California Wildfire Prediction

Course: ISTM 6217 Internet of Things Management Project

Group Members: Saif Alneyadi, Pushti Chawala, Ayla Karimova, Sanjay Devang, Quang Tran

Table of Contents

Introduction

Wildfires are becoming one of the most devastating challenges facing California today. Each year, thousands of acres of land go up in flames, leaving behind destroyed homes, damaged ecosystems, and communities struggling to recover. In the past five years alone, wildfires have burned more than 10 million acres across the state. The scale of the destruction is overwhelming, costing billions of dollars in property loss and taking lives that can never be replaced.

As climate change makes extreme weather conditions more common, the risk of wildfires continues to grow. Dry landscapes, strong winds, and rising temperatures create the perfect conditions for fires to spark and spread quickly. Unfortunately, the current methods used to detect and respond to wildfires are often too slow. In many cases, the first sign of danger comes from a 911 call or a satellite image, which means precious time has already been lost. When it comes to wildfires, every second matters.

Our project is built around the idea that we can do better. We believe that technology can help us see the signs of danger earlier, respond more effectively, and ultimately save lives and resources. That's why we're developing an early warning and risk prediction system that uses the power of the Internet of Things, or IoT. By placing smart sensors in the environment, we can monitor important factors like temperature, humidity, and wind speed in real time. This live data gives us a much clearer picture of the conditions that could lead to a wildfire.

But sensors alone aren't enough. To truly understand the risks, we need to look at the bigger picture. That's where big data platforms like Apache Spark come in. These tools allow us to process huge amounts of information quickly and find patterns that might otherwise go unnoticed. On top of that, we're using machine learning models to make sense of the data. These models learn from past wildfires and help predict where and when a fire might occur.

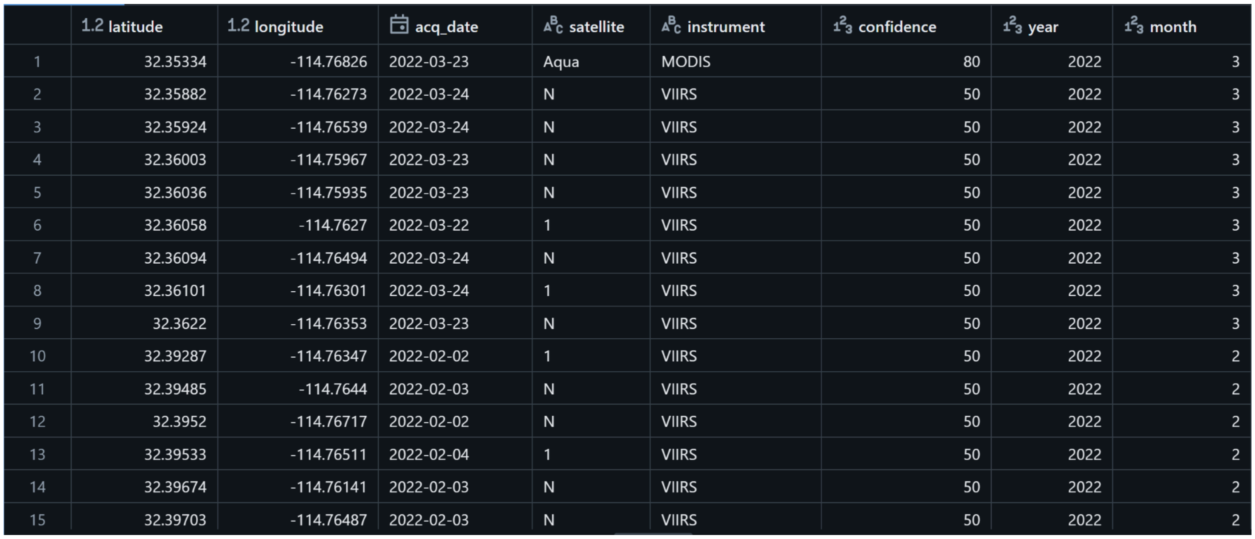
The goal of our system is to move from reacting to wildfires after they start to preventing them before they do. We want to enable early detection, reduce false alarms, and give emergency teams the information they need to act fast. By providing real-time risk assessments, we also hope to protect communities that are especially vulnerable to wildfire threats.

In this paper, we will walk through the technologies behind our system, the way it works, and the difference it can make. At its heart, this project is about using technology in a thoughtful and meaningful way to address a real-world problem. We believe that with the right tools and a proactive mindset, we can help build a safer and more resilient California.

Data Management

Dataset

California is one of the places having the most deadliest and destructive wildfire seasons.We have combined two different datasets to reach accurate results. The dataset contains the list of Wildfires that have occurred in California between 2013 and 2020. The dataset contains the location where wildfires have occurred including the County name, latitude and longitude values and also details on when the wildfire has started. This data helps to generate insights on what locations in California are under fire threat, what time do Wildfires usually occur and how frequent and devastating they are!



