A COMPARATIVE STUDY OF FOREST FIRE PREDICTION USING MACHINE LEARNING MODELS

by

Mr. Sachin Malego (st125171) & Mr. Sila N. Mahoot (st125127)

A Project Proposal Submitted in Partial Fulfillment of the Requirements for the Degree of Master of Science in Data Science & AI

Submitted to: Dr. Chantri Polprasert
Assistant Professor
CSIM, Department of ICT

Nationality: Nepalese and Thai

Asian Institute of Technology
CSIM, Department of Data Science & AI
Thailand
October 2024

ABSTRACT

Forest fires are one of the most pressing incidents that have a significant impact on both the environment and human life, making early prediction is crucial for minimizing the damage. This proposal presents a comparative study of various machine learning models for forest fire prediction leveraging the historical data from various dataset. Techniques such as XGBoost, Decision Tree, Random Forest, and Logistic Regression are explored, with models evaluated based on accuracy and other performance metrics. The study highlights the strength and weakness of each approach with geographic maps of the country (Portugal and Algeria), demonstrating that machine learning can significantly enhance predictive capabilities, providing a valuable tool for fire prevention and disaster management. This project does not determine if a forest fire will take place or not in any place, however we are predicting the confidence probability of the forest fire based on some attributes.

CONTENTS

LIST OF FIGURES

Figures **Page**

CHAPTER 1
INTRODUCTION

With the high incident rates of forest fires, its prediction holds a significant environmental and scientific importance.[1] In recent years, forest fires have emerged as a critical concern, posing a recurring threat to vast forested areas worldwide. Observing the data from Nepal alone, it has experienced 965 forest fire incidents in the last one year (08-10-2023 to 07-10-2024) with an estimated economic loss of around USD. 644,696. Such is not only the scenario of Nepal but at a global level. Here in this project, we review such dataset of forest fire from Forest fire dataset of Algeria, and Portugal and run different models to make the best possible prediction.

1.1 Background

Forest fire forecasting is critical for mitigating the environmental, economic, and public health impacts of wildfires. According to existing literature, forest fire prediction methods can be broadly categorized into three main approaches: physics-based models, statistical models, and machine learning models.[2] Physics-based models rely on simulating fire behavior based on physical laws, such as fluid dynamics and thermodynamics. Statistical models focus on identifying correlations between past fire events and contributing factors, often using regression techniques. However, machine learning (ML) models have recently gained prominence due to their ability to learn complex, non-linear relationships from large datasets, making them highly suitable for dynamic and uncertain phenomena like forest fires.

This project leverages data-driven algorithms in the machine learning domain to predict the likelihood of forest fire occurrences. Our goal is to train ML models on historical weather and environmental data and use these models to predict fire risks based on current conditions. The data, sourced from regions in Portugal and Algeria, contains key variables such as temperature, relative humidity, wind speed, and climate events like El Niño. These variables are integral to capturing the complex interplay of factors that contribute to fire ignition and spread.

Before delving into the technicalities of forest fire prediction, it's crucial to explore the climatic conditions that exacerbate wildfire risks. El Niño is a particularly significant

factor, characterized by the abnormal warming of sea surface temperatures in the central and eastern Pacific Ocean. Although El Niño is a Pacific-based event, it has far-reaching implications for global weather patterns, leading to drier or wetter conditions in various regions, including Europe and North Africa. In both Portugal and Algeria, El Niño can disrupt typical weather patterns, resulting in reduced rainfall, prolonged dry spells, and increased temperatures - conditions that are conducive to the ignition and spread of forest fires.

In addition to El Niño, several other environmental and climatic variables influence fire risk, including rainfall patterns, wind speed, relative humidity, fine fuel moisture, duff moisture, and drought conditions. These factors, when integrated into machine learning models, can provide a powerful predictive tool for early fire detection and risk assessment.

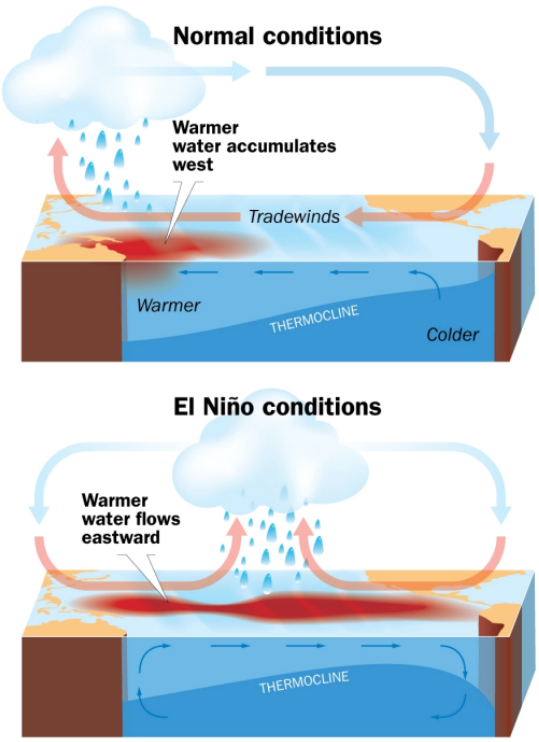

Figure 1: Understanding El Nino

1.2 Objective

The objective part in these answers the why part of need of forest fire prediction. The objective of forest fire prediction is to improve early detection and risk assessment, a crucial need globally. Despite the experience of forest departments, human limitations in processing multiple variables hinder accurate predictions. Hence in such situations Machine learning can analyze numerous factors simultaneously, making fire prediction more efficient and effective. The major objectives of this project are:

i. **Evaluate Model Performance**: Compare the accuracy, precision, recall, F1 score and overall predictive performance of various machine learning models in predicting forest fires.

ii. **Identify Key Features**: Analyze the importance of various environmental and weather-related factors (e.g., temperature, humidity, wind speed, El Niño) across different models to identify which features contribute most to fire prediction accuracy.

iii. **Assess Generalization Capability**: Investigate how well each machine learning model generalizes to unseen data, particularly when predicting forest fires in different geographic regions or under varying climate conditions.

iv. **Model Interpretability and Usability**: Evaluate the interpretability of each model, determining how easily stakeholders (e.g., forest management agencies) can understand and utilize the model's predictions for decision-making.

v. **Predictive Alerts for Health Precautions**: Evaluate how machine learning models can provide early fire risk alerts, enabling timely health interventions, such as air quality warnings or evacuation notices, to protect vulnerable populations (e.g., children, elderly, people with pre-existing conditions).

vi. **Support Sustainable Forest Management**: Explore how machine learning predictions can aid in sustainable forest management by allowing for preemptive actions like controlled burns or forest thinning to mitigate the severity of future fires, reducing long-term environmental damage.

1.3 Business Understanding

Climate change is intensifying droughts and unpredictable weather, making forest fires more frequent and destructive. This project focuses on predicting fires in Portugal and Algeria using weather and environmental data. By comparing multiple machine learning models, it aims to evaluate the impact of variables like temperature, humidity, wind, and El Niño on fire risk prediction.

1.4 Impact

The proposed study on forest fire prediction using machine learning models has the potential to generate significant and far-reaching impacts across multiple sectors. These impacts can be categorized as environmental, social, economic, technological, policy-related, and scientific, all contributing to a more informed and effective approach to forest fire management.

The comparative study of forest fire prediction using machine learning models has the potential to significantly impact environmental conservation, public safety, economic stability, technological progress, policy development, and scientific advancement. By addressing a critical global challenge, this research will contribute to a safer and more resilient world in the face of increasing forest fire risks due to climate change.

CHAPTER 2
PROBLEM STATEMENT

The increasing frequency and intensity of forest fires due to climate change pose significant threats to ecosystems, human health, and economies worldwide. To address this pressing issue, there is a need to develop accurate predictive models that can forecast forest fire occurrences based on historical data and environmental conditions. This study aims to investigate the relationships between various factors contributing to forest fires, including:

- **Temporal Trends**: Analyzing seasonal, and monthly trends in forest fire occurrences to identify patterns and peak risk periods.
- **Correlations Between Environmental Parameters**: Examining the correlations between key weather parameters (e.g., temperature, humidity, wind speed) and the incidence of forest fires to understand how these factors influence fire risk.
- **Impact of Climate Phenomena**: Assessing the effects of significant climate phenomena, such as El Niño, on forest fire occurrences, providing insights into how broader climatic trends affect fire behavior.
- **Comparative Analysis of Machine Learning Models**: Evaluating the performance of different machine learning algorithms in predicting forest fire risk, determining which model offers the best accuracy and reliability for real-time decision-making.
- **Visualization**: Visualize impact areas in maps for prediction and effective decision making.

By comprehensively analyzing these elements, the project seeks to enhance forest fire prediction capabilities, enabling more effective prevention and response strategies to mitigate the impacts of forest fires. The goal is to create a reliable machine learning model that can predict the chances of forest fires based on weather and environmental data. By using data from Portugal and Algeria, this project aims to improve traditional prediction methods and provide a tool to help authorities make better decisions for managing fires. The challenge is to make a model that can handle different weather patterns, ultimately reducing the damage caused by forest fires.

