Predicting Wildfire Susceptibility in Napa County, California using Machine Learning

Stefan Shakeri[1] and Krti Tallam[2#]

[1] Vintage High School
[2]Stanford University
[#]Advisor

ABSTRACT

Wildfires have long been a part of the natural environment, however through climate change and increased human activity, they have become a significant problem to both humans and wildland. Stopping the expansion of wildfires would be critical in mitigating the dangerous outcomes of them. Firefighters stopping the spread of wildfires must know which parts of the environment are most vulnerable to the spread of wildfires, and vegetation is one of the key determining factors in the wildfire susceptibility of a given area. Previous works have used several different machine learning algorithms for the purpose of determining wildfire susceptibility. The algorithm used in this study for wildfire susceptibility prediction is a random forest applied to a vegetation dataset of Napa County, California provided by the California Department of Fish and Wildlife (CDFW). The random forest works by creating a set of decision trees to get an overall probability for each vegetation area. The model has a 91.7% accuracy in predicting wildfire burn probability in a vegetation area.

Introduction

Wildfires are a part of the natural cycle of Northern California, contributing to increased biodiversity, a healthier ecosystem, and several other benefits (Pausas & Keeley, 2019). It is important for some wildfires to happen in order to maintain the natural order of the environment. For example, the chaparral vegetation that makes up large parts of Napa County and Northern California has a typical fire cycle of 10 to 40 years, in which wildfires clear old plants and shrubs to let new life grow in the cleared environment (Muller et al., 1968). However, climate change and other human activities in wildfire-prone areas has led to an increase in wildfires in California to reach unsustainable levels (Miller & Safford, 2012). Increased wildfires put both humans and wildlands in danger, especially in the wildland urban interface (WUI) (Kramer et al., 2019), where humans live in or near wildland areas. In order to prevent damage caused by increased wildfires, people countering wildfires need to be able to identify the most vulnerable places to a wildfire. Identifying these wildfire susceptible areas can be done by examining the vegetation they contain.

Determining wildfire susceptibility has been done through different types of machine learning algorithms. These include a classification and regression tree (CART) algorithm (Amatulli et al., 2006), artificial neural networks, regression trees (Jain et al., 2020), etc. One algorithm that is effective for wildfire susceptibility is the random forest. The random forest outputs several probabilistic results from decision trees to create a final decision for the algorithm (Bustillo Sánchez et al., 2021). Two examples of random forests being used in relation to wildfires is by Ma et al. (2020) to identify the causes of wildfires in China and by Collins et al. (2018) to identify wildfire severity. Furthermore, Malik et. al. in 2021 used a random forest model to predict wildfire susceptibility in Northern California, specifically in the area surrounding Winters, California (Malik et al., 2021).

Despite these previous uses of the random forest model for predicting wildfire susceptibility, the algorithm has yet to be used in Napa County, California. Napa County offers a wide array of vegetation types to be analyzed (Thorne, 2020) and has experienced many devastating fires, especially in the past decade. This high volume of wildfires provides plentiful data for different types of wildfires with regards to size, vegetation burned, and location. This combination of variables makes Napa County useful to predict wildfire susceptibility in vegetation types because of the excess of data, creating a better understanding of what types of vegetation are most prone to wildfires.

Methods

Data Collection

The data used to examine the vegetation of Napa County is a vegetation geodatabase of Napa County provided by the California Department of Fish & Wildlife (CDFW) Vegetation Classification and Mapping Program (Thorne, 2020). The original vegetation map of Napa County was created in 1993 and gathered from 3 meter per pixel, monochrome, digital orthophoto quadrangles. This data was subsequently updated in 2004 through field work to adjust the boundaries of vegetation categories. The current vegetation map from 2016 is gathered from 1 meter per pixel color satellite imagery, which was taken by the National Agriculture Imagery Program (NAIP). The satellite imagery is then delineated by Aerial Services Incorporated (ASI) (Barrette et al., 2000) to output polygons with vegetation and landcover attributes in accordance with the CDFW Manual of California Vegetation (Sawyer et al., 2009). The polygons were compared to time series imagery from Google Earth and imagery from ArcMap to ensure their accuracy. Furthermore, the vegetation types were reviewed with those from two other counties in California to ensure accuracy.

