Analyzing the Prominent Environmental Factors that Cause Forest Fires in the Algerian and Montesinho Forests

By Sneha Narayan

Abstract

Forest fires are becoming increasingly common due to rising temperatures from global warming. According to the National Interagency Fire Center (NIFC), in 2022, 66,225 fires in the United States burned 7,534,403 acres of land. Data from the Global Forest Watch (GFW) indicates that in the last 10 years, around 82 million hectares of forests have been destroyed by wildfires worldwide. In this paper, I use a random forest classifier to predict the occurrence of a forest fire given a set of environmental conditions (Fine Fuel Moisture Code, Duff Moisture Code, Drought Code, Initial Spread Index, temperature, relative humidity, wind speed, rain, and the month in which the data was collected). A random forest classifier uses an ensemble of decision trees—which ask a series of binary questions to split the data—to determine whether a fire is likely, and displays which factors are most important in making this decision. My model achieves 100% accuracy in predicting whether a forest fire occurs given a set of environmental conditions from the Algerian Forest Fires dataset, and achieves 55.78% accuracy when given the Montesinho Forest Fires dataset. Knowing the prominent factors responsible for wildfire formation is useful for devising measures to offset the conditions and prevent harm. In the future, we can compare environmental data from more forests around the world to form a holistic view of the environmental factors that cause forest fires.

Introduction

In the last 10 years, around 82 million hectares of forests have been destroyed by wildfires worldwide [1]. Forest fires are common in hot, dry, and windy areas. They pose great harm to infrastructure in surrounding areas and can release harmful contaminants, sediments, and heavy metal that leach into nearby water sources [2]. Machine learning models can be used to analyze the occurrence of these forest fires and the factors that influence their formation. Specifically, this paper uses a random forest classifier to identify the factors which most affect wildfire ignition in the Algerian and Montesinho forests.

Literature Review: Forest Fires

Forest fires are large, uncontrolled fires that burn in vegetation more than six feet in height. Ground fires typically ignite in soil with thick organic matter and can last a full season. Surface fires burn what is on the ground, such as dead leaves, parched grass, and other types of dry vegetation. Crown fires are the biggest type of forest fire and burn tree canopies [3]. Forest fires are greatly influenced by temperature, humidity, precipitation, and wind [4]. High temperatures cause vegetation to dry out, which provides more fuel for fires [5]. Wind increases the supply of oxygen to the fires, which accelerates the ignition and spread of the forest fire. Precipitation and humidity can prevent fire ignition because it dampens the fuel—the air exchanges moisture with the dry vegetation, causing it to become more moist [6].

Forest fires have a damaging effect on wildlife and ecosystems. High severity fires can burn tree canopies and scorch the soil and tree roots [7]. However, the effect is even greater than destroying ecosystems. Smoke from forest fires include a mixture of toxic pollutants such as PM2.5 (fine particulate matter), nitrogen dioxide, ozone, and lead—which contaminate the air. In addition, forest fires release carbon dioxide and other greenhouse gasses into the air, increasing the impact of climate change. With global warming leading to higher temperatures, the risk of forest fires is increased, leading to a vicious cycle [8]. Therefore, it is important to accurately identify the factors that lead to forest fire ignition in order to be better prepared to combat the fires early on and devise mitigation strategies.

According to the National Interagency Fire Center, in 2022, 66,225 fires in the U.S. burned 7,534,403 acres of land [9]. Data from the Global Forest Watch indicates that in the last 10 years, around 82 million hectares of forests have been destroyed by wildfires worldwide [10]. Climate change, which results in global warming and extended drought, has increased the risk of forest fires in the United States over the last two decades [11]. Data from researchers at the University of Maryland shows that between the years 2001 and 2023, the area burned by forest fires increased by around 5.4% per year. Now, forest fires result in nearly 6 million more hectares of tree cover loss per year than they did in 2001—an area roughly the size of Croatia [12]. Data from the Global Fire Watch shows that fires were responsible for 74% of tree cover loss in Algeria between 2001 and 2023 [13]. In contrast, fires were responsible for 34% of tree cover loss in Portugal between 2001 and 2023 [14].

