Predicting Characteristics of Forest Fires in California

A Regression and Classification Analysis

Lucas Atayde
Computer Science
Rice University
Houston, TX United States
lsa4@rice.edu

Manaal Khan
Computer Science
Rice University
San Antonio, TX United States
mk90@rice.edu

Julie Street
Cognitive Sciences
Rice University
Atlanta, GA United States
jds21@rice.edu

ABSTRACT

Our research aim was to see whether we could predict characteristics such as a fire's cause, size, or duration by using a series of regression and classification models including Linear Regression, Ridge Regression, Neural Network Regression, Random Forest Regression, Random Forest Classification, Decision Tree Classification, and Neural Network Classification. Because worsening climate change and increasingly damaging forest fires in California are more relevant now than ever, being able to make predictions and pull insights from existing data to help us deal with our changing climate is worthwhile. Though our regression models showed that our hypothesis was wrong in terms of being able to predict a fire's size and duration using the past sixteen months of weather data, our classification models were able to determine the features that most accurately predict a fire's cause. Just as other relevant research has found, human factors are the most frequent cause of fires and the proximity of a fire to human-occupied land is the most significant factor for making this prediction.

1 INTRODUCTION

Forest fires are among the most devastating natural disasters in California. They cause both monetary and human losses, and can burn for months at a time. Due to factors such as global warming and severe droughts, forest fires have increased drastically with regard to the amount of time they burn, how quickly they are able to be extinguished, and the damage they cause in the last decade. For some context, NASA reports that eight out of ten of California's most devastating fires on record have occurred within the past five years[4]. Due to the pressing nature of this topic, we decided that it would be of interest to attempt to predict features of forest fires in California. We hope that our findings can aid in preventing forest fires or decreasing the amount of damage done by them.

When we began this topic, however, we wanted to see if we could make connections between global climate trends and forest fires in the United States. We quickly found, though, that this was a topic of too broad a scope, and that in order to achieve more meaningful results, we would need to narrow down our focus. We switched then, to simply predicting attributes of forest fires in California, specifically aiming to predict the cause of a fire, the number of acres the fire burned, and for how long it burned, given that it is in existence and certain factors surrounding the fire. Upon reviewing prior research, we found that there has been a lot of work done in the way of predicting whether or not a forest fire will occur given some features such as previous climate, etc, which differs from the goal of our project.

The first paper we looked at aimed to predict forest fires in Slovenia using data mining techniques such as logistic regression, random forests, decision trees, bagging and boosting ensemble methods[2]. This group found that bagging of decision trees resulted in predicting the occurrence of a forest fire with the highest accuracy, with their random forest model coming in second in terms of accuracy. From this paper, we learned the importance of bagging, which we used in one of our classification models.

The second paper we looked at used artificial neural networks to predict the existence of a forest fire[6]. In their paper, they spoke in length about the backpropagation step of their algorithm, and how they determined how many hidden layers their neural network used, which was helpful to us when we were designing and testing our neural networks. Their model was able to predict a fire's existence with a five percent error rate.

The third paper we looked at tested the Bayes Network, Naïve Bayes, Decision Trees, and Multivariate Logistic Regression to model and predict characteristics of forest

fires[1]. The authors used information like the fire's elevation and historical climate information such as temperatures, drought indices, river density, land cover, and the distance to human structures like roads and cities. They found that human factors were, by far, the most important cause of fires and predictor of where fires will start. In terms of models, they found Bayes networks and decision trees the most effective for predicting fires based on the data and features they used. Their paper inspired us to select decision trees as one of our classification-task models.

The final paper we looked at aimed to simplify their model for ease of use[5]. It uses an SVM that is only reliant on four aspects of the weather, i.e. temperature, rain, relative humidity and wind speed, to predict the existence of a fire, however, they found that this model was best suited for predicting smaller fires. Although we didn't end up using SVMs for any of our models, the paper still provided good insights into the topic.

