# Predictive Modeling Approach for California’s Wildfire Threat

**Sneha Manjunath Chakrabhavi**

# Table of Contents

## A. Problem Setting

- Devastating impacts on residents, environment, and economy.
- Increased frequency and intensity due to climate change and human activities.
- Immediate threats to lives and property, with long-lasting effects on ecosystems and public health.
- Necessity for understanding wildfire patterns for effective management and mitigation.

## B. Problem Definition

- Develop a predictive model to analyze the presence and trends in wildfire frequency and severity using data mining techniques and machine learning models.
- Identify primary causes and geographical patterns using various natural and human-induced factors.
- Assess impacts on property, and lives also examine patterns across regions for better risk-reduction strategies.
- Provide timely insights to communities and authorities for efficient decision-making.

## C. Data Sources

- Main data source: Kaggle's compilation of historical wildfire records (2013-2020).
- Additional data from Fire.CA.gov for comprehensive insights.

  Kaggle/California_Wildframes/2013-2020

  Fire.CA.gov/incidents

## D. Data Description

- Initial dataset: 40,780 rows across 40 columns, detailing six wildfire occurrences from 2013 to 2020.
- Refined to 18 columns and 10,988 unique wildfire events for more focused analysis.
- Key variables: Date, County, Fire names, Maximum/Minimum/Average temperature, Cause of fire, Latitude, Longitude, and Acres burned.
- Involved in removing duplicates, handling missing values, and standardizing data formats for consistency.
- Streamlined dataset supports predictive modeling to forecast future wildfire incidents.

# Description of Variables

| Columns | Type | Meaning |
|---|---|---|
| Date | Object | The month and year of when the fire took place. |
| Count | Object | The county the fire started in. |
| Maxtemp | Float | The average maximum temperature of that month (°F). |
| Mintemp | Float | The average minimum temperature of that month (°F). |
| Avgtemp | Float | The average temperature of that month (°F). |
| Snow | Float | The total snow for that month. |
| Humid | Float | The average humidity for that month. |
| Wind | Float | The average wind for that month. |
| Precip | Float | The average precipitation for that month. |
| q_avgtemp | Float | The quarterly average temperature (°F). |
| q_avghumid | Float | The quarterly average humidity. |
| q_sumprecip | Float | The quarterly average precipitation. |
| Sunhour | Float | The average hours of sun for that month. |
| Name | Object | The name of the fire. |
| Cause | Float | The cause of the fire. |
| Latitude | Float | The latitude coordinate of the fire's location. |
| Longitude | Float | The longitude coordinate of the fire's location. |
| Acresburned | Float | The total number of acres burned. |

**Numeric Summary:**

- Temperature variables (maxtemp, mintemp, avgtemp) have their averages around 72.79°F, 49.04°F, and 64.68°F respectively.
- Totalsnow has an average of 0.09 inches, indicating that snowfall is rare in the dataset.
- Humidity averages around 54.41% with a standard deviation of 16.93%.

**Categorical Summary:**

- Los Angeles County has the highest number of entries (425), followed by Kern and Mariposa.
- A large number of entries are labeled as no_fire, implying no significant fire event.
- Among the actual fires, names like Canyon and Creek appear frequently.

# E. Data Exploration

**Data cleaning and preparation**

Irrelevant columns and duplicate rows were removed to focus on more impactful variables. Part of the initial cleaning involved addressing missing values, along with standardization to maintain consistency.

**Dimension reduction and Feature selection**

Created features from existing ones to improve performance. No major dimension reductions like PCA are done as most of the essential information was relevant and hence retained.

**Data Transformation**

The spatial and temporal transformation was done on variables along with the introduction of binary variable fire_occurred, derived from the acres_burned for the prediction of wildfire presence.

**Exploratory Data Analysis**

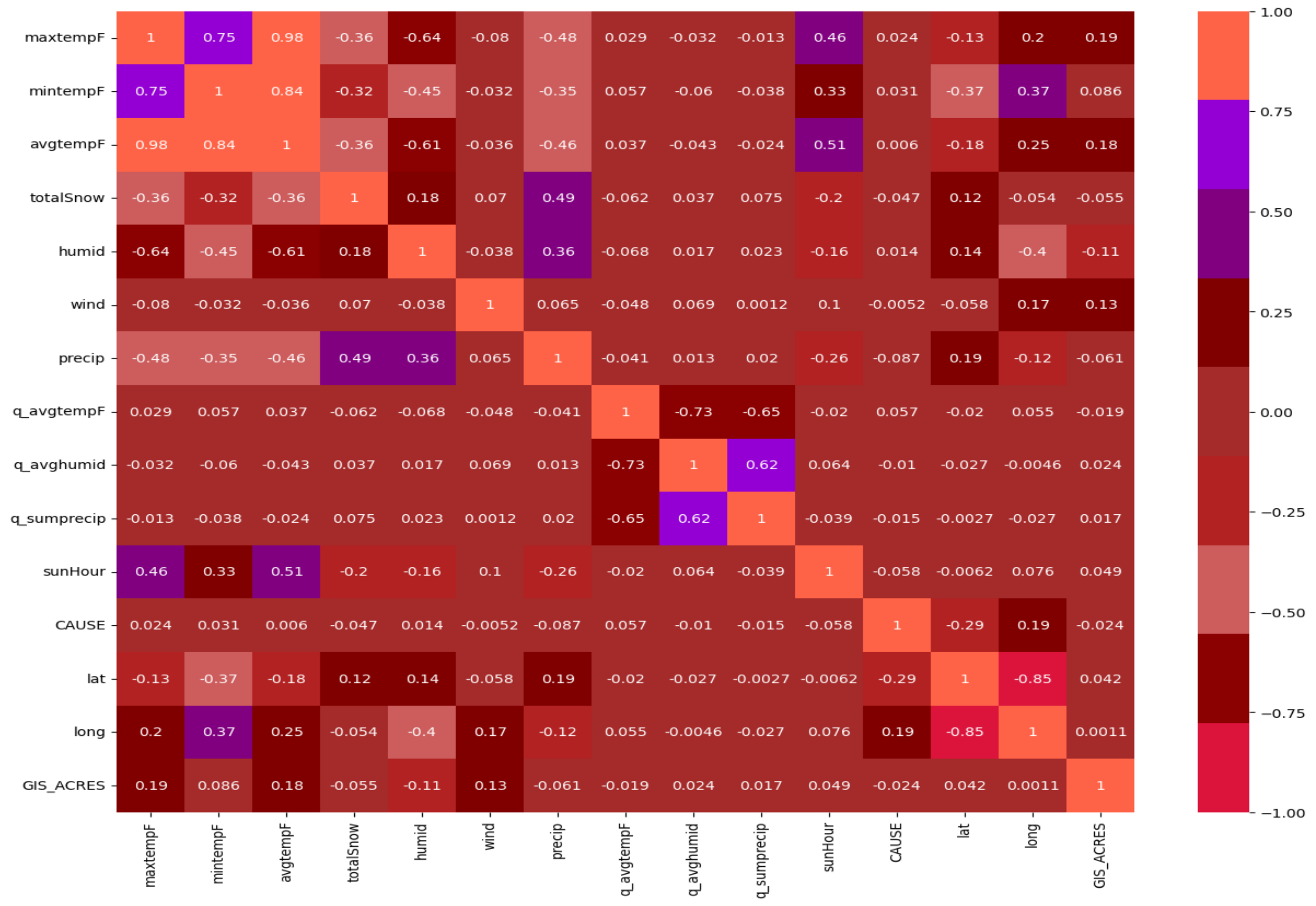
Heatmaps, Bar Graphs, Pair Plots, Boxplots, and Scatter Plots visualized the relationships and influence of these on wildfires showing unique features and correlations.

