California Fire Predictor Model

Purpose / Objective

- Build a machine learning model that predicts whether a **wildfire will start on a given day**, using historical weather conditions.
- Understand which environmental factors (temperature, wind, precipitation, seasonality, lagged conditions) influence wildfire risk the most.
- Use **time-aware evaluation** to avoid data leakage and simulate real-world forecasting.
- Provide interpretable visualizations (correlation heatmaps & feature importance)

Dataset

- Gathered from NOAA Climate Data Online with fire incident data from CAL FIRE.

Each row includes:

- Precipitation, Max / Min / Avg Temperature, Avg Wind Speed, Season, Lagged weather metrics (previous day's wind & precipitation), Derived features (temperature range, wind–temperature ratio), **fire_start_day** (True = fire occurred)

After cleaning: **~15,000 usable rows**
Class distribution:

- **No Fire:** 10005
- **Fire:** 4971

CA_Weather_Fire_Dataset_1984-2025_new.csv > data

```
date,precipitation,max_temp,min_temp,avg_wind_speed,fire_start_day,year,temp_range,wind_temp_ratio,month,season,lagged_precipitat
1984-01-07,0.0,59.0,54.0,5.82,False,1984,5.0,0.0986440677966101,1,Winter,0.0,5.528571428571429,7,66.0,6.113333333333333,0.0,7.0,3
1984-01-08,0.0,59.0,55.0,3.36,False,1984,4.0,0.0569491525423728,1,Winter,0.0,5.337142857142857,8,61.0,5.52,0.0,7.0,198.2399999999
1984-01-09,0.0,61.0,54.0,6.71,False,1984,7.0,0.11,1,Winter,0.0,5.497142857142857,9,59.666666666666664,5.296666666666667,0.0,7.0,4
1984-01-10,0.0,70.0,47.0,4.7,False,1984,23.0,0.0671428571428571,1,Winter,0.0,5.40142857142857,10,63.333333333333336,4.9233333333
1984-01-11,0.0,68.0,46.0,5.82,False,1984,22.0,0.0855882352941176,1,Winter,0.0,5.561428571428571,11,66.33333333333333,5.7433333333
1984-01-12,0.0,69.0,47.0,5.14,False,1984,22.0,0.0744927536231884,1,Winter,0.0,5.561428571428571,12,69.0,5.22,0.0,7.0,354.65999999
1984-01-13,0.0,62.0,48.0,6.93,False,1984,14.0,0.1117741935483871,1,Winter,0.0,5.497142857142857,13,66.33333333333333,5.9633333333
1984-01-14,0.0,59.0,48.0,6.04,False,1984,11.0,0.1023728813559322,1,Winter,0.0,5.528571428571429,14,63.333333333333336,6.036666666
1984-01-15,0.0,59.0,43.0,6.71,False,1984,16.0,0.113728813559322,1,Winter,0.0,6.007142857142857,15,60.0,6.56,0.0,7.0,395.89
1984-01-16,0.39,55.0,45.0,6.71,False,1984,10.0,0.122,1,Winter,0.39,6.007142857142857,16,57.666666666666664,6.486666666666667,0.39
1984-01-17,0.0,63.0,41.0,5.59,False,1984,22.0,0.0887301587301587,1,Winter,0.39,6.134285714285714,17,59.0,6.336666666666667,0.39,6
1984-01-18,0.0,61.0,44.0,5.59,False,1984,17.0,0.091639344262295,1,Winter,0.39,6.1014285714285705,18,59.666666666666664,5.9633333
1984-01-19,0.0,60.0,47.0,4.92,False,1984,13.0,0.082,1,Winter,0.39,6.069999999999999,19,61.333333333333336,5.366666666666666,0.39,
1984-01-20,0.0,70.0,44.0,5.59,False,1984,26.0,0.0798571428571428,1,Winter,0.39,5.878571428571429,20,63.666666666666664,5.3666666
1984-01-21,0.0,64.0,53.0,5.82,False,1984,11.0,0.0909375,1,Winter,0.39,5.847142857142857,21,64.66666666666667,5.4433333333333325,0
