PROJECT PROGRESS REPORT

A COMPARATIVE STUDY OF FOREST FIRE PREDICTION USING MACHINE LEARNING MODELS

by

Mr. Sachin Malego (st125171) & Mr. Sila N. Mahoot (st125127)

A Project Proposal Submitted in Partial Fulfillment of the Requirements for the Degree of Master of Science in Data Science & AI

Submitted to: Dr. Chantri Polprasert
Assistant Professor
CSIM, Department of ICT

Nationality: Nepalese and Thai

Asian Insitute of Technology
CSIM, Department of Data Science & AI
Thailand
November 2024

CONTENTS

LIST OF FIGURES

Figures **Page**

CHAPTER 1
PROJET OVERVIEW

This project, titled "A Comparative Study of Forest Fire Prediction Using Machine Learning Models," aims to enhance forest fire prediction capabilities by evaluating various machine learning models based on a wide range of influential factors. The primary objective is to assess the performance of different models in predicting forest fires, taking into account not only traditional indices like the Fine Fuel Moisture Code (FFMC) and Duff Moisture Code (DMC) but also a broader set of environmental and meteorological factors. These include temperature, humidity, wind speed, and Sea Surface Temperature (SST) fluctuations, such as those influenced by El Niño.

The study systematically explores the impact of these diverse features on prediction accuracy and precision across various machine learning models, aiming to identify the factors that contribute most significantly to effective fire prediction. Additionally, the project seeks to evaluate each model’s ability to generalize to unseen data, examining its predictive performance across different geographic regions and varying climatic conditions.

A crucial component of the project is assessing the interpretability of each model. The study aims to determine the extent to which stakeholders, including forest management agencies, can understand and effectively utilize each model's predictions to support informed decision-making and foster sustainable forest management practices.

CHAPTER 2
EXECUTIVE SUMMARY

The project is advancing well, with several critical milestones achieved, including data collection, exploratory data analysis (EDA), data preprocessing, and initial model training and comparison. For model training, two primary datasets related to forest fires in Algeria and Portugal were sourced from Kaggle. Additionally, relevant weather data and El Niño information for the same period were web-scraped and processed to align with the Kaggle datasets. These datasets have been merged, and thorough EDA has been conducted, revealing key insights and relationships among various features.

Challenges with data quality from Kaggle and other sources were encountered but have been addressed. During the dataset merging process, class imbalance issues were identified and effectively mitigated using the Synthetic Minority Over-sampling Technique (SMOTE).

The upcoming tasks include further testing and deployment of the machine learning models, along with additional data validation efforts to ensure reliable analysis.

Findings and Analysis

The violin plot in figure 1 provides a comprehensive visual representation of the distribution of various factors potentially influencing forest fires. Temperature and humidity, while displaying relatively normal distributions, might not be the strongest indicators of fire risk individually. In contrast, wind and rain exhibit skewed distributions, suggesting that high wind speeds and low rainfall could significantly increase fire risk. The fuel-related factors, FFMC, DMC, DC, and ISI, show skewed distributions towards higher values,

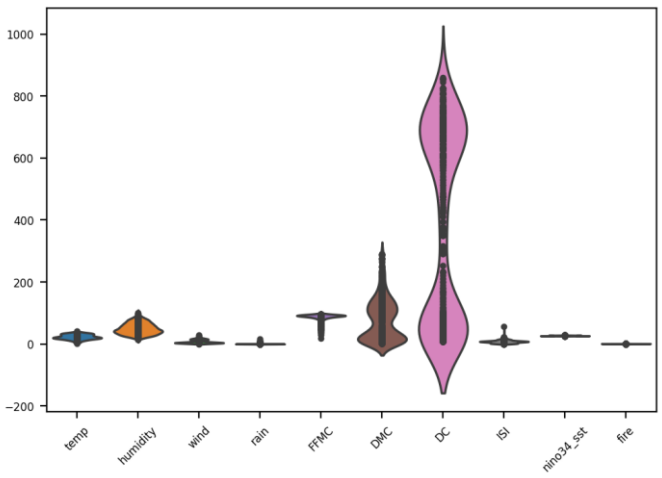

Figure 1: Violin plot of various features related to forest fires

indicating that areas with dry fuels are more prone to fires. The distribution of nino34_sst, related to sea surface temperatures, appears relatively normal, while the fire occurrences are heavily skewed, with most days having no fires and a few experiencing significant fire activity.

