Predicting Wildfire Risk in Southern California Using Machine Learning Algorithms

By Rose W., Tianyang Y., David Z.

Thousand Oaks High School

Abstract

As wildfires can destroy homes and pollute the air in Southern California, wildfire risk assessment is necessary for the welfare of Californian communities. Compared to currently used statistical models for wildfire prediction, which is best suited for determining relationships, a machine learning model allows a large amount of data to be considered to make a prediction and does not require assumptions over data distribution. We determined whether a random forest machine learning model is effective in predicting wildfires in Southern California. We created a random forest model, which has been found to be accurate in similar predictions, to predict wildfires in Ventura County using data from 2015 to 2020. Our machine learning model predicted the occurrence of a wildfire occurring with a high accuracy. The normalized difference vegetation index, surface pressure, and volumetric soil water were the three most important predictor variables in the model's prediction-making, but predictor variables may correlate to the season in addition to wildfires. The ability to predict wildfires using a machine learning model allows people in predicted higher risk areas to better prepare themselves in the event of a fire, reducing a fire's spread and damage.

Table of Contents

Introduction

Wildfires negatively affect almost all aspects of Southern Californian communities through the destruction of property and air pollution.

For example, across downwind regions in the December 2017 wildfires in Southern California, fine particulate matter (PM2.5) levels increased by up to 231.2 $\mu g/m^3$ over a 14-day period, above safe levels set by the US government (Shi et al., 2019). Air pollution can spread up to thousands of miles from wildfire locations; pollutants from a Canadian wildfire were found to spread to even South Dakota and Iowa (NASA, 2015). Pollutants are dangerous as they cause breathing difficulties and health issues in individuals with pre-existing health conditions (American Lung Association, 2016). In particular, during instances of wildfire smoke exposure, hospital admissions have been observed to increase for symptoms of respiratory illness as well as respiratory illness (Black et al., 2017), creating a burden on the healthcare system. More directly, during the Camp Fire of November 2018 in California, 153,336 acres of land were burned down and there were 85 deaths (California Department of Forestry and Fire Protection, 2021).

In addition to health risks and deaths, economic impacts of wildfires include the destruction of property, healthcare costs from treating symptoms due to pollution and fires, firefighting costs, and more. For example, 2018 wildfires in California resulted in estimated damages of $148.5 billion, or 1.5% of the state's annual gross domestic product, due to factors such as capital losses and healthcare costs (Wang et al., 2020).

Furthermore, the damages of wildfires can result in emotional effects as well. In one study of wildfires, it was discovered that "individuals found it difficult to cope with their losses." The study detailed many responses that people have due to wildfires, while there were some positives discovered such the newly perceived closeness of the community in the aftermath of wildfires, many more people described feeling helplessness and terror due to the idea of another fire occuring (Kulig et al, 2013). The damages that can be caused by wildfires are exceptionally traumatizing.

There are multiple causes of wildfires. In Southern California, very large wildfires are caused by multiple moderately-severe factors, which has occurred increasingly in the past four decades, instead of single highly-severe factors (Khorshidi et al., 2020). Therefore, it is necessary to consider multiple factors when predicting wildfire risk. Some examples of wildfire risks considered in the aforementioned study include the moisture of vegetation, temperature, wind speeds, and pressure (Khorshidi et al., 2020).

In comparison to statistical methods to predict wildfire risk, machine learning may be more effective. Statistical methods are used primarily to determine relationships between two variables. Since a linear relationship between risk factors and wildfires cannot be assumed, machine learning, which can analyze nonlinear relationships, can improve the accuracy of wildfire prediction and reduce negative impacts on communities (Malik et al., 2021).

Researchers have previously modeled wildfire risk using machine learning in multiple locations around the world. For example, studies have been conducted in Australia (Dutta et al., 2016), Amol County, Iran (Gholamnia et al., 2020), Monticello and Winters, California (Malik et al., 2021), Canada (Sayad et al., 2019), and Spain (Rodrigues & de la Riva, 2014), to name a few.

