**California Blaze Predictor: Forecasting Wildfire Intensity with Environmental Insights**

By:

Iyabode Ogunbode, Yusuf Saadu, Sufia Nooreen, Jordan Bernot

University of North Texas

**Table of Contents**

[Abstract 2](#_Toc1561458770)

[Introduction 3](#_Toc1511128382)

[Stakeholders 4](#_Toc1262622556)

[Research Objectives 5](#_Toc603936321)

[Research Questions 5](#_Toc2062896671)

[Literature Review 6](#_Toc852081580)

[Relevant Theories on California’s Wildfire 6](#_Toc804486212)

[Other Relevant Research Studies 7](#_Toc1964648904)

[Relevant Organizations Countering Model Generalization Issue 9](#_Toc1354363189)

[Research Methods 10](#_Toc1294632621)

[Data Description 13](#_Toc1675789219)

[Data Analysis 18](#_Toc1628925365)

[1. Exploratory Data Analysis 19](#_Toc1494997287)

[1.1 Distribution Analysis 19](#_Toc1570323758)

[1.2 Outliers Analysis 21](#_Toc90841836)

[1.3 Correlation Analysis 21](#_Toc905179445)

[2. Assumptions Tests 23](#_Toc1969744455)

[2.1 Multicollinearity Test 23](#_Toc200762235)

[2.2 Normality Test - Shapiro-Wilk Test 24](#_Toc1814918608)

[2.3 Homoscedasticity Test - Breusch-Pagan Test 25](#_Toc53068905)

[2.4 Significance Test 25](#_Toc1721899953)

[3. Adding New Target ‘Intensity Level’ 26](#_Toc2005668613)

[3.1 Categories of Intensity Levels 27](#_Toc672984503)

[3.2 Distribution of the Intensity Levels 27](#_Toc184074843)

[4. Model Evaluation 28](#_Toc1164650891)

[4.1 Model Selection 28](#_Toc1452651627)

[4.2 Characteristics of the Selected XG Boost Model 29](#_Toc1373615206)

[Discussion of Results 29](#_Toc432246531)

[Conclusion 33](#_Toc2013853537)

[List of References 35](#_Toc1951827385)

**List of Tables and Figures**

[Table 1: Interpretation of Results of the Histogram Plots of Environmental Variables 20](#_Toc152078269)

[Table 2: Interpretation of Results of the Box Plots of the Environmental Variables 22](#_Toc152078270)

[Table 3: Interpretation of the Multicollinearity Test Results 24](#_Toc152078271)

[Table 4: Interpretation of the Shapiro-Wilk’s Test Results 25](#_Toc152078272)

[Table 5: Interpretation of Breusch-Pagan Test Results 26](#_Toc152078273)

[Table 6: Probability Thresholds for the Different Intensity Levels 28](#_Toc152078274)

[Table 7: Performance Metrics of the XG Boost Model 30](#_Toc152078275)

[Figure 1: Correlation Heat Map 16](#_Toc152077530)

[Figure 2: Histogram Distribution of Target = 1 in Station, Bishop 17](#_Toc152077531)

[Figure 3: Environmental Factors Time Series Color Coded by Wildfire Occurrence 18](#_Toc152077532)

[Figure 4: Histogram Plots of the Environmental Variables 20](#_Toc152077533)

[Figure 5: Correlation Heat Map After Data Cleaning 23](#_Toc152077534)

[Figure 6: Multicollinearity Test Result After Dropping Wind Run (miles) 25](#_Toc152077535)

[Figure 7: OLS Regression Results 27](#_Toc152077536)

[Figure 8: Generating the Predicted Probabilities after Applying Logistic Regression 28](#_Toc152077537)

[Figure 9: Summary Statistics of the Probabilities 28](#_Toc152077538)

[Figure 10: Distribution of Environmental Variables Categorized by Intensity Levels 29](#_Toc152077539)

[Figure 11: Model Comparisons between XG Boost, RF Classifier, and SVM 29](#_Toc152077540)

[Figure 12: ANOVA F-Value and Mutual Information Scores for Each Environmental Variable 31](#_Toc152077541)

Abstract

This research focuses on predicting wildfire intensity levels in California by leveraging 14 environmental factors. In addition to predicting intensity this research study also identifies the key environmental factors of wildfires. There is a growing concern about the worsening of wildfires, significantly impacting California residents, the local economy, businesses, firefighters, and governmental agencies. This study aims to develop a predictive model crucial for providing timely wildfire warnings that can be used to optimize resource allocation and enhance environmental preservation in response to climate change. The dataset, sourced from Kaggle, was originally collected from the California Irrigation Management Information System. Feature engineering was conducted to create a new variable named 'Intensity Level' using logistic regression-derived probabilities, subsequently categorized into 'Low', 'Medium', and 'High' intensity levels. The supervised learning algorithms Extreme Gradient Boost, Random Forest Classifier, and Support Vector Machine were explored as part of the analysis. XG Boost was eventually chosen for its robustness in handling diverse data types and exceptional overall performance compared to the other models. Additionally, Anova F-value and Mutual Information scores were employed, revealing that Precipitation, Average Air Temperature, Soil Temperature, and Humidity are the most influential factors affecting wildfire intensity. Our research is an extension of previous research on climate change and its impact on the occurrence of wildfires (Lukovic et al., 2021) and machine learning for wildfire prediction (Sayad et al., 2019). Our research model offers a more nuanced approach to predicting wildfires with the addition of wildfire intensity as an output variable.

**Keywords:** California Wildfires, Wildfire Intensity Prediction, Environmental Variables, XG Boost, SVM

Introduction

Over the years, wildfires have become more prevalent and intense globally, inflicting significant impacts on society, the environment, and the economy. The United States has seen 344 big wildfires in the last five years, which have resulted in significant financial losses as well as 180 fatal wildfires and 431 injuries (Bishop, 2023). The total burned acreage has quadrupled since the 1980s, primarily attributed to climate-induced factors such as fuel accumulation and an increase in fuel aridity (Burke et al., 2020). While human activities play a significant role, natural forces like lightning strikes, seasonal droughts, and other spontaneous ignition sources also contribute. Extreme temperatures and dry air in California show how wildfires are escalating due to climate change (Mulkern, 2023).

The state has witnessed the highest number of wildfires in the United States and continues to face recurrent and devastating fires that displace residents and cause extensive damage to properties and lives. According to CAL FIRE's wildfire data for 2023, 6,114 fires in California consumed 318,272 acres of land, which is like the numbers from 2022. But the wildfires of 2020 were even worse off, consuming an incredible 4.3 million acres of land and resulting in the deaths of 33 people due to nearly 8,600 fires. Although 2023 offers a brief window of relief, it is critical to recognize that California's current favorable conditions could change quickly. Looking ahead, a significant concern arises from a study indicating that dry mountainous forests in California and other Western states are poised to witness progressively worsening fires in the coming years (Elbein, 2021).

## Stakeholders

Government agencies shoulder the critical task of formulating robust emergency response strategies, ensuring efficient allocation of limited resources, and overseeing the safety of firefighters and first responders. Meanwhile, California residents and homeowners bear the immediate brunt, facing potential loss of property and life. The local economy and businesses also confront substantial financial losses. CAL FIRE reports an average of 7,366 annual fires leading to the damage and destruction of approximately 16,619 structures over the years from 2019 to 2023 (present). The collective dollar damage amounted to approximately $5.48 billion, which quantifies the monetary value of property and content loss, excluding firefighting expenses and indirect impacts like business interruption.

