A Project Report
On

Predictive Modelling of Forest Fires in India

By

Rohan Jha
2021A7PS2721G

Under the supervision of

Dr. Rajiv Kumar Chaturvedi

Submitted For

BITS F329: Project on Social and Environmental Applications of Data Science

BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE PILANI
GOA CAMPUS

DECEMBER 2024

TABLE OF CONTENT:

ABSTRACT

This paper presents a comprehensive framework for predicting forest fire probabilities using advanced machine learning techniques, diverse data integration, and an intuitive web platform. Leveraging data from Google Earth Engine (GEE) for vegetation and Open Meteo for real-time weather and elevation, the system captures the complex factors influencing wildfire risks. The study evaluates Neural Networks (88% accuracy), Linear Regression (72%), Random Forest Regressor (93%), and XGB Regressor (94%), utilizing 20 key features, including weather parameters, vegetation indices, and geographical data. Rigorous preprocessing, hyperparameter tuning, and testing ensure a balance between accuracy and efficiency. The PyroPredictor website demonstrates the real-world application of the trained models, offering users an interactive and user-friendly platform to explore wildfire probabilities across India. Features such as color-coded risk maps, location-specific predictions, and customizable parameter weights empower researchers, policymakers, and communities to make data-driven decisions for wildfire prevention and management. PyroPredictor sets a benchmark for combining data-driven insights with practical tools to address pressing environmental challenges.

I. INTRODUCTION

Forest fires in India have become a major environmental challenge, posing serious threats to biodiversity, human lives, and local economies. In recent years, the frequency and intensity of these fires have increased, largely driven by a complex interplay of climatic, geographical, and anthropogenic factors. The destruction of forests not only results in the loss of valuable ecosystems and wildlife habitats but also contributes to air pollution, carbon emissions, and the disruption of local water cycles. With vast stretches of forested land, varying climatic conditions, and growing human encroachment, predicting and managing forest fires in India has become more challenging than ever.

Addressing this issue requires advanced methods for forecasting forest fires, enabling timely intervention and more effective resource allocation. Recent advancements in data science and machine learning offer an opportunity to proactively predict forest fires based on real-time environmental conditions. This paper presents a predictive model aimed at forecasting the likelihood and severity of forest fires across different regions of India. By integrating 20 environmental features, such as temperature, humidity, wind speed, and vegetation type, the model seeks to identify the underlying patterns that drive fire occurrences.

The website aims to provide actionable insights to guide decision-making in fire prevention and mitigation strategies. It assists in identifying fire-prone areas, optimizing resource allocation, and contributing to long-term forest management plans. By reducing risks, protecting biodiversity, and supporting sustainable conservation, it plays a critical role in addressing the growing threat of forest fires. The platform integrates advanced analytics and cutting-edge algorithms to deliver data-driven predictions tailored to specific regions. Its user-friendly interface ensures accessibility for policymakers, researchers, and forest management teams alike. This initiative underscores a commitment to leveraging technology for ecological preservation and proactive environmental stewardship.

II. DATA COLLECTION FOR FOREST FIRES

A. GDAC Wildfire Incident Data

The Global Disaster Alert and Coordination System (GDACS) is a real-time web-based information system that provides alerts about natural disasters worldwide, including earthquakes, tsunamis, tropical cyclones, and other major events. GDACS consolidates data from various sources, such as national and international monitoring agencies, to deliver real-time information, which is critical for response coordination and disaster management. GDACS is considered a credible source of information due to its backing by authoritative international organizations like the UN and the European Commission, its use of official monitoring networks (such as seismological and meteorological data from trusted global agencies), and its commitment to providing timely and accurate disaster-related data. This makes it a trusted resource for decision-makers in disaster preparedness, response, and recovery.

For the development of the forest fire prediction model, comprehensive data was gathered from the Global Disaster Alert and Coordination System (GDACS), which provided a detailed record of over 500 significant wildfire incidents across India over the past several years. This dataset includes essential information such as the dates of each wildfire, their geographical locations, and the severity levels of the fires. The inclusion of such data offers a robust foundation for understanding the dates, the frequency, and intensity of forest fires within the Indian subcontinent. By analyzing the spatiotemporal distribution and severity of these incidents, the model could identify trends and correlations that are pivotal in predicting future wildfire occurrences.

Using data from the past five years is appropriate for this model, as it reflects current population densities, vegetation types, and climate conditions. Going further back would introduce inaccuracies due to significant changes in human activity, land use, and environmental factors, such as shifts in vegetation, climate patterns, and urban development, which could skew the results and make predictions less reliable.

Figure 1: Forest Fire Incidents in India from 2023 to 2024 according to GDAC

B. Open Meteo Weather and Elevation Data

Weather data for the days when forest fires started was sourced from Open Meteo, a global weather service that provides freely accessible historical, current, and forecast data. Open Meteo leverages a wide array of publicly available weather data sources and advanced weather models to offer accurate and timely climate information. Its data spans over 45 years, making it a credible and valuable resource for historical weather analysis, particularly for research and forecasting applications related to natural disasters like forest fires. Open Meteo's commitment to transparency and the accuracy of its models has established it as a reliable source for weather data in a variety of fields, including climate studies, disaster management, and environmental monitoring.

To enrich the predictive capabilities of the forest fire model, I incorporated a range of weather variables that may directly influence fire behavior. These variables, which describe atmospheric and environmental conditions on the specific days when fires occurred, include:

- Relative Humidity: The amount of moisture in the air relative to its maximum capacity. Low humidity increases the risk of wildfires by drying out vegetation.
- Dew Point: The temperature at which air becomes saturated with moisture. A lower dew point suggests drier conditions, which can increase fire risk.
- Surface Pressure: The atmospheric pressure at ground level, influencing wind patterns and weather systems, that can impact fire behavior.
- Cloud Cover: The fraction of the sky covered by clouds. Clear skies often correlate with higher temperatures and drier conditions, increasing fire risk.
- Wind Speed: Strong winds can spread fires rapidly by carrying embers, making wind speed a critical factor in fire dynamics.
- Soil Temperature: The temperature of the soil surface, affecting vegetation dryness and flammability.
- Soil Moisture: The water content in the soil, with dry soils contributing to increased fire risk by reducing vegetation moisture.
- Direct Radiation: The amount of solar radiation hitting the earth’s surface. Higher radiation enhances fire risk.
- Weather Code: A categorical indicator of weather conditions (e.g., clear, rainy), providing a summary of atmospheric conditions.
- Temperature: Air temperature, which directly influences the moisture content of vegetation and soil, impacting fire potential.
- Rain Sum: Total precipitation over a given period. Rainfall reduces fire risk by increasing moisture in vegetation and soil.
- Evapotranspiration: The combined process of water evaporation and plant transpiration, which depletes moisture in soil and vegetation, increasing fire susceptibility.

In addition to the weather data, elevation data was also obtained from Open Meteo. Elevation plays a significant role in wildfire behavior, as higher altitudes often have different climate patterns, vegetation types, and fire dynamics compared to lower elevations. Areas at higher elevations may experience cooler temperatures, reduced atmospheric pressure, and variations in vegetation that influence fire risk. By incorporating elevation data, the model accounts for these geographic differences, helping to better understand how terrain can affect the likelihood and severity of forest fires. This data adds an essential layer to the predictive model, improving its ability to capture the complexities of fire behavior across diverse landscapes in India.

C. MODIS/061/MOD13A1 Vegetation Data

For this study, I utilized Google Earth Engine (GEE), a powerful cloud-based platform for geospatial analysis, which provides access to an extensive archive of satellite imagery and geospatial datasets. GEE offers tools for processing large-scale environmental data, allowing researchers to perform analysis without needing to manage complex computational infrastructure. It is widely used in the scientific community for applications such as land use monitoring, environmental modeling, and disaster management due to its reliable and up-to-date datasets from global satellites.

One of the primary datasets used in this study is MODIS (Moderate Resolution Imaging Spectroradiometer), a key instrument aboard NASA's Terra and Aqua satellites. MODIS provides global, high-resolution data on various environmental variables, including vegetation indices like NDVI (Normalized Difference Vegetation Index) and EVI (Enhanced Vegetation Index). These indices are crucial for assessing the health, density, and greenness of vegetation, which are key indicators in fire risk assessment. MODIS has been collecting data for over two decades, making it a valuable source for historical and current vegetation analysis.

