Wildfire Prediction K-Means Clustering

Ashley Deibler

Data Science | Bellevue University

DSC 680 – T302 Applied Data Science

Professor Amirfarrokh Iranitalab

December 22, 2024

Business Problem

Forest wildfires, while ecologically necessary for some biological processes, pose threats and create large amounts of danger to natural areas, human property, and wildlife habitat when they occur in excess. I would like to build a model that uses various conditions and variables of previous wildfires to predict the location and nature of future fires to aid in allocating preventative resources and measures.

Wildfires pose many significant threats to humans, wildlife, and essential natural ecosystems. Some of these threats include destroying homes, businesses, and infrastructure, destroying animal habitat, impacting air quality, and contributing to increased soil erosion which further leads to poor water quality. While controlled fires are beneficial for certain ecosystems to flourish, uncontrolled wildfires can be detrimental.

Background/History

Modeling wildfires was established in the 1940s by mechanical engineer Wallace Fons during the turmoil of World War II. He reasoned that fire spread was majorly fueled by "successively heating neighboring fuel particles up to ignition temperature," and that the rate of fire spread is heavily reliant upon the amount of time it takes for different fuel types to ignite, and how far apart those fuel particles are (*The history of wildfire modeling*). Using this idea, Fons created and published a mathematical model for wildfire spread which was validated by experiments using pine needles. However, the purpose of this research and understanding the spread of fire was, at the time, primarily used as a weapon, as well as to better understand how to minimize catastrophe of fire in a wartime scenario. Fons, with the United States Forest Service,

explored what types of cover may be most effective from nuclear blasts, the way trees bent or broke, and blast effects on different tree species.

Jumping forward to the use of wildfire modeling today, the reasoning, method, and use of wildfire models is much different. Today, the Incident Command System, a department established after the California wildfires of the 1970s, runs and observes the models. The Incident Commander coordinates emergency responses based on the information gathered from the fire model output, while also considering the safety of crews, structures at the highest risk, how and from where the fire can be accessed, the nearest water sources, weather, and terrain.

The California Department of Forestry and Fire Protection (CalFire) uses an AI and machine-learning based tool called Wildfire Analyst Enterprise to predict the behavior of wildfires and to compare current and historical fires. This information is used to accurately allocate resources to the most at-risk sites and understand the most effective ways to fight a wildfire.

Data Explanation

The dataset used for this project was obtained from the University of California, Irvine archives. It consists of wildfire data from the Bejaia region in northeastern Algeria and the Sidi Bel-abbes region in northwestern Algeria between the months of June and September of 2012.

The dataset includes the following variables, environmental conditions and measurements as follows (Vizzuality):

1. Region: classifies observation based on location – Bajaia (1) or Sidi Bel-abbes (2).
2. Date: (DD/MM/YYYY) Day, month (between June and September), and year (2012).
3. Temperature: maximum temperature in Celsius; each observation measured at noon.

4. RH: Relative Humidity in percentage (%).
5. Ws: Wind Speed in kilometers per hour (km/h).
6. Rain: total rainfall per day in millimeters (mm).
7. FFMC: Fine Fuel Moisture Code index from the FWI system; a numerical rating indicating the moisture content of cured fine fuels and litter, used to define how flammable fine fuels are (Range: 0 - 101, with values greater than 91 classified as extreme).
8. DMC: Duff Moisture Code index from the FWI system; a numerical rating indicating the average moisture content of the organic layers in the forest floor, used to predict the likelihood of lightning ignition (Range: 0 - 1000, with values greater than 60 indicating extreme conditions).
9. DC: Drought Code index from the FWI system; a numerical rating that indicates moisture level of deep, compacted layers of soil, acting as a measure how easily fire can smolder within deep duff or large logs (Range: 0 - 1000, with values higher than 425 indicating more severe drought conditions).
10. ISI: Initial Spread Index index from the FWI system; a numerical rating indicating how quickly fire is likely to spread immediately after ignition. Based on the combination of Ws and FFMC (Range: no limit, although values higher than 10 are considered high, and higher than 15 are considered extreme).
11. BUI: Buildup Index index from the FWI system; a numerical rating of the total amount of fuel available for combustion. It is a combination of DMC and DC, reflecting the combined effects of daily drying and precipitation on fuels (Range: no limit, although values higher than 100 indicate extreme fire danger).

12. FWI: Fire Weather Index; numerical rating of fire intensity, based on the ISI and BUI, and is a general index of fire danger (Range: 0 - 99, with values greater than 50 classified as extreme).
13. Classes: separates observations into ‘Fire’ and ‘Not Fire,’ describing whether or not a fire was observed at that location on that date.

