Bhagya Ram
Katelyn Wong
Sebin Lee
Yiming Gao
Dr. Siddharth Vishwanath
Math 189
3/16/2025

# Analyzing Forest Fire Severity in California

## 1. Introduction

Wildfires pose a significant threat to ecosystems, economies, and human lives, especially in California, where climate conditions and human activity make wildfires a recurring disaster. In January 2025, major wildfires swept through Los Angeles, destroying tens of thousands of acres and causing economic losses estimated between $250 billion and $275 billion [1]. Given the severe impact of such events, understanding the factors that influence fire severity is critical for prevention and mitigation strategies. This study aims to analyze historical wildfire and weather data to investigate which combination of meteorological factors (temperature, pressure, humidity, wind speed, wind direction) most strongly predict wildfire severity in the United States as measured by acres burned between 2012-2015. Our hope is that this will provide insights to help predict future fire risk.

## 2. Data Sources:

We utilized two primary datasets:

1. Kaggle: Historical Hourly Weather Data 2012-2017
   https://www.kaggle.com/datasets/selfishgene/historical-hourly-weather-data?select=wind_speed.csv.
2. Kaggle: 1.88 Million US Wildfires
   https://www.kaggle.com/datasets/rtatman/188-million-us-wildfires/data.

## 3. Data Description and Processing

1. Weather Data:
   While we discovered this dataset on Kaggle, the original data was acquired using the Weather API on the OpenWeatherMap website. The weather dataset contains seven individual datasets that each represent an important meteorological attribute: city attributes, humidity, pressure, temperature, weather description, wind direction and wind speed. The city_attributes.csv file contains latitude and longitude information on 36 different cities across the U.S., Canada and Israel.

| | City | Country | Latitude | Longitude |
|---|---|---|---|---|
| 0 | Vancouver | Canada | 49.249660 | -123.119339 |
| 1 | Portland | United States | 45.523449 | -122.676208 |
| 2 | San Francisco | United States | 37.774929 | -122.419418 |
| 3 | Seattle | United States | 47.606209 | -122.332069 |
| 4 | Los Angeles | United States | 34.052231 | -118.243683 |

Fig. 1: Snapshot of city_attributes.csv

The remaining 6 datasets each contain five years' of hourly measurements, each with 45252 observations and 37 variables. Of the 37 variables, one is the datetime measure while the other 36 correspond to the weather conditions in each city at a specific time.

| | datetime | Vancouver | Portland | San Francisco | Seattle | Los Angeles | San Diego | Las Vegas | Phoenix | Albuquerque | ... |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 2012-10-01 13:00:00 | 0.0 | 0.0 | 150.0 | 0.0 | 0.0 | 0.0 | 0.0 | 10.0 | 360.0 | ... |
| 2 | 2012-10-01 14:00:00 | 6.0 | 4.0 | 147.0 | 2.0 | 0.0 | 0.0 | 8.0 | 9.0 | 360.0 | ... |
| 3 | 2012-10-01 15:00:00 | 20.0 | 18.0 | 141.0 | 10.0 | 0.0 | 0.0 | 23.0 | 9.0 | 360.0 | ... |
| 4 | 2012-10-01 16:00:00 | 34.0 | 31.0 | 135.0 | 17.0 | 0.0 | 0.0 | 37.0 | 9.0 | 360.0 | ... |
| 5 | 2012-10-01 17:00:00 | 47.0 | 44.0 | 129.0 | 24.0 | 0.0 | 0.0 | 51.0 | 8.0 | 360.0 | ... |

Fig. 2: Snapshot of wind_direction.csv

- Wildfire data:

The wildfire dataset is a spatial SQLite database that contains 27 individual datasets. The data was originally generated to support the now retired Fire Program Analysis (FPA) program. The database contains 1.88 million records of wildfires that occurred in the United States from 1992 to 2015. Given how voluminous the database is, we are focusing on the dataset Fires, which contains information acquired from US federal, state and local reporting systems.

Our first dataset contains information across the U.S., Canada and Israel, while our second dataset is limited to the U.S. states. Further, our first dataset contains weather information collected from 2012 through 2017, while our second dataset includes data from 1992 to 2015. Considering the size of the Fires dataset, we began by randomly sampling a subset of 1800 rows from the data. We merged the two datasets based on shared time frame and geographic location.

- ❖ Data processing, merging and subsetting:

We have 2 separately generated datasets, so in order to find relationships between multiple variables (columns) across the datasets, we first need to merge the datasets based on certain variables. Specifically, in the weather dataset we have the weather conditions in certain cities across time, while in the fire dataset dataframe, we have fire re4cords at certain latitude and longitude coordinates across time. Based on the location and time similarity, we can merge the two datasets. So, we first extracted the cities in the weather dataset, and retrieved their latitude and longitude coordinates. Then for each fire (row) in the fire dataset subset, and for each city in the weather dataset, we calculate their great circle distance using the haversine distance method. We set the threshold to 100 miles to determine correlation between fire and city. If the distance is smaller than 100 miles, we add a row in the merged data frame with both the fire and the city weather condition at the same time. As the entire fire dataset has 1.8 million fire records (rows), during the pipeline construction we first used its subset to accelerate the running process. We randomly sampled 1800 rows from the fire dataset as a subset, and the run time of the merging process was 31.5 seconds on a MacBook computer.

- ❖ Correlation matrix (heatmap):

Among the many variables from the fire dataset, we selected FIRE_SIZE_CLASS for the prediction purpose. FIRE_SIZE_CLASS_encoded is transferred from standard FIRE_SIZE_CLASS definition, in which A is least severe fire, G is most severe fire. Note that the only fire sizes in the final dataset are A, B, C because of subsetting and merging dataframes. The weather condition variables we used in the correlation example are: Humidity, Wind_Direction, Temperature, Pressure, Wind_Speed.