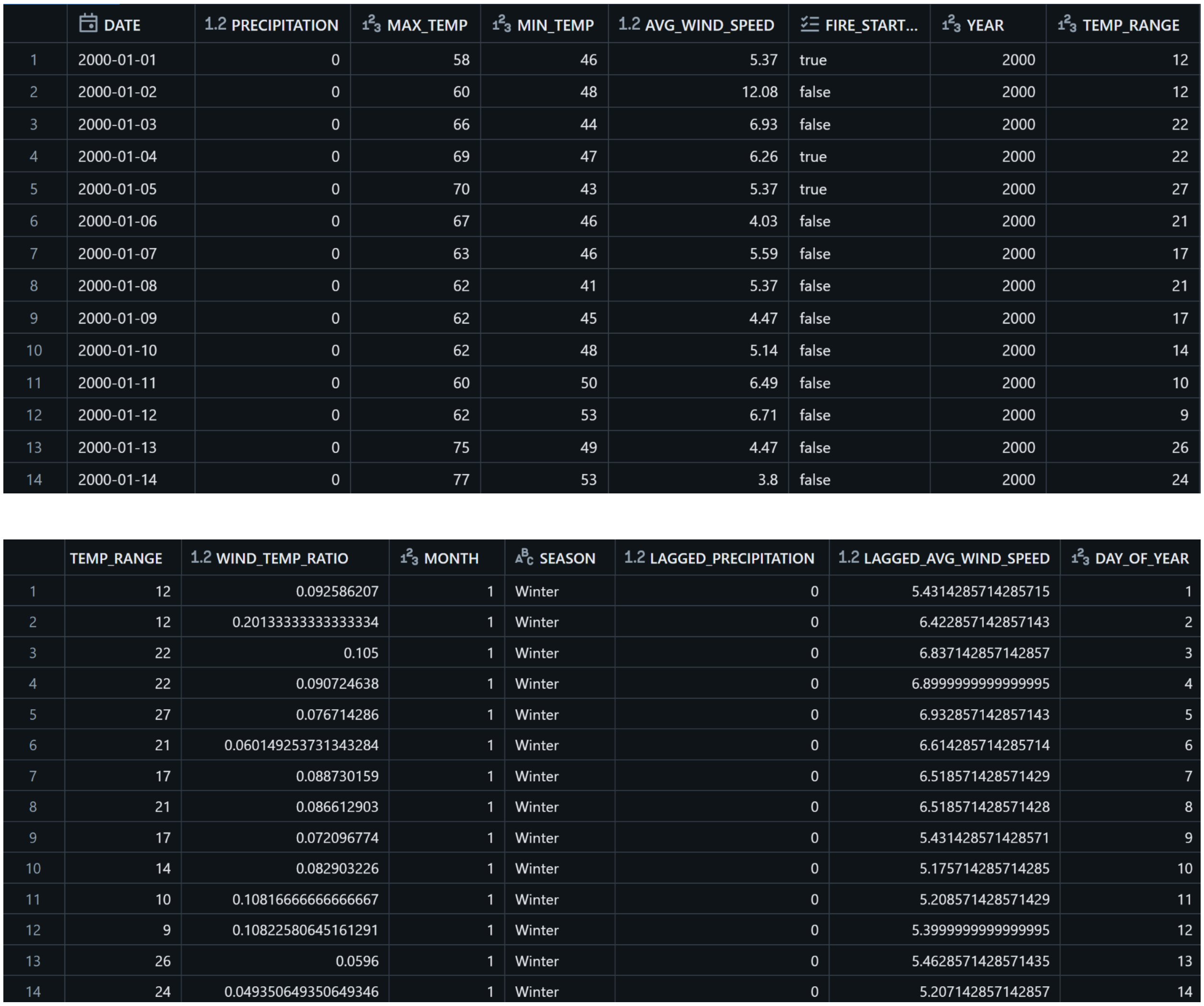
	latitude	longitude	acq_date	satellite	instrument	confidence	year	month
1	32.35334	-114.76826	2022-03-23	Aqua	MODIS	80	2022	3
2	32.35882	-114.76273	2022-03-24	N	VIIRS	50	2022	3
3	32.35924	-114.76539	2022-03-24	N	VIIRS	50	2022	3
4	32.36003	-114.75967	2022-03-23	N	VIIRS	50	2022	3
5	32.36036	-114.75935	2022-03-23	N	VIIRS	50	2022	3
6	32.36058	-114.7627	2022-03-22	1	VIIRS	50	2022	3
7	32.36094	-114.76494	2022-03-24	N	VIIRS	50	2022	3
8	32.36101	-114.76301	2022-03-24	1	VIIRS	50	2022	3
9	32.3622	-114.76353	2022-03-23	N	VIIRS	50	2022	3
10	32.39287	-114.76347	2022-02-02	1	VIIRS	50	2022	2
11	32.39485	-114.7644	2022-02-03	N	VIIRS	50	2022	2
12	32.3952	-114.76717	2022-02-02	N	VIIRS	50	2022	2
13	32.39533	-114.76511	2022-02-04	1	VIIRS	50	2022	2
14	32.39674	-114.76141	2022-02-03	N	VIIRS	50	2022	2
15	32.39703	-114.76487	2022-02-03	N	VIIRS	50	2022	2

Daily Wildfire Dataset

The second dataset provides in-depth information about weather related factors that play a major role in starting wildfires and the route of spreading. This dataset ranges from 1984 to 2025, January. This dataset consists of 14,988 daily records covering temperature, precipitation, wind speed, and other meteorological factors.