The specific product used in this study is MODIS/061/MOD13A1, which provides 16-day composite images of NDVI and EVI at a spatial resolution of 500 meters. NDVI is a widely used index to monitor vegetation health, where higher values indicate more robust and healthier vegetation. EVI, on the other hand, corrects for atmospheric influences and provides more accurate measurements in areas with dense vegetation, making it especially useful in tropical and subtropical regions like India. By examining the NDVI and EVI values just before forest fires occurred, I was able to assess the condition of vegetation and its potential flammability.

The methodology employed to collect vegetation data using Google Earth Engine (GEE) is detailed in the following steps, focusing on the extraction of NDVI (Normalized Difference Vegetation Index) and EVI (Enhanced Vegetation Index) values for specific wildfire locations prior to the occurrence of fires.

1. Accessing MODIS Data: The MODIS dataset (MODIS/061/MOD13A1) is filtered by date using the ee.Filter.date(START_DATE, END_DATE) function to focus on the time period prior to each wildfire.
2. Sampling NDVI and EVI: The function sample_indices(image) uses the reduceRegion method to calculate the mean NDVI and EVI values for the specific location of each fire. This is done at a 500-meter resolution, and the date, NDVI, and EVI values are extracted for each image.
3. Mapping the Sampling Function: The sampling function is applied to the filtered image collection using the map() function. This processes each image to extract the required vegetation indices.
4. Creating a Feature Collection: The resulting data is stored as a FeatureCollection, where each feature contains the date and corresponding NDVI/EVI values for that fire location.
5. Extracting the Data: The data is retrieved using getInfo(), which converts the feature collection into a list of date-specific NDVI and EVI values for further analysis.

This methodology is effective because it provides precise, location-specific data on vegetation health (through NDVI and EVI) just before the forest fires. By using Google Earth Engine, we are able to process large volumes of satellite data efficiently, ensuring that the vegetation indices accurately reflect the conditions that may have contributed to the ignition or spread of wildfires.

D. MODIS/061/MOD13A Vegetation Data

The process for collecting Leaf Area Index (LAI) and Fraction of Photosynthetically Active Radiation (FPAR) data is similar to the previous methodology but focuses on different vegetation indices. Here's the essential explanation:

1. Accessing MODIS LAI and FPAR Data: The MODIS MCD15A3H dataset provides LAI and FPAR values, which are important indicators of vegetation health and productivity. The dataset is filtered by date using the ee.Filter.date(START_DATE, END_DATE) to focus on the period before each wildfire event.
2. Sampling LAI and FPAR: The function sample_indices(image) calculates the mean LAI and FPAR values for a specific point location using the reduceRegion method. This is done at a 500-meter resolution. The function extracts the values of LAI and FPAR for each image, along with the date.
3. Mapping the Sampling Function: The map() function is used to apply the sample_indices function across the filtered ImageCollection. This extracts the LAI and FPAR values for each image in the dataset for the specified time period.
4. Creating a Feature Collection: The results are stored in a FeatureCollection where each feature contains the date and the corresponding LAI and FPAR values for that fire location.
5. Extracting the Data: The data is retrieved using getInfo(), which converts the feature collection into a list of LAI and FPAR values for each relevant date, ready for analysis.

LAI (Leaf Area Index): LAI is a measure of the total leaf area per unit ground area. It is an important parameter for understanding vegetation density and health, as a higher LAI indicates more leaf cover, which can impact wildfire risk.

FPAR (Fraction of Photosynthetically Active Radiation): FPAR is the fraction of solar radiation absorbed by vegetation, which reflects the plant's ability to carry out photosynthesis. Higher FPAR values suggest healthier vegetation that may have higher fuel potential for wildfires. This methodology allows for the integration of LAI and FPAR data into the wildfire prediction model, providing additional insights into vegetation health and its role in fire dynamics.

Together, these indices offer a comprehensive understanding of vegetation health, fuel availability, and ecological stress, which are directly correlated with wildfire risk. By incorporating these parameters, the model can better predict areas with high wildfire potential, considering not just the vegetation type and density but also its overall health and photosynthetic activity. Furthermore, the combination of these indices allows the model to capture different aspects of vegetation dynamics and how they interact with environmental factors.

In addition to their direct role in fire prediction, these indices also provide insights into human establishment and activity. For example, areas with low NDVI or EVI values may indicate urbanized regions, agricultural land, or areas heavily impacted by human activity. These indices can, therefore, help infer regions with significant human establishment, where land use and land cover changes, such as deforestation, urban sprawl, or agricultural development, could influence fire risks. By understanding the relationship between human activities and these vegetation parameters, the model can identify how changes in land use patterns contribute to the overall wildfire risk, enhancing its ability to make predictions for both natural and human-affected areas.

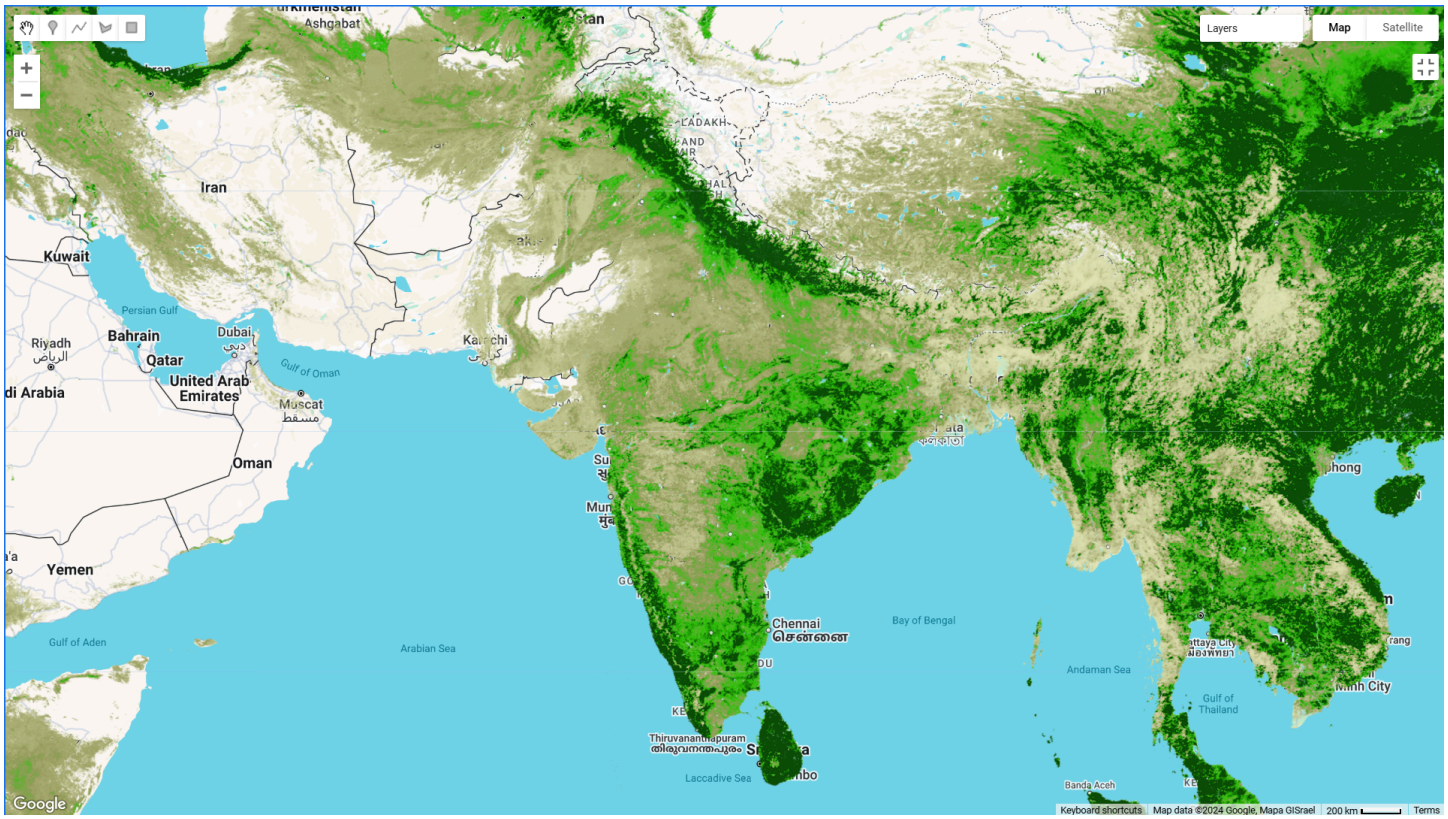


Figure 2: Vegetation Coverage in India as of 30th Nov 2024, based on MODIS/061/MCD15A3H Data

