

Wildfire Prediction Modeling using R*

James McSweeney

University of Miami

CSC 411

March 11, 2022

Table of Contents

1. Introduction	3
2. Methods	6
3.1 Dataset	7
3.2 Introduction to important variables	8
3.3 Data acquisition	9
3.3.2 Remote sensing.....	9
Remote sensing advantages.	9
Remote sensing disadvantages.	9
3.3.3 Satellites.....	10
3.3 Database Design	12
3.4 Proposed solution.....	13
3.4.2 Classification methods	14
3.4.3 Hyperparameter tuning	15
4. Choice of Technology.....	16
4.1 Language	16
4.2 Packages	16
4.3 IDE.....	17
4. Results	18
5.1 Individual results	18
5.2 Comparison	21
5. Conclusion	21
6. References	23

1. Introduction

1.1.Motivation

Why Wildfires?

Wildfires plague the western United States and many other areas globally. Over the past 30 years wildfire rates and severity have been rapidly increasing. Projections indicate that wildfires are going to increase by 50% in the next 75 years.(1) In 2020, the top .1% of wildfires caused over half the acreage burned.(2) By finding areas susceptible to wildfires, government agencies can adapt to changing climate patterns and learn how to best employ their resources to prevent future wildfires. My project intends to predict potential wildfires based off remotely sensed data.

Are Wildfires Random?

Contradictory to popular belief, wildfires do not occur randomly. Many environmental factors work to create an ideal area for fire to spark and spread. This project will identify these factors that lead to a higher risk of wildfires.

The United States Federal Government spends approximately 1 billion each year in wildfire prevention.(3) This funding is spread out between 4 agencies: the Bureau of Indian Affairs, the Bureau of Land Management, the National Park Service, and the U.S. Fish and Wildlife Service. Although these agencies have the same goal, their different outlooks and tactics

have varying rates of success. For example, the National Park Service and the U.S. Fish and Wildlife Service refuse to cut down trees or shrubs to reduce highly flammable vegetation. In contrast, other bureaus sell timber to raise money to fight fires and combat risk. These preventive measures dwindle in comparison to the 5 billion spent yearly fighting active fires which leaves large rural areas vulnerable for the next wildfire. My model hypothesizes that increasing proactive funds and prevention can decrease the total damage of wildfires. Currently, the government focuses on reducing home damage and casualties caused by wildfires. My model distinguishes from this by estimating areas to understand the likelihood of wildfires. Wildfire distance to human population centers will not be considered.

1.2.State of the Art.

Wildfires have been a topic of interest for many researchers. In 2020 researchers at University of Tabriz in Iran published *Comparisons of Diverse Machine Learning Approaches for Wildfire Susceptibility Mapping*.⁽⁴⁾ Their goal was to find areas susceptible to wildfires in Amol County, Iran. They used one of the satellites proposed in my research for their data collection, MODIS. The analysis was performed with a large array of classification methods including artificial neural network (ANN), DM neural, least angle regression (LARS), multi-layer perceptron (MLP), random forest (RF), and Support Vector Machines (SVM).

Another paper has a slightly different objective but uses many methods seen in my project. These researchers from Cadi Ayyad University, Morocco published *Predictive modeling*

of wildfires: A new dataset and machine learning approach in 2019.(5) Like the previous paper they received data from the MODIS satellite. The data they used were satellite images instead of specific remote sensing data points. In their paper they stress how large the amount of data they could extract from these images was crucial to their success in predicting wildfires.

2. Definitions

Some relevant terms to the project are defined subsequently.

AppEARS: Application for Extracting and Exploring Analysis Ready Samples.

LOOCV: Leave One Out Cross-Validation. This is a type of cross-validation approach in which each observation is considered as the validation set and the rest (N-1) observations are considered as the training set.

CSV: A comma-separated values (CSV) file is a delimited text file that uses a comma to separate values.

GEOTIFF: Public domain metadata standard file type that enables georeferencing information to be embedded within an image file

Equirectangular projection: The projection maps meridians to vertical straight lines of constant spacing, and circles of latitude to horizontal straight lines of constant spacing.

Red reflectance: the amount of red light that is reflected off a surface. Highly correlated with plant growth.

