+	+	+
Campfire	+	−	+	−	−	0	−	−	−
Children	−	−	+	−	−	0	−	−	−
Debris Burning	+	+	−	−	+	0	+	+	+
Equipment Use	+	+	−	+	+	+	+	+	+
Fireworks	−	−	+	−	−	−	−	−	−
Lightning	+	−	+	−	−	−	−	−	−
Miscellaneous	+	+	−	+	+	+	+	+	+
Powerline	+	−	+	−	−	−	−	−	−
Railroad	−	−	+	−	−	−	−	−	−
Smoking	−	−	+	−	−	−	−	−	−
Structure	−	−	+	−	−	−	−	−	−

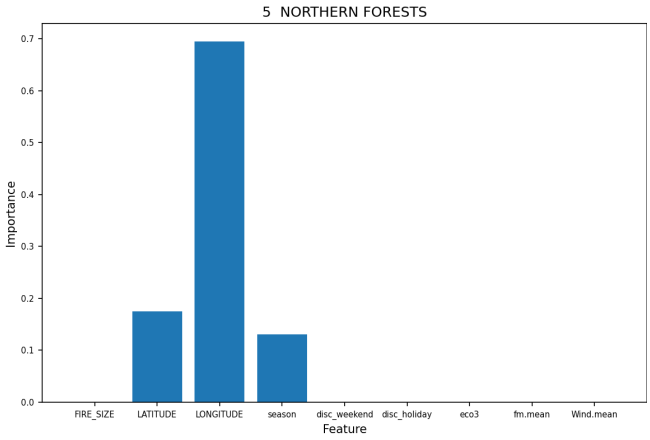
12 SOUTHERN SEMIARID HIGHLANDS	FIRE_SIZE	LATITUDE	LONGITUDE	season	disc_weekend	disc_holiday	eco3	fm_mean	Wind_mean
Arson	−	+	−	+	+	+	+	+	+
Campfire	+	+	−	−	+	+	+	+	+
Children	−	+	−	+	+	0	−	+	+
Debris Burning	−	−	+	−	−	0	−	−	−
Equipment Use	−	−	+	−	−	−	−	−	−
Fireworks	−	−	+	−	−	−	−	−	−
Lightning	+	+	−	+	−	−	−	−	−
Miscellaneous	+	+	−	+	+	+	+	+	+
Powerline	−	−	+	−	−	−	−	−	−
Railroad	−	−	+	−	−	−	−	−	−
Smoking	−	−	+	−	−	−	−	−	−
Structure	−	−	+	−	−	−	−	−	−

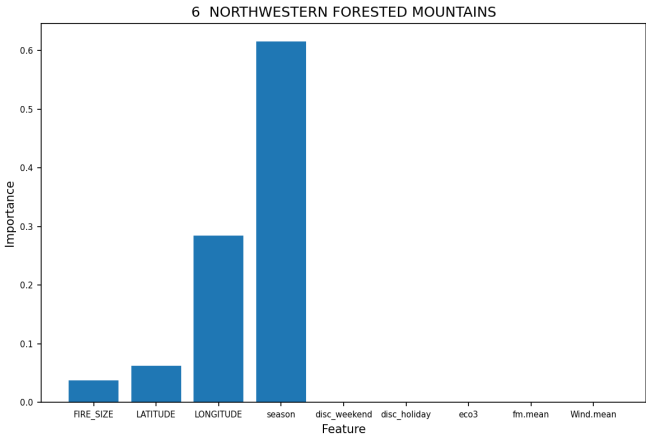
13 TEMPERATE SIERRAS	FIRE_SIZE	LATITUDE	LONGITUDE	season	disc_weekend	disc_holiday	eco3	fm_mean	Wind_mean
Arson	+	−	+	−	−	0	−	−	−
Campfire	+	+	−	+	+	+	+	+	+
Children	−	−	+	−	−	0	−	−	−
Debris Burning	−	−	+	−	−	0	−	−	−
Equipment Use	−	−	+	−	−	0	−	−	−
Fireworks	−	−	+	−	−	0	−	−	−
Lightning	+	+	−	+	+	+	+	+	+
Miscellaneous	+	+	−	+	+	0	+	+	+
Powerline	−	−	+	−	−	0	−	−	−
Railroad	−	−	+	−	−	0	−	−	−
Smoking	−	−	+	−	−	0	−	−	−
Structure	−	−	+	−	−	0	−	−	−

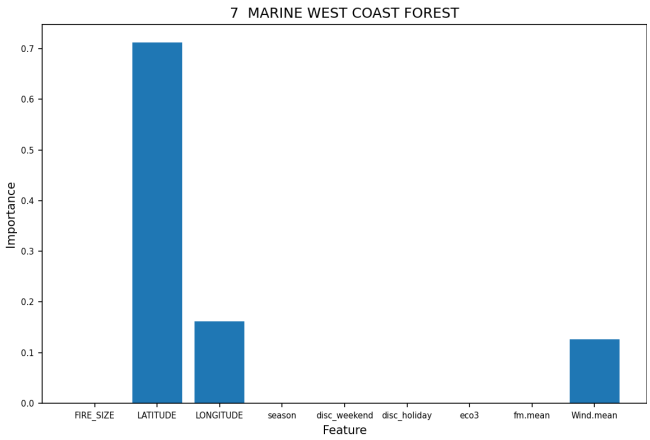
15 TROPICAL WET FORESTS	FIRE_SIZE	LATITUDE	LONGITUDE	season	disc_weekend	disc_holiday	eco3	fm_mean	Wind_mean
Arson	+	+	−	+	+	0	+	+	+
Campfire	−	−	+	−	−	0	−	−	−
Children	−	−	+	−	0	0	−	−	−
Debris Burning	+	−	+	−	0	0	−	−	−
Equipment Use	+	+	−	+	0	0	+	+	+
Fireworks	−	−	+	−	−	0	−	−	−
Lightning	+	+	−	+	+	0	+	+	+
Miscellaneous	+	+	−	+	+	0	+	+	+
Powerline	−	−	+	−	−	0	−	−	−
Railroad	+	+	−	+	+	0	+	+	+
Smoking	−	−	+	−	−	0	−	−	−
Structure	−	−	+	−	−	0	−	−	−

