County-Specific Wildfire Prediction Using Machine Learning: A Southern California Case Study

Mariah Reilley

Department of Data Science, University of Wisconsin – La Crosse

DS 785: Capstone

July 30, 2024

Abstract

Wildfires present a growing threat to public safety, environmental sustainability, and economic stability in California, with climate change threatening to amplify these risks. Accurate prediction of wildfire occurrences is essential for effective management and mitigation strategies. This project evaluates the performance of various machine learning models, including Random Forests, Decision Trees, k-Nearest Neighbors (KNN), and Support Vector Machines (SVM), in predicting wildfire occurrences in Southern California's most vulnerable counties: Los Angeles and Riverside. The study utilized a comprehensive dataset from 2016 to 2020, including meteorological variables, vegetation indices, and fire occurrence records. The findings indicated that the Random Forest model outperformed other models, achieving the highest accuracy and F1 scores in both counties, thus providing the most reliable predictions of wildfire occurrences. Key variables such as the Normalized Difference Vegetation Index (NDVI) and average temperature were identified as the most significant predictors of wildfire risk, highlighting the importance of vegetation density and climatic conditions in wildfire occurrence. The results emphasize the need for localized, county-specific models to capture the nuanced factors influencing wildfire risk. The study concludes with a discussion of the implications of these findings for improving wildfire risk assessment tools, proposing that more accurate and region-specific models can aid in developing targeted and effective wildfire management strategies in Southern California. This research contributes to the ongoing efforts to improve wildfire prediction models and offers valuable insights for policymakers, fire response teams, and environmental scientists.

Keywords: wildfire prediction, machine learning, Random Forests, Decision Trees, k-Nearest Neighbors, Support Vector Machines, NDVI, Southern California, regional analysis

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

Background

Wildfire management in California is a critical issue with significant implications for public health and safety, environmental sustainability, and economic stability. In 2020 alone, wildfire emissions in California were "seven times the 2003-2019 mean", negating two times the amount of California's total greenhouse gas emission reductions for that year (Jerrett et al, 2022). As reported by CAL Fire, these fires blazed over 4.3 million acres and caused well over $4 billion worth of damages (CAL Fire, n.d.). Wildfires not only cause significant damage to infrastructure and natural resources but also result in severe health consequences and long-term ecological instability. Effective wildfire risk assessment is important now more than ever, as recent studies suggest the frequency and severity of wildfires will increase, due to drought and rising temperatures driven by climate change (NIDIS, n.d.).

There are a wide variety of factors that influence wildfire risk. Those most commonly cited include increased air temperature, decreased humidity, higher wind speed, and drier vegetation (Khorshidi et. al, 2020). However, each affected area has a unique set of risk factors that contribute to wildfire ignition and spread. For instance, Santa Ana winds and dense populations, specific to Southern California, are often the main contributing factors to more aggressive wildfires compared to Northern California, where lightning strike density and vegetation prevail as the highest drivers (Center for Ecosystem Climate Solutions; Cal OES). This regional specificity underscores the need for distinct, localized predictive models to provide accurate risk assessments based on unique regional factors.

In recent years, data science approaches have rapidly improved in their application to solving environmental science problems, often involving vast datasets with many parameters.

Likewise, machine learning in wildfire risk and behavioral prediction is increasingly recognized for its ability to handle large, diverse datasets and capture complex patterns that lead to accurate predictions. Machine learning models, which rely on pattern recognition rather than linear assumptions, improve the accuracy and lead time of wildfire risk predictions, providing valuable insights for early warning systems and strategic planning.

Problem Statement

The application of machine learning approaches to wildfire risk assessment has been explored in several prominent studies(Castrejon et al. 2023; Malik et al. 2021). However, these studies were primarily conducted at a state level, often resulting in low accuracy scores due to the complex and diverse nature of wildfire risk factors across different regions. The problem with applying a single machine-learning model to a large geographic area, such as an entire state or country, lies in the variability of localized conditions. Wildfires are caused by a wide variety of interdependent factors, including climate, topography, vegetation type, and human activity, which can vary significantly even within small geographic areas. As a result, universal models may fail to capture these localized conditions, leading to inaccurate fire predictions and misleading risk assessments.

The specific problem this project aimed to address is the need for more accurate, localized wildfire risk prediction models that can account for the unique environmental and meteorological factors specific to different regions. By focusing on county-level data in Southern California, this study aimed to develop machine learning models that can provide more precise and reliable predictions of wildfire risk. This localized approach was expected to enhance the predictive power of the models and provide valuable insights for targeted wildfire management and prevention strategies.

Purpose of the Project

The next generation of data science applications in wildfire risk assessment needs a new approach as we work towards building accurate predictive models that inform response strategies. This project aimed to identify a more accurate method of predicting wildfire risk with machine learning. Unlike existing models that often use a generalized approach, this report focused on identifying the best-performing model on a county-by-county basis by analyzing localized data and risk factors. Through the comparison of multiple machine learning techniques with data within the bounds of a singular county, this study seeks to explore the utility of regional-specific analysis. These models will allow for more powerful wildfire management strategies by providing teams with more preparation time and informed targeted responses.

Project Objectives

The first objective of this project was to determine the counties in Southern California that are most highly impacted by wildfires. Counties included in this analysis were those with the largest population including San Bernadino, Los Angeles, San Diego, and Riverside.

The second objective was to compare the accuracy of several machine-learning methods to predict wildfire risk in highly impacted counties. These machine learning methods include random forests, decision trees, Naïve Bayes, and neural networks, selected based on their efficacy demonstrated in previous studies. With comprehensive evaluation using precision, recall, and F1 score accuracy metrics, this project will identify the most accurate methods for each county included in the study.

The third objective was to identify the county-specific factors within the dataset that are the most significant contributors to wildfire risk. This involves determining which variables - such as wind speed, air temperature, precipitation levels, and NDVI (Normalized Vegetation

Index) - have the strongest impact on model accuracy in each county. This information will reveal important location-specific risk factors that allow local fire management organizations to implement targeted mitigation efforts.

Finally, with important data-driven information, the final objective of the project was to explore how region-specific wildfire insights can be applied and provide actionable recommendations for California wildfire management.

Limitations

While this report aimed to provide a novel approach to wildfire risk predictions, there were limitations to consider. The main limitation was the availability and quality of the data. This project used historical data from 2010 to 2021 gathered by NASA satellites and NOAA, which contained gaps and inconsistencies. While fire history data was mostly complete and validated by the Fire and Resource Assessment Program (FRAP), weather data obtained through NOAA contains large gaps. To address this, missing data was estimated with the value of recent existing records.

Another potential limitation is that the scope of the project is confined to the counties of Riverside, San Diego, San Bernardino, and Los Angeles in Southern California. While this regional focus is necessary for detailed and localized analysis, the results from this study will not have a generalized accuracy baseline to compare to. Future research can expand on this work by applying similar methodologies to datasets encompassing the entire state of California.

Conclusion

In conclusion, this project aimed to develop and compare machine learning models for predicting wildfire risk on a county-by-county basis in Southern California and provide valuable data-driven insights and recommendations for improving wildfire management practices. By addressing gaps and limitations in current research on this topic, the main goal was to contribute to the ongoing efforts to mitigate the impacts of wildfires in California.

Organization of the Paper

The structure of the paper will be as follows:

I. Introduction
   - A. Overview of wildfire management importance
   - B. Problem statement
   - C. Project objectives, purpose, and limitations

II. Literature Review
   - A. Factors influencing wildfire risk
   - B. Challenges of traditional prediction systems
   - C. Machine learning methods in wildfire predictions
   - D. Current gaps in research

III. Methodology
   - A. Data collection, preparation, and preprocessing
   - B. Modeling methods

IV. Results
   - A. Modeling results evaluation
   - B. Feature selection
   - C. Important drivers of wildfire
   - D. Discussion of final model results
   - E. Project objectives evaluation

V. Summary and Conclusions

A. Overall summary of the project

B. Recommendations for the future

Chapter 2: Literature Review

Recent studies in the areas of wildfire risk factors, current challenges faced by traditional wildfire prediction systems, machine learning methods utilized by wildfire prediction research, and California-specific models were researched to advise the management of this project. This review summarizes key findings and contextualizes the significance of the project's objectives.

Factors Influencing Wildfire Risk

Various factors including vegetation characteristics (density, moisture content, type), meteorological variables (humidity, air temperature, wind speed), and human activity, have been attributed to wildfire risk. MacDonald et al. (2023), published in the International Journal of Wildland Fire, provided a comprehensive analysis of the significant drivers of California's changing wildfires, identifying anthropogenic climate change as a major factor in increasing air temperature and lack of humidity. Additionally, the study highlighted the role of Indigenous fire practice underutilization, changes in logging management, and lack of fire suppression as causes for significant vegetative fuel accumulation in high-risk areas.

Regional Specificity

California, spanning ten degrees of latitude, encompasses three variations of the Mediterranean climate (Kauffman et al., 2003). Li et al. (2021) investigated the spatial and temporal patterns of wildfires in California and found that significant climatic differences result in varying wildfire occurrence factors from region to region. For instance, changes in vegetation and lightning are more likely to ignite wildfires in Northern California compared to Southern California, where low humidity and Santa Ana winds have accelerated wildfire occurrence. Regional specificity in wildfire risk factors signifies the practicality of assessing wildfire risk within boundaries defined by similar physical and climatic conditions.