## 4. Exploratory Data Analysis (EDA)

### 4.1) Missingness Analysis

We began by exploring missing values in our dataset.

| | Missing Values | Percentage (%) |
|---|---|---|
| ICS_209_INCIDENT_NUMBER | 1669 | 100.000000 |
| ICS_209_NAME | 1669 | 100.000000 |
| MTBS_ID | 1669 | 100.000000 |
| MTBS_FIRE_NAME | 1669 | 100.000000 |
| COMPLEX_NAME | 1669 | 100.000000 |
| FIRE_CODE | 1524 | 91.312163 |
| LOCAL_FIRE_REPORT_ID | 1475 | 88.376273 |
| CONT_TIME | 499 | 29.898143 |
| DISCOVERY_TIME | 372 | 22.288796 |
| FIRE_NAME | 350 | 20.970641 |
| CONT_DATE | 350 | 20.970641 |
| CONT_DOY | 350 | 20.970641 |
| COUNTY | 219 | 13.121630 |
| FIPS_CODE | 219 | 13.121630 |
| FIPS_NAME | 219 | 13.121630 |
| LOCAL_INCIDENT_ID | 72 | 4.313960 |

Fig. 3

As is evident in Figure 1, of the 53 columns in our merged dataset, 16 columns contain missing values. Upon further investigation, we realized that all 16 of these columns were of the object data type. Our dataset did not have any missing values in the key weather-related numeric columns that we planned on utilizing for analysis. We conducted a missingness analysis on the columns that did have missing values:

- We believe that the values in several of these columns, such as FIRE_CODE (91.3% is missing) and LOCAL_FIRE_REPORT_ID (88.3% of data missing) are missing at random (MAR). Data is said to be MAR when its missingness depends on other observed variables and not the data itself. Here, these values could be MAR because they might be related to the state or source reporting unit variables, since certain states or reporting units might be less likely to report this information.
- Values in a few columns such as FIRE_NAME are likely to be not missing at random (NMAR). Certain smaller fires that were extinguished almost immediately may have been recorded but not named. Data is considered NMAR when the cause for the missingness is to do with the data itself, and not any of the other observed variables.

4.2) Outlier detection

Next, we utilized Z-scores to check our dataset for outliers. We used a threshold of 3, such that values with z-scores greater than 3 or less than -3 were considered outliers. We did not have any negative values in our dataset, and only investigated the numeric columns for outliers.

```
Outliers in each column (Z-score > 3):
 Humidity                 0
Wind_Direction            0
Temperature               0
Pressure                 22
Wind_Speed               30
FIRE_SIZE_CLASS_encoded    0
dtype: int64
```

Fig. 4

As shown in Figure 2, most of the columns contained values that were reasonable. Pressure and Wind_Speed appeared to have 22 and 30 values respectively that were outliers. We visualized these using a boxplot to understand them better.

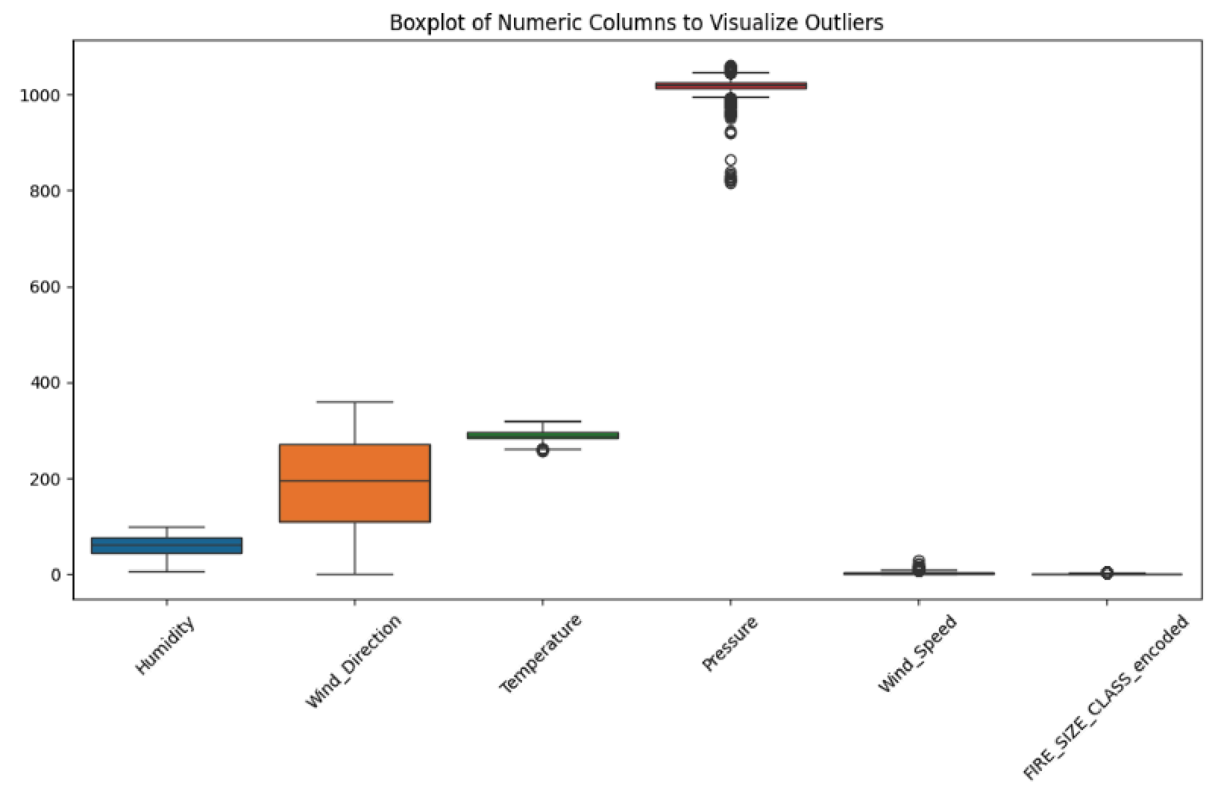


Fig. 5

Wind_Direction appears to have a wide range of values, evidenced by the long whiskers. Pressure, Wind_Speed and Temperature seem to contain outliers, given the individual circles on and outside of the whiskers. However, according to Temperature's z-score, all its values are in a reasonable range. It is understandable for weather attributes to have extreme values. They're an important part of the story that helps highlight the cause behind fires, so we are not eliminating the outliers.