Literature Review: Machine Learning and Random Forest Classifier

Machine learning is the ability to train a machine to imitate human behavior. It is a subset of artificial intelligence. There are two major types of machine learning methods—supervised learning and unsupervised learning. Supervised learning uses labeled data to train the model. On the other hand, learning is unsupervised when the dataset does not have output labels. This means that the model forms clusters based on patterns identified in the data. Within supervised learning, there can be a classification model (where the predictive variable is categorical) or a regression model (where the predictive variable is continuous and numerical).

There are two stages when creating a supervised machine learning model—training and testing. In the training stage, the model creates a function for which the value from that function most closely relates to the true label of the data. In the testing stage, the function previously created is used to predict the labels of new data. Overfitting occurs when the model too closely adapts to the intricacies of the training data without generalizing results. The data is split into two categories, training and testing. Some common training/testing splits are 70% training and 30% testing, 80% training and 20% testing, and 90% training and 10% testing.

A decision tree uses a series of binary questions to split the data. The questions at each node of the decision tree are carefully selected during the training phase to split the data in the best way possible. These same questions are then used during the testing phase on a new set of data. A random forest is an ensemble of decision trees, all of which are used to make the final decision. In a decision tree, each point where a decision is made is called a “decision node”. Each terminal node is called a “leaf node”. When using a random forest algorithm, the importance of each feature in making the final decision can be shown. Some of the benefits of using a random forest algorithm include: accuracy for small datasets, ability to identify feature importance, ability to handle numerical and categorical data, and ease to interpret and visualize results. Some downsides are that random forest algorithms can create biased trees, overfit the data, and have greedy algorithms. However, these negatives can be mitigated by using an ensemble of trees to make the decision.

Methods

I created a Random Forest Classifier to predict the occurrence of a forest fire. I used methods from the pandas library in Python to clean the data, scikit-learn methods to split the data and create the model, and matplotlib for data visualization.

I ran my model on two different labeled datasets— “Algerian Forest Fires Dataset”[1] and “Montesinho Forest Fire Prediction Dataset”[2]. Each dataset recorded the environmental conditions during times of a fire and normal conditions in a different area—the Bejaia region and Sidi Bel-abbes region in Algeria, and the Montesinho forest in Portugal. The Montesinho forest fires dataset recorded the burn area, so I cleaned the data and sorted it into two classes—whether a forest fire occurred or not. I used the burn area to classify the data into these two classes and made the assumption that a “0” burn area indicated that there was no forest fire.

I further cleaned the datasets so that they included the same environmental factors, making it an apples-to-apples comparison. Each dataset has the following features: Fine Fuel Moisture Code (FFMC), Duff Moisture Code (DMC), Drought Code (DC), Initial Spread Index (ISI), temperature, relative humidity, wind speed, rain, and the month in which the data was collected. The Fine Fuel Moisture Code is a numeric rating of the moisture content of litter and other fine fuels such as small twigs, grasses, and ferns [15]. It indicates the relative ease of ignition and flammability of fine fuels. The higher the FFMC, the greater the risk of a fire. The Duff Moisture Code is a rating of the average moisture content of moderately deep loosely compacted layers of decomposing organic matter. The Drought Code Index measures the average moisture content of deep, compact organic layers. The Initial Spread Index is calculated by taking the product of the Fine Fuel Moisture Code and Wind Speed. It is the expected rate of fire spread [16].

After cleaning the data, I used scikit-learn methods to split it into training and testing sets. For both the Algerian Forests and Montesinho region datasets, I split the data into 80% training and 20% testing. I used the RandomizedSearchCV scikit-learn method to randomly choose different combinations of hyperparameters. I then used the best_estimator_ method to find the best combination of hyperparameters. This resulted in a high accuracy when the model was trained on the Algerian Forest Fires dataset. However, since the Monteinho Forest Fires dataset still yielded a low accuracy, I further tuned the n_estimators (number of decision trees that are used in the random forest) hyperparameter (shown in Figure 1). After doing so, I found the optimal hyperparameters for the Montesinho forest fires dataset. When tuning the hyperparameters, I used accuracy as the evaluation metric. Table 1 shows the distributions I used to tune the hyperparameters.