MODIS: Moderate Resolution Imaging Spectroradiometer

3. Methods

3.0.1 Data

To start building the dataset both wildfire and non-wildfire points must be collected. Wildfire points were sourced from a Kaggle database.(6) Non-wildfire points were non water coordinates within the United States. I generated these with a random number generator using the United States boundaries as the boundaries of the generator. Each of the coordinate points had a randomly generated date assigned to it. It is important to note that the data before the start of the wildfire was used to understand areas that are susceptible. Once there were dates and coordinate pairs for all points. These points were inputted into the AppEEARS Database.(7) Once inputted the MODIS and DAYMET satellites needed to be selected. Following that the variables of interest were selected. This would complete the request for each point. Once each request was complete the data needed to be combined. All of the DAYMET points and MODIS points combined into two different .csv files. Then in R the values of the coordinates and dates would be used as a key to join the two tables. A new column had to be added to this new data frame to encode which point was a wildfire. Points were assigned a wildfire value if their coordinate date pair was found in the initial wildfire csv file. Once points were encoded the classification could begin.

3.0.2 Modeling

Modeling started with splitting the data into training and testing sets. The ratio of 80:20 was used. Each model was ran using 5-fold Cross validation. While modeling not all variables

were used. Some columns were removed because of NAs or irrelevance to the project. See feature selection under data preprocessing for more information. The target value for each method was the coding for if a point was a wildfire or not. The data was run through several classification methods.

3.0.3 Analysis

The results were put into confusion matrixes to make each of the predictions understandable. Each method was compared by accuracy of prediction and total runtime. This project was demoed in front of a live audience using new data to show the model being ran and tested with different data. This new data was procured in the same fashion as the previous data. The findings are shared through a presentation.

3.1 Dataset

In this project, a subset of the existing data on wildfires for the United States was employed. The original, full dataset consisted of over 100,000 entries. Data is recorded either bi-weekly or daily. To get prediction data, the day before the wildfire's start date had to be used. This was to understand what an area looks like that is prime for a wildfire to spark and spread. Class A, B, C fires were removed from the dataset. These are the smallest size of fires, ranking from one 1/8 of an acre to 100 acres. These were removed because detecting these small of a scale of fires has a lot of uncertainty. Most data were in the D, E, F, and G classifications, 100 acres and above. Non-wildfire data was collected by randomly receiving coordinates that fall within the continental United States. This led to an issue of points on lakes, rivers, and oceans. Fortunately, NASA data has a descriptive variable called land/water mask. This classified if

points were land or water. Naturally water has very high light reflectance and low vegetation indexes. These outliers are detrimental in training the model therefore they were all removed. Wildfires also do not happen on water so it would not help in classification.

3.2 Introduction to important variables

Vegetive indexes. There are two kinds of vegetative indexes used in the model. The main difference between the two is that the normalized difference vegetation index (NDVI) is combined with red reflectance and its normalized. Low values of NDVI generally correspond to barren areas of rock, sand, exposed soils, or snow. Increasing NDVI values indicate greener vegetation, including things like forests, croplands, and wetlands. The Vegetative indexes are recorded via MODIS satellites.

Weather. Different aspects of weather were also collected to aid in the model. These include precipitation, vapor pressure, cloudiness, and presence of ice. All weather-related information was obtained via the DAYMET satellites.

Daylight. While the amount of sunlight does not directly contribute to wildfires. Daylight is representative of the season and latitude. Where the values would be lower in the winter and higher in the summer because all data and analysis were from the United States.

Descriptive variables. One of the most important variables is Quality_Land/ Water_Mask_Description. Quality_Land/ Water_Mask_Description states if the data point is on land or on water. This is used to filter out points on water as there are no wildfires on the water.

These points also would skew the data since water has low vegetation and high reflectance.

Therefore, removing points that were not land was crucial in building a model.

3.3 Data acquisition

3.3.1 Introduction

Wildfire coordinates were obtained through a database which combined local, state, and federal wildfire data. Non wildfire points were generated through a random number generator.

Both of these points were run through the AppEEARS database which aggregates data from different satellites

3.3.2 Remote sensing

Remote sensing advantages.

- I. The satellites can access remote areas of the United States that are not normally reachable through the collection of ground measuring systems. These satellites can upload information about an area in as little as 3 hours.

Remote sensing disadvantages.

