Wildfire Size Forecasting: A Machine Learning Approach

Abstract - Wildfires have burned an average of 7 million acres per year between 2001 and 2020 (NCEI et al., 2023). A National Park Service article claims that 85% of wildfires are caused by humans and the rest by extremely long and hot lightning bolts (Park Service et al). Predicting and mitigating forest fires is critical for environmental preservation, human safety, and economic well-being. This project aims to develop a comprehensive forest fire prediction system using machine learning algorithms, including SVM, Naive Bayesian, Decision Tree, Random Forest, and Logistic Regression. The dataset, encompassing factors like wind, humidity, temperature, and vegetation, is analyzed to predict the severity of wildfires. The project addresses sustainability goals outlined by the United Nations, emphasizing the impact of wildfires on agriculture, health, infrastructure, consumption patterns, climate, and biodiversity. The objective is to contribute to Sustainable Development Goals by developing a model capable of predicting forest fire size based on various environmental factors. The literature review highlights existing studies utilizing machine learning for forest fire prediction and emphasizes the importance of machine learning in this domain. Data engineering involves collection, description, exploratory data analysis, and preprocessing. The project explores different classifiers, and the evaluation metrics include accuracy, precision, recall, F1 score, AUC-ROC, and confusion matrix. Random Forest emerges as the best-performing model with an accuracy of 75.44%, demonstrating its capability to handle complex relationships in the data. The study concludes that forest fire prediction is a multifaceted problem requiring a combination of machine learning models and real-world strategies for effective prevention and containment.

1 INTRODUCTION

1.1 Project Background

Predicting forest fires is essential for maintaining ecosystems and reducing damage to people and their property. Devastating effects of forest fires can include destruction of habitat, declines in biodiversity, and release of toxic pollutants into the atmosphere. Furthermore, forest fires have high financial costs because they cause billions of dollars' worth of property damage and fire fighting expenses. We can anticipate these dangers and lessen the impact of such tragedies by making forest fire prediction and prevention a priority. For instance, forest fires in the US have caused ecological damages estimated at $16 billion. This highlights the importance of investing in strategies that can help predict forest fires, such as monitoring weather conditions, managing vegetation, and promoting sustainable forest management practices. Furthermore, the effects of forest fires go beyond monetary losses. They can also lead to the loss of human life, displacement of communities, and damage to cultural heritage.

The project aims to develop a comprehensive forest fire prediction system that leverages advanced technologies on the dataset consisting of different factors like wind, humidity,

temperature, vegetation and applying machine learning algorithms namely - SVM, Naive Bayesian, Decision Tree, Random Forest, Logistic Regression. After which, we will make use of evaluation metrics to determine which top three models perform the best.

The goal of this project is to accurately predict the severity of a wildfire and to classify them at a very early stage of their development. Our final goal of this project is to help mitigate the forest fires and keep the damage to a minimum. The output of our research which takes several factors into consideration, can help and aid the research which is ongoing.

Also, it is significantly important to understand the problems which are associated with forest fires as Forest fires can spread quickly, making it difficult to detect them at first, particularly in isolated areas. This may result in a delayed reaction time, which might let the fire get bigger and do more harm.Also, a lot of forests are situated in isolated locations, making it challenging for emergency responders and other staff to get there fast. This may cause a lag in the reaction time and complicate fire containment. The frequency and intensity of forest fires are predicted to rise due to climate change, especially in areas with arid vegetation. Because of this, it's imperative to create practical plans for anticipating and putting out forest fires in an environment that is changing.

1.2 SUSTAINABILITY

According to the article on UNEP (UN et al., 2020) wildfires have a great impact on the following six SDGs:
SDG 2: Zero hunger - Wildfires can damage agricultural land and forest resources, leading to food insecurity and loss of livelihoods for rural communities.
SDG 3: Good health and well-being - Wildfires can release harmful pollutants into the air, posing health risks to people in the surrounding area.
SDG 9: Industry, innovation and infrastructure - Wildfires can damage infrastructure such as power lines and homes, requiring costly repairs.
SDG 12: Responsible consumption and production - Unsustainable consumption patterns and pollution contribute to global heating, which in turn increases the likelihood of wildfires.
SDG 13: Climate action - Wildfires release greenhouse gasses and contribute to global heating, accelerating climate change.
SDG 15: Life on land - Wildfires can destroy habitats and biodiversity, affecting ecosystems and the livelihoods of communities that depend on them.

1.3 OBJECTIVE

This project aims at developing a classification model that is capable of predicting the size of forest fires based on vegetation factors as well as meteorological factors, which is an important contribution to achieving the Sustainable Development Goals.

1.4 LITERATURE REVIEW

This work (Coffield et al., 2019) shows how machine learning can be used to estimate the size of a fire at the point of ignition. The State of Alaska provided the authors with data on vegetation, topography, fires, and weather. A decision tree model that effectively divides ignition events into small, medium, and large fires was created by them. Vapor-pressure deficit (VPD) and a vegetation variable that indicates the presence of a particular tree species were used to train the model. With an accuracy of about 49%, the results indicate that VPD was the best predictor of fire size at ignition. Accuracy was raised to about 50% by including the vegetation variable.

In order to predict the susceptibility of a forest fire in the Lao Cai province of Vietnam, this paper (Tien Bui et al., 2019) presents a machine learning approach that uses Differential Flower Pollination optimization (DFP) and Multivariate Adaptive Regression Splines (MARS). The model takes into account ten factors, such as topography, vegetation, climate, and human activity. With MARS-DFP, the model achieves an accuracy of approximately 95%, outperforming other algorithms like BPANN, Random Forest, and MARS alone. The Normalized Difference Vegetation Index (NDVI), a measure of the health of the vegetation, is found to be the most significant factor.