Challenges with Traditional Wildfire Prediction Systems

Current wildfire prediction approaches display limitations in their ability to accurately detect where and when a fire may occur. Fire teams in California typically employ models that rely on historical fire data and basic weather variables to estimate risk, lacking the modern capabilities of advanced data science tools. Fire Danger Indices (FDIs) are often used in fire management to indicate wildfire occurrence and spread. Feng et al. (2023) evaluated the accuracy of four common FDIs, highlighting the importance of predictive variables such as wind speed and daily minimum relative humidity for forecasting accuracy. Composite risk scores offered by FDIs use multiple weather variables to assess the potential of fire ignition, helping fire management teams prioritize high-risk areas. However, Feng et al. (2023) suggested that the low granularity of FDIs indicates the need for a scalable model to improve wildfire risk assessment at a more localized level. This paper emphasized the need for prediction tools for detailed, region-specific predictions.

Recognizing recent advancements in data technologies, California has made large investments in data-driven approaches for modeling wildfires. In partnership with Aerospace and Defense company Lockheed Martin, CAL FIRE is exploring the use of AI, data analytics, and drone software to model wildfire risk and behavior (Hooper, 2023). This initiative creates an opportunity for the state to utilize advanced tools to monitor and control wildfires with machine learning which leverages advanced algorithms to detect complex patterns within large datasets.

Review of Machine Learning Methods in Wildfire Prediction

Machine learning algorithms have emerged as powerful tools for wildfire risk prediction, addressing several limitations in traditional forecasting methods, due to their capacity to process large datasets and uncover intricate, nonlinear patterns. Jain et al. (2020) discussed the most prominent algorithms used in previous scientific literature in their review of the application of machine learning techniques in wildfire science and management. Within the domain of fire occurrence prediction, the review identified several commonly used classification methods: Support Vector Machine (SVM), Bayesian Networks (BN), Artificial Neural Networks (ANN), Decision Trees (DT), Random Forests (RF), Maximum Entropy (MaxEnt), and Naive Bayes (NB).

Performance Comparison of BN, NB, and DT Methods

Pham et al. (2020) conducted a study comparing Bayes Network (BN), Naïve Bayes (NB), and Decision Tree (DT) machine learning methods for predicting fire risk in Pu Mat National Park, Vietnam. Their study found that the BN model outperformed the DT and NB models, with an AUCROC value of 0.96. This analysis revealed that while all three classification models yielded high accuracy(AUCROC) values, the Bayes Network model was the most equipped to handle complex stochastic relationships between variables. The study attributed fire occurrence to human-related variables, highlighting the importance of these factors in fire prediction. These findings suggest that while BN demonstrated a higher accuracy score, DT and NB are valuable tools in wildfire management strategies as well.

Performance Comparison of ANN and SVM

Artificial Neural Networks (ANN) and Support Vector Machines (SVM) were established in the review by Jain et al. (2020) as two of the most widely used machine learning algorithms for wildfire prediction. A study by Sayad et al. (2019) compared ANN and SVM for predicting wildfire occurrences in Canada using a simulated dataset containing vegetation, climate, and thermal anomaly variables based on remote sensing data. In terms of predictive accuracy, ANN slightly outperformed SVM, scoring 98.3% and 97.48% respectively. Additionally, findings from this study highlighted that the ANN model tends to perform better in capturing non-linear relationships, while SVM is better at classifying binary fire and no-fire data.

Performance Comparison of MaxEnt and Random Forest

Janiec and Gadal (2020) compared the effectiveness of Maximum Entropy (MaxEnt) and Random Forest (RF) models in their paper on predicting wildfire risk in northeastern Siberia. The study defines fire risk variables as bioclimatic factors, satellite images, and vector data. The results showed that the RF method performed better at the macroscale, while the MaxEnt model was more effective at the microscale. When compared on the same scale, RF provided more accurate results at narrower risk prediction margins and MaxEnt identified broader, less precise classifications. These findings suggest that when applied to datasets of different scales, both RF and MaxEnt displayed distinct strengths. MaxEnt is preferable for localized, small-scale predictions, and RF is useful for larger-scale risk assessments.

Case Study within California

Acknowledging the region-specific nature of wildfire risk data, it is important to examine machine learning methods tested on wildfire data within California, as evidenced by the study conducted by Pham et al. (2022). This study compared the accuracy of five machine learning methods–ANN, SVM, KNN, and Gaussian Naive Bayes–in predicting wildfire occurrence in California. The researchers used two datasets with different-sized regions, one treating the whole state as one region, and another broken down into distinct counties for localized analysis. Pham et al. applied the machine learning methods to a dataset of features including binary past fire occurrence (Fire/No_Fire), NDVI, thermal anomalies, burned area, and land surface temperature. For the state-level dataset, the highest prediction accuracy was achieved by the ANN model, at

89%. The accuracy of the five tested models improved significantly when applied to the county-specific dataset, with the KNN model achieving 97% accuracy.

The results of Pham et al.'s paper underscore several key implications for the future of wildfire prediction using machine learning techniques. Integrating precise remote-sensing environmental data with historical fire report records can improve model prediction accuracy. Additionally, ANN and KNN models are valuable for detecting complex patterns within datasets with many variables. The study's most novel finding is that machine learning models applied to wildfire data perform better on localized data. This suggests that region-specific models constrained by the boundaries of a single county can better capture the nuances of local wildfire risk factors.

Gaps in Current Research

While much research has been done in predicting wildfire risk, most of the work has focused on predicting risk from datasets encompassing large areas. Since the variables most highly correlated with wildfire risk can change drastically from region to region, there is a need for research comparing machine learning methods within smaller areas with more homogenous climate and land distribution variables. Hernandez and Hoskins (2024) conducted research in California's Central Valley supporting this theory. They compared the accuracy of machine learning algorithms to predict wildfire risk from data encompassing Yosemite Valley, Kings Canyon, and the Sequoias, some of the most at-risk areas in the state. Their analysis yielded low-accuracy results, with the best F1 score being 0.689. They concluded that the most appropriate way to improve these results would be to separate their dataset into distinct regions and run the algorithms individually. This study emphasizes the current gaps in wildfire prediction research, highlighting a need for further development in localized models to enhance model accuracy.

Conclusion

The existing literature reveals significant findings in the application of machine learning methods for wildfire risk prediction, emphasizing the value of environmental and meteorological factors and the limitations of traditional prediction systems. While extensive studies have utilized large datasets, the findings from Hernandez and Hoskins (2024) and Pham et al. (2022) identify the need for developing localized models. Addressing these gaps is crucial, as it supports the objectives of the research done in this project to integrate machine learning methods and region-specific models for exploring more effective wildfire risk prediction strategies.

Chapter 3: Methodology

The primary objective of this project was to compare the efficacy of various machine learning modeling techniques to predict wildfire occurrence in high-risk Southern California counties using various environmental and meteorological data. This chapter outlines the methods used in preparing this data for modeling including data collection, identifying the most at-risk counties in Southern California for wildfire, data preprocessing, feature engineering, handling missing values, and model selection.

Data Collection

The data collected for this analysis spanned from January 1, 2016, to December 31, 2020, and included several key variables: fire occurrence data (latitude, longitude, date of fire, and fire radiative power), NDVI (Normalized Difference Vegetation Index), average wind speed (AWND), average temperature (TAVG), and precipitation (PRCP). Fire occurrence data was collected from both NASA MODIS and VIIRS. Meteorological data was collected from the National Centers for Environmental Information open-access database. NDVI numbers were calculated from satellite imagery data made available by NASA Landsat 8 satellite.

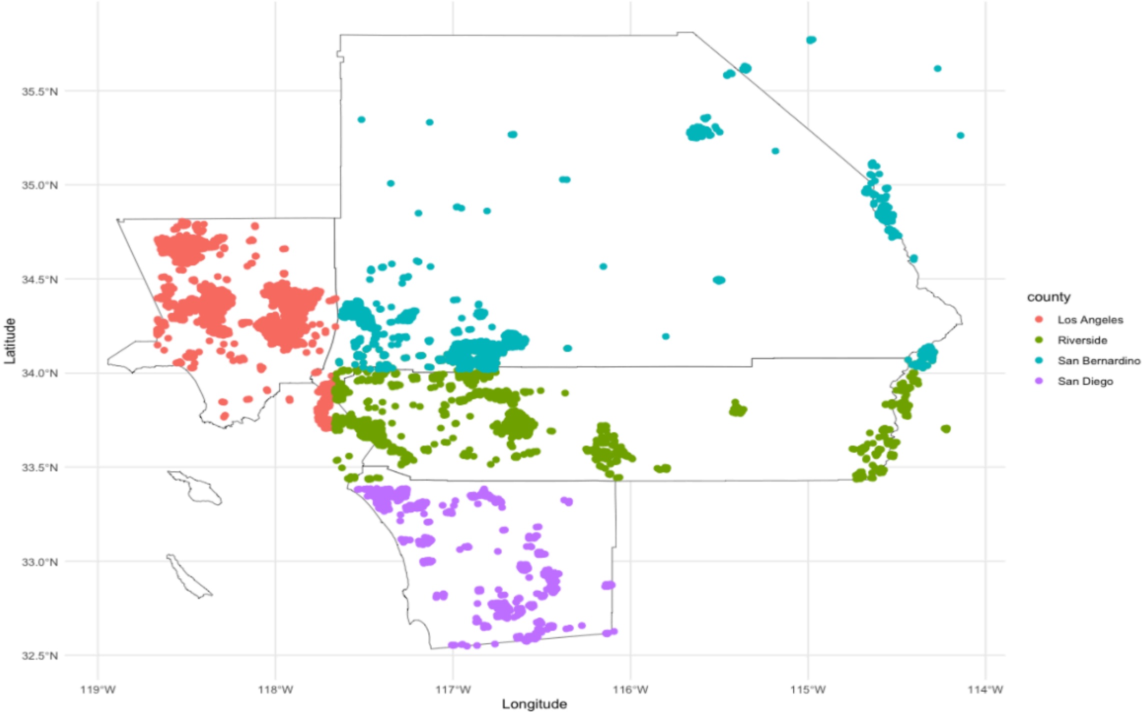
The Normalized Difference Vegetation Index (NDVI) represents the amount of chlorophyll reflected from a spatial composite image and was used as a marker for the presence of live vegetation in the studied areas. Google Earth Engine code was used to call specific regions included in the dataset and calculate NDVI using a Raster Calculator tool. The wavelengths of images retrieved from the Landsat 8 satellite included the Near Infrared Band (NIR) and the surface reflection (Red). The NDVI calculation of these wavelengths yielded values ranging from -1 to 1, where positive values indicate the presence of vegetation. This data is highly important in determining wildfire risk as Nguyen et al. (2018) found that NDVI was the most reliable vegetation index predictor of fire occurrence.