The results will empower California residents to make informed safety decisions, aid government agencies in effective emergency response planning and policymaking and assist businesses in preparing for the economic impact of fires. The model will significantly benefit firefighters and first responders, enhancing their ability to allocate resources effectively. Predicting wildfire intensity levels in risk areas enables timely and targeted deployment, potentially saving lives and reducing damage. It will contribute to a safer and more resilient environment and enhance recovery from wildfires.

## Research Objectives

This research aims to create a model for predicting wildfire intensity, enabling efficient resource allocation, and minimizing wildfire damage. Overall, the project “California Blaze Prediction” primarily focuses on aiding communities and authorities in California in their battle against wildfires. Through this research, we will be able to build a model that will help achieve the following research objectives:

* Understand environmental factors that most influence wildfire behavior and occurrences.
* Enable prediction of wildfire intensity levels in high-risk regions based on environmental variables.
* Facilitate the strategic allocation of resources for firefighting efforts and emergency response, optimizing efficiency and minimizing response times.
* Enhance safety efforts, aiming to protect both residents and firefighters from wildfires, saving lives.
* Promote wildfire preparedness and curtail the extent of damage inflicted on the structures, natural habitats, and local economies.

## Research Questions

1. What is the relationship between environmental factors and wildfire intensity and frequency? Which environmental factors affect wildfire occurrences more?
2. How can a data-driven approach to risk assessment assist in the mitigation of wildfire emergencies in California?
3. How do we measure or certify the accuracy of the model? What is the best evaluation metric for success?

# Literature Review

As of today, wildfires are still a persistent issue in the western United States, particularly in California. Wildfires started becoming a major issue in California in 2000. In 2000, nearly 300,000 acres were burnt, about $140 million was spent on recovery, and 130 properties were destroyed (California Department of Forestry and Fire Protection, 2001). Historically, wildfires have caused a lot of direct and indirect effects on California’s ecosystem and its population. An example is the August 2020 Complex Fire, which was caused by lightning strikes and burned over a million acres and is hard to predict by experts (California Department of Forestry and Fire Protection, 2020). That is why our project will be crucial to predicting wildfire intensity levels and further examining weather factors that may influence fire behavior while also overcoming the present challenges. Since wildfires are also known to be caused by spontaneous and unpredictable events such as arson, fireworks, lightning strikes, etc., this has presented challenges in building an accurate predictive model. Despite the challenges, some opportunities identified would present us with a better way of approaching our project. Reviewing previous related studies would help us identify both key biases and limitations and explore new opportunities to consider while building our predictive model for wildfire intensity levels in California regions.

## Relevant Theories on California’s Wildfire

Climate change is currently a trending topic all over the globe. Some believe it’s real while others don’t due to misinformation, fear, and perception. The increase in wildfire occurrence in California has been supported by climate change theories, causing increased temperatures and varying precipitation patterns that create environmental conditions more conducive to wildfires occurring (Wibberly, 2021). Lukovic et al. emphasize how the timing of the onset of the rainy season in California has grave consequences for wildfire risk levels. Over the past years, California’s rainy season has been noted to shift, attributable to climate change. In the study, Lukovic et al. noticed that the rainy season in California has been shifting since the 1960s, leading to shorter and sharper rainy seasons. The researchers examined the rainy season by analyzing precipitation data collected from 407 stations from 1960-2019. Through their analysis, they determined a statistically significant decrease in precipitation during the Fall season, especially in November. The rainy season is shown to start later, typically shifted by 27 days from the 1960s to the 2010s. This shift is caused by the continuous summer circulation patterns extending into the month of November. With the increased summer season, the theory surrounding the environmental lapse rate can further explain how rising temperatures can contribute to abnormal shifts in the rainy season, thus increasing wildfire risk (Wikipedia Contributors, 2019). The lapse rate could also explain how fire can spread through heat transfer and convection. In conclusion, the relationship between the decreased rainfall and increased dry weather conditions would increase the risks for California’s wildfire season. Thus, it demonstrates how decreased rainfall precipitation due to climate change is consistent with the increasing wildfire seasons in California.

Although Lukovic et al. study provided insights concerning wildfire risks in California, there are some potential biases and limitations observed. The researchers heavily relied on the precipitation data obtained from the 407 stations. These data from the different stations may have contained varying spatial coverage and inconsistencies over the time periods, from the 1960s-2019. This could potentially raise doubt about the reliability of the precipitation data collected. Another bias would be the criteria used to define the onset of the rainfall precipitation. This may introduce subjectivity into the research as it may not have captured the gradual transition into the different seasons.

## Other Relevant Research Studies

In the Sayad et al study, a wildfire occurrence predictive model was built using machine learning models and remote sensing data. The researchers were successfully able to collect their data in Canada using the MODIS (Moderate Resolution Imaging Spectroradiometer) satellite instrument. The data contained information regarding normalized difference vegetation index (NVDI), land surface temperature (LST), thermal anomalies, and other wildfire indicators. The researchers were able to obtain these data in Canadian regions, particularly in British Columbia and Quebec from the years 2013-2014. They created a labeled dataset of fire and no fire instances. The researchers then evaluated different models by testing and training them, with 70% of the data used for training and the remaining percentage for testing. Some of the models evaluated included neural networks, support vector machines, and random forests for predicting wildfire occurrence. They were able to obtain a prediction accuracy of 98.32% for artificial neural networks and 97.48% for support vector machines. The model was then validated using methods such as classification metrics, regularization, and cross-validation. To evaluate the performance of the model, other systems such as STIFF (spatiotemporal forecasting framework), ISTFF (integrated spatiotemporal forecasting framework), and ARIMA (autoregressive integrated moving average) were used. In conclusion, the study emphasizes the importance of using machine learning models, remote sensing data, and model comparison in accurately predicting wildfire occurrence. A limitation identified in this study is that only wildfire occurrence was predicted, and not the intensity or other fire spread behavior. This would potentially limit the application of this model beyond wildfire occurrence and cannot be used when devising contingency plans after a wildfire has occurred. Because wildfires could occur spontaneously, further studying the fire spread behavior or intensity levels would present opportunities for better wildfire management in Canada and other regions.

Liang et al study focused on building a predictive model of forest wildfire scale at ignition in Alberta, Canada during the preliminary stages. The data was collected from the Canada National Fire Database (CNFDB) which contained information about the size of the burned area, fire duration, and other meteorological conditions to predict the wildfire scale between the years 1990 and 2018. The data was pre-processed using multicollinearity testing and feature normalization, then it was split into the training and testing datasets. Different models such as recurrent neural network (RNN), backpropagation neural network (BPNN), and long short-term memory (LTSM) were evaluated. The researchers selected the long short-term memory model because it had the highest accuracy score of 90.9%. Their study emphasized the relationship between different meteorological data and the wildfire scale in Alberta, Canada. The study was able to demonstrate the significance of neural networks in predicting wildfire scale at ignition. In conclusion, the study provides valuable insights for more efficient wildfire risk management and contingent plans. Although the research presents a step forward in producing an accurate predictive model for wildfire scale, the researchers acknowledged some limitations. The data was collected from only Alberta, Canada. This could potentially limit the generalization of the model beyond Alberta, thus not accounting for varying environmental factors and geography. Thus, it is possible that the predictive model cannot be used in other regions with different environmental characteristics and topography.

