US Forest Fire Size Prediction using Machine Learning

By

Padma Prabagaran

Fall 2022

Ryan Meuth Director

Robert McCulloch Second Committee Member

Abstract or Executive Summary

The number of extreme wildfires is on the rise globally, and predicting the size of a fire will help officials make appropriate decisions to mitigate the risk the fire poses against the environment and humans. This study attempts to find the burned area of a fire based on attributes such as time, weather, and location of the fire using machine learning methods. The study is split into two parts: the first is a regression approach to fire size prediction that follows the methodology in the paper published by Paulo Cortez and Anibal Morais called *A Data Mining Approach to Predict Forest Fires using Meteorological Data.* The best configuration for this part of the study was using a Support Vector Machine with all three spatial, temporal, and weather variables which had an MAE of 2676.55. The second part of the study uses classification techniques to find the fire size class of wildfire; this resulted in models that accurately predicted around 30% of fires and 60% of the fires within one class boundary. The models that performed the best were the Random Forest and XGBoost models; they also were able to predict smaller fires much better compared to larger fires. The results of this study indicate that there are opportunities for improvement but are promising in improving resource management for fire mitigation.

Introduction

Climate Change and the complications that arise from it have been a matter of great importance among concerned citizens worldwide. Although wildfires occur naturally in ecosystems, fire seasons have been consistently getting more extreme due to the drier, hotter weather brought on by climate change. Along with inadequate land management, the conditions are ideal for bigger, high-intensity fires [3]. The consequences of a forest fire can be quite

devastating; besides the possible destruction of human habitats, forest fires also disturb the natural cycle of forest ecosystems, causing some species to disappear while also allowing some invasive species to thrive. Forest fires also contribute to increases in carbon dioxide levels, thus exacerbating the greenhouse effect, which in turn increases the chance of more fires [4], leading to an endless loop.

Uncontrolled fires have the ability to cause billions of dollars in economic damage. In the last 5 Years alone (2017-2021) the United States has spent over $79.8 billion on forest fire mitigation and damages [4]. This value is predicted to rise in the coming years. Forest fires in the United States have displaced thousands of families [5] and destroyed homes, wildlife habitats, timber, and polluted the air with emissions that are detrimental to human health. Thus, being able to predict the size of the fire can help with resource allocation for forest fire mitigation in the country.

Background

This paper aims to benchmark the results found in the paper by Paulo Cortez and Anibal Morais [1]. In their paper, Cortez and Morais use five machine learning techniques; Support Vector Machines, Random Forests, Multiple Regression, Decision Trees, and Neural Nets to predict the burned area of fires that occurred in Montesinho natural park in Portugal.

Their study utilizes spatial, temporal, weather, and FWI attributes in their experiment. The spatial attributes X and Y track the discovery location of the fires; this was a substitute for vegetation data. Month and Day are temporal variables. The four meteorological data attributes used were the average temperature, relative humidity, wind speed, and rain when the fire was detected. Cortez and Morais also use the Forest Weather Index (FWI), an indicator of the

intensity of the fire. The distribution of the area attribute was skewed heavily to the right, and to combat this, a logarithmic transformation of $y = ln(x+1)$ was employed.

Attribute	Description
X	x-axis coordinate (from 1 to 9)
Y	y-axis coordinate (from 1 to 9)
Month	Month of the year
Day	Day of the week
FFMC	FFMC code
DMC	DMC code
DC	DC code
ISI	ISI index
Temp	Outside temperature (in °C)
RH	Outside relative humidity (in %)
Wind	Outside wind speed (in km/h)
Rain	Outside rain in (mm/m^2)
area	Total burned area (in ha)

Temporal variables were encoded using 1-to-N encoding, and all attributes were then scaled using Standard Scaler. After fitting the models, the outputs were post-processed using the inverse logarithm transformation. The negative values were rounded to 0.

The model performance was evaluated using Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). RMSE depicts how much the predicted value deviates from the actual value; MAE is similar, but instead of squaring the difference, the absolute value is taken. RMSE assigns a larger penalization to more significant errors due to the squaring, whereas MAE treats all errors equally. Either way, the model with the smallest error value is the best performing.

$$MAE = 1/N \times \sum_{i=1}^{N} |y_i - \hat{y_i}|$$

$$RMSE = \sqrt{\sum_{i=1}^{N} (y_i - \hat{y}_i)^2 / N}$$

Cortez and Morais tested out different combinations of the different types of variables; specifically, they tried Spatial-Temporal-FWI (STFWI), Spatial-Temporal-Meteorological (STM), FWI, and Meteorological (M) in their experiments. The best model they achieved for the MAE was under the M setup with the SVM model, but under the RSME criterion, the best-performing model was the naïve average predictor. This contradiction was reasoned to be due to the nature of the error criteria. That is, the RMSE is more sensitive to outliers compared to MAE.

Data

The forest fire data used in this study was obtained from the data published by Short [2]. This dataset contains spatial wildfire occurrences in the United States from 1992 to 2018. There was a total of 963,640 observed fires in the dataset. Each fire had its discovery date, longitude and latitude coordinates, year, state, size in acres, and size class recorded. The fire size class is determined by the National Wildfire Coordinating Group (NWCG) standard depicted in Table 1. Any missing values in the data set were dropped using listwise deletion as they were missing at random.

Table 1:

Size Class	Area of fire
A	< 0.25 acres
B	0.25 – 10 acres
C	10 – 100 acres
D	100 – 300 acres
E	300 – 1000 acres
F	1000 – 5000 acres
G	> 5000 acres

Using the coordinate and discovery date data of the fires in Short [2], the VisualCrossing Weather API was used to collect historical weather data. Visual Crossing sources the weather data from the National Digital Forecast Database provided by NOAA and MADIS. Some locations did not have any weather station nearby to record the data, and those forest fire instances were dropped from the dataset. The API was used to collect the weather data near the discovery location on the discovery date. In the end, a total of 31,361 fire instances had all weather data. The weather metrics provided by the API are summarized in Table 2.

For this study, fire size is the dependent variable or variable being predicted. Similar to the study by Cortez and Morais, the independent variables will be varied to find the best model for predicting the size of forest fires in the United States. The spatial variables in this study are the longitude and latitudes of the fire discovery site. Since it is unlikely for there to be a true linear relationship, a zoning or clustering of fires by location is employed.

