Analysis & Results

Name : Pooja Patel

Overview :

Wildfires in the United States have become increasingly frequent and destructive, posing significant threats to communities, ecosystems, and the economy. Understanding the main reason (causes) of wildfires is essential for effective prevention.

Objective : The main objective of this project is to analyze the historical wildfire data (FPA dataset) between 1992 to 2020 and identify the primary causes of wildfires in the USA.

Limitations :

There are a few limitations to consider in this wildfire analysis project:

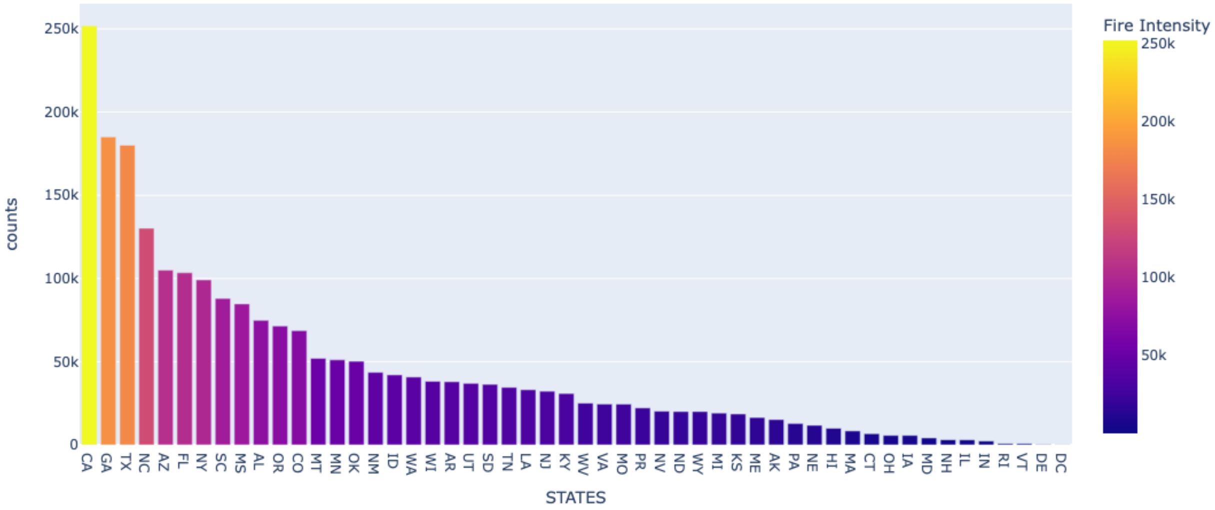
- The available FPA dataset may not encompass all factors influencing wildfires. Variables such as changes in land use or population density might not be adequately captured in the dataset.
- External factors, including climate change, policy alterations, or socioeconomic dynamics, can impact wildfire causes. These factors may not be fully captured in the FPA dataset, potentially affecting the identification of the primary cause of wildfires.
- The patterns and causes of wildfires are subject to change over time. While my analysis covers the period up to 2020, it is important to acknowledge the evolving nature of wildfire dynamics and consider the relevance of historical patterns to current and future scenarios.

Analysis :

Exploring Wildfire :

Before diving into machine learning for the FPA(wildfire) dataset, it's essential to do Exploratory Data Analysis (EDA). By analyzing these relationships, we gain insights that help interpret the results later on. This section focuses on examining how data is distributed and illustrating the influence of wildfires across the country.