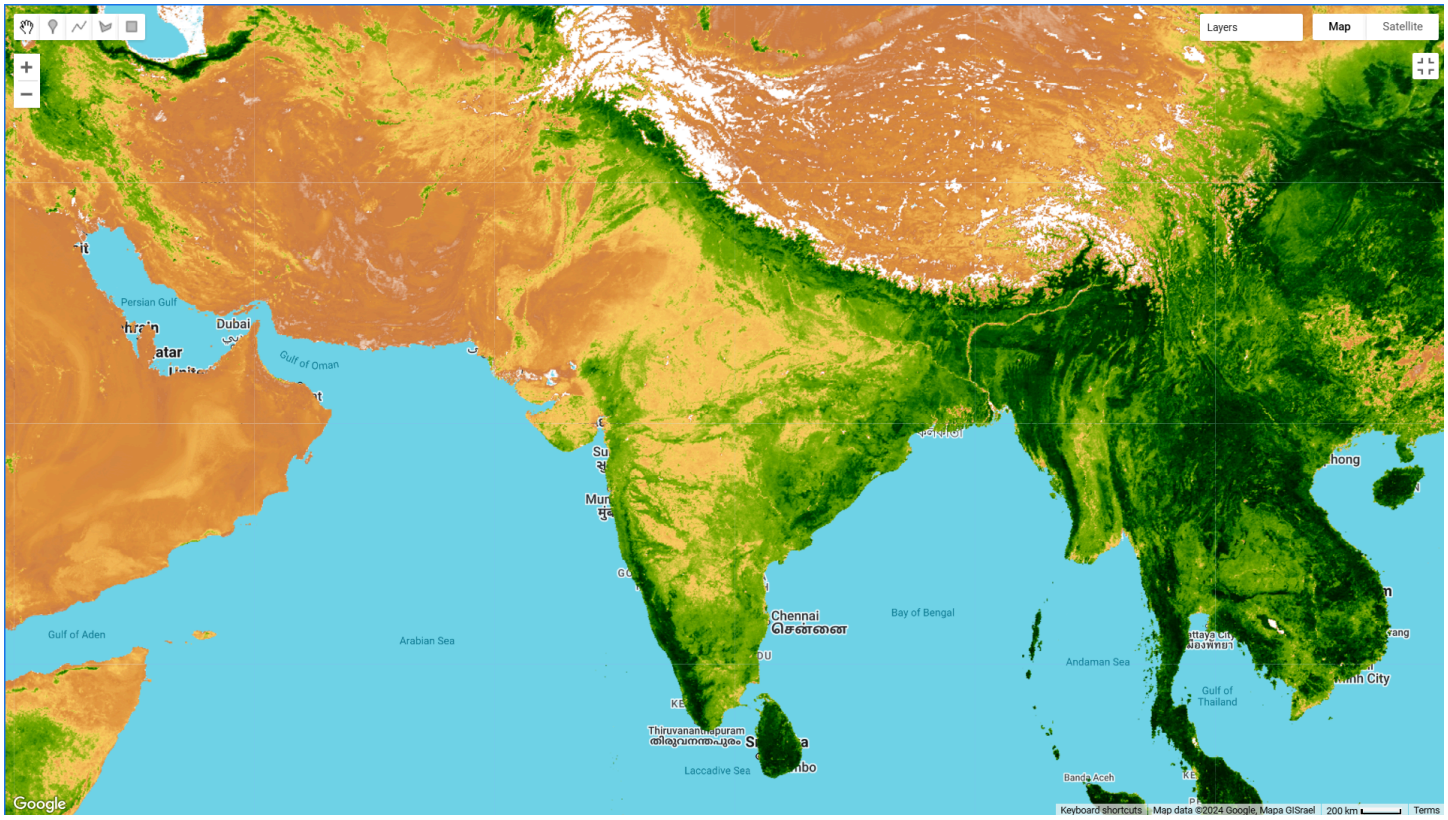


Figure 3: Vegetation Coverage in India as of 30th Nov 2024, based on MODIS/061/MCD15A3H data

III. DATA COLLECTION FOR NON FIRES

To complement the forest fire data, it was crucial to identify regions that have not been affected by forest fires, creating a control group for comparison in the analysis. These non-fire regions serve as a control group, providing a baseline against which the characteristics of fire-prone areas can be compared. The non-fire regions were selected using a systematic procedure involving satellite data on fire activity, geographical boundaries, and random point generation. The procedure for selecting these non-fire regions was carried out systematically, ensuring high accuracy and relevance to the research objectives.

1. Fire Hotspot Identification and Buffer Creation

The first step in identifying non-fire regions was to use satellite data to determine the locations and intensity of forest fires. To do this, the FIRMS (Fire Information for Resource Management System) dataset, which provides real-time fire data from the MODIS satellite, was utilized. This dataset includes thermal anomalies, which are indicative of areas with fire activity. By aggregating fire activity data over the specified period (from January 1, 2019, to November 20, 2024), significant fire hotspots across the study area of India were identified.

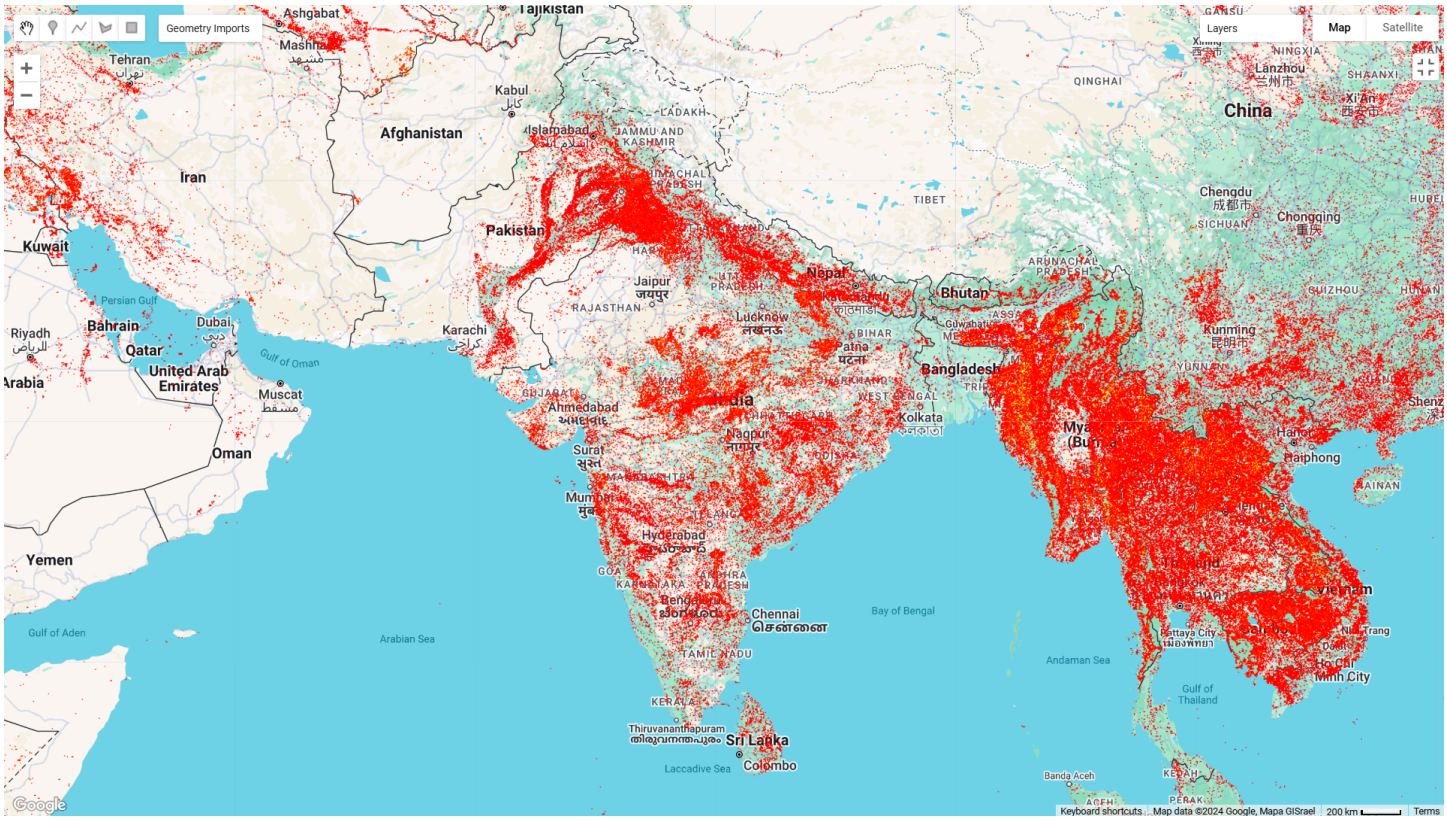


Figure 4: Fire Hotspots in India from 2019 to 2024, based on MODIS MOD14/MYD14

Once fire hotspots were determined, a buffer zone was created around these areas. The buffer zone serves to account for regions that may have been affected by fire but did not necessarily show up as hotspots in the dataset. These buffer zones extended 10 kilometers around the fire-prone areas, recognizing the possibility of post-fire environmental changes or ongoing fire risks in the vicinity. This was achieved through a focal analysis that expanded the areas of known fire activity to include adjacent regions likely to have been influenced by the fire.