Table 2:

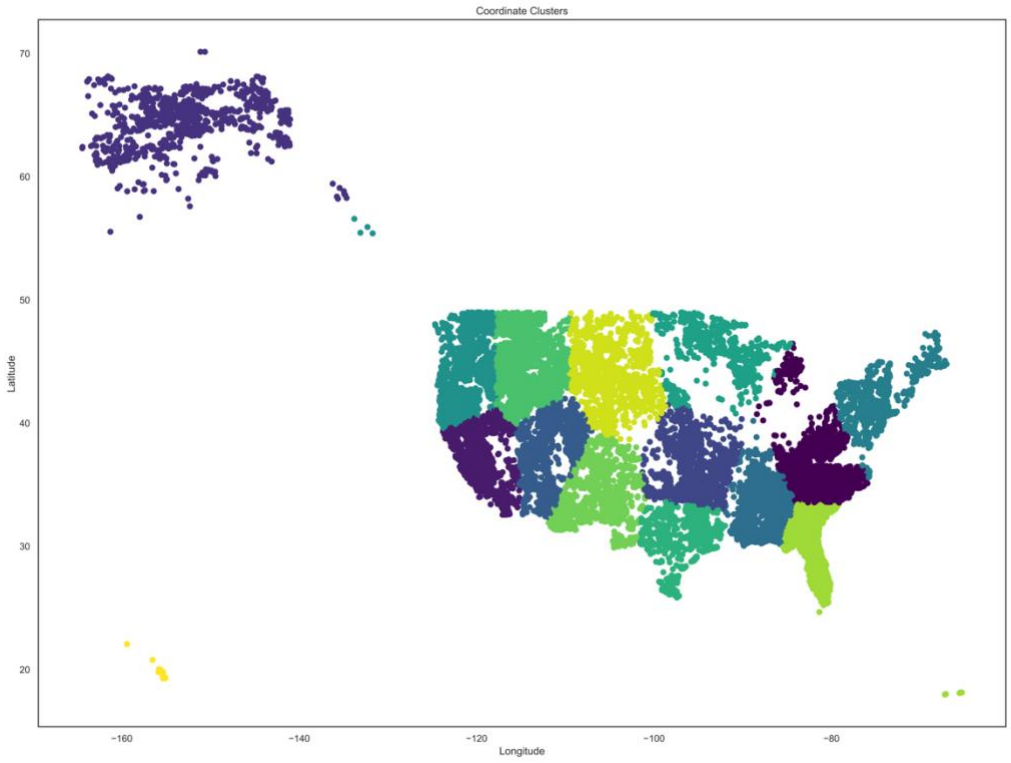
Metric	Description	Units
Temperature	Daily average temperature	F
Precipitation	The amount of precipitation	inches
Wind Speed	The maximum wind speed	mph
Relative Humidity	Amount of water vapor present in the air	%

The clustering of the coordinate data was done using the method of K-means. K-means is an unsupervised machine learning algorithm where a group of K centroids is randomly placed in the data space then the algorithm iteratively calculates the optimal position of the centroids form the clusters in the data. The number of clusters was determined based on domain knowledge. According to the EPA, there is a total of 15 Level I Ecological Regions in America [6]. Thus, k was set to 15 to replicate these regions. Since there is no ground truth for the vegetation in the

data, an intrinsic measure was employed to validate the clustering quality, more specifically, the Calinski-Harabasz Index [7]. Since the number of clusters is fixed, the location of the centroids was varied, and the optimal model was chosen based on its Calinski-Harabasz Index. The optimal model had a score of 73694.85. The clusters are depicted in *Figure 1*.

The temporal variables considered in this study were the day of the week, and the month the fire started. The month variable can indicate if there is a seasonal cause that affects the size of wildfires, and the day of the week indicates if there is an association between the day and the size of the fire. Thus, the day of the week may indicate if it is a weekend, thus increasing the possibility of human activity. For example, if we look at the data, more fires appear to be initiated by human activity *(Fig 2)*. Most of the forest fires, in general, tend to be on the smaller side (Fig 3) but looking further into the class distribution of fires caused by humans and naturally occurring ones, there is a noticeable difference in sizes. Naturally occurring fires tend to be much

Figure 1:



larger with class sizes of F, G, and E, whereas human-caused fires are smaller with class sizes of B, C, and D (*Fig 4-5*). Table 3 summarizes the attributes that are to be used in this study.

Figure 2:

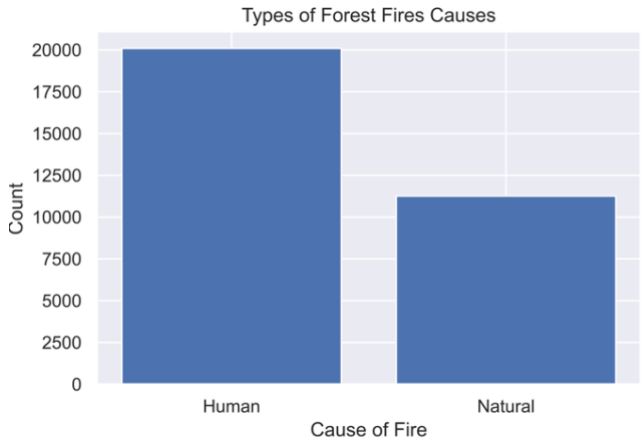


Figure 3:

Figure 4:

Figure 5:

Figure 5

Table 3:

	Attribute Description
Temp	Outside temperature (in °F)
RH	Outside relative humidity (in %)
Wind	Outside wind speed (in mph)
Rain	Outside rain in (inches)
Cluster	Cluster the coordinate belongs to (1-15)
Month	Month of the year
Day	Day of the week
area	Total burned area (in acres)

Data Preprocessing

Before the variables could be used in building models, some preprocessing was done to the attributes. The distribution of the burned area depicted in Figure 6 is heavily skewed to the right, similar to the distribution in Cortez and Morais's study. Likewise, a logarithmic function of $y = ln(x + 1)$ was applied to the dependent variable to reduce the skewness (Figure 7); this transformed variable is the target output of the models.

The spatial variable, *cluster*, is nominal in nature, and thus it was transformed using one-hot encoding. One-hot encoding is where each categorical value is converted to a column, and a value of 1 is assigned to the column based on the initial value. This transformed variable will be

Figure 6:

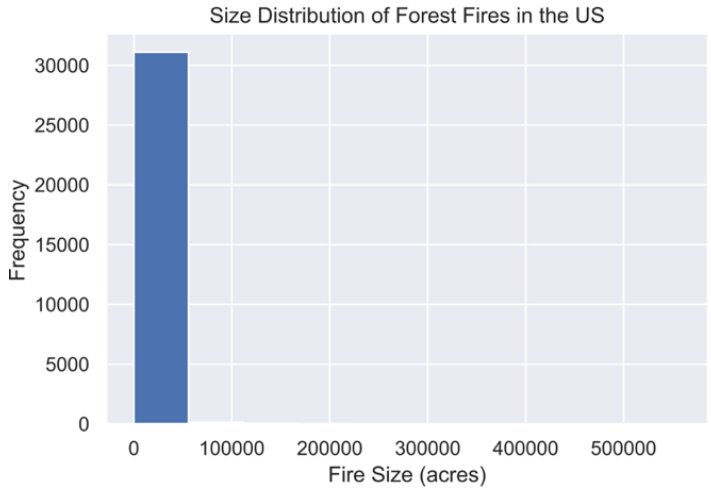
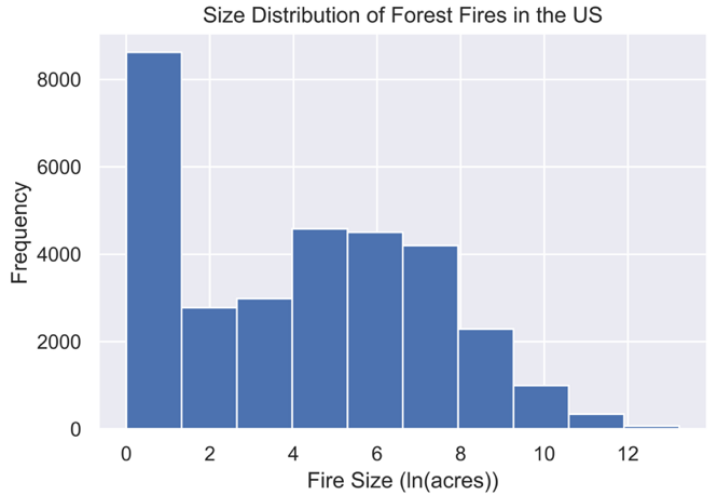


Figure 7:



used for all the models. The temporal variables are cyclical, and thus they were encoded using a sine and cosine transformation [8]. This transformation will allow us to see how impactful seasonality is on the size of wildfires.