Methods

In order to accurately predict the occurrence of wildfires, we implemented two machine learning models to analyze the conditions and measurements present in the dataset. The first model implemented was a Decision Tree Classifier, evaluated using accuracy, precision, recall, F1-score, confusion matrix, and Gini index. The second model implemented was a Random Forest Regression model, evaluated using R2 score and mean absolute value. Before creating the models, however, an exploratory data analysis (EDA) was performed to gain valuable insights into the dataset and understand what values we are working with.

The EDA consisted of a series of visualizations including the following:

- Basic histograms to visualize the distributions for each numerical variable in the dataset;
- Shaded histograms to visualize the distribution of each numerical variable, but separated by “Class” to compare distributions;
- Grouped bar plots to observe count of ‘Fire’ and ‘Not Fire’ observations per month for both the Bejaia and Sidi-Bel Abbes regions;
- Boxplot to visualize the distribution of each non-categorical variable and identify the minimum, maximum, median, and quartile values;

- Individual bar charts for temperature, precipitation, FFMC, and humidity analysis.

The Decision Tree Classifier was chosen for this project because of its ability to handle complex, non-linear relationships between environmental factors contributing to wildfire risk prediction. It also provides the ability to see and interpret the decision rules behind said predictions, allowing for deeper understanding of which variables are most important in determining the occurrence of wildfire in an area.

The Random Forest Regressor was chosen because it is also able to assess and handle complex relationships between multiple environmental conditions and measurements and provides high accuracy in predicting wildfire risk. In addition, it is effective in identifying the most important contributing factors to fire occurrence, allowing us to accurately hone in on high-risk areas to focus preventative measures.

Analysis

Exploratory Data Analysis

The exploratory data analysis provided some very useful information to allow us to better understand our dataset's contents. The insights gained through the EDA presented the range of values during which wildfire counts, or the risk for wildfires, were highest, as shown in Table 1.

Table 1. EDA results for fire occurrence; reported values indicate conditions where risk of wildfires is highest for this dataset.

Variable	Conditions of Highest Fire Occurrence
Temperature (C)	30 - 37
Rain (mm)	0.0 - 0.1

Wind Speed (km/h)	13 - 19
Humidity (%)	50 - 80
FFMC	> 75
DMC	10 - 30
DC	> 25
ISI	> 3
BUI	> 10
FWI	3 - 25

Decision Tree Classifier

The initial decision tree classifier produced an accuracy score value of 0.9726. Feature importance and selection determined that the most important variables for the decision tree classifier were *Temperature, Ws, FFMC,* and *ISI*. The final decision tree model consisting of only those four features resulted in an accuracy score of 0.9726, and an excellent confusion matrix consisting almost entirely of True Positive and True Negative values (Figure 1 and Table 2).

The final decision tree classifier was then plotted (Figure 2) to visualize the decision rules implemented into the model. The tree classifies the decisions into ‘Fire’ and ‘Not Fire’ classes based on the variables’ likelihood to indicate the existence of a wildfire given stated conditions.

