Machine Learning: Mini Project

Wildfire Prediction

Name : Alviya Ali
PRN : 24070243005

- ## Introduction

  Wildfires are the most devastating natural disasters with disastrous environmental, economic, and social consequences. Wildfires lead to deforestation, wildlife and vegetation loss, air pollution, and severe damage to human life and infrastructure. As the temperature of the world is increasing and the environment is experiencing climate changes, wildfires are on the rise and increasing in intensity, which calls for accurate prediction to manage the disaster effectively. Wildfire propagation is regulated by meteorological parameters, i.e., temperature, humidity, wind, and rain, and fire danger indices such as FFMC, DMC, DC, and ISI. Geographical information is also used towards mapping regions of high hazard. The research builds a machine learning model for the prediction of forest fire burned area as a function of meteorological and geographic variables. The aim is to enhance wildfire hazard determination and provide pre-emptive chances for fire management.

- ## Problem Statement

  Wildfires pose a major threat to the environment, human life, and infrastructure. Forecasting the spread of fires is critical for efficient prevention and response planning. Conventional methods rely on past trends and weather conditions, which might not be accurate in quantifying the fire effect area. The aim of this project is to develop a predictive model based on machine learning to forecast the area burned based on meteorological and spatial variables, which will aid proactive wildfire management.

- ## Objective

- To analyze key meteorological and spatial factors influencing wildfire spread.
- To develop and compare multiple machine learning models for predicting the burned area of wildfires.
- To evaluate the performance of these models using standard regression metrics.
- To provide insights that enhance wildfire risk assessment and emergency preparedness.

- ## Dataset Description

The dataset used in this study originates from Portugal’s Montesinho Park and contains meteorological, temporal, and spatial attributes related to wildfire occurrences. It includes the following features:

- **Meteorological Variables:**
  - Temperature (°C): Measures ambient temperature, influencing fire ignition and spread.
  - Humidity (%): Represents atmospheric moisture, affecting fire persistence.
  - Wind Speed (km/h): Affects fire intensity and direction.
  - Rainfall (mm): Reduces fire risk by increasing moisture levels.
- **Fire Risk Indices:**
  - Fine Fuel Moisture Code (FFMC): Indicates the moisture content of fine fuels, affecting fire ignition probability.
  - Duff Moisture Code (DMC): Represents moisture levels in loosely compacted organic matter, influencing fire sustainability.
  - Drought Code (DC): Measures deep-layer moisture content, indicating long-term dryness.
  - Initial Spread Index (ISI): Estimates potential fire spread based on wind speed and FFMC values.
- **Temporal Features:**
  - Month & Day: Represent seasonal variations in wildfire occurrences.
- **Spatial Variables:**
  - X and Y Coordinates: Indicate the fire’s location within Montesinho Park.

- **Target Variable:**
  - Burned Area (ha): Represents the size of land affected by the wildfire. Due to its highly skewed distribution, a log transformation is applied to improve model performance.

  This dataset provides crucial information for predicting wildfire spread, aiding in risk assessment and fire management strategies.

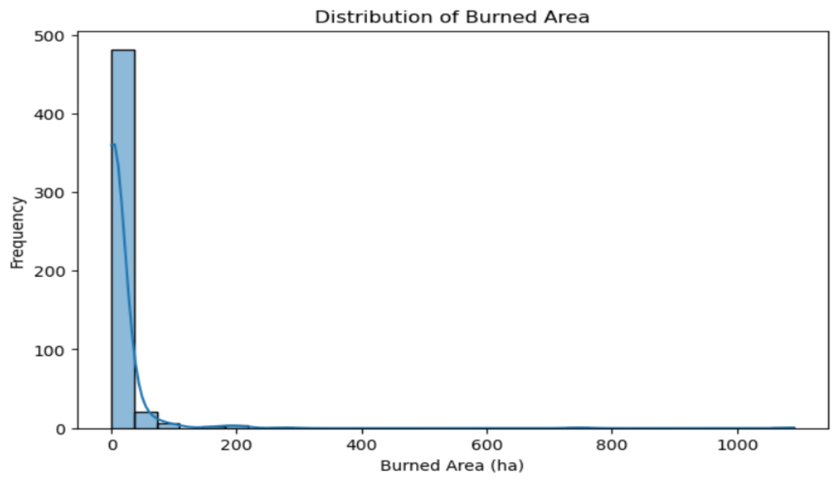
Data Preprocessing

The dataset was preprocessed to ensure data quality and improve model performance. The following steps were performed:

1. **Data Loading:** The dataset was imported into a Pandas DataFrame for analysis.
2. **Handling Missing Values:** The dataset was checked for missing values.
3. **Encoding Categorical Variables:** The categorical features (month and day) were converted into numerical format using one-hot encoding to ensure compatibility with machine learning models.
4. **Feature Scaling:** Numerical features such as temperature, humidity, wind speed, rainfall, and fire indices (FFMC, DMC, DC, ISI) were standardized using scaling techniques to improve model efficiency.
5. **Log Transformation of Target Variable:** The burned area variable was highly skewed, so a logarithmic transformation was applied to normalize the distribution and enhance prediction accuracy.

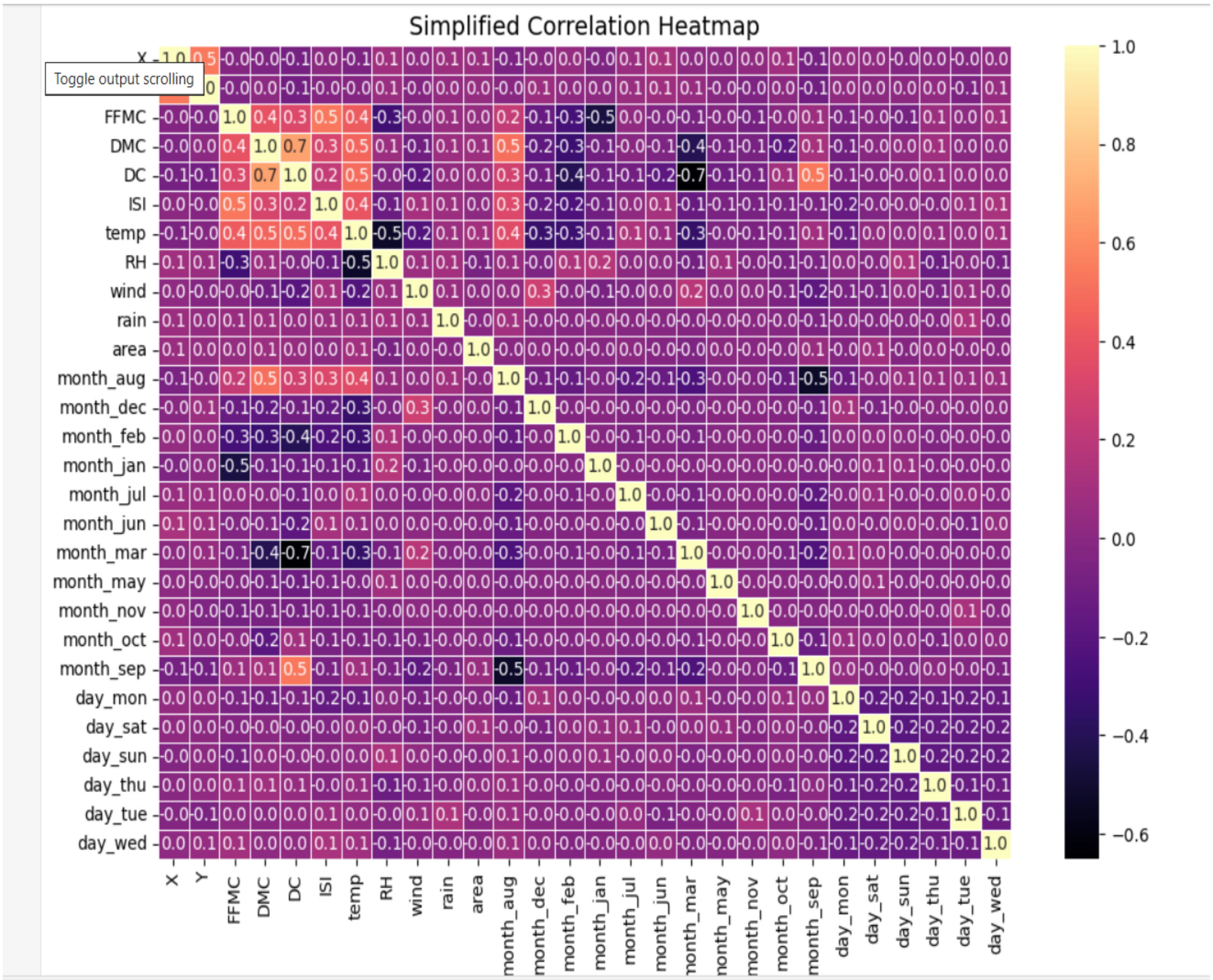
Exploratory Data Analysis

1. Distribution of Fire Area (Burned Area)



Insight: Most fires have small burned areas, but there are extreme outliers with large fires.

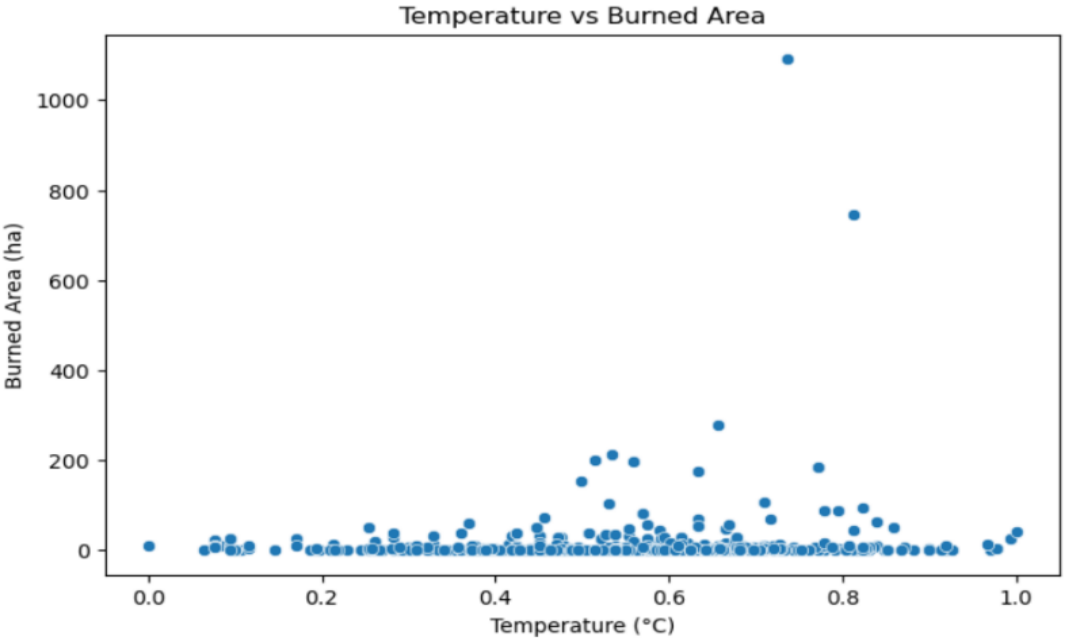
2. Correlation Heatmap



Insight: - Temp & RH have a strong inverse relationship.

