POLITECNICO
MILANO 1863

M.Sc. Geoinformatics Engineering
Geoinformatics Project

Project Technical Report

Machine Learning Analysis of Factors Influencing Wildfire Distribution

Author: Alqrinawi Feras Younis Mahmoud
ferasyounis.alqrinawi@mail.polimi.it

Supervisors: Prof. Venuti Giovanna,
Dr. Daniela Stroppiana

July 8, 2024

GitHub repo:
https://github.com/Ferasqr/wildfires_factors_analysis

Table of Contents

Abstract

This project investigates the factors that affect wildfire distribution across the seven continents, utilizing geoinformatics and machine learning techniques. For this purpose, Google Earth Engine (GEE) and Python were employed to gather, preprocess, and analyze data, and to train predictive models. The study identifies critical environmental and human-related variables that contribute to wildfire occurrences. By employing a hexagon tiling-sampling method, it extracts data points from specific global regions based on the density of burned pixels. This innovative method enhances flexibility in data extraction and significantly improves the accuracy of the predictions.

1 Introduction

Climate change is projected to alter the geographical spread of wildfires in the future. In 2024, numerous regions witnessed unprecedented wildfire activity. The environmental and societal impacts of these fires extend beyond direct flame damage, encompassing smoke pollution and CO2 emissions. Given that 2023 marked the hottest year in the past 150 years, it suggests global temperature as a potential catalyst for wildfires. However, combustion is a highly intricate physical process, and the combination of factors necessary to initiate wildfires is far from straightforward. This document will provide a comprehensive account of the materials, concepts, and methodologies used to develop the project, including the implementation of necessary code. Additionally, it will present the final outcomes, future prospects for development and enhancement, while addressing challenges encountered during the project's execution. [1]

1.1 Background on wildfires

Wildfires are a natural phenomenon shaped by a combination of climate, vegetation, and human activities. They play a crucial role in ecosystem dynamics, influencing biodiversity and landscape diversity. [2] However, human-induced factors like land use changes and climate change have intensified their frequency and severity in recent decades. Wildfires vary in scale from small, controlled burns to large, destructive infernos, impacting habitats, economies, and communities worldwide. Understanding their causes, behavior, and ecological effects is essential for effective management and mitigation strategies in fire prone regions. Scientific research continues to explore ways to predict, prevent, and manage wildfires sustainably.

1.2 Understanding wildfires distribution

Wildfire distribution is influenced by a combination of environmental and human-related factors. Climatic conditions such as temperature, precipitation, and humidity directly affect the likelihood of fire ignition and spread. [3] Vegetation type and density, as well as topographic features like slope and aspect, are also critical factors. Additionally, human activities such as land-use changes, urban expansion into wildland areas, and fire management practices play a significant role. By integrating geoinformatics and machine learning, this study aims to analyze these factors in detail and understand their impact on wildfire distribution globally.

1.3 Objectives of the study

The primary objective of this study is to investigate the multifaceted factors that influence wildfire distribution across the seven continents. This research aims to achieve the following specific objectives:

- **Identify Key Factors:** Determine the significant environmental and human-related factors that contribute to wildfire occurrences. Environmental factors include climatic conditions such as temperature, precipitation, and humidity, as well as vegetation indices and topographic features. Human-related factors encompass land-use changes, population density, and fire management practices.
- **Develop Predictive Models:** Create robust models to predict wildfire risks based on the identified factors. The study employs advanced machine learning techniques to analyze the data and develop accurate predictive models that can forecast wildfire occurrences.

1.4 Scope of the project

This project encompasses a comprehensive analysis of wildfire distribution across the seven continents, leveraging the capabilities of Google Earth Engine (GEE) and Python. The scope of the study includes:

- **Global Scale Analysis:** This project covers all seven continents, focusing on areas prone to wildfires due to various environmental and human-related factors. Specific regions include savannas and forests in Africa [4], diverse ecosystems in Asia [5], the Mediterranean in Europe [6], western regions in North America, bushfire-prone areas in Australia [7], the Amazon in South America [8], and potential future risks in Antarctica.

- **Africa**
- **Asia**
- **Europe**
- **Australia**
- **South America**
- **North America**
- **Antarctica**

Figure (1): World Map

2 Literature review

2.1 Introduction to wildfire studies

Wildfires are a fundamental ecological process in many ecosystems [2], influencing vegetation patterns, nutrient cycles, and habitat dynamics. However, human activities and climate change have significantly altered wildfire regimes, increasing their frequency, intensity, and distribution. Understanding the multifaceted factors influencing wildfire occurrences is crucial for effective prevention, mitigation, and management strategies.

2.2 Environmental Factors Influencing Wildfires

Environmental factors, such as climate, vegetation, and topography, play a significant role in wildfire behavior and distribution. Climate conditions, including temperature, precipitation, and humidity, directly affect the likelihood of fire ignition and spread. Studies have shown that prolonged droughts and heatwaves increase the probability of wildfires by creating favorable conditions for combustion [2]. Vegetation type and density also influence wildfire dynamics [9], as some plant species are more flammable than others. For instance, forests with high fuel loads, such as those containing dense underbrush or deadwood, are more prone to intense fires. Topographic features, such as slope and aspect, can affect fire spread by influencing wind patterns and the movement of flames. [2]

Environmental Factors	Description
Precipitation	The amount of rainfall or snowfall, measured in millimeters.
Relative Humidity	The amount of moisture in the air compared to what the air can hold at that temperature, expressed as a percentage.
Solar Radiation	The amount of solar energy received by a specific area, measured in watts per square meter.
NDMI	Normalized Difference Moisture Index, used to determine vegetation water content. ((Band 5 – Band 6) / (Band 5 + Band 6))
NDVI	Normalized Difference Vegetation Index, used to assess whether the target being observed contains live green vegetation. (Band 5 – Band 4) / (Band 5 + Band 4))
Slope	The steepness or incline of the land, usually expressed as a percentage.
Aspect	The compass direction that a slope faces, usually measured in degrees from north. (Categorical)
Elevation	The height of the land above sea level, measured in meters.
Soil Moisture	Soil moisture content, indicating the amount of water contained in the soil.
evapotranspiration	The sum of evaporation and plant transpiration from the Earth's land and ocean surface to the atmosphere.

Table (1) : Environmental Factors

2.3 Human-Related Factors Influencing Wildfires

Human activities have significantly altered wildfire regimes through land-use changes, population growth, and fire suppression practices. Urban expansion into wildland areas, known as the wildland-urban interface (WUI) [10], has increased the risk of wildfires and their potential damage to human infrastructure [11]. Agricultural practices, such as slash-and-burn farming, and deforestation for timber or agricultural land have also contributed to increased fire occurrences [12]. Additionally, the introduction of fire suppression policies has led to the accumulation of fuel loads in forests, making them more susceptible to large, uncontrollable fires when they do occur.

Factors	Description
Population Density	The number of people living per unit of area, usually measured in people per square kilometer. (people/km2)
Human Impact Index	A measure of human impact on the environment, averaged over a specific area. (Numerical)

Table (2) : Human-Related Factors

2.4 Other Factors Influencing Wildfires

By considering these additional factors alongside environmental and human-related variables, the study provides a comprehensive analysis of wildfire distribution and risks. This holistic approach helps in developing more accurate predictive models and effective wildfire management strategies.

Factors	Description
Fuel Load	Global Aboveground and Belowground Biomass Carbon Density measured in tons per hectare.
Fuel Moisture Content	The amount of moisture in the fuel, expressed as a percentage.
Lightning Frequency	The frequency of lightning strikes in a given area, measured in strikes per year.

Table (3): Other Factors

3 Methodology

3.1 Data collection

The data collection process is critical for accurately analyzing the factors influencing wildfire distribution. This study utilizes various sources of remote sensing images and datasets to obtain comprehensive environmental and human-related information. The primary sources of these images are:

1. **IDAHO_EPSCOR/TERRACLIMATE**
    - TerraClimate is a high-resolution global dataset that provides monthly climate and climatic water balance data from 1958 to the present. This dataset includes various climate variables such as precipitation, solar radiation, and temperature, essential for ecological and hydrological studies. The data is derived using climatically aided interpolation, combining high-resolution climatological normals from the WorldClim dataset with time-varying data from other sources to produce comprehensive monthly datasets [13].
2. **NASA/GLDAS/V021/NOAH/G025/T3H**
    - The Global Land Data Assimilation System (GLDAS) integrates satellite and ground-based observational data to provide high-resolution land surface datasets. This dataset includes variables such as relative humidity, temperature, and soil moisture, which are critical for understanding land surface conditions and their impact on wildfire risks. [14]
3. **LANDSAT/LC08/C02/T1_TOA**
    - Landsat 8 provides high-quality optical imagery used to monitor and assess various land surface characteristics. The dataset includes top-of-atmosphere reflectance data, which can be used to calculate vegetation indices like the Normalized Difference Vegetation Index (NDVI) and the Normalized Difference Moisture Index (NDMI) [14]
4. **MODIS/006/MOD16A2**
    - The Moderate Resolution Imaging Spectroradiometer (MODIS) dataset includes evapotranspiration data, representing the sum of evaporation and plant transpiration. This factor is important for understanding the water cycle's impact on vegetation dryness and fire risks. [14]
5. **NASA SRTM Digital Elevation 30m**
    - The Shuttle Radar Topography Mission (SRTM) provides high-resolution digital elevation data used to derive slope and aspect information. These topographic features influence fire spread by affecting wind patterns and fuel distribution. As these variables are static, no temporal aggregation is applied. [14]
6. **MCD12Q1.061 MODIS Land Cover Type Yearly Global 500m**
    - This MODIS dataset provides annual land cover classifications at a 500-meter resolution. It includes different land cover types such as forests, grasslands, and urban areas, which are important for analyzing the distribution and types of vegetation that can fuel wildfires. [14]