Artificial intelligence (AI) has been used in wildfire science and management since the 1990s, with machine learning (ML) methods being widely adopted. This scoping review (Jain et al., 2020) aims to improve awareness of ML methods among researchers and managers, illustrating the diverse range of problems available to ML data scientists. ML approaches are categorized into six problem domains: fuels characterization, fire detection, fire weather, fire occurrence, fire behavior prediction, fire effects, and fire management. The review identified 300 publications mentioning random forests, MaxEnt, artificial neural networks, decision trees, support vector machines, and genetic algorithms. Opportunities for ML methods in wildfire science include deep learning and agent-based learning.

Climate change is posing a significant threat to forests, with wildfires being a major driver of loss. To address this issue, a wildfire data inventory was created (Ghorbanzadeh et al., 2019) by integrating GPS polygons with data from the MODIS thermal anomalies product. Different machine learning (ML) approaches were applied to predict wildfire susceptibility, including artificial neural networks (ANN), support vector machines (SVM), and random forest (RF). The resulting maps showed CV accuracies of 74%, 79%, and 88% for ANN, SVM, and RF, respectively.

Remote Sensing, Machine Learning, and big data are crucial in predicting and preventing wildfires. These technologies provide a vast amount of data for monitoring and preventing these disasters. This paper (Sayad et al., 2019) combines these techniques to process satellite images

and extract insights to predict wildfire occurrences. The dataset, derived from MODIS, includes data on crop state, meteorological conditions, and fire indicators. The results show high prediction accuracy (98.32%), demonstrating the potential of these technologies in preventing and preventing wildfires.

Machine learning models are being developed to predict wildfire spread, a crucial aspect of disaster preparedness. The Next Day Wildfire Spread dataset (Huot et al., 2022), a large-scale, multivariate dataset combining remote-sensing data from the United States, is used to develop models. The dataset, compared to logistic regression and random forest, provides a feature-rich dataset for machine learning, enabling the development of models that can predict wildfire spread with a one-day lead time.

California has been plagued by wildfires for years, causing economic and environmental loss. Recent machine learning models have been introduced to predict wildfire risks, but their accuracy has been limited. This paper (Malik et al., 2021) proposes two data-driven approaches based on random forest models to predict wildfire risk near Monticello and Winters, California. The combined model uses spatial and temporal parameters as a single dataset, while the ensemble model uses separate parameters. The combined model achieved a 92% accuracy in predicting wildfire risks, including ignition, using regional spatial and temporal data along with standard data parameters in Northern California. The models were validated using Receiver Operating Characteristic (ROC) curves, learning curves, and evaluation metrics such as accuracy, confusion matrices, and classification report.

2. Data Engineering

Figure 1 Flow chart of the project

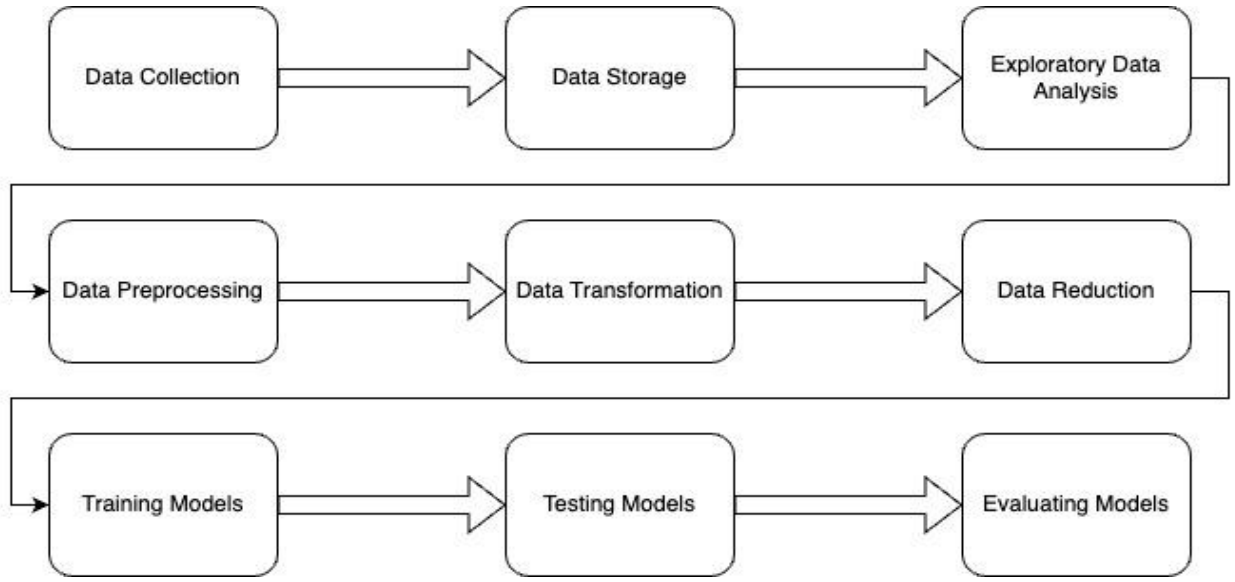
2.1 Data Collection

The dataset utilized in this project (Coder, 2020) is a subset of a larger collection of data encompassing 1.88 million fires in the United States and ranges from 1991 - 2015. This subset was created by combining historical weather and vegetation data sharing the same latitude and longitude as 50,000 randomly selected fire samples. Given below are some of the definitions of the important variables used in our project.