A comma-separated values (CSV) file is extracted from the geodatabase to provide a workable dataset. This dataset containing 31 different attributes is reduced to 8 attributes that are necessary for the algorithm. These are: vegetation_category = the overarching vegetation category in a polygon, area_acres = the area of a polygon in acres, size_class = a classification for the sizes of trees in a polygon, density_class = a classification for the density of a polygon, WUI = a classification for if a polygon is part of the wildland-urban interface (WUI), with 5 classifications for different levels of WUI, burn_coverage = if a polygon has been partially, fully, or never burned by a wildfire, year_burned = the latest year that a polygon was burned by a wildfire, if ever, and CalVeg_name = the exact type of vegetation in a polygon.

To allow for the model to be more effective, the burn_coverage attribute is converted from a “String” datatype to a binary “Integer” datatype. A value previously classified as “null” is converted to 0, and “partial” and “full” values are converted to 1 to signify that they had been burned by wildfires. Additionally, the vegetation_category variables are also converted to Integers, with: “Wetlands” = 0, “Shrubland” = 1, “Riparian woodland” = 2, “Oak woodlands” = 3, “Grassland” = 4, and “Coniferous forest” = 5.

TABLE 1. Sample data from the edited “Vegetation - Napa County” dataset (Thorne, 2020). All attributes except for CalVeg_name are numerical values and can be used in the model.

	vegetation_category	area_acres	size_class	density_class	WUI	burn_coverage	year_burned	CalVeg_name
1	1	1.804458	9	1	0	0	0	Lower Montane Mixed Chaparral
2	1	2.724046	9	1	0	0	0	Chamise
3	4	1.016912	9	1	0	1	2008	Annual Grasses and Forbs
4	3	6.185103	4	2	0	0	0	Blue Oak

The dataset was run through a correlation matrix, which provides the positive and negative correlation between variables of a dataset. This can be useful in determining if certain variables should be removed because of too high of a correlation with another variable. This resulted in a perfect correlation of 1 between burn_coverage and year_burned. As a result, the year_burned attribute was removed from the dataset to produce a better model that would avoid giving too much weight to those two variables. Other variables had a high negative correlation; however, it was not enough to warrant their removal from the dataset.