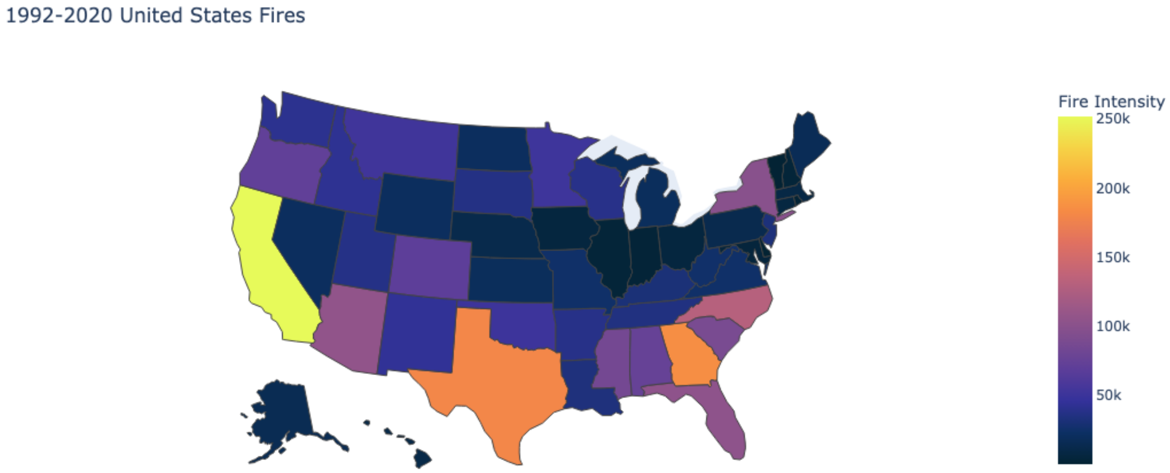
Fire Prone States :



(Figure 1 : Fire Intensity in USA)

Looking at the top 10 most fire prone states of the USA in figure 1, 6/10 are from southeastern states. And most Fire Prone states are the District of Columbia (DC), Delaware (DE) and Vermont (VT).

Wildfire Locations :



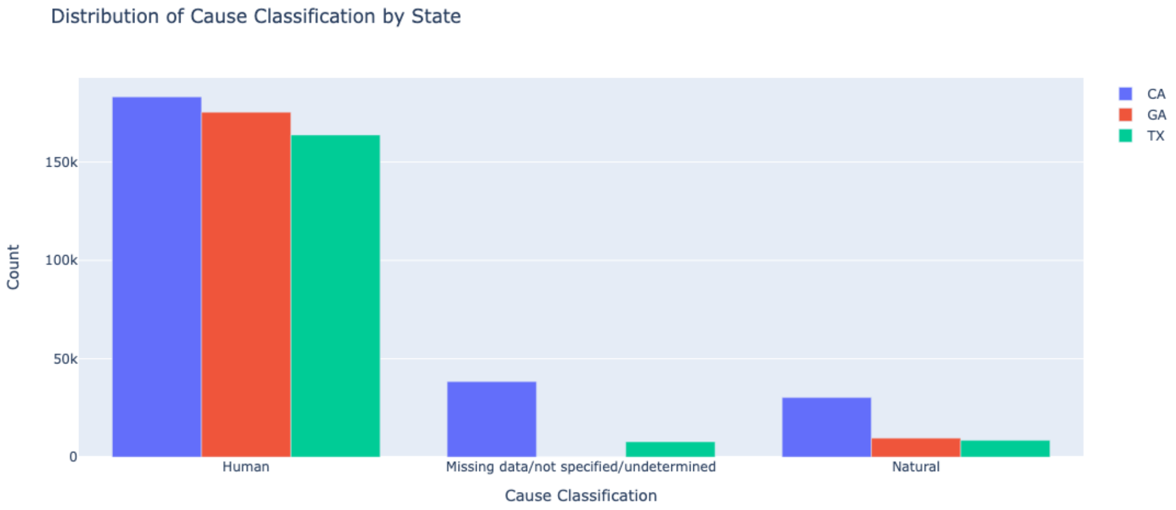
(Figure 2 : Fire Intensity in USA MAP)

After looking into data from all USA states (Figure 1 and Figure 2), it's evident that California, Texas, and Georgia face the most wildfire impact. California tops the list with a Fire Intensity of 251.88k, followed by Georgia at 185.04k and Texas at 180.08k.

Further Analysis with most wildfire impacted states :

Since California, Texas, and Georgia states stand out for their high wildfire intensities, I've chosen to focus my analysis on them to understand the main reasons behind the wildfires.