In our research, the environmental data was only collected from California regions. This could potentially introduce challenges during our analysis when we are building models across different regions. These different California regions could have diverse environmental conditions and flora that may influence wildfire risk levels. Testing multiple models could be used in improving model portability for California.

## Relevant Organizations Countering Model Generalization Issue

The California Department of Forestry and Fire Protection (CAL FIRE) has improved its wildfire risk management plan to mitigate wildfires in California. They utilize data-integrated analytical tools that provide geospatial data about the wildfires’ spread and environmental risk factors. Their integrated tool can provide real-time data visualization of wildfires that can be used for future forecasting. Real-time data will provide different training data for models across various locations and fuel types in California. This would aid in increasing the generalizability of models in other locations. Also, CAL FIRE contains historical wildfire data that could be used for testing and training models for future purposes. These advanced tools enable effective communication and smooth decision-making processes. CAL FIRE plans on implementing drones that would further support disaster recovery processes in California (ESRI Blog, 2021).

California Governor’s Office of Emergency Services (Cal OES) has implemented a tool called Next-Gen SCOUT which would be used to ensure effective communication and coordination for firefighters and other first responders during wildfire season. The Next-Gen SCOUT is integrated with the Fire Integrated Real-Time Intelligence System Program (FIRIS) which provides real-time data of wildfire events. This tool would have the capacity to train new wildfire data and keep the models adaptable to changing conditions. In turn, it should promote model generalization. In summary, this tool would enhance situational awareness for first responders which would support a faster and safer fire operation (Cal OES News, 2023).

# Research Methods

To sufficiently analyze our data we chose three distinct, but similar methods given the structure and variables available in our dataset. One is the XG Boost classifier and the others are the random forest classifier and support vector machine (SVM). The selected methods are supervised learning algorithms. Due to both algorithms being supervised, it requires a labeled or characterized dataset, one which has both the input and outcome variables correctly identified. Supervised algorithms learn by trying to understand the parameters of each input that are associated with a particular output, in our case wildfires. It is important to train supervised algorithms with accurate data and to determine the accuracy of our algorithm training, we deploy precision, recall, and F1 score to evaluate and compare our two methods.

SVM is a supervised machine learning method typically used for classification, but it can commonly be applied to data analysis requiring a regression approach (Vapnik, 1998). While SVM is commonly used in situations where the DV is continuous it is effective with data that requires a separation of the data into various classes. The concept of kernels is a significant factor that improves SVM over other supervised learning algorithms. The inclusion of the kernel and what is called a global optimum improves its performance over other algorithms that may converge on a local optimum (Vapnik, 1995). While computationally intense SVM is an ideal choice for our analysis because the goal of the SVM algorithm is to delineate an optimal separation between the various categories of data, determining a hyperplane on the decision boundary. This requires drawing a line that optimized the distance of data points between the two defined groups. Provided our dataset, SVM will be an effective tool to determine the elements of wildfires.

Extreme Gradient Boosting (XG Boost) is a supervised learning method that uses a gradient boosting framework. XG Boost is a decision-tree ensemble model used for complex data where traditional machine learning models may fail (Chen and Guestrin, 2016). This model is capable of handling diverse data structures and large datasets and can prevent the problem of overfitting (Natekin and Knoll, 2013). It's a scalable and accurate implementation of gradient boosting machines, which are ensemble learning models known for their predictive power. It is a model that is mostly used for classification purposes though it can also be used for tasks like regressions, forecasting, feature importance, and anomaly detection. Due to its regularization techniques it tends to be robust to overfitting making it a compelling model selection for our wildfire intensity level prediction project. It stands out in scenarios where the dependent variable involves categorizing instances into distinct classes. One of the method’s key benefits is in its use of a gradient boosting framework, which basically iteratively corrects the mistakes of previous models and adds them to the ensemble, ensuring that the errors decrease over time. Moreover, regularization terms are incorporated in its objective function to avoid overfitting, enhancing model generalization compared to traditional boosting algorithms.

For our wildfire intensity prediction project, XGBoost was selected due to its robustness in handling varied data types and its effectiveness in classification problems. The algorithm's ability to manage imbalanced datasets common in wildfire occurrences makes it particularly suitable for this task. XGBoost's feature importance metric offers valuable insights into which factors most significantly influence wildfire intensity. Given these characteristics, XGBoost serves as an ideal tool to discern the intricate patterns within our wildfire dataset, classifying each instance of environmental factors into a predicted risk level with high accuracy and interpretability.

Random forest is also another useful method to classify data as it is an extension of decision tree classifier. Decision tree classifiers separate data based on specific features and use the separation of each feature to make predictions about the DV. While decision trees are effective, they can suffer from overfitting, are sensitive to variations in the training dataset, and have several issues during implementation (Kotsiantis, 2013). Random forests overcome the problem of overfitting and apply multiple machine learning models to give it robustness which provides more accurate predictions. Random forest uses decision tree elements to generate a single model and, in this way, random forest attempts to reduce the correlation of each tree through random iterations. Random forest also implements random subsets of data to reduce data correlation. Importantly, random forests can identify the most important variables in a classification which is useful for our purposes to better understand each IV of wildfires. Like XG Boost and SVM, random forests are robust and not sensitive to outliers or noisy data. These approaches were selected for our dataset because they have some key differences that offer alternative analysis for our wildfire dataset. Through this diversity we discover the most accurate and useful model.

# Data Description

The dataset utilized for this wildfire intensity prediction project consists of 18 variables in total. Only 14 of them which are numerical variables are to be used as independent variables to build the predictive model. Of the original 18 variables, we are only using 14 of the numerical variables as independent variables while synthesizing a new categorical variable for the dependent variable that displays the intensity and risk level of the wildfire. The 4 categorical and 1 temporal variable are not being used to build the predictive model because we found them irrelevant to our goal and thus deemed them useless. They are listed and described as follows:

* ***Stn ID (Station ID)***: The ID of the station where the data was recorded
* ***Stn Name (Station Name)****:* The name of the station that recorded the data
* ***CIMIS Region***: California Irrigation Management Information System region where the station is located.
* ***Date***: Date that the readings were recorded
* ***ETo (in) (Evapotranspiration)*:** measure of the water vapor lost from the ground due to evaporation and from plants due to transpiration. The lower it is the less likely wildfire would intensify.
* ***Precip (in) (Precipitation/Rainfall)***: Amount of rain received in a specific area. The higher it is the less likely wildfire will intensify.
* ***Sol Rad (Ly/day) (Solar Radiation)***: Amount of solar energy received per day. The higher it is, the higher the risk and potential intensity of wildfire.
* ***Avg Vap Pres (mBars) (Average Vapor Pressure)***: measure of the pressure exerted by water vapor in the atmosphere. The lower it is, the drier it is which would mean the more likely and the higher the intensity of the wildfire.
* ***Max Air Temp (F) (Maximum Air Temperature)***: This represents the highest air temperature. The higher it is, the higher the chances for increased wildfire intensity.
* ***Min Air Temp (F) (Minimum Air Temperature)***: This represents the lowest air temperature. The higher it is, the higher the chances for increased wildfire intensity.
* ***Avg Air Temp (F) (Average Air Temperature)***: This represents the average air temperature. The higher it is, the higher the chances for increased wildfire intensity.
* ***Max Rel Hum (%) (Maximum Relative Humidity)***: the highest recorded amount of moisture in the air compared to the maximum amount of moisture the air could hold at the same temperature. The lower it is the higher the likelihood of increased wildfire intensity.
* ***Min Rel Hum (%) (Minimum Relative Humidity)***: the lowest recorded amount of moisture in the air compared to the maximum amount of moisture the air could hold at the same temperature. The lower it is the higher the likelihood of increased wildfire intensity.
* ***Avg Rel Hum (%) (Average Relative Humidity)***: the mean recorded amount of moisture in the air compared to the maximum amount of moisture the air could hold at the same temperature. The lower it is the higher the likelihood of increased wildfire intensity.
* ***Dew Point (F), (Dew Point Temperature)***: The temperature at which air becomes saturated and water vapor condenses into dew. The lower it is, the drier the air likely is which would mean that there is a higher likelihood of increased wildfire intensity.
* ***Avg Wind Speed (mph) (Average Wind Speed)***: the average speed of the wind. The higher it is the more likely wildfire intensity will increase as the wildfire would spread faster.
* ***Wind Run (miles) (Wind Run)***: the total distance the wind has traveled in a specific period. The larger this is the more likely the wildfire risk level or intensity would increase as the wildfire could travel or spread farther.
* ***Avg Soil Temp (F) (Average Soil Temperature)***: the average temperature of the soil. The higher it is the more likely things are to dry out which would lead to higher wildfire intensity and risk level.

