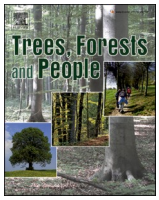
Trees, Forests and People 15 (2024) 100516

Contents lists available at ScienceDirect

Trees, Forests and People

journal homepage: www.sciencedirect.com/journal/trees-forests-and-people



Machine learning algorithms applied to wildfire data in California's central valley☆

Kassandra Hernandez, Aaron B. Hoskins *

Department of Mechanical Engineering, California State University, Fresno, USA

ARTICLE INFO

Keywords:
Wildfire prediction
Machine learning
Decision trees
Random forest
Neural networks
Naïve Bayes

ABSTRACT

This study focuses on using Machine Learning methods to predict wildfires within California's Central Valley. The specific areas within the Central Valley were Yosemite Valley, Sequoias, and Kings Canyon since these areas can be considered wildfire hotspots. This topic is relevant since California has seen an increase in wildfires with an increase in annual forest burned areas to +172 % from 1996 to 2021 (ABC 2024). The algorithms selected were based on previous research that conducted similar studies. From this research it is hypothesized that the best performing algorithm for predicting wildfires would be Random Forest. The novelty in this study stems from focusing on the specific areas mentioned above, which is where many wildfires have occurred throughout the years. The overall goal is to determine the best machine learning algorithm to predict wildfires in the Central Valley and take the results to improve upon wildfire prevention within these regions. The methods implemented included Decision Trees, Random Forest, Naïve Bayes, and Neural Networks. The dataset was gathered from the following satellite data which include MERRA-2 and USGS Landsat 8 along with fire history from 2012 to 2023 within these regions. Utilizing the dataset in the following two variations were a random split and a chronological split of training and testing sets. The best-performing algorithm using this dataset was Decision Trees at 550 maximum splits with an F1-Score of 0.689. The F1-Score ranges between 0 and 1 with a score of 0.7 or higher being deemed a good model to be used for predictions. The conclusion that could be determined from this result is that the randomized data has better predicting power over a chronologically split dataset. This can be seen in the confusion matrices for the chronological split dataset having zero true positive values in all the methods except for Naïve Bayes. Overall, the results show that Decision trees with a larger maximum split in the leaf nodes result in a more accurate prediction of whether a fire will occur within the given regions. The conclusion that can be made from this result is that Decision Trees can be a useful tool in predicting wildfires in California's Central Valley. The applications of this research would be the ability to use the information gained in this study to aid in optimizing resources in wildfire prevention within these areas.

Introduction

California has had an increase in wildfires in recent years, which can be linked to climate change (Predicting and planning for forest fires | PreventionWeb 2023) and (ABC 2024). Climate change factors include drought, lightning, temperature, wind speed, and lack of precipitation (Mansoor et al., 2022). California has also been named for having large destructive wildfires that have claimed many lives, and these concerns are a focus of this research (Wildfires Kill Unprecedented Numbers of Large Sequoia Trees (U.S. National Park Service 2023). California can be broken into Northern and Southern California where each region has its own designated fire season (California Department of Forestry and Fire Protection | CAL FIRE 2023). The California wildfire season is between May and October for the Southern California region based on the historical trends of past wildfires. The Northern California wildfire season is from July to October (wfca_teila 2022). Wildfires continue throughout the years, focusing on 2013 to 2022, showing that the total acres burned each year exceed 114,000 (When Is California Fire Season?, 2023). The trend also indicates that July has the most fire incidents, while September and October include the most burned acres (California

☆ This article is part of a special issue entitled: "Fire and environment. Issues and challenges" published at the journal Trees, Forests and People. Hyperlink : Trees, Forests and People Link: https://www.sciencedirect.com/journal/trees-forests-and-people/special-issue/10MWNM0C08H

* Corresponding author.
E-mail address: ahoskins@csufresno.edu (A.B. Hoskins).

https://doi.org/10.1016/j.tfp.2024.100516

Available online 8 February 2024

Department of Forestry and Fire Protection | CAL FIRE 2023). This is when factors such as heat, drought, wind, and human activity increase the risk of fires. Measures are being taken to prevent wildfires, such as vegetation management, wildfire prevention grants, increase in fire training, etc. (wfca_teila 2022; California Department of Forestry and Fire Protection | CAL FIRE 2023) and (When Is California Fire Season?, 2023). Even with these measures put in place, it is hard to predict where a fire will pop up since many of the factors stated are variable. This is where Machine Learning (ML) has been utilized to predict wildfire locations based on the provided variable features (Sulova and Jokar Arsanjani, 2021). Other research has focused on different areas within California or the whole region of California (Pham et al., 2022; A. Malik et al., 2021). Focusing on the Central Valley as seen in Fig. 1 will advance research and provide modeling for this region.

ML algorithms are a helpful tool in classifying forest fire activity by using data gathered from various satellites, which consists of meteorological and vegetation parameters. One main concern brought up by Vasconcelos et al. was the high number of false alarms with the ML models that only predict fires within the fire season (Vasconcelos et al., 2001). This study addresses this concern by testing multiple models and finding which model minimizes the false alarms. This research focuses on the specific areas in the Central Valley of Yosemite Valley, Sequoias, and Kings Canyon. Wildfires can occur anytime throughout the year, but this research focuses on the California wildfire season. The data will go through pre-processing of extracting data based on the specific region and time. Then, the data can be used on various classification algorithms utilized by other researchers that have shown accuracy in binary classification.

ML algorithms have been utilized to predict wildfires in regions where large amounts of wildfires are present. The research done by Arif et al. showed the importance of how necessary ML can be in predicting wildfires and the target areas where this can be applied (Arif et al., 2021). Of the many algorithms to choose from, the most common algorithms used were Support Vector Machine (SVM), Naïve Bayes, Decision Trees, Artificial Neural Networks (ANN), and Random Forest.