County Selection

The wildfire data was collected from four counties in Southern California including Riverside, Los Angeles, San Diego, and San Bernadino. These counties were deemed to have the highest risk for wildfire based on their expected net financial loss, population size, and historic wildfire record (Rodriguez, 2024). The locations of fires occurring between 2013 and 2020 displayed in Figure 1 show that each county experienced a sizeable amount of wildfires typically concentrated in fire hotspots.

Figure 1

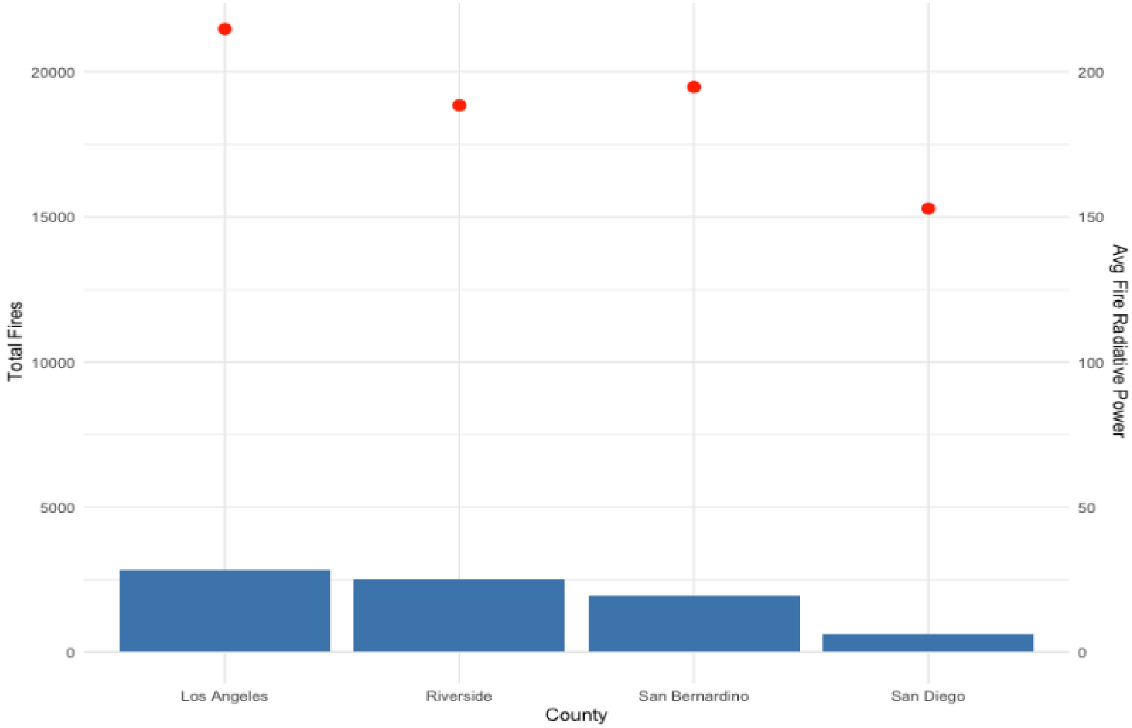
Map of Wildfire Locations Across Four Counties in Southern California



The wildfire occurrence data from all counties was combined into a single dataset and analyzed to find the two counties with the highest risk of wildfire to be selected for predictive modeling and comparison. Figure 2 depicts the average fire radiative power (FRP), measuring the radiative energy emitted from wildfires, and the total number of fires for each county. Los Angeles County experienced the most wildfires and the highest average FRP out of the four counties included in the comparison, with an average FRP of 215 watts per fire and 354 fires per year. San Bernadino County experienced the second-highest FRP with an average of 195 watts per fire. Although Riverside County's fires had a slightly lower average FRP, the county experienced 73 more fires per year compared to San Bernadino County. Based on this comparative analysis, Los Angeles County and Riverside County were chosen for modeling.

Figure 2

Comparing Wildfire Occurrence and Average Intensity Across Four Counties



Data Preparation and Preprocessing

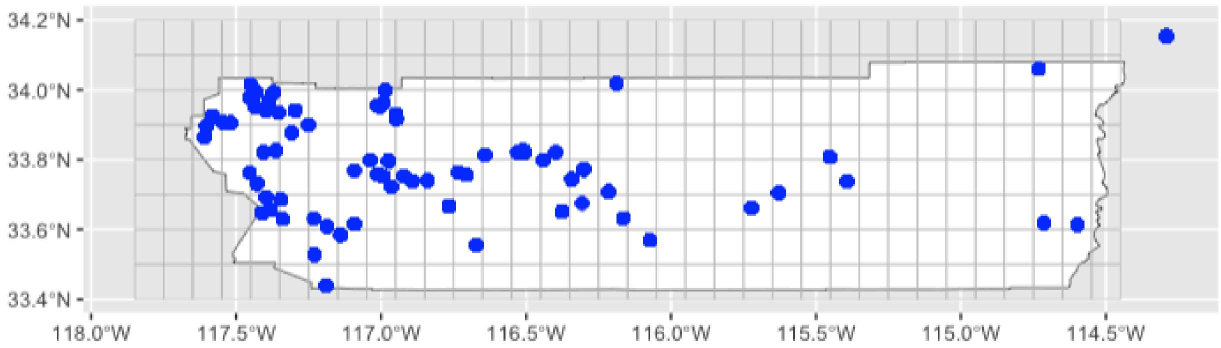
Data preparation and preprocessing are crucial for achieving high accuracy in predictive models. This process involved several important steps including data subdivision for spatial organization, handling missing data, addressing data imbalances, and data normalization.

Grid Tile Division

The weather data included in this project was collected from weather stations dispersed across their respective county, each with unique coordinates. Los Angeles has 115 weather stations, while Riverside has 81. To accurately match fire record data with relevant weather conditions collected from these stations, Los Angeles and Riverside counties were divided into evenly sized grid tiles of approximately 10 km x 10 km. Figure 3 shows the tiles generated for Riverside County with weather stations plotted in blue. This method, inspired by Pham et al. (2022), facilitated the creation of a dataset with a unique coordinate pair for each tile and associated weather data from the nearest weather station.

Figure 3

Weather Stations and Grid Tiles in Riverside County



The rationale for utilizing this method was to increase the spatial reliability and accuracy of the dataset. The division of each county into a grid of evenly sized tiles allowed for systematic organization and data matching from different sources. This approach not only assisted in the smooth handling of varying locations of weather stations and fire occurrences but also ensured that the grid tiles were matched with stations with complete data, as many stations did not record data for all of the days included in this analysis.

Missing Data

The missing data obtained from weather stations was controlled by the grid tile matching system which matched each tile coordinate to the nearest weather station with complete data. For Riverside County, the number of stations with complete data were 9, 13, and 6 for average temperature (TAVG), precipitation (PRCP), and average wind speed (AWND), respectively. For Los Angeles County, the number of stations with complete data was 23, 94, and 13 for TAVG, PRCP, and AWND, respectively. While the truncated list of stations may imply that weather data is generalized, this method ensured that each point within county grids was supplied with the most accurate and complete data available.

Normalized Difference Vegetation Index (NDVI) data was collected by Landsat 8 every 16 days, while the collection frequency of weather data obtained through NCEI was collected daily. To handle the missing NDVI data caused by this sampling discrepancy, the Last Observed Carried Forward (LOCF) method was used. This method was chosen over interpolation due to its common usage in related wildfire prediction and environmental research. Additionally, it can be reasonably assumed that the vegetation index will not change drastically within a given 16-day period.

Imbalanced Data

The combined datasets for wildfire instances and environmental data were imbalanced for both Riverside and Los Angeles counties. Only 2% of the collected data for Los Angeles County contained fire instances, while the Riverside County dataset contained only 0.3% of fire instances. To address this, the Synthetic Minority Over-sampling Technique (SMOTE) was applied using the R Studio ROSE package to generate synthetic samples for the minority class (presence of fire). After SMOTE balancing, datasets for both Riverside and Los Angeles counties had a near 50/50 split for fire and no fire data points. Other techniques considered included oversampling the minority class and undersampling the majority class. However, these methods were not chosen due to their potential to result in either model overfitting or valuable data loss.

Data Normalization

Environmental data must be normalized before training a machine learning model to predict wildfire occurrence. Several features in the dataset were heavily skewed. Specifically, environmental variables AWND and PRCP were positively skewed for both Los Angeles and Riverside counties, suggesting that the majority of data fall within the lower range of values. Modeling techniques included in this project perform best when variables are normalized and balanced. To do this, square root transformation was chosen to correct skew based on its ability to handle zero values.

Modeling Methodology

The modeling techniques selected for this project were k-Nearest Neighbors (KNN), Support Vector Machines (SVM), Random Forests, and Decision Trees. These methods were chosen based on their high accuracy in similar modeling contexts (Pham et al., 2022; Sayad et al., 2019). Maximum Entropy (MaxEnt) was considered but not chosen due to its varied performance with smaller datasets (Janiec and Gadal, 2020). Artificial Neural Networks (ANN) were also considered, however, Sayad et al. (2019) found that SVM performed better at classifying binary fire and no-fire data.