The correlation matrix in figure 2 reveals a complex interplay between various factors influencing forest fire occurrence. Fuel-related variables, such as FFMC, DMC, DC, and ISI, exhibit strong positive correlations, suggesting their interconnectedness in contributing to fire risk. Temperature and humidity show a moderate negative correlation, indicating that higher temperatures often coincide with lower humidity, potentially leading to drier conditions and increased fire risk. Similarly, wind and rain exhibit a moderate negative correlation, implying that areas with higher wind speeds tend to experience lower rainfall, creating conditions favorable for fires. Interestingly, fire has strong negative correlations with FFMC, DMC, DC, and ISI, suggesting that these factors might be associated with more mature forests with higher moisture content, which are less prone to fire. While temperature has a weak negative correlation with fire, other factors might play a more significant role in determining fire occurrence. nino34_sst, related to sea surface temperatures, appears to have a relatively weak correlation with other variables, suggesting its impact on fire risk might be less direct. It's important to remember that correlation does not imply causation, and further analysis is needed to establish causal relationships between these variables.

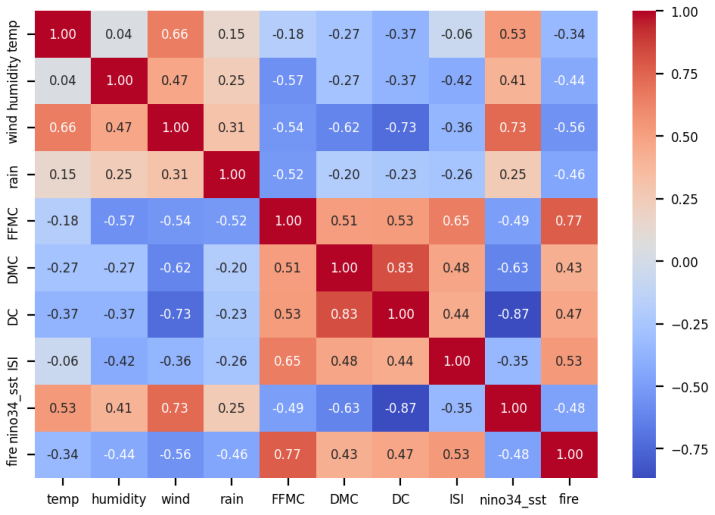

Figure 2: Correlation matrix exploring the factors influencing forest fires

The pair plot in figure 3 provides a comprehensive visual representation of the relationships between different variables in the dataset. Each subplot in the matrix

represents the scatter plot between two variables, while the diagonal plots showcase the distribution of each individual variable. Some variables exhibit linear relationships, meaning that as one variable increase or decreases, the other tends to follow a similar trend. For instance, the relationship between FFMC, DMC, DC, and ISI appear linear. However, other variables display non-linear relationships, where the connection between the two variables is not straightforward, suggesting a more complex relationship that might not be easily captured by linear models. Additionally, some plots reveal the presence of outliers, which are data points that deviate significantly from the general trend. Outliers can influence the analysis and modeling process, making it crucial to identify and handle them appropriately. The diagonal plots offer insights into the distribution of each variable, with some variables appearing normally distributed and others exhibiting skewed distributions or other patterns.