A. Sulova and J. Jokar Arsanjani used ML algorithms to predict wildfires in Australia from the summer season of 2019–2020. This research used a binary classification problem of fire vs. no fire occurrences (Sulova and Jokar Arsanjani, 2021). The specific variables used within many of the ML models were wind speed, temperature, soil moisture, soil depth, precipitation, population, Normalized Difference Vegetation Index (NDVI), land cover, global human modification, elevation, drought, and topography. Another classification method approach was used by Qiu et al., where the classification features were categorized into small, medium, large, and extreme wildfires (Qiu et al., 2022). Knowing what factors play a key role in predicting the location and time a wildfire will occur can be used to mitigate future wildfires. Sayad et al. took a similar approach in using the binary classification approach, with their region of focus being Canada's forest between 2013 and 2014 (Sayad et al., 2019).

Xie et al. used a similar approach of ML to determine the wildfire risk in the Liangshan Prefecture in China (Xie et al., 2022). These algorithms were utilized with a 70–30 split of training and testing data (Xie et al., 2022). The dataset was gathered from the Moderate Resolution Imaging Spectroradiometer (MODIS) satellite (Xie et al., 2022). The classification features were categorized into groups of a percentage of risk levels from

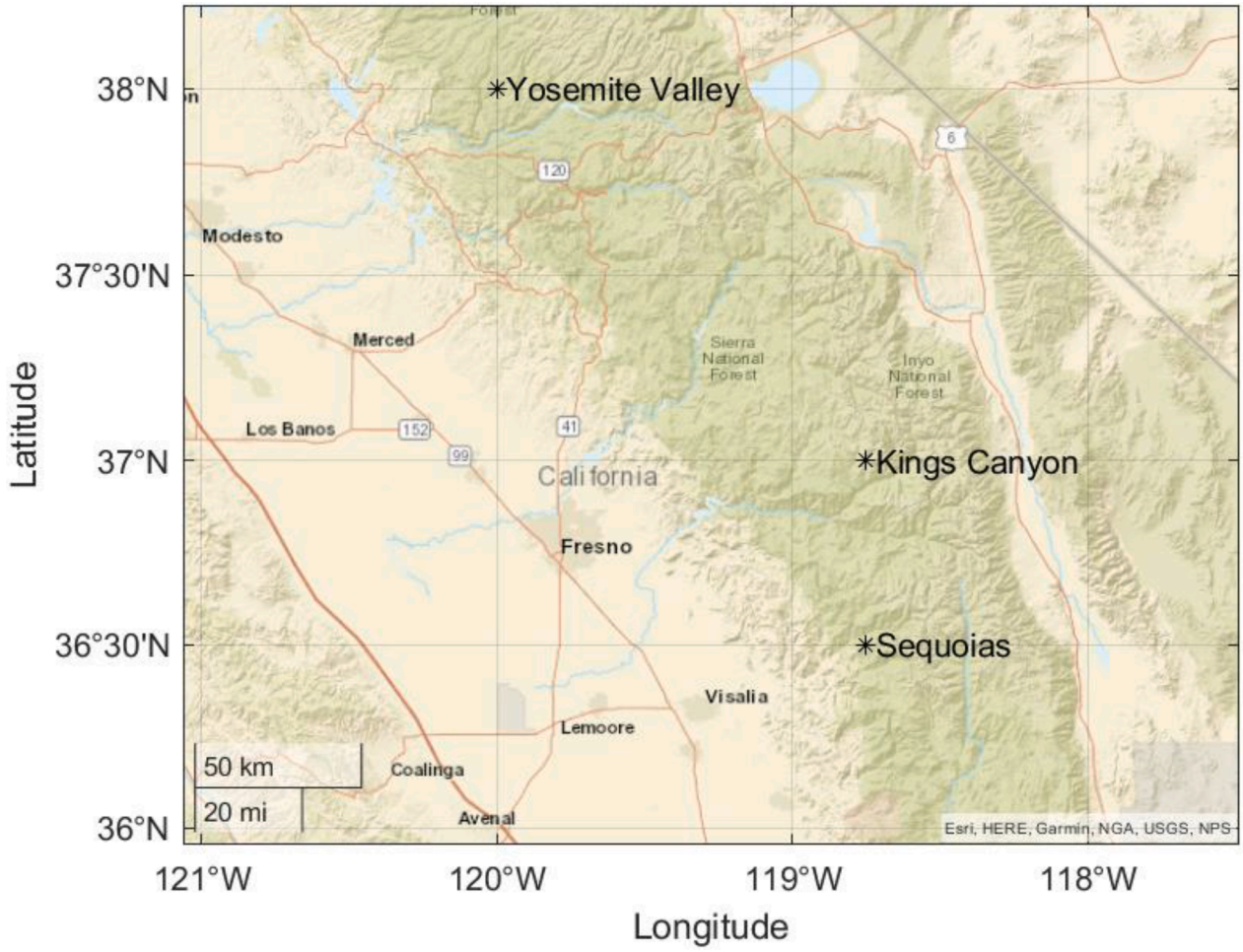


Fig. 1. Map of studied area.

very low to very high (Xie et al., 2022). Unlike the Sulova research of fire or no fire areas, this result showed an overall reduction in accuracy (Oliveira et al., 2012). This would suggest that the binary dependent variable is better than the percentage model (Oliveira et al., 2012). The research by Arif et al. highlighted that these models require up-to-date forecasting of weather, drought predictions, and socio-economic situations for these models to be of use (Arif et al., 2021). This does not diminish the need for such models to be created and explored, but rather the increasing need for data accuracy for better short- and long-term prediction (Wang et al., 2021). However, the research done by Wang et al. showed another method taken to predict wildfires [196]. This method varies from other methods researched by predicting the monthly burned area within the United States (Wang et al., 2021). The selected variables for this method were like other research, but this one tackled a more extensive area (Qiu et al., 2022).

Aside from similar classification methods, Preeti et al. used regression ML methods to predict wildfires. Artificial Neural Networks (ANN) and Linear Regression (Preeti et al., 2021) were the two methods. The linear regression approach was not considered because the stochastic nature of wildfires is generally hard to linearize. However, the research done by Abid determined that the ANN method has the unique ability to classify noisy data (Abid, 2021). Pang et al. noted that ANN is a better model with larger datasets to speed up computation time (Pang et al., 2022). Oliveira et al. compared a Multiple Linear Regression and Random Forest and showed that the classification model worked better than the regression model (Oliveira et al., 2012). While there is the possibility of using both classification and regression ML methods, this study will focus on the classification methods to predict wildfires in California's Central Valley. The research by Bhowmik et al. showed the important factors for predicting wildfires include geological factors such as elevation and the Normalized Difference Vegetation Index (NDVI) (Bhowmik et al., 2023).