- I. The spatial resolution of the satellites makes it hard to record small events. For example, a small bush fire would not be recognized by a satellite until it got to a certain size.
- II. Temporary resolution is another limiting factor for data uploaded via satellite. Certain areas of the earth are recorded every 2 days to as seldom as 16 days.
- III. Spectral resolution affects passive sensors found in satellites. This means if there is dense cover from clouds or vegetation instruments that collect data via

bouncing waves off the earth will not work. Spectral resolution does not affect thermal sensors.

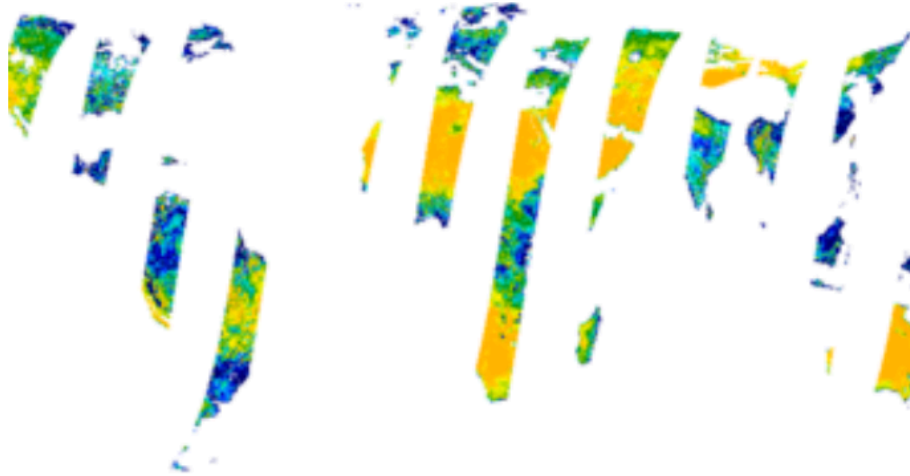


Figure 1: this represents how satellites collect data from the globe. (Source: “Wildfires Data Pathfinder.” NASA, NASA, 5 May 2022, <https://earthdata.nasa.gov/learn/pathfinders/wildfire-data-pathfinder>.)

3.3.3 Satellites

MODIS. MODIS has a viewing swath width of 2,330 km, and it images the Earth in 36 spectral bands, or groups of wavelengths, ranging from 0.405 to 14.385 μm . It collects data at three spatial resolutions: 250, 500, and 1,000 meters. The average rate of data collection is 6.1 megabits each second.(8)

DAYMET. DAYMET is a data product derived from a collection of algorithms and computer software designed to interpolate and extrapolate from daily meteorological observations to produce gridded estimates of daily weather parameters. Weather parameters generated include daily surfaces of minimum and maximum temperature, precipitation, vapor pressure, radiation, snow water equivalent, and day length produced on a 1 km x 1 km gridded surface.(9)

Coordinate System. It is important to note that the coordinates that NASA uses are sourced from the Equirectangular projection. The advantage of using an equirectangular projection is that it preserves equal distance between latitude and longitude lines.

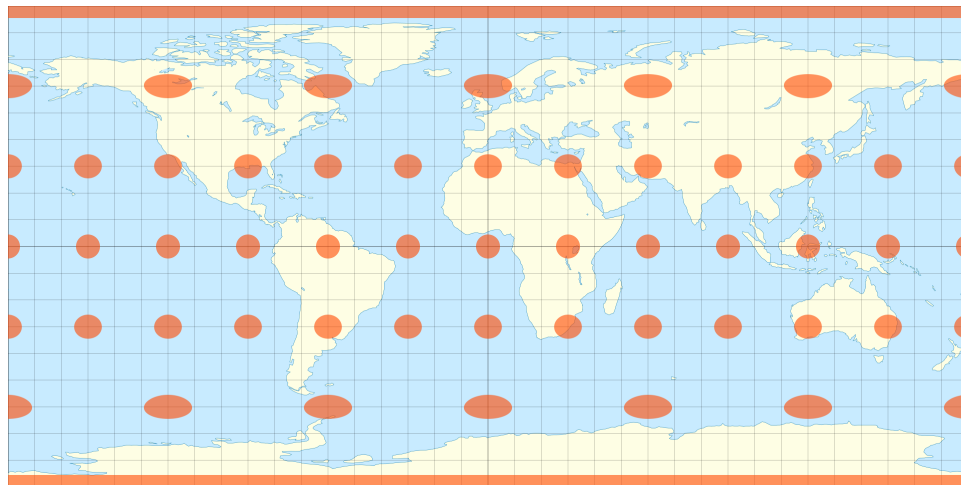


Figure 2: Equirectangular projection of the world. Circles are used to show how some space is altered.(Source: “Equirectangular Projection.” Wikipedia, Wikimedia Foundation, 4 Apr. 2022, https://en.wikipedia.org/wiki/Equirectangular_projection#/media/File:Plate_Carr%C3%A9_with_Tissot's_Indicatrices_of_Distortion.svg.)