When it comes to the dependent variable the dataset has the 19th variable **Target** which is a binary variable indicating whether a wildfire occurred on that date or instance with 1 indicating a wildfire and 0 indicating otherwise. However, that is not the focus of our project. We aim to predict the wildfire intensity of an already occurring wildfire based on the environmental conditions. Thus, we decided to fashion a categorical variable based on probability called ***Intensity Level*** that consists of three brackets, Low, Medium, and High. The previously mentioned and explained environmental conditions independent variables will be used to predict this Risk level variable. Standouts of interest based on our hypothesis of which environmental conditions would play a significant role in predicting potential wildfire intensity or risk level include Precipitation, Air Temperature, Relative Humidity, Wind speed, and Wind run.

We started our analysis of the data focusing on getting a summary for the data and checking for null or missing values. We dropped the column Stn ID and found our dataset to have 128,125 rows. There were also a couple of columns with missing or null values in the dataset of differing degrees. A summary found the mean, min, max, standard deviation, 25, 50, and 75 percentiles for each numerical variable column. It was there we noticed that some variables had a minimum of zero which would not make sense utilizing domain knowledge. We’re currently working on handling those null values that do not make sense. We plan to drop the null values of the identified records because they represent an exceedingly small proportion of our dataset. After that came correlation analysis which we hoped to use to help us identify any highly correlated independent variables.

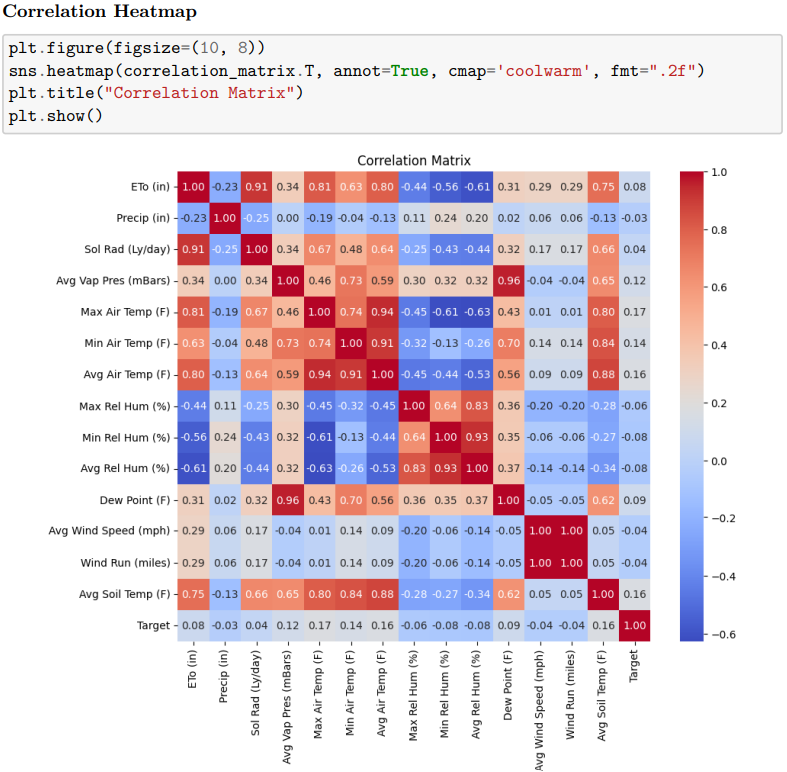


Figure 1: Correlation Heat Map

As can be seen above there are a few variables in the dataset that are highly correlated as we went with the basis of anything over 0.8 is highly correlated. The decision of which correlated variables to drop from the model has yet to be made. Moving on, we also ran a feature importance ranking on our independent variables to decipher which ones would be more relevant in making prediction on the Target variable, whether a wildfire happened or not. Another one will be running once we compute the new dependent variable. We have identified some outliers in our data and are currently working on data transformation methods to mitigate their effect on the model while noting the shape of their histogram distribution and utilizing box plots.

The next step of our analysis included deciphering the historical trend of each of the variables and overlaying it with whether the wildfire occurred or not to see what conditions were consistently present when a wildfire occurred. When carrying this out rather than looking at California as a whole for the readings, we decided to drill down to specific cities to see how the environmental conditions played a part in their wildfire occurrences. The city of Bishop, California was chosen because they had the most wildfire occurrences in the dataset.