**Data mining techniques**

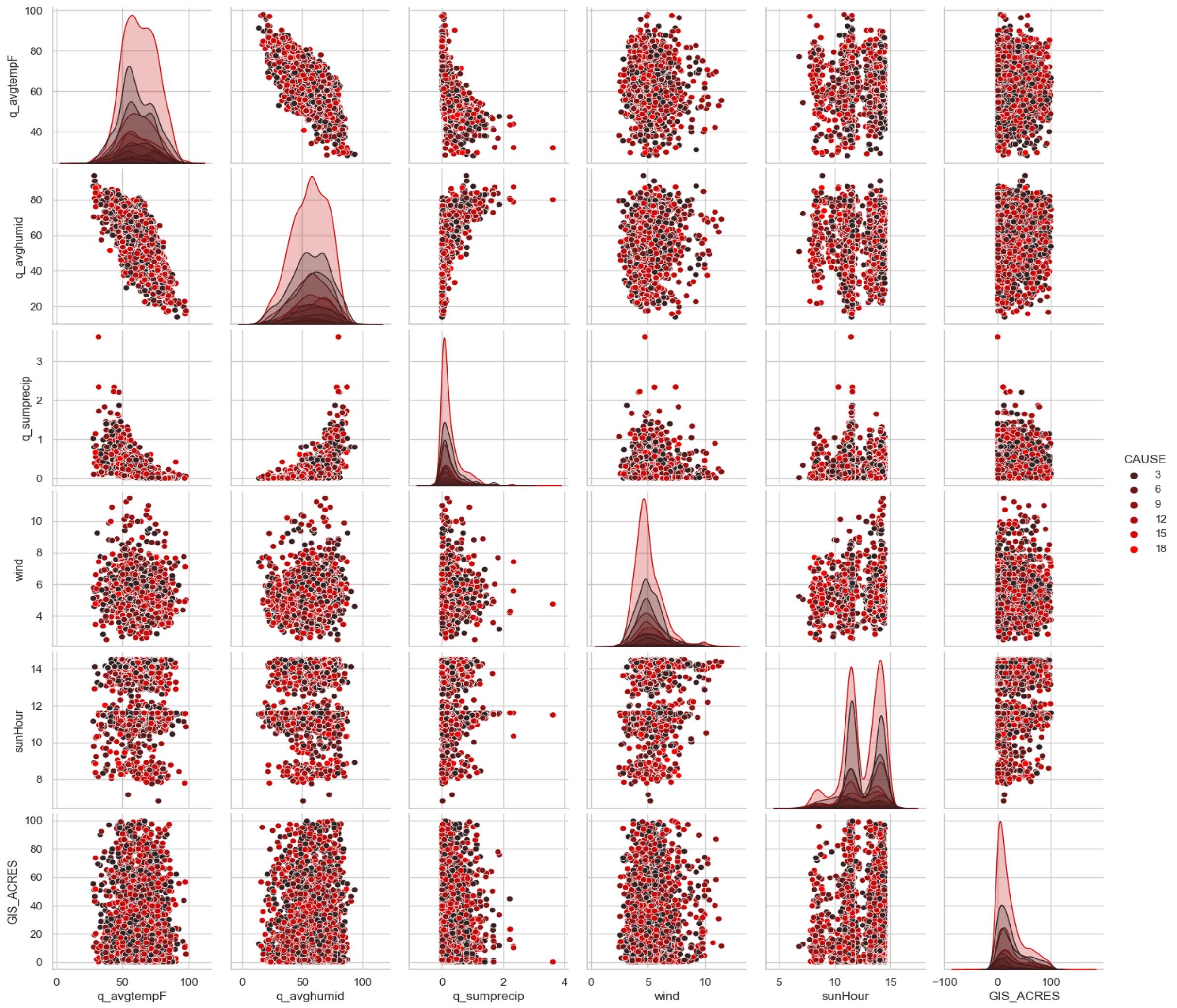
Performs various data mining tasks like model training, testing, model selection and evaluation, now that the dataset is clean and preprocessed it is easy to perform further steps.