4.3) Choropleth Map showing varying fire sizes across the country

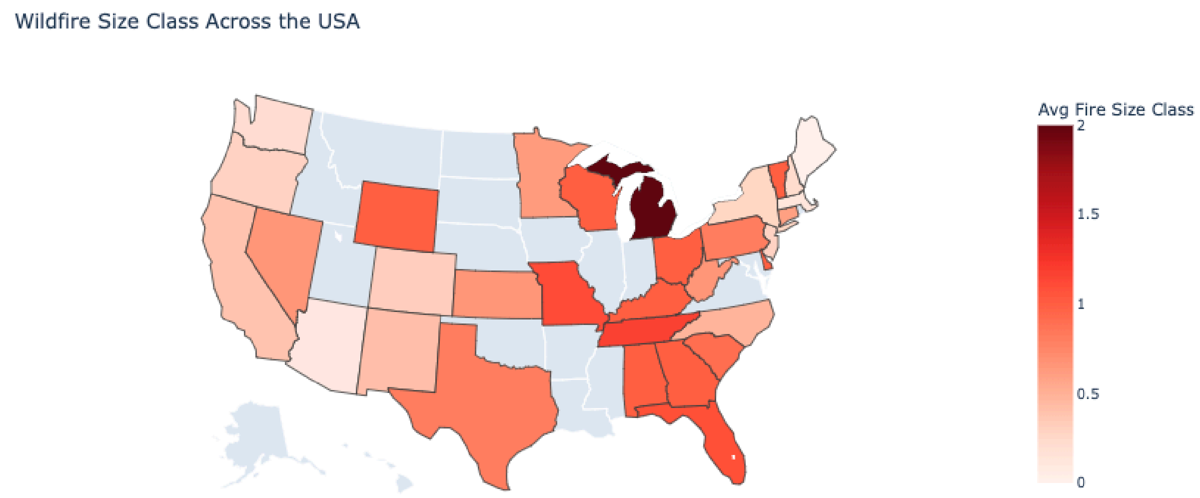


Fig. 6

After investigating the outliers, we were curious to see how the sizes of fires varied across the United States. The 'FIRE_SIZE_CLASS_encoded' column has 3 categories that fires are classified into depending on the size of the fires, with the 3 most common being:

- 0: (0, 0.25] acres i.e, fires that are greater than 0 but less than or equal to 0.25 acres
- 1: [0.26-9.9] acres
- 2: [10.0-99.9] acres

We built a choropleth map that shows the average size of the fires per state across the country using these 3 fire categories. There are 21 U.S. states represented in our dataset, and Figure 4 portrays the average fire sizes in those states. We chose a continuous color scale to emphasize how Florida, for instance, has larger fires on average than Colorado.

4.4) Bubble chart

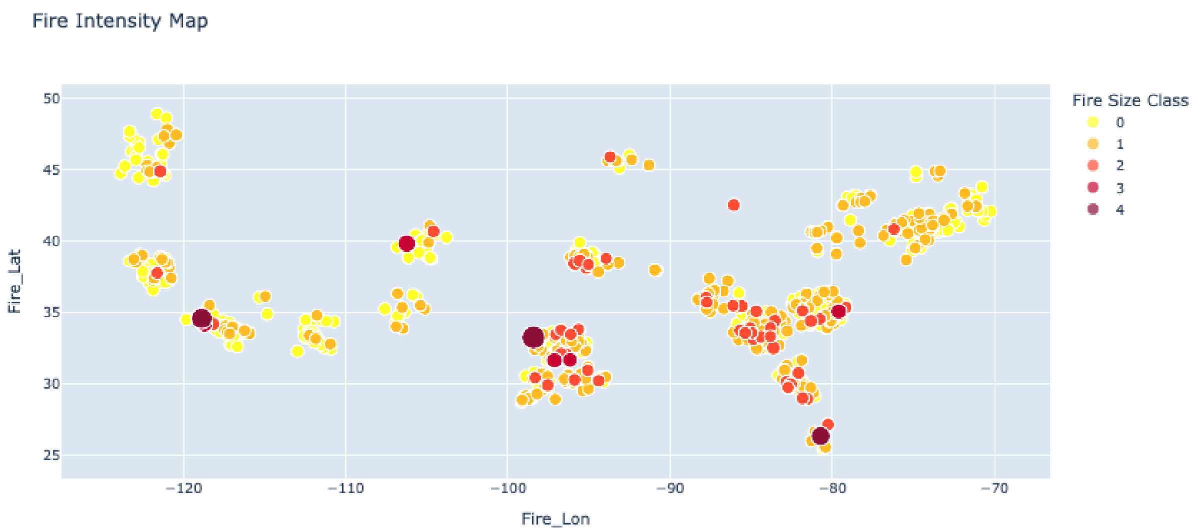

Fig. 7

Similar to the U.S. map in Figure 4, we wanted to create a map that would portray how the fire sizes vary by latitude and longitude. We ensured the color choices for the legend would align with reader expectations, and so we chose yellow, orange and red to portray the increasing fire sizes.

4.5) Pairplot, including scatterplot and Kernel Density Estimation (KDE) plots of key variables