Modeling Preparation

The chosen models were trained and evaluated using the prepared datasets. The datasets for both Los Angeles and Riverside counties were split into training (80%) and testing (20%) to ensure effective validation. Additionally, the fire classifier was set to a binary variable with fire presence coded as 1. The evaluation metrics included precision, recall, F1 score, and accuracy, which provided a comprehensive understanding of each model's performance.

K-Nearest Neighbors

The K-Nearest Neighbors (KNN) classification algorithm works under the assumption of proximity in similarity, identifying the closest data points to make predictions. This model has demonstrated particularly accurate results in fire occurrence prediction on county datasets, likely due to the homogeneity of region-specific factors (Pham et al. 2022). For this analysis, the default parameter of (k = 5) was utilized. The model achieved a higher F1 score for the Los Angeles County data of 0.88, compared to Riverside County of 0.79. However, KNN produced a higher recall score for Riverside, compared to Los Angeles County.

Support Vector Machine

The second supervised learning approach used to predict wildfire occurrence was SVM. Hyperparameter tuning was performed to find the best C regularization term which was determined to be 1 with a radial kernel for both datasets. However, this tune did not improve the performance results of the model. SVM performed less effectively in nearly every performance metric category for both counties. There are several potential reasons for this model’s consistently poor performance, with the most likely cause being a lack of data variable diversity resulting in model underfitting.

Decision Trees

Decision Tree modeling was also utilized and compared on both datasets. This method works through an iterative process of splitting the dataset into smaller subsets based on variable power to make predictions. For predicting wildfire occurrences in Riverside County Decision Trees performed well achieving the second-highest F1 score of 0.79. Furthermore, this method produced superior recall scores for both Riverside and Los Angeles counties, suggesting that this method excelled in correctly identifying actual fire instances.

Random Forest

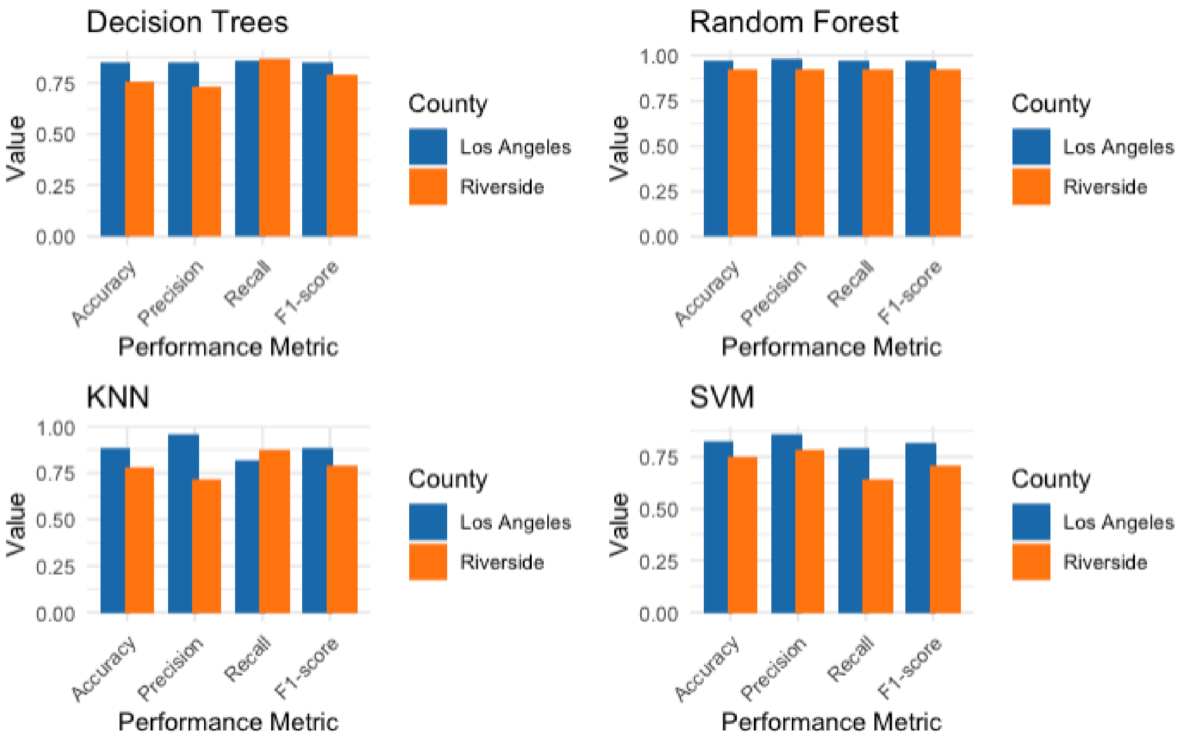
The final modeling approach utilized was Random Forests. This ensemble learning method combines a “forest” of decision trees to boost predictive accuracy, with a random subset of dataset features at each split. The Random Forests model with ntree = 500 produced the highest performance metrics compared to all other methods included in this project, for both Riverside and Los Angeles counties. Specifically, the accuracy and F1 scores for Riverside County were 0.924 and 0.92 for Riverside County, respectively, and 0.975 for Los Angeles County. The Random Forests models also exhibited superior precision and recall metrics for both counties. These results indicate that this modeling method was highly effective at minimizing false positives and negatives while identifying fire occurrences.

Model Comparison

Figure 4 shows the performance results of each model across the four key metrics for both Los Angeles and Riverside counties.

Figure 4

Model Performance Comparison for Los Angeles and Riverside Counties



From the graph, it is evident that Random Forest outperformed all other models, while KNN also achieved high-performance metrics.

Conclusion

In summary, Chapter 3 details the methodology used to prepare wildfire data in Los Angeles and Riverside counties for predictive modeling. The steps included comprehensive data collection from various sources, data preprocessing such as grid tile division, handling missing data, addressing data imbalances with SMOTE, and data normalization. Four machine learning models –KNN, SVM, Decision Trees, and Random Forests–were trained and then compared using recall, accuracy, F1 score, and precision performance metrics. Chapter 4 will discuss the results of this comparison as well as identify the most important variables for predicting wildfire occurrence in Los Angeles and Riverside counties.

Chapter 4: Results

In Chapter 3, four machine-learning models were trained and tested on wildfire data from Los Angeles and Riverside counties to predict wildfire occurrence. Chapter 4 discusses the results of these models, which were evaluated on the F1 score and AUCROC. Additionally, feature importance is discussed and the best model is selected for predicting wildfire occurrence in each county. This chapter concludes with an assessment of project objective achievement.

Model Results Evaluation

As discussed earlier, machine learning models included in this project were assessed using several metrics including recall, accuracy, F1 score, and precision performance metrics. AUCROC is also an important metric as it represents the degree to which a model can distinguish between classes, summarizing the trade-off between sensitivity(recall) and false positive rate. The F1 score provides a single metric balancing precision, the proportion of true positive predictions, and recall. Both AUCROC and F1 are considered in this project's selection of the best model. Additionally, model complexity and training time were examined, as resource-intensive models with low accuracy trade-offs may not outweigh the utility of their counterparts from a business perspective.

Table 1 presents the F1 and AUCROC scores for the four models tested on Los Angeles and Riverside County wildfire data in the methodology stage of this project. These scores were collected after the model training phase and reflect the accuracy and performance of each model to predict the positive class of fire occurrence with variables wind speed, precipitation, average temperature, and NDVI. Together, these metrics offer a well-rounded understanding of each model's predictive power and reliability.

Table 1

Comparing Model Performance Metrics for Los Angeles and Riverside Counties

County	Method	AUCROC	F1 Score
Los Angeles	Random Forest Classifier	0.990	0.964
	KNN	0.865	0.874
	Decision Trees	0.883	0.843
	SVM	0.860	0.818
Riverside	Random Forest Classifier	0.959	0.911
	KNN	0.893	0.791
	Decision Trees	0.829	0.768
	SVM	0.848	0.727

Best Model Selection Los Angeles County

Random Forest Classifier was the best-performing model for Los Angeles County with an AUCROC score of 0.990 and an F1 Score of 0.964. These scores were significantly higher than other models included in the project. These results suggest that the Random Forest Classifier was particularly effective at distinguishing between fire and no-fire occurrences, striking a balance between precision and recall. The high F1 score also demonstrated that this model was good at minimizing false positives and negatives, leading to accurate predictions.

While not as high, the K-Nearest Neighbors (KNN) model showed promising performance metrics as well, with an AUCROC score of 0.865 and an F1 Score of 0.874. Decision Trees and Support Vector Machines (SVM) had lower AUCROC and F1 Scores (Decision Trees: 0.883 and 0.843; SVM: 0.860 and 0.818), suggesting their inadequacy for Los Angeles County data.

Additionally, the computational efficiency of the Random Forest model, despite its complexity, made it a practical choice for high-stakes applications where accuracy is of the utmost importance. Although they are simpler and faster to train, KNN and SVM compromise on predictive accuracy. Likewise, Decision Trees, a less complex model compared to Random Forests, also fell short in predictive performance. Therefore the Random Forest Classifier was the optimal choice for Los Angeles County.