CHAPTER 3
RELATED WORKS

The project goal is to train the model to be able to predict the forest fire. Here the focus is on the use of different factors along with climate change phenomenon of El Niño to generalize and predict the fire occurrence scenario for better decision making.

Jing et al., in the research paper "***Toward a more resilient Thailand: Developing a machine learning-powered forest fire warning system***" describes a machine learning-based forest fire warning system in Thailand. It uses satellite data and gas measurements to predict forest fires, with models like linear classifiers, gradient boosting classifiers, and neural networks.[3] The XGBoost model had the best performance with an accuracy of 99.6%. Our project aims to improve this approach by adding climate factors like El Niño and testing more models across different regions.

Abdelhamid Zaidi in the paper "***Predicting wildfires in Algerian forests using machine learning models***" focuses on developing a predictive system for forest fires in Algeria, a region that has seen an increase in fire occurrences over recent years.[4] The study aims to create an accessible, low-budget system for predicting wildfires based on climatic data and machine learning algorithms. The research highlights challenges in forest fire prediction, such as fire behavior complexity and the need for accurate datasets. Machine learning algorithms are used to analyze weather data and predict potential fires. The study concludes that a well-designed machine learning system could benefit local authorities by enabling early detection and timely interventions.

In another paper from IEEE Viswa et al., "***Comparison of Forest Fire Prediction System using Machine Learning Algorithms***" presents a comparative analysis of various machine learning algorithms for predicting forest fires. It emphasizes the importance of early detection in minimizing the devastating effects of wildfires on ecosystems and human settlements.[5] The study evaluates the performance of different algorithms based on accuracy, precision, recall, and other relevant metrics. The conclusion iterates the importance of machine learning in forest fire prediction and suggests future research directions, including the integration of additional data sources and advanced algorithms.

CHAPTER 4
DATASETS

4.1 Description

In this project, we are using two main datasets from Kaggle and one weather data from another public source via web scraping. These datasets represent key information about forest fire occurrences and influencing factors. The datasets are structured in the Excel spreadsheet format.

- Forest Fires Dataset – Algeria
- Forest Fires Dataset – Portugal
- Weather (Meteorological) Dataset

Along with this dataset if feasible we are also trying to study El Niño if it has any significane in forest fire occurrence although the period for this is very low.

Figure 2: Area of Interest

4.2 Features

The dataset contains following features:

1) *Algerian Forest Fires Dataset*: This dataset will be used for training the model. The dataset includes 244 instances that regroup a data of two regions of Algeria, namely the Béjaïa region located in the northeast of Algeria and the Sidi Bel-abbes region located in the northwest of Algeria. 122 instances for each region. The period from June 2012 to September 2012. The dataset includes 11 attributes and 1 output attribute (class). The 244 instances have been classified into fire (138 classes) and not fire (106 classes) classes.
    a) Date: The date of the observation.
    b) Temperature: Daily average temperature in degrees Celsius.
    c) Humidity: Daily average relative humidity as a percentage.
    d) Wind Speed: Daily average wind speed in km/h.
    e) Rainfall: Daily rainfall in mm.
    f) Region: The specific region in Algeria where the data was collected.
    g) Fire Occurrence: Binary indicator (Yes/No) for whether a fire occurred.

2) *Forest Fires Dataset - Portugal*: This dataset will be used for **testing** the model. The dataset includes 517 instances. The dataset includes 13 attributes and 1 output attribute (class).

   a) Date: The date of the observation.
   b) Temperature: Daily average temperature in degrees Celsius.
   c) Humidity: Daily average relative humidity as a percentage.
   d) Wind Speed: Daily average wind speed in km/h.
   e) Rainfall: Daily rainfall in mm.
   f) FFMC (Fine Fuel Moisture Code): Index representing the moisture content of surface litter.
   g) DMC (Duff Moisture Code): Index indicating the moisture content of loosely compacted organic layers.
   h) DC (Drought Code): Index indicating long-term drying conditions.
   i) ISI (Initial Spread Index): Index for predicting the rate of fire spread.
   j) Fire Occurrence: Binary indicator (Yes/No) for whether a fire occurred.

3) *Weather and Environmental Data*: These are the additional factors which may have some role in the fire. This dataset will be mapped with the other two on basis of timestamp and location.

   a) Temperature: Daily average temperature in degrees Celsius.
   b) Humidity: Daily average relative humidity as a percentage.
   c) Wind Speed: Daily average wind speed in km/h.
   d) Precipitation: Daily rainfall in mm.

4) El Niño Data (Sea Surface Temperature): This dataset is used only for observing if it has any significance or not. This is not a required dataset however we wanted to have a look into this as well to see its significance if any.

   a) Niño 3.4 Index: Represents Sea surface temperature anomalies in the central equatorial Pacific, which is a critical indicator of El Niño events.

The figure below depicts the El Nino occurrence and average temperature of the 3 months of data from the Algerian dataset, it tries to map the El Nino occurrence with temperature observed in that period.

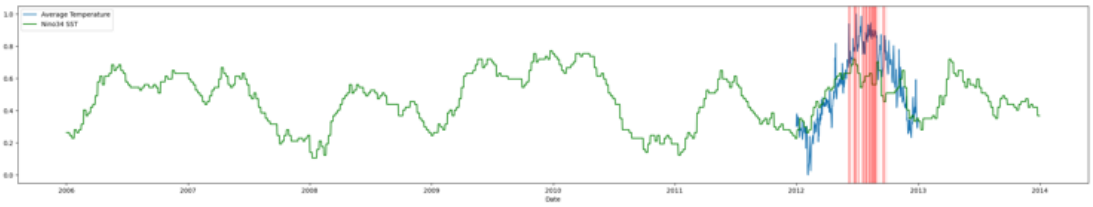

Figure 3: El Nino occurrence and Average Temperature

CHAPTER 5
METHODOLOGY

5.1 Data Acquisition

The data for this project was collected from publicly available sources on Kaggle and other weather data platforms. The main datasets are the Algerian Forest Fires Dataset and the Forest Fires Dataset - Portugal, which have valuable information on past forest fires and weather conditions. The data is provided in CSV format, making it easy to load and work with using Python libraries like Pandas.

We also got weather data for Algeria and Portugal from Meteostat and Weather Underground. Meteostat provides historical weather metrics, while Weather Underground allows us to gather more recent daily weather data using web scraping. This weather data is very important for understanding the factors that contribute to forest fires.

To understand the role of climate events like El Niño, we also used El Niño Sea surface temperature data from the Climate Prediction Center (CPC). This helps us see how larger climate patterns might influence forest fire occurrences. All the data was downloaded or accessed through APIs, then integrated for further exploration and analysis.

5.2 Exploratory Data Analysis

Exploratory Data Analysis (EDA) helps us understand the patterns and relationships in our data. For this project, we are doing EDA on the Algerian and Portugal datasets, as well as weather and El Niño data, to find trends that might predict forest fires.

The following steps and best practices will be applied during the EDA process:

1) Data Cleaning

   We will check for missing values, mistakes, and inconsistencies. Missing data will either be filled in or removed based on its importance.

2) Data Transformation

   We will turn date features into more useful information, like year or month, to help find seasonal trends.

3) Data Visualization

   We will use different types of plots to understand the data better:

   a) Line Plots to show trends over time, like changes in temperature and humidity.
   b) Histograms show how individual features are distributed.
   c) Scatter Plots to look at relationships between key features, like temperature and fire occurrences.
   d) Correlation Heat Maps to find which features are most important for predicting fires.
   e) Box Plots to find and deal with outliers in the data.

EDA will give us important insights that will help us prepare the data, choose features, and build better models.

The histogram in Figure 4 provides us with some idea on the Algerian fire dataset.