Although the independent variables, *temperature* and *relative humidity* have similar data ranges, *precipitation* and *wind speed* do not (Figure 8.2). To ensure that the variables are able to contribute equally to the fit of the model, a standard scaler transformation is applied to the weather variables. Standard Scaler is used to adjust the distribution of values in the dataset so that the mean of the values is 0 and the standard deviation is 1. These variables are also the only quantitative variables in this study and looking at Figure 8.1, we can see that there are no features that are highly correlated, thus there is no issue of multicollinearity among the features in this study.

Figure 8.1:

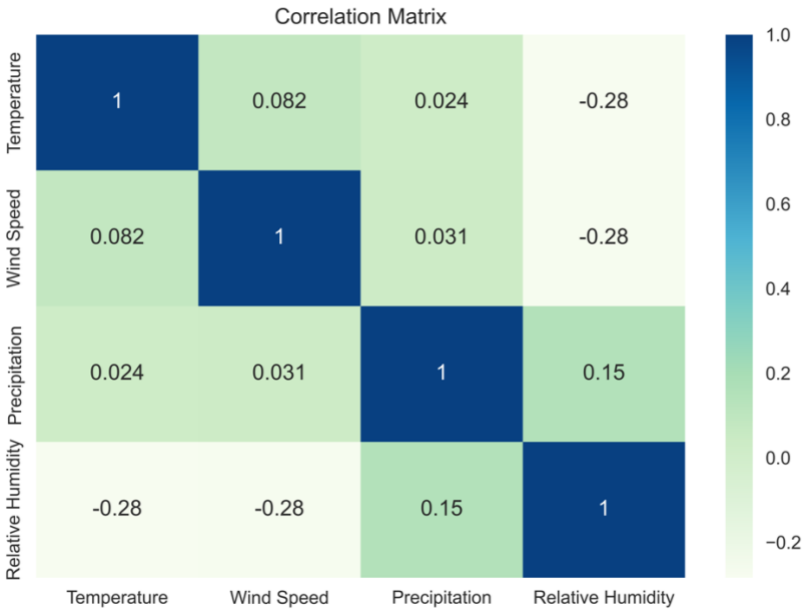
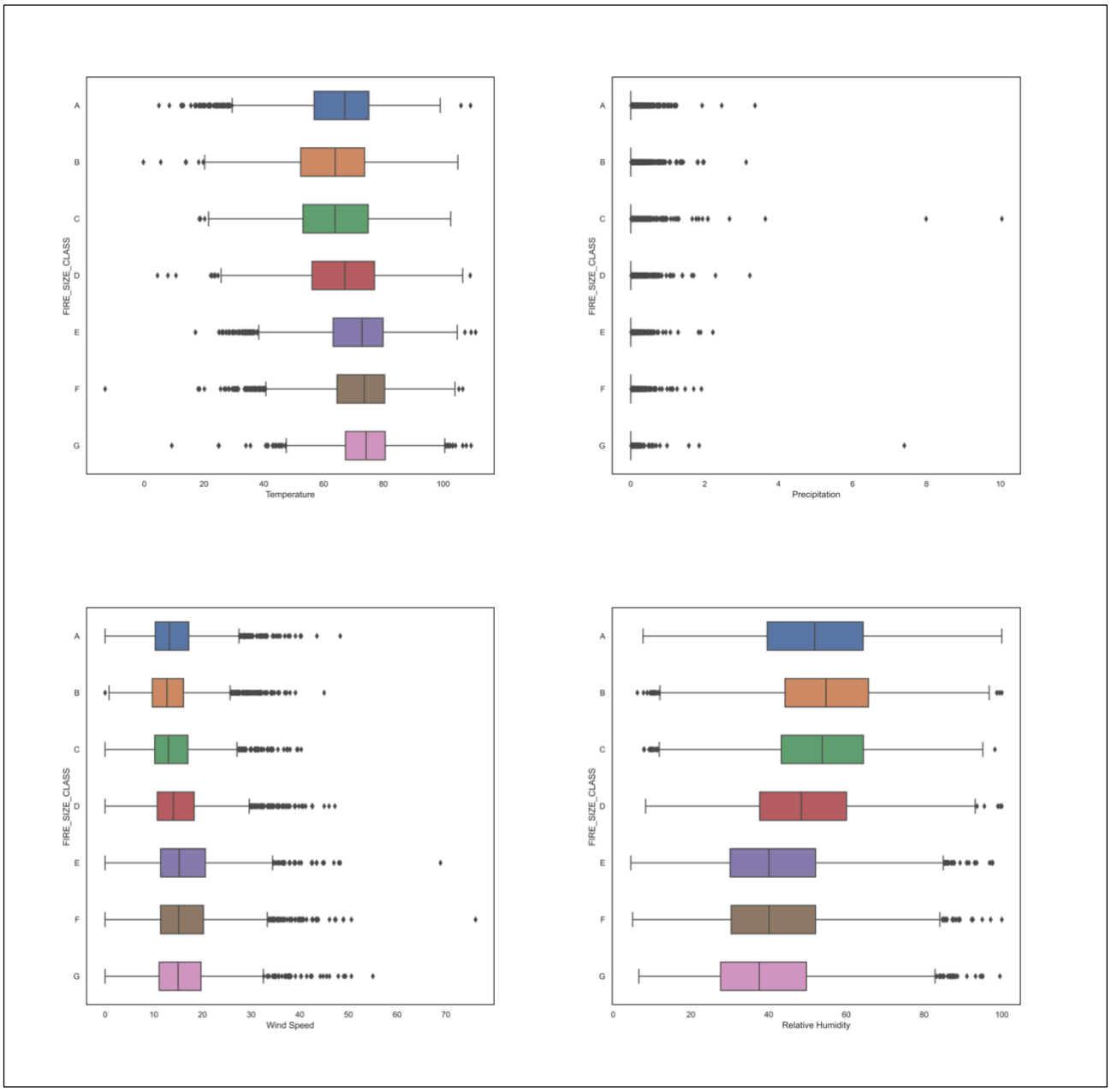


Figure 8.2:



Approach

This study consists of two parts; the first part uses regression techniques to predict the size of the burned area of fires in the United States. The method of approach used in this part

closely follows the one in the paper by Cortez and Morias. The best performing model and predictors are chosen based on the evaluation metrics of RSME and MAE. Once the best predictors are found, the second part will use them in different classification models to find which method best predicts the *size class* of forest fires. Predicting the size class would increase the odds of being accurate since the margin of error is bounded by the minimum and maximum size for each class. This would also help authorities to prepare the appropriate amount of resources for a given *fire class size*.