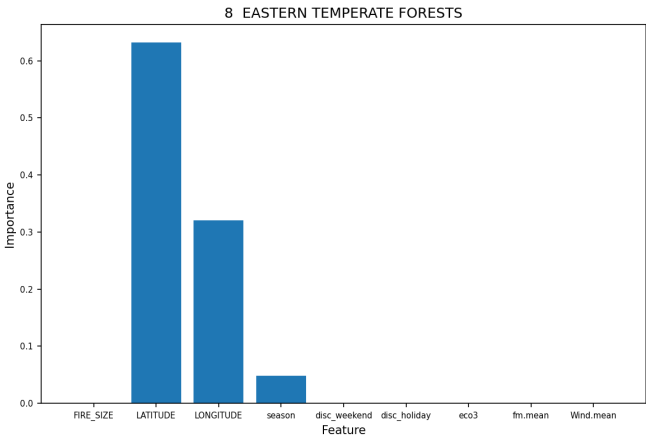
B.2 Variable Importance

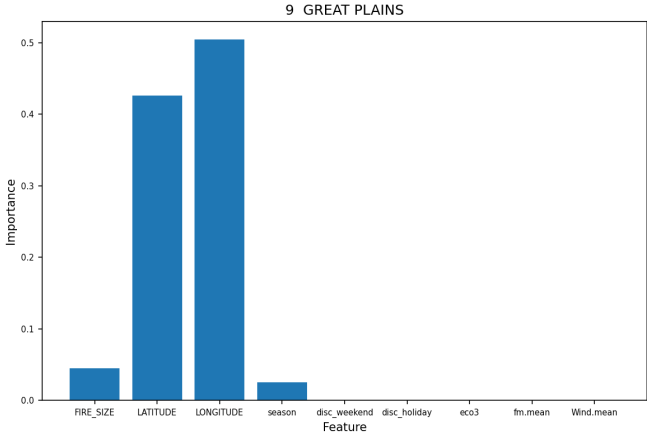


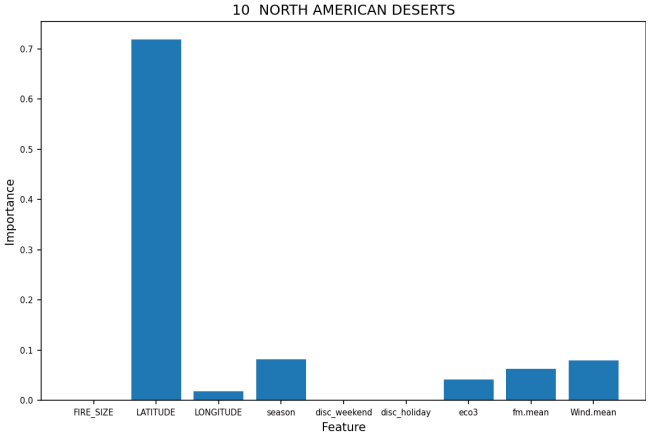


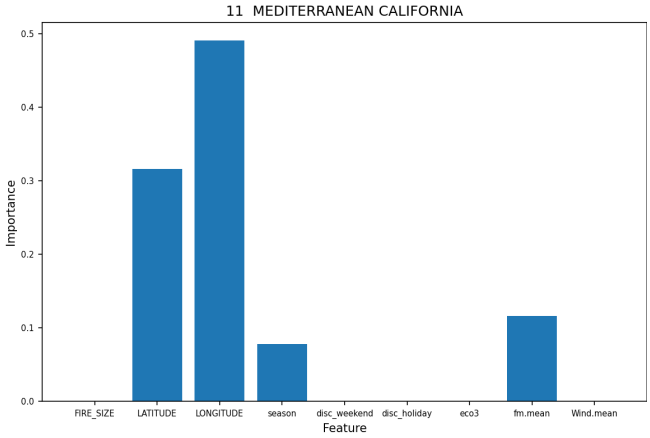


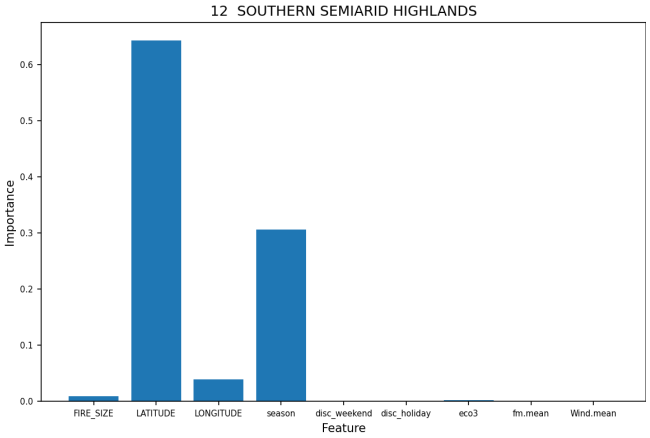


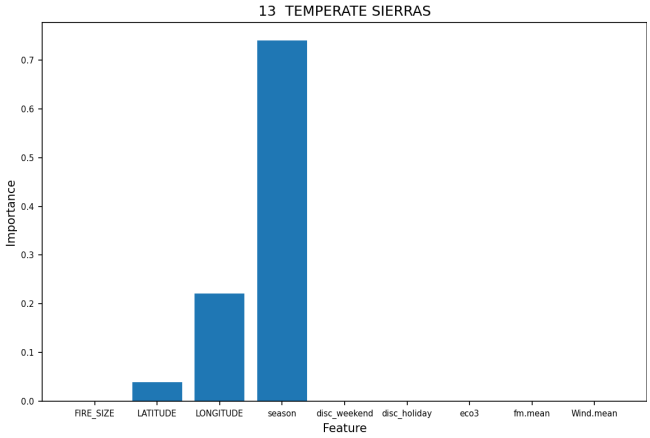


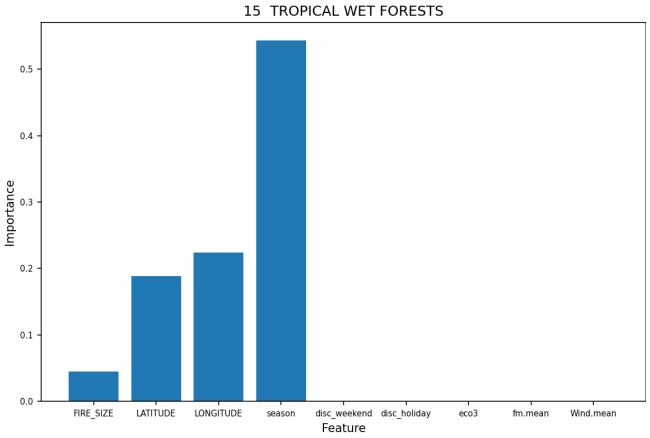

9 GREAT PLAINS
Importance
0.0
0.1
0.2
0.3
0.4
0.5
FIRE_SIZE
LATITUDE
LONGITUDE
season
disc_weekend