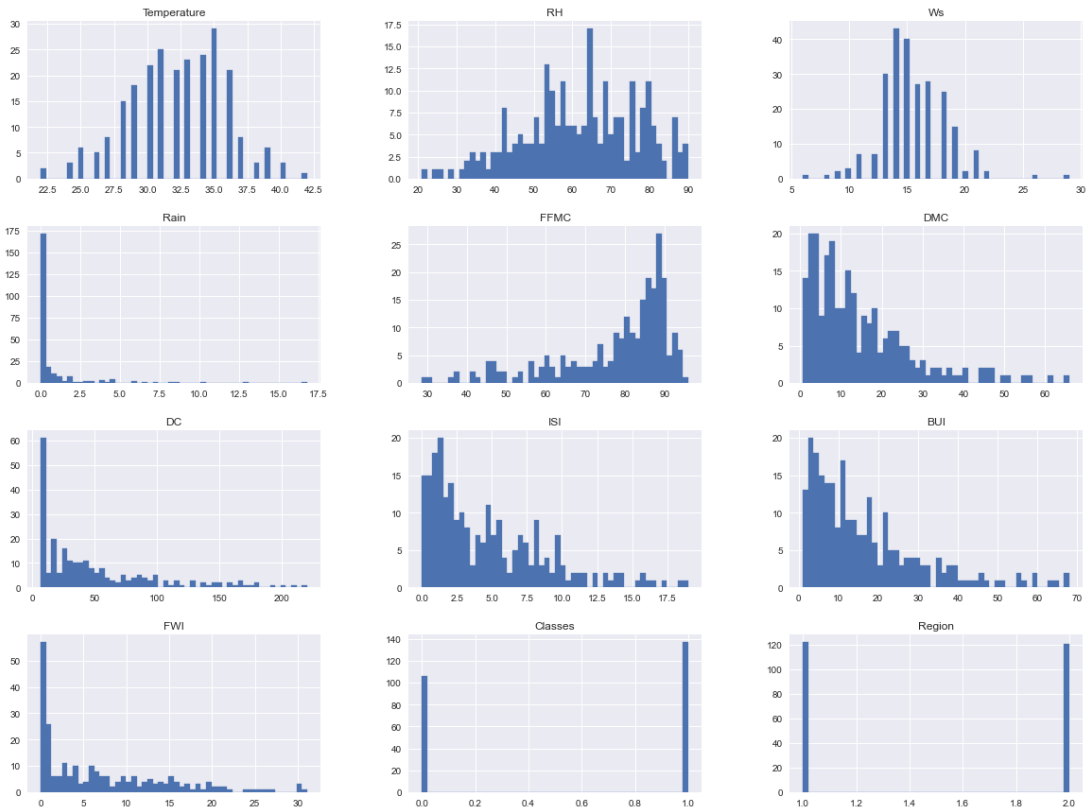

Figure 4: Histogram plot of Algerian fire dataset

a) *Temperature*: The distribution of temperature values appears roughly symmetric, with the most frequent values between 30 and 35 degrees Celsius.

b) *RH* (Relative Humidity): The distribution is more spread out, with the highest frequency around 50-60% relative humidity. There are still some values between 30 and 90%.

c) *Ws* (Wind Speed): This has a peak around 15, meaning wind speed typically clusters around this value, and fewer data points appear at lower and higher speeds.

d) *Rain*: The data for rain is highly skewed to the left, with most values close to zero, indicating very few instances of significant rainfall.

e) *FFMC* (Fine Fuel Moisture Code): The distribution shows that most values are concentrated around 80 to 90, meaning most of the observations have high FFMC values.

f) *DMC* (Duff Moisture Code): The distribution is skewed to the left with a peak around 0 to 10, meaning most of the observations have low DMC values, with fewer instances having higher values.

g) *DC* (Drought Code): The DC data also exhibits a skewed left distribution, with most values clustering around the lower end of the scale but extending up to 200.

h) *ISI* (Initial Spread Index): This distribution is left-skewed, with most values between 0 and 5. There are few instances of higher ISI values.

i) *BUI* (Buildup Index): This shows a similar left-skewed distribution, with a peak in lower values (0 to 20), indicating lower BUI is common.

j) *FWI* (Fire Weather Index): The distribution shows most values are concentrated between 0 and 10, with a steep drop-off after that.

k) *Classes*: This likely represents some form of classification (e.g., fire severity), and the histogram shows that one class (possibly 0) dominates the data, while other classes have fewer occurrences.

l) *Region*: This variable has only two distinct categories, with one region (possibly represented by 1.0) dominating the data.

Key Observations in the histograms:

- Many of the features (Rain, DC, ISI, BUI, etc.) show a left-skewed distribution, meaning most values are low, with a few high outliers.
- Temperature and wind speed (Ws) show more symmetric distributions.
- The dominance of one class in the "Classes" and "Region" histograms could imply an imbalanced dataset for those categorical variables.

The **heatmap** of multicollinearity check in the Algerian fire dataset provides some insight information in the relationship between features in the dataset.

Key Observations in the heatmap:

- Strong positive correlations between fire-related indices (e.g., BUI, DC, FWI, ISI) suggest that

Figure 5: Multicollinearity check in Algerian fire dataset

these factors increase together, affecting fire severity.

- Negative correlations between relative humidity and fire indices (FWI, ISI, FFMC) highlight the role of moisture in reducing fire spread risk.
- Temperature and certain moisture indices (DMC, FFMC) show moderate correlations, indicating a relationship between weather conditions and fire risk.

This **bar plot** displays a month-wise fire analysis for the Béjaïa region, comparing the occurrence of fire and non-fire incidents across the months of June, July, August, and September.

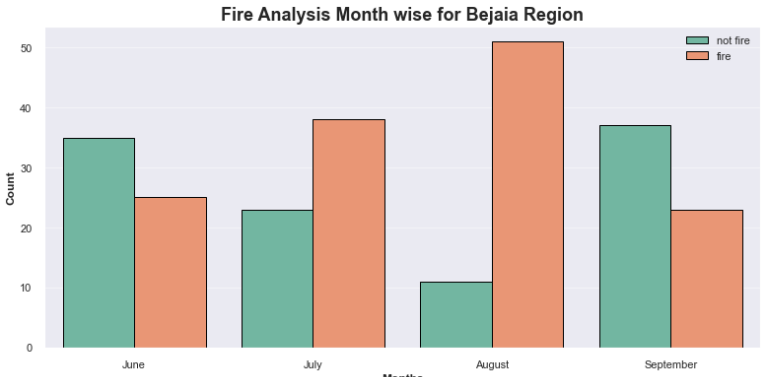

Figure 6: Monthly fire analysis of Algerian fire dataset

Observation in the bar graph:

- August stands out as the peak month for fire incidents in the Béjaïa region, with significantly more fire events compared to non-fire events.
- In contrast, June and September have more non-fire events, while July shows a similar count for both fire and non-fire events.

This plot helps understand the seasonal variation of fire occurrences, showing that fire incidents are more likely in the late summer (July and August) in the Béjaïa region. This could indicate that environmental factors, like temperature and dry conditions, are more conducive to fires during these months.

In this **bar graph** there is a clear increase in fire occurrences as temperatures rise, peaking between 34°C and 35°C. After this peak, the number of fires decreases, indicating that the highest fire risk occurs at these temperature levels.

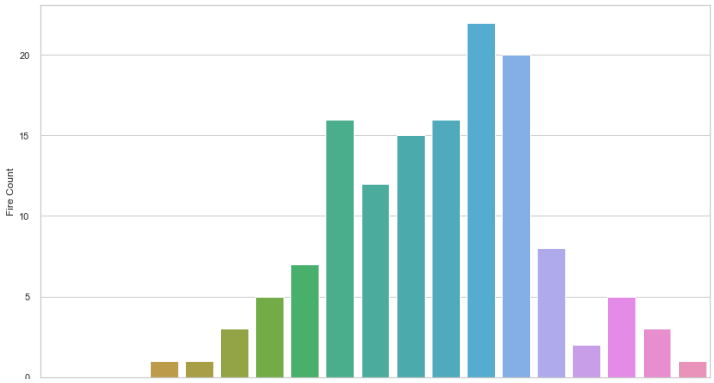

Figure 7: Temperature data representation of Algerian fire dataset

From this we can conclude that, high temperatures, especially around 34-35°C, significantly contribute to fire incidents, likely due to the drying out of vegetation and increased flammability.

We can observer that almost all fires occur when there is **very little to no rainfall** (below 0.2 mm). After this point, the fire count drops dramatically, with almost no fires occurring as rainfall increases. Hence, it is conclusive that drier conditions (low rainfall) are the most critical factor for fire outbreaks. Rainfall appears to have a strong dampening effect on the number of fires, as more moisture makes ignition less likely.

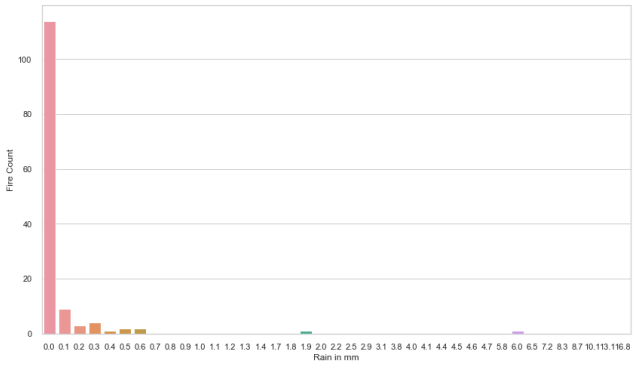

Figure 8: Rainfall in mm in Algerian fire dataset

Here, fire activity peaks when **wind speeds** are between 13 km/h and 16 km/h. Beyond this range, the fire count decreases. Moderate wind speeds seem to enhance fire spread by providing enough airflow to fuel the flames, while lower and higher wind speeds may limit fire occurrences. Very high winds might disperse fire sources, while lower winds may not provide enough oxygen for the fire to spread rapidly.