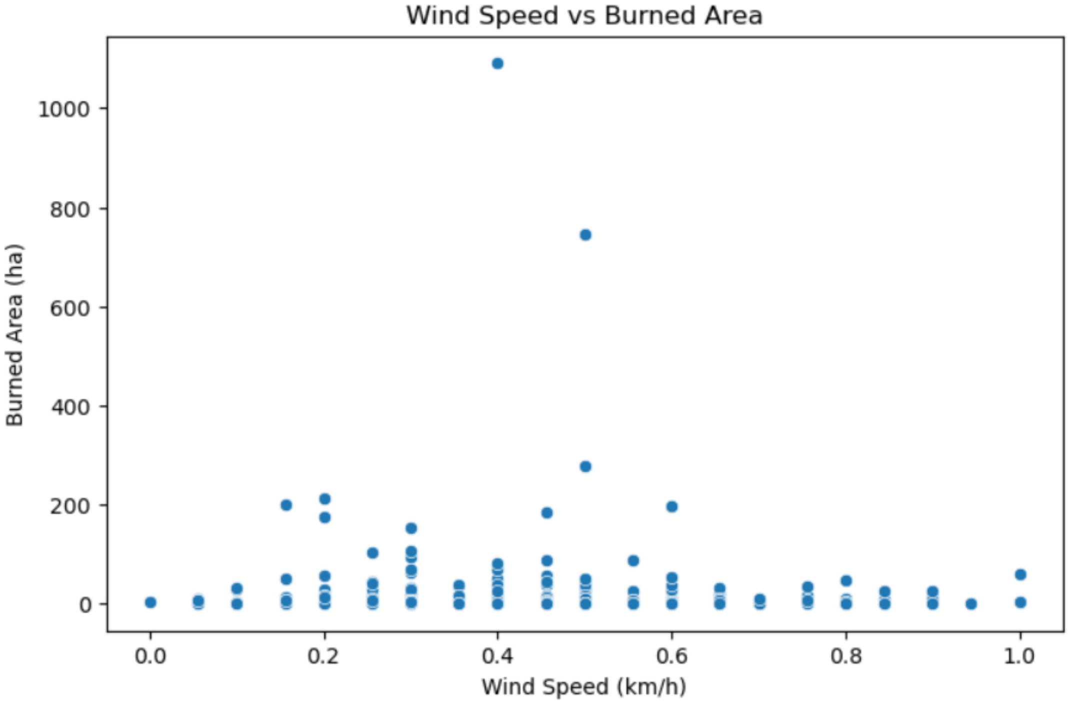
- Wind has a minor impact on area, but still relevant.
- Rain has almost no effect (since most values are zero).
- area has weak correlation with features suggesting non-linear relationships.

3. Scatter plot of Temperature vs Burned Area



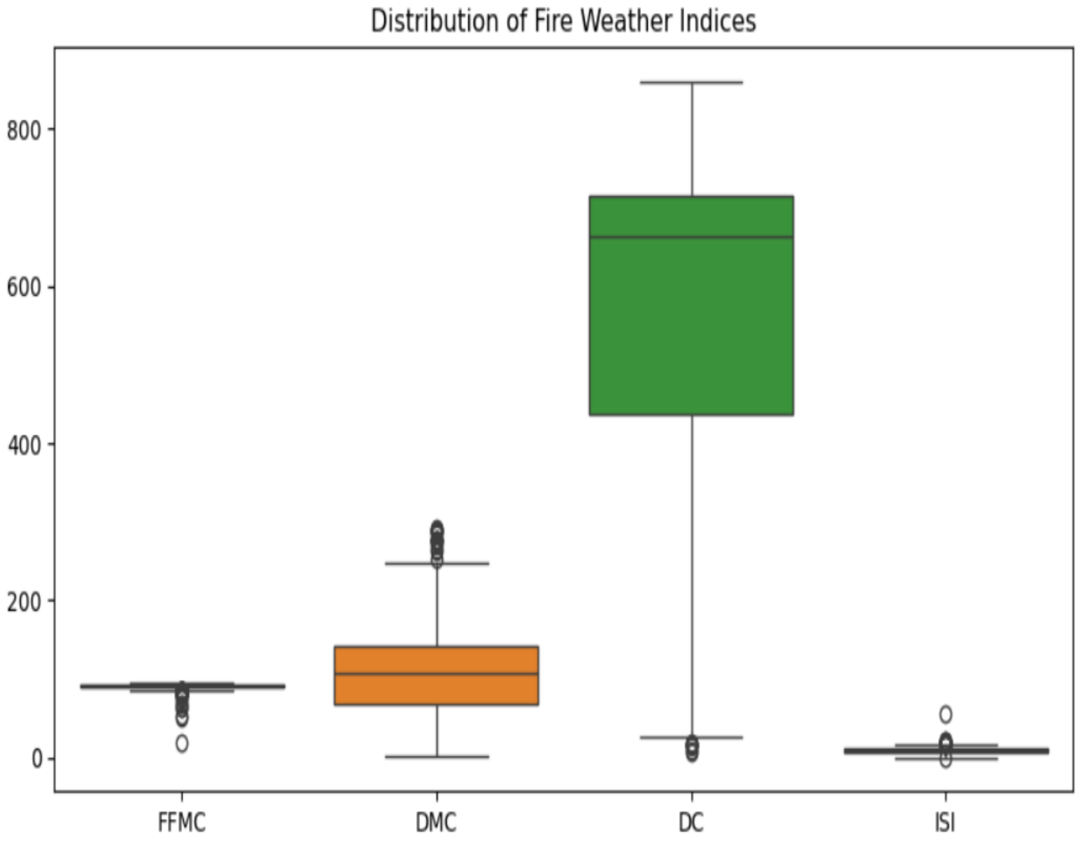
Insight: Higher temperatures may slightly increase the burned area.