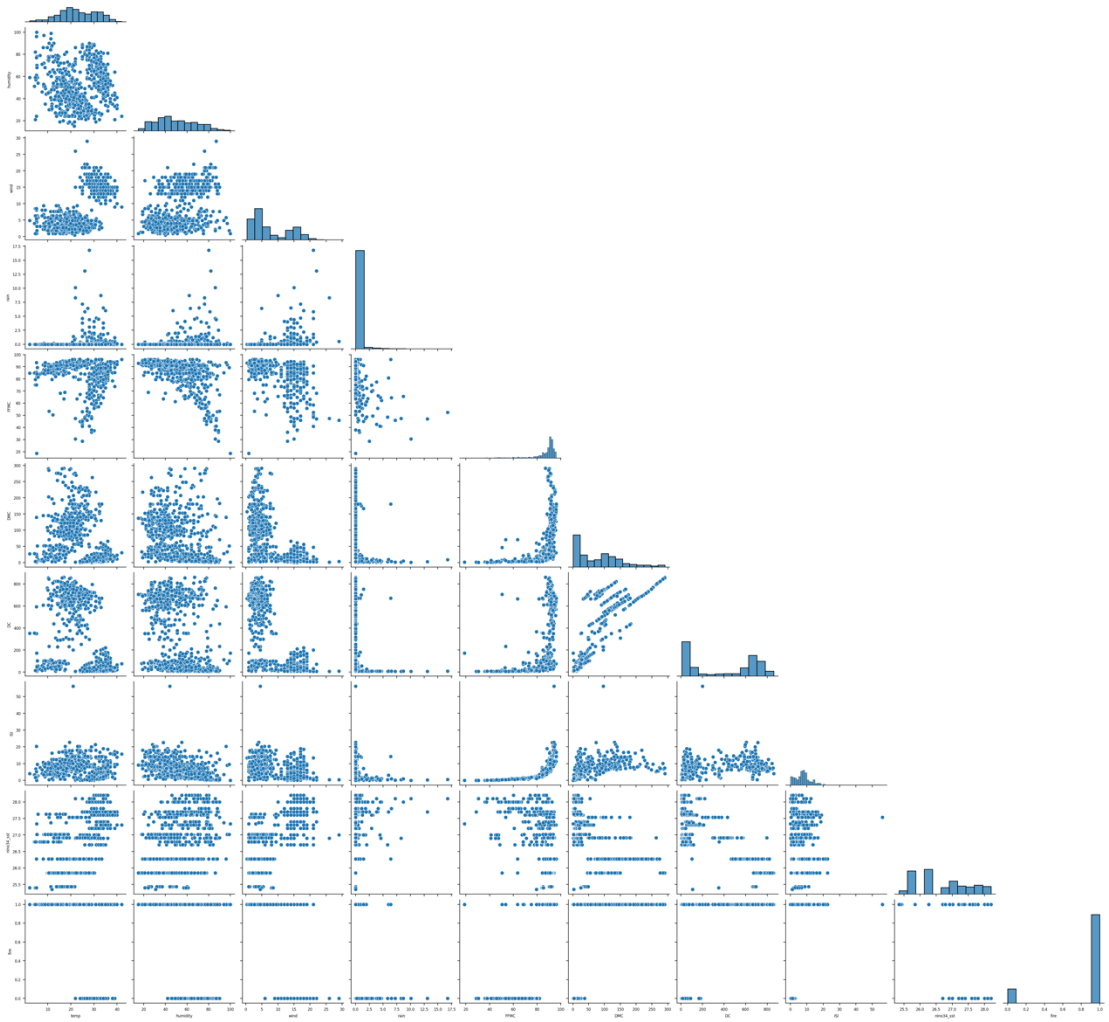

Figure 3: Pair plot between different features

Class Imbalance

The count plot in figure 4 reveals a significant class imbalance in the dataset, with the majority of instances belonging to the class labeled "1" (fire). This imbalance poses a challenge for machine learning models, as they may become biased towards the majority class, leading to poor performance in predicting the minority class (not fire). To address this issue, we have used down-sampling, up-sampling, and SMOTE up-sampling technique (figure 5). By addressing the class imbalance, we are looking forward to improve the model's ability to accurately predict fire occurrences and make more informed decisions about fire prevention and management strategies.