In studies comparing the accuracy of multiple machine learning algorithms, the random forest algorithm was highly effective (Gholamnia et al., 2020; Rodrigues & de la Riva, 2014), thus, in this project, a random forest algorithm will be used. Overall, previous machine learning models have included data such as fire history, weather (such as temperature, humidity, wind speed, solar radiation, precipitation, pressure), topography, soil moisture, land use, vegetation levels, and power infrastructure (Malik et al., 2021). In this study, similar publicly-available data, listed in the materials and methods section, was used to produce the machine learning model.

A random forest model involves the training of a large number of decision trees consisting of parts of the data from a dataset. A decision tree involves comparing predictor variables to a branching set of conditions until a conclusion can be made about an item's category. Each decision tree in the random forest is trained using a random subset of the data available; one random subset is excluded from all decision trees and is used to test the accuracy of the tree on new data. The item is classified into the class that the most trees agree it belongs to (Biau & Scornet, 2016).

In addition, XGBoost is used in the project for hyperparameter tuning. XGBoost is a gradient boosting algorithm, which has been successfully used in many large-scale machine learning applications (Chen & Guestrin, 2016). In contrast with a random forest, gradient boosting involves the creation of decision trees consecutively instead of independently, using the information from prior trees (Natekin & Knoll, 2013). Hyperparameter tuning is necessary to set hyperparameters, preset values that influence the training of the machine learning model, like the number of observations taken in each tree from the dataset, to create the best model possible (Probst et al., 2019).

In this project, we will create a machine learning model that accurately predicts the likelihood of wildfires in Southern California over a time period through the random forest algorithm.

Research Question

Is a random forest machine learning model effective in predicting wildfires in Southern California?

Hypothesis

Alternative hypothesis: A random forest machine learning model is effective in predicting wildfires in Southern California over a time period.

Null hypothesis: A random forest machine learning model is not effective in predicting wildfires in Southern California over a time period.

Materials and Methods

1. Data was collected from public datasets. In addition to fire history, data that was used as predictor variables, such as climate and vegetation indices, were obtained. Examples of data that was used are listed in Table 1.

Table 1. A list of data and the sources that will be used in the study. The in-text citation is listed below; the full citations are listed in the bibliography.

Data type	Citation	Format
Fire history	California Department of Forestry and Fire Protection (2021)	GeoJSON
Weather	Hersbach et al. (2018)	NetCDF4
Vegetation	Didan (2021)	GeoTIFF

2. The Python programming language (version 3.9.7) was used to process the data and train the model. The code was written using Jupyter Notebook. Table 2 lists the libraries used to build the model.

Table 2. A list of Python libraries used to build the model. The purpose is listed to the right of each library.

Library	Version	Purpose
shapely	1.8.1.post1	Polygon and point data types
pandas	1.3.4	Data manipulation
geopandas	0.10.2	Support for geospatial data types, namely GeoJSON files, within pandas
rasterio	1.2.10	Read GeoTIFF files
netCDF4	1.5.7	Read .nc (NetCDF) files
numpy	1.20.3	Mathematical functions
scikit-learn	0.24.2	Machine learning model training
seaborn	0.11.2	Data visualization
matplotlib	3.4.3	Data visualization
shap	0.38.1	Hyperparameter tuning
xgboost	1.4.0	Hyperparameter tuning

3. The dataset files were opened. GeoJSON files were opened through the geopandas library, GeoTIFF files were opened through the rasterio library, and NetCDF files were opened through the netCDF4 library. GeoJSON files are used to store geographical features. GeoTIFF files are image files with Earth coordinates. NetCDF files are array-based data storage files.
4. The datasets were processed and then concatenated to create a combined dataset. To process data, collected data was normalized when needed into standard data types, unnecessary data outside the scope of the project was removed, and data entries with errors and null values were corrected or removed. Specifically, the fire history dataset was processed to account for null values, data entry errors, and data outside the scope of the project. Dates were converted from the string data type to Python built-in data types, entries with invalid or null alarm dates or dates outside of the scope were removed, dates that were improperly entered (such as inconsistencies between the year column and the year in the date column) were corrected, and null fire containment dates were set to five days after the alarm date. The variables used in the combined dataset are listed in Table 3.

