Data-driven Wildfire Risk Prediction with augmented spatial accuracy

A Project Report

Presented to

The Faculty of the Department of Applied Data Science

San Jose State University

In Partial Fulfillment

Of the Requirements for the Degree

Master of Science in Data Analytics

By

Prathusha Koouri, Qiao Liu, Nandini Puppala,

Megha Rajam Rao and Venkata Anil Kumar Thota

May 2020

APPROVED FOR DEPARTMENT OF APPLIED DATA SCIENCE

Dr. Jerry Gao, Project Advisor

Dr. Lee C. Chang, Department Chair

ABSTRACT

Data-driven Wildfire Risk Prediction with augmented spatial accuracy

By

Prathusha Koouri, Qiao Liu, Nandini Puppala, Megha Rajam Rao, Venkata Anil Kumar Thota

In this research project, we target the problem of wildfire risk prediction in Northern California. Rampant wildfires have plagued the state of California time and again. In 2018, wildfires have cost nearly 800 million dollars in economic loss and claimed more than 100 lives in California [1]. Over 1.6 million acres of land have been burned and caused huge environmental damage. Many researchers have touched upon this subject. Conventional solutions implement mathematical or statistical models to predict the risk of fire. Recent studies, based on machine learning, are restricted to a limited number of parameters. They are not necessarily temporal or spatial specific either. To fill this gap, we are introducing a comprehensive model with diverse parameters to augment the spatial accuracy in predicting the wildfire risk in Northern California. Cutting-edge techniques such as machine learning and neural networks were implemented before optimization and evaluation. The plan was to conduct research in an incremental manner, starting with subsets of parameters and machine learning algorithms such as Weighted Decision Trees, Random Forest, Adaboost, and Gradient boosting before progressing to complicated Neural Networks such as Long short-term Memory (LSTM). The models were integrated to generate ensemble models, whereas the parameters were merged to create combined models, both of comparable accuracy. At the culminating point, we created a user interface aptly named the 'Spartan Wildfire Risk Prediction system (SWiPS)' with augmented spatial sensitivity. It can be accessed as a web portal by users. This robust model can aid relevant agencies in wildfire risk prediction.

ACKNOWLEDGMENTS

We would like to express our sincere gratitude to mentor and project guide Dr. Jerry Gao for his relentless supervision, wisdom, and guidance. With him at our side, we never ran out of ideas to generate a better version at each step. An esteemed Professor at the Department of Computer Engineering at San Jose State University, his constant support and faith in our abilities drove this noble undertaking to fruition.

We express our utmost gratitude to Dr. Lee Chang, Professor and Chair of the Department of Applied Data Science (DADS) for his ceaseless enthusiasm and encouragement. His succinct expectations, communication, and deadlines made the project clear and hassle-free. We thank the Department in its entirety, including Dr. Chang and dedicated staff such as Jefferson Armanini and Danny Weiner, for the continued support and engagement. We are immensely grateful for the opportunity to pursue our master's program in Data Analytics.

We express our gratitude from the bottom of our hearts to all the experts, who shared their knowledge and information during the course of this project. Below experts were extremely generous with their time and wisdom.

- Dr. Loic Dutrieux - Noted French Geospatial specialist and post-doctoral fellow answered our queries on anomalies in Landsat Satellite data extraction.
- Dr. Thomas J. Vandal - AI and Earth Science Research Scientist at NASA Ames Research Center shared his insights.
- Dave Sapsis - Seasoned expert from Cal Fire/California Dept of Fire and Forestry provided guidance regarding data collection.

We wholeheartedly thank National Aeronautics and Space Administration (NASA) and Cal Fire personnel for their willingness to offer unofficial advice. Finally, we would like to thank the following groups for their time and technical consultation.

- Justlight Technology Inc., Fremont, CA
- Industrial Technology Research Institute (ITRI), Taiwan
- Research Center of Smart Technology and Computing for Complex Systems, San Jose State University, CA

TABLE OF CONTENTS

Table of Figures

Table of Tables

1. Introduction

1.1 Project Goals and Background

In this era of climate change and rising temperatures, blazing wildfires are becoming a year-round phenomenon [2]. Wildfires, also known as Wildland fires, forest fires or brush fires, are uncontrolled fires sweeping across millions of acres of land, causing severe and extensive damage to our ecosystem [2]. Often triggered by extreme heat, stormy weather, dry fuel and human factors, efforts are made to contain its rapid spread, evacuate the human population and mitigate the losses. These fires, at an insurmountable pace of 14 miles per hour, are inescapable for the average human. According to National Geographic, an average of 7 million acres of land in the United States(US) was torched by 72,400 wildfires in the 21st century, which is more than double the numbers from the previous century. In fact, the nation encountered the largest fire season in the country's history in 2015 when wildfires scorched an estimated 10 million acres of land [3].

Record-breaking conflagrations have wreaked havoc on the vegetation, wildlife, atmosphere and human population alike, leading to loss of life and property, the extinction of flora and fauna, plummeting air quality, global warming, among others [4]. Specifically, in the State of California, rampant wildfire outbreaks are an eminent concern. Fire season has extended due to drought, insufficient precipitation and human activities. As per California Department of Forestry and Fire Prevention (Cal Fire), the year 2018, deemed one of the worst years in California history, witnessed 7571 fires that burned across 1.6 million acres of land and claimed more than 100 lives [5]. Figure 1.1 provides a glimpse of the long history of Wildfires in California, while Figure 1.2,1.3 and 1.4 illustrate the Death toll, Cost of wildfires and Acreage burned by wildfires,

respectively [4]. It is evident that vast expanses of land have been ravaged by wildfires in recent years, which lead to a spike in death toll and heightened costs to suppress the wildfires due to a rise in the acreage devoured. 974 million dollars were expended for fire suppression by the state of California in the fiscal year 2017-2018, according to Cal Fire.

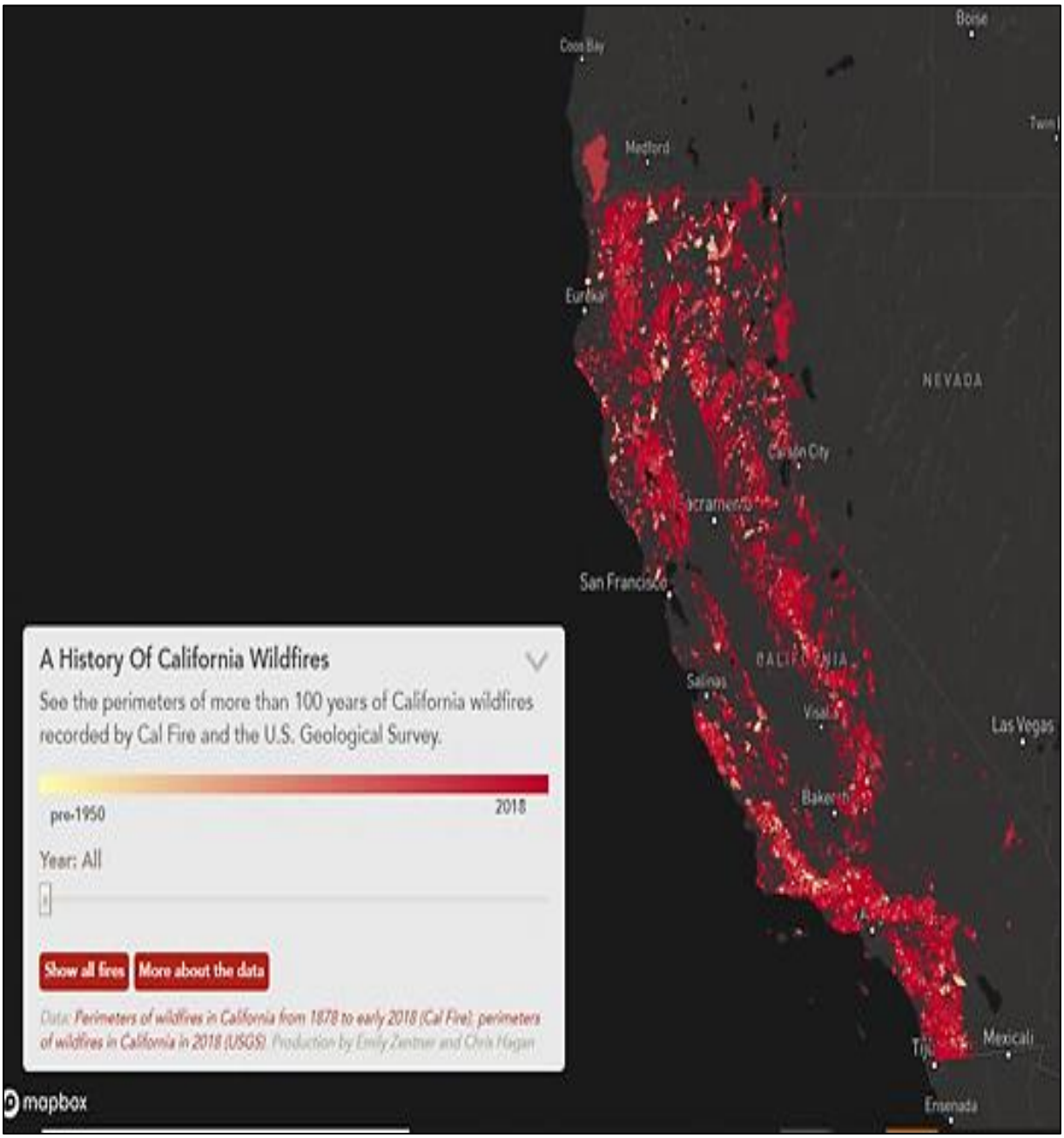


Figure 1.1. *History of Wildfires in California (Years 1878 to 2018)*[6]

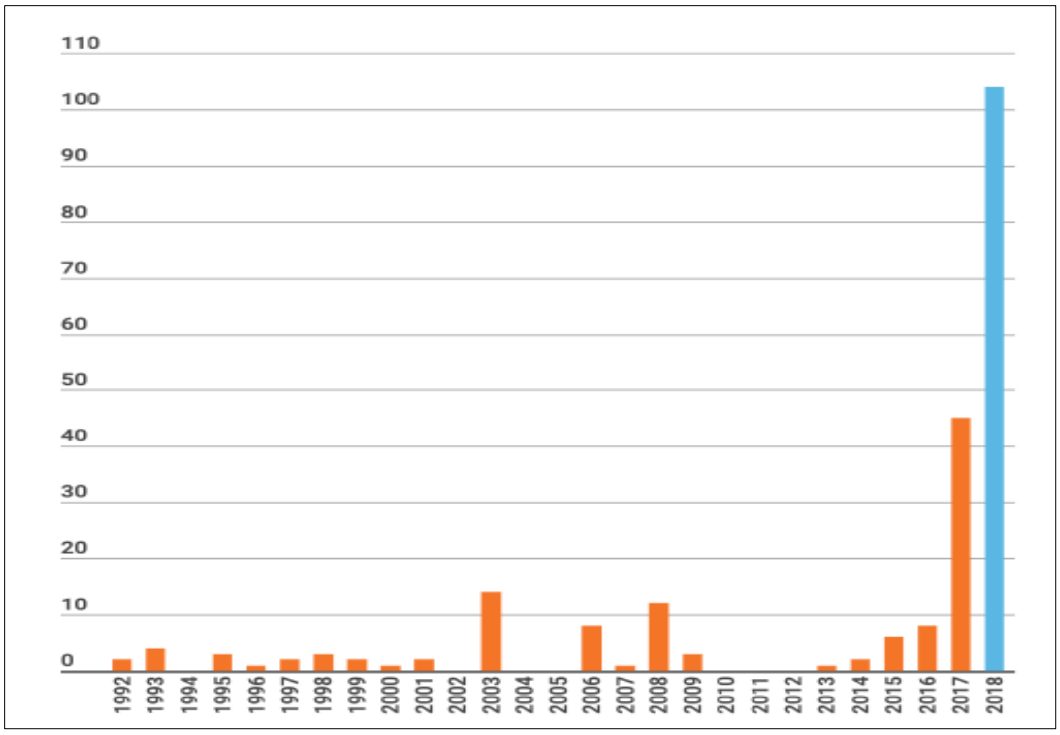


Figure 1.2. *Yearly death toll in California [4]*

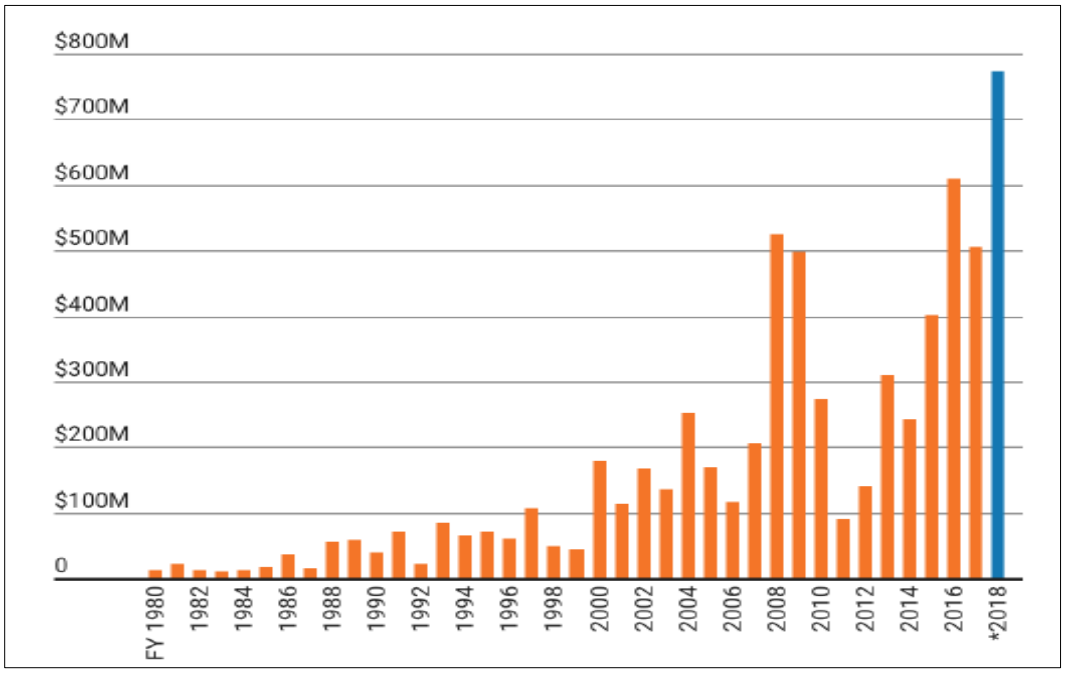


Figure 1.3. *Cost of Wildfires from the year 1980 to 2018 [4]*

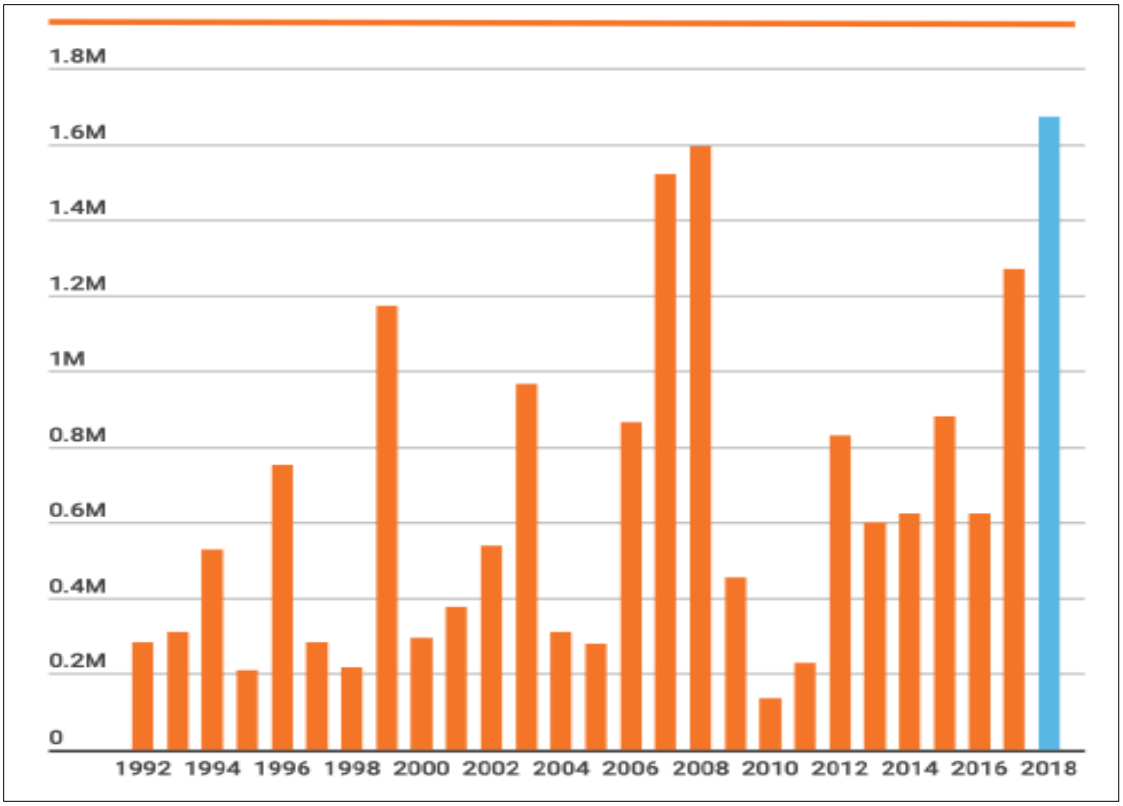


Figure. 1.4. *Yearly Estimation of acreage burned in California [4]*

1.1.2 Project Goals

The aforementioned statistical revelations can be considered as ample evidence to prove that recent years have been marked by a stark increase in duration and level of destruction caused by wildfires. As a result, Wildfire prediction has been at the epicenter of various studies pertaining to fire prevention, management and response. A facet of paramount significance, Prediction plays a vital role in resource allocation, containment and mitigation by public and private agencies. As per our reading [7], conventional approaches employ mathematical and statistical methods with a heavy reliance on equations and calculated metrics. These traditional techniques suffer from lower accuracy, abysmal efficiency, unclear cut-offs, the complexity of equations and lack the processing capabilities to support real-time decision making. With the advent of computer-assisted fire

prediction, Machine Learning and Neural Networks have been utilized to improve this lagging outcome. However, only a few parameters were employed in the earlier studies, which were limited in accuracy. To cope up with this deficiency, we have incorporated a wide range of parameters and indices and focused on improving the spatial and temporal accuracy of the outcome. This will eventually facilitate the smart management of wildfires. Hence, we defined the goal of this study as the successful completion of a robust model-backed system with augmented spatial and temporal sensitivity, to efficiently predict wildfires in Northern California.

1.1.3 Elements of a Fire

To understand the elements of a fire, let us look at the fundamental model widely known as the Fire Triangle in Figure 1.5 [8]. It showcases the interdependent ingredients of a fire in the three sides of a triangle. Flames happen to be a visual indicator of the onset of a fire and form the simplest model in the multi-scale representation. Below are the three legs of this triangle [9].

- Heat is an essential component responsible for ignition, spread and fire sustenance. It dries up surrounding fuel and warms the temperature.
- Fuel is any combustible material with a conducive shape, size and placement. Moisture content must be on the lower side.
- Oxygen is a crucial component as most fires require a minimum of 16 percent oxygen content, which is way below the 21 percent oxygen in the atmosphere. Hence, Air can act as an oxidizing agent in the oxidation process.

Therefore, ignited fuel(s) in the presence of oxygen at high temperatures turn into larger fires. Even smoldering fires at lower temperatures turn into destructive fires when they gradually attain the ignition temperature.

An evolved form of the fire triangle is a fire tetrahedron that represents the chemical chain reaction in the center of this representation [10].

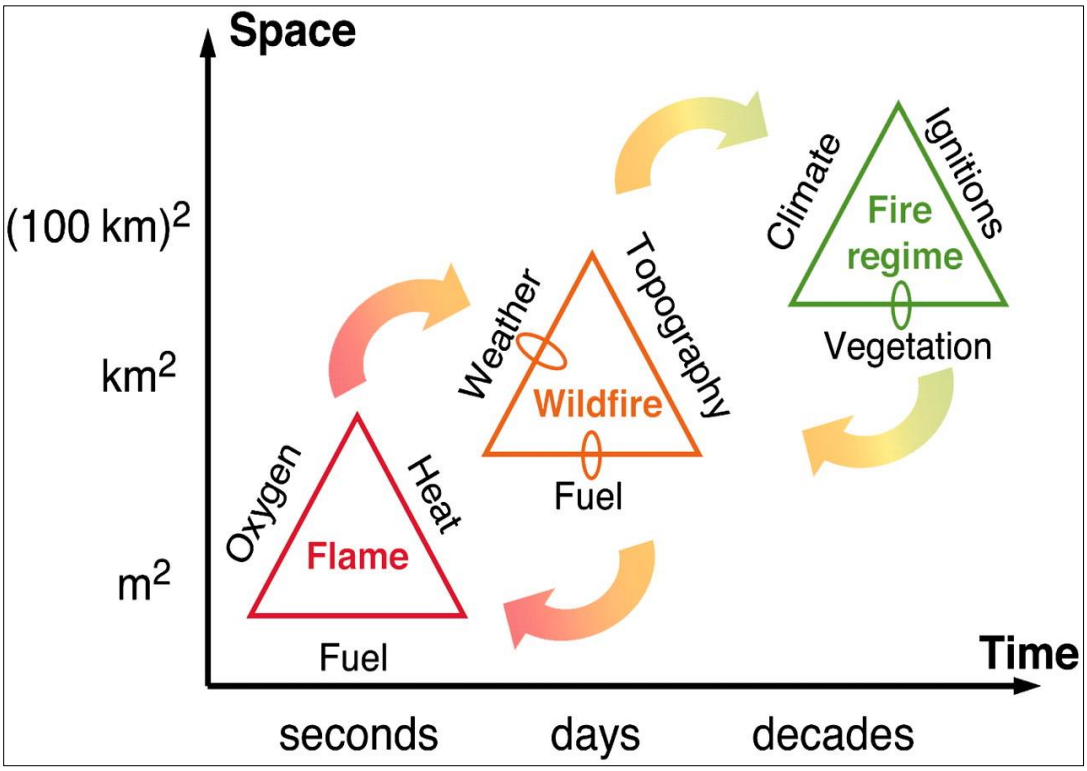


Figure 1.5. *Multi-scale Fire triangles* ***[8]***

When flames burn for days, they spread across the land and turn into wildfires. Presence of fuel, conducive weather and topography aids in fire propagation. Over the decade, the proliferation of wildfires forms a fire regime, which is driven by the climate, ignition and vegetation factors [8].

1.1.4 Wildfire Types & Regional Factors

Wildfires can be broadly classified into three categories namely crown fires, surface fires and ground fires [3].

- Crown fires affect the upper foliage and spread rapidly.

- Ground fires are limited to the forest floor and underneath.
- Surface fires fall in between these two categories and burns the intermediate segment

Few regional factors need to be highlighted. California drought, insufficient rainfall along with fire mismanagement played a contributing role in triggering wildfires in the past few decades. Natural fires must be contained rather than extinguished, as it is an integral process beneficial to the ecosystem for forest cleansing and disease eradication, seed propagation and soil fertilization, giving rise to new vegetation [1]. Otherwise, dense vegetation will lead to the creation of abundant fuels, ready to spur and ignite at any moment. Prescribed fires are planned fires designed to tackle this issue. They aid in hazard fuel reduction, such as dry grasslands and bushes [11]. However, if not careful, these can turn into full-blown wildfires. Often, human settlements constructed close to forests fall in the path of natural wildfires.

It is key to understand the causes of ignition. Sparks are caused by dry fuels with low moisture content, hot weather, lightning, power line displacement due to strong winds and human factors. Although there is a correlation between vegetation and ignition, the water content in vegetation decreases the likelihood of ignition.

1.1.5 Composite Design

The composite design of this project is multi-layered and elaborated in detail in the subsequent sections. The initial analysis involves a study of the performance of conventional indices used in the current systems that mainly utilize fire history data. Thereafter, we will employ contemporary techniques such as Machine Learning and Neural Networks in an incremental manner. The research will involve traditional Machine Learning algorithms including Logistic, Random forest, Adaboost and Gradient boosting in the preliminary stages of the second stage.

Later, we will venture into sophisticated techniques in Deep Learning such as Long short-term Memory (LSTM) in Recurrent Neural Network. Separate models will be created for vegetation and Weather with terrain and powerline data. Eventually, it will be compared and averaged into a single model. Also, one more model will be created by combining Weather, Vegetation, terrain and powerline data. The final model will be validated using a receiver operating characteristic (ROC) curve, Confusion matrices, spatial and temporal accuracy, and computation speed.

1.2 Analysis of Requirements

The research aims to build a comprehensive wildfire prediction system that considers diverse parameters such as weather, terrain, powerlines, vegetation and fire history data. Using a test-driven development approach, we ran a series of experiments. The area of study has to be divided into square grids for a diligent data collection. Individual models have to be ensembled to cover deficiencies of the weak models. Further, all parameters have to be combined to build a complete modeling system. Subsets of data and experimental target labeling lead to a series of experiments. We included all the scenarios based on Cal Fire expert opinions, with the intent to cover all-natural environments of our study area. All parameters, train/test sets and machine learning models will be thoroughly optimized and evaluated to isolate the best model. All the results will be checked based on the below criteria.

1. Juxtaposing spatial and temporal accuracy against traditional methods.
2. Confusion matrix, Classification report and ROC curve with the best threshold value.
3. Learning curves that show the training and validation scores, scalability and performance of the model.

Figure 1.6 shows the workflow of our testing process. Test suites were created using the above experiments and cross-validation techniques such as gridsearchCV.

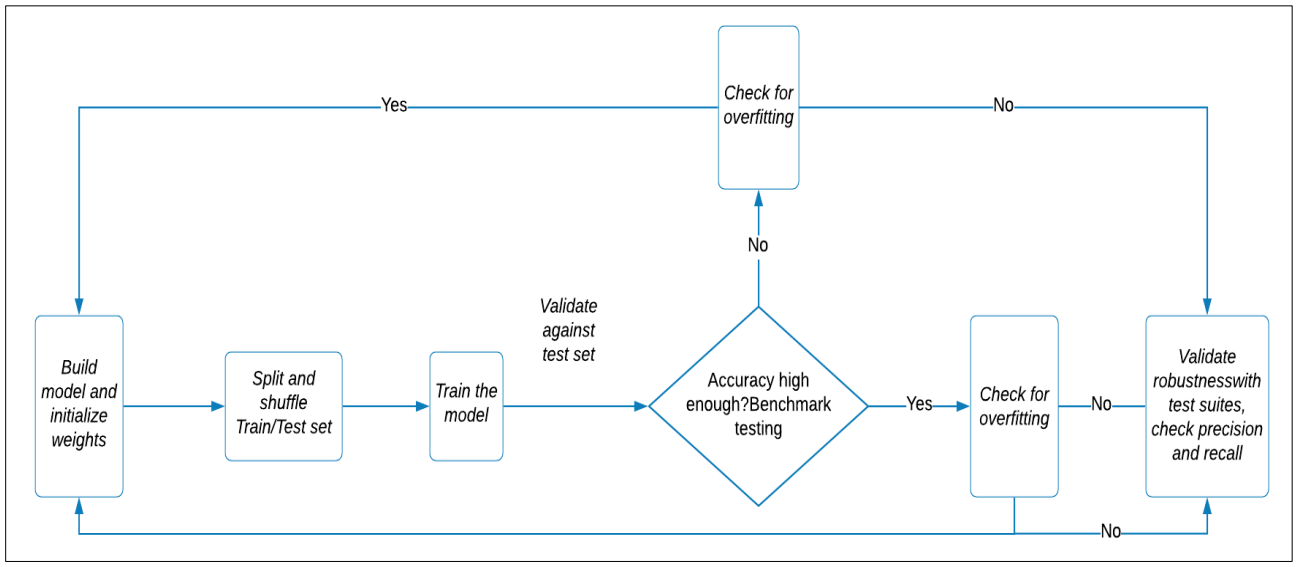


Figure 1.6. *Test-Driven Development*

Ultimately, we need to create an interactive user interface backed by the machine learning algorithm wherein the user can select the grid and learn the wildfire risk probability for the selected area. The user interface has to be interactive and should show the area of study, grids, parameters such as weather, vegetation and fire history along with statistical analysis.

1.3 Project Deliverables

For the wildfire prediction system with enhanced spatial sensitivity, the deliverables include backend model creation and frontend user interface creation. The best model after several iterations of experimentation is chosen. The user interface displays area-specific parameter analysis along with the fire risk probability for the study area as well as the selected grid. The data science lifecycle is provided in Table 1.1.

Table 1.1. *Data Science Life Cycle [12]*

Task	Proportion of Work	Subtasks
Problem Understanding	5%	Determine Objective
		Define Success Criteria
		Assess Constraints
Data Understanding	25%	Access Data Situation
		Obtain Data (Access)
		Explore Data
Data Preparation	30%	Filter Data
		Clean Data
		Feature Engineering
Modeling	20%	Select Model Approach
		Build Models
Evaluation of Results	5%	Select Model
		Validate Model
		Explain Model
Deployment	15%	Deploy Model

		Monitor and Maintain
		Terminate

These parameters are weather, vegetation, terrain and Human activity. Spartan Wildfire risk prediction system (SWiPS) will receive inputs from all the variables and display the probability of fire in the selected area. The resultant wildfire risk probability will be displayed in a layered interactive map. Other options include study area satellite and street view, California fire history, fire history analysis, integrated weather dashboard, vegetation dashboard and final model evaluation metrics.

1.4 Technology and Solution Survey

The fire prediction and detection systems in different countries across the globe were extensively researched. While some countries continue to practice traditional methods of fire detection such as cameras and sensors, authorities in countries, namely Australia and Canada, have adopted statistical algorithms to predict fire risk across the country. Newer technologies, such as satellite-based detection, have emerged in the past few decades. They have been adopted by governments such as the U.S and Japan. The tables below describe the national fire risk prediction methods used in major countries at a federal level. Table 1.2 studies the purpose, data acquisition, approach and methods used in each system. Table 1.3 details the parameters used in each of these systems.

Table 1.2. *Systems Survey*

System name		MFFDI	NFDRS	CFFDRS	FFRFS	NCMSSD
Country		Australia	US	Canada	Japan	Russia
Purpose	Prediction	Yes	Yes	Yes	Yes	No
	Detection	No	No	No	No	No
	Simulation	No	No	No	No	No
	Management	No	Yes	Yes	No	No
Data Acquisition	Satellite	No	Yes	No	Yes	Yes
	Sensor	Yes	Yes	Yes	Yes	No
	Manual	No	Yes	Yes	Yes	No
	Camera	No	No	No	No	No
Approach	Data-driven	Yes	Yes	Yes	Yes	Yes
	ML	No	No	No	Yes	No
	IR	No	No	No	No	No
Methods	Mathematical	Yes	Yes	Yes	No	Yes
	ANN	No	No	No	Yes	No

Note:
MFFDI - McArthur Forest Fire Danger Index (McArthur,1967)
USNFDRS - US National Fire Danger Rating System (Schlobohm et al. (2002))
CFFDRS - Canadian Forest Fire Danger Rating System (Stocks et al. (1989))

FFRFS - Forest Fire Risk Forecast System (Sawada, 2005)
NCMSSD- Nesterov + combination of multi spectral satellite data (Škvarenina et al. (2004))
SWiPS - Spartan Wildfire Risk prediction system
ANN - Artificial Neural Networks
IR - Image Recognition

Table 1.3. *Parameters Table*

System Name		MFFDI	NFDRS	CFFDRS	FFRFS	NCMSSD	SWiPS
Weather	Relative Humidity	Yes	Yes	Yes	Yes	No	Yes
	Temperature	Yes	Yes	Yes	Yes	Yes	Yes
	Precipitation	No	Yes	Yes	Yes	Yes	Yes
	Pressure	No	Yes	No	No	No	No
	Days Since Last Rainfall	No	Yes	No	Yes	Yes	No
	Wind Speed	Yes	Yes	Yes	Yes	No	Yes
	Dew-Point Temperature	No	Yes	No	No	Yes	No
	Elevation	No	Yes	Yes	No	No	No
	Slope, Aspect & Gradient	No	Yes	Yes	No	No	Yes
	Soil Moisture	Yes	Yes	Yes	Yes	No	No

Terrain, Land Cover & Fuel	Terrain Model	No	Yes	No	No	No	Yes
	Fuel Moisture	Yes	Yes	No	Yes	No	No
	Land Cover	No	Yes	Yes	No	No	Yes
	Vegetation Type	Yes	Yes	Yes	Yes	No	Yes
Fire History	Historic Fire Occurrence	Yes	Yes	Yes	Yes	No	Yes

These systems have the following issues:

- Algorithms such as Fire Weather Index (FWI) and Nesterov Index use numerical coefficients derived from past statistics which are not spatially sensitive.
- Most indexes focused on fewer parameters while wildfire is a complex combination of weather, land cover and topology.
- Current systems do not provide sufficient temporal accuracy because of limited data availability.

1.5 Literature Survey of Existing Research

Sophisticated techniques were introduced in wildfire risk prediction, utilizing the power of machine learning and neural networks. Some papers used Machine Learning to predict wildfire occurrences [13] uses Support Vector Machine to develop an algorithm for fire risk classification

over four classes based on the historical number of fires and certain weather conditions. [13] implements a Multilayer Perceptron approach based on a back-propagation algorithm, for mapping forest fire probability in the Upper Seyhan Basin area of Turkey. [14] builds predictive models to estimate the risk of fire outbreaks in Slovenia. It models both fire risk probabilities and fire detection. The single classifier methods include k-Nearest Neighbors, Naive Bayes, J48 Decision Trees, jRIP classification rules, Weighted Decision Trees, Support Vector Machine, Bayesian Networks, Tree Augmented Naive Bayes while the ensemble methods include Boosting, Bagging and Random Forest of decision trees [15] used Deep Learning algorithm based on data-driven unsupervised deep belief neural network along with conventional supervised ensemble machine learning to predict bush-fire hot spots for the continent of Australia. [15] compares several models using three machine learning algorithms: Random Forest, Boosted Regression Trees and Support Vector Machine, with the traditional method of Weighted Decision Trees [12] outlines a hybrid approach that combines meteorological data with Fire Weather Index. Table 1.4 shows the comparison of models in Wildfire risk prediction research and table 1.5 shows the comparison of Machine Learning Models in Wildfire Risk Prediction Research.

Table 1.4. *Comparison of Models in Wildfire Risk Prediction Research*

		Paper No.			
	Parameters	[13]	[14]	[16]	[15]
Fire History	Daily Number of Fire	Yes	No	No	No
	Fire History	No	Yes	Yes	Yes
Land	Elevation	No	Yes	No	No
	Land Usage	No	No	Yes	No

	Altitude	No	No	Yes	No
	Soil Moisture	No	No	Yes	Yes
Weather	Temperature	Yes	Yes	Yes	No
	Humidity	Yes	Yes	Yes	Yes
	Wind Speed	No	Yes	Yes	Yes
	Solar Radiation	Yes	No	Yes	Yes
	Precipitation	Yes	No	Yes	No
	Long-Term Climatology	No	Yes	No	No
	Transpiration & Evaporation	No	No	Yes	Yes
	Heat Flux	No	No	No	Yes
	Vapor Pressure	No	No	No	Yes
	Weather Forecast	No	No	Yes	No
Other	Human Activity	No	Yes	No	No
	Traffic Corridor	No	Yes	Yes	No
	Settlement Map	No	Yes	No	No

Table 1.5. *Comparison of ML Models used in Wildfire Risk Prediction studies*

Paper No.	Study area	Model	Accuracy
[13]	Lebanon	SVM	96% for binary

			classification
[14]	Upper Seyhan Basin, Turkey	MLP	83%
[16]	Slovenia	iBk	80.5 ± 1.1
		NB	81.0 ± 1.0
		J48	78.6 ± 1.2
		JRip	81.5 ± 1.2
		LogR	83.0 ± 0.8
		SVM	83.0 ± 0.7
		AdaBoost	83.3 ± 1.2
		BagJ48	84.9 ± 1.9
		RF	82.5 ± 1.2
		BNet	81.7 ± 0.9
[15]	Australia	kNN	91.76%
		Bagging Tree	94.53%
		ensemble method based on a two-layered machine learning model	91%

Note:
SVM - Support Vector Machine
MLP - Multilayer Perceptron
KNN - k-Nearest Neighbors
NB - Naive Bayes
J48 – J48 Decision Trees

LogR - Weighted Decision Trees
SVM - Support Vector Machine
BN - Bayesian Networks
Bnet - Tree Augmented Naive Bayes

Without considering local conditions and specialized scenarios derived from extensive analysis, Machine Learning models tend to underperform, especially when large and complex data is processed. Valuable information is lost and the fire risk prediction methodology is subpar. Hence, we propose a system that puts places constraints, based on expert opinions, on machine learning models and neural networks.

2. Data Exploration

2.1 Data Exploration Strategy and Planning

The basic strategy was to utilize the diverse datasets to create the below models. After evaluation, the best model was to be selected as the backed model for the user interface.

1) Ensemble Model
2) Combined Model

2.1.1 Ensemble Model

We trained separate models for each type of data and ensembled them using an innovative index-based ensemble method. Stacking is known to yield exceptionally high accuracy.

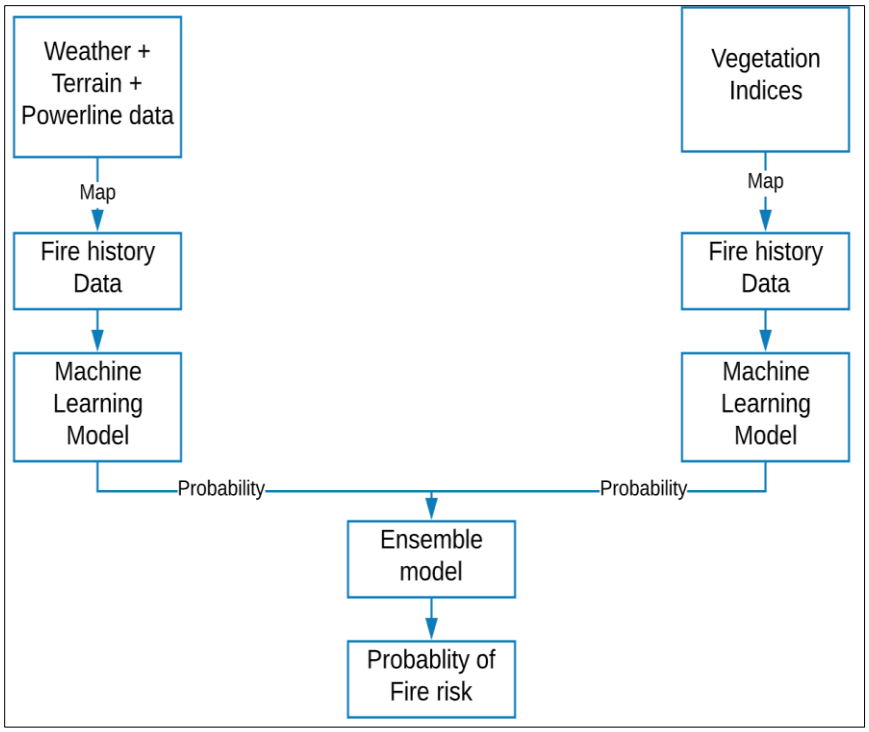


Figure 2.1. *Data Model I – Ensemble Strategy*

Our plan was to train a machine learning model and ensemble the same by stacking. Thereafter, it was fed to a classification algorithm of choice. Figure 2.1 shows the basic design of the ensemble strategy.

2.1.2 Combined Model

The second plan is to combine all the parameters, map it to the target fire history data before applying a suitable machine learning algorithm for predicting the fire risk probability for the given dates. Figure 2.2 shows the basic design for the combined model.

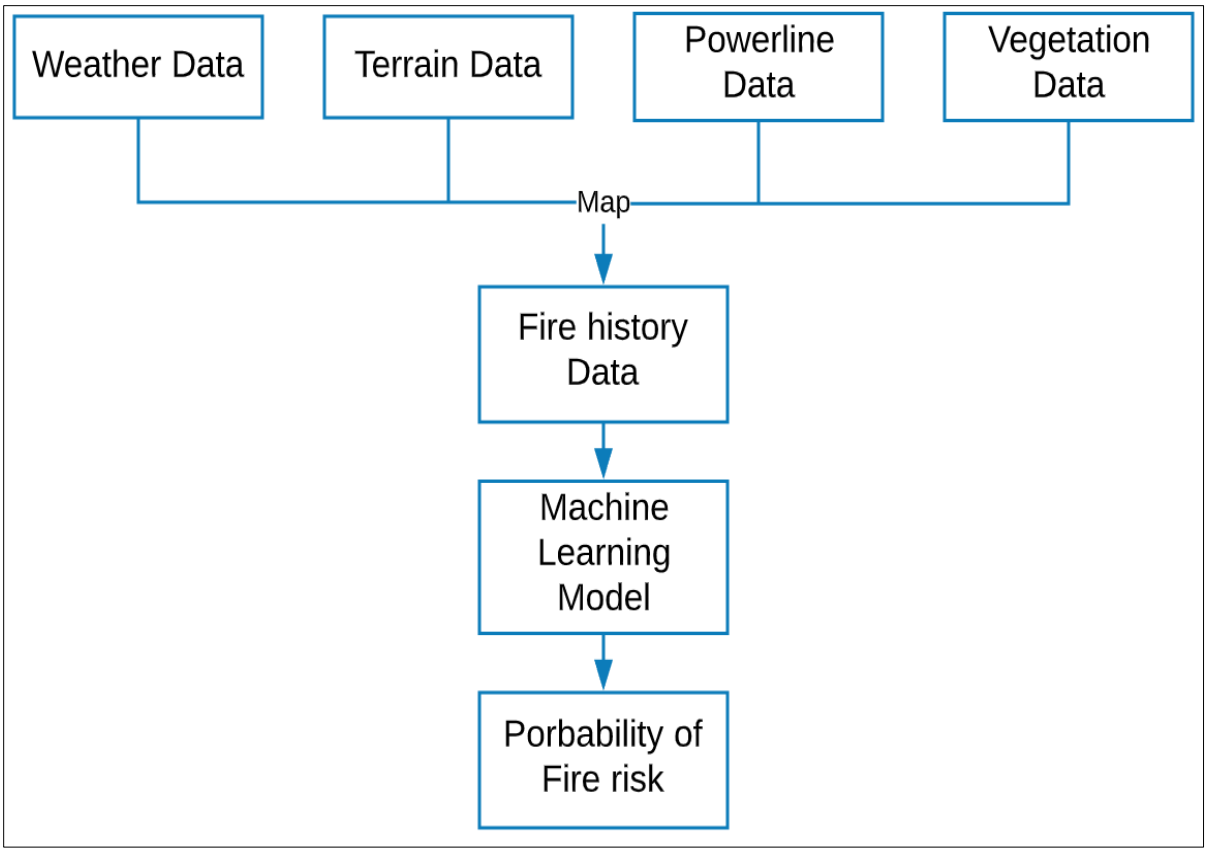


Figure 2.2. *Data Model II – Combined Strategy*

2.2 Data Sources and Dataset Parameters

Due to advancements in Big Data technologies, fire systems can be modeled using complex data such as satellite data. Below are the data sources considered in this research.

- Fire History Data.
- Weather Data.
- Vegetation Data.
- Terrain Data
- Powerline data.

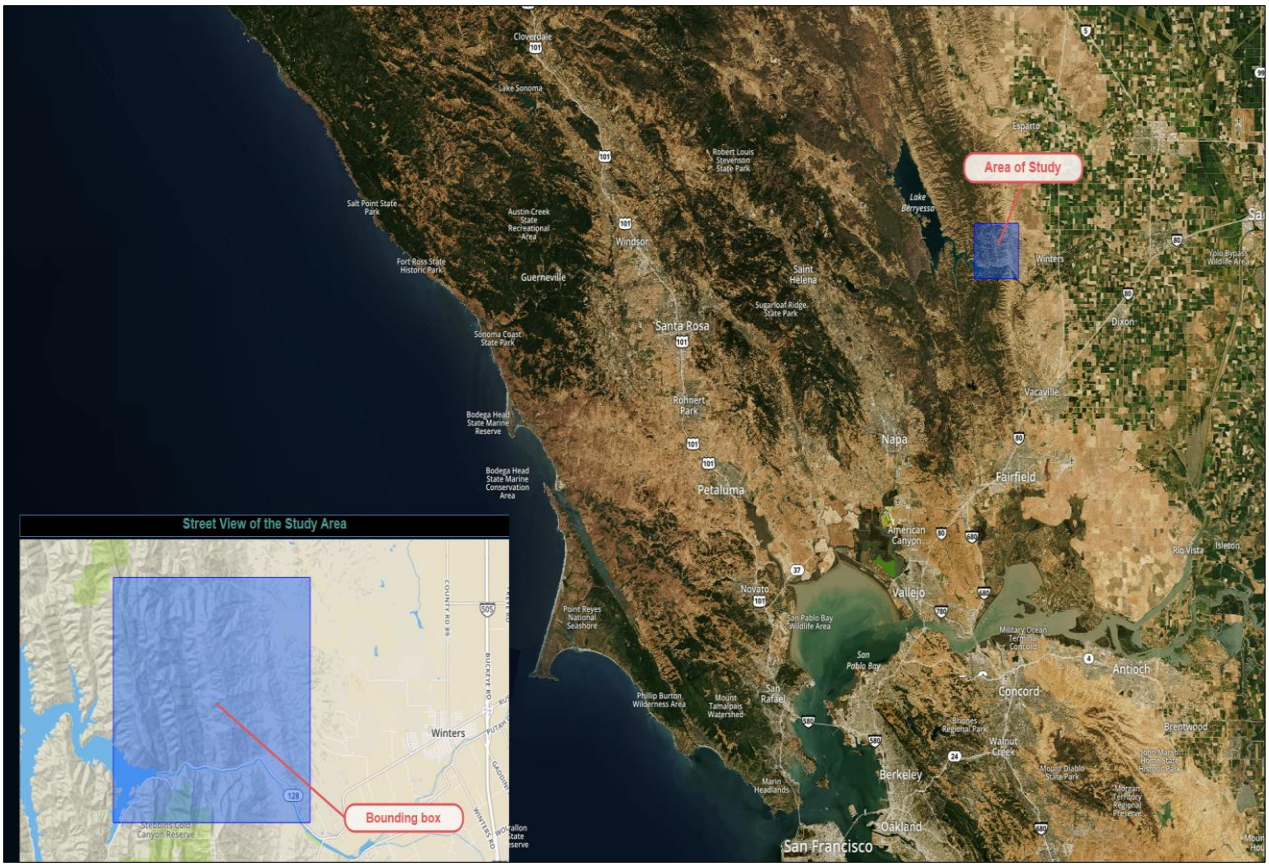


Figure 2.3. *Map of Study Area*

The date range for the study is 01/01/2015 to 12/31/2019. The study area near Monticello and Winters, California, is the bounded area in Figure 2.3. It is enveloped between Davis and Napa. Due to the intent to perform an area-based wildfire risk analysis, our study area was divided into 63 grids of dimension 1 x 1 km grids.

The total area covered is around 3969 kilometers with bounding shape longitude between -122.1241886091637 and 122.03418860916369 and the latitude between 38.49925978424475 and 38.56925978424475.

2.2.1 Fire History Data

Fire history data from Fire and Resource Assessment Program (FRAP) helps in identifying the location and relevant information about the fire incident [17]. CAL Fire, United States Forest

Service region, Bureau of Land Management and National park service developed a GIS layer providing the boundaries of historical fires in California.

Parameters of fire history data are explained in Table 2.1. The fire perimeters database provides ESRI ArcGIS files. The geodatabase has the following three data layers, also known as feature classes.

- A layer depicting wildfire perimeters from contributing agencies for the current fires and fires from previous fire years.
- A layer depicting prescribed fires supplied from contributing agencies for the current fires and fires from previous fire years. When there is a wildfire risk, excess vegetation is burned in a controlled environment by fire management agencies.
- A layer representing non-prescribed fire fuel reduction projects.

Table 2.1. *Parameters in Fire History Dataset*

Column Name	Data Type	Description	Range	Unit
Year	Object	Year of Fire occurred.	(2016, 2018)	year
State	Object	State where the fire occurred	'CA'	NA
Fire_Name	Object	Name of the fire	NA	NA
Alarm_Date	Datetime	Alarm date of the Fire.	(01/01/2016 ,12/31/2018)	NA
Cont_Date	Datetime	Containment date for fire.	(01/01/2016 ,12/31/2018)	NA
Cause	Integer	This column contains the Cause code which gives the reason for the fire.	(1,19)	NA
Report_Ac	Float	Area Consumed in the fire	(0, 25)	acre
GIS_Acres	Float	Area Calculated by GIS	(8.266294,	acre

			26.002495)	
C_Method	Float	Collection of data method coding	(1, 1)	NA
Objective	Float	Shows whether the fire is suppression or resource benefit	(1,1)	NA
Fire_Num	Object	Number assigned to the fire	(00001890 1716, 00000825)	NA
Shape_Length	Float	Length of the area burnt	(9.093210, 445282.4447 98)	meter
Shape_Area	Float	Area burnt	(6.130331,1. 660030e+09)	Square meter
Geometry	Geometry	Shape of the area burnt	Within study area	degrees

A sample of the data is shown in Figure 2.4. The dataframe is shown in Figure 2.5.

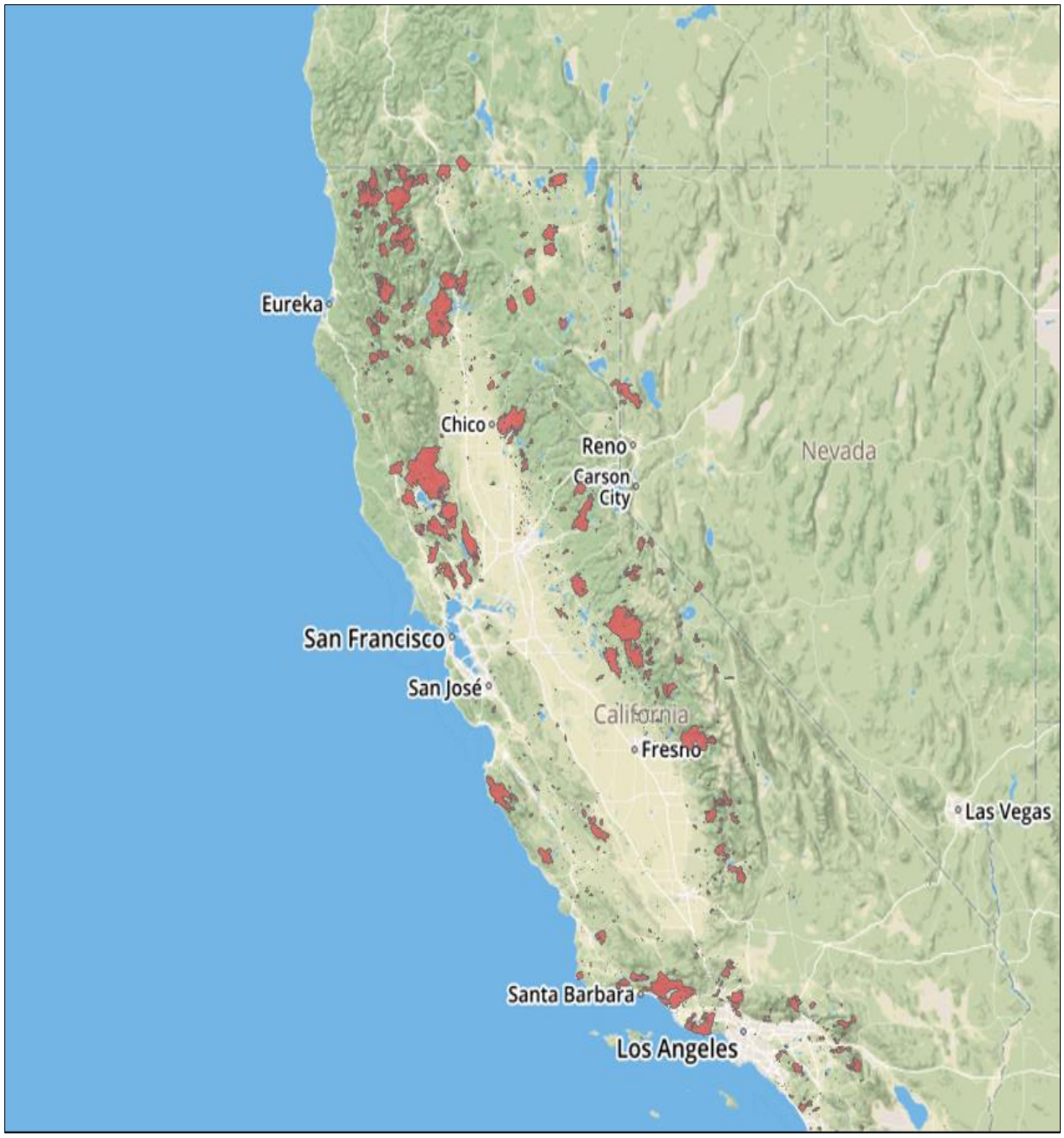


Figure 2.4. *Map of Fire History in California*

	YEAR_	FIRE_NAME	ALARM_DATE	CONT_DATE	CAUSE	REPORT_AC	GIS_ACRES	FIRE_NUM	geometry	Shape_Leng	Shape_Area
0	2007	OCTOBER	2007/10/21 00:00:00.000	2007/10/23 00:00:00.000	14.0	NaN	25.736713	00233414	POLYGON ((138036.906 -402646.363, 138086.986 -...	1902.439051	1.041528e+05
1	2007	MAGIC	2007/10/22 00:00:00.000	2007/10/25 00:00:00.000	14.0	NaN	2824.877197	00233077	POLYGON ((130072.487 -398622.842, 130094.237 -...	20407.965662	1.143187e+07
2	2007	RANCH	2007/10/20 00:00:00.000	2007/11/15 00:00:00.000	2.0	54716.0	58410.335938	00000166	POLYGON ((114013.974 -379231.746, 114190.835 -...	169150.715690	2.363782e+08
3	2007	EMMA	2007/09/11 00:00:00.000	2007/09/11 00:00:00.000	14.0	NaN	172.214951	00201384	POLYGON ((176902.236 -388673.082, 176907.996 -...	6117.777086	6.969292e+05
4	2007	CORRAL	2007/11/24 00:00:00.000	2007/11/27 00:00:00.000	14.0	NaN	4707.997070	00259483	POLYGON ((115905.006 -436381.137, 115926.897 -...	22907.182174	1.905259e+07

Figure 2.5. *Snippet of Fire History Data Frame*

2.2.2 Weather Data

Weather parameters from Local Climatology Data (LCD) are maintained by the National Centers for Environmental Information (NCEI). It provides atmospheric and geospatial data across the United States [18]. LCD summarizes hourly weather observations for 68 weather stations in Northern California. Each weather station is identified by a WBAN number. Dataset had county, latitude, longitude and elevation information from the respective weather stations. The relevant columns in the dataset are explained in Table 2.2.