Methods

The methods of fulfilling the objectives of creating a working model included data mining, preprocessing of the data, running classification models, and validation of the results. The data mining process included downloading the subsets of data within the period from 2012 to 2023 from the regions listed in Table 1.

The data was gathered from the Modern-Era Retrospective analysis for Research and Applications version 2 (MERRA-2) satellite obtained from NASA's EarthData website (A. GES DISC Dataset 2023). The MERRA-2 satellite was also used to gather rainfall data (A. GES DISC Dataset 2023). CAL FIRE was the source of the fire history from 2013 to 2022 (Incidents | CAL FIRE 2023) and (Wildfires Kill Unprecedented Numbers of Large Sequoia Trees (U.S. National Park Service 2023). Google Earth Engine was the source of NDVI (USGS Landsat 8 Level 2 2023). The data collection included the variables listed in Table 2. Since the data had various ranges depending on the source of the data, there needed to be a processing step for each source.

Table 1
Specific California regions being examined.

Longitude	Latitude	Region
−120	37.5	Yosemite
−120	38	Yosemite
−119.375	37.5	Yosemite
−119.375	38	Yosemite
−118.75	36	Sequoia
−118.75	36.5	Sequoia
−118.75	37	Kings Canyon
−118.75	37.5	Kings Canyon
−118.75	38	Yosemite
−118.125	36.5	Sequoia
−118.125	37	Kings Canyon
−118.125	37.5	Kings Canyon

Data processing

The cleaning of all the data was done at the largest temporal resolution, which was 16 days. The Normalized Difference Vegetation Index (NDVI) is the value with the largest temporal resolution. NDVI was calculated using the Google Earth Engine code to call the specific regions outlined in Table 1 and retrieve those specific data point values. The values range between −1 and 1 from the wavelengths of the image taken in that location from the Landsat 8 satellite. The wavelengths utilized for the equation include the Near Infrared band (NIR) and the surface reflection (Red). Eq. (1) was used to determine the NDVI values from the Landsat 8 satellite (A. Malik et al., 2021) and (Xie et al., 2022).

$$NDVI = \frac{(NIR - Red)}{(NIR + Red)} \quad (1)$$

The cleaning of the NDVI data included removing any data that consisted of blank cells where no data was collected on that day in that region. CAL FIRE data ranges of days where a fire was present were processed by inserting a zero or a one in that place. The zero represented no fire took place on that day in that area, and the one represented there was a fire. This was done to keep in line with the binary problem as seen in past research (Pham et al., 2022; A. Malik et al., 2021; Sulova and Jokar Arsanjani, 2021; Xie et al., 2022), and (A. Malik et al., 2021). The fire data was then reduced to match the exact times of the cleaned NDVI data. The MERRA-2 data needed to be summarized for a daily collection of data points. After this, the data could be matched to the 16 days of temporal resolution in the NDVI data format. This dramatically reduced the data from the original data having a resolution of every day in 2013–2022 to every 16th day from 2013 to 2022. This resulted in the final dataset used for the algorithms to have 73,770 records with 21 factors listed in Table 2 as variables.

Data sorting

The data was then sorted into a 70 % training and 30 % testing set in two different ways for modeling. This resulted in the testing set having 22,131 records of the 21 factors used. The training set had 51,639 records for the 21 factors used. It should be noted that the time range was not reduced in any way when separating into testing and training datasets. The modeling included distinguishing between the factors and the label. The label was determined to be the presence of fire in the region and the factor names can be seen in Table 2. This was done to ensure that the model would not use the presence of fire as a factor in the algorithms. The first way of separating the data into testing and training was to randomly distribute it to fit the 70–30 model. This was done once, and the same randomly generated matrices were used throughout the different ML algorithms. The second way of separating the data was by selecting the first 70 % of the data for training and the last 30 % for testing. This meant that the training data was 70 % of the sorted data by date and year, while the last 30 % was the remaining data from the years used. The purpose of separating the data in this fashion was to test the random set to the year set and ensure that the random did not have any bias.

ML methods

The ML methods used in this study were based on what was done in other research relating to this topic. The final methods used include Random Forest, Decision Trees, Naïve Bayes, and Neural Networks (Chandramouli et al., 2018). The Decision Tree method uses a tree structure to have easily interpretable results. The breakdown of the tree structure as seen in Fig. 2 comes from three terms which are the root node, branch node, and leaf nodes. The root node (A) is built from the data separated based on class and determines the feature where the most strongly determined class is considered first. The branch nodes (B) are features that are considered next to be the most valuable in determining

Table 2
Data collection.

Source	Variable	Description	Temporal	Units	Label ■ or Factor ●
CAL FIRE (Incidents | CAL FIRE 2023) and US National Park Services (Wildfires Kill Unprecedented Numbers of Large Sequoia Trees (U.S. National Park Service 2023)	Fire History	Fire incident days and location	Incident Days	days	■
	Latitude			Degrees North	1●
	Longitude			Degrees East	2●
MERRA-2 Instantaneous 3D 3-hourly data collection (A. GES DISC Dataset 2023)	Vertical Level	Vertical Pressure Level	3 hrs. Instantaneous	hPa	3●
	Air Temperature			K	4●
	Mass fraction of cloud liquid water	Ratio of water in the air		kg^{-1}	5●
	East Wind			$m\ s^{-1}$	6●
	Mass fraction of cloud ice water	Ratio of ice in the air		kg^{-1}	7●
	North Wind			$m\ s^{-1}$	8●
	Relative Humidity			kg^{-1}	9●
	Specific Humidity			kg^{-1}	10●
	Ozone Ratio			kg^{-1}	11●
MERRA-2 Time Averaged three hourly data collection [243]	Cumulative mass flux	Rate of mass flow	3 hrs. time average	$kg\ m^{-2}\ s^{-1}$	12●
	Convective Rainwater			$kg^{-1}\ s^{-1}$	13●
	Large scale rainwater			$kg^{-1}\ s^{-1}$	14●
	3D flux of ice convective precipitation			$kg\ m^{-2}\ s^{-1}$	15●
	3D flux of ice nonconvective precipitation			$kg\ m^{-2}\ s^{-1}$	16●
	3D flux of liquid convective precipitation			$kg\ m^{-2}\ s^{-1}$	17●
	3D flux of liquid nonconvective precipitation			$kg\ m^{-2}\ s^{-1}$	18●
	Evap subl convective precipitation			$kg^{-1}\ s^{-1}$	19●
	Evap subl nonconvective precipitation			$kg^{-1}\ s^{-1}$	20●
USGS Landsat 8 Level 2, Collection 2, Tier 1 (USGS Landsat 8 Level 2 2023)	Normalized Difference Vegetation Index (NDVI)		16 days	unitless	21●

