JOURNAL OF
EMERGING INVESTIGATORS

Article

Utilizing meteorological data and machine learning to predict and reduce the spread of California wildfires

Sejal Bilwar[1*], Siddharth Taneja[1*], Saarth Gaonkar[1*], Sahil Mehta[1*], Yash Chanchani[1*], Larry McMahan[1]

[1] Aspiring Scholars Directed Research Program, Fremont, California

[*] These authors contributed equally to this work

SUMMARY

With the increase in the number and severity of California wildfires, firefighters need more efficient tools to predict potential wildfires so that they can efficiently cease these devastating events, preserve the safety of the human population, and the environment. With nearly eight million acres burned between 2010 and 2019, there is a clear and immediate issue. We hypothesized that a machine learning model could be developed to accurately predict the severity of California wildfires and determine the most influential meteorological factors. In this study, we used machine learning to analyze historical fire behaviors and then accurately predict the severity (defined by acres burned) of California wildfires. We created a custom dataset by using a combination of information from the World Weather Online API and a Kaggle dataset of wildfires in California from 2013-2020. The resulting dataset consisted of information (including temperature, humidity, dew point, and wind speeds) on 1,462 wildfires. We used three classification algorithms—logistic regression, support vector machine, and random forest regression—to classify these fires. Devising a system categorizing fires based on acres burned, our algorithms were able to classify fires into seven categories with promising accuracy (around 55 percent). We generated a correlation matrix, providing insight into the most concerning factors and allowing firefighters to choose the best course of action. We found that higher temperatures, lower humidity, lower dew point, higher wind gusts, and higher wind speeds are the most significant contributors to the spread of a wildfire. This machine learning tool could vastly improve the efficiency and preparedness of firefighters as they deal with these crises.

INTRODUCTION

California is facing an extreme wildfire problem. Over the last two decades, millions of acres of land have been burned, and there has been a constant increase in the severity and rate of these fires, with over 7,802,985 acres being burned across 81,943 wildfires between 2010 and 2019 (1). The state is especially prone to frequent wildfires for several reasons (1). Worsening drought has brought a lack of soil moisture. Soil moisture is not only necessary to provide water to vegetation, but to protect vegetation from becoming easily ignitable in the summer months. Combined with increasingly drastic temperature increases, California has become incredibly susceptible to these disasters (1). Finally, the increased level of human presence is playing a significant role as well. Regions with intensive man-made structures and human activities are historically common points of ignition (1). California's 2020 wildfire season was especially unprecedented, with 4,397,809 acres burned across 9,639 fires – more than double the previous record set in 2018. The rapidly increasing intensity of forest fires has made prediction a growing necessity.

The process of problem-solving is thoroughly intertwined with reflection and learning from the past. In order to create an effective crisis prevention system, it is imperative that we look to the past and analyze California's history of wildfires. Machine learning and predictive algorithms are practical manifestations of that idea of learning from the past. According to a comprehensive review of machine learning applications, there are six distinguishable categories in this realm of disaster/fire prediction approaches, the one most relevant to our work being "fire occurrence, susceptibility, and risk" (2). Within this category, the review lays out four major types: "fire occurrence prediction," "landscape-scale burned-area prediction," "fire-susceptibility mapping," and "landscape controls on fire." Our niche sits between the first two types in the list. Fire occurrence prediction methods attempt to either predict the number of fires in an area or make a more specific binary decision between "fire" and "no fire", providing relatively simple, quick information which can be used to dictate the allocation of preparatory resources. Landscape-scale burned-area prediction methods, the most recent and least explored type, attempt to obtain accurate predictions of the area burned by the fire. Our project aligns with area-burned-prediction conceptually but, in an effort to prioritize reaction time, uses an ideology more in line with fire occurrence prediction, aiming for quick and slightly more general information (severity ranges instead of exact values) for efficient usage and implementation.