KEYWORDS

Forest Fires, Prediction, Machine Learning, Linear Regression, Ridge Regression, Neural Network Regression, Random Forest Regression, Random Forest Classification, Decision Tree Classification, Neural Network Classification

Reference format:

Lucas Atayde, Manaal Khan and Julie Street. 2021. Predicting Characteristics of Forest Fires in California: A Regression and Classification Analysis. *Houston, TX, USA, 7 pages.*

2 DESCRIPTION OF THE DATA

We built up a hand-crafted dataset by merging two different sources: The first gave a comprehensive list of every documented wildfire in the United States from 1992 to 2015, reported by the National Fire Program. The data set included basic facts about the fire like fire size, fire start and end time, and fire causes. We merged it with a second dataset from the National Oceanic and Atmospheric Administration: They have weather stations set up all across the United States, providing public weather summaries for each state's subregions.

2.1 Data Cleaning. To construct the dataset we modeled off of, we took each wildfire in the dataset and paired it with the closest (in number of kilometers) NOAA weather station. For each fire, we then matched it with the global monthly weather summaries for the station for the 16 months prior to the fire. This will help us directly test our hypothesis that the prior year or more of weather in a region has a relationship with fire size, duration, and classification.

2.2 Feature Summary. Out of the over three hundred features we had, a majority of them were related to the prior weather near the site of the fire. Each fire was matched with climate data recorded from 16 months prior to the fire, such as minimum and maximum temperatures, the number of days the region got above or below certain temperature thresholds, average temperatures, and latitude and longitude. In addition to the fire-specific data, which included the day of the year the fire was discovered, the date the fire was discovered, the time the fire was discovered, the size of the fire in acres, the year, the day the fire was contained, the time the fire was contained, and the cause of the fire, each fire also had geographic data, like elevation and distance to weather stations which recorded the data on the weather.

3 REGRESSION MODELS

For the regression tasks, we aimed to predict the values of two variables, "FIRE_SIZE" and "DURATION," which detail the number of acres burned by the fire and the amount of time elapsed between when the fire was first discovered and when it was extinguished in days, respectively. Before running any models, we determined which features were the most relevant to these two variables by ranking them using their Pearson correlation coefficient. We found that the features that most affected the "FIRE_SIZE" all indicated that extreme temperatures during the same months of a year strongly influenced the size of the fire, and that the features that most affected the "DURATION" variable dealt with how cold the last six months had been at the weather station closest to the fire. We used a number of models to try and predict these variables, but unfortunately, the outcome of this task was unsatisfactory.

3.1 Linear Regression. Linear regression and ridge regression are simple models used to predict linear relationships between a feature vector and a response variable. Both forms of regression use a loss function based on the convex mean squared error function, and the models are among the fastest to train. Given that we have very high

dimensional data, linear regression and ridge regression were our first choices due to their low performance overhead and relative ease of implementation. Additionally, we have hypothesized and observed that past weather statistics correlate with fire size and duration. For example, how hot it was a month prior and a year prior to a fire is positively correlated with fire size. We started off the regression models with a simple linear regression to predict both aforementioned variables. We found that replacing the values for the area burned to their log helped to control the range of values better after running the model without doing so and getting accuracy levels worse than these, and then we split the data and ran a regression model for both variables, but the accuracy we were getting was still poor. For the “FIRE_SIZE” variable, our error rates were

Test mean absolute error: 0.70065056
R2 Score training set: 0.049671639
R2 Score testing set: 0.449282757,

and for the “DURATION” variable we had

Test mean absolute error: 0.723566208
R2 Score training set: 0.05807320576
R2 Score testing set: 0.05106583388.

We then moved to a ridge regression model to see if the reason our models were performing so badly was because of high correlation between the features we used.

3.2 Ridge Regression. Ridge regression distinguishes itself from linear regression by incorporating a regularization term in the model’s loss function. The regularization term penalizes the model for having abnormally large coefficients in its weight vector, which aids in preventing the model from overfitting the data. For the models, we used the same setup process as the previous linear regression models—the accuracy for the model predicting the fire duration did increase, but our accuracy for the fire size model was still just as bad. For the “DURATION” variable, we had

Test mean absolute error: 0.403486581
R2 Score training set: 0.0486411799
R2 Score testing set: 0.0544865846.

To better see the inaccuracy of the predictions made by these models, refer to Figure 4.2.1, a simple plot which shows the values from the test set against what value our model predicted for the “DURATION” variable. We can see that at a high level, the predictions follow an upward trend as the true values increase in size, but for fires that didn’t last very long, the predictions are scattered across the entire y-axis, meaning that the model was not able to predict those values with any significance. We believe however, that because past weather is linearly correlated with fire size and fire duration, there may be interfering variables. Due to human intervention (mainly fire departments) limiting the scope of fires with better efficiency each year, we hypothesize that past weather influences the scale and duration of wildfires less than expected. Even when only training a model over a given month of the year, the linear models struggle to achieve a R^2 score beyond .05. That being said, the mean absolute error of each model is decent.