3.3 Database Design

Data Source. This data is sourced from NASA Earth Data. The conventional way to receive data is in the form of GEOTIFF files. GEOTIFF files are a type of raster file. They are raster files because the data originally comes from remote sensing satellites which scan large areas of data. But instead of the user having to convert GEOTIFF files NASA has set up the AppEEARS which enables users to extract small point samples based off equirectangular projection Latitude and Longitude coordinates. When performed in this way the user receives a .CSV file. This file can be manipulated in R, while a GEOTIFF file can only be manipulated in applications like ARC GIS.

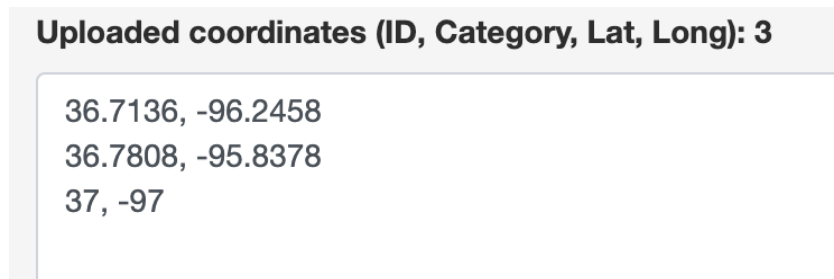


Figure 3: Inputting coordinates to AppEEARS database (Source: “AppEEARS.” NASA, NASA, <https://appeears.earthdatacloud.nasa.gov/task/point>.)



Selected layers		
 dayl	1000m, Daily	—
 prcp	1000m, Daily	—
 srad	1000m, Daily	—
 swe	1000m, Daily	—
 vp	1000m, Daily	—

Figure 4: Selecting features from the AppEEARS database. (Source: “Appeears.” NASA, NASA, <https://appeears.earthdatacloud.nasa.gov/task/point.>)

3.4 Proposed solution

3.4.1 Data Preprocessing

K-fold Cross validation. K-Fold Cross validation is a resampling process that helps make classifications by cycling through different training and testing sets. The K in K-folds means how many subdivisions the data should be split into. To keep an 80:20 ratio of training to testing, 5-Fold was used. Cross validation aids in preventing overfitting of training data. The only downside for running K-fold cross validation is it has a higher computational cost because the methods are running multiple times.

Feature selection. I did not use all the columns that I extracted. For each variable there was an accompanying quality column which represents the accuracy of the variable. This was factored into the data preprocessing in order to use high quality data but for classification purposes it would not be relevant. Other columns that were deleted were the encodings for some variables such as precipitation. There were ranges associated with High, medium, low and no precipitation. Since precipitation was covered in my model through the actual inch value per day, the coding was not needed. Lastly there were

nine columns associated with the number of clouds covered over a point. These columns had a lot of repeat information and a lot of missing data; therefore, they were not used.

3.4.2 Classification methods

SVM. Support vector machines are a supervised classification method that uses hyperplanes to classify data. The hyperplane is placed in between the average between the two closest points that separate each grouping of points. Having this line would be a very high bias and low variance. Therefore the hyperplane is moved according to the soft margin. Instead of using the two closest points it will start to use slightly further points. This leads to misclassification in most models. Cross validation is applied to the soft margin in order to have the correct balance between bias and variance. SVM hyperplanes can either be straight lines or using the Kernel trick to produce a circular boundary.

Naïve Bayes. Naïve Bayes is a probabilistic classifier which applies bayes theorem with independent assumptions between the data features. Naïve Bayes assigns a conditional probability to each feature in each class. To make a classification the likelihood for every feature seen in the instance of the test data is multiplied together and compared to each other predicting class. Whichever has the largest value becomes the classification. Because of the fact that likelihoods are multiplied together, adding pseudo counts is very important. A pseudo count could be any number, but it needs to be evenly added to each feature. This is to avoid if a likely hood is zero, that zero value would automatically make the classification zero.