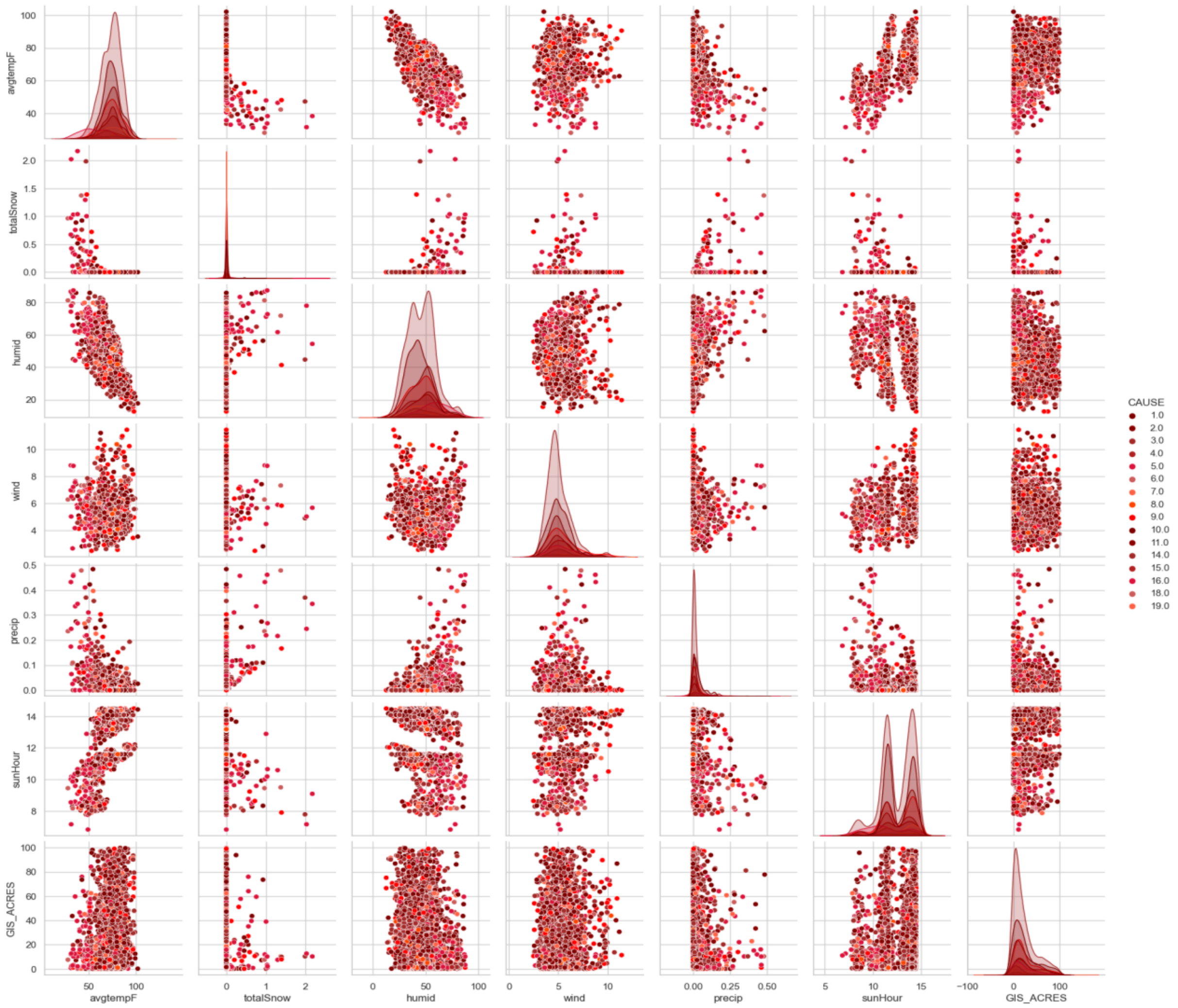
## Exploratory Data Analysis

- Heatmap depicts correlations of weather conditions, geographical coordinates, and wildfire data.
- High temperatures show a strong positive correlation, while humidity has a negative correlation with wildfire occurrence measured by acres_burned.