Table 2.2. *Parameters in Weather Dataset*

Fields	Data Type	Description	Range	Unit
WBAN	Integer	WBAN is a five-digit number used for digital data storage and weather station identification.	(117, 94299)	NA
Latitude	Float	Latitude of the weather station	(32.81667, 41.78139)	Degree

Longitude	Float	Longitude of the weather station	(-115.57861, -124.23667)	Degree
Elevation	Float	Elevation of the weather station	(-17.7, 2172.6)	Meter
Date	Datetime	Time at which different weather parameters are captured.	(01/01/2016, 12/31/2018)	NA
Hourly DryBulb Temperature	Float	Dry Bulb Temperature of air at the weather station.	(-22.0, 115.0)	Fahrenheit
Hourly Relative Humidity	Float	Relative humidity at the weather station.	(1.0, 100.0)	Percentage
Hourly WindSpeed	Float	Wind speed at the station.	(0.0, 57.0)	Miles per Hour (mph)
Hourly Precipitation	Float	Precipitation at the station.	(0.0, 10.31)	Inches to Hundredths

A sample of the python dataframe is shown in Figure 2.6. Hourly dry bulb temperature, relative humidity, wind speed and precipitation are the weather parameters.

	STATION	DATE	HourlyDryBulbTemperature	HourlyRelativeHumidity	HourlyWindSpeed	HourlyPrecipitation	Year	day
0	72483793216	2015-01-01 00:58:00	33	75	6.0	NaN	2015	2015-01-01
1	72483793216	2015-01-01 01:58:00	33	74	5.0	NaN	2015	2015-01-01
2	72483793216	2015-01-01 02:58:00	33	76	8.0	NaN	2015	2015-01-01
3	72483793216	2015-01-01 03:58:00	32	76	9.0	NaN	2015	2015-01-01
4	72483793216	2015-01-01 04:58:00	31	79	8.0	NaN	2015	2015-01-01

Figure 2.6. *Snippet of the Weather Dataframe*

2.2.3 Vegetation Data

After careful contemplation and research, we proceeded with data extraction from Landsat 8 satellite using Google Earth Engine (GEE) for our study. Initially, we considered Landsat 8, Sentinel, Proba-V and Terra/Aqua satellites, as well as sources such as Harris Geospatial and National Agriculture Imagery Program (NAIP) for vegetation data. Further, we processed the data, compared the resolution and checked data availability from sources, namely Landsat 8, Proba-V and Terra/Aqua satellite-based MODIS (Moderate Resolution Imaging Spectroradiometer) sensor data. The reasons that favored Landsat as the final data source are as follows.

- Landsat 8 data has an impeccable resolution of 30m, although the temporal frequency is not consistent as the satellite is not systematic. However, we have 8-day composites and once in 16 days sensor data available in Google Earth Engine (GEE).
- Sentinel is a European satellite well-suited when the area of study is in Europe.
- Proba-V data was hard to process due to the data structure and format of data. The resolution was similar to Landsat data.
- Harris Geospatial is a commercial source with exceptionally high spatial resolution data of 0.7m and 1m. For educational purposes, the discounted price for the data was above $100,000. Due to a lack of funding, we decided to forgo this resource.
- National Agriculture Imagery Program (NAIP) [63] is an annual aerial imagery program that captures images for estimating the land cover and agricultural plantation and yield. Although the resolution is high at 1-2 m, temporal resolution is poor because it is acquired once in a year and unsuitable for our wildfire study.
- Terra/Aqua satellite-based MODIS (Moderate Resolution Imaging Spectroradiometer) sensor data was reliable and consistent but with moderate resolution as the name suggests.

The best spatial resolution is 250m, which was far inferior compared to the 30m resolution promised by Landsat data.

Landsat 8, launched jointly by the United States Geological Survey (USGS) and the National Aeronautics and Space Administration (NASA), offered better clarity than its predecessors. It is equipped with an Operational Land Imager (OLI) and Thermal Infrared Sensor (TIRS). The spatial resolution is remarkable at 30 meters (visible, NIR, SWIR), 100 meters (thermal) and 15 meters (panchromatic). The 100 m TIRS data is registered to the OLI data to create radiometrically, geometrically, and terrain-corrected 12-bit data products. It captures around 400-725 scenes each day and returns it to the USGS data archive. Although it achieves global coverage once in 16 days, Landsat 7 and Landsat 8 together provide ample coverage of images once in 8 days for a region under study.

The spacecraft is positioned at a 705km altitude [43]. Satellite images of Camp fire are shown in Figure 2.7, one of the deadliest fires in the history of the State of California, which claimed 85 lives and destroyed 18,793 structures in the year 2018 [19].

Although Landsat 8 provides once in 16 days raw images as well as Tier 1 data, 8-day composites were found to be more reliable and accurate, due to lesser cloud cover, aerosol, atmospheric conditions and temporary fluctuations. From Landsat 8, we initially downloaded bands and calculated indices. The data was highly inconsistent as Landsat8 is not systematic. Further, the frequency was once in 16-days. Hence, we chose the 8-day composite data product from the Google Earth engine (GEE) and extracted the indices Normalized Difference Vegetation Index (NDVI), Enhanced Vegetation Index (EVI) and Normalized Difference Water Index (NDWI). These indices quantify greenness and classify vegetation types.

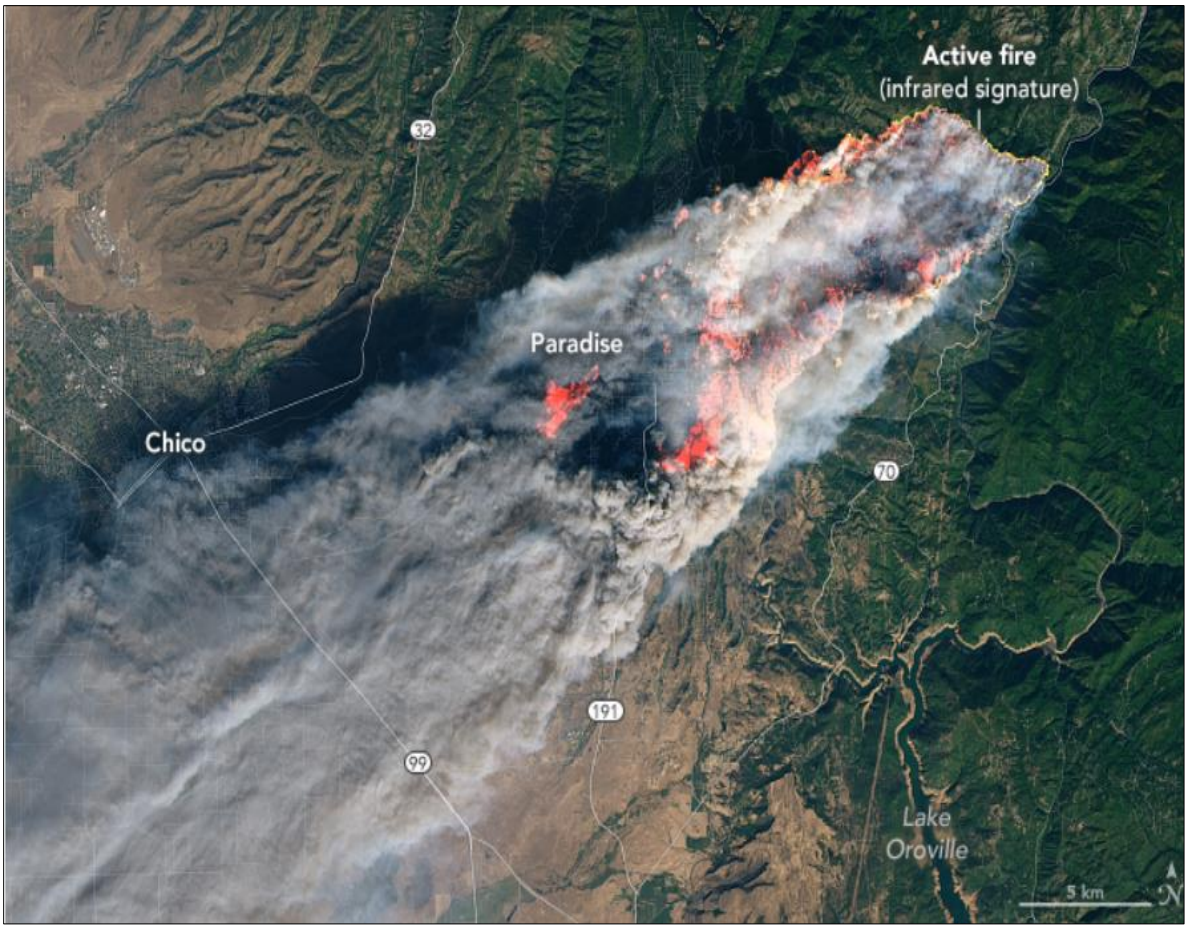


Figure 2.7. *Landsat image of Campfire wildfire [20]*

From Table 2.3 it is evident that Landsat 8 has a good spatial and temporal resolution. Although NDVI [63][45][47][24] is the standard vegetation index with simple calculation and reliable classification, it is highly sensitive to chlorophyll content. EVI is an enhanced version of NDVI with increased sensitivity to biomass and canopy type as it has lesser distortion due to atmospheric conditions such as cloud cover and background noise such as soil reflection. NDWI aids in water content analysis, including the leaf water content. Table 2.4 depicts the parameters in the vegetation dataset.

Table 2.3. *Landsat 8 [21]*

Satellite	Landsat 8
Launch date	February 2013
Sensor	Operational Land Imager (OLI) and the Thermal Infrared Sensor (TIRS)
Spatial resolution	30 m (multispectral resolution) 15 m (panchromatic resolution)
Temporal resolution	16 days

Table 2.4. *Parameters in the Vegetation Dataset*

Fields	Data Type	Description	Range	Unit
Left	Float	Leftmost coordinate of the inner grid	NA	Degree
Right	Float	Rightmost coordinate of the inner grid	NA	Degree
Top	Float	Topmost coordinate of the inner grid	NA	Degree
Bottom	Float	Bottom-most coordinate of the inner grid	NA	Degree
id	Integer	Unique id for the inner grid	1-63	NA
Centroid Latitude	Float	Latitude of the inner grid centroid	(38.4992, 38.5692)	Degree
Centroid Longitude	Float	Longitude of the inner grid centroid	(-122.1241, -122.0341)	Degree
Polygon/Geometry	Polygon	Grid coordinates of the inner grid	((38.4992, -122.1241), (38.5692, -122.0341))	Degree

start_date	Datetime	Start date for fetching the 8-day composite indices	(01/01/2015, 12/27/2019)	NA
end_date	Datetime	7 days added to the start date provides the end date for 8-day composite indices	(01/08/2015, 1/3/2020)	NA
NDVI	Float	Calculated from red and infrared bands.	-1 to +1	NA
EVI	Float	Calculated from visible and near-infrared bands.	-1 to +1	NA
NDWI	Float	Calculated from short-wave infrared and near infrared bands.	-1 to +1	NA
topLeft_coords, topRight_coords, bottomLeft_coords, bottomRight_coords, centroid_coords, midLeft_coords, midRight_coords, midTop_coords, midBottom_coords, diagonal, diagonal2, diagonal3, diagonal4	Float	13 columns corresponding to the 13 points sampled from each inner grid. The points are Centroid, 4 diagonal points, 4 midpoints of corners and corner points. The values are the latitude and longitude coordinates.	((38.4992, -122.1241), (38.5692, -122.0341))	Degrees

The final snippets of the vegetation data with relevant parameters such as NDVI, EVI and NDWI, are shown below in Figure 2.8.

	left	top	right	bottom	id	geometry	Centroid Longitude	Centroid Latitude	Start Date	End Date	NDVI	EVI	NDWI
0	-122.124189	38.56926	-122.114189	38.55926	1	POLYGON ((-122.12419 38.56926, -122.11419 38.5...	-122.119189	38.56426	2014-01-01	2014-01-08	0.416393	0.340703	0.094683
1	-122.124189	38.56926	-122.114189	38.55926	1	POLYGON ((-122.12419 38.56926, -122.11419 38.5...	-122.119189	38.56426	2014-01-09	2014-01-16	0.416393	0.340703	0.094683
2	-122.124189	38.56926	-122.114189	38.55926	1	POLYGON ((-122.12419 38.56926, -122.11419 38.5...	-122.119189	38.56426	2014-01-17	01-24 2014-	0.337617	0.310645	0.110320

topLeft_coords	topRight_coords	bottomLeft_coords	bottomRight_coords	centroid_coords	midLeft_coords	midRight_coords
(-122.1241886091637, 38.56925978424475)	(-122.11418860916369, 38.56925978424475)	(-122.1241886091637, 38.55925978424475)	(-122.11418860916369, 38.55925978424475)	(-122.11918860916371, 38.56425978424475)	(-122.1241886091637, 38.56425978424475)	(-122.11418860916369, 38.56425978424475)
(-122.1241886091637, 38.56925978424475)	(-122.11418860916369, 38.56925978424475)	(-122.1241886091637, 38.55925978424475)	(-122.11418860916369, 38.55925978424475)	(-122.11918860916371, 38.56425978424475)	(-122.1241886091637, 38.56425978424475)	(-122.11418860916369, 38.56425978424475)
(-122.1241886091637, 38.56925978424475)	(-122.11418860916369, 38.56925978424475)	(-122.1241886091637, 38.55925978424475)	(-122.11418860916369, 38.55925978424475)	(-122.11918860916371, 38.56425978424475)	(-122.1241886091637, 38.56425978424475)	(-122.11418860916369, 38.56425978424475)

midTop_coords	midBottom_coords	diagonal1	diagonal2	diagonal3	diagonal4
(-122.11918860916371, 38.56925978424475)	(-122.11918860916371, 38.55925978424475)	(-122.12143860916372, 38.566509784244744)	(-122.11693860916371, 38.566509784244744)	(-122.11693860916371, 38.56200978424475)	(-122.12143860916372, 38.56200978424475)
(-122.11918860916371, 38.56925978424475)	(-122.11918860916371, 38.55925978424475)	(-122.12143860916372, 38.566509784244744)	(-122.11693860916371, 38.566509784244744)	(-122.11693860916371, 38.56200978424475)	(-122.12143860916372, 38.56200978424475)
(-122.11918860916371, 38.56925978424475)	(-122.11918860916371, 38.55925978424475)	(-122.12143860916372, 38.566509784244744)	(-122.11693860916371, 38.566509784244744)	(-122.11693860916371, 38.56200978424475)	(-122.12143860916372, 38.56200978424475)

Figure 2.8. *The dataframe with Vegetation Data*

2.2.4 Powerline Data

The California Energy Commission created an Electric transmission line geospatial data layer to display the electric transmission grids in California. Using these layers, we can analyze the geographic relationships between the electric transmission grids across utilities, counties and states. Table 2.5 shows the parameters extracted from the data layer.

Table 2.5. *Parameters in Powerline Dataset*

Parameter	Description
kV	Voltage of the powerline
Owner	Owner of the powerline
Status	Whether the line is operating or not
Circuit	Type of circuit
Length (Mile)	Length of the line in miles
Length (Feet)	Length of the line in Feet

2.2.5 Terrain Data

The United States Geological Survey (USGS) provides topography information, which is available in the form of Digital Elevation Models (DEM). DEM are classified into project-based and seamless. The project-based DEMs are available for the full area when produced from light detection and ranging (lidar), or as one-degree blocks with overedge. Seamless DEMs are produced by blending only the highest quality project data into a continuous terrain surface for the U.S. These data are distributed in tiles that can be merged to support analysis across large geographic areas. 1/3 arc-second, which is the highest resolution seamless DEM dataset for the U.S. with full coverage of the 48 conterminous states was used for our project. Ground spacing is

approximately 10 meters north/south, but variable east/west due to the convergence of meridians with latitude. From the DEM, parameters given in Table 2.6 can be computed [22].

Table 2.6. *Parameters in Terrain Dataset*

Parameter	**Description**	**Range**	**Unit**
Slope	Slope of the terrain	(0, 90)	degree
Local upslope	Upslope of the terrain	(0, 52,012)	Square kilometers
Local aspect	It is a compass direction that a slope faces	(0, 360)	azimuth
Hillshade	Terrain surface with sun relative position.	(0,360)	azimuth

2.3 Collection of Training Datasets

2.3.1 Fire History Data

Even though layers in ArcGIS files contain mostly digital records of wildfire, it has some incomplete data. Fire perimeter data are improved constantly by the Fire and Resource Assessment Program (FRAP). The data capture process tries to identify duplicates by considering data from the United States Forest Service and Cal Fire. The updated fire perimeters are released in April every year. Hence, they update the data layer annually in different iterations due to the below reasons.

- Standardizing and combining existing fire perimeters.
- Eliminating repeated fires.
- Completing the gaps in data.
- Collecting and standardizing previous agency fire perimeters and adding them to the GIS layer.

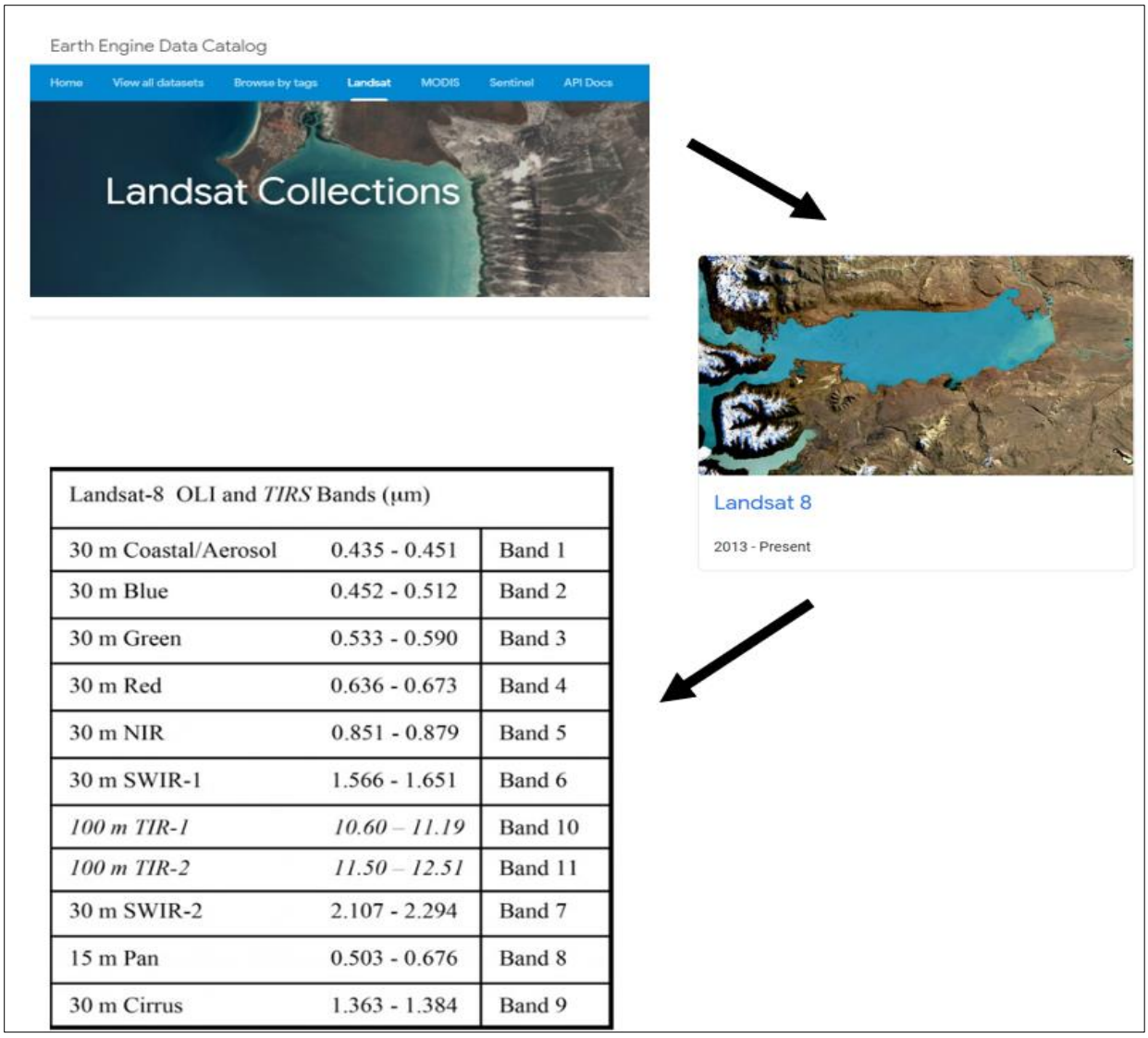
2.3.2 Weather Data

Weather stations collect data using sensors which include hygrometers and barometers, among others. Climate Data Online (CDO) allows you to access data with several tools. Local Climatology Data (LCD) is one of the tools to access historical climate data. Weather data can be obtained by selecting state, county and date range in this tool. Observations from the weather stations include parameters such as temperature, dew point, relative humidity, precipitation, wind direction and speed. In addition to these observations, they provide services that include data collection, quality control, removal of biases by considering factors like urbanization.

Dataset is divided into the train and test set. Training data is used by the model to learn and the resultant model is evaluated against the test set. Training data is fitted with our target data by geographic proximity and time.

2.3.3 Vegetation Data

Google earth engine (GEE) has a cloud repository of satellite data, free for non-commercial use. We fetched the Landsat 8 data from the data catalog using a package named GEE extract. After exploring several methods to fetch the image collection, we finalized this method, which is deemed ideal for time series data [43].



Landsat-8 OLI and *TIRS* Bands (μm)		
30 m Coastal/Aerosol	0.435 - 0.451	Band 1
30 m Blue	0.452 - 0.512	Band 2
30 m Green	0.533 - 0.590	Band 3
30 m Red	0.636 - 0.673	Band 4
30 m NIR	0.851 - 0.879	Band 5
30 m SWIR-1	1.566 - 1.651	Band 6
100 m TIR-1	*10.60 – 11.19*	Band 10
100 m TIR-2	*11.50 – 12.51*	Band 11
30 m SWIR-2	2.107 - 2.294	Band 7
15 m Pan	0.503 - 0.676	Band 8
30 m Cirrus	1.363 - 1.384	Band 9

Figure 2.9. *Data collection from Google Earth Engine (GEE) [46]*

Two custom functions were written for fetching data from GEE using python language in Jupyter notebook. The initial function which fetches the image collection for a single sensor within the time frame (2014-2019) was utilized in the subsequent function to fetch NDVI, EVI and NDWI vegetation indices from 3 separate data products GEE and process the data into a pandas Data frame, through a series of complex steps involving concatenation and formatting. Using this method, we generated the one-point dataset with only centroid, the five-point dataset with centroid and corners and the 13-points dataset with centroid, 4 corners, 4 midpoints of edges and

4 diagonals. However, the downloaded 8-day composite data had troubling periodic missing values, although it is a known issues in free publicly available remote sensing satellite data. Winter data goes missing due to outages. Figure 2.9. shows the data collection from GEE.

An elaborate and exhaustive data cleansing process was followed by visualizations to validate our outcome as explained in section 2.4.3.

2.3.3 Terrain Data

Data collected from the United States Geological Survey (USGS) is in the form of a DEM map. After exploring various ways of processing DEM maps, we found that QGIS is one of the best tools to analyze spatial statistics. After merging the DEM map with fire history data and our bounding box grid, using the inbuilt zonal statistics algorithm in QGIS, we calculated the statistics for each feature (slope, hill shade, aspect) in each grid. Figure 2.10 shows the raster analysis parameters calculated from the DEM map.

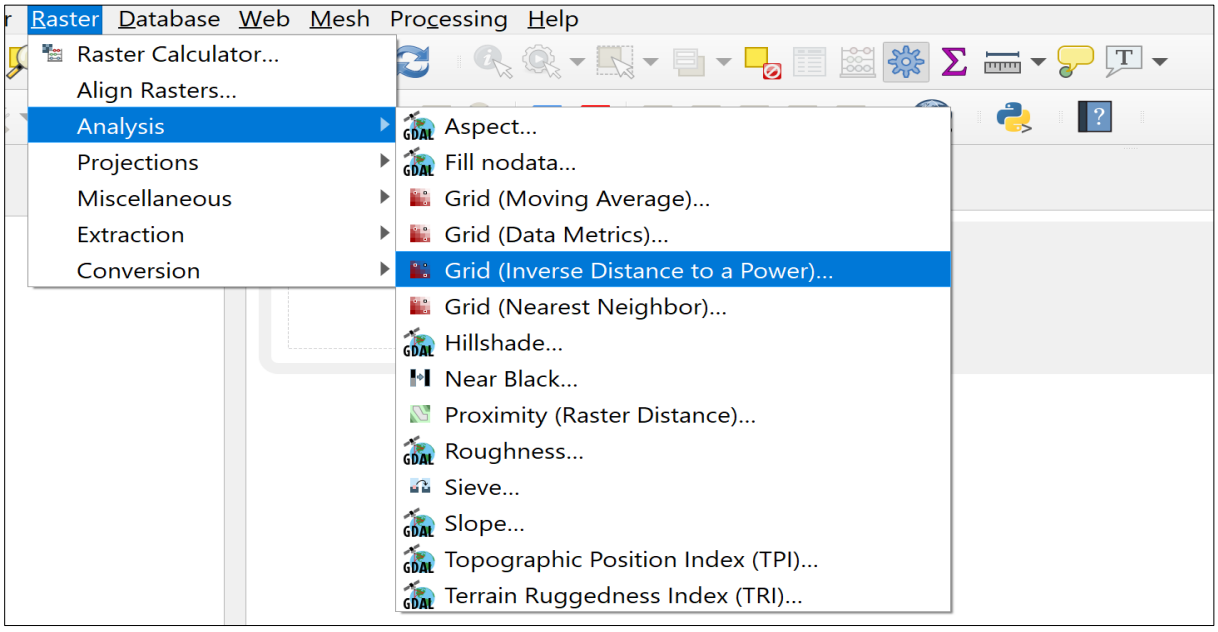


Figure 2.10. *Feature extraction process for Terrain data*

2.3.4 Powerline Data

Powerline data collected from the California Energy Commission is available in shapefile format. We merged the spatial file with fire history and bounding box grid using the tools in QGIS software, under the data management category. The resultant data can be exported as a flat file. Figure 2.11 shows the data management tools that are available in the QGIS application, to merge spatial vector layers.

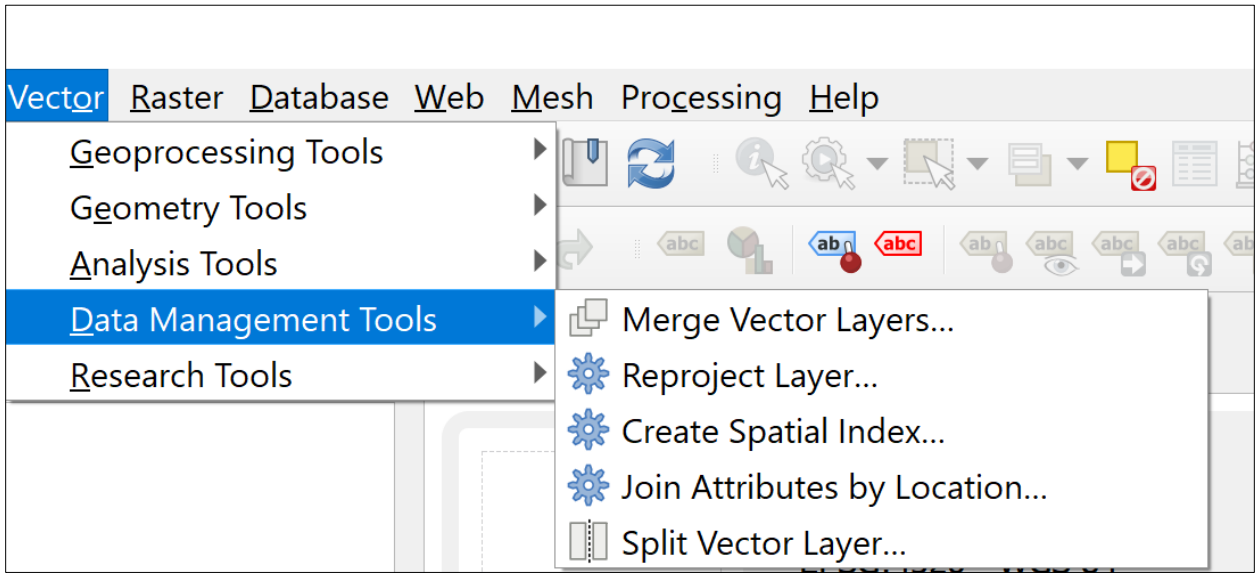


Figure 2.11. *Feature extraction process for Powerline data*

2.4 Data Cleansing and Validation

2.4.1 Fire History

The data source is pre-cleaned and validated by Fire and Resource Assessment Program (FRAP). For our machine learning purpose, we will divide the geographic locations of historical fires into 1 x 1 km grids and denote the presence of fire with binary variables. Figure 2.12 shows the 63 grids (9 columns and 7 rows) and Figure 2.13 shows fire history data in our study area.

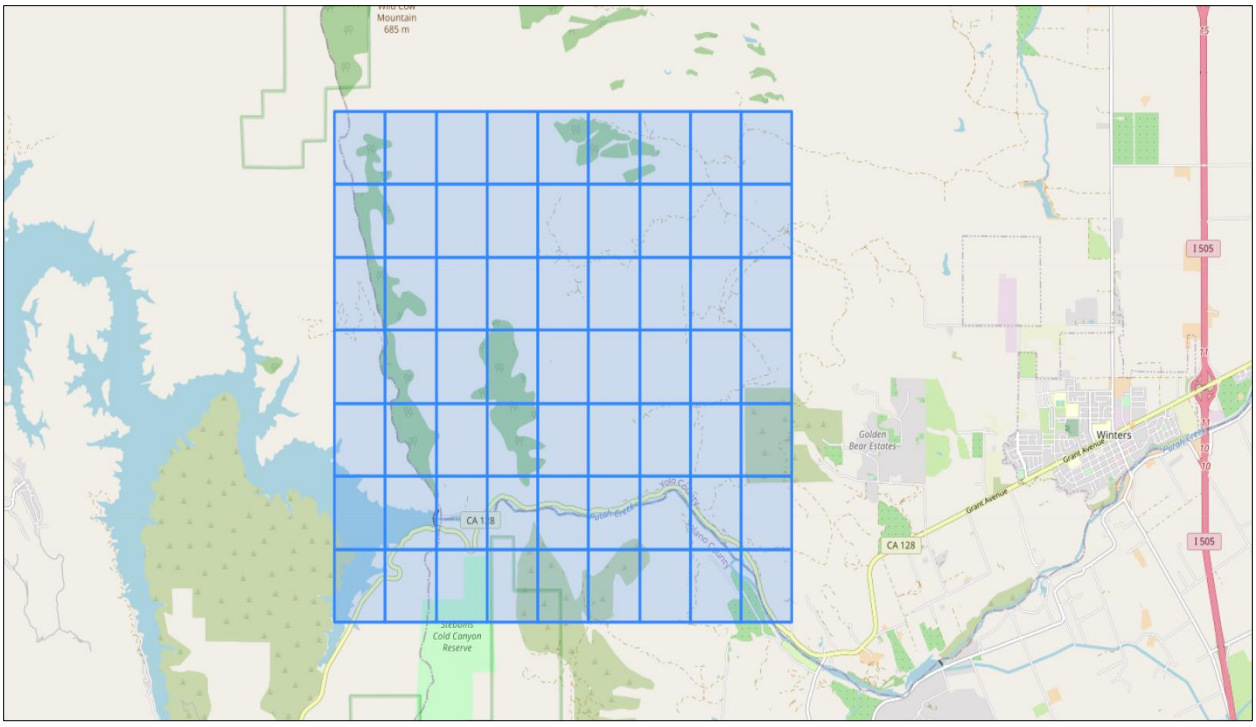


Figure 2.12. *Grids in the Study Area*

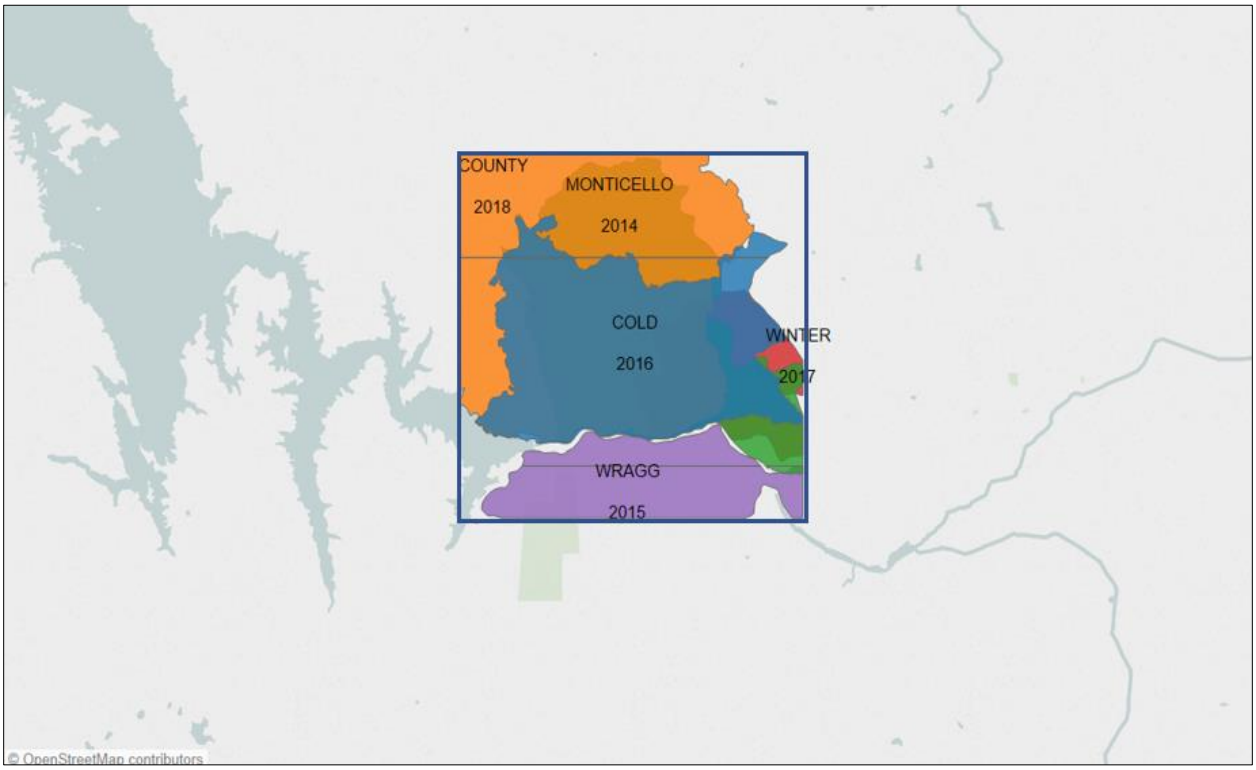


Figure 2.13. *Fire history data exclusively in the Study Area*

2.4.2 Weather

Preprocessing is required for two major issues. It is the essential step that refines the dataset and converts it to a workable format.

- Data losses due to hardware malfunctioning or transmission interruption
- Abnormal values caused by human recording.

For the above data, we will first filter outliers and abnormalities. Then, we will impute the null values based using the mean of the hourly measurements before and after the missing value at first. If the data is not available, we will search for 5 days ahead in the subsequent week and replace the missing values with measurements of the same hour. If all those methods fail, we will fill in the missing data from records of previous years. We used null value imputation to fill in the missing data with similar values as well.

To validate our preprocessed data, we compared the cleaned data with the raw data to check if they follow the same distribution.

2.4.3 Vegetation

For vegetation data, we created a custom function integrating all activities required to clean the data. In the initial data collection step, a data frame with NDVI, EVI and NDWI indices for each of the points was created. However, for a single inner grid, 13 points were sampled in the final dataset although we validated the results with a 1-point centroid only dataset and 5 point centroid and corners dataset. 13 points dataset comprises of the centroid, 4 corners, 4 diagonals and 4 midpoints of edges.

For data cleaning, we grouped points based on point id, start and end date and calculated the mean value such that a single value/row remains for a single inner grid at a given time. At this point, the latitude and longitude columns were dropped as the mean value does not give any useful information. Instead, we integrated the original grid table with geometry, left, right, top, bottom coordinates. Additionally, the table was formerly updated with centroid latitude and longitude, and co-ordinates of the 13 points under study. Thereafter, the missing values were imputed.

As advocated by experts [24] we considered using 'time' and 'linear' interpolation deemed most suitable for the time-series null value imputation, reliable as per our reference[24]. Further, we considered 'rolling mean' but it gives rise to low R2 square.

Although 'time' method captures the seasonality of the data and compares former and later years for data interpolation, it is not suitable for fire-prone areas such as our area of study with huge variability in data.

Seasonality may not be ideal as frequent fires disrupt the seasonal cycles. Hence, we are considering Linear interpolation which fits a line between available points and imputes the value of the points in between. The final dataset had 7749 rows and 26 columns. Column details are provided in Table. 2.7. As shown in subsequent Figure 2.14, a thorough statistical study and visualizations are the tools used to validate the dataset.

	left	top	right	bottom	id	Centroid Longitude	Centroid Latitude	NDVI	EVI	NDWI
count	17388.000000	17388.000000	17388.000000	17388.000000	17388.000000	17388.000000	17388.000000	17388.000000	17388.000000	17388.000000
mean	-122.084189	38.539260	-122.074189	38.529260	32.000000	-122.079189	38.534260	0.303082	0.303386	0.106917
std	0.025821	0.020001	0.025821	0.020001	18.184765	0.025821	0.020001	0.137473	0.107192	0.127875
min	-122.124189	38.509260	-122.114189	38.499260	1.000000	-122.119189	38.504260	-0.200841	-0.013835	-0.264864
25%	-122.104189	38.519260	-122.094189	38.509260	16.000000	-122.099189	38.514260	0.212849	0.231374	0.021941
50%	-122.084189	38.539260	-122.074189	38.529260	32.000000	-122.079189	38.534260	0.297970	0.283860	0.091828
75%	-122.064189	38.559260	-122.054189	38.549260	48.000000	-122.059189	38.554260	0.391561	0.367870	0.188576
max	-122.044189	38.569260	-122.034189	38.559260	63.000000	-122.039189	38.564260	0.742485	0.767890	0.600332

Figure 2.14. *Statistical study of 13 points dataset*

Figure 2.15 is a plot of the distribution of datasets wherein vegetation indices NDVI, EVI and NDWI are mapped to their corresponding counts to study the data. All the datasets are forming a normal distribution.

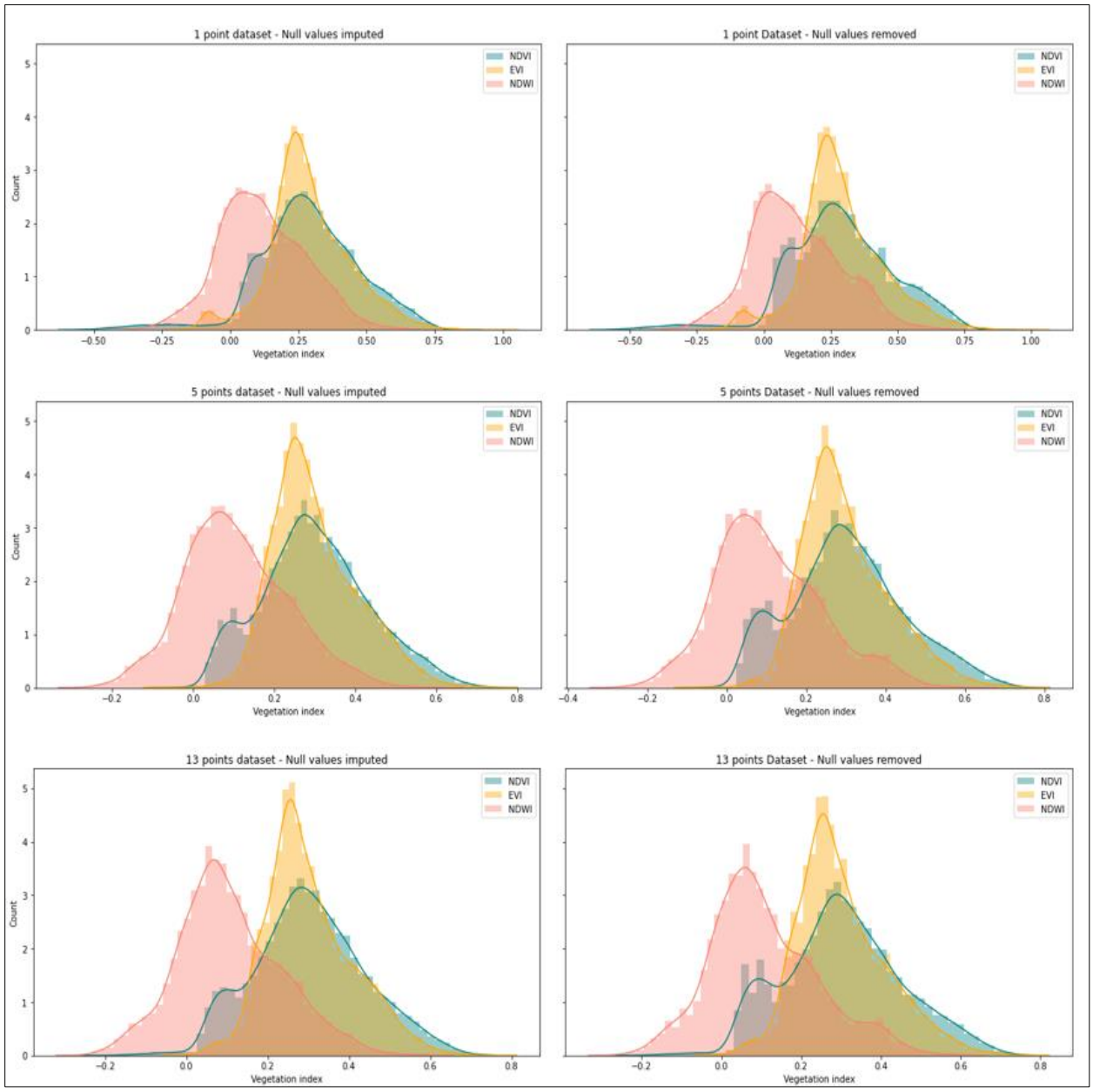


Figure 2.15. *Null value imputation vs. Null value removal*

When null values are removed as shown in column 2, the rows with no indices are simply dropped created gaps in data. However, it is evident that null value imputation, in column 1, leads

to a smoother data distribution without making any drastic changes to the distribution. This informs us that the imputation or adding more sampling points does not alter the basic behavior of the dataset or introduce newer patterns or noise.

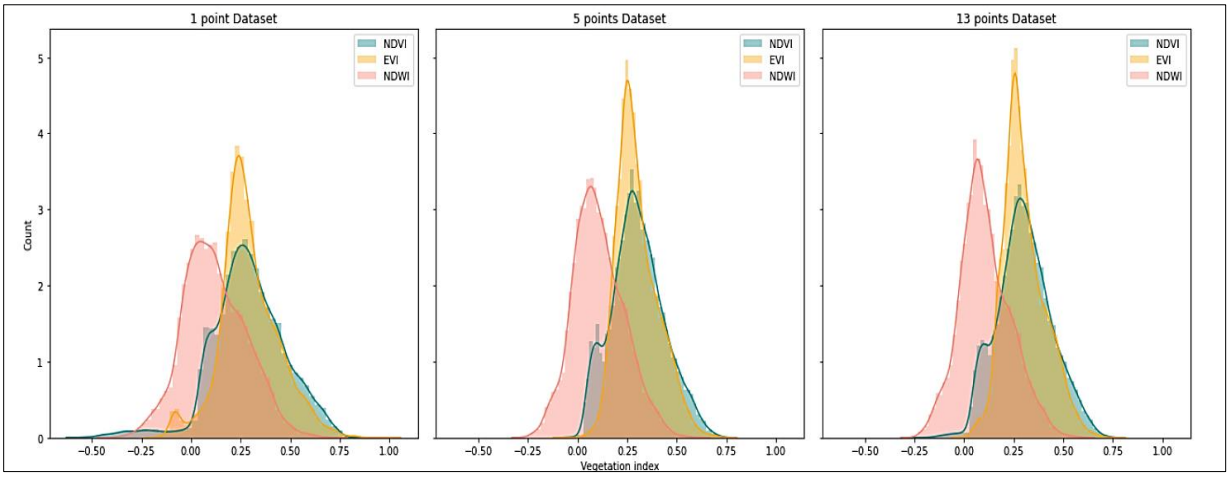


Figure 2.16. *Comparison of 1- point, 5- points and 13- points sampling vegetation datasets*

The purpose of this study is to validate the data and confirm proper data sampling. Upon comparing 1-point dataset with only the centroid of each inner grid, 5 point dataset with centroid and corners and 13 points dataset with centroid, 4 corners, 4 midpoints of edges and 4 diagonals in Figure 2.16 and 2.17, it is evident that 1 point dataset does not capture all the values correctly. The count near the mean value is lesser with greater standard deviation and variance. The count progressively improves as we sample more points. However, the difference between 13 points and 5 points datasets is negligible. In fact, NDVI value distributions for 5 points and 13 points are identical. Hence, we have decided to forgo sampling more points as we may have an accurate representation of the vegetation in the area with the 13 points dataset. 13 points dataset has a higher peak near the mean value, indicating more indices with the mean value and seems to be the most suitable dataset for our study.

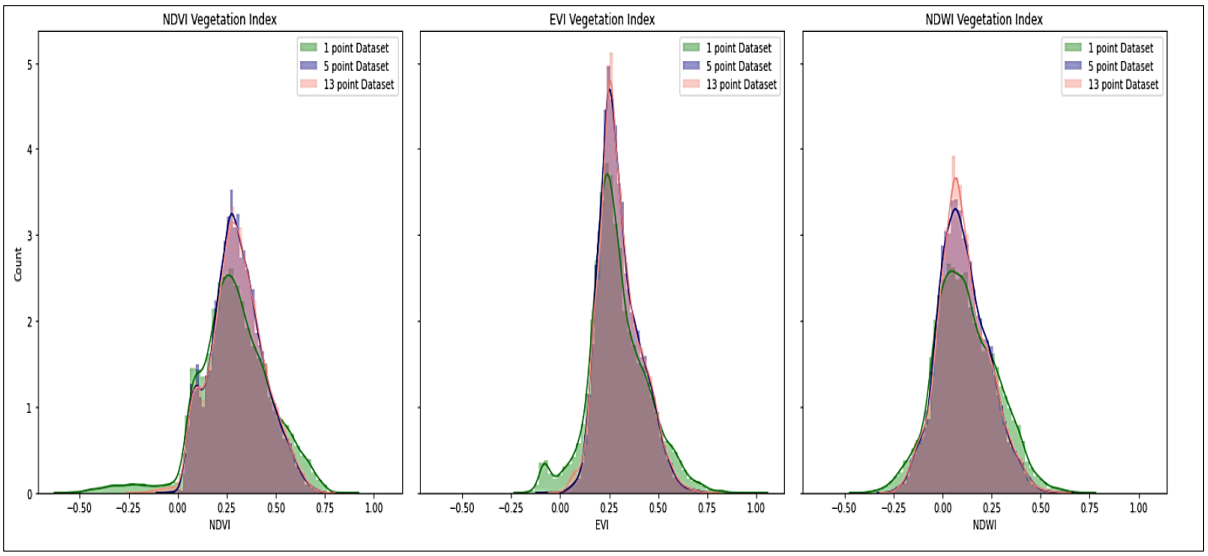


Figure 2.17. *Comparison NDVI, EVI and NDWI values in the 3 dataset*

Visualizations in Figure 2.18 and 2.19 enabled us to view the index fluctuations from 2014 to 2019. The plot is not smooth because it's a fire-prone area. When there are fires, there may be a dip in value followed by a spike in vegetation due to rapid post-fire flora regeneration. The ashes from the fires serve as the nutrient-rich substrate for vegetation growth, although it may take a month or more to grow back the lost vegetation. Hence, in our visualizations, we will see a pre-fire plot with good vegetation, followed by a dip during the fire which continues after the fire, as we are not plotting the data for two weeks post-fire when the vegetation has not recuperated from the fire damage. Water content may experience a surge as firefighting agencies may use water to suppress the fire.

The trend analysis shows how the indices, which were initially lower at the beginning of the year, picked up during the spring and gradually declined as the year progressed with the lowest values towards the end of the year. NDVI and EVI are closely related. The graphical study further validates the correctness of the data.

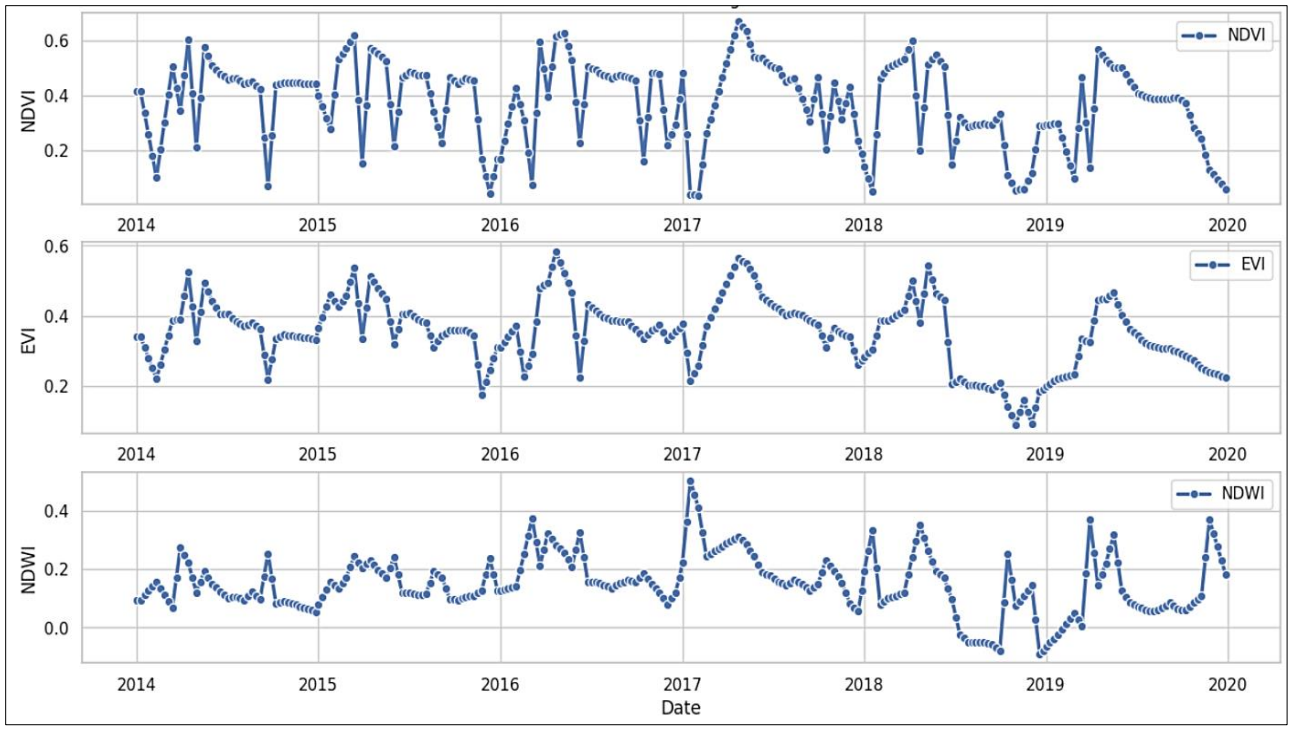


Figure 2.18. *Landsat 8 time-series Vegetation data*

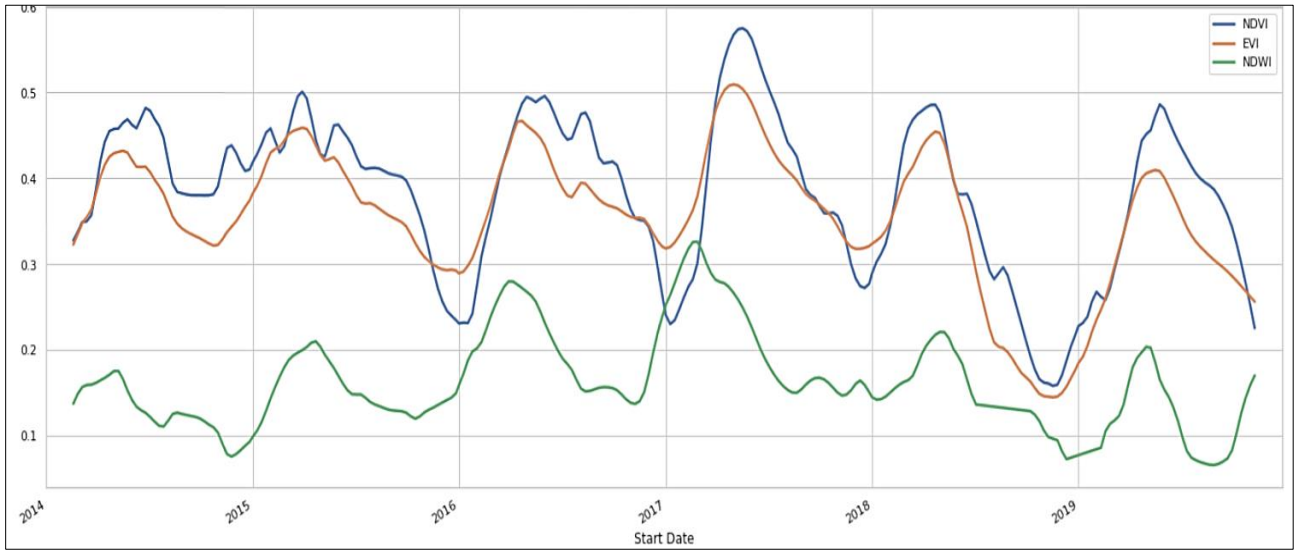


Figure 2.19. *Trend analysis of Vegetation indices (2014 to 2019)*

Further, we have plotted the seasonality of the data over the years for a single point in the area of study in Figure 2.20, 2.21 and 2.22. NDVI, EVI and NDWI seem to follow a seasonal trend wherein the value spikes in spring (green in color) and drops in winter (blue). After spring, the

value seems to steadily decline towards summer (yellow in color) and fall (brown in color). As it is a fire-prone region, there are dips and peaks due to fires and subsequent rapid vegetation regeneration.