This dataset also contains air pollution levels, including:

- PM2.5 (fine particulate matter)—a major pollutant from wildfire smoke.
- CO (carbon monoxide): Indicates combustion-related pollution.
- NO_2 (Nitrogen Dioxide) & O_3 (Ozone) contribute to smog and respiratory issues.
- Daily Air Quality Index (AQI): Measures overall pollution levels



	DATE	PRECIPITATION	MAX_TEMP	MIN_TEMP	AVG_WIND_SPEED	FIRE_START...	YEAR	TEMP_RANGE
1	2000-01-01	0	58	46	5.37	true	2000	12
2	2000-01-02	0	60	48	12.08	false	2000	12
3	2000-01-03	0	66	44	6.93	false	2000	22
4	2000-01-04	0	69	47	6.26	true	2000	22
5	2000-01-05	0	70	43	5.37	true	2000	27
6	2000-01-06	0	67	46	4.03	false	2000	21
7	2000-01-07	0	63	46	5.59	false	2000	17
8	2000-01-08	0	62	41	5.37	false	2000	21
9	2000-01-09	0	62	45	4.47	false	2000	17
10	2000-01-10	0	62	48	5.14	false	2000	14
11	2000-01-11	0	60	50	6.49	false	2000	10
12	2000-01-12	0	62	53	6.71	false	2000	9
13	2000-01-13	0	75	49	4.47	false	2000	26
14	2000-01-14	0	77	53	3.8	false	2000	24

	TEMP_RANGE	WIND_TEMP_RATIO	MONTH	SEASON	LAGGED_PRECIPITATION	LAGGED_AVG_WIND_SPEED	DAY_OF_YEAR
1	12	0.092586207	1	Winter	0	5.4314285714285715	1
2	12	0.20133333333333334	1	Winter	0	6.422857142857143	2
3	22	0.105	1	Winter	0	6.837142857142857	3
4	22	0.090724638	1	Winter	0	6.8999999999999995	4
5	27	0.076714286	1	Winter	0	6.932857142857143	5
6	21	0.060149253731343284	1	Winter	0	6.614285714285714	6
7	17	0.088730159	1	Winter	0	6.518571428571429	7
8	21	0.086612903	1	Winter	0	6.518571428571428	8
9	17	0.072096774	1	Winter	0	5.431428571428571	9
10	14	0.082903226	1	Winter	0	5.175714285714285	10
11	10	0.10816666666666667	1	Winter	0	5.208571428571429	11
12	9	0.10822580645161291	1	Winter	0	5.3999999999999995	12
13	26	0.0596	1	Winter	0	5.4628571428571435	13
14	24	0.049350649350649346	1	Winter	0	5.207142857142857	14

Weather Wildfire Dataset

Both the datasets are joined using Date as Primary Key and Foreign Key considering Date as the unique identifier in both the datasets. The major analysis is performed considering data from the year 2000 to 2021 as these years are the overlapping years in both datasets.

Data Cleaning

The weather dataset included entries from 1984 to 2025, while the satellite fire dataset covered 2000 to 2021. To ensure accurate time-series alignment and avoid null joins during joining, we synchronized the data for analysis by filtering the weather dataset to include dates from 2000–2021.

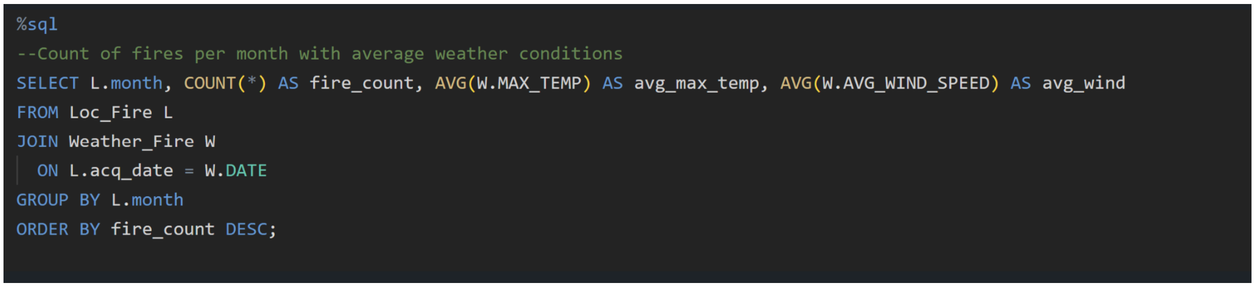
Next, we standardized the date formats. Both datasets contained date information in string form (DATE and acq_date), so we converted these to a common DateType using Spark SQL functions. This enabled accurate joins between the datasets on the date field.

After cleaning and joining both datasets, we handled null values with zeros replacement by using the COALESCE() function to maintain numeric consistency. We also checked for nulls or imputed missing values using rolling averages. By using Spark SQL, we queried environmental factors (temperature, wind speed, precipitation) to identify under what conditions wildfires likely occur.

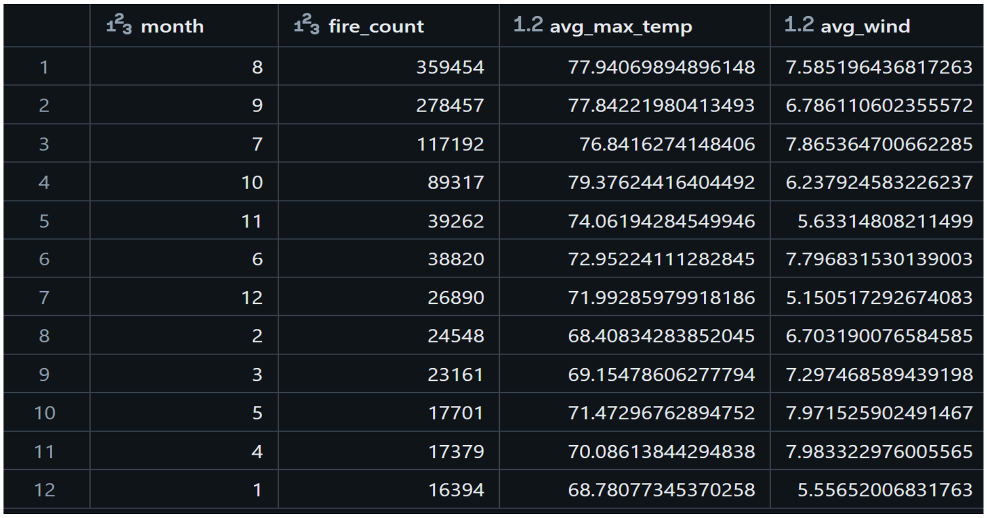
Data Retrieval

The data retrieval process served as the backbone for this project, enabling the integration of Wildfire Location data with corresponding weather conditions in California. The datasets were stored in relational tables, allowing structured querying and seamless data manipulation. Using SQL, we performed several essential operations to extract meaningful insights.

