# Machine Learning Project Report: Predicting Forest Fires in Portugal’s Montesinho Natural Park

## Executive Summary

Throughout the world, forest fires cause major damage to both economies and ecosystems. They often spread quickly and can grow out of control in the absence of immediate human intervention. This study addresses the issue of predicting areas at the highest risk of burning due to forest fires by exploring the influence of environmental and weather-related factors. Through a dataset sourced from a 2007 Portuguese study, this study utilizes several machine learning algorithms, such as Logistic Regression, Classification and Regression Trees (CART), Random Forests, and machine learning preprocessing techniques such as Box-Cox and Yeo Johnson transformations, as well as Min-Max Feature Scaling. Through a high-performing Random Forests model, this study highlights several key factors crucial for predicting the presence of forest fires in a given area. This approach shines light on the potential for machine learning as a tool in the early detection and control of forest fires.

## Introduction

Background Information:Forest fires are extremely detrimental to ecosystems and the economy, threatening natural resource loss and ecosystem degradation through biodiversity loss. Through forest fire and its aftermath is an important part of many ecosystems, and is a driving force of the system of ecological disturbance and succession, it quickly becomes extremely detrimental to an ecosystem if the fires are too frequent and too intense, past the point of tolerance where the ecosystem will no longer quickly recover on its own. Over the past few decades, as climate change has progressed, the world has experienced a continuing trend of more frequent, more intense, and more damaging wildfires.

Problem Statement: Early detection and action is key in curbing the intensity and affected area of forest fires. As climate change continues to progress, it will only exacerbate this problem, as changing weather patterns point to growing frequency and intensity of such fires. This study seeks to combat this growing issue by aiding in the early prevention and control of forest fires, allowing the allocation of often scarce resources to where they are most needed, in the extremely rural and remote setting of many National Forests and Parks. This will be carried out by predicting the presence of fire based on the values of carefully selected meteorological and Fire Weather Index (FWI) metrics.

Scope of the Report: This report details the methodology and reasoning of each action throughout the study. From an overview of relevant variables, to exploratory data analysis (EDA), feature and model selection, hyperparameter tuning, and model analysis, this report covers each individual step of the project process, and an analysis of its real-world uses and implications.

## Dataset Description

Data Source: The data for this study was sourced from the Montesinho Natural Park, in the Tr´as-os-Montes northeast region of Portugal. The data used in the experiments was collected from January 2000 to December 2003, and it was built using two sources. The first database was collected by the inspector that was responsible for the Montesinho fire occurrences. On a daily basis, every time a forest fire occurred, several features were registered, such as the time, date, spatial location, the type of vegetation involved, the six components of the FWI system, and the total burned area. The second database was collected by the Bragança Polytechnic Institute, containing several weather observations (e.g. wind speed) that were recorded within a 30 minute period by a meteorological station located in the center of the Montesinho park. The two databases were stored in tens of individual spreadsheets, under distinct formats, and a substantial manual effort was performed to integrate them into a single dataset with a total of 517 entries.

Data Characteristics: The dataset includes 517 observations, each of which identify a specific spatial location in the Montesinho Natural Park at a certain time and date. Each entry features an area target variable, denoting the burned area of the forest sector in hectares. Besides the target variable, there are 12 additional features that indicate location, temperature, wind speed, rain amount, and other environmental metrics established by the Canadian forest Fire Weather Index.

Suitability: Since our study is concerned with forest fire prevention, analyzing this data on forest fires in the Montesinho Natural Park can give us insights on how to detect forest fires in other locations. Although Montesinho Natural Park is not representative of every forested area in the world, the climate features and environmental metrics identified for each data observation are applicable to other global areas. What we learn from this specific forest is of the utmost importance to preventing forest fires as a whole.