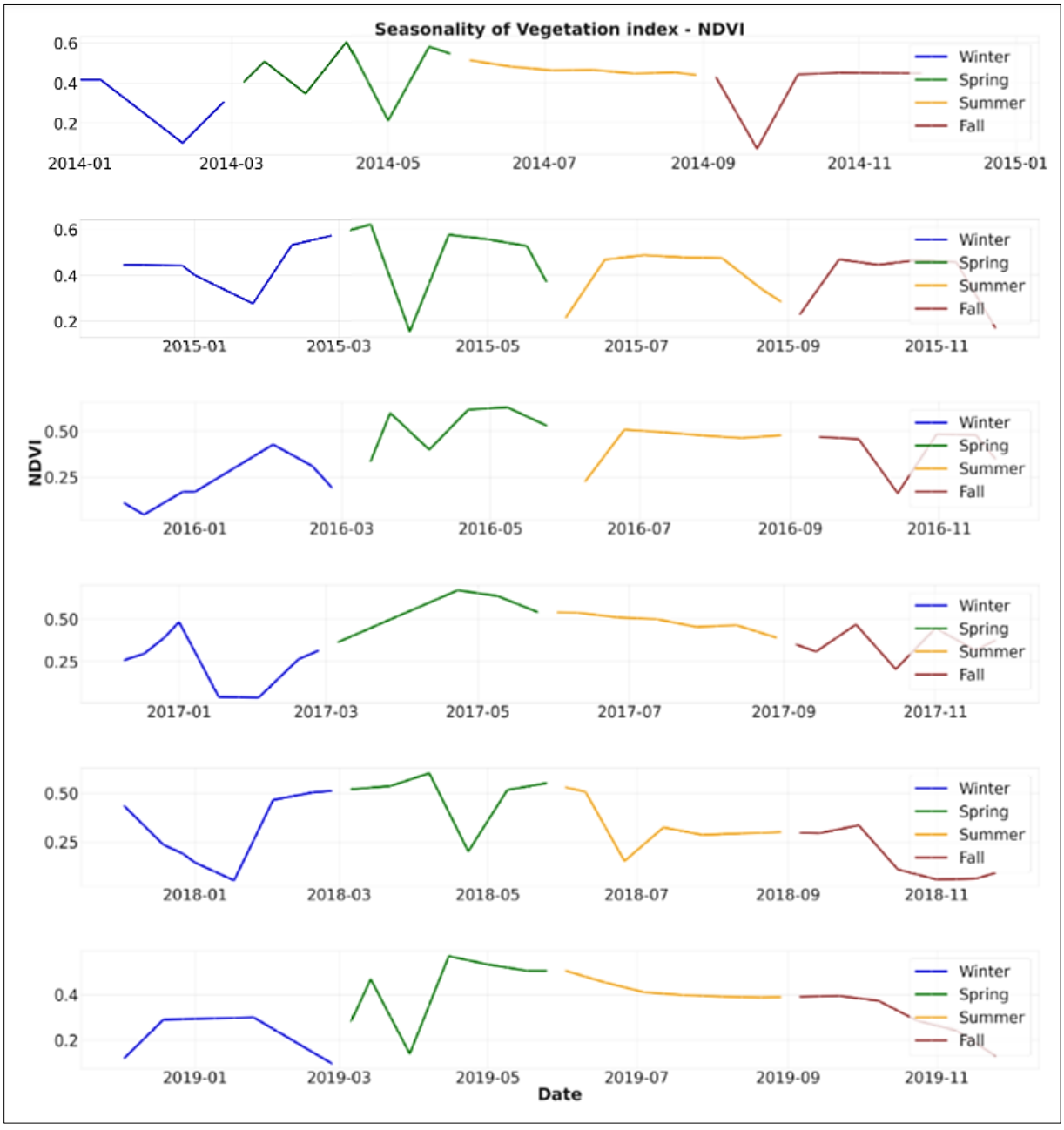


Figure 2.20. *Seasonality of NDVI index from 2014-2019 (Winter, Spring, Summer and Fall)*

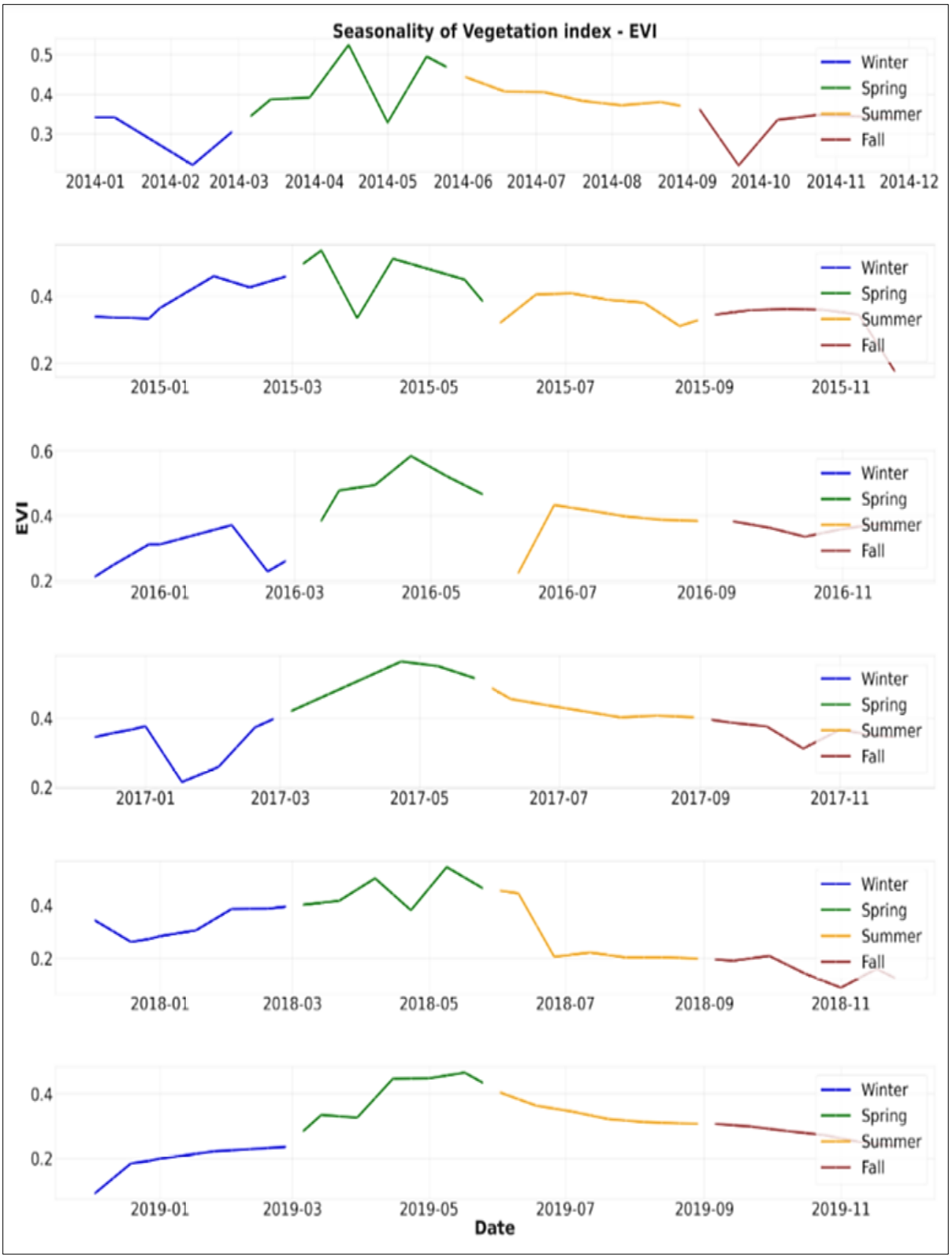


Figure 2.21. *Seasonality of EVI index from 2014-2019 (Winter, Spring, Summer and Fall)*

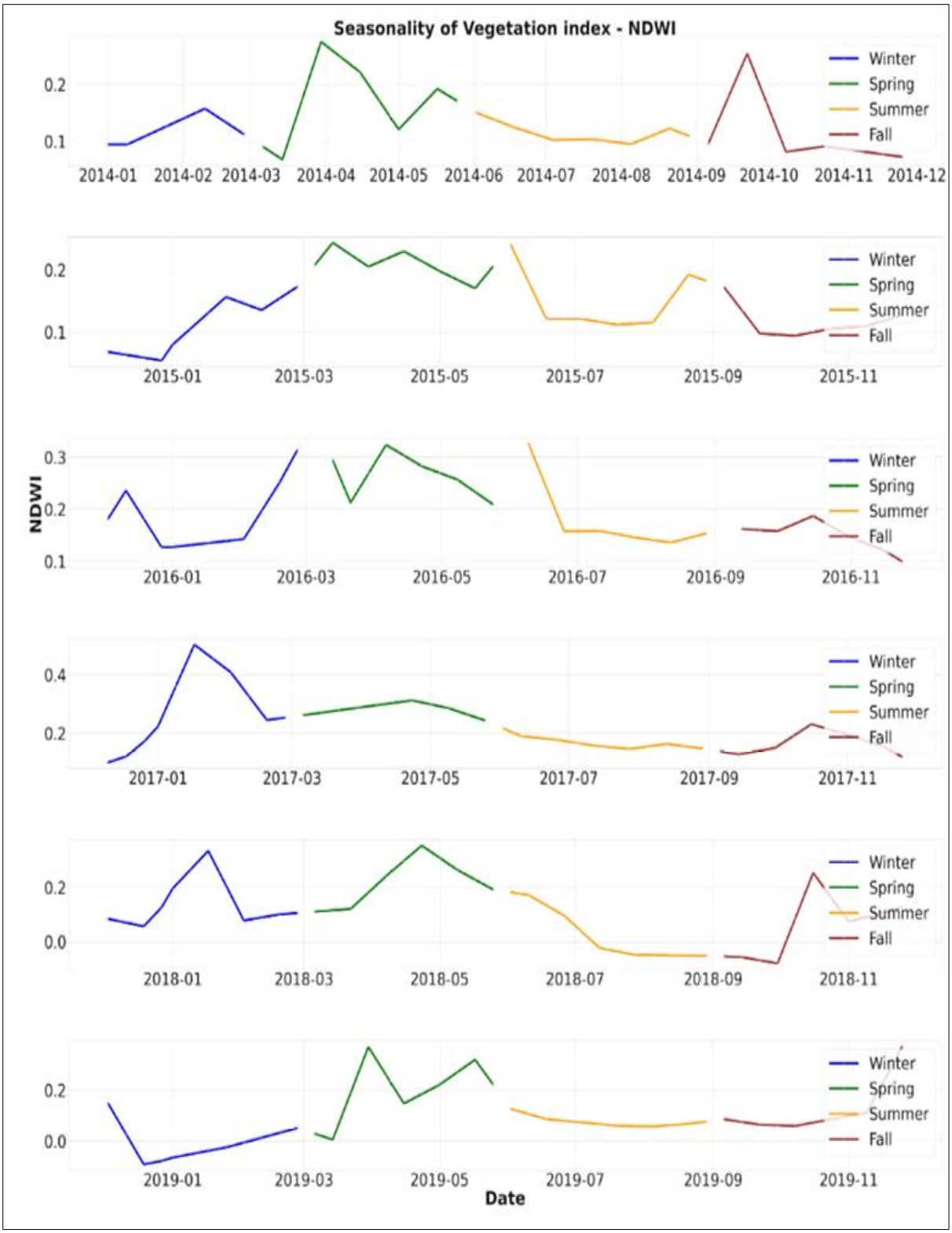


Figure 2.22. *Seasonality of NDWI index from 2014-2019 (Winter, Spring, Summer and fall)*

2.4.4 Terrain data

As discussed in section 2.3.3, using the DEM, we calculated slope, hill shade and aspect and classified each parameter as shown in Table 2.7,2.8 and 2.9.

Table. 2.7. *Classification of Slope*

Slope	Classification
0 to 30	Low Slope
30 to 60	Moderate Slope
60 to 90	High Slope

Table. 2.8. *Classification of Hill Shade*

Hill Shade	Direction
0 to 90	North
90 to 180	East
180 to 270	South
270 to 360	West

Table. 2.9. *Classification of Aspect*

Aspect	Direction of Slope
0 to 22.5	North
22.5 to 67.5	North East
67.5 to 112.5	East
112.5 to 157.5	South East
157.5 to 202.5	South
202.5 to 247.5	South West
247.5 to 292.5	West
292.5 to 337.5	North East

337.5 to 360	North

2.4.5 Powerline data

As mentioned in 2.3.4, after exporting the flat file grids with no powerline passing through it end up having null values. The figure shows the data frame with null values. For all the columns except Status and Circuit, we replaced null values with zero as the powerline is not going through those grids. Whereas for Status column null values are replaced as “Not operating” and for Circuit column they are replaced with “Other” as they are categorical variables. Figure 2.23 and Figure 2,24 shows the resultant data frame before and after cleansing.

	left	top	right	bottom	id	OBJECTID	Name	kV	kV_Sort	Owner	Status	Circuit	Type	Legend	Length_Mil	Length_Fee
	-122.124189	38.55926	-122.114189	38.54926	2	0	NaN	NaN	0	NaN	NaN	NaN	NaN	NaN	0	NaN
	-122.064189	38.50926	-122.054189	38.49926	49	0	NaN	NaN	0	NaN	NaN	NaN	NaN	NaN	0	NaN

Figure 2.23. *Dataset before cleansing*

	left	top	right	bottom	id	kV	Status	Circuit	Length_Mil	Length_Fee
	-122.124189	38.51926	-122.114189	38.50926	6	115.0	Operational	Single	25	133951.91011
	-122.094189	38.53926	-122.084189	38.52926	25	0.0	Not Operating	Other	0	0.00000
	-122.104189	38.52926	-122.094189	38.51926	19	0.0	Not Operating	Other	0	0.00000
	-122.074189	38.52926	-122.064189	38.51926	40	0.0	Not Operating	Other	0	0.00000
	-122.104189	38.51926	-122.094189	38.50926	20	115.0	Operational	Single	25	133951.91011

Figure 2.24. *Dataset after cleansing*

2.5 Data Transformation and Tools

For weather, the raw data is in a machine readable format. We will fit the records with geographic locations of each weather station to correlate with our target data.

For topological data, we will extract latitude and longitude information, then derive the local slope, local aspect and hill shade from the digital elevation models (DEMs). The values will be put into a machine-readable format.

For Vegetation data, vegetation indices are calculated and available as a data product in google earth engine (GEE). 3 data products corresponding to 3 8-day composite vegetation indices are downloaded one image at a time from GEE, concatenated into a data frame and further transformed to get a mean value from 13 points sampled in each grid. Initially, we experimented with 1 and 5 points before deciding on a 13-point dataset. Further, we merged the data structure with the geodataframe from the grid shapefile, which was already updated with centroid, corners, midpoints of edges and diagonals. Thereafter, we imputed null values as explained in the previous section. Alternatively, we tried calculating the indices from the bands and the results were similar although the 8-day composites were more reliable and accurate. The mathematical formulae and Landsat bands-based formulae for calculating Vegetation Index such as Normalized Difference Vegetation Index (NDVI), Enhanced vegetation index (EVI) and Normalized Difference Water Index (NDWI) are given below. Below are the formulas and range [48]. NIR stands for Near-infrared band whereas SWIR is a shortwave infrared band. L in EVI formula adjusts canopy background whereas C value is the coefficient of atmospheric resistance. B is a blue band and G is the gain factor [64].

- Formula for NDVI
    - **NDVI = (NIR – Red) / (NIR + Red)** [48]
    - Landsat 8: NDVI = (Band 5 – Band 4) / (Band 5 + Band 4) [48]
    - Range: -1 to 1
- Formula for EVI

- **EVI = G * ((NIR - Red) / (NIR + C1 * Red – C2 * B + L))** [64]
- Landsat 8: EVI = 2.5 * ((Band 5 – Band 4) / (Band 5 + 6 * Band 4 – 7.5 * Band 2 + 1)).
- Range: -1 to 1

- Formula for NDWI
  - **NDWI = (NIR – SWIR) / (NIR + SWIR)** [48]
  - Landsat 8: NDWI = (Band 5 – Band 6) / (Band 5 + Band 6)
  - Range: -1 to 1

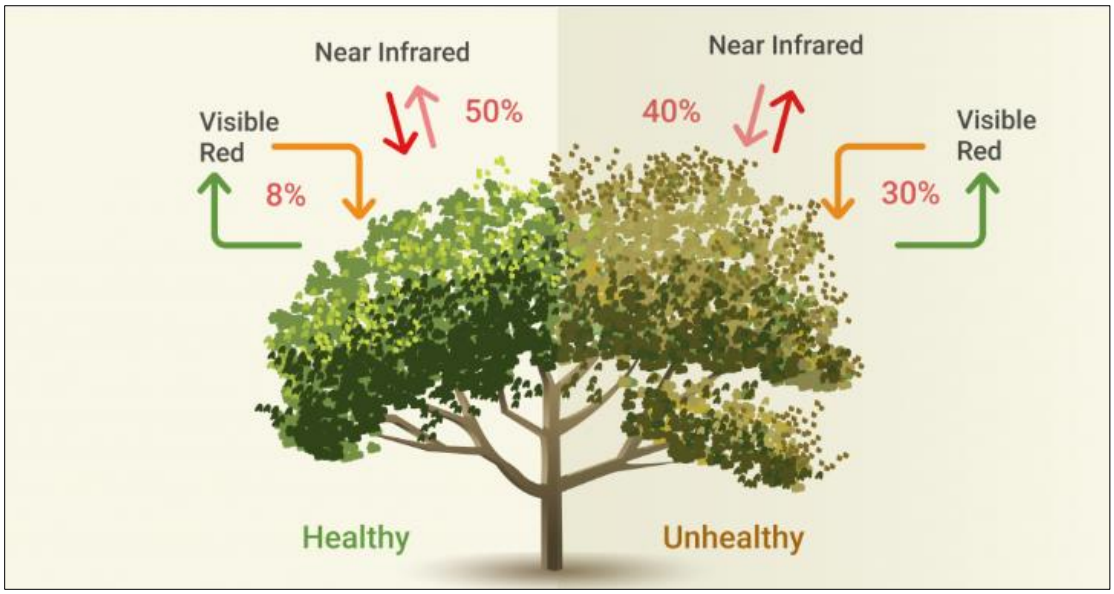


Figure 2.25. *NDVI index for healthy and unhealthy/dry plant [61]*

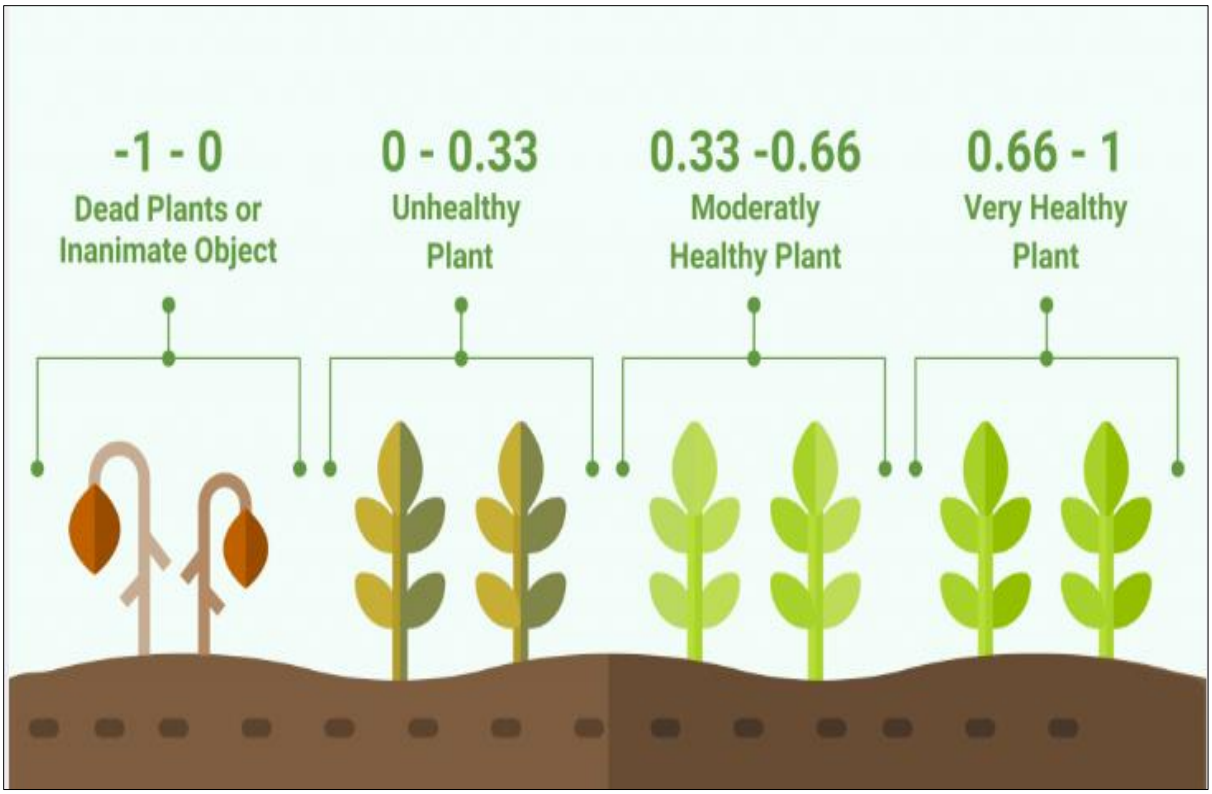


Figure 2.26. *NDVI index range [61]*

Figure 2.25 illustrates the calculation of NDVI for healthy and unhealthy plants, as well as the index range and Figure 2.26 shows the NDVI vegetation index visualization.

Using vegetation indices, we can quantify the greenness and classify the vegetation types. To reiterate, NDWI aids in water content analysis, including the leaf water content. Although NDVI [63][45][47][24] is the standard vegetation index with simple calculation and reliable classification, it is highly sensitive to chlorophyll content. EVI is an enhanced version of NDVI with increased sensitivity to biomass and canopy type as it has lesser distortion due to atmospheric conditions such as cloud cover and background noise such as soil reflection. Table 2.10,2.11 and 2.12 show the NDVI, EVI and NDWI range respectively.

Table 2.10. *NDVI Range and type of vegetation [65]*

NDVI	Type of vegetation/water
-1 to -0.7	Water bodies
-0.7 to -0.2	Barren rocks, sand, or snow
0.2 to 0.6	Shrubs, grasslands or senescing crops
0.6 to 1.0	Dense vegetation, tropical or temperate rainforest

Table 2.11. *EVI Range and type of vegetation [65]*

EVI	Type of vegetation/water
-1 to -0.1	Water bodies
-0.1 to 0.3	Barren Rocks, Sand, or Snow
0.3 to 0.6	Shrubs, Grasslands, or Senescing crops
0.6 to 1.0	Dense Vegetation, Tropical or Temperate Rainforest

Table 2.12. *NDWI Range and type of vegetation [65]*

NDWI	Type of vegetation/water
<= -0.1	Dry
-0.1 to 0.3	Moist
>= 0.3	Wet

Figure 2.27 displays the plotted NDVI indices from Landsat8 extracted data for our focused area. Figure 2.28 is a time-series plot for a single point.

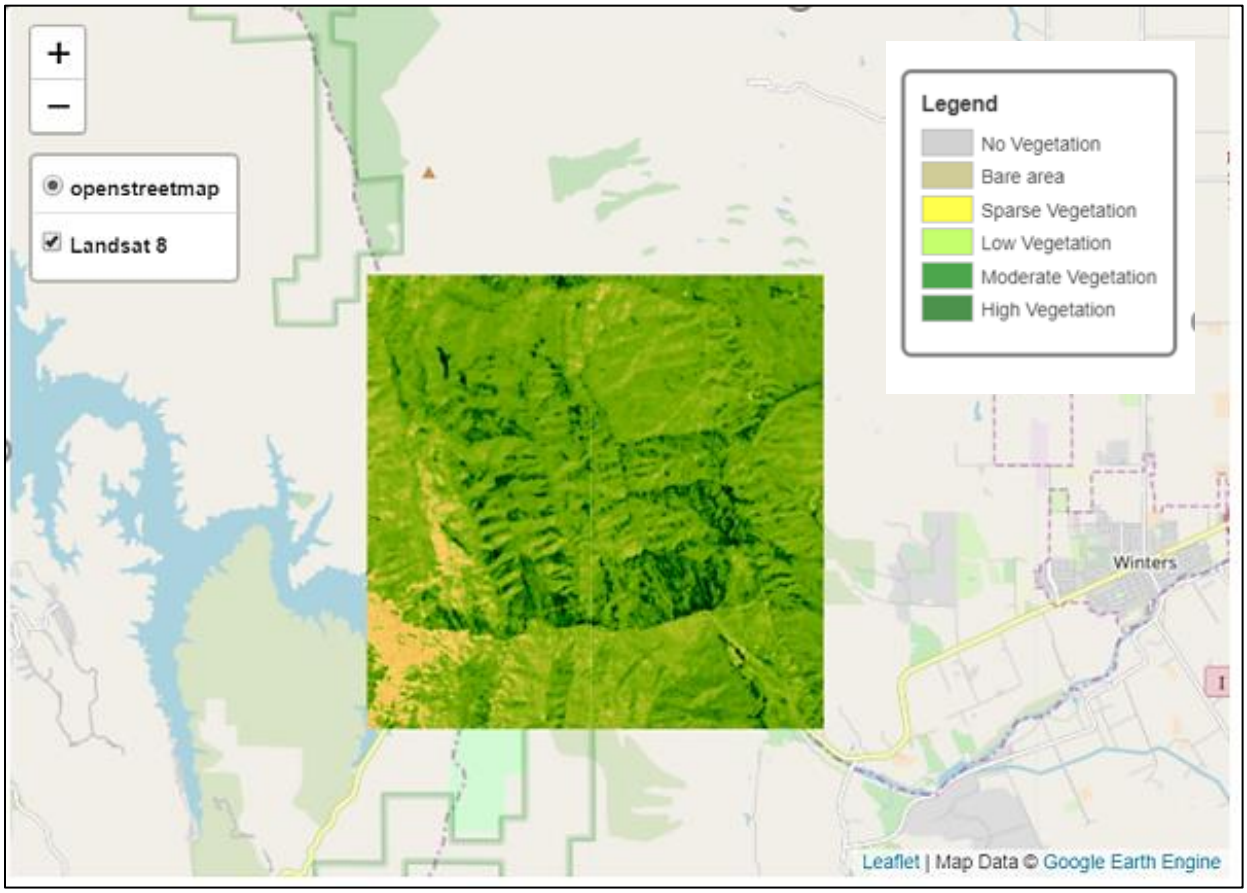


Figure 2.27 *NDVI values from Google Earth Engine (GEE) for the study area*

The tools we will use include Quantum Geographic Information System (QGIS), Google Earth Engine (GEE), ArcGIS, Python libraries, Jupyter Notebook, Google Colab, AWS and other services.

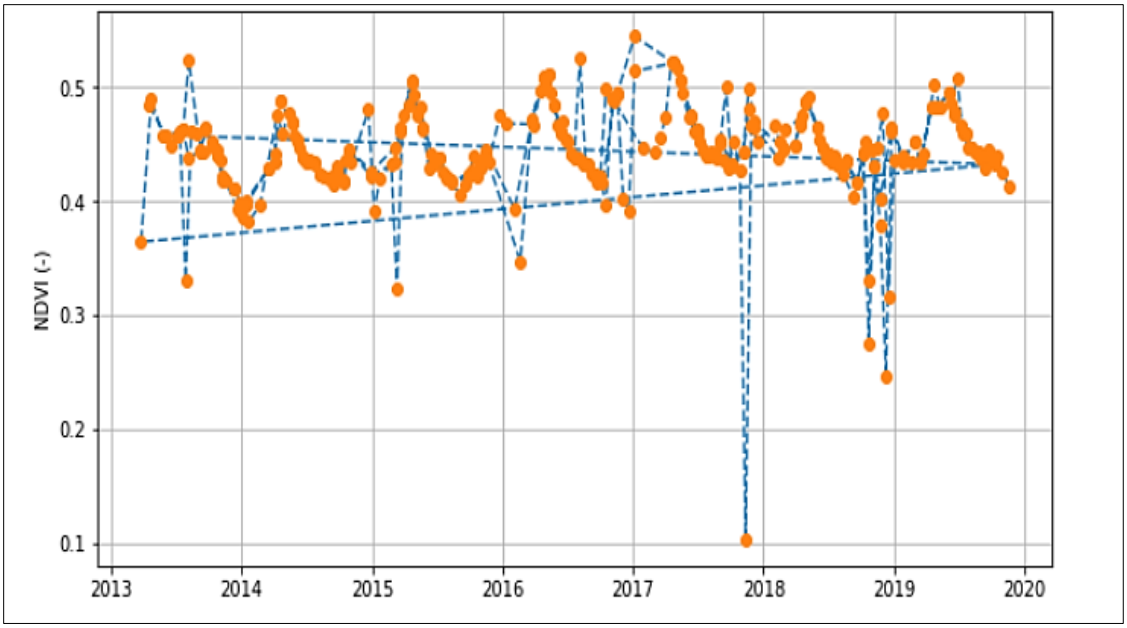


Figure 2.28. *Landsat 8 NDVI time series data for a single grid*

2.6 Raw Data Visualization

Figure 2.29 and 2.30 shows the bands derived from Landsat 8 can be visualized by layering over a map. We can use matplotlib or folium to perform this task.

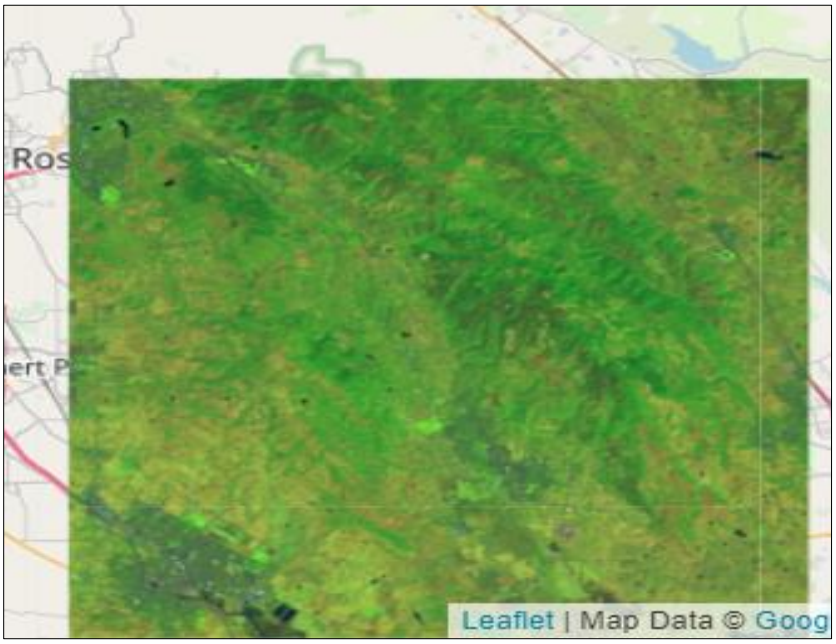


Figure 2.29. *Landsat Raw data visualization*

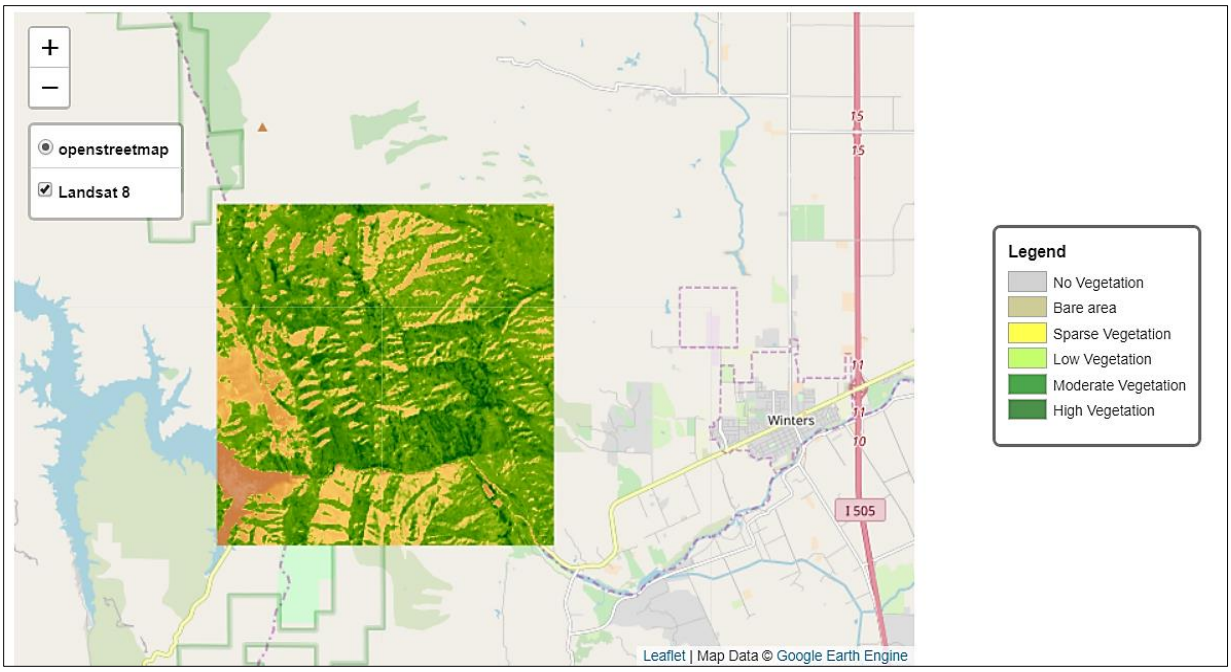


Figure 2.30. *Landsat 8 NDVI visualization*

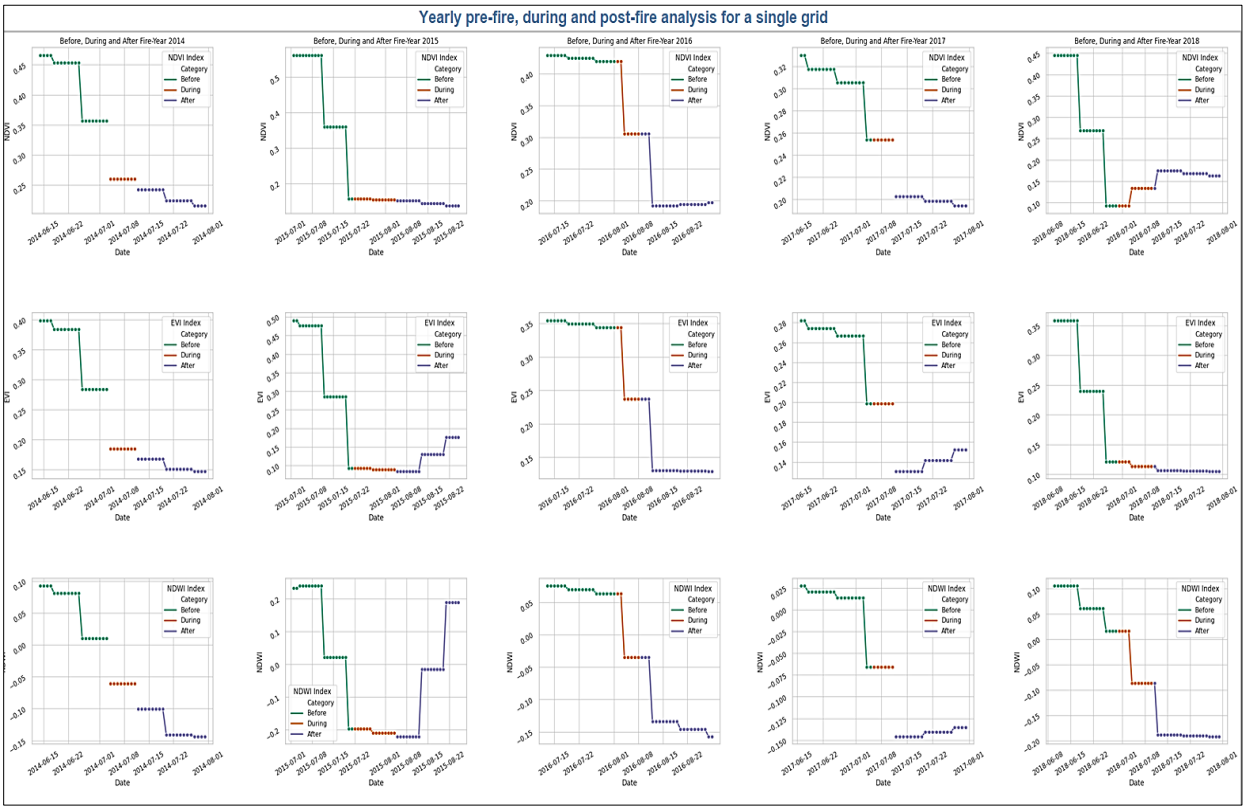


Figure 2.31. *Vegetation data for individual grids - Before, during and after Fire*

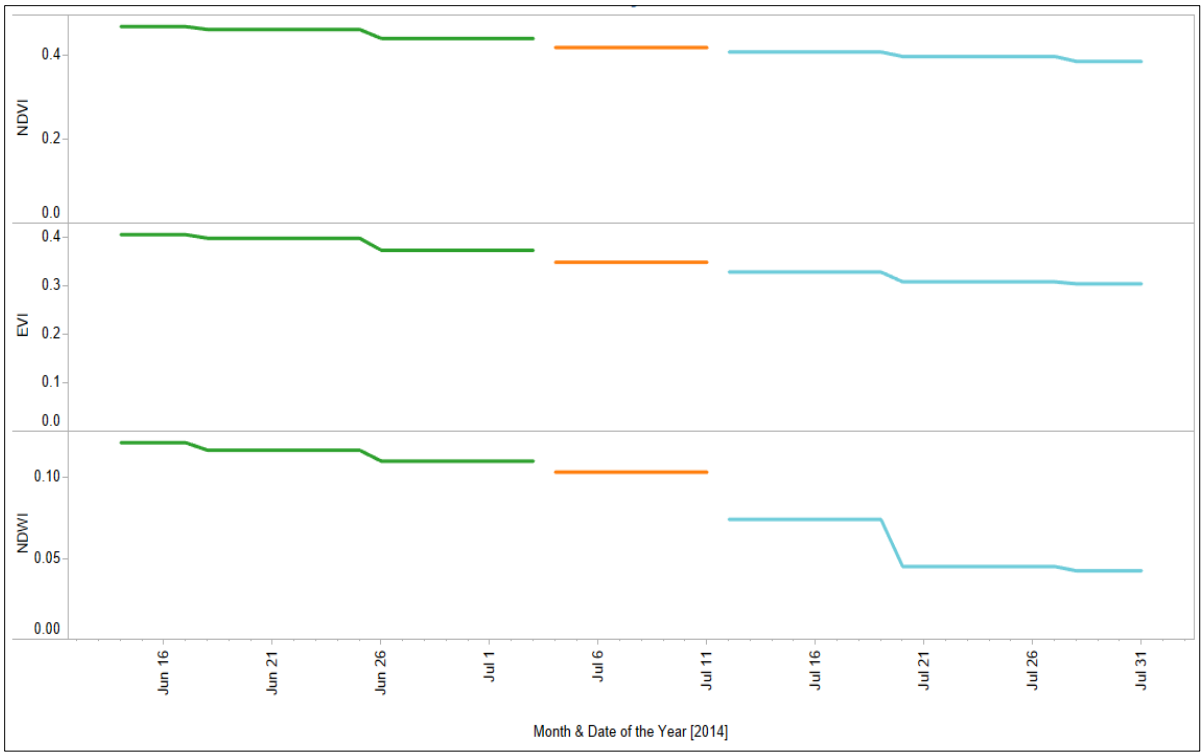


Figure 2.32. *Vegetation Data - Pattern Before, During and After Fire for a single grid*

Figure 2.31 and 2.32 shows the variation of vegetation data before, during and after Fire. We are trying to capture the same for individual grids.

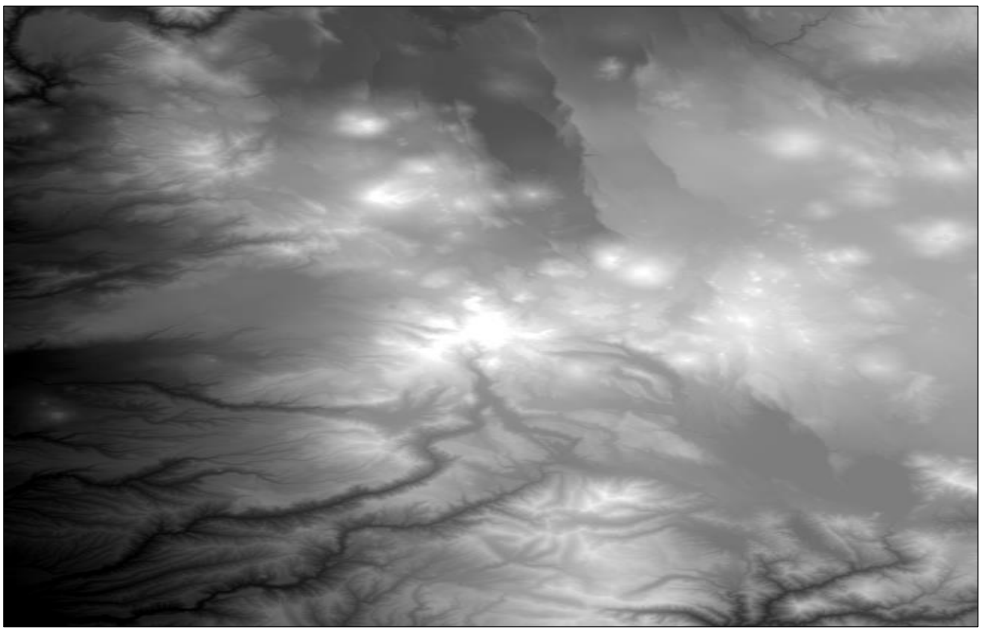


Figure 2.33. *DEM Map*

Figure 2.33 shows the raw data visualization for ⅓ arc-second DEM map. Using this DEM map, raster analysis parameters slope, hill shade and aspect are calculated.

Figure 2.34 shows the powerlines in California. With this data, we can determine whether there is a correlation between power lines and wildfires.

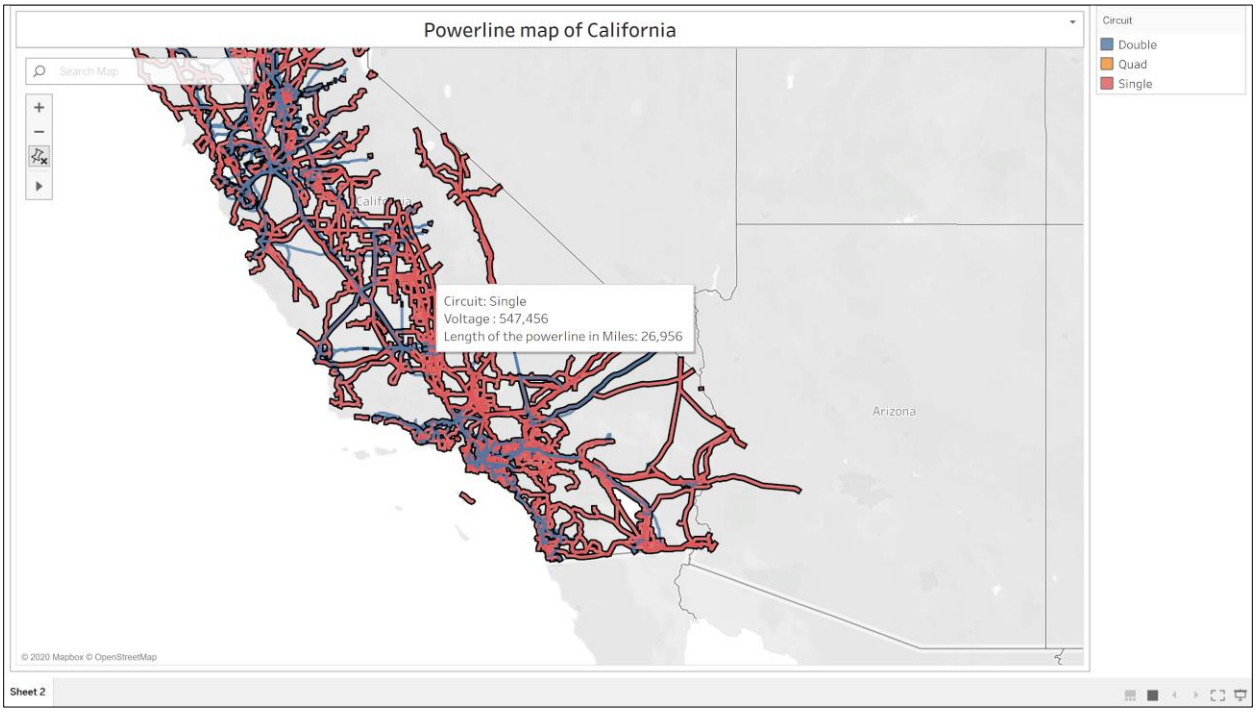


Figure 2.34. *Powerline map*

Figure 2.35 shows the visualization of the categorical variable “Circuit” in powerline data. Powerlines with single circuits are relatively more when compared with double and other (Hughson-Grayson, LaGrange - Don Pedro-Hawkins, Quad, Liberty Energy) circuits..

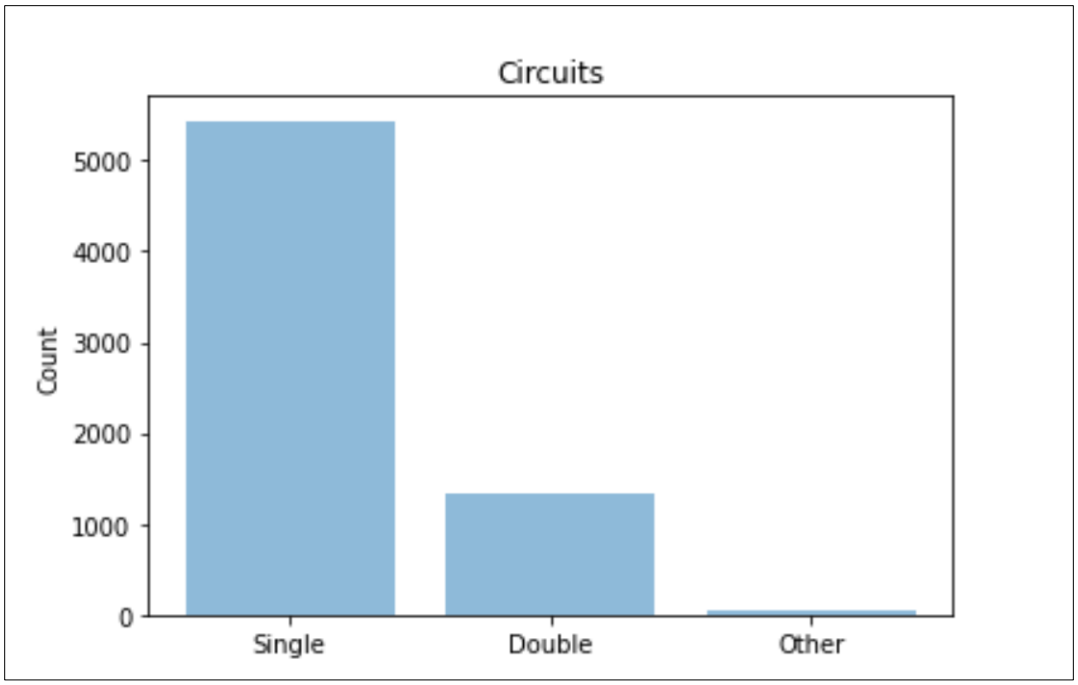


Figure 2.35. *Comparison of powerline circuits*

Figure 2.36 of the geographic distribution of weather stations that provides us the weather data for predicting wildfires.

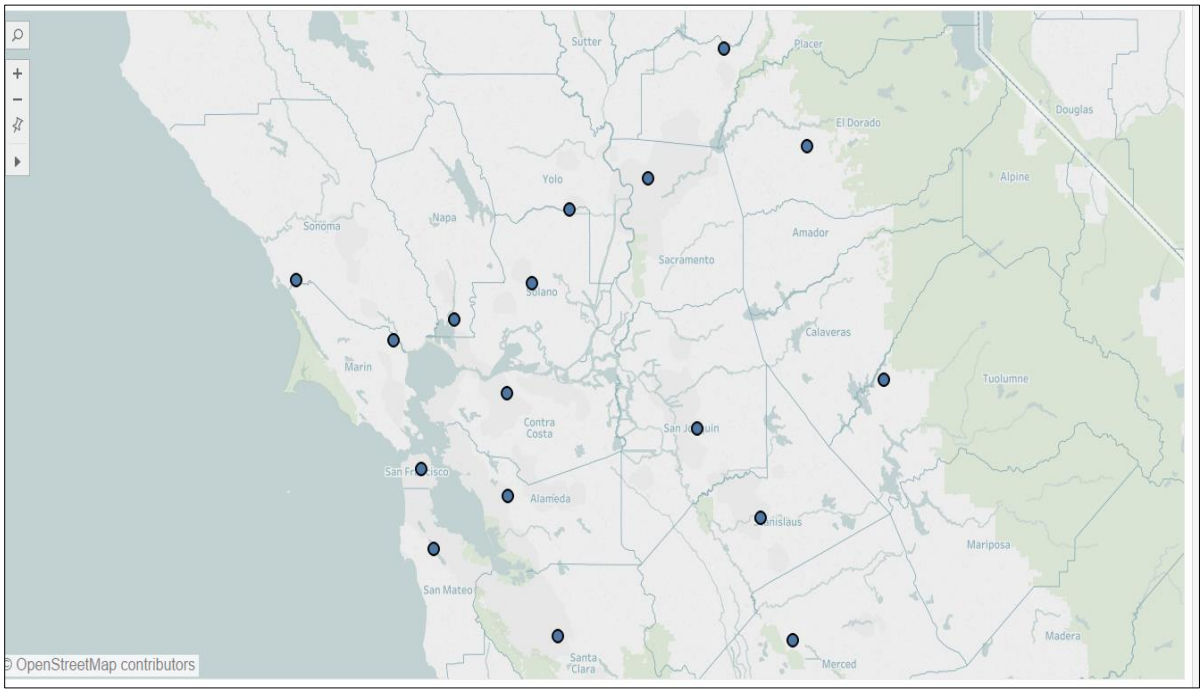


Figure 2.36. *Weather Station Map*

We divided weather data into four groups for better analysis. This helped us to isolate the relevant subset of the dataset.

- Before – weather records one week before the start of the fire
- Start – weather records on the day of fire started
- During – weather records during the fire
- After – weather records after the end of the fire

Figure 2.37 shows the comparison of Dry bulb temperature between the aforementioned groups. It is evident that the temperature is high at the start of the fire. Figure 2.38 represents how relative humidity changed over time. It is high during and after the fire. Wind speed is high on the day when the fire is started, this can be depicted from Figure 2.39.

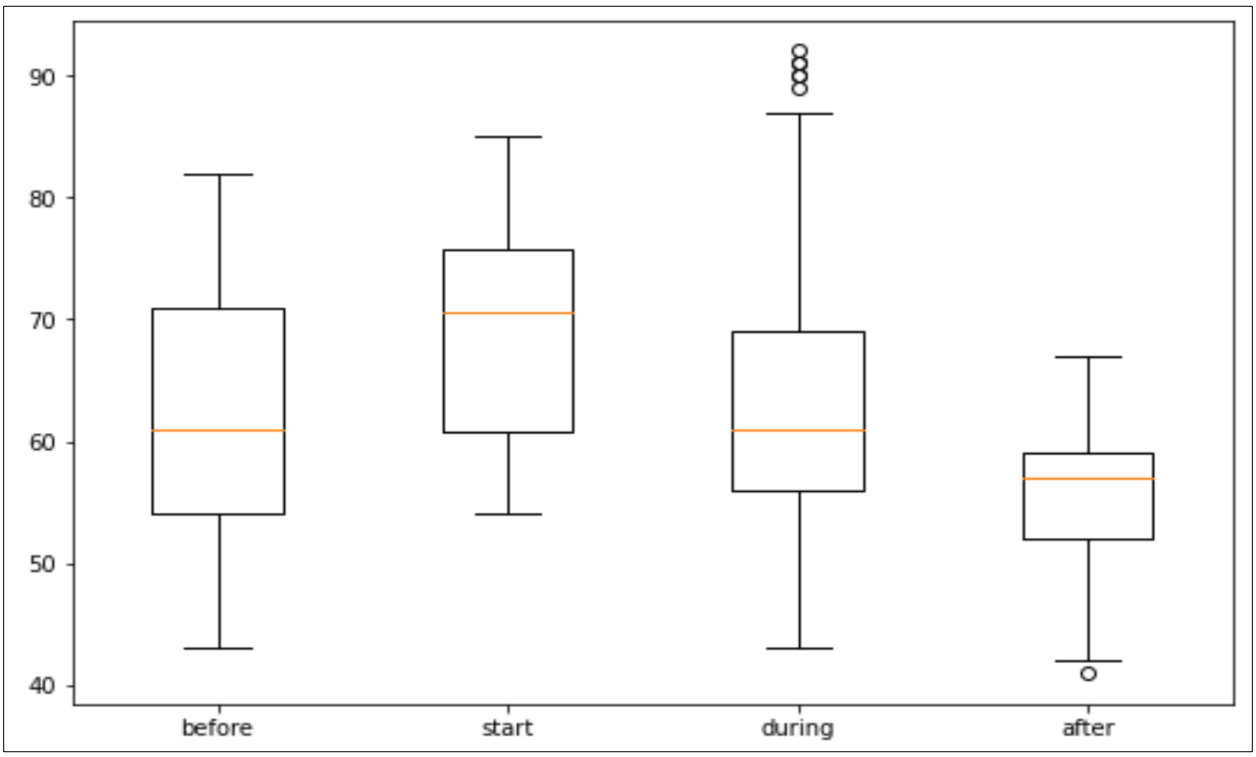


Figure 2.37. *Comparison of Hourly dry-bulb temperature*

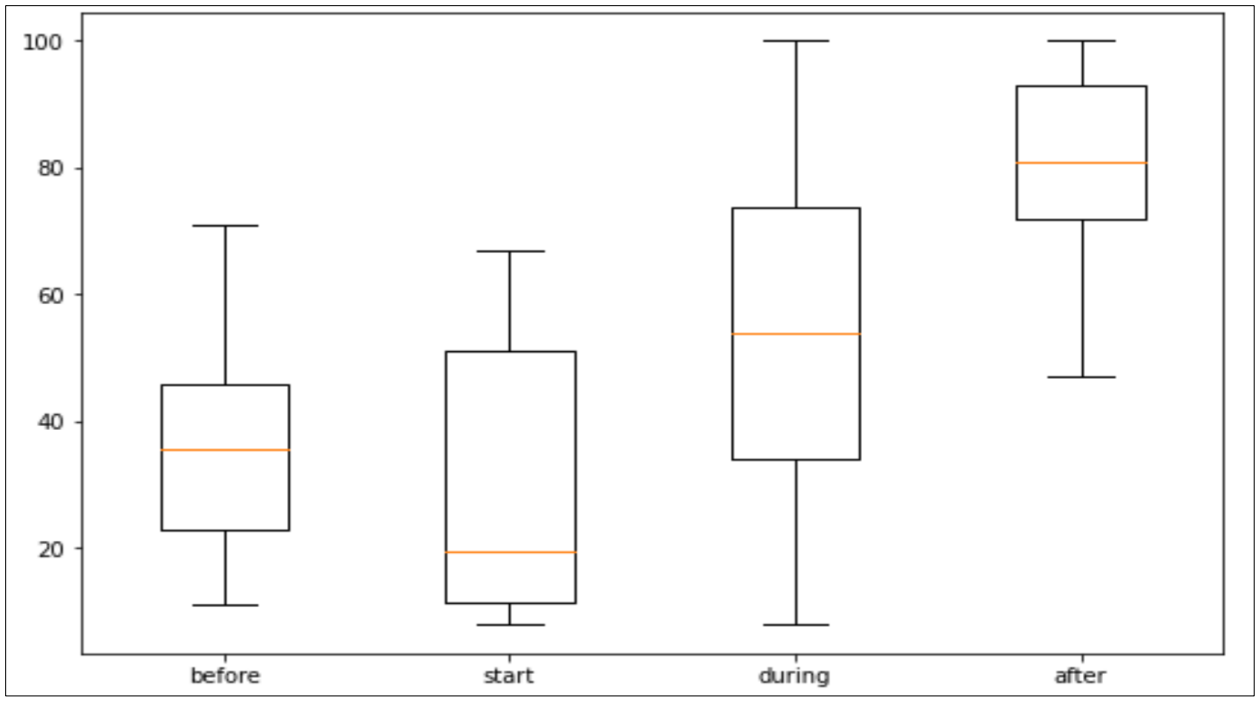


Figure 2.38. *Comparison of Hourly relative humidity*

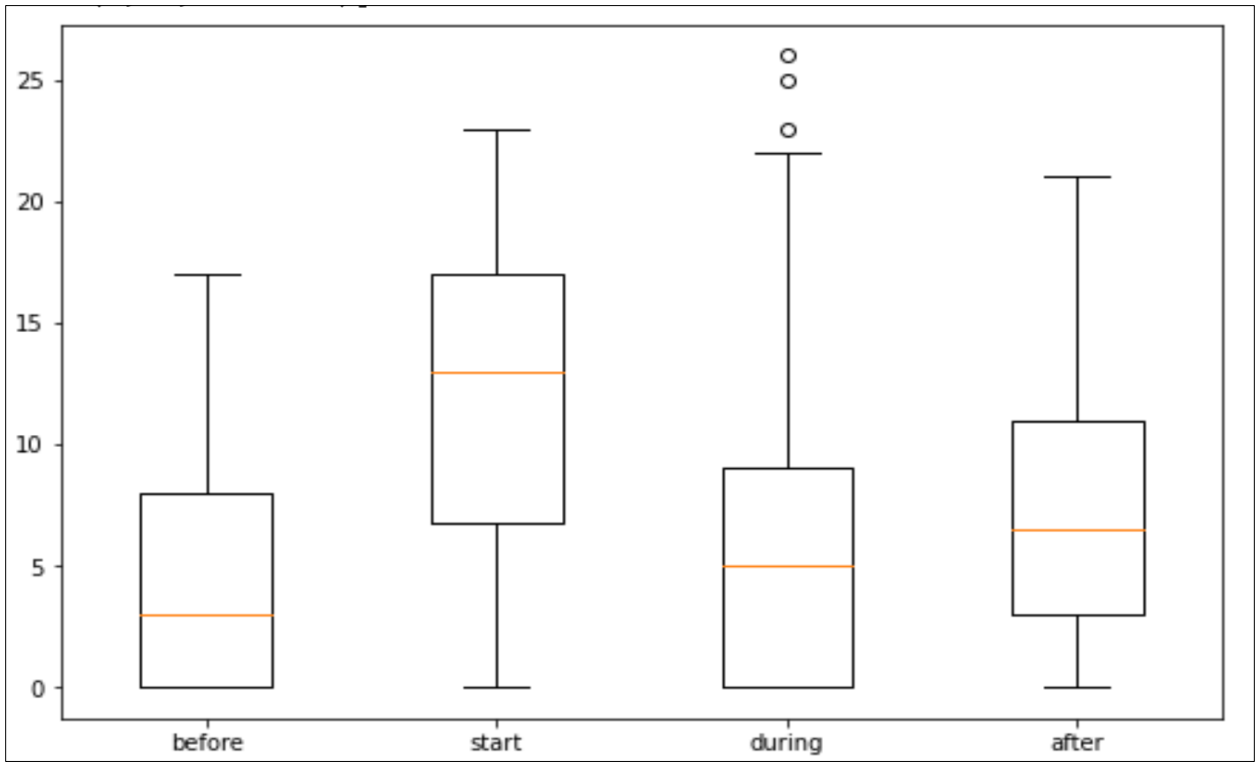


Figure 2.39. *Comparison of Hourly wind speed*

3. Project Management

3.1 Project Organization

Figure 3.1 shows the cross-industry standard process for data mining (CRISP-DM) model used to organize the team's work into manageable sections.