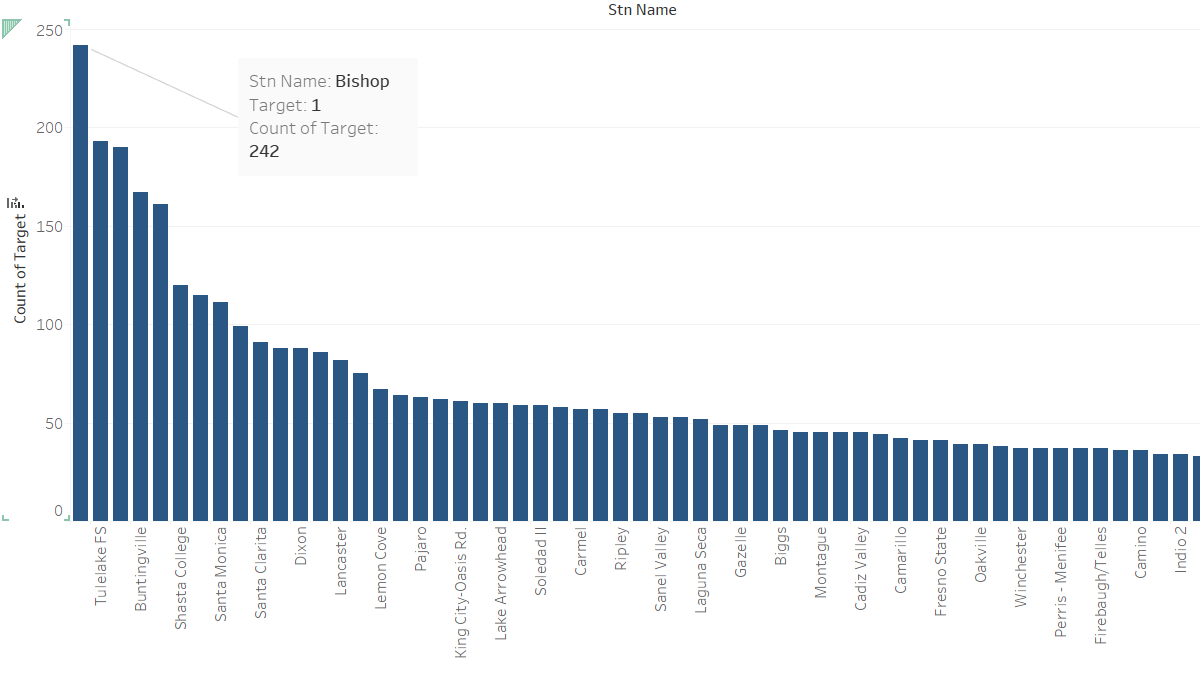


Figure 2: Histogram Distribution of Target = 1 in Station, Bishop

The selected variables of focus for now are Wind Speed, Wind run, Precipitation, Air Temperature and Relative Humidity. We feel these environmental factors play a key part in predicting wildfire occurrence and intensity. By looking at the historical data consisting of environmental readings captured by the bishop station, overlaying it with a color tint of when a wildfire occurred and not, we hope it will send us in the right direction of what variables to focus on. By analyzing the data this way, we hope to catch and handle any nulls or outliers, decipher which variables are superfluous or leave us at risk for multicollinearity, and finally which variables will be important in the predictive model especially as it pertains to wildfires and their intensity.

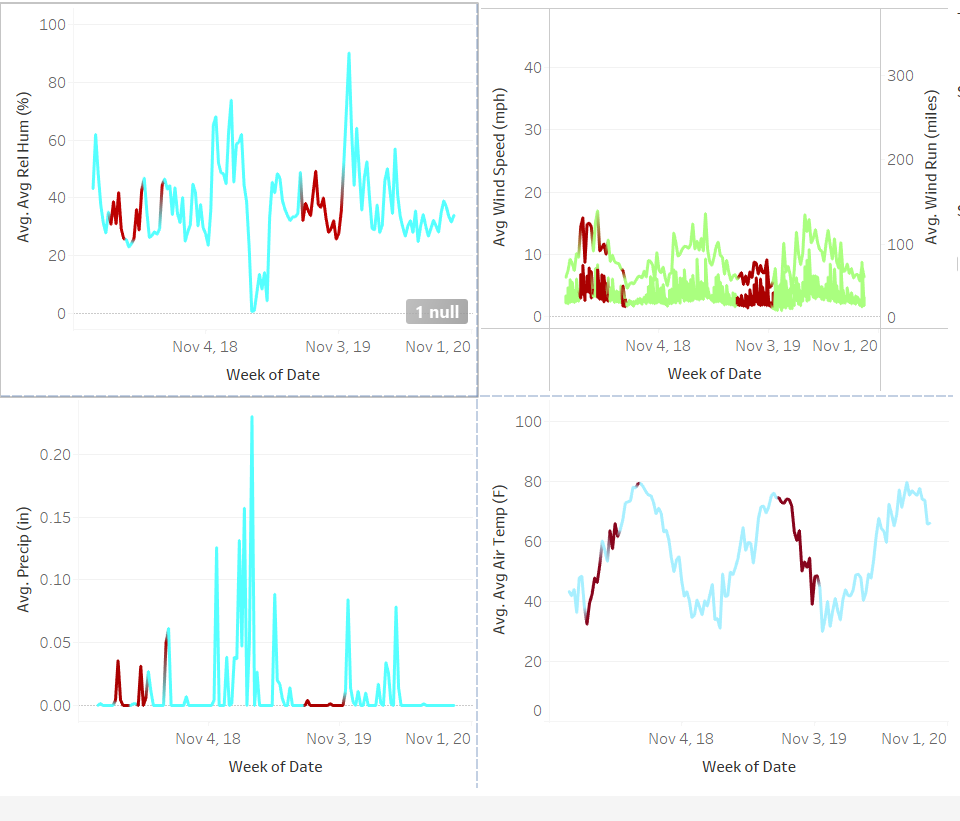


Figure 3: Environmental Factors Time Series Distribution Color Coded by Wildfire Occurrence

# **Data Analysis**

The purpose of analyzing the data is to explore the relationships that exist between the independent variables and the dependent variable, visualize the distribution of the data, examine the correlation between the independent variables, and make assumptions such as homoscedasticity, normality, significance, and multicollinearity tests. These tests and analyses would help prepare our data for further analysis and make predictions on California’s wildfire intensity levels that would help in making data-driven decisions when it comes to resource allocation and proactive measures. Our data has different environmental conditions that could potentially influence the wildfire intensity levels.

First, we prepared the data by performing preprocessing steps that would make it suitable for building a suitable model. These preparations included imputing missing values, handling extreme values, data exploration, encoding categorical variables, dropping columns that were significant to determining wildfire intensity, and comparing the different models based on their performance.

## 1. Exploratory Data Analysis

### *1.1 Distribution Analysis*

A group of graphs showing different sizes of data

Description automatically generated with medium confidence

Figure 4: Histogram Plots of the Environmental Variables

Table 1: Interpretation of Results of the Histogram Plots of Environmental Variables

|  |  |  |
| --- | --- | --- |
| **Features** | **Distribution** | **Interpretation of Results** |
| ETo (in) | Right skewed | Most days with lower evapotranspiration. |
| Precip (in) | Right skewed | Most days fall in the dry season, few days fall in the wet season. ***Expected*** results might be attributable to climate change. |
| Sol Rad (Ly/day) | Bimodal | Seasonal variation (wet and dry) in solar radiation. ***Expected*** results. |
| Avg Vap Pres (mBars) | Normal, slightly right skewed | Seasonal variation (wet and dry) in vapor pressure. ***Expected*** results. |
| Avg Wind Speed (mph) | Right skewed | Lower wind speeds are common and high wind speeds are uncommon. ***Expected*** results since the higher wind speeds might be attributable only to severe weather. |
| Avg Air Temp (°F) | Normal | Seasonal variation (wet and dry) in air temperature. ***Expected*** results. |
| Wind Run (miles) | Right skewed | Lower wind runs are common and high wind runs are uncommon. ***Expected*** results since the higher wind runs might be attributable only to severe weather. |
| Avg Rel Hum (%) | Right skewed | Most days fall in the dry season, few days fall in wet season. ***Expected*** results might be attributable to climate change. |
| Dew Point (°F) | Normal | ***Expected*** results because it indicates consistency in the dew point on most days. |
| Avg Soil Temp (°F) | Bimodal | Seasonal variation (wet and dry) in soil temperature. ***Expected*** results. |