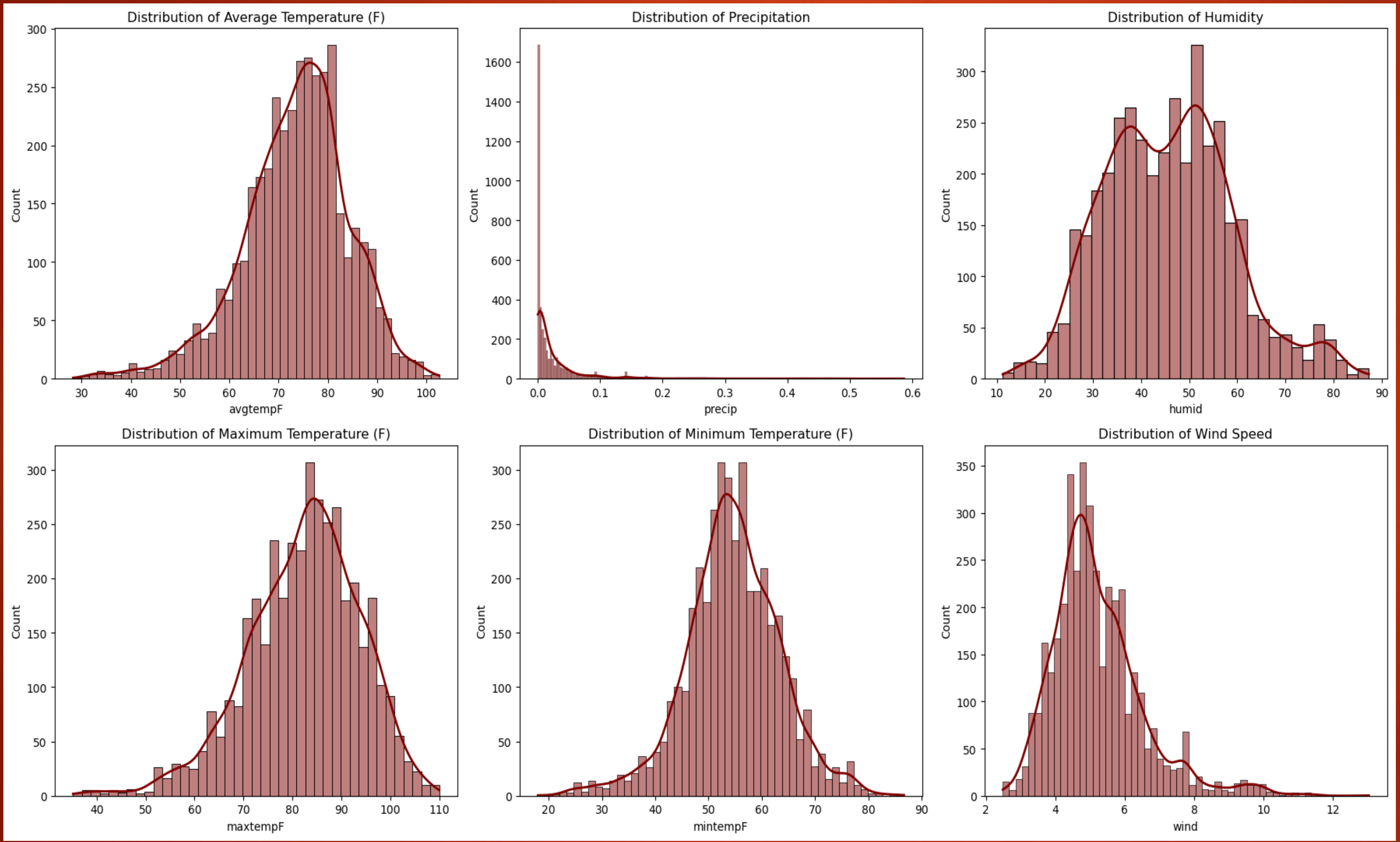


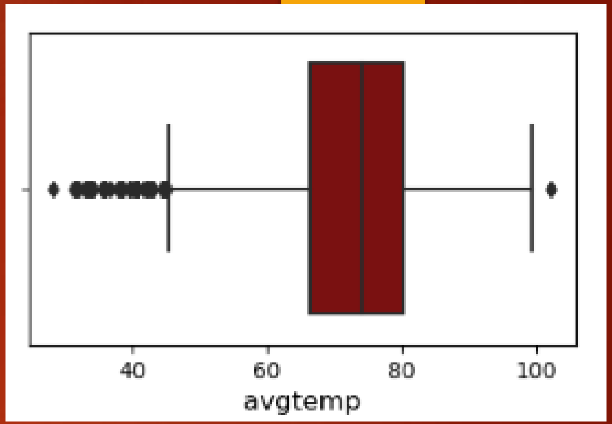
- These pair plots show relationships between different variables and wildfire causes, with distribution density on the diagonal quarterly and monthly.
- Data points are color-coded by cause, illustrating patterns and clusters that may indicate correlations between variables and wildfire triggers.

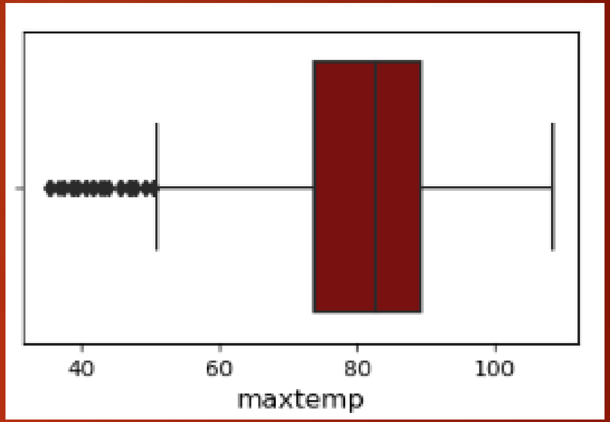


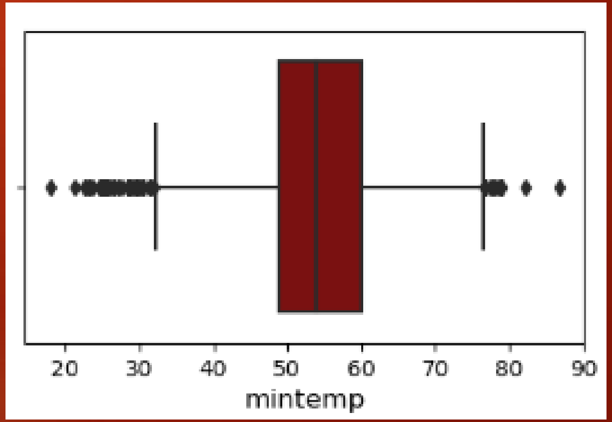



- To explore the relationship between temperature and wildfire occurrences, and to gauge the severity of wildfires in terms of acres burned.
- To explore the distribution of the numeric variables and their correlation.