The aim of this project is to create a machine learning algorithm that can provide significant aid in the process of mitigating forest fires in California. Our research utilizes machine learning to accurately predict the damage of a

DOI: https://doi.org/10.59720/22-163

```
def grabData(day, lonLat):
    frequency=1
    start_date = day
    end_date = day
    api_key = '4023ffda504d4cc286011902211208'
    location_list = [lonLat]

    hist_weather_data = retrieve_hist_data(api_key,
                                location_list,
                                start_date,
                                end_date,
                                frequency,
                                location_label = False,
                                export_csv = True,
                                store_df = True)
```

Figure 1: Python script to access weather data through an API. A simple API call allows us to access vast amounts of weather data at specific locations and times, spanning decades. Several of the 1600 wildfires from the Kaggle dataset were lacking longitude and latitude values, so ultimately only 1462 of those data points were useful.

```
for i in range(0, len(locList)):
    up = False
    num = 0
    list = []
    try:
        grabData(dateValues[i], locList[i])
        if int(timeValues[i][3:5]) >= 31:
            up = True
        with open('C:\\Users\\sidta\\Downloads\\WeatherData\\' + locList[i] + '.csv') as
            csv_reader = csv.reader(f)
            next(csv_reader)
            for row in csv_reader:
                temp = row[0][11:16]
                if up:
                    if int(temp[0:2]) == int(timeValues[i][0:2])+1:
                        num = int(temp[0:2])
                        list = row
                        break
                elif up == False:
                    if temp[0:2] == timeValues[i][0:2]:
                        num = int(temp[0:2])
                        list = row
                        break

        cell_list = sheet_instance.range('I' + str(i+2) + ":AG" + str(i+2))

        for cell in range(len(list)):
            cell_list[cell].value = list[cell]

        sheet_instance.update_cells(cell_list, value_input_option='USER_ENTERED')
```

Figure 2: Python script to modify data set. The script is able to access all of World Weather Online's information at the location and time of each fire.

new or potential fire by categorizing it into certain severity ranges. It also generates meaningful patterns/trends to help first responders develop the best way to fight a fire, and the best actions to prevent a potential one. By providing firefighters with all of this information, we may be able to increase the efficiency of their operation as they encounter and neutralize wildfires. Our model utilizes past incidents and meteorological data to predict severity by classifying fires into one of seven categories (A-G) based on acres burned (3). We also generated a correlation matrix to determine which specific meteorological factors had the highest effect on the spread of wildfires. The results of our algorithms indicate the viability of the use of machine learning to predict and manage forest fires, showing us that a machine learning program can, with notable accuracy, predict the severity (and existence) of a future wildfire, and identify the most important contributing factors to look out for. The importance of such an algorithm cannot be overstated; this project builds upon the existing field of research yielding very promising results, pointing towards a future where the damage done by wildfires is greatly reduced (4).

RESULTS

We tested three algorithms—logistic regression, support vector machine (SVM), and random forest regression—based on their accuracy in predicting the severity of wildfires based on a set of nine factors. We used a dataset from Kaggle consisting of all recorded California wildfires between 2013 and 2020 (3). We utilized the location information from this dataset to make API calls and obtain meteorological data at the time and place of every fire (**Figures 1 and 2**), creating a comprehensive dataset (**Figure 3**) to train our model with. The resulting data had entries for 1462 fires. The three models we used, chosen with a consultation to research in the field, were logistic regression, support vector machine, and random forest (RF) classifier (4). Logistic regression predicts a value for the dependent variable based on some number of independent variables. SVM creates a line to separate data points, classifying them based on which side they end up on. RF classifiers run an ensemble of decision trees, choosing the majority output as the correct classification. Logistic regression was chosen due to our focus on fire severity prediction (measured in acres burned), and SVM and RF classifiers were chosen for classification and detection. We defined accuracy by how many fires were placed in the correct category (A-G) by the model. We also tested a correlation matrix using the same set of factors to reveal which features impact results the most relative to each other.

Of the three models, the SVM performed the best with a classification accuracy of 55%. Logistic regression had an accuracy of 52%, and the random forest classifier had an accuracy of about 50%. Based on the correlation matrix, temperature, humidity, dew point, wind gust, and wind speed were the most important factors that determined the type of wildfire and how many acres a wildfire will burn once it starts (**Figure 4**). Temperature had a 0.12 correlation coefficient with the type of wildfire, humidity had a -0.24 correlation with the type of wildfire, dew point had a -0.15 correlation with the type of wildfire, wind gust had a 0.12 correlation with the type of wildfire, and wind speed had a 0.15 correlation with the type of wildfire (p = 0.239, 0.0173, 0.140, 0.239, 0.140, Pearson Correlation Coefficient; **Figure 4**).