Updated Attribute Correlation Matrix

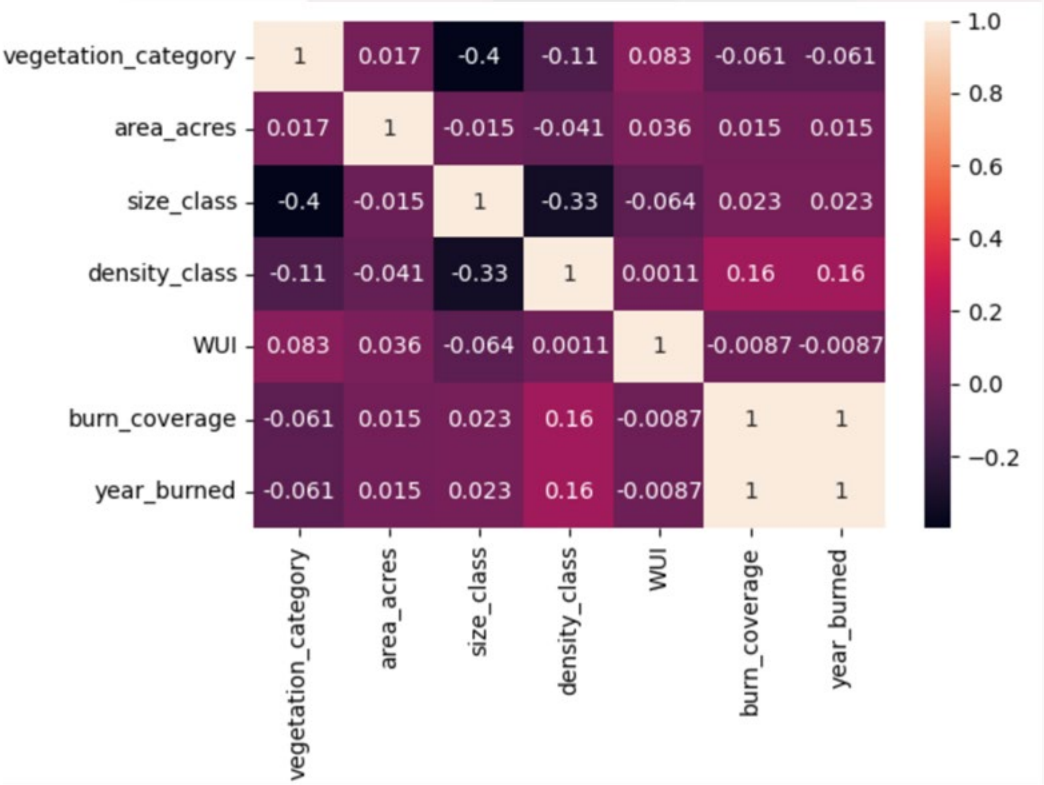

FIG. 1. The correlation now contains burn_coverage. Because of the correlation of 1 between burn_coverage and year_burned, the year_burned attribute was removed to prevent a damaged model. The attributes of size_class and vegetation_category as well as size_class and density_class had high negative correlations, but not enough to damage the model.

Random Forest Model

The basis of the random forest model is the decision tree. A decision tree is a model for obtaining an output in which one decision leads to several subsequent decisions to get a precise output. The decision tree starts with a small set of options to choose from for a decision, with each option presenting more options that are related to the previously chosen option. This model allows for a specific output based on what the result of the decision tree path taken is. In the case of this study, the output would either be 0 or 1, with 0 indicating no wildfire burn and 1 indicating a wildfire burn. What a random forest does is creates several decision trees that all output their responses and then takes the mode (majority) response of the decision trees as a final output (Belgiu & Drăguț, 2016). This aggregation of the decision trees allows the random forest model to be more reliable than just a single decision tree.