Counting Fires Per Month with Average Weather Conditions:
This query analyzes wildfire activity by month and examines the associated weather conditions. It joins two datasets — one containing wildfire location data (Loc_Fire) and another containing weather and fire data (Weather_Fire) — based on matching dates. For each month, it calculates the total number of fires, the average maximum temperature, and the average wind speed. The results are then sorted in descending order by the number of fires, highlighting the months with the highest wildfire activity. This analysis helps to identify seasonal trends in fire occurrences and the typical weather patterns during high-fire periods.



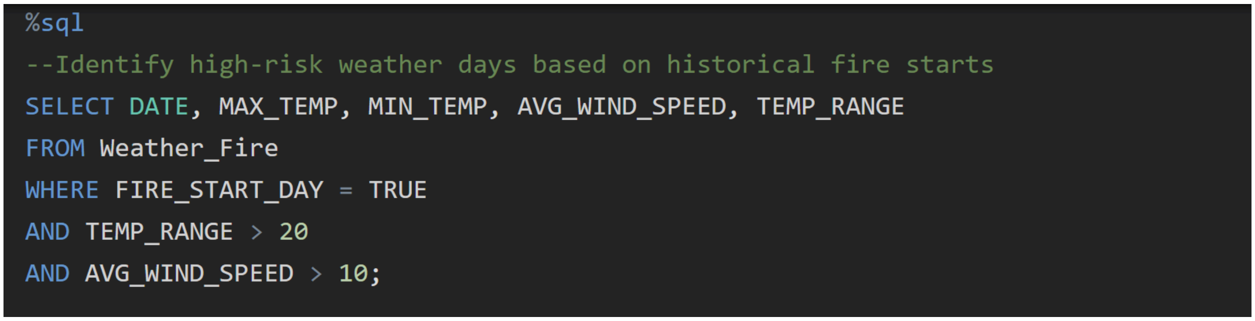
```sql
%sql
--Count of fires per month with average weather conditions
SELECT L.month, COUNT(*) AS fire_count, AVG(W.MAX_TEMP) AS avg_max_temp, AVG(W.AVG_WIND_SPEED) AS avg_wind
FROM Loc_Fire L
JOIN Weather_Fire W
  ON L.acq_date = W.DATE
GROUP BY L.month
ORDER BY fire_count DESC;
```



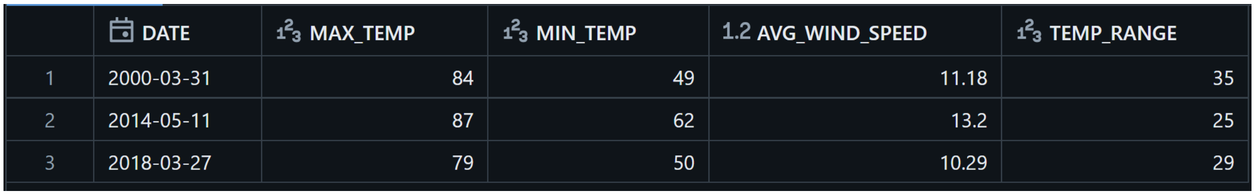
	month	fire_count	avg_max_temp	avg_wind
1	8	359454	77.94069894896148	7.585196436817263
2	9	278457	77.84221980413493	6.786110602355572
3	7	117192	76.8416274148406	7.865364700662285
4	10	89317	79.37624416404492	6.237924583226237
5	11	39262	74.06194284549946	5.63314808211499
6	6	38820	72.95224111282845	7.796831530139003
7	12	26890	71.99285979918186	5.150517292674083
8	2	24548	68.40834283852045	6.703190076584585
9	3	23161	69.15478606277794	7.297468589439198
10	5	17701	71.47296762894752	7.971525902491467
11	4	17379	70.08613844294838	7.983322976005565
12	1	16394	68.78077345370258	5.55652006831763

Identifying High-Risk Weather Days Based on Historical Fire Starts:

This query focuses on isolating historical days that had particularly dangerous weather conditions associated with the ignition of wildfires. It selects records from the Weather_Fire dataset where a fire start is confirmed (FIRE_START_DAY = TRUE). Further filtering is applied to capture only those days where the temperature range exceeded 20 degrees and the average wind speed was above 10 units. The query outputs the date, maximum temperature, minimum temperature, average wind speed, and temperature range for these high-risk days. The goal is to characterize weather patterns that are strongly correlated with the start of wildfires, potentially informing fire risk warning systems.