1984-01-22,0.0,62.0,45.0,6.71,False,1984,17.0,0.1082258064516129,1,Winter,0.39,5.847142857142856,22,65.33333333333333,6.04,0.39,6
1984-01-23,0.0,66.0,45.0,5.59,False,1984,21.0,0.0846969696969697,1,Winter,0.0,5.687142857142858,23,64.0,6.04,0.0,7.0,368.94
1984-01-24,0.0,70.0,48.0,4.25,False,1984,22.0,0.0607142857142857,1,Winter,0.0,5.4957142857142856,24,66.0,5.516666666666667,0.0,7.
1984-01-25,0.0,74.0,51.0,6.04,False,1984,23.0,0.0816216216216216,1,Winter,0.0,5.5600000000000005,25,70.0,5.293333333333333,0.0,7.
1984-01-26,0.0,73.0,48.0,13.65,False,1984,25.0,0.186986301369863,1,Winter,0.0,6.807142857142858,26,72.33333333333333,7.98,0.0,7.0
1984-01-27,0.0,79.0,57.0,8.5,True,1984,22.0,0.1075949367088607,1,Winter,0.0,7.222857142857143,27,75.33333333333333,9.396666666666
1984-01-28,0.0,76.0,50.0,5.59,False,1984,26.0,0.0735526315789473,1,Winter,0.0,7.19,28,76.0,9.246666666666668,0.0,7.0,424.84
1984-01-29,0.0,78.0,51.0,5.59,False,1984,27.0,0.0716666666666666,1,Winter,0.0,7.030000000000001,29,77.66666666666667,6.56,0.0,7.0
1984-01-30,0.0,78.0,54.0,5.82,False,1984,24.0,0.0746153846153846,1,Winter,0.0,7.062857142857142,30,77.33333333333333,5.666666666666
1984-01-31,0.0,73.0,54.0,5.14,False,1984,19.0,0.0704109589041095,1,Winter,0.0,7.19,31,76.33333333333333,5.516666666666667,0.0,7.0
1984-02-01,0.0,62.0,53.0,7.61,False,1984,9.0,0.1227419354838709,2,Winter,0.0,7.4142857142857155,32,71.0,6.19,0.0,7.0,471.82
1984-02-02,0.0,62.0,51.0,6.71,False,1984,11.0,0.1082258064516129,2,Winter,0.0,6.422857142857142,33,65.66666666666667,6.486666666666
1984-02-03,0.0,72.0,52.0,5.82,False,1984,20.0,0.0808333333333333,2,Winter,0.0,6.039999999999999,34,65.33333333333333,6.7133333333
1984-02-04,0.0,67.0,53.0,6.49,False,1984,14.0,0.096865671641791,2,Winter,0.0,6.168571428571428,35,67.0,6.34,0.0,7.0,434.830000000
1984-02-05,0.0,65.0,48.0,7.16,False,1984,17.0,0.1101538461538461,2,Winter,0.0,6.392857142857144,36,68.0,6.489999999999999,0.0,7.0
1984-02-06,0.0,73.0,49.0,5.37,False,1984,24.0,0.0735616438356164,2,Winter,0.0,6.328571428571429,37,68.33333333333333,6.34,0.0,7.0
1984-02-07,0.0,75.0,49.0,5.82,False,1984,26.0,0.0776,2,Winter,0.0,6.425714285714286,38,71.0,6.116666666666667,0.0,7.0,436.5
1984-02-08,0.0,68.0,48.0,5.37,False,1984,20.0,0.0789705882352941,2,Winter,0.0,6.105714285714286,39,72.0,5.5200000000000005,0.0,7.
1984-02-09,0.0,62.0,52.0,6.71,False,1984,10.0,0.1082258064516129,2,Winter,0.0,6.105714285714286,40,68.33333333333333,5.966666666666
```

Methodology

Overview

- Encode FIRE-START-DAY as binary target (1 = risk of fire, 0 = no risk of fire)
- Perform a 90/10 train-test split while maintaining class balance (risk of fires = rare)

Models Evaluated

- Decision Tree – non-linear patterns, interpretable structure
- Random Forest – robust non-linear modeling, provides feature importance

Analysis Techniques

- Feature importance – identify key environmental drivers of fire occurrence
- Ablation Testing – assess the impact of removing specific variables

Preprocessing

One-Hot Season Encoding: Improved model understanding of seasonal fire patterns.