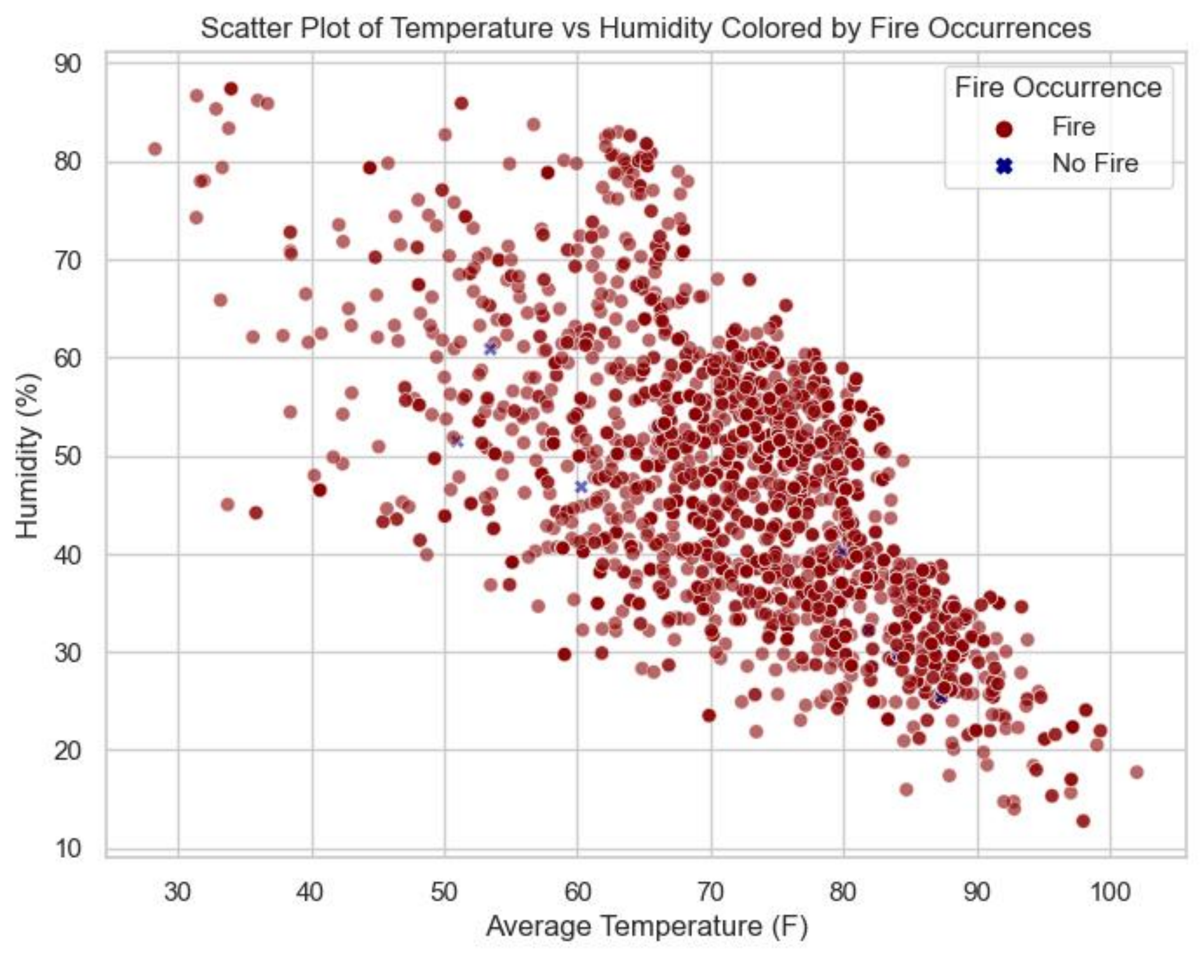


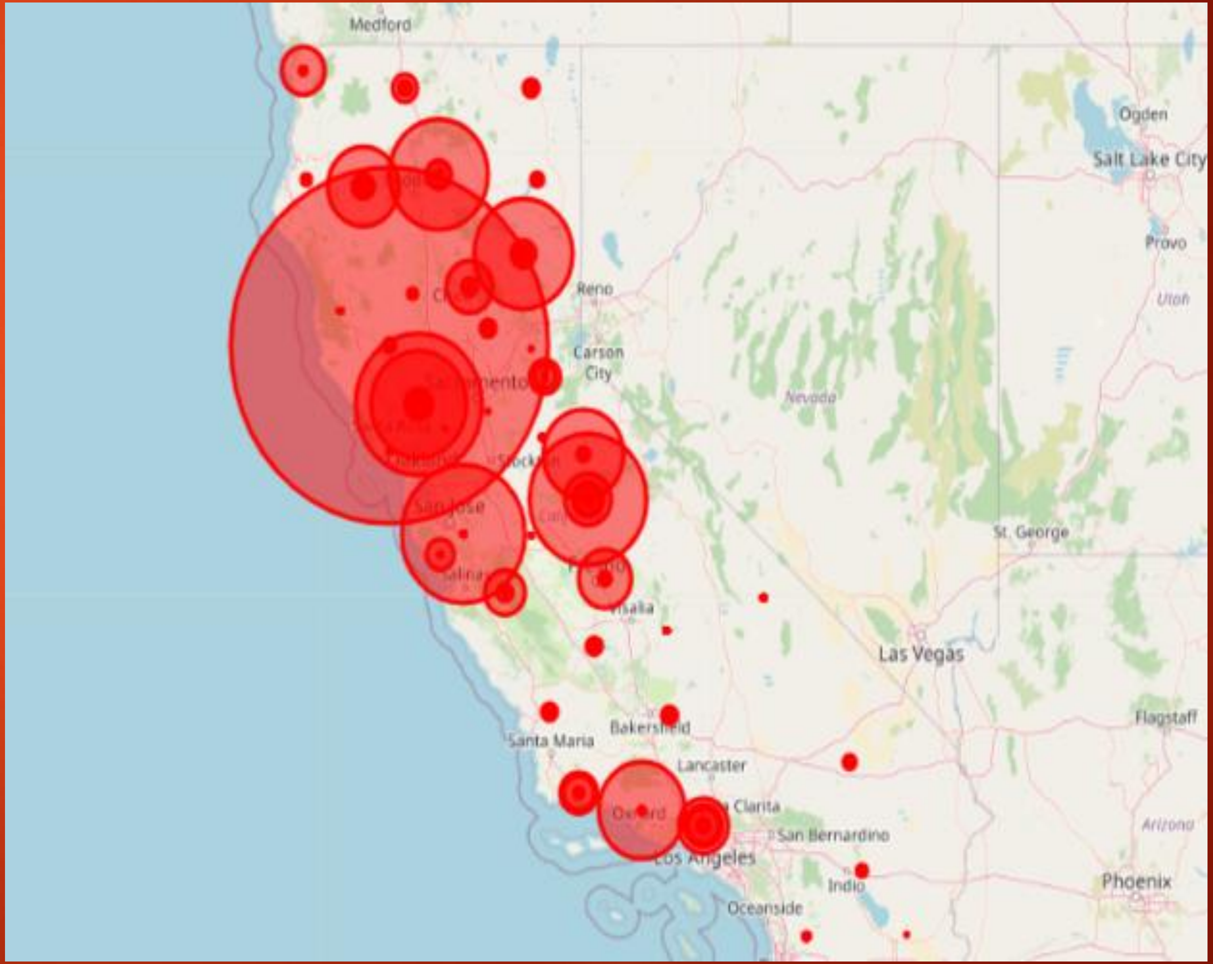



- The map displays a concentration of data points with higher intensities around cities like SF and LA.
- Provide a summary of complex datasets, allowing for immediate identification of hotspots and patterns over geographical locations.
- The scatter plot visualizes the relationship between average temperature and humidity, with the data points colored to indicate whether a fire occurred (red for fire, blue for no fire). It appears that fires tend to occur under a variety of humidity conditions but more frequently at higher temperatures.



# F. Data Mining Tasks

**Data splitting**

Once the data cleaning and preprocessing is done the dataset is divided into a feature matrix, X, and a target vector, y, to prepare for modeling, further, this data is split into training, testing, and validation datasets.

**Model selection**

As we are predicting the presence (binary classification) of wildfires we choose 4 algorithms that are, Random Forest, Logistic Regression, KNN, and Decision Tress and perform techniques on these to choose the best.

**Data Training**

The majority (80%) is used for training and the remainder (20%) is reserved for testing. We split the data into training and testing sets using train_test_split().

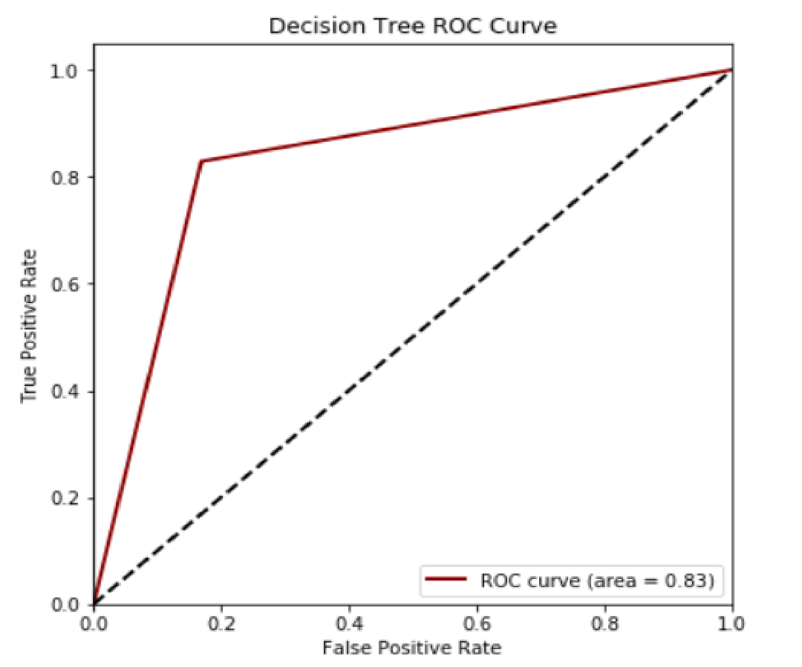
**Model Testing and Evaluation**

We use the classification performance evaluations like the F1 score, Accuracy, Error rate, ROC AUC, and Recall to test which algorithm performs the best after training the model.