Best Model Selection Riverside County

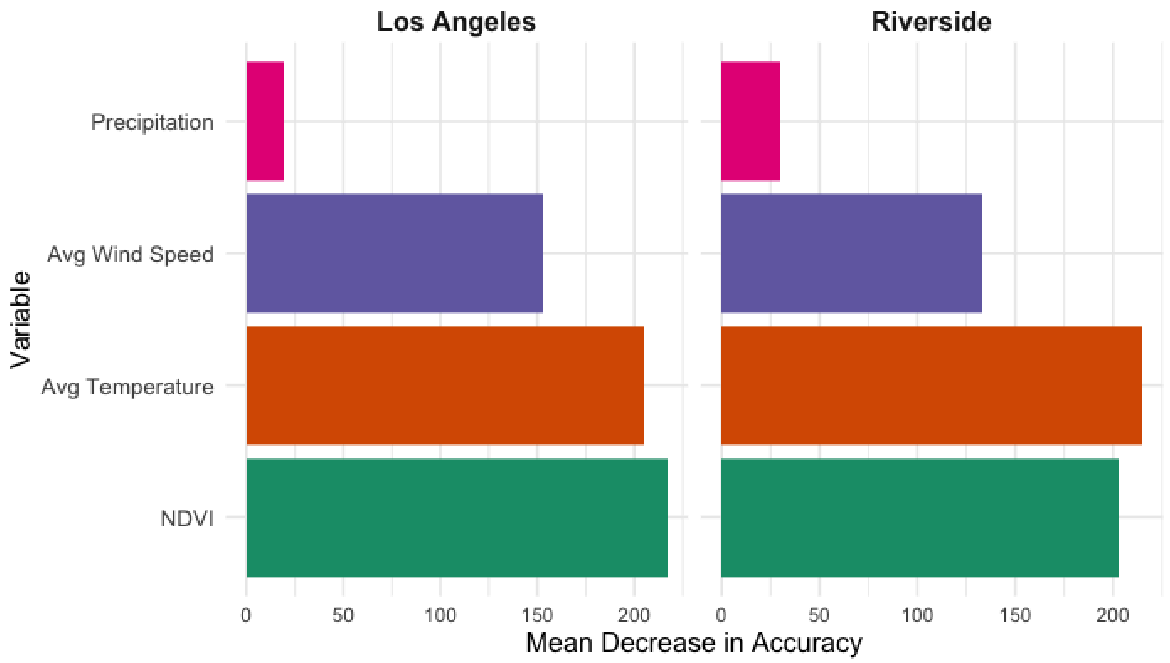
For Riverside County, the Random Forest Classifier also proved to be the model with the highest performance metrics with an AUCROC score of 0.959 and an F1 Score of 0.911. The KNN model, with an AUCROC score of 0.893 and an F1 Score of 0.791, also showed excellent performance but not as high as the Random Forest Classifier. Decision Trees and SVM perform even less effectively (Decision Trees: 0.829 and 0.768; SVM: 0.848 and 0.727). In terms of model complexity and training time, the Random Forest Classifier, although complex and resource-intensive, provided the best balance of predictive accuracy and robustness.

Feature Importance

Feature importance information was obtained by finding the features that held the most relative predictive power in the trained Random Forest Classifier model, our best model for both Riverside and Los Angeles. Features were ranked on Mean Decrease in Accuracy (MDA) which measures the average decrease in model accuracy, or the proportion of predictions that are classified incorrectly, when a variable is permuted. An alternative method for evaluating variable importance, Mean Decrease Gini, was also considered however it was not chosen for this project due to the potential for biases in datasets with fewer variables. Figure 5 displays the feature importance scores measured in MDA for the Random Forest Classifier model applied to all features included in the wildfire datasets.

Figure 5

Random Forest Model Feature Importance Scores For Both Counties



Feature Selection

Typically in the model-building process, selecting the best subset of features based on feature importance scores is advantageous. Features with low importance can be removed from the model, potentially increasing accuracy and decreasing model complexity. In this study, precipitation had the lowest Mean Decrease Accuracy (MDA) scores out of all four variables, with scores of 19 and 30 for Los Angeles and Riverside counties, respectively. However, due to the small number of variables in the datasets, the model's computational costs and complexity are already low. Thus, the benefits of removing a variable are minimal compared to models with a larger number of variables.

Environmental variables often have hidden interactions that can act unpredictably. For example, even though precipitation showed low individual importance, it may still contribute to significant interactions with other variables. On a windy and hot day, the model might predict a fire, but the presence of precipitation could inform this prediction. Given the small number of variables in the datasets and the potential for important interactions, feature selection was deemed unnecessary. The inclusion of all variables ensures that the model captures these complex interactions, maintaining predictive accuracy and reliability.

Important Drivers of Wildfire

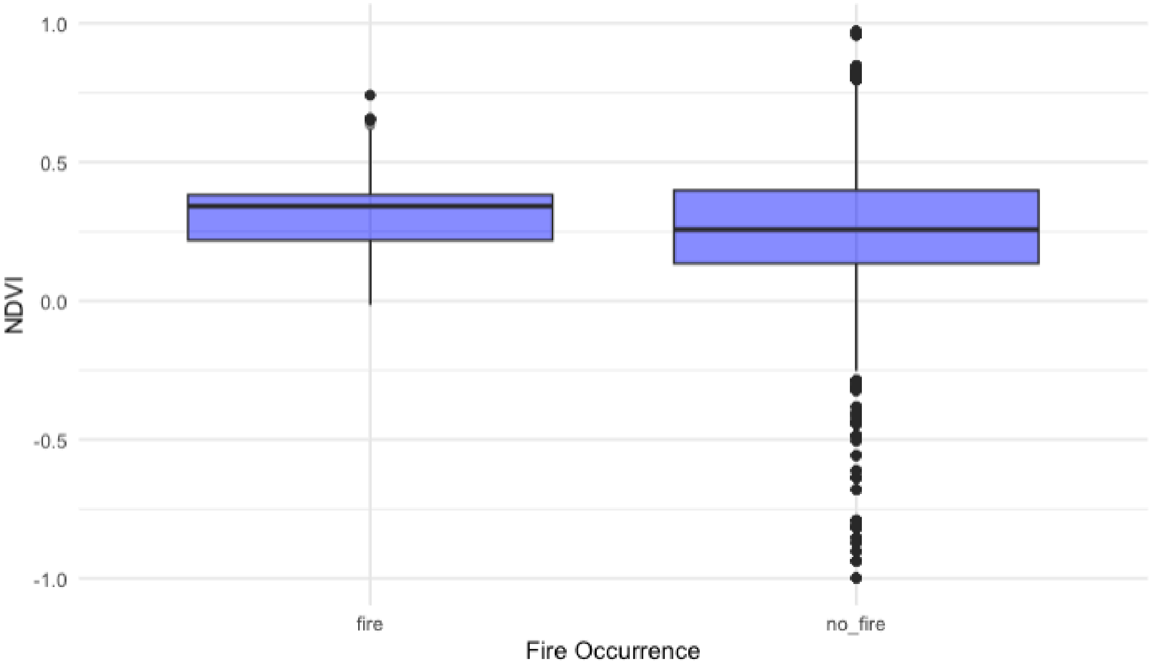
One of the main objectives of this project was to determine which variables were the most important in predicting fire occurrence for both Los Angeles and Riverside counties. The feature importance scores provide a model-based metric for interpreting the greatest drivers of wildfire. This information can be very useful for fire management authorities who look for ways to anticipate fire occurrences and plan their hazard abatement management strategies accordingly. Understanding the relative importance of these variables can help prioritize resource allocation and intervention strategies in the areas within each county that are most susceptible to wildfires.

Los Angeles County Wildfire Drivers

The most important feature in the Random Forest Classifier model for predicting wildfire occurrence in Los Angeles County was the Normalized Difference Vegetation Index (NDVI), with a Mean Decrease Accuracy (MDA) of 217. This result aligns with previous wildfire prediction literature discussed in Chapter 2. It is also intuitive as NDVI measures the amount of live vegetation on the land. Figure 6 compares the average NDVI value for fire versus no-fire occurrences. Los Angeles County typically experienced fires when NDVI levels were higher on average compared to no fire occurrences, with little to no fire ignitions at NDVI less than 0.25. This finding is consistent with the notion that vegetation serves as fuel for wildfires, so areas with high NDVI values are more likely to experience intense fires when ignition occurs.

Figure 6

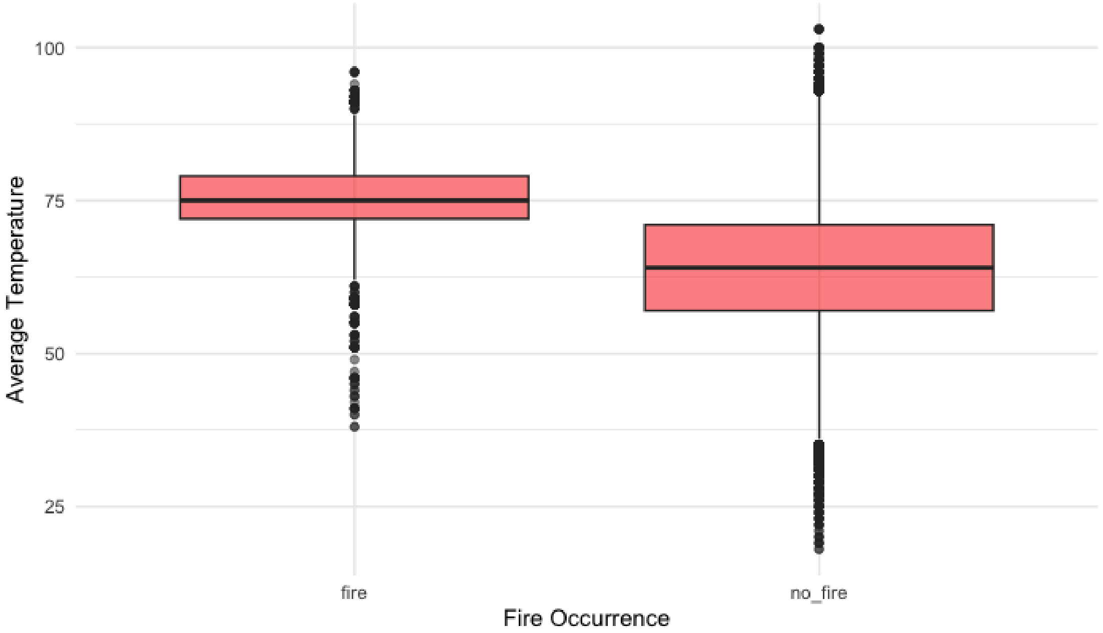
Boxplot of NDVI Distribution by Fire Occurrence in Los Angeles County



The second most important feature in Los Angeles County was average temperature, with an MDA score of 205. This score is only slightly lower than that of NDVI, indicating that average temperature is also highly important. Figure 7 shows a boxplot of Riverside County's average temperature for fire and no-fire occurrences. The figure suggests that fires typically happen at higher temperatures, with a median temperature of around 75 degrees Fahrenheit, compared to no-fire occurrences, which have a lower median temperature of around 64 degrees Fahrenheit. Higher temperatures increase the rate of soil moisture evaporation, increasing the likelihood of dry vegetation, thus escalating flammability. This correlation between temperature and wildfire risk is well-documented in wildfire studies, as hotter, drier conditions typically accelerate fire hazards (Environmental Defense Fund, n.d.).

