Forest Fire Risk Prediction System

Course Information:

Course Title: Geospatial Data Analytics
Course Code: CS321
Instructor: Dr.Dubacharla Gyaneshwar

Submitted by:

K.Akshay Kumar (CS22B1028)
and
S.Chidvilasini (CS22B1056)

Institution/Department:

Indian Institute of Information Technology,Raichur

Abstract

Forest fires are one of the most significant natural disasters, causing severe environmental destruction and economic losses. Early prediction of such disasters is crucial to saving lives, resources, and ecosystems. This project aims to predict forest fires using machine learning algorithms and visualize spatial data for better decision-making.

The dataset used in this project includes environmental factors like FFMC, DMC, DC, ISI, Temperature, RH, Wind, and Rain, along with spatial coordinates and fire occurrences. Several machine learning models, including Logistic Regression, Support Vector Machines (SVM), and Random Forest, were tested, with SVM achieving the highest accuracy of 62%.

Geospatial processing techniques were applied to convert spatial coordinates into Latitude and Longitude, enabling advanced visualization. The Kepler tool was employed to map fire-prone areas, providing a clear representation of high-risk zones. This project offers valuable insights for mitigating fire risks, enhancing environmental monitoring, and supporting disaster management efforts.

Table of Contents

Introduction

1.1 Problem Statement

Forest fires are highly destructive and unpredictable, causing significant environmental, social, and economic damage. Identifying fire-prone areas and predicting fire occurrences in advance is essential for mitigating risks, safeguarding ecosystems, and allocating resources effectively.

1.2 Objectives

- Prediction: Use machine learning models to accurately predict forest fires based on environmental conditions.
- Visualization: Leverage geospatial tools to visualize fire-prone areas and environmental changes over time.
- Insights: Provide actionable insights for stakeholders to make informed decisions on forest management and fire prevention.

Dataset Description

2.1 Source of Data

The dataset used for this project was sourced from the UCI Machine Learning Repository, a renowned source for datasets used in academic research and machine learning tasks. The dataset includes environmental and spatial data related to forest fire occurrences.

2.2 Key Features in the Dataset

The initial dataset contained the following attributes:

- Environmental Factors:
    - FFMC (Fine Fuel Moisture Code): Indicates the dryness of surface fuels.
    - DMC (Duff Moisture Code): Represents the moisture level of the organic layer in the forest.
    - DC (Drought Code): Indicates the long-term dryness in the forest.
    - ISI (Initial Spread Index): Measures the potential fire spread.
    - Temperature: Ambient temperature at the location.
    - Relative Humidity (RH): Percentage of atmospheric moisture at the location.
    - Wind Speed: The intensity of wind at the location.
    - Rainfall: Precipitation levels before fire occurrence.
    - Month and Day: These columns were removed during preprocessing as they did not contribute to the final predictions.
    - Area: Represents the area burned by the fire (also removed after preprocessing).
- Spatial Data:
    - X and Y Coordinates: These represent the locations of the data points.
- Target Variable:
    - y: Binary indicator of fire occurrence (1: Fire, 0: No Fire).

2.3 Data Preprocessing and Enhancements

The raw dataset underwent preprocessing to better fit the model and facilitate more effective predictions:

- The Month, Day, and Area columns were removed as they were deemed irrelevant to the prediction task.
- The X and Y spatial coordinates were retained for geospatial analysis, though no conversion to Latitude and Longitude was done in this case.
- Each data point was assigned a Point ID to track individual records.

After preprocessing, the dataset was organized as follows:

X	Y	FFMC	DMC	DC	ISI	Temp	RH	Wind	Rain	y
7	5	86.2	26.2	94.3	5.1	8.2	51	6.7	0.0	0

2.4 Example Data Point (Post-Processed)

- The X and Y columns represent the original spatial coordinates of the data point.
- FFMC, DMC, DC, ISI, Temp, RH, Wind, Rain correspond to the environmental features that were used for fire prediction.
- y is the target variable indicating whether a fire occurred (1) or not (0).