Part I

Some differences between Cortez and Morias's approach and the one in this study are the exclusion of the FWI attribute and the inclusion of clustered spatial data. The experiment space is split in five different ways, where the fifth experiment, which excludes the weather variables, is the naïve predictor or the benchmark metric. The first experiment uses spatial, temporal, and weather data to predict the burned area. The second and third experiments use weather and switch between temporal and spatial variables, respectively. The last experiment uses weather data to predict the fire size.

For each experiment, all the data was used in five different machine learning models that were fitted and evaluated. There was 75-25 split for the training and test datasets from a total of 31361 datapoints. All five methods used in this study are similar to Cortez and Morias's approach. The models used in this study are Multiple Regression, Support Vector Machines, Decision Trees, Random Forests, and Neural Networks. After the models are computed they are compared against each and the best model is chosen based on its RSME and MAE values.

For the multiple regression models there is one dependent variable and two or more independent variables. It is a linear model that fits coefficients $\omega = (\omega_1,\ldots\omega_n)$ (where n is the number of independent variables) to minimize the residual sum of squares between the actual and predicted values. The regression models were set up with the default settings similar to the approach by Cortez and Morais.

Support Vector Machines for regression is a nonparametric technique (no assumption is being made about the data) since it depends on kernel functions [14]. The use of kernels helps in using a linear method to solve a non-linear problem. Compared to MR the goal of SVR is to minimize the coefficients rather than the squared error. In the experiments the RBF kernel was used to fit the model similar to Cortez and Morais.

The next modeling methods used are Decision Trees and Random Forests, these are both nonparametric methods. The objective of a Decision Tree is to predict the size of the fire by learning rules from the data features. Random Forests on the other hand are an ensemble of many decision trees. The method of bagging is used to build a random forest, this is when the individual decision tree is used are as parallel estimators. In the case of regression, the value predicated at the leaf node is mean of the target values for that leaf [9]. All the models used the default hyperparameters for both the methods.

The last method used is a neural network, there are three main parts to a neural net the input layer, processing or hidden layers and the output layer. For the model used in this study the input layer consists of a dense layer with 64 units and *relu* activation. This layer then connects to 3 similar hidden layers which the connects to the dense output layer with one unit. The model is compiled with a loss function of *mean squared error*, optimized with *adam*, and uses *mean*

absolute error as a metric. Each experiment in the study uses this same model to compute the size of the forest fire.

Table 4:

Experiments	
Experiment 1	Spatial, Temporal and Weather
Experiment 2	Temporal and Weather
Experiment 3	Spatial and Weather
Experiment 4	Weather
Benchmark	Spatial and Temporal

Part II

The approach for the second part of this study is to create different machine-learning classification models. After the models are computed, metrics such as precision score, recall score, f1-score, and accuracy will be used to decide on the best model. A confusion matrix and a ROC-AUC analysis is also computed for each model. The confusion matrix summarizes the prediction results and shows the true positives, true negatives, false positives, and false negatives of the model prediction. The Area Under the Curve Receiver Operating Characteristic analysis is an evaluation metric that measures a classifier's ability to differentiate between the various classes. It is plotted using the True Positive and False Positive rates at different threshold values. The higher the value, the better the classifier predicts the fire class size. Since the problem at hand is a multiclass classification problem, the macro-average of each metric will be considered since this metric gives equal importance to all classes [11].

This part of the study employs classification models such as Support Vector Machines, K Nearest Neighbors, Decision Trees, Random Forests, Multinomial Logistic Regression and XGBoost. The features used in the models are from the experiment in part 1 that performed the

best (Experiment 1). The data in this part was balanced to ensure that each class is prioritized equally, and no bias is added. Balancing the dataset resulted in 2460 data points for each class, with a total of 17220 data points. Each model was trained and tested on the same randomly sampled train and test datasets and evaluated with the abovementioned metrics.

For classification with KNN, the class value is determined by the average of its k neighbors. The value of k in the KNN model is significant; thus, the grid search technique was utilized to find the best optimal k for the model. The optimal k from the grid search was 53. The grid search also yielded the best weight hyperparameter of *distance;* the default values were used for the other hyperparameters. The support vector machine model also utilized grid search to determine the best kernel for predicting the fire classes. The results from the search indicated that the RBF kernel was the best choice for this dataset. Since this is a multiclass classification problem and the support vector machine is designed for binary problems, the approach of One-vs-One is used to fit the data to the model. One-vs-One is a strategy that divides a multiclass classification task into one binary classification for each pair of classes [10]. The default was used for the rest of the hyperparameters.

Since tree-based algorithms are not distance-based the transformations done on the features in experiment 1 is not required [12]. Thus, the features used in the decision tree, random forest and XGBoost models are the untransformed spatial, weather, and temporal variables. A grid search was run to find the *max_depth* values for decision tree and random forest, and the search resulted in a value of 10 for both models. For the XGBoost model *objective* was set to 'multi:softmax' and the hyperparameters *max_depth, learning_rate and subsample* were found using gridsearch with values 0.05, 5 and 0.6. The default value was used for all the other hyperparameters. The last method used in the classification task is the multinomial logistic

regression this algorithm uses a One-vs-Rest approach and cross-entropy metric for loss. One-vs-Rest is a strategy where a multi-class classification is split into one binary classification by class [11]. The *solver* and *penalty* used in the model set up are *saga* and *l2*. The rest of the hyperparameters were set to the default value.

The best model is determined by looking at their performance metrics, especially the F1-score and the ROC-AUC score. After the best model is found, the model is then interpreted by looking at its feature importance scores. The permutation feature importance metric will be used to get a better understanding about the feature that the models think are more important. This metric is a model agnostic method and measures the prediction error increase of the model after the values in a feature column are permuted [15]. One disadvantage of using this method is if there is some correlation among the features, correlation between features can cause a decrease in the importance of the feature. But this is not an issue in the case of this study as indicated inFigure8.1.

Results

Part I

The results from all the experiments are summarized in *Table 5*. The first value depicts the Mean Absolute Error, followed by the Root Mean Square Error and that last metric is the percent of the values that were accurately predicted in the fires size classes. This metric is used to get a better insight into how well the regression predictions fall within the bounds of the different fire sizes.

Experiment 1: Spatial, Temporal, Weather variables

In this experiment the features used were the clusters from the K-means, the sine and cosine transformed month and day variables and the four different weather variables. The best

model based on the lowest RMSE and MAE metrics is the SVM model with values of 1.66×10^4 and 2676.55 acres respectively. The second-best model was the Multiple Linear Regression model which had a MAE of 2693.24 and a RSME of 1.66×10^4. The other models also performed relatively well beating the benchmark results of experiment 5 except for the decision tree, random forest and neural network models where the benchmark performed better with respect to the MAE metric but not accuracy.