2.2 Data Description

Fire_size_class - Class of Fire Size (A-G)
Stat_cause_descr - Cause of Fire
Latitude - Latitude of Fire
Longitude - Longitude of Fire
Discovery_month - Month in which Fire was discovered
Vegetation - Dominant vegetation in the areas (can save some factors of vegetation)
Temp_pre - temperature in deg C at the location of fire up to 30, 15 and 7 days prior

Temp_cont - temperature in deg C at the location of fire up to day the fire was



Wind_pre - wind in deg C at the location of fire up to 30, 15 and 7 days prior
Wind_cont - wind in deg C at the location of fire up to day the fire was
Prec_pre - Precipitation in deg C at the location of fire up to 30, 15 and 7 days prior
Prec_cont - Precipitation in deg C at the location of fire up to day the fire was
Hum_pre - Humidity in deg C at the location of fire up to 30, 15 and 7 days prior
Hum_cont - Humidity in deg C at the location of fire up to day the fire was
Remoteness - non-dimensional distance to closest city

2.3 Exploratory Data Analysis

Exploratory data analysis (EDA) is an iterative process that involves summarizing and analyzing datasets in order to uncover potential problems, better understand the underlying structure and patterns of the data, and produce hypotheses for additional research.

We implemented EDA in our project by plotting several charts namely, bar, choropleth, box and whisker plots to better understand the data and find out the outliers which can be removed in the data pre-processing phase

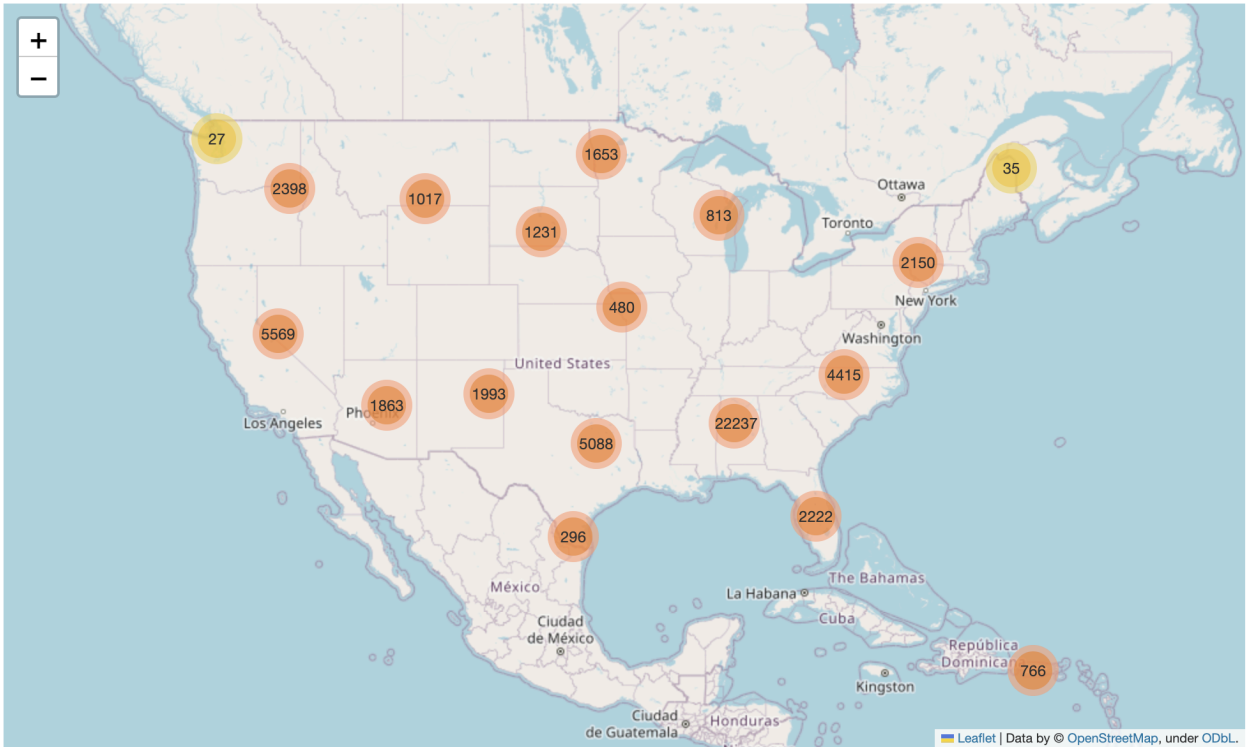


Figure2 Choropleth map of the fire occurrences in different regions of the USA.

