Forest Fires

By Madhuri, Will, Jinnie, Emir, Tianze
Mentor: Ethan

What is a forest fire?

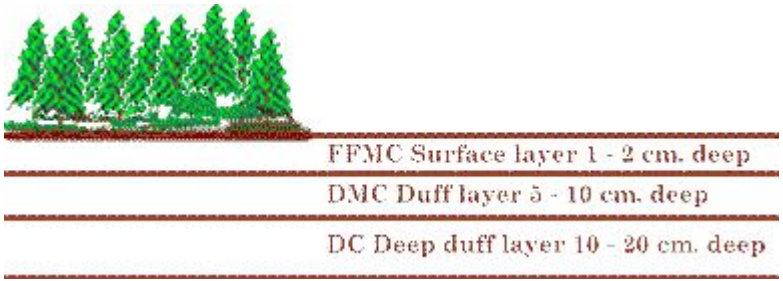
- Also known as wildfires or bush fires.
- Uncontrolled, unplanned and unpredictable fire in an area of combustible vegetation.
- Often caused by lightning
- Impacts on ecosystems and human communities.
- Common in dry arid areas with high temperatures.

Background

- Why we chose this project?
  - We live in California.
  - Good modeling project.
- Why is it important?
  - Wildfires have significant economic, environmental, and social impacts.
  - Better risk assessment models can improve resource allocation, public awareness, and decision-making.
  - Climate change

Data Understanding

- **Dataset:** Forest Fires Dataset: the burned area of forest fires, in the northeast region of Portugal
- **Why we choose this dataset:** We chose this particular forest fire dataset because it includes all the essential variables that are critical for accurately predicting the risk of forest fire.



Features information:

- **X** - x coordinate within the Montesinho park map
- **Y** - y coordinate within the Montesinho park map
- **month** - "jan" to "dec"
- **day** - "mon" to "sun"
- **temperature** - in °C
- **Relative Humidity** - in %
- **wind speed** - in km/h
- **rain** - mm/m2
- **area** - burned area (in ha)

Fire Weather Indices: numbers that tell us how likely a wildfire is to start and spread based on weather and fuel conditions

- **FFMC (The Fine Fuel Moisture Code)** - moisture content of litter and other fine fuels that are on or near the surface of the ground.
- **DMC (The Duff Moisture Code)** - average moisture content of organic material such as dead leaves or twigs in moderate layer of the soil.
- **DC** - (**The Drought Code)** moisture content of deeper organic layers in the soil.
- **ISI** (**The Initial Spread Index)** - a numerical rating of the expected rate of fire spread.

P. Cortez and A. Morais. A Data Mining Approach to Predict Forest Fires using Meteorological Data In J. Neves, M. F. Santos and J. Machado Eds., New Trends in Artificial Intelligence, Proceedings of the 13th EPIA 2007 - Portuguese Conference on Artificial Intelligence, December, Guimaraes, Portugal, pp. 512-523, 2007. APPIA, ISBN-13 978-989-95618-0-9. Available at: http://www.dsi.uminho.pt/~pcortez/fires.pdf

Transition from Linear to Logistic regression

- Challenges with linear regression:
  - Predicting area could result in very small values (e.g., 0.5 hectares), which may not be meaningful for fire risk assessment.
- Moving from predicting area to high/low risk:
  - Logistic regression enables us to classify fires as high or low risk based on a chosen threshold, providing a more actionable and interpretable output.
- High risk classification criteria:
  - Classified as high risk if the area affected is greater than 10 hectares, allowing us to focus on the most impactful fires.

```
train, test = train_test_split(forestfires_encoded, test_size=0.2, random_state=42)

# Split the data into features and labels
X_train = train[['season_summer', 'season_fall', 'season_winter', 'season_spring', 'FFMC_sd', 'DC_sd', 'temp_sd
Y_train = train['high risk']

X_test = test[['season_summer', 'season_fall', 'season_winter', 'season_spring', 'FFMC_sd','DC_sd', 'temp_sd',
Y_test = test['high risk']

lr = LogisticRegression(fit_intercept=True, solver='lbfgs')
lr.fit(X_train, Y_train)
```

```
▸ LogisticRegression
```