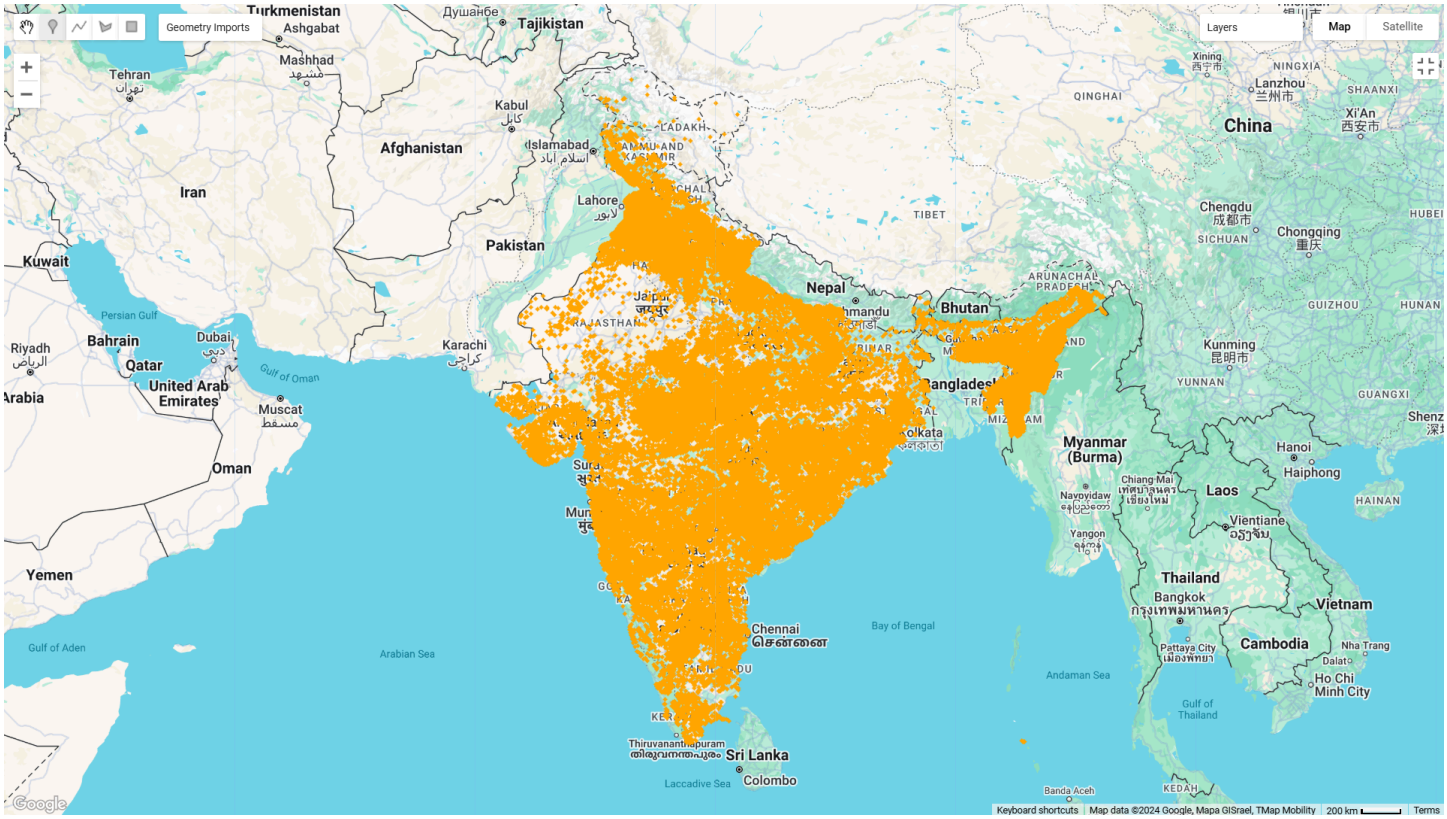


Figure 5: Fire Buffer Zones identified in India from 2019 to 2024

2. Generation of Random Points in Non-Fire Regions

With the fire hotspot and buffer areas established, the next step was to generate random data points in regions not affected by fire. To ensure the selection of valid non-fire regions, random points were generated across the entire study area of India using a random point generation method. A total of 1,000 points were selected, providing a comprehensive sample for the analysis.

However, not all of these random points fell outside the identified fire-affected areas. To address this, the generated points were filtered to exclude any that intersected with the buffer zones created around the fire hotspots. This process effectively removed points located in areas of significant fire activity or post-fire impact, ensuring that only points in regions free from fire influence were included in the dataset.