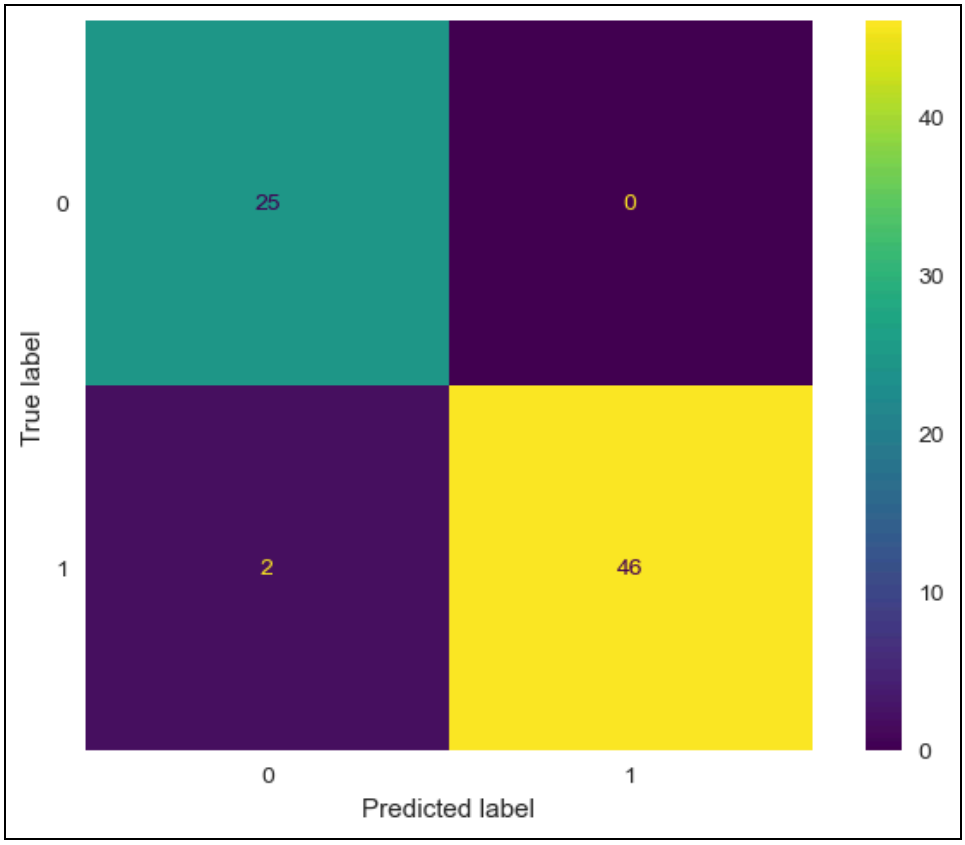

Figure 1. Confusion Matrix for finalized decision tree classifier consisting of variables *Temperature, Ws, FFMC*, and *ISI*.

Table 2. Classification report for final decision tree classification model.

	Precision	**Recall**	**F1-Score**	**Support**
Positive (0)	0.93	1.00	0.96	25
Negative (1)	1.00	0.96	0.98	48