disc_holiday
eco3
fm.mean
Wind.mean
Feature


10 NORTH AMERICAN DESERTS
Importance
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
FIRE_SIZE
LATITUDE
LONGITUDE
season
disc_weekend
disc_holiday
eco3
fm.mean
Wind.mean
Feature

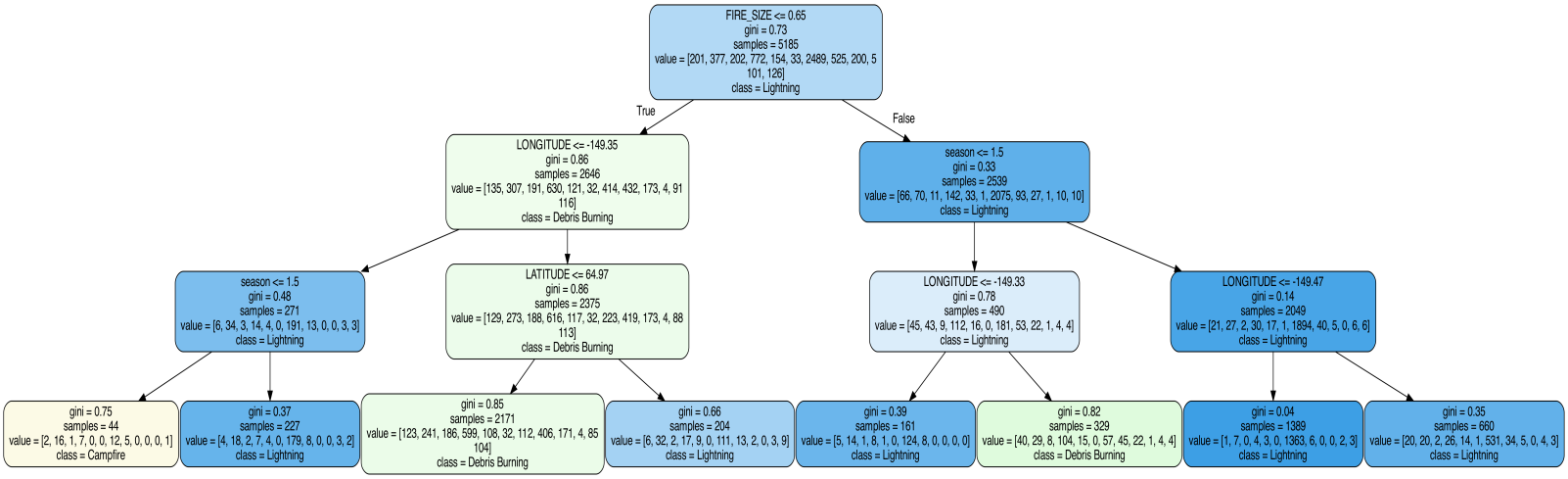

11 MEDITERRANEAN CALIFORNIA
Importance
0.0
0.1
0.2
0.3
0.4
0.5
FIRE_SIZE
LATITUDE
LONGITUDE
season
disc_weekend
disc_holiday
eco3
fm.mean
Wind.mean
Feature