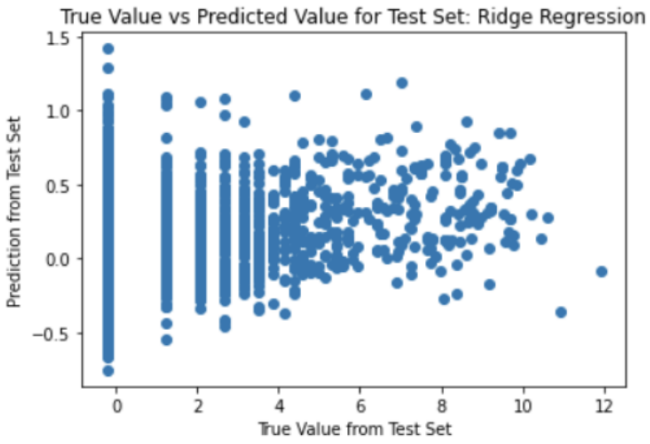

Figure 3.2.1: **A graph of the true versus predicted values for the ridge regression model**

3.3 Neural Network Regression. We then moved to a more complex model to see if that would increase our accuracy values and ran this model on only the “FIRE_SIZE” variable. Unlike linear regression and ridge regression, neural networks abide by the Universal Approximation Theorem, meaning that they are capable of estimating arbitrary functions, making them well suited for complex classification and regression tasks[3]. The models accomplish this by training a set of weights and biases, used as parameters in a deeply composed network of functions, trained using the backpropagation algorithm. Seeing as our linear models resisted fitting to the data, even when using polynomial features and hyperparameter tuning, we felt that a neural network would be suited for our regression task.

Out of the different network layouts we tried, we found that a sequential neural network with two hidden layers and

a dropout layer performed best on the dataset. Compared to our ridge regression model, the mean absolute error of the neural network demonstrated a slightly smaller mean absolute error:

mean_squared_error: 0.9087 - mean_absolute_error = 0.6778.

Despite this, the model still wasn't particularly strong based on the plot:

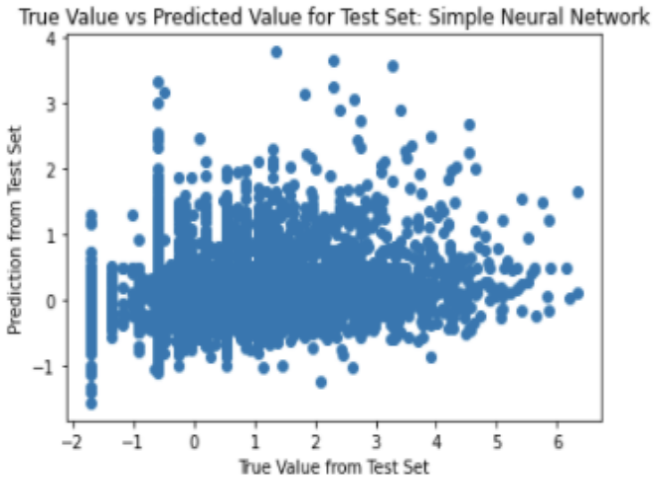

Figure 3.3.1: **A graph of the true versus predicted values for the simple neural network model**

Unfortunately, we see very little to no correlation between the true and predicted values, despite the more involved model.

3.4 Random Forest Regression. As a final effort before we deemed that our hypothesis was incorrect, we attempted a random forest model for both outcome variables. Random forest models are primarily used for classification tasks, but in a regression setting they are capable of estimating functions using a piecewise approximation, composed of multiple constant functions. When diagnosing our data, we observed that for smaller fire sizes and duration, the variables were often rounded to the nearest tenth, or hundreth. This indicated to us that a random forest model may be capable of more accurately predicting the rounded values because intervals of the function could capture the rounding function accurately. We set up and ran the random forest model the same way as the previous two models--split the data into training and testing sets with an eighty/twenty split and then used the sklearn package for a RandomForestRegressor and set the number of estimators to be 100, which we found was the optimal value. Out of the regression models we tried, the random forest models performed best for predicting both fire duration and fire size. Both outperformed their counterparts in both the ridge regression models and neural network models, below are the error values we calculated:

"FIRE_SIZE" Test mean absolute error: 0.64353722131
"DURATION" Test mean absolute error: 0.32429512347

Despite the multiple models tried and the various tuning efforts, our accuracy values weren't as high as we hoped they would be. We saw that the "DURATION" variable models yielded better accuracy values than the models for the "FIRE_SIZE" outcome variable, but overall, the models don't seem to support the hypothesis that past weather influences the amount of land burned by a forest fire and how long it burned. As mentioned earlier, however, this may be a product of human intervention.

4 CLASSIFICATION MODELS

Because the data contains information on how each fire was started, we figured this could use it to predict a fire's cause. The classification models we ran all aimed to predict the cause of a forest fire, given certain attributes of the already existing fire. This variable encodes the different causes for the fire in integers, and examples of causes could be lightning, arson, etc. Our data set listed thirteen options for the cause of the fire, and two of the thirteen options were "Missing" and "Other." Because our classification models have to differentiate between about eleven classes, our accuracy rates will not be as high as a binary classification task's accuracy may be, and this was reflected in our models, but we were still able to attain quite high accuracy measures, especially with the decision tree and random forest models.

4.1 Random Forest Classifier Preparation. Random forest classification models can be viewed as an extension to the classic decision tree. A decision tree is a machine learning model used for classification. The model is trained by successively dividing a feature space into smaller and smaller partitions that maximize label homogeneity within each partition, while not violating any model constraints to prevent overfitting. Decision tree models make these space splitting decisions using metrics like Gini impurity or entropy-based information gain. Random forests are simply a collection of decision trees, each tree trained with slightly different feature subsets or importances. Aggregation algorithms, like majority vote, determine the output of a

random forest based on the guesses of each of its constituent decision trees.

In order to determine the optimal number of trees for our classifier model, we ran a series of tests and graphed the resulting out-of-bag scores. We started with twenty trees and worked our way up, ultimately landing on 182 trees as the optimal number because we saw a decrease in the out-of-bag scores after that. We ran these tests in three parts because each of the first two showed the out-of-bag scores continuing to go up, but as the number of trees went up, the time it took to run each trial became more significant. Hence, in order to save time, we had each trial pick up where the last one left off.