7. **Human Impact Index (HII) by Wildlife Conservation Society**
    - The Human Impact Index (HII) measures human influence on the environment, averaged over specific areas. This index includes factors like population density, infrastructure, and land use, providing a comprehensive measure of human impact on wildfire risks. [14]
8. **NASA/ORNL/biomass_carbon_density/v1**
    - This dataset includes global estimates of aboveground and belowground biomass carbon density, measured in tons per hectare. It provides critical information on the amount of available fuel for wildfires. [14]
9. **MODIS/006/MOD13Q1**
    - The MODIS dataset includes data on vegetation indices such as NDVI and NDMI, as well as fuel moisture content. These indices are used to monitor vegetation health and moisture levels, which influence wildfire susceptibility. [14]
10. **NASA/GLDAS/V021/NOAH/G025/T3H**
    - This dataset includes the frequency of lightning strikes, which are a natural ignition source for wildfires. The data is provided by the Global Land Data Assimilation System, offering insights into the natural causes of wildfire ignition. [14]

Environmental Factors	Image Collection
Precipitation, Solar Radiation	**"IDAHO_EPSCOR/TERRACLIMATE"**
Relative Humidity	**"NASA/GLDAS/V021/NOAH/G025/T3H"**
NDMI , NDVI	**"LANDSAT/LC08/C02/T1_TOA"** • Removed Cloud Cover and Shadows from Landsat images using Qa band (Their presence can introduce noise and errors in the results.)
Evapotranspiration	**"MODIS/006/MOD16A2"**
Slope , Aspect	**NASA SRTM Digital Elevation 30m**
Land Cover	**MCD12Q1.061 MODIS Land Cover Type Yearly Global 500m**

Table (4) : Environmental Factors Image Collections

Factor	Image Collection
Population Density , Human Impact Index	**Human Impact Index (HII) by Wildlife Conservation Society**
Fuel Load	**"NASA/ORNL/biomass_carbon_density/v1"**
Fuel Moisture Content	**"MODIS/006/MOD13Q1"**
Lightning Frequency	**"NASA/GLDAS/V021/NOAH/G025/T3H"**

Table (5) : Human-Related Factors & Other Factors Image Collections

3.2 Data preprocessing

It is a critical step in ensuring the quality and reliability of the data used for analyzing wildfire distribution. The preprocessing involves several key steps to clean, transform, and prepare the data for further analysis and modeling. The following describes the main preprocessing tasks carried out in this study:

3.2.1 Filtering and Time Range Selection

- **Temporal Consistency:** To maintain consistency across all datasets, a specific time range of 13.5 years (2010-2024) is selected. This period provides a balanced dataset with recent and relevant data for the analysis.
- **Filtering:** Each dataset is filtered to include only the data within this selected time range, ensuring that the analysis covers the same temporal period across different variables.

Filtered to 2010-2024, Median Aggregation	Mean Aggregation	No Processing Required
Precipitation, Solar Radiation, Relative Humidity, NDVI, NDMI, Evapotranspiration, Fuel Moisture Content, Lightning Frequency, Fuel load	Human Impact Index (HII), Population Density	Slope, Aspect, Land Cover (Reprojected to match the CRS of other products.)

Table (6) : Data preprocessing

3.2.2 Data Cleaning

- **Cloud Cover and Shadow Removal:** For satellite imagery, particularly from the Landsat 8 dataset, it is crucial to remove any cloud cover and shadows that can obscure the underlying land surface. This is done using the Quality Assurance (QA) band to identify and mask out these areas, resulting in clearer images for calculating vegetation indices and other variables. [15]
- **Noise Reduction:** Any anomalous or noisy data points are identified and corrected or removed. This step is vital for ensuring the accuracy of the extracted information.

3.2.3 Clipping the Area

The study area were clipped by continent using shapefiles, assigning each continent its own shapefile to process individually and exclude the oceans. This approach reduces computational strain and speeds up processing.

3.2.4 Hexagon grids

The hexagon grid for this study is created using QGIS tools. For this purpose, a 500 km horizontal and vertical spacing is chosen. This means that moving horizontally from the center of one hexagon to the center of its neighboring hexagon covers a distance of 500 kilometers. Similarly, moving vertically from the center of one hexagon to the center of its adjacent hexagon also covers a distance of 500 kilometers.

This approach is suitable for representing geographic areas at a broader scale. Given that the analysis considers a global scale, it allows for the visualization and analysis of data with a lower level of detail. This method strikes a balance between capturing the spatial complexity of wildfire distribution and ensuring computational efficiency.

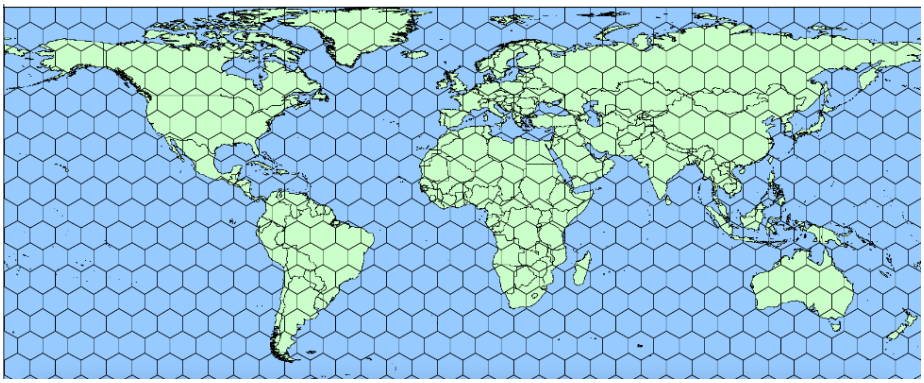

Figure (2) : Hexagon Grid

4. Data Ingestion and Implementation

These are critical steps that prepare the collected datasets for building the models upon them. These processes involve loading data into an appropriate platform, creating necessary masks, and implementing efficient sampling methods to manage the large study area effectively. Here's a detailed explanation of the data ingestion and implementation steps:

4.1 Loading Data into Google Earth Engine (GEE)

Google Earth Engine (GEE) is a robust cloud-based platform designed for processing and analyzing large geospatial datasets. In this study, the collected datasets [see Table 4 & Table 5], which include satellite imagery and various environmental and human-related factors, are imported into GEE. To optimize computational efficiency, each continent is loaded as an asset into GEE. This segmentation reduces the overall computational power needed for data processing and allows for more manageable analysis. [14]

Figure (3) : Loading Australia to GEE

4.2 Binary Mask Creation for Burned and Not Burned Areas

The process of creating a binary mask to distinguish burned areas involves several steps:

1. **Load MODIS Burned Area Collection**: The MODIS burned area dataset is filtered by a specific start and end date to retrieve relevant data. [14]
2. **Define Burn Date Threshold**: A threshold value (e.g., 1) is set to identify burned areas.
3. **Classify Pixels**: A function is applied to classify pixels as burned if the burn date exceeds the defined threshold.
4. **Create Binary Mask per Image**: For each image in the collection, a binary mask is created where burned pixels are marked.
5. **Create Cumulative Burned Area Image**: Binary masks from the collection are summed to generate a cumulative burned area image.
6. **Create Pure Fires Binary Mask**: The final binary mask is produced, representing all the burned areas identified during the period.

***All the codes are in the GitHub repo.**

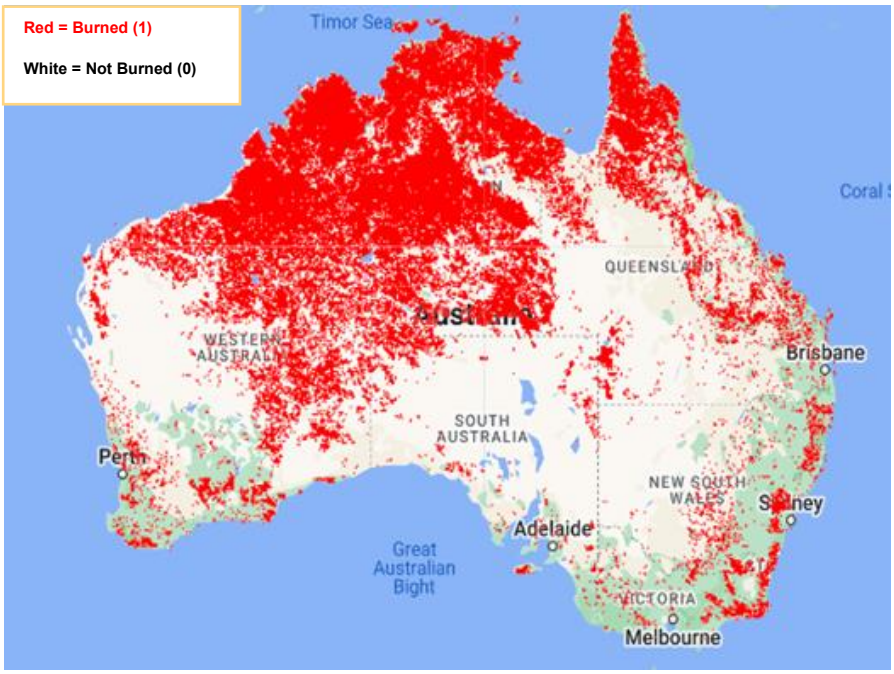

Figure (4) :Fire Binary Mask (Australia)

4.3 Hexagon-Based Sampling

To efficiently manage the study area, a hexagon-based sampling method is employed:

1. **Create Binary Burned Mask**: The burned mask is multiplied by the pixel area to obtain an area image.
2. **Perform Zonal Statistics**: Using the reduceRegions function with a sum() reducer, zonal statistics are performed on the area image within each hexagon.
3. **Calculate Hexagons**: The hexagons are divided into two categories: those with a burned pixel sum of 0 and those with a burned pixel sum greater than 0. Each hexagon is assigned a unique ID for further analysis.