[1] Algerian Forest Fires Dataset: https://www.kaggle.com/datasets/sudhanshu432/algerian-forest-fires-cleaned-dataset

[2] Montesinho Forest Fires Dataset: https://www.kaggle.com/datasets/elikplim/forest-fires-data-set

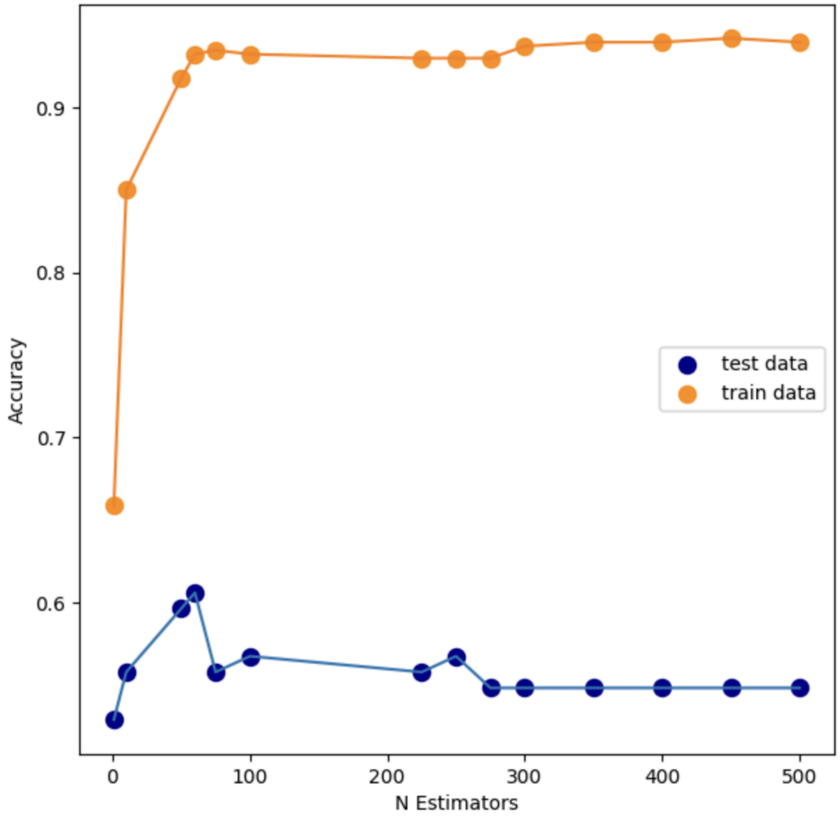


Figure 1: N Estimators vs Accuracy Plot for the model trained on the Montesinho Forest Fires Dataset.

As seen in Figure 1, when the number of estimators increases, the accuracy of the model on the training data also increases. The number of estimators refers to the number of decision trees in the random forest. There is a peak in the test data accuracy at around 50-60 estimators, but this is likely due to a fluke. The accuracy of the model levels off around 300 estimators which means including over 300 estimators will result in marginal benefits.

Hyperparameter Name	Description	Type and Range	Decided Value
n_estimators	The number of decision trees in the random forest.	Integer (50-3000)	350
max_depth	The maximum depth of the decision tree. If it is set to none, then the nodes will expand until all leaves are pure (where all data in that node belongs to a single class) or until all leaves contain less samples than min_samples split.	Integer (1-20)	8
max_features	The number of features to consider when looking for the best fit. If “sqrt” then the features considered will be the square root of the total features. If “log2”, the number of features considered will be $\log_2$(number of features). If “none” then all features will be considered.	sqrt, log2, none	None
oob_score	Whether to test the model on random samples of data to estimate how well the model generalizes to new data.	True, False	True
min_samples_split	The minimum number of samples required to split an internal node	Integer (2-10)	7

Table 1: Distributions used to tune the hyperparameters for the model trained using the Montesinho Forest Fires dataset. The hyperparameters I tuned are “n_estimators”, “max_depth”, “max_features”, “oob_score”, and “min_samples_split”.

Results

The model works well on the Algerian Forest Fires dataset, with 100% accuracy. The confusion matrix in Figure 2 shows that each piece of data is correctly classified.