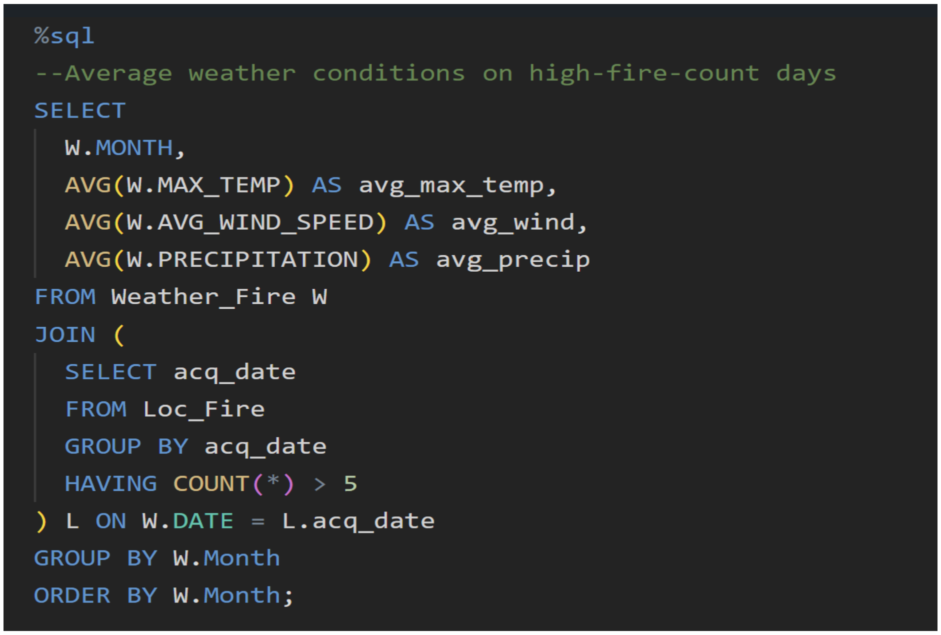
```sql
%sql
--Identify high-risk weather days based on historical fire starts
SELECT DATE, MAX_TEMP, MIN_TEMP, AVG_WIND_SPEED, TEMP_RANGE
FROM Weather_Fire
WHERE FIRE_START_DAY = TRUE
AND TEMP_RANGE > 20
AND AVG_WIND_SPEED > 10;
```



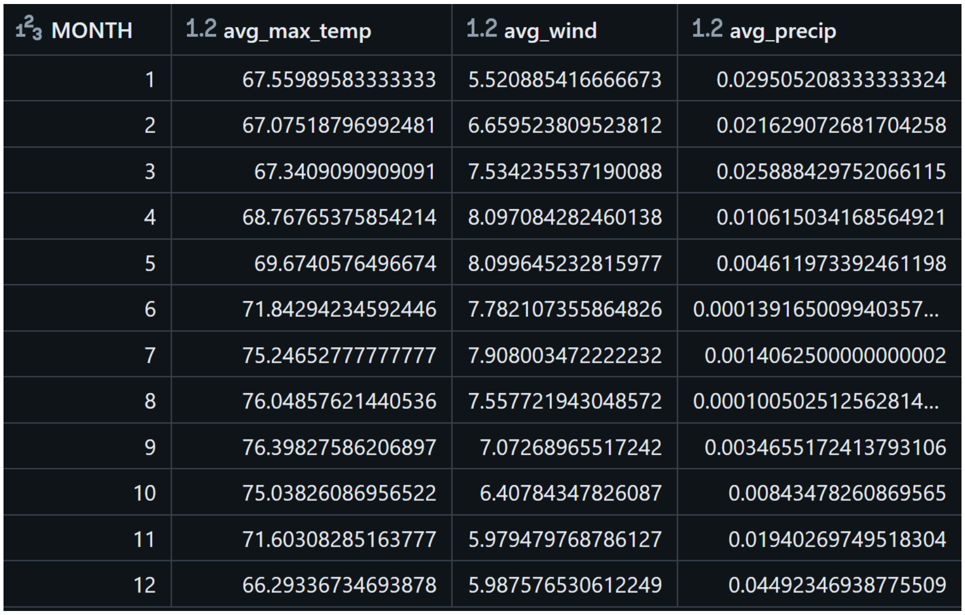
	DATE	MAX_TEMP	MIN_TEMP	AVG_WIND_SPEED	TEMP_RANGE
1	2000-03-31	84	49	11.18	35
2	2014-05-11	87	62	13.2	25
3	2018-03-27	79	50	10.29	29

Average Weather Conditions on High-Fire-Count Days:
This query investigates the weather conditions on days that experienced a particularly high number of fires. A subquery first identifies the dates from the Loc_Fire dataset where more than

five fire incidents occurred. These dates are then joined with the Weather_Fire dataset to access the corresponding weather information. The main query calculates the average maximum temperature, average wind speed, and average precipitation for each month, considering only those high-fire-count days. By grouping and ordering the results by month, this analysis provides insight into the weather characteristics that commonly coincide with extremely active fire days across different times of the year.



```
%sql
--Average weather conditions on high-fire-count days
SELECT
  W.MONTH,
  AVG(W.MAX_TEMP) AS avg_max_temp,
  AVG(W.AVG_WIND_SPEED) AS avg_wind,
  AVG(W.PRECIPITATION) AS avg_precip
FROM Weather_Fire W
JOIN (
  SELECT acq_date
  FROM Loc_Fire
  GROUP BY acq_date
  HAVING COUNT(*) > 5
) L ON W.DATE = L.acq_date
GROUP BY W.Month
ORDER BY W.Month;
```