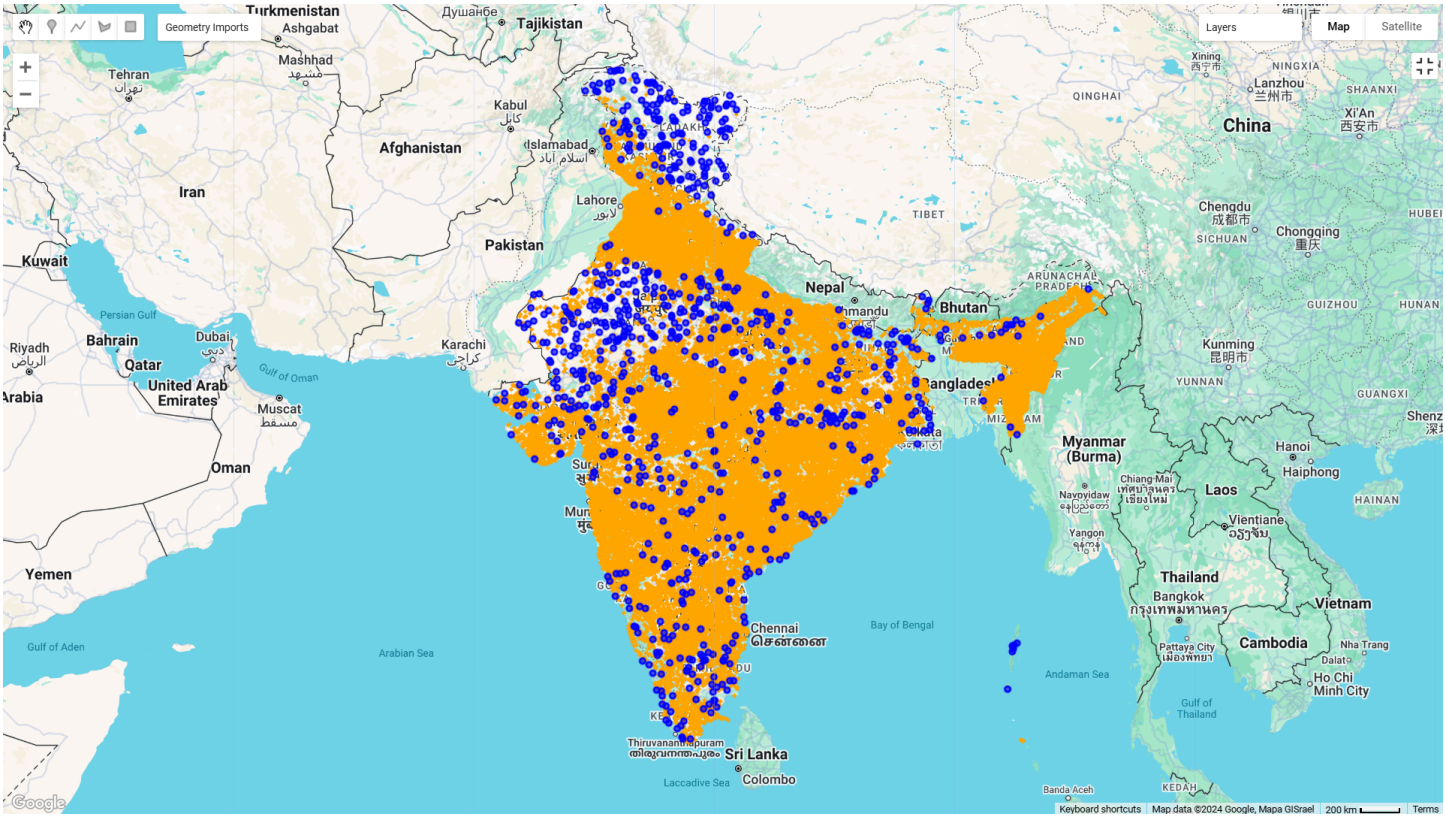


Figure 6: Some Non-Fire Points sampled from places in India outside the buffer zone

3. Random Dates Assignment

Each of the randomly selected points was assigned a random date within the specified time frame (from January 1, 2019, to November 20, 2024). This randomization of dates ensured that the data collection process was not biased by seasonal patterns or specific fire events. It also helped to create a diverse set of data points that were spread across different times, providing a more representative sample of non-fire conditions throughout the study period.

This method of collecting non-fire data is highly accurate due to the use of reputable and up-to-date satellite data for fire activity, along with well-defined geographical boundaries. The use of buffer zones around fire-prone areas is particularly important, as it accounts for environmental effects that may extend beyond the immediate fire locations, such as changes in vegetation, soil conditions, and air quality. By filtering out points that intersect with fire-affected regions, the methodology ensures that only genuinely unaffected areas are selected. Furthermore, the random point generation within the well-defined boundaries of India ensures that the non-fire regions are widely distributed across the study area. This diversity helps in avoiding any regional or temporal biases that could skew the results.

In conclusion, the methodology for selecting non-fire data points provides a reliable and accurate control group for comparison with fire-affected areas, ensuring that the model can make valid distinctions between the characteristics of regions with and without forest fires. This comprehensive approach to data collection enhances the overall quality of the research and contributes to the model's ability to predict forest fire occurrences with high precision.

4. Vegetation, Elevation and Weather Data

For the non-fire points, I similarly collected the weather conditions, vegetation indices, and elevation data for the random dates assigned to each point. The weather data, including parameters like temperature, humidity, wind speed, and radiation, was sourced from Open Meteo's historical database for the specific dates of the non-fire points. This allowed me to capture the environmental conditions that prevailed during those times. I retrieved satellite-derived vegetation indices—NDVI, EVI, LAI, and FPAR—from MODIS satellite products to assess the vegetation health and photosynthetic activity of the regions on the given dates. These indices are important for understanding how the vegetation in non-fire regions differs from fire-prone areas in terms of growth and vitality. Additionally, elevation data was collected from Open Meteo, providing insights into the topography of each point, which is essential for evaluating how landscape features might influence fire behavior. This comprehensive data collection for the non-fire points allowed for a detailed comparison with fire-affected regions, supporting the accuracy of the forest fire prediction model.

IV. DATASET PREPARATION

After collecting data from various sources, I proceeded with combining the datasets to create a comprehensive dataset consisting of 1,500 data points. Each data point included 20 distinct features, such as weather conditions, vegetation indices (NDVI, LAI, FPAR), and geographical information (elevation, latitude, longitude).

In this study, I used the severity of wildfires as a continuous target variable, with non-fire incidents assigned a severity score of 0. This approach enables the model to capture a wide range of wildfire severity levels, from the absence of fire to varying degrees of intensity. By treating the severity as a continuous variable, the model can predict not only whether a fire will occur but also the expected intensity of the fire, which provides a more nuanced understanding of wildfire risks. Assigning a severity score of 0 to non-fire incidents allows the model to distinguish between fire and non-fire events, while still accommodating a continuous spectrum of severity in predicting future wildfires. This method enhances the model's predictive power by focusing on the likelihood of varying fire intensities, which is crucial for developing early warning systems and mitigation strategies. Furthermore, by incorporating environmental and meteorological features as predictors, the model can generalize effectively to new locations and times, thereby improving its ability to forecast future wildfires and their potential severity based on observed conditions.

To ensure the integrity and quality of the data, I removed any duplicate entries that may have occurred during the data merging process. Following this, the dataset was organized, with the features arranged in a consistent manner across all records. To minimize bias and ensure randomness in the analysis, I scrambled the data entries, providing a well-balanced and representative sample for training and evaluation. Since outliers constituted only a small percentage (26 out of 1,500), they were removed. Additionally, feature outliers (precipitation and severity), which exhibited larger values of skewness and kurtosis, were transformed using logarithmic values to base 10. Consequently, the models predict the log-transformed severity as the target variable. After these preprocessing steps, I split the dataset into training and testing sets in an 80-20 ratio. This prepared dataset served as the foundation for subsequent model development and analysis.