***All the codes are in the GitHub repo.**

Figure (5) :Burned/Not Hexagons (Australia)

4.4 Extracting Sample Points Over Hexagons

- **Create Feature Collection**: A Feature Collection is created representing hexagons with burned pixels. The 'id' of each hexagon is used to access hexagons with a non-zero burned pixel sum.
- **Filter Hexagons by 'id'**: Iterating over a subset of hexagons, filter them based on specific 'id'.
- **Generate Stratified Sample Points**: Generate 500 stratified sample points within the burned areas of each hexagon. These points are stratified over the Pure Fire mask to ensure representative sampling And other 500 stratified sample points within the Unburned areas of each hexagon.
- **Extract Predictor Values**: Extract values from the predictor image variables for each sampled point.
- **Adjust Sampling Based on Area**: The number of hexagons and sample points varies per continent, depending on the area covered by burned pixels. Larger areas require more points to ensure comprehensive sampling, considering memory limitations.

***Two codes were developed one to extract the samples from the unburned areas and the other from burned areas the codes are in the GitHub repo.**

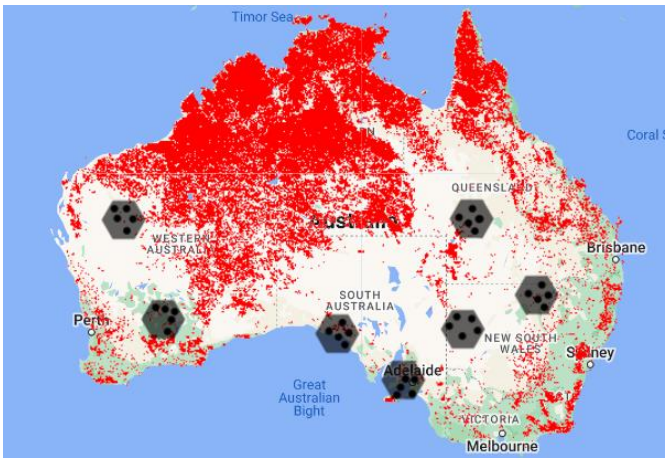

Figure (6): Unburned Area Sampling (Australia)

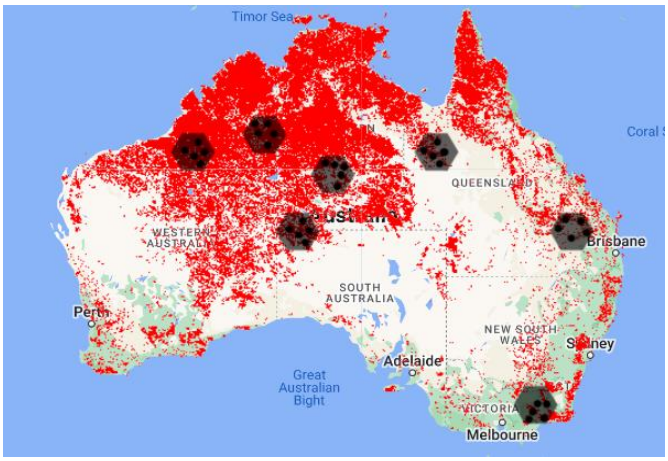

Figure (7): Burned Area Sampling (Australia)

Continent	# of Hexagons	# of points within each hexagon	# of samples
Asia	37	4	148
Africa	20	5	100
North America	16	5	80
South America	12	5	60
Europe	7	7	49
Australia	7	5	35

Table (7) - Hexagon-based sampling

4.5 Exporting the Final Product

- The final Feature Collection, named 'finalresults', is exported as a CSV file to a designated Google Drive folder. This data is then further processed using Python in a Jupyter Notebook.

Column Name	Data Type
Aspect	float64
HumanImpactIndexMean	float64
LandCover	float64
NDMI	float64
NDVI	float64
Precipitation	float64
Slope	float64
SoilMoist	float64
TempMax	float64
TempMin	float64
WaterDeficit	float64
WindSpeed	float64
idx	float64
.geo	object
Burned (Target)	int64
RelativeHumidity	float64
Elevation	float64
FuelLoad	float64
FuelMoistureContent	float64
SolarRadiation	float64
LightningFrequency	float64
Evapotranspiration	float64
PopulationDensity	float64

Table (8) - Csv Table Contents

4. Model Building

Model building is a critical phase in the analysis process, involving various steps to ensure the development of robust and accurate predictive models. This section covers the key stages, from exploratory data analysis to model evaluation. Using the features we got in the steps before (Table (8) - Csv Table Contents).

4.1 Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) involves investigating the datasets to discover patterns, spot anomalies, test hypotheses, and check assumptions through summary statistics and graphical representations. EDA helps in understanding the underlying structure of the data and provides insights that guide the subsequent steps in model building.

- **Descriptive Statistics**:

  The dataset exhibits a wide range of values across various environmental and geographical variables.

Variable	Mean	Median	Standard Deviation	Min	Max
Aspect	177.84	180.00	104.01	0.00	354.83
HumanImpactIndexMean	961.52	769.21	845.58	0.00	2941.20
LandCover	12.02	12.29	4.03	2.00	17.00
NDMI	-0.002	-0.057	0.154	-0.235	0.431
NDVI	0.246	0.175	0.173	0.034	0.705
Precipitation	43.43	9.16	64.18	0.00	233.33
Slope	4.27	1.92	5.11	0.00	19.74
SoilMoist	656.54	3.25	1142.85	0.00	4174.50
TempMax	303.02	311.63	62.83	128.00	409.00
TempMin	164.13	172.44	56.39	6.50	246.00
RelativeHumidity	-350.54	-399.76	317.25	-842.50	99.90
Elevation	420.32	175.50	500.34	0.00	1974.00
FuelLoad	4.70	5.00	2.57	1.00	10.00
FuelMoistureContent	660.90	3.50	1143.64	0.00	4174.50
SolarRadiation	235.33	245.79	54.80	78.75	315.75
LightningFrequency	4.11	0.88	6.12	0.00	23.33
Evapotranspiration	232.92	242.98	55.91	75.70	315.75
PopulationDensity	97.06	77.50	84.02	0.00	294.12

Table (9): Summary Statistics of data

- **Aspect**: The Aspect variable has a mean of 177.84 and a median of 180.00, indicating a nearly uniform distribution around its central value, with a standard deviation of 104.01.

- **HumanImpactIndexMean**: Shows a high variability with a mean of 961.52 and a substantial standard deviation of 845.58, reflecting significant differences in human impact across the sampled regions.

- **LandCover**: Has a mean of 12.02 and a relatively small standard deviation of 4.03, suggesting a moderate variability in land cover types.

- **NDMI and NDVI**: Values, which are indicators of vegetation health and moisture, have means of -0.002 and 0.246, respectively, with NDMI exhibiting a negative skew.

- **Precipitation and Slope**: Precipitation varies widely across the dataset, with a mean of 43.43 mm and a large standard deviation of 64.18 mm, indicating diverse climatic conditions. The Slope has a mean of 4.27 degrees, with the majority of data points clustered around low values, as indicated by the median of 1.92 degrees.

- **SoilMoist, TempMax, TempMin, and RelativeHumidity**: Soil moisture levels, represented by SoilMoist, show high variability, with a mean of 656.54 and a standard deviation of 1142.85, highlighting significant differences in soil moisture across different regions. The temperature variables, TempMax and TempMin, have means of 303.02 and 164.13, respectively, with considerable standard deviations, indicating a wide range of temperatures in the dataset. RelativeHumidity presents a negative mean value of -350.54, suggesting potential anomalies or specific conditions in the data collection process.

- **Elevation**: Shows a mean of 420.32 meters, with a high standard deviation of 500.34 meters, indicating a diverse topography.

- **FuelLoad and FuelMoistureContent**: Important for understanding fire risk, have means of 4.70 and 660.90, respectively, with significant variability.

- **SolarRadiation**: Has a mean of 235.33, indicating moderate solar exposure on average.

- **LightningFrequency and Evapotranspiration**: Exhibit means of 4.11 and 232.92, respectively, reflecting diverse climatic and environmental conditions.

- Finally, **PopulationDensity** shows a mean of 97.06, with a wide spread as indicated by its standard deviation of 84.02, pointing to varied human settlement patterns across the dataset.

- **Visualization**:

  Utilized plots, specifically histograms, to effectively visualize and analyze the distribution and relationships between various variables within the dataset. By examining these histograms, we can gain valuable insights into the central tendencies, variability, and frequency distributions of each variable. This graphical representation helps to identify patterns, trends, and potential anomalies in the data, thereby enhancing our understanding of the underlying relationships and characteristics of the variables. These visual tools are crucial for interpreting complex datasets and making informed decisions based on the observed data distributions.

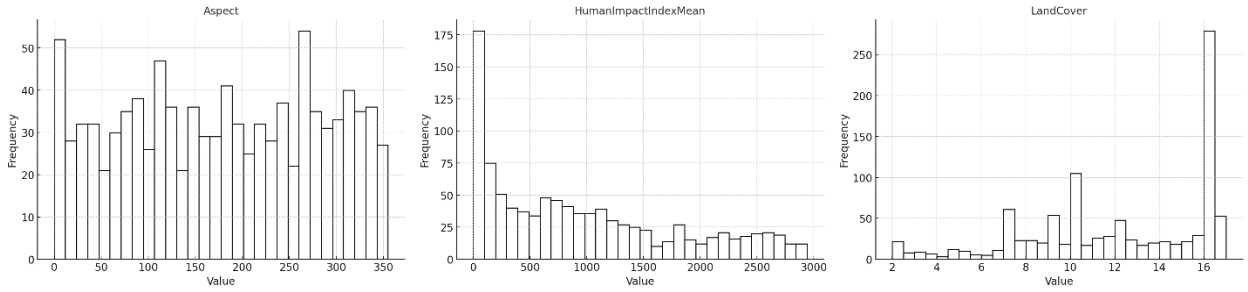

Figure (8): Aspect, HumanImpactIndexMean, LandCover Histograms

  - **Aspect**: The distribution is nearly uniform, with values spread across the range.
  - **HumanImpactIndexMean**: Shows a right-skewed distribution with most values concentrated at lower values and a long tail extending towards higher values.
  - **LandCover**: The distribution appears to be multimodal, indicating distinct categories within the land cover data.