## Methodology

### Data Preprocessing:

Data Cleaning: The data cleaning process was mostly straightforward, as the dataset did not have any missing values. The main strategy employed was row deletion, which was used to remove an unexplained extreme outlier in the ISI column. Additionally, the X, Y, day, and rain columns were removed as part of this process. The X and Y columns were removed on account of being positional columns specific to this dataset, which lack usefulness in generalizing the model’s findings to more broad real-world settings. The day column was removed due to an unnecessary temporal column: data was in some areas collected weekly, so the distribution of days were heavily skewed. The rain column was removed on account of it having only eight total non-zero values, in order to combat overfitting by reducing dimensionality.

Feature Engineering: Two key features were created during the feature engineering process. The month column was replaced with a binary "fire season" column, displaying whether the observation was taken during fire season in Portugal, which is from July through September. This avoids creating relationships that do not actually exist in the data via label encoding, or increased dimensionality due to one hot encoding, which would have resulted in the creation of 12 additional columns. Additionally, the area column, which was the target variable, was mostly comprised of zeroes. Therefore, we engineered a new categorical target variable, displaying whether or not there is any burned area in the forest sector.

Data Transformation: None of the continuous variables in the dataset were adequately normally distributed (though some more so than others). To ensure that every columns was near normally distributed, we utilized Box-Cox and Yeo Johnson transformations. The data was also scaled using Min Max Scaling to ensure equal variable consideration in the model.

Train-Test Split: The dataset was split into training and testing sets with a 80:20 ratio, using stratified sampling to maintain the same proportion of burned to non-burned samples in each set. This approach ensures that the target variable’s distribution in the training and testing sets is representative of the overall dataset.

### Model Selection and Development:

Model Selection: Logistic Regression was one of the first models we learned for classification, and because it is relatively computationally simple, we decided that it would be a good first step to developing a robust set of models. With Logistic Regression, we not only develop a binary classification model that definitively classifies observations into two classes, but the model also makes classification decisions based on a percentage on how confident it is about a certain classification. This way, we can set our own decision threshold that determines the percentage at which we believe it is best to distinguish our two classes.

The second model used in this study is Random Forest. Random Forest consists of multiple CARTs (Classification and Regression Trees) and is a common, robust model for classification. The benefits of Random Forest is that we can apply feature importance (to limit dimensionality and prevent overfitting), hyperparameter tuning (to optimize the various hyperparameters influencing our model), and pruning (also for preventing overfitting). Random Forest is also an ensemble method, meaning that we are taking the average results of our several Trees and providing comprehensive predictions and insights.

Model Training and Tuning: For the basic logistic regression model, we are setting ‘solver’ to liblinear because liblinear is best at dealing with small datasets, such as the one we are analyzing. To determine the best decision threshold, we evaluated the performance of the model on various thresholds with accuracy, precision, recall, F1 score, ROC-AUC, and log loss. All of these metrics are common for classification problems. In terms of hyperparameter tuning, we are looking at penalty and C. Penalty determines whether the model will be regularized using L1 (Lasso) or L2 (Ridge) methods. C is the inverse of regularization strength. Both are useful in preventing overfitting.

As mentioned, we are applying feature importance, hyperparameter tuning, and pruning because these are all useful methods associated with Random Forest for optimizing the model and preventing overfitting. We first created a random forest model without setting any hyperparameters (with the exception of n\_estimators for computational efficiency). We did this in order to take advantage of the built in variable importance methods, so we can eliminate unneeded features. The hyperparameters we are optimizing for are criterion, max depth, min samples split, and max features. Criterion determines which metric we will use to measure the purity of a node, max depth determines how many levels of nodes each tree can have, min samples split determines the minimum number of samples required to split an internal node, and max features determines the number of features to consider when looking for the best split. We also apply pruning methods on max depth and min samples leaf to prevent overfitting.