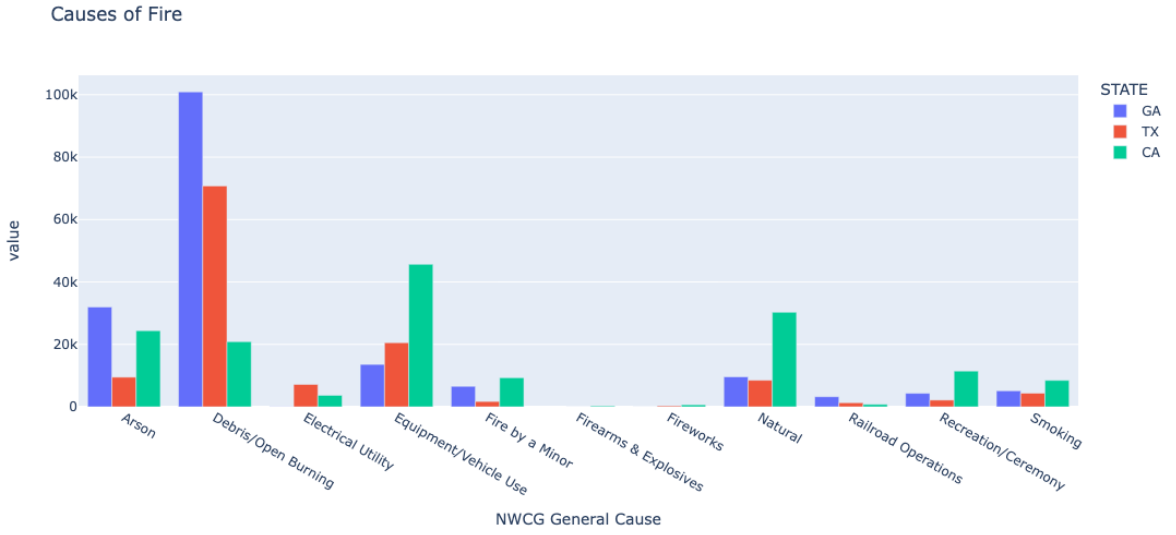
Overall Causes Classification of Wildfire :



(Figure 3 : Cause Classification)

Across all three states(CA, TX and GA), Human activity is the main contributing factor to the wildfires. That's why, I have decided to further analyze the primary cause within 3 states.

Causes of Fire :



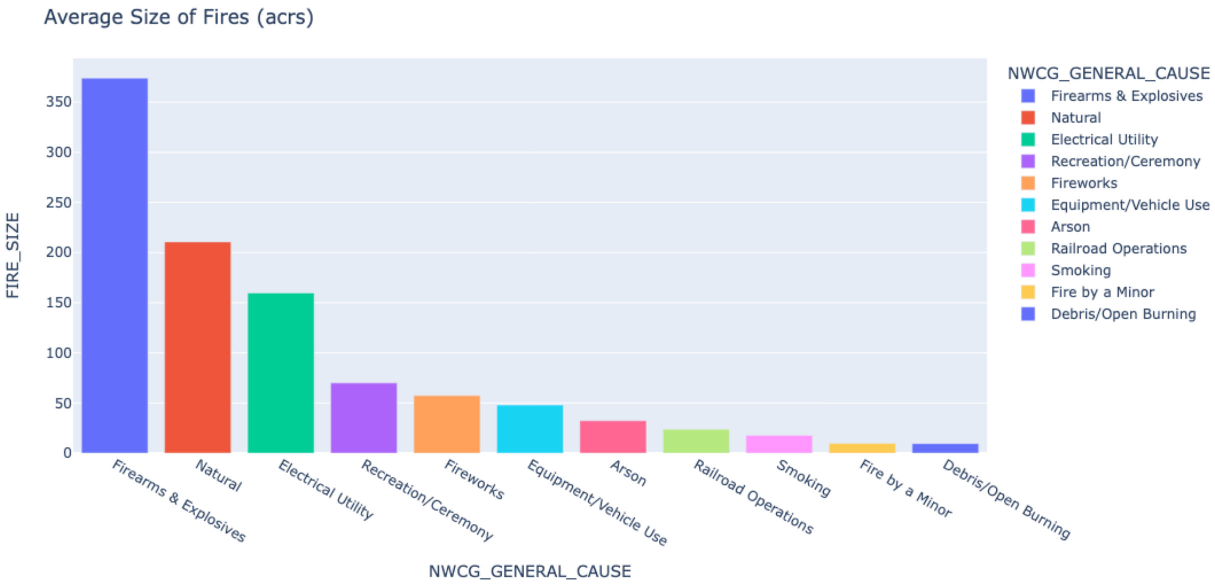
(Figure 4 : Causes of fire)