DISCUSSION

The five factors that show a stronger association to wildfire severity are temperature, humidity, dew point, wind gust, and wind speeds according to the correlation matrix we generated (**Figure 4**). The columns for these four factors all have correlations with higher absolute values, demonstrating that wildfires spread the most under these factors. Our SVM model performed the best with a classification accuracy

DOI: https://doi.org/10.59720/22-163

AcresBurned	Counties	Type	Latitude	Longitude	Started	MaxtempC	MintempC	SunHour	UvIndex	DewPointC	FeelsLikeC
257314	Tuolumne		37.857	-120.086	2013-08-17T15:	32	14	11.6	7	2	29
38274	Los Angeles		34.585596	-118.423176	2013-05-30T15:	34	9	14.5	7	2	31
27531	Riverside		33.7696	-116.72885	2013-07-15T13:	32	18	14.1	7	6	28
27440	Placer		39.12	-120.65	2013-08-10T16:	26	6	11.6	6	5	21
22992	Fresno		37.279	-119.318	2013-07-22T22:	25	15	14.5	6	4	15
20292	Riverside		33.86157	-116.90427	2013-08-07T14:	27	15	10.4	6	-1	24
14754	Siskiyou		41.32	-123.176	2013-07-31T22:	24	1	11.6	5	0	1
12503	Humboldt		41.836	-123.488	2013-08-10T11:	21	4	10.5	4	13	21
11429	Tehama		40.04263	-121.85397	2013-08-23T14:	33	16	11.6	7	7	31
8873	Shasta		40.498332	-122.535496	2013-09-09T12:	37	17	11.6	7	2	35
7855	San Diego		32.95435	-116.47381	2013-07-06T12:	29	19	14.1	6	14	29
6565	Tehama		40.190662	-121.595555	2013-05-01T09:	16	1	11.6	4	-11	11
4346	Kern		34.7861	-118.9411	2013-05-15T12:	25	13	14	6	6	24
4346	Ventura		34.7861	-118.9411	2013-05-15T12:	25	13	14	6	6	24
3505	Sonoma		38.8167	-122.8498	2013-11-22T02:	21	7	8.7	4	-16	4
3166	Riverside		34.288877	-116.941311	2013-05-01T12:	24	11	13.8	6	-24	23
3111	Contra Costa		37.90757	-121.882582	2013-09-08T13:	34	19	11.6	7	10	32
2781	San Diego		33.12111	-116.62579	2013-05-23T12:	27	6	14	6	0	26
2462	Butte		39.446268	-121.38236	2013-08-16T16:	32	16	11.6	7	5	26
2236	San Diego		33.341	-117.3092	2013-10-05T12:	30	15	11.6	6	2	28
2060	Tulare		36.208	-118.444	2013-08-24T14:	30	12	11.6	6	-5	23
1904	Santa Barbara		34.55048	-119.82425	2013-05-27T14:	25	13	14.5	6	9	24

Figure 3: Resulting data after API calls to incorporate weather data. Sample of the resulting data table, which includes cohesive meteorological information for each forest fire and is used with our classification algorithms.

of 55%, while in a previous study predicting the scale of California wildfires, the top performing model only predicted the correct scale of an occurring fire 27.4% of the time (5). The SVM model had the highest accuracy level compared to the linear regression model because SVM has an optimization task to find the best margin which reduces the risk of error on the data. Meanwhile, linear regression is more vulnerable to overfitting. Overfitting is when a model begins predicting outputs *extremely* close to the actual values. When overfitting occurs, the model is simply learning how to predict the outputs for this specific training data rather than how to predict outputs in general. This overfitting causes the algorithm to perform poorly on the test data. When training the models, we used 80% of the custom dataset, while using the other 20% of the dataset to test the algorithm. The percentages of the dataset used for training and testing the algorithms affect the accuracy level, so if we use more relevant data to train the model, we can improve the accuracy and avoid overfitting.