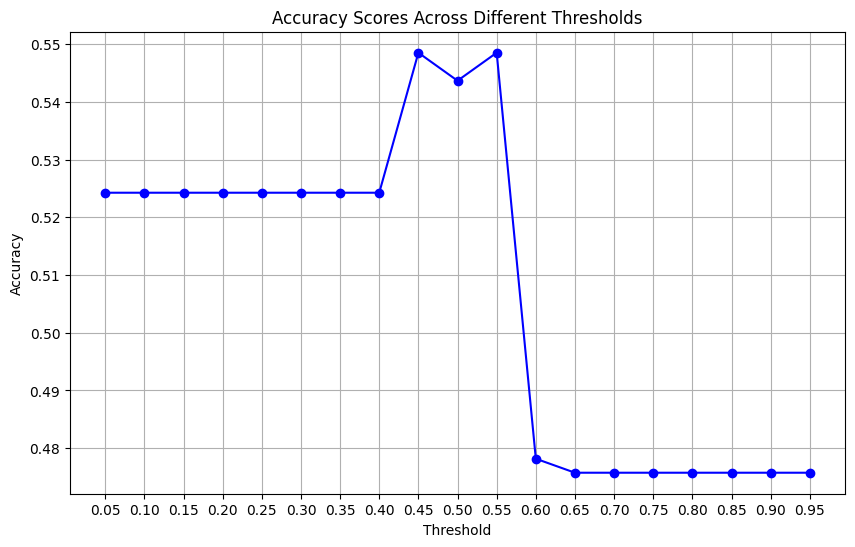
## Results

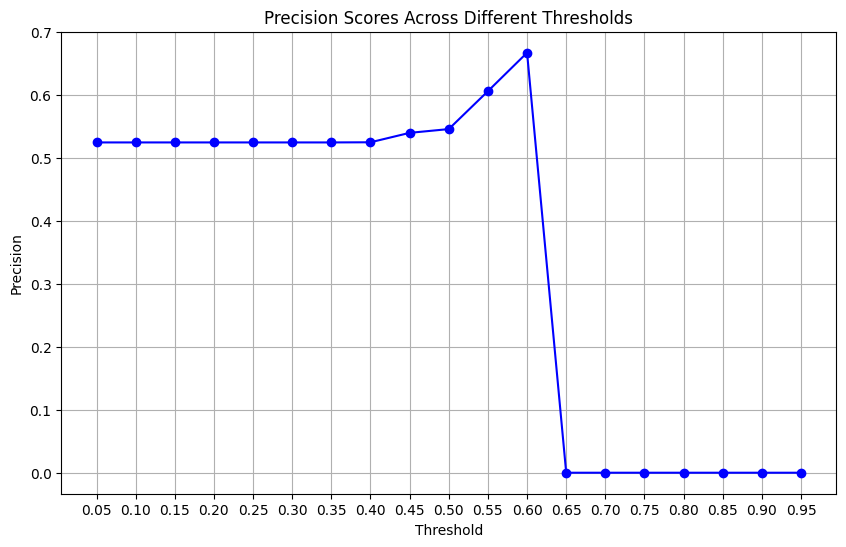
Model Evaluation: After close analysis of the logistic regression’s evaluation metrics, 0.45 resulted in the best threshold. Even though this does not seem very far from 0.5 (the default threshold), a difference of 0.05 can make a significant difference in the model. We are trying to optimize everything as much as possible. After hyperparameter tuning, C was 0.001, and penalty was L2 (ridge). In terms of the overall model, accuracy, precision, and ROC-AUC are roughly 0.5, and log loss is around 17 (for both train and test). The best performing metric is recall (1 for both train and test). Although we do care about recall because wrongly predicting an area as not burnt can be adverse to the goals of this study, the unfavorable results of the other metrics demonstrate that the model is not performing well.