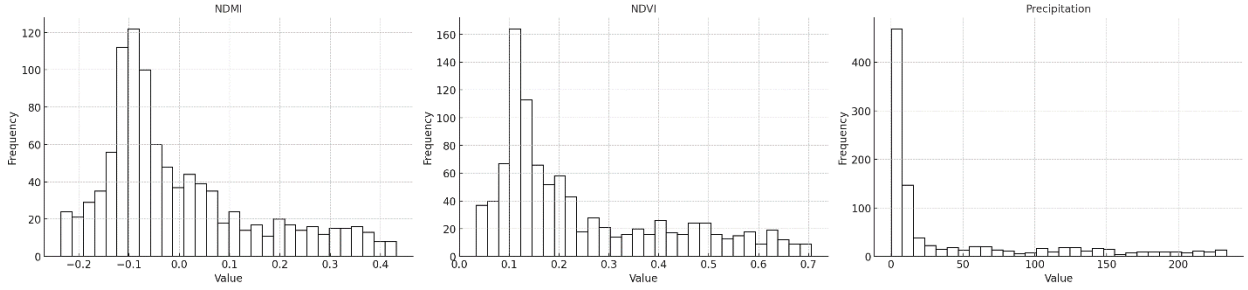

Figure (9): NDMI, NDVI, Precipitation Histograms

  - **NDMI:** The distribution is centered around zero, with a slight skew towards negative values.
  - **NDVI:** The distribution is right-skewed, with most values concentrated around lower to mid-range values.
  - **Precipitation:** Displays a right-skewed distribution, with a large number of observations at lower values and a long tail extending towards higher values.

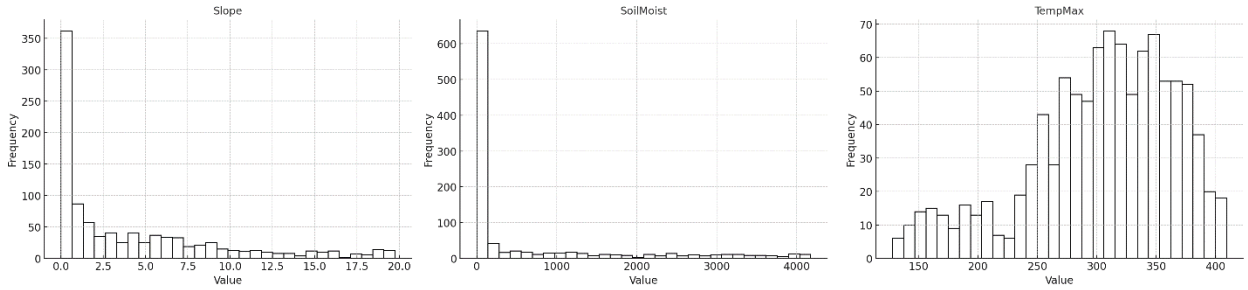

Figure (10) : Slope, SoilMoist, TempMax Histograms

- **Slope:** The distribution is right-skewed, with most values at lower ranges and few high values.
- **SoilMoist:** Shows a right-skewed distribution with most values concentrated at lower levels and a few extremely high values.
- **TempMax:** The distribution appears bimodal, indicating two distinct temperature regimes.

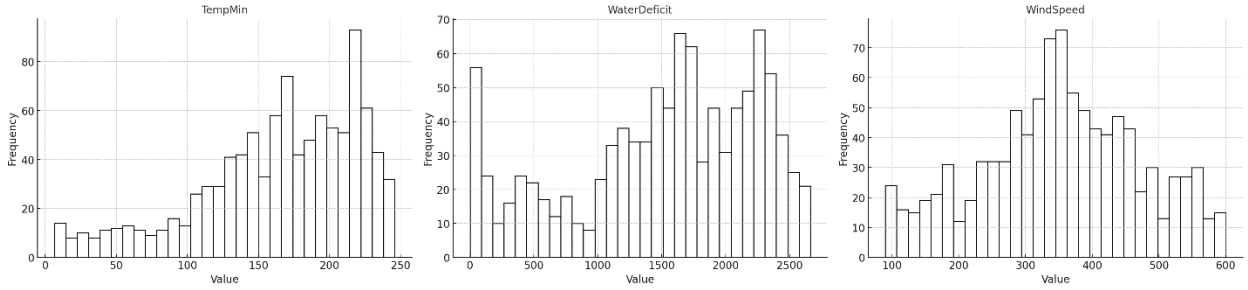

Figure (11) : TempMin, WaterDeficit, WindSpeed Histograms

- **TempMin**: The distribution of minimum temperature values shows a relatively uniform spread with a slight skew towards lower values. There are noticeable peaks around 150 and 200, indicating common minimum temperatures in these ranges.
- **WaterDeficit**: This variable displays a bimodal distribution with two prominent peaks around 1500 and 2000. The values are spread across a broad range, indicating variability in water deficit across different regions.
- **WindSpeed**: The histogram for wind speed exhibits a central peak around 300 to 400, suggesting this is the most common range of wind speeds. The distribution has a slight right skew, with fewer occurrences of higher wind speeds.

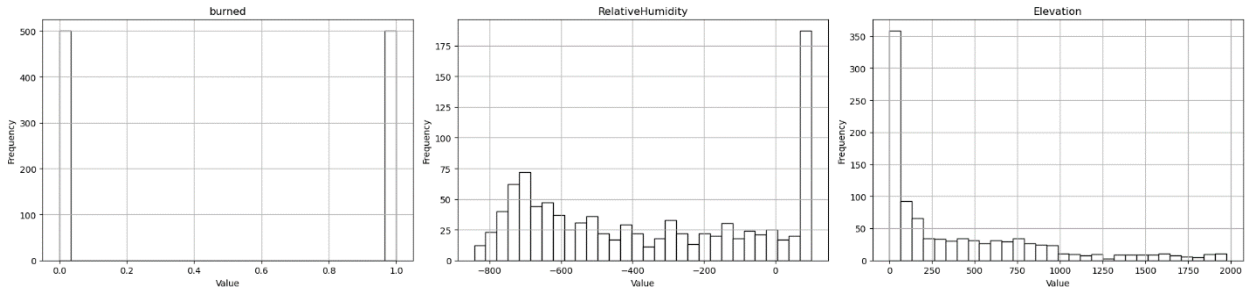

Figure (12) :Burned, RelativeHumidity, Elevation Histograms

- **burned**: The histogram for the 'burned' variable shows a binary distribution, indicating that the data is categorical with only two possible values (0 and 1). The frequencies of both categories are approximately equal, showing that the dataset is evenly split between burned and unburned areas.
- **RelativeHumidity**: This histogram displays a left-skewed distribution with most values concentrated at lower (more negative) humidity levels. There is a significant number of observations around -800 to -400, and a noticeable spike near zero, suggesting an anomaly or specific environmental condition affecting relative humidity.
- **Elevation**: The elevation data shows a right-skewed distribution with most values clustered at lower elevations. The frequency gradually decreases as elevation increases, with very few observations at higher elevations, indicating that the majority of the data points are from lower altitude regions.

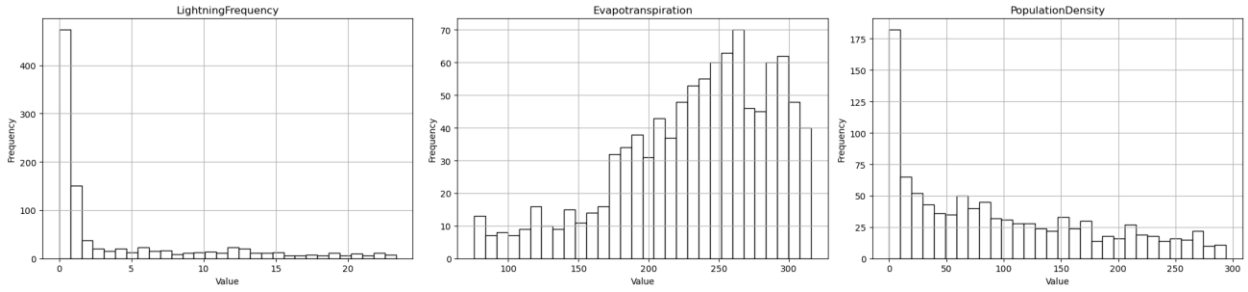

Figure (13) : LightningFrequency, Evapotranspiration, PopulationDensity Histograms

- **LightningFrequency**: The histogram for lightning frequency shows a right-skewed distribution with a significant concentration of values near zero. This indicates that low lightning frequency is common, while high frequency occurrences are rare.
- **Evapotranspiration**: This variable exhibits a somewhat normal distribution centered around 250, with values spreading from about 75 to 315. The frequency peaks between 200 and 300, indicating this is the most common range for evapotranspiration values.
- **PopulationDensity**: The population density data shows a right-skewed distribution with most values concentrated at lower densities. There is a gradual decrease in frequency as the population density increases, with few high-density observations.

- **Correlation Analysis**: Identify the correlation between predictor variables and the target variable (wildfire occurrence). This helps in understanding which variables are most strongly associated with wildfires.