**Hyperparameter Tuning and Validation**

The dataset is further tuned to get the best results and is validated accordingly, each of these methods played a role in getting the best model.

# G. Performance Evaluation
# Decision Tree Evaluation




Advantages

- Works well with datasets having mixed data types without the need for extensive preprocessing.
- Capable of dealing with missing data points effectively.
- The tree structure is easy to visualize and understand, aiding in explaining the model's logic.

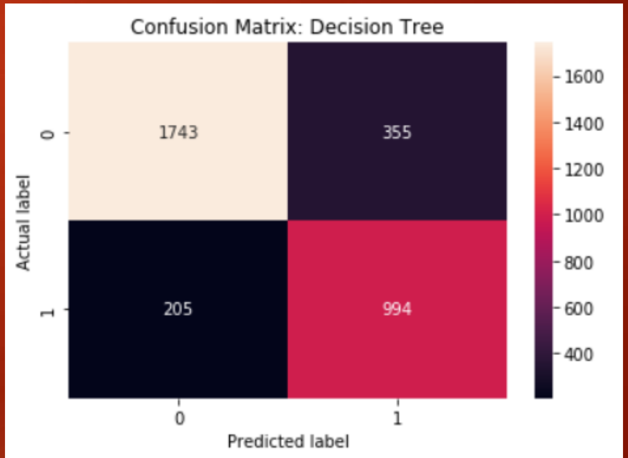
Disadvantages

- Prone to overfitting, especially with complex or deep trees.
- Minor changes in data can significantly alter the tree structure.
- Can become unwieldy with large datasets, leading to long training times and decreased interpretability.

Performance Metrics

- **Accuracy:** 0.8301
- **Precision:** 0.7368
- **Recall:** 0.8290
- **F1 Score:** 0.7802
- **ROC AUC:** 0.8298

- Suitable for datasets with mixed data types without extensive preprocessing.
- Prone to overfitting and sensitive to minor data changes.



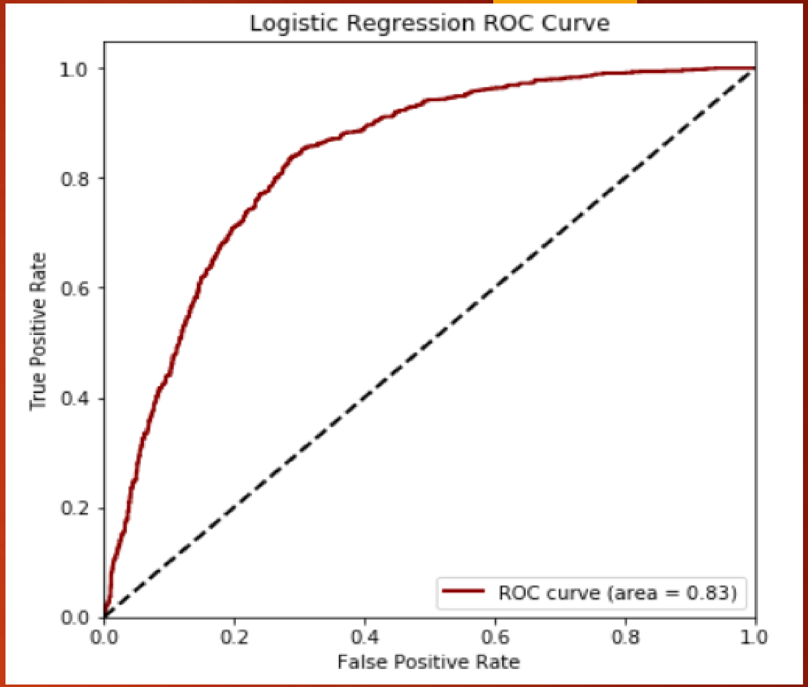
# Logistic Regression Evaluation

Advantages

- Quick to train and can handle large datasets efficiently.
- Provides clear insights into how each feature influences the outcome.
- Techniques like L1 and L2 regularization help in preventing overfitting and improving model generalization.

Disadvantages

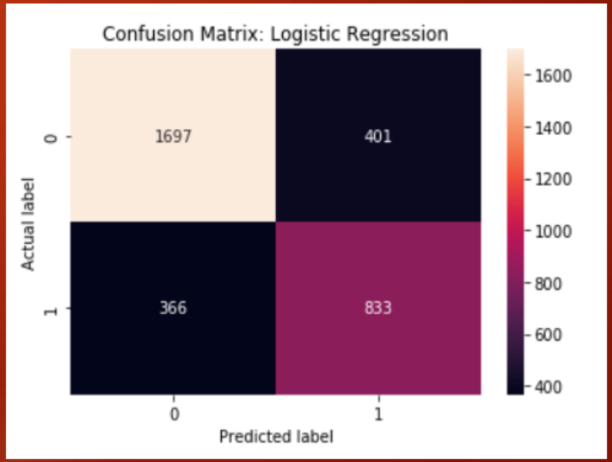
- Assumes a linear relationship between features of the outcome, limiting to capture of complex relationships.
- Outliers can significantly affect model coefficients and predictions, leading to skewed results.




## Performance Metrics

- **Accuracy:** 0.7673
- **Precision:** 0.6750
- **Recall:** 0.6947
- **F1 Score:** 0.6847
- **ROC AUC:** 0.8322

- Efficient for large datasets and provides interpretable coefficients.
- Assumes a linear relationship between features and outcomes, may not capture complex relationships.