Figure 7

Boxplot of Temperature Distribution by Fire Occurrence in Los Angeles County



Despite their lower MDA scores, other variables such as wind speed and precipitation still played a critical role in the overall predictive model. While they did not rank as highly individually, their interaction with NDVI and temperature can significantly influence fire occurrence. For instance, high winds can spread fires rapidly, and the absence of precipitation can lead to drier conditions, further facilitating fire spread (Fendell and Wolff, 2001). The significant roles of NDVI and temperature in predicting wildfires highlight the need for targeted vegetation management and climate adaptation strategies in Los Angeles County.

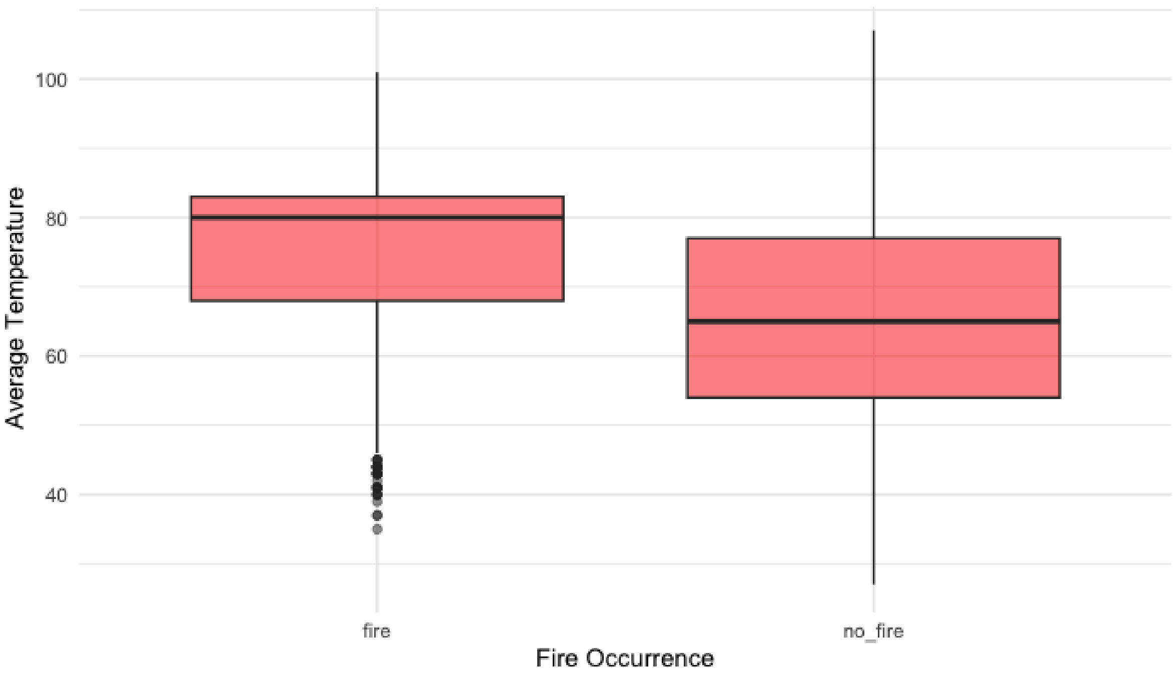
Riverside County Wildfire Drivers

In Riverside County, the most important feature in the Random Forest Classifier model was average temperature, with an MDA score of 215. This result is consistent with findings in

Los Angeles County, emphasizing the role of temperature in wildfire risk. Figure 8, a boxplot of temperature distribution across fire scenarios, shows a similar trend to that of Los Angeles, where higher average temperatures are typically associated with fire occurrences, with average temperatures of 80 degrees Fahrenheit and 65 degrees Fahrenheit for fire and no-fire occurrences respectively. The higher average temperatures in Riverside County compared to Los Angeles County contribute to the increase in its susceptibility to fire ignition, likely making temperature a more indicative variable for wildfire occurrence.

Figure 8

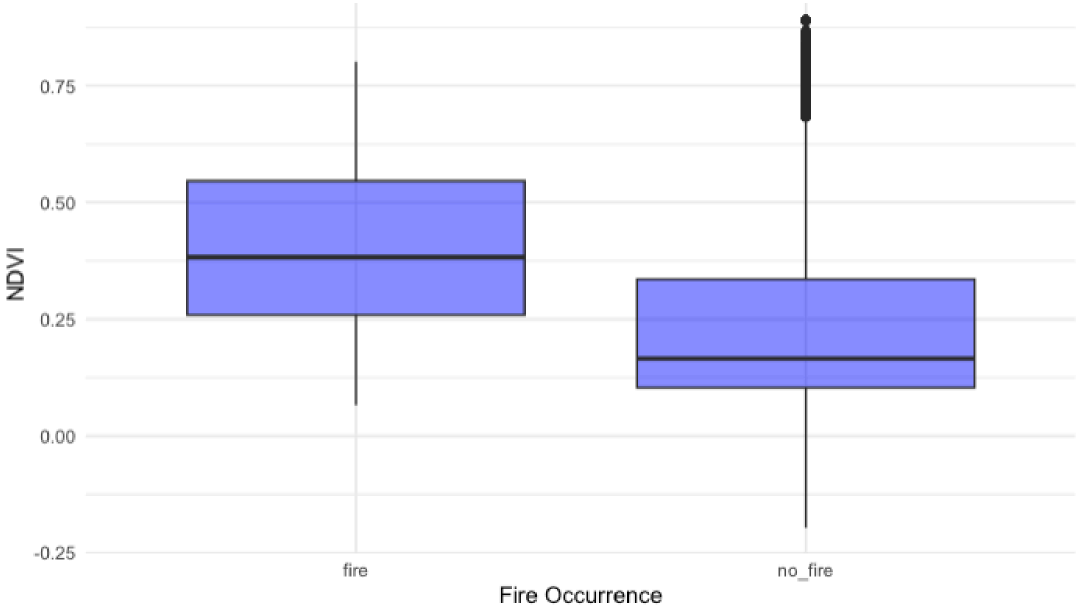
Boxplot of Temperature Distribution by Fire Occurrence in Riverside County



The second most important feature for Riverside County was NDVI, with an MDA score of 203. Similar to Los Angeles, vegetation density (measured by NDVI) was a crucial factor in determining wildfire risk. Figure 9 compares the average NDVI value for fire versus no-fire occurrences. Riverside County typically experienced fires when NDVI levels were above 0.25, with little to no fire ignitions at lower NDVIs. High NDVI values indicate more fuel and desirable conditions for wildfires. Additionally, the secondary, but significant roles of wind speed and precipitation are also present in the results for Riverside County. While these features had lower individual importance scores, they can be valuable in understanding the overall risk for fire occurrence and spread when evaluated alongside temperature and NDVI. The results suggest that wildfire mitigation strategies in Riverside County could benefit from monitoring temperature variation and managing vegetation.

Figure 9

Boxplot of NDVI Distribution by Fire Occurrence in Riverside County



Discussion of Results

The analysis revealed that NDVI and average temperature were the most critical drivers of wildfire occurrence in both Los Angeles and Riverside counties from 2016 to 2020. These findings are consistent with previous research and an intuitive understanding of the dynamics of wildfire occurrence. NDVI's importance highlights the significant role of vegetation as a primary fuel source for fires, while average temperature's significance highlights the impact of climatic conditions on fire risk.

The slightly different rankings of these features between the two counties may be attributed to regional variations in climate and vegetation types. For instance, Los Angeles County's varied terrain and microclimates could cause NDVI to have a marginally higher influence (LaDochy and Witiw, 2023). Riverside County's generally higher temperatures might increase the importance of temperature in fire prediction.

Additionally, the interaction between the features of primary importance (NDVI and average temperature) and secondary features (wind speed and precipitation) demonstrates the nuanced nature of wildfire prediction. Even though wind speed and precipitation individually had lower importance scores, their role in conjunction with NDVI and temperature is crucial. Intuitively, the absence of rain and the presence of high wind speed can exacerbate dry conditions required for fire ignition. This nuanced understanding of wildfire drivers underscores the importance of integrated management approaches that consider a multitude of environmental factors and their interactions with one another.

Objectives Evaluation

The first objective of this project was to determine the counties in Southern California that are most highly impacted by wildfires. This was achieved through the comparison of average fire radiative power and fire occurrence across counties as outlined in Chapter 3. The second objective was to construct models to predict fire occurrence using various machine-learning methods and compare their accuracy. This objective was achieved with k-Nearest Neighbors (KNN), Support Vector Machines (SVM), Random Forests, and Decision Trees models using the methods outlined in Chapter 3. Random Forests Classifier was determined to be the best-performing model for both Los Angeles and Riverside counties. The third and final objective of this project was to identify county-specific factors within the datasets that were the most significant contributors to wildfire risk. This objective was achieved and discussed in Chapter 4.