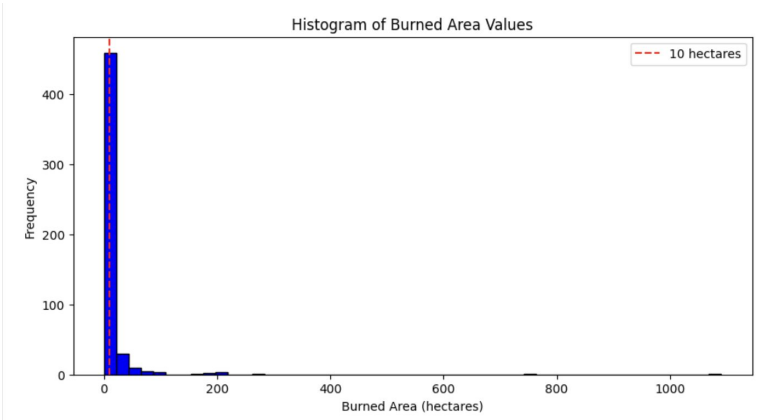
Determining the 10 Hectares Threshold

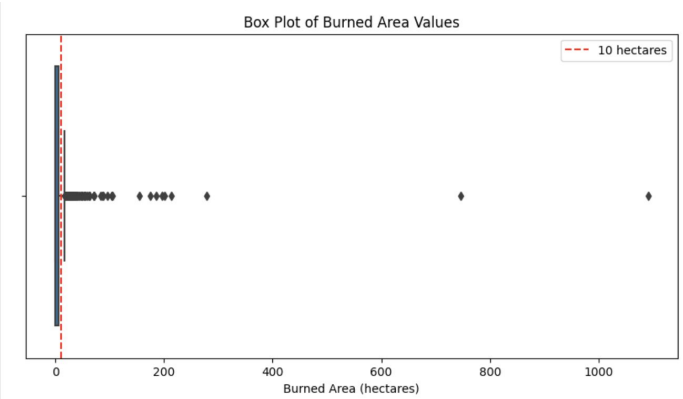
1. Online Research:
   a. Moderate/High risk forest fire
   b. Sources varied, but 10 hectares identified as a reasonable value
2. Data Analysis:
   a. 10 hectares in the 81.92th percentile of burned area
   b. Provides a clear distinction between high and low risk

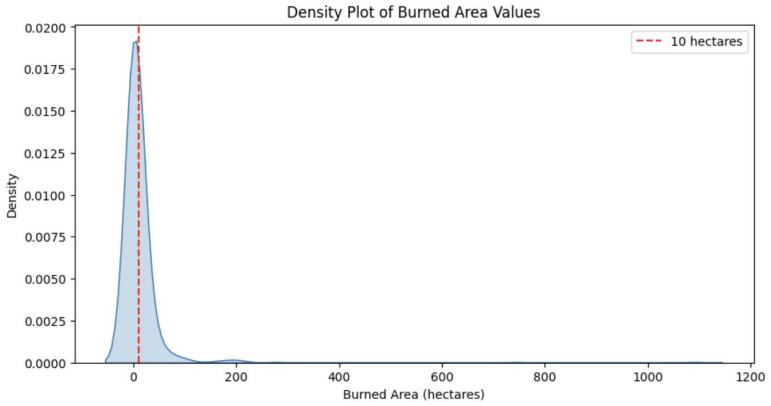
Conclusion: 10 hectares chosen as the threshold for high vs low risk forest fires

```
25th percentile: 0.0
50th percentile: 0.52
75th percentile: 6.57
90th percentile: 25.262000000000043
95th percentile: 48.713999999999984

The value 10 is at the 81.92th percentile.
```