Experiment 2: Temporal and Weather variables

In experiment 2 the spatial variables were dropped and only the weather features and the temporal feature were used as predictors. This resulted in models that performed much worse compared the last experiment. The best model for this set of features was the Support Vector Machine which had a MAE of 2719.22 and RSME of 1.67×10^4. The SVM model also had the best accuracy compared to the others, guessing 21.83% of the values into the right size class range. This experiment also yielded in models that performed worse than their benchmark counterparts thus suggesting that the importance of including the spatial variable.

Experiment 3: Spatial and Weather variables

In this experiment the features used to predict the burned area size were the spatial and weather variables. The predictors in this experiment resulted in models that are very close to the results from experiment 1. The best model is the SVM with a RSME of 1.66×10^4 and a MAE of 2676.65. The percent of the outputs that fell with the right class boundaries was 25.50% which is an improvement from experiment 2 but not better than the result in experiment 1. Similarly, all the models besides the decision tree, random forest and neural network beat the benchmark metrics. This indicates that perhaps weather might help with predicting the size of wildfires.

Experiment 4: Weather variables

The last experiment consisted of just the four weather variables as the predictors in models. The best model in this feature selection setup is the SVM which had a MAE of 2721.32 and RSME of 1.68×10^4. But compared to the benchmark results none of the models within this feature selection performed comparably. This result does not agree with the results from Cortez and Morias, where the weather variables were the best performing features.

From the results shown in Table 5, we can see that SVM overall produces the best models regardless of the features. The best performing features are STW, but SW is very close, with only a difference of 0.10 acres for the MAE. To check if this difference is significant enough to conclude that the model with STW as features is better, a paired t-test was conducted. If the null hypothesis is not rejected, then the difference in the metrics is not significant and the feature space with fewer features would be preferred but if the null is rejected then we can conclude that the feature space of STW are the best predictors of burned area size.

Table 5: MAE, RSME, Percent of predictions within Fire Class Size

Models	**Feature Selection Setup**				
	STW	TW	SW	W	ST
MR	2696.29 1.67×10^4 22.57%	2733.75 1.68×10^4 19.41%	2696.58 1.67×10^4 22.52%	2735.17 1.68×10^4 19.23%	2710.71 **1.67×10^4** 20.66%
SVM	**2676.55** **1.66×10^4** *25.86%*	**2719.22** **1.67×10^4** *21.83%*	**2676.65** **1.66×10^4** *25.50%*	**2721.32** 1.68×10^4 *21.05%*	**2700.49** **1.67×10^4** *24.63%*
DT	4644.87 2.16×10^4 23.90%	4788.91 2.24×10^4 21.00%	4704.70 2.31×10^4 23.68%	4912.00 2.22×10^4 19.39%	2712.40 **1.67×10^4** 22.36%
RF	2723.73 **1.66×10^4** 24.23%	2735.74 **1.67×10^4** 21.28%	2726.46 **1.66×10^4** 23.68%	2758.16 **1.67×10^4** 19.90%	2710.56 **1.67×10^4** 22.19%
NN	2738.01 **1.66×10^4** 25.36%	2733.72 **1.67×10^4** 20.57%	2710.88 1.67×10^4 23.59%	2732.50 1.68×10^4 19.05%	2708.46 **1.67×10^4** 22.16%

With a null hypothesis where the results of the model are equal and an alternative hypothesis of them not being equal, we reject the null as the p-value of 0.00145 is much smaller than $\alpha = 0.05$. Thus, there is sufficient evidence to claim that there is a difference between the results of the models. The SVM model with STW as its features is the best model out of all the ones tested.

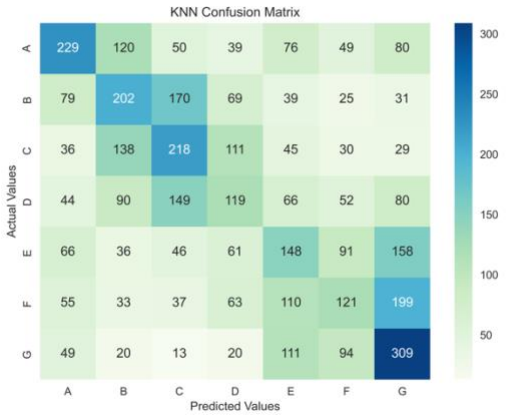
Cortez and Morais state that spatial and temporal variables were irrelevant in predicting fires, and their best model only had meteorological variables as its features. However, that is not the case with wildfires in the United States. The model that included the spatial and temporal variables performed better than the ones without them. A reason for this difference may stem from the fact that in this study, fires from across the US are being considered rather than concentrating on one specific region. Thus, there is bound to be a difference in vegetation and terrain affecting the fires' size across the US.

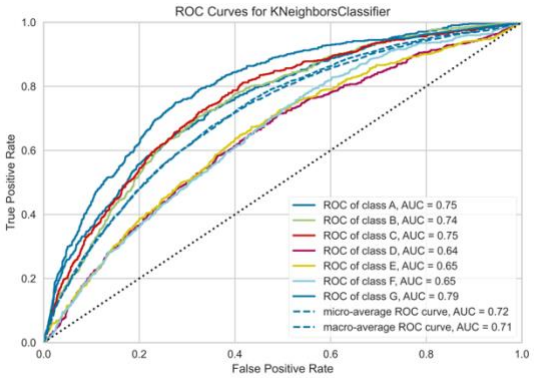
Part II

KNN

With k = 53 and weights set to ‘distance’ the KNN model accurately predicted the classes 31.26% of the time. Looking at the confusion matrix inFigure*9*. We can see that, for the most part, the matrix diagonal indicates that the model classifies the data in the correct category. For the cases of misclassified points, the model usually classifies them into a class close to the actual. This is probably due to the ordinal nature of the classes. The model was better at predicting smaller fires than larger ones (except for fires of class G), as indicated by the AUC values in Figure*10*. The KNN model guessed around 63.50% of the points within 1 class boundary of its actual size class.

Figures 9 and 10:

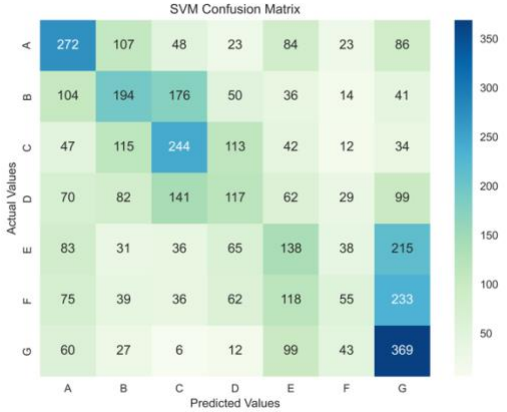


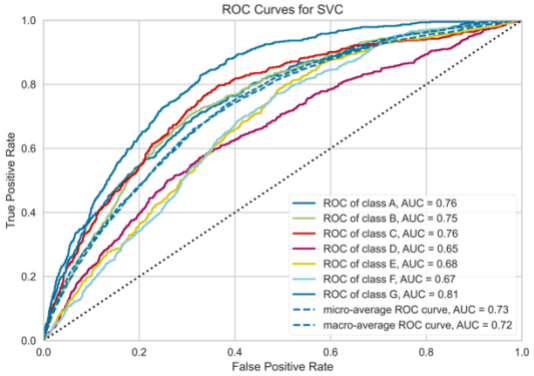


SVM

The support vector machine model with the RBF kernel and a *decision function shape* of OVO resulted in a model with an accuracy of 0.3189 and a ±1 class boundary accuracy of 0.6278. The macro-average precision, recall, and f1-score were 0.31, 0.32,0.30 respectively, and the model had a macro-average ROC-curve AUC value of 0.72. The SVM was best at predicting classes G, A, B, and C. the AUC values of classes D, E, and F did not break 0.70.