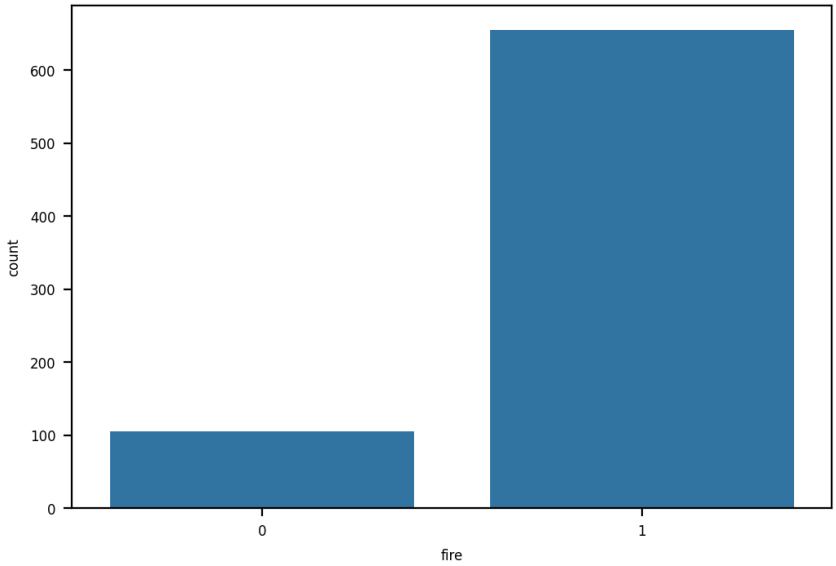

Figure 4: Class imbalance in fire status

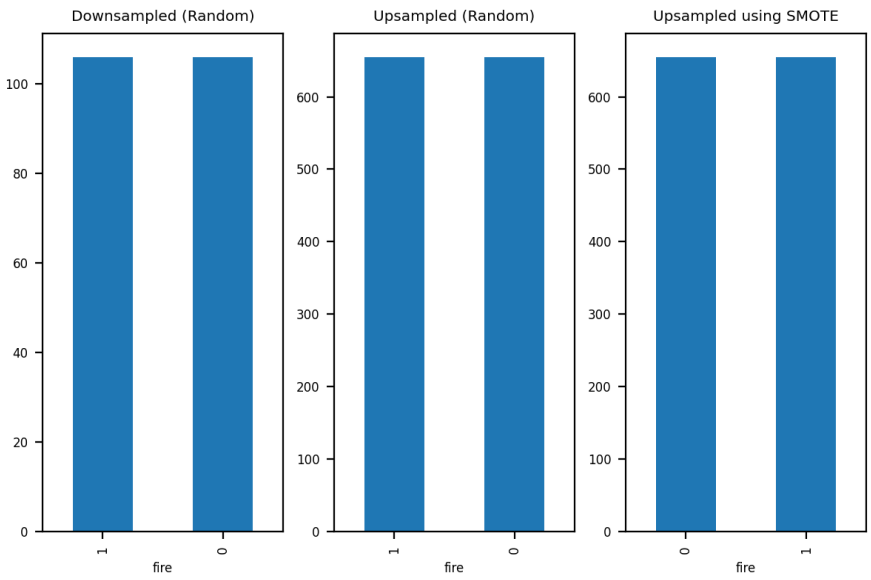

Figure 5: Down-sampled, Up-sampled, and SMOTE up-sampled

Cross-validation

All models demonstrated high accuracy (figure 6), with XGBoost, SVM, and Random Forest consistently achieving near-perfect performance regardless of the data balancing method.

Default hyperparameters

	Accuracy	Precision	Recall	F1
Random Forest	0.995229	0.996154	0.994212	0.995150
XGBoost	0.995224	0.998058	0.992252	0.995141
SVM	0.974204	1.000000	0.947647	0.972771

Best hyperparameters

	Accuracy	Precision	Recall	F1
Random Forest	0.996186	0.998077	0.994212	0.996116
XGBoost	0.997134	0.998077	0.996135	0.997092
SVM	0.994267	0.996190	0.992270	0.994174

Figure 6: Model performance - Random Forest, XGBoost, and SVM

Down-sampling, however, led to slightly lower accuracy than SMOTE or upsampling, due to its reduction in data volume. Precision was high across most models, particularly with XGBoost and Random Forest, and was further improved with SMOTE, which helped models accurately identify fire days with minimal false positives (figure 7). Conversely, simpler models like Decision Tree and KNN showed moderately lower precision, especially when applied to down-sampled data.

In terms of recall, which is critical for identifying actual fire days, SMOTE significantly enhanced recall in models like XGBoost, SVM, and Random Forest, while downsampling tended to reduce recall due to the reduced dataset size. The F1-score, which balances precision and recall, highlighted XGBoost, SVM, and Random Forest as top performers across all sampling methods, with SMOTE consistently providing the best results. Although Logistic Regression and Naive Bayes delivered steady performance, they fell short of the advanced models in terms of precision and recall.

Overall, XGBoost, SVM, and Random Forest stand out as the leading models, especially when paired with SMOTE, which proves to be the most effective sampling

method for improving fire detection accuracy. While simpler models like Decision Tree and KNN performed reasonably, they did not match the precision and recall of the more sophisticated models. Moving forward, SMOTE is recommended for dataset balancing, with XGBoost, SVM, and Random Forest prioritized for further tuning and application, especially with a focus on achieving high recall and F1-scores for reliable fire detection.