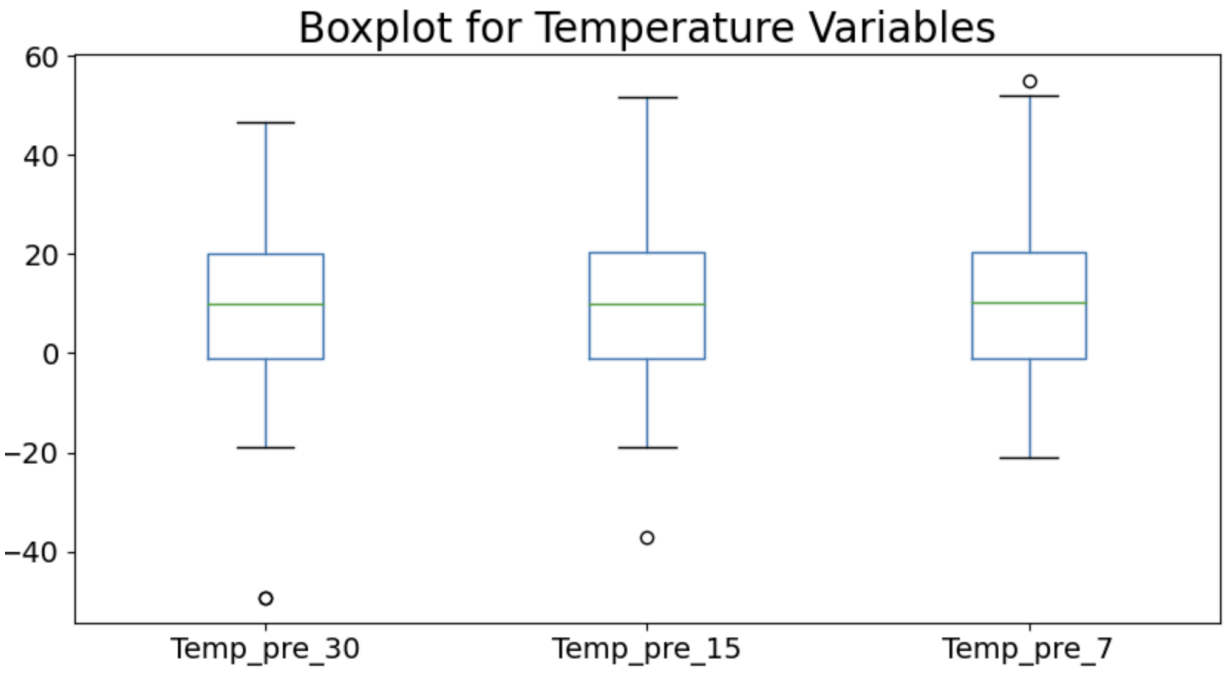


Figure3 Box plot of Temperature values for 30, 15 and 7 days prior to the fire.

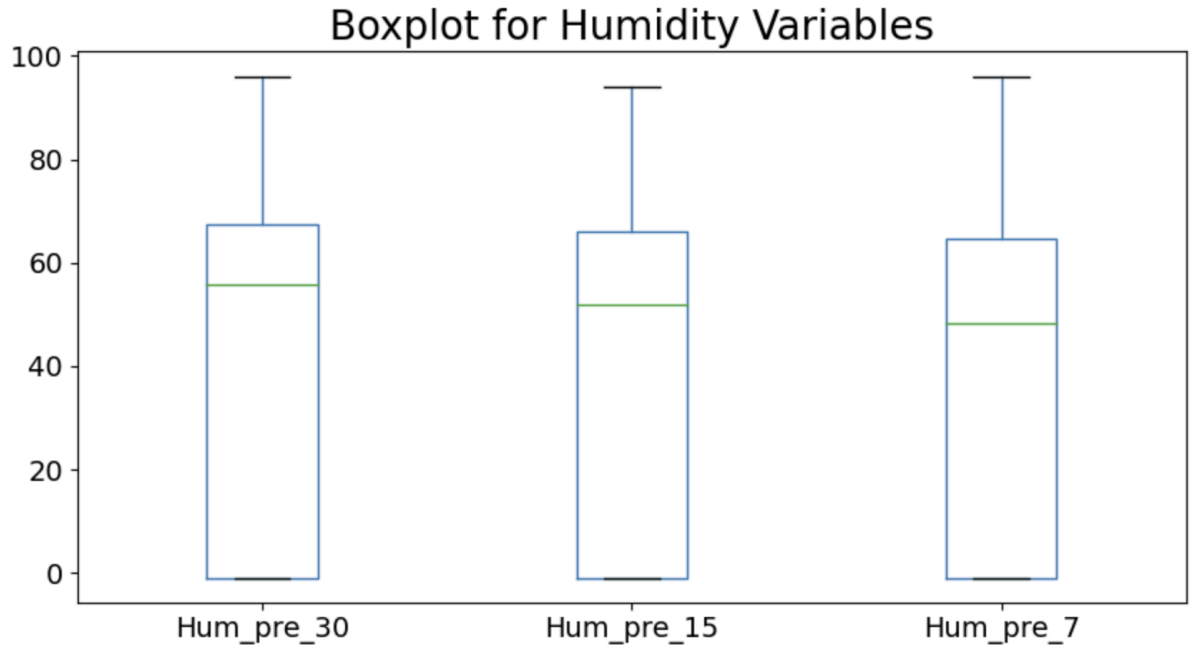


Figure4 Box plot of Humidity values for 30, 15 and 7 days prior to the fire.