Conclusion

In summary, Chapter 4 discussed the model selection process and feature importance of the machine learning models trained on wildfire data included in this report. Random Forest Classifier was determined to be the best model for predicting wildfire occurrence in both Los Angeles and Riverside counties. With this model selected, average temperature and NDVI were found to be the most important variables in predicting fire, with slight differences in ranking between the two counties. Lastly, the project's objectives were assessed for completion and all three goals were met. The findings discussed in Chapter 4 provide a strong foundation for valuable insights into developing more accurate and localized wildfire prediction models which can significantly enhance wildfire management strategies in Southern California. In Chapter 5, the limitations of the project and suggestions for future research and wildfire management will be discussed.

Chapter 5

This chapter synthesizes the findings from the previous chapters, discussing the implications for wildfire management and providing recommendations for future research. The discussion will be structured around the project's objectives, highlighting the practical applications of the results and acknowledging the limitations of the study.

Summary of Findings

This study aimed to accomplish several objectives related to understanding how machine learning techniques can be applied to wildfire occurrence prediction. The following sections will summarize the findings of this report as organized by the objectives.

Objective 1: Determine the Southern California Counties Most Impacted by Wildfires

A comparative analysis of fire occurrences and average fire radiative power, discussed in Chapter 3, revealed that Los Angeles and Riverside counties were the most significantly affected by wildfires. Figures 10 and 11 show the distribution of fires from 2016 to 2020 across Los Angeles and Riverside counties, respectively. These maps reveal wildfire hotspots. In Los Angeles County, fires were concentrated in the Angeles National Forest. San Bernadino County displayed a similar trend, with the majority of fires occurring in areas with less development and more vegetation such as the San Bernadino National Forest. Additionally, temporal analysis showed season patterns, with notable spikes in fire ignition during the late summer months and early autumn. Both counties experienced the largest amount of fires in 2020, which may be attributed to extreme heat and prolonged periods of drought.

Figure 10

Distribution of Fire Occurrences in Los Angeles County From 2016 to 2020

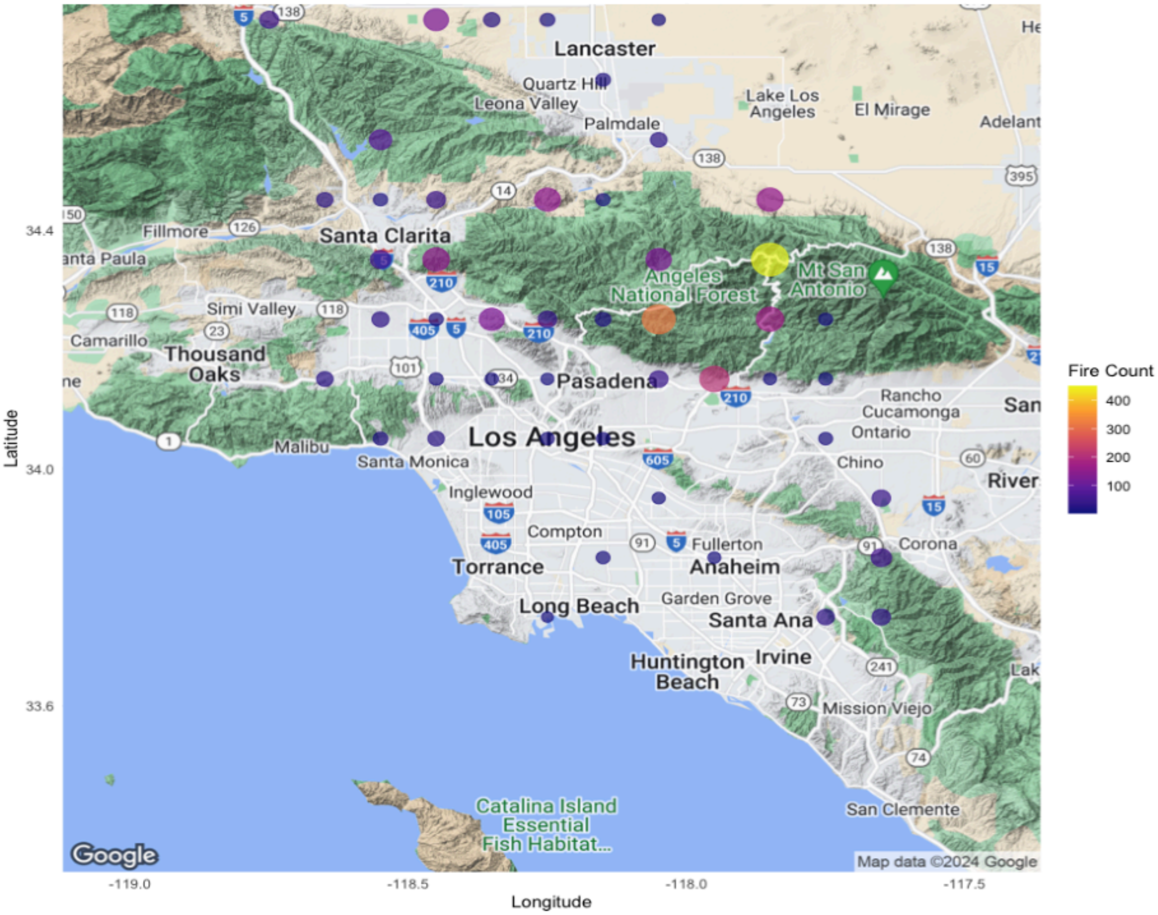
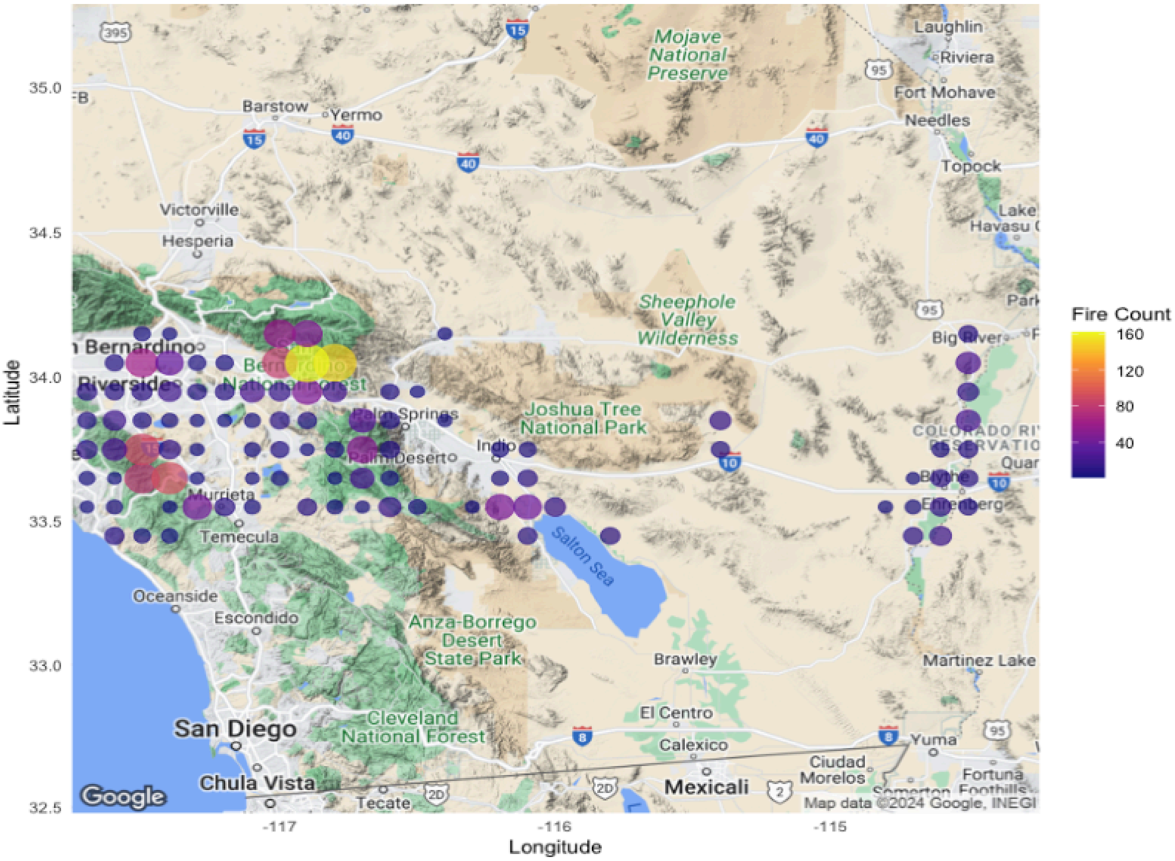


Figure 11

Distribution of Fire Occurrences in Riverside County From 2016 to 2020



Objective 2: Construct and Compare Machine Learning Models

The modeling selection process included the comparison of four different machine-learning techniques. The Random Forest Classifier model emerged as the best-performing, demonstrating the highest AUROC and F1 scores compared to K-Nearest Neighbors (KNN), Support Vector Machines (SVM), and Decision Trees. The Random Forest model's superiority in handling complex wildfire datasets was evident from its high-performance metrics: an AUCROC score of 0.990 an F1 score of 0.964 for Los Angeles County, and an AUCROC score of 0.959 and F1 score of 0.911 for Riverside County. These results highlight the

Random Forest model's ability to accurately discriminate between fire and no-fire occurrences, providing reliable predictions that are suitable for effective wildfire management.

Objective 3: Identify County-Specific Factors Influencing Wildfire Risk

The third objective was to identify county-specific factors influencing wildfire risk. The feature importance analysis highlighted NDVI and average temperature as the most critical drivers of wildfire occurrences in both counties for the given datasets. NDVI, representing vegetation density, increases the availability of fuel for wildfires, making areas with higher NDVI values more susceptible to fire. Average temperature also plays a significant role, as higher temperatures contribute to drier conditions, increasing the likelihood of fire ignition and spread. Our analysis revealed that while NDVI was slightly more influential in the Los Angeles County mode, the opposite was true for Riverside County. This may be attributed to the hotter, drier climate present in Riverside County compared to Los Angeles County. Wind speed and precipitation had lower individual importance scores. However, it could be argued that their interaction with NDVI and temperature is still significantly influential in wildfire dynamics.