Let's understand the meaning of the cause.

- Arson : person deliberately set a fire.
- Debris/Open Burning : Vegetation, dead plants, and other organic materials are on the forest floor. These ground materials on the ground acts like fuel for the fire, making it spread. This fire can start due to lighting or human activity.
- Electrical utility : Use electrical power lines which cause a fire.
- Natural : During the storm, lightning strikes a tree, creating a spark. This spark ignites a fire.
- Recreation/Ceremony : Recreation means Campfire and Ceremony means Certain cultural or religious ceremonies involve the use of fire.

As per the analysis, I have found out that Equipment/Vehicle Use, Debris/Open burning and Arson are the most frequent causes of wildfire.

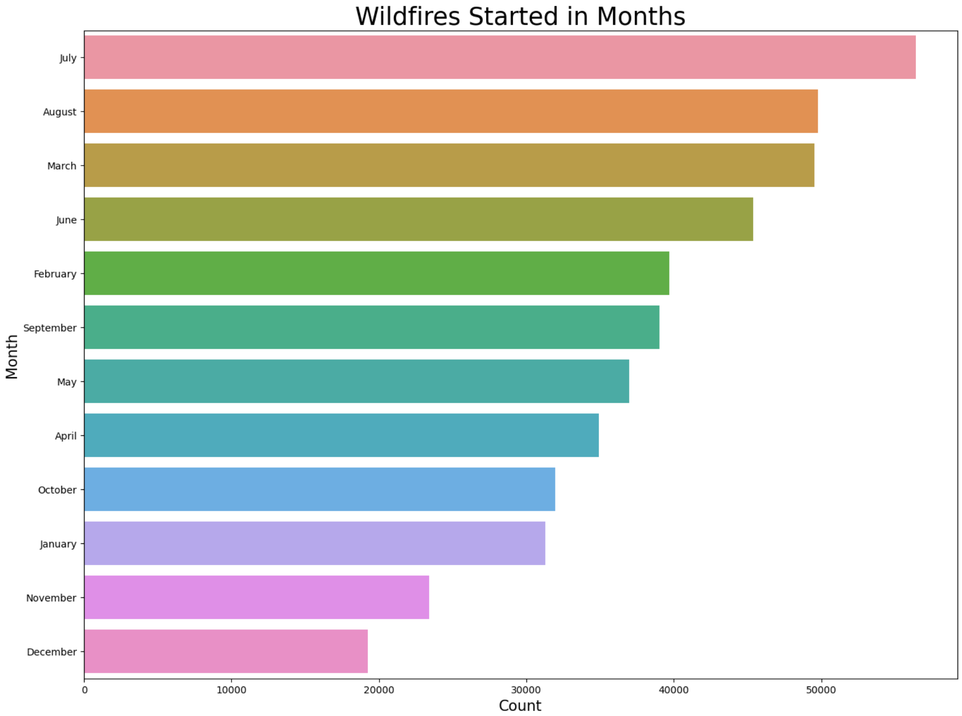
Average Size of Fires (Arcs)



(Figure 5 : Avg Size of Fires by Cause)

The average size of fires in arcs is influenced by various causes. Factors such as the use of firearms and explosives, natural events, and electrical utilities, including power lines, contribute significantly to larger fires.