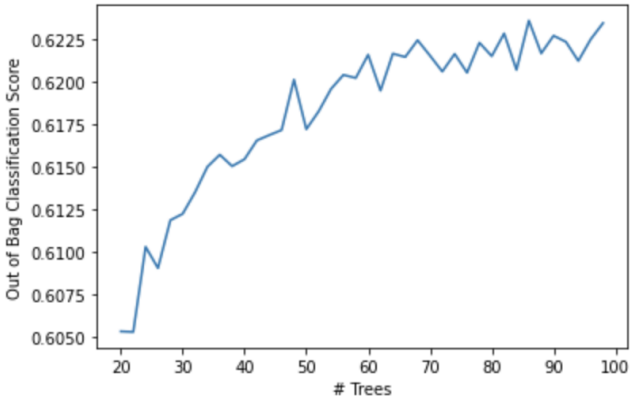

Figure 4.1.1: **Out of Bag Classification Scores for 20-100 Decision Trees**

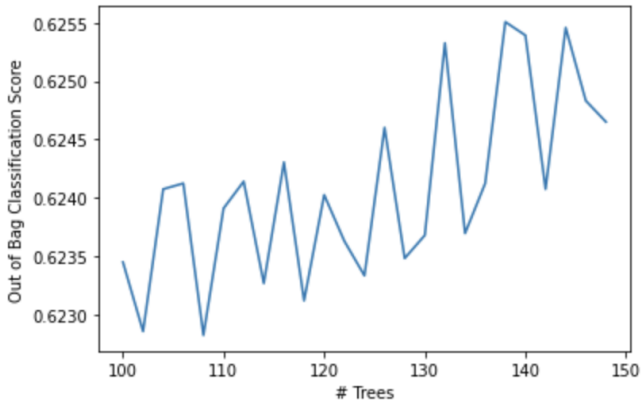

Figure 4.1.2: **Out of Bag Classification Scores for 100-150 Decision Trees**

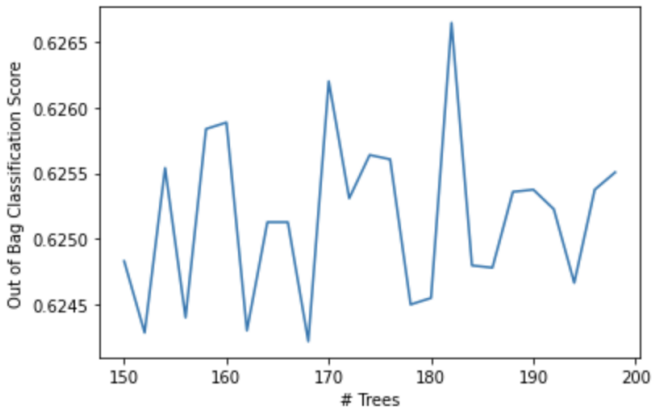

Figure 4.1.3: **Out of Bag Classification Scores for 150-200 Decision Trees**

4.2 Random Forest Classifier Results. We ran a Random Forest Classifier model on all of the features to determine which features were the most strongly correlated with the "STAT_CAUSE_CODE" variable, or how the fire was started. We were able to determine the features that were the most correlated, and have listed them below in addition to the forest score and out-of-bag classification score for a model using 182 trees.

1.0 (forest score)
0.6264445395581603 (oob score)

Feature 0: 0.214687 (BODIE)
Feature 4: 0.186169 (CONT_DATE)
Feature 3: 0.172743 (CONTAINMENT_MONTH)
Feature 1: 0.166481 (BROOKS)
Feature 2: 0.165398 (COHASSET)
Feature 5: 0.051952 (CONT_DOY)
Feature 6: 0.042570 (CONT_TIME)

Based on this, we saw that features 0, 1, 2, and 3 were the most important (each got roughly ¼ of the importance), which just means that the distance a fire is from the nearest station is the most important for classifying how the fire started. In other words, all of the fires closest to a given location are likely started the same way as each other, whereas fires closest to another location are more likely to be started the same way as each other (though the two ways could be different from one another).

4.3 Decision Tree Classification. To begin the classification process, we started with a simple decision tree to gauge what our accuracy values may be. A decision tree does what the name implies: it constructs a structure similar

to a flow-chart where each node is a "question" and the branches are the "responses" to the question. After splitting the data into a training set and testing set with an eighty/twenty split, we ran the model using the default options recommended by the sklearn package, and we were presented with these values:

Training Set Evaluation F1-Score=>0.997304304881060
Testing Set Evaluation F1-Score=>0.5464197530864198

Considering that the probability of choosing the cause of the fire correctly is a little less than 0.1, this model does a remarkably good job.

4.4 Neural Network Classification. We kept the first network pretty simple— it had two hidden layers of density 400, an input layer of density 233, and an output layer of size 14 (which corresponded to the 13 different outcomes for this variable), and 60 epochs. The result was a 47.6% accuracy, but we wondered if we could do better. We then created a more complex architecture, using the same sizes for the input and output layers, but using smaller hidden layers and one more "reset" for the data in the form of "Dropout" to prevent overfitting. The accuracy for this model was essentially the same at 47.4%, and in comparison to average accuracies for neural network classification models as a whole, ours performed pretty well.

It is amusing to see that, despite the more complex and involved nature of a neural network model, the simpler decision tree model was able to output a higher accuracy. This serves as a good reminder of Occam's razor, and the idea that oftentimes the simplest models are the most effective.

5 RESULTS

Regression Model Type	**Average Error**	**Mean Absolute Error**
Linear Regression	Size*: 0.538 Duration**: 0.56	Size: 0.7 Duration: 0.733
Ridge Regression	Size: 0.539 Duration: 0.559	Size: 0.7 Duration: 0.403
Neural Network Regression	Size: 0.9087	Size: 0.6778
Random Forest Regression	Size: 0.445 Duration: 0.043	Size: 0.644 Duration: 0.324

*** 'Size' refers to the fire size**
**** 'Duration' refers to the fire duration**

Figure 5.0.1: **Results table for regression models. Includes average errors and mean absolute errors for each model.**

Classification Model Type	**Measurement Type and Result**
Random Forest Classification	Forest score: 0.8585 OOB Score: 0.852
Decision Tree Classification	F1 Score: 0.546
Neural Network Classification	Accuracy: 0.4761 (60 epochs) 0.4741 (400 epochs)

Figure 5.0.2: **Results table for classification models. Includes measurements appropriate to each type of model.**

All of the code that we ran for this project has been uploaded to GitHub, and the repository is accessible at this link: https://github.com/lukeatayde/dsci_303_final_project.