Methodology

3.1 Overview of Methodology

This project follows a machine learning-based approach to predict forest fire occurrences using environmental data and spatial coordinates. The methodology consists of the following major steps:

Data Collection and Preprocessing
Feature Selection
Model Training and Evaluation
Prediction and Geospatial Visualization

3.2 Data Preprocessing

To prepare the dataset for modeling, several preprocessing steps were carried out:

Handling Missing Values:
Any missing data points were handled by either removing or imputing missing values based on the context of the dataset.

Normalization and Scaling:
The data was normalized to ensure all features contributed equally to the model's predictions. Feature scaling was performed on numerical values like temperature, humidity, and wind speed to ensure consistency.

Feature Engineering:
We derived new features like Point ID for each data point, and spatial coordinates (X, Y) were preserved for geospatial analysis. Other columns like Month, Day, and Area were excluded as they did not significantly impact the fire prediction.

3.3 Model Selection and Training

For predicting forest fire occurrences, multiple machine learning algorithms were tested:

Logistic Regression:
A basic binary classification model used for predicting fire (1) or no fire (0).

Support Vector Machine (SVM):
An effective classifier for separating classes, especially when there is a large margin between them.

Random Forest Classifier:
An ensemble learning method that combines multiple decision trees to improve prediction accuracy and avoid overfitting.

XGBoost (Extreme Gradient Boosting):
A powerful, optimized implementation of gradient boosting that handles both classification and regression problems efficiently.

3.4 Model Evaluation

Each model's performance was evaluated based on accuracy and other metrics like precision, recall, and F1-score.

3.5 Geospatial Visualization

The final predictions were mapped to geographic locations using Kepler to provide stakeholders with interactive visual insights.

Fire-prone areas were marked using red dots, and no fire regions were marked with yellow dots.
The geospatial coordinates were visualized for a better understanding of where fire occurrences are more likely to happen.

3.6 Process Flow Overview

The process followed in the project can be outlined as:

Data Collection → Preprocessing → Model Training → Evaluation → Prediction → Geospatial Visualization.

Results and Discussion

4.1 Model Performance

The performance of the implemented machine learning models was evaluated using accuracy as the primary metric. The results indicate that different models varied in their ability to predict forest fire occurrences:

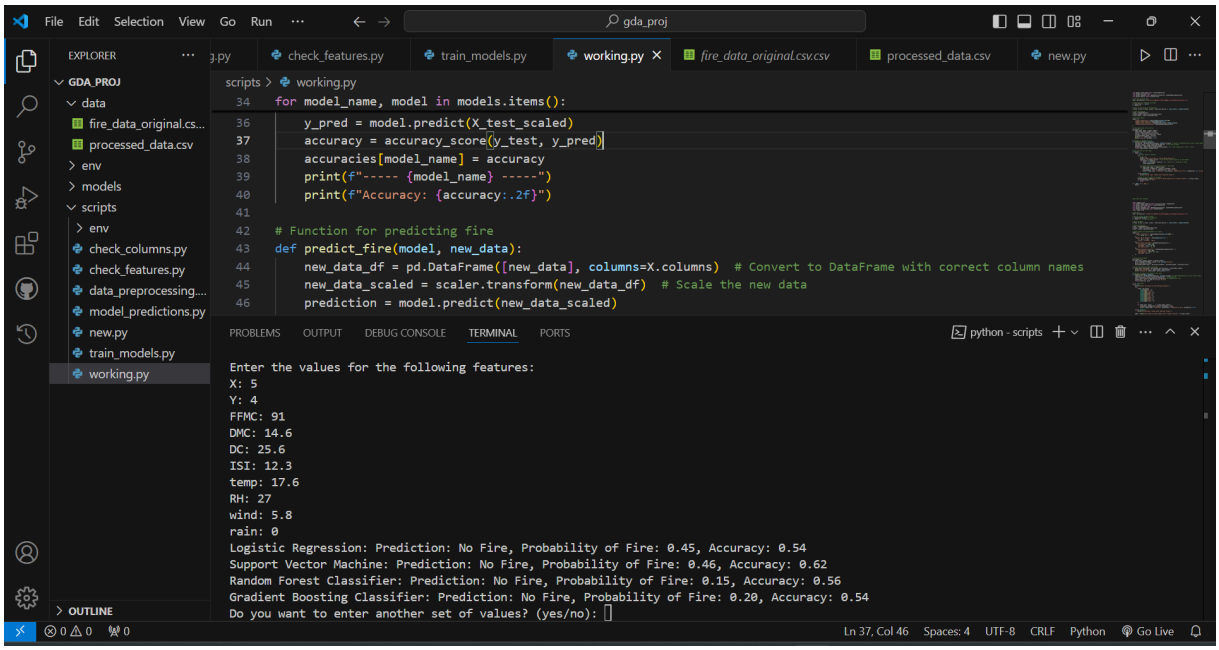
- **Random Forest Classifier:**
  - Achieved an accuracy of 56%, showcasing moderate performance.
  - The ensemble approach provided stable results but could not outperform other models significantly.
- **Gradient Boosting Classifier (XGBoost):**
  - Achieved an accuracy of 54%.
  - While XGBoost is generally robust, the limited accuracy reflects the challenges posed by the dataset.