### *1.2 Outliers Analysis*

Table 2: Interpretation of Results of the Box Plots of the Environmental Variables

|  |  |
| --- | --- |
| **Features** | **Observations** |
| ETo (in) | Presence of a few outliers. This is ***expected***. |
| Precip (in) | The presence of extreme outliers indicates days with higher precipitation. This is ***expected*** because California has a Mediterranean climate. |
| Sol Rad (Ly/day) | The presence of outliers on both ends. This is ***expected*** because it indicates days with higher and lower solar radiation. California has only wet and dry seasons. |
| Avg Vap Pres (mBars) | Presence of few outliers. This is ***expected.*** |
| Avg Wind Speed (mph) | The presence of a lot of outliers on the higher end. This is ***expected*** because the wind is naturally variable. |
| Avg Air Temp (°F) | The presence of outliers on both ends. This is ***expected*** because it possibly indicates days with extreme temperatures. Findings consistent with wildfire events between the years 2018-2020. |
| Wind Run (miles) | The presence of a lot of outliers on the higher end. This is ***expected*** because the wind is naturally variable. |
| Avg Rel Hum (%) | This data is highly skewed. This is ***unsettling*** because there are a lot of extreme values which may indicate abnormal weather patterns. |
| Dew Point (°F) | The presence of outliers on the higher end. This is ***expected*** because it indicates that some days (likely due to the wet season) had higher air moisture. |
| Avg Soil Temp (°F) | The presence of outliers on both ends. This is ***expected*** because it possibly indicates days with extreme temperatures. Findings consistent with wildfire events between the years 2018-2020. |

### *1.3 Correlation Analysis*

A screenshot of a graph

Description automatically generated

Figure 5: Correlation Heat Map After Data Cleaning

The correlation heat map visually provides the correlation between the environmental conditions and ‘Target’. As key observation is the presence of a strong correlation between Eto (in) and Sol Rad (Ly/day). This strong and positive correlation is expected since the condition, solar radiation is a major factor in evapotranspiration. There is also a strong correlation between soil and air temperatures, which is expected since air temperature is expected to influence soil temperature.

The weak correlations between the individual environmental conditions and ‘Target’ seems unsettling. We expected at least one of the variables to have a very strong correlation with ‘Target’. However, this might suggest the environmental variables may have a compound effect on wildfires. Thus, it is likely that there is an interplay between the independent variables to result in the wildfire occurrence.

## 2. Assumptions Tests

### *2.1 Multicollinearity Test*

Table 3: Interpretation of the Multicollinearity Test Results

|  |  |  |
| --- | --- | --- |
| **Features** | **VIF Score** | **Interpretation of Results** |
| ETo (in) | 27.35 | Moderate VIF score. A level of correlation is ***expected*** with other environmental variables. |
| Precip (in) | 1.12 | Low VIF score. This is ***expected*** for a distinct environmental variable. |
| Sol Rad (Ly/day) | 13.82 | Moderate VIF score. A level of correlation is ***expected*** with other environmental variables. |
| Avg Vap Pres (mBars) | 16.37 | Moderate VIF score. A level of correlation is ***expected*** with other environmental variables. |
| Avg Wind Speed (mph) | 4937.79 | A very high VIF score. This is ***unsettling*** because it could influence our results. |
| Avg Air Temp (°F) | 32.01 | Moderate VIF score. A level of correlation is ***expected*** with other environmental variables. |
| Wind Run (miles) | 4938.44 | A very high VIF score. This is ***unsettling*** because it could influence our results. |
| Avg Rel Hum (%) | 21.03 | Moderate VIF score. A level of correlation is ***expected*** with other environmental variables. |
| Dew Point (°F) | 34.42 | Moderate VIF score. A level of correlation is ***expected*** with other environmental variables. |
| Avg Soil Temp (°F) | 5.50 | Slightly elevated VIF score. A level of correlation is ***expected*** with other environmental variables. |

Wind Run (miles) had the highest VIF score, thus the column was dropped to improve the results. Here are the multicollinearity test results after dropping Wind Run (miles):

A screenshot of a computer

Description automatically generated

Figure 6: Multicollinearity Test Result After Dropping Wind Run (miles)

There is a significant improvement as there is no longer any extremely high multicollinearity. The extent of the reduction in the VIF score for Avg Wind Speed was unexpected.

### *2.2 Normality Test - Shapiro-Wilk Test*

Table 4: Interpretation of the Shapiro-Wilk’s Test Results

|  |  |  |
| --- | --- | --- |
| **Feature** | **p-value** | **Assumption Met?** |
| ETo (in) | < 0.001 | No |
| Precip (in) | < 0.001 | No |
| Sol Rad (Ly/day) | < 0.001 | No |
| Avg Vap Pres (mBars) | < 0.001 | No |
| Avg Wind Speed (mph) | < 0.001 | No |
| Avg Air Temp (°F) | < 0.001 | No |
| Avg Rel Hum (%) | < 0.001 | No |
| Dew Point (°F) | < 0.001 | No |
| Avg Soil Temp (°F) | < 0.001 | No |

***Interpretation of Results***:

In the result, all environmental features have p-values less than 0.05. This demonstrates that all the environmental data do not follow a normal distribution. Shapiro-Wilk's test and all other normality tests are quite sensitive to large datasets. This might be a reason why all features were indicated as not normally distributed. Another plausible explanation is the environmental variables may not exhibit a normal distribution because of the natural variability and/or extreme weather conditions. However, some researchers have observed non-normal distributions frequently in environmental data (Andersson, 2021). Thus, it is expected based on other studies for environmental data to not exhibit a normal distribution.

### *2.3 Homoscedasticity Test - Breusch-Pagan Test*

Table 5: Interpretation of Breusch-Pagan Test Results

|  |  |  |
| --- | --- | --- |
| **Feature** | **p-value** | **Assumption Met?** |
| ETo (in) | < 0.001 | No |
| Precip (in) | < 0.001 | No |
| Sol Rad (Ly/day) | < 0.001 | No |
| Avg Vap Pres (mBars) | < 0.001 | No |
| Avg Wind Speed (mph) | < 0.001 | No |
| Avg Air Temp (°F) | < 0.001 | No |
| Avg Rel Hum (%) | < 0.001 | No |
| Dew Point (°F) | < 0.001 | No |
| Avg Soil Temp (°F) | < 0.001 | No |

***Interpretation of Results****:*

In the result, the p-values for all the environmental variables were 0.0 indicating the presence of heteroscedasticity. This indicated that the residuals do not have a constant variance across the data. The presence of heteroscedasticity in environmental data is not unexpected due to natural variability and other factors. Some researchers explained that the presence of heteroscedasticity in environmental data is not uncommon because of its complex nature and variability (Dutilleul and Carriere, 1998).

### *2.4 Significance Test*

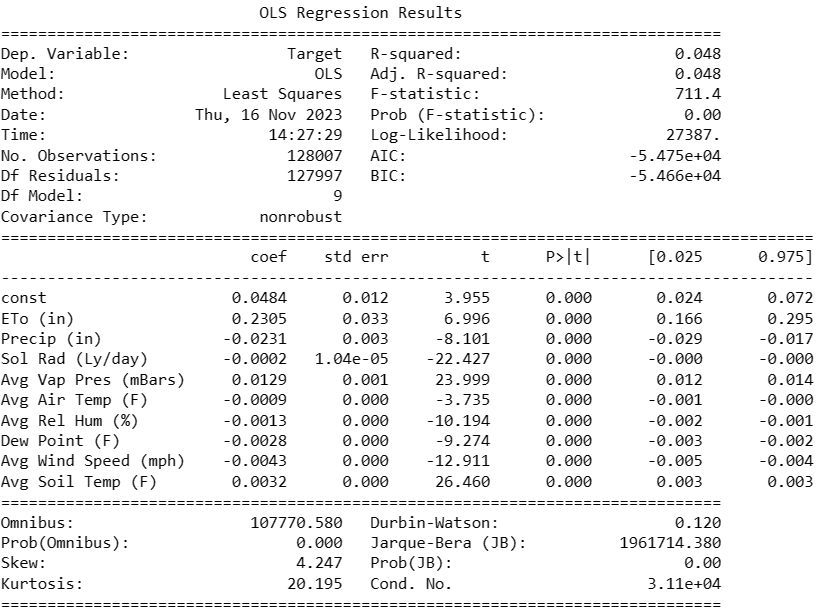


Figure 7: OLS Regression Results

***Interpretation of Results***:

In the results, the p-values for all the environmental variables were recorded as 0.000. This indicates that all the environmental variables are significantly related to the target variable. This is expected since environmental conditions can influence the occurrence of wildfires.