12 SOUTHERN SEMIARID HIGHLANDS
Importance
0.0
0.1
0.2
0.3
0.4
0.5
0.6
FIRE_SIZE
LATITUDE
LONGITUDE
season
disc_weekend
disc_holiday
eco3
fm.mean
Wind.mean
Feature

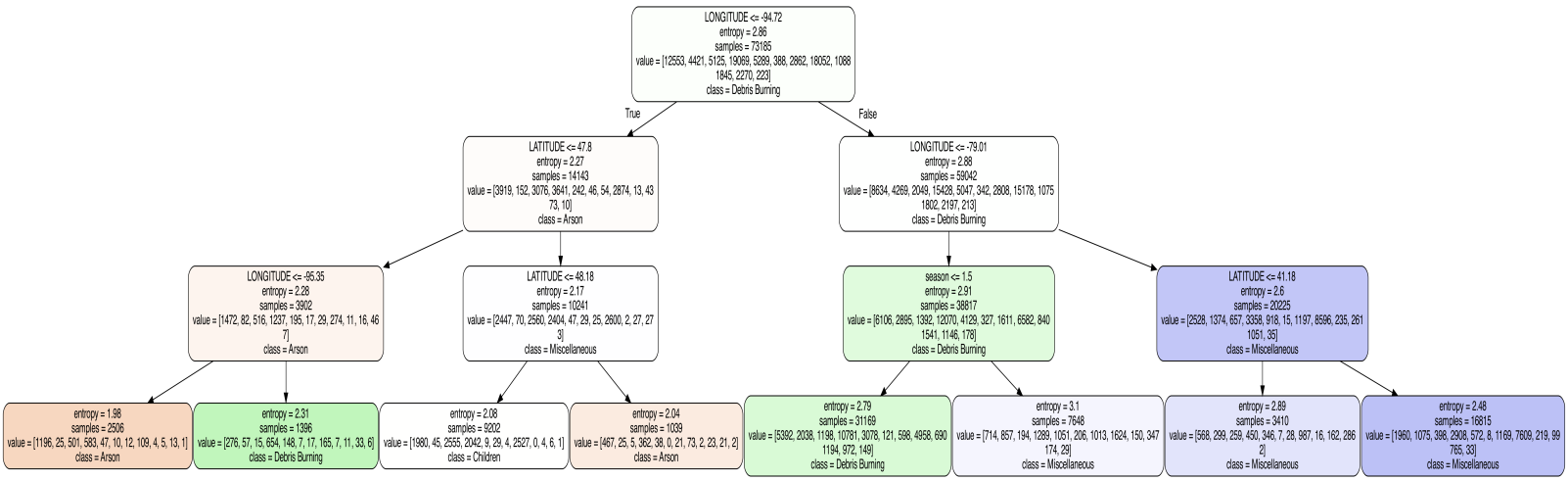

13 TEMPERATE SIERRAS
Importance
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
FIRE_SIZE
LATITUDE
LONGITUDE
season
disc_weekend
disc_holiday
eco3
fm.mean
Wind.mean
Feature


15 TROPICAL WET FORESTS
Importance
0.0
0.1
0.2
0.3
0.4
0.5
FIRE_SIZE
LATITUDE
LONGITUDE
season
disc_weekend
disc_holiday
eco3
fm.mean
Wind.mean
Feature