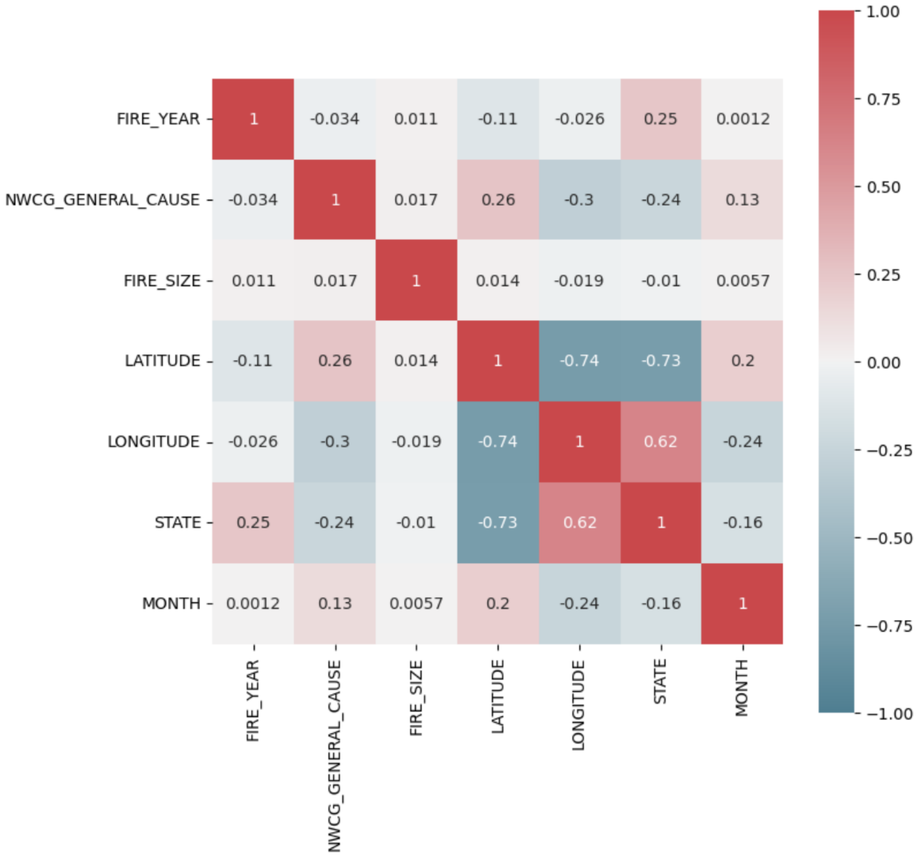
Month analysis of Fire :



(Figure 6 : Wildfire Begins in Months)

The majority of fires happen from late spring to summer each year, and there are a couple of reasons for this. Firstly, the higher temperatures and drier conditions make it easier for fires, especially those caused by 'debris burning.' Secondly, as people spend more time outdoors during these seasons, the likelihood of accidental fires(campfires, Equipment/ Vehicle use etc) increases.

Correlation Matrix :



When analyzing this correlation matrix, there are no features that have strong correlations, or at least correlations greater than 0.3. It's interesting to note that the fire year has a slight correlation with STATE. It's clear that FIRE YEAR should relate to STATE which is 0.25, but it's nice to see that represented visually. Also, GENERAL CAUSE has correlation with STATE and LONGITUDE & LATITUDE.

Based on the above analysis of different features, let's move to Machine Learning Models.

Results :

Machine Learning :

There are various ways that can be categorized into three main subsections of machine learning: supervised learning, unsupervised learning and reinforcement learning. Here, we will be using supervised learning. A supervised learning algorithm takes labeled data and creates a model that can make predictions given new data.

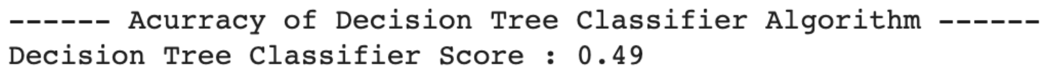
For this project, predicting the cause of wildfire is a Supervised machine learning problem. I would like to try to predict the causes of wildfires in the highest fire intensity states which are California, Georgia and Texas. In doing so, this could help response units pinpoint the causation and understand whether a fire was started through natural causes or malicious human action. To accomplish this, I will need to predict the cause of the fire using the target variable ‘NWCG_GENERAL_CAUSE’ which includes several categorical causes.