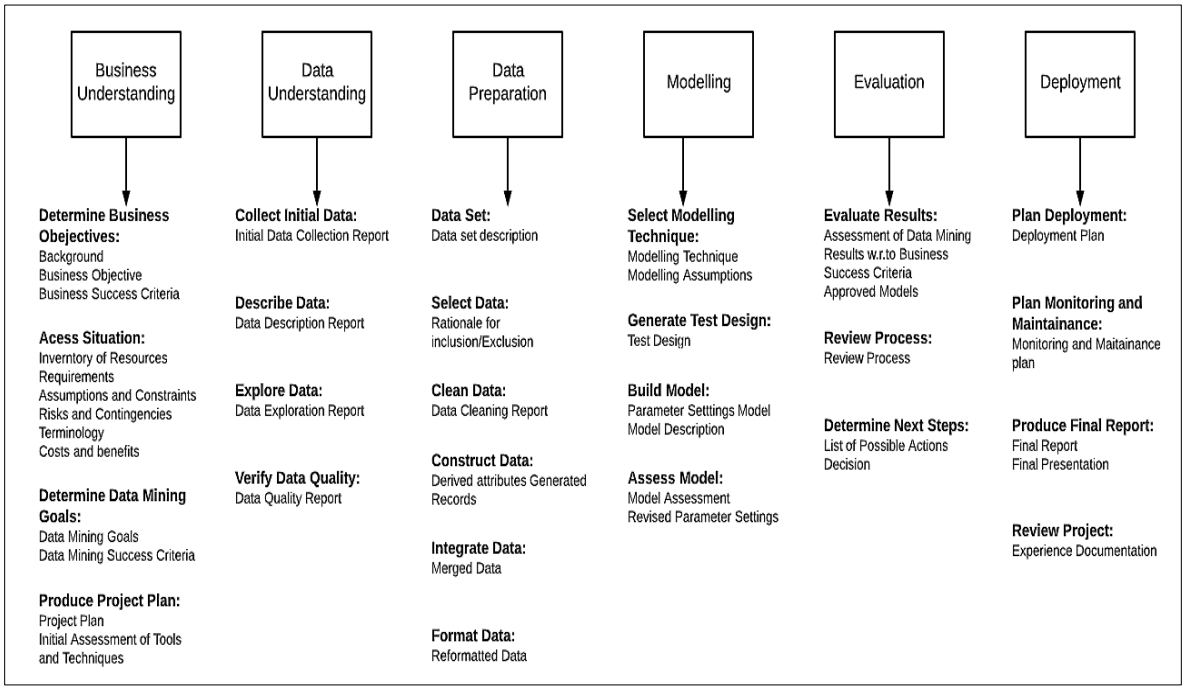


Figure 3.1. *CRISP-DM [26]*

Business understanding, data understanding, data preparation, modeling, evaluation and deployment are the major steps of this undertaking. We start with the determination of business objectives and access the situation before defining the goals and project plan. For data understanding, we collect, describe, explore and verify the data. In preparation stage, we create the dataset by selecting, cleaning, constructing, integrating and formatting the data. Modeling involves technique selection, test design and model building and assessment. Further, we evaluate and deploy the model in a user interface designed for area-specific wildfire risk prediction.

3.2 Resource Requirements

3.2.1 Data Requirements

We are planning to outcome work with the publicly available datasets, as it is the most cost-effective method, given the dearth in funding. Below are the types of datasets and their sources,

- Weather: National Oceanic and Atmospheric Administration (NOAA).
- Vegetation: Remote sensing satellite data (Landsat 8) from Google Earth Engine (GEE).
- Topology: United States Geological Survey (USGS) DEM data from the USGS 3D Elevation Program (3DEP).
- Equipment Data: California Energy Commission (CEC) Electric Transmission Line geospatial data.
- Forest Fire History: Fire history data from Fire and Resource Assessment Program (FRAP).

3.2.2 Intellectual Property Rights

- For Vegetation data, we attempted to contact a private entity (Harris geospatial entity) for high-resolution satellite data. They replied with a quotation of close to 200k dollars for the images with the high image resolution. Hence, we decided to go with the publicly available data that has been allowed for academic purposes.
- If we use NASA AppEEARS data, as a supplementary data source in the future, it is essential to provide the citation for both Software and Data products before publishing the satellite data and imagery, to avoid any Intellectual Property (IP) infringement.

- Several pictures were used in the User interface and presentation. Although only royalty-free images were hand-picked for this task, we have provided the website details in the appendix to avoid legal hassles in the future.

3.3 Project Schedule

An Enhanced Project Evaluation and Review Technique (PERT) chart was used to organize the dependent serial tasks in the schedule. Figure 3.2 shows the PERT chart for this research project.

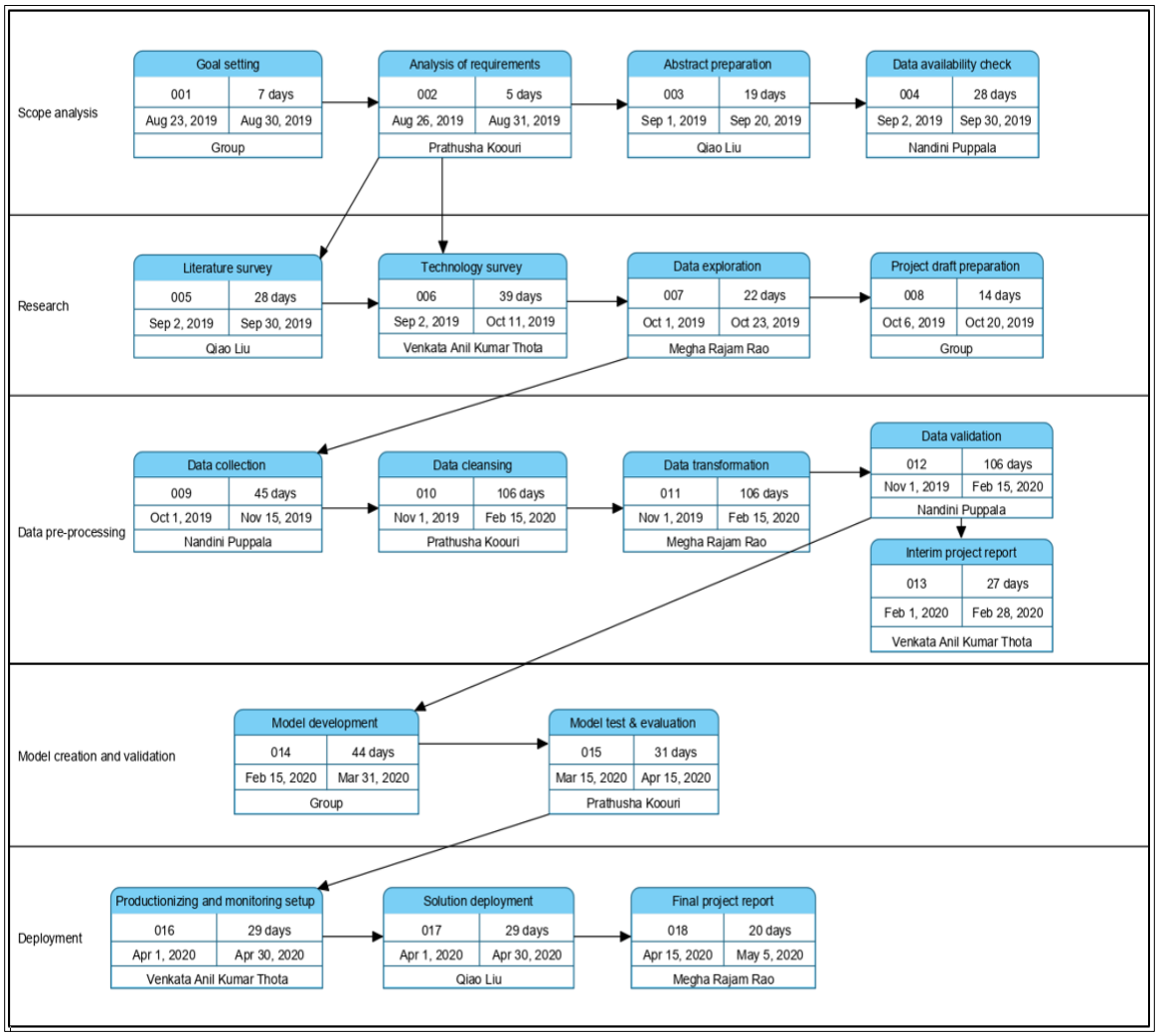


Figure 3.2. *PERT chart*

4. Problem Formulation and Model Selection

4.1 Problem formulation

4.1.1 Motivation and goals

Wildfires have caused extensive damage to the ecosystem, including plants, animals, human beings and the environment. Reiterating the statistics from National Geographic, an average of greater than 100000 wildfires burns 4-5 million acres of land in the United States every year, clearing up to 9 million acres of land in recent years. With a maximum speed of 14 miles per hour, the environmental impact is substantial as it consumes everything in its way. The surrounding area is subject to the risk of mudslides. The fire history data from 2018 from the state of California reveals damages of around $800 million with more than 100 lives lost. The recent devastating wildfires in Australia [49] and Amazon [50] bring serious attention, and urgent demands among the research community and fire emergency agencies on how to address the challenges in wildfire detection, prediction, prevention as well as fire emergency management, rescue responses. To address this challenge, the timely wildfire detection and real-time prediction of wildfire risk is the need of the hour. Our goal is to address the current wildfire crisis in Northern California by leveraging modern big data analytics technologies and state-of-the-art machine learning models for discovering, managing and analyzing large and complex data sets.

4.1.2 Demerits of the traditional approach

Many of the former research papers addressing wildfire detection, simulation, and prediction in the past years, have relied on applied traditional mathematical and statistical

approaches based on physical models with mathematical equations and historical data [51]. These models and methods have below drawbacks.

- Limited perspective data parameters and location-specific historical data,
- Inadequate regional and location-oriented wildfire detection accuracy and prediction,
- Lack of large-scale and concurrent real-time computation support due to their dependency on a single computation process.

This causes serious challenges in real-time information processing and intelligent decision-making for firefighters to respond and rescue operations to proceed.

According to our system survey (refer Table 1.2), we have concluded that most of the deployed wildfire emergency systems [52][53][54][55][56] worldwide use conventional wildfire detection and prediction approaches and technologies, and depend on limited data sources and parameter technology solutions. There is a lack of intelligent real-time solutions and systems to cope with large-scale wildfire risk using big data-driven machine learning models based on cloud technologies. This proposed research project focuses on wildfire challenges in risk prediction by developing innovative integrated machine learning models for wildfire prediction.

4.2 Foundation of Proposed Solutions

4.2.1 Overview

Based on our research on the existing Wildfire prediction systems, most of the government agencies are using statistical models based on the pre-calculated indices. There has been immense interest in newer hybrid methods and ongoing studies using machine learning approaches for wildfire detection. Albeit, there are no scalable solutions to address this problem.

Figure 4.1 and Figure 4.2 gives an overview of our proposed real-time scalable location-based machine learning solutions with the ensemble and combined models.

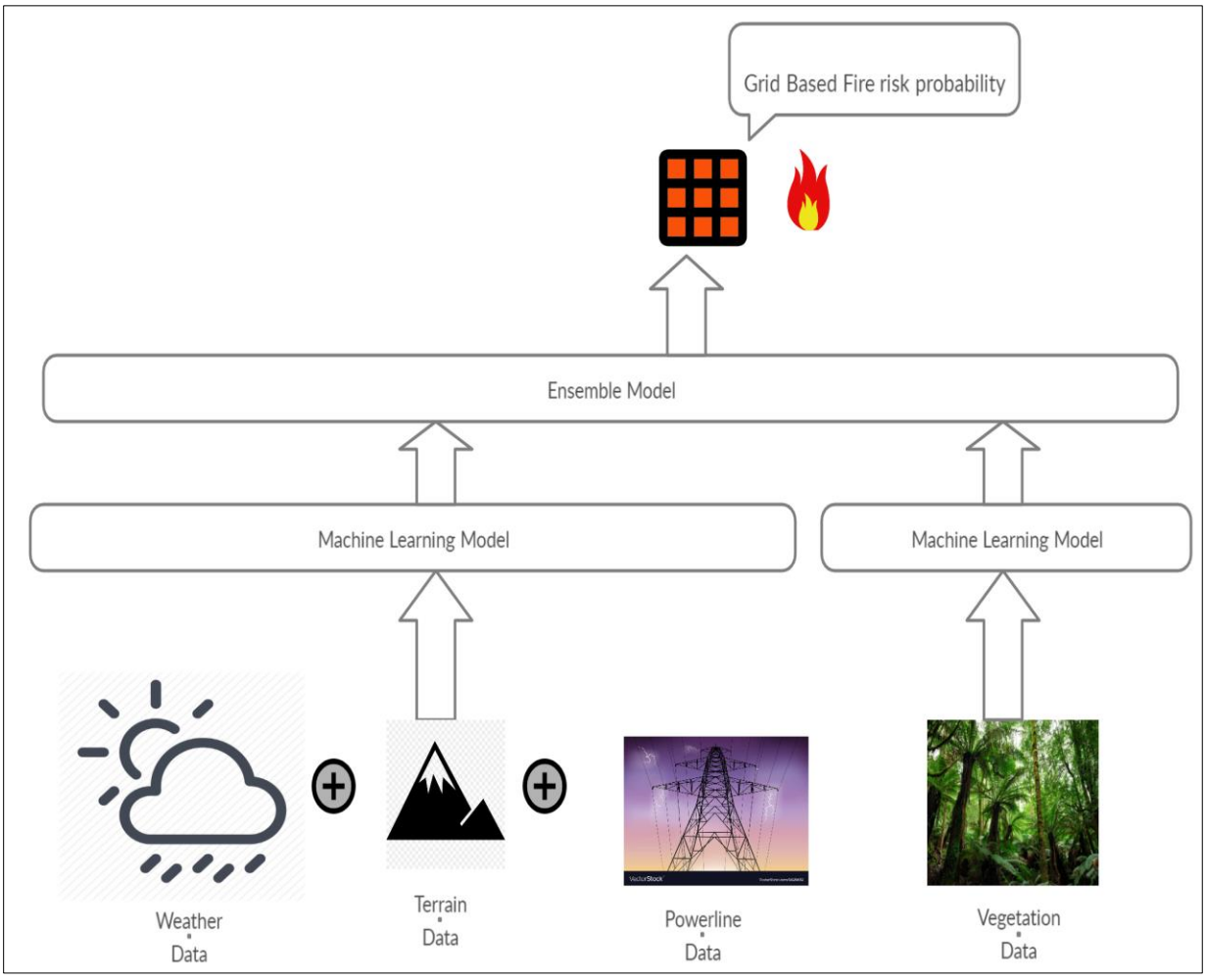


Figure 4.1. *Proposed Solution with Ensemble Model*

In ensemble models, individual parameter-based models were ensembled to convert weak learners into strong learners. Generally, ensembling by stacking results in exceptionally high accuracy models. As terrain and powerline data did not produce good standalone models, we decided to combine it with the weather data, whereas there was a separate model for Vegetation data. Both groups were subjected to exhaustive experimentation with different subsets of data and target labeling. Thereafter, the best models were ensembled by stacking.

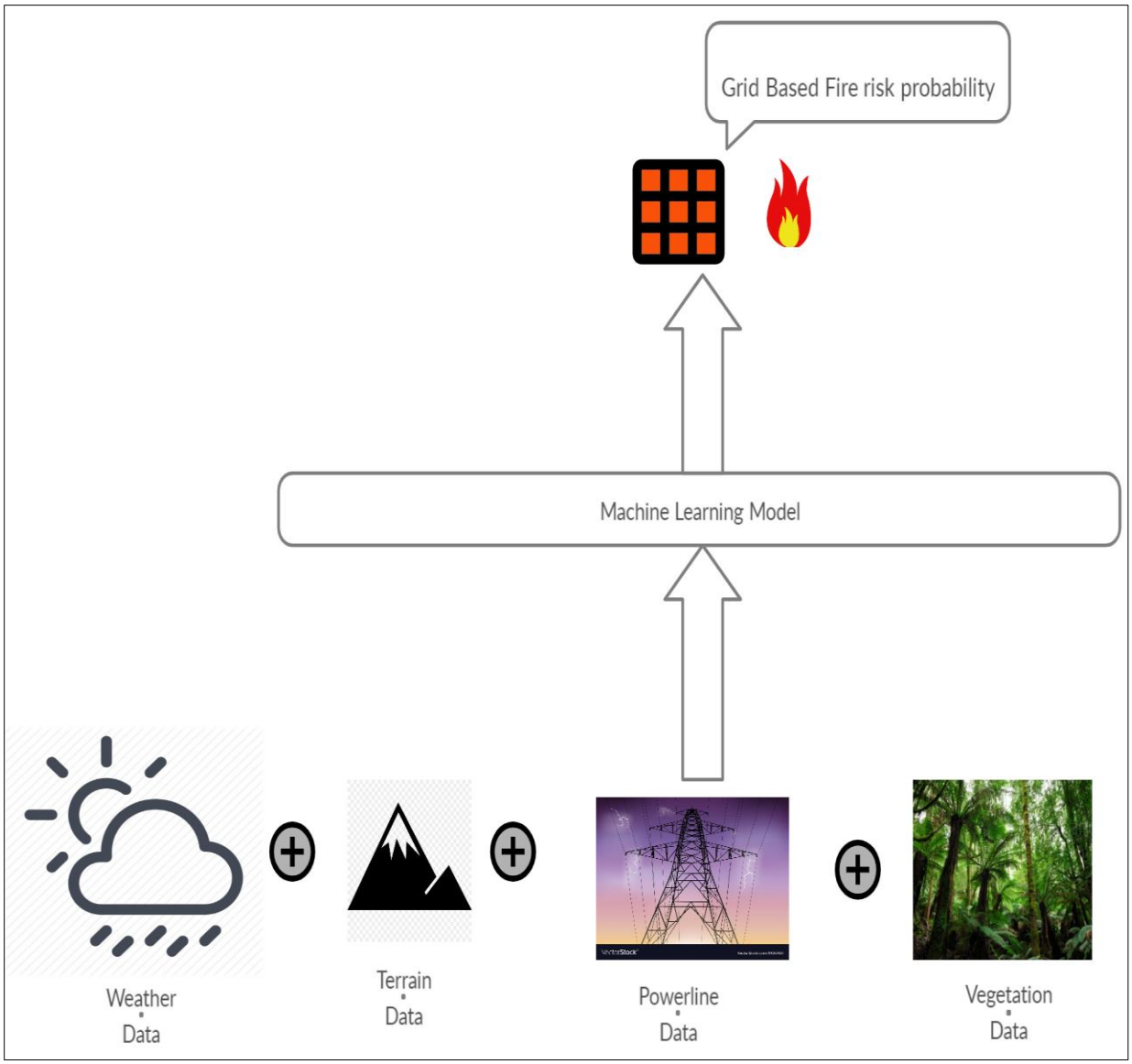


Figure 4.2. *Proposed Solution with Combined Model*

In the combined model, we merge all parameter-based datasets to generate a model with the best-performing algorithm with reliable accuracy and least false negatives.

Creating a machine learning model to detect the fire is a major challenge especially due to the sparse and unstructured nature of publicly available data. Data discovery, access to data, training, cleaning and feature engineering were very challenging owing to the nature of the data

we have been dealing with. Real-time wildfire detection requires a location-based solution supporting large scale monitoring.

4.2.2 Theory of proposed solution

While addressing any wildfire problem, there are four major facets to be considered. They are as follows.

- Wildfire risk prediction and warning
- Wildfire monitoring and detection
- Wildfire spread/progression prediction and responses
- Post-fire analysis and estimation

As a part of this project we are dealing with the first issue in addressing wildfire, that is, wildfire risk prediction and warning. Until now, many research papers have been published to address this issue. According to our literature survey [Section 1.5], existing researchers have used three different approaches in wildfire study, analysis, and prediction.

- First is the statistics-based approach. For example, G. Bianchinia *et. al* [57] combines statistical analysis with parallel evolutionary algorithms to improve the quality of the model output.
- The other approach is the simulation-based approach, which usually depends on a large number of input parameters with uncertainty in real-time of wildfires[58][59].

Both approaches have the limitations and difficulty in modeling and presenting real-time dynamic wildfire risk prediction and predicting the changes based on risk pattern changes due to environmental factors in land-cover, landscape, and climate changes. Recently, few researchers

have begun to use the data-driven machine learning approach, in which machine learning models are trained using combined big data sets to predict wildfire risk.

This proposed project has two major intellectual merits which are given below. Further, it also provides a starting point for exhaustive research in the future.

- The project will develop innovative models with hourly or biweekly data for area-based wildfire risk prediction with augmented spatial sensitivity. Unlike current wildfire models, these models are multi-columned machine learning models, which are trained and developed to address urgent needs in location-based wildfire risk prediction. We plan to use Landsat satellite data to derive vegetation Indices such as Normalized difference vegetation index (NDVI), Normalized vegetation water index (NDWI) and Enhanced vegetation index (EVI). Topology, weather and powerline datasets along with fire history data were extracted. The tradeoff among spatial coverage, spatial resolution, and temporal update frequency was evaluated before the final datasets were determined. As a result, for weather, the dataset has exceptional temporal frequency along with considerable spatial accuracy. For Vegetation data, the dataset has moderate temporal frequency along with exceptional spatial accuracy. 13 points from each grid were sampled as mean was not deemed an ideal measure for the aggregate vegetation in a single grid.
- The project will develop and deliver data-driven intelligent wildfire risk prediction backed by machine learning. The system will use large-scale multi-dimensional data, including satellite images for remote sensing data (vegetation and terrain), weather sensor data, power line data and fire history data.

Research paper in [60] presents a model for predicting the scale of forest wildfires of Alberta, Canada. This prediction model uses wildfire and meteorological data for Alberta, Canada. Taking the meteorological factors as input values, a backpropagation neural network (BPNN), a recurrent neural network (RNN), and long short-term memory (LSTM) were implemented to establish prediction models. From the above-cited research, we decided to use Weighted Decision Trees, AdaBoost, Random Forest, Gradient boost and LSTM for our models considering the merits and demerits of various models used.

4.2.3 Model

The machine learning models utilized satellite data from Landsat, Weather data from the weather stations, Terrain, Powerlines, and fire history data for the study area near Monticello and Winters in California with the specified location coordinates (-122.1241886091637, -122.03418860916369, 38.49925978424475, 38.56925978424475). The model was developed based on the above parameters to predict the fire risk for each targeted grid on the given map. Table 4.1 shows the comparison of the existing solutions and the proposed solution. The criteria for selecting Machine Learning (ML) Algorithms for our Model are as follows.

- Accuracy of the model.
- Minimize False Negatives.
- Interpretability of the model.
- Algorithms with the least assumptions about the data.
- Scalability of the model.
- Infrastructural resources at hand.

Random Forest, Weighted Decision Trees, Gradient Boosting, Adaboost, Multi-layer perceptron (MLP) and Long-short Term Memory (LSTM) were the algorithms used for the individual models. The best models for each of the parameters were ensembled whereas a combined model was built after merging all the data using the best-performing algorithm with the highest accuracy and least false negatives.

Table 4.1. *Comparison of the existing systems and the proposed Machine learning model*

Paper ID	Focus Area	ML Model	Targeted Region
Proposed Approach	To predict location-based wildfire risk	Grid-Based Temporal Wildfire Risk Prediction Model using Random Forest, Weighted Decision Trees, Gradient Boosting, Adaboost, Multi-layer perceptron (MLP) and Long-short Term Memory (LSTM)	Grids in Monticello and Winters, California (enveloped between Davis and Napa)
[13]	To predict forest fires	Support Vector Machine (SVM)	Lebanon
[14]	To predict firs risk based on satellite data	Multi-layer perceptron (MLP)	Upper Seyhan Basin, Turkey
[16]	To predict the size of forest fires	iBK, Naïve Bayes (NB), J48,JRip,Logistic Regression,SVM,Adaboost,BagJ48,Random Forest, BNet	Slovenia
[15]	To assess human-caused wildfire occurrence	K-Nearest neighbors (KNN), Bagging Tree, ensemble method based on a two-layered machine learning model	Australia

4.3 Feature Engineering

4.3.1 Vegetation

We processed the fire history shapefiles for each fire into a data frame and intersected the geological dataframe with vegetation data to isolate the days with fire. During fire days, the target was set to True, else it was set to False.

Below is the snippet of the final merged data frame when the target (fire or no fire) was merged to the features (indices, dates and co-ordinates), along with the column details in Figures 4.3 and 4.4.

	diagonal4	fire_2018_COUNTY2	fire_2017_WINTER2	fire_2016_COLD2	fire_2015_WRAGG2	fire_2014_MONTICELLO2	target
(-122.12143860916372, 38.56200978424475)		False	False	False	False	False	False
(-122.12143860916372, 38.56200978424475)		False	False	False	False	False	False

Figure 4.3. *Snippet of the final dataframe*

At a more advanced stage of model building, the target was converted to numerical values 0 and 1 using an appropriate label encoding method.

```
gee_geodata.columns

Index(['left', 'top', 'right', 'bottom', 'id', 'geometry',
       'Centroid Longitude', 'Centroid Latitude', 'Start Date', 'End Date',
       'NDVI', 'EVI', 'NDWI', 'topLeft_coords', 'topRight_coords',
       'bottomLeft_coords', 'bottomRight_coords', 'centroid_coords',
       'midLeft_coords', 'midRight_coords', 'midTop_coords',
       'midBottom_coords', 'diagonal1', 'diagonal2', 'diagonal3', 'diagonal4',
       'fire_2018_COUNTY2', 'fire_2017_WINTER2', 'fire_2016_COLD2',
       'fire_2015_WRAGG2', 'fire_2014_MONTICELLO2', 'target'],
      dtype='object')
```

Figure 4.4. *Columns in the data frame*

Pair plot maps the relationship between the features and the target in Figures 4.5. The proportion of fire data against No fire data is shown in Figure 4.6.

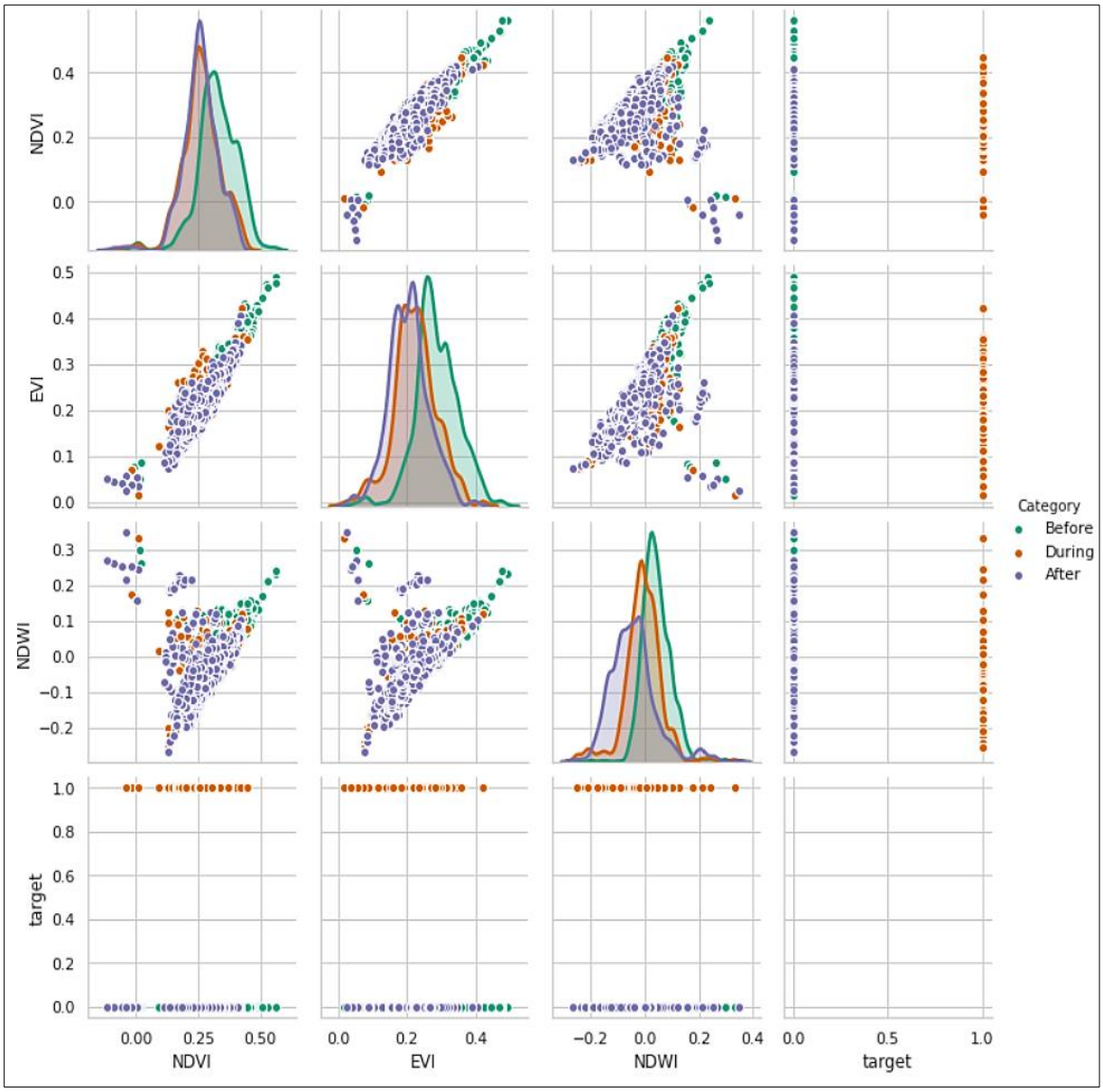


Figure 4.5. *Pair plot of the features and target before, during and after Fire*

The data is highly unbalanced with a disproportionately large number of 'No Fire' grids as shown in the pie chart in Figure 4.6. The number of affected grids are shown in Figure 4.7.

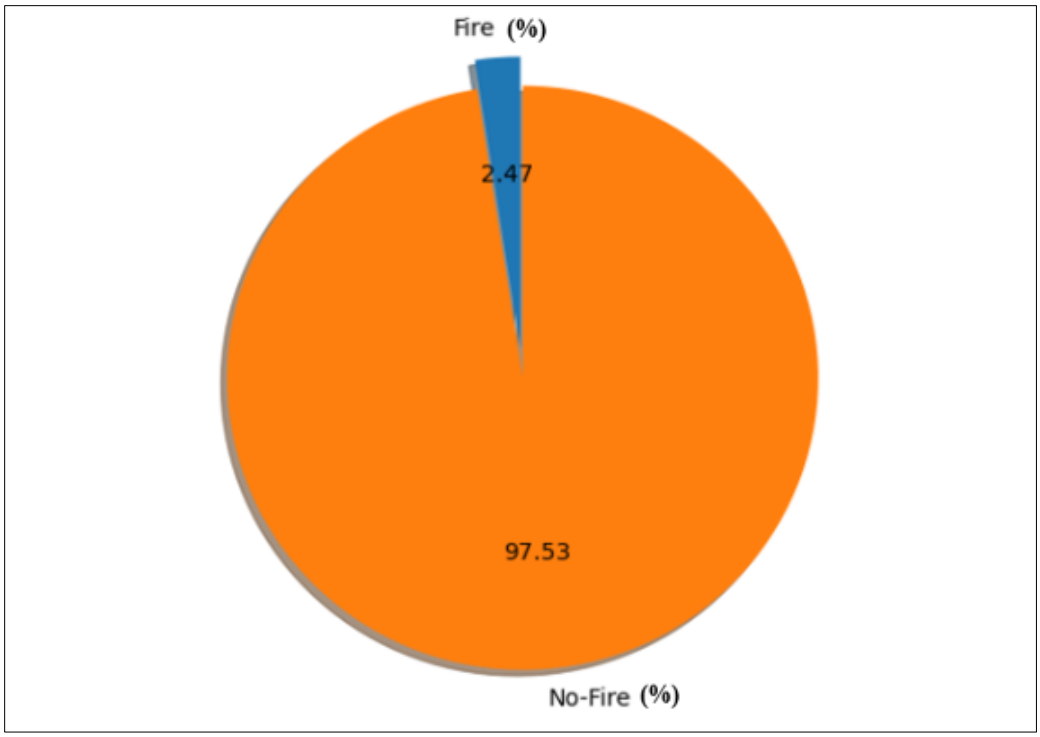


Figure 4.6. *Fire versus No-Fire targets in the dataset (in percentages)*

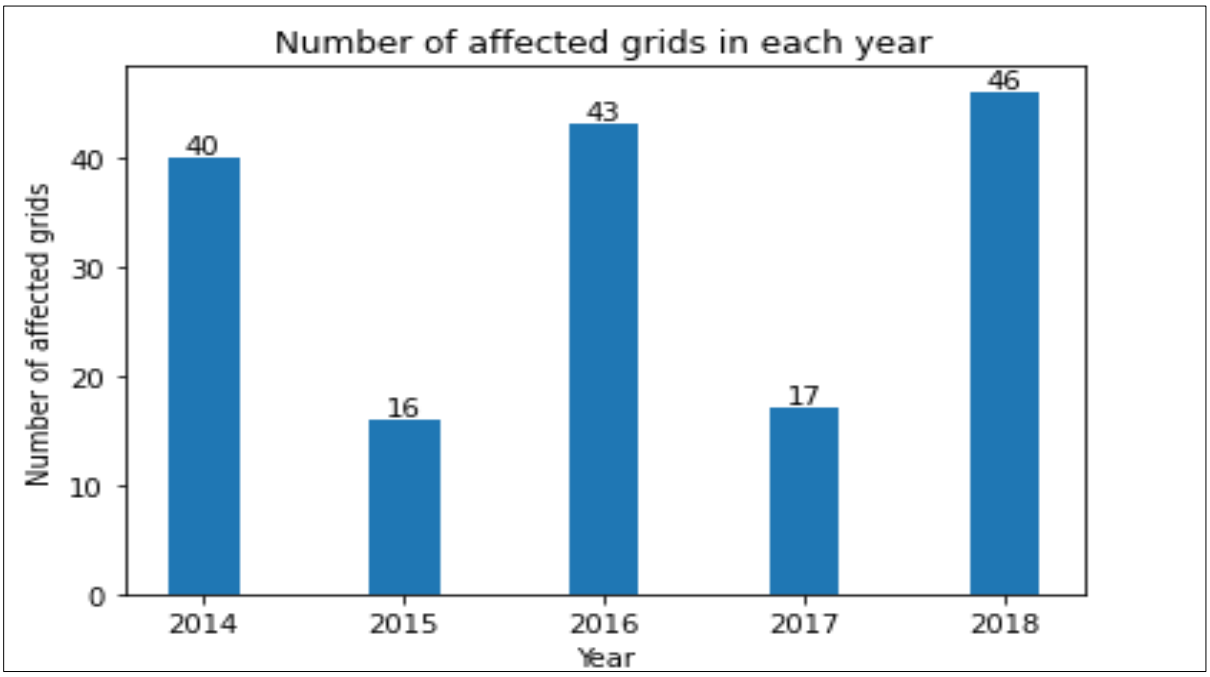


Figure 4.7. *Yearly count for fire-affected grids*

- As the data was highly unbalanced with an exceptionally high number of no-fire days, we used Synthetic Minority Oversampling Technique (SMOTE) [66] to oversample and generate samples of the minority class.
- Null value imputation discussed in Data cleansing is a necessary component of feature engineering for missing gaps in feature indices. Using linear interpolation, a line was fit between the available points and missing points in between were interpolated.
- Columns were dropped, filtered before label encoding the categorical target value to binary values 0 and 1 for fire and no-fire days.

Figure 4.8 shows the distribution of indices in the final dataset with aggregated index values from 13 sample points, formerly validated with the 5 points and 1-point dataset.

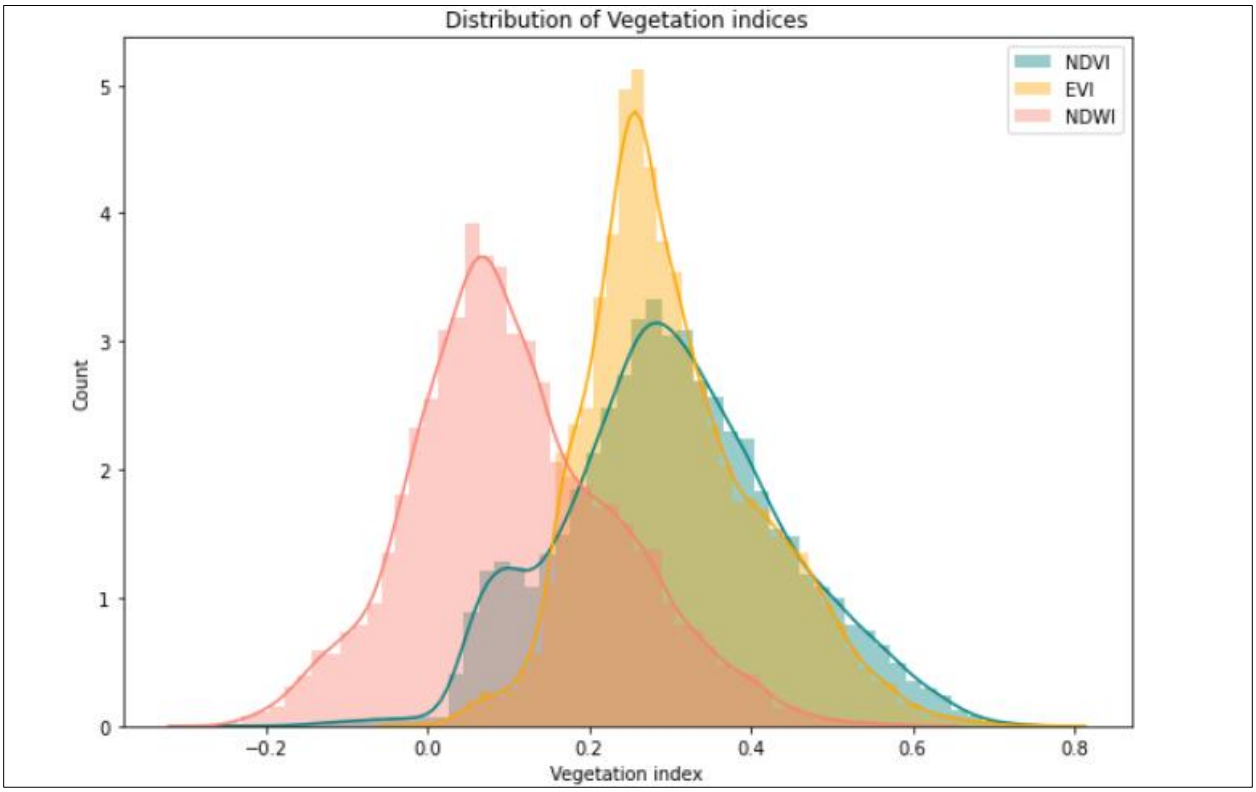


Figure 4.8. *NDVI, EVI and NDWI data distribution in the final dataset*

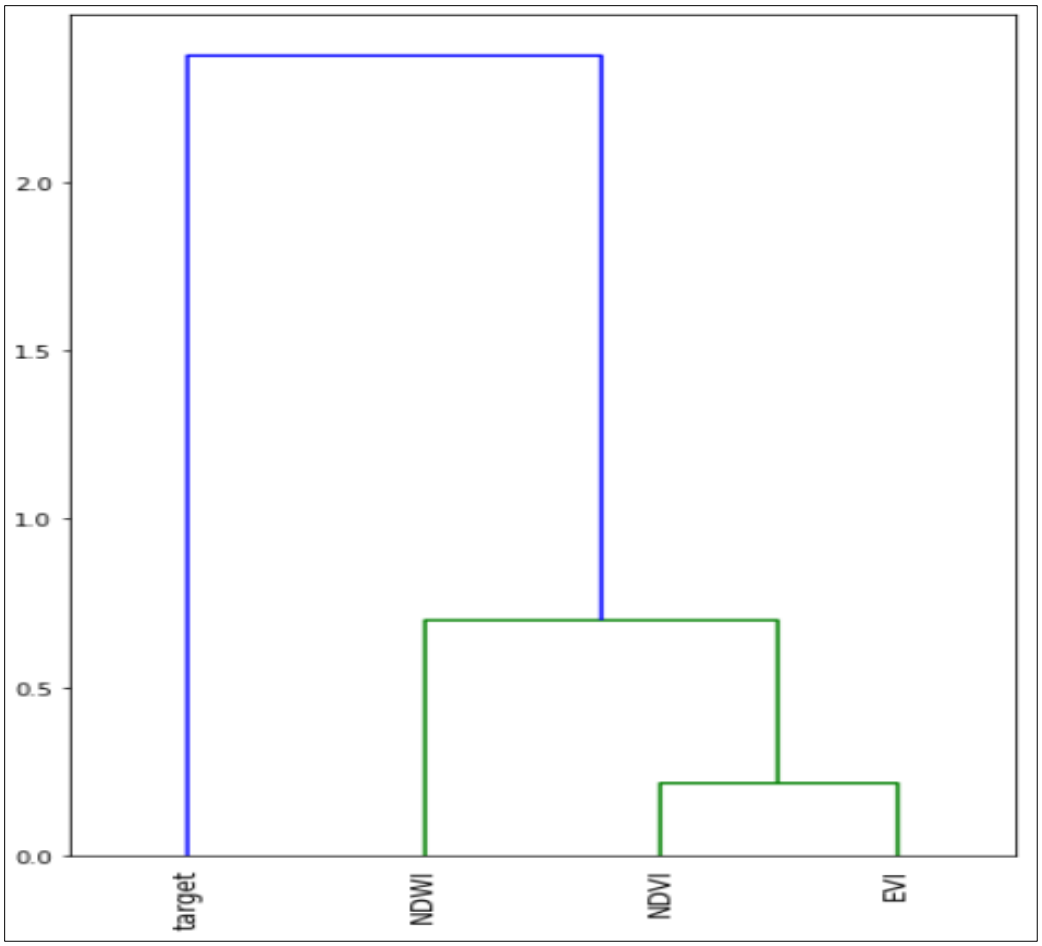


Figure 4.9. *Dendrogram of features and target in Vegetation dataset*

Figure 4.9 shows a dendrogram of the features and the target in the vegetation dataset. Figure 4.10 shows the negative correlation between good vegetation and water content (as indicated by healthy NDVI, EVI and NDWI values) and target values. To display the fire probability of each of these indices, we choose a small subset of data before and after the fire and set the data points before fire to 1 (fire expected) and after fire to 0 (No Fire). The heatmap of features and the modified Fire/No Fire parameter is shown in Figure 4.11. NDVI and EVI are reliable indicators of the likelihood of fire.

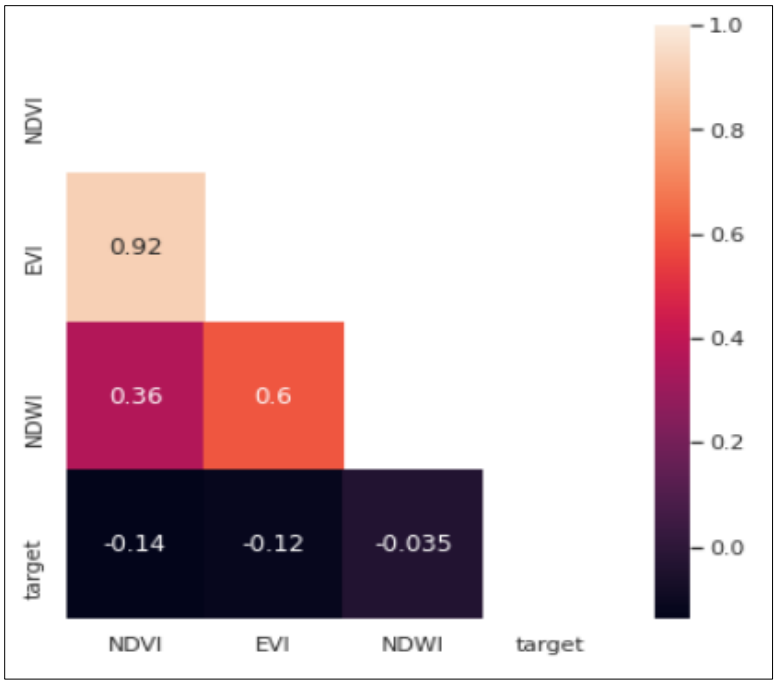


Figure 4.10. *Correlation heatmap of features and target in Vegetation Dataset*

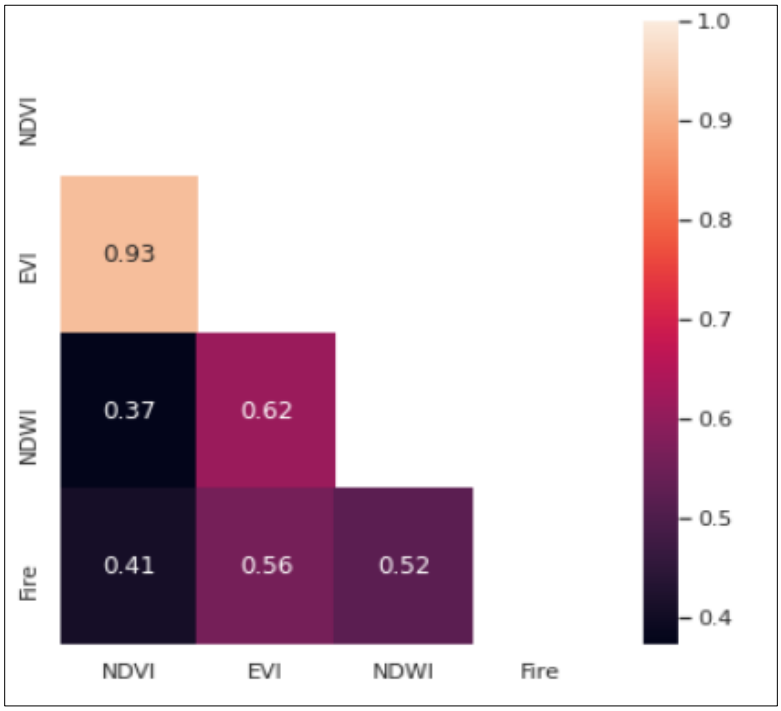


Figure 4.11. *Correlation heatmaps with subset of Vegetation Dataset*

	left	top	right	bottom	id	Centroid Longitude	Centroid Latitude	NDVI	EVI	NDWI
count	7749.000000	7749.000000	7749.000000	7749.000000	7749.000000	7749.000000	7749.000000	7749.000000	7749.000000	7749.000000
mean	-122.084189	38.539260	-122.074189	38.529260	32.000000	-122.079189	38.534260	0.304432	0.302413	0.107344
std	0.025822	0.020001	0.025822	0.020001	18.185416	0.025822	0.020001	0.148405	0.111154	0.135167
min	-122.124189	38.509260	-122.114189	38.499260	1.000000	-122.119189	38.504260	-0.200841	-0.013835	-0.264864
25%	-122.104189	38.519260	-122.094189	38.509260	16.000000	-122.099189	38.514260	0.206271	0.227681	0.016323
50%	-122.084189	38.539260	-122.074189	38.529260	32.000000	-122.079189	38.534260	0.299359	0.281683	0.087096
75%	-122.064189	38.559260	-122.054189	38.549260	48.000000	-122.059189	38.554260	0.398841	0.368185	0.191471
max	-122.044189	38.569260	-122.034189	38.559260	63.000000	-122.039189	38.564260	0.742485	0.767890	0.600332

Figure 4.12. *Vegetation data Statistics*

Vegetation data statistics are listed in Figure 4.12. The mean NDVI is 0.304 with a standard deviation of 0.148, whereas the mean EVI is 0.302 with a standard deviation of 0.111. Mean NDWI is 0.107 with a standard deviation of 0.135.

4.3.2. Terrain

We selected the following three features from the terrain DEM maps as the input to the machine learning algorithm,

1) **Slope:** Slope represents the rate of change of elevation for each digital elevation model (DEM) cell. It's the first derivative of a DEM. By default, the slope appears as a grayscale image. You can add the Colormap function to specify a color scheme. The inclination of the slope can be output as either a value in degrees or percent rise. There are three options:

   - Degree - The inclination of the slope is calculated in degrees. The values range from 0 to 90.
   - Scaled - The inclination of the slope is calculated the same as degrees, but the z-factor is adjusted for scale.

- Percent Rise - The inclination of the slope is output as percentage values. The values range from 0 to essentially infinity. A flat surface is 0 percent and a 45-degree surface is 100 percent, and as the surface becomes more vertical, the percent rise becomes increasingly larger. Figure 4.13 shows the slope of our study area.

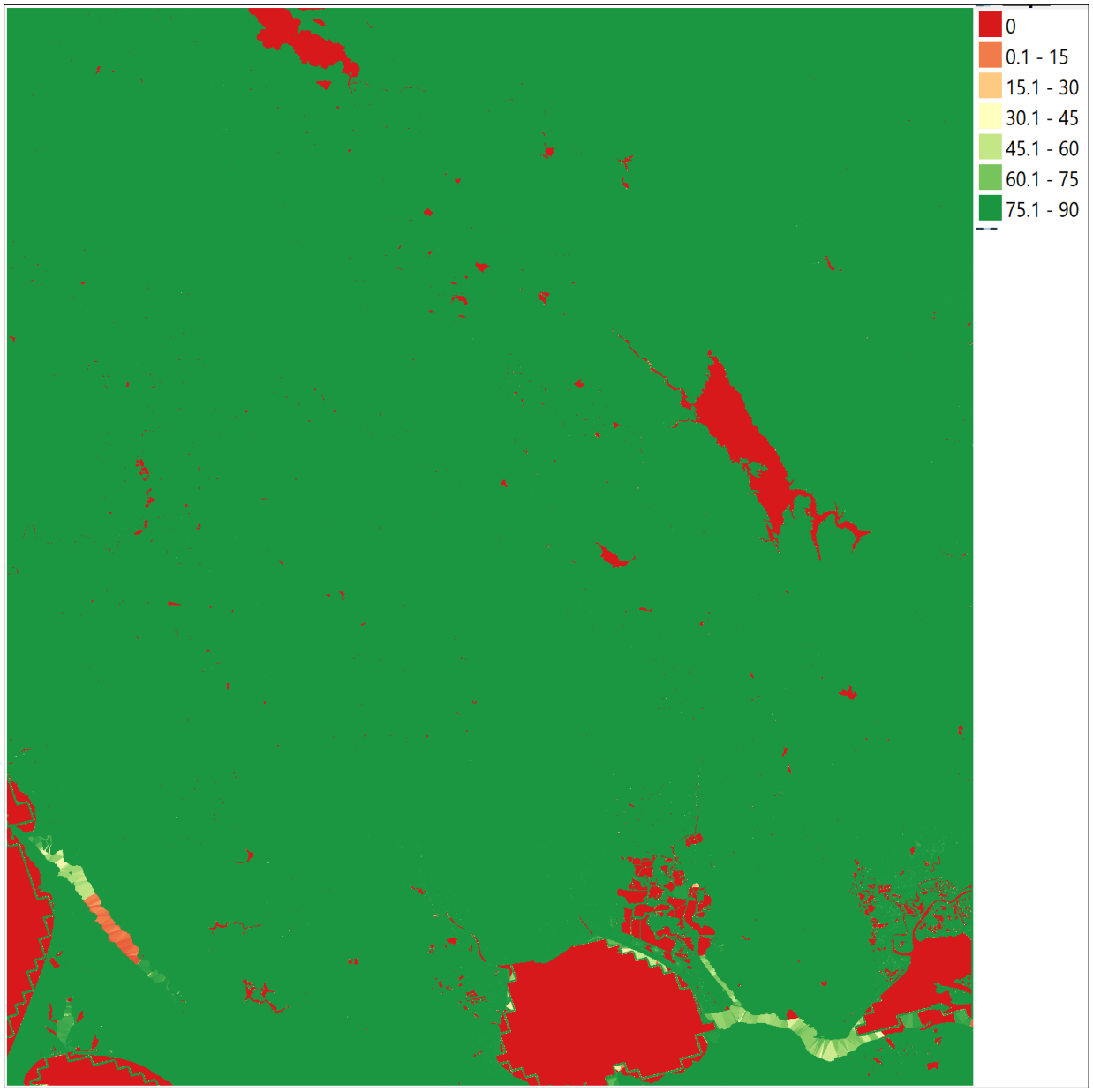


Figure 4.13. *Slope of our study area*

2) **Hill Shade:** To get a better look at the terrain, it is possible to calculate a hill shade, which is a raster that maps the terrain using light and shadow. A hill shade can provide very useful information about the sunlight at a given time of day. But it can also be used for aesthetic purposes, to make the map look better. Figure 4.14 shows the hill shade of our study area.