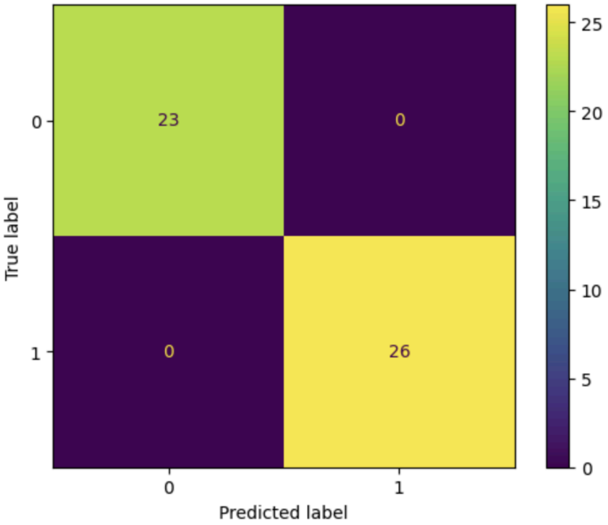
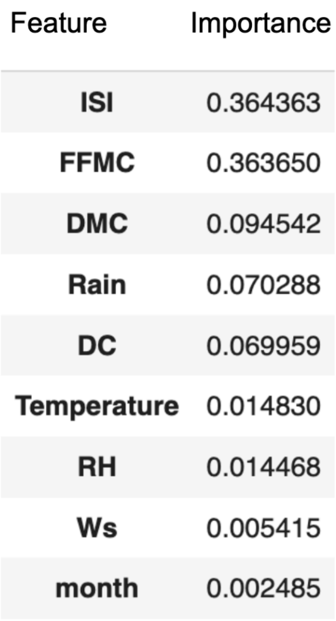


Figure 2: Confusion matrix of the model trained on the Algerian Forest Fires Dataset. The predicted label and true label matched for each piece of training data, meaning all the data was correctly classified. Here the label “0” represents that there was no forest fire and “1” indicates the presence of a forest fire.

The feature importance table in Table 2 indicates that the Initial Spread Index (ISI) and Fine Fuel Moisture Code (FFMC) have the greatest impact in making the decision for whether a fire will occur or not. On the other hand, features such as the month in which the data was recorded, wind speed, temperature, and relative humidity do not have a big impact on the final decision, as seen by their low feature importance values.



Feature	Importance
ISI	0.364363
FFMC	0.363650
DMC	0.094542
Rain	0.070288
DC	0.069959
Temperature	0.014830
RH	0.014468
Ws	0.005415
month	0.002485

Table 2: Feature importance table of model trained on the Algerian Forest Fires dataset. The Initial Spread Index (ISI) and Fine Fuel Moisture Code (FFMC) have the greatest importance. Features such as the month in which the data was recorded, wind speed, temperature, and relative humidity do not have a big impact on the final decision as shown by their feature importance values of less than .015 (or about 1.5%).

The model was trained with 500 estimators—which means that there were 500 decision trees in the random forest, from which the results were averaged together to determine whether a fire would emerge given the parameters. Figure 3 shows an example of one of the 500 decision trees. The tree is simple and does not have too many layers. Near the top of the tree, the model is making splits with decisive criteria, which split the data roughly in half. All the leaves in the decision tree are pure.