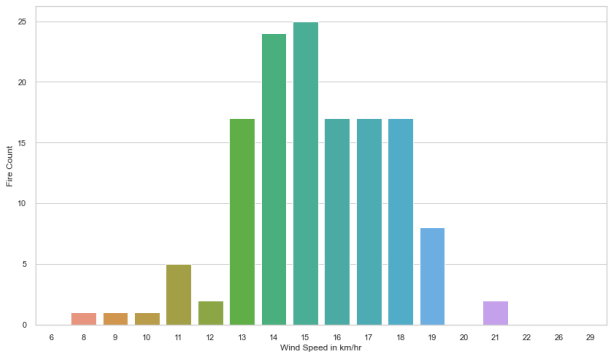

Figure 9: Wind speed

The graph suggests that relative humidity is a significant factor in determining the number of fires. There seems to be a general inverse relationship between relative humidity and the number of fires. As the relative humidity increases, the number of fires tends to decrease. This suggests that higher humidity levels may have a suppressing effect on fires. There are noticeable clusters of fires at certain relative humidity levels, such as around 30-35% and 60-70%. This might indicate that specific weather conditions or factors associated with these humidity levels are more conducive to fires. The bar corresponding to 55% relative humidity stands out with an exceptionally high number of fires. This could be due to a specific event or circumstance that caused a significant spike in fires at that particular humidity level.

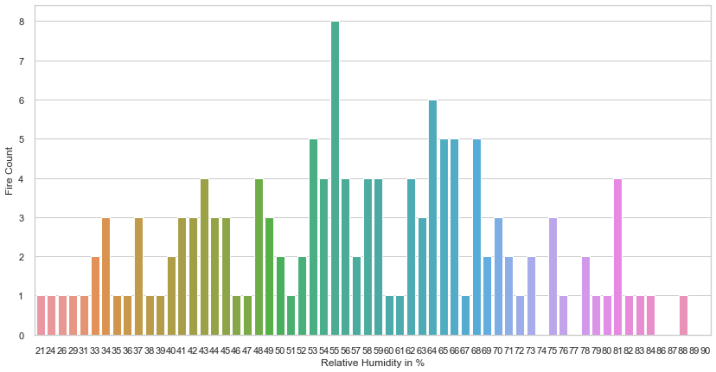

Figure 10: Relative humidity

5.3 Pre-processing

Data preprocessing is important for making sure our datasets are clean and ready for training. Here are the steps we will take:

1) Handling Missing Values: Missing values will be filled in or removed, depending on their impact.
2) Encoding Categorical Variables: We will turn categorical features, like region, into numerical data using one-hot or label encoding.
3) Feature Scaling: We will normalize or standardize features like temperature and wind speed to make sure they're on a similar scale.
4) Date Feature Extraction: Dates will be transformed to get useful information like month or day of the week.
5) Handling Outliers: We will find and handle outliers so they don't affect our model badly.
6) Data Integration: We will combine all datasets — fire data, weather data, and El Niño data — into one comprehensive dataset for modeling.

7) Feature Engineering: We will create new features that could improve the model's performance, like combining temperature, humidity, and wind speed.
8) Data Balancing: Since forest fires are rare, we will balance the dataset so the model can learn well from both fire and non-fire cases.
9) Splitting Data: The data will be split into training, validation, and test sets to make sure our model performs well on new data.

These preprocessing steps will help make sure our data is suitable for training a reliable machine learning model.

5.4 Modeling

To find the best model for predicting forest fires, we will use **cross-validation** and grid search to compare different machine learning algorithms and tune their settings. Cross-validation will help make sure the models can work well on new data, and grid search will help find the best settings for each model.

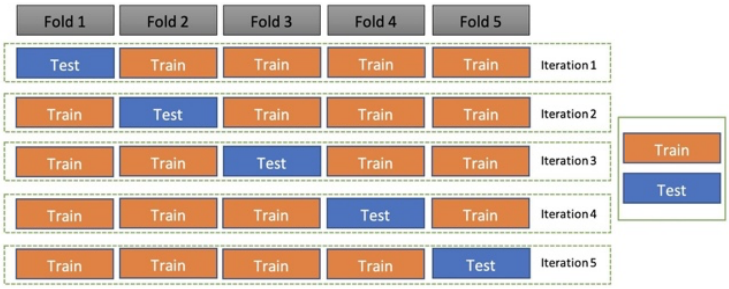

Figure 11: K-Fold cross validation

1) Random Forest and XGBoost are good options because they handle many features and complex relationships well.
2) Gradient Boosting is also strong for modeling interactions between features.
3) Regression models will be used as a simple baseline model to compare with the others.

We will compare these models based on accuracy, precision, recall, and F1 score, then choose the best one for predicting forest fires.

5.5 Training

Training the model means using the training data to teach the model how to predict forest fires. We will use the scikit-learn library to train our models. The model will learn by finding the relationships between input features, like temperature and wind speed, and the target feature, fire occurrence.

Steps in Training:

1) Model Initialization: Start the model with the best settings found in grid search.
2) Fitting the Model: Train the model with the *.fit()* method.
3) Monitoring Performance: Keep track of how well the model is learning by checking the loss function and metrics like accuracy.
4) Addressing Overfitting: Use methods like early stopping to prevent the model from overfitting.

After training, we will evaluate the model with the validation set. If needed, we will adjust the settings to improve performance, then test the final model with new data.

5.6 Evaluation

We will evaluate the model to make sure it accurately predicts forest fires. We will use these metrics:

- Accuracy: How often the model makes the correct prediction.
- Precision: How often the model correctly predicts a fire when there is one.
- Recall: How many actual fires the model can predict.
- F1 Score: A balance between precision and recall.

These metrics will help us understand how well the model works and what areas need improvement.

5.7 Deployment

The forest fire prediction model will be deployed using AWS Machine Learning for hosting and Dash for the web-based interface. AWS Machine Learning provides a safe and scalable way to host machine learning models.

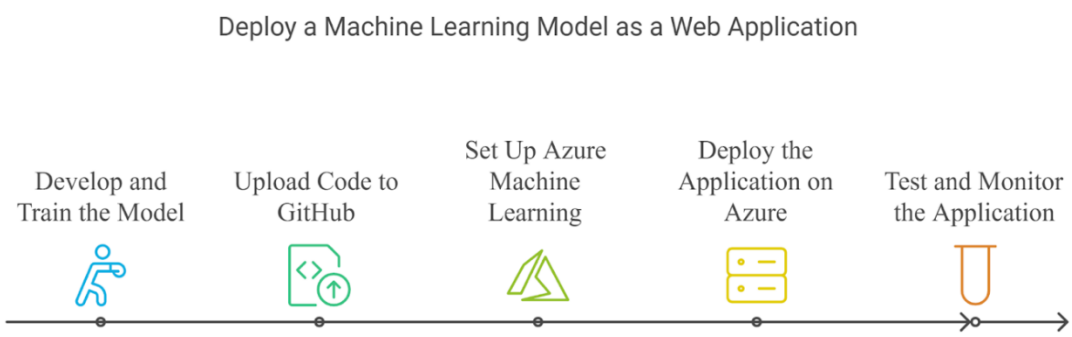

Figure 12: Web Application Deployment

Here are the steps for deployment:

1) Create Prediction Model and Web Application Locally
    a) Develop ML Model (*model.py*): Build the model in a Python script named *model.py*.
    b) Convert Model to Deployable Format: Save the trained model using *pickle* or *joblib* to prepare it for deployment.
    c) Create Dash Web Application (*app.py*): Make a Dash web app that lets users input data and see predictions, including maps showing fire risk.
2) Upload Code to GitHub: The relevant codes, including the *model.py* and *app.py* scripts, along with dependencies, will be uploaded to GitHub for version control and collaboration.
3) AWS Machine Learning Setup
    a) Create AWS Account: Use AWS to access Machine Learning services.
    b) Register Model on Azure: Register the model in AWS to manage it easily.
    c) Create an Inference Endpoint: Create an endpoint that the Dash app can use to get predictions from the model.
    d) Containerize the Model: Use Docker to containerize the model, making it easier to deploy.
4) Deploy Dash Application on AWS Web App
    a) Create AWS Web App Service: Use AWS App Service to host the Dash app.
    b) Configure AWS App Service: Connect the web app to the model endpoint.
    c) Clone Repository and Deploy: Clone the GitHub repository to AWS and deploy the app.
5) Testing and Monitoring
    a) Test the Deployment: Make sure the deployed app works well and interacts properly with the model.
    b) Monitor the Application: Use Azure's monitoring tools to track performance and fix any issues.

By using AWS Machine Learning and Dash, we will make the forest fire prediction system accessible online. This system will help users input data and see predictions in real time, including maps that show areas at risk of forest fires.