Firstly, I have converted all the categorical features into numerical labels so that machine learning models can interpret the data and identify how the fire began. Since I'll be using supervised learning, I have divided the dataset into training and testing sets. Initially, I have decided to use a Logistic Regression algorithm and Support Vector Machine. Since it's a multi classification problem, both Logistic Regression and Support Vector Machine work on binary classification problems. Thus, I have decided to use classification algorithms like Decision Trees, Random Forest and XGBoost which can perform well on multi classification problems.

Method 1 : Default ML Models

Decision Tree :

A decision tree is a predictive model that organizes information into a tree-like structure. It makes decisions by recursively breaking down a dataset into subsets based on the most influential features, creating a branching structure. Each leaf node in the tree represents a predicted outcome. Decision trees are known for their interpretability and ability to handle both classification and regression tasks.

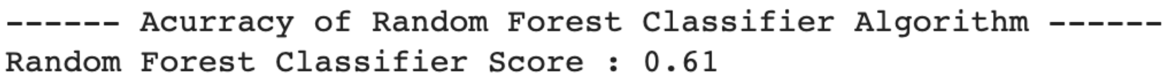


```
------ Acurracy of Decision Tree Classifier Algorithm ------
Decision Tree Classifier Score : 0.49
```

Here, I have used a decision tree on a wildfire dataset and the accuracy of the decision tree algorithm is around 0.49

Random Forest :

Random Forest performs well compared to the Decision Tree algorithm. Because in decision trees only trees work. Random forest, as its name suggests, is like having a bunch of decision trees working together. Each tree is trained on a random subset of the data, and they collectively vote on predictions. By combining diverse tree predictions, Random Forest enhances accuracy and generalization while mitigating overfitting.



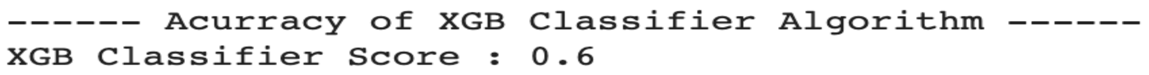
```
------ Acurracy of Random Forest Classifier Algorithm ------
Random Forest Classifier Score : 0.61
```

Here, I have used the default Random Forest algorithm in which all the default parameters are used. And The accuracy of Random Forest is around 0.6 which is better than the decision tree algorithm.

XGBoost :

XGBoost is an ensemble learning algorithm that combines the predictions of multiple weak learners(decision trees) to create a strong learner. The XGBoost algorithm helps to control the complexity of the individual trees and prevent overfitting.

Here, I have used the XGBoost algorithm in which all the default parameters are used. And the accuracy of the XGBoost algorithm is around 0.60



```
------ Acurracy of XGB Classifier Algorithm ------
XGB Classifier Score : 0.6
```

As per earlier discussion, predicting the cause of wildfire is a multi-classification problem, and we have around 12 different cause (categories) classes. That's why machine learning models are not able to perform well.

Method 2 : Improved Machine Learning Models

The Machine Learning models encounter difficulties distinguishing between numerous classes, with getting very low accuracy results. To refine model performance, I plan to consolidate these classes and reevaluate the results.

I have combined some cause Label from 12 cause to 4 cause labels (Target variable) :

Label 0 = Natural

- Natural

Label 1 = Accidentally (Accidentally refers to actions which spread a wildfire due to carelessness, recklessness, or failure to take reasonable precautions.)