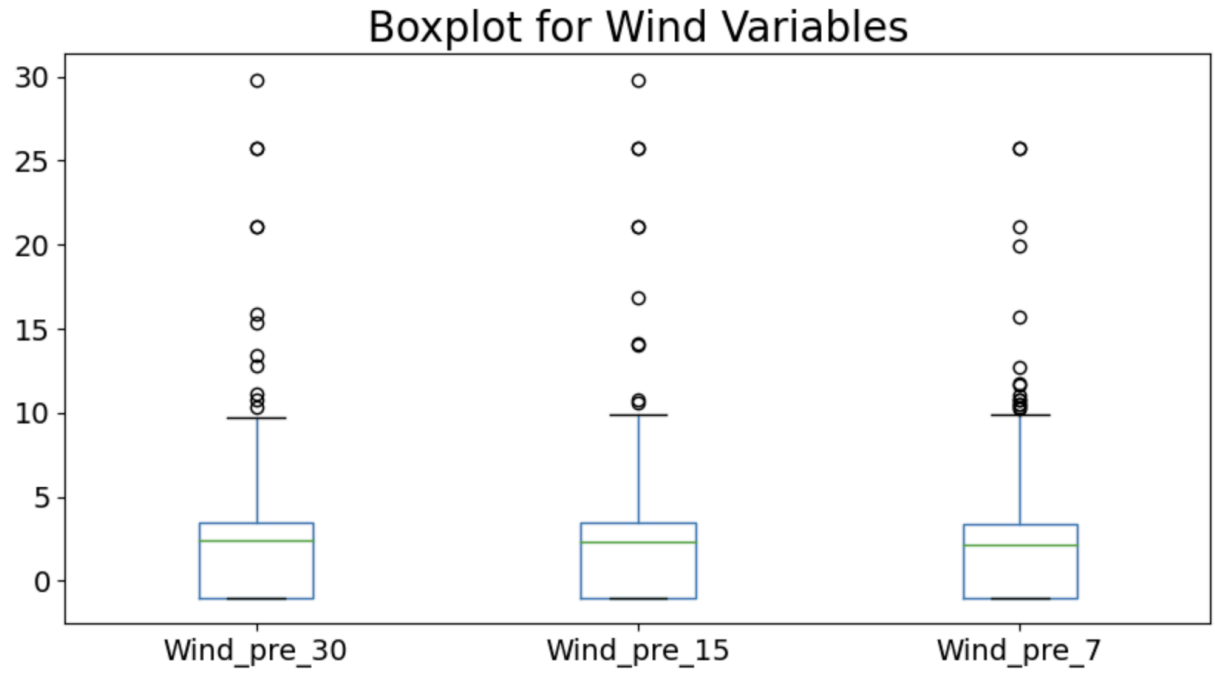


Figure5 Box plot of Wind values for 30, 15 and 7 days prior to the fire.

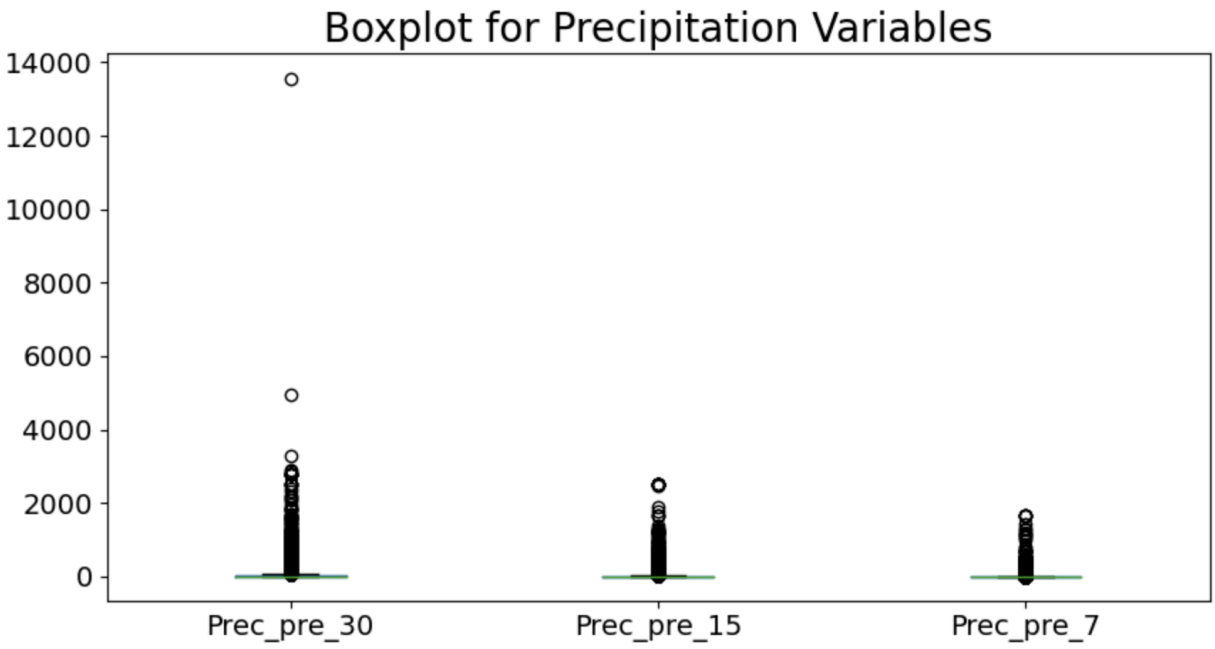


Figure6 Box plot of Precipitation values for 30, 15 and 7 days prior to the fire.