KNN. K Nearest Neighbors is a classification method that's uses majority voting of surrounding points to generate a consensus to classify new points. The points that are allowed to vote are the K closest points to the test point. There are many different approaches to how to calculate distance to find the closest K points, in this project Euclidean distance will be used.

Classification trees. Classification tree analysis is a type of machine learning algorithm for classifying based off of binary decisions. The trees are built using binary recursive partitioning. This

iterative process splits data into 2 groups. Splitting into branches continues until a node that can classify the points are reached.

Random Forest. Random Forest is an ensemble learning method for classification. It starts by constructing multiple classification trees at training time. The output is the class selected by the majority of trees. Random forest provides additional value when compared to classification trees because it is less likely to over fit the training data.

3.4.3 Hyperparameter tuning

A hyper parameter is a characteristic of a model that is external therefore not estimated from the data. Hyper parameters guide the learning process of an algorithm. Common examples of hyperparameters are K in K Nearest Neighbors, number of variables used in tree-based models, and cost in Support Vector models.

Search Methods. There were 2 different kinds of hyperparameter tuning used. The first method performed was grid search tuning. Grid search tuning computes optimum values by exhaustively searching on specific parameter values. This is done by creating a grid of hyper parameters and testing each combination. The second method performed was random search. Similar to grid search, random search also creates a grid but instead of testing it selects random combinations to train the model. When comparing the two, grid search is shown to use less resources than grid search, while grid search is more thorough.

Random Forest. The random forest model was the model that received both kinds of hyperparameter optimization. The hyperparameter that was optimized was the number of variables sampled as candidates at each split. This hyperparameter is important because it's there to balance between time complexity and accuracy.

KNN. The hyper parameter that was tuned for K Nearest Neighbors was K, the number of neighbors to consider in the classification process. The K value is intended to be the smallest number

while still correctly classification new points. Not only is a large value of K computationally expensive but also is counter intuitive towards the idea of K Nearest Neighbors.

SVM. Support vector machines tuned for the cost variable. Cost in support vector machines is the parameter which determines how hard or soft the boundary should be. A soft boundary would lead to several misclassifications in order to reduce the variation in the test set. While a hard boundary has small bias on the training data, which leads to a higher variation in the test set.

4. Choice of Technology

4.1 Language

R. R was chosen as the language for data analysis. One advantage of R is the access to 10,000 packages. Most functions and analysis for this project will be from the equations in Base R. However, packages like GGLOT provide extra analysis tools for data visualization. R is a platform-independent language. It is a cross-platform programming language, meaning that it can be run quite easily on Windows, Linux, and Mac.(10)

SQL. Standard Query language is used to receive and filter data from large datasets. The only time SQL was used was to select data from the wildfire dataset.(11)

4.2 Packages

Stats. This is included in Base R. Stats contains the classification method equations for R.
(12)

DYPLR. DYPLR will simplify making a data frame from CSV data. Along with other methods the Filter method is crucial in the data cleaning process.(13)

GGPLOT. GGPLOT is a very powerful package for visually representing data. GGPLOT is a part of the Tidyverse family of packages in which DYPLR is also part of.(14)

Random Forest. Random forest package holds methods for creating classification trees and random forest models.(15)

Class. holds functions for K nearest neighbors' classification. In addition, to the models it provides hyperparameter tuning methods.(16)

Caret. Classification and regression training (CARET) is a package that aids in splitting of training and testing sets.(17)

E1071. Package that includes tuning and classification methods. Methods include SVM and Naïve Bayes.(18)

4.3 IDE

R Studio. The Interactive development environment used for this project is R studio. R studio seamlessly lets users download new packages from the internet. The layout makes it easy to locate variables and their values. (19)

Oracle SQL Developer. This was necessary to extract wildfire coordinates through an SQL database. (20)

4. Results

5.1 Individual results

Importance of Variables. The variables are ranked in order of importance. The more asterisks present the higher the influence the variable has on the model. The variables in order of significance: SRAD, Daylight, Vapor Pressure, NDVI, EVI, precipitation, NIR reflectance and SWE