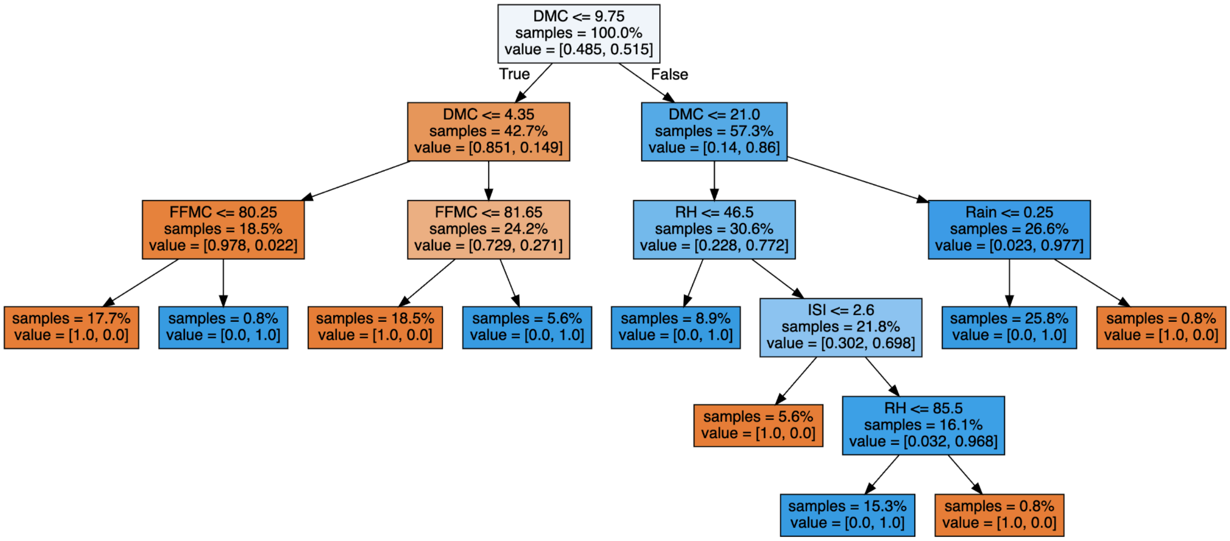


Figure 3: Example of one decision tree from the random forest model trained on the Algerian Forest Fires dataset. All the leaves in the decision tree are pure. The tree is simple and does not have too many layers. Near the top of the tree, the model is making splits with good criteria, which split the data roughly in half. All the leaves in the decision tree are pure.

In contrast, the model has a much lower accuracy for the Montesinho Forest Fires dataset. After hyperparameter tuning (shown in Figure 1), the model achieves an accuracy of 55.78%. The confusion matrix in Figure 4 shows that the model is not good at predicting when the true label for the data is “0”, meaning that the conditions would not cause a forest fire. As seen in the confusion matrix, when the true label is “0”, the model predicts correctly 20 out of 42 times, and predicts incorrectly 22 out of 42 times. This indicates that the model cannot find good patterns in the data when there are no forest fires, and therefore is unable to accurately make predictions.

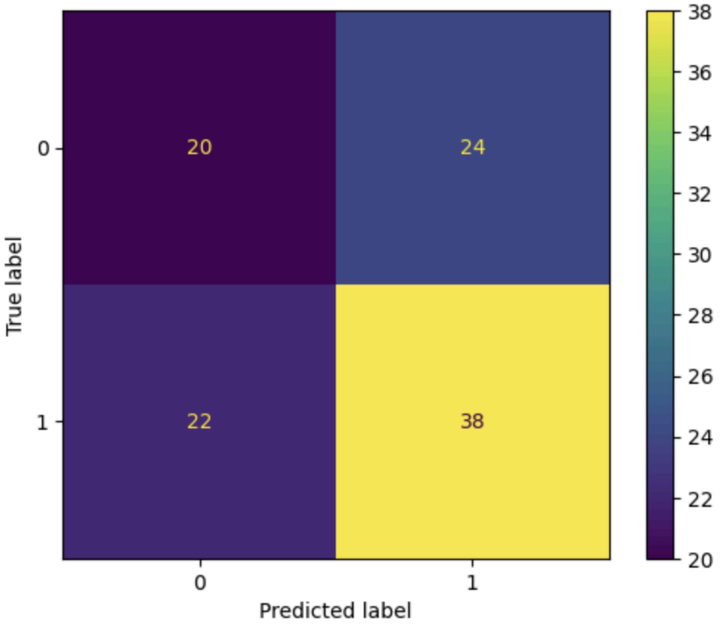


Figure 4: Confusion matrix of the model trained on the Montesinho Forest Fires Dataset. Here 0 represents no forest fire and 1 indicates the presence of a forest fire. When the true label is 0, the model cannot accurately predict whether there is a forest fire and essentially guesses—there is a 50% chance of the model predicting either label. This means that the model can not find good patterns in the data when there are no forest fires, and therefore is unable to accurately make predictions. When the true label is 1, the model is more accurate in its predictions but still not as accurate as in the Algerian Forest Fires dataset.

