# Predicting Wildfire Risk Using Machine Learning Algorithms

Rose W., Tianyang Y., David Z. | Thousand Oaks High School

## 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.

## Abstract

Wildfires can destroy homes and pollute the air in Southern California; wildfire risk assessment is necessary for the welfare of Californian communities.

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.

In this paper, we create 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.

This project was successful in producing a machine learning model that predicts the occurrence of a wildfire occurring. 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.

## Research

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).

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, topography, soil moisture, land use, vegetation levels, and power infrastructure (Malik et al., 2021).

A random forest model involves the training of a large number of decision trees trained using a random subset of the data available (Biau & Scornet, 2016).

XGBoost, also used in this project, is a gradient boosting algorithm, where trees are trained consecutively instead of concurrently, which has been successfully used in many large-scale machine learning applications (Chen & Guestrin, 2016).

## Procedure

1. Collect public datasets: fire history, weather, and vegetation
2. Python was used to read and process the data.
3. Random generation was used to make a testing and training dataset for the machine learning model. A random position and time was chosen and the relevant data points were retrieved from the datasets.
4. Random forest was used for training and XGBoost and SHAP were used for hyperparameter tuning.

## Results

```
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: Precision is the number of true positives over the number of true and false positives. Recall is the number of true positives over the number of true positives and false negatives. The F1 score is a harmonic mean of the precision and recall value.

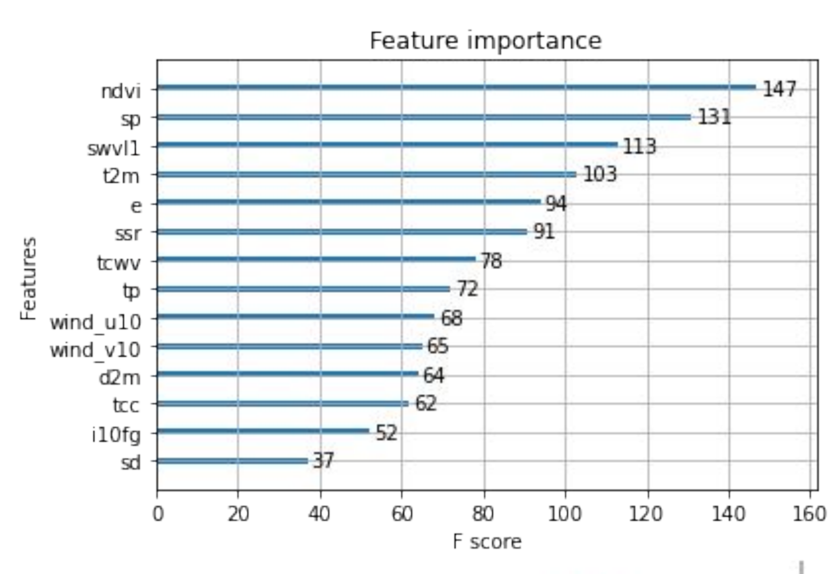



Figure 2: Predictor variables, called features, are listed in order of most predictive at the top and least predictive at the bottom.

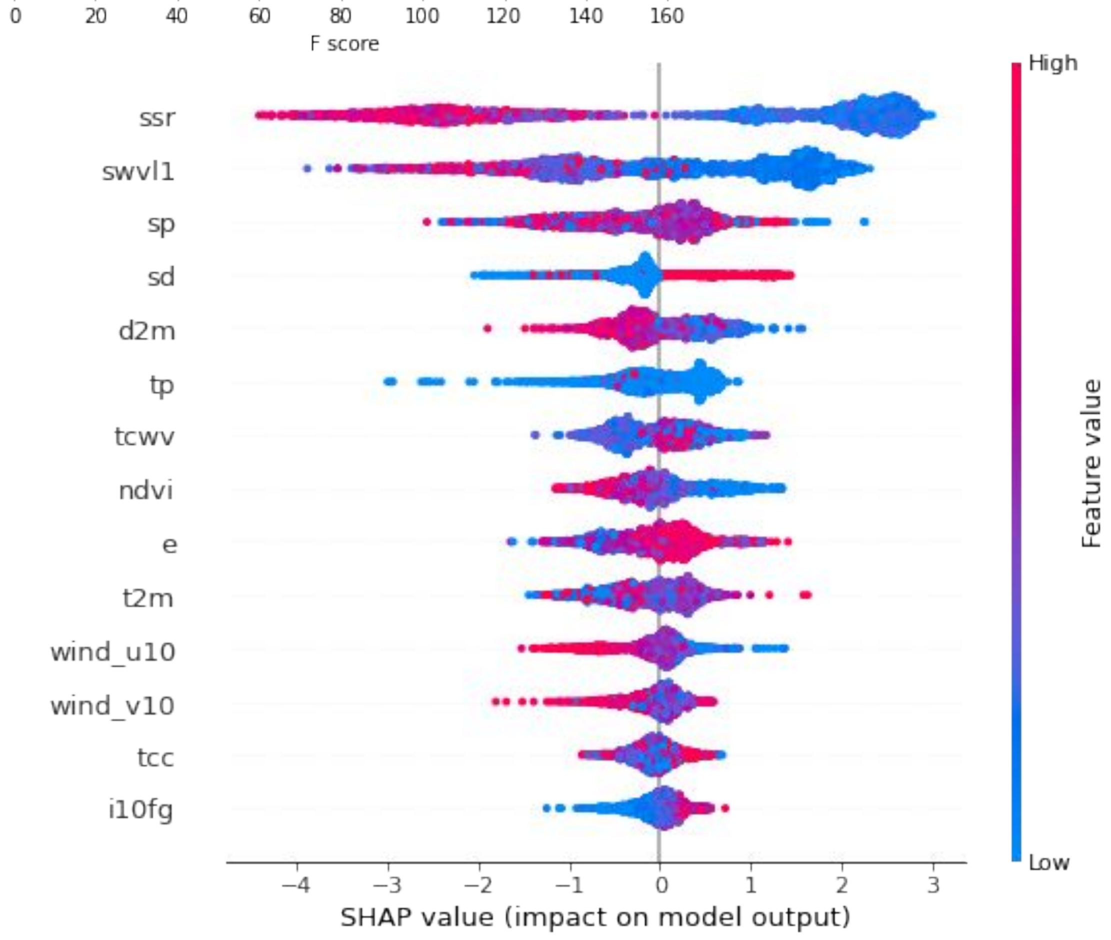
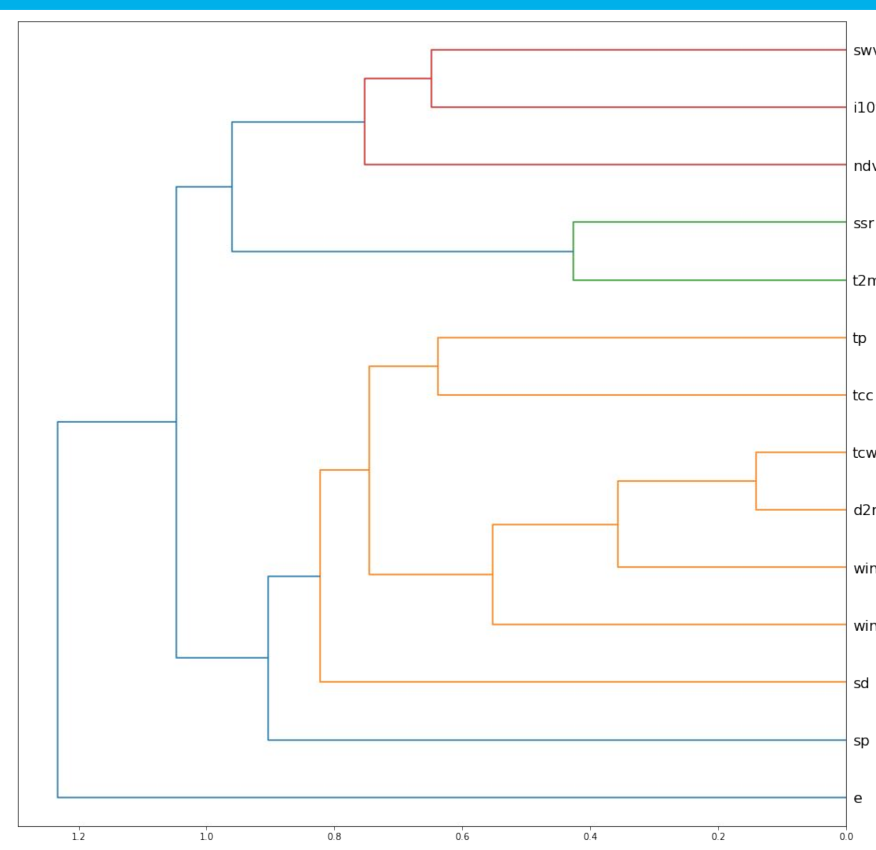



Figure 3: The SHAP value represents the impact of each value on the model's prediction. If the feature value becomes less red and more blue moving from left to right, it is inversely related to wildfire occurrence, since as the value becomes lower (bluer) the wildfire occurrence value will become higher, and vice versa.

## Results cont.

Figure 4:

Cross-correlation chart showing the relationships between predictor variables. Related factors are connected by lines.




## Conclusion

This project was successful in producing a machine learning model that is able 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 includes training a model that considers weather conditions before the fire when creating the combined dataset is resistant to seasonal changes in weather.

## Works Cited

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