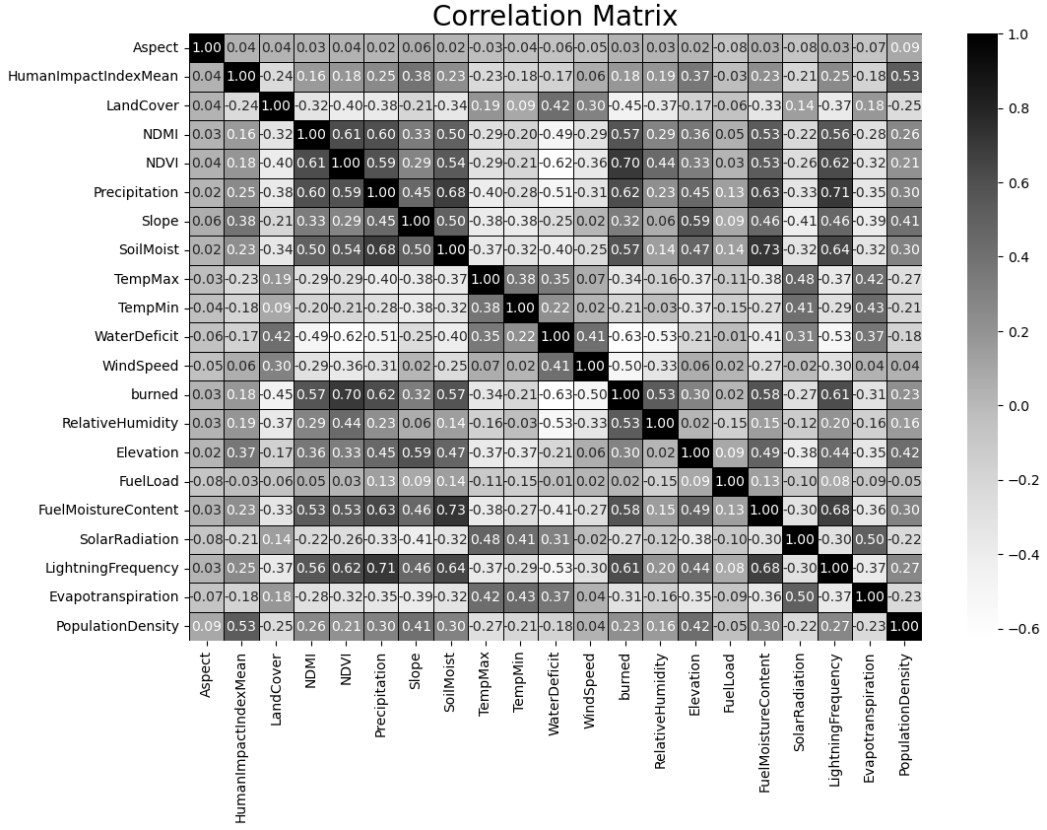

Figure (14): Correlation Matrix

1. **Strong Positive Correlations**:
    - **NDVI and Precipitation** (0.59): Higher vegetation density is associated with higher precipitation.
    - **NDVI and SoilMoist** (0.73): Areas with higher vegetation density tend to have higher soil moisture.
    - **WaterDeficit and Precipitation** (0.51): Areas with higher precipitation tend to have lower water deficit.
    - **TempMax and TempMin** (0.81): Maximum and minimum temperatures are strongly correlated.
    - **FuelMoistureContent and SoilMoist** (0.58): Higher soil moisture is associated with higher fuel moisture content.
    - **SolarRadiation and Evapotranspiration** (0.50): Higher solar radiation is associated with higher evapotranspiration.
    - **PopulationDensity and HumanImpactIndexMean** (0.53): Higher population density is associated with higher human impact.

2. **Strong Negative Correlations**:
    - **NDVI and TempMax** (-0.29): Higher vegetation density is associated with lower maximum temperatures.
    - **WaterDeficit and SoilMoist** (-0.62): Higher water deficit is associated with lower soil moisture.
    - **RelativeHumidity and TempMax** (-0.37): Higher relative humidity is associated with lower maximum temperatures.
    - **Evapotranspiration and SoilMoist** (-0.39): Higher evapotranspiration is associated with lower soil moisture.
3. **Interesting Patterns**:
    - **HumanImpactIndexMean and LandCover** (0.24): Areas with higher human impact tend to have distinct land cover types.
    - **NDMI and NDVI** (0.61): Strong positive correlation, indicating a relationship between vegetation indices.
    - **FuelLoad and SoilMoist** (0.09): Relatively weak correlation, suggesting fuel load is less dependent on soil moisture.

4.2 Data Preparation

Data preparation involves transforming raw data into a suitable format for analysis. This step is crucial for ensuring that the data fed into the models is clean, consistent, and appropriately scaled.

- **Separating Features and Target Variable**: The initial step in preparing the data for modeling involves separating the features (independent variables) from the target variable (dependent variable). In this context, the burned column is the target variable indicating whether an area was burned, and the remaining columns are the features used to predict this outcome.
- **Normalization and Scaling**: Normalizing or scaling the features ensures that they contribute equally to the analysis and improve the model's convergence during training. Features often have different units and scales, which can lead to biases in the model if not addressed. In the provided code, the StandardScaler from scikit-learn is used to standardize the features.So, a mean of zero and a standard deviation of one, making sure that each feature contributes proportionately to the model.
- **Data Splitting**: Splitting the data into training and testing sets is a critical step in preparing the data for model development. The training set is used to train the model, allowing it to learn the underlying patterns and relationships within the data. The testing set is then used to evaluate the model’s performance, providing an unbiased assessment of its predictive accuracy. This separation ensures that the model's performance is validated on unseen data, reducing the risk of overfitting and improving its generalizability to new data. In the provided code, the train_test_split function from scikit-learn is used to split the data. So, 80% of the data is used for training and 20% for testing, with a fixed random state to ensure reproducibility of the results.

4.3 Model Selection

This is a binary classification problem, where the goal is to classify areas as either burned or unburned. Different algorithms may be tested to determine which one performs best based on the nature of the data and the specific requirements of the study. The algorithms considered for this task include:

Model	Description
Logistic Regression	A statistical model that estimates the probability of a binary outcome based on one or more predictor variables. It is simple to implement and interpret, making it a good starting point for binary classification problems. [16]
Random Forest Classifier	An ensemble learning method that constructs multiple decision trees during training and outputs the mode of the classes for classification tasks. It is robust to overfitting and can handle large datasets with higher dimensionality. [17]
Gradient Boosting Classifier	An ensemble technique that builds models sequentially, with each new model attempting to correct errors made by the previous ones. It often provides high predictive accuracy and is effective for various types of data. [18]
Support Vector Machine (SVM)	A powerful classifier that works by finding the hyperplane that best separates the classes in the feature space. SVM is effective in high-dimensional spaces and is used for both linear and non-linear classification. [19]
XGBoost Classifier	An optimized implementation of gradient boosting that is designed for speed and performance. It includes various regularization techniques to prevent overfitting and often yields superior results in classification tasks. [20]
Neural Network	A computational model inspired by the human brain, capable of capturing complex patterns in the data through multiple layers of interconnected neurons. Neural networks are particularly useful for handling large amounts of data and capturing non-linear relationships. [21]

Table (10): Selected Models Description

4.4 Model Evaluation

Model evaluation is a crucial step in the machine learning workflow, as it helps to determine the effectiveness and reliability of the trained models. This step involves using the trained models to make predictions on the test data and then assessing their performance using various metrics. The evaluation process ensures that the models generalize well to unseen data and are capable of making accurate predictions in real-world scenarios.

1.4.1 Predictions

Once the models are trained, they are used to make predictions on the test data. This involves feeding the test features into the model and obtaining the predicted labels. The predictions are then compared to the actual labels in the test set to evaluate the model's performance.

1.4.2 Performance Metrics

To comprehensively evaluate the performance of the models, several metrics are used. Each metric provides different insights into the model's strengths and weaknesses:

- **Accuracy**: The proportion of correctly classified instances among the total instances. It gives a general sense of the model's performance but can be misleading if the classes are imbalanced.
- **Precision**: The proportion of true positive predictions among all positive predictions. It indicates how many of the predicted positive instances are actually positive.
- **Recall**: The proportion of true positive predictions among all actual positive instances. It measures the model's ability to identify all relevant instances.
- **F1 Score**: The harmonic mean of precision and recall. It provides a single metric that balances both precision and recall, especially useful when the classes are imbalanced.
- **ROC-AUC**: The area under the Receiver Operating Characteristic (ROC) curve. It represents the model's ability to distinguish between positive and negative classes across different threshold values. A higher ROC-AUC value indicates better model performance.

By using these performance metrics, we can gain a detailed understanding of how well the models are performing. For instance, high precision but low recall indicates that the model is conservative in its positive predictions, whereas high recall but low precision indicates that the model is more inclusive but at the cost of more false positives. The F1 score provides a balanced metric to evaluate the overall performance, while the ROC-AUC score gives insight into the model's capability to distinguish between classes. [22]

1.4.3 Cross-Validation

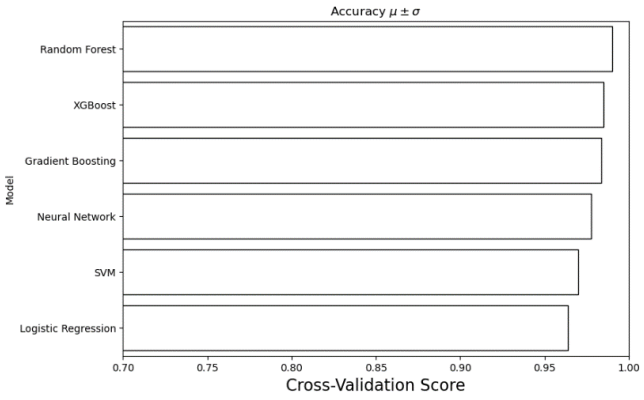

Figure (15): Cross-Validation Accuracy

To ensure the robustness and generalizability of the models, cross-validation is employed. This technique involves dividing the dataset into multiple subsets or folds. The model is trained on a combination of these folds and validated on the remaining fold. This process is repeated several times, with each fold being used as the validation set once. Cross-validation provides a more reliable estimate of model performance by reducing the variance associated with a single train-test split and ensuring that the model is evaluated on multiple subsets of the data. Common cross-validation techniques include k-fold cross-validation, where the data is divided into k equal-sized folds, and stratified k-fold cross-validation, which maintains the class distribution across folds. [23]

1.4.4 Visualization of Model Performance

Visualizing the performance of models provides a more intuitive understanding of how well they are performing. One effective way to do this is by plotting the ROC curves and confusion matrices for each model. The ROC curve shows the trade-off between the true positive rate and the false positive rate for different threshold values, while the confusion matrix provides a detailed breakdown of the true positives, true negatives, false positives, and false negatives. These visualizations help in comparing the performance of different models and understanding their strengths and weaknesses.