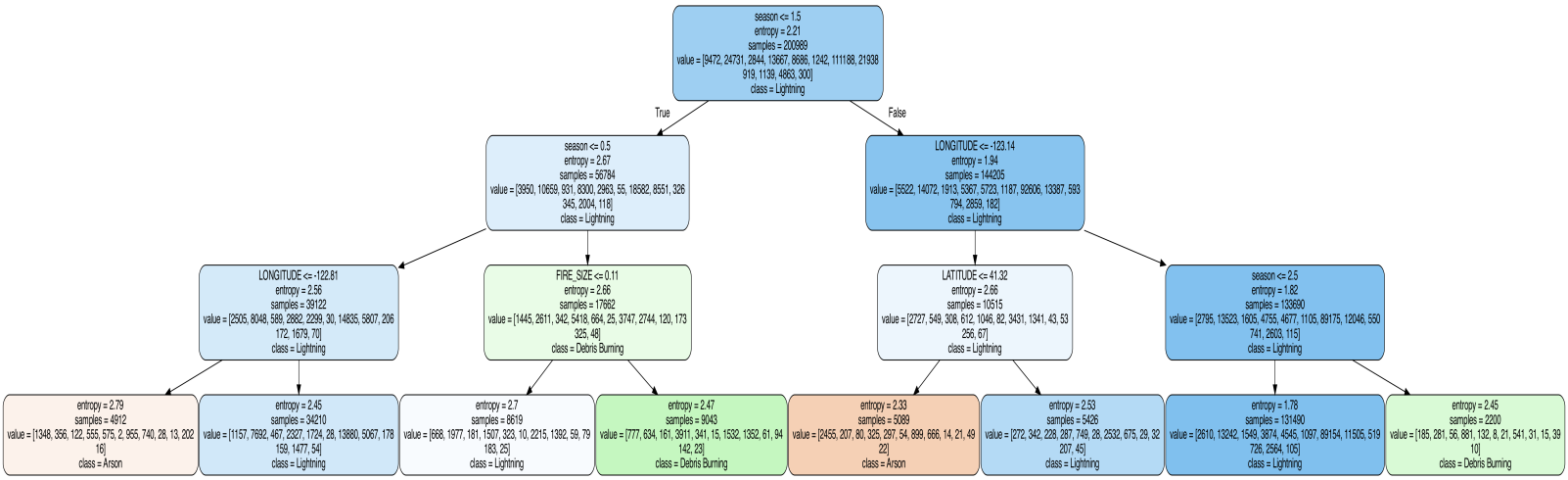
B.3 Decision Trees



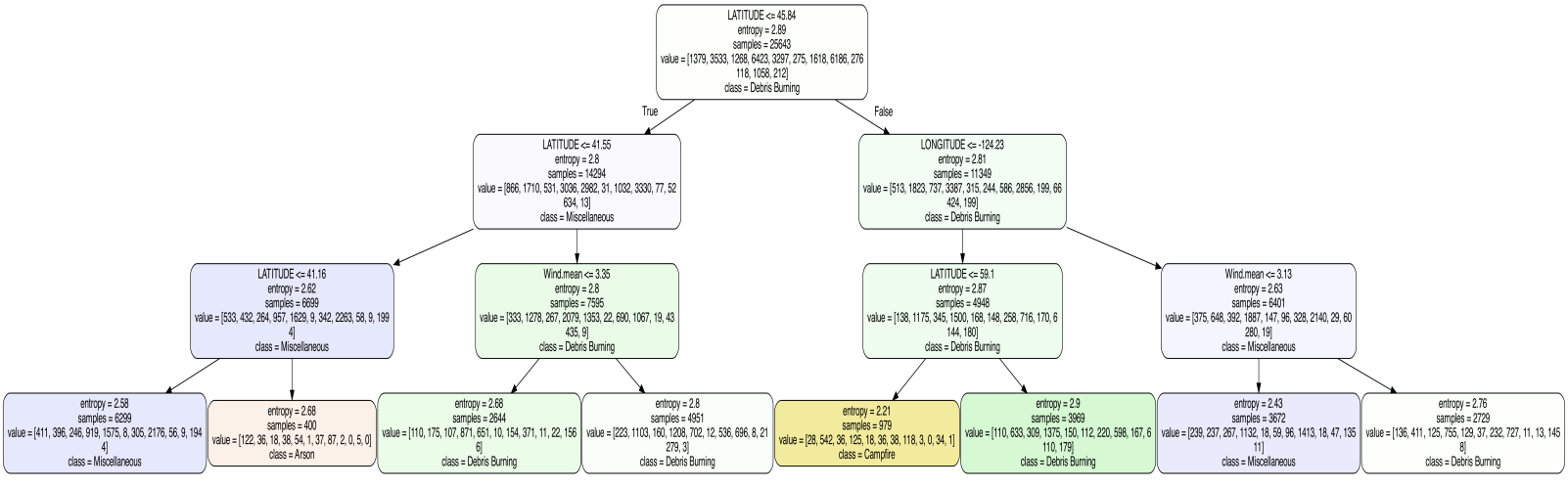
Region 3 TAIGA



Region 5 NORTHERN FORESTS



Region 6 NORTHWESTERN FORESTED MTNS