Table 3. A list of variable names used in data analysis, descriptions of data, and the sources.

Variable name (as used in data analysis)	**Description**	**Source**
date	Date for which the data will be retrieved, from January 1, 2015 to December 31, 2020	Randomly generated
point	Coordinate for which the data will be retrieved, within the bounds of Ventura County, California	Randomly generated
fire_occurred	Boolean data recorded as 0 if a fire did not occur at the date and point; 1 if a fire did occur	California Department of Forestry and Fire Protection (2021)
ndvi	Normalized difference vegetation index	Didan (2021)
wind_u10	10 meter U wind component	Hersbach et al. (2018)
wind_v10	10 meter V wind component	Hersbach et al. (2018)
d2m	2 meter dewpoint temperature	Hersbach et al. (2018)
t2m	2 meter temperature	Hersbach et al. (2018)

e	Evaporation	Hersbach et al. (2018)
i10fg	Instantaneous 10 meter wind gust	Hersbach et al. (2018)
sd	Snow depth	Hersbach et al. (2018)
ssr	Surface net solar radiation	Hersbach et al. (2018)
sp	Surface pressure	Hersbach et al. (2018)
tcc	Total cloud cover	Hersbach et al. (2018)
tcwv	Total column water vapor	Hersbach et al. (2018)
tp	Total precipitation	Hersbach et al. (2018)
swvl1	Volumetric soil water layer 1	Hersbach et al. (2018)

5. For each data entry in the combined dataset, a random coordinate point and date was generated within the scope of the project. Then, each predictor variable for the specified location and time was retrieved from the datasets. This procedure was repeated until 10,000 data entries were generated for training the model, with half being entries where a fire occurred. Each predictor variable was obtained through the following methods:
    a. For the presence of wildfires, the wildfire dataset was searched. If a wildfire's burn area contained the generated point and the generated date was between the alarm date and containment date, a wildfire was present at that point.
    b. For the vegetation data, the file corresponding to the correct date was opened. The files are named with the respective day number; each file is valid from that day

number to fifteen days following that day. The NDVI for the correct coordinate was then obtained.

c. For the climate data, the file corresponding to the correct one-year interval was opened. The correct indices for the time, latitude, and longitude were calculated. For each variable, the data were accessed using the calculated indices and the twenty-four hour mean of the data was used when generating the combined dataset.

6. The generated dataset was split randomly into a training and testing dataset; the testing dataset consisted of 30% of the original dataset, while the training data consisted of 70%. The wildfire data and the predictor variables above in the newly-split training dataset were used to train a random forest model, as it was found to be effective in previous wildfire risk modeling projects (Gholamnia et al., 2020; Malik et al., 2021). For each point, if a fire has occurred at the time and location specified in the point, then the point was considered presence data; otherwise, the point was considered absence data.
7. XGBoost, explained previously in the introduction, was used for hyperparameter tuning, or choosing the best features to build a more predictive model with (see Figure 1). SHAP, a method of analyzing data based on game theory, was also used to provide better insight into the predictiveness of features, or predictor variables (see Figure 2). As the last step in hyperparameter tuning, a cross-correlation chart was generated (see Figure 3) showing the correlation between features. Among related features, the more predictive features were selected to create another training and testing dataset. The dataset was once again split into 30% testing and 70% training data. By utilizing SHAP and XGBoost, a new random forest model with better results was created and trained using this data.

8. The testing data was used to test the accuracy of the new model, which is measured as a percentage. The accuracy of the model represents the number of correctly predicted data values out of all data values.
9. Outputs, including the F score, or the predictiveness of each feature, were analyzed and graphed. These values will be interpreted in the discussion section of this paper.

The source code used to process and analyze the data and the datasets used are available through the following link: https://github.com/tyu012/wildfire-risk-project.