Figures 11 and 12:

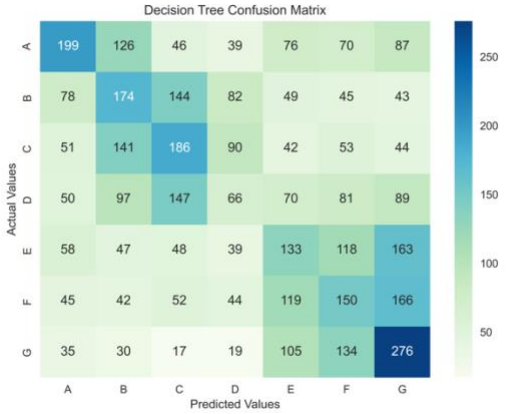


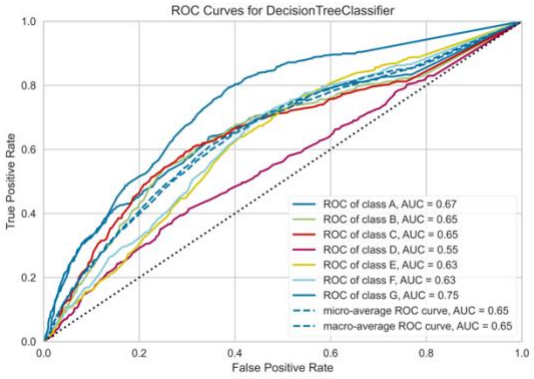


Decision Tree

The decision tree model did not measure up to the performance of the previous two models and had a general accuracy of 0.2750 and 0.5937 for the ± 1 class accuracy. The model yielded a recall score of 0.27, a precision score of 0.27, an F1 score of 0.27, and a ROC-AUC score of 0.65. The classes that were classified the best were classes A, B, C, and G, as indicated by the ROC-AUC scores in Figure 14.

Figures 13 and 14:

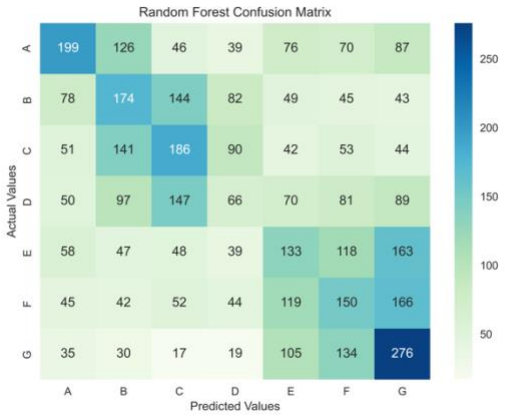


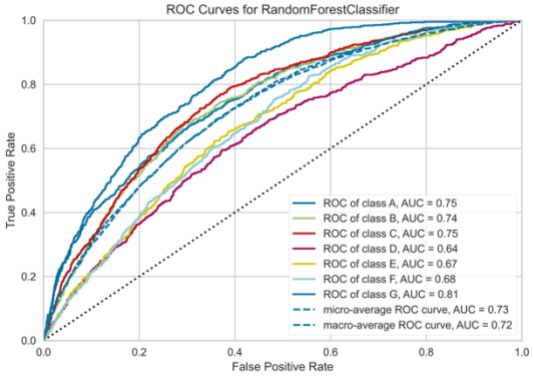


Random Forest

The random forest model accurately predicted 31.59% of the data in the test set into their correct class and 62.13% of within a one class boundary. The model also yielded a recall score of 0.31, a precision score of 0.31, a F1-score of 0.30 and a ROC-AUC score of 0.72. This model has a better chance of correctly segregating data points from classes G, A, B, and C compared to classes D, E, and F.

Figures 15 and 16:

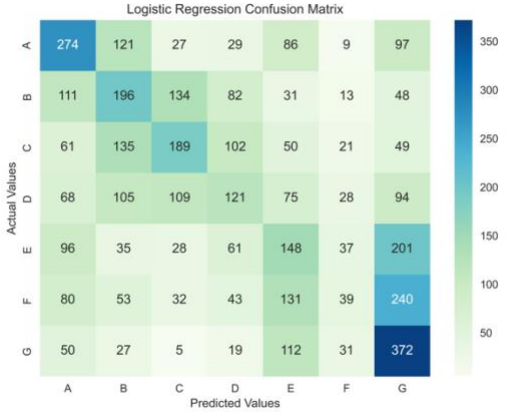


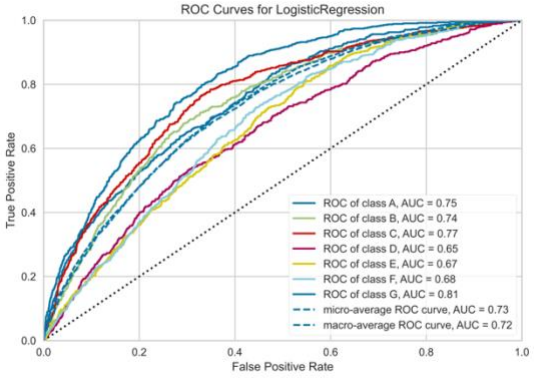


Multinomial Logit

The multinomial logistic regression model has an accuracy of 0.3126 and a ± 1 class error accuracy of 0.6153. The model had a precision score of 0.30, a recall score of 0.31, a F1-score of 0.29 and a ROC-AUC score of 0.72. Similar to the previous models the classes that were most accurately predicted were A, B, C, and G. All of these classes had an individual AUC score above 0.70 whereas classes D, E, and F and score around 0.65-0.68.

Figures 17 and 18:

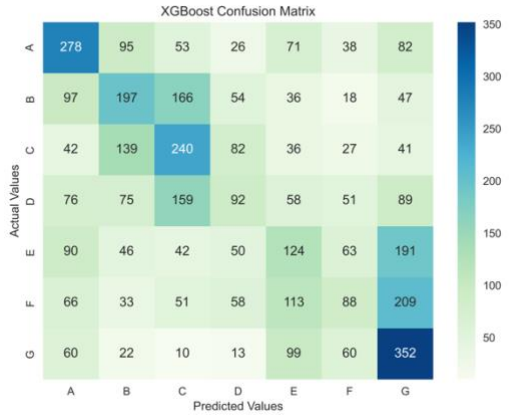




XGBoost

The XGBoost model performed on par with the other models with an accuracy of 31.84%. The classification accuracy of this model with a ± 1 class error is 61.84%. The model yielded a precision score of 0.30, a recall score of 0.32, a F1-score of 0.30 and a ROC-AUC score of 0.72. Looking at the confusion matrix, we can see that classes A, B, C, and G were more often labeled correctly compared to classes D, E, and F.