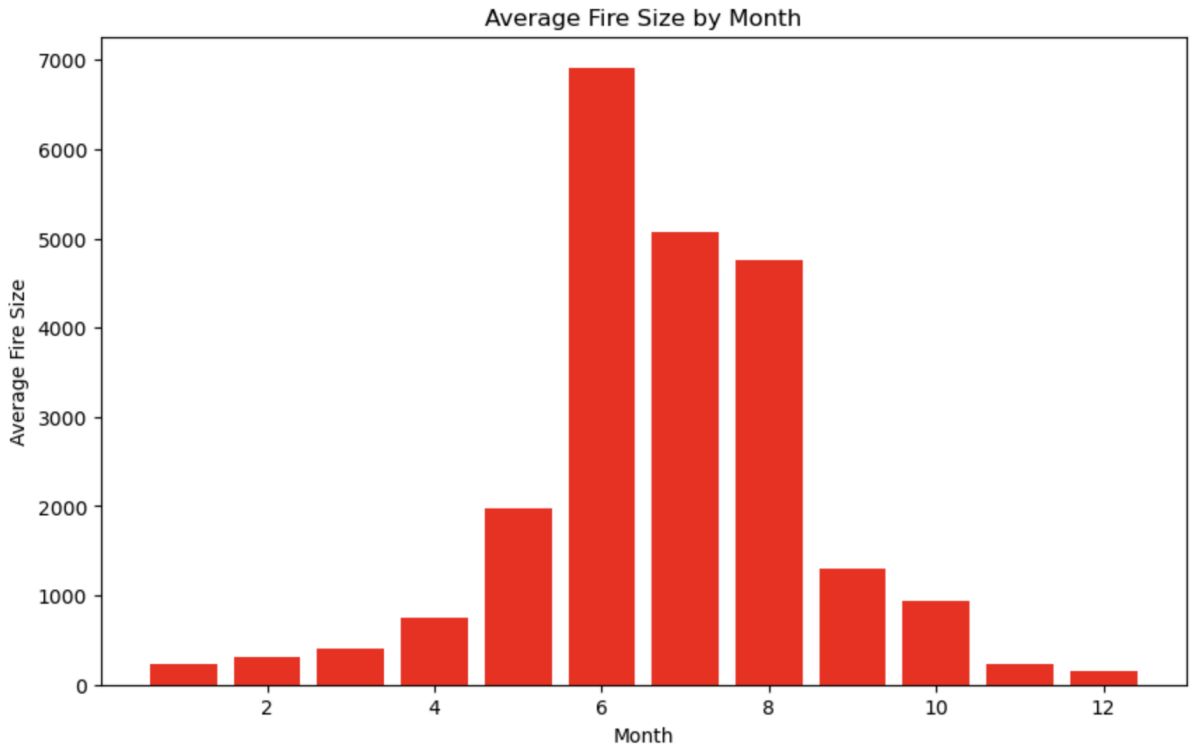


Figure7 Bar Chart of average fire size over the months.

2.4 Data pre-processing

A data analysis method called "data pre-processing" is used to transform manageable, raw data into a format that is both practical and well-organized.

2.4.1 Data Cleaning

In addition to having a lot of redundant and irrelevant information, data can occasionally be empty. In order to address this, data cleansing is used. Three of the columns were dropped which were having null values and were not important for the analysis. Here certain irrelevant columns like state, putout_time, disc_pre_year were also dropped.

2.4.2. Data Transformation

Data transformation is the process of converting raw data into a usable format that can be analyzed and used for decision-making. We addressed this in our project by using the MinmaxScaler which scaled the range of wind, temperature, humidity and precipitation columns to be between 0 and 1.

The target variable which was multiclass was imbalanced, so we decided to club the classes into two, one for fire size who's magnitude was less than 9.9 acres and other greater than 9.9 acres.

Table 1. Converting Multiclass Classification Problem to Binary Classification

Fire_Size_Class	Spread	Target Variable
A,B	<9.9 acres	0
C,D,E,F,G	>9.9 acres	1

In order to convert categorical variables into numerical values based on the mean (or other aggregation) of the target variable corresponding to each category, we employed a machine learning technique called target encoding. In a supervised learning environment, this technique aids in capturing correlations between categorical features and the target variable.

2.5 Model Selection

Accurately predicting forest fires is crucial for preventing environmental damage and safeguarding human lives. Machine learning algorithms have emerged as powerful tools for forest fire prediction, offering promising results in various studies. Among the numerous models available, Support Vector Machines (SVMs), K-Nearest Neighbors (KNN), Random Forest, Decision Trees, and Logistic Regression have demonstrated notable performance in forest fire prediction applications. SVMs excel in handling complex nonlinear relationships between features and the target variable, making them well-suited for forest fire prediction, where environmental factors interact in intricate ways. KNN's simplicity and robustness to noise make

it a viable choice, particularly when dealing with heterogeneous data sets. Random Forest, with its ensemble of decision trees, effectively captures complex interactions and reduces overfitting, leading to improved predictive performance. Decision Trees, while prone to overfitting, offer interpretability, providing insights into the factors influencing forest fire occurrence. Logistic Regression, though limited to linear relationships, provides clear probabilistic predictions and is computationally efficient.

2.6 Model Validation

It was essential to thoroughly validate the performance of our machine learning models after selecting them in order to guarantee their reliability and generalizability. Model validation involves evaluating the model's ability to accurately predict forest fires using data that was not used during training. This process helps identify potential biases or limitations in the model and provides insights into its true predictive power.

Several techniques are commonly employed for model validation in forest fire prediction. One approach is to split the available data into training and testing sets. Randomly splitting data for model training and testing can lead to a class imbalance, where certain target classes are overrepresented in one set compared to the other. This imbalance can introduce bias into the training process, resulting in a model that performs poorly on the underrepresented classes. To prevent this issue and ensure a fair representation of target classes in both the training and testing sets, a technique called StratifiedKFold can be employed. StratifiedKFold repeatedly divides the data into k folds while maintaining the same proportion of target classes in each fold. This ensures that the model is trained on a balanced representation of the data, leading to more accurate and unbiased predictions. The model is trained on each fold except one, and its performance is evaluated on the excluded fold. This process is repeated for each fold, providing a more comprehensive assessment of the model's generalizability. Here we are using 5 fold cross validation for our models.

2.7 Model Evaluation

Evaluating the performance of machine learning models is crucial for ensuring their effectiveness in real-world applications.Several metrics are commonly used to assess the performance of forest fire prediction models. These metrics provide insights into the model's ability to correctly identify forest fires and avoid false alarms.

Some of the metrics which we have used are Accuracy, Precision, Recall, F1 score, AUC (Area under the curve).

3. METHODOLOGY

3.1 Model Development

In the modeling phase of this project, a number of classification models were built for the prediction of the target feature “Fire_Size_Class”.

3.1.1 SVM Classifier -SVM classifiers are classification techniques that use supervised learning. They determine the best hyperplane for categorizing data into discrete groups. SVMs can handle high-dimensional data, detect non-linear correlations, and perform binary and multi-class classification problems.
Our project employs SVM's to construct a hyperplane that separates the data points belonging to different classes, effectively classifying new data points. This classification capability can be used to predict whether a wildfire will spread rapidly or remain contained.

3.1.2 Decision Tree Classifier
Decision Trees are tree-structured machine learning algorithms that recursively partition the data into subsets based on predefined decision rules, leading to a classification or prediction for each data point.
Decision trees are used in our project to create a tree-like model that splits the data into smaller subsets based on decision rules, with each leaf node representing a predicted risk category.
We could learn more about the elements that lead to a large wildfire spread by dissecting the decision-making criteria and the features that are utilized to make these decisions.