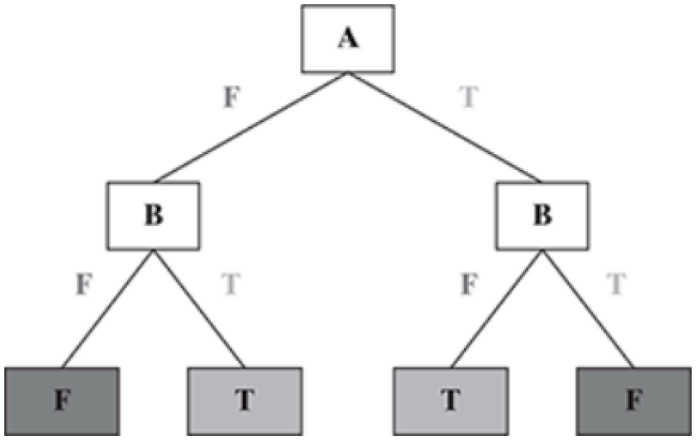


Fig. 2. Example decision tree (Chandramouli et al., 2018).

the class. The leaf node (gray blocks) is where the tree terminates and is where the label is determined (Chandramouli et al., 2018).

The Random Forest method is an ensemble classifier in where it creates several classifiers and combines them together. The breakdown of this method can be seen in Fig. 3. This method utilizes many uncomplicated trees to determine the final label. The basic steps that the algorithm goes through are to select subsets of M out of N features and uses a splitting principle of the M features for the D nodes. The tree will keep splitting until it grows to the maximum which can be defined or undefined. Once the tree has grown to its maximum there will be a selection of different subset of training data randomly with replacement N times. The final step is to have the N trees vote on the classifier (Chandramouli et al., 2018).

The basic steps in the Naïve Bayes classifier start by the construction of a frequency table and the algorithm calculates the cumulative probabilities through normalization. The understanding of Neural Networks stems from the biological neuron which is replicated into a digital form of a neuron. The neuron has three basic components which are the dendrites, soma, and axon. The dendrites receive signals, and the soma accumulates signals then fires when sufficient signal is received. The axon acts as a terminal for the information received from the soma to pass through. The basic steps of the Neural Network can be seen in Fig. 4

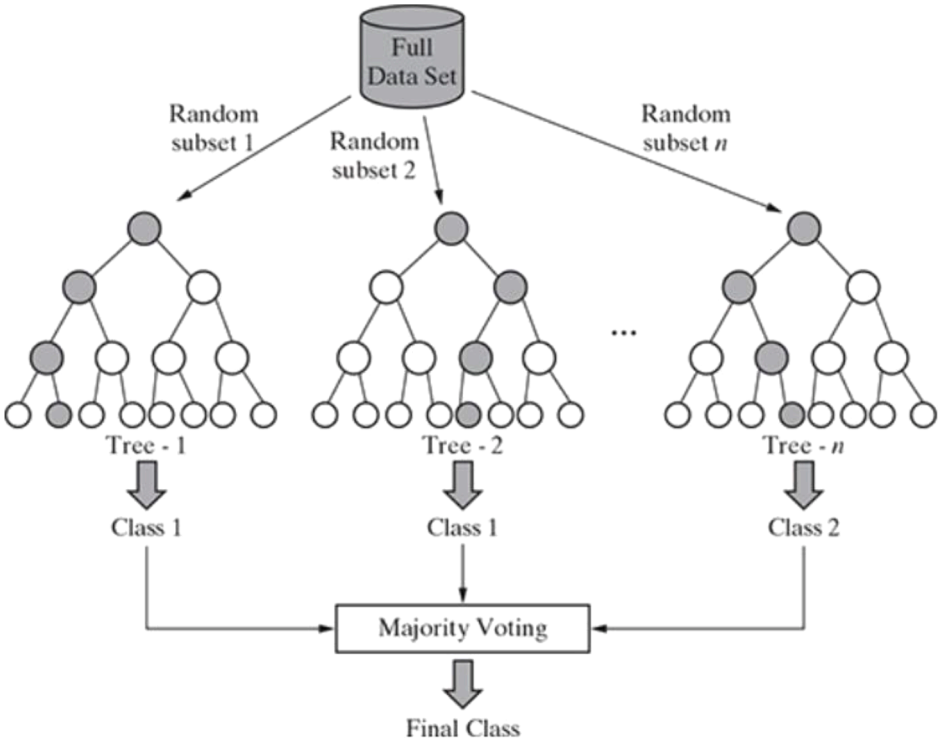


Fig. 3. Example random forest model (Chandramouli et al., 2018).