5.8 Machine Learning Model

To predict forest fires effectively, we employed multiple machine learning algorithms, including **Logistic Regression**, **Support Vector Machines (SVM)**, **Random Forest**, **Decision Tree**, and **XGBoost**. The objective was to evaluate the performance of each model in capturing the underlying relationships in the dataset and identifying the most suitable model for accurate prediction in this context.

Each of these models was selected based on their respective strengths:

1. **Logistic Regression**: Logistic regression was utilized as a baseline model to establish a reference for the performance of more complex algorithms. It assumes a linear relationship between the independent variables (e.g., temperature, wind speed, humidity) and the dependent variable (fire intensity or occurrence).
2. **Support Vector Machines (SVM)**: SVM was included due to its ability to handle high-dimensional data and separate classes effectively using a hyperplane. By employing kernels, SVM can capture non-linear relationships, which are often present in forest fire prediction scenarios.
3. **Decision Tree**: This algorithm provides an interpretable model, making it easier to identify the key features contributing to forest fire prediction. Decision Trees split data recursively based on feature thresholds, which is advantageous for capturing non-linear patterns.
4. **Random Forest**: Random Forest, an ensemble method, was selected for its robustness against overfitting and its ability to handle complex interactions between features. By averaging the predictions from multiple decision trees, it reduces variance and enhances generalizability.
5. **XGBoost**: Known for its efficiency and accuracy, XGBoost (Extreme Gradient Boosting) builds models iteratively by minimizing errors from previous iterations. It is particularly suitable for structured data and often outperforms other models in prediction tasks due to its ability to capture intricate patterns.

CHAPTER 6
FINDINGS & ANALYSIS

This chapter presents results of the analysis conducted on various environmental factors that influence forest fire occurrence. The findings provide insights into the conditions under which forest fires are most likely to occur, helping to identify critical thresholds for different parameters that contribute to fire risk.

- *Temperature*: The highest fire counts occurred between 30-37 degrees Celsius.
- *Rain*: The highest fire counts occurred when there was no rain or very little rain, i.e., between 0.0 to 0.3.
- *Wind Speed*: The highest fire counts occurred when the wind speed was between 13 to 19 km/hr.f
- *Relative Humidity*: The highest fire counts occurred when the relative humidity (RH) was between 50 to 80%.
- *Fine Fuel Moisture Code* (FFMC) index, which ranges between 28.6 to 92.5, shows that above 75 there is a higher chance of forest fires.
- *Duff Moisture Code* (DMC) index, which ranges between 1.1 to 65.9, indicates that values between 1.1-10 have a lower chance of forest fires, whereas above 10-30 DMC has very high evidence of forest fires in the past.
- *Drought Code* (DC) index, which ranges between 7 to 220.4, shows that values between 0-25 are safe and have a lower chance of forest fires, whereas values above 25 DC have a higher chance of forest fires.
- *Initial Spread Index* (ISI) index, which ranges between 0 to 18, shows that values between 0-3 have lower fire chances, and values above 3 ISI have a higher chance of forest fires.
- *Buildup Index* (BUI) index, which ranges between 1.1 to 68, shows that values between 1.1 to 10 have a lower fire chance, and values above 10 BUI have a higher chance of forest fires.
- *Fire Weather Index* (FWI) index, which ranges between 1 to 31.1, shows that values between 0-3 have a lower chance of forest fires, and values between 3-25 FWI have a higher chance of forest fires.

Findings and Analysis

The violin plot in Figure 13 provides a comprehensive visual representation of the distribution of various factors potentially influencing forest fires. Temperature and humidity, while displaying relatively normal distributions, might not be the strongest indicators of fire risk individually. In contrast, wind and rain exhibit skewed distributions, suggesting that high wind speeds and low rainfall could significantly increase fire risk. The fuel-related factors, FFMC, DMC, DC, and ISI, show skewed distributions towards higher values, indicating that areas with dry fuels are more prone to fires. The distribution of nino34_sst, related to sea surface temperatures, appears relatively normal, while the fire occurrences are heavily skewed, with most days having no fires and a few experiencing significant fire activity.

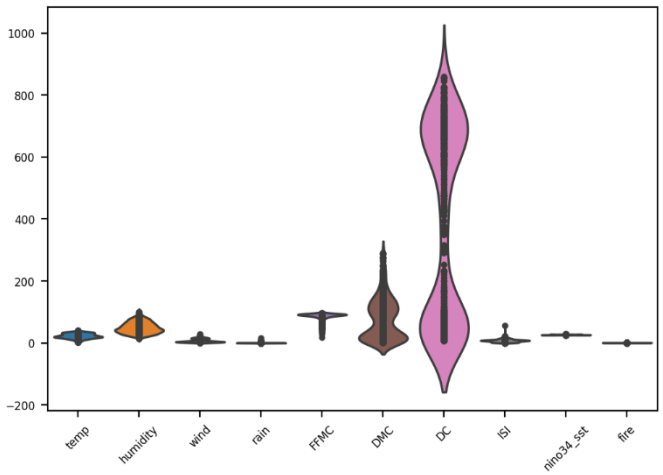

Figure 13: Violin plot of various features related to forest fires

The correlation matrix in Figure 14 reveals a complex interplay between various factors influencing forest fire occurrence. Fuel-related variables, such as FFMC, DMC, DC, and ISI, exhibit strong positive correlations, suggesting their interconnectedness in contributing to fire risk. Temperature and humidity show a moderate negative correlation, indicating that higher temperatures often coincide with lower humidity, potentially leading to drier conditions and increased fire risk. Similarly, wind and rain exhibit a moderate negative correlation, implying that

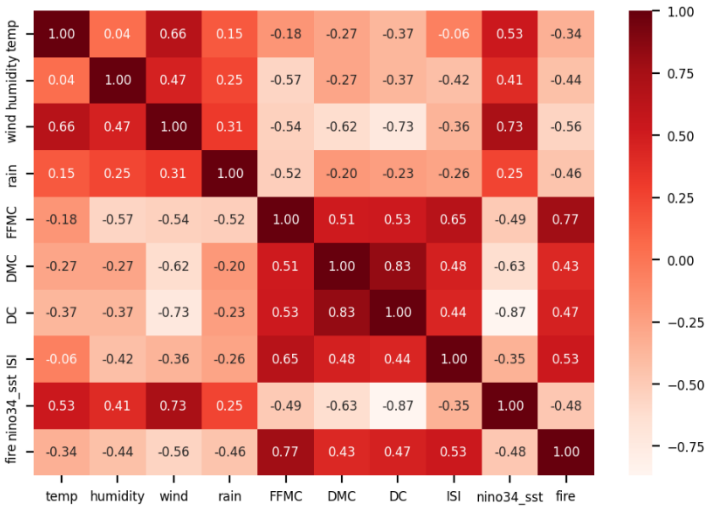

Figure 14: Correlation matrix exploring the factors influencing forest fires

areas with higher wind speeds tend to experience lower rainfall, creating conditions favorable for fires. Interestingly, fire has strong negative correlations with FFMC, DMC, DC, and ISI, suggesting that these factors might be associated with more mature forests with higher moisture content, which are less prone to fire. While temperature has a weak negative correlation with fire, other factors might play a more significant role in determining fire occurrence. nino34_sst, related to sea surface temperatures, appears to have a relatively weak correlation with other variables, suggesting its impact on fire risk might be less direct. It's important to remember that correlation does not imply causation, and further analysis is needed to establish causal relationships between these variables.

The pair plot in Figure 15 provides a comprehensive visual representation of the relationships between different variables in the dataset. Each subplot in the matrix represents the scatter plot between two variables, while the diagonal plots showcase the distribution of each individual variable. Some variables exhibit linear relationships, meaning that as one variable increase or decreases, the other tends to follow a similar trend. For instance, the relationship between FFMC, DMC, DC, and ISI appear linear. However, other variables display non-linear relationships, where the connection between the two variables is not straightforward, suggesting a more complex relationship that might not be easily captured by linear models. Additionally, some plots reveal the presence of outliers, which are data points that deviate significantly from the general trend. Outliers can influence the analysis and modeling process, making it crucial to identify and handle them appropriately. The diagonal plots offer insights into the distribution of each variable, with some variables appearing normally distributed and others exhibiting skewed distributions or other patterns.