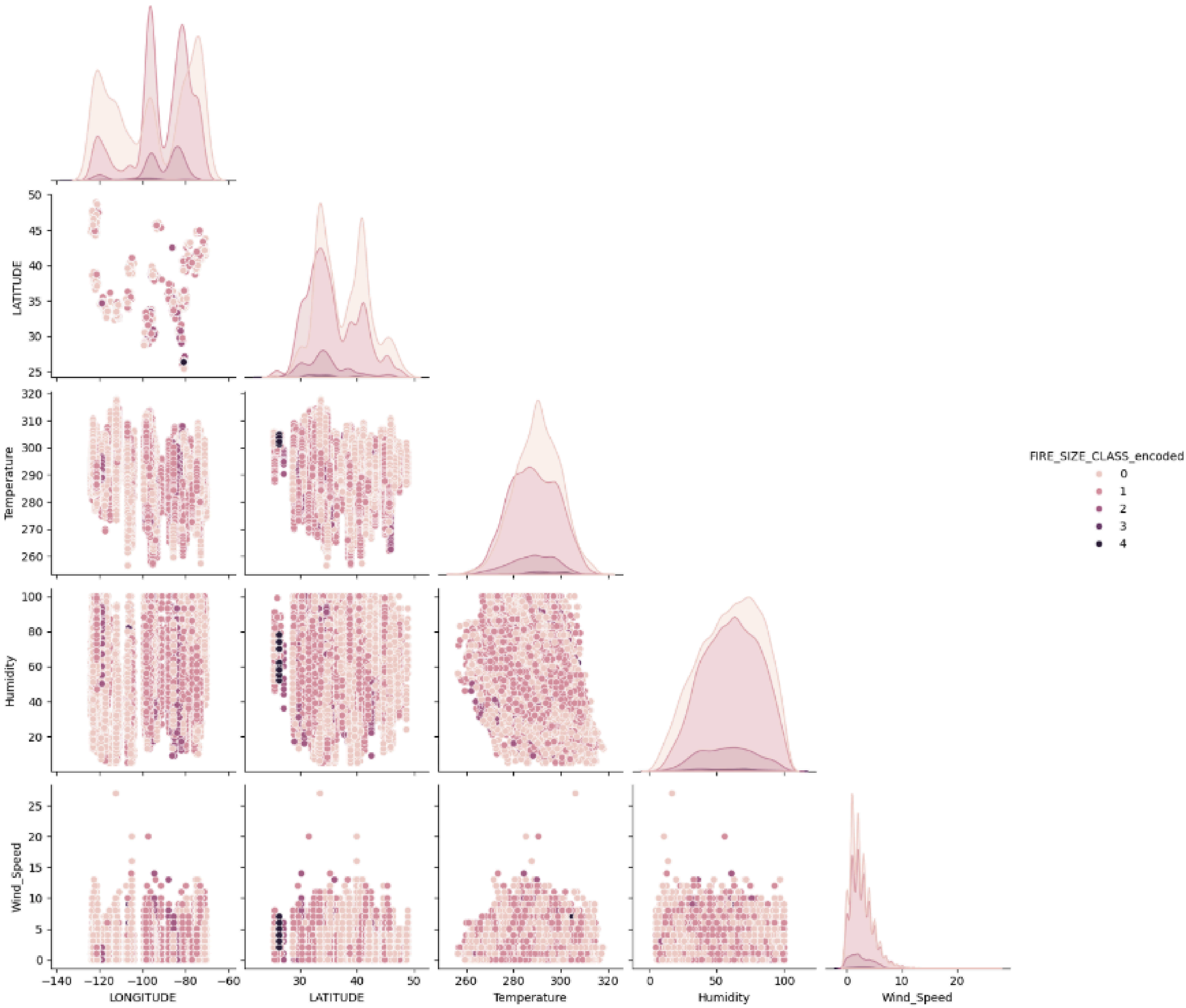


Fig. 8

This pairplot displays the relationship between key variables and our response variable, FIRE_SIZE_CLASS_encoded. We observe this relationship through a scatter plot, as well as a KDE plot. We observe from the Temperature plot that higher fire sizes i.e, the darker points are more concentrated in higher temperature ranges.

## 5. Statistical Analyses and Modeling

- Each weather parameter dataset was first reshaped using a melt operation to standardize the format. These reshaped datasets were then merged on 'Datetime' and 'City'. We computed a correlation matrix, which helped us identify significant relationships among the weather variables.

- ### 5.1) Logistic Model Building

In our initial prediction model, we randomly sampled from our dataset (n = 20,000) and employed an ordinal logistic regression model (Proportional Odds Model) to predict fire size, an ordinal variable with three levels, based on various weather and location features (all continuous variables). Specifically, we modeled fire size (FIRE_SIZE_CLASS_encoded) as a function of humidity, wind direction, wind speed, temperature, pressure, wind, latitude (City_Lat), and longitude (City_Lon).

Before constructing the regression model, we observed that certain predictor variables exhibited high Variance Inflation Factors (VIFs), indicating multicollinearity issues. To address this, we standardized these variables, which successfully mitigated the multicollinearity problem. However, we chose not to standardize latitude and longitude, despite their higher VIFs, because their interpretability relies on maintaining their original scale.

Instead of standardizing longitude, we opted to remove it from our feature set. This decision was based on the reasoning that latitude is generally more predictive of temperature, as cities closer to the equator tend to have higher temperatures on average. By retaining latitude and excluding longitude, we aimed to reduce redundancy while preserving meaningful geographic information.

After standardizing the remaining variables and removing longitude, the updated predictor variables have the following corresponding VIFs:

```
predictor variables:
                   Variable       VIF
0     Humidity_standardized  1.087349
1  Temperature_standardized  1.151861
2     Pressure_standardized  1.116165
3            Wind_Direction  4.562096
4                Wind_Speed  2.920621
5                  City_Lat  5.033826
```

Fig. 9

At this point, we ran the ordered logistic regression and obtained the following results:

```
Optimization terminated successfully.
         Current function value: 0.893174
         Iterations: 30
         Function evaluations: 34
         Gradient evaluations: 34
                             OrderedModel Results
==================================================================================
Dep. Variable:     FIRE_SIZE_CLASS_encoded   Log-Likelihood:            -17410.
Model:                        OrderedModel   AIC:                     3.484e+04
Method:                 Maximum Likelihood   BIC:                     3.492e+04
Date:                     Sat, 15 Mar 2025
Time:                             23:33:34
No. Observations:                    19492
Df Residuals:                        19482
Df Model:                                6
===========================================================================================
                             coef    std err          z      P>|z|      [0.025      0.975]
-------------------------------------------------------------------------------------------
Humidity_standardized     -0.0592      0.015     -3.960      0.000      -0.089      -0.030
Temperature_standardized  -0.2472      0.016    -15.559      0.000      -0.278      -0.216
Pressure_standardized      0.1881      0.017     10.944      0.000       0.154       0.222
Wind_Direction            -0.0007      0.000     -4.709      0.000      -0.001      -0.000
Wind_Speed                 0.0590      0.007      7.907      0.000       0.044       0.074
City_Lat                  -0.1216      0.003    -35.244      0.000      -0.128      -0.115
0.0/1.0                   -4.3367      0.126    -34.489      0.000      -4.583      -4.090
1.0/2.0                    0.9553      0.011     88.595      0.000       0.934       0.976
2.0/3.0                    0.7283      0.032     22.655      0.000       0.665       0.791
3.0/4.0                   -0.0127      0.093     -0.136      0.892      -0.195       0.169
===========================================================================================
```

Fig. 10

7. Interpretation of Model Results

1. **Lower humidity is associated with larger fire sizes**:

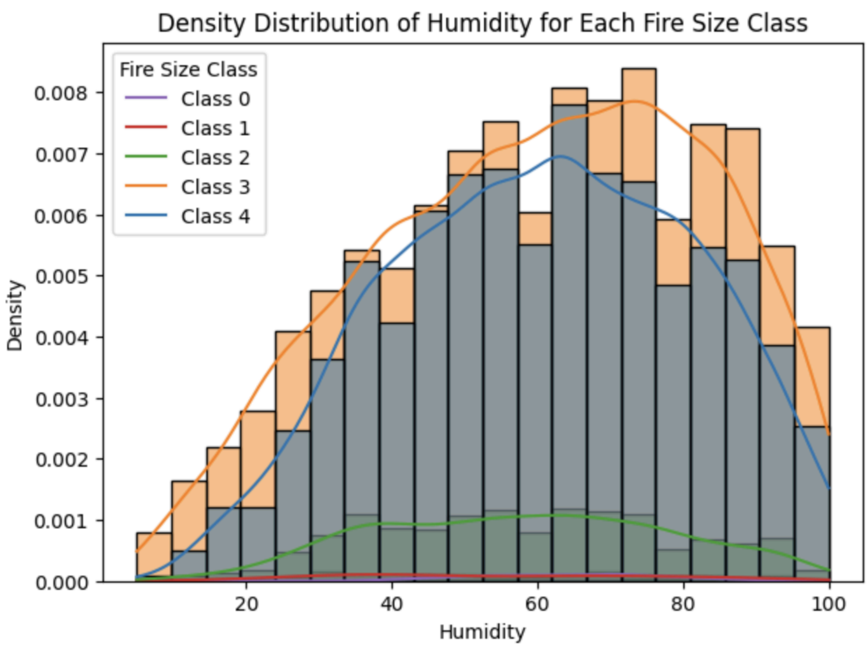


Fig. 11

Holding all other variables constant, a one standard deviation increase in humidity likely decreases the log odds of being in a larger fire size category by 0.0592 ($p < 0.001$ indicates a strong effect).

a. These results align with scientific observations as lower humidity levels indicate drier air, which can lead to drier environmental conditions that can lead to rapid fire spread.

2. **Lower temperature is associated with larger fire sizes:**

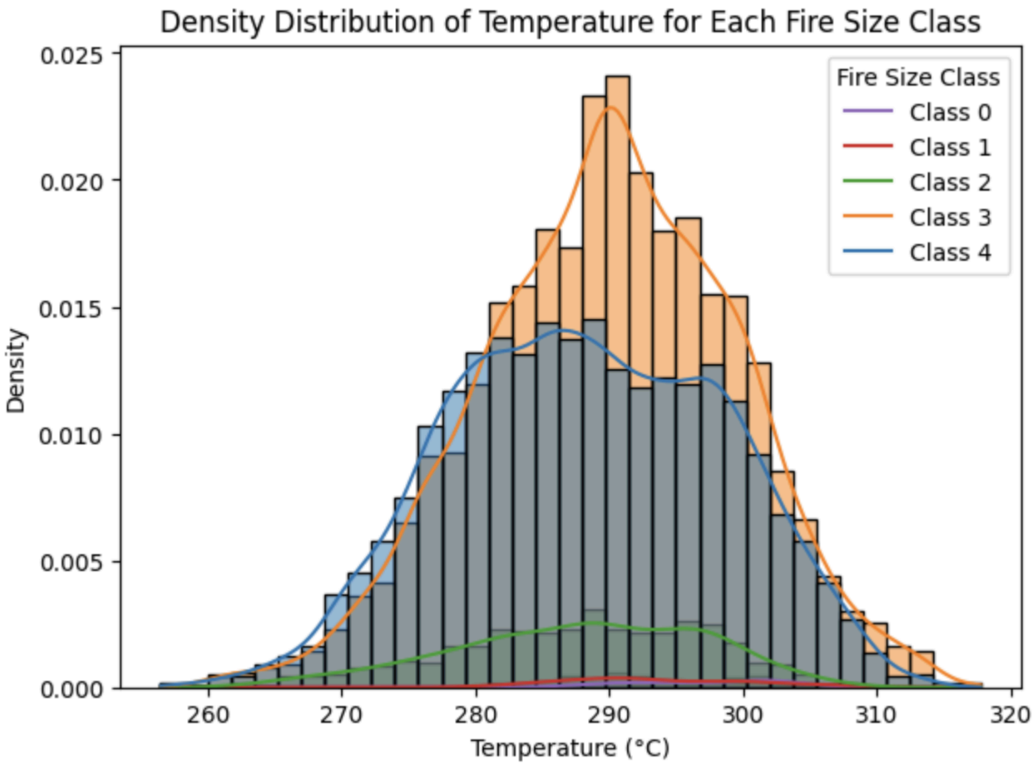


Fig. 12

Holding all other variables constant, a one standard deviation increase in temperature likely decreases the log odds of being in a larger fire size category by 0.2472 ($p < 0.001$ indicates a strong effect).

a. While this may seem counterintuitive, temperature is likely correlated with other environmental variables that directly affect fire spread, resulting in **omitted variable bias.** For example, cold temperatures are correlated with colder seasons and dry conditions, which could promote larger fires. In addition, in certain areas, higher temperatures may be associated with humid, fire-suppressing conditions, which could reduce fire spread and size. **This could be a potential area for further research.**

3. **Higher pressure is associated with larger fire sizes:** Holding all other variables constant, a one standard deviation increase in pressure likely increases the log odds of being in a larger fire size category by 0.1881 ($p < 0.001$ indicates a strong effect).

a. According to scientific observations, higher atmospheric pressure is generally associated with low humidity, clear skies, and prolonged dry spells/heat waves, leading to conditions that also promote the start and spread of wildfires.