Results

The model had an accuracy of 97.0%. A confusion matrix is shown in Figure 1. Each row represents the actual classes, and each column represents the class predicted by the model. For the absence data, the precision, recall, and F1 score were 0.96, 0.98, and 0.97 respectively (out of 1.00). For the presence data, the precision, recall, and F1 score were 0.98, 0.96, and 0.97 respectively.

```
97.0 %
[[1481   31]
 [  55 1433]]
              precision    recall  f1-score   support

           0       0.96      0.98      0.97      1512
           1       0.98      0.96      0.97      1488

    accuracy                           0.97      3000
   macro avg       0.97      0.97      0.97      3000
weighted avg       0.97      0.97      0.97      3000
```

Figure 1. From top to bottom, the percent accuracy, confusion matrix, and model statistics (precision, recall, and F1 score) are displayed. Accuracy is the percent of correct predictions with respect to total predictions. Precision is the number of true positives over the number of true and false positives, meaning it measures the fraction of correct positive identifications. Recall is the number of true positives over the number of true positives and false negatives, meaning it measures the fraction of positive data identified as positive by the model. The F1 score is a harmonic mean of the precision and recall values, combining the two values into one indicator.

The combined dataset used for training and testing was generated such that 50% of the data was presence data, or data when a wildfire occurred, as presence data is only generated 0.5% of the time if the data is randomly generated. The additional presence data is used to increase the accuracy of predicting presence data.

Figure 2 displays the importance of each of the predictor variables for the trained model when making predictions, measured as a F score. According to the below graph generated by the model, ndvi, or vegetation cover data, is the most predictive factor in predicting wildfires, with a significant F-score of 147. Surface pressure, labeled sp, is also significantly predictive of wildfire occurrence with a F-score of 131. Snow depth, labeled sd, is the least predictive with a F-score of 37, and may not be significantly related to wildfire occurrence.

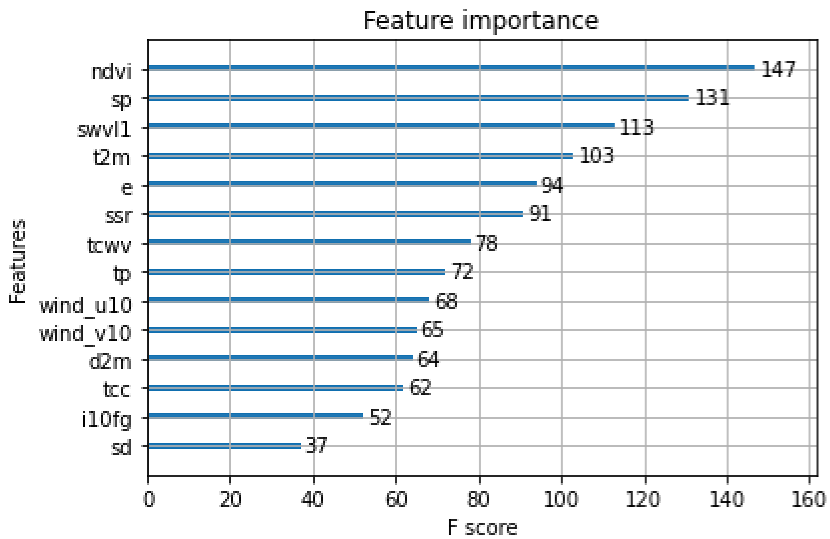

Figure 2. Factors, called features, are listed on the left-hand side of the graph in order of most predictive at the top and least predictive at the bottom. The F-score on the x-axis measures how strong each factor is in predicting whether a fire has occurred or not, or the feature importance.

Figure 3 displays the results when the data was analyzed using SHAP, a method of analyzing data based on game theory. Essentially, SHAP measures a feature's importance in model performance according to how poorly the model performs without a feature's data. Unlike XGBoost results, SHAP found surface net solar radiation (ssr) to be most predictive of wildfire occurrence. SHAP results also differed in that snow depth (sd) was found to be predictive, while in Figure 2 snow depth is the least predictive feature. Similarly to XGBoost, however, SHAP found vegetation cover (ndvi), surface pressure (sp), and volumetric soil water layer 1 (swvl1) to be significantly predictive.