Figures 19 and 20:



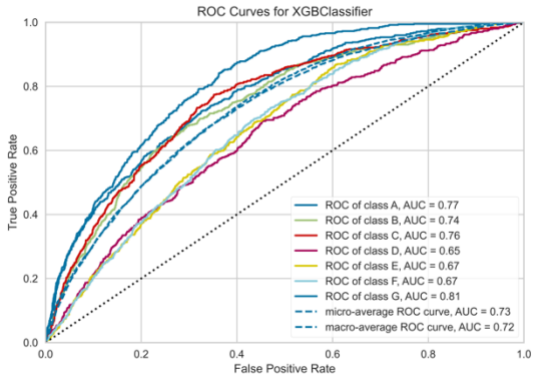


Table 6:

Model	Recall	Precision	F1-Score	ROC-AUC	Accuracy	Accuracy (± 1)
Decision Tree	0.27	0.27	0.27	0.65	0.2750	0.5937
Random Forest	0.31	0.31	0.30	0.72	0.3159	0.6213
K Nearest Neighbor	0.31	0.31	0.31	0.71	0.3126	0.6350
Support Vector Machine	0.32	0.31	0.30	0.72	0.3189	0.6278
Multinomial Logit	0.31	0.30	0.29	0.72	0.3126	0.6153
XGBoost	0.32	0.30	0.30	0.72	0.3184	0.6184

The table above summarizes the performance metric scores for all the models tested in this study. All the models had trouble accurately identifying fires of classes D, E, and F compared to classes A, B, C, and G. The models were more likely to predict class G if the fire was of size E and F. For the size D fire, all the models were most likely to classify it as one of its neighbor classes. A possible reason for this misclassification among these classes could be because the fire size classes are ordinal in nature. However, this can only partially be why the model misclassifies the larger fire classes because they can predict smaller fire sizes much better in comparison. All the models expect the Decision Tree had ROC-AUC scores above 0.7, this means that these models are relatively good at distinguishing between the different classes.

By first looking at the best ROC-AUC and F1 scores, the best models for this classification task are the Random Forest, Support Vector Machine, and XGBoost models. Since all of these models appear to be on par with each, another metric we can look into is interpretability. Interpretability in machine learning is the ability to explain a model in human-understandable terms [13]. The metric used to interpret the models is Permutation Feature Importance; this will show the most essential features in making predictions for the model. Figure 22 shows the feature importance for the SVM model; we see that Relative Humidity, Temperature is among the most critical variables, and variables such as day of the week and Precipitation, on the other hand, are not as important. Since the features for this model were encoded, some clusters might appear to be more important than others. Even if this is the case, all the cluster variables must be kept in the model. For the temporal variables, both the sine and cosine transformed day variables do not appear to be that important of a feature; thus, if we remove one of them, the other has to be removed as well.

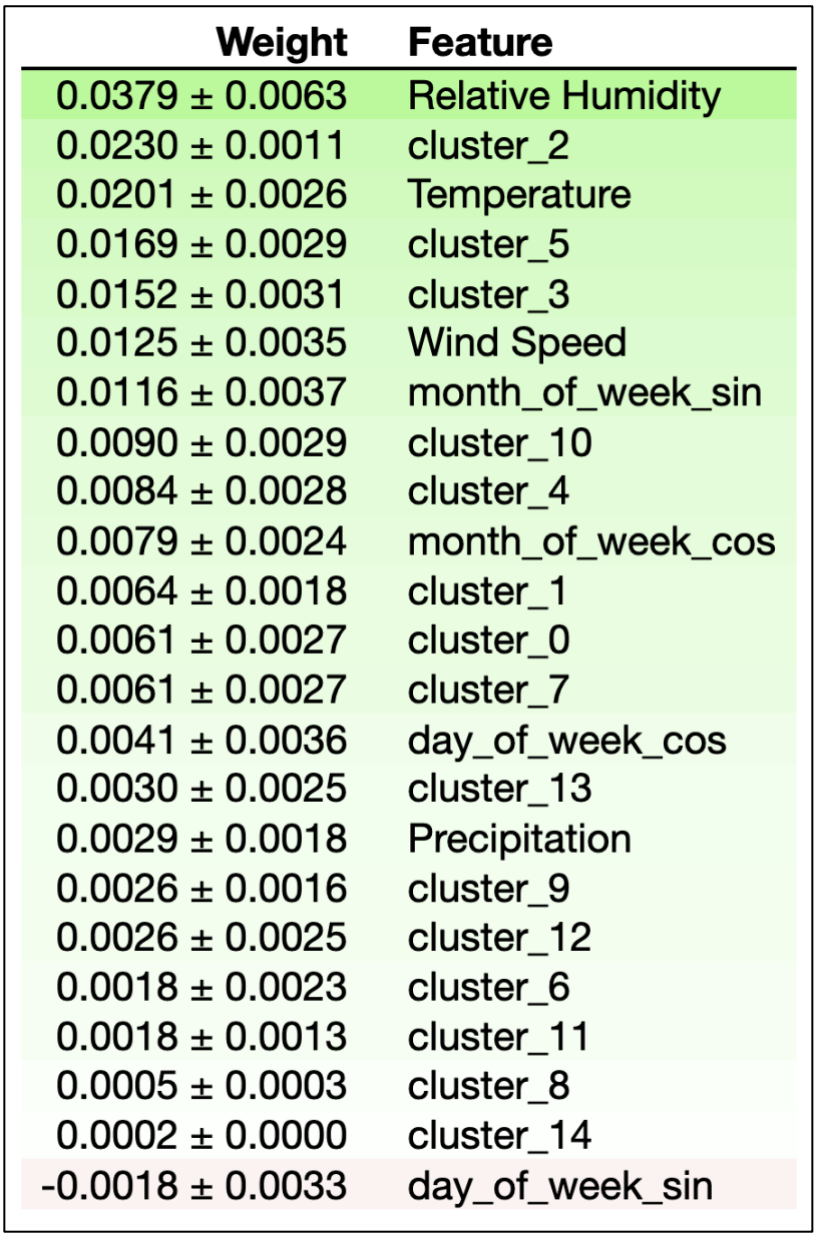
The permutation feature importance tables for the Random Forest (Figure 22) and XGBoost models are much easier to interpret since they do not employ any transformed features. Looking at the table for Random Forest, we see that the most important features were *cluster, month,* and *Relative Humidity* compared to features *Precipitation* and *day*. The XGBoost model, on the other hand, also places heavy importance on *cluster, month,* and *Relative Humidity* and lesser importance on *day* and *Precipitation*. In fact, the table (Figure 23) indicated that permutated *Precipitation* values did not increase the error but rather decreased it. The feature *day* also had a low impact on the models, disproving the assumption that the day of the week will help predict human involvement in starting the fires and thus help with fire size prediction.

Based on these results, the better model would be one that is much easier to understand, and in this study, that would be either the Random Forest or XGBoost models. The feature transformation of the SVM model makes it harder to interpret compared to the other two.