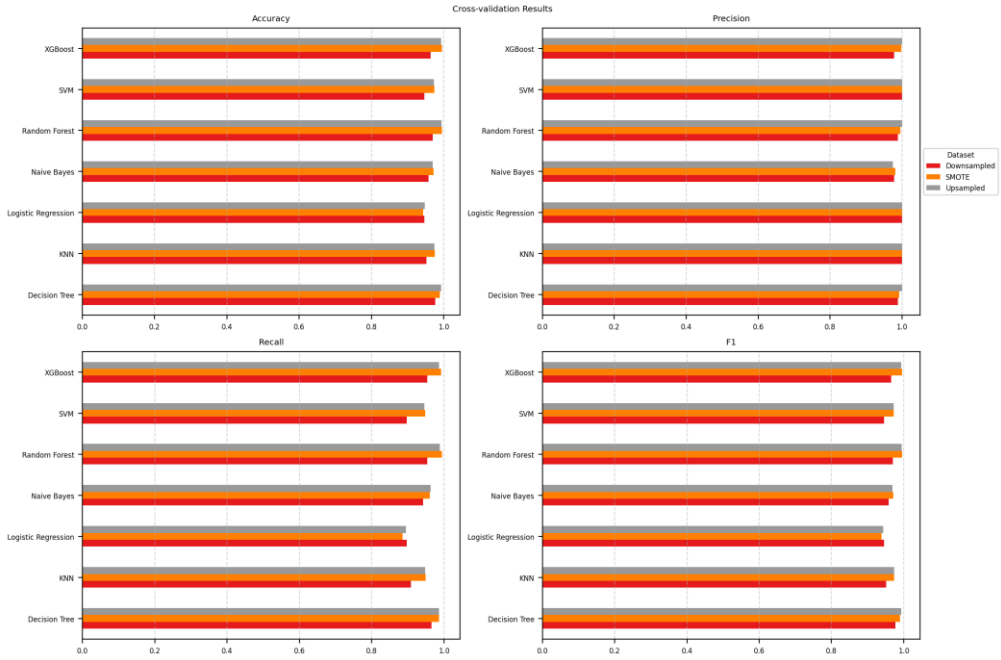

Figure 7: Cross Validation results across different models

Figure 8: Model Performance Before and After Hyperparameter Tuning

Observations

The three classifiers—SVM, XGBoost, and Random Forest—demonstrate consistently high performance across all metrics, suggesting they are well-tuned. Among them, SVM achieves the highest recall, making it an optimal choice if the priority is to capture all potential fire events and minimize false negatives. XGBoost and Random Forest, however, have very close scores in precision, F1-score, and accuracy, indicating their reliability in both detecting actual fire events and avoiding false positives.

Choosing the Best Classifier

In selecting the best classifier, recall is a critical metric, as it measures the model's ability to catch all actual fire cases, reducing the risk of missing a true fire event, which could allow fires to grow unchecked and lead to dangerous situations. This emphasis on recall aligns with the importance of avoiding false negatives, as missing an actual fire could result in significant costs and damage, whereas a false positive would primarily result in additional inspections, which pose less risk. Although recall is prioritized, maintaining a balance with precision is also essential to prevent an excessive number of false alarms; however, recall remains the primary focus to ensure reliable detection of fire risks.

Choosing SVM

1. Simplicity and Interpretability

- SVM is simpler and easier to understand, especially if it's a linear SVM. It creates a clear boundary between fire and non-fire cases.
- XGBoost and Random Forest combine many decisions, which can make it harder to explain why a specific prediction was made.

2. Less Tuning Needed

- SVM generally performs well with fewer settings (hyperparameters) to adjust, making it quicker to set up.
- XGBoost and Random Forest have many settings that need fine-tuning to work well, which can take extra time and effort.

3. Efficient and Fast

- SVM can be more memory-efficient, using less computer power, especially if you have limited resources.
- XGBoost and Random Forest need more memory and processing power, especially with larger datasets.

4. Lower Risk of Overfitting

- SVM has a regularization setting (C) that controls complexity, which helps it avoid overfitting (memorizing the data too closely).
- XGBoost and Random Forest can sometimes overfit, especially if they're not carefully tuned, meaning they might perform worse on new data.

5. Good Generalization

- SVM is known to generalize well to new data, meaning it can be stable and consistent even when used on data it hasn't seen before.
- XGBoost and Random Forest are powerful but can sometimes be more sensitive to changes in the data.