Region 7 MARINE WEST COAST FOREST

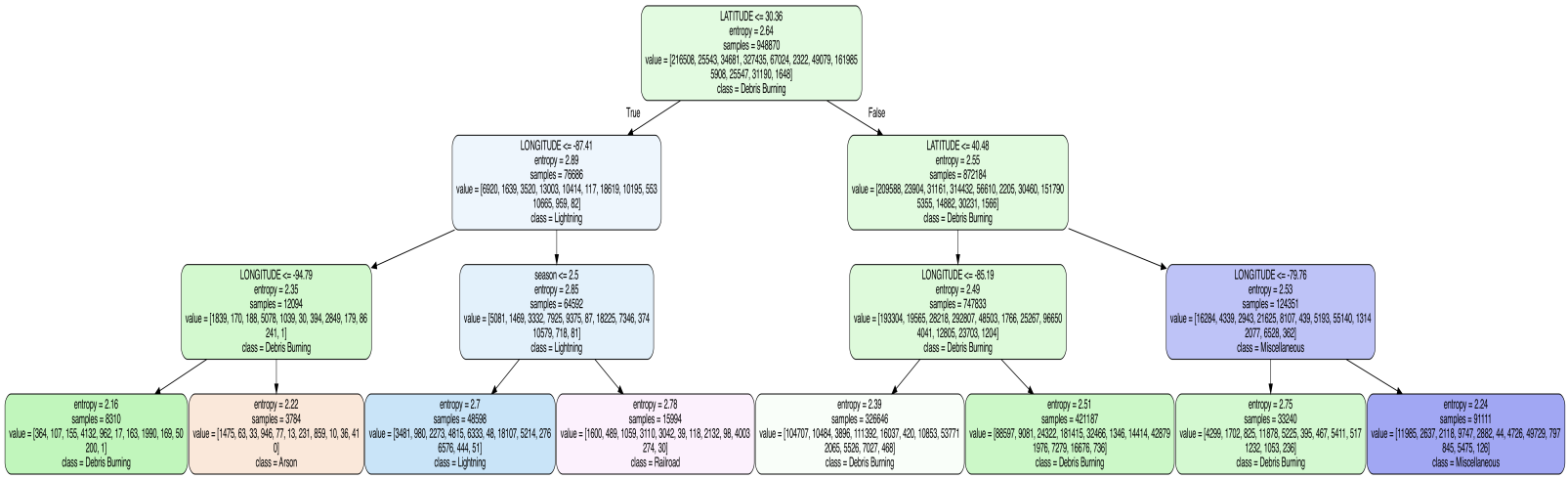
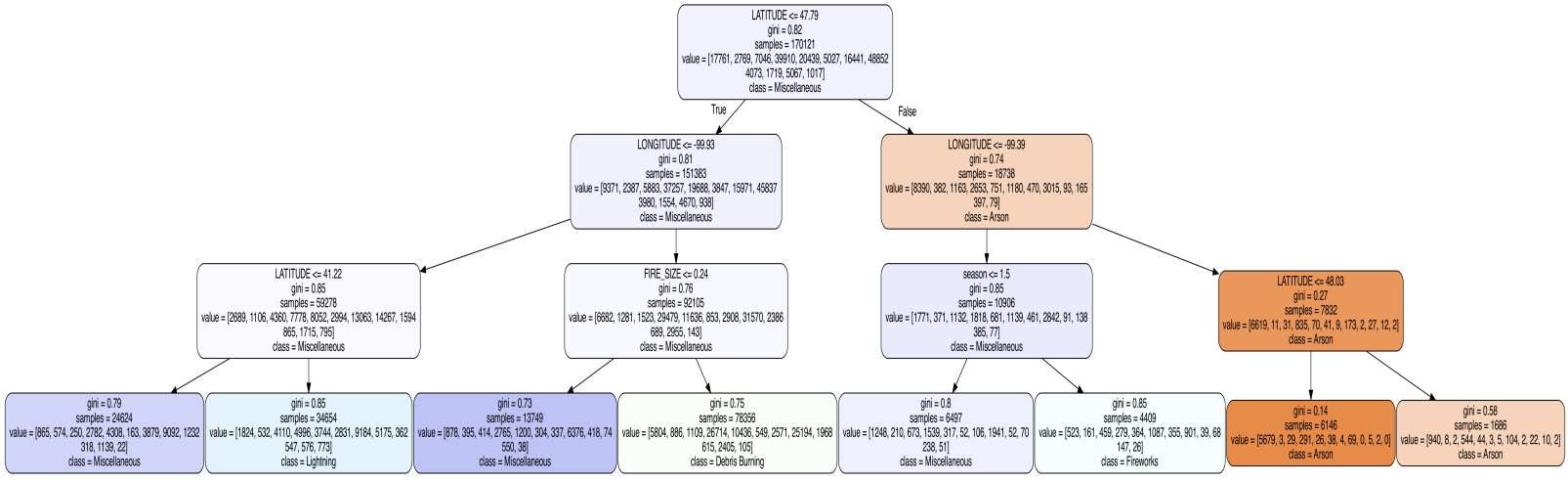
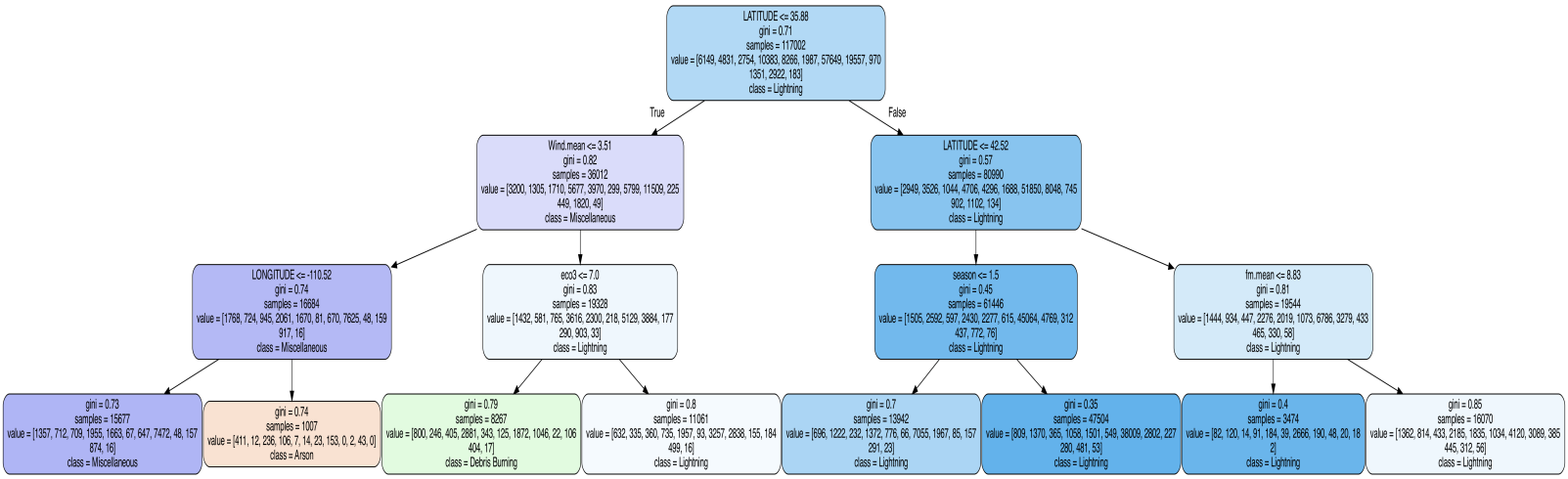