4. Scatter plot of Wind vs Burned Area



Insight: Wind speed does not show a strong relationship with burned area.

5. Boxplot of Fire Weather Index (FWI) Components



Insight: - FFMC (Fine Fuel Moisture Code) is high for most cases which means Fires start easily.
- ISI (Initial Spread Index) has some extreme values, meaning rapid fire spread in some cases.

- ## Exploratory Data Analysis Findings

  - Most fires burn small areas, but some extreme cases exist.
  - Fires are most frequent in August & September.
  - Temperature & humidity play major roles in fire occurrence.
  - FWI components (FFMC, DMC, DC, ISI) are important for fire risk assessment.
  - Fire area does not strongly correlate with any single variable, suggesting complex interactions.

- ## ML Models

  The accuracy of the machine learning models was measured in terms of Root Mean Squared Error (RMSE). Lower RMSE reflects higher predictive accuracy. The RMSE for each model is as follows:

1. **Linear Regression (RMSE: 1.5171)**
   - Performed similarly to the Decision Tree model.
   - Unable to capture non-linear relationships in the data, leading to limited predictive power.
   - Despite its simplicity, it provided a decent baseline for comparison.
2. **Decision Tree Regressor (RMSE: 1.5171)**
   - Matched the performance of Linear Regression but with better adaptability to complex patterns.
   - However, prone to overfitting, which may reduce generalization to unseen data.
3. **Random Forest Regressor (RMSE: 1.5252)**
   - Expected to outperform Decision Trees due to ensemble learning, but the RMSE was slightly higher.
   - Provided more robust predictions compared to a single Decision Tree.
4. **XGBoost Regressor (RMSE: 1.5928)**
   - Surprisingly had the highest RMSE among all models.

- ## Findings

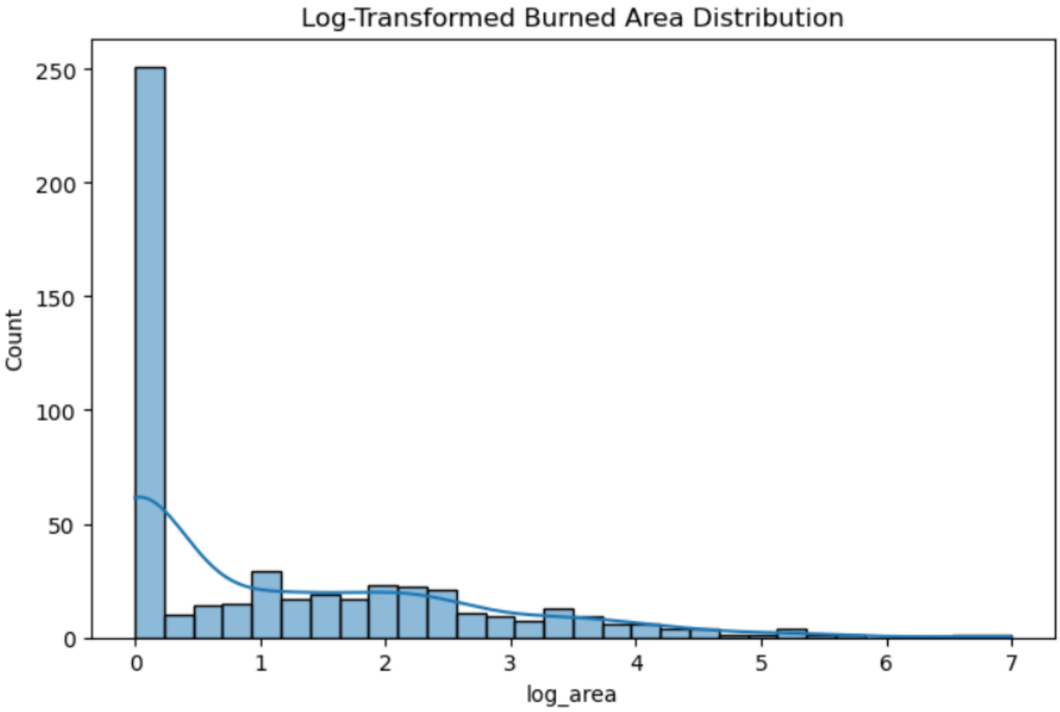
  - Linear Regression and Decision Tree had the same performance with the minimum RMSE (1.5171).
  - Random Forest recorded a marginally greater RMSE (1.5252), reflecting that it did not particularly outperform the more basic models.
  - XGBoost, being a sophisticated model, recorded the highest RMSE (1.5928), indicating that more tuning was needed for improved performance.