MONTH	avg_max_temp	avg_wind	avg_precip
1	67.55989583333333	5.520885416666673	0.029505208333333324
2	67.07518796992481	6.659523809523812	0.021629072681704258
3	67.3409090909091	7.534235537190088	0.025888429752066115
4	68.76765375854214	8.097084282460138	0.010615034168564921
5	69.6740576496674	8.099645232815977	0.004611973392461198
6	71.84294234592446	7.782107355864826	0.000139165009940357...
7	75.24652777777777	7.908003472222232	0.0014062500000000002
8	76.04857621440536	7.557721943048572	0.000100502512562814...
9	76.39827586206897	7.07268965517242	0.0034655172413793106
10	75.03826086956522	6.40784347826087	0.008434782608695652
11	71.60308285163777	5.979479768786127	0.019402697495183048
12	66.29336734693878	5.987576530612249	0.044923469387755099

Exploratory Analysis

Temperature and Fire Intensity

To understand the influence of temperature, we grouped fire days by maximum temperature ranges and calculated the average number of fires per day:

Temp Range	Fire Days	Avg Fires/Day
< 60°F	295	16.10
60–79°F	6,037	125.06
80–99°F	827	348.68
100°F+	6	84.33

The 60–99°F band supports wildfire activity suggesting that is the dangerous range, likely due to vegetation dryness and low humidity.This also suggests moderate warmth should not be underestimated.

Wind Speed and Fire Propagation

Wind plays a crucial role in the spread of wildfires. And our second query grouped fire activity by average daily wind speed and calculated the average number of fire detections per day:

Wind Speed (mph)	Fire Days	Avg Fires/Day
6	1,535	168.64
7	1,479	185.62
8	1,310	165.30
9	828	155.82
10	249	147.65
< 5	~1,380	47–100

Fire activity peaked in the 6–9 mph wind range. This finding indicates that even moderate winds are impactful on promoting fire spread, likely due to their ability to distribute flames. However, days with very low or very high winds saw low fire activity, possibly due to stagnation or excessive turbulence.

Precipitation and Fire Suppression

Rainfall is a natural suppressant for wildfires. Our third query analyzed the relationship between daily precipitation and total fire detections:

Precipitation (inches)	Days	Total Fires
0.0	7,591	1,040,075
0.1	170	4,856
0.2–0.5	211	~2,000
>1.0	<20	~100

The finding shows that 99% of all wildfires occurred on dry days with no measurable rain. Even light rainfall was likely to reduce wildfire.

Logistic regression

What is logistic regression?

Logistic regression is a simple method used to help predict yes or no outcomes. In the California fires project, we used logistic regression to predict the start day of a fire based on weather conditions like temperature, wind speed, and precipitation. The model looks at those weather features and calculates the chance of a fire happening. If that chance is high, it predicts “yes.” If it’s low, it predicts “no” This helps identify high-risk days beforehand, so that fire crews and surrounding communities can prepare early.

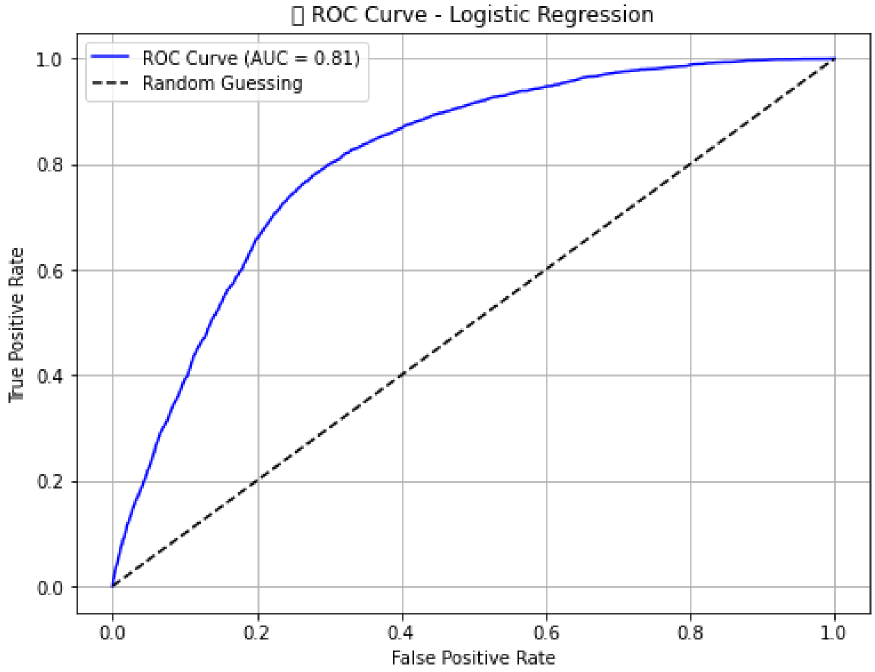
How was it used in this project?

In this project, logistic regression was used to predict whether a fire would start on a specific day based on daily weather data. The dataset included weather features such as maximum temperature, minimum temperature, precipitation, wind speed, and other related indicators. Each day was labeled as either a fire start day (1) or non-fire day (0).

We trained the logistic regression model using these weather features to learn patterns that are common on fire days. Once trained, the model could look at the weather for a new day and predict whether a fire was likely to start or not. This approach helps provide early warning for fire risk, allowing fire management teams to take action before a fire happens.

What was the result & model performance?

The ROC Curve AUC of 0.81 indicates that the model is able to correctly predict fire start days 81% of the time, showing strong overall performance. The curve rises sharply toward the top-left corner, which reflects the model's high sensitivity and a low false positive rate at early decision thresholds. This means the model is effective at identifying true fire events while minimizing false alarms. Overall, the logistic regression model performs well in predicting the occurrence of fires using weather data alone. It has practical value and can be integrated into early-warning systems or fire risk dashboards to help anticipate and prepare for fire events in advance.