- **Support Vector Machine (SVM):**
  - Provided an accuracy of 62%, performing better than other models.
  - Showcased its strength in binary classification, though it faced limitations with imbalanced data.
- **Logistic Regression:**
  - Served as the baseline model with an accuracy of 54%.
  - Highlighted the limitations of linear models in capturing complex patterns within the data.

4.2 Test Case Predictions

Test Case 1 - Prediction: No Fire

- **Input Features:**
  - X: 5, Y: 4, FFMC: 91, DMC: 14.6, DC: 25.6, ISI: 12.3, temp: 17.6, RH: 27, wind: 5.8, rain: 0

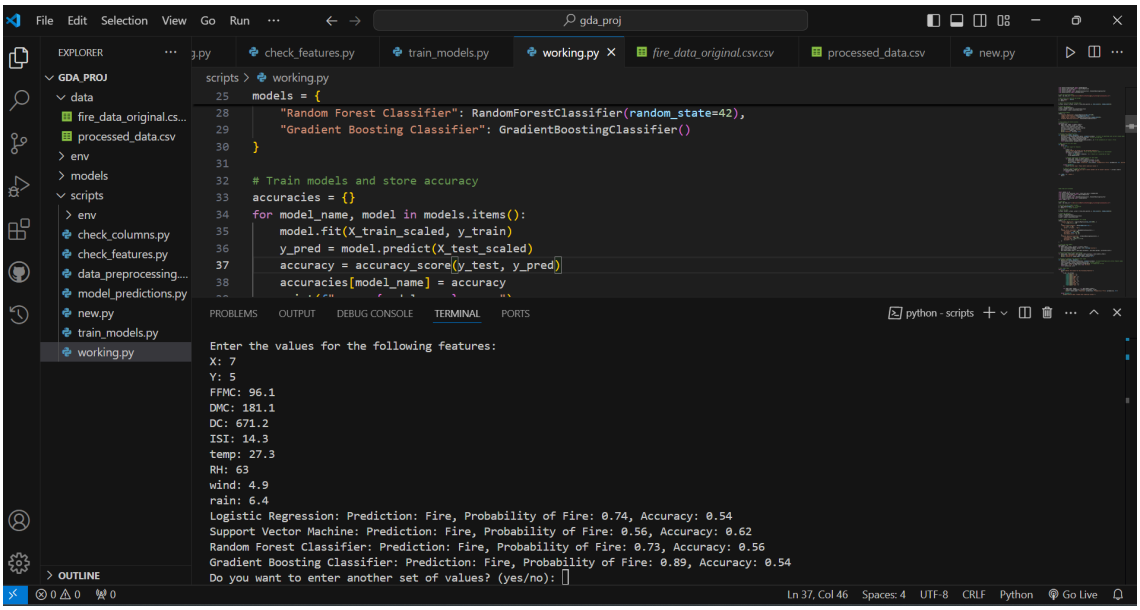


- **Model Predictions and Accuracies:**

- Logistic Regression: Prediction: No Fire, Probability of Fire: 0.45, Accuracy: 0.54
- Support Vector Machine: Prediction: No Fire, Probability of Fire: 0.46, Accuracy: 0.62