Random Forest Model Diagram

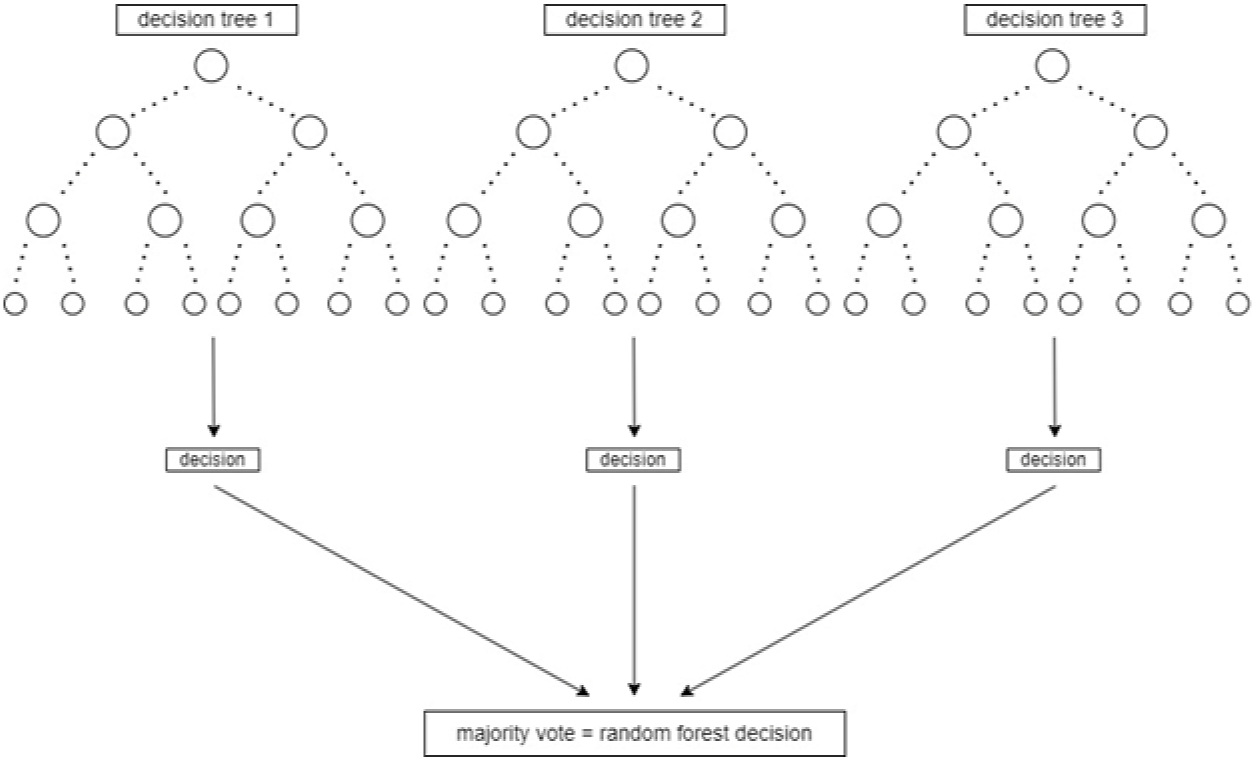

FIG. 2. A random forest takes its input as the output of several decision trees, resulting in an output for the random forest that is the majority vote of the decision trees.

The random forest model used for this study is based off the model used by Husted (2022) to predict the causes of wildfires in the United States. The first step is to split the dataset into a train group and a test group. The test size is set to 0.3, meaning that 30% of the data was designated to the test group and 70% to the train group. The data split is to provide sufficient unique data for the algorithm to train on and leaving enough unique data to test the algorithm on. The train data is used by the random forest algorithm to create a model that can accurately predict the wildfire susceptibility of each input value. The random forest algorithm used is the RandomForestClassifier function from SkLearn. The estimators, or number of decision trees, is set to 50 to have that many generated decision trees that the random forest takes the majority of. The other hyperparameters of the random forest are at their default values.

Once the training stage of the model has been completed, the remaining 30% of the data designated for testing is run through the model to get an output value. In order to verify the accuracy of the random forest model, a confusion matrix is then created with the data to examine the ratio of true positives, false positives, false negatives, and true negatives. As a final verification of the accuracy of the model, the precision, recall, and negative predictive value scores are also calculated. The equations are:

$$P = \frac{T_p}{T_p + F_p}$$

$$R = \frac{T_P}{T_p + F_n}$$

$$NPV = \frac{T_n}{T_n + F_n}$$

Where P = precision, R = recall, NPV = negative predictive value, T_p = true positive, F_p = false positive, F_n = false negative, and T_n = true negative. Precision represents the number of correctly identified values predicted out of the total wildfire susceptibility values predicted as positive. Recall represents the number of positive values out of the correctly identified values. Negative predictive value represents the number of correctly identified values out of the values predicted as negative. out of the wildfire susceptibility values predicted as negative, how many were correctly identified.

Results

By training the algorithm and testing it, I was able to examine the accuracy of the model for predicting the wildfire susceptibility of vegetation areas. In total the accuracy, also known as the model score, was 91.7%. This means that the random forest model was correct on its predictions 91.7% of the time. The test data was 30% of the original dataset, which equates to 8263 data entries. The first test data entry contained: vegetation_category = 3 (Oak woodlands), area_acres = 5.329091, size_class = 4, density_class = 1, and WUI = 0. The predicted burn_coverage variable by the random forest model was 0 (no burn), which is the same as the actual burn_coverage value from the dataset.

Test Data Results

	vegetation_category	area_acres	size_class	density_class	WUI	burn_coverage (prediction)
1	3	5.329091	4	1	0	0
2	5	4.475406	4	1	0	0
3	4	1.421915	9	1	3	0
4	1	1.194808	9	4	0	1

TABLE 2. A sample of the test data results from the random forest model. This indicates that the model predicts a high number of burn_coverage variables as 0 even with differing input data, reflecting the dataset.