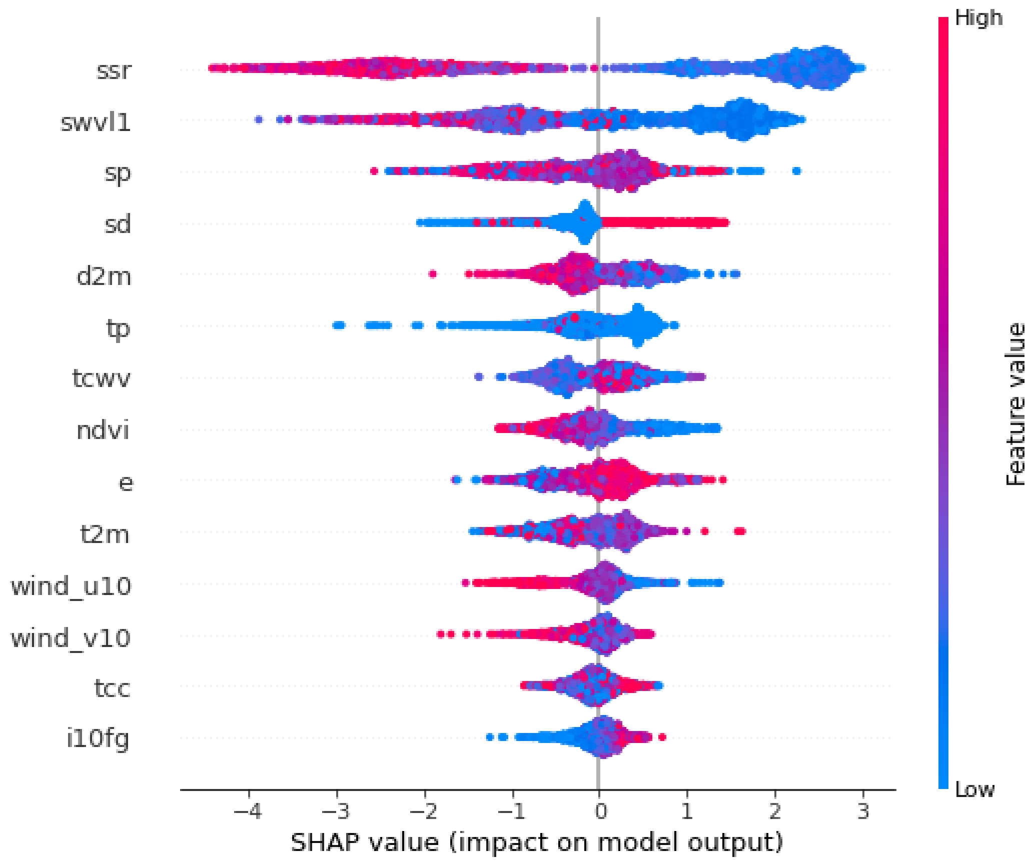

Figure 3. SHAP, a method of analyzing data based on game theory, was used to generate the above chart. All factors used to predict wildfire occurrence are listed on the left, with the most predictive (according to SHAP) at the top and the least predictive at the bottom. If the feature value becomes less red and more blue moving from left to right, it is inversely related to wildfire occurrence, since as the value becomes lower (bluer) the wildfire occurrence value will become higher. The opposite is also true. For example, surface net solar radiation is inversely related to wildfire occurrence, while snow depth is directly related to wildfire occurrence.

In Figure 4, a cross-correlation chart displays the relationships between each predictor variable. If the line connecting two variables is close, they are more closely related. For example, total column water vapor (tcwv) and 2 meter dewpoint temperature (d2m) are closely related while evaporation (e) and volumetric soil water layer 1 (swvl1) are not related.