When looking at an example decision tree (Figure 5), it is evident that the model does not make good splits in the data early on. Instead of using conditions to split the data roughly in half, the model overfits to the data, isolating individual samples in the beginning rather than towards the middle/end.his means that the model is not able to accurately identify patterns in the data and instead is fitting very closely to the intricacies of the training data rather than producing generalizable results.

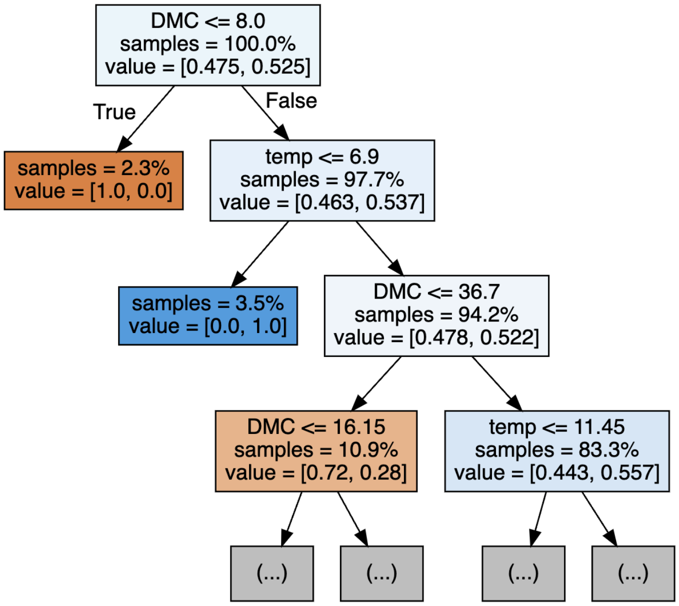
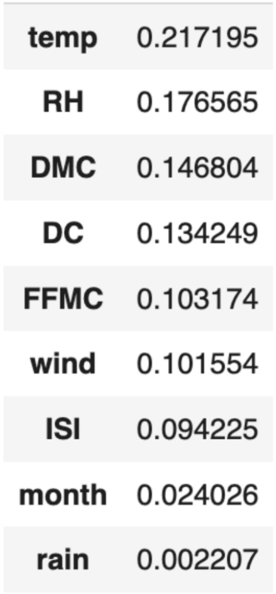


Figure 5: Example of the first three layers from one decision tree from the random forest model trained on the Montesinho Forest Fires dataset. The model overfits the data. Instead of using conditions to split the data roughly in half, the model isolates samples early on. For example, the initial split partitions the data into 97.7% and 2.3%, creating a pure leaf. If the model were learning the patterns in the data, the split would be closer to 50% and 50%.

The feature importance table in Table 3 shows that temperature, relative humidity, and the Duff Moisture Code are the most prominent factors in making the decision for whether a forest fire will occur or not. This means that a higher priority should be given to offset these factors. Unlike the feature importances for the Algerian Forest Fires dataset, where there is a clear indication of the features that are the most vital in making the final decision, the feature importances for each feature in the Portugal dataset are relatively similar to one another.



Feature	Importance
temp	0.217195
RH	0.176565
DMC	0.146804
DC	0.134249
FFMC	0.103174
wind	0.101554
ISI	0.094225
month	0.024026
rain	0.002207

Table 3: Feature importance table of model trained on the Montesinho Forest Fires dataset. Temperature, relative humidity, and the Duff Moisture Code are the most prominent factors in making the decision for whether a forest fire will occur or not. Unlike the feature importances for the Algerian Forest Fires dataset, where there is a clear indication of the features that are the most vital in making the final decision, the feature importances for each feature in the Portugal dataset are relatively similar.

Discussion

As seen in Figure 6, the decision trees for the model trained on the Algerian Forest Fires dataset were much simpler than those for the model trained on the Montesinho Forest Fires Dataset.