## Adding New Target ‘Intensity Level’

Probabilities used to define the new target variable; ‘Intensity Level’ were because of the logistic regression. The probabilities were used to classify each instance into the categories: ‘Low’, ‘Medium’, and ‘High’. The categories would assist in identifying the likelihood of wildfire events, which would be essential for resource allocation and risk management.

**A screenshot of a computer code

Description automatically generated**

Figure 8: Generating the Predicted Probabilities after Applying Logistic Regression

### *Categories of Intensity Levels*

A screenshot of a computer

Description automatically generated

Figure 9: Summary Statistics of the Probabilities

Table 6: Probability Thresholds for the Different Intensity Levels

|  |  |
| --- | --- |
| **Intensity Level** | **Probability Threshold** |
| Low | 0.000 – 0.013 |
| Medium | > 0.013 – 0.055 |
| High | > 0.055 – 0.844 |

### *3.2 Distribution of the Intensity Levels*

A collage of graphs

Description automatically generated

Figure 10: Distribution of Environmental Variables Categorized by Intensity Levels

## Model Evaluation

**XG Boost Classifier**  **Random Forest Classifier** **Support Vector Machine**

A screenshot of a report

Description automatically generated A white paper with black text

Description automatically generated A screenshot of a report

Description automatically generated

Figure 11: Model Comparisons between XG Boost, RF Classifier, and SVM

### *4.1 Model Selection*

After a thorough evaluation of the different models, we decided to select the XG Boost classifier model. This is because the model demonstrated the best performance metrics out of the 3 models. The model is capable of making predictions for the wildfire intensity levels in California based on the environmental data. This was expected because the model can handle complex and heterogeneous data.

### *4.2 Characteristics of the Selected XG Boost Model*

Table 7: Performance Metrics of the XG Boost Model

|  |  |  |
| --- | --- | --- |
| **Intensity Levels** | **Performance Metrics** | **Value** |
| ***Low*** | Precision | 0.98 |
| Recall | 0.96 |
| F1-score | 0.97 |
| ***Medium*** | Precision | 0.97 |
| Recall | 0.98 |
| F1-score | 0.98 |
| ***High*** | Precision | 0.97 |
| Recall | 0.97 |
| F1-score | 0.97 |
|  | **Accuracy** | **0.97** |

# Discussion of Results

*Final Post-Study Review of Research Questions*

The conclusion of our project would dictate we reassess our initial research questions based on our individual model and overall findings. Our project's goal was to create a predictive model to better understand and gauge wildfire intensity. Our model utilized readily available and established environmental factors as a data source. The aim of our research was to support more effective decision-making in firefighting and resource allocation when managing and fighting wildfires. Having constructed, tested, and validated our model we can now provide answers to our initial research questions. We can also draw inferences regarding the relationship between wildfire intensity levels and the specific environmental factors associated with each occurrence.

Starting out with our first research question:

*“What is the relationship between environmental factors and wildfire intensity and frequency? Which environmental factors affect wildfire intensity more?”*

We tackled this by utilizing several feature selection statistical methods to determine feature importance for our classification model, namely the Anova F-Value and Mutual Information scores. Both methods help to determine how correlated the independent variables were to our dependent variable “Intensity Level”. Upon running both tests, we found some interesting parallels in their results:

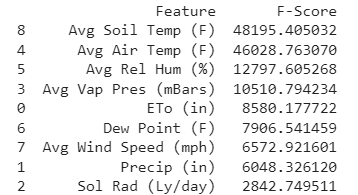
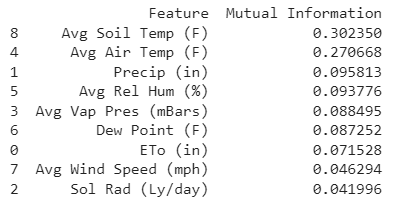
 

Figure 12: ANOVA F-Value and Mutual Information Scores for all Environmental Variables

As you can see by the ranking, the higher the score the better the variable is at predicting the dependent variable. The top two environmental factors when it comes to predicting a wildfire’s intensity level are Average Soil Temperature and Average Air Temperature with Average Relative Humidity taking the third and fourth spot. These findings provide support to answer our research questions as it shows by rank which environmental factors are most influential and provide confidence in each variable in our classification model. We are inclined to give more weight to the mutual info score because it tackles both linear and non-linear relationships as opposed to the f-score which treated each variable independently. We feel that our research has shown that Precipitation along with Average Air temp, Soil temp and Humidity affects Wildfire intensity the most. Our findings align with Lukovic et al.’s study which demonstrated climate change’s impact on wildfire risks, with more emphasis on increased temperatures and lower precipitation (Lukovic et al., 2021). However, unlike Lukovic et al.’s study, we used a different environmental condition to predict wildfire intensity, which would reduce the risk of overly relying on one single data type.

Moving on to the next research question:

*“How do we measure or certify the accuracy of the model? What is the best evaluation metric for success?”*

Regarding predicting the intensity level of a wildfire, we want to create a scenario where our model accurately and reliably predicts the likelihood of highly intense wildfires. While our model can accurately identify wildfires of low, medium, and high intensity wildfires, we place practical significance on the identification of high intensity wildfires to prevent situations where avoidable losses occur. A situation where lives and property were lost due to not enough resources being allocated due to our model assigning a low intensity level when it is meant to be high would be catastrophic. At the same time, we want a scenario where our model does not just label everything as a high intensity because that wastes resources and spreads the firefighting effort thin leading to an overwhelming inefficiency. Therefore, while Recall as a metric is an important evaluator for our model, Precision must be factored in too at least if not just as much. We want our model to aid in the efficient management of firefighting efforts in California which means correctly identifying where the risks are every time and efficiently utilizing the resources where they are required. Inevitably, we decided to use Recall and Precision as our model’s evaluation metric for performance as well as the F1 score which could be seen as a combination of both. The use of these different metrics can be supported by Zhou et al.’s study where they recommended using multiple evaluation metrics for a comprehensive analysis of their unbalanced data for fire point prediction (Zhou et al., 2021).

Finally, on to our last research question:

*“How can a data-driven approach to risk assessment assist in the mitigation of wildfire emergencies in California?”*

As previously stated, our goal in this project was to build a predictive model that could accurately take in environmental factors at the site of a fire outbreak and correctly estimate the threat or intensity level for efficient firefighting measures. Correctly estimating the threat of each situation would allow firefighting organizations such as CALFIRE (California Department of Forestry and Fire Protection) and local governments to make informed decisions in allocating resources to each incident. They could save the environment, human/animals lives and property from harm all while conserving the limited resources they must carry out such actions. There would not be situations where they are caught off guard by a fire, they are not ready for or a fire they overestimated and wasted resources on. The efficient decision-making and resource allocation when it comes to wildfire fighting and management in California will only be boosted by our model as it optimizes operations removing the uncertainty and doubt that comes with tackling each wildfire occurrence as well as the crisis as a whole.