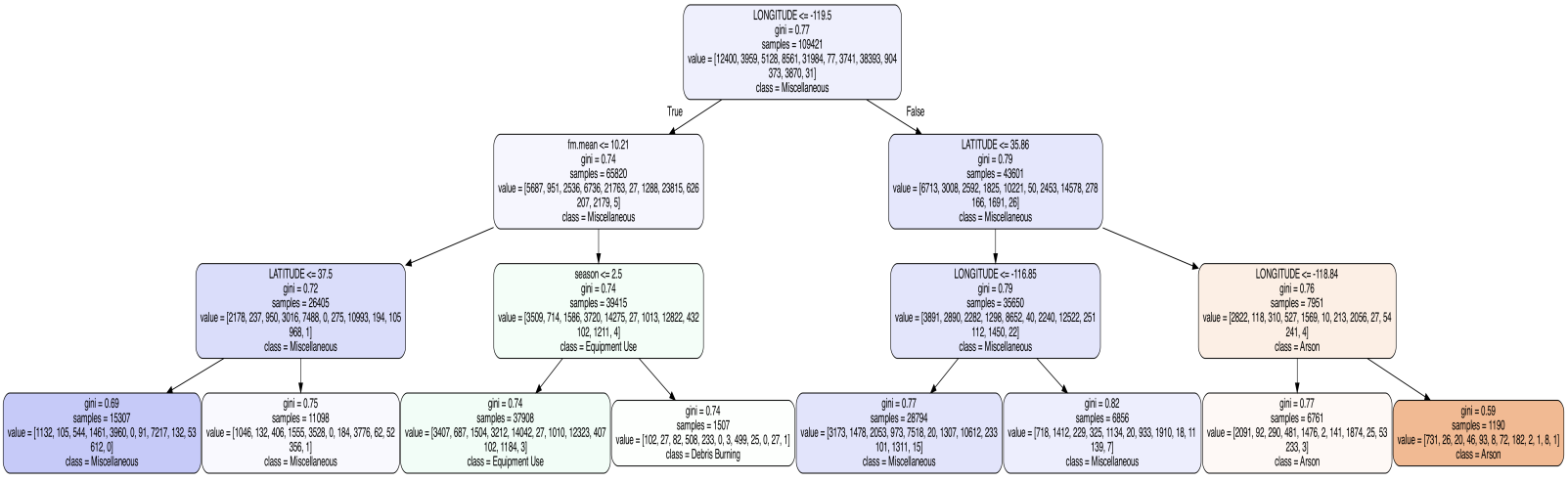
Figure 19: Region 8 EASTERN TEMPERATE FORESTS