- Random Forest Classifier: Prediction: No Fire, Probability of Fire: 0.15, Accuracy: 0.56
- Gradient Boosting Classifier: Prediction: No Fire, Probability of Fire: 0.20, Accuracy: 0.54

Test Case 2 - Prediction: Fire

- **Input Features:**
  - X: 7, Y: 5, FFMC: 96.1, DMC: 181.1, DC: 671.2, ISI: 14.3, temp: 27.3, RH: 63, wind: 4.9, rain: 6.4



Model Predictions:

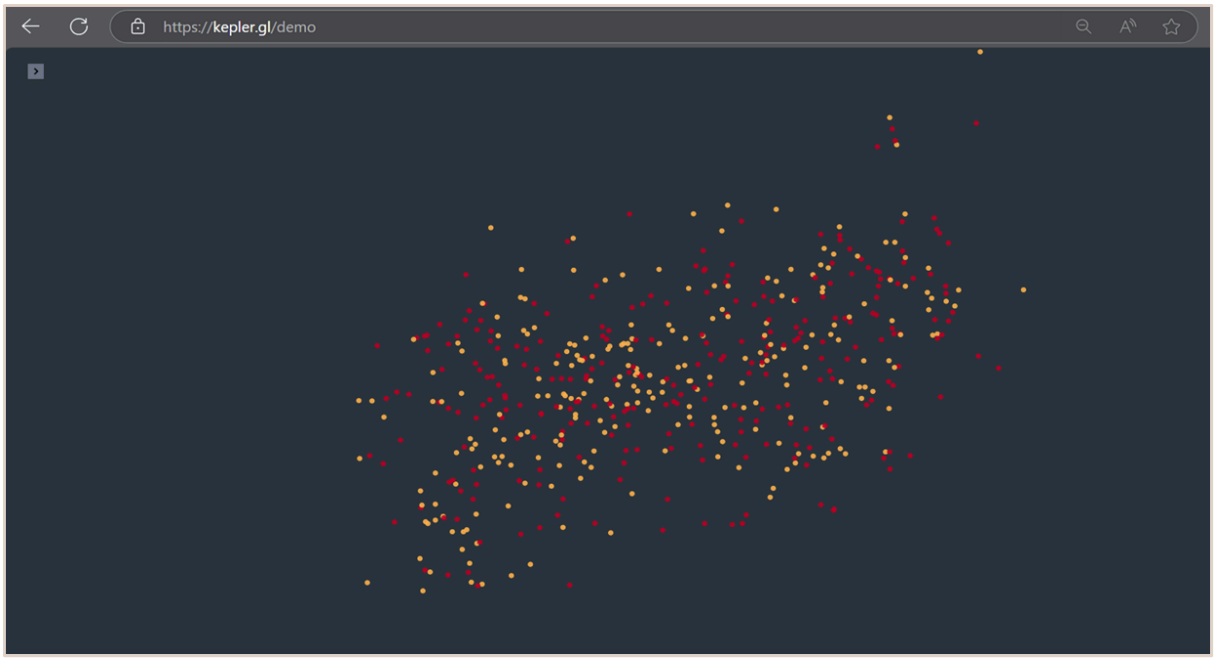
- Logistic Regression: Prediction: Fire, Probability of Fire: 0.74, Accuracy: 0.54
- Support Vector Machine: Prediction: Fire, Probability of Fire: 0.56, Accuracy: 0.62
- Random Forest Classifier: Prediction: Fire, Probability of Fire: 0.73, Accuracy: 0.56
- Gradient Boosting Classifier: Prediction: Fire, Probability of Fire: 0.89, Accuracy: 0.54