To verify the accuracy of the model, a confusion matrix was used to show the true positives, false positives, false negatives, and true negatives of the model. There were in total: 63 true positives, 330 false positives, 360 false negatives, and 7510 true negatives. The model predicted by far the most true negatives, or data entries with a burn_coverage value of 0. This is likely because the majority of the dataset contained true negatives, totaling at 94.8% of the dataset. The model predicted 90.9% astrue negatives, showing the similarity between the true negative percentage of the dataset and model. Furthermore, a confusion matrix with percentages reveals the percentage breakdown of each category of the confusion matrix. A limitation of the model was its low accuracy in predicting true positives; however, this is likely due to the low number of true positives in the dataset.

Percentage-Based Confusion Matrix

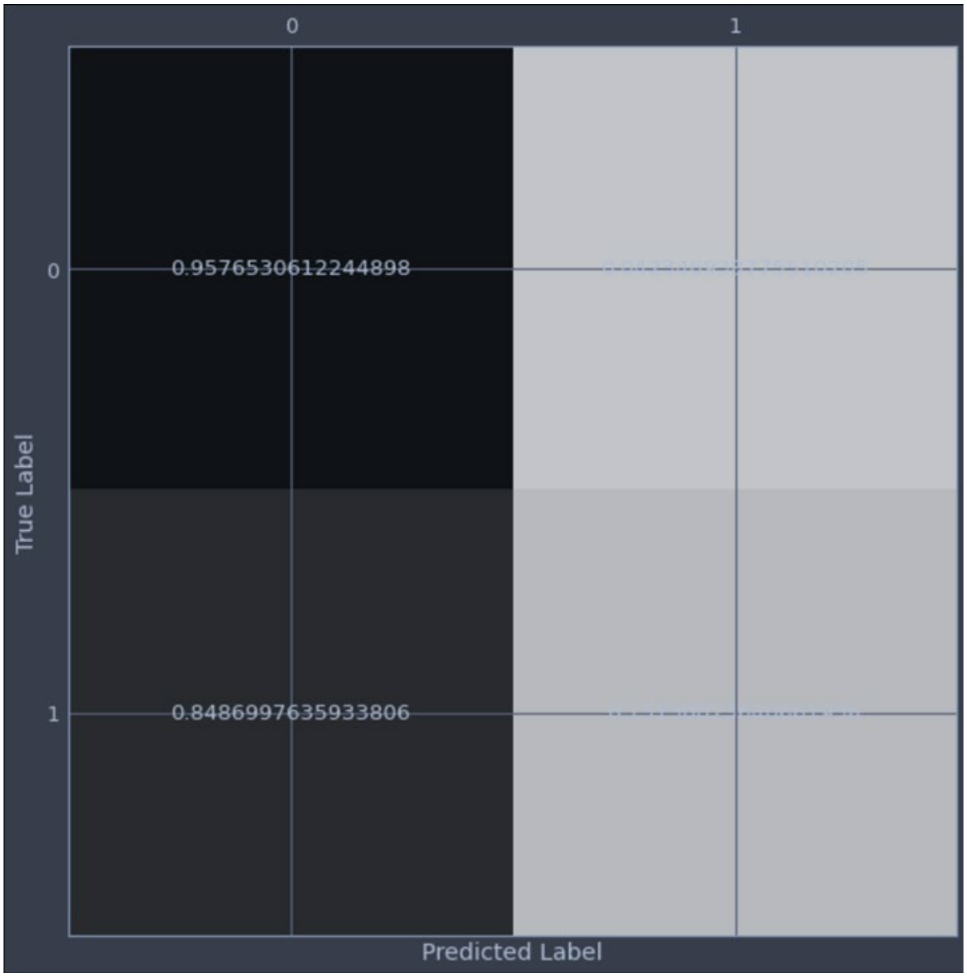

FIG. 3. A percentage-based confusion matrix of the model. The top left = true negative (95.7% of true negatives), top right = false positive (4.23% of true negatives), bottom left = false negative (84.9% of true positives), and bottom right = true positive (15.1% of true positives).