Data Cleaning: Removed corrupt rows to maintain consistent numeric input (missing values).

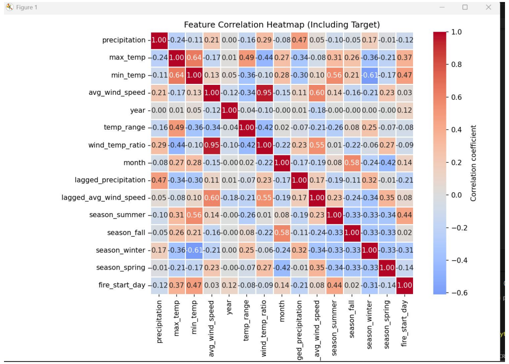
Feature Name Construction: Generated readable names (season_summer, season_fall, ...) for interpretability.

Correlation Analysis

Constructed heatmaps to identify which features correlate with fire occurrence.

Strongest correlations:

- High temperatures
- Low precipitation
- Seasonal patterns
- Lagged weather conditions



Preprocessing Continued

Feature Additions: Engineered several new weather-based features designed to capture patterns that the raw dataset could not represent.

- Features added: a 3 day rolling average of maximum temperature, 3 day rolling average of wind speed, a 7 day cumulative precipitation sum, number of dry days within the last 7 days (precipitation = 0), and the interaction of temperature and wind (max_temp * avg_wind_speed).
- Interaction between temperature and wind + max temperature proved to have a top 5 feature importance.
- 3 day average temperature also made it in the top 5 features, for feature importance.

Features Removed: Removed low-importance features in order to reduce noise in the data set and improve the model's ability to focus and find true patterns and variables that may influence fire risk.

- Trained the random forest model using all features, and after training we printed out the feature importance.
- Defined a cutoff, any feature with an importance less than .025 would be removed.
- The following values were removed: precipitation, lagged_precipitation, number of dry days.

Machine Learning

Class Balance:

- No fire risk (0) = 9999
- Fire risk (1) = 4971

Addressing the class imbalance:

- Weight for No: 0.7486
- Weight for Yes: 1.5057
- Improves recall for fire days, which is the most important objective
- We calculated class weights using scikit-learn tools, which assign higher weight to the minority class (fire days) so the model does not ignore rare fire events.

Model Used

RandomForestClassifier

- 300 trees
- Handles nonlinear relationships
- Provides built-in feature importance
- Robust to noise and real-world variability

Evaluation Metrics

- Accuracy
- Precision, Recall, F1-Score

Post-Processing / Results

Model Performance:

- Accuracy: 71.2%
- Fire Risk Precision 0.5419
- Fire Risk Recall: 0.6360
- Fire Risk F1-Score: 0.5862

The model correctly identifies ~63% of real

fire-start days

Top 5 Important Features:

- Min_temp, temp_3day_avg, year, lagged_avg_wind_speed, wind_3day_avg

```
Number of instances with fire_start_day = 0 (No) and 1 = (Yes)
No:  9999
Yes:  4971
Class weights:
Weight for No: 0.7486
Weight for Yes: 1.5057
Classification report:
              precision    recall  f1-score   support

           0     0.8141    0.7478    0.7795      1019
           1     0.5419    0.6360    0.5852       478

    accuracy                         0.7121      1497
   macro avg     0.6780    0.6919    0.6824      1497
weighted avg     0.7272    0.7121    0.7175      1497


Top 5 features by importance:
                feature  importance
               min_temp    0.137239
          temp_3day_avg    0.117377
                   year    0.084914
 lagged_avg_wind_speed    0.083779
          wind_3day_avg    0.075930
```

Takeaways / Future Improvements

- Wildfires are often anomalies
    - Had to implement class weights
    - SMOTE wouldn't be possible because these had chronological order
        - Can't synthesize a fire day on a non-fire day

- Weather variables are consistently **top predictors**
    - Things related to temperature, humidity, and wind were always on top

- Possible Additional Features
    - Adding human factors (population density, proximity to roads/power lines)
    - Vegetation moisture
    - Find actual areas of California instead of an overall estimate

- Pick a data set that has a defined class
    - The model only predicts if the day is at high risk for a fire
        - High Risk Fire Day *does not equal* an actual fire