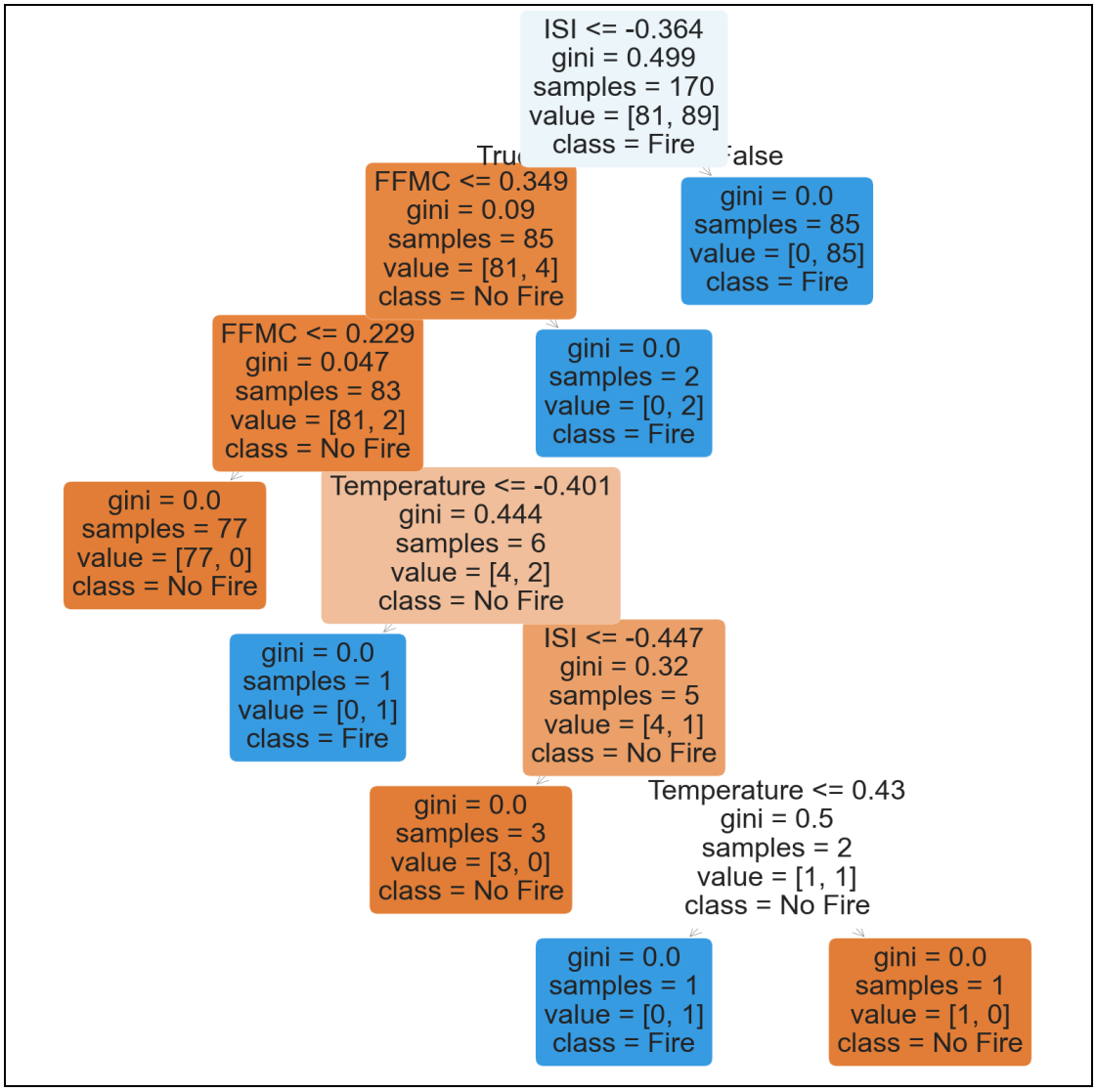

Figure 2. Visualized decision tree classifier including gini index values and classifications for variables *Temperature, Ws, FFMC,* and *ISI.*

Random Forest Regressor

The finalized, tuned random forest regressor model produced an R2 score value of 0.8847, and a mean absolute error (MAE) value of 0.0405. Feature importance and selection established that the most important features include *ISI, FFMC, DMC,* and *Rain* (Figure 3).

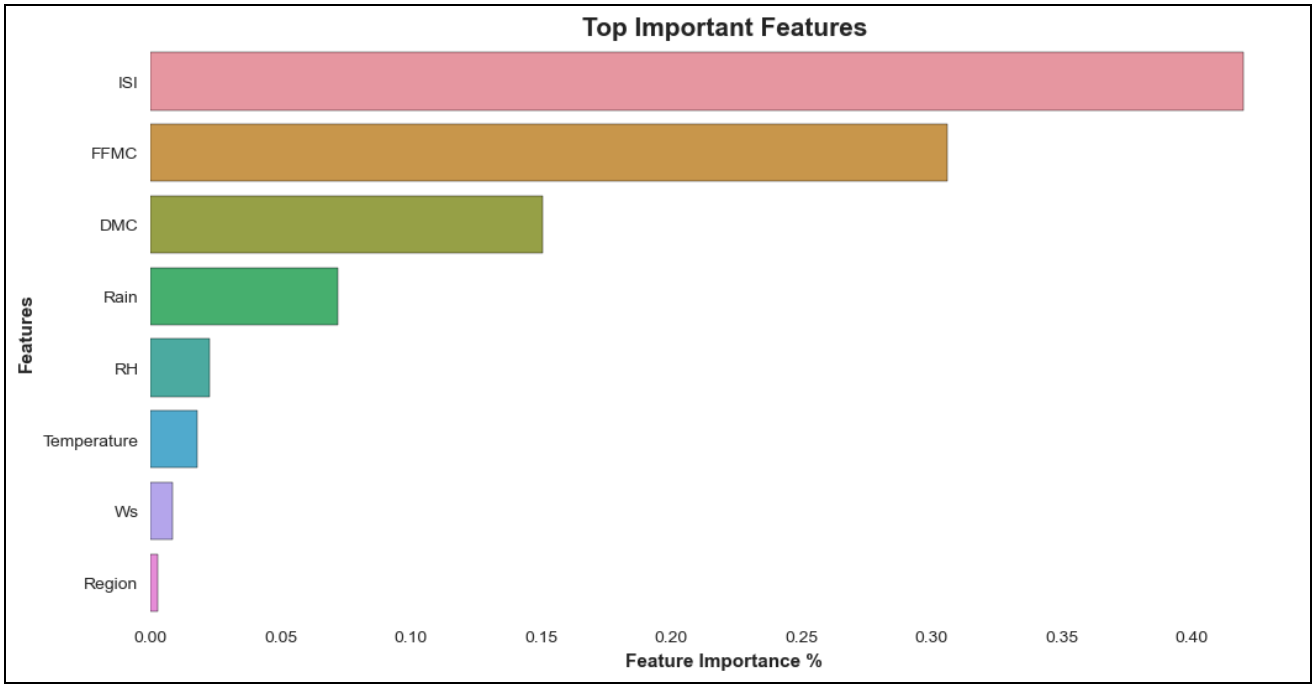

Figure 3. Feature importance for random forest regressor.