3.1.3 Random Forest Classifier
Random Forest classifiers are ensemble learning methods that mix numerous decision trees to improve the robustness and accuracy of the model. They improve the model's ability to generalize to new data, reduce overfitting, and are widely employed because of their excellent accuracy and ability to handle high-dimensional data.
While Decision Trees offer simplicity and ease of interpretation, Random Forest's enhanced predictive performance and reduced overfitting make it a more robust and reliable choice in our project.

3.1.4 Naive Bayes Classifier
Naive Bayes classifiers are Bayes theorem-based probabilistic classifiers. They divide data into groups based on a set of characteristics. They are common in applications like spam filtering, sentiment analysis, and medical diagnosis.
It utilized Bayes' theorem to estimate the probability of a wildfire spreading to a large size given its observed features, such as wind, temperature, humidity, and vegetation. The algorithm constructed a probabilistic model that represented the relationships between various factors and the likelihood of a large wildfire occurrence.

3.1.5 KNN Classifier
K-Nearest Neighbors (KNN) is a classification and regression machine learning technique. KNN is employed to classify areas as high or low risk for forest fires based on their similarity to historical fire occurrences.

KNN identified the k nearest neighbors, or the wildfire events that share the most similar environmental characteristics to the new event. Based on the characteristics of these nearest neighbors, KNN could infer the likelihood and potential spread of the new wildfire for our project.

3.1.6 Logistic Regression Classifier

Logistic regression is a machine learning approach that is used to solve classification problems. It forecasts the possibility of an instance falling into a certain class using a linear regression model. It computes the probability of each class given the input features to predict the class of future examples after training.

It estimates the relationship between a set of independent variables (e.g., meteorological conditions, topography, vegetation) and the dependent variable (fire_size_class). By analyzing historical data on wildfires and their associated environmental factors, logistic regression can generate a model that predicts the probability of a small or large wildfire spread based on new observations of these factors.

3.2 Evaluation metrics

The classification models can be evaluated using several metrics such as

3.2.1 Accuracy: The proportion of correctly classified instances out of all instances in the test dataset.

3.2.2 Precision: The proportion of true positives out of all positive predictions made by the model.

3.2.3 Recall: The proportion of true positives out of all actual positive instances in the test dataset.

3.2.4 F1-score: The harmonic mean of precision and recall, providing a balanced measure of both.

3.2.5 AUC-ROC: The area under the ROC curve, indicating the model's ability to distinguish between positive and negative classes.

3.2.6 Confusion matrix: A table that summarizes the predictions made by the model against the actual class labels, allowing for analysis of the model's performance and identification of issues such as misclassification.

3.3 Model Comparison

For 5 fold Cross Validation:

Models	Accuracy	Precision	Recall	F1 Score

Logistic Regression	72.99	71.75	72.25	68.81
Decision Tree Classifier	68.35	68.11	67.95	68.02
Random Forest Classifier	75.44	74.64	75.18	73.34
K Nearest Neighbors Classifier	71.18	69.73	70.95	69.88
Naive Bayes	67.70	66.15	67.81	66.43
Support Vector Machine	67.50	74.79	67.48	56.07