4.3 Geospatial Insights and Visualization

After converting spatial coordinates (X, Y) to latitude and longitude, the fire-prone areas were mapped using **Kepler** for enhanced interpretability:

- **Visualization Insights:**
  - **Red Points:** Represented areas with high fire probability, assisting in identifying zones requiring urgent attention.
  - **Yellow Points:** Represented areas with low fire probability, indicating safer regions.

This visualization provided actionable insights for decision-makers to prioritize monitoring and resource allocation.



4.4 Discussion

The results demonstrate that Support Vector Machine (SVM) outperformed other models with a modest accuracy of 62%, highlighting its effectiveness for binary classification in the given dataset. Logistic Regression and Gradient Boosting served as foundational models, while Random Forest provided stable, interpretable results.

The integration of geospatial data and machine learning enhanced the project's applicability. By using Kepler for visualization, stakeholders were empowered to interpret the results intuitively and identify critical fire-prone zones for proactive measures.

Conclusion and Future Scope

5.1 Conclusion

This project successfully demonstrates the application of machine learning models and geospatial data for predicting forest fire occurrences. The integration of environmental factors such as temperature, humidity, and wind with spatial coordinates enabled the identification of high-risk zones. The results showed:

- Support Vector Machine (SVM) as the most accurate model for this dataset, with an accuracy of 62%.
- Logistic Regression and Gradient Boosting served as baseline models but highlighted dataset limitations.
- Visualization using Kepler provided intuitive, actionable insights for fire risk management.

This work emphasizes the importance of combining data-driven models with geospatial tools to address real-world challenges like forest fires.

5.2 Future Scope

The project lays the groundwork for further advancements:

1. **Enhanced Dataset Quality:**
    - Expand the dataset to include more years and additional features like vegetation type and soil moisture.
2. **Improved Model Performance:**
    - Experiment with deep learning models (e.g., Neural Networks) to capture complex relationships.
    - Address data imbalance using oversampling techniques like SMOTE.
3. **Dynamic Visualization Tools:**
    - Integrate dynamic, real-time fire prediction visualizations for better monitoring and quicker response.
4. **IoT Integration:**
    - Employ IoT-based sensors for real-time data collection on temperature, wind, and humidity.
5. **Policy Development:**
    - Collaborate with policymakers to develop strategies for proactive forest fire management based on predictive insights.

References

This section should include all the sources you referred to during the project, including datasets, libraries, and tools used. Proper citation adds credibility and acknowledges the original creators.

6.1 References List

1. **Dataset:**
    - UCI Machine Learning Repository: Forest Fires Dataset.
2. **Libraries and Tools:**
    - Pandas, NumPy, Scikit-learn, Matplotlib, PyProj, Shapely.
    - Kepler.gl for geospatial visualization.

Acknowledgments

We would like to express our heartfelt gratitude to **Dr. Gyaneshwar Sir** for his invaluable guidance, mentorship, and continuous support throughout the course of this project. His expertise and constructive feedback played a pivotal role in shaping the project and ensuring its successful completion.

We would also like to thank our institution for providing the resources and environment necessary to undertake this project, as well as our peers and colleagues for their encouragement and suggestions during challenging times.

Finally, we acknowledge the use of tools and platforms such as Python, scikit-learn, Kepler.gl, and the UCL Machine Learning Repository, which were instrumental in the analysis and implementation of this project.