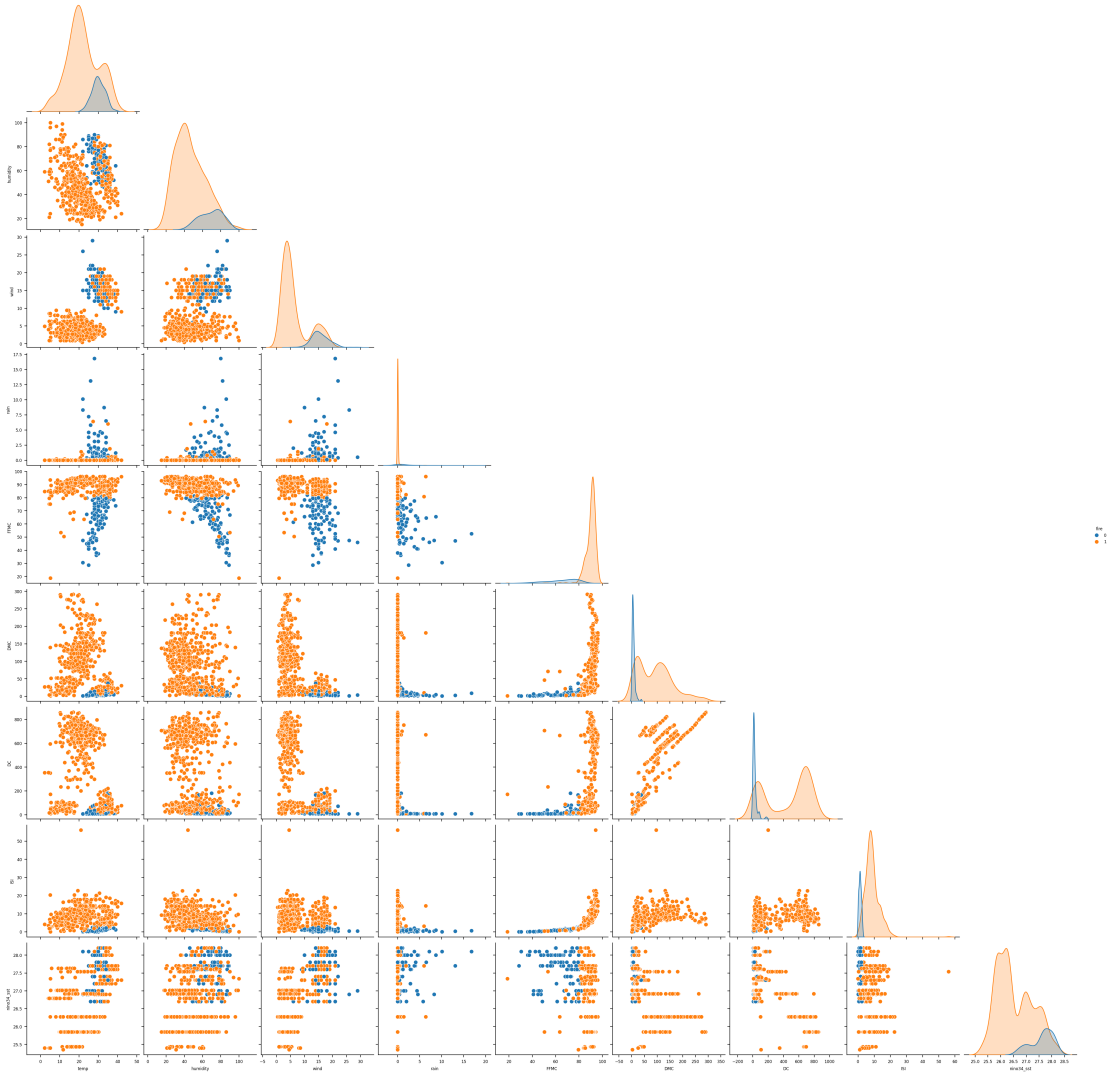

Figure 15: Pair plot between different features

Class Imbalance

The count plot in Figure 16 reveals a significant class imbalance in the dataset, with the majority of instances belonging to the class labeled "1" (fire). This imbalance poses a challenge for machine learning models, as they may become biased towards the majority class, leading to poor performance in predicting the minority class (not fire). To address this issue, we have used down-sampling, up-sampling, and SMOTE up-sampling technique (Figure 17). By addressing the class imbalance, we are looking forward to improve the model's ability to accurately predict fire occurrences and make more informed decisions about fire prevention and management strategies.

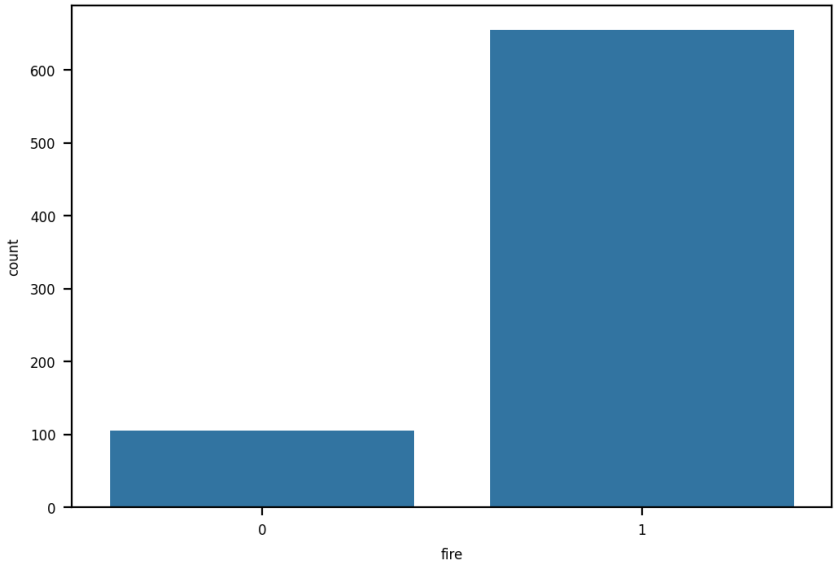

Figure 16: Class imbalance in fire status

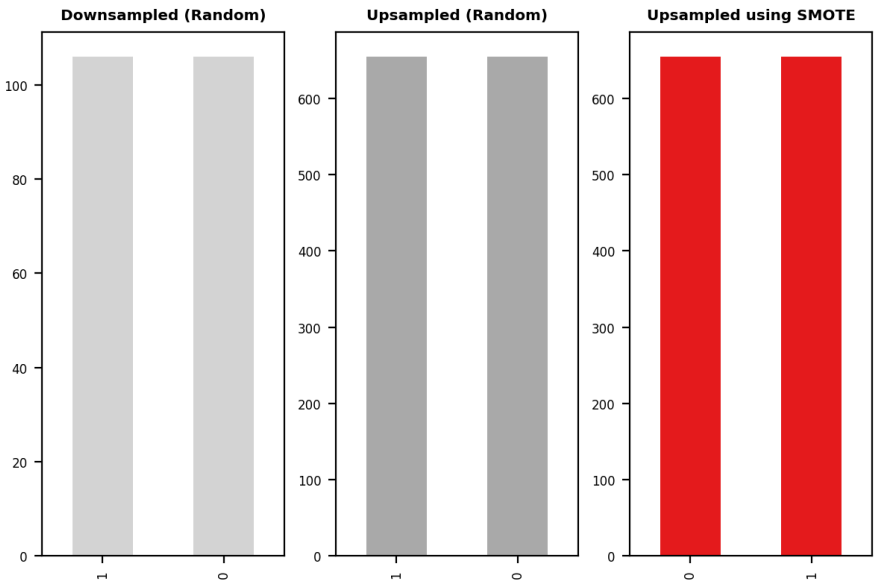

Figure 17: Down-sampled, Up-sampled, and SMOTE up-sampled

CHAPTER 7
MODEL EVALUATION RESULTS

Each model was trained and validated using a combination of cross-validation techniques to ensure robustness and reduce the risk of overfitting. Performance metrics such as Accuracy, Precision, Recall, and F1-score for classification scenarios were used to evaluate the models.

Cross-validation

All models demonstrated high accuracy (Figure 18), with XGBoost, SVM, and Random Forest consistently achieving near-perfect performance regardless of the data balancing method.

Default hyperparameters

	Accuracy	Precision	Recall	F1
Random Forest	0.995229	0.996154	0.994212	0.995150
XGBoost	0.995224	0.998058	0.992252	0.995141
SVM	0.974204	1.000000	0.947647	0.972771

Best hyperparameters

	Accuracy	Precision	Recall	F1
Random Forest	0.996186	0.998077	0.994212	0.996116
XGBoost	0.997134	0.998077	0.996135	0.997092
SVM	0.994267	0.996190	0.992270	0.994174

Figure 18: Model performance - Random Forest, XGBoost, and SVM

Down-sampling, however, led to slightly lower accuracy than SMOTE or up sampling, due to its reduction in data volume. Precision was high across most models, particularly with XGBoost and Random Forest, and was further improved with SMOTE, which helped models accurately identify fire days with minimal false positives (Figure 19). Conversely, simpler models like Decision Tree and KNN showed moderately lower precision, especially when applied to down-sampled data.

In terms of recall, which is critical for identifying actual fire days, SMOTE significantly enhanced recall in models like XGBoost, SVM, and Random Forest, while down

sampling tended to reduce recall due to the reduced dataset size. The F1-score, which balances precision and recall, highlighted XGBoost, SVM, and Random Forest as top performers across all sampling methods, with SMOTE consistently providing the best results. Although Logistic Regression and Naive Bayes delivered steady performance, they fell short of the advanced models in terms of precision and recall.