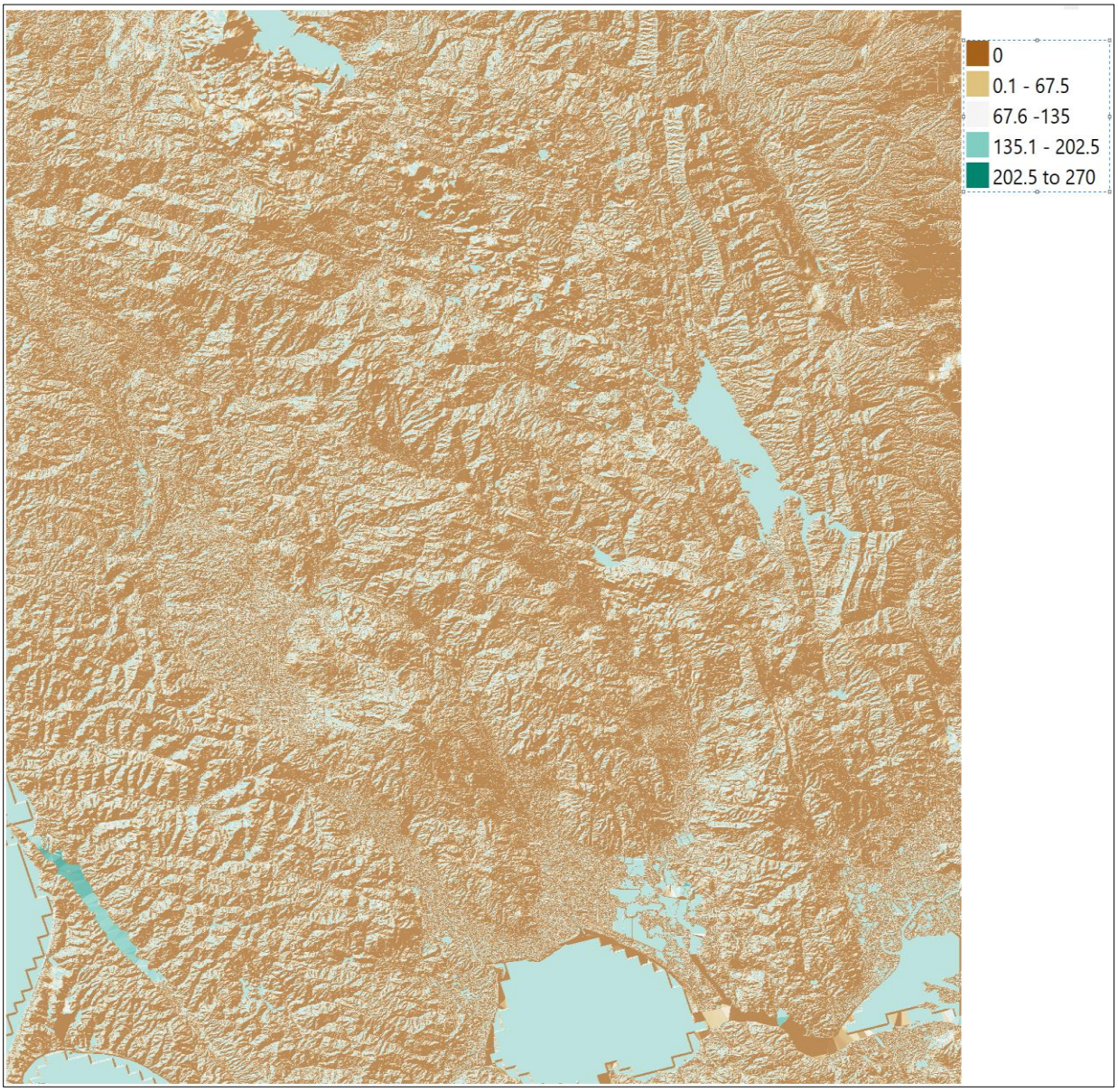


Figure 4.14. *Hill shade of study area*

3) **Aspect:** Aspect is the compass direction that a slope faces. Figure 4.15 shows the aspect calculated from the DEM map.

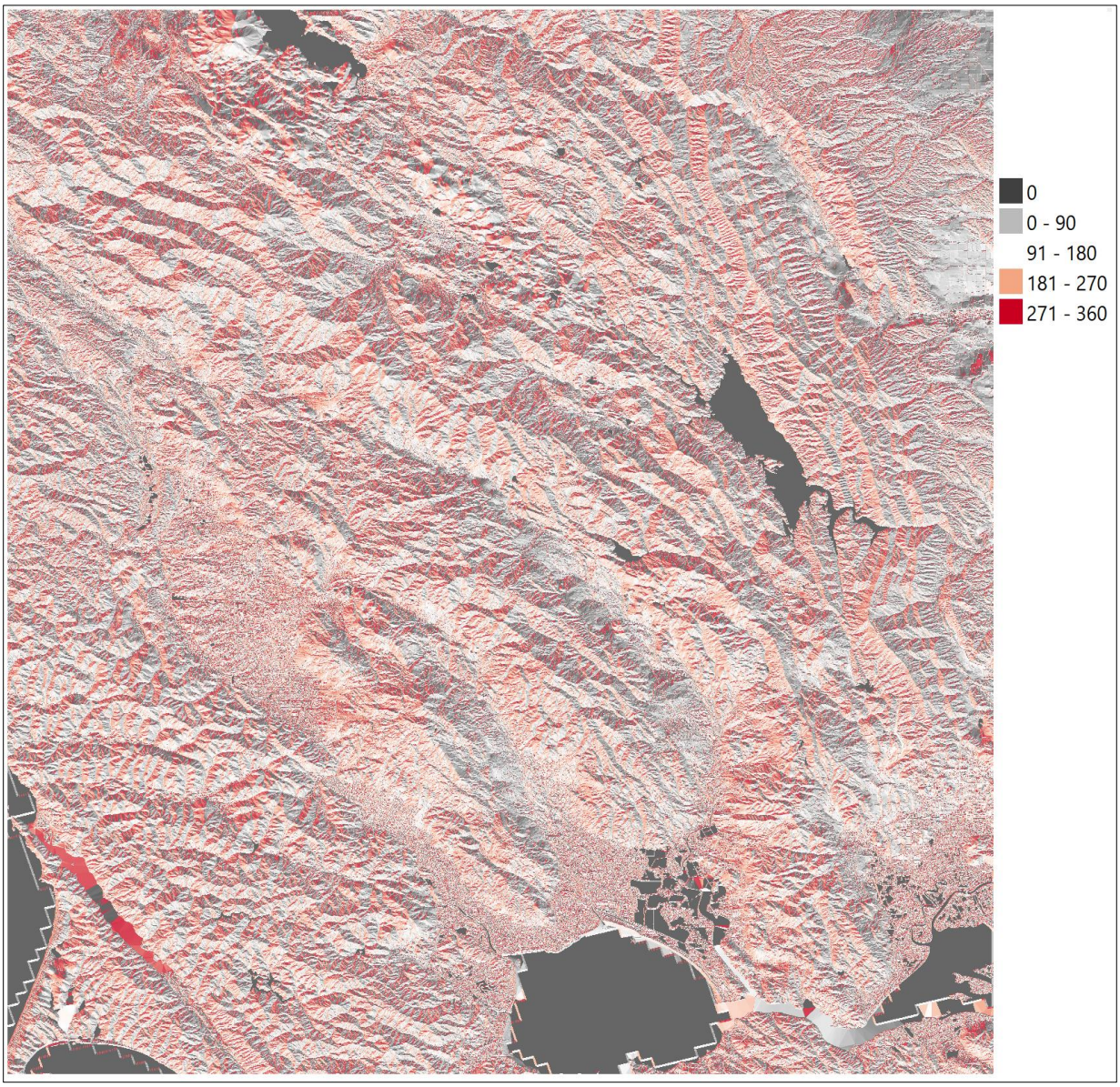


Figure 4.15. *Aspect of Study area*

After applying 1 x 1 km grid statistics of the slope, hill shade and aspect are calculated for each grid on the map. Overall there are 63 grids for our study area. Figure 4.16 shows statistics terrain features in each grid with respective location coordinates.

	left	top	right	bottom	id	slope_mean	hillshade_mean	aspect_mean
0	-122.58104	38.39748	-122.57204	38.38848	351	89.993114	36.770007	142.761488
1	-122.58104	38.38848	-122.57204	38.37948	352	89.991473	45.377761	137.358713
2	-122.58104	38.41548	-122.57204	38.40648	349	89.996934	37.346264	144.130817
3	-122.58104	38.40648	-122.57204	38.39748	350	89.994110	23.676267	132.775934
4	-122.58104	38.43348	-122.57204	38.42448	347	89.965223	21.990217	168.291628

Figure 4.16. *Statistics of Terrain Parameters in each grid*

4.3.3. Weather

- For feature selection, out of the 22 features on hourly weather information, we chose the 4 features: dry bulb temperature, relative humidity, wind speed and hourly precipitation. Figure 4.17 shows the features selected for weather data.

DATE	HourlyDryBulbTemperature	HourlyRelativeHumidity	HourlyWindSpeed	HourlyPrecipitation	left	top	right	bottom	id
2015-07-22 17:53:00	76.0	39.0	9.0	0.0	-122.054189	38.55926	-122.044189	38.54926	51
2015-07-22 12:53:00	83.0	37.0	14.0	0.0	-122.104189	38.50926	-122.094189	38.49926	21
2015-07-22 02:53:00	65.0	66.0	14.0	0.0	-122.074189	38.54926	-122.064189	38.53926	38
2015-07-22 20:53:00	64.0	65.0	18.0	0.0	-122.074189	38.50926	-122.064189	38.49926	42
2015-07-22 00:53:00	67.0	61.0	10.0	0.0	-122.114189	38.56926	-122.104189	38.55926	8

Figure 4.17. *Selected weather features*

- Missing values: After examination, we discovered that there was a lot of missing data. We plot the stations on the y-axis and data availability on the x-axis, data coverage from 2016 to 2018 is shown in Figures 4.18, 4.19 and 4.20.

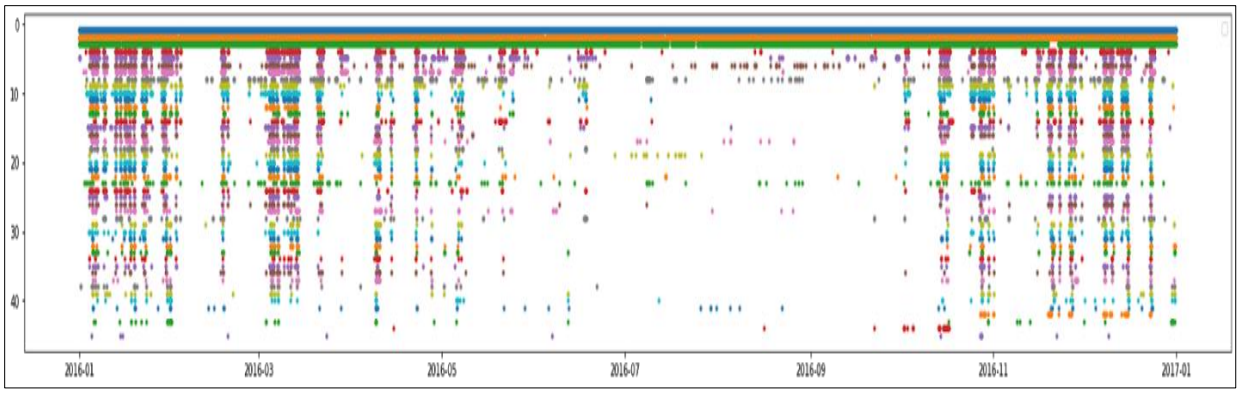


Figure 4.18. *Weather data coverage for 2016*

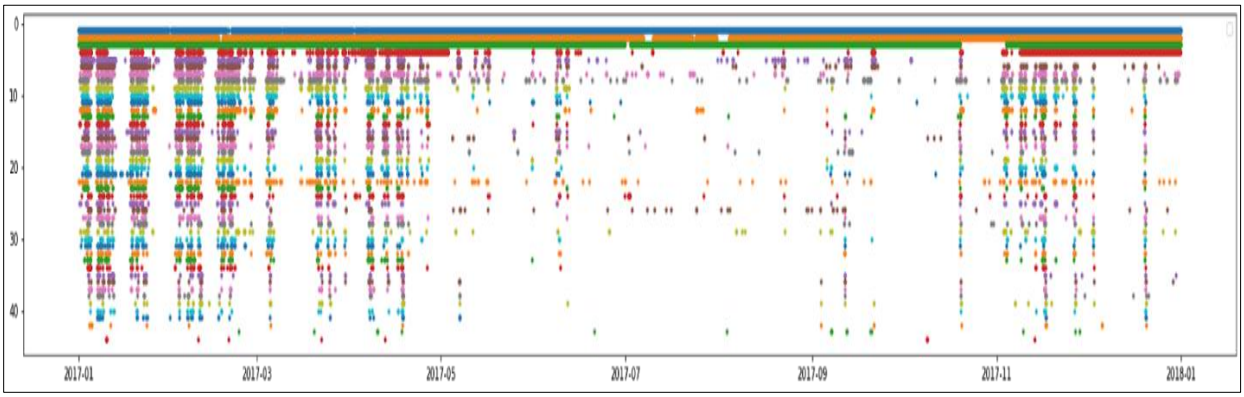


Figure 4.19. *Weather data coverage for 2017*

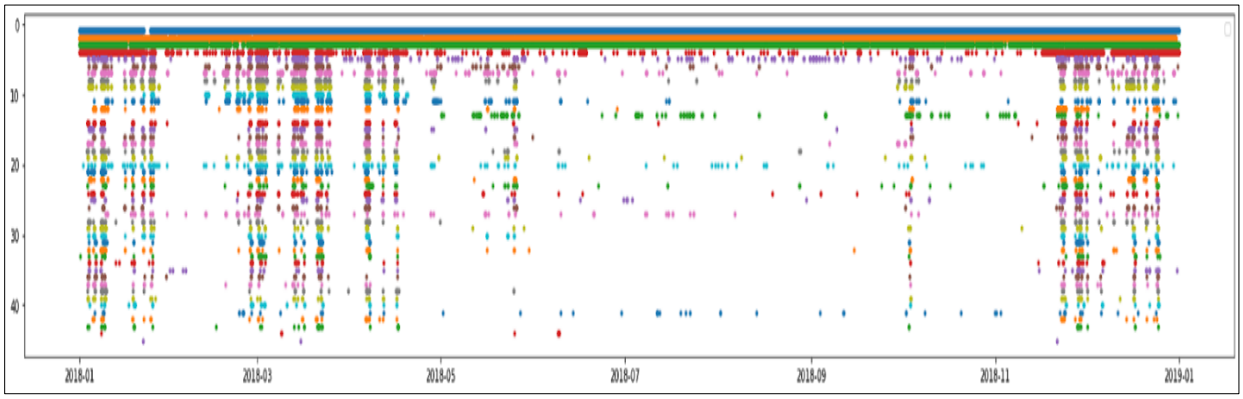


Figure 4.20. *Weather data coverage for 2018*

The formerly defined scenarios along with their corresponding null-filling techniques are as follows.

- For a single station, if the previous and next hour record is present for the missing value, we took the mean of the previous and next hour record and filled the missing value.

- For a single station, if there's consecutive data missing for up to 10 days, we took the previous years' data of the same date.
- If there's a multiple station failure across a relatively long time period, we used machine learning methods to simulate the missing data.

To fill the missing values for better performance, we applied different null-filling techniques for different scenarios.

- Numerical Imputation: Our dataset contained non-numerical marker values such as 'T' to indicate trace amounts of precipitation. We converted all records with 'T' into 0.0001.
- Removing outliers: For each feature, there was a theoretical domain. We removed all outliers outside this range.
- Normalization: We normalized our training dataset before training.

4.3.4. Fire History

Feature selection: Out of the 19 features from our dataset, we chose the 'YEAR_', 'ALARM_DATE', 'CONT_DATE', 'geometry', 'CAUSE', 'REPORT_AC', 'GIS_ACRES' for statistical analysis. Figure 4.21 shows the fire history data with 1 x 1 km grids.

A total of 63 grids are available for our area of study. After integrating the individual shapefiles for the fires, we created five new fields for fires in each year and integrated the information into the Target field. If there was a Fire at a certain timeframe, it was set to 1 ('Fire'). Otherwise, it was set to 0 ('No Fire').

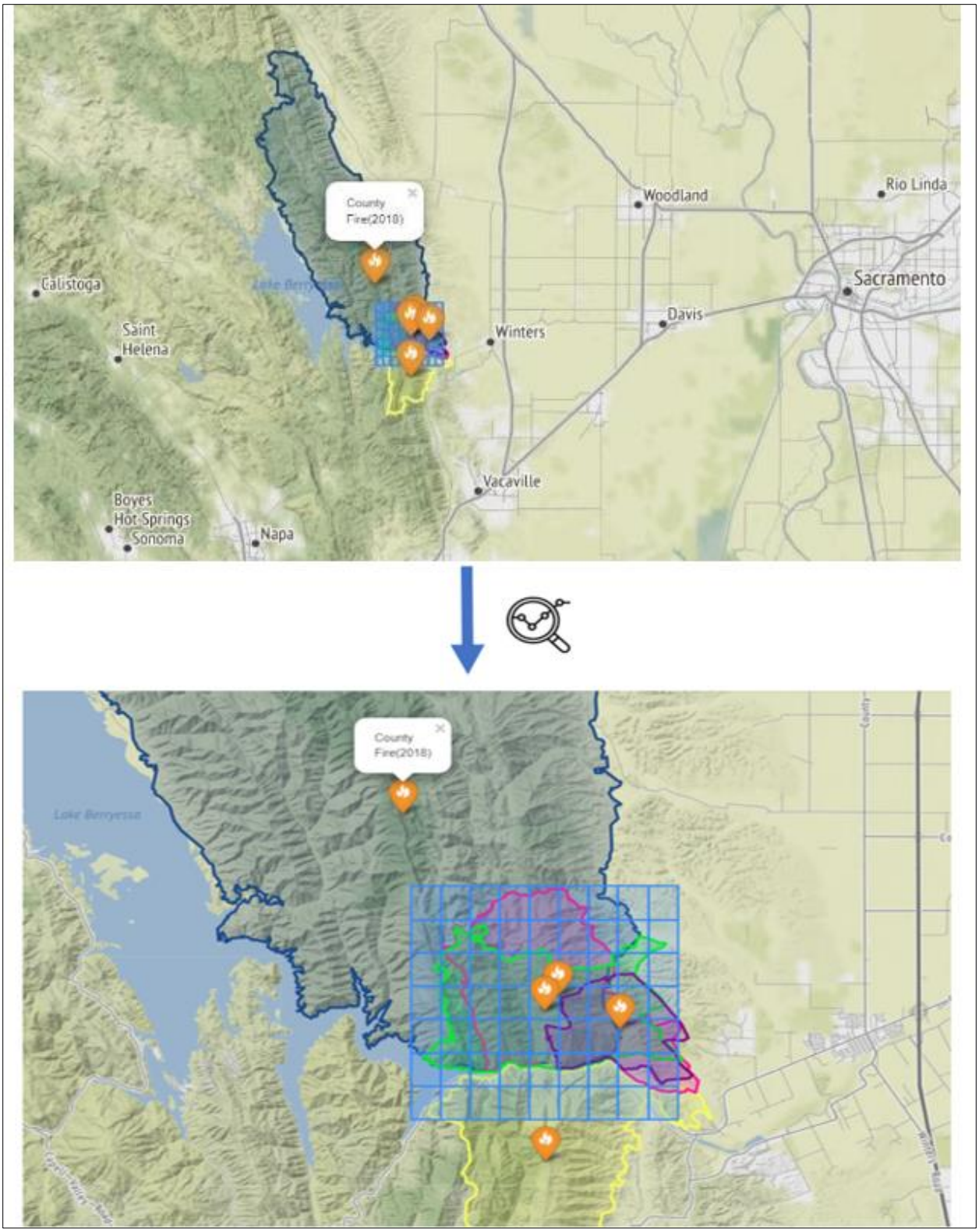


Figure 4.21. *Fire history data and the Grids*

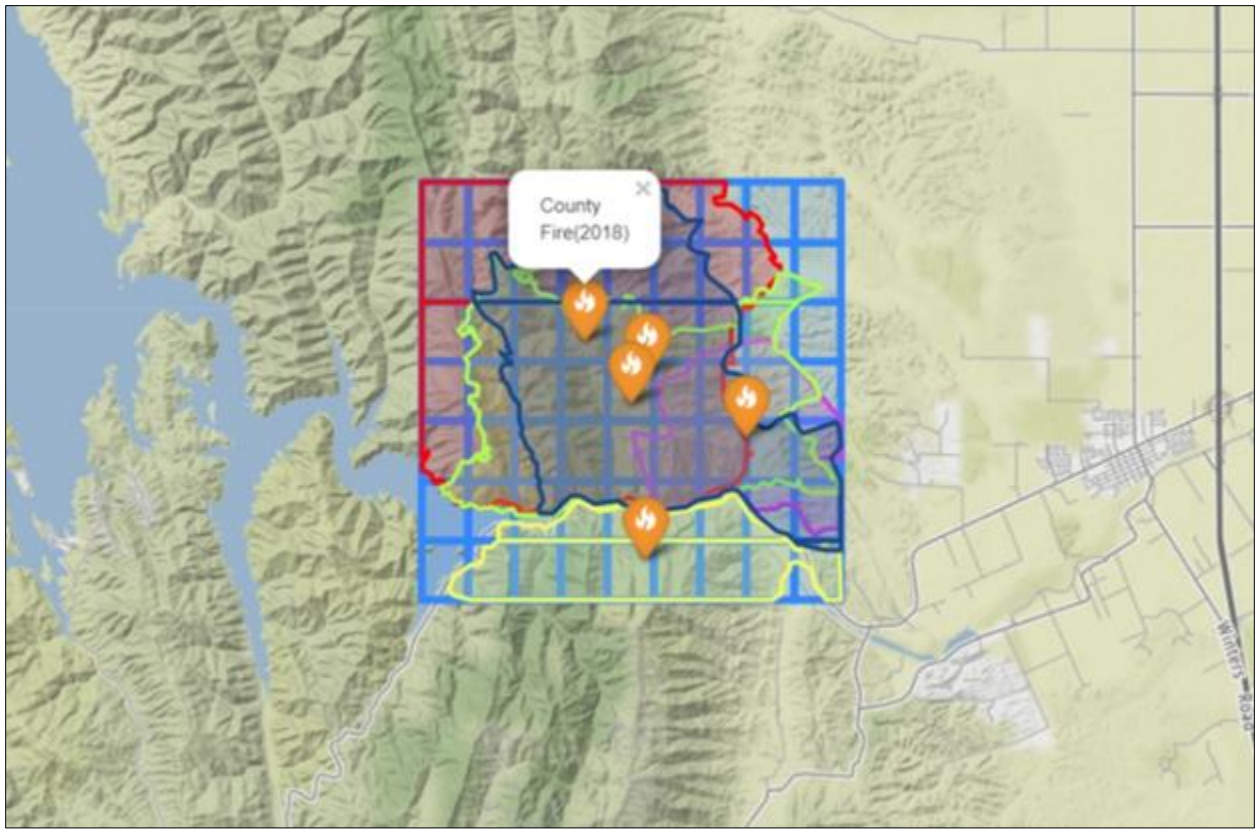


Figure 4.22. *Integration of Study Area and Fire history*

Figure 4.22 shows the final dataset that was created by intersecting the fire history data with the grids. Only the common areas were retained in the dataset.

4.3.5. Powerlines

- Feature selection: Source data has undergone thorough quality assurance procedures from the original provider. We chose the geographic distribution of power lines for our model.
- Feature extraction: We intersected the powerline shapefile with our grid and got the IDs of the powerline crossing each box in the grid. Further, we performed statistical analysis to study the relationship between powerlines and wildfires in our study area. Figure 4.23 shows the powerline details in each grid. Figure 4.24 shows the single major powerline across the grids in our study area.

left	top	right	bottom	id	kV	Status	Circuit	Length_Mil	Length_Fee
-122.074189	38.50926	-122.064189	38.49926	42	0.0	Not Operating	Other	0	0.00000
-122.124189	38.51926	-122.114189	38.50926	6	115.0	Operational	Single	25	133951.91011
-122.114189	38.50926	-122.104189	38.49926	14	0.0	Not Operating	Other	0	0.00000
-122.064189	38.51926	-122.054189	38.50926	48	115.0	Operational	Single	25	133951.91011
-122.054189	38.53926	-122.044189	38.52926	53	0.0	Not Operating	Other	0	0.00000

Figure 4.23. *Powerline features in each grid*

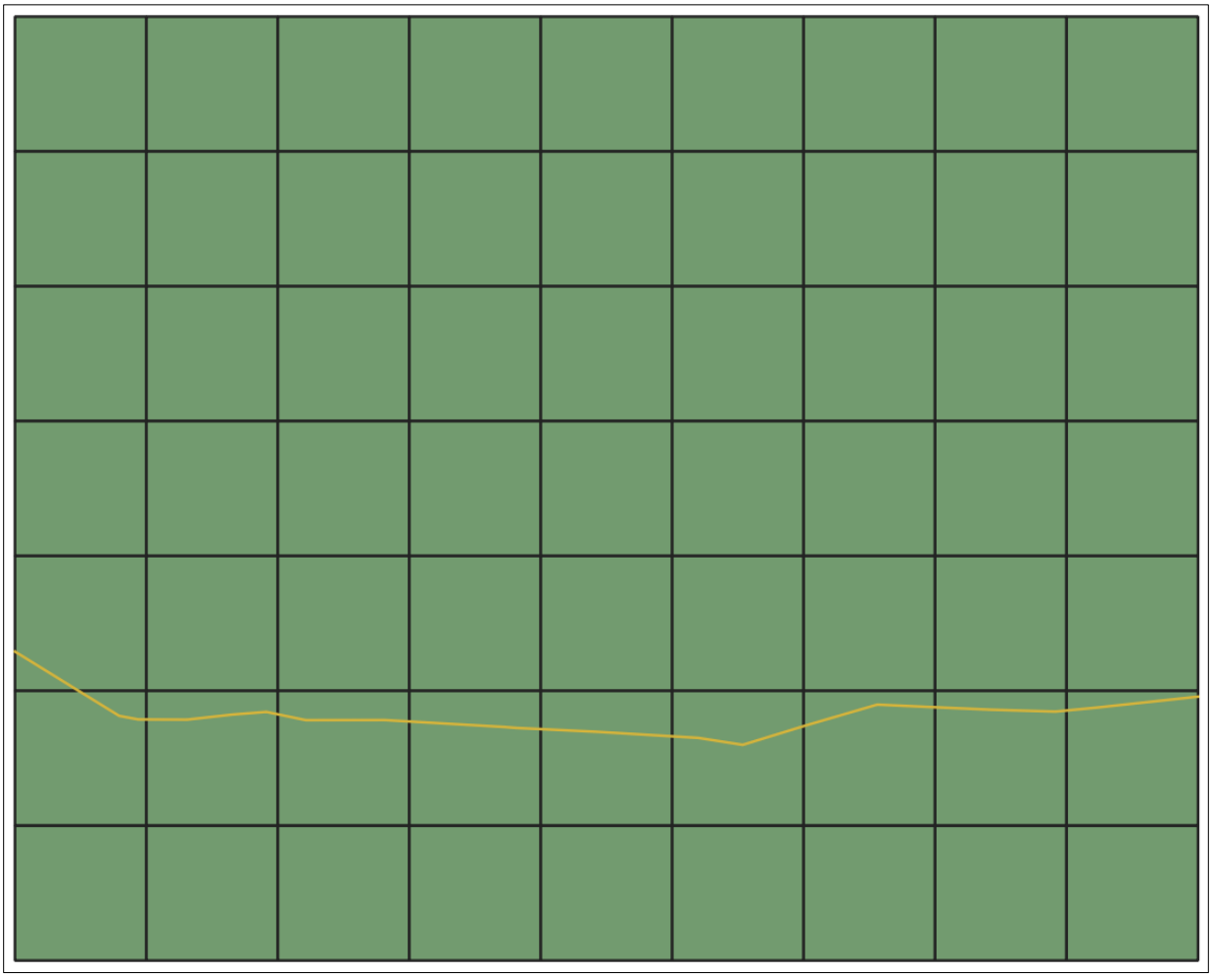


Figure 4.24. *Powerline crossing through our study area*

4.4 Select and Justify Solution Model

Addressing the practical wildfire challenges posed by the frequent fluctuations in weather forecast and location-based risk prediction, we need a model that deals with both the spatial and temporal dimensions of our data. Hence, we used machine learning models to predict fire risk in our study area based on satellite data.

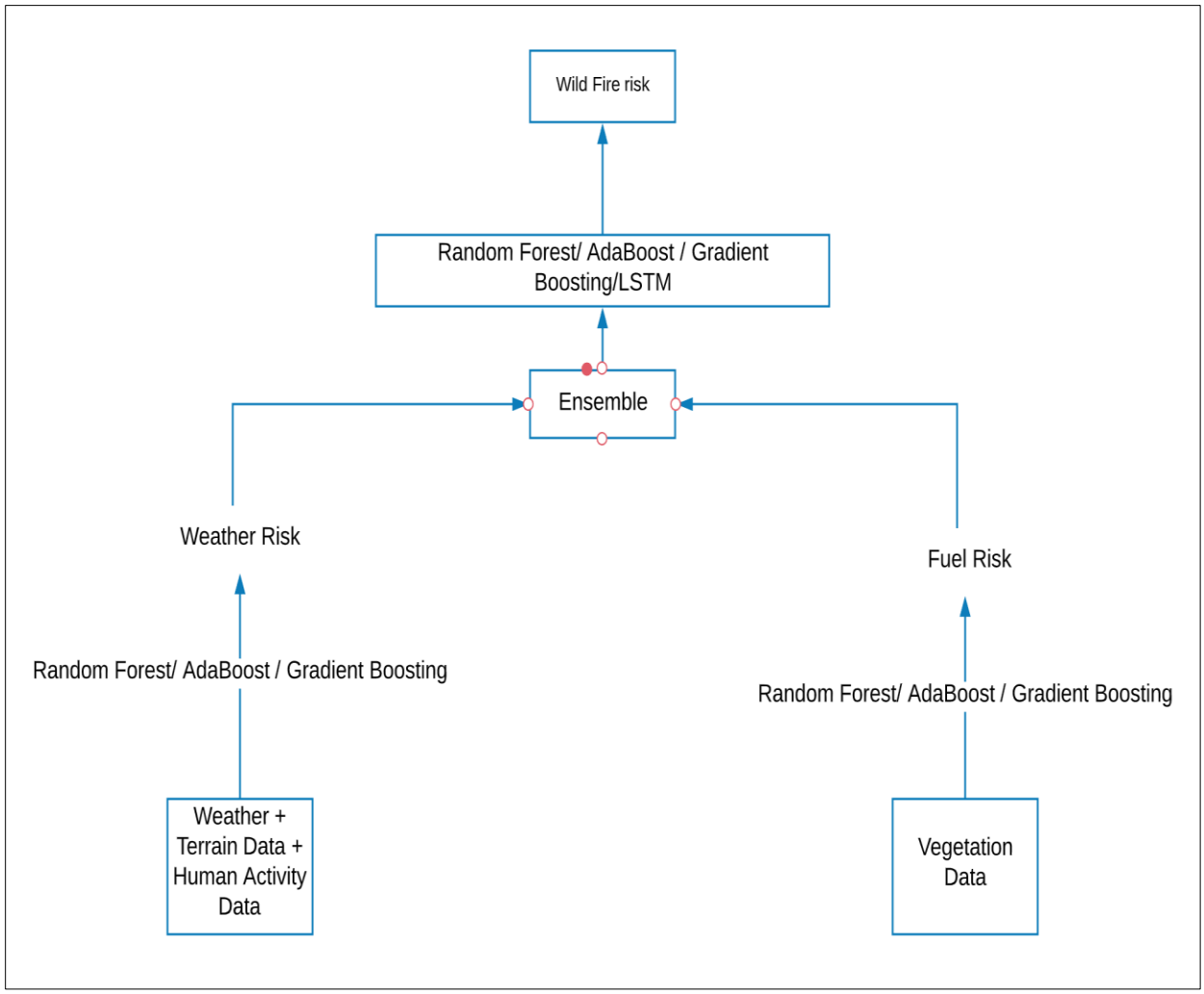


Figure 4.25. *High-level architecture of the ensemble model*

For addressing this challenge, we need to enlist innovative data-driven machine learning models that run on comprehensive datasets with multiple parameters including location-based weather, terrain, vegetation, and powerlines data, along with the fire history data. Figures 4.1 and

4.2 presents our proposed temporal machine learning solutions that ingests grid-based values. Unlike other existing models, this model is an integrated model powered by a suitable algorithm such as Adaboost, Decision trees, Gradient descent, Multi-layered perceptron, Random Forest Tree (RF) and Long Short-Term Memory (LSTM) to address convoluted location-specific wildfire risk prediction. Finally, we created a holistic model that explores the relationship between the parameters and the fire occurrences, including the pre-fire conditions. Figures 4.25 and Figure 4.26 shows the high-level architecture of the ensemble and combined models. Ensemble model stacks the best models whereas the combined model runs on a combined dataset using the best-performing algorithm.

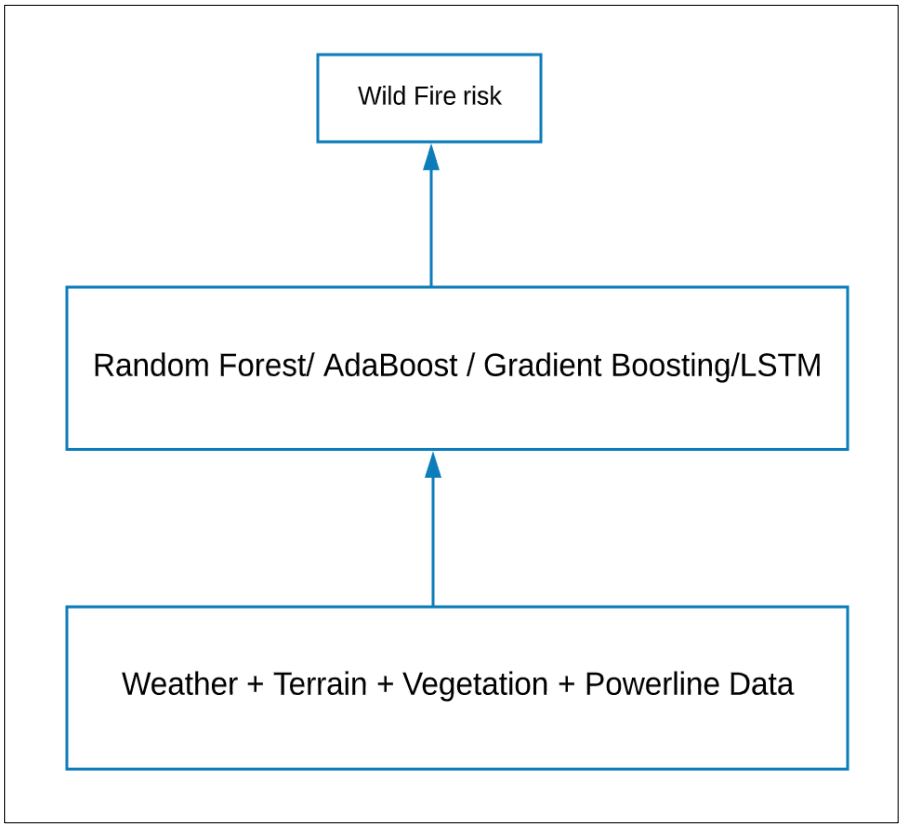


Figure 4.26. *High-level architecture of the combined model*

As we have multiple layers of complex data sources, we decided to extract grid-specific data for the parameters in our study area. This makes it easier to integrate the datasets at a later point. Few of the algorithms in the initial plan are as follows.

- Random Forest [62]: Random Forest consists of a large number of individual decision trees that operate as an ensemble. It improves on bagging by decorrelating the trees with the introduction of splitting on a random subset of features. Each individual tree in the random forest provides a class prediction and the class with the most votes becomes our models prediction. As it can handle binary, categorical and numerical features, random forest algorithm can be used for regression and classification tasks.
- LSTM [62]: Long Short-Term Memory (LSTM) is one of the most widely used recurrent structures in sequential modeling. It overcomes the fundamental problem of recurrent networks that do not capture long-term dependencies in a sequence. They are powerful enough to learn the most important past behaviors and understand whether or not those past behaviors are important features in making future predictions. LSTMs may benefit from transfer learning techniques even when applied to standard classification problems.
- Adaboost [68] : An Adaboost classifier is a meta-estimator that begins by fitting a classifier on the original dataset and then fits additional copies of the classifier on the same dataset but where the weights of incorrectly classified instances are adjusted such that subsequent classifiers focus more on difficult cases. It combines multiple classifiers to increase the accuracy of classifiers. Any machine learning algorithm can be used as a base classifier if it accepts weights on the training set. The methodology is explained below.
    - The classifier is trained interactively on various weighted training examples.

    - In each iteration, it tries to provide an excellent fit for these training set by minimizing the training error.
- Gradient Boosting [69]: Boosting is a technique where the error of one predictor is passed as input to the next in a sequential manner. Gradient Boosting uses a gradient descent procedure to minimize the log loss for each subsequent classification tree added one at a time that, on their own. These classification trees are weak decision models. It builds an additive model in a forward stage-wise fashion; it allows for the optimization of arbitrary differentiable loss functions.

Weighted decision trees and MLP were included in model algorithm comparison.

4.5 Justify Solution Model

Considering various factors [70], we selected our solution model. Some of them are specified here.

- Interpretability: Algorithms like Random forest, Adaboost and Gradient generate models that are easily interpretable. In certain situations, we may want to understand what our model is doing to the data in order to make better decisions. Hence, interpretability is one of the key factors to be considered while selecting a model. Weighted Decision Trees and SVM lag in this regard since they seem like a “black box” to the user.
- Model Assumptions: Algorithms like Weighted Decision trees assumes that the data is linearly separable whereas Decision trees do not assume data to be linearly separable. Instead, it assumes those decision boundaries lie parallel to the coordinate axes. Random forests assume that averaging outperformance on multiple random classifiers is safer.
- Infrastructural Resources at hand: Certain algorithms tend to be more CPU and/or memory intensive than the others. This can be attributed to a large number of computations they do

or the number of intermediate results that they try to store. In such cases, an algorithm that has minimum storage requirements is preferred.

- Nature of data: This is the most important factor for deciding on which algorithm to deploy. When there are a lot of features compared to data points, decision trees are often not a good choice. Especially for unbalanced datasets, the foremost aim is to maximize the correct classification of the lower frequency class. In such cases, overall accuracy is not the topmost priority. Hence, algorithms like Weighted Decision Trees that tend to maximize the overall accuracy cannot be used in such circumstances, while decision trees can be due to their ability to incorporate class bias.
- Evaluation metrics: High accuracy, lower false negatives, scalability and performance of the models as well as the training and validation scores served as prime indicators of model efficiency. These metrics are plotted and visualized for an in-depth comparison in the coming sections.

The tools and technologies used for the project include:

- **QGIS –** Quantum Geographic Information System (QGIS) is a desktop software that allows its users to analyze and edit spatial information, in addition to composing and exporting graphical maps. QGIS supports both raster (geotiff format) and vector (shapefile) layers.
- **GEE –** Google earth engine is a multi-petabyte cloud-based data catalog of satellite data from multiple sources, available for free for public use, for non-commercial purposes. It allows academics and researchers to map the trends around the world using its analytical capabilities.

- **Aeronautical Reconnaissance Coverage Geographic Information System (ArcGIS) –** It is used for creating and using maps, compiling geographic data, analyzing mapped information, sharing and discovering geographic information, and managing geographic information in a database.
- **FileZilla –** It is the most popular File transfer protocol (FTP) client. It is mainly used to upload and download files from any web hosting server.
- **Python libraries** (such as ee, geopandas, folium, bokeh) **–** These libraries are installed and used for accessing and working with geospatial data files. Advanced python visualization packages such as folium and bokeh were used to generate leaflet map visualizations with map overlap, cascading style sheets (CSS), hypertext markup language (HTML) and popups.
- **Jupyter Notebook in Google Collaboratory (colab) –** It is a free cloud service used for Python programming language. We can also develop deep learning applications using popular libraries such as Keras, TensorFlow, PyTorch, and OpenCV which are preinstalled in colab.
- **Amazon Web Services (AWS)** – It is a secure cloud services platform that offers many functionalities including computing power, database storage, content delivery. It uses databases like MySQL, PostgreSQL, Oracle or SQL Server to store information.
- **Tableau** – It is a popular and powerful data visualization tool. It can process shapefiles with spatial data and plot the coordinates on an interactive map.

5. Model Development and Presentation

5.1 Model Building and Training

Due to lack of reliable weather data in 2014, we exempted that particular year from our model experimentation. With the four years of data available, we experimented with subsets of data for fitting and building the model as shown in Table 5.1.

Table 5.1. *Subsets of Data used*

Data Used	Positive target data	Negative Target data
Type I	Data during fire	7 Days before fire
Type II	Data on fire start date	Data before the fire
Type III	Data before fire	Data after fire
Type IV	Data on fire start date, excluding no-fire grids	Data before the fire, excluding no-fire grids

Combination of different years from the above types along with stratified samples were considered in model building after statistically analyzing the samples. We randomly split the dataset into 80% training data and 20% testing data, although we had initially tested with 2018 as the test year. After weather data for 2014 was deemed inconsistent and unreliable, we shifted to the 4:1 random train-test split methodology. Among the datasets, Type I and Type II datasets are imbalanced. To overcome this problem, we used Synthetic Minority Oversampling Technique (SMOTE) Oversampling [66] for Imbalanced Classification in Python. SMOTE works by generating synthetic samples that lie close to the feature space. By fitting a line among the samples

in the feature space, the technique captures new samples that lie on that line. Figure 5.1 shows the code for generating SMOTE samples.

```
oversample = SMOTE()
X, y = oversample.fit_resample(X_1, y_1)
```

Figure 5.1. *SMOTE Oversampling*

5.1.1 Base model I - Weather, Terrain and Powerline

Figure 5.2 shows the data mapping high-level design. To create training sets, we divide the selected area to 1 x 1 km grid. After fitting weather data to each cell, we picked one day of weather data from each fire start day and another sample from a no-fire day to generate a balanced data. Thereafter, the fire data is combined with the no-fire data. We considered various combinations of years in train and test sets. The stratified dataset produced the best results for both training and testing purposes.

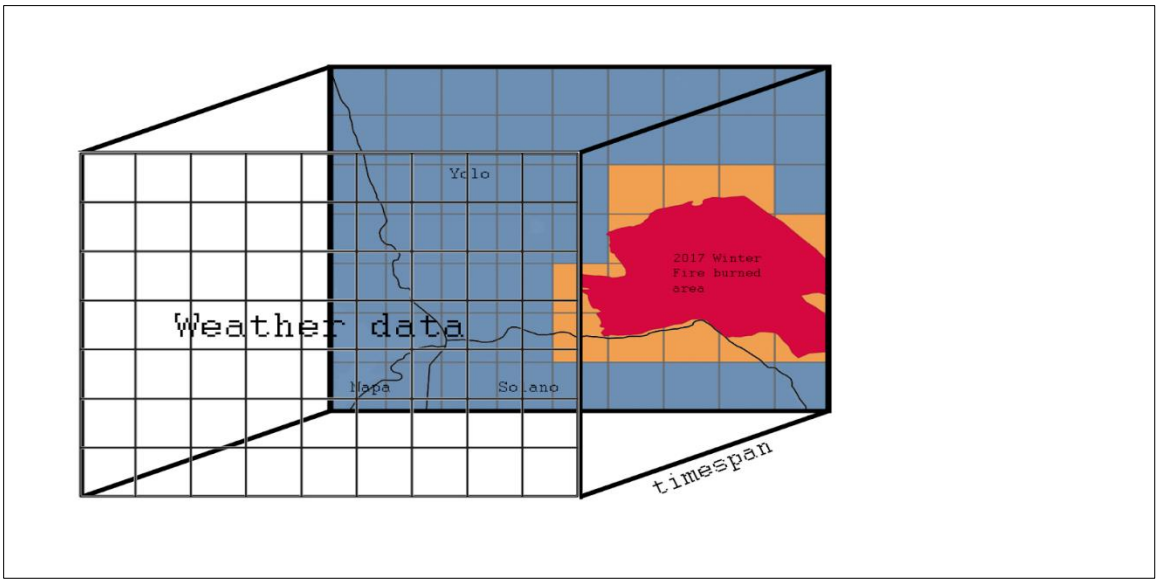


Figure 5.2. *Data mapping design*

Weather, terrain and powerline parameters are combined to form a single dataset for model ingestion because terrain and powerline data were sparse and lacked sufficient temporal variation.

Hence, we could not generate time-series data for standalone models. After conducting numerous experiments with various machine learning algorithms, varied target labeling (as in Table 5.1) and subsets of the datasets, the Random forest model with the type II data in the stratified format emerged the winner. Further, it produced the best accuracy for the problem as shown in Figure 5.3. The hyperparameters used for the Random forest model are shown in Figure 5.4. Feature importance for the model is shown in Figure 5.5.

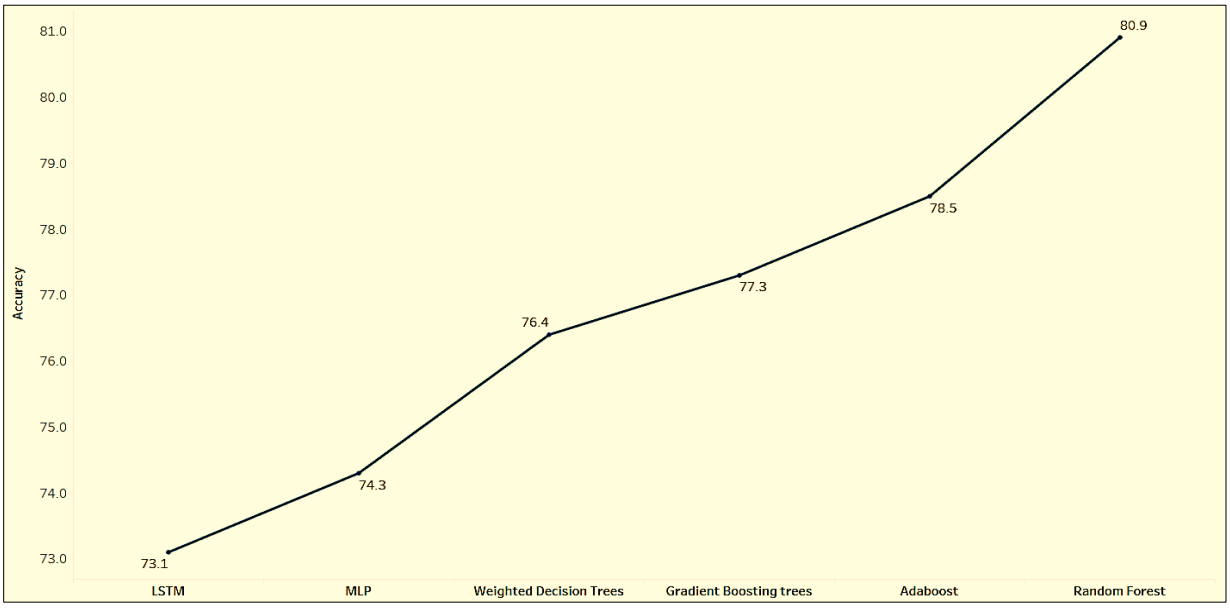


Figure 5.3. *Accuracy comparison for weather, powerline and terrain dataset*

```
rf_weather = RandomForestClassifier(n_estimators = 200, random_state = 42)
rf_weather.fit(X_train, y_train)
```

```
RandomForestClassifier(bootstrap=True, ccp_alpha=0.0, class_weight=None,
                       criterion='gini', max_depth=None, max_features='auto',
                       max_leaf_nodes=None, max_samples=None,
                       min_impurity_decrease=0.0, min_impurity_split=None,
                       min_samples_leaf=1, min_samples_split=2,
                       min_weight_fraction_leaf=0.0, n_estimators=200,
                       n_jobs=None, oob_score=False, random_state=42, verbose=0,
                       warm_start=False)
```

Figure 5.4. *Random Forest model hyperparameters for weather, powerline and terrain dataset*

0	HourlyDryBulbTemperature	0.175645
1	HourlyRelativeHumidity	0.174377
2	HourlyWindSpeed	0.149343
11	Ascpect_Range_South	0.022450
12	Ascpect_Range_South East	0.019919
9	Ascpect_Range_East	0.009592
13	Ascpect_Range_South West	0.008348
14	hillshade_direction_East	0.007076
15	hillshade_direction_North	0.006978
20	Status_Operational	0.003899
8	Length_Fee	0.003361
19	Status_Not Operating	0.002769
16	slope_range_High Slope	0.002607
22	Circuit_Single	0.002524
7	Length_Mil	0.002411
6	kV	0.002252
21	Circuit_Other	0.002130
18	slope_range_Moderate Slope	0.001445
17	slope_range_Low Slope	0.001095
10	Ascpect_Range_North East	0.000885
3	HourlyPrecipitation	0.000000

Figure 5.5. *Feature importance for weather, powerline and terrain dataset*

5.1.2 Base model II - Vegetation

After several iterations of dataset generation, we chose 8-day composites of vegetation indices available in Google Earth Engine (GEE) cloud catalogs as image collections. Thereafter, we created the training, test and validation sets following a similar methodology as the weather dataset.

After conducting numerous experiments with various machine learning algorithms, varied target labeling (as in Table 5.1) and subsets of the datasets, the Random forest model with the type II data in the stratified format fared best. Figure 5.6 shows the accuracy comparison of various machine learning models performed on vegetation dataset.

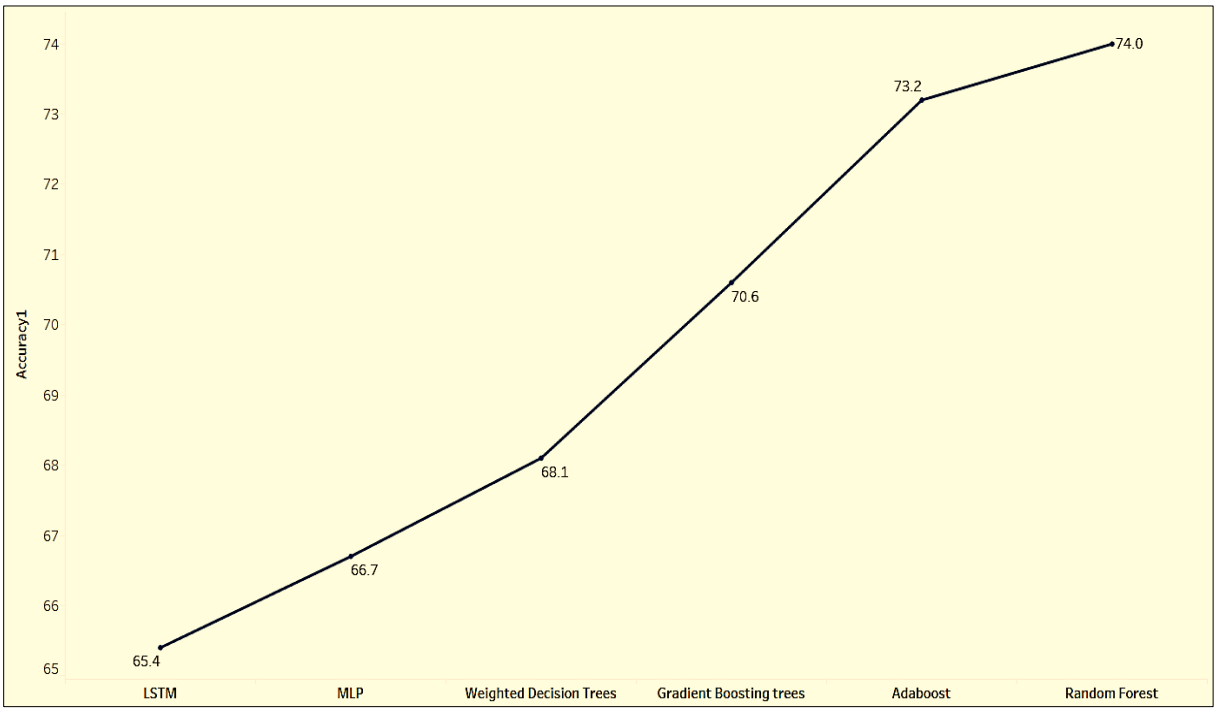


Figure 5.6. *Accuracy comparison for Vegetation dataset*

In the final model, we took stratified samples of the dataset. The same was cross validated to generate a Random forest model. Figure 5.7 shows the hyperparameters used for random forest model and figure 5.8 displays the feature importance.

```
rf_veg = RandomForestClassifier(n_estimators = 200, random_state = 42)
rf_veg.fit(X_train, y_train)
```

Figure 5.7. *Random Forest model hyper parameters*

	A	B
2	NDVI	0.268818
3	EVI	0.227308
4	NDWI	0.217428
1	Centroid Latitude	0.154895
0	Centroid Longitude	0.131551

Figure 5.8. *Feature importance for vegetation dataset*

Algorithms, such as Weighted Decision Trees, Decision trees, Gradient Boosting, Adaboost, LSTM and MLP, backed by previous research were implemented to figure out the best model. Random Forest produced the best accuracy as shown in Figure 5.7. Therefore, we will be considering the Random Forest model for combined and ensemble modeling. Figure 5.8 shows the feature importance sorted by importance.