Conclusion

The decision tree classifier showed to be an effective and accurate model to predict the occurrence of wildfires based on particular conditions and factors. With an accuracy score of 0.9726, we can confidently assume that by inputting a dataset, the model will be able to tell us whether a wildfire is likely to occur in that region. These decisions will allow professionals to accurately allocate resources for preventative measures to the areas that have the highest likelihood for wildfire occurrence.

The random forest regressor model also produced a relatively efficient model with an R2-score of 0.8847 and a mean absolute error value of 0.0405. An R2-score closer to 1.0 indicates a better model fit, so our R2-score indicates that the model can partially predict outcomes, but we cannot expect all estimates to be perfect. A mean absolute error value closer to 0 indicates better model performance, so our value of 0.0405 tells us that our model performance

is very good. Therefore, we can conclude that the random forest regressor is another good model choice to predict wildfire occurrence.

Assumptions

One of the biggest assumptions in using a decision tree classification model is that the data can be and is effectively split based on its features, with the goal of maximizing the information gained at each decision node with each split. We are assuming that features can be used to make clear distinctions between classes, with a preference for easily-split categorical features over continuous, numerical features. Additionally,we are assuming that features are independent of each other, and that the decision made at one node does not heavily influence the decision at another node.

There are more assumptions in using random forest regression models. First, we are assuming that input data is continuous, the target variable is discrete, and the data is normally distributed. We are also assuming that the decision trees in the forest are independent of each other, and that they are grown deep to capture underlying patterns in the data.

Challenges & Limitations

Some challenges involved in creating prediction models include the typical challenges one must be aware of when working with decision tree classification and random forest regression models. Challenges for decision tree classifiers include overfitting the model to training data, handling continuous variables through binning, and addressing small changes and instability in the data that can lead to drastic changes in the model structure. In addition,

challenges for random forest regressors include difficulty of interpretation of the model's decision-making process and potentially overfitting poorly-tuned parameters.

The original project proposal included a plan to implement a k-means clustering model to determine location of wildfire hotspots, however the dataset in use did not provide the necessary data in order to build this specific type of model. For this reason, the k-means clustering model was replaced by the random forest regressor.

Future Uses/Additional Applications

These prediction models may be used to accurately predict the likelihood of wildfire occurrence for other regions around the world based on the features identified as most important in this project. Furthermore, it allows professionals to assess the most effective allocation of resources for preventative measures, providing more aid to regions with higher risk.

Recommendations

When using these models for wildfire prediction in the real-world, one recommendation would be to obtain very detailed weather and environmental data. While the training dataset was sufficient in building a working model, there are still many factors that play into wildfire occurrence and likelihood that were not included in this dataset. For example, one might also want to obtain data on vegetation, terrain topography, and fire history in an area to provide the model more variables to find patterns from. Additionally, different geographic regions will produce different "important features" that play more pivotal roles within the decision trees in the models. For that reason, it's important to assess each global region independently to ensure each site is being considered for its unique traits and features.

Implementation Plan

An effective implementation plan of these models for real-world wildfire prediction would include customizing variables and datasets for each region being focused on. First, it would involve collecting and processing data from sources such as weather stations, satellite imagery, topography maps, and vegetation indices to optimize a dataset with sufficient, cleaned data for the models. Next, the models would be developed and deployed for real-time predictions, risk assessment, and regular monitoring.

Ethical Assessment

In working toward wildfire prevention there are a few ethical considerations to take note of. First of all, when determining wildfire hotspots to focus preventative measures on, it is important to consider the health and safety of all communities surrounding the hotspot, equally. Each community that should be directly impacted by wildfires should have equal access to the research and preventative methods implemented in order to ensure their safety. Another ethical consideration to note includes focusing on ecological sensitivity and reaction to prevention methods. Conserving biodiversity and ecosystem health is important when determining which prevention methods would be most effective in a particular area, as some habitats and ecosystems may be more sensitive to different management strategies and it may throw off ecosystem balance.

References

Faroudja Abid. (2019). *Algerian forest fires* [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C5KW4N

The history of wildfire modeling. (2020, October 28). DEV Community. https://dev.to/jenciarochi/the-history-of-wildfire-modeling-5anl

Wildfires | cisa. (n.d.). Retrieved November 30, 2024, from https://www.cisa.gov/topics/critical-infrastructure-security-and-resilience/extreme-weather-and-climate-change/wildfires

Vizzuality. (n.d.). *Resource watch*. Retrieved December 10, 2024, from https://resourcewatch.org/data/explore/for012-Fire-Risk-Index