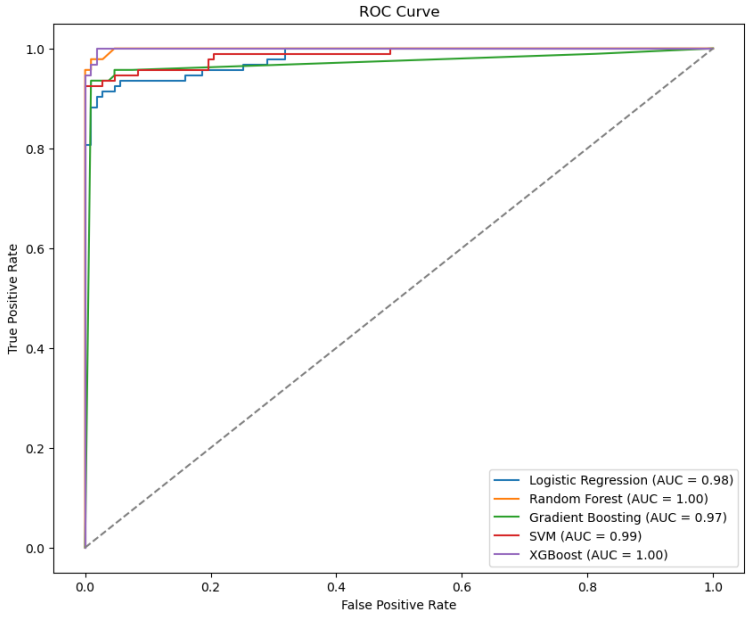

Figure (16): ROC Curve

The ROC curve compares the performance of different models in distinguishing between positive and negative classes. A higher curve closer to the top-left corner indicates better performance. Logistic Regression achieved an AUC of 0.98, showing strong performance. Random Forest and XGBoost both achieved perfect AUC scores of 1.00, indicating they perfectly separate the classes. Gradient Boosting also performed well with an AUC of 0.97, and SVM had an excellent performance with an AUC of 0.99. Overall, all models showed good performance, with Random Forest and XGBoost being the top performers.

The tables below summarize the performance of various models used for predicting wildfire distribution, grouped by type of model. The evaluation metrics include accuracy, classification reports for each model and the confusion matrix. [See Tables 11,12,13,14]

Linear Models

Model	Accuracy	Precision (Class 0)	Precision (Class 1)	Recall (Class 0)	Recall (Class 1)	F1 Score (Class 0)	F1 Score (Class 1)
Logistic Regression	0.9350	0.91	0.98	0.98	0.88	0.94	0.93

Table (11): Linear Models Performance

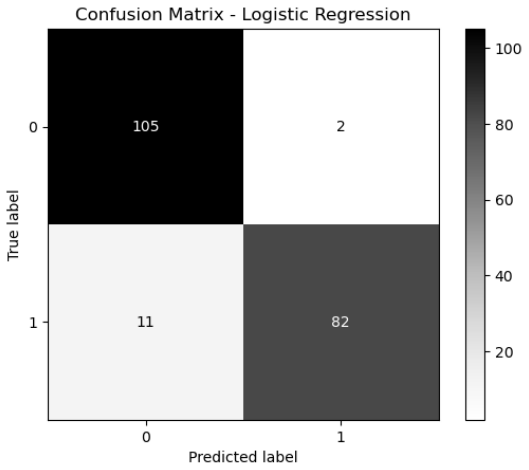

Figure (17): Confusion Matrix-Logistic Regression

By evaluating these metrics and visualizing the confusion matrix, we can conclude that the Logistic Regression model performs well for this binary classification task, making it a reliable choice for predicting wildfire distribution.

Ensemble Models

Model	Accuracy	Precision (Class 0)	Precision (Class 1)	Recall (Class 0)	Recall (Class 1)	F1 Score (Class 0)	F1 Score (Class 1)
Random Forest	0.9750	0.96	0.99	0.99	0.96	0.98	0.97
Gradient Boosting	0.9500	0.96	0.94	0.94	0.96	0.95	0.95
XGBoost	0.9800	0.98	0.98	0.98	0.98	0.98	0.98

Table (12): Ensemble Models Performance

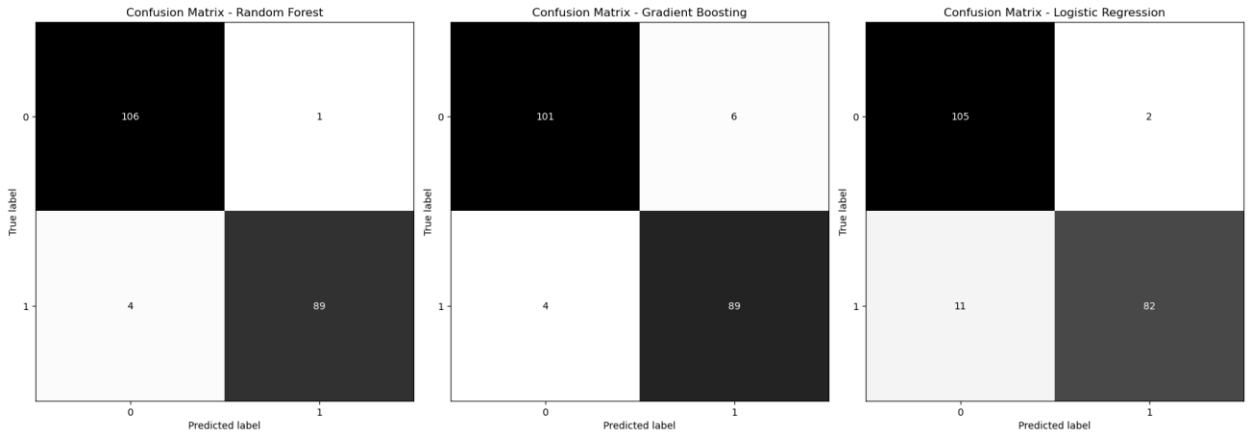

Figure (18): Ensemble Models Confusion Matrices

The confusion matrices illustrate the models' abilities to correctly predict both burned and unburned areas. High true positive and true negative counts indicate the models' effectiveness. Low false positive and false negative counts highlight their accuracy in distinguishing between the two classes. These ensemble models perform well for this binary classification task, making them reliable choices for predicting wildfire distribution.

Support Vector Machines

Model	Accuracy	Precision (Class 0)	Precision (Class 1)	Recall (Class 0)	Recall (Class 1)	F1 Score (Class 0)	F1 Score (Class 1)
SVM	0.9550	0.92	1.00	1.00	0.90	0.96	0.95

Table (13): Support Vector Machines Performance

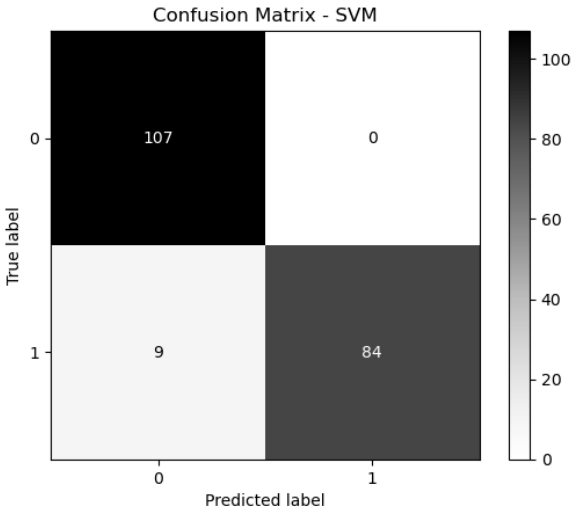

Figure (19): Confusion Matrix-SVM

Neural Networks

Model	Accuracy	Precision (Class 0)	Precision (Class 1)	Recall (Class 0)	Recall (Class 1)	F1 Score (Class 0)	F1 Score (Class 1)
Neural Network	0.9750	0.98	0.97	0.97	0.98	0.98	0.97

Table (14): Neural Networks Performance

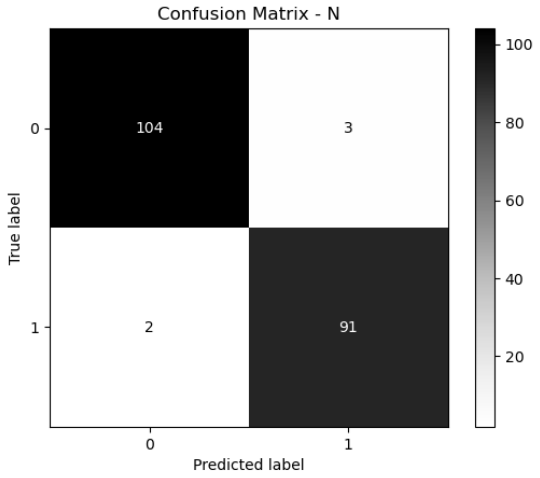

Figure (20): Confusion Matrix-Neural Network

*Both the SVM and Neural Network models perform well for this binary classification task.

5 Results

After evaluating each model, the top four models based on accuracy were selected for further analysis. These models include all the ensemble methods—Random Forest, Gradient Boosting, and XGBoost—as well as the Neural Network.

To better understand the factors that most significantly affect wildfire distribution, we will analyze the most important features from these top-performing models. By examining these feature importance graphs, we can identify which variables have the greatest impact on predicting wildfires. This analysis will provide valuable insights into the key drivers of wildfire occurrence and distribution, aiding in more effective wildfire management and prevention strategies.

- **Random Forest**

  In the Random Forest model, Wind Speed is the most significant factor affecting wildfire distribution. Fuel Moisture Content also plays a crucial role. Solar Radiation and Elevation contribute to wildfire behavior, though to a lesser extent. Evapotranspiration also influences wildfire occurrence, reflecting the impact of water loss and vegetation health. These factors together highlight the complexity of wildfire risk.