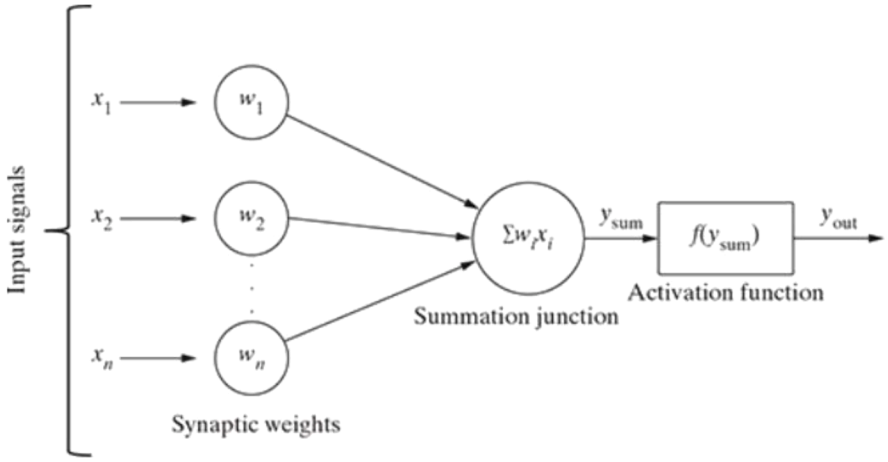


Fig. 4. Example of a neural network (Chandramouli et al., 2018).

(Chandramouli et al., 2018).

The specifications for each method used through MATLAB are defined as follows. Random Forest utilized the ADA Boost and Bag methods with the standard settings provided by MATLAB. The decision tree's hyperparameter tuning process included using the maximum number of splits varied from 250 to 550 in increments of 50. The Naïve Bayes model was created using the training features, with the dependent variable being the fire or no fire label. The Neural Network method was chosen based on the research conducted by Kondylatos et al. and Khanmohammadi to use the deep learning aspect of ML (Khanmohammadi et al., 2022) and (Kondylatos et al., 2022). This method utilized the training features with the dependent variable of the fire no fire label. The package and toolbox information to run these algorithms in MATLAB include the Deep Learning Toolbox along with the Statistics and Machine Learning Toolbox.

Performance analysis

To understand the results of any of the classification ML methods described was achieved by using a confusion matrix. This can help visualize the number of correctly classified data compared to the incorrectly classified data. A confusion matrix compares the predicted class to the true class. The true class in the case of predicting wildfires is based on the historical data of when a fire occurred. The predicted class is based on the randomization of the data in the training and predicting datasets to how the algorithm did on correctly choosing when a fire occurred based on the variables given. The results for a binary classification have four different possible classes, which are true positive (TP), true negative (TN), false positive (FP), and false negative (FN). To compare all the methods used, Eqs. (2)-(5) are listed below (Pérez-Porras et al., 2021).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (2)$$

$$Precision = \frac{TP}{TP + FP} \quad (3)$$

$$Recall = \frac{TP}{TP + FN} \quad (4)$$

$$F1score = 2\frac{Precision \times Recall}{Precision + Recall} \quad (5)$$

Eqs. (2)-(5) are implemented to understand the performance of the models and to compare them.

Results

The final comparison of determining the best model was based on the F1-Score. A good F1-Score is determined by balancing the precision and recall, resulting in a value of 0.7 or greater. The results for the models and the assessment of the models are shown in Tables 3 and 4. For understanding purposes, the results are put into two main categories of predicted class and true class. The class in this case is fire vs no fire so there are four possible outcomes. The "predicted no fire" means the algorithm predicted a fire not to occur while "actual no fire" means there was no fire that occurred based on the data. The "predicted fire" means the algorithm predicted a fire to occur and "actual fire" means a fire did occur based on the data.

Upon inspection of each of the methods chosen from the random and chronological datasets, it was shown that the best method was the Decision Trees at 550 max splits. The F1-Score was at a value of 0.689, which was the highest compared to any other method. While rounded, this barely meets the requirements for a good result; the confusion matrix shows the least false positives and false negatives. With regards to utilizing this method to predict wildfires and allocating resources for wildfire prevention, the number of false alarms must be minimal. Thus, the Decision Trees at 550 max splits the overall best method from the other trials. The worst performing models are the chronological datasets in Table 4 which show in all models the algorithm predicted no fires to occur when there were fire occurrences. This results in an F1-Score of NaN which means not a number. This is directly related to the predicted no fires and actual fire class containing all zeroes. With a zero in this cell the calculation for the F1-Score will result in a division of zero by zero thus giving the NaN result. These results are important to note since the objective is to predict when a fire is going to occur, and the chronological set fails to predict one occurrence in all algorithms used. Compared to the other research conducted, this result was vastly different since the better methods using the binary problem were the Random Forest method (A. Malik et al., 2021) and (Sulova and Jokar Arsanjani, 2021). The probable reasons for this type of result stem from including multiple separate regions that were used based on Table 1. Other research focused on one specific geographic region, which could be a reason why the other methods prevailed. The decision trees also provided an insight into what variables played the most significant roles in the algorithm. Those variables that played the most crucial role in the decision trees were NDVI, east wind, north wind, longitude, and latitude which can be seen in Figs. 5. These figures show the predictor importance of the first and last Decision trees. The x axis is labeled from 1 to 21 based on the order of feature names used which can be determined by examining the order number in column 6 of Table 2.

Discussion

The implication of these results is that the best working model to predict wildfires in the Central Valley is Decision Trees. Some challenges throughout developing the best model for prediction included narrowing down the maximum number of splits for the Decision Trees. The range was determined through trial and error and capping the region based on how the performances changed the F1-Score. Upon further investigation many features used in the predictive models have little to

Table 3
Trial using a randomized dataset from 2013 to 2022.