4. **Higher wind speed is associated with larger fire sizes:**

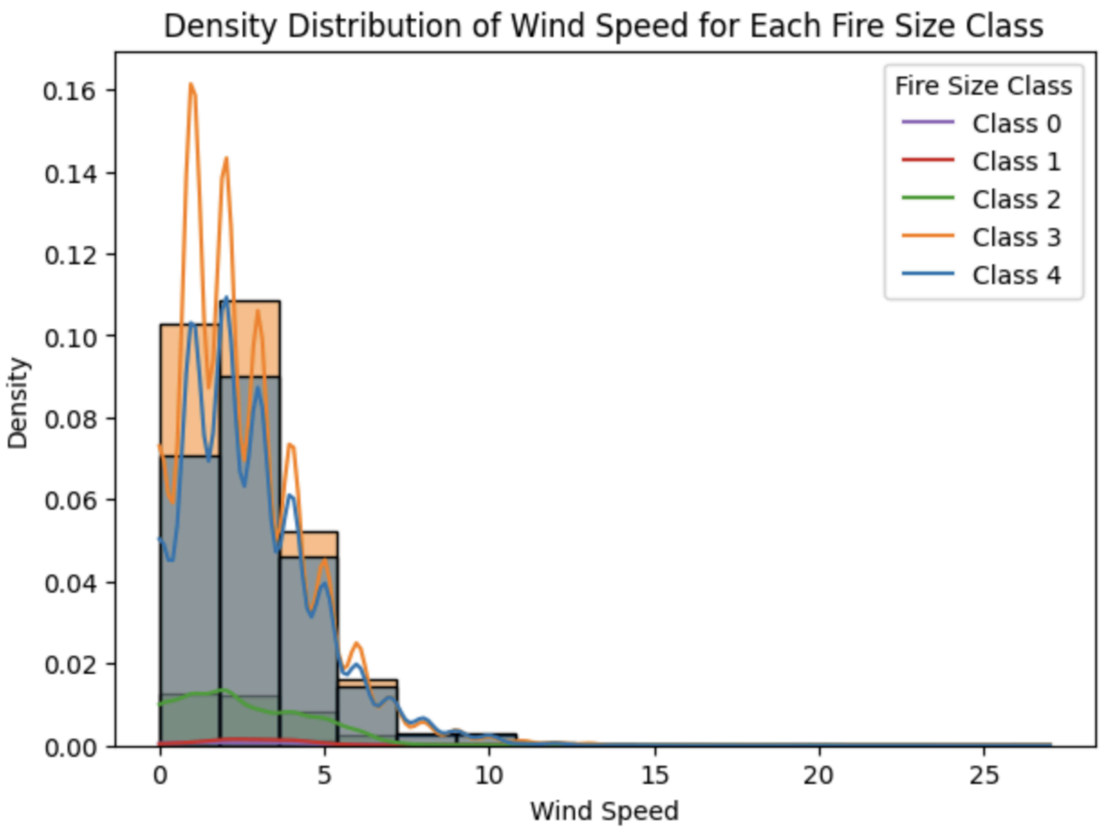


Fig. 13

Holding all other variables constant, a one unit increase in wind speed likely increases the log odds of being in a larger fire size category by 0.0590 ($p < 0.001$ indicates a strong effect).

a. These results also align with scientific observations. Because wind supplies fresh oxygen to fire, wildfires in higher wind conditions tend to have higher intensity and combustion efficiency. In addition, stronger winds can direct flames horizontally, leading to faster spread of fires.

5. **Wind direction has a significant but very small effect on larger fire sizes:** Holding all other variables constant, a one unit increase in wind direction likely decreases the log odds of being in a larger fire size category by 0.0007 (likely not a significant prediction feature for fire size).

6. **Locations with lower latitudes are associated with larger fire sizes:** Holding all other variables constant, a one degree decrease in latitude likely increases the log odds of being in a larger fire size category by 0.1216 ($p < 0.001$ indicates a strong effect).

  a. Because latitude is strongly correlated with temperature and regions with lower latitudes are closer to the equator and receive more direct sunlight year round, they may have more frequent droughts and lower humidity levels, leading to conditions that are prone to large wildfires.

7. **Threshold (Cut-Off) Values Interpretation**
   a. 0/1: The log-odds of a fire being in size class 1 or higher compared to class 0 is **-4.3367 ($p < 0.001$)**. Because this value is highly negative, it suggests that **class 0 fires are very common**, and the probability of moving to class 1 or higher is relatively low.
   b. 1/2: The log-odds of being in size class 2 or higher compared to class 1 is **0.5553 ($p < 0.001$)**. Since this value is greater than the 0/1 threshold, it indicates that transitioning from **class 1 to class 2 is easier than moving from class 0 to class 1**.
   c. 2/3: The log-odds of being in size class 3 or higher compared to class 2 is **0.7283 ($p < 0.001$)**. This suggests that the likelihood of moving into class 3 continues to increase as fire size grows.
   d. 3/4: The log-odds of being in size class 4 or higher compared to class 3 is **-0.0127 ($p = 0.892$)**, indicating an **insignificant effect**. Since this value is close to zero and the p-value is above the **0.05 significance level**, it suggests that **the model struggles to distinguish between class 3 and class 4 fires**. Even under extreme weather conditions, the transition between these two largest fire sizes may be less clearly defined in the dataset.

❖ Ordinal Logistic Regression Model Summary:
The model identifies several significant predictors of fire size, with lower humidity, lower temperature, higher pressure, higher wind speed, and lower city latitude all increasing the odds of a larger fire. While wind direction is statistically significant, its small coefficient suggests a negligible practical effect on fire size. Additionally, an analysis of threshold values indicates that smaller fire sizes are more common, but as fire size increases, the model is more likely to classify fires into higher size categories, reflecting a natural progression in fire growth.

8. Model Experimentation with Machine Learning (Random Forest Classifier and Neural Networks)

After experimenting with the logistic regression model, we tried to use other machine learning models for predicting the fire.

8.1) Random Forest Classification

First we selected the Random Forest model, which is a traditional and interpretable machine learning model based on decision trees. We have a standard 8:2 split of training and testing set, and run on 18000 samples that have 6 fire size categories.

- ❖ Pipeline

```
Pipeline(steps=[('preprocessor',
                 ColumnTransformer(transformers=[('num', StandardScaler(),
                                                  ['Humidity', 'Temperature',
                                                   'Pressure', 'Wind_Speed']),
                                                 ('cyclical_hour',
                                                  FunctionTransformer(func=<function <lambda> at 0x3493e3c40>),
                                                  ['hour']),
                                                 ('cyclical_wind',
                                                  FunctionTransformer(func=<function <lambda> at 0x3493e18a0>),
                                                  ['Wind_Direction'])])),
                ('classifier',
                 RandomForestClassifier(class_weight='balanced'))])
In a Jupyter environment, please rerun this cell to show the HTML representation or trust the notebook.
On GitHub, the HTML representation is unable to render, please try loading this page with nbviewer.org.
Pipeline
Pipeline(steps=[('preprocessor',
                 ColumnTransformer(transformers=[('num', StandardScaler(),
                                                  ['Humidity', 'Temperature',
                                                   'Pressure', 'Wind_Speed']),
                                                 ('cyclical_hour',
                                                  FunctionTransformer(func=<function <lambda> at 0x3493e3c40>),
                                                  ['hour']),
                                                 ('cyclical_wind',
                                                  FunctionTransformer(func=<function <lambda> at 0x3493e18a0>),
                                                  ['Wind_Direction'])])),
                ('classifier',
                 RandomForestClassifier(class_weight='balanced'))])
preprocessor: ColumnTransformer
ColumnTransformer(transformers=[('num', StandardScaler(),
                                 ['Humidity', 'Temperature', 'Pressure',
                                  'Wind_Speed']),
                                ('cyclical_hour',
                                 FunctionTransformer(func=<function <lambda> at 0x3493e3c40>),
                                 ['hour']),
                                ('cyclical_wind',
                                 FunctionTransformer(func=<function <lambda> at 0x3493e18a0>),
                                 ['Wind_Direction'])])
num
['Humidity', 'Temperature', 'Pressure', 'Wind_Speed']
StandardScaler
StandardScaler()
cyclical_hour
['hour']
FunctionTransformer
FunctionTransformer(func=<function <lambda> at 0x3493e3c40>)
cyclical_wind
['Wind_Direction']
FunctionTransformer
FunctionTransformer(func=<function <lambda> at 0x3493e18a0>)
RandomForestClassifier
RandomForestClassifier(class_weight='balanced')
```