V. DATA ANALYSIS

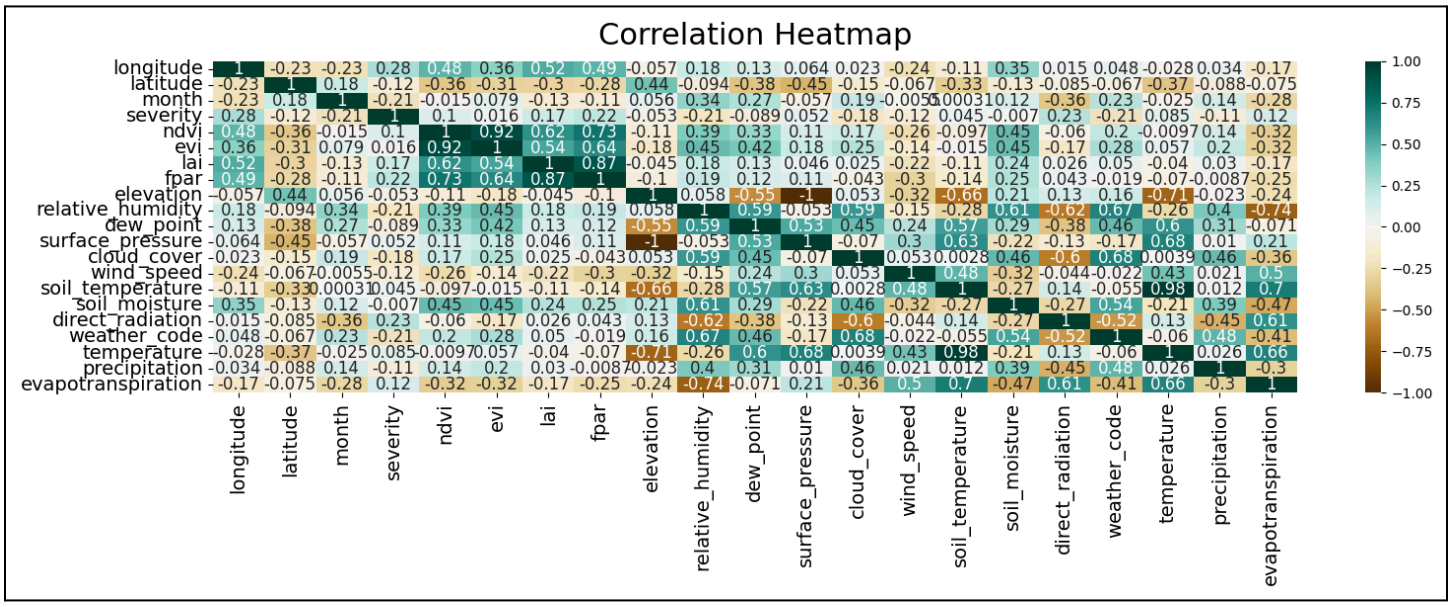


Figure 7: Heatmap to show correlation between all data points collected

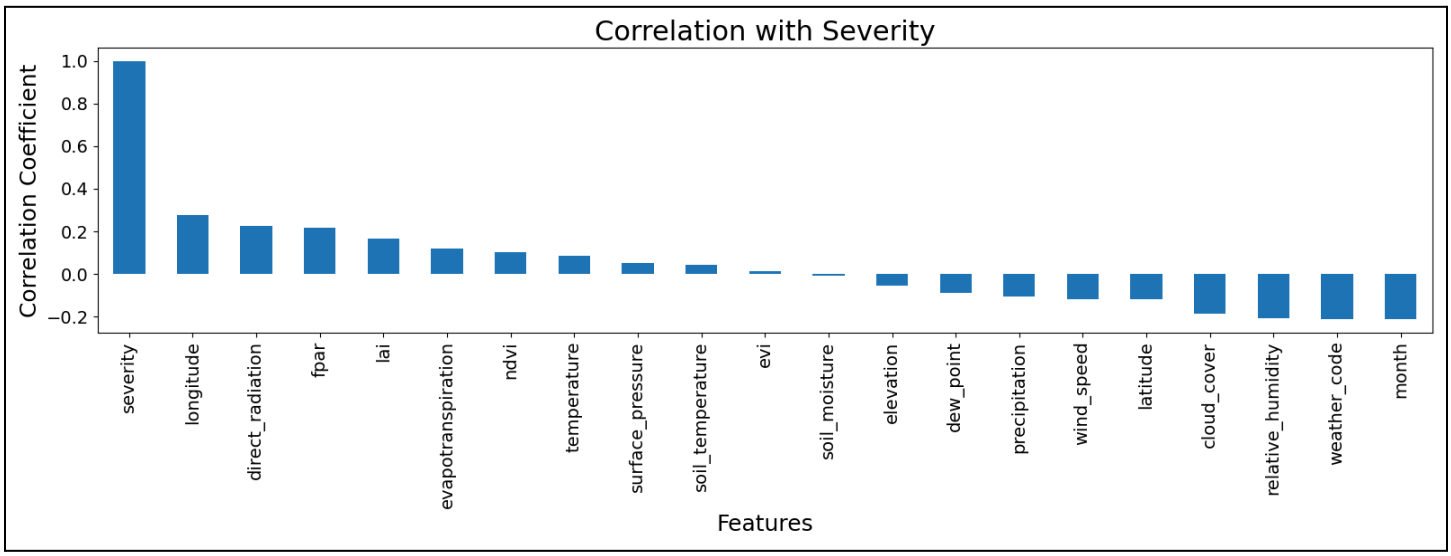


Figure 8: Bar graph to show correlation of all features collected with severity of the Forest Fires

Based on the analysis, it was observed that the majority of forest fires occurred in the eastern region of India. Several environmental and meteorological factors exhibit varying degrees of correlation with the severity of forest fires.

Strong Positive Correlations:

1. **Direct Radiation**: A significant positive correlation with fire severity suggests that higher levels of direct solar radiation contribute to increased fire intensity. This is likely due to the higher heat exposure promoting ignition and fueling the spread of fires.
2. **FPAR (Fraction of Photosynthetically Active Radiation)**: FPAR is closely linked to vegetation health and productivity. A higher FPAR indicates greater vegetation growth, which can provide more fuel for fires. Thus, areas with higher FPAR tend to experience more severe fires.
3. **LAI (Leaf Area Index)**: LAI, which quantifies the leaf cover of vegetation, also shows a strong positive correlation with fire severity. A larger leaf area indicates denser vegetation, which can support larger fires due to increased biomass.
4. **Evapotranspiration**: This is the sum of evaporation and plant transpiration, which is influenced by temperature and moisture availability. High evapotranspiration rates are associated with drier conditions, making vegetation more flammable and contributing to more severe fires.
5. **NDVI (Normalized Difference Vegetation Index)**: NDVI, which measures vegetation greenness, is another indicator of vegetation health and fuel availability. Higher NDVI values correlate with more intense fires, as lush vegetation offers abundant fuel.
6. **Temperature**: Temperature has a strong positive relationship with fire severity, where higher temperatures increase the likelihood and intensity of fires. Elevated temperatures dry out vegetation, making it more susceptible to combustion.

Weak Positive Correlations:

1. **Surface Pressure**: Surface pressure exhibits a weak positive correlation with fire severity, suggesting that slight variations in atmospheric pressure might have a minimal impact on fire intensity, though this relationship is not as pronounced as with other factors.
2. **Soil Temperature**: Similar to surface pressure, soil temperature shows a weak positive correlation. While warmer soils can dry out vegetation, making it more prone to ignition, this relationship is less direct compared to other environmental variables like temperature and evapotranspiration.
3. **EVI (Enhanced Vegetation Index)**: EVI, which is another vegetation index, has a weak positive correlation with fire severity, indicating that areas with more vigorous vegetation growth may experience more intense fires, although this correlation is weaker than other vegetation indices like NDVI. However, healthy forests are less likely to have Forest Fires.

Weak Negative Correlations:

1. **Soil Moisture**: Soil moisture shows a strong negative correlation with fire severity, meaning that as soil moisture increases, fire severity decreases. Higher moisture content in the soil helps to maintain vegetation hydration, making it less susceptible to fire.
2. **Elevation**: Similarly, elevation shows a strong negative correlation with fire severity. Fires tend to occur less frequently or with less intensity at higher altitudes, possibly due to cooler temperatures and higher humidity levels in these regions, which are less conducive to fire outbreaks.
3. **Dew Point:** Dew point shows a weak negative correlation with forest fire severity, indicating that higher dew points, which correspond to increased moisture in the air, reduce the likelihood and intensity of fires. As the dew point rises, the increased atmospheric moisture makes vegetation less susceptible to ignition.

Strong Negative Correlations:

1. **Precipitation**: Precipitation is strongly negatively correlated with fire severity. Higher rainfall reduces the availability of dry fuel and lowers fire risk by increasing moisture in vegetation and soil.
2. **Wind Speed**: Wind speed shows a strong negative correlation with fire severity, likely because high winds may hinder the fire's ability to move and intensify.
4. **Latitude**: Latitude is negatively correlated with fire severity, indicating that areas at higher latitudes (away from the equator) tend to experience less severe forest fires, likely due to cooler, wetter climates compared to lower latitudes.
5. **Cloud Cover**: Cloud cover shows a negative correlation with fire severity, as cloudy conditions can lower temperatures and increase humidity, reducing the likelihood of fire outbreaks.
6. **Relative Humidity**: Higher relative humidity is strongly negatively correlated with fire severity. Increased humidity levels in the atmosphere reduce the flammability of vegetation, making fires less likely to start and spread.
7. **Weather Code**: Weather conditions coded to represent specific weather types show a negative correlation with fire severity, with certain weather conditions, such as rain or cloudy weather, being less conducive to forest fires.

Temporal Analysis:

Forest fires tend to occur more frequently during the **initial months** of the year, suggesting that environmental conditions early in the year, such as lower rainfall and higher temperatures, create favorable conditions for fire outbreaks.

In summary, environmental factors such as direct radiation, temperature, and vegetation indices (FPAR, LAI, NDVI) show strong positive correlations with the severity of forest fires. In contrast, factors like soil moisture, precipitation, and relative humidity exhibit strong negative correlations, helping to mitigate fire severity. Additionally, geographical and temporal factors, such as elevation and the time of year, further influence fire occurrence and intensity.

VI. MACHINE LEARNING MODELS

1. Linear Regressor

In this study, first a Linear Regression model was implemented to predict forest fire severity using a dataset comprising around 1200 data points and 20 features. The model was trained on the training set to establish a linear relationship between the independent variables (predictors) and the dependent variable (severity). This approach provides a baseline for comparing the performance of more complex models.

After training, the model was evaluated on the test set by predicting the severity values. The Root Mean Squared Error (RMSE) and Mean Squared Error (MSE) were calculated as performance metrics to quantify the model's accuracy. The model gave me an RMSE score of 1.18 and MSE of 1.4. These metrics highlight the predictive capabilities of the Linear Regression model and serve as a reference for evaluating more sophisticated methods like Random Forests. To find the accuracy of the model, the results were sorted into two categories: values above 1 were classified as fires and those below were classified as non-Fires. Under these conditions the model gave me an accuracy of 77% in predicting Forest Fires in the test dataset.