According to the RMSE scores, Linear Regression and Decision Tree models were best suited for the prediction of wildfire burned area, refuting the hypothesis that highly complex models would always provide improved results.

- ## Feature Engineering

  Feature engineering was done to improve the predictive capability of machine learning models by converting raw data into more useful representations. The most important steps are :

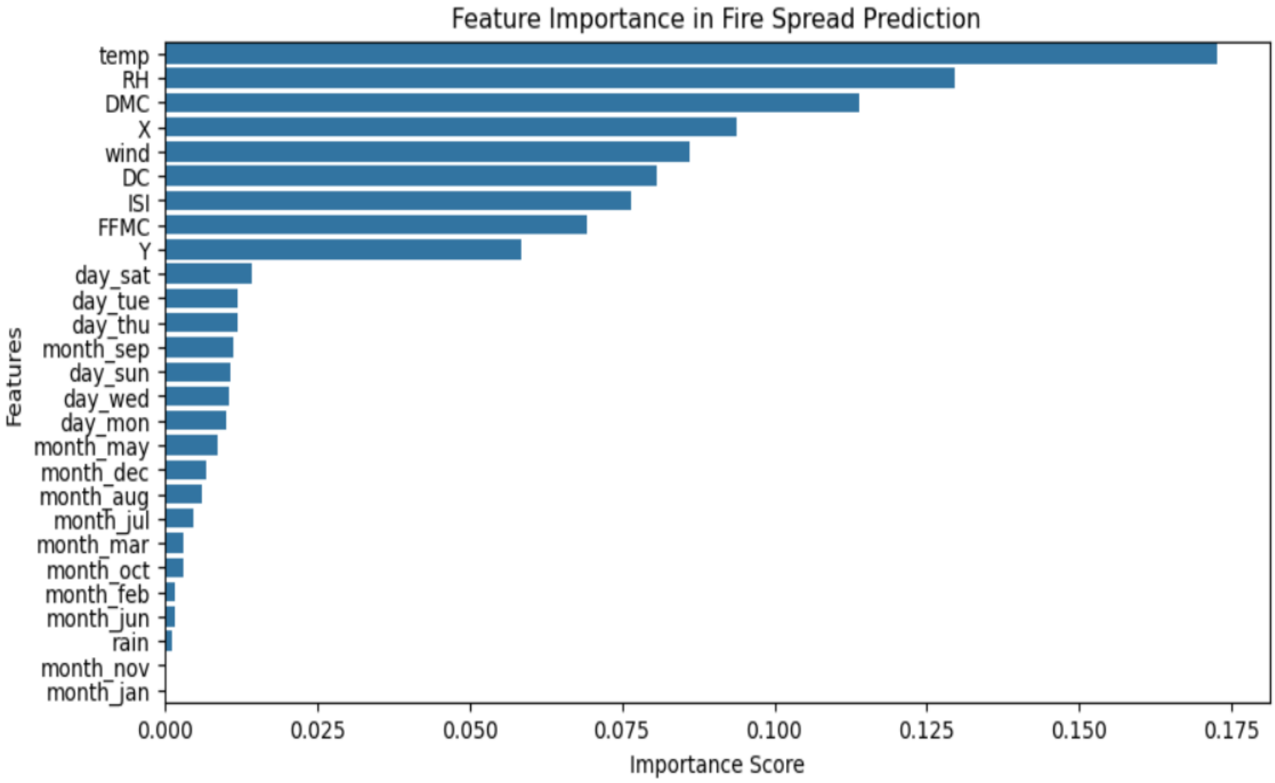
  1. Encoding Categorical Variables
     - Month and Day were encoded into numerical values using one-hot encoding to prepare them for machine learning models.
  2. Feature Scaling
     - Standardization was used for numerical attributes like temperature, humidity, wind speed, rainfall, and fire danger indices (FFMC, DMC, DC, ISI) through StandardScaler to enhance model convergence.
  3. Log Transformation of Target Variable
     - The variable burned area was very skew, and therefore a log transformation was used to normalise its distribution and dampen the effect of outliers.