Confusion Matrix – SVM

The confusion matrix reveals that the SVM model is highly accurate in predicting forest fires, with 137 true positives and 123 true negatives, indicating that most predictions are correct. There are no false positives, demonstrating high precision and reliability, as the model avoids raising unnecessary false alarms. Additionally, the model has good recall, with only two missed fire cases (false negatives), meaning it successfully detects nearly all actual fire events. Overall, this SVM model is effective at accurately identifying fire occurrences with minimal errors, ensuring both high reliability and strong sensitivity in its predictions.

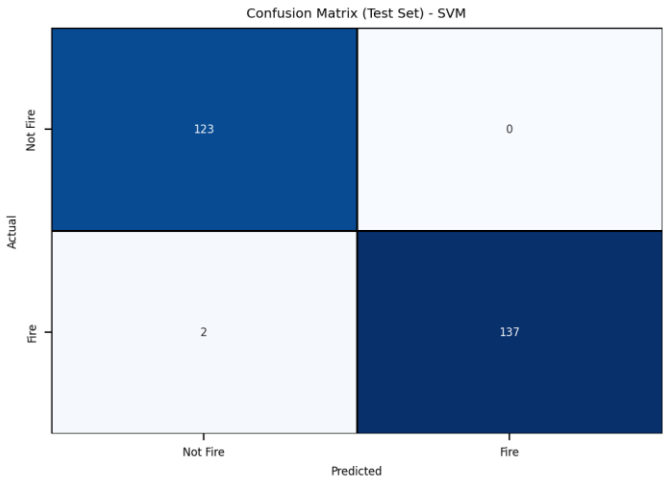

Figure 9: Confusion Matrix (Test set) - SVM

CHAPTER 3

MILESTONES AND DELIVERABLES

Milestones	Status
Data Collection	Completed
Data Cleaning	Completed
Data Merge	Completed
EDA	Completed
Data Preprocessing	Completed
Model Selection and Development	Completed
UI Development	In progress
Map Integration	In progress
Other improvements	In progress
Deployment	To be started
Final Report	In progress

CHAPTER 4
CHALLENGES AND ISSUES

The major challenges in the application development are acquiring additional datasets for testing and integrating an interactive map. The lack of diverse, high-quality datasets is a significant hurdle, as more data is needed to ensure the model's accuracy across different regions, conditions, and environmental factors. Efforts are underway to find new data sources, but harmonizing them for consistent use in the model is time-consuming. Additionally, integrating an interactive map that displays fire predictions and environmental data is complex. Developing a smooth, user-friendly map interface that allows users to interact with various data layers requires advanced front-end development and seamless connectivity with the model, adding technical complexity to the project. Both challenges need to be addressed to ensure a comprehensive and functional tool for forest fire prediction.

CHAPTER 5
NEXT STEPS

To complete this project, the remaining steps focus on UI development, map integration, and deployment. The user interface will present model predictions in an accessible dashboard, allowing users to interact with fire risk data, and visualize trends. Map integration will add an interactive geographic layer, enabling users to view predictions and environmental factors across specific areas. The deployment will involve hosting the application on a cloud platform (e.g., AWS EC2), establishing the UI, map, and model, and ensuring data scalability, ensuring reliable performance and usability in forest fire prediction and management.

CHAPTER 6
CONCLUSION

In conclusion, this project demonstrates significant progress in advancing forest fire prediction through the application of machine learning models. The evaluation of various models, including XGBoost, SVM, and Random Forest, has shown their potential in accurately predicting forest fires by leveraging a wide range of environmental and meteorological factors. Notably, the study highlights the importance of recall in minimizing false negatives, ensuring that potential fire events are detected early. The integration of techniques like SMOTE has helped address data imbalances, improving model performance across different metrics.

While the project has made considerable strides in data collection, preprocessing, and initial model testing, challenges remain in acquiring more diverse datasets for further testing and refining the model's capabilities. Additionally, integrating an interactive map and ensuring smooth deployment are key tasks that need to be completed. Once these hurdles are overcome, the project aims to provide a comprehensive, user-friendly tool for forest fire prediction and management. The final deployment will enable stakeholders, such as forest management agencies, to make informed decisions based on accurate fire risk predictions, ultimately contributing to more effective forest conservation and disaster prevention efforts.