Fig. 14

- ❖ We achieved a decent ~71% prediction accuracy using the random forest model. Then we checked the feature importance of the model. We can see that the longitude is the most important variable, and humidity is the second. This actually made a lot of sense. In Fig. 7, we could see that the fires are clustered across the country and the most significant fires occur in certain places such as Florida and Texas, which have a different longitude compared to other places. Also, humidity is certainly closely related to fires.

```
print("Accuracy:", model.score(X_test, y_test))
Accuracy: 0.7058220056424724
```

Fig. 15

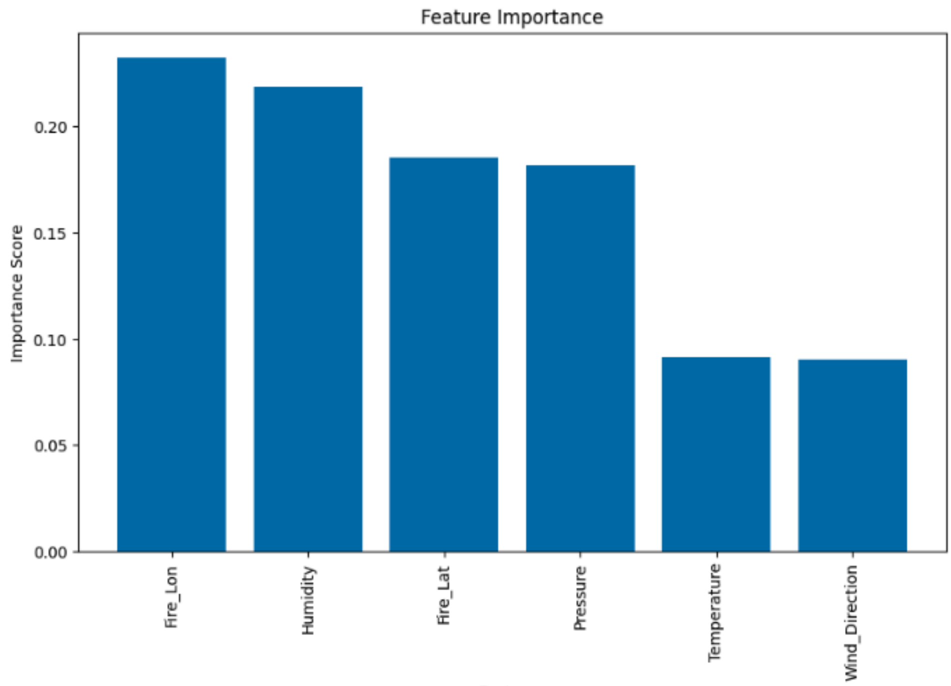


Fig. 16

❖ Confusion matrix

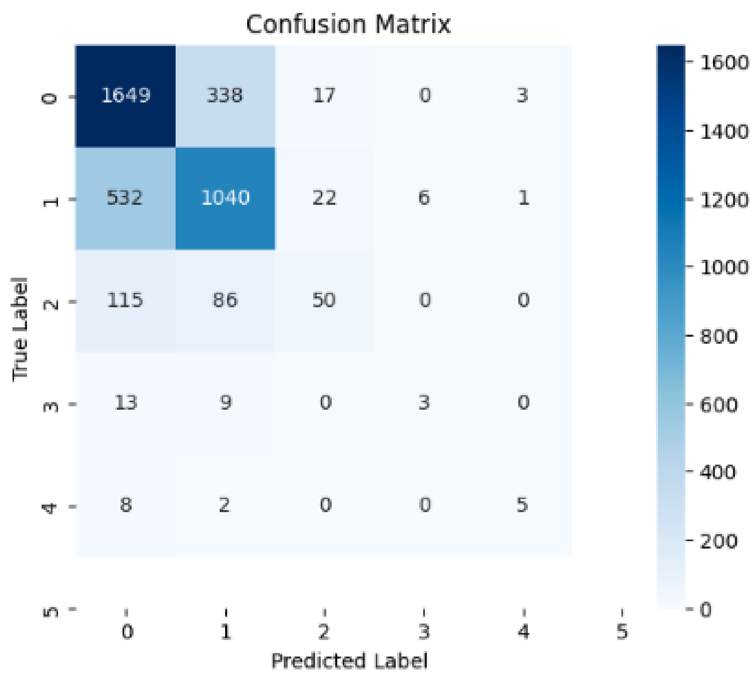

Fig. 17

8.2) Neural Networks

❖ Secondly, we tried to train a more complicated neural network model using the same data as the random forest model. It is a 5-layer fully connected network, adding Batch Normalization and Dropout for each layer separately.

```
Epoch 10: Train Loss = 1.2441, Val Loss = 1.1968
Epoch 20: Train Loss = 1.1136, Val Loss = 1.0815
Epoch 30: Train Loss = 1.0934, Val Loss = 1.0632
Epoch 40: Train Loss = 1.0364, Val Loss = 1.0221
Epoch 50: Train Loss = 1.0457, Val Loss = 1.0494
Epoch 60: Train Loss = 1.0589, Val Loss = 1.0005
Epoch 70: Train Loss = 1.0213, Val Loss = 0.9808
Epoch 80: Train Loss = 0.9939, Val Loss = 0.9620
Epoch 90: Train Loss = 1.0088, Val Loss = 0.9842
Early stopping triggered
<All keys matched successfully>
```