Using this model, we are able to predict the size and severity of wildfires based on various factors, giving firefighters vital information on how to approach potentially deadly wildfires. The model can also provide insight into how influential these different factors are on the fire, enabling firefighters to identify the ideal preventative measures in any given situation and minimize the damage done to the environment and property. These measures can range vastly in scope and timing. There are steps to fight wildfires depending on how large the fires can potentially be. Firefighters control the spread of wildfires by removing one of these 3 elements necessary for fire: heat, oxygen, or fuel. Firefighters remove heat by using water or by distributing fire retardants. Fire retardants can be distributed on the ground through pumps or in the air through airplanes or helicopters. Firefighters remove fuel by removing vegetation using various techniques such as bulldozers or intentionally burning it before the wildfire reaches it. Lastly, firefighters remove oxygen by using a fog nozzle that produces fog to smother fires (6). Landowners and local authorities can also make an effort to control the amount of vegetation growing in an area using machines, grazing animals, or carefully managed fires. Fire breaks or control lines should be strategically placed to divide up the land to control the spread of wildfires (6). These can include roads, rivers, railway lines, or areas with no vegetation. A similar method is establishing fire lines, which are boundaries that contain no combustible material. Fire lines are constructed by removing combustible material from an area. After fire lines are established, the next step is mopping up, which is when firefighters look for places near the fire lines that are still hot and burning and put them out, and then check the burned area as well. The combination of such methods and reliable prediction algorithms is the path forward in the fight against California's troubling wildfire problem. The five major factors we identified (temperature, humidity, dew point, wind gust, and wind speed) are factors we can already predict in advance. Using that information and our model, firefighters will be able to prepare resources and mitigation methods accordingly. For example, a given combination of meteorological factors could cause a fire severe enough that fire blocking, control lines, and even minimized vegetation in the area will not efficiently contain it. This could lead to a false sense of security and insufficient allocation of resources. Our model would consolidate otherwise inconclusive information to give firefighters a better idea of what to expect. As of right now, firefighters could attempt to correlate the meteorological factors at play with the most efficient course of action, but this process can be improved upon. An expanded version of this model which includes the mitigation methods used for each fire would be even more efficient, so firefighters can adapt to each situation with extensive knowledge of what combination of firefighting methods and meteorological conditions will be optimal.

JOURNAL OF
EMERGING INVESTIGATORS
DOI: https://doi.org/10.59720/22-163

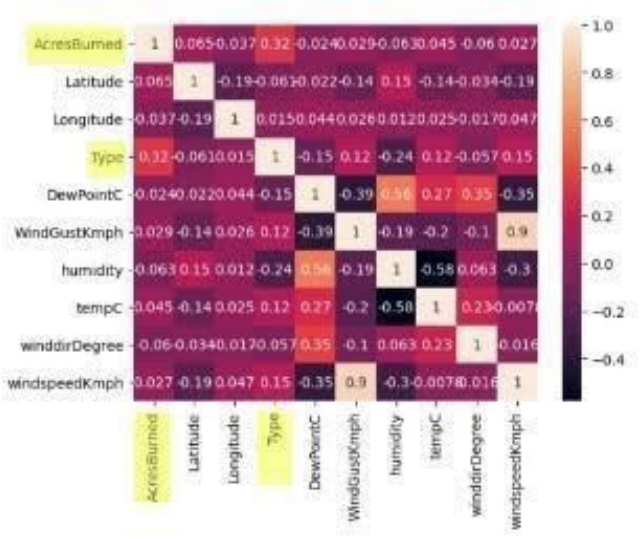

Figure 4: Type of wildfire is positively correlated with temperature and wind speeds, but negatively correlated with dewpoint and humidity. Correlation matrix of meteorological factors that contribute to a wildfire, generated by Matplotlib and Seaborn libraries using the Kaggle dataset. Type of wildfire is significantly correlated to temperature, humidity, dew point, wind gust, wind speed (p = 0.239, 0.0173, 0.140, 0.239, 0.140, Pearson Correlation Coefficient).

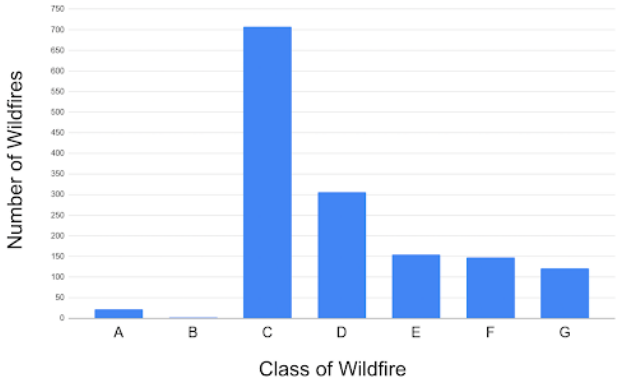

Figure 5: Almost half of all wildfires are class C. Frequency of wildfires of each size class, in acres, (A-G) in the Kaggle dataset. A < 0.25; 0.25 ≤ B < 10; 10 ≤ C < 100; 100 ≤ D < 300; 300 ≤ E < 1000; 1000 ≤ F < 5000; G ≥ 5000.