Key Takeaways:

Wildfires are no longer unpredictable – with data and smart systems, we can act before it’s too late.

Traditionally, wildfires have been seen as sudden and uncontrollable natural events. However, this project shows that by leveraging historical weather data and applying machine learning models, we can detect patterns that signal increased fire risk. These predictive insights empower fire management teams to respond proactively rather than reactively, making wildfire events more manageable and less destructive.

Regression models can be used to predict wildfire occurrence and support operational strategy.

Using logistic regression, we demonstrated how weather features such as temperature, wind speed, and precipitation can predict the likelihood of a fire starting on any given day. This enables the development of data-driven operational strategies—such as staging firefighting resources in high-risk areas or issuing public alerts—before fires occur.

The models developed can serve as a foundation for resource planning and policy decisions.
The predictive models built in this project can support long-term planning by informing when and where fire prevention efforts are needed most. Policymakers and emergency planners can use these insights to prioritize budgets, schedule controlled burns, plan evacuations, and allocate equipment more efficiently. As climate conditions shift, these models can be retrained and improved, ensuring California remains one step ahead in wildfire preparedness.

XGBoost Regressor: Predicting Fire Count from Weather

What is XGBoost Regressor?

XGBoost (Extreme Gradient Boosting) is a powerful machine learning algorithm that uses an ensemble of decision trees. Unlike logistic regression, which predicts a yes/no outcome, XGBoost Regressor is used for predicting continuous numeric values — in our case, the daily number of fires observed based on weather conditions.

It handles non-linear relationships well and is optimized for speed and performance, making it ideal for large, complex datasets like fire activity combined with weather records.

How was it used in this project?

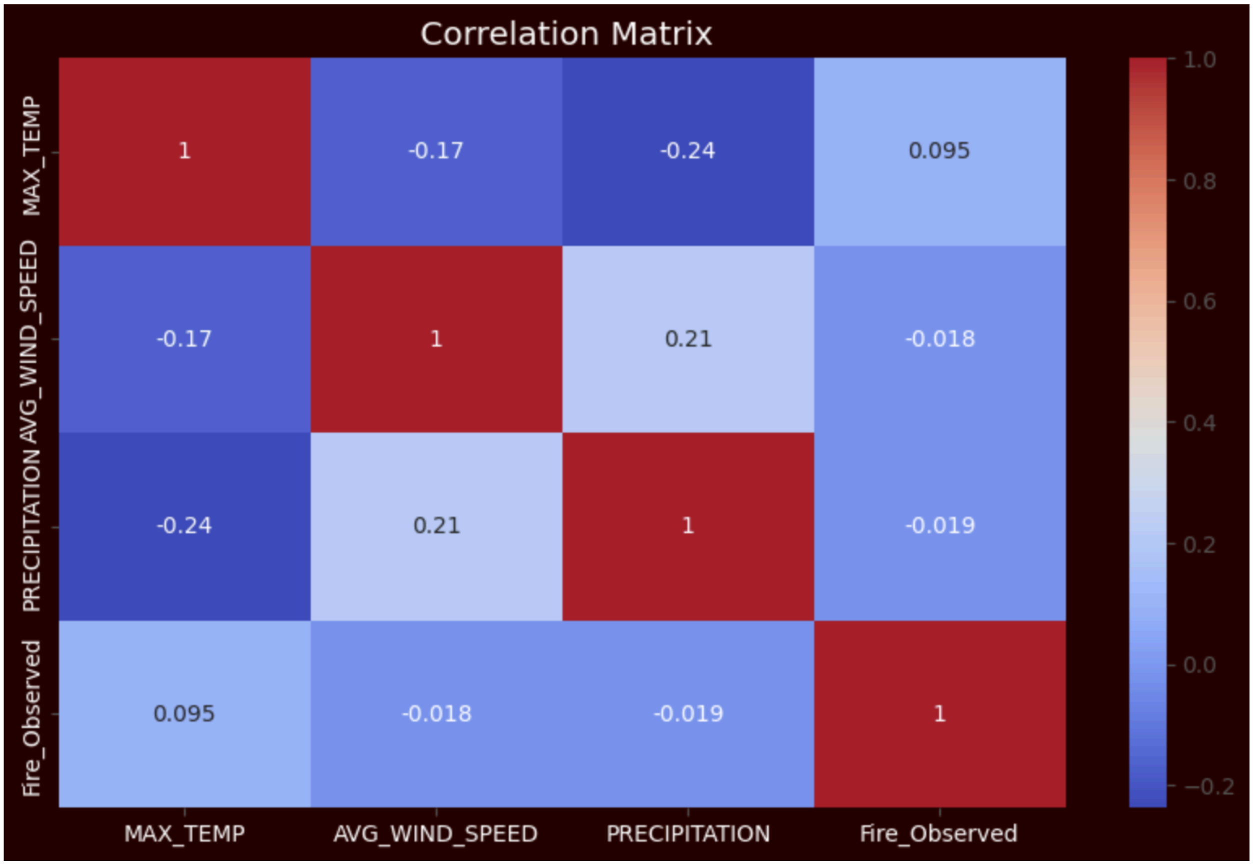
In this project, the XGBoost Regressor was used to predict the number of fires per day using weather features such as:

- Maximum Temperature
- Wind Speed
- Precipitation
- and other engineered variables like lagged values and temperature range

The model was trained using 80% of the dataset and tested on the remaining 20%, ensuring the model could generalize and didn’t overfit to training data.

What insights were found?