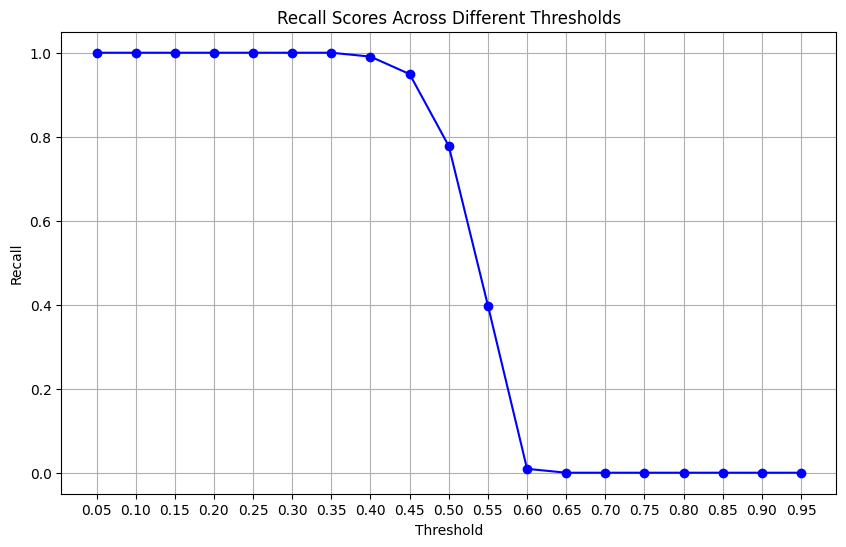
Random forest, on the other hand, yielded impressive results. Feature importance allowed us to remove the fire season column. After hyperparameter tuning, criterion was gini, max depth was 2, max features was None (meaning there would be no limit to the max depth), and min samples split was 3. Since hyperparameter tuning resulted in a max depth of 2 (which seems odd), we wanted to apply methods of pruning to see whether there are other max depth values that would benefit the model without causing overfitting. We also performed pruning for the hyperparameter min\_samples\_leaf. Unfortunately, pruning had no effect on the model (with respect to the parameters max\_depth and min\_samples\_leaf). Therefore, we set both to their default values: there is no max\_depth (the model is most likely not overfitting and can deal with a large max\_depth), and min\_samples\_leaf is at its minimum value, 1. For the overall model, accuracy, precision, recall, F1 score, and ROC-AUC are all very near 1, and log loss is very near 0, meaning that our random forest model is performing very, very well. Since the results are similar for both the train and test, it means that the model is not overfitting.

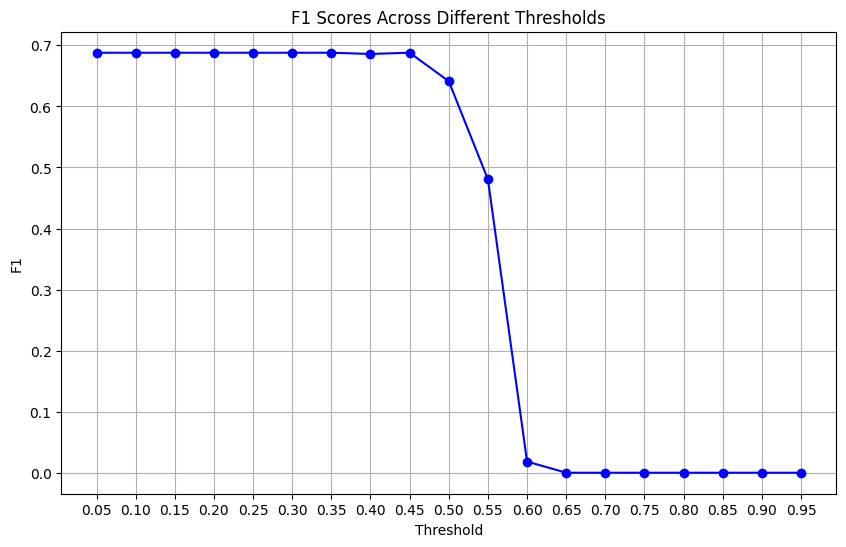
Visualizations:

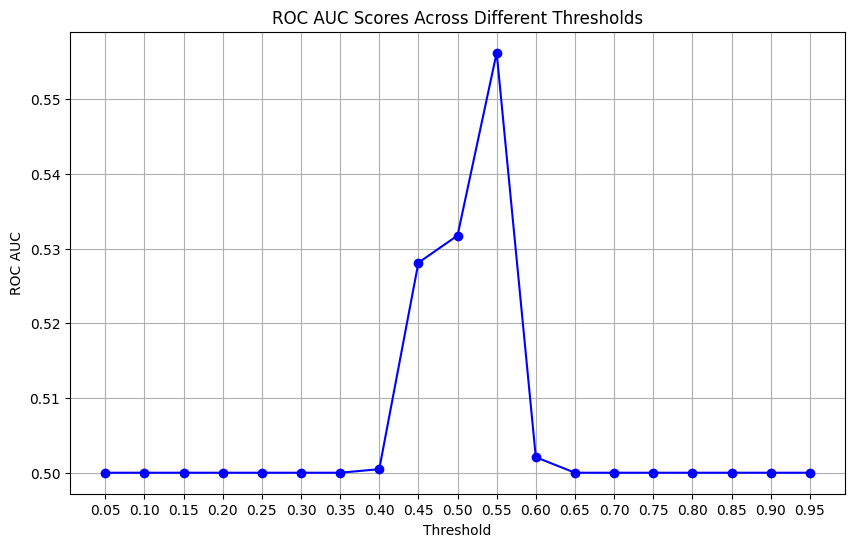
Plots for each evaluation metric, for determining the best threshold for our logistic regression model