		Predicted No Fire	Predicted Fire	Accuracy	Precision	Recall	F1 Score
Decision Tree 250	Actual No Fire	21,266	93	0.974	0.757	0.377	0.503
	Actual Fire	480	290				
		Predicted No Fire	Predicted Fire				
Decision Tree 300	Actual No Fire	21,255	106	0.976	0.764	0.445	0.563
	Actual Fire	427	343				
		Predicted No Fire	Predicted Fire				
Decision Tree 350	Actual No Fire	21,249	112	0.976	0.759	0.458	0.571
	Actual Fire	416	352				
		Predicted No Fire	Predicted Fire				
Decision Tree 400	Actual No Fire	21,240	121	0.977	0.757	0.490	0.595
	Actual Fire	393	377				
		Predicted No Fire	Predicted Fire				
Decision Tree 450	Actual No Fire	21,239	122	0.977	0.763	0.509	0.611
	Actual Fire	378	392				
		Predicted No Fire	Predicted Fire				
Decision Tree 500	Actual No Fire	21,224	137	0.979	0.764	0.577	0.657
	Actual Fire	326	444				
		Predicted No Fire	Predicted Fire				
Decision Tree 550	Actual No Fire	21,200	161	0.980	0.752	0.635	0.689
	Actual Fire	281	489				
		Predicted No Fire	Predicted Fire				
Random Forest Bag	Actual No Fire	21,359	2	0.969	0.977	0.112	0.200
	Actual Fire	684	86				
		Predicted No Fire	Predicted Fire				
Random Forest ADA Boost	Actual No Fire	21,337	24	0.968	0.774	0.106	0.187
	Actual Fire	688	82				
		Predicted No Fire	Predicted Fire				
Naive Bayes	Actual No Fire	20,995	366	0.949	0.034	0.017	0.023
	Actual Fire	757	13				
		Predicted No Fire	Predicted Fire				
Neural Network	Actual No Fire	21,339	22	0.965	0.241	0.009	0.018
	Actual Fire	763	7				

Table 4
Trial using dataset in chronological order from 2013 to 2022.

		Predicted No Fire	Predicted Fire	Accuracy	Precision	Recall	F1 Score
Decision Tree 250	Actual No Fire	20,867	904	0.943	0.000	0.000	NaN
	Actual Fire	360	0				
		Predicted No Fire	Predicted Fire				
Decision Tree 300	Actual No Fire	20,859	912	0.943	0.000	0.000	NaN
	Actual Fire	360	0				
		Predicted No Fire	Predicted Fire				
Decision Tree 350	Actual No Fire	20,706	1065	0.936	0.000	0.000	NaN
	Actual Fire	360	0				
		Predicted No Fire	Predicted Fire				
Decision Tree 400	Actual No Fire	20,670	1101	0.934	0.000	0.000	NaN
	Actual Fire	360	0				
		Predicted No Fire	Predicted Fire				
Decision Tree 450	Actual No Fire	20,655	1116	0.933	0.000	0.000	NaN
	Actual Fire	360	0				
		Predicted No Fire	Predicted Fire				
Decision Tree 500	Actual No Fire	20,652	1119	0.933	0.000	0.000	NaN
	Actual Fire	360	0				
		Predicted No Fire	Predicted Fire				
Decision Tree 550	Actual No Fire	20,652	1119	0.933	0.000	0.000	NaN
	Actual Fire	360	0				
		Predicted No Fire	Predicted Fire				
Random Forest Bag	Actual No Fire	21,715	56	0.981	0.000	0.000	NaN
	Actual Fire	360	0				
		Predicted No Fire	Predicted Fire				
Random Forest ADA Boost	Actual No Fire	21,713	58	0.981	0.000	0.000	NaN
	Actual Fire	360	0				
		Predicted No Fire	Predicted Fire				
Naive Bayes	Actual No Fire	21,326	445	0.964	0.002	0.003	0.002
	Actual Fire	359	1				
		Predicted No Fire	Predicted Fire				
Neural Network	Actual No Fire	21,718	53	0.981	0.000	0.000	NaN
	Actual Fire	360	0				

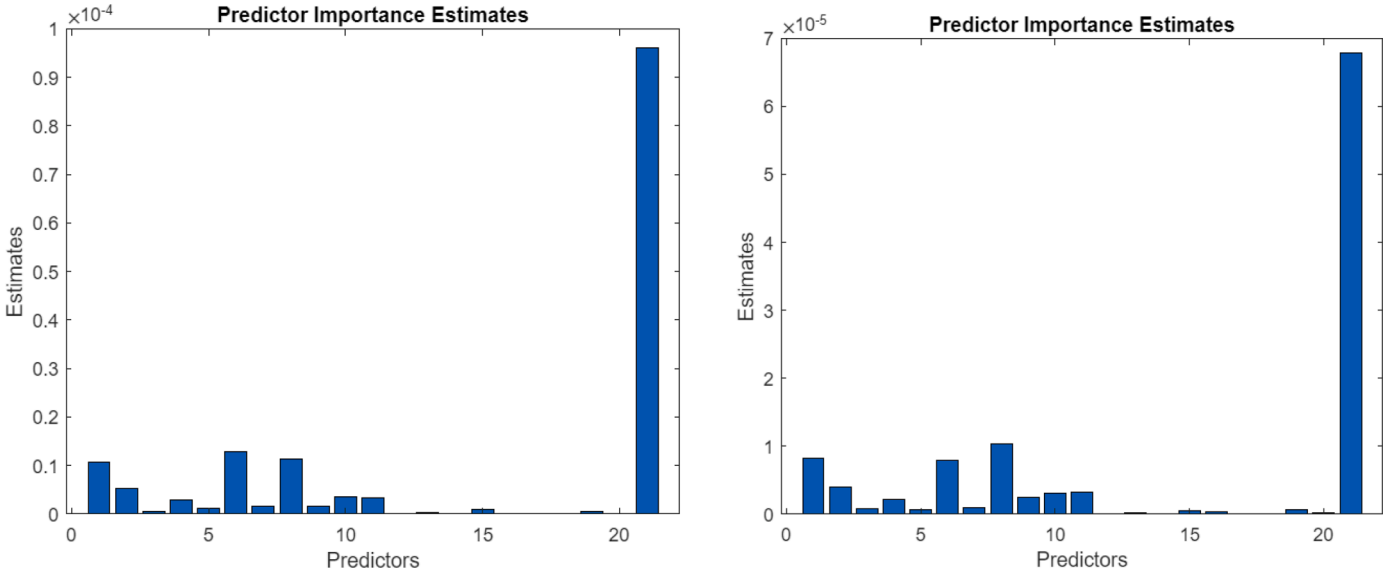


Fig. 5. Factor importance for decision tree 250 (left) and decision tree 550 (right).

no effect on the model's performance. In terms of how this information relates to practical use for the field of wildfire management involves where to focus the research and data collection. For example, if there was a better resolution on the NDVI data there would be more information to add to potentially increase the predictive power of the models. Since NDVI is by far the most important predictor. The conclusion that could be made from the NDVI significance is that the health of the vegetation is the important factor for the field of wildfire management to focus on. From these results the irrefutable evidence to mitigate wildfires lies in where the resources go to improve the health of the vegetation in the wildfire hotspot areas. The insight that can also be gained from determining the wildfire risk as seen in Fig. 6 which shows what regions have the highest risk. This can further justify where to allocate resources for wildfire prevention, especially in the high-risk areas. The limitations of further research going into focusing on the specific regions at a time based on the risk can dramatically reduce the performance by having smaller datasets.