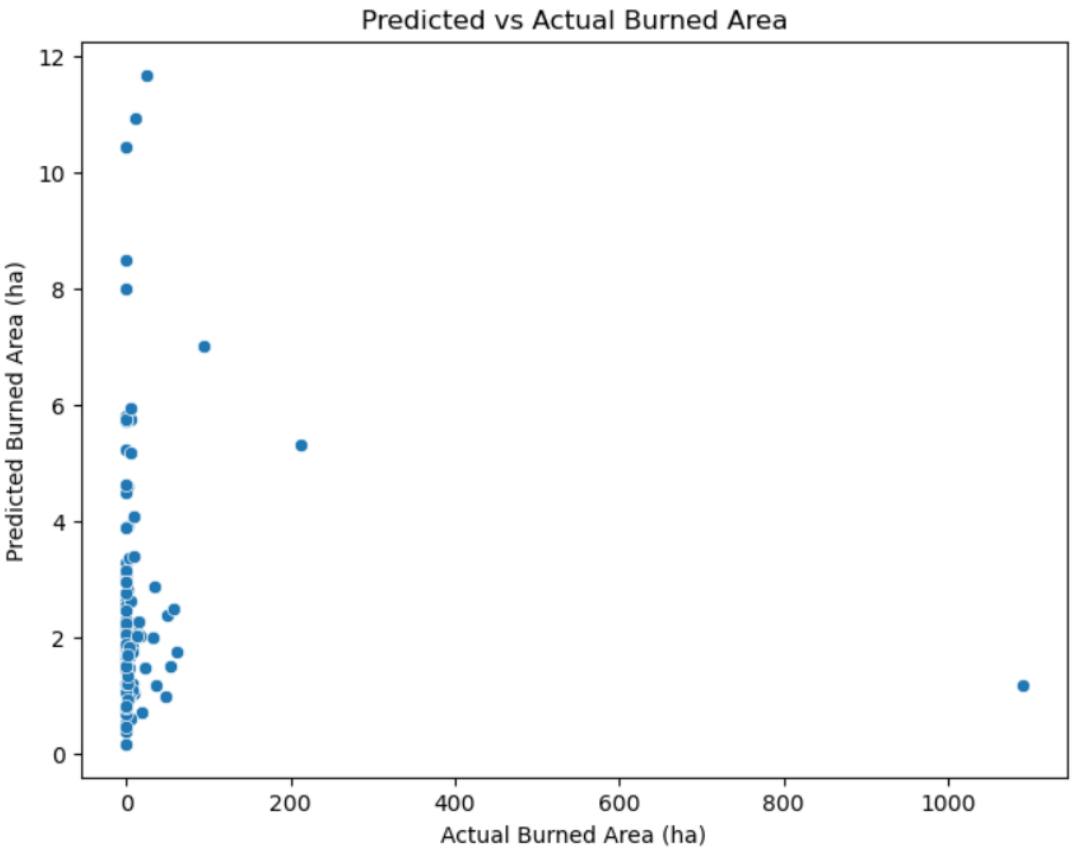
- **Insight :** - The original burned area data was highly right-skewed, indicating that most fires affected small areas, while a few caused significantly large burns.

  -After log transformation, the distribution became more normalized, making it more suitable for regression models.