# Conclusion

This research project is focused on building a predictive model that forecasts wildfire intensity levels in California by analyzing key environmental variables. We constructed a model capable of categorizing wildfire intensity into three classes low, medium, and high. Our research further delved into investigating the relationship between environmental conditions and wildfire intensity levels, and through our analysis, we understand that soil and air temperatures, relative humidity, and precipitation, in that order, have the highest impact on wildfire occurrences and intensity levels. We have employed supervised learning algorithms throughout our research, among which the XG Boost model yielded better results. Model evaluation was done not just based on accuracy rate, but also other metrics such as recall, precision, and F1-score, and the same has been supported by previous studies.

*Practical Applications and Implications*

As discussed, our predictive model will have far-reaching implications in optimizing firefighting strategies. Our model can accurately predict the intensity of wildfires, which will help agencies like CAL FIRE and local governments in California. It will guarantee the effective allocation of resources, minimizing wastage and facilitating more effective firefighting strategies in regions with higher risk levels. The model will enhance wildfire preparedness by providing early wildfire warnings, aiding in reducing any potential losses of lives, properties, and businesses. This will eventually result in cost savings, economic benefits, and increased community safety. Aside from the above benefits, the model will ensure environmental conservation in the region by preventing the widespread destruction caused by intense wildfires. *Recommendations for Future Model Development*

The current study relied on historical datasets due to time constraints and data availability. We believe that the inclusion of real-time data in the future could significantly enhance the model’s accuracy and timeliness of wildfire intensity predictions. Updated data on ever-changing environmental conditions during active fires could be incorporated to improve the model's adaptability and reliability.

While our primary focus was on just the California regions, future studies could extend to other locations and cover a much larger area. By extending the analysis to different geographical areas with distinct environmental conditions, we can assess the model's generalizability and adaptability in diverse settings. We would be able to determine whether the predictive capacities developed in California apply to different climates, landscapes, and ecosystems.

Incorporating diverse datasets having ecological, vegetation, and socioeconomic data along with the environmental conditions may improve the predictive power of wildfire risk assessment. This approach would provide a deeper knowledge of the complex relationships between environmental factors, human activity, and vegetation behavior. By including these variables in the model, the potential for enhancing wildfire risk assessment strategies and facilitating improved management efforts across different regions can be significantly increased.

# List of References

Andersson, A. (2021). Mechanisms for log normal concentration distributions in the environment. *Scientific Reports*, *11*(1). <https://doi.org/10.1038/s41598-021-96010-6>

Asanjan, Ata & Memarzadeh, Milad & Lott, Paul & Rieffel, Eleanor & Grabbe, Shon. (2023). *Probabilistic Wildfire Segmentation Using Supervised Deep Generative Model from Satellite Imagery.* Remote Sensing. 15. 2718. 10.3390/rs15112718.

Bishop, L. (2023). *Wildfire Statistics: Damage, Fatalities and Insurance Rates*. Value Penguin. <https://www.valuepenguin.com/homeowners-insurance/wildfire-statistics>

Burke, M., Driscoll, A., Heft-Neal, S., Xue, J., Burney, J., & Wara, M. (2021). The changing risk and burden of wildfire in the United States. *Proceedings of the National Academy of Sciences*, *118*(2). <https://doi.org/10.1073/pnas.2011048118>

California Department of Forestry and Fire Protection. (2001). *Statistics* | CAL FIRE. <https://www.fire.ca.gov/our-impact/statistics>

California Department of Forestry and Fire Protection. (2020). *August Complex (includes Doe fire)*. Cal FIRE. <https://www.fire.ca.gov/incidents/2020/8/16/august-complex-includes-doe-fire/>

Cal OES News. (2023). *Cal OES Launches Innovative Real-Time Tool for California’s First Responder Community*. <https://news.caloes.ca.gov/cal-oes-launches-innovative-real-time-tool-for-californias-first-responder-community/>

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD ’16*, 785–794. <https://doi.org/10.1145/2939672.2939785>

Dutilleul, P., & Carrière, Y. (1998). Among-environment heteroscedasticity and the estimation and testing of genetic correlation. *Heredity*, *80*(4), 403–413. <https://doi.org/10.1046/j.1365-2540.1998.00267.x>

Elbein, S. (2021). Researchers paint bleak picture of forest fires beyond 2030. *The Hill*. <https://thehill.com/policy/equilibrium-sustainability/565773-researchers-paint-bleak-picture-of-forest-fires-beyond-2030/>.

ESRI Blog. (2021). *CAL FIRE Applies New Predictive and Tactical Tools to Fight Ferocious Fire*. ESRI. <https://www.esri.com/about/newsroom/blog/cal-fire-puts-entire-team-on-same-map/>

Igini, M. (2023). *What causes wildfires?*. Earth.Org. <https://earth.org/what-causes-wildfires/>

Internet archive: Wayback Machine. (2001). *CDF 2000 Fire Season Summary*. <https://archive.org/web/>

Kotsiantis, S. B. (2013). Decision trees: a recent overview. *Artificial Intelligence Review*, 39, 261-283.

Liang, H., Zhang, M., & Wang, H. (2019). A Neural Network Model for Wildfire Scale Prediction Using Meteorological Factors. *IEEE Access*, *7*, 176746–176755. <https://doi.org/10.1109/ACCESS.2019.2957837>

Lukovic, J., Blagojević, D., Chiang, J. C. H., & Sekulić, A. (2021). *A Later Onset of the Rainy Season in California*. *Geophysical Research Letters*, 48, e2020GL090350. <https://doi.org/10.1029/2020GL090350>

Malik, A., Rao, M. R., Puppala, N., Koouri, P., Thota, V. A. K., Liu, Q., Chiao, S., & Gao, J. (2021). Data-Driven Wildfire Risk Prediction in Northern California. *Atmosphere*, *12*(1), 109. <https://doi.org/10.3390/atmos12010109>

Mulkern, A. C. (2023). *Climate change is escalating California’s wildfires*. Scientific American. <https://www.scientificamerican.com/article/climate-change-is-escalating-californias-wildfires/>

Natekin, A., & Knoll, A. (2013). Gradient boosting machines, a tutorial. *Frontiers in Neurorobotics*, *7*(21). <https://doi.org/10.3389/fnbot.2013.00021>

Sayad, Y. O., Mousannif, H., & Al Moatassime, H. (2019). Predictive modeling of wildfires: A new dataset and machine learning approach. *Fire Safety Journal*, *104*, 130–146. <https://doi.org/10.1016/j.firesaf.2019.01.006>

Swain, D. L. (2021). *A Shorter, Sharper Rainy Season Amplifies California Wildfire Risk*. *Geophysical Research Letters*, 48, e2021GL092843. <https://doi.org/10.1029/2021GL092843>

Vapnik, V. (1998). Statistical Learning Theory. Wiley, New York, New York.

Vapnik, V. (1995). The Nature of Statistical Learning Theory. Springer, New York, New York.

Wibberly, A. (2021). *How Does Climate Change Affect the Seasons?* Smartenergy.com. <https://www.smartenergy.com/how-does-climate-change-affect-the-seasons/>

Wikipedia Contributors. (2019). *Lapse rate*. Wikipedia; Wikimedia Foundation. <https://en.wikipedia.org/wiki/Lapse_rate>