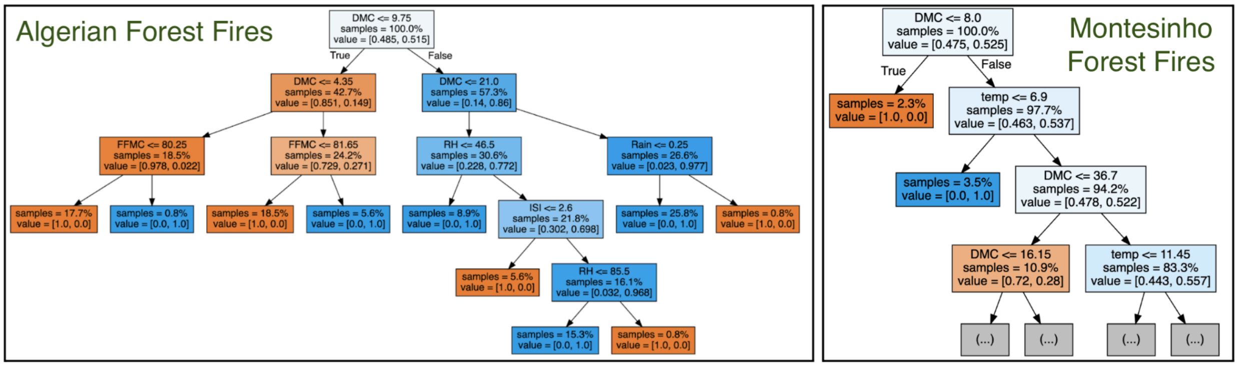


Figure 6: Side by side comparison of an example decision tree from the model trained on the Algerian Forest Fires dataset and the first three layers of an example decision tree from the model trained on the Montesinho Forest Fires dataset.

For the model trained on the Algerian Forest Fires dataset, the model made good splits early on in the data—splitting the data roughly in half in the beginning few splits. On the other hand, the decision trees for the Montesinho Forest Fires dataset isolated samples in the first few splits, which is an indication that the model was overfitting the data. Since the model worked well for the Algerian Forest Fires dataset, this can be an indication that there is no significant pattern in the features that correlates with forest fire presence in the Montesinho region (since the same features are used in both datasets). As stated in the literature review, forest fires are much more common in Algeria than Portugal—forest fires resulted in 74% [13] of tree cover loss in Algeria between 2001 and 2023 and only 34% [14] of tree cover loss in Portugal. This could potentially explain the low accuracy of the model trained on the Portugal dataset. Additionally, the data used for forest fires in the Montesinho Forest was originally a burn area dataset which I manipulated to serve as whether a forest fire occurred or not. I made an assumption that a 0 burn area indicated that a fire did not occur, but it is possible that a 0 burn area could have represented a very small fire.

Further, the feature importance tables for both datasets are different which could mean that the environmental factors that cause forest fires are specific to the region, meaning that results cannot necessarily be generalized to all areas. For the Algerian Forest Fires dataset, the important features were the initial spread index and Fine Fuel Moisture Code. This makes sense because high content of fine fuels leads to ease of fire ignition. For the Montesinho Forest Fires dataset, the temperature and relative humidity were the most dominant factors. The Montesinho Forest is located in northern Portugal, where areas experience higher humidity levels. Summers are humid and winters are not overly humid [17]. Due to the variation of humidity, it plays a large role in forest fire occurrence. On days when it is more humid, fires are less likely to occur due to the dampening effect the moisture in the air creates.

Conclusion

I built a random forest classifier to predict whether a forest fire is likely to form given a set of environmental conditions. I trained the model on two datasets—Algerian Forest Fires and fires in the Montesinho region. The model achieved a 100% accuracy when trained on the Algerian Forest Fires dataset, and achieved 55.78% accuracy when trained on the Montesinho Forest Fires dataset. By looking at the example decision trees for both datasets, it is likely that the model overfitted the data from fires in the Montesinho forest. The decision trees were far more complex for the Montesinho Forest Fires dataset compared to the Algerian Forest Fires dataset. The feature importance for both datasets are significantly different. The key features for the Algerian Forest Fires dataset are Initial Spread Index and Fine Fuel Moisture Code. For the Monteinsho Forest Fires dataset, the dominant features are temperature, relative humidity, and Duff Moisture Code. This is an indication that the results of important features cannot necessarily be generalized to fires in all regions.