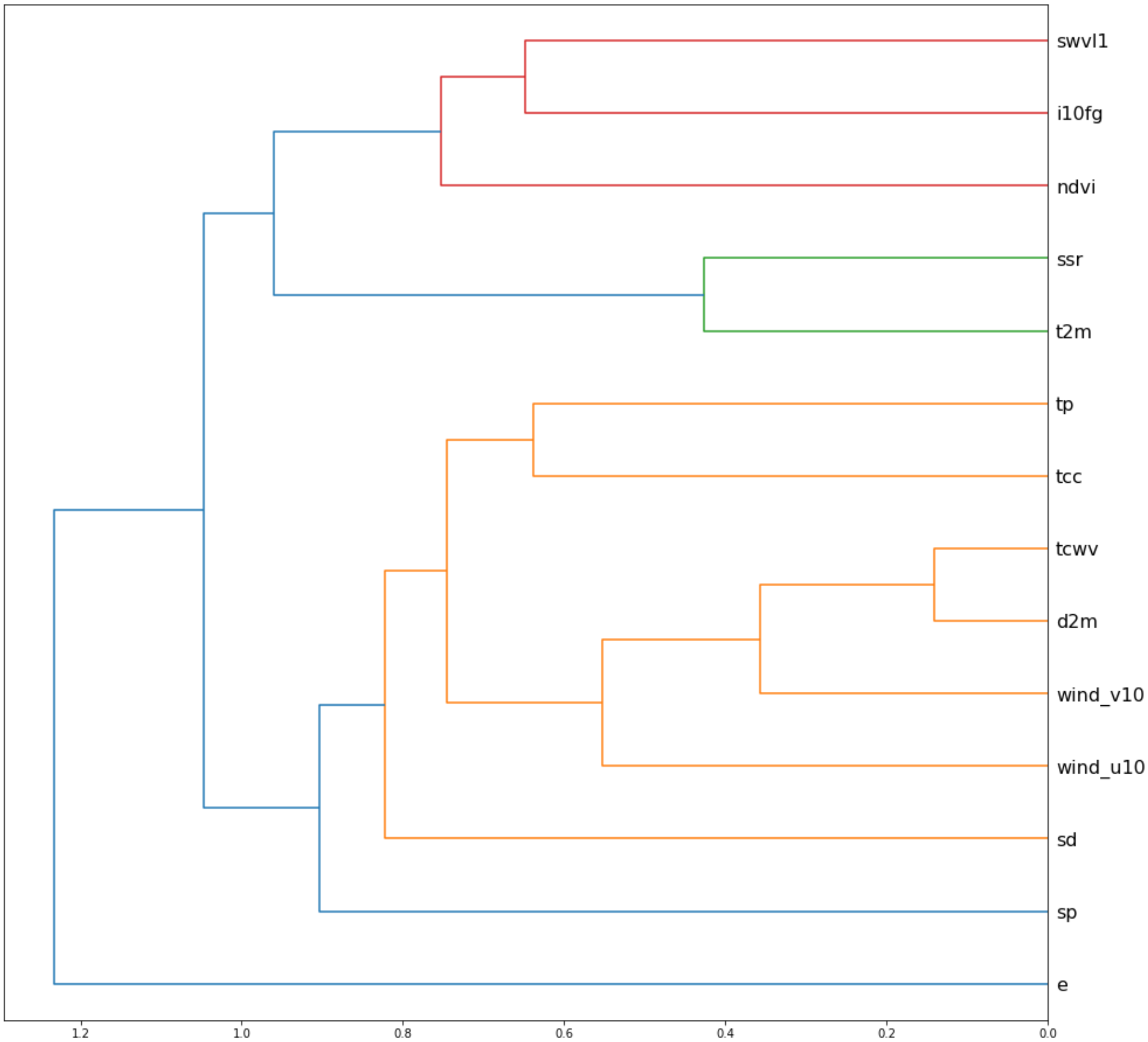

Figure 4. The above is a cross-correlation chart showing the relationships between features. Related features are connected by lines. The closer the line connecting the features is to the right-side of the graph, the more the features are related.

Discussion

Precision is the number of true positives over the number of true and false positives, meaning it measures the fraction of correct positive identifications. For the presence data, since precision is 0.98, 98% of test data identified as positive were true positives. Recall is the number of true positives over the number of true positives and false negatives, meaning it measures the fraction of positive data identified as positive by the model. For the presence data, since recall is 0.96, 96% of all positive test data was identified as positive by the model. The F1 score is a harmonic mean of the precision and recall values, combining the two values into one indicator. Since the precision and recall for both the presence and absence data were high, at least 0.96, and similar, the model is accurate in predicting wildfires between the start of 2015 and the end of 2020. Therefore, we reject the null hypothesis.