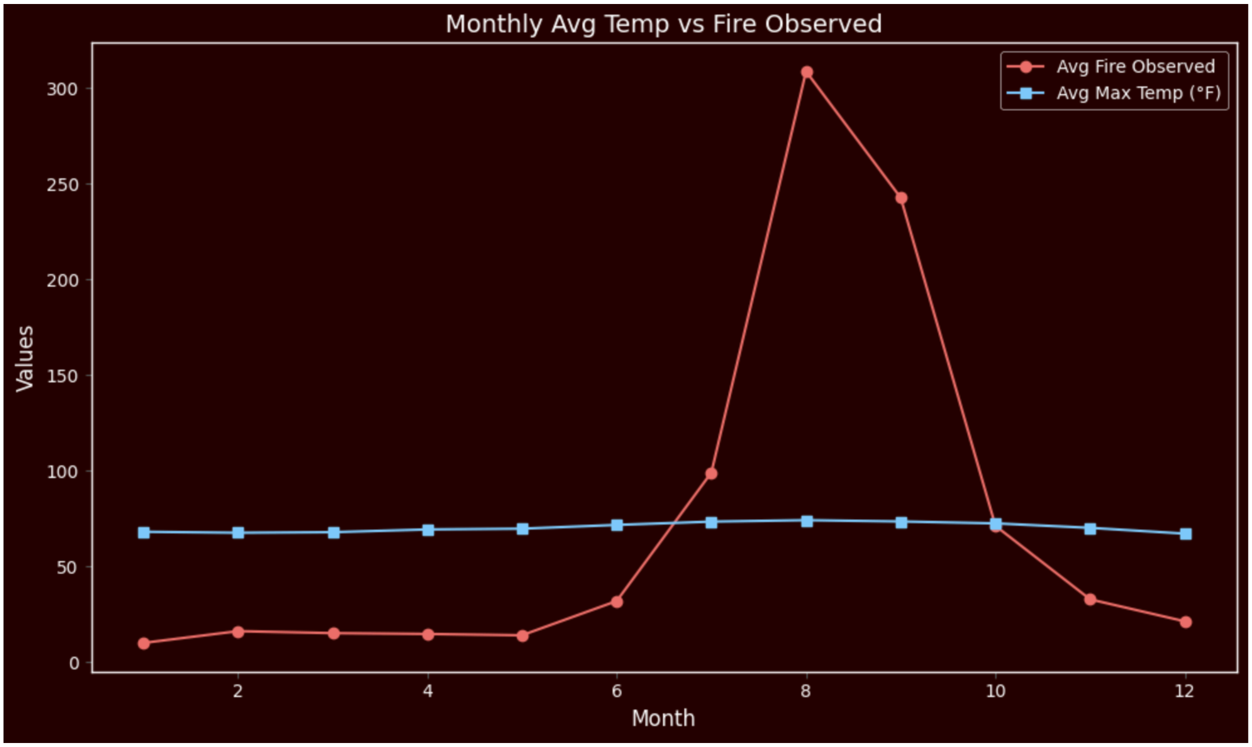
- Correlation analysis showed only weak relationships between individual weather variables and fire count.
    - For example, MAX_TEMP had a weak positive correlation (0.095) with fire count, indicating that while temperature may play a role, it's not the only factor.
- A scatter plot revealed that most fires occurred between 65°F and 85°F, especially clustering around 75°F–80°F.
- Monthly trend visualization showed fire activity peaks in late summer (August) while temperature remains more stable.



Model Performance:

- The XGBoost Regressor was evaluated using RMSE (Root Mean Squared Error) and R² (coefficient of determination).
- While the correlation between variables was weak individually, the model successfully captured multi-feature patterns that affect fire activity.

- ➢ This means that even though no single weather feature strongly predicts fire activity, their combined influence can be effectively modeled using XGBoost.



Why this matter?

This model enables fire response teams to estimate the expected number of fires daily, rather than just risk level. Such quantitative forecasting can help with resource allocation, planning patrols, and launching awareness campaigns on high-risk days.

Key Takeaways:

- ➢ Fire activity peaks in late summer (July–September), aligning with higher temperatures and lower precipitation.
- ➢ Individual weather features had weak correlations with fire activity, but combining them using machine learning improved predictive power.
- ➢ The models we developed can support early warning systems, seasonal preparedness plans, and emergency response strategies.

Conclusion

Our project combined historical fire records with daily weather data to understand and predict wildfire activity in California. We explored both classification and regression models to address different perspectives of fire prediction:

- ➢ Logistic Regression helped identify whether a fire is likely to occur on a given day using weather indicators — offering a binary risk assessment.
- ➢ XGBoost Regressor provided quantitative predictions of how many fires may occur — supporting better planning and resource deployment.

References

"Wildfires." Office of Environmental Health Hazard Assessment, 2024, https://oehha.ca.gov/climate-change/epic-2022/impacts-vegetation-and-wildlife/wildfires.

"Statistics." Cal Fire, 2024, https://www.fire.ca.gov/our-impact/statistics.

"Ultra Early Wildfire Detection." Dryad Networks, 2024, https://www.dryad.net/.

"Internet of Things (IoT) and Forest Fires Detection [1/2]." Kineis, 2023, https://www.kineis.com/en/iot-satellite-forest-fire-detection/.

"Predicting Forest Fires with Apache Spark." LinkedIn, 2017, https://www.linkedin.com/pulse/predicting-forest-fires-apache-spark-ian-downard.

"Analysis of Wildfire Danger Level Using Logistic Regression Model." MDPI, 2023, https://www.mdpi.com/1999-4907/14/12/2352.