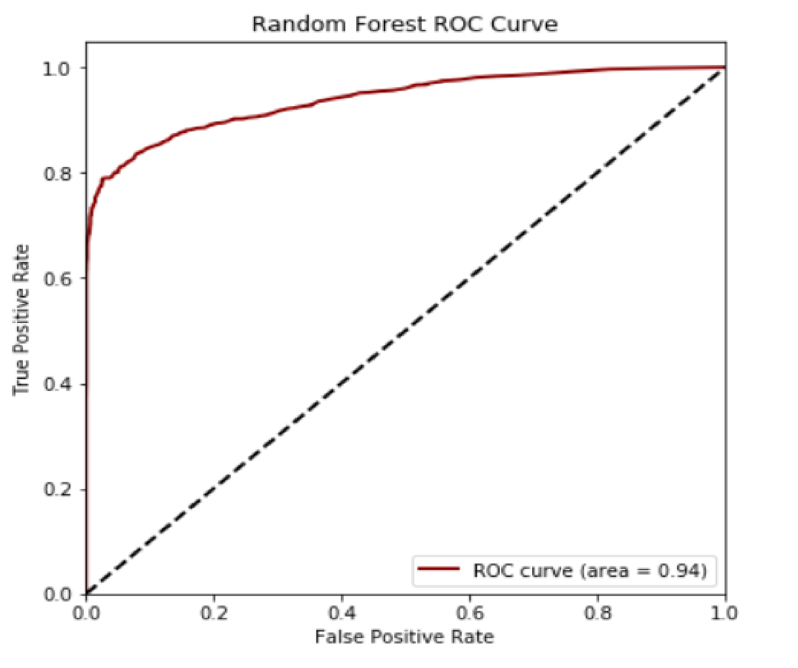
# Random Forest Evaluation

Advantages

- Utilizes trees to minimize overfitting, enhancing model generalization.
- Averages out biases, making it less sensitive to noisy data.
- Capable of dealing with missing values and outliers effectively without extensive preprocessing.

Disadvantages

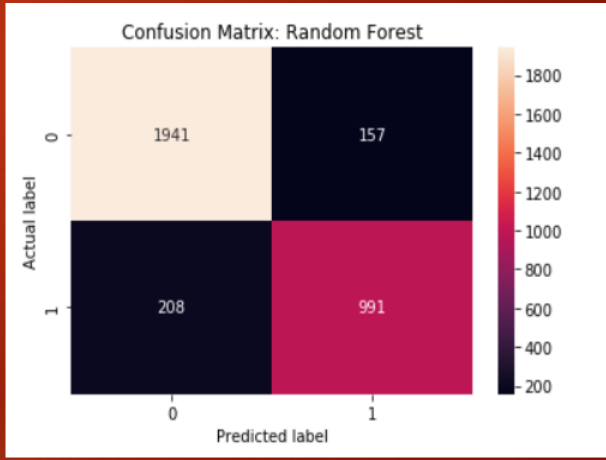
- High accuracy comes with complexity, makes it difficult to interpret.
- Requires substantial computational resources for training multiple trees, especially with large datasets.
- Memory Usage: Increased memory consumption due to the storage of numerous decision trees.




## Performance Metrics

- **Accuracy:** 0.8892
- **Precision:** 0.8632
- **Recall:** 0.8265
- **F1 Score:** 0.8444
- **ROC AUC:** 0.9535

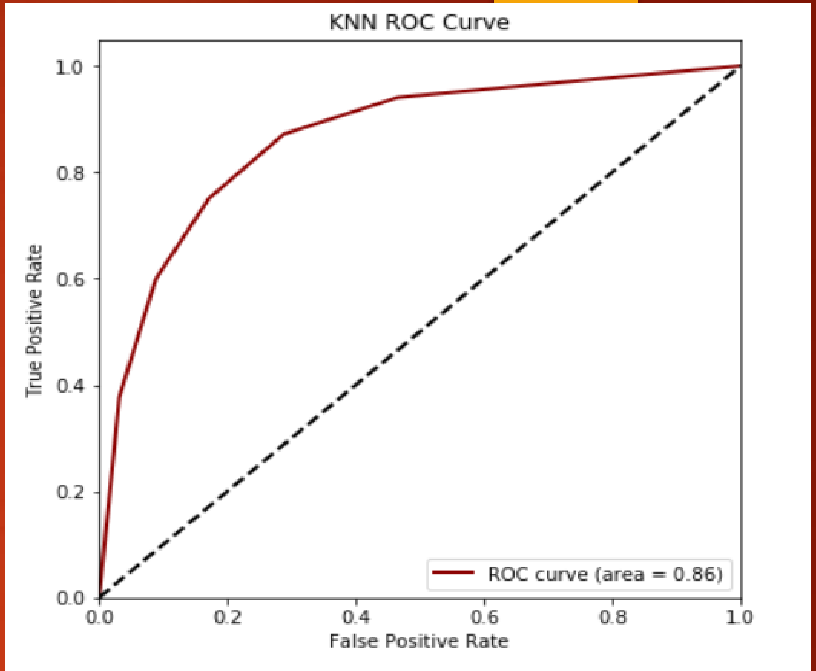
- Higher performance and less prone to overfitting than individual Decision Trees.
- Can handle large datasets with complex structures but is computationally intensive.



# K Nearest Neighbours Evaluation

Advantages
- Easy to understand and implement.
- Works without assumptions about the underlying data distribution.
- Uses stored training instances directly for predictions, enhancing speed and simplicity.
- Applicable for both classification and regression tasks.

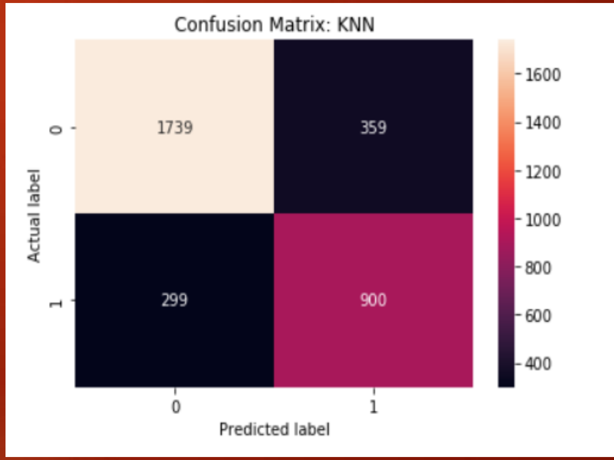
Disadvantages
- Time to find neighbors increases with dataset size, impacting efficiency.
- Performance can degrade with noise, outliers, or irrelevant features.
- Choosing the right K value is crucial and can require experimentation.
- May not perform well on imbalanced datasets, favors the majority class.