The conducted research has left us with an incredibly promising model that is capable of, with increased accuracy compared to some previous models, predicting the severity of forest fires and breaking down the contextual importance of different meteorological factors in this severity. However, our model does have some limitations. This model is trained with data for California wildfires specifically, so this exact algorithm could not be used for wildfires in other regions without modifications. Our three algorithms can be applied to better forecast wildfires and control the situation before too much destruction occurs, which would help firefighters use their resources more efficiently. Iterations of this concept, in tandem with the plethora of firefighting resources and strategies already at hand, will ultimately be the best way to mitigate and eliminate the threat of forest fires. This project can be built upon by including more meteorological factors to improve accuracy, such as vapor pressure, radiation, day length, and landscape data (i.e., soil). Further studies could also determine if there is a correlation between wildfire class and meteorological factors in the week leading up to the fire. These limitations and potential improvements to our SVM model aside, there is also the matter of algorithm choice; others have found neural networks to be the most promising direction to take wildfire severity prediction in (5). A neural network may in fact perform better than our SVM model, but the type and amount of data required would be a bit more difficult to acquire. For example, one could feed satellite imagery into a convolutional neural network and train it to recognize potential risks. The data collection process for that is more complicated but could yield interesting results. Nevertheless, an ANN, RNN, or CNN (artificial, recurrent, or convolutional neural network)-based classifier certainly has the potential to be more accurate.

MATERIALS AND METHODS

A Kaggle dataset containing information on 1600 California wildfires was utilized as the base of our data (3). This dataset alone was not sufficient to supply our model, as our research heavily focused on meteorological data. Using the World Weather Online API, a python script was devised to fill in the Kaggle dataset with hourly weather statistics for each fire (**Figures 1 and 2**). A change from daily to hourly data significantly boosted the accuracy of our model. The data which was ultimately used to train our models contained 1462 fires, due to some fires missing location data in the original dataset.

One more column was added to classify the fires by the number of acres burned (**Figure 5**). Class A had acres < 0.25, Class B had 0.25 ≤ acres < 10, Class C had 10 ≤ acres < 100, Class D had 100 ≤ acres < 300, Class E had 300 ≤ acres < 1000, Class F had 1000 ≤ acres < 5000, and Class G had acres ≥ 5000. The data was read through the Pandas library before the above conditions were used to add an extra column in the Pandas data frame of values from 1-7, mapping the letters A-G.

Finally, the classification models were ready for training. The three models chosen were logistic regression, SVM, and random forest classifier. This choice was made with a consultation on previous research on various machine learning applications to wildfire research (4). Logistic regression was chosen due to our focus on fire severity prediction (measured in acres burned), and SVM and RF classifiers were chosen for classification and detection.

The models were implemented with the Scikit learn machine learning library, trained with 80% of the data, and tested with 20%. A correlation matrix was also created using the Matplotlib and Seaborn libraries to better understand the most important meteorological factors that determine

DOI: https://doi.org/10.59720/22-163

how much a wildfire will spread. The complete list of factors used is Latitude, Longitude, Avg Temp, High Temp, Low Temp, Precipitation, Max Wind Speed, Sea Level, Pressure, Visibility, Lowest Humidity, Highest Humidity, Type, Avg Acres Burned, Drought Monitor Index, and Avg Dew Point.

Received: July 8, 2022
Accepted: December 3, 2023
Published: January 15, 2023

REFERENCES

1. Li, Shu, and Tirtha Banerjee. "Spatial and Temporal Pattern of Wildfires in California from 2000 to 2019." *Nature News*, Nature Publishing Group, 22 Apr. 2021, doi.org/10.1038/s41598-021-88131-9.
2. Piyush Jain, et al. "A Review of Machine Learning Applications in Wildfire Science and Management." *Environmental Reviews*, 28 July 2020, doi.org/10.1139/er-2020-0019.
3. Ares. "California Wildfires (2013-2020)." Kaggle, 9 Feb. 2020, www.kaggle.com/datasets/ananthu017/california-wildfire-incidents-20132020. Accessed 10, Aug. 2021.
4. Rodrigues, Marcos, and Juan De la Riva. "An insight into machine-learning algorithms to model human-caused wildfire occurrence." *Environmental Modelling & Software*, vol. 57, 2014, pp. 192–201, doi.org/10.1016/j.envsoft.2014.03.003.
5. Walters, Matthew. "Predicting the Likelihood and Scale of Wildfires in California using Meteorological and Vegetation Data." *ScholarWorks at University of Arkansas*, May 2022, scholarworks.uark.edu/etd/4521.
6. *Firefighting techniques to prevent the spread of wildfires*. WFCA. (2023, July 7). wfca.com/articles/prevent-the-spread-of-wildfires/. Accessed 10, Aug. 2021.