10 Hectares Threshold Visualized

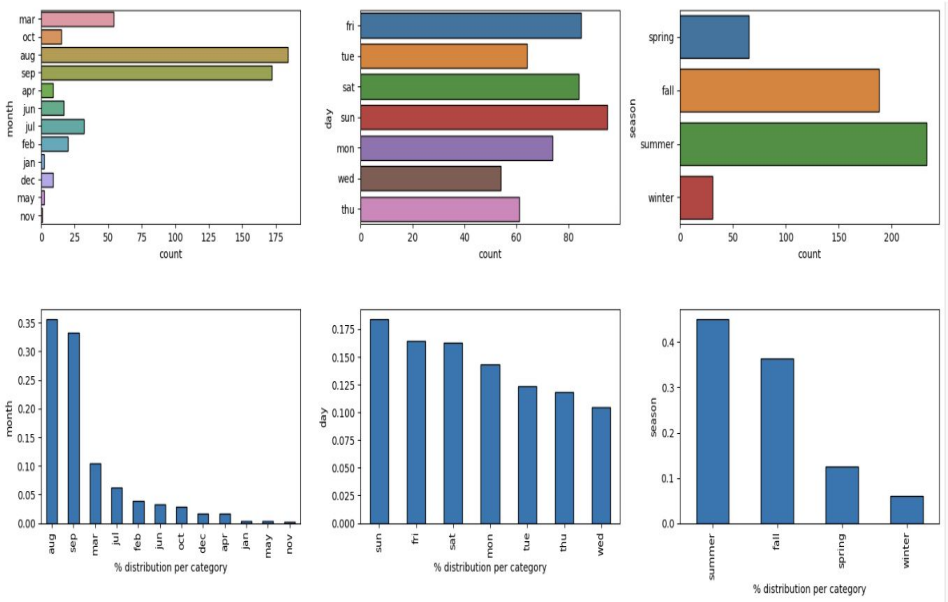






Modeling

- **Model**: Binary Logistic Regression (1 = high risk, 0 = low risk)
- Issues faced:
  a. Different unit features / Categorical variables
  b. Feature selection (avoiding multicollinearity)
  c. Class imbalance



Modeling Issues Solution (a)

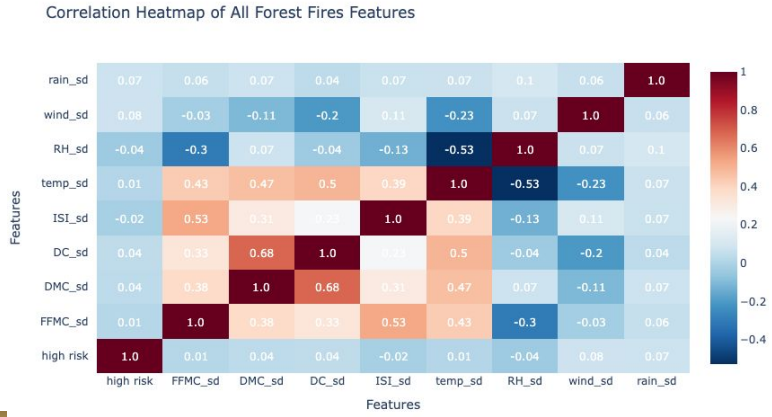
- Used **helper functions** get_risk to identify the risk of burned area (1's and 0's) and get_season to create a season column.
- **One-hot encoded** the season column, dropped other categorical variables since ‘days’ column didn't show any noticeable distribution, dropped the X and Y (coordinates) columns, and simplified months to seasons.
- **Standardize**d all numerical features because they were in different units, ensuring that all features are on a similar scale.
- **Standardization** is the process of transforming a variable to have a mean of 0 and a standard deviation of 1. This is achieved by subtracting the mean of the variable and dividing the result by the standard deviation.