The precision score of the model was 16.0% and the recall score was 14.9%. The unusually low precision score and recall scores are due to the nature of the dataset. The high value of true negatives in relation to true positives in the dataset results in low precision and recall scores, even with a high accuracy score. The negative predictive value score was 95.4%, indicating that the model is reliable in predicting negative values. Because of its increased performance in accurately predicting negative values, the random forest model is likely more useful as a means of determining which vegetation areas do not need as much precautions as compared to determining which ones do.

To test the proficiency of the random forest model and the overall performance of the dataset, two subsequent predictions were made. The first was predicting the fire year of a vegetation region. For this model, the target attribute was year_burned, and the burn_coverage attribute was removed so as to not skew the data. Unfortunately, the accuracy of this model was only 33.4%, proving that there was simply not enough data to accurately predict the year of fire. Another challenge with this model is that the year_burned attribute only identifies the most recent year the vegetation region was burned, meaning that in previous years, that same region could have burned, but the model would have no way of knowing this. The other prediction made was to predict the vegetation category, with vegetation_category being the target attribute. This model performed much better compared to the year_burned model, with an accuracy of 64.5%. This model is still not nearly as accurate as the original burn_coverage model, but shows the proficiency of the random forest algorithm and the range of information in the dataset.

As a visual aid for the vegetation region dataset (Thorne, 2020), another dataset showing the origins of wildfires in Napa County and surrounding areas (USDA Forest Service, 2022) was used. This dataset shows the origin points of all the fires to have occured in the area with color-coding for the fire size in acres, allowing for a better view of which type of areas receive lots of fires. Many of the fires with a size larger than 1 acre occured in the mountainous regions as opposed to in the valleys. The valleys are generally more populated and contain more agriculture as compared to pure wildlands, potentially resulting in less fires.