4. Correlation Analysis for Feature Selection
   - Correlation matrix was employed to determine relationships among variables.
5. Spatial Feature Analysis
   - X and Y coordinates were investigated to gain insights into fire-risk areas, perhaps affecting risk prediction.
6. Feature Importance
   - Features like FFMC (Fine Fuel Moisture Code), DMC (Duff Moisture Code), and ISI (Initial Spread Index) had the highest importance, indicating their strong impact on wildfire spread.
   - These indices directly relate to fire ignition and spread conditions.



7. Predicted vs Actual Burned Area
   - Predictions were more accurate for small fires but had larger errors for extensive wildfires.
   - The model struggled with extreme values, indicating potential data imbalance.



- ## Classification Performance Analysis

  In addition to regression models, classification models were tested to determine if a wildfire would lead to a large burned area. The models were tested using accuracy, precision, recall, and F1-score to determine their performance. The following is the performance summary of each model:

  1. Linear Regression
     - **Accuracy:** 0.51
     - **Precision:** 0.51
     - **Recall:** 1.00
     - **F1-score:** 0.68

Insights:

- Achieved a perfect recall (1.00), which means it classified all actual wildfire events correctly.
- However, precision was low, indicating a high false positive rate.

2. Decision Tree Classifier
   - **Accuracy:** 0.51
   - **Precision:** 0.51
   - **Recall:** 1.00
   - **F1-score:** 0.68

Insights:

- Similar performance to Linear Regression with high recall but poor precision.

3. Random Forest Classifier
   - **Accuracy:** 0.51
   - **Precision:** 0.51
   - **Recall:** 1.00
   - **F1-score:** 0.68

Insights:

- Did not provide significant improvements over Decision Tree classification.
- High recall suggests good sensitivity to wildfire prediction, but precision issues remain.

4. XGBoost Classifier
   - **Accuracy:** 0.50
   - **Precision:** 0.50
   - **Recall:** 0.98
   - **F1-score:** 0.67

Insights:

- Slightly lower accuracy compared to other models.
- Recall of 0.98 indicates that it missed very few actual fire events.
- Precision was lower, meaning many false positives were classified as wildfire events.

5. Logistic Regression
   - **Accuracy:** 0.51
   - **Precision:** 0.52
   - **Recall:** 0.57
   - **F1-score:** 0.54

Insights:

- Achieved the best balance between precision (0.52) and recall (0.57).

- Did not over-predict wildfire events as much as other models.

- ## Results and Findings

1. **Data Insights**
    - The burned area distribution was highly skewed, with most fire incidents being small and a few causing extensive damage.
    - Fire risk indices (FFMC, DMC, DC, ISI) had the highest feature importance, confirming their strong influence on wildfire spread.
    - Temperature and wind speed were critical meteorological factors, while humidity and rainfall had lower predictive value.
    - Temporal (Month, Day) and spatial (X, Y coordinates) features had minimal impact compared to fire risk and meteorological factors.
2. **Regression Model Performance**
    - Linear Regression and Decision Tree models performed the best, challenging the assumption that complex models always yield better results.
    - Random Forest did not significantly outperform Decision Trees, despite being an ensemble model.
    - XGBoost, though powerful, had the highest RMSE, indicating potential overfitting or suboptimal hyperparameters.

- ## Conclusion

Wildfires constitute a major danger to the economy, environment, and human life, and their precise prediction becomes a necessity in disaster management. The project employed machine learning models to forecast forest fire burned area using meteorological and spatial features. The feature analysis found fire risk indices (FFMC, DMC, ISI), temperature, and wind speed to be the most significant predictors of wildfire spread, with temporal and spatial attributes having lesser influence. Among the regression models tested, Linear Regression and Decision Tree yielded the best performance (RMSE: 1.5171), contradicting the hypothesis that more complex models such as XGBoost would always be better. Log transformation of the area burned enhanced model performance by decreasing skewness and increasing interpretability. Yet, larger fire events continued to be challenging to predict with accuracy, indicating that more sophisticated modeling methods and real-time data incorporation are required. This project illustrates how machine learning can be an effective tool for wildfire forecasting but also points to areas of improvement, including inputting real-time weather conditions, improving model parameters, and investigating deep learning methods.