Overall, XGBoost, SVM, and Random Forest stand out as the leading models, especially when paired with SMOTE, which proves to be the most effective sampling method for improving fire detection accuracy. While simpler models like Decision Tree and KNN performed reasonably, they did not match the precision and recall of the more sophisticated models. Moving forward, SMOTE is recommended for dataset balancing, with XGBoost, SVM, and Random Forest prioritized for further tuning and application, especially with a focus on achieving high recall and F1-scores for reliable fire detection.

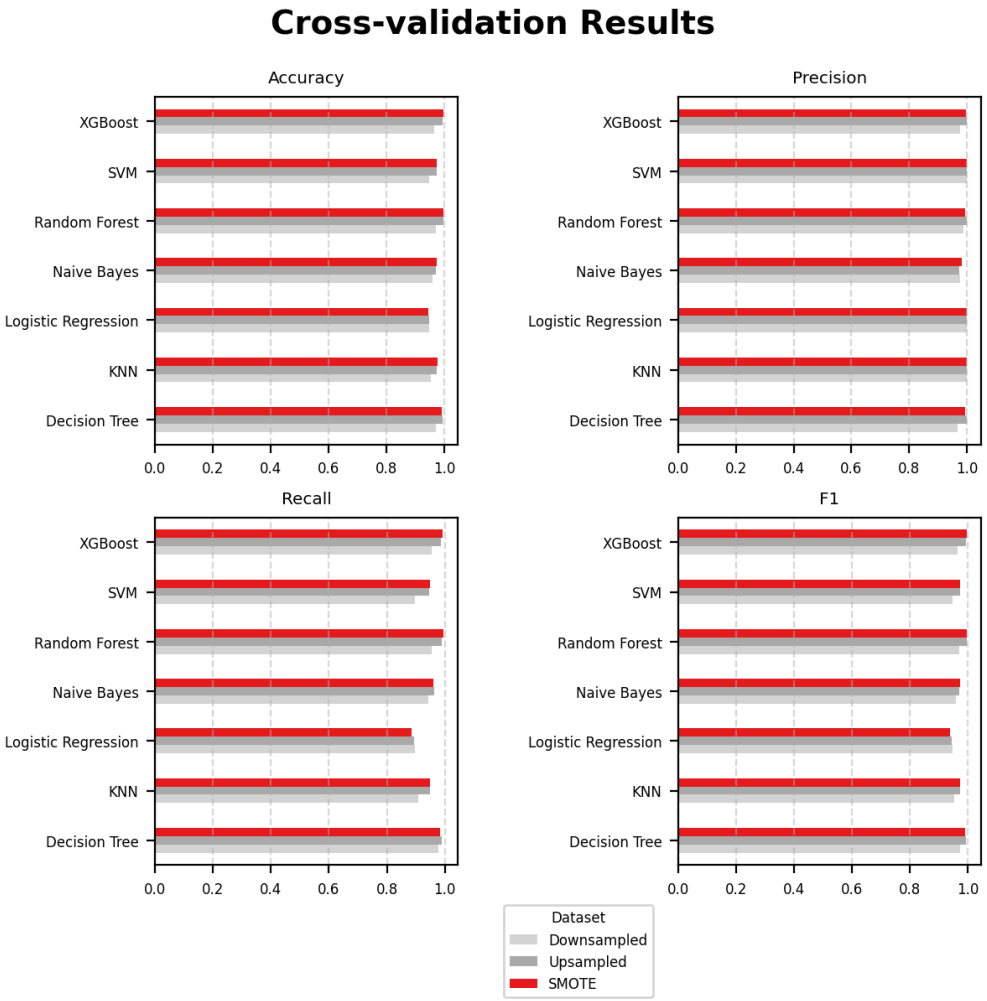

Figure 19: Cross Validation results across different models

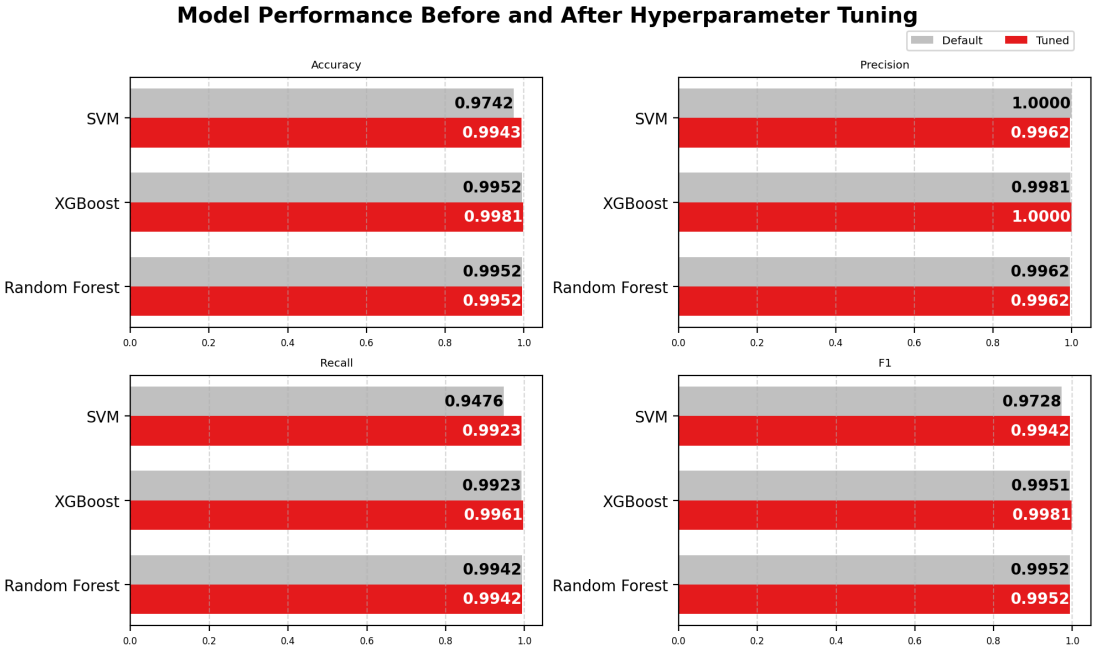

Figure 20: Model Performance Before and After Hyperparameter Tuning

Observations

The three classifiers—SVM, XGBoost, and Random Forest—demonstrate consistently high performance across all metrics (Figure 20), suggesting they are well-tuned. Among them, SVM achieves the highest recall, making it an optimal choice if the priority is to capture all potential fire events and minimize false negatives. XGBoost and Random Forest, however, have very close scores in precision, F1-score, and accuracy, indicating their reliability in both detecting actual fire events and avoiding false positives.

Choosing the Best Classifier

In selecting the best classifier, recall is a critical metric, as it measures the model's ability to catch all actual fire cases, reducing the risk of missing a true fire event, which could allow fires to grow unchecked and lead to dangerous situations. This emphasis on recall aligns with the importance of avoiding false negatives, as missing an actual fire could result in significant costs and damage, whereas a false positive would primarily result in additional inspections, which pose less risk. Although recall is prioritized, maintaining a balance with precision is also essential to prevent an excessive number of false alarms; however, recall remains the primary focus to ensure reliable detection of fire risks.

Choosing SVM

For the prediction of forest fires, we evaluated several machine learning models, including Logistic Regression, Support Vector Machines (SVM), Decision Trees, Random Forest, and XGBoost. Initial results revealed that while Logistic Regression offered insights into general trends, it was insufficient for capturing the intricate, non-linear, and interactive relationships inherent in forest fire datasets. Decision Trees provided better interpretability but were prone to overfitting without additional constraints. Similarly, Random Forest and XGBoost showed strong predictive capabilities; however, they were resource-intensive and required extensive tuning.

Based on our analysis, SVM emerged as the most suitable model for this scenario due to the following reasons:

1. Simplicity and Interpretability
    - SVM is simpler and easier to understand, especially if it's a linear SVM. It creates a clear boundary between fire and non-fire cases.
    - XGBoost and Random Forest combine many decisions, which can make it harder to explain why a specific prediction was made.

2. Less Tuning Needed
    - SVM generally performs well with fewer settings (hyperparameters) to adjust, making it quicker to set up.
    - XGBoost and Random Forest have many settings that need fine-tuning to work well, which can take extra time and effort.

3. Efficient and Fast
    - SVM can be more memory-efficient, using less computer power, especially if you have limited resources.
    - XGBoost and Random Forest need more memory and processing power, especially with larger datasets.

4. Lower Risk of Overfitting
    - SVM has a regularization setting (C) that controls complexity, which helps it avoid overfitting (memorizing the data too closely).
    - XGBoost and Random Forest can sometimes overfit, especially if they’re not carefully tuned, meaning they might perform worse on new data.