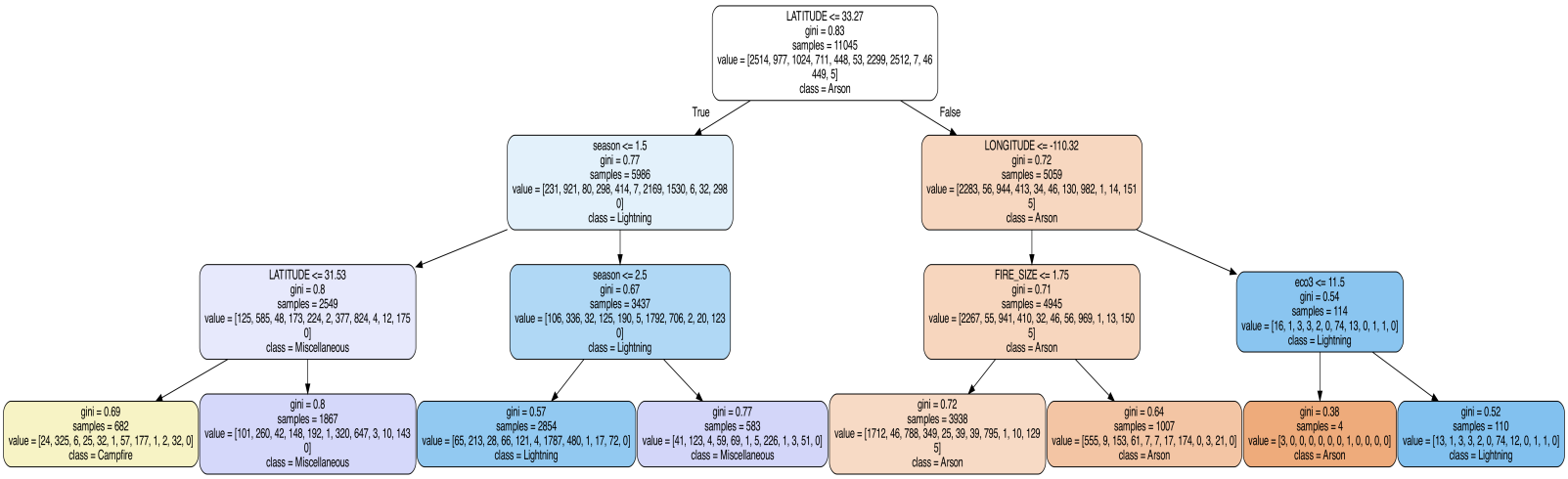
Region 9 GREAT PLAINS



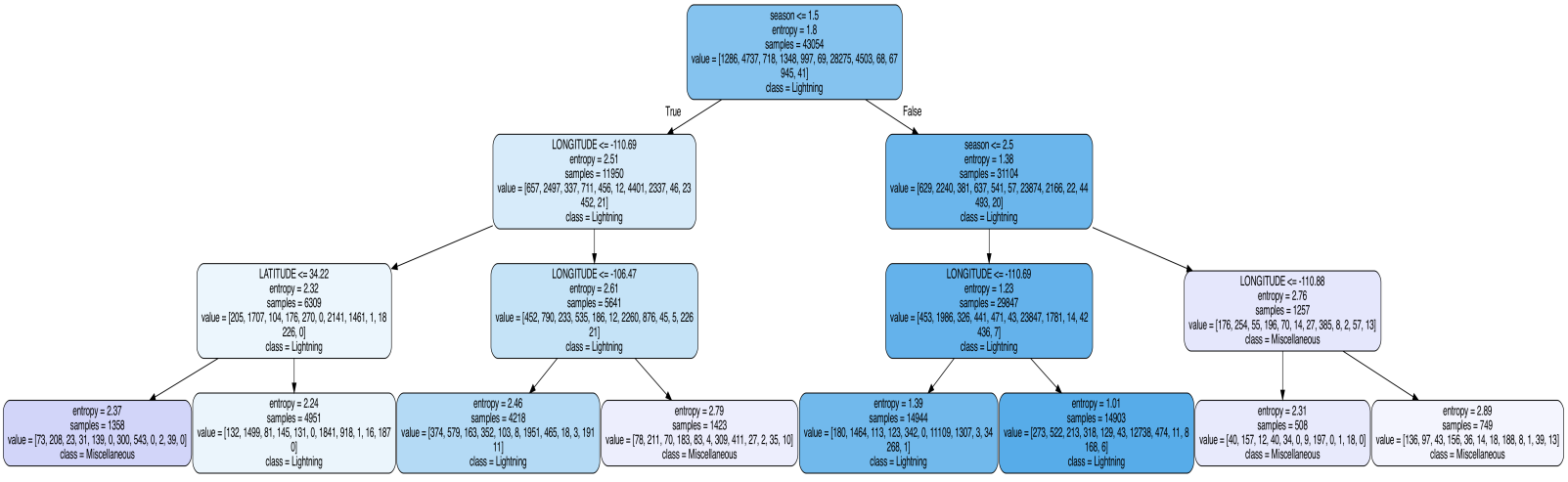
Region 10 NORTH AMERICAN DESERTS



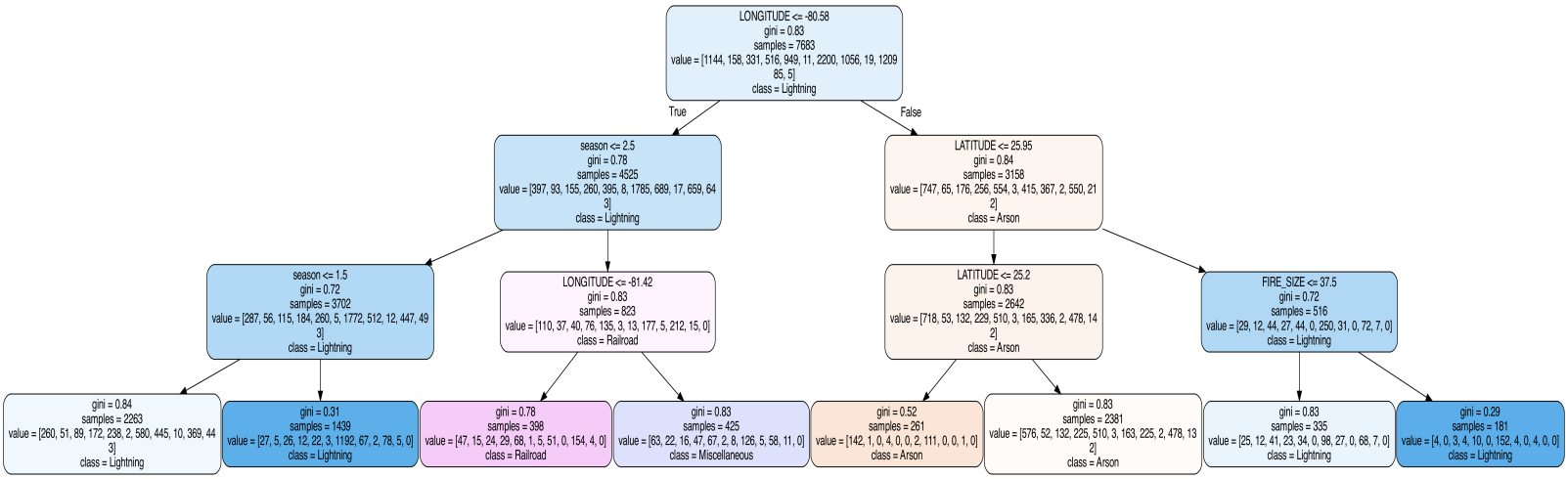
Region 11 MEDITERRANEAN CALIFORNIA



Region 12 SOUTHERN SEMIARID HIGHLANDS



Region 13 TEMPERATE SIERRAS



Region 15 TROPICAL WET FORESTS

References

[1] Jennifer K Balch, Bethany A Bradley, John T Abatzoglou, R. Chelsea Nagy, Emily J Fusco, and Adam L Mahood. "Human-started wildfires expand the fire niche across the United States". eng. In: *Proceedings of the National Academy of Sciences - PNAS*. From the Cover 114.11 (2017), pp. 2946–2951. ISSN: 0027-8424.

[2] R. Chelsea Nagy. *Short Large Fires*. Version 1.0. GitHub, June 2018. URL: https://github.com/nagyrc/Short_large_fires.

[3] R.C. Nagy, Emily Fusco, Bethany Bradley, and John T Abatzoglou. "Human-Related Ignitions Increase the Number of Large Wildfires across U.S. Ecoregions". eng. In: *Fire (Basel, Switzerland)* 1.1 (2018), pp. 4–. ISSN: 2571-6255.