Table 2: Model Comparison

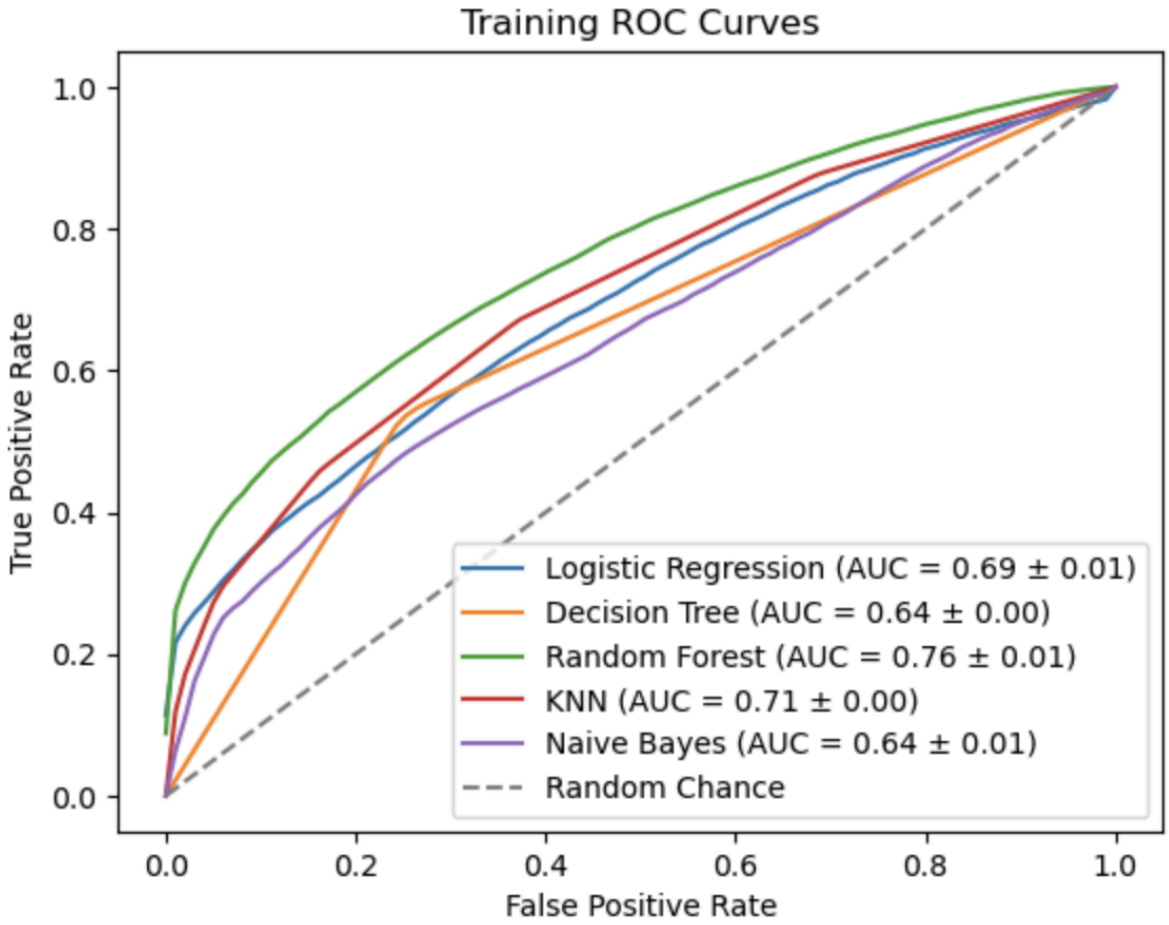


Figure8: AUC-ROC Curves for all the models

3.4 Tools Used - Google Colab and Jupyter Notebook were utilized for implementing the model. Data preprocessing is done with sklearn; visualizations are done with matplotlib and seaborn; performance comparison is done with sklearn models. As communication tools, Trello and

GitHub were used to facilitate teamwork, code development, and task delegation. GitHub offered a code sharing and version control platform that let team members collaborate on code development and monitor changes over time.We were able to assign and create tasks, monitor progress, and share updates by using Trello. The team was able to work productively and efficiently to complete the project by making use of these tools.

5 Acknowledgement

We would like to express our gratitude to Professor Dr. Vishnu Pendyala for his valuable assistance and support during the course of this project. His guidance has been instrumental in helping us achieve our goals.

6. KEY LEARNINGS

Our key learnings from the project are:

1) Data Preprocessing is essential: Data preprocessing is essential in improving the performance of the models. We preprocessed the data by removing missing values, handling outliers, and normalizing the data. This helped improve the accuracy of the models.
2) When we apply one hot encoding to some feature, it increases the dimensionality of the dataset by the number of the categories. Additionally, it makes the dataset sparse in nature which does not work well with traditional ML algorithms. This is the reason we move to other encoding techniques such as target encoding.
3) Accuracy is not the only metric that can be considered for model performance, there are other metrics namely, Precision, Recall, F1 Score, AUC/ROC. However, on comparing all of the metrics we found that Random Forest outperformed other models such as SVM, KNN, Naive Bayes, and Decision Trees in terms of accuracy. This could be due to the ability of Random Forest to handle complex relationships between features and target variables.
4) Ensemble models outperformed individual models such as SVM, KNN, and Naive Bayes. This could be due to the ability of ensemble models to combine multiple models and reduce the effect of noise in the data.
5) Forest Fire Prediction is a complex problem: that requires careful consideration of multiple factors. We used machine learning models to predict the likelihood of a fire spread. The deployment methods like public awareness, coordination and collaboration with groups also play an important role, which can help prevent the spread of forest fires more quickly.

7 CONCLUSION -

After training and evaluating several models, Random Forest turned out to be the best classifier with Accuracy of 75.44% and a recall of 75.18%. Its precision turned out to be 74.64%. KNN and Logistic Regression were the second best performing models helping classify the fire_size class efficiently. These models took meteorological, vegetation and topographical factors into consideration. The output from these models can aid in the research and help the authorities concerned to act accordingly by creating public awareness, coordination and collaboration with several organizations so as to minimize the fire spread.

REFERENCES

Coder, C. (2020, October 6). *U.S. wildfire data (plus other attributes)*. Kaggle. https://www.kaggle.com/datasets/capcloudcoder/us-wildfire-data-plus-other-attributes/data

Environment Programe, U. N. (2020). *The effect of wildfires on sustainable development*. UNEP. https://www.unep.org/news-and-stories/story/effect-wildfires-sustainable-development

Ghorbanzadeh, O., Valizadeh Kamran, K., Blaschke, T., Aryal, J., Naboureh, A., Einali, J., & Bian, J. (2019). Spatial prediction of wildfire susceptibility using Field Survey GPS data and machine learning approaches. *Fire*, *2*(3), 43. https://doi.org/10.3390/fire2030043

Huot, F., Hu, R. L., Goyal, N., Sankar, T., Ihme, M., & Chen, Y.-F. (2022). Next day wildfire spread: A machine learning dataset to predict wildfire spreading from remote-sensing data. *IEEE Transactions on Geoscience and Remote Sensing*, *60*, 1–13. https://doi.org/10.1109/tgrs.2022.3192974

Jain, P., Coogan, S. C. P., Subramanian, S. G., Crowley, M., Taylor, S., & Flannigan, M. D. (2020). A review of Machine Learning Applications in wildfire science and management. *Environmental Reviews*, *28*(4), 478–505. https://doi.org/10.1139/er-2020-0019

Malik, A., Rao, M. R., Puppala, N., Koouri, P., Thota, V. A., Liu, Q., Chiao, S., & Gao, J. (2021). Data-driven wildfire risk prediction in Northern California. *Atmosphere*, *12*(1), 109. https://doi.org/10.3390/atmos12010109

Sayad, Y. O., Mousannif, H., & Al Moatassime, H. (2019). Predictive modeling of wildfires: A new dataset and Machine Learning Approach. *Fire Safety Journal*, *104*, 130–146. https://doi.org/10.1016/j.firesaf.2019.01.006