5. Good Generalization

- SVM is known to generalize well to new data, meaning it can be stable and consistent even when used on data it hasn't seen before.
- XGBoost and Random Forest are powerful but can sometimes be more sensitive to changes in the data.

The performance of SVM was validated through various evaluation metrics such as accuracy, precision, recall, and F1-score. While ensemble methods such as Random Forest and XGBoost demonstrated slightly higher predictive power under specific tuning conditions, SVM was chosen for its simplicity, efficiency, and interpretability, especially in scenarios where computational resources and time are constrained. Additionally, its lower risk of overfitting ensures reliable predictions across varying environmental datasets.

This selection aligns with our goal of building a robust, interpretable, and resource-efficient model for forest fire prediction while balancing complexity and accuracy. Further chapters detail the experimental setup, hyperparameter optimization, and performance comparisons for SVM.

Confusion Matrix – SVM

Figure 21 illustrates the confusion matrix showing that the SVM model is highly accurate, with 137 true positives and 123 true negatives, indicating that most predictions are correct. There are no false positives, demonstrating high precision and reliability. Additionally, the model has good recall, with only two missed fire cases (false negatives), meaning it successfully detects nearly all actual fire events. Overall, this SVM model effectively identifies fire occurrences with minimal errors, ensuring strong sensitivity in its predictions.

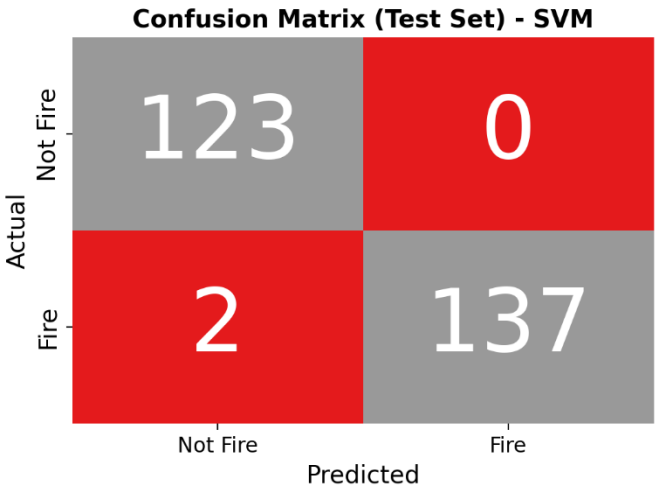

Figure 21: Confusion Matrix (Test set) - SVM

CHAPTER 8
DISCUSSIONS

The present study aimed to investigate the potential of machine learning models in predicting forest fires specifically in Portugal and Algeria. The findings of this research not only contribute to the growing body of literature on forest fire prediction but also provide practical implications for forest management strategies.

Our analysis indicated that the Support Vector Machine (SVM) model significantly generalized better than other machine learning algorithms, including Logistic Regression, Decision Trees, Random Forest, and XGBoost, with respect to accuracy, precision, and recall metrics. This suggests that SVM is particularly effective in handling the complexity and nonlinearities often present in environmental data. The successful application of SVM underscores the importance of selecting appropriate algorithms tailored to the specific characteristics of the dataset and the predictive task at hand.

The implications of our findings are profound for forest fire management and mitigation strategies. The ability to predict forest fires with high accuracy enables timely interventions that can significantly reduce the potential for catastrophic wildfires. By implementing machine learning models like SVM into existing early warning systems, forest managers can allocate resources more efficiently and improve response times during critical wildfire season. Moreover, these predictive models can assist in risk assessment and planning, allowing for targeted preventive measures in high-risk areas.

Our results align with previous research that has explored the use of machine learning for forest fire prediction. For instance, a study by Tang et al. (2024) acknowledged the effectiveness of machine learning models, particularly SVM, in regions prone to wildfires, such as Thailand. This consistency across different geographic contexts suggests that machine learning techniques can serve as a universal tool in forest fire prediction. Additionally, our findings complement the work of Zaidi (2023), which highlighted the utility of machine learning in predicting wildfires in the Algerian

landscape, reinforcing the notion that such methodologies can be adapted to diverse ecological environments.

Limitations of the Study

Despite the promising results, this study is not without its limitations. The analysis was constrained to specific geographic regions, which may limit the generalizability of the findings. Future studies should aim to validate these results across different climates and ecosystems to assess the robustness of the SVM model. Furthermore, the dataset utilized did not account for certain variables, such as human-induced factors like arson or land-use changes, which can significantly influence fire occurrence. Future research should incorporate these elements to enhance the predictive accuracy of machine learning models.

Future Research Directions

This study lays the groundwork for future investigations into forest fire prediction using machine learning. Future research could explore the incorporation of additional data sources, such as satellite imagery and real-time environmental sensors, to augment the predictive capabilities of the models. Moreover, integrating human activity data could provide a more holistic understanding of fire dynamics, allowing for improved model performance.

Other avenues for exploration include:

1. **Development of Hybrid Models**: Combining different machine learning techniques or integrating deep learning models may yield even more accurate predictions and insights regarding forest fire behavior.
2. **Real-time Predictive Systems**: Developing real-time monitoring systems that utilize machine learning could prove indispensable in forest fire management, offering decision-makers immediate data and forecasts during high-risk periods.
3. **Community Engagement**: Research could also examine how community involvement and awareness programs could complement machine learning predictions to create a more comprehensive approach to wildfire prevention.

CHAPTER 9
CONCLUSION

The present study aimed to investigate the potential of machine learning models in predicting forest fires in Portugal and Algeria. Through a comprehensive analysis of various machine learning algorithms, our research demonstrated the efficacy of Support Vector Machines (SVM) in predicting forest fires with high accuracy, precision, and recall. The findings of this study contribute significantly to the growing body of literature on forest fire prediction, highlighting the importance of selecting appropriate algorithms tailored to the specific characteristics of the dataset and the predictive task at hand.

The successful application of SVM in this study underscores the potential of machine learning models in enhancing forest fire management and prevention strategies. By integrating these models into existing early warning systems, forest managers can allocate resources more efficiently, improve response times during critical wildfire season, and ultimately reduce the risk of devastating wildfires. Moreover, the predictive capabilities of these models can assist in risk assessment and planning, allowing for targeted preventive measures in high-risk areas.

While this study provides valuable insights into the application of machine learning models in forest fire prediction, it also highlights the need for further research in several areas. Future studies should aim to validate these results across different climates and ecosystems, incorporate additional data sources, and explore the development of hybrid models and real-time predictive systems. Furthermore, integrating human activity data and community engagement programs could provide a more holistic understanding of fire dynamics and enhance the predictive accuracy of machine learning models.

In conclusion, this study demonstrates the potential of machine learning models in predicting forest fires and highlights the importance of continued research in this domain. As the frequency and intensity of wildfires continue to escalate globally, the integration of advanced predictive analytics will be paramount in enhancing forest fire management and prevention strategies. Our findings provide a foundation for future

research and have significant implications for policymakers, forest managers, and researchers seeking to mitigate the impacts of wildfires on ecosystems and human communities.

Ultimately, this study contributes to the development of a more comprehensive approach to forest fire management, one that leverages the strengths of machine learning models to predict and prevent wildfires, while also acknowledging the complexities and uncertainties inherent in these events. By continuing to advance our understanding of forest fire dynamics and the predictive capabilities of machine learning models, we can work towards creating a safer and more sustainable future for ecosystems and human communities alike.

REFERENCES

[1] Sanjeev Sharma, Puskar Khanal 2024. Forest Fire Prediction: A Spatial Machine Learning and Neural Network Approach. Fire, Available at: https://www.mdpi.com/2571-6255/7/6/205 [Accessed 7 Oct. 2024].

[2] Abdelhamid Zaidi, 2023. Predicting wildfires in Algerian forests using machine learning models. Heliyon. Available at: https://www.sciencedirect.com/science/article/pii/S2405844023052726 [Accessed 7 Oct. 2024]

[3] Tang, J. et al. (2024). Toward a more resilient Thailand: Developing a machine learning-powered forest fire warning system', Heliyon. Available at: https://www.sciencedirect.com/science/article/pii/S2405844024100527?via%3Dihub [Accessed 7 Oct. 2024]

[4] Zaidi, A. (2023). Predicting wildfires in Algerian forests using machine learning models. Heliyon. Available at: https://www.sciencedirect.com/science/article/pii/S2405844023052726?via%3Dihub [Accessed 7 Oct. 2024]

[5] Bharathi, V. and Pedda Reddy, C. (2023). Comparison of forest fire prediction system using machine learning algorithms, in Proceedings of the 2023 3rd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), Greater Noida, India. Available at: https://ieeexplore.ieee.org/document/10182818 [Accessed 7 Oct. 2024]

[6] Datasets:

A. Algerian Forest Fires Dataset: https://www.kaggle.com/datasets/nitinchoudhary012/algerian-forest-fires-dataset

B. Forest fires Dataset Portugal: https://www.kaggle.com/datasets/ishandutta/forest-fires-data-set-portugal