The justification of using these models within this research comes down to the performance of the model and how this can be improved upon with further research. With this knowledge, the next steps in

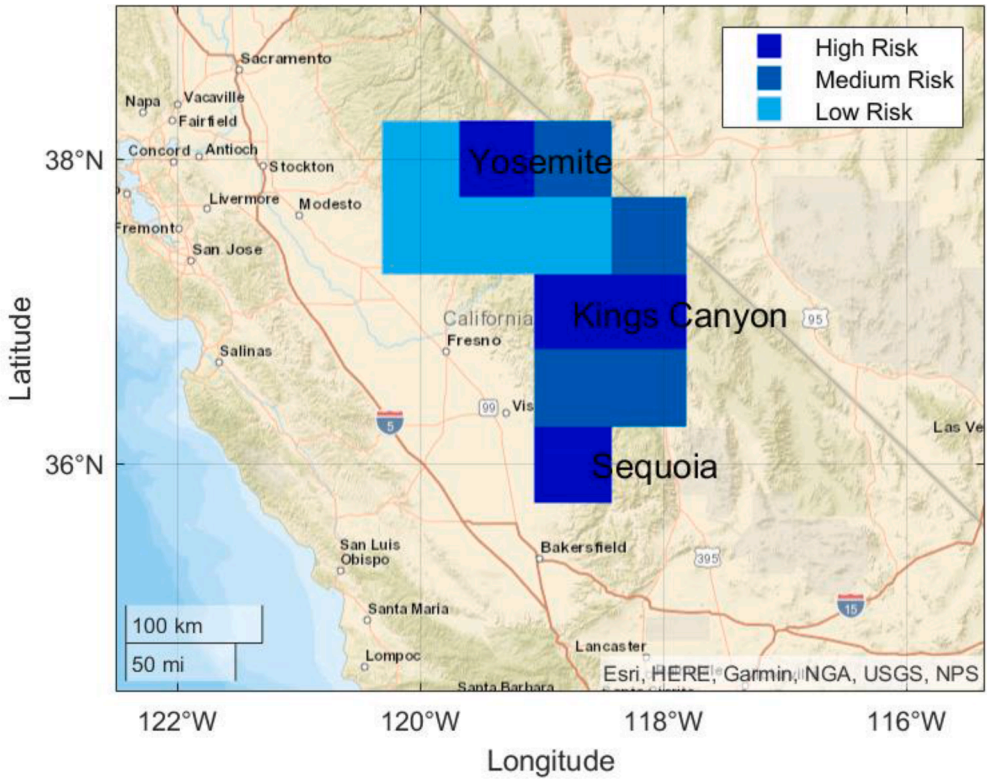


Fig. 6. Mapping of wildfire risk by region.

further research would stem from focusing on a single region at a time to produce more accurate results. The new research that can be done next includes focusing on a single area and including other types of fire. This would give a more area specific result that can be utilized in wildfire, criminal, and accidental fire prevention resources. With fine tuning the models to achieve better F1-Scores another next step would be to utilize these models in other wildfire prone areas.

Conclusion

This research aimed to apply known ML methods to predict wildfires within the Central Valley. The dataset was gathered from various satellites to collect weather, vegetation, and previous fire history data. The best method from the multiple methods tested was the Decision tree at 550 maximum splits with an F1-Score of 0.689. While this is an acceptable result, future research in this study would be to improve upon the number of false alarms reported within the confusion matrix. Some approaches that could prove useful would be to separate the longitude and latitude regions into their own dataset and run the algorithms individually, then combine the results and rerun the algorithms. This could help determine why the location is an important factor in the decision-making process.

CRediT authorship contribution statement

Kassandra Hernandez: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Aaron B. Hoskins:** Writing – review & editing, Supervision, Software, Resources, Project administration, Methodology, Investigation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

All data is publicly available and cited. Code available on request.

References

Malik, A., et al., 2021a. Data-driven wildfire risk prediction in northern California. Atmosphere (Basel) 12 (1). https://doi.org/10.3390/atmos12010109. Art. no. 1Jan.

Abid, F., 2021. A survey of machine learning algorithms based forest fires prediction and detection systems. Fire Technol. 57 (2), 559–590. https://doi.org/10.1007/s10694-020-01056-z. Mar.

Arif, M., et al., 2021. Role of machine learning algorithms in forest fire management: a literature review. J. Robot. Automat. 5 https://doi.org/10.36959/673/372. Feb.

Bhowmik, R.T., Jung, Y.S., Aguilera, J.A., Prunicki, M., Nadeau, K., 2023. A multi-modal wildfire prediction and early-warning system based on a novel machine learning framework. J. Environ. Manage. 341, 117908 https://doi.org/10.1016/j.jenvman.2023.117908. Sep.

Chandramouli, S., Dutt, S., Das, A.K., 2018. Machinel Learning. Pearson Education India [Online]. Available: https://learning.oreilly.com/library/view/machine-learning/9789389588132/xhtml/chapter007.xhtml#ch7_5.

Khanmohammadi, S., Arashpour, M., Golafshani, E.M., Cruz, M.G., Rajabifard, A., Bai, Y., 2022. Prediction of wildfire rate of spread in grasslands using machine learning methods. Environ. Modell. Software 156, 105507. https://doi.org/10.1016/j.envsoft.2022.105507. Oct.

Kondylatos, S., et al., 2022. Wildfire danger prediction and understanding with deep learning. Geophys. Res. Lett. 49 (17), e2022GL099368 https://doi.org/10.1029/2022GL099368.