Forest fires are extremely destructive and detrimental to the environment, so it is crucial to prevent them. Understanding the environmental factors that cause forest fires is the first step in order to offset such conditions to reduce the likelihood of fires.

Future work includes analyzing datasets from forests around the world to develop a more cohesive understanding of the factors which greatly affect forest fire formation. It is also vital to understand the key environmental factors that cause forest fire formation and take preventative measures to prevent the conditions from reaching a point of harm.

References

[1] “Wildfires Destroy 82m Hectares of Forest Worldwide in Decade.” *Anadolu Ajansı*, www.aa.com.tr/en/world/wildfires-destroy-82m-hectares-of-forest-worldwide-in-decade/2954918. Accessed 6 July 2024.
[2] “5 Negative Effects of Wildfires.” *WFCA*, 4 Mar. 2024, wfca.com/wildfire-articles/negative-effects-of-wildfires/.
[3] “Forest Fire.” Encyclopædia Britannica, Encyclopædia Britannica, inc., 30 July 2024, www.britannica.com/science/forest-fire.
[4] “Wildfires and Climate Change.” Center for Climate and Energy Solutions, 14 July 2023, www.c2es.org/content/wildfires-and-climate-change/.
[5] “Wildfires.” Education, education.nationalgeographic.org/resource/wildfires/. Accessed 30 July 2024.
[6] “How Does Humidity Affect a Fire?” WFCA, 4 Mar. 2024, wfca.com/wildfire-articles/how-does-humidity-affect-wildfire.
[7] Meghan Snow. “How Does Wildfire Impact Wildlife and Forests?: U.S. Fish & Wildlife Service.” FWS.Gov, 11 Oct. 2022, www.fws.gov/story/2022-10/how-does-wildfire-impact-wildlife-and-forests.
[8] “Wildfires.” World Health Organization, World Health Organization, www.who.int/health-topics/wildfires#tab=tab_1. Accessed 30 July 2024.
[9] “NCEI Monthly Fire Report.” National Centers for Environmental Information (NCEI), www.ncei.noaa.gov/access/monitoring/monthly-report/fire/. Accessed 11 July 2024.
[10] “Wildfires Destroy 82m Hectares of Forest Worldwide in Decade.” Anadolu Ajansı, www.aa.com.tr/en/world/wildfires-destroy-82m-hectares-of-forest-worldwide-in-decade/2954918. Accessed 17 July 2024.
[11] “Wildfire Climate Connection.” National Oceanic and Atmospheric Administration, www.noaa.gov/noaa-wildfire/wildfire-climate-connection. Accessed 17 July 2024.
[12] MacCarthy, James, et al. “The Latest Data Confirms: Forest Fires Are Getting Worse.” World Resources Institute, 13 Aug. 2024, www.wri.org/insights/global-trends-forest-fires.
[13] Vizzuality. “Algeria Deforestation Rates & Statistics: GFW.” Forest Monitoring, Land Use & Deforestation Trends, www.globalforestwatch.org/dashboards/country/DZA?category=fires. Accessed 5 Oct. 2024.
[14] Vizzuality. “Portugal Deforestation Rates & Statistics: GFW.” Forest Monitoring, Land Use & Deforestation Trends, www.globalforestwatch.org/dashboards/country/PRT?category=fires. Accessed 5 Oct. 2024.
[15] “LESSON 3 - EFFECTS OF WEATHER, TOPOGRAPHY AND FUELS ON FIRE BEHAVIOR.” Government of Nova Scotia, Canada, novascotia.ca/natr/forestprotection/wildfire/bffsc/lessons/lesson3/fuels.asp. Accessed 15 Sep. 2024.
[16] Canada, Natural Resources. “Canadian Wildland Fire Information System: Canadian Forest Fire Weather Index (FWI) System.” Canadian Wildland Fire Information System | Canadian Forest Fire Weather Index (FWI) System, cwfis.cfs.nrcan.gc.ca/background/summary/fwi. Accessed 15 Aug. 2024.
[17] “Climate: North Portugal.” Worlddata.Info, www.worlddata.info/europe/portugal/climate-north.php. Accessed 5 Oct. 2024.