DISCUSSION

As shown above, we used a number of models to try and predict three characteristics of forest fires given features about them and that they are already in existence-- how much time elapsed between when a fire was discovered and when it was extinguished, how many acres the fire burned, and what the cause of the fire was.

Beginning with the regression models, we found that they predicted the time and destroyage performed less optimally than we would have hoped. The error rates for those models ranged from around 0.7 to 0.4, and from the plots we included that depict the predicted value against the true value, we can see that there isn't much of a correlation between those values. This came as a surprise, since during the exploratory data phases of this project, we were able to see a correlation between previous weather and climate data and the existence of forest fires. Since we tried multiple models and tuned them to reach the highest accuracy possible, we can determine that either the features that were included in our dataset were not as

impactful as we had predicted they would be, or as was alluded to earlier, that human intervention has decreased the damage that forest fires have imparted on California per each fire, specifically, the number of fires have drastically increased with climate change, but the damage that each fire causes has been curbed in recent years due to technological advancements, which can cause issues for the models, since the dataset included forest fires from more than two decades ago. Despite these difficulties, we arrived upon the Random Forest Regression model as the most effective for both outcome variables.

The classification models did perform much better than the regression models. Here, we attempted to predict how a fire was started; some of the causes included lightning, children, electrical, etc. We used three methods to predict the causes, and we found that the simple decision tree yielded the best accuracy values, although they were not as high as a binary classification task's accuracy may be, since our model had to differentiate between eleven classes.

CONCLUSION

In the last decade, forest fires have increased dramatically, which is why we chose this topic. However, we wanted to look into a section of the problem that we didn't find much prior research on, as most papers that dealt with this topic aimed to predict whether or not a forest fire would occur given certain factors surrounding it. Not all of our models performed as well as we would have liked them to, but this is a virtue of the data science field. The process of this project showed us the virtue of persistence when it came to tuning models, that it is okay when the models perform suboptimally, and that oftentimes, the simplest models are the most effective. Overall, we found that the cause of a fire was easier to predict, but perhaps with a new dataset that included different features, we may be able to accurately predict damage done by a forest fire.

ACKNOWLEDGMENTS

We thank Dr. Akane Sano for her assistance and advice throughout this project.

REFERENCES

[1] Binh Thai Pham, Abolfazl Jaafari, Mohammadtaghi Avand, Nadhir Al-Ansari, Tran Dinh Du, Hoang Phan Hai Yen, Tran Van Phong, Duy Huu Nguyen, Hiep Van Le, Davood Mafi-Gholami, Indra Prakash, Hoang Thi Thuy. and Tran Thi Tuyen, 2020. Performance evaluation of machine learning methods for forest fire modeling and prediction. *Symmetry* 12, 6 (Jun, 2020), 1022. DOI: https://doi.org/10.3390/sym12061022

[2] Daniela Stojanova, Panče Panov, Andrej Kobler, Sašo Džeroski, and Katerina Taškova, 2006. Learning to predict forest fires with different data mining techniques. *Information Society.* (Jan, 2006). http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.116.2555&rep=rep1&type=pdf

[3] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. Autoencoders: Representational Power, Layer Size, and Depth. In Deep Learning. Cambridge (EE. UU.): MIT Press.

[4] NASA. (Sept, 2021). *What's behind California's surge of large fires?* EarthObservatory. https://earthobservatory.nasa.gov/images/148908/whats-behind-californias-surge-of-large-fires

[5] Paulo Cortez and Aníbal Morais, 2007. A data mining approach to predicting forest fires using meteorological data. (Jan, 2007). http://www3.dsi.uminho.pt/pcortez/fires.pdf

[6] Youssef Safi and Abdelaziz Bouroumi, 2013. Prediction of forest fires using artificial neural networks. *Applied Mathematical Sciences* 7, 6 (2013), 271-286. http://m-hikari.com/ams/ams-2013/ams-5-8-2013/safiAMS5-8-2013.pdf