Fire Occurrence Map - Napa County

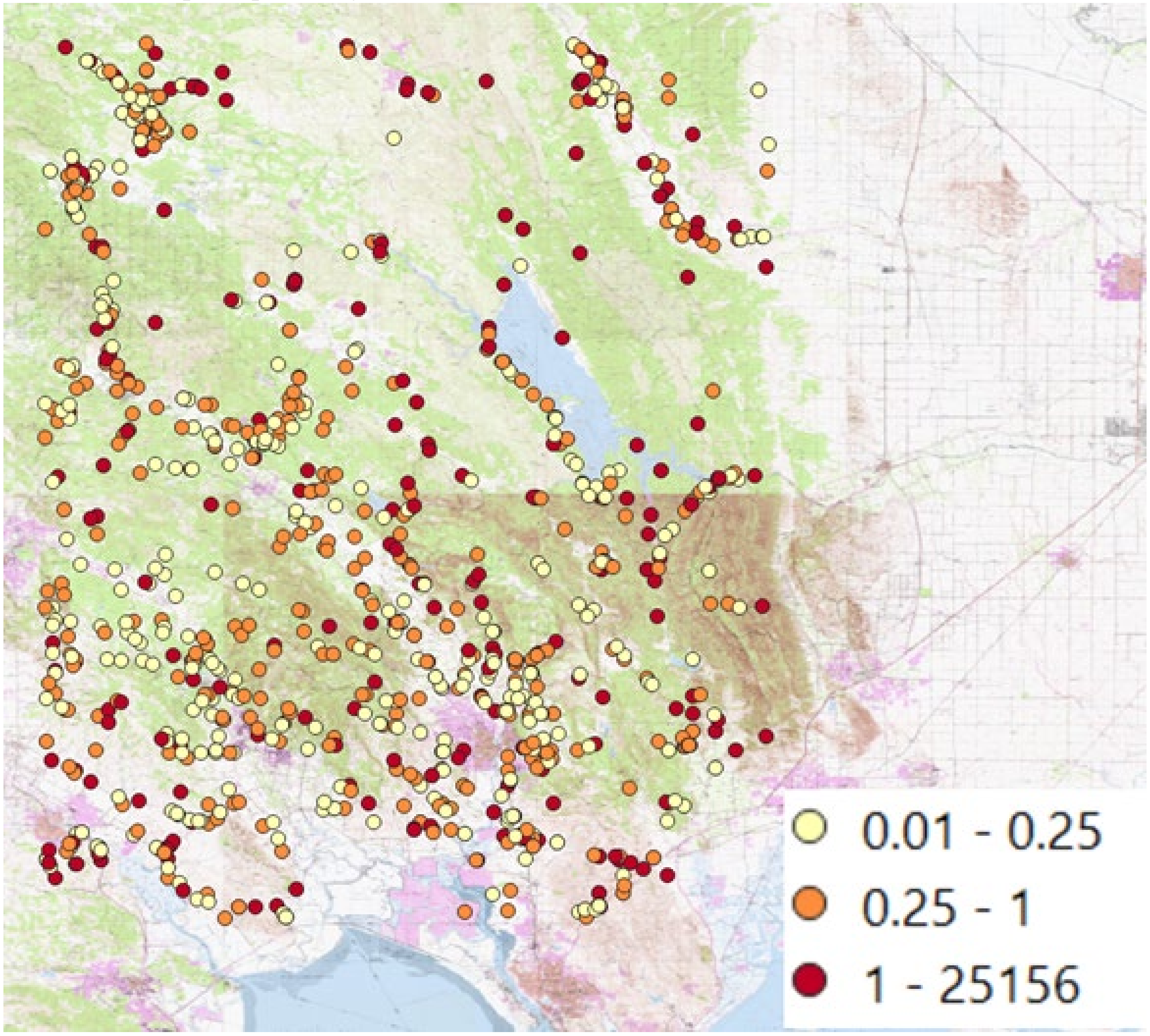

FIG. 4. A map of fire occurrence points in Napa County and surrounding areas. The points are color-coded by the size of the fire in acres. Many of the wildfires large than 1 acre tend to occur in the mountains rather than the valleys. Additionally, many of the fires form in clusters, potentially indicating that certain areas are very fire prone.

The information provided from this map in conjunction with the random forest model could be a great asset for determining which areas in Napa County are most at risk of wildfires. Viewing an overlay of the fire occurrence map on the vegetation dataset could outline which vegetation types are most likely to start fires compared to which ones are most likely to have fires spread in them, as well as which ones have a high rate of both. Furthermore, the random forest model's high negative predictive value score indicates that overlaying the true negatives of the dataset with this fire occurrence map could be valuable in determining which areas of Napa County might be classified as low risk in terms of wildfires.

Discussion

The purpose of the study was to develop a method of predicting the wildfire susceptibility of vegetation in Napa County. I achieved this by using a random forest algorithm (Husted, 2022) with a vegetation dataset of Napa County (Thorne, 2020) to create a model that would accurately determine the wildfire susceptibility of a vegetation area. The random forest algorithm was made of a series of decision trees that assigned a binary value to the wildfire susceptibility of a vegetation area. The random forest then took the majority decision from the decision trees as the prediction for wildfire susceptibility. The model had an accuracy of 91.7%, identifying mostly true negatives, in accordance with the dataset. More specifically, the model had a precision of 16.0%, a recall of 14.9%, and a negative predictive value of 95.4%. The high negative predictive value indicates that the model performed well at detecting which vegetation areas have not been burned by wildfires. The low precision and recall scores indicate the model's limitations in differentiating between vegetation areas that have been burned by wildfires when compared to other vegetation areas that have not been burned.

This study can be useful in the future for both additional research and in fighting wildfires. The model's high negative predictive value indicates that it can be used in determining which vegetation areas are least likely to be afflicted by wildfires, thereby allowing fire prevention methods to be used more effectively by not targeting low-risk areas. A common wildfire prevention technique employed by firefighters is clear-cutting the forest floor (Francos et al., 2018), and having a more reliable method of knowing where to do these clearings could greatly benefit fire prevention methods. Knowing which vegetation types are more wildfire-prone than others will be able to direct firefighters to the right areas in preventing the spread of wildfires because they would know which areas would be most devastating if the wildfire spread to. In addition, the random forest model could be improved through future research by applying datasets with more data of burned vegetation areas or applying the model to other areas in Northern California outside of Napa County. The diverse vegetation of Napa County means that the model might be able to be accurately applied to other counties and regions in the area. This model for predicting wildfire susceptibility in vegetation areas can be part of the front to tackle California's wildfire crisis.