Factor	Importance
WindSpeed	0.113000
FuelMoistureContent	0.039833
SolarRadiation	0.007000
Elevation	0.007000
Evapotranspiration	0.006667

Table (15): Top 5 Factors-Random Forest

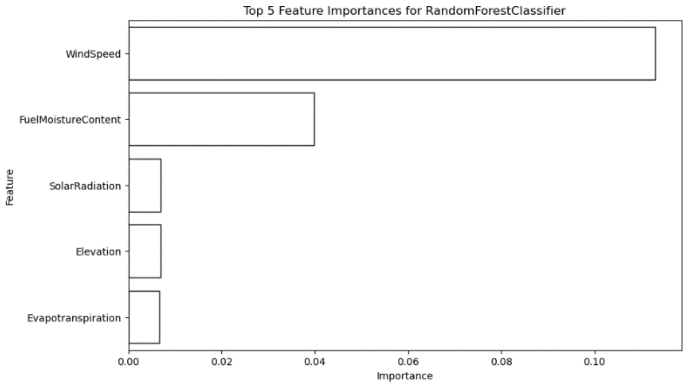

Figure (21): Top 5 Factors-Random Forest

- **Neural Network**

  In the Neural Network model, Wind Speed is the most significant factor influencing wildfire distribution. NDMI and NDVI, reflecting vegetation health and moisture, are also important. Fuel Moisture Content and Human Impact Index Mean further contribute to wildfire dynamics, highlighting the roles of environmental conditions and human activities.

Feature	Importance
WindSpeed	0.116000
NDMI	0.047000
NDVI	0.032500
FuelMoistureContent	0.028500
HumanImpactIndexMean	0.024667

Table (16): Top 5 Factors-Neural Network

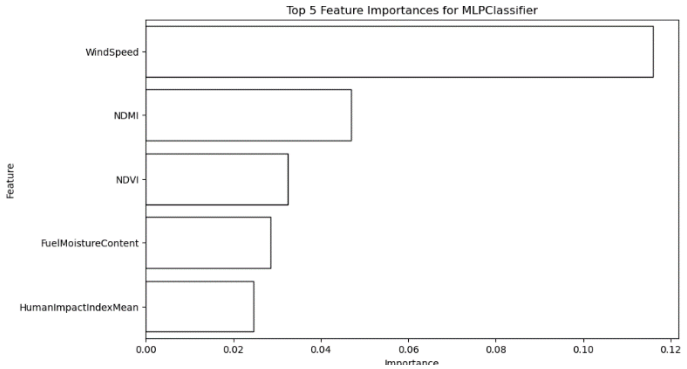

Figure (22): Top 5 Factors-Neural Network

- **Gradient boosting**

  In the Gradient Boosting model, Soil Moisture is the most significant factor influencing wildfire distribution. Wind Speed and Minimum Temperature are also crucial. Elevation and Solar Radiation contribute to a lesser extent, highlighting the roles of soil moisture, weather, and topography in wildfire dynamics.

Factor	Importance
SoilMoist	0.221667
WindSpeed	0.099333
TempMin	0.079833
Elevation	0.020833
SolarRadiation	0.007667

Table (17): Top 5 Factors-Gradient Boosting

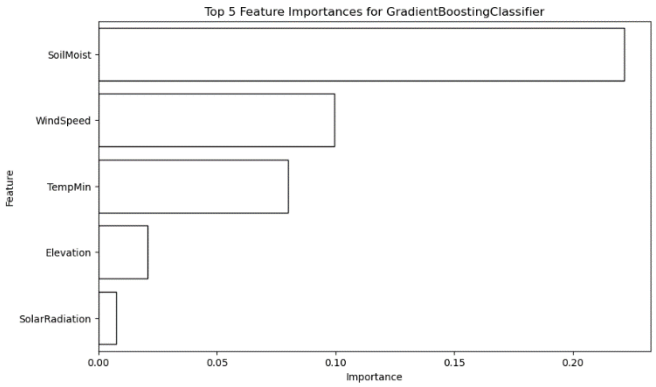

Figure (23): Top 5 Factors-Gradient Boosting

- **Xgboost**

  In the XGBoost model, Soil Moisture is the most significant factor influencing wildfire distribution. Wind Speed also plays a crucial role. Minimum Temperature, Solar Radiation, and Human Impact Index Mean contribute to a lesser extent, highlighting the importance of soil moisture, weather conditions, and human activities in wildfire dynamics.

Factor	Importance
SoilMoist	0.263167
WindSpeed	0.159500
TempMin	0.013500
SolarRadiation	0.010500
HumanImpactIndexMean	0.006833

Table (18): Top 5 Factors-Xgboost

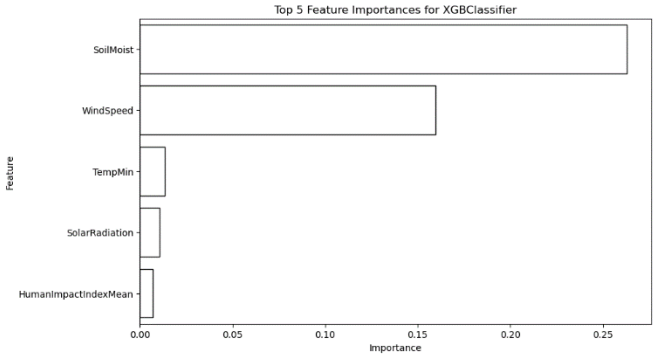

Figure (24):Top 5 Factors-Xgboost

Most Important Factors Affecting Wildfire Distribution

All four models—Random Forest, Neural Network, Gradient Boosting, and XGBoost—agree on the crucial factors affecting wildfire distribution, with wind speed consistently identified as the most significant factor.

[1] **Wind Speed**: All models highlight wind speed as the primary factor influencing wildfire distribution. This is because higher wind speeds:
- Increase fire spread and intensity.
- Enhance spotting and change fire direction.
- Promote the drying of fuels, increasing fire risk.

[2] **Soil Moisture**: Soil moisture is another critical factor identified by the models, particularly in Gradient Boosting and XGBoost.
- Soil moisture levels influence vegetation flammability and wildfire intensity.
- Dry soil conditions extend fire seasons and increase fire frequency and spread.

[3] **Solar Radiation**: Solar radiation is also an important factor across the models.
- Dries out vegetation, making it more susceptible to burning.
- Higher radiation levels can lead to more intense fires by increasing fuel flammability and combustion rates.

[4] **Elevation**: Elevation is recognized as a significant factor in the Random Forest and Gradient Boosting models.
- Elevated areas may experience drier conditions and stronger winds, facilitating fire spread.

[5] **Aspect**: Although not explicitly highlighted in all models, aspect is generally known to affect wildfire risk.
- Affects how much sunlight slopes receive, influencing vegetation dryness and fire risk.

These factors collectively underline the complex interplay between environmental conditions and wildfire dynamics. Wind speed, soil moisture, solar radiation, and elevation are crucial for understanding and predicting wildfire distribution, highlighting the importance of considering multiple variables in wildfire risk assessment and management.

6 Conclusion

This project thoroughly investigated the factors affecting wildfire distribution across different regions using geoinformatics and machine learning techniques. By employing models such as Logistic Regression, Random Forest, Gradient Boosting, SVM, XGBoost, and Neural Networks, we identified key variables like wind speed, soil moisture, solar radiation, and elevation as critical influencers of wildfire occurrences.

All models consistently highlighted wind speed as the most crucial factor, underscoring its significant impact on fire spread, intensity, and fuel drying. Soil moisture was also a vital variable, affecting vegetation flammability and fire seasons. Solar radiation and elevation further contributed to wildfire risk, emphasizing the roles of environmental conditions and topography.

The ensemble models and neural network demonstrated high accuracy and reliability in predicting wildfire distribution, with Random Forest and XGBoost achieving perfect AUC scores. These findings provide valuable insights for wildfire management and prevention strategies, emphasizing the importance of considering multiple environmental and human-related factors.

Overall, the integration of geoinformatics and machine learning proved effective in analyzing wildfire distribution, offering a robust framework for future research and practical applications in wildfire risk assessment and mitigation.

7 Limitations

Despite the promising results and insights gained from this project, several limitations were encountered during the analysis and modeling process:

1. **Computational Resources**: The complexity and size of the dataset required significant computational resources. While using Google Earth Engine (GEE) facilitated the handling of large geospatial data, the limited computational capacity and runtime constraints of GEE sometimes hindered the efficiency and speed of data processing.
2. **Model Training Time**: Training sophisticated models such as Neural Networks and ensemble methods like Gradient Boosting and XGBoost was time-consuming. This was particularly challenging when performing cross-validation and hyperparameter tuning, which are essential for optimizing model performance.
3. **Data Quality and Availability**: The quality and granularity of the input data can significantly affect the model's performance. Some variables may have missing values or inconsistencies, which can introduce bias and reduce the accuracy of the predictions. Additionally, certain environmental and human-related factors influencing wildfire risk may not be adequately captured in the available data.
4. **Feature Importance Interpretation**: While the models can identify important features influencing wildfire distribution, interpreting these features and understanding their interactions can be challenging. The complex relationships between variables may not always be straightforward, requiring further domain-specific knowledge and analysis.
5. **Geographical and Temporal Generalization**: The models were trained and validated on specific geographical regions and time periods. Consequently, their ability to generalize to other regions or different temporal scales may be limited. Local environmental conditions, vegetation types, and climate variability can differ significantly, impacting the models' predictive performance in new contexts.
6. **Simplified Assumptions**: Some models may rely on simplified assumptions that do not fully capture the complexities of wildfire dynamics. For instance, assuming independence between certain variables or linear relationships can oversimplify the real-world interactions influencing wildfires.
7. **Run-time Constraints**: The execution time for processing large datasets and training complex models was often prolonged, which could delay the analysis and limit the ability to quickly iterate and refine models.
8. **External Factors**: Wildfire occurrence is influenced by numerous external factors such as government policies, human activities, and unexpected climatic events. These factors are difficult to quantify and incorporate into predictive models, yet they play a crucial role in real-world wildfire dynamics.