Fig. 18

❖ Training loss

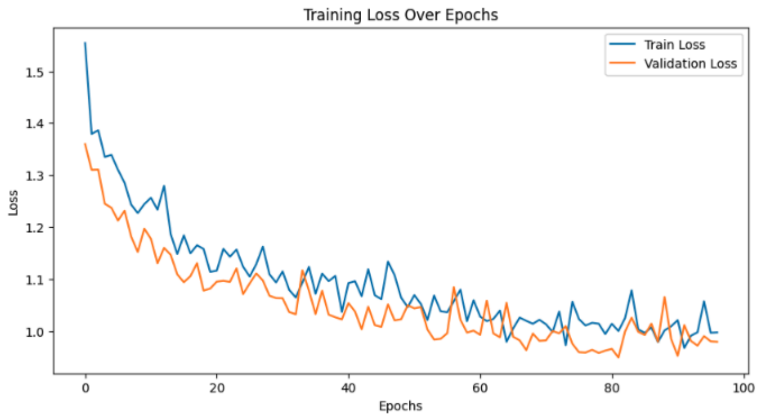


Fig. 19

❖ Results: However, this more complicated model did not perform well compared to the random forest model. From the confusion matrix, we can see that the model did not catch the skewed data distribution (majority of the data has fire level 0), and predicted a lot of false positives.

```
test_acc = correct / total
print(f"Test Accuracy: {test_acc:.4f}")

Test Accuracy: 0.2598
```

Fig. 20

❖ Confusion matrix

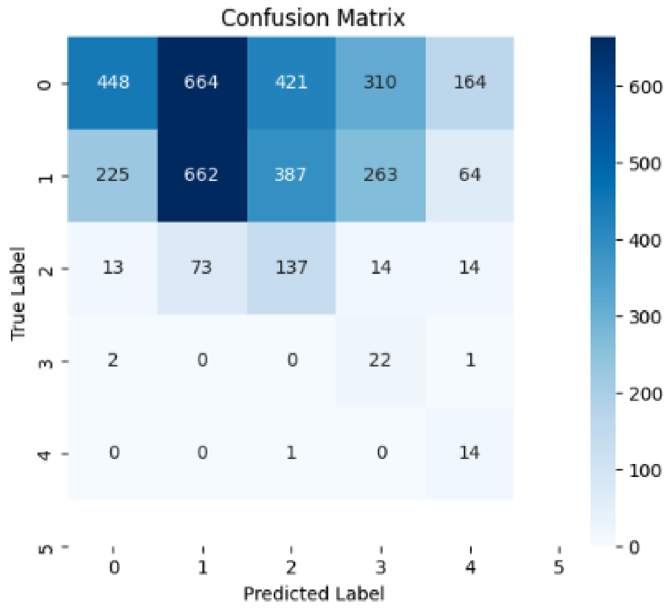


Fig. 21

## 8. Limitations and Future Work

Limitations:

- ❖ Incomplete Data: Our data set is limited to fire information in 21 U.S. states. This means our predictions may not generalize well to other states.
- ❖ Temporal Predictions: Our model has been trained on data from 2012 - 2015. However, the wildfire patterns observed in the historical data may not hold true for future predictions due to climate change. This means the model has to be constantly updated with real-time weather and wildfire data to ensure it results in reliable predictions.
- ❖ Omitted Variable Bias: Because our data set is limited to certain predictor variables, our model may suffer from omitted variable bias, meaning that certain key factors that influence wildfire size are not included in the analysis due to the lack of sufficient wildfire data. One major example of this is temperature interactions; we included temperature as a standalone predictor, but temperature generally interacts with other environmental conditions, such as fuel load, lightning strikes, and land cover type. As such, the omission of these variables could introduce bias in our predictions, making our estimates less reliable in certain conditions.
- ❖ Class Imbalance: Our dataset is the disproportionate representation of small fires (Class 0) compared to larger fires (Classes 3 and 4). As a result, the model may become biased toward predicting smaller fires accurately while underperforming in identifying and classifying the relatively rare large fires. This imbalance can lead to skewed evaluation metrics and reduce the model's real-world applicability.

- ❖ Outliers: Pressure and Wind speed have more outliers than Humidity, Wind direction, or Temperature.

## 9. Relation to Project Proposal

Since the proposal, we've streamlined our research question to make it more focused. We conducted the statistical analyses mentioned in our proposal, and incorporated additional modeling.

## 10. Conclusion

In this study, we investigated the relationship between meteorological variables and wildfire severity using a dataset that merges wildfire records (2012–2015) with weather observations from the United States. Humidity and temperature emerged as key environmental factors, with low humidity and unexpectedly low temperatures increasing the odds of larger wildfires. Wind speed and pressure were also associated with larger fires, suggesting that windy, high-pressure weather systems can accelerate fire spread. Additionally, regions closer to the equator tended to have larger fires due to their generally warmer and drier climate. This project has several limitations. Expanding the dataset to include additional states, updated years, and other environmental factors (fuel moisture, topography, land use) could improve predictive ability. We also anticipate that continued improvements in machine learning models and incorporating domain-specific data (e.g., lightning) will further improve accuracy and practical utility. The results of this project provide a preliminary framework for identifying high-risk conditions that can lead to large fires. We can apply these insights to more effectively allocate resources and focus them on areas or weather conditions most suited to severe wildfires.

**Citations**

1. Danielle, Monica. “AccuWeather Estimates More than $250 Billion in Damages and Economic Loss from La Wildfires.” *AccuWeather Estimates More than $250 Billion in Damages and Economic Loss from LA Wildfires*, AccuWeather, 16 Jan. 2025, www.accuweather.com/en/weather-news/accuweather-estimates-more-than-250-billion-in-damages-and-economic-loss-from-la-wildfires/1733821#:~:text=News%20%2F%20Weather%20News-,AccuWeather%20estimates%20more%20than%20%24250%20billion%20in%20damages%20and%20economic,the%20entire%202020%20wildfire%20season.
2. Selfish Gene. (2018). *Historical hourly weather data*. Kaggle. Retrieved from https://www.kaggle.com/datasets/selfishgene/historical-hourly-weather-data?select=wind_speed.csv
3. Tatman, R. (2020). *188 million US wildfires*. Kaggle. Retrieved from https://www.kaggle.com/datasets/rtatman/188-million-us-wildfires