Malik, A., Jalin, N., Rani, S., Singhal, P., Jain, S., Gao, J., 2021b. Wildfire risk prediction and detection using machine learning in San Diego, California. In: 2021 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/IOP/SCI), pp. 622–629. https://doi.org/10.1109/SWC50871.2021.00092. Oct.

Mansoor, S., et al., 2022. Elevation in wildfire frequencies with respect to the climate change. J. Environ. Manage. 301, 113769 https://doi.org/10.1016/j.jenvman.2021.113769. Jan.

Oliveira, S., Oehler, F., San-Miguel-Ayanz, J., Camia, A., Pereira, J.M.C., 2012. Modeling spatial patterns of fire occurrence in Mediterranean Europe using multiple regression

and random forest. For. Ecol. Manage. 275, 117–129. https://doi.org/10.1016/j.foreco.2012.03.003. Jul.
Pang, Y., et al., 2022. Forest fire occurrence prediction in china based on machine learning methods. Remote Sens. (Basel) 14 (21). https://doi.org/10.3390/rs14215546. Art. no. 21Jan.
Pérez-Porras, F.-J., Triviño-Tarradas, P., Cima-Rodríguez, C., Meroño-de-Larriva, J.-E., García-Ferrer, A., Mesas-Carrascosa, F.-J., 2021. Machine learning methods and synthetic data generation to predict large wildfires. Sensors 21 (11). https://doi.org/10.3390/s21113694. Art. no. 11Jan.
Pham, K., et al., 2022. California wildfire prediction using machine learning. In: 2022 21st IEEE International Conference on Machine Learning and Applications (ICMLA), pp. 525–530. https://doi.org/10.1109/ICMLA55696.2022.00086. Dec.
Preeti, T., Kanakaraddi, S., Beelagi, A., Malagi, S., Sudi, A., 2021. Forest fire prediction using machine learning techniques. In: 2021 International Conference on Intelligent Technologies (CONIT), pp. 1–6. https://doi.org/10.1109/CONIT51480.2021.9498448. Jun.
Qiu, L., Chen, J., Fan, L., Sun, L., Zheng, C., 2022. High-resolution mapping of wildfire drivers in California based on machine learning. Sci. Total Environ. 833, 155155 https://doi.org/10.1016/j.scitotenv.2022.155155. Aug.
Sayad, Y.O., Mousannif, H., Al Moatassime, H., 2019. Predictive modeling of wildfires: a new dataset and machine learning approach. Fire Saf. J. 104, 130–146. https://doi.org/10.1016/j.firesaf.2019.01.006. Mar.
Sulova, A., Jokar Arsanjani, J., 2021. Exploratory analysis of driving force of wildfires in Australia: an application of machine learning within Google earth engine. Remote Sens. (Basel) 13 (1). https://doi.org/10.3390/rs13010010. Art. no. 1Jan.
Vasconcelos, M., Silva, S., Tomé, M., Alvim, M., Pereira, J., 2001. Spatial prediction of fire ignition probabilities: comparing logistic regression and neural networks. Photogramm. Eng. Remote Sensing. 67, 73–81. Jan.
Wang, S.S.-C., Qian, Y., Leung, L.R., Zhang, Y., 2021. Identifying key drivers of wildfires in the contiguous US using machine learning and game theory interpretation. Earth's Future 9 (6), e2020EF001910. https://doi.org/10.1029/2020EF001910.
wfca_teila, 2022. California Fire Season: In-Depth Guide. WFCA. Jul. 06. https://wfca.com/articles/california-fire-season-in-depth-guide/. accessed Jun. 26, 2023.
Xie, L., et al., 2022. Wildfire risk assessment in liangshan prefecture, China based on an integration machine learning algorithm. Remote Sens. (Basel) 14 (18). https://doi.org/10.3390/rs14184592. Art. no. 18Jan.
"California Department of Forestry and Fire Protection | CAL FIRE." https://www.fire.ca.gov/(accessed Jun. 04, 2023).
"GES DISC Dataset: MERRA-2 inst3_3d_asm_Np: 3d,3-hourly,instantaneous,pressure-level,assimilation,assimilated meteorological fields V5.12.4 (M2I3NPASM 5.12.4)." https://disc.gsfc.nasa.gov/datasets/M2I3NPASM_5.12.4/summary (accessed Aug. 18, 2023).
"GES DISC Dataset: MERRA-2 tavg3_3d_mst_Np: 3d,3-hourly,time-averaged,pressure-level,assimilation,moist processes diagnostics V5.12.4 (M2T3NPMST 5.12.4)." https://disc.gsfc.nasa.gov/datasets/M2T3NPMST_5.12.4/summary?keywords=precipitation%20and%20drought (accessed Aug. 18, 2023).
"Incidents | CAL FIRE." https://www.fire.ca.gov/incidents (accessed Jun. 04, 2023).
"USGS Landsat 8 Level 2, Collection 2, tier 1 | earth engine data catalog," *Google for Developers*. https://developers.google.com/earth-engine/datasets/catalog/LANDSAT_LC08_C02_T1_L2 (accessed Aug. 18, 2023).
"When Is California Fire Season?," *Frontline*. https://www.frontlinewildfire.com/wildfire-news-and-resources/california-fire-season/(accessed Jun. 26, 2023).
"Wildfires Kill Unprecedented Numbers of Large Sequoia Trees (U.S. National Park Service)." https://www.nps.gov/articles/000/wildfires-kill-unprecedented-numbers-of-large-sequoia-trees.htm (accessed Jun. 04, 2023).
"Predicting and planning for forest fires | PreventionWeb." Accessed: Dec. 21, 2023. [Online]. Available: https://www.preventionweb.net/news/predicting-and-planning-forest-fires-requires-modelling-many-complex-interrelated-factors.
"Study finds climate change to blame for record-breaking california wildfires | August 8, 2023 | Drought.gov." Accessed: Jan. 23, 2024. [Online]. Available: https://www.drought.gov/news/study-finds-climate-change-blame-record-breaking-california-wildfires-2023-08-08.