5.1.3 Ensemble model

After generating base models for vegetation and weather with power lines and terrain, combining these two weak models and creating a single ensembled model gave better results. The two weak models were stacked by ensembling. Figure 5.9 shows the accuracy comparison of models used in ensemble model experimentation.

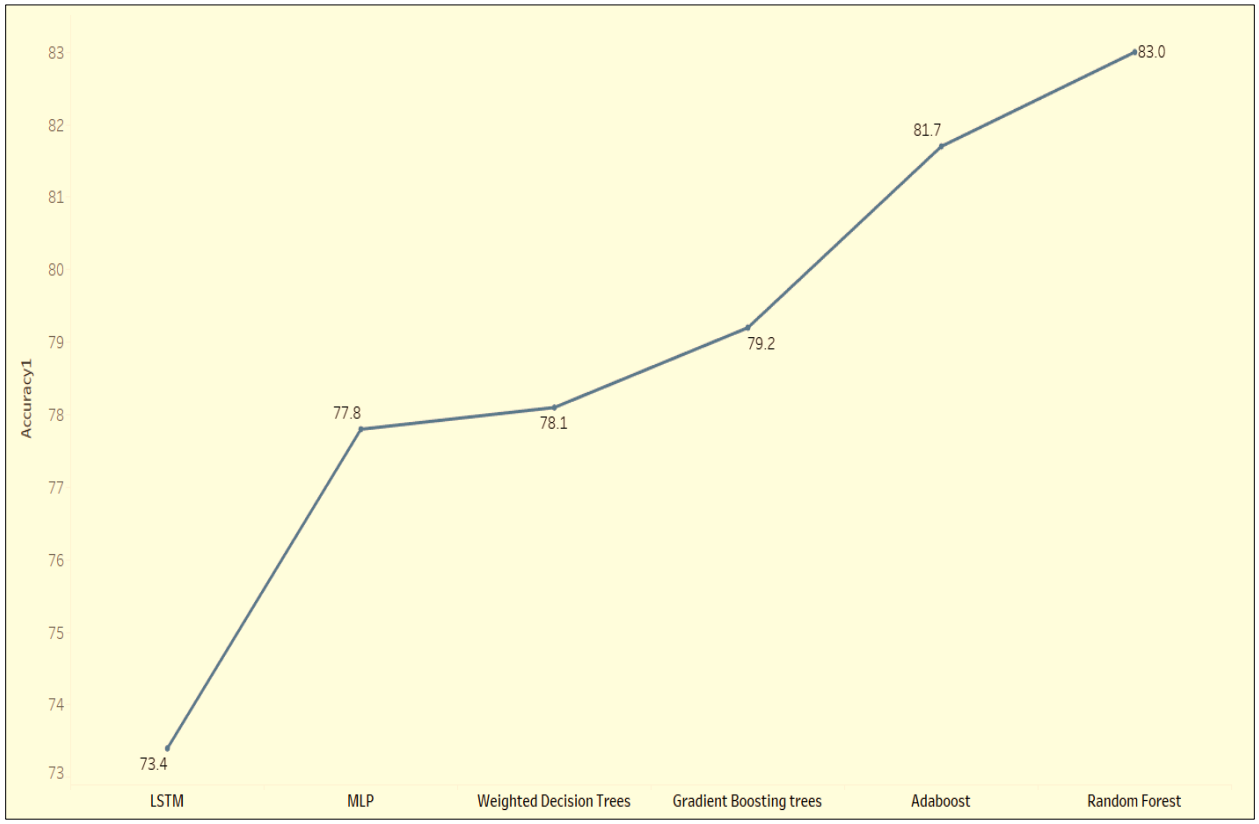


Figure 5.9. *Accuracy comparison for ensemble model*

In the stacked ensemble model, the outputs from the two weak models (weather with powerline, terrain and vegetation), called the base models, were inputted to the meta classifier model resulting in a robust ensemble model. During our experiments, Adaboost classifier consistently gave high accuracy as the second-layer learning algorithm for ensemble model, although Random forest fared best in terms of accuracy. However, the results from Random forest were inconsistent and unreliable. Hence, we ensembled the models using Adaboost meta estimator for the two base models with random forest classifier. Figure 5.10 shows the hyperparameters used for the same.

```
from sklearn.ensemble import AdaBoostClassifier
from sklearn.tree import DecisionTreeClassifier
classifier = AdaBoostClassifier(
    RandomForestClassifier(max_depth=1),
    n_estimators=200
)
classifier.fit(X_train_ensemble, y_train_ensemble)
```

Figure 5.10. *Hyperparameters used for Adaboost classifier in ensemble model*

5.1.4 Combined model

The combined dataset with all the parameters such as weather, vegetation, terrain and power lines, in a python dataframe format, was utilized for training various models. Yet again, type II dataset (from Table 5.1) with Random forest classifier and stratified train test split fared best. Figure 5.11 shows the accuracy comparison for various combined modeling experiments and Figure 5.12 shows the hyperparameters used for the final Random forest classifiers. Figure 5.13. shows the feature importance of the parameters in the combined model, sorted by importance.

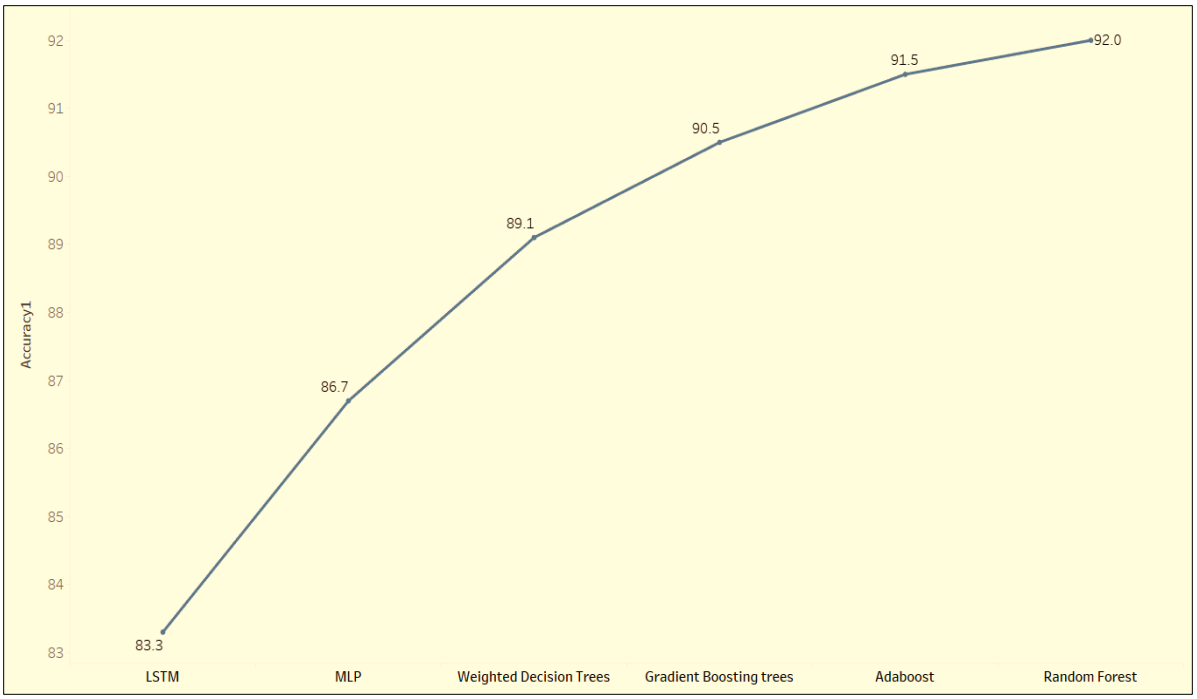


Figure 5.11. *Accuracy comparison for combined model*

```
rf_combined = RandomForestClassifier(n_estimators = 200, random_state = 42)

rf_combined.fit(X_train, y_train)

RandomForestClassifier(bootstrap=True, ccp_alpha=0.0, class_weight=None,
                       criterion='gini', max_depth=None, max_features='auto',
                       max_leaf_nodes=None, max_samples=None,
                       min_impurity_decrease=0.0, min_impurity_split=None,
                       min_samples_leaf=1, min_samples_split=2,
                       min_weight_fraction_leaf=0.0, n_estimators=200,
                       n_jobs=None, oob_score=False, random_state=42, verbose=0,
                       warm_start=False)
```

Figure 5.12. *Hyperparameters used for Random forest in combined model*

6	NDVI	0.238528
8	EVI	0.169975
4	Longitude	0.142837
7	NDWI	0.134486
5	Latitude	0.076802
2	HourlyWindSpeed	0.058784
1	HourlyRelativeHumidity	0.056988
0	HourlyDryBulbTemperature	0.048333
15	Ascpect_Range_South East	0.013469
14	Ascpect_Range_South	0.012470
16	Ascpect_Range_South West	0.007184
17	hillshade_direction_East	0.005387
18	hillshade_direction_North	0.004580
24	Circuit_Other	0.003174
22	Status_Not Operating	0.003138
12	Ascpect_Range_East	0.003025
9	kV	0.003012
19	slope_range_High Slope	0.002944
10	Length_Mil	0.002909
11	Length_Fee	0.002894
25	Circuit_Single	0.002796
23	Status_Operational	0.002720
21	slope_range_Moderate Slope	0.002167
13	Ascpect_Range_North East	0.000801
20	slope_range_Low Slope	0.000598

Figure 5.13. *Feature importance for combined model*

5.2 Model Execution or Evaluation

5.2.1 Base model I - Weather, Terrain and Powerline

After fitting the final random forest model and cross validating within our training data, we tested it against test data and evaluated the results using metrics. Figure 5.14 shows the classification report, accuracy and confusion matrix for the weather, powerline and terrain test data. We obtained an accuracy of 80% for the weather model.

```
Confusion Matrix :
[[1464  370]
 [ 330 1504]]
Accuracy Score : 0.8091603053435115
              precision    recall  f1-score   support

           0       0.82      0.80      0.81      1834
           1       0.80      0.82      0.81      1834

    accuracy                           0.81      3668
   macro avg       0.81      0.81      0.81      3668
weighted avg       0.81      0.81      0.81      3668
```

Figure 5.14. *Evaluation metrics for weather, terrain and powerline data*

5.2.2 Base model II - Vegetation

We continued the same series of experiments on vegetation data. Figure 5.15 shows the classification report, accuracy and confusion matrix for the Vegetation model. We obtained an accuracy of 73%.

```
Confusion Matrix :
[[59 17]
 [23 54]]
Accuracy Score : 0.738562091503268
Report :
              precision    recall  f1-score   support

           0       0.72      0.78      0.75        76
           1       0.76      0.70      0.73        77

    accuracy                           0.74       153
   macro avg       0.74      0.74      0.74       153
weighted avg       0.74      0.74      0.74       153
```

Figure 5.15. *Evaluation metrics for vegetation data*

5.2.3 Ensemble model

Figure 5.16 shows the classification report, accuracy and confusion matrix for the Ensemble model. An accuracy of 83% was obtained for our model.

```
Confusion Matrix :
[[1590  105]
 [ 297  428]]
Accuracy Score : 0.8338842975206612
Report :
              precision    recall  f1-score   support

           0       0.84      0.94      0.89      1695
           1       0.80      0.59      0.68       725

    accuracy                           0.83      2420
   macro avg       0.82      0.76      0.78      2420
weighted avg       0.83      0.83      0.83      2420
```

Figure 5.16. *Evaluation metrics for ensemble model*

5.2.4 Combined model

Figure 5.17. shows the classification report, accuracy and confusion matrix for the combined model. Clearly, this model has better evaluation results. The model accuracy is 91%. Further, we tested it on a newer validation dataset and obtained good results. Thereafter, this robust model was considered as the final Machine learning model for this project.

```
Confusion Matrix :
[[1497  198]
 [  74 1620]]
Accuracy Score : 0.9197403363824137
Report :
              precision    recall  f1-score   support

           0       0.95      0.88      0.92      1695
           1       0.89      0.96      0.92      1694

    accuracy                           0.92      3389
   macro avg       0.92      0.92      0.92      3389
weighted avg       0.92      0.92      0.92      3389
```

Figure 5.17. *Evaluation metrics for combined model*

5.3 Model Validation

Throughout our final model experimentation phase, we cross-validated the models. This technique enhanced the model accuracy and prevented overfitting. A brief summary of the cross-validation results are shown in the figures 5.14, 5.15, 5.16, 5.17, 5.18, 5.19, 5.20 and 5.21. The final model accuracy is 91% with the combined model as shown in figure 5.17.

SMOTE technique was used to generate synthetic samples of the underrepresented class in the unbalanced dataset. The newly generated balanced dataset was diligently stratification by cross-validation with the intent to split the dataset into homogeneous subpopulations. The random train-test ratio is 4:1 (80% training data and 20% testing data). We adopted the following techniques to ensure model correctness, consistency and reproducibility.

- Exhaustive experimentation with numerous combinations of datasets and algorithms.
- Fix the random forest classifier random state, which in turn freezes the tree upon creation. Hence, our results are reproducible.
- Specialized data sampling and stratification of the dataset. Our sampling represented relevant information, and the results were consistent and reproducible.

The robust models with comparable accuracy, that is, the combined and ensemble model, were tested thoroughly. Below are the results. Learning curves display the comparison between the training and validation scores, scalability and performance of the models.

5.3.1 Weather, Powerline and Terrain model

Apart from the results in figure 5.14, learning curves for the model fitted with weather, powerlines and terrain data are shown in figure 5.18.

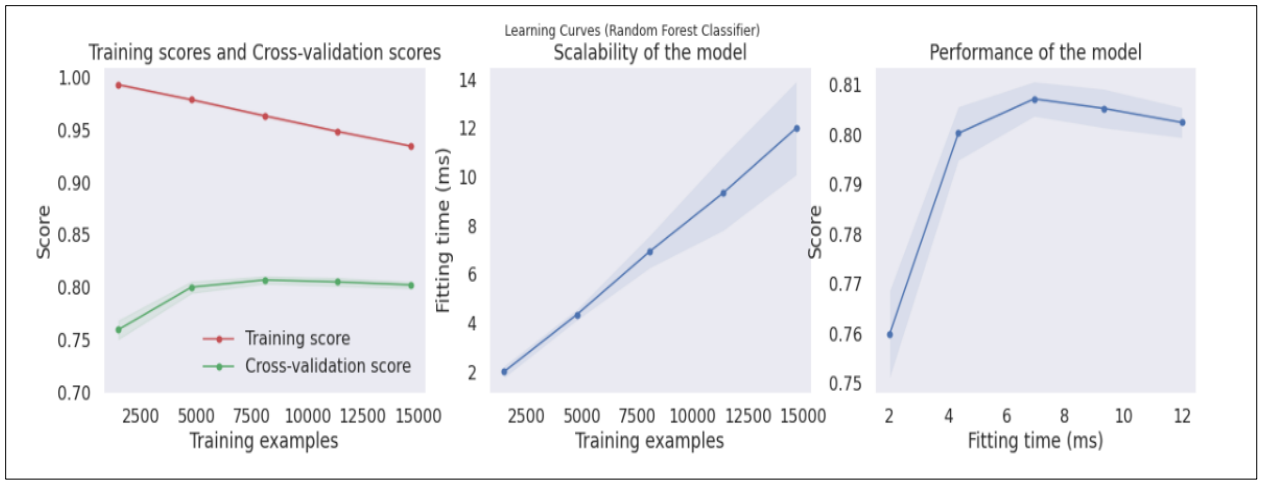


Figure 5.18. *Learning curves for Weather, Powerline and Terrain model*

5.3.2 Vegetation model

Apart from the results in figure 5.15, learning curves for the model fitted with vegetation data are shown in Figure 5.19.

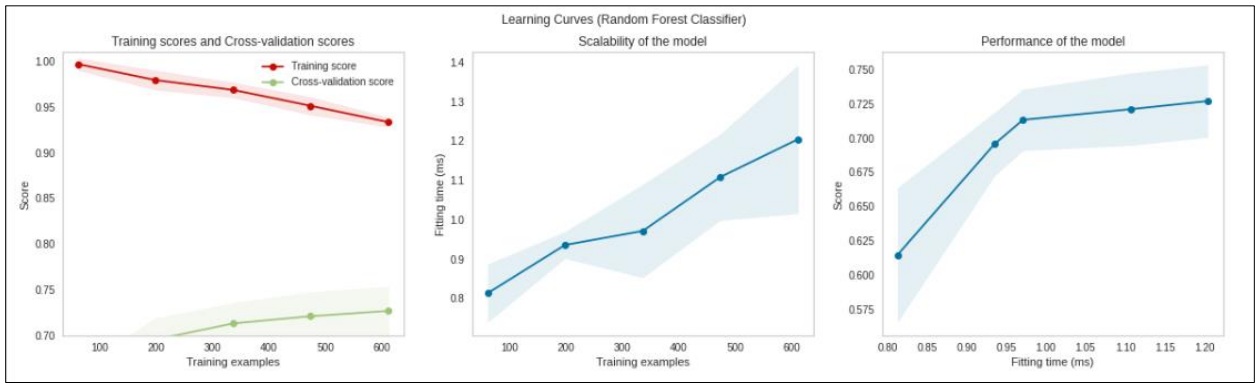


Figure 5.19. *Learning curves for Vegetation model*

The dataset was cross validated using 5 segments, best model parameters were gauged, and threshold values were tweaked to yield an accuracy of 73%.

5.3.3 Ensemble model

Apart from the results in figure 5.16, learning curves for the ensemble model generated from the base models (weather, terrain and powerline model, and vegetation model) are shown in Figure 5.20.

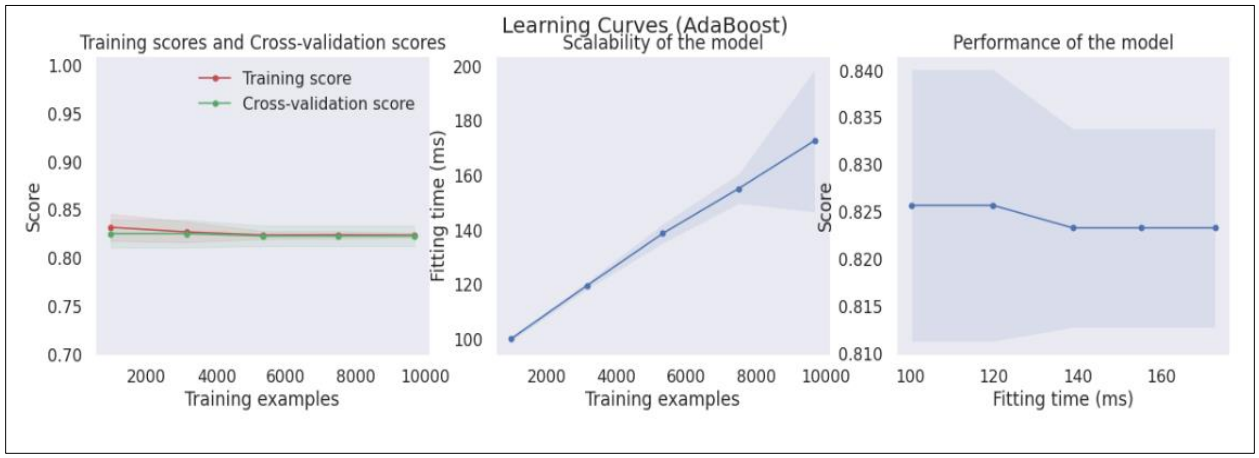


Figure 5.20. *Learning curves for Ensemble model*

The dataset was cross validated using 5 segments, best model parameters were gauged, and threshold values were tweaked to yield an accuracy of 83%. Precision recall and f1-scores are similar but leaning towards negative targets.

5.3.4 Combined model

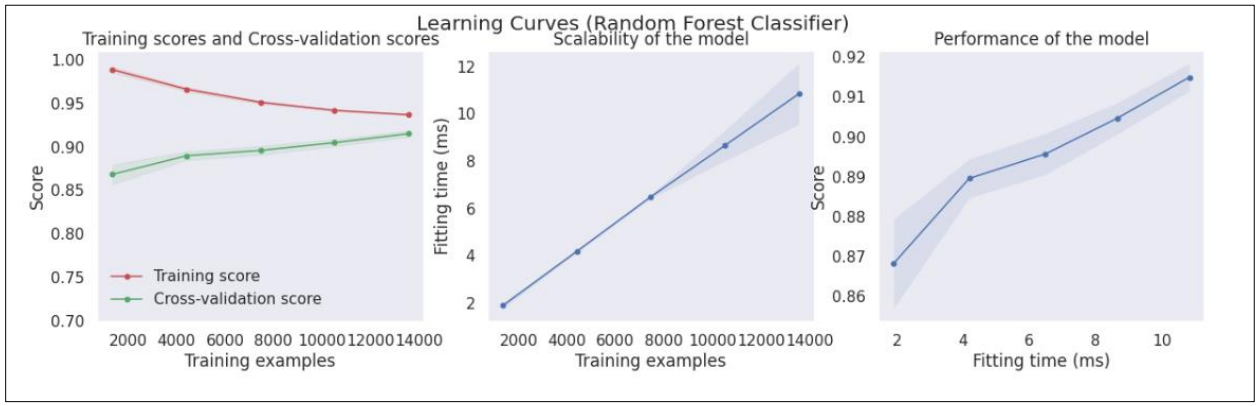


Figure 5.21. *Learning curves for Combined model*

Apart from the results in figure 5.17, learning curves for the ensemble model generated from the base models (weather, terrain and powerline model, and vegetation model) are shown in

Figure 5.21. The Classification report in figures 5.17, 6.8 and 6.9 shows an accuracy of 92% for the combined model with the combined dataset. The ROC curve was plotted and carefully evaluated, with a best threshold value. By far, this model fared best, given the complexity of the combined data.

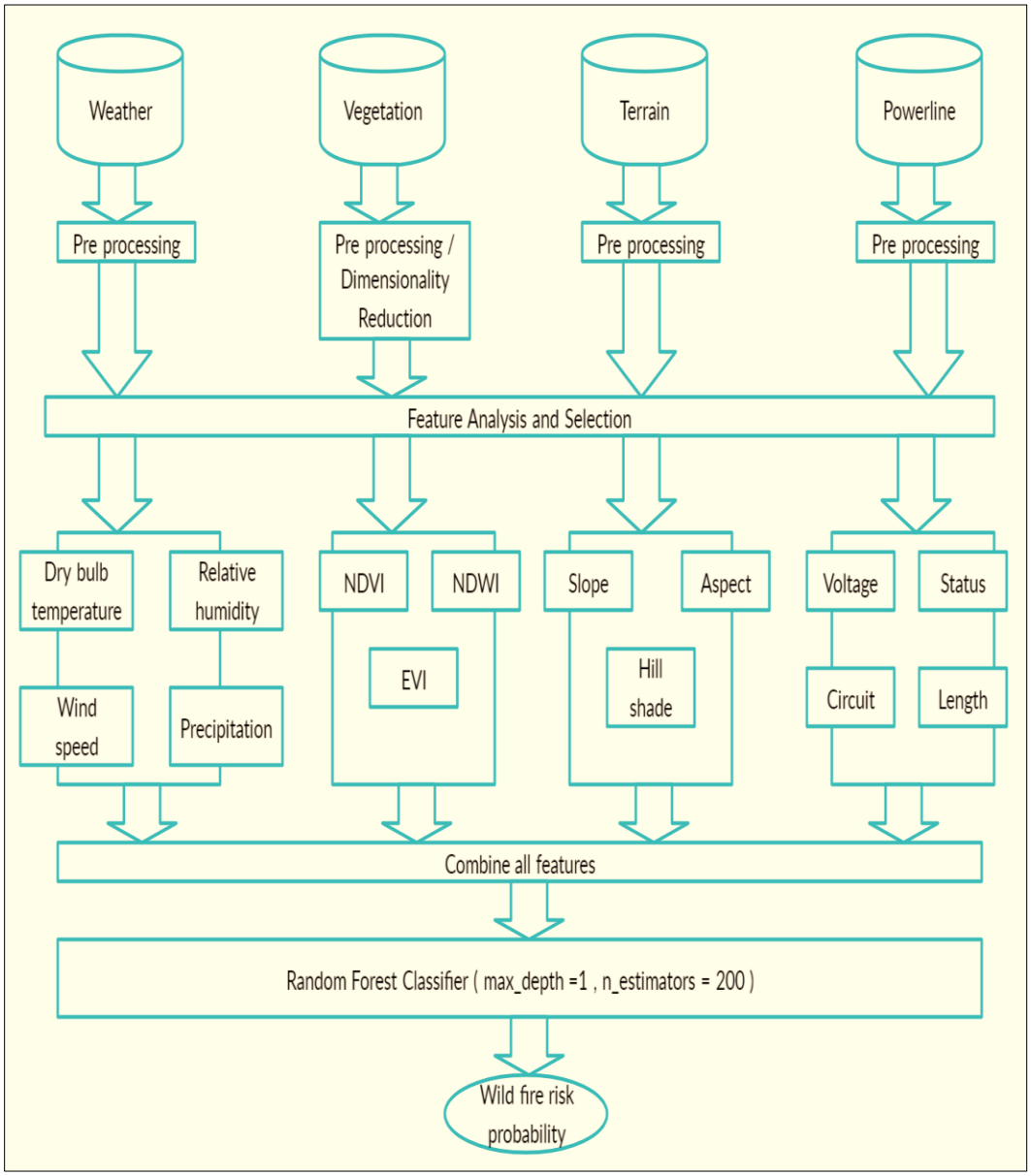


Figure 5.22. *Final combined model architecture*

Both ensemble and combined model had comparable accuracy, although the ensemble model accuracy dropped after few testing cycles, when the algorithm was re-executed. Figure 5.22

and 5.23 show the combined and ensemble model architectures in detail. Combined model was chosen as the final model for this project.

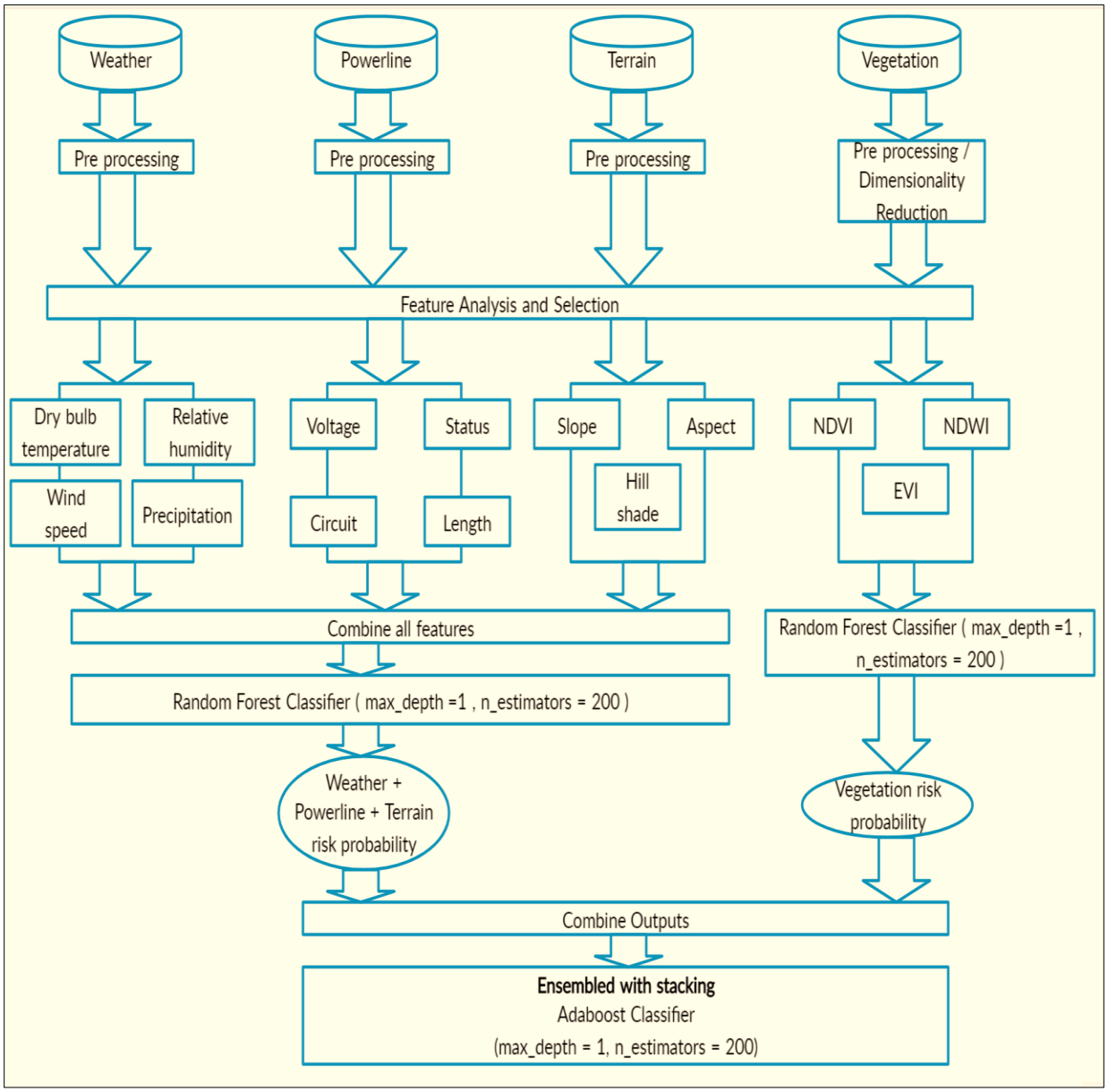


Figure 5.23. *Final ensemble model architecture*

6. EVALUATION AND VISUALIZATION

6.1 Analysis of Model Execution/Evaluation Results

As shown in Figure 5.22, we trained two separate base models for vegetation as well as weather, terrain and powerline data. These models were ensembled by stacking as shown in Figure 5.23. In the combined model, we merged all the datasets to generate a combined dataset before applying machine learning algorithms. Random forest algorithm was the overall best performer in our experiments. Evaluation metrics are described here.

- **Confusion Matrix**

The confusion matrix [67] is a table representing the performance of your model and its ability classify labels correctly. Figure 6.1 elaborates the components in a confusion matrix.

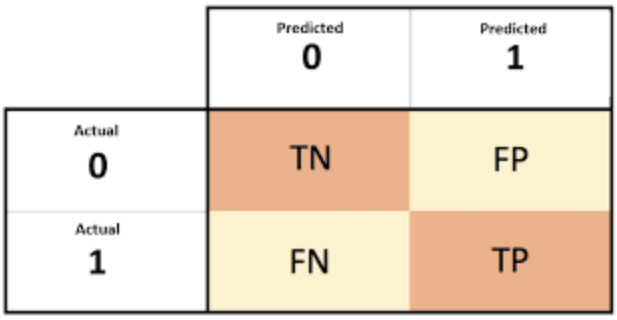


Figure 6.1. *Confusion matrix*

In a binary classifier, the "true" class is typically labeled as 1 and the "false" class is labeled as 0.

- TP: true positives (classifier correct; classifier guessed 1).
- FP: false positives (classifier incorrect; classifier guessed 1).
- TN: true negative (classifier correct; classifier guessed 0).
- FN: false negative (classifier incorrect; classifier guessed 0).

- **Accuracy**

With the total population as the sum of the above components in confusion matrix, below are the formulae.

- Total samples = TP + TN + FP + FN
- The accuracy can be calculated as:

$$Accuracy = \frac{truepositives + truenegatives}{totalexamples}$$

- **Sensitivity /Recall /True Positive Rate**

The true positive rate is the percent of times that the model correctly predicted 1 when the label was in fact 1. This is alternatively known as the sensitivity or recall.

- The sensitivity or recall can be calculated as:

$$Recall = \frac{truepositives}{truepositives + falsenegatives}$$

- **Precision /Positive Predictive Value**

Precision measures the percent of times the classifier was correct when it was predicting the true (1) class.

- The precision can be calculated as:

$$Precision = \frac{truepositives}{truepositives + falsepositives}$$

The idea of the classifier being *precise* is subtly different than it being *accurate*. Precision is a measure of correctness only for its positive class predictions, whereas accuracy is a measure of correctness for all predictions.

- **F1 Score**

The F1 score is the harmonic mean of the precision and recall metrics. Blending the two is useful; precision measures how effectively the classifier performs when it is predicting a '1', whereas recall measures how many of the total '1' classes out of all the 1-labeled observations were predicted correctly.

  - The F1 score can be calculated as:

$$F1 = 2 \times \frac{precision \times recall}{precision + recall}$$

By combining the two, we have a measure of the classifier's ability to find the positive-labeled observations, as well as how permissive it is of identification errors on those labels.

The Receiver Operating Characteristic (ROC) Curve

The ROC curve is a popular visual of the performance of a classifier. Below are its properties.

- It compares the true positive rate to the false positive rate as the threshold for predicting '1' changes.
- When the area under the curve (AUC) is 0.50 or 50%; the model is equivalent to the baseline (chance) prediction.
- When the area under the curve is 1.00 or 100%, the model makes perfect predictions.

The area under the ROC curve is inherently related to the accuracy, but the AUC-ROC is preferred because it is automatically adjusted to the baseline and gives a robust picture of how the classifier performs at different threshold choices.

6.1.1 Vegetation Model

Evaluation metrics for the vegetation models are shown in Figure 6.2 and Figure 6.3.

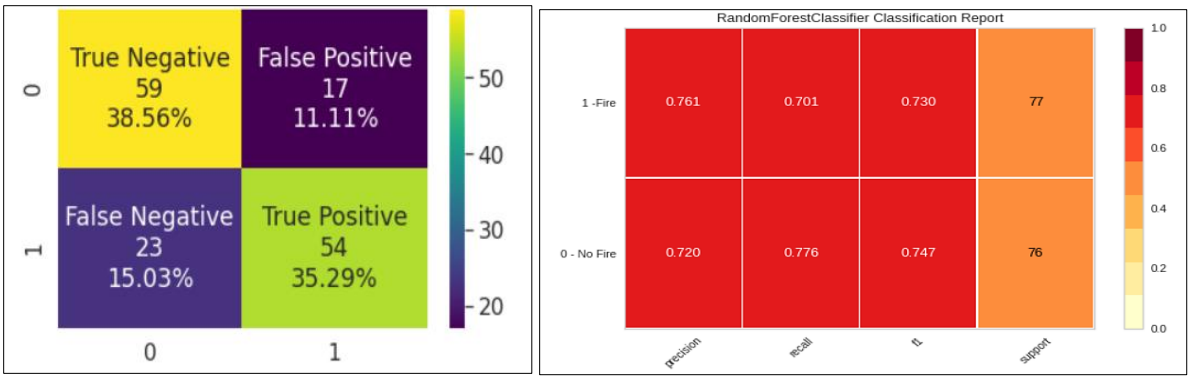


Figure 6.2. *Evaluation metrics for vegetation model*

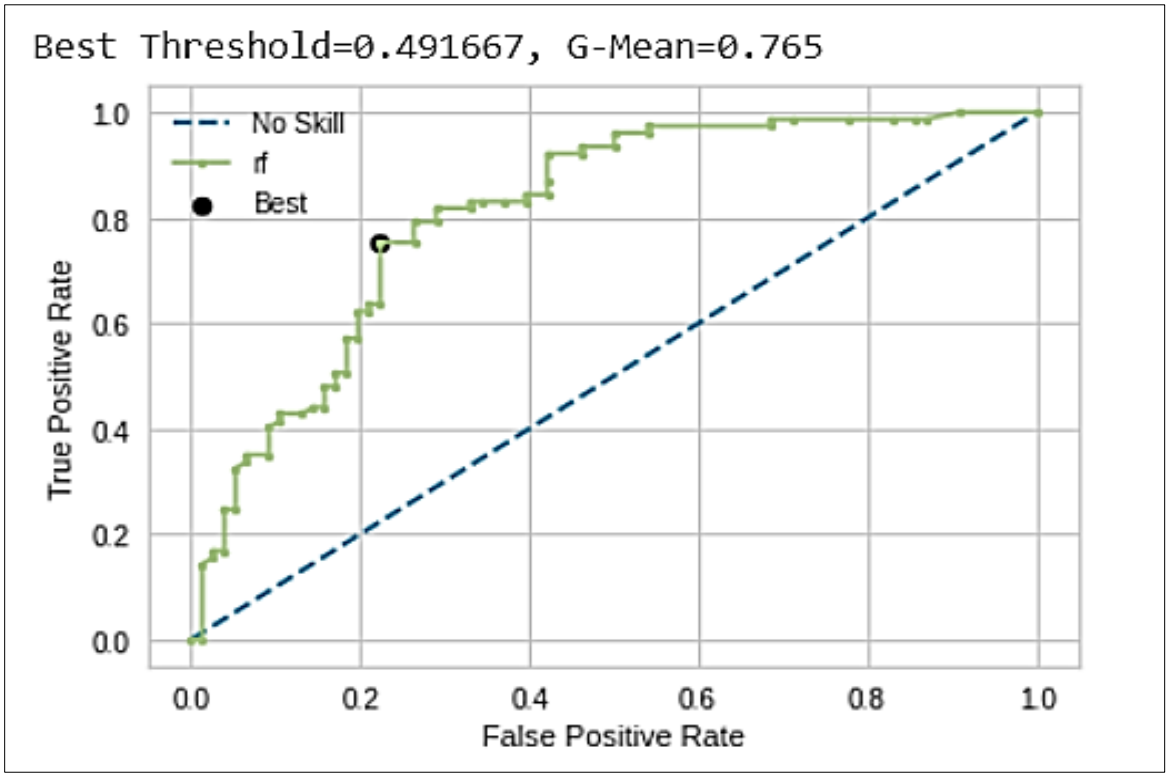


Figure 6.3. *ROC curve for vegetation model*

6.1.2 Weather, Terrain and Powerline Model

Evaluation metrics for weather, terrain and powerline model are shown in Figure 6.4 and Figure 6.5.

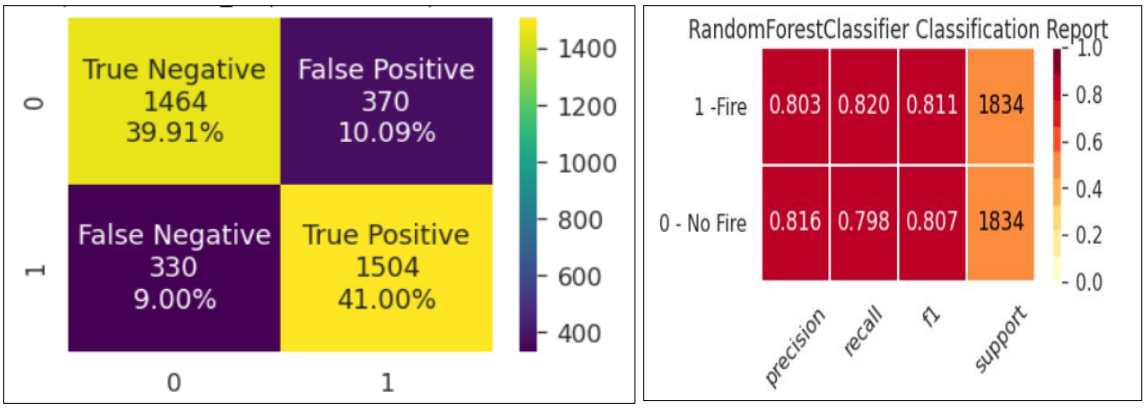


Figure 6.4. *Evaluation metrics for weather, terrain and powerline model*

Figure 6.5. *ROC curve for weather, terrain and powerline model*

6.1.3 Ensemble Model

Evaluation metrics for Ensemble model are shown in Figure 6.6 and Figure 6.7.

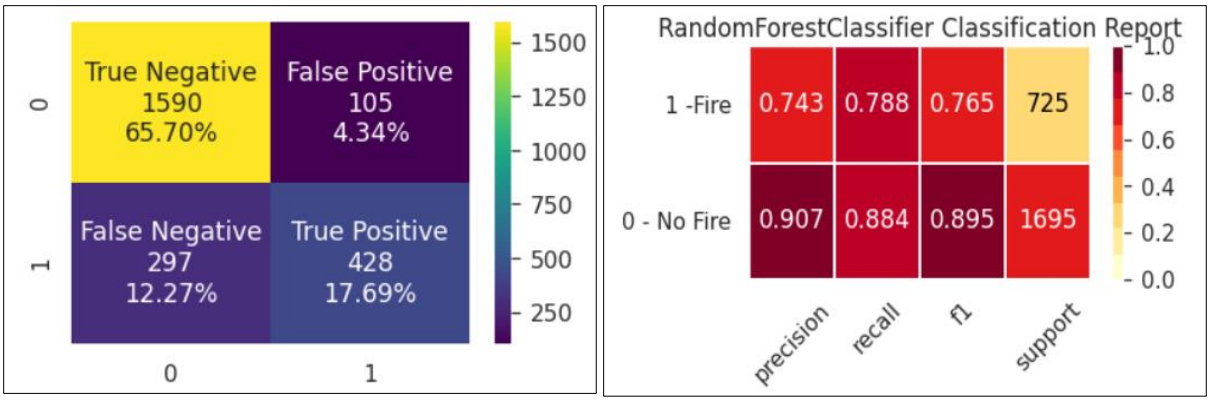


Figure 6.6. *Evaluation metrics for ensemble model*

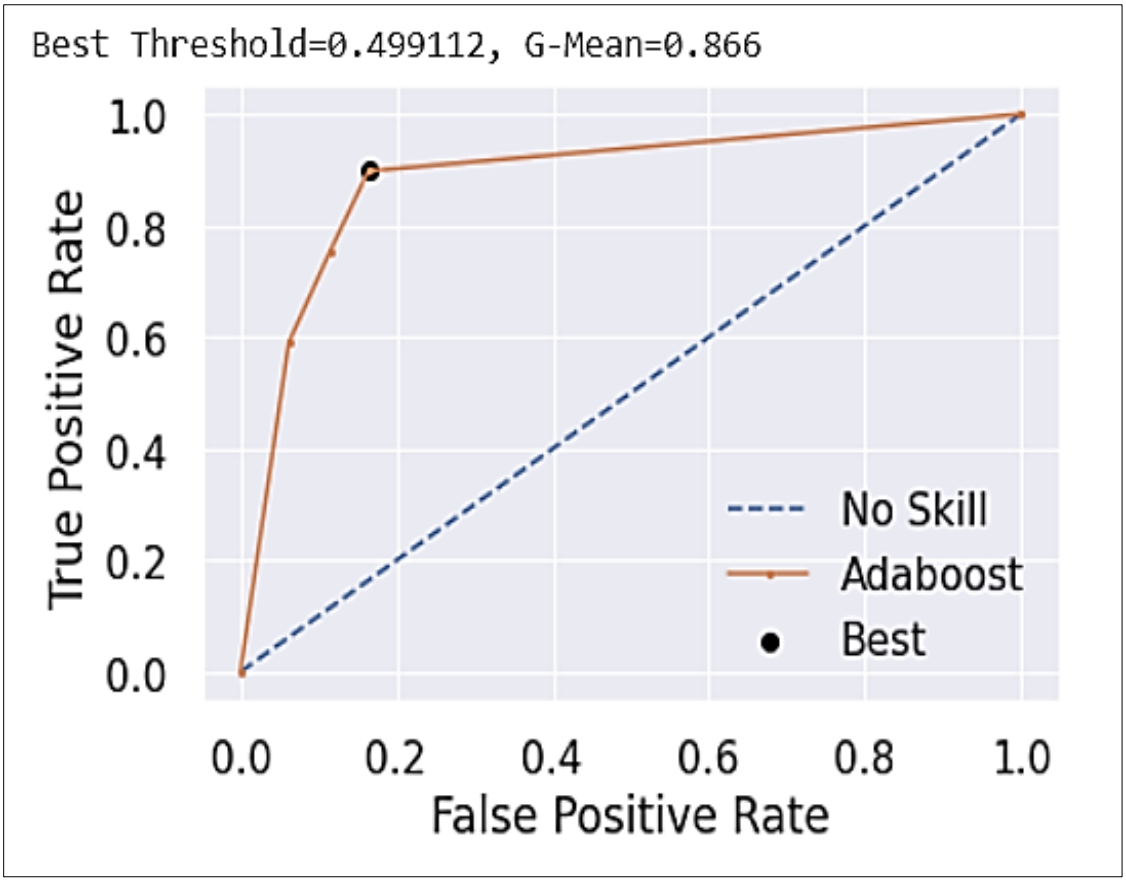


Figure 6.7. *ROC curve for ensemble model*

6.1.4 Combined Model

Evaluation metrics for combined models are shown in Figure 6.8 and Figure 6.9.

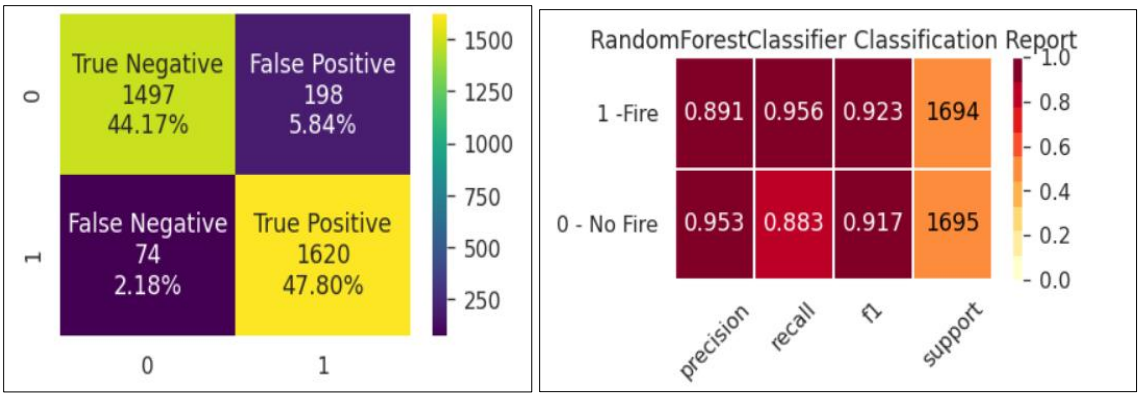


Figure 6.8. *Evaluation metrics for combined model*

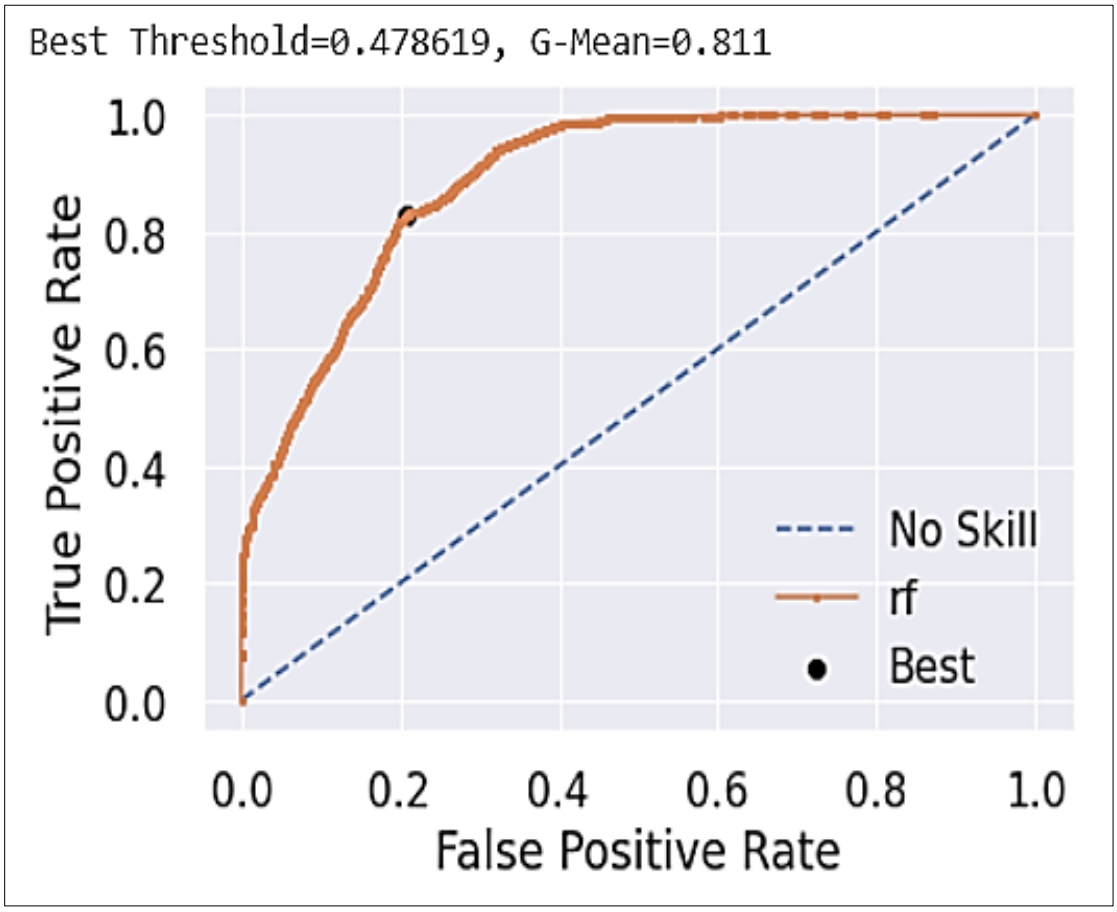


Figure 6.9. *ROC curve for combined model*

6.2 Achievements and Constraints

After reviewing thoroughly both the traditional prediction systems and latest data-driven methods, we concluded the following shortcomings of existing solutions:

- For traditional methods, they use statistical methods based on simplified mathematical models. Coefficients were constants derived from past fire studies, they can only represent a statistical value in a fixed condition. Therefore, they cannot dynamically change with different areas.
- For data-driven machine learning methods, they lack spatial and temporal accuracy and cannot predict real-time.

Our model has better temporal accuracy than previous machine learning models and can achieve real-time prediction. However, only paid data has high temporal resolution. We also trained our model to be location specific so it has high spatial accuracy. It can pinpoint an incident in real-time if we can fetch the most recent data.

The limitations of our models are:

- Requires high quality data which are not readily available.
- It is confined to our study area until further trained.

6.3 Quality Evaluation of Model Functions and Performance

For all the classification problems, default threshold value is 0.5. Sometimes the default threshold value can result in poor performance. The performance of a classifier that predicts probabilities can be tuned using ROC Curves and Precision-Recall Curves. Using these methods,

we experimented with different threshold values which will reduce false negatives by reducing accuracy to only some extent. Improved results for all the models are shown below.

6.3.1 Vegetation Model

After tuning the threshold value to 0.491667 we got less false negatives. Figure 6.10 shows the improved results for the vegetation model.

```
0.7516339869281046
Confusion Matrix :
[[59 17]
 [21 56]]
Accuracy Score : 0.7516339869281046
              precision    recall  f1-score   support

           0       0.74      0.78      0.76        76
           1       0.77      0.73      0.75        77

    accuracy                           0.75       153
   macro avg       0.75      0.75      0.75       153
weighted avg       0.75      0.75      0.75       153
```

Figure 6.10. *Results for vegetation model with best threshold value*

6.3.2 Weather, Terrain and Powerline Model

After modifying the threshold to the best obtained threshold value of 0.478619, the results have improved reducing false negatives for the predictions. But to obtain a better model we further reduced the threshold and captured the best threshold which had a good balance of false negatives and model accuracy. Improved results for weather, terrain and powerline model are shown in Figure 6.11. The new improved threshold value has decreased the number of false negatives.

```
0.8107960741548528
Confusion Matrix :
[[1454  380]
 [ 314 1520]]
Accuracy Score : 0.8107960741548528
              precision    recall  f1-score   support

           0       0.82      0.79      0.81      1834
           1       0.80      0.83      0.81      1834

    accuracy                           0.81      3668
   macro avg       0.81      0.81      0.81      3668
weighted avg       0.81      0.81      0.81      3668
```

Figure 6.11. *Results for weather, terrain and powerline model with best threshold value*

6.3.3 Ensemble Model

We have tested with multiple threshold values which are below 0.5 to decrease the number of false negatives which is most important for fire risk. The final threshold value is 0.491667 has increased the model accuracy along with reducing false negatives. Modified results are shown in Figure 6.12.

```
0.8458677685950413
Confusion Matrix :
[[1500  195]
 [ 178  547]]
Accuracy Score : 0.8458677685950413
Report :
              precision    recall  f1-score   support

           0       0.89      0.88      0.89      1695
           1       0.74      0.75      0.75       725

    accuracy                           0.85      2420
   macro avg       0.82      0.82      0.82      2420
weighted avg       0.85      0.85      0.85      2420
```

Figure 6.12. *Results for ensemble model with best threshold value*

6.3.4 Combined Model

Though the best threshold is 0.5, we tried with different threshold values to reduce false negatives and with 0.478619 as threshold the false negatives are minimum with improved accuracy score. The results are shown in Figure 6.13.

```
0.9223959870168191
Confusion Matrix :
[[1497  198]
 [  65 1629]]
Accuracy Score : 0.9223959870168191
Report :
              precision    recall  f1-score   support

           0       0.96      0.88      0.92      1695
           1       0.89      0.96      0.93      1694

    accuracy                           0.92      3389
   macro avg       0.93      0.92      0.92      3389
weighted avg       0.93      0.92      0.92      3389
```

Figure 6.13. *Results for combined model with best threshold value*

6.4 Evaluation of Models vs. Requirements

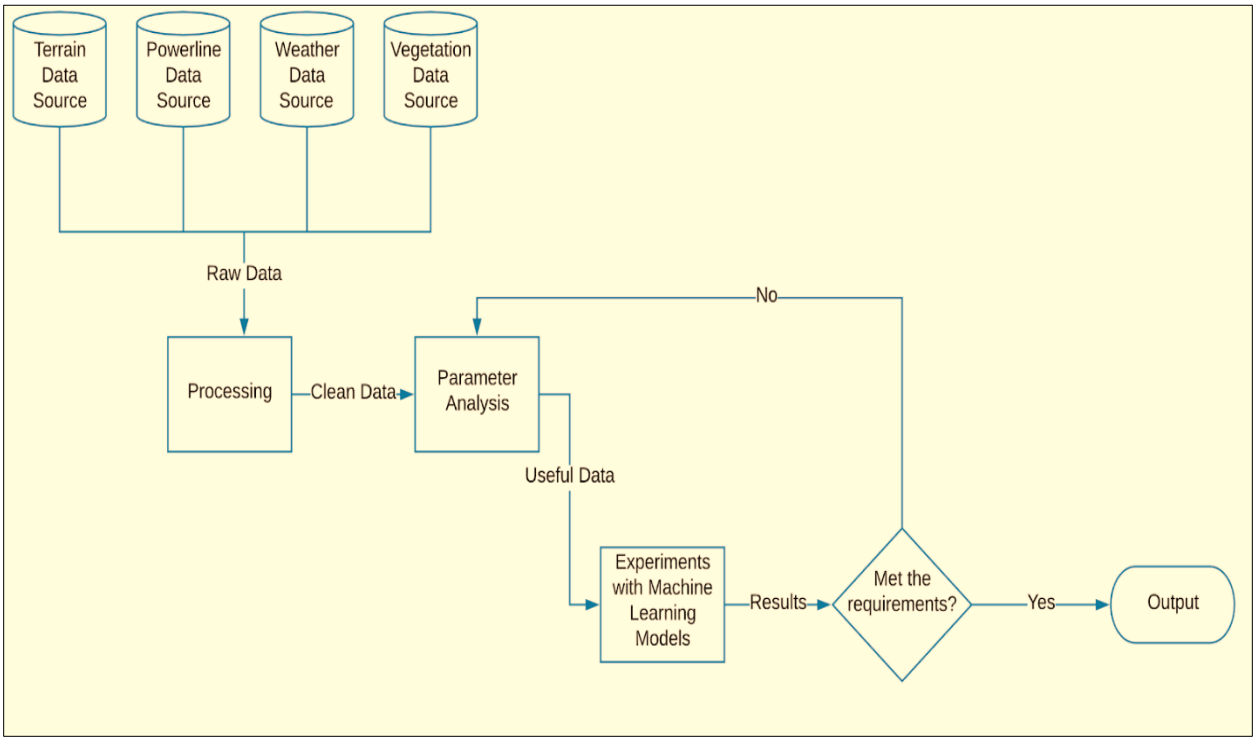


Figure 6.14. *Evaluation Strategy*

Figure 6.14. shows the flow diagram of our process. First the raw data from the source are processed and the parameters are analyzed. Considering our results from parameter analysis, we applied machine learning algorithms on different types of data as discussed in chapter 5.1. Parallelly for each experiment we optimized our model, evaluated the results with classification metrics, and compared them with following requirements.