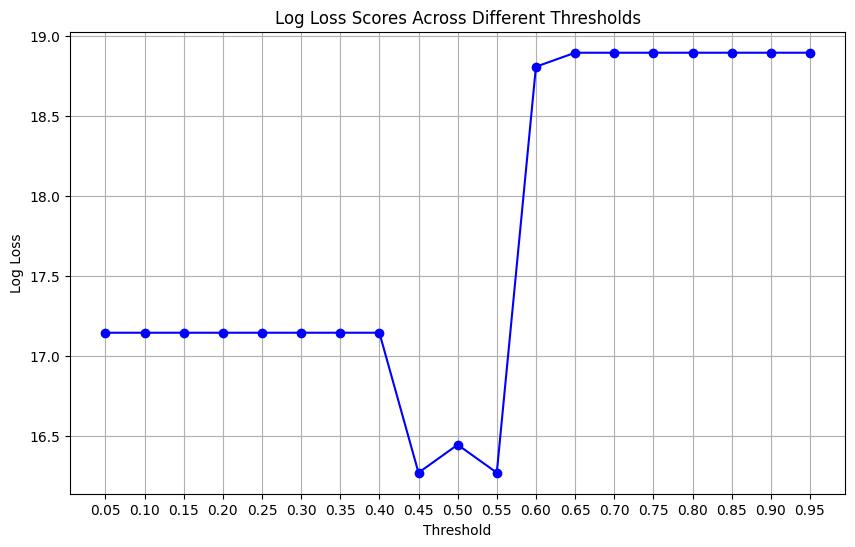




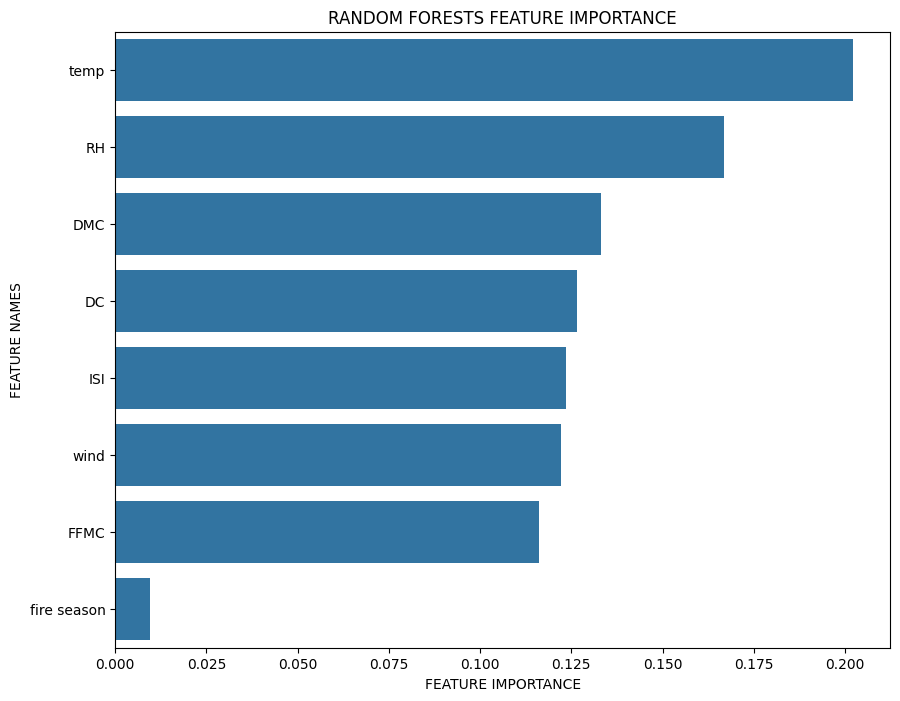








Feature importance plot for random forest



## Discussion

Insights and Analysis: This model and our findings can be generalized to other large forests of similar composition and climate, though some analysis should be done to conclusively study how well the model performs in areas with drastically different climate and/or forest composition. From the work we have done, we cannot confidently assess this particular question without further research and testing on other datasets.

We concluded that this model would be of particular use in remote areas with limited fire prevention and response resources. In these scenarios, determining which areas of a large forest are likely to be burned based on weather and FWI data can be used a useful tool for smart resource allocation, allowing local and government agencies to prioritize areas of forest that are more likely to need human intervention in order to prevent, slow, and eradicate forest fires.

Limitations: As discussed above, this model cannot necessarily be applied to forests or other ecosystems that are very different in climate or composition compared to the forest analyzed in this exploration. This is due to how fire spread metrics and weather conditions may differ in their interactions in an ecosystem with drastically different flora and fauna. In particular, the presence or absence of duff and brush layers would likely be heavy contributors to these differences, due to their roles in early stage wildfire spread.

The other main limitation of our model is in how our data was collected. As mentioned above, data was collected whenever a forest fire was present somewhere in the park. This means that we lack data from times when no fire was present anywhere in the park, so we cannot generalize our findings to data where no fire is present in the immediate geographic area.

Future Work: In the future, we would like to examine the impact of similar metrics in other datasets featuring different climates and forest compositions. This would allow us to better generalize our findings to a wide variety of forests worldwide, and understand how interactions between our environmental variables change or remain constant in different areas. We would also like to explore the temporal aspect of this data more in depth, though this would require a different method of data collection, such as using environmental (weather and FWI) data taken at regular intervals throughout a long period of time. We would like to explore the impact of different times of year (which have different weather patterns) on forest fires, but also analyze how forest fire frequency and severity has changed over the past few decades. This would allow us to look at climate patterns as a whole, confronting the overall impact of climate change on forest fire frequency and intensity.

## Conclusion

Summary of Findings: We determined that weather factors such as temperature, relative humidity, and windspeed, combined with FWI metrics, are very effective in predicting whether an area will be burned in a forest fire. Temporal metrics proved to be mostly irrelevant, likely in part due to the data collection method of our dataset: each time a forest fire was recorded somewhere in the park, data was collected in every single forest quadrant. In other words, the frequency of fires during different seasons was not relevant because all data was taken from a point in time during which there was an active fire in some part of the park.

## Appendices

Code: Reference the GitHub repository containing all code, a detailed README, and an environment file for project reproducibility.

## References

Bibliography:

1. Dataset on Forest Fires, UC Irvine Machine Learning Repository

<https://archive.ics.uci.edu/dataset/162/forest+fires>

1. Cortez, Paulo & Morais, A.. (2007). A Data Mining Approach to Predict Forest Fires using Meteorological Data.

<http://www3.dsi.uminho.pt/pcortez/fires.pdf>

1. Information on the Canadian Forest Fire Weather Index (FWI) System

<https://cwfis.cfs.nrcan.gc.ca/background/summary/fwi>