As previously described, the normalized difference vegetation index, surface pressure, and volumetric soil water layer 1 are the three most important predictor variables. Furthermore, low levels of NDVI and soil water led to the model more likely predicting a wildfire has occurred. As a result, low vegetation and soil water levels had the strongest association with wildfires. The three least important predictor variables are snow depth, instantaneous 10 meter wind gust, and total cloud cover. NDVI's inverse relationship to wildfire occurrence could possibly be explained by attributing more green cover to lower risk of wildfires, potentially due to the increased water in the environment that allowed more plants to grow. Additionally, seasonal changes could increase chances of wildfires occurring.

The effect of some variables on the prediction on the SHAP chart may appear counterintuitive. For example, when the value of snow depth (sd) is high, or when the value of surface net solar radiation (ssr) is low, the model is more likely to predict that a fire occurred. However, in the training dataset, 4760 out of the 5000 data entries where a fire has occurred take place in either December or January, which are usually wet months in Ventura County. As a result, these anomalies are likely caused by the model associating wildfires with winter weather conditions when training. For example, surface net solar radiation (ssr) may be inversely related to fire occurrence because during wet months, higher radiation could help with green vegetation growth, preventing fires from catching. Due to these results and the seasonality of wildfires in Ventura County, machine learning models trained with weather data are only applicable locally as they associate wildfire presence with seasonal changes.

When given weather conditions for a location at a specific date, the model can predict whether there is a wildfire occurring. As a result, through obtaining weather and vegetation data, which may be more easily accessible, potential wildfires, especially in remote areas, may be detected more easily due to the knowledge of what locations are more likely to have fires. Overall, machine learning models, especially random forest and XGBoost, hold promise to provide more accurate wildfire predictions and detections. With the improvement of prediction technology, machine learning models could play an important role in saving possible lives and minimizing damage to Southern Californian communities.

Limitations

One of the biggest issues faced in this project was the varied formats of the datasets. In addition to the differing temporal resolutions, the dataset used for wildfire perimeters used a format where the location was described by polygons, the dataset used for vegetation used an image-based format, while the dataset used for weather used an array-based format. To account for this, the model generated random coordinate positions to create a dataset composed of all the points generated instead of a static dataset. This introduces a greater degree of randomness due to the large range of potential data points.

Another limitation to the project's method was that it did not consider data before and after the random date generated meaning that a fire could occur the day after and the program would have no way of knowing. Furthermore, the fire history dataset only includes the area burned by the wildfire over a time interval, but it does not include the specific areas burned as part of a wildfire on a particular day. As a result, it cannot be determined whether at a certain point and time, a fire has occurred or is about to occur. A point representing presence data is only a location near (potentially before or after) a time where a fire has occurred.

Conclusion

This project was successful in producing a machine learning model to predict the occurrence of a wildfire in Ventura County during a six-year period due to its accuracy of 97.0%. Out of all predictor variables used, the normalized difference vegetation index, surface pressure, and volumetric soil water were the three most important predictor variables in the model's prediction-making. Due to the seasonality of wildfires in Ventura County, seasonal shifts in the values of predictor variables may be associated with wildfires.

Further work may be pursued to reduce the impact of this study's limitations. To improve the capability for the model to predict wildfires before they occur, one area of further work is training a model that considers weather conditions before the fire when creating the combined dataset. In addition, to address the seasonality of local wildfires, it is necessary to develop means to reduce the impact of seasonal weather shifts when training a model to enable more general applications of machine learning to predict wildfires. Finally, further studies could create monthly "lag" features, analyzing data for factors within a certain number of months before a fire occurs.

Bibliography

American Lung Association. (2016, January 1). *How wildfires affect our health*. Retrieved February 10, 2022, from https://www.lung.org/blog/how-wildfires-affect-health#:~:text=Wildfires%20threaten%20lives%20directly%2C%20and,COPD%20and%20other%20lung%20diseases.