Addressing these limitations in future work could involve leveraging more advanced computational resources, improving data quality and coverage, and developing more sophisticated models that better capture the complex interactions between variables. Additionally, expanding the analysis to cover more diverse geographical regions and time periods would enhance the generalizability and robustness of the models.

8 Future Development

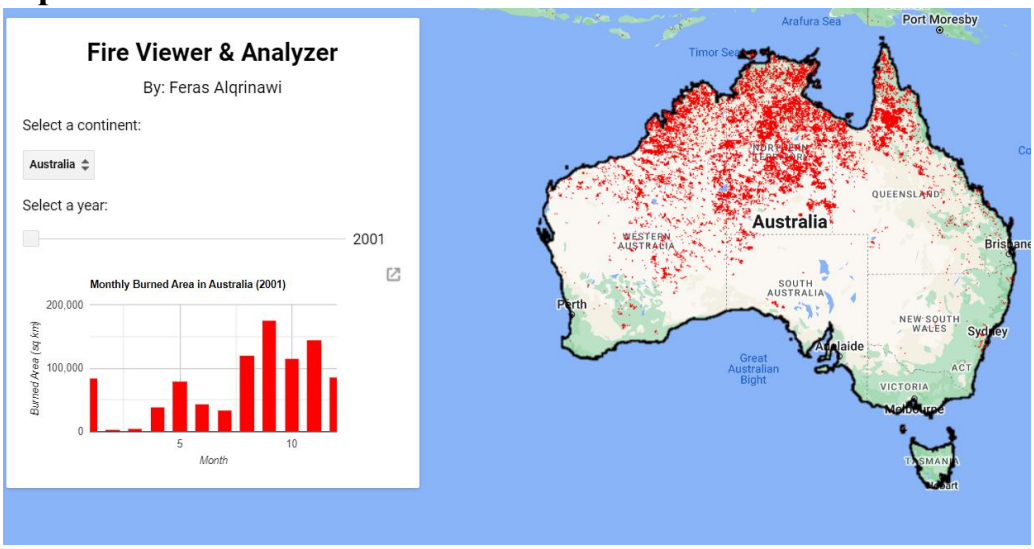

Figure (25):Fire Viewer and analyzer

The "Fire Viewer & Analyzer" application, developed by Feras Alqrinawi using Google Earth Engine (GEE), extends this project by visualizing and analyzing wildfire data. Future enhancements can further improve this tool and the overall project, offering deeper insights and more robust wildfire management solutions.

Future Enhancements for the "Fire Viewer & Analyzer" Application

1. **Real-Time Data Integration**:
    - Incorporate real-time fire detection data from satellite sources.
    - Provide up-to-date information on active fires.
    - Enhance the application's responsiveness to emerging fire events.
2. **Predictive Modeling**:
    - Integrate machine learning models for wildfire prediction.
    - Utilize historical data and environmental conditions for accurate forecasts.
    - Offer advanced warning systems for potential wildfire occurrences.
3. **User-Defined Analysis**:
    - Allow users to select specific regions and time periods for analysis.
    - Enable custom visualizations based on user-defined parameters.
    - Facilitate targeted investigations and localized risk assessments.

Broader Project Enhancements

1. **Expanded Data Sources**:
    - Incorporate high-resolution satellite imagery and ground-based sensor data.
    - Improve data accuracy and comprehensiveness.
2. **Improved Data Quality and Coverage**:
    - Extend the temporal scope to analyze long-term trends.
3. **Interdisciplinary Collaboration**:
    - Collaborate with experts in ecology, meteorology, and firefighting.
    - Refine models and validate findings with expert insights.

Acknowledgments

I am deeply grateful to my supervisors, Prof. Stroppiana and Prof. Venuti, for their invaluable guidance and support throughout this project. Their expert advice and direction were crucial in navigating the research and ensuring its success. I would also like to extend my heartfelt thanks to Prof. Fugini for organizing this course. Her efforts in creating such an engaging and educational experience have provided me with the opportunity to get into this fascinating project and gain profound insights into wildfire management and data analysis.

9 References

[1] R. L. Hutto, "THE ECOLOGICAL IMPORTANCE OF SEVERE WILDFIRES: SOME LIKE IT HOT," *Ecological Applications,* vol. 18, no. 8, pp. 1827-1834, 2008.

[2] K. . Krajick, "Fire in the hole," *Smithsonian Magazine,* vol. , no. , p. , .

[3] R. L. Wade, A. . Jokar, K. . Cydzik, A. . Dershowitz, R. . Bronstein and R. . Bronstein, "Wildland fire ash and particulate distribution in adjacent residential areas," *International Journal of Wildland Fire,* vol. 22, no. 8, pp. 1078-1082, 2013.

[4] M. . Roy, M. . Roy, R. D. Zinck, M. J. Bouma, M. . Pascual and M. . Pascual, "Epidemic cholera spreads like wildfire," *Scientific Reports,* vol. 4, no. 1, pp. 3710-3710, 2015.

[5] L. . Tacconi, "Fires in Indonesia: Causes, Costs, and Policy Implications (CIFOR Occasional Paper No. 38)," *Occasional Paper,* vol. , no. , p. , .

[6] O. . Rackham, "Fire in the European Mediterranean: History," *AridLands Newsletter,* vol. 54, no. , p. , .

[7] R. A. Bradstock, R. A. Bradstock, M. M. Boer, M. M. Boer, M. M. Boer, G. J. Cary, G. J. Cary, O. . Price, R. J. Williams, R. J. Williams, D. . Barrett, G. D. Cook, A. M. Gill, A. M. Gill, L. B. Hutley, H. . Keith, S. W. Maier, M. . Meyer, S. H. Roxburgh and J. . Russell-Smith, "Modelling the potential for prescribed burning to mitigate carbon emissions from wildfires in fire-prone forests of Australia," *International Journal of Wildland Fire,* vol. 21, no. 6, pp. 629-639, 2012.

[8] J. M. Silveira, J. . Barlow, R. B. d. Andrade, J. . Louzada, L. A. M. Mestre, L. A. M. Mestre, S. . Lacau, R. . Zanetti, I. . Numata and M. A. Cochrane, "The responses of leaf litter ant communities to wildfires in the Brazilian Amazon: a multi-region assessment," *Biodiversity and Conservation,* vol. 22, no. 2, pp. 513-529, 2013.

[9] L. B. Lentile, P. . Morgan, A. T. Hudak, M. J. Bobbitt, S. A. Lewis, A. M. S. Smith and P. R. Robichaud, "Post-fire burn severity and vegetation response following eight large wildfires across the western United States," *Fire Ecology,* vol. 3, no. 1, pp. 91-108, 2007.

[10] "Wildland Fire: History Timeline," , . [Online]. Available: http://www.nps.gov/fire/wildland-fire/learning-center/fireside-chats/history-timeline.cfm. [Accessed 7 7 2024].

[11] A. D. Syphard, V. . Butsic, A. . Bar-Massada, J. E. Keeley, J. A. Tracey and R. N. Fisher, "Setting priorities for private land conservation in fire-prone landscapes: Are fire risk reduction and biodiversity conservation competing or compatible objectives?," *Ecology and Society,* vol. 21, no. 3, p. , 2016.

[12] D. M. J. S. Bowman, B. P. Murphy, B. P. Murphy, M. M. Boer, R. A. Bradstock, G. J. Cary, M. A. Cochrane, R. J. Fensham, M. A. Krawchuk, O. . Price and R. J. Williams, "Forest fire management, climate change, and the risk of catastrophic carbon losses," *Frontiers in Ecology and the Environment,* vol. 11, no. 2, pp. 66-68, 2013.

[13] "U.S. Climate Data," , . [Online]. Available: http://www.usclimatedata.com. [Accessed 7 7 2024].

[14] "Google Earth," , . [Online]. Available: https://www.google.com/earth/. [Accessed 7 7 2024].

[15] X. . Kong, Y. . Qian and A. . Zhang, "Cloud and shadow detection and removal for Landsat-8 data," , 2013. [Online]. Available: https://spiedigitallibrary.org/conference-proceedings-of-spie/8921/1/cloud-and-shadow-detection-and-removal-for-landsat-8-data/10.1117/12.2031120.full. [Accessed 7 7 2024].

[16] V. . Bewick, L. . Cheek and J. . Ball, "Statistics review 14: Logistic regression," *Critical Care,* vol. 9, no. 1, pp. 112-118, 2005.

[17] M. . Pal, "Random forest classifier for remote sensing classification," *International Journal of Remote Sensing,* vol. 26, no. 1, pp. 217-222, 2005.

[18] L. . Lin, W. . Yue and Y. . Mao, "Multi-class Image Classification Based on Fast Stochastic Gradient Boosting," *Informatica (lithuanian Academy of Sciences),* vol. 38, no. 3, pp. 145-153, 2014.

[19] S. . Suthaharan, "Support Vector Machine," , 2016. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-030-17989-2_8. [Accessed 7 7 2024].

[20] T. . Chen and C. . Guestrin, "XGBoost: A Scalable Tree Boosting System," *arXiv: Learning,* vol. , no. , pp. 785-794, 2016.

[21] T. . Hill, L. . Marquez, M. . O'Connor and W. . Remus, "Artificial neural network models for forecasting and decision making," *International Journal of Forecasting,* vol. 10, no. 1, pp. 5-15, 1994.

[22] B. . Yu and J. L. Gastwirth, "HOW WELL DO SELECTION MODELS PERFORM? ASSESSING THE ACURACY OF ART AUCTION PRE-SALE ESTIMATES," *Statistica Sinica,* vol. 20, no. 2, pp. 837-852, 2010.

[23] M. W. Browne, "Cross-validation methods," *Journal of Mathematical Psychology,* vol. 44, no. 1, pp. 108-132, 2000.

10 Appendices