DAYMET_004_dayl	1.920e-05	8.285e-07	23.172	< 2e-16 ***
DAYMET_004_prcp	-1.604e-03	1.481e-03	-1.083	0.280
DAYMET_004_srad	1.287e-03	1.296e-04	9.928	< 2e-16 ***
DAYMET_004_swe	-6.987e-05	2.042e-03	-0.034	0.973
DAYMET_004_vp	1.171e-04	2.516e-05	4.655	4.51e-06 ***
MOD13A1_006__500m_16_days_EVI	-1.367e-03	1.137e-04	-12.016	< 2e-16 ***
MOD13A1_006__500m_16_days_NDVI	2.489e-04	2.059e-03	0.121	0.904
MOD13A1_006__500m_16_days_NIR_reflectance	-1.503e-04	3.280e-05	-4.583	6.24e-06 ***

Figure 5: Importance of Variables found through linear regression (Source: rStudio)

SVM. Support vector machines had a very good performance. SVM only misclassified six points. Giving it an accuracy of 93.8%. While the accuracy was very high, so was the computational cost. In my model Support vector machines were hyper tuned on the parameter of cost. The best model was found to have a cost of .1. The costs that were tested were 0.001, 0.01, 0.1, 1,5,10, and 100. Future work could include testing different types of kernels.

	0	1
0	44	2
1	4	46

Figure 6: Confusion matrix for Support Vector Machines. The columns are the truth values while the rows are the prediction values.

Naïve Bayes. Naïve Bayes was the worst performing classification method. Naïve Bayes misclassified 35 pieces of data from the test set. Giving it a performance of 73.5%. It is interesting to note that Naïve Bayes favored classifying points as wildfires. Almost 70 % of predictions were fire points where the actual amount was around 50 %. Naïve Bayes also made no incorrect non wildfire prediction seen as the bottom left corner of the confusion matrix.

	0	1
0	31	35
1	0	66

Figure 7: Confusion matrix for Naïve Bayes. The columns are the truth values while the rows are the prediction values.

KNN. KNN was the second-best classification method. It misclassified 4 out of 97 data points in the test set. Giving KNN a 95.8% accuracy rate. The optimal value for K was found to be 5.

	0	1
0	50	4
1	0	43

Figure 8: Confusion matrix for KNN. The columns are the truth values while the rows are the prediction values.

Classification trees. Classification trees misclassified seven predictions. This gives Classification trees a 92.7% accuracy rate. Classification trees were unlikely to be the best model because random forest is a collection of classification trees.

	0	1
0	40	6
1	1	50

Figure 9: Confusion matrix for Classification trees. The columns are the truth values while the rows are the prediction values.

Random Forest. Random Forest was the best model. Only misclassifying 1 piece of data in the test set. While Random Forest was the best model it was also the most computationally expensive. There were 500 trees in the random forest models. In addition, grid search was used to optimize the parameters so random forest was run many times in order to compare the success of different hyper parameters. The optimal value for the number of variables considered per split was found to be 3.

	0	1
0	50	1
1	0	46

Figure 10: Confusion matrix for Random Forest. The columns are the truth values while the rows are the prediction values.

5.2 Comparison

A summary is provided showing the different values for accuracy and the time it took to run each algorithm on a 2019 model MacBook Air. Important specs include a CPU: 1.6GHz Intel Core i5-8210Y (dual-core, 4 threads, 4MB cache, up to 3.6GHz) and RAM: 8GB (2,133MHz LPDDR3). As noted in the previous section, random forest had the best performance and showed good compromise in terms of runtime. In turn, the fastest method – Naïve Bayes – achieved the lowest accuracy.

Classification Method	Accuracy	Runtime (s)
KNN	95.8%	3.1
Classification Tree	92.7%	.7
Random Forest	98%	40
Naive Bayes	73.5%	.4
SVM	93.8%	1.09

Figure 11: Accuracy and runtime for each method

5. Conclusion

5.1.Challenges

Data cleaning proved to be very difficult. The issue was that there was too much free NASA data available. While too much data is easier to work with than with too little data, finding usable data proved very challenging. This colossal amount of data is because NASA

saves data either in GEOTIFF or raster files. These files can get as big as a Gigabyte each week per satellite. Trying to work with a large subset of that data would crash my R Studio. The solution was to randomly sample points in the United States. The main difference was that point data was in comma separated value format instead of the GEOTIFF file type. I received this guidance from a data professional at AppEEARS whom I contacted with my problem. Once the data was in a comma separated value format, I was able to perform analysis in R.