Performance Metrics

- **Accuracy:** 0.8004
- **Precision:** 0.7148
- **Recall:** 0.7506
- **F1 Score:** 0.7323
- **ROC AUC:** 0.8638

- Simple and effective, especially when the assumption of linearity does not hold.
- No training phase is required, but computationally expensive with large datasets.



# H. Project Results

Model Comparison

- Introduction to model comparison based on ROC curves and performance metrics.
- Models Evaluated: Decision Tree, Random Forest, Logistic Regression, KNN.
- Key Focus: Accuracy, F1 Score, ROC AUC.

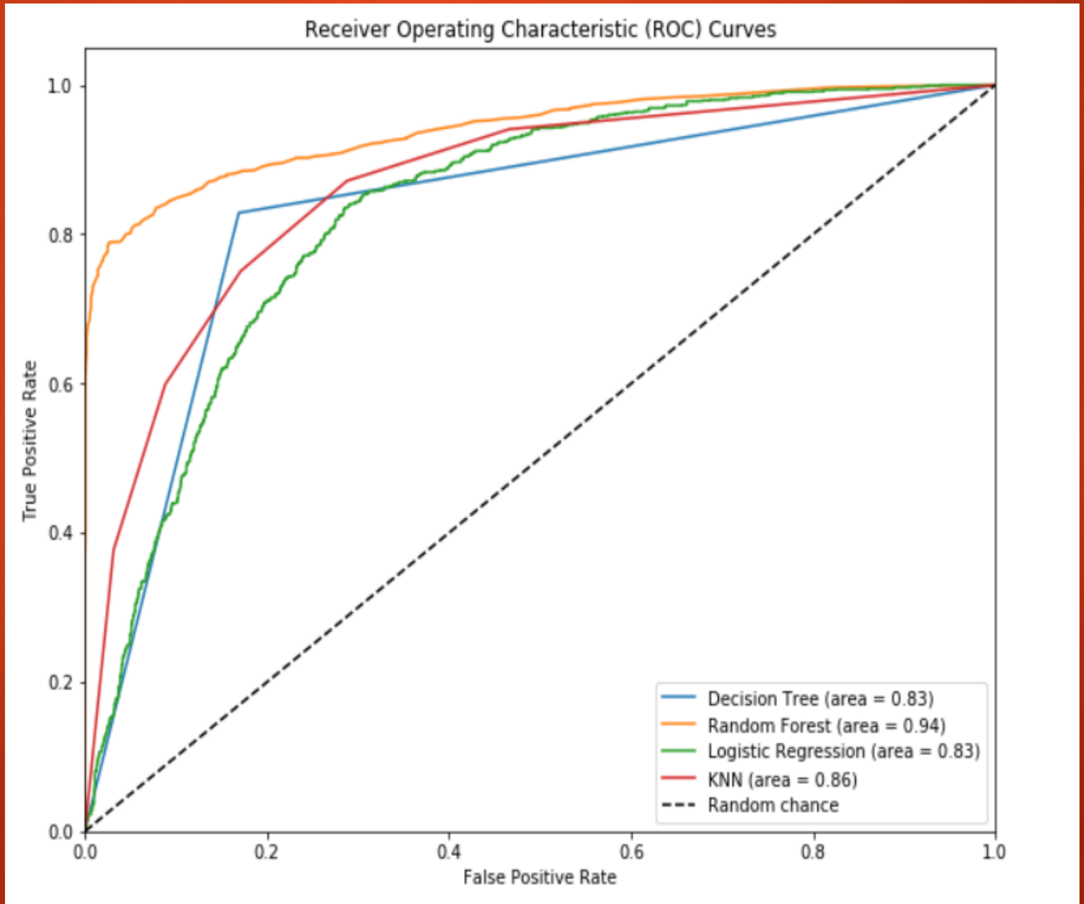
Comparative Insights

- Decision Tree and KNN show comparable accuracy.
- KNN outperforms in F1 Score and ROC AUC, securing second place.
- Logistic Regression trails with lower scores across metrics.

| | Model | Accuracy | Precision | Recall | F1 Score | ROC AUC |
|---|---|---|---|---|---|---|
| 0 | Decision Tree | 0.830149 | 0.736842 | 0.829024 | 0.780220 | 0.829842 |
| 1 | Random Forest | 0.889293 | 0.863240 | 0.826522 | 0.844482 | 0.939592 |
| 2 | Logistic Regression | 0.767364 | 0.675041 | 0.694746 | 0.684751 | 0.832250 |
| 3 | KNN | 0.800425 | 0.714853 | 0.750626 | 0.732303 | 0.863856 |

## Top Performer - Random Forest

- Highest AUC, showcasing superior class distinction.
- Leads in Accuracy and F1 Score.
- Best model for balancing Precision and Recall.
- Its ROC curve stays consistently above others, demonstrating fewer false positives and more true positives.
- This implies better performance in distinguishing between actual wildfires and non-wildfires.



## I. Conclusion

- Successfully developed a predictive model to analyze the presence and trends of wildfires. This was accomplished using data mining techniques and machine learning models.
- The model identified primary causes and geographical patterns affecting wildfires. This includes natural factors like temperature and human-induced factors, thus providing critical insights for managing and mitigating wildfire risks.

## Challenges faced

- The project faced challenges in data cleaning and preparation, including dealing with a large initial dataset that required significant refinement to focus on relevant variables for the model.
- Selecting and tuning appropriate models to accurately predict wildfires presented difficulties. We tested various algorithms ( Random Forest, Logistic Regression, KNN, Decision Trees), each with its own set of challenges, including handling overfitting and ensuring good generalization.
- While models like Random Forest showed high accuracy, they were computationally intensive and less interpretable, which can complicate their implementation in real-world settings where explainability is crucial for decision-makers.

## J. Insights for Decision Making

- The Random Forest model, with its high ROC AUC, has proven to be exceptional in classification tasks, providing reliable performance for predictive purposes.
- These insights allow for the strategic allocation of resources to high-risk areas and the crafting of proactive measures, enhancing both the focus of risk-reduction strategies and the efficacy of resource distribution.

## Impact of Project Outcomes

- Predictive models have bolstered early warning capabilities, facilitating timely evacuation and response, while the integration of new data ensures their continuous refinement to address evolving environmental conditions.
- This advancement aids operational decisions, equipping wildfire management to better prepare for the diverse challenges presented by varying geographic and environmental scenarios.

# THANK YOU