Recommendations for Wildfire Management

Based on the findings of this report, several recommendations can be made for wildfire management. First, integrating county-specific models into existing wildfire management systems can enhance prediction accuracy. Localized models that account for unique regional factors provide tailored risk assessments, improving preparedness and response strategies. Wildfire management strategies can be significantly improved by developing protocols for using these models in real-time and training emergency fire responders on interpreting the model’s outputs.

Targeted vegetation management is also crucial. Implementing controlled burns and vegetation maintenance in high NDVI areas identified by region-specific models can improve environmental conditions by reducing the fuel available for wildfires, thus mitigating risk. Regular monitoring of vegetation density and health in wildfire hotspots using satellite imagery and remote sensing technologies, such as the ones utilized by this project, can help pinpoint areas that need immediate intervention before the peak of the fire season.

Given the potential of climate change to exacerbate environmental conditions that lead to wildfires, climate adaptation strategies should be prioritized. Monitoring temperature trends and using climate projections to anticipate periods of higher wildfire risk can inform long-term wildfire management resource planning. Optimizing the deployment of fire abatement resources based on model predictions is another important recommendation. Prioritizing high-risk areas, such as Angeles National Forest and San Bernadino National Forest, for resource deployment and environmental remediation during peak fire seasons and strategically positioning wildland fire teams based on model predictions can ensure rapid response times and enhance wildfire management.

Limitations of the Study

Despite the valuable insights gained, the study has several limitations. Data quality and completeness were significant issues, particularly with weather and NDVI records containing gaps and inconsistencies. Although sufficient efforts were made to address missing data, the accuracy of predictions could be impacted. Considering the critical role of NDVI and temperature in wildfire prediction, it is important to enhance the quality and frequency of data collection. Additionally, the study focused on data from four years, 2016 to 2020. A longer period could provide more comprehensive insights into wildfire trends and patterns.

Recommendations for Future Research

Future research should consider expanding the geographic scope to include more counties. Conducting cross-regional analyses to identify commonalities and differences in wildfire risk factors could provide insights into the intricacies of fire behavior. Using larger datasets to capture trends over decades and conducting longitudinal studies can help us understand how wildfire risk factors evolve so we can anticipate changes and adjust management strategies accordingly. Future research could also incorporate a wider range of environmental and meteorological variables such as wind direction, soil moisture, relative humidity, and ozone ratio.

Many modern data science tools were not utilized in this report. Future research should explore more advanced machine learning and AI techniques, such as deep learning models that could utilize satellite imagery to detect small fires before they spread. Incorporating real-time data from satellite imagery, weather stations, and other sources to create dynamic prediction models can also improve model adaptability and prediction accuracy.

Conclusion

This case study has demonstrated the value of localized, data-driven approaches in enhancing wildfire prediction and management. By focusing on county-specific factors and employing robust machine learning models, this study achieved high wildfire prediction accuracy with F1-Scores of 0.964 and 0.911 for Los Angeles and Riverside Counties respectively using Random Forest Classifier models. The report provides actionable insights that can significantly improve wildfire preparedness and response strategies within Southern California as well as a path forward for future research. Continued studies and collaboration across disciplines are essential for developing more accurate and effective wildfire management practices, ensuring the safety and resilience of communities in the face of increasing wildfire threats posed by climate change.

References

Author. (n.d.). *Statistics | CAL FIRE*. https://www.fire.ca.gov/our-impact/statistics

California, S. O. (2023, October 14). *Impending Santa Ana winds bring increased fire risk to southern California | Cal OES News*. https://news.caloes.ca.gov/impending-santa-ana-winds-bring-increased-fire-risk-to-southern-california/

Castrejon, D. J., Wang, C., Osmak, D., Kukadiya, B., Liu, L., Giraldo, M., & Jiang, X. (2023). Machine Learning-based California Wildfire Risk Prediction and Visualization. *2023 International Conference on Machine Learning and Applications (ICMLA)*. https://doi.org/10.1109/icmla58977.2023.00182

Constance. (2022, April 20). *New research uncovers the complex drivers of wildfire ignition in California*. California Ecosystem Climate Solutions. https://california-ecosystem-climate.solutions/new-research-uncovers-the-complex-drivers-of-wildfire-ignition-events-in-california/

Fendell, F., & Wolff, M. (2001). Wind-Aided fire spread. In *Elsevier eBooks* (pp. 171–223). https://doi.org/10.1016/b978-012386660-8/50008-8

Hernandez, K., & Hoskins, A. B. (2024). Machine learning algorithms applied to wildfire data in California's central Valley. *Trees, Forests and People*, *15*, 100516. https://doi.org/10.1016/j.tfp.2024.100516

Hooper, C. (2023, February 21). Defense giant Lockheed Martin eyes new opportunities in wildfire fighting. *Forbes*. https://www.forbes.com/sites/craighooper/2023/02/21/defense-giant-lockheed-martin-corporation-eyes-new-opportunities-in-wildfire-fighting/

Jain, P., Coogan, S. C., Subramanian, S. G., Crowley, M., Taylor, S. W., & Flannigan, M. D. (2020). A review of machine learning applications in wildfire science and management. *Environmental Reviews*, *28*(4), 478–505. https://doi.org/10.1139/er-2020-0019

Janiec, P., & Gadal, S. (2020). A comparison of two machine learning classification methods for remote sensing predictive modeling of the forest fire in the North-Eastern Siberia. *Remote Sensing*, *12*(24), 4157. https://doi.org/10.3390/rs12244157

Jerrett, M., Jina, A. S., & Marlier, M. E. (2022). Up in smoke: California's greenhouse gas reductions could be wiped out by 2020 wildfires. *Environmental Pollution*, *310*, 119888. https://doi.org/10.1016/j.envpol.2022.119888

Kauffman, E. (n.d.). Atlas of the Biodiversity of California. In *A Remarkable Geography*. https://www.coastal.ca.gov/coastalvoices/resources/Biodiversity_Atlas_Climate_and_Topography.pdf

Khorshidi, M. S., Dennison, P. E., Nikoo, M. R., AghaKouchak, A., Luce, C. H., & Sadegh, M. (2020). Increasing concurrence of wildfire drivers tripled megafire critical danger days in Southern California between1982 and 2018. *Environmental Research Letters*, *15*(10), 104002. https://doi.org/10.1088/1748-9326/abae9e

LaDochy, S., & Witiw, M. (2023). The Most Climatically Diverse State in the United States. In *Fire And Rain* (pp. 1–9). https://doi.org/10.1007/978-3-031-32273-0_1

Li, S., & Banerjee, T. (2021). Spatial and temporal pattern of wildfires in California from 2000 to 2019. *Scientific Reports*, *11*(1). https://doi.org/10.1038/s41598-021-88131-9

MacDonald, G., Wall, T., Enquist, C. a. F., LeRoy, S. R., Bradford, J. B., Breshears, D. D., Brown, T., Cayan, D., Dong, C., Falk, D. A., Fleishman, E., Gershunov, A., Hunter, M., Loehman, R. A., Van Mantgem, P. J., Middleton, B. R., Safford, H. D., Schwartz, M. W.,

& Trouet, V. (2023). Drivers of California's changing wildfires: a state-of-the-knowledge synthesis. *International Journal of Wildland Fire*, *32*(7), 1039–1058. https://doi.org/10.1071/wf22155

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

Pham, B. T., Jaafari, A., Avand, M., Al-Ansari, N., Du, T. D., Yen, H. P. H., Van Phong, T., Nguyen, D. H., Van Le, H., Mafi-Gholami, D., Prakash, I., Thuy, H. T., & Tuyen, T. T. (2020b). Performance evaluation of machine learning methods for forest fire modeling and prediction. *Symmetry*, *12*(6), 1022. https://doi.org/10.3390/sym12061022

Pham, K., Ward, D., Rubio, S., Shin, D., Zlotikman, L., Ramirez, S., Poplawski, T., & Jiang, X. (2022). California Wildfire Prediction using Machine Learning. *2022 21st IEEE International Conference on Machine Learning and Applications*. https://doi.org/10.1109/icmla55696.2022.00086

Sayad, Y. O., Mousannif, H., & Moatassime, H. A. (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

Study finds climate change to blame for Record-Breaking California wildfires | August 8, 2023 | Drought.gov. (2023, August 8). Drought.gov. https://www.drought.gov/news/study-finds-climate-change-blame-record-breaking-california-wildfires-2023-08-08

Thach, N. N., Ngo, D. B., Xuan-Canh, P., Hong-Thi, N., Thi, B. H., Nhat-Duc, H., & Dieu, T. B. (2018). Spatial pattern assessment of tropical forest fire danger at Thuan Chau area

(Vietnam) using GIS-based advanced machine learning algorithms: A comparative study. *Ecological Informatics*, *46*, 74–85. https://doi.org/10.1016/j.ecoinf.2018.05.009

Which California counties are most at risk for wildfires? (2024, March 13). GovTech. https://www.govtech.com/em/preparedness/which-california-counties-are-most-at-risk-for-wildfires

Wildfires. (n.d.). Environmental Defense Fund. https://www.edf.org/climate/heres-how-climate-change-affects-wildfires#:~:text=Increasing%20severe%20heat%20and%20drought,and%20fallen%20branches%20into%20kindling

Yu, G., Feng, Y., Wang, J., & Wright, D. B. (2023). Performance of fire danger indices and their utility in predicting future wildfire danger over the conterminous United States. *Earth's Future*, *11*(11). https://doi.org/10.1029/2023ef003823

Appendix

Link to code: https://github.com/reilley-m/CapstoneProjectMSDS2024