Figure 21: Random Forest PFI

Weight	Feature
0.0706 ± 0.0096	cluster
0.0650 ± 0.0068	month
0.0602 ± 0.0081	Relative Humidity
0.0388 ± 0.0099	Temperature
0.0151 ± 0.0086	Wind Speed
0.0040 ± 0.0074	day
0.0006 ± 0.0012	Precipitation

Figure 22: SVM PFI



Weight	Feature
0.0379 ± 0.0063	Relative Humidity
0.0230 ± 0.0011	cluster_2
0.0201 ± 0.0026	Temperature
0.0169 ± 0.0029	cluster_5
0.0152 ± 0.0031	cluster_3
0.0125 ± 0.0035	Wind Speed
0.0116 ± 0.0037	month_of_week_sin
0.0090 ± 0.0029	cluster_10
0.0084 ± 0.0028	cluster_4
0.0079 ± 0.0024	month_of_week_cos
0.0064 ± 0.0018	cluster_1
0.0061 ± 0.0027	cluster_0
0.0061 ± 0.0027	cluster_7
0.0041 ± 0.0036	day_of_week_cos
0.0030 ± 0.0025	cluster_13
0.0029 ± 0.0018	Precipitation
0.0026 ± 0.0016	cluster_9
0.0026 ± 0.0025	cluster_12
0.0018 ± 0.0023	cluster_6
0.0018 ± 0.0013	cluster_11
0.0005 ± 0.0003	cluster_8
0.0002 ± 0.0000	cluster_14
-0.0018 ± 0.0033	day_of_week_sin

Figure 23: XGBoost PFI

Weight	Feature
0.0779 ± 0.0086	cluster
0.0514 ± 0.0112	Relative Humidity
0.0476 ± 0.0117	month
0.0279 ± 0.0100	Temperature
0.0096 ± 0.0051	Wind Speed
0.0009 ± 0.0028	day
-0.0015 ± 0.0013	Precipitation

Conclusion

Predicting the size of wildfires has many advantages; it will allow officials to prepare better and allocate appropriate resources to mitigate the fire and also warn residents who live within the predicted area to evacuate before it is too late.

The best model from Part I of the study was the Support Vector Machine model with spatial, temporal, and weather features. This model had an RSME of 1.66×10^4, an MAE of 2676.55 acres, and predicted 25.86% of the fires within the correct class boundary. This result differs from the one presented in Cortez and Morais' experiment, where they concluded that the best configuration was the SVM model with the four meteorological variables. This difference can be attributed to the fact that their model was trained on fires from one area, Montesinho natural park, whereas in this experiment, fires from all across the United States were used. This difference resulted in lower accuracies in predicting fire sizes compared to Cortez and Morais, who were able to accurately predict 46% of the fires within 2.47 acres of the actual value.

Perhaps a way to improve the models would be to concentrate on fires in a specific region, for example, the Pacific Northwest, and check if the model has any improvements in prediction.

The best models for Part II of the study were the Random Forest and XGBoost models. These models had similar accuracies and ROC-AUC scores. The Random Forest had a ROC-AUC score of 0.72 and predicted 31.26% of the fires accurately and 62.13% within one class boundary. The XGBoost model accurately predicted 31.84% of fires and 61.84% within one class boundary. The most essential features in predicting these fire classes were cluster, month, and Relative Humidity. One way to improve the classification models is to remove the less critical features indicated in the Permutation Feature Importance metrics results. The results also indicated that the cluster feature was significant in predicting values. Since clusters were used as a substitute for vegetation data, getting the proper vegetation data for the fire regions may improve the outcome of predictions.

Another area for possible improvement is to look at ways to build models that deal with classes that are ordinal in nature. Using a standard classification approach loses the ordering information of each class since they are treated as unordered values.

Overall, the results are promising but also indicate much room for improvement in predicting the burned area of a fire or the size class. Another way the results can be improved is if more data is used for model building, less than 4% (31,361 fire instances) of the original 963,640 observed fires had all the weather variables needed from the API. Using a different method to collect the weather data may yield more data to train and improve the model.

References

[1] Cortez, P., & Morais, A. (n.d.). *A Data Mining Approach to Predict Forest Fires using Meteorological Data*. 12.

[2] Short, Karen C. 2021. Spatial wildfire occurrence data for the United States, 1992-2018 [FPA_FOD_20210617]. 5th Edition. Fort Collins, CO: Forest Service Research Data Archive. https://doi.org/10.2737/RDS-2013-0009.5

[3] Vizzuality. (n.d.). *Forest fires & climate change: Effects of deforestation on wildfires: GFW*. Global Forest Watch. Retrieved November 23, 2022, from https://www.globalforestwatch.org/topics/fires/#intro

[4] NOAA National Centers for Environmental Information (NCEI) U.S. Billion-Dollar Weather and Climate Disasters (2021). https://www.ncdc.noaa.gov/billions/, DOI: 10.25921/stkw-7w73

[5] BBC. (2015, September 15). California wildfires: 23,000 people displaced from homes. BBC News. Retrieved January 14, 2022, from https://www.bbc.com/news/world- us-canada-34238228

[6] Environmental Protection Agency. (n.d.). *Ecoregions of North America* . EPA. Retrieved November 23, 2022, from https://www.epa.gov/eco-research/ecoregions-north-america

[7] Caliński, T., & Harabasz, J. (1974). "A Dendrite Method for Cluster Analysis". Communications in Statistics-theory and Methods 3: 1-27.

[8] *Three approaches to encoding time information as features for ML Models*. NVIDIA Technical Blog. (2022, August 21). Retrieved November 23, 2022, from https://developer.nvidia.com/blog/three-approaches-to-encoding-time-information-as-features-for-ml-models/

[9] Soner. (2020, February 11). *Decision tree and Random Forest-explained*. Medium. Retrieved November 23, 2022, from https://towardsdatascience.com/decision-tree-and-random-forest-explained-8d20ddabc9dd

[10] Brownlee, J. (2021, April 26). *One-vs-rest and one-vs-one for multi-class classification*. MachineLearningMastery.com. Retrieved November 23, 2022, from https://machinelearningmastery.com/one-vs-rest-and-one-vs-one-for-multi-class-classification/

[11] Allwright, Stephen. "Micro vs Macro F1 Score, What's the Difference?" *Stephen Allwright*, Stephen Allwright, 19 Aug. 2022, https://stephenallwright.com/micro-vs-macro-f1-score/.

[12] *Does the random forest algorithm need normalization?* KDnuggets. (n.d.). Retrieved November 23, 2022, from https://www.kdnuggets.com/2022/07/random-forest-algorithm-need-normalization.html

[13] Doshi-Velez, & Kim, B. (2017). Towards A Rigorous Science of Interpretable Machine Learning.

[14] *1.4. Support Vector Machines*. scikit. (n.d.). Retrieved November 24, 2022, from https://scikit-learn.org/stable/modules/svm.html#classification

[15] Sand, M. (2020, May 23). *Model interpretability : Eli5 & Permutation Importance*. Medium. Retrieved November 24, 2022, from https://medium.com/analytics-vidhya/why-should-i-trust-your-model-bdda6be94c6f