2. Random Forest Regressor

In this study, a Random Forest Regressor was employed to predict forest fire severity using a dataset. To enhance the model's performance, GridSearchCV was applied for hyperparameter tuning, systematically testing various combinations of parameters such as tree depth, leaf nodes, and the minimum samples required for splitting and leaf nodes. This approach helps in identifying the most optimal set of hyperparameters for the model, improving its ability to generalize to unseen data. The model's training process involved cross-validation, where the data was split into five subsets to assess performance more robustly. By evaluating 135 combinations of the chosen hyperparameters, the best configuration was selected, and the final model was retrained on the entire dataset. This methodology ensures that the Random Forest model is finely tuned, reducing overfitting and providing more accurate predictions of forest fire severity.

Just like before, the results were sorted into two categories: values above 1 were classified as fires and those below were classified as non-Fires. Under these conditions the model gave me an accuracy of 93% in predicting Forest Fires in the test dataset. The model gave me an RMSE score of 0.711.

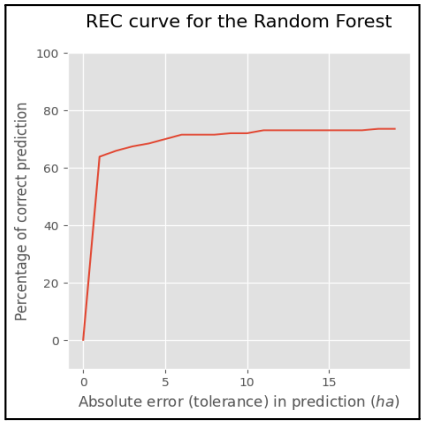


Figure 9: REC Curve for Random Forest

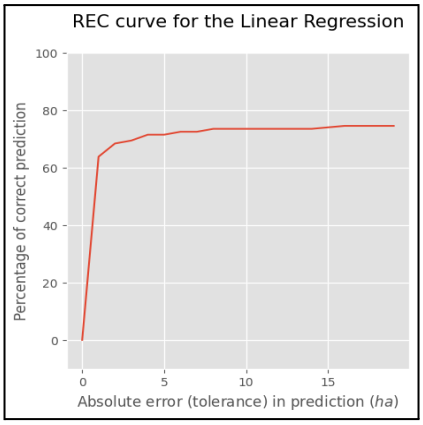


Figure 10: REC Curve for Linear Regression

3. Neural Network

In this study, a Deep Neural Network (DNN) was implemented using the Keras library to predict forest fire severity. The network architecture consists of multiple dense layers, with a total of 9 fully connected (Dense) layers and 2 dropout layers to mitigate overfitting. The input data, comprising 1200 data points with 20 features, was processed through these layers to learn complex non-linear relationships between the predictors and the target variable.

The first layer has 100 neurons with a ReLU (Rectified Linear Unit) activation function, which introduces non-linearity and helps the model capture intricate patterns. Subsequent layers expand the network's capacity with 150 neutrons in four layers, followed by two additional layers with 100 neurons each. Dropout layers are included after every few dense layers, with a rate of 0.3, randomly deactivating 30% of neurons during training to prevent overfitting and enhance generalization. The final layer has a single neuron with a linear activation function, providing the predicted continuous severity value. This architecture allows the network to process complex feature interactions, improving prediction accuracy for forest fire severity.

The model gave me an RMSE of 1.01 and had an accuracy of 88% found using the same method as above.

4. XGBoost Regressor

An XGBoost Regressor was utilized to predict forest fire severity, leveraging its ability to model complex, non-linear relationships effectively. The model was initially configured with 100 estimators, a learning rate of 0.1, and the gradient tree boosting algorithm, with additional hyperparameter tuning performed using GridSearchCV. Key parameters such as min_child_weight, gamma, subsample, colsample_bytree, and max_depth were systematically evaluated across a range of values using 5-fold cross-validation to ensure robust model performance. A total of 180 parameter combinations were tested to identify the optimal configuration, after which the final model was retrained on the entire training dataset. This approach ensured a finely tuned model capable of providing accurate and generalized predictions of forest fire severity.

The model gave me an RMSE of 0.519 and had an accuracy of 94% found using the same method as above.

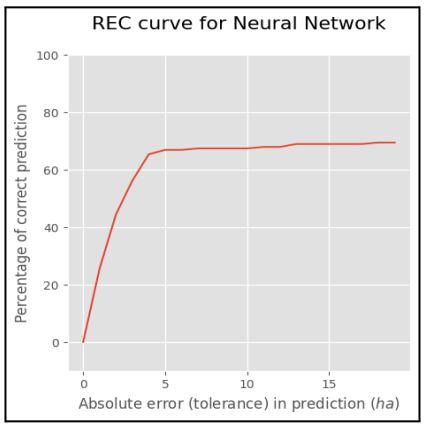


Figure 9: REC Curve for Random Forest

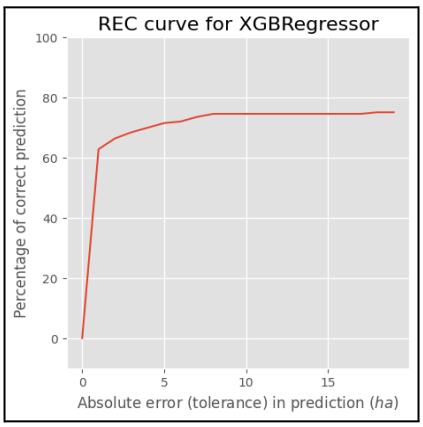


Figure 10: REC Curve for Linear Regression

VIII. THE PYRO PREDICTOR WEBSITE

PyroPredictor is an advanced web application aimed at providing accurate and actionable predictions for forest fire risks across India. By leveraging a custom-trained neural network model, the platform analyzes critical environmental factors such as weather conditions, vegetation indices (e.g., NDVI, LAI), historical fire data, and geographical information (e.g., elevation, latitude, and longitude). The predictions generated by the neural network offer valuable insights for a wide range of users, including researchers, policymakers, and local communities, enabling them to make informed decisions to mitigate wildfire risks and protect natural ecosystems.

A. How the Website Works:

At the core of PyroPredictor's prediction capabilities is a tensorflow neural network model trained on a diverse set of environmental data. This model is designed to understand complex patterns in weather, vegetation, and historical wildfire data, which are key factors in determining the likelihood of forest fires. The model uses these features to output a probability score for forest fire risk in a given location, which is then visualized on the platform for easy interpretation by users. This model was then converted to TensorflowJS so it could be setup on a website.

B. PyroPredictor Architectural Design:

The PyroPredictor system is designed to provide real-time forest fire risk predictions by integrating data from multiple external sources, including Google Earth Engine (GEE) and Open Meteo. The architecture consists of both a client-side website and a server-side backend that work together to process and present fire probability information to users efficiently.

1. Server-Side Infrastructure

The backend server is responsible for handling API requests from the client-side website and facilitating the retrieval of environmental data. The server connects to GEE and Open Meteo to collect real-time data on vegetation, elevation, and weather conditions. Once the data is fetched, the server processes it and returns the relevant information to the client in response to API calls.

To optimize data retrieval, the server utilizes Firebase as its backend, which stores the current reports on the conditions of all districts. Instead of querying GEE and Open Meteo for each request, the server directly retrieves the stored data from Firebase, ensuring faster response times and reducing the load on external services. A scheduled function hosted on Netlify is configured to refresh the Firebase database nightly, ensuring that the data provided to the client is always up-to-date.

2. Client-Side Operation

Upon initialization, the website requests the latest environmental data for all districts from the server. This data is then used to compute the fire probability for each district based on predefined parameters, including weather conditions, vegetation health, and elevation. Users can adjust the weights assigned to these parameters to influence the probability calculation. The resulting fire probabilities are visually represented on the website through a color-coded map, where each district is colored according to its risk level.

To improve performance and minimize redundant API calls, the website implements a local caching mechanism. When the website is first loaded, it stores the retrieved district data in the browser's local storage. This allows the client to quickly retrieve and display the data on subsequent visits, provided the data was retrieved on the same day as the current session. This local caching ensures that the website operates efficiently, even if the user revisits the platform within a short timeframe.

Additionally, the website allows users to input their specific locations. When a user submits a location, the client queries the server for the relevant environmental parameters of that region and subsequently calculates the fire probability for the specified area. This localized query ensures that users receive accurate, up-to-date fire risk assessments for their desired locations.