5.2.Future Work

Wildfires plague the United States. Without proper detection and conservation techniques the problem will only get worse. There have been 17,299 wildfires in the United States since the beginning of the semester. A total of 831,842 acres burned. That is almost the size of the state of Rhode Island. My project completed its goal of assessing areas which are vulnerable wildfires in United States. Future work would be valuable in making my model more accurate and functional for helping agencies focus their efforts on areas that need it the most. For example, adding extra weather data such as lightning strikes or distance to power lines could be useful. These are the next two largest causes of wildfires that were not included in the model. Additionally, if broad satellite data could be used, then it would be possible to have a real time heat map of areas that are highly susceptible to wildfires.

5.3.GitHub

All materials are found in a GitHub repository. This includes all R files for building the dataset and Classification methods. Also has the report and presentation for further information in the project steps. McSweeney, James. “Jamesmcsweeney/Wildfires: Machine and Statical Learning Project about Understanding What Variables Contribute to Wildfires and How We Might Be Able to Predict Them.” GitHub, <https://github.com/JamesMcSweeney/WildFires>.

6. References

1. Vetter, David. “Wildfires Could Increase 50% This Century. Here's What to Do about It.” Forbes, Forbes Magazine, 25 Feb. 2022, <https://www.forbes.com/sites/davidrvetter/2022/02/23/wildfires-could-increase-50-this-century-heres-what-to-do-about-it/>.
2. Wildfire Statistics - Federation of American Scientists. <https://sgp.fas.org/crs/misc/IF10244.pdf>.
3. “Budget.” U.S. Department of the Interior, 22 Apr. 2022, <https://www.doi.gov/wildlandfire/budget>.
4. Gholamnia, Khalil, et al. “Comparisons of Diverse Machine Learning Approaches for Wildfire Susceptibility Mapping.” *Symmetry*, vol. 12, no. 4, Apr. 2020, p. 604. Crossref, <https://doi.org/10.3390/sym12040604>
5. Sayad, Younes Oulad, et al. “Predictive Modeling of Wildfires: A New Dataset and Machine Learning Approach.” *Fire Safety Journal*, Elsevier, 24 Jan. 2019, <https://www.sciencedirect.com/science/article/pii/S0379711218303941>.
6. “1.88 Million US Wildfires.” Kaggle, <https://www.kaggle.com/datasets/rtatman/188-million-us-wildfires>.
7. “Appeears.” NASA, NASA, <https://appeears.earthdatacloud.nasa.gov/>.
8. “Modis Web.” NASA, NASA, <https://modis.gsfc.nasa.gov/about/>.
9. “Daymet.” NASA, NASA <https://daymet.ornl.gov/>
10. “The R Project for Statistical Computing.” R, <https://www.r-project.org/>.
11. “SQL.” Wikipedia, Wikimedia Foundation, 4 May 2022, <https://en.wikipedia.org/wiki/SQL>.
12. “The R Stats Package.” R: The R Stats Package, <https://stat.ethz.ch/R-manual/R-devel/library/stats/html/00Index.html>.
13. “A Grammar of Data Manipulation.” A Grammar of Data Manipulation •, <https://dplyr.tidyverse.org/>.
14. “Create Elegant Data Visualisations Using the Grammar of Graphics.” Create Elegant Data Visualisations Using the Grammar of Graphics •, <https://ggplot2.tidyverse.org/index.html>.

15. Randomforests RC - Cran.r-Project.org. <https://cran.r-project.org/web/packages/randomForestSRC/randomForestSRC.pdf>.
16. Class - Cran.r-Project.org. <https://cran.r-project.org/web/packages/class/class.pdf>
17. Caret - Cran.r-Project.org. <https://cran.r-project.org/web/packages/caret/vignettes/caret.html>
18. E1071 - Cran.r-Project.org. <https://cran.r-project.org/web/packages/e1071/index.html>
19. “Open Source & Professional Software for Data Science Teams.” RStudio, 4 May 2022, <https://www.rstudio.com/>.
20. SQL Developer, <https://www.oracle.com/database/technologies/appdev/sqldeveloper-landing.html>.