Biau, G., & Scornet, E. (2016). A random forest guided tour. *TEST, 25*, 197-227. https://doi.org/10.1007/s11749-016-0481-7

Black, C., Tesfaigzi, Y., Bassein, J. A., & Miller, L. A. (2017). Wildfire smoke exposure and human health: Significant gaps in research for a growing public health issue. *Environmental Toxicology and Pharmacology, 55*, 186-195. https://doi.org/10.1016/j.etap.2017.08.022

California Department of Forestry and Fire Protection (2021). *California fire perimeters (all)* [Data set]. California State Geoportal. https://gis.data.ca.gov/datasets/CALFIRE-Forestry::california-fire-perimeters-all/

California Department of Forestry and Fire Protection. (2021, October 22). *Top 20 most destructive California wildfires.* Retrieved February 10, 2022, from https://www.fire.ca.gov/media/t1rdhizr/top20_destruction.pdf

Chen, T. & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785-794. https://doi.org/10.1145/2939672.2939785

Didan, K (2021). *MODIS/Terra Vegetation Indices 16-Day L3 Global 250m SIN Grid V061* [Data set]. https://doi.org/10.5067/MODIS/MOD13Q1.061

Dutta, R., Das, A., & Aryal, J. (2016). Big data integration shows Australian bush-fire frequency is increasing significantly. *Royal Society Open Science, 3*, 150241. https://doi.org/10.1098/rsos.150241

Gholamnia, K., Nachappa, T. G., Ghorbanzadeh, O., & Blaschke, T. (2020). Comparisons of diverse machine learning approaches for wildfire susceptibility mapping. *Symmetry, 12*, 604. https://doi.org/10.3390/sym12040604

Hersbach, H., Bell, B., Berrisford, P., Biavati, G., Horányi, A., Muñoz Sabater, J., Nicolas, J., Peubey, C., Radu, R., Rozum, I., Schepers, D., Simmons, A., Soci, C., Dee, D., & Thépaut, J-N. (2018). *ERA5 hourly data on single levels from 1979 to present* [data set]. https://doi.org/10.24381/cds.adbb2d47

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 between 1982 and 2018. *Environmental Research Letters, 15*, 104002. https://doi.org/10.1088/1748-9326/abae9e

Kulig, J., Townshend, I., Reimer, W., Edge, D., & Lightfoot, N. (2013). Impacts of wildfires: Aftermath at individual and community levels? *The Australian Journal of Emergency Management, 28*(3), 29–34. https://doi.org/10.3316/agispt.20132328

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

NASA. (n.d.). *Canadian wildfires produce river of smoke*. Retrieved February 10, 2022, from https://earthobservatory.nasa.gov/images/86151/canadian-wildfires-produce-river-of-smoke

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

Probst, P., Wright, M. N., & Boulesteix, A.-L. (2019). Hyperparameters and tuning strategies for random forest. *WIREs Data Mining and Knowledge Discovery, 9*(3), e1301. https://doi.org/10.1002/widm.1301

Rodrigues, M. & de la Riva, J. (2014). An insight into machine-learning algorithms to model human-caused wildfire occurrence. *Environmental Modelling & Software, 57*, 192-201. https://doi.org/10.1016/j.envsoft.2014.03.003

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

Shi, H., Jiang, Z., Zhao, B., Li, Z., Chen, Y., Gu, Y., Jiang, J. H., Lee, M., Liou, K.-N., Neu, J. L., Payne, V. H., Su, H., Wang, Y., Witek, M., & Worden, J. (2019). Modeling study of the air quality impact of record-breaking Southern California wildfires in December 2017. *Journal of Geophysical Research: Atmospheres, 124*, 6554-6570. https://doi.org/10.1029/2019JD030472

Wang, D., Guan, D., Zhu, S., Kinnon, M. M., Geng, G., Zhang, Q., Zheng, H., Lei, T., Shao, S., Gong, P., & Davis, S. J. (2021). Economic footprint of California wildfires in 2018. *Nature Sustainability, 4*, 252-260. https://doi.org/10.1038/s41893-020-00646-7