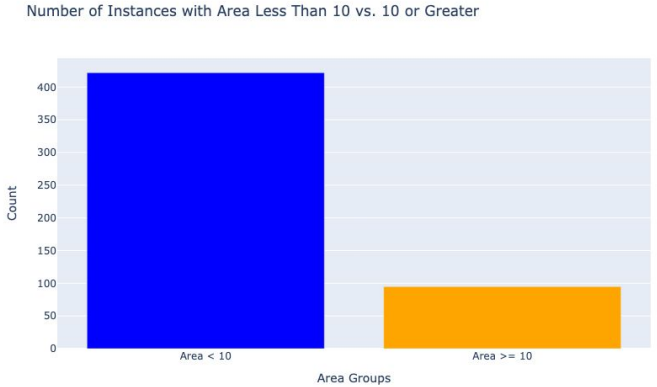
```
forestfires_df['high risk']  = forestfires_df['area'].apply(get_risk)
forestfires_df['season'] = forestfires_df['month'].apply(get_season)

one_hot_encoded = pd.get_dummies(forestfires_df['season'], prefix='season')
forestfires_encoded = pd.concat([forestfires_df, one_hot_encoded], axis=1)
forestfires_encoded.drop('season', axis=1, inplace=True)

standardize(forestfires_encoded, ['FFMC', 'DMC', 'DC', 'ISI', 'temp', 'RH', 'wind', 'rain', 'area'])
forestfires_encoded
```

```
def standardize(df, lst):
    for var in lst:
        df[var+'_sd'] = (df[var] - np.mean(df[var])) / np.std(df[var])
```

Modeling Issues Solutions (b and c)

- To avoid **multicollinearity**, we removed features that had a correlation of 0.5 or higher between them.
- Given the **class imbalance** in the dataset, we decided to lower the **probability threshold** so our model would prioritize recall over precision. This helps minimize the number of **false negatives** (i.e., model predicting fires as low risk when they are actually high risk).

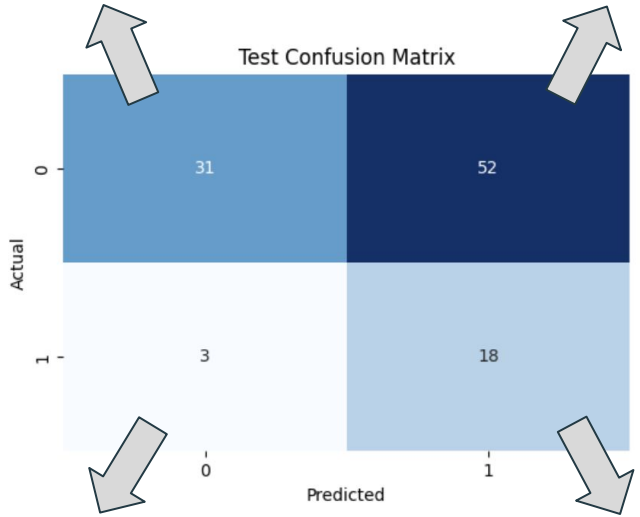




Evaluation

True Negative

False Positive



False Negative

True Positive

0 - low risk
1 - high risk

trade-off

```
Recall = TP / (TP + FN)
Recall
```

```
0.8571428571428571
```

```
Precision = TP / (TP + FP)
Precision
```

```
0.25714285714285712
```

Conclusion

- Hyperparameters:
  - Classified as high risk if area burned was above 10 hectares.
    - An alternative approach: classify based on the presence of fire (any fire vs. no fire) to avoid class imbalance.
  - Threshold 0.15:
    - Prioritized recall over precision: "Better safe than sorry."
- Limitations:
  - Model built for fires in the National Forest in Portugal, not directly applicable to other locations like California.
  - Risk levels may vary depending on location (e.g., a small fire in a densely populated area might cause more damage).