- Accuracy of the model.
- Minimal False Negatives
- Interpretability of the model.
- Complexity of the model.
- Scalability of the model.

- Time taken to train and test the model
- Time taken to make predictions using the model

On completion of this iterative process model with all the parameters (vegetation, weather, powerlines, terrain) combined into a single dataset gave better results.

6.5 Information Visualization

6.5.1 Area of Study and Grids

Below are few suitable visualizations. The final user interface visualizations are provided at the end of this section.

There are 63 grids in our study area with 7 rows and 9 columns. Figure 6.15 and 6.16 display the map-based visualizations, which include both Satellite images and Street View of the location -based information.

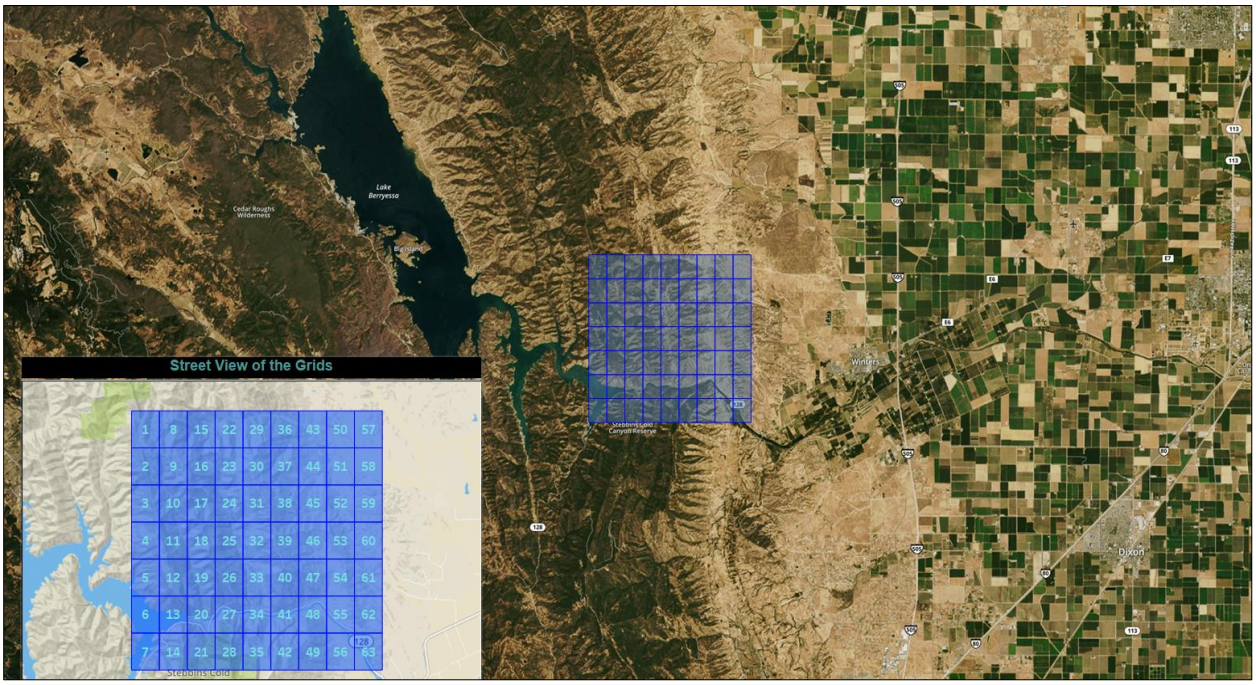


Figure 6.15. *Satellite/Street view of the grids in the Study Area*

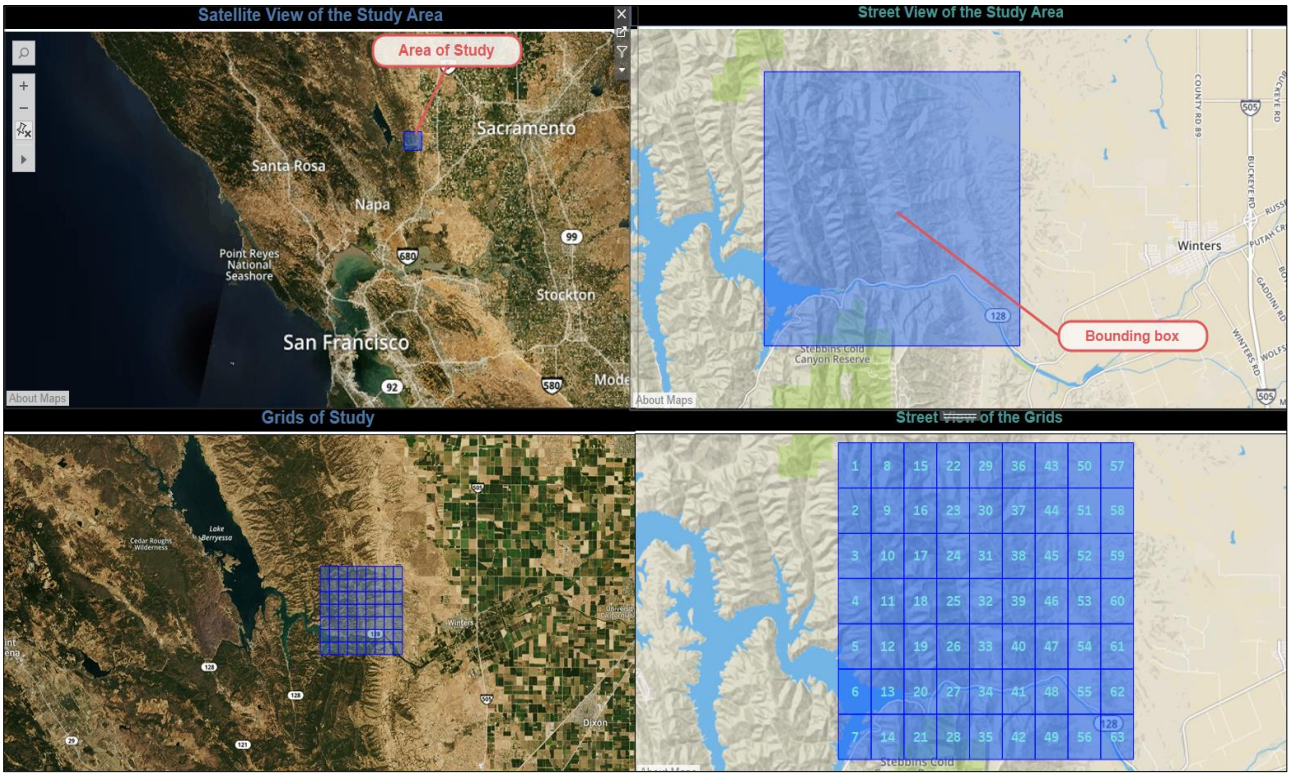


Figure 6.16. *Satellite and Street View of Study Area and the Grids*

6.5.2 Fire History Visualization

Figure 6.17. shows the relevant fire history visualizations such as California state history from 2014 to 2018, relevant fires for our area of study and its overlap with the grids.

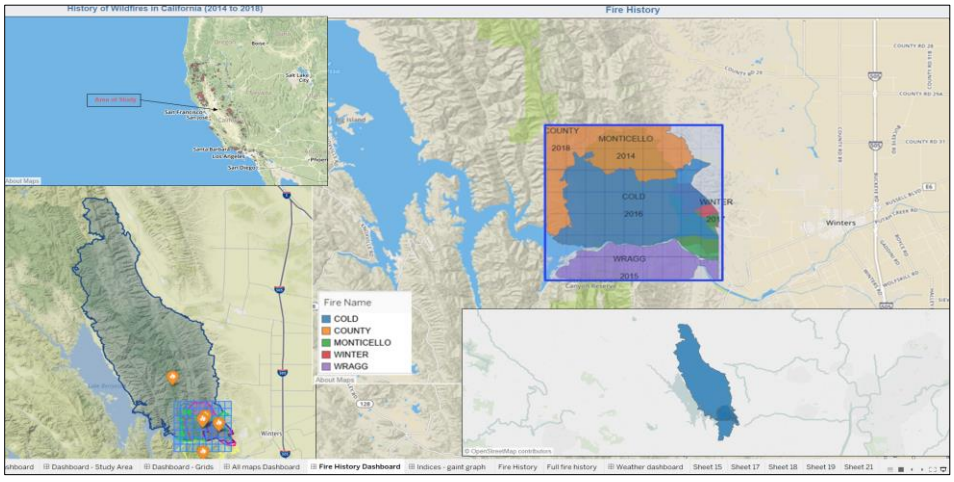


Figure 6.17. *Fire History Information*

6.5.3 Code-based Model Visualizations

Figure 6.18. shows the results of the model in python using a folium visualization. For each of the 63 grids, the icon is red if the target is '1', corresponding to a Fire event. Else, it is marked blue.

The important parameter for vegetation turned out to be NDVI. Hence, we have mapped the indices in each of these grids. Figure 6.19 shows the visualization of results for combined data with an emphasis on weather data. Hourly Dry Bulb temperature turned out to be a determining feature for fire related to weather.

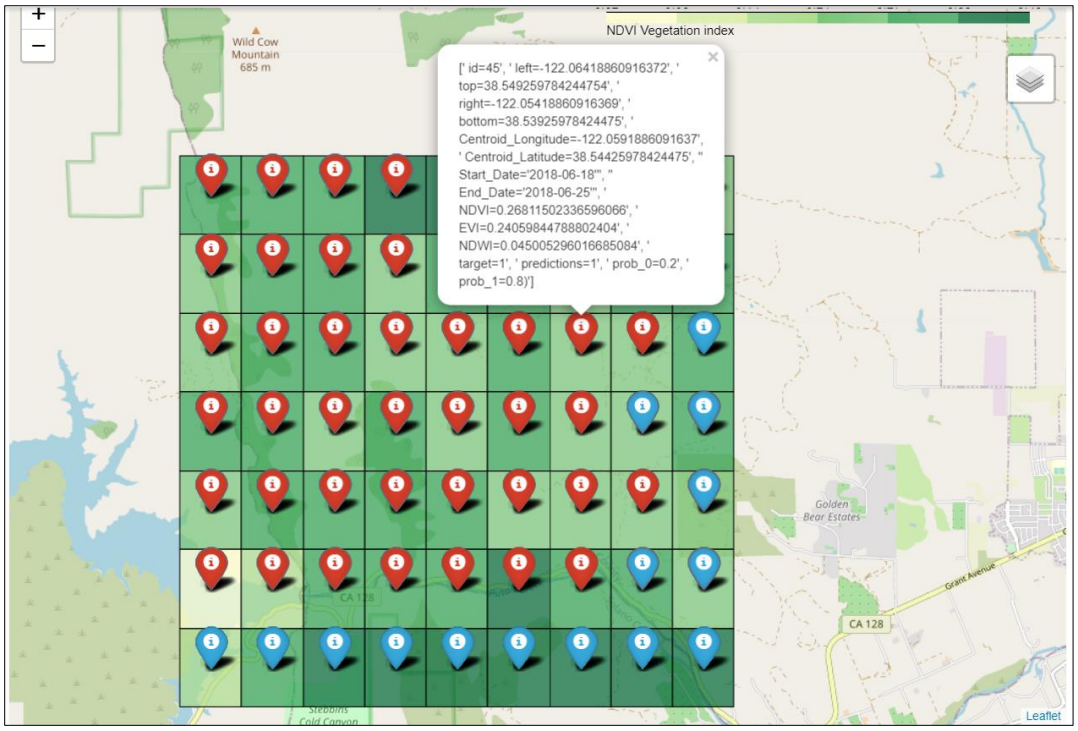


Figure 6.18. *Vegetation model results visualizations*

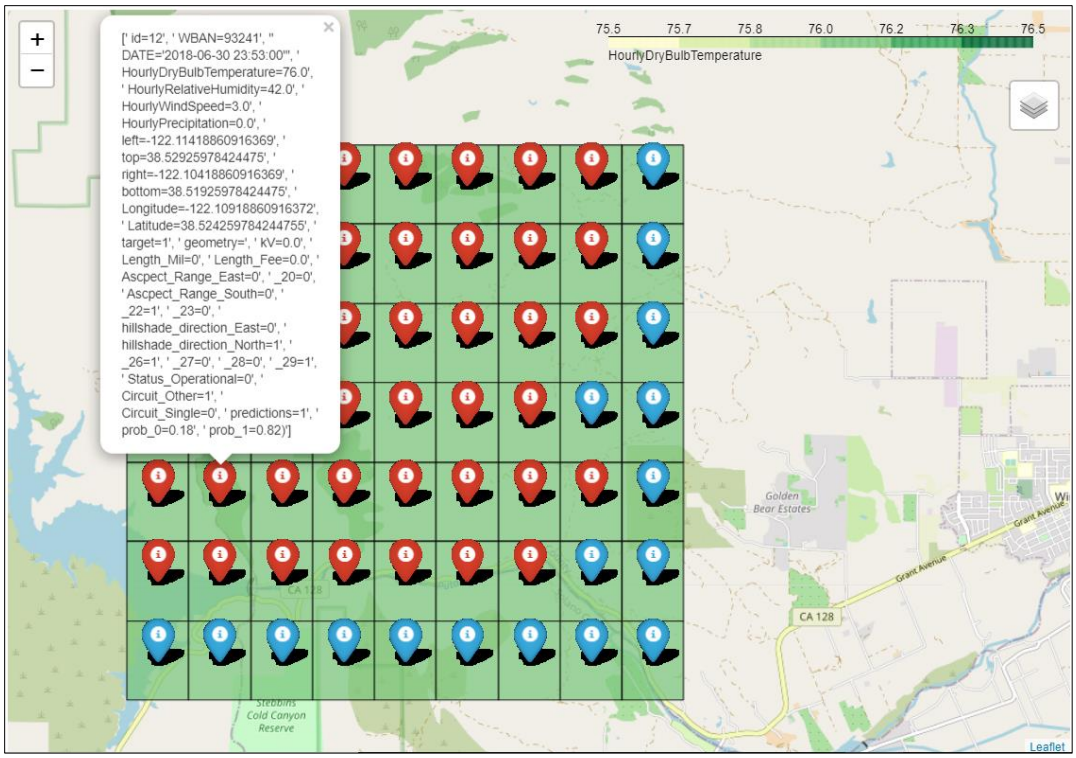


Figure 6.19. *Combined model results visualizations*

6.5.4 Vegetation

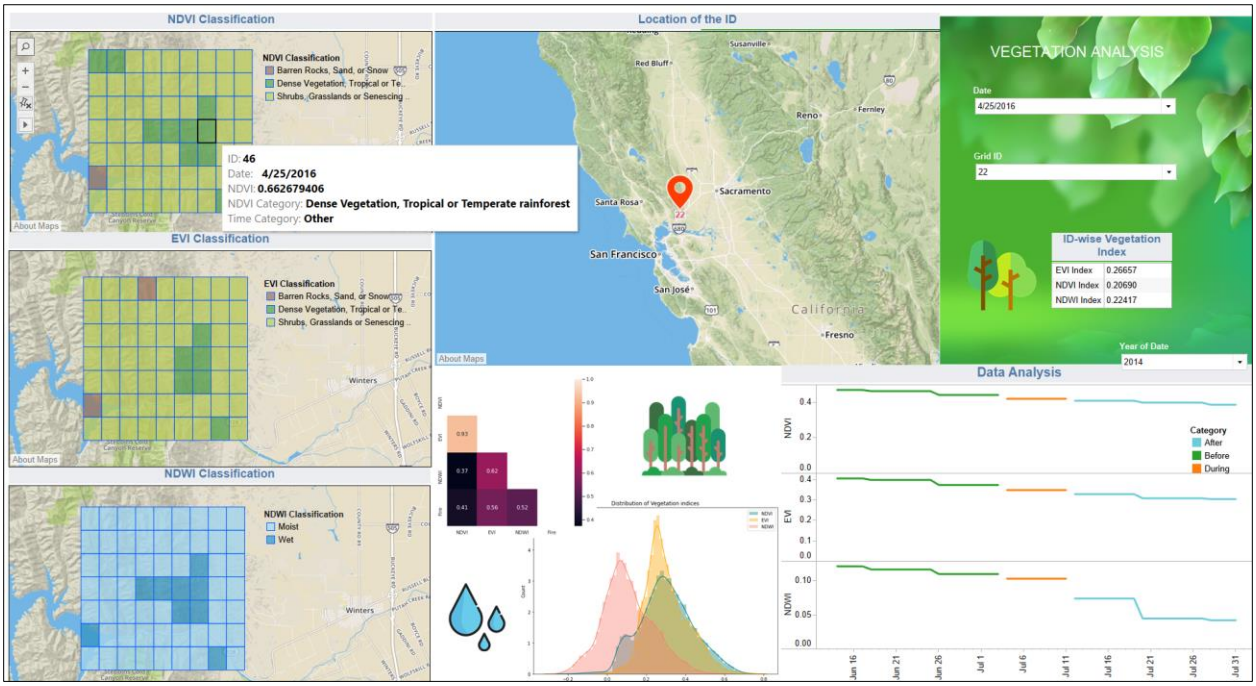


Figure 6.20. *Initial Vegetation dashboard*

Figure 6.20 displays the Vegetation dashboard. The design includes dynamic and static infographics. Filters to select Date, Grid ID and Year of Date are provided for the user. Whenever we choose date and ID, grid-wise NDVI, EVI and NDWI classification grids change, along-with the ID on map and Data Analysis. Further, a table shows the ID-specific values of the indices for that particular date. Data analysis section maps the ID-specific variation of the indices for 3 categories, namely Before, During and After fire, revealing that vegetation indices differ as expected. A normal distribution plot of each of the indices reveals that the data follows a normal bell curve. Heatmap correlation further shows the correlation between indices and fire.

6.5.5 Weather

Figure 6.21 shows the Weather dashboard. Similar to the former dashboard, it shows ID-specific weather information such as Dry Bulb Temperature, Precipitation, Humidity and Wind Speed. We can choose the ID as well as Date and Time. Pie chart shows the unbalanced nature of the dataset. Hence, we filtered out the relevant Area, isolating the fire-ridden grids. Furthermore, we chose to display the ID-specific variation of parameters for 2-time frames, namely During fire and otherwise. From a cumulative standpoint, a filtered plot maps the relationship between the average weather features and target.

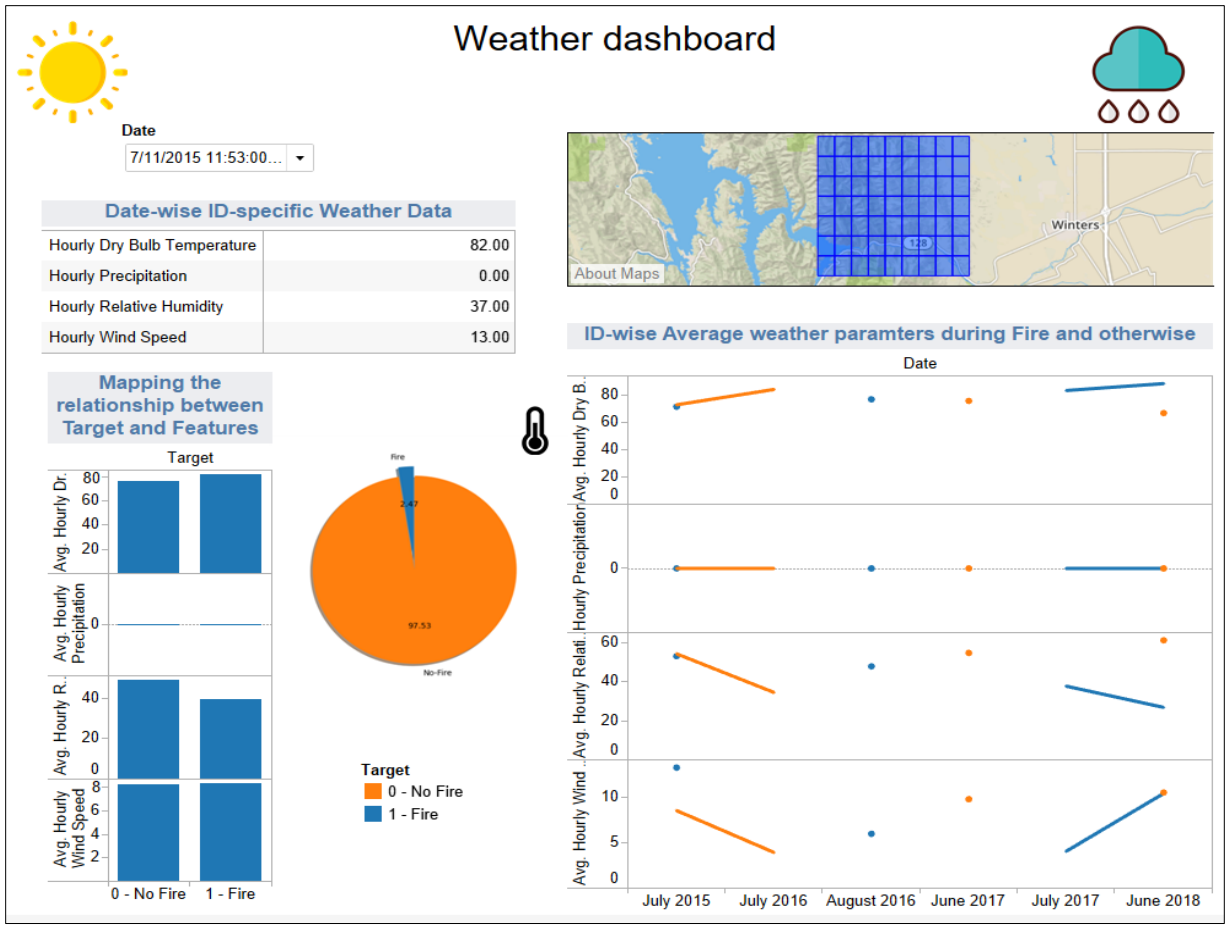


Figure 6.21. *Weather Dashboard*

6.5.6 Results Visualization

The Figure 6.22 shows the final model prediction results for the selected date filter. We can also see the parameters for each grid which resulted in the fire prediction.

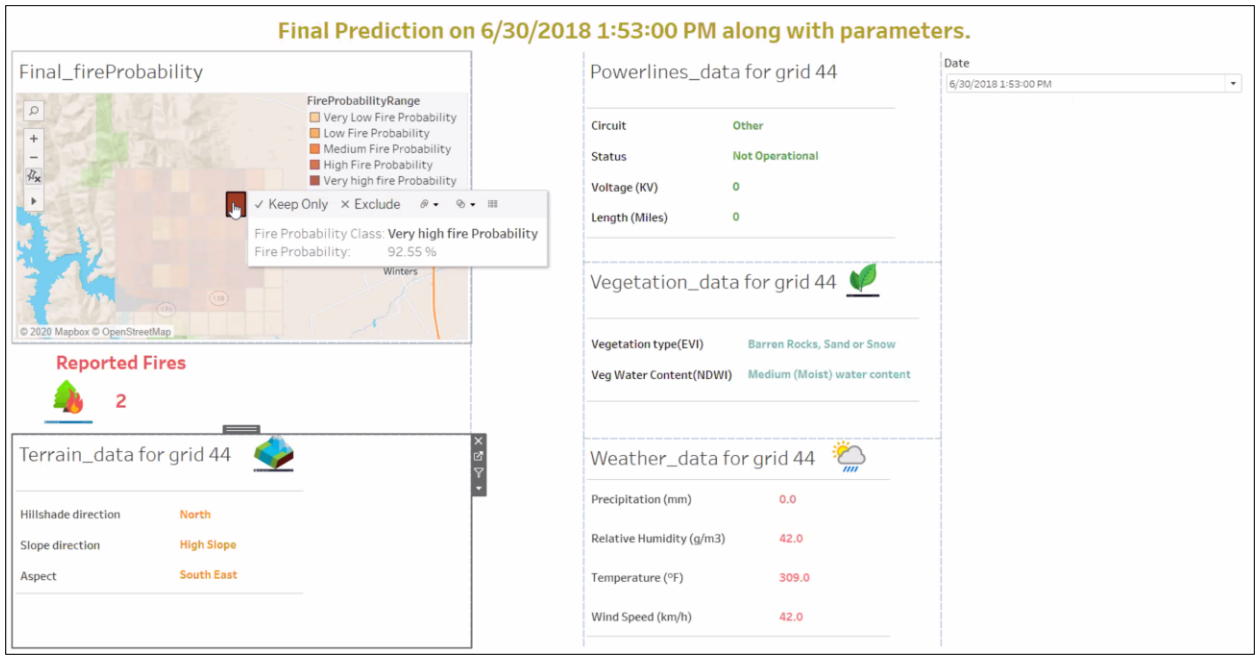


Figure 6.22. *Results Dashboard*

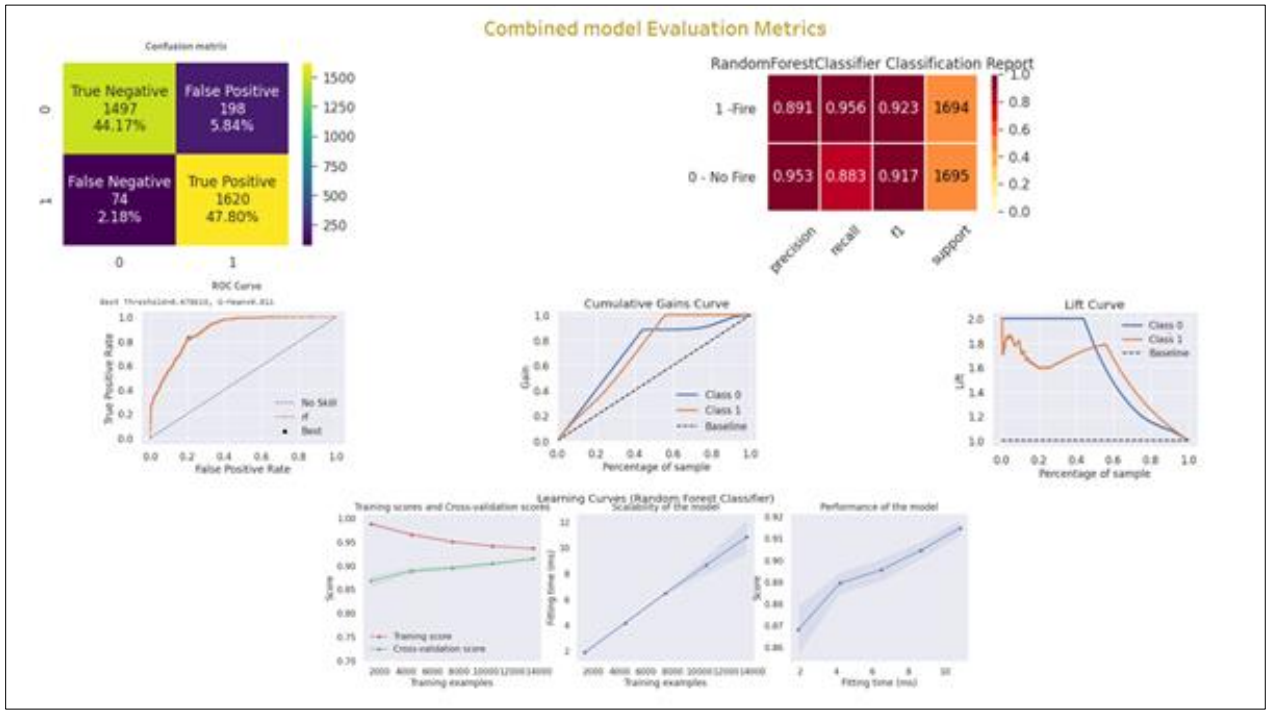


Figure 6.23. *Final Model Evaluation Results*

Figure 6.23 shows the various types of model evaluation results which are suitable for classification problems. Each one is described as follows:

- The confusion matrix is a tabular representation of the performance of classification models by analyzing the model based on the number of true and false predictions.
- Classification Report is based on confusion matrix numbers and shows the percentage for Precision, recall and F1 Score for all classes.
- ROC Curve plots False positive rates vs True positive rates. It is deemed as an essential tool for comparing and tuning models based on threshold values. Area under the curve is a true mark of model efficiency and its ability to classify.
- Cumulative gain curve [71] evaluates the model performance of the best model against a model highly random in feature selection.
- Cumulative Lift curve [72] measures the effectiveness of a predictive calculated as the ratio between the results obtained with and without the predictive model. It evaluates the likelihood of target or prediction probability of the model against a random data used as a baseline model.
- Learning curves include 3 graphs which display the comparison between training and cross validation scores, scalability of the model and performance of model. They are immensely helpful in evaluating the feasibility of the model.

6.5.7 User Interface (UI) visualizations

Below are the screenshots of the user interface and the final dashboards. Using flask, we built a front-end interface powered by the backed model. Final model was the combined model. The above visualizations were refined to create succinct and interactive visualizations in Tableau. The same was imported to the UI and displayed to the users.

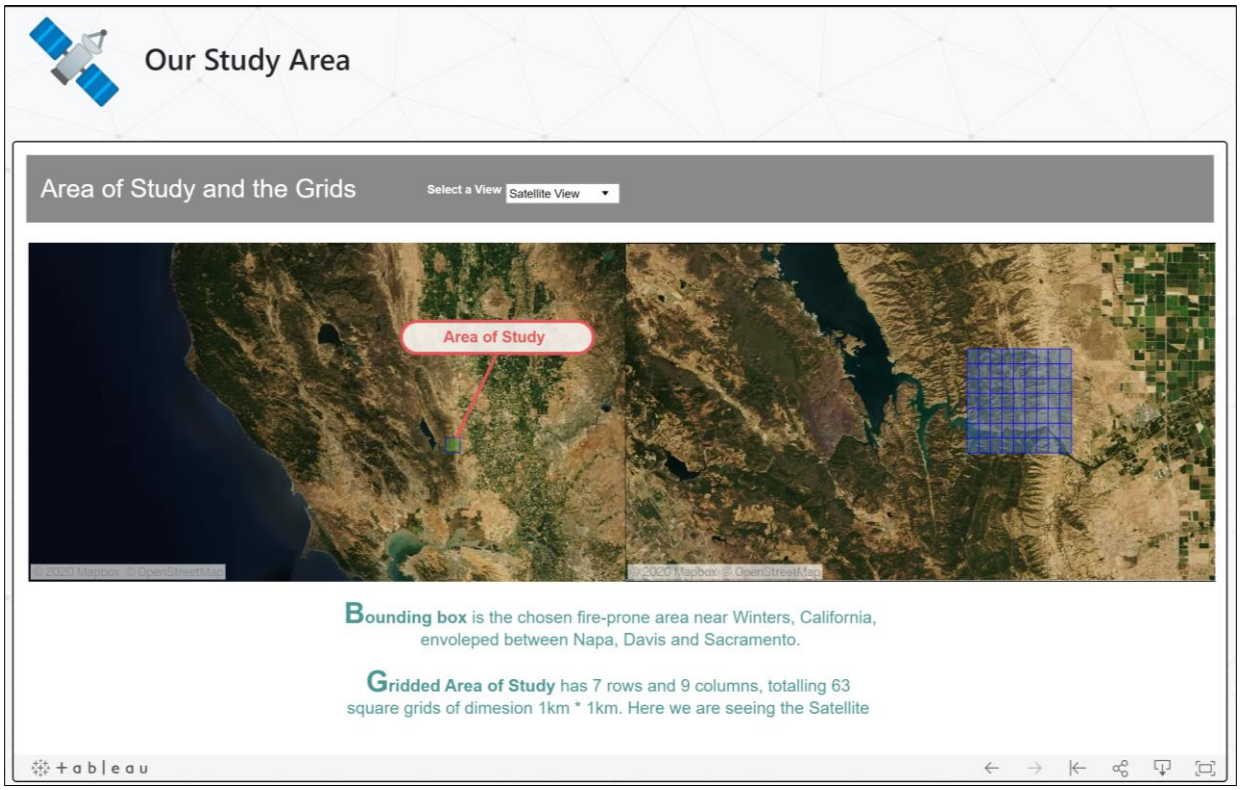


Figure 6.24. *Study Area and grids – Satellite View*

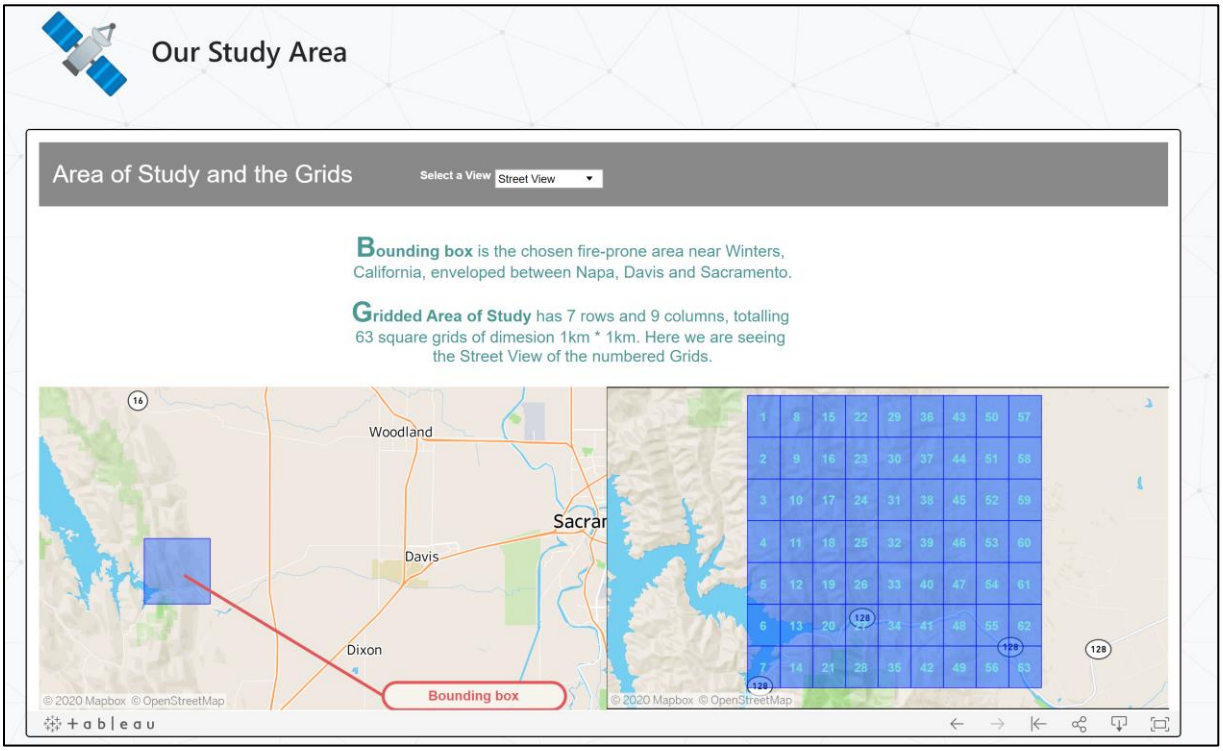


Figure 6.25. *Study Area and grids – Street View*

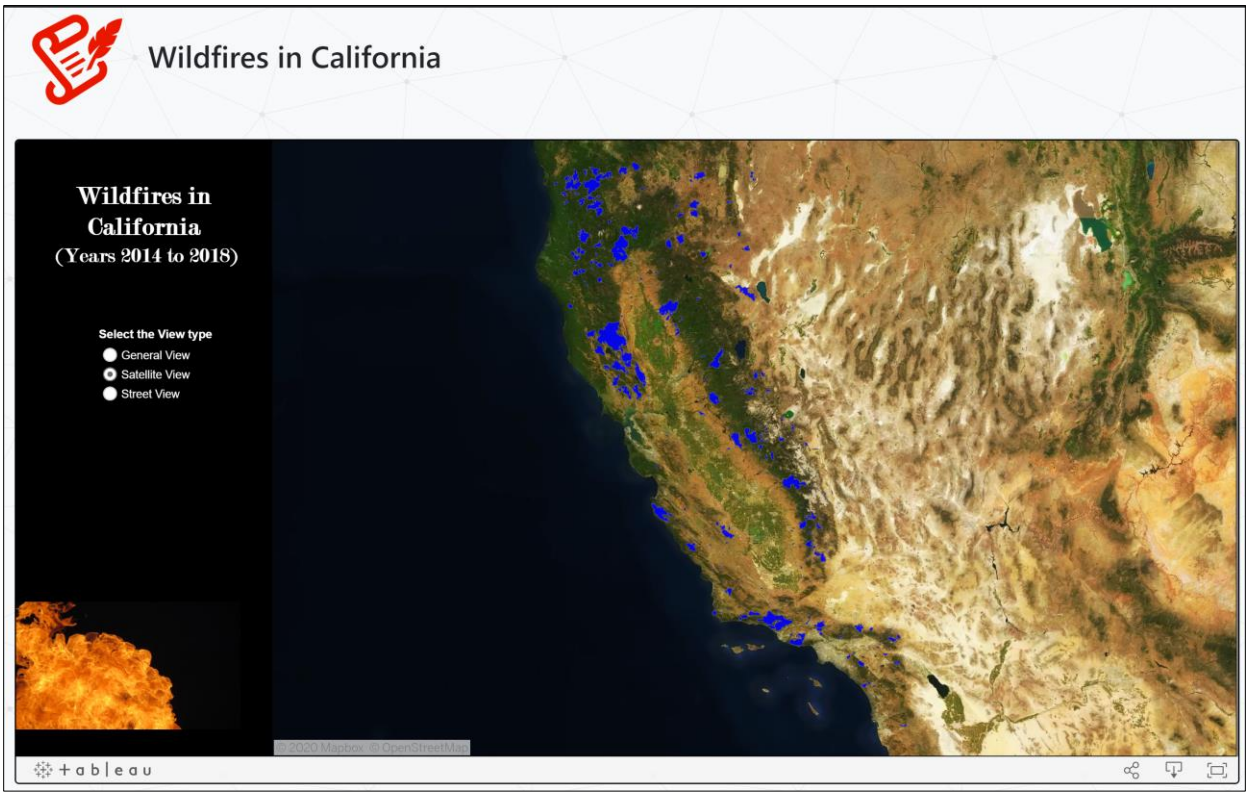


Figure 6.26. *Fires in California – Satellite View*

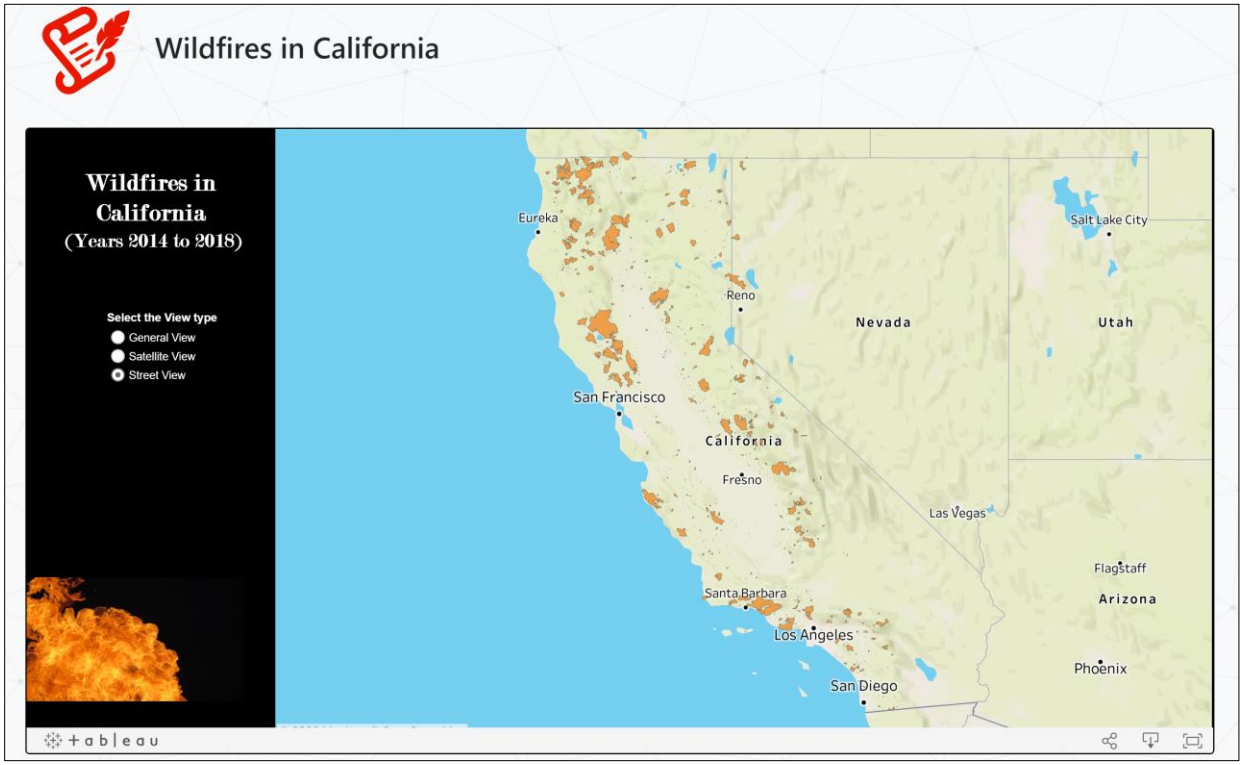


Figure 6.27. *Fires in California – Street View*

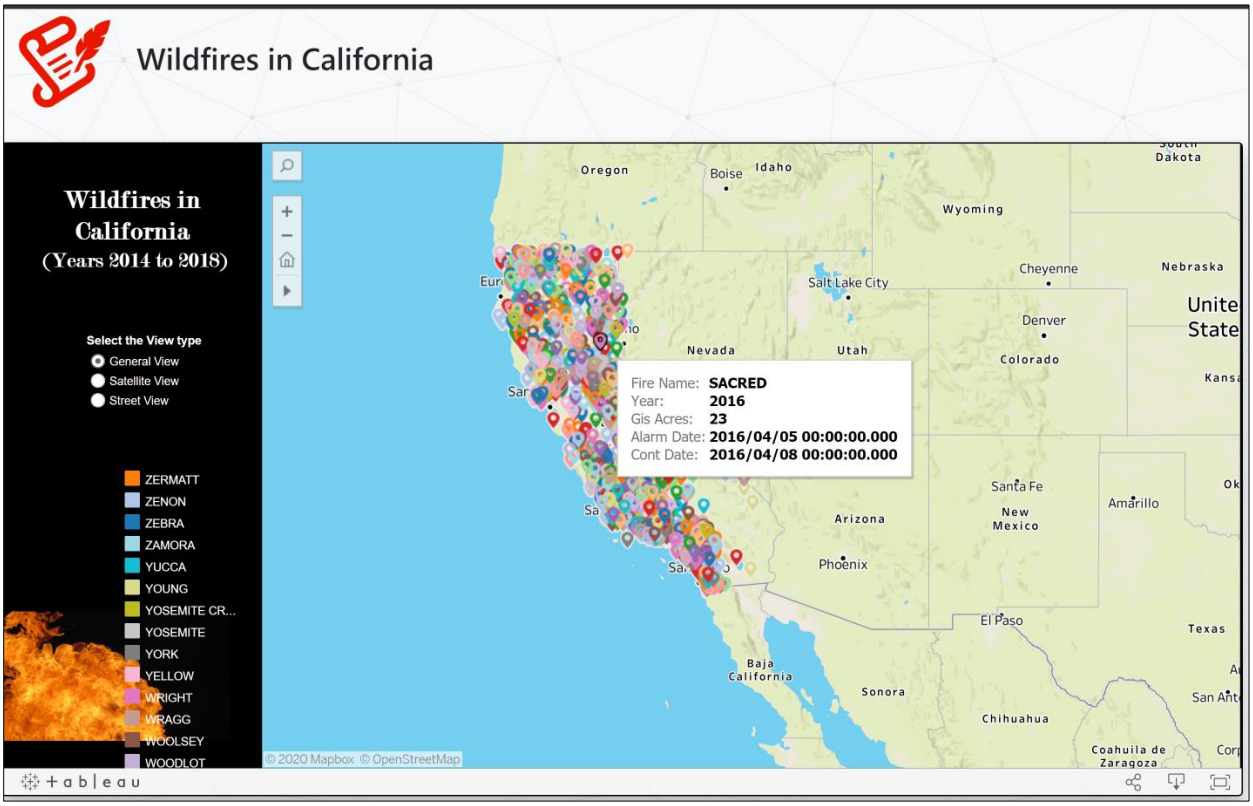


Figure 6.28. *Fires in California – General View*

Figures 6.24 and 6.25 shows the satellite and street view of the study area and grids in the study area. Figures 6.26, 6.27 and 6.28 show the fire history in California state from the years 2014 to 2018.

Figures 6.29 and 6.30 shows the satellite and street view of the fire history data along with the grids.

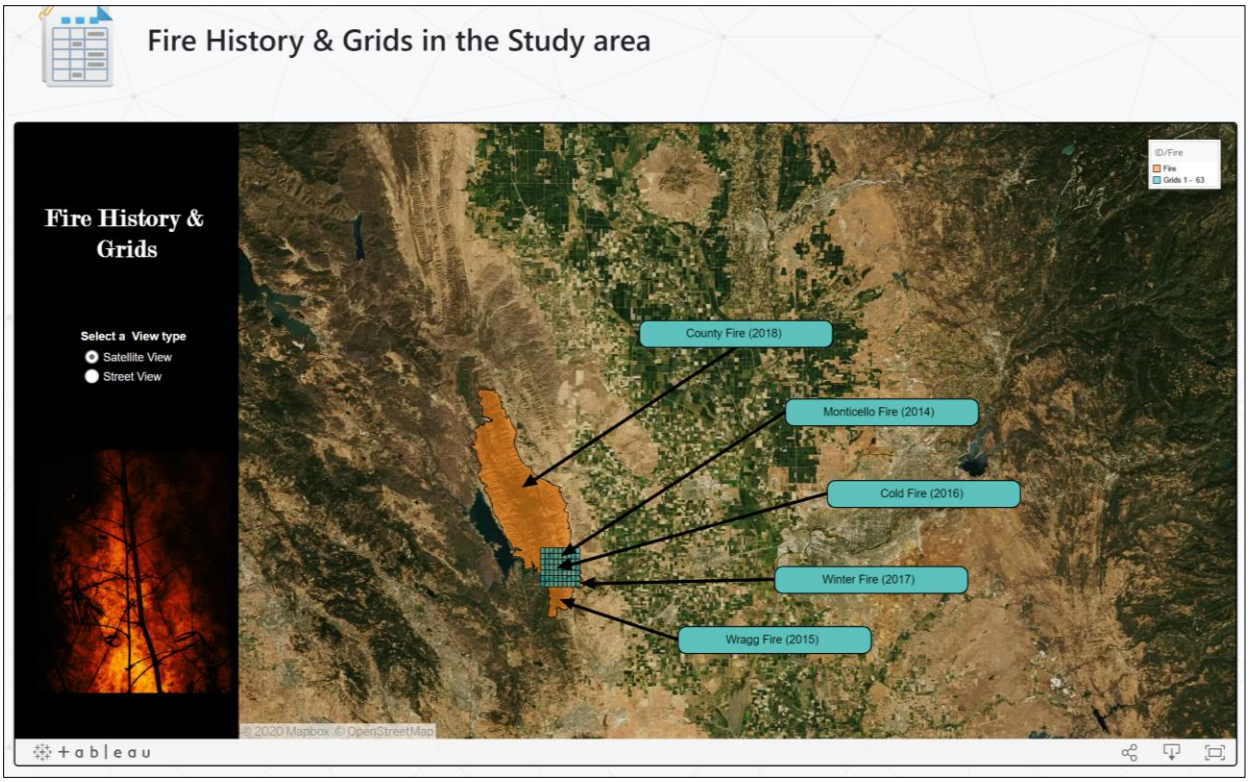


Figure 6.29. *Fire history and grids – Satellite View*

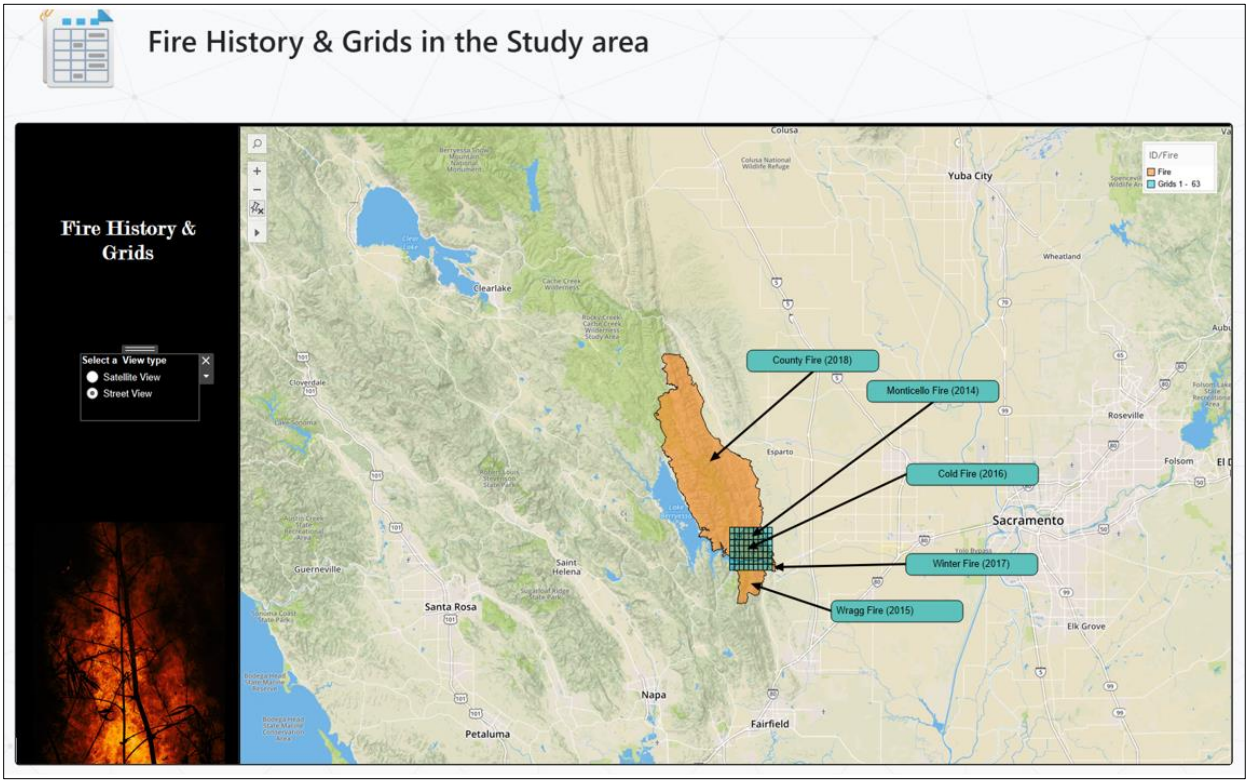


Figure 6.30. *Fire history and grids – Street View*

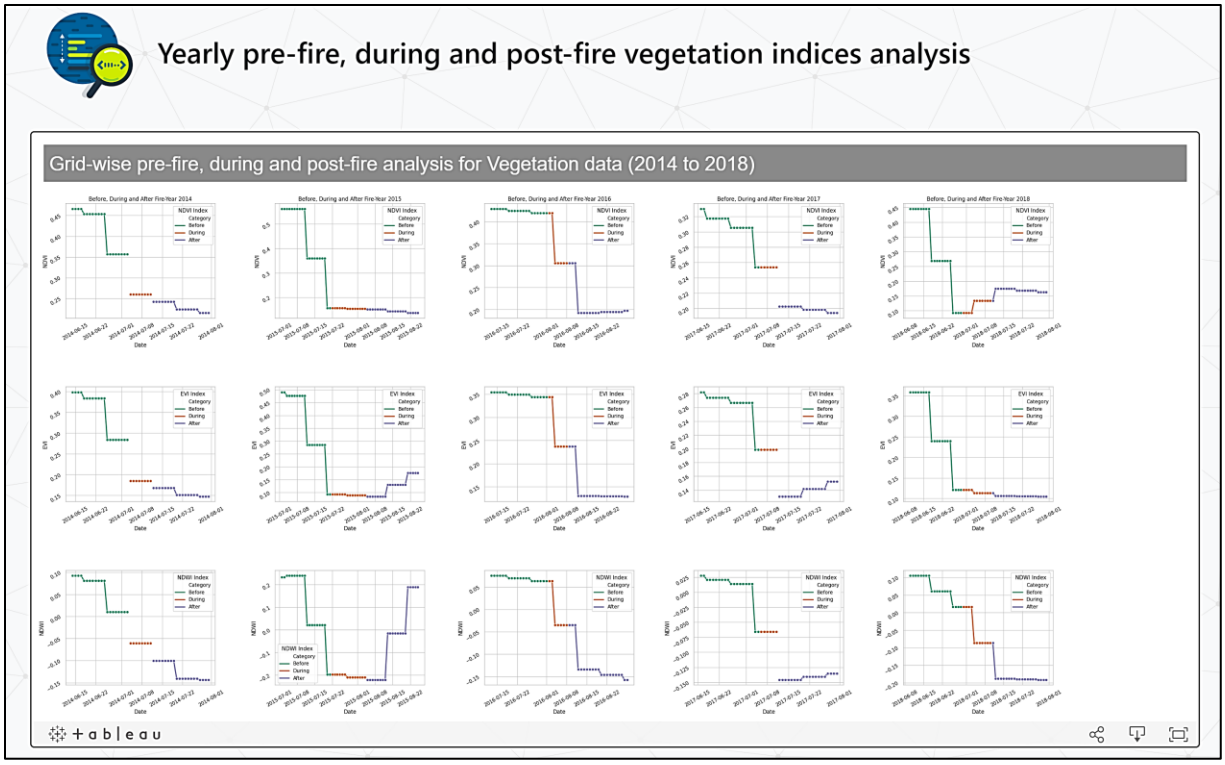


Figure 6.31. *Vegetation dataset analysis – Before, During and After fire (year 2014 to 2018)*