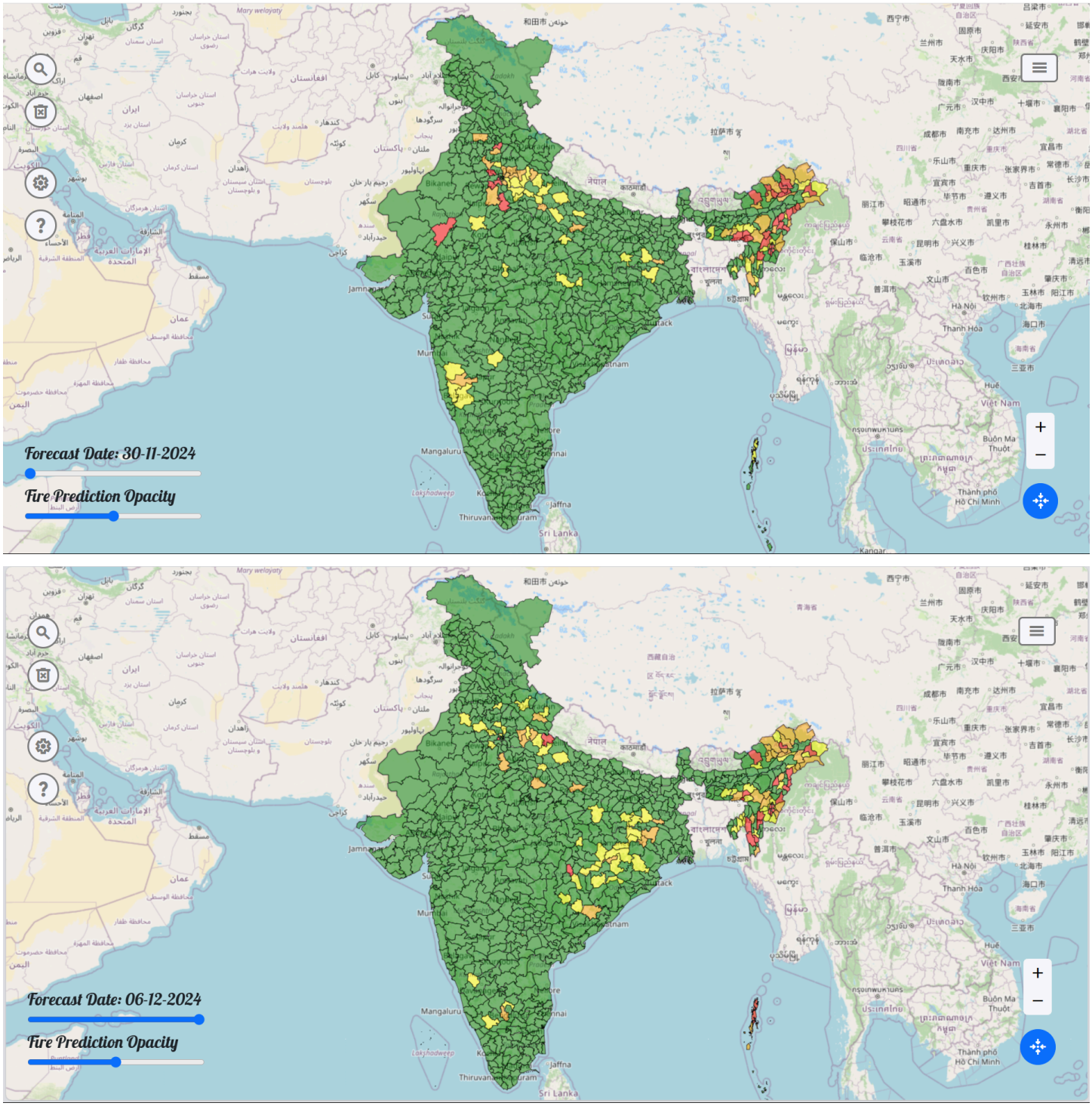
C. User-Friendliness and Interactive Features:

PyroPredictor is designed with the end-user in mind, ensuring that even complex, data-driven predictions are accessible and easy to interpret. The platform incorporates several user-friendly features that make it intuitive and engaging for users across different backgrounds:

1. **Interactive Map Visualization:** One of the key features of PyroPredictor is the dynamic, interactive map that allows users to explore forest fire risk levels across India. For mapping and geospatial visualization, Leaflet allows users to interact with the map in real-time. It enables smooth zooming, panning, and information display for district-specific risk levels. The map is color-coded, with districts represented in different colors based on their predicted risk levels. This visual representation allows users to quickly identify high-risk areas and focus their attention on regions that need immediate action.
2. **Customizable Predictions:** The platform allows users to adjust the parameters used in the predictions, such as selecting specific weather conditions or historical wildfire data, to tailor the results to different scenarios. For instance, users can simulate future fire risks under different climate change scenarios or view predictions for specific dates based on past weather trends. This level of customization helps users make predictions that are relevant to their unique needs, whether they are conducting research, planning policy interventions, or assessing community safety.
3. **Data-Driven Insights and Trends:** In addition to the map, users can view trends over time, such as how risk levels fluctuate seasonally or in response to changes in environmental factors. This feature enables policymakers and researchers to identify patterns and predict potential fire outbreaks based on historical data and ongoing trends.
4. **Responsive Design:** PyroPredictor is built with Bootstrap, a responsive design framework, ensuring the platform is accessible across different devices, including desktops, tablets, and smartphones. Whether users are at a research station, in a government office, or in the field, they can access the application from anywhere with an internet connection, making it highly versatile and convenient.

D. Empowering Users with Actionable Information:

The primary objective of PyroPredictor is to empower its users to make informed decisions. Whether it's researchers identifying areas that need further study, policymakers planning wildfire prevention strategies, or local communities preparing for fire season, the platform provides the tools to make proactive, data-driven decisions. By offering accurate risk assessments based on the latest environmental data and machine learning predictions, PyroPredictor helps mitigate wildfire risks and protect both human and ecological health.



Figures 11 and 12: Predictions of Forest Fires made by Pyro Predictor for 30th Nov and 6th Dec 2024

CONCLUSION

This paper outlines the successful development and implementation of a comprehensive system for forest fire risk prediction, combining cutting-edge machine learning models with real-time environmental data from trusted sources. The approach ensures that predictions are both accurate and actionable, enabling stakeholders to make informed decisions in wildfire prevention and management.

The PyroPredictor platform exemplifies the practical application of these technologies. Designed with user accessibility in mind, it offers an engaging and intuitive interface where users can explore fire probabilities dynamically. Features such as color-coded district maps, location-specific predictions, and parameter customization empower users to analyze and respond to wildfire risks effectively. The integration of caching mechanisms and efficient server-client communication ensures seamless performance, providing users with up-to-date and reliable information with minimal latency. By leveraging state-of-the-art data collection methods, robust machine learning models, and a thoughtfully designed interface, PyroPredictor bridges the gap between advanced technology and real-world application. It serves as a valuable tool for a diverse audience, including researchers, policymakers, and local communities, helping them take proactive measures to mitigate wildfire risks and protect natural habitats.

PyroPredictor stands at the intersection of advanced machine learning and accessible user interfaces, offering a powerful tool for understanding and predicting forest fire risks. This project demonstrates how innovative solutions can address pressing environmental challenges, showcasing the potential of technology in promoting sustainability and resilience. Looking ahead, PyroPredictor establishes a strong foundation for future developments, such as trend analysis, integration with disaster response systems, and expanded geographic coverage. These enhancements will further its mission to support communities and foster data-driven decision-making in the face of climate challenges.

REFERENCES

https://www.gdacs.org/Alerts/default.aspx
https://developers.google.com/earth-engine/datasets/catalog/MODIS_061_MCD15A3H
https://developers.google.com/earth-engine/datasets/catalog/MODIS_061_MOD13A1
https://developers.google.com/earth-engine/datasets/catalog/FIRMS
https://open-meteo.com/
https://www.visualcrossing.com/
https://www.sciencedirect.com/science/article/pii/S2666719324000244
https://www.sciencedirect.com/science/article/pii/S0379711218303941
https://www.mdpi.com/1999-4907/13/7/1050
https://cdnsciencepub.com/doi/full/10.1139/er-2020-0019
https://www.kaggle.com/datasets/elikplim/forest-fires-data-set
https://fsi.nic.in/forest-fire-activities
https://www.globalforestwatch.org/dashboards/country/IND?category=fires