References

Pausas, J. G., & Keeley, J. E. (2019). Wildfires as an ecosystem service. *Frontiers in Ecology and the Environment*, *17*(5), 289-295. doi:10.1002/fee.2044

Muller, C. H., Hanawalt, R. B., & McPherson, J. K. (1968). Allelopathic control of herb growth in the fire cycle of California chaparral. *Bulletin of the Torrey Botanical Club*, 225-231. doi:10.2307/2483669

Miller, J. D., & Safford, H. (2012). Trends in wildfire severity: 1984 to 2010 in the Sierra Nevada, Modoc Plateau, and southern Cascades, California, USA. *Fire ecology*, *8*(3), 41-57. doi:10.4996/fireecology.0803041

Kramer, H. A., Mockrin, M. H., Alexandre, P. M., & Radeloff, V. C. (2019). High wildfire damage in interface communities in California. *International journal of wildland fire*, *28*(9), 641-650. doi:10.1071/WF18108

Amatulli, G., Rodrigues, M. J., Trombetti, M., & Lovreglio, R. (2006). Assessing long-term fire risk at local scale by means of decision tree technique. *Journal of Geophysical Research: Biogeosciences*, *111*(G4). doi:10.1029/2005JG000133

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

Bustillo Sánchez, M., Tonini, M., Mapelli, A., & Fiorucci, P. (2021). Spatial assessment of wildfires susceptibility in Santa Cruz (Bolivia) using random forest. *Geosciences*, *11*(5), 224. doi:10.3390/geosciences11050224

Ma, W., Feng, Z., Cheng, Z., Chen, S., & Wang, F. (2020). Identifying forest fire driving factors and related impacts in china using random forest algorithm. *Forests*, *11*(5), 507. doi:10.3390/f11050507

Collins, L., Griffioen, P., Newell, G., & Mellor, A. (2018). The utility of Random Forests for wildfire severity mapping. *Remote Sensing of Environment*, *216*, 374-384. doi:10.1016/j.rse.2018.07.005

Malik, A., Rao, M. R., Puppala, N., Koouri, P., Thota, V. A. K., Liu, Q., ... & Gao, J. (2021). Data-driven wildfire risk prediction in northern california. *Atmosphere*, *12*(1), 109. doi:10.3390/atmos12010109

Thorne, J. (2020, Jul. 22). Vegetation - Napa County Update 2016 [ds2899]. CDFW Vegetation Classification and Mapping Program. Retrieved July 29, 2022 from http://bios.dfg.ca.gov

Barrette, J., August, P., & Golet, F. (2000). Accuracy assessment of wetland boundary delineation using aerial photography and digital orthophotography. *Photogrammetric Engineering and Remote Sensing*, *66*(4), 409-416. doi:00099-1112/00/6504-409$3.00/0

Sawyer, J.O., Keeler-Wolf, T., & Evens, J.M. (2009). *A Manual of California Vegetation, Second Edition.* Sacramento: California Native Plant Society

Belgiu, M., & Drăguţ, L. (2016). Random forest in remote sensing: A review of applications and future directions. *ISPRS journal of photogrammetry and remote sensing*, *114*, 24-31. doi:10.1016/j.isprsjprs.2016.01.011

Husted, A. (2022, Sep. 20). U.S. Wildfire Prediction. https://github.com/jalexander03/dsc-5-capstone-project-online-ds-ft-041519

USDA Forest Service (2022, Jul. 14). Fire Occurrence FIRESTAT Yearly. USDA Forest Service. Retrieved July 29, 2022 from https://data.fs.usda.gov

Francos, M., Pereira, P., Mataix-Solera, J., Arcenegui, V., Alcañiz, M., & Úbeda, X. (2018). How clear-cutting affects fire severity and soil properties in a Mediterranean ecosystem. *Journal of environmental management*, *206*, 625-632. doi:10.1016/j.jenvman.2017.11.011