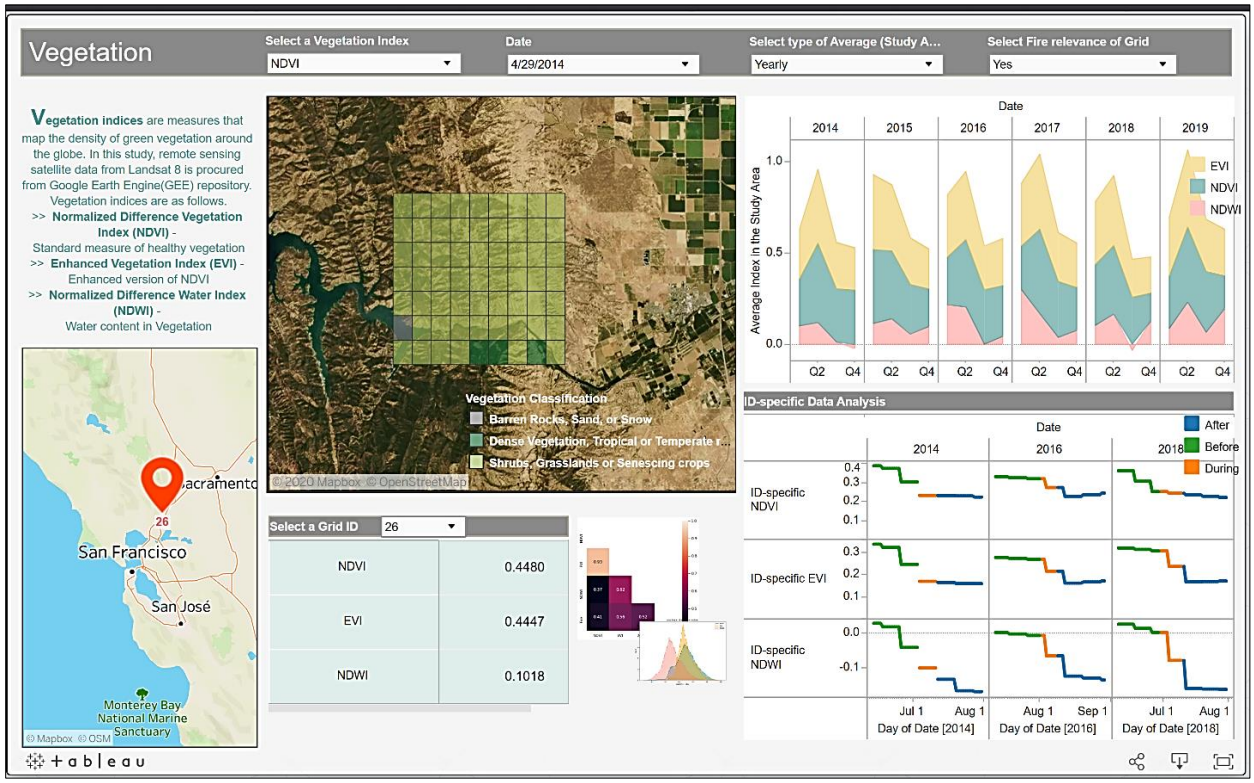


Figure 6.32. *Final Vegetation dashboard*

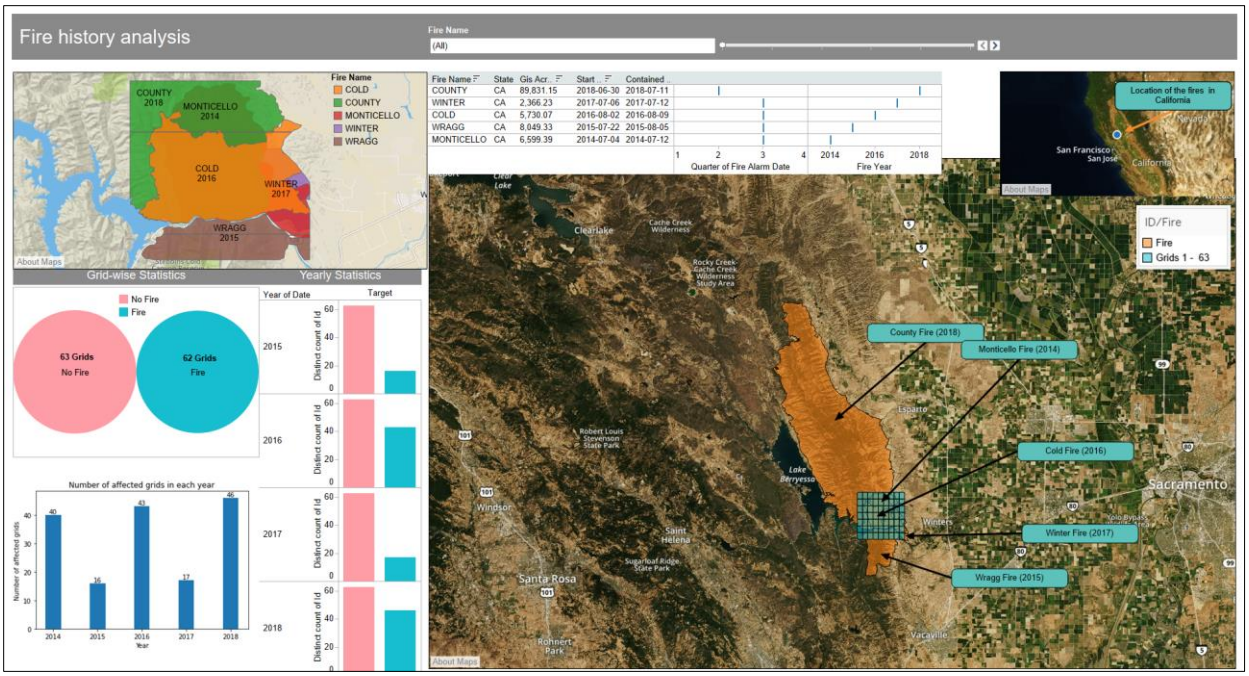


Figure 6.33. *Fire history analysis*

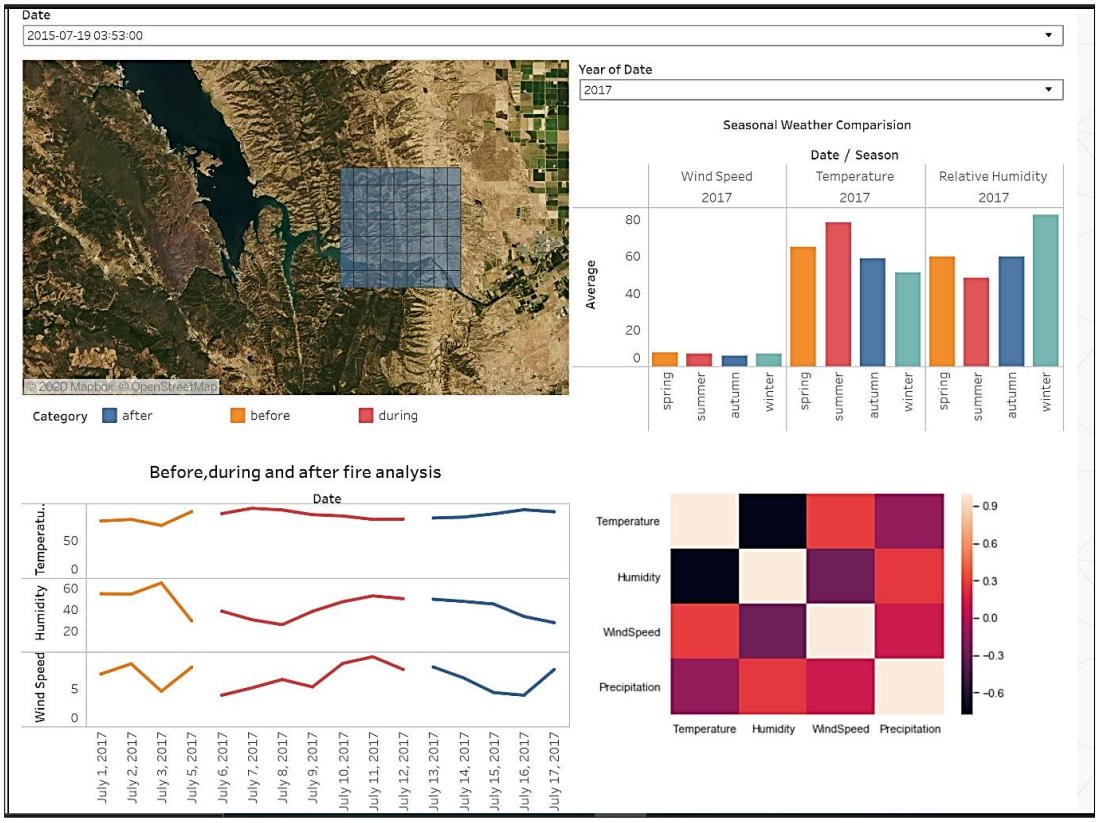


Figure 6.34. *Final Weather dashboard*

Figure 6.31 shows the vegetation dataset-based statistical analysis for the timeline before, during and after fire for the years 2014 to 2018. Figures 6.32, 6.33 and 6.34 display the final vegetation, fire history analysis and weather dashboard respectively. The visualizations were formerly explained in some of the previous chapters.

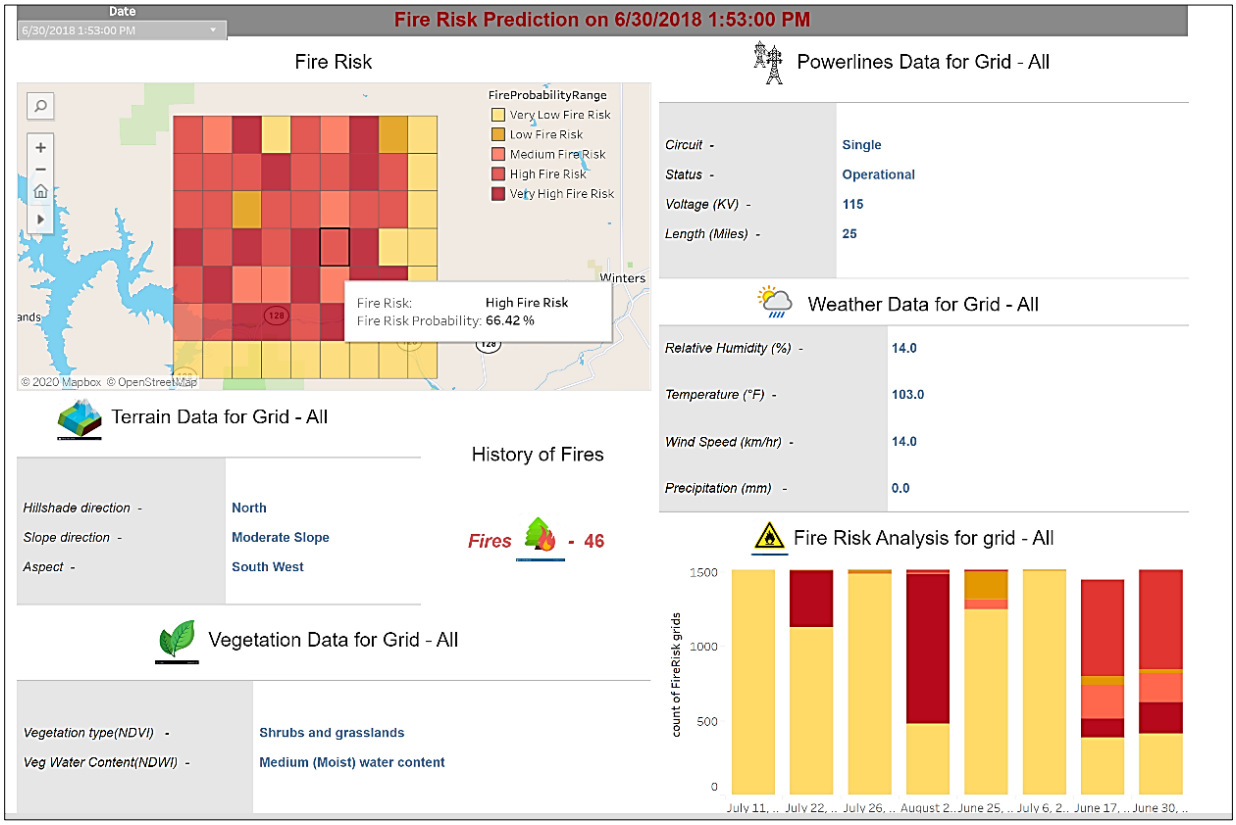


Figure 6.35. *Final Prediction dashboard*

Figure 6.34 shows the final prediction dashboard whereas figure 6.35 displays the evaluation metrics of the final combined model using Random forest classifier.

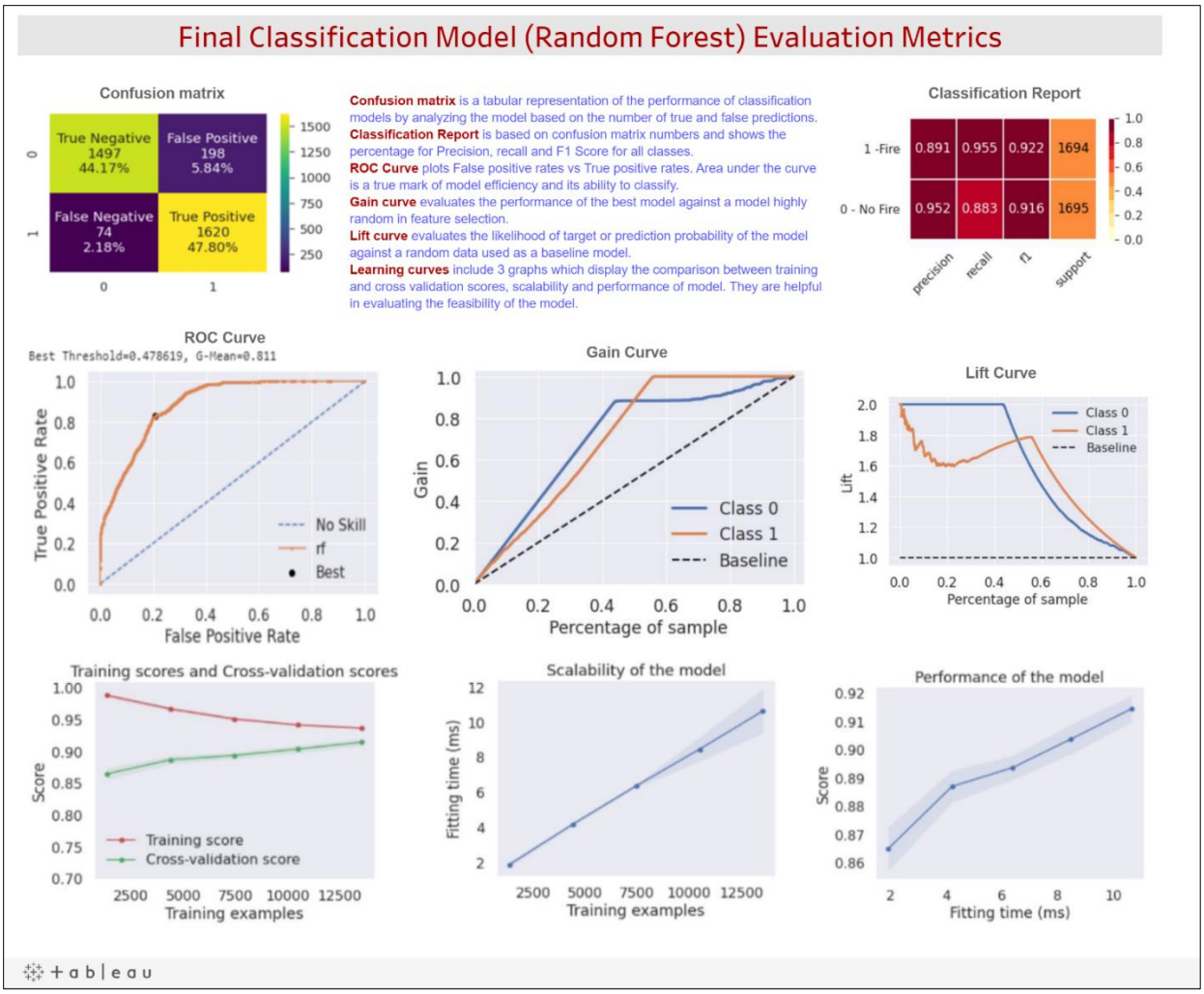


Figure 6.36. *Evaluation metrics dashboard*

7. CONCLUSION

7.1 Summary

In this system, we model the precursor fire conditions with an integration strategy to mimic the wildfire dynamics leading to a robust data-driven fire prediction system named 'Spartan Wildfire Prediction system (SWiPS)'. It comprises a powerful backend algorithm connected to a front-end interface using Flask that displays the data analyses and results of our projects along with an add-on option for expert users to determine the fire probability at a specified location.

Figure 7.1. *Homepage of the User Interface*

Figure 7.2 *SWiPS logo*

System name	Country	Purpose				Data Acquisition				Approach			Methods	
		Prediction	Detection	Simulation	Management	Satellite	Sensor	Manual	Camera	Data-driven	ML	IR	Mathematical	ANN
MFFDI	Australia	√	×	×	×	×	√	×	×	√	×	×	√	×
NFDRS	US	√	×	×	×	√	√	√	×	√	×	×	√	×
CFFDRS	Canada	√	×	×	√	×	√	√	×	√	×	×	√	×
FFRFS	Japan	√	×	×	×	√	√	√	×	√	√	×	×	√
NCMSSD	Russia	×	×	×	×	√	×	×	×	√	×	×	√	×
SWiPS	(Ours)	√	×	×	×	√	√	×	×	√	√	×	√	√

Figure 7.3. *Comparison of our system with existing systems*

Figures 7.1 and 7.2 show the homepage and logo of the final user interface. Figure 7.3 compares our system with the existing systems in countries such as Australia, US, Canada, Japan and Russia. SWiPS is a data-driven robust comprehensive Fire Risk prediction system that uses satellite-based vegetation data and sensor-based weather data along with powerline and terrain data to create a hybrid model based on mathematical formulae, machine learning and neural networks.

Below is the brief overview of the subsets of data and final algorithm used before the model was generated. Vegetation model is a standalone model whereas the remaining parameters were closely related and combined to create an integrated model. Combining vegetation data with weather, terrain and powerlines we created a combined dataset and used this for training the "Combined" Machine learning model. Although the models were stacked using an ensemble model, we found that a tuned Random Forest Classifier fared better with the combined dataset. The steps followed are as follows.

- For vegetation study, we fetched Vegetation indices from Landsat 8 using Google Earth Engine (GEE) before preprocessing the data.

- For weather, we collected hourly data from Local Climatology Data (LCD), maintained by National Centers for Environmental Information (NCEI).
- For terrain, we collected ⅓ arc second Digital Elevation Map (DEM) data from the United States Geological Survey (USGS) database.
- For powerline, we gathered powerline shapes data from California Energy Transmission ArcGIS public database.

When all the datasets were merged, year 2014 was dropped due to lack of weather data. We also fine-tuned our dataset based on feature importance. Random forest, Ada Boost, Long Short-Term memory (LSTM) and Gradient boosting algorithms were used in a combined model. This model was validated using a new dataset from another area. Accuracy of the combined model was 70% whereas the accuracy of the validation dataset was 50%.

7.1.1 Vegetation

7.1.1.1 Data collection

For vegetation study, our team perused numerous research studies and evaluated the usage of remote-sensing satellite data for determining the density of vegetation in a region. Likewise, we analyzed the formulae as well as applications of various vegetation indices in the initial phase. Normalized Difference Vegetation Index (NDVI), Enhanced Vegetation Index (EVI) and Normalized Difference Water Index (NDWI) were selected as the possible indices to concentrate on. Simultaneously, we checked the feasibility, calculations and data collection of these indices.

We contacted commercial data providers, who put a hefty price on high resolution data of exceptional spatial and temporal resolution. Nevertheless, we considered free and reliable sources of acceptable resolution such as Landsat 8, Proba V powered by IBM and Moderate Resolution

Imaging Spectroradiometer (MODIS) from National Aeronautics and Space Administration (NASA) Terra and Aqua satellites. Landsat 8 launched by NASA and United States Geological Survey (USGS). We pre-processed a small subset of data juxtaposed the resolution, feasibility and ease of processing of each of the data sources before zeroing in on Landsat8. Null values were imputed using the neighboring values during an exhaustive pre-processing and statistical analysis using visualizations and metrics.

7.1.1.2 Model creation

Multiple experiments with various algorithms were performed before deciding on Random forest classifier. The data was unbalanced with a large number of negative targets. Hence, we used Synthetic Minority Oversampling technique (SMOTE) to generate synthetic samples to recreate a similar feature for the target of '1' or 'Fire'. The vegetation indices NDVI, EVI and NDWI of the synthetic samples correspond to fire-prone regions. Thereafter, the model was built and cross-validated with 5 segments, wherein different segments are interchangeably used as training and testing sets before deciding on the final segment.

7.1.1.3 Model Evaluation

With a model accuracy of 75%, the Classification report reveals a precision of 77% precision along with 75% f1-score and 77 as the support value for positive test data. Negative data has a precision of 74% precision along with 76% f1-score and 76 as the support value for positive test data. Recall for 'no-fire' test data is 73% and 78% for 'fire' test data. We got 21 false negative results and 17 false positive results, which means that our model does not lean towards a specific type of error. 59 and 56 are true positive and true negative results. The ROC curve was plotted and carefully evaluated, with a best threshold of 0.491667 and a G-Means of 0.811.

7.1.2 Weather, Powerline and Terrain

7.1.2.1 Data collection

After literature review and studying conventional systems, we decided to combine weather, powerline and terrain data together to train our model. The theoretical reason for this decision is: terrain data on its own does not provide enough information on wildfire prediction, but it is a major factor in modeling wind direction and micro weather conditions. Powerline data faces the same issue of lacking temporal information since it does not change frequently. But powerline related fire usually occurs during windy weathers. So, we combine these 3 datasets together.

For weather, we got our data from Local Climatology Data (LCD), maintained by National Centers for Environmental Information (NCEI) which provides atmospheric and geospatial data across the United States [18]. We picked 4 weather parameters: temperature, wind speed, humidity and precipitation, which were also key predictors used in Fire Weather Index (FWI).

For terrain, we used Digital Elevation Models (DEM) provided by the United States Geological Survey (USGS). We derived slope, hill shade and aspect from DEM. We calculated the zonal statistics and got aspect ranges, hill shade directions and slope ranges to represent different terrain conditions.

For the powerline, we got a powerline map from the California Energy Commission and picked voltage, circuit, whether the powerline is operational or not and the length of the powerline as our parameter. We combined them into a single dataset as our training set, merged by cells. We used a Random Forest Classifier with hyperparameter tuning. We used bootstrapping with 200 estimators and a random state of 42. We did not limit a maximum depth, but restricted minimum samples split to be 2. Test result gave us 80.91% accuracy.

7.1.2.2 Model Evaluation

Classification report shows balanced results with 0.82 precision and 0.82 f1-score with 1834 support for positive and 0.80 precision and 0.82 f1-score with 1834 support for negative test data. Recall for ‘no-fire’ test data is 80% and 82% for ‘fire’ test data. We got 330 false negative results and 370 false positive results, which means that our model does not lean towards a specific type of error. The ROC curve was plotted and carefully evaluated, with a best threshold of 0.478691 and a G-Means of 0.811.

Also, the above-mentioned individual vegetation and weather models were stacked and used the output from these models as input to the stacked ensemble model which fetched us 83% model accuracy.

Classification report for the ensemble model gave results with 0.84 precision and 0.89 f1-score with 1695 support for positive and 0.80 precision and 0.68 f1-score with 725 support for negative test data. The ROC curve was plotted and carefully evaluated, with a best threshold of 0.499112 and a G-Means of 0.866.

7.1.3. Combined model

With the weather, vegetation, power lines and terrain data being major parameters used in predicting fire risk, we tried different combinations of the data and also models to get the best fit.

Dataset with all the above-mentioned parameters together was used to build one variation of the Machine learning model. We called it a combined model which could fetch a model with 92% accuracy. This is by far the best model we got in terms of the evaluation results.

Classification report for the combined model shows balanced results with 0.95 precision and 0.92 f1-score with 1695 support for positive and 0.89 precision and 0.92 f1-score with 1694 support for negative test data. The ROC curve was plotted and carefully evaluated, with a best threshold of 0.478619 and a G-Means of 0.811.

7.1.4 Outcome

We were successful in assimilating data and modeling all relevant parameters and metrics related to Vegetation, Weather, Terrain and Human activities -Powerline to build a robust and reliable fire prediction system named 'Spartan Wildfire Prediction system (SWiPS)'. The final model has an accuracy of 92%. It was built using a combined model powered by Random Forest algorithm. The user interface displays grids, data analyses and results of this project. An additional option is exclusively available for expert users, such as researchers and fire management personnel. It enables fire evaluators to input metrics, choose algorithms and determine the fire probability at a prescribed location with a map overlay in the background.

In this research, we studied the individual and combined influence of 4 features: vegetation, weather, terrain and powerlines to wildfire occurrences. We achieved 3 main results:

- Analyzed the importance of each feature to the occurrence of wildfire.
- Trained spatial specific algorithm for wildfire prediction with 92% accuracy and low false-negative results.
- Built a wildfire prediction alert system.

In chapter 1, we reviewed existing models for wildfire prediction. After thoroughly studying both the fire study aspect and data-driven model aspect, we discovered that current

systems lack spatial and temporal accuracy in wildfire prediction. In chapter 2, we explored multiple data sources. Based on our previous research, we determined that 4 factors were the top influencers in wildfire prediction: weather, vegetation, terrain and powerlines. In chapter 3 and 4, we conducted feature engineering to each data source and converted them into machine-learning ready data. In chapter 5, we experimented with both individual factor models and combined factor models and reached a 92% accuracy in predicting the wildfire occurrence for a specific location and time.

This research successfully solved past issues of lacking spatial accuracy in wildfire prediction. Our model was able to pinpoint an incident within 1 x 1 km spatial accuracy and 1-hour temporal accuracy. This will greatly improve wildfire prediction, planning and resource management for the California Fire department.

7.2 Benefits and Shortcoming

The benefits of our research are as follows.

- We established a machine learning-ready database for future studies.
- We explored the key impacting features for wildfire prediction from a machine learning point of view.
- We tested multiple hypotheses and ruled out the wrong ones.
- Our model reached cut-edge accuracy and near real-time effectiveness.
- We established a model selection, feature extraction and model building workflow for location specific wildfire prediction research.

The shortcomings of our solution are:

- Solution relies heavily on the quality of data and is unsuitable for cases with limited data.

- Our model is location specific; it cannot be generalized to all areas.

7.3 Potential Model Applications

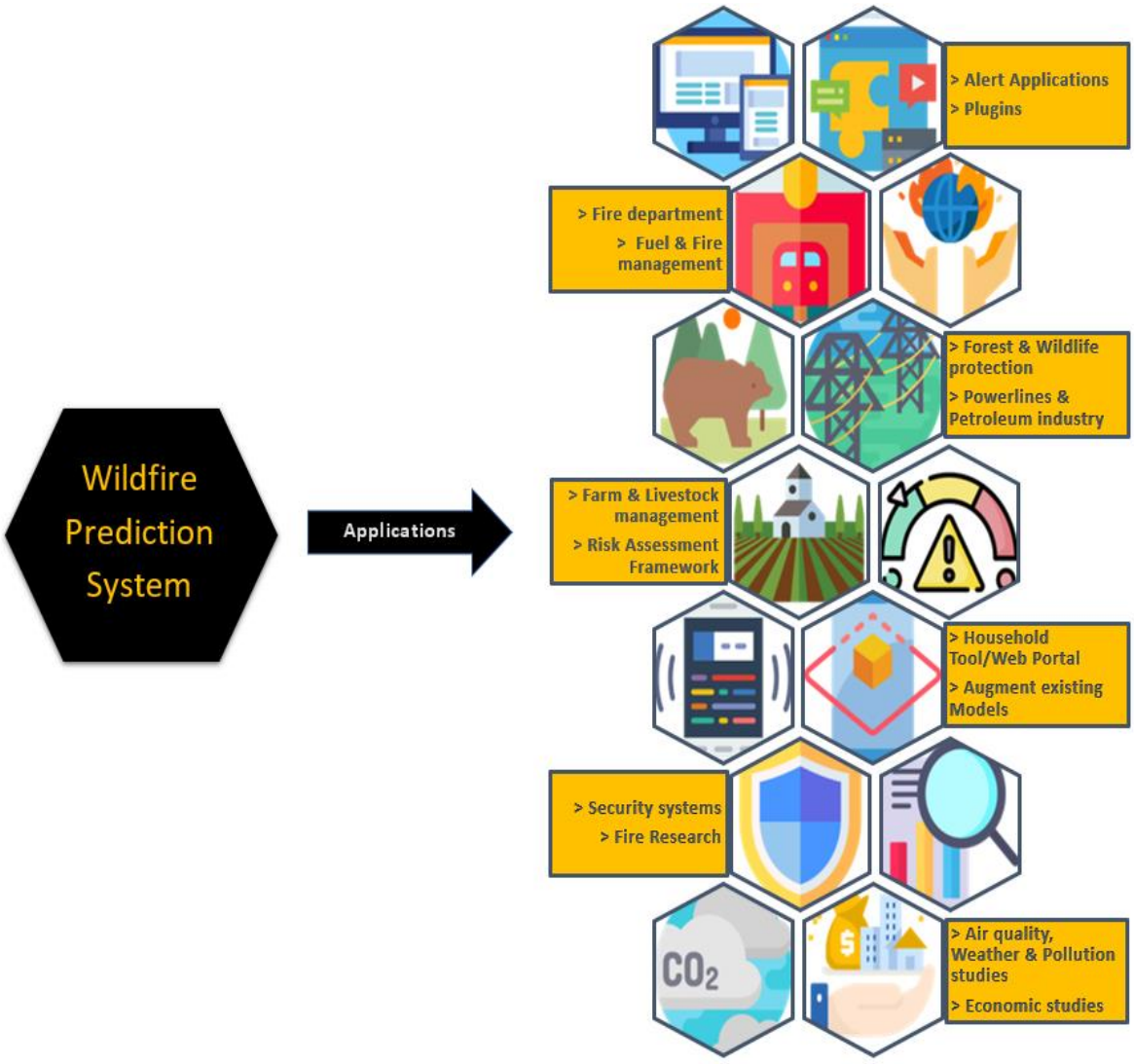


Figure 7.4. *Applications of Wildfire Prediction system*

- **Alert application creation:** The advanced modelling system can be used for a wildfire app(application) creation, accessible from devices such as mobile phones, kiosks or computers.

- **Plugin for applications:** Travel, vacation, hiking and social media applications can integrate Wildfire prediction in their native apps.
- **Fire department:** Firefighters and the fire department can greatly benefit from the adoption of a precise and accurate modeling system that provides focused spatial data analyses with commendable temporal resolution.
- **Fuel and fire management planning:** Quintessential component of fuel and fire management planning, both government and private agencies can be beneficiaries of a scientific approach to Wildfire prediction. For example, local fire management authorities need to plan activities such as prescribed fires to reduce the density of fuel.
- **Forest and Wildlife protection:** Organizations involved in forest and wildfire preservation can utilize the system to predict the likelihood of fire and mitigate the fire event.
- **Powerline and Petroleum companies:** The system can be used to monitor or relocate Power Lines or oil and gas reserves from fire-prone regions.
- **Farm and livestock management:** Precautionary safety measures and timely relocation will be possible with an accurate Wildfire prediction system.
- **Risk assessment frameworks:** The system serves as a vital component in risk assessment frameworks and identification of fire-prone areas by Government organizations. It harmonizes the information needed to enable inter-agency collaboration on wildfire prevention and containment in both intra-boundary and trans-boundary fire events. [73]
- **Household tool (web portal):** Provides a viable and easy-to-use wildfire prediction system for the general population.

- **Augment existing modelling systems:** Models, backed by real-time satellite data and precise scientific parameters, provide great flexibility with multiple modelling options and faster model execution. Hence, it can be used to fortify or supplement existing risk assessment and modelling systems.
- S**ecurity systems:** Powerline danger zones can be monitored and preventative measures as well as distancing protocols can be enforced by security agencies. Further, the system can be used to inform citizens and curb human activities in fire-prone areas.
- **Fire research:** Enable evaluations in Fire, fuel and smoke science programs and research studies that investigate the fire dynamics, behavior and spread.
- **Air quality, weather and Pollution studies:** The system can serve as a predictor of an upcoming descent in air quality and pollution. Further, it can also predict a change in weather, such as temperature in these regions.
- **Economic studies:** The prediction system can supplement and inform economic studies as an indicator of economic volatility.

The system, as shown in figure 7.4, aids in intelligent data-driven decision making, backed with multiple and reliable machine learning algorithms. Further, it will greatly augment preparedness and response time, and minimize fatalities, loss of property and destruction of flora and fauna.

7.4 Lessons Learned

7.4.1 Data collection

- In this era of cloud computing, we can directly access Satellite data in cloud catalogs such as Google Earth Engine (GEE) instead of downloading the data and extinguishing storage space.
- Although remote-sensing Satellite data is highly complex, we can maneuver it using Python Application programming interface (API).
- Always research and check feasibility of data collection, processing and modelling before deciding on the parameters.
- Although there are many sources of satellite data available, most of them are not publicly available and there were commercial sites with better clarity and frequency of data.
- Null value imputation will fortify the dataset at hand by interpolating the null values based on neighboring values. Visualizing data is an ideal method for statistical analysis of data. It is easier to isolate anomalies.
- QGIS a Geographic information system software with Python plugin, better analyze and edit spatial information in addition to composing and exporting geographical maps.
- For sensor data quality assurance procedure, there are 3 main possible failures to be looked out for: mechanical failure, data transformation failure and collection failure. They exhibit different statistical patterns and should be checked carefully.

7.4.2 Model creation

- Hyperparameter tuning is the holy grail of model performance improvement. Various techniques like Gridsearchcv can be used which makes the process easier.

- Creating subsets of data and tweaking the parameters for modelling experiments can improve the modeling results.
- Visualizations help in precisely gauging and juxtaposing the efficacy of the models.
- Always check for overfitting and underfitting.
- Use a confusion matrix to determine the characteristics of your prediction results. Whether it prunes towards generating false-negative errors or false-positive errors. Adjust your model according to your research purpose.

7.4.3 User interface generation

- Folium can be used for map generation and plotting geological data in Python.
- Tableau can be used to build comprehensive dashboards with exceptional granular resolution. Tableau and GIS can easily process geological data.
- We learned the nuances of creating a User Interface (UI) in the form of a webpage as well as integration of Python-based models, visualizations and tableau plots in the final UI.

7.5 Recommendations for Future Work

Our recommendations are as follows.

- Fire behavior and spread studies using relevant metrics and fire history data of greater temporal resolution is an ideal next step to this undertaking.
- Fuel and fire dynamics can be explored further to isolate and formulate their relationship before modelling the same using Machine learning.
- Additional metrics, parameters and algorithms can supplement and possibly augment the model at hand.

- If funding is available, commercial data sources, with exceptional spatial and temporal resolution, can be accessed. For example, Satellite data from Harris Space and Intelligence systems has a higher resolution.
- Different parameters like Lighting, Debris, Campfire, smoking can be considered if relevant data is made available by concerned authorities.
- Data quality assurance standards are important for studying traditional areas with machine-learning methods.

7.6 Contributions and Impacts on Society

Wildfire prediction system is a reliable, fast and flexible modeling system with extensive social applications. Figure 7.5 highlights the various implications of the prediction system.

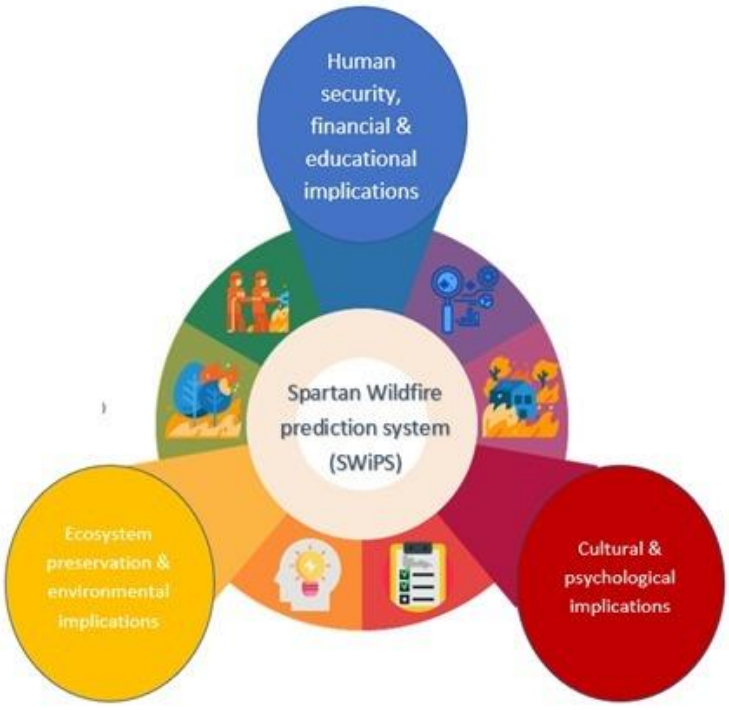


Figure 7.5. *Implications of prediction system*

The educational, cultural and socio-economic contributions and impacts are cited below.

- ➔ The system will fortify the planning, management and containment efforts by Government and private agencies.
- ➔ Data-driven intelligent decision making will ensure safety and security of all stakeholders, prevent fatalities and minimize loss of property.
- ➔ Enhanced organization, funding and preparedness can be coordinated at local, regional, national and global scales, with enhanced cooperation between agencies at each level.
- ➔ Adoption of the system will help minimize the damage to our ecosystem, flora and fauna alike.
- ➔ Concerned agencies such as fire department, fire management personnel and healthcare professionals can improve their response time and augment fire event preparedness.
- ➔ The web portal, which can be developed into an application or plugin, can educate the general population and lead them to safety.
- ➔ From a psychological standpoint, preparedness will reduce anxiety, avoid trauma and promote social wellbeing.
- ➔ From a research standpoint, the accurate prediction with commendable spatial and temporal resolution will simplify the complicated problem of prediction.
- ➔ The system has the potential to trigger further educational advancements in the field of fire dynamics, behavior and spread study.
- ➔ From a cultural standpoint, we envision and anticipate a cultural shift towards data-driven technologically-advanced scientific systems of Wildfire prediction.

➔ A precise system will advocate and inspire a planned approach towards prevention and management of other potentially dangerous natural and predictable disasters.

This holistic modeling system, backed by a host of machine learning algorithms, integrates multiple metrics and indices related to weather, vegetation, terrain and human activities, such as powerlines, to build an accurate and robust prediction system.

Given the vast social applications, this undertaking can be deemed vital for our society.

References

[1]. Morgan Hill Life Editorial. (2019, April 24). Editorial - Wildfire season is becoming a year-round phenomenon. Retrieved from https://morganhilllife.com/2018/08/24/editorial-wildfire-season-is-becoming-a-year-round-phenomenon/.

[2]. Phillips, C. (n.d.). Climate change is creating catastrophic wildfires. Retrieved from https://www.weforum.org/agenda/2019/05/the-vicious-climate-wildfire-cycle.

[3]. California's worsening wildfires, explained. (n.d.). Retrieved from https://calmatters.org/explainers/californias-worsening-wildfires-explained/.

[4]. Guoyan, Dr. Xu, Forest Fire Prediction: Big Data and Machine Learning Approaches. Submitted for Publication

[5]. California Department of Forestry and Fire Protection. (n.d.). Stats and Events. Retrieved from https://www.fire.ca.gov/stats-events/.

[6]. Watts, A., Halla, T., North, M. of the, Erren, H., JoeShaw, MilwaukeeBob, … SteveTa. (2019, January 26). California's Wildfire History – in one map. Retrieved from https://wattsupwiththat.com/2019/01/26/californias-wildfire-history-in-one-map/.

[7]. S. W. Taylor, Douglas G. Woolford, C. B. Dean and David L. Martell, "Wildfire Prediction to Inform Management: Statistical Science Challenges", Statistical Science, Vol. 28, No. 4, Special Issue on Mathematics of Planet Earth (November 2013), pp. 586-615. Published by Institute of Mathematical Statistics Stable. URL: https://www.jstor.org/stable/43288437?seq=1#page_scan_tab_contents

[8]. Moritz, M. A., Morais, M. E., Summerell, L. A., Carlson, J. M., & Doyle, J. (2005, December 13). Wildfires, complexity, and highly optimized tolerance. Retrieved from https://www.pnas.org/content/102/50/17912.

[9]. Elements of Fire. (n.d.). Retrieved from https://smokeybear.com/en/about-wildland-fire/fire-science/elements-of-fire.

[10]. Information about the Fire Triangle/Tetrahedron and Combustion. (n.d.). Retrieved from https://www.firesafe.org.uk/information-about-the-fire-triangletetrahedron-and-combustion/.

[11]. (n.d.). Wildland Fire: What is a Prescribed Fire? (U.S. National Park Service). Retrieved from https://www.nps.gov/articles/what-is-a-prescribed-fire.htm.

[12]. E. Sapp, C. (2017). Preparing and Architecting for Machine Learning. Gartner.

[13]. Sakr, G. E., Elhajj, I. H., Mitri, G., & Wejinya, U. C. (2010). Artificial intelligence for forest fire prediction. 2010 IEEE/ASME International Conference on Advanced Intelligent Mechatronics. doi:10.1109/aim.2010.5695809

[14]. Satir, O., Berberoglu, S., & Donmez, C. (2015). Mapping regional forest fire probability using artificial neural network model in a Mediterranean forest ecosystem. Geomatics, Natural Hazards and Risk, 7(5), 1645-1658. doi:10.1080/19475705.2015.1084541

[15]. Ritaban, D., Ritaban Dutta Ritaban Dutta, Dutta, R., Ritaban Dutta CSIRO Data61, Das, A., Aryal, J., … Jagannath Aryal School of Land. (2016, February 1). Big data integration shows Australian bush-fire frequency is increasing significantly. Retrieved from https://royalsocietypublishing.org/doi/10.1098/rsos.150241.

[16]. Stojanova, D., Kobler, A., Ogrinc, P., Ženko, B., & Džeroski, S. (2011, March 2). Estimating the risk of fire outbreaks in the natural environment. Retrieved from https://link.springer.com/article/10.1007/s10618-011-0213-2.

[17]. California Department of Forestry and Fire Protection. (n.d.). Fire Perimeters. Retrieved from https://frap.fire.ca.gov/frap-projects/fire-perimeters/.

[18]. National Centers for Environmental Information, & Ncei. (n.d.). Data Tools: Local Climatological Data (LCD). Retrieved from https://www.ncdc.noaa.gov/cdo-web/datatools/lcd.

[19]. California Department of Forestry and Fire Protection. (n.d.). Camp Fire. Retrieved from https://www.fire.ca.gov/incidents/2018/11/8/camp-fire/.

[20]. Landsat Image Gallery - Camp Fire Rages in California. (n.d.). Retrieved from https://landsat.visibleearth.nasa.gov/view.php?id=144225.

[21]. Retrieved (n.d.).from https://modis.gsfc.nasa.gov/about/specifications.php.

[22]. Retrieved from https://www.usgs.gov/core-science-systems/ngp/tnm-delivery.

[23]. AppEEARS Team. (2019). Application for Extracting and Exploring Analysis Ready Samples (AppEEARS). Ver. 2.30. NASA EOSDIS Land Processes Distributed Active Archive Center (LP DAAC), USGS/Earth Resources Observation and Science (EROS) Center, Sioux Falls, South Dakota, USA. Accessed October 10, 2019. https://lpdaacsvc.cr.usgs.gov/appeears

[24]. El-Nesr, M. (2019, December 8). Filling gaps of a time-series using python. Retrieved from https://medium.com/@drnesr/filling-gaps-of-a-time-series-using-python-d4bfddd8c460

[25]. US Census Bureau. (2019, May 6). TIGER/Line Shapefiles. Retrieved from https://www.census.gov/geographies/mapping-files/time-series/geo/tiger-line-file.html

[26]. Prabhala, R. P. R. (2019, April 9). Role of Data Management (and MDM) in Analytics. Retrieved from https://www.intellitide.com/role-of-data-management-and-mdm-in-analytics/.

[27]. California Department of Forestry and Fire Protection (2018, November 9) Camp Fire Incident Update 🔲 AM Accessed November 9, 2018.

[28]. Canada, N. R. (n.d.). Background Information. Retrieved from http://cwfis.cfs.nrcan.gc.ca/background.

[29]. Stocks, B.J., B.D. Lawson, M.E. Alexander, M.E., C.E. Van Wagner, R.S. McAlpine, T.J. Lynham, D.E. Dube. 1989. The Canadian Forest Fire Danger Rating System: An Overview. Forestry Chronicle Vol. 65 issue 6 : 450-457.

[30]. Canada, N. R. (n.d.). Canadian Wildland Fire Information System: Canadian Forest Fire Danger Rating System (CFFDRS). Retrieved from https://cwfis.cfs.nrcan.gc.ca/background/summary/fdr.

[31]. Coastal Fire Center - hot topics in fire center on coast. (n.d.). Retrieved from https://www2.gov.bc.ca/gov

[32]. Li, Liming & Song, W.G. & Ma, Jian & Satoh, Kohyu. (2009). Artificial neural network approach for modeling the impact of population density and weather parameters on forest fire risk. International Journal of Wildland Fire. 18. 640-647. 10.1071/WF07136.

[33]. MODIS/Terra Vegetation Indices 16-Day L3 Global 250m SIN Grid V006. NASA EOSDIS Land Processes DAAC. Accessed 2019-10-10 from https://doi.org/10.5067/MODIS/MOD13Q1.006. Accessed October 10, 2019.

[34]. McArthur, A. G. (1967, January 1). Fire behaviour in eucalypt forests (1967 edition). Retrieved from https://openlibrary.org/books/OL5764346M/Fire_behaviour_in_eucalypt_forests.

[35]. Retrieved (n.d.).from http://www.interfire.org/termoftheweek.asp?term=1240.

[36]. Sawada, H. (2005, December 5). Japanese Forest Monitoring - 12/05/2005. Retrieved from https://www.gim-international.com/content/article/japanese-forest-monitoring.

[37]. Schlobohm, P., & Brain, J. (Eds.). (2002, July). Gaining an Understanding of the National Fire Danger ... Retrieved from https://www.nwcg.gov/sites/default/files/products/pms932.pdf.

[38]. Škvarenina J, Mindáš J, Holécy J, Tuček J (2004) An analysis of the meteorological conditions during two largest forest fire events in the Slovak Paradise National Park. Meteorological Journal 7:167–171. Retrieved from https://www.researchgate.net/profile/Jan_Tucek/publication/268288327_Analysis_of_the_natural_and_meteorological_conditions_during_two_largest_forest_fire_events_in_the_Slovak_Paradise_National_Park/links/54be741e0cf218da9391ed2e.pdf

[39]. Thiessen, M. (2019, August 9). Wildfires. Retrieved from https://www.nationalgeographic.com/environment/natural-disasters/wildfires/.

[40]. The Editors of Encyclopaedia Britannica. (2017, October 16). Wildfire. Retrieved from https://www.britannica.com/science/wildfire.

[41]. Terra/Aqua MODIS - Earth Online. (n.d.). Retrieved from https://earth.esa.int/web/guest/missions/3rd-party-missions/current-missions/terraaqua-modis..

[42]. Watts, A., Halla, T., North, M. of the, Erren, H., JoeShaw, MilwaukeeBob, … SteveTa. (2019, January 26). California's Wildfire History – in one map. Retrieved from https://wattsupwiththat.com/2019/01/26/californias-wildfire-history-in-one-map/.

[43]. "Landsat 8 Overview " Landsat Science," NASA. [Online]. Available: https://landsat.gsfc.nasa.gov/landsat-8/landsat-8-overview/. [Accessed: 29-Feb-2020].

[44]. Proba-V - ESA EO Missions - Earth Online. (n.d.). Retrieved from https://earth.esa.int/web/guest/missions/esa-operational-eo-missions/proba-v

[45]. Staff, S. X. (2017, June 22). Proba-V images Portuguese forest fire. Retrieved from https://phys.org/news/2017-06-proba-v-images-portuguese-forest.html

[46]. Landsat Collections in Earth Engine | Earth Engine Data Catalog. (n.d.). Retrieved from https://developers.google.com/earth-engine/datasets/catalog/landsat

[47]. Retrieved from https://www.vito-eodata.be/PDF/portal/Application.html#Home

[48]. Asian Institute. (n.d.). NDVI, NDBI & NDWI Calculation Using Landsat 7, 8. Retrieved from https://www.linkedin.com/pulse/ndvi-ndbi-ndwi-calculation-using-landsat-7-8-tek-bahadur-kshetri

[49]. Christina Maxouris, CNN, "Here's just how bad the devastating Australian fires are -- by the numbers", Updated 6:30 AM ET, Monday, January 6, 2020, https://www.cnn.com/2020/01/06/us/australian-fires-by-the-numbers-trnd/index.html.

[50]. Alejandra Borunda, "See how much of the Amazon is burning, how it compares to other years", PUBLISHED by National Geographic, Environment Column, AUGUST 29, 2019.

[51]. Taylor S. W., Woolford D. G., Dean C. B. and Martell D. L.: Wildfire Prediction to Inform Management: Statistical Science Challenges, Statistical Science, 28(4), 586-615, 10.1214/13-STS451, 2013.

[52]. Schlobohm, P., & Brain, J. (Eds.), Gaining an Understanding of the National Fire Danger, July, 2002, July, Retrieved from https://www.nwcg.gov/sites/default/files/products/pms932.pdf

[53]. McArthur, A. G. (1967, January 1). Fire behaviour in eucalypt forests (1967 edition). Retrieved from https://openlibrary.org/books/OL5764346M/Fire_behaviour_in_eucalypt_forests

[54]. Stocks, B.J., B.D. Lawson, M.E. Alexander, M.E., C.E. Van Wagner, R.S. McAlpine, T.J. Lynham, D.E. Dube. The Canadian Forest Fire Danger Rating System: An Overview. Forestry Chronicle Vol. 65 issue 6 : 450-457, 1989.

[55]. Sawada, H. Japanese Forest Monitoring, December 5, 2005. Retrieved from https://www.gim-international.com/content/article/japanese-forest-monitoring.

[56]. Škvarenina J, Mindáš J, Holécy J, Tuček J., An analysis of the meteorological conditions during two largest forest fire events in the Slovak Paradise National Park. Meteorological Journal 7:167–171, 2004, Retrieved from https://www.researchgate.net/profile/Jan_Tucek/publication/268288327_Analysis_of_the_natural_and_meteorological_conditions_during_two_largest_forest_fire_events_in_the_Slovak_Paradise_National_Park/links/54be741e0cf218da9391ed2e.pdf

[57]. Bianchinia G., Caymes-Scutariab P., and Méndez-Garabettiab M.: Evolutionary-Statistical System: A parallel method for improving forest fire spread prediction, Journal of Computational Science, 6, 58-66, https://doi.org/10.1016/j.jocs.2014.12.001, 2015.

[58]. Andrés C., Ana C., and Tomàs M.: Applying Probability Theory for the Quality Assessment of a Wildfire Spread Prediction Framework Based on Genetic Algorithms, the Scientific World Journal, 2013, 728414, http://dx.doi.org/10.1155/2013/728414, 2013.

[59]. Andrés C., Ana C., and Tomàs M.: Response time assessment in forest fire spread simulation: An integrated methodology for efficient exploitation of available prediction time, Environmental Modelling & Software, 54, 153-164, https://doi.org/10.1016/j.envsoft.2014.01.008, 2014

[60]. Retrieved (n.d.). from https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=8924693

[61]. Home. (n.d.). Retrieved from https://eos.com/blog/ndvi-faq-all-you-need-to-know-about-ndvi/

[62]. Heller, M. (2019, May 9). Machine learning algorithms explained. Retrieved from https://www.infoworld.com/article/3394399/machine-learning-algorithms-explained.html

[63]. Retrieved (n.d.). from https://www.fsa.usda.gov/programs-and-services/aerial-photography/imagery-programs/naip-imagery/

[64]. Landsat Surface Reflectance-Derived Spectral Indices. (n.d.). Retrieved from https://www.usgs.gov/land-resources/nli/landsat/landsat-enhanced-vegetation-index?qt-science_support_page_related_con=0#qt-science_support_page_related_con

[65]. Global Enhanced Vegetation Index. (n.d.). Retrieved from https://earthobservatory.nasa.gov/images/1863/global-enhanced-vegetation-index

[66]. Brownlee, J. (2020, January 14). SMOTE Oversampling for Imbalanced Classification with Python. Retrieved from https://machinelearningmastery.com/smote-oversampling-for-imbalanced-classification/

[67]. Zuhaib, M. (2019, December 9). Demystifying the Confusion Matrix Using a Business Example. Retrieved from https://towardsdatascience.com/demystifying-confusion-matrix-29f3037b0cfa

[68]. AdaBoost Classifier in Python. (n.d.). Retrieved from https://www.datacamp.com/community/tutorials/adaboost-classifier-python

[69]. Singh, H. (2018, November 4). Understanding Gradient Boosting Machines. Retrieved from https://towardsdatascience.com/understanding-gradient-boosting-machines-9be756fe76ab

[70]. Demsar, J. (2006). Statistical Comparisons of Classifiers over Multiple Data Sets. Retrieved from http://www.jmlr.org/papers/volume7/demsar06a/demsar06a.pdf

[71]. Sahakyan, R. (2019, November 7). Meaningful Metrics: Cumulative Gains and Lyft Charts. Retrieved from https://towardsdatascience.com/meaningful-metrics-cumulative-gains-and-lyft-charts-7aac02fc5c14

[72]. University of Regina DBD. (n.d.). Cumulative Gains and Lift Charts. Retrieved from http://www2.cs.uregina.ca/~dbd/cs831/notes/lift_chart/lift_chart.html

[73]. Rodriguez-Aseretto, D., Rigo, D. de, Leo, M. D., Cortés, A., & San-Miguel-Ayanz, J. (2013, June 1). A Data-driven Model for Large Wildfire Behaviour Prediction in Europe. Retrieved from https://www.sciencedirect.com/science/article/pii/S1877050913004985

Appendices

Appendix 1 – Attribution

Tools, technologies and online sources have played a pivotal role in helping us complete this ambitious endeavor. Hereby, we list the main resources we used, free of cost due to funding constraints. Further, we acknowledge the numerous unlisted tools that made this project possible.

- Jupyter notebook in Google Collaboratory
- Python libraries
- Google Earth Engine
- Quantum Geographic Information System (QGIS)
- Aeronautical Reconnaissance Coverage Geographic Information System (ArcGIS)
- NASA-based data sources
- Tableau
- Lucid Chart
- Creately
- Pixabay
- Flaticon.com
- Freepik.com
- Weather stations
- FileZilla
- Amazon Web services (AWS)
- Google docs
- Email and internet service providers
- SJSU Canvas tool

- Plagiarism Checker - In-house tool and smallseotools.com
- Nuwallpaperhd.info
- Unsplunk.com

These resources enabled us to successfully complete this project and build the Spartan Wildfire Prediction system (SWiPS).