- Recreation/Ceremony :- Recreation : Campfire, Ceremony : Certain cultural or religious ceremonies involve the use of fire.
- Railroad Operations
- Fireworks

- Smoking
- Fire by a Minor
- Equipment/Vehicle Use
- Debris/Open Burning

Label 2 = Arson

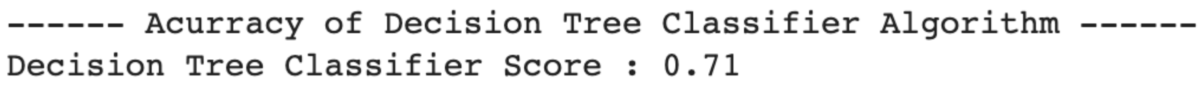
- Arson

Label 3 = Other

- Firearms & Explosives
- Missing/Undefined,
- Miscellaneous

Decision Tree Improved version

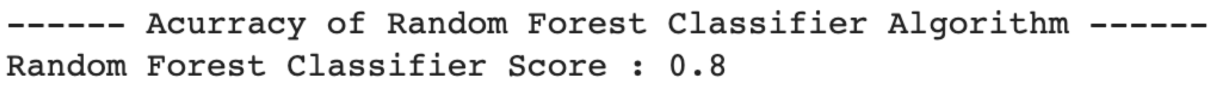
Here, the Decision Tree model has been applied to categorize data originally labeled with 12 cause categories into 4 distinct cause labels. And the model achieves an accuracy of 0.71 which is better than the previous decision tree model.



```
------ Acurracy of Decision Tree Classifier Algorithm ------
Decision Tree Classifier Score : 0.71
```

Random forest Improved version

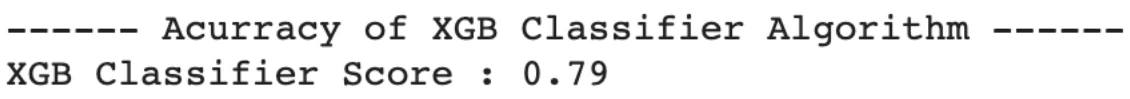
Here, I have used the default Random Forest algorithm in which all the default parameters are used. But Random Forest is classified from 12 cause category into 4 cause category. And The accuracy of Random Forest is around 0.8 which is better than previous the Random forest algorithm(accuracy 0.6).



```
------ Acurracy of Random Forest Classifier Algorithm ------
Random Forest Classifier Score : 0.8
```

XGBoost Improved version

Here, I have used the default XGBoost algorithm in which all the default parameters are used. The improved XGBoost algorithm is classified from 12 cause categories into 4 cause categories. And The accuracy of XGBoost is around 0.79 which is better than the previous XGBoost algorithm(accuracy 0.6).



```
------ Acurracy of XGB Classifier Algorithm ------
XGB Classifier Score : 0.79
```

Hyperparameter Tuning on improved ML models :

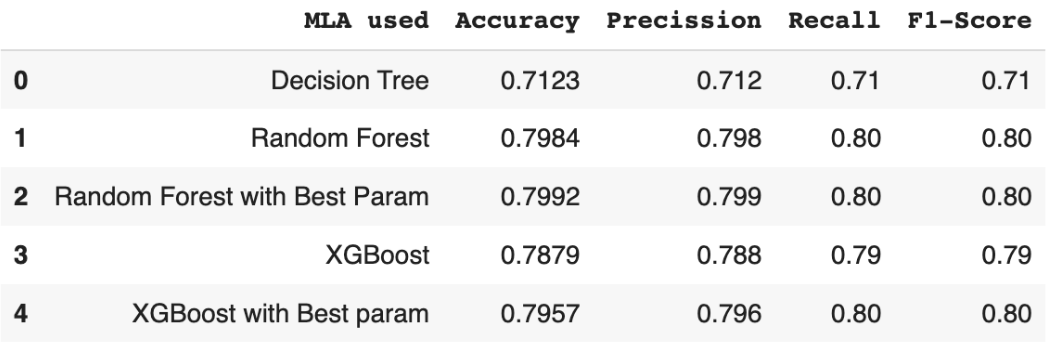
Hyperparameter tuning is the process of finding the optimal configuration for a machine learning model's hyperparameters. And the goal is to find the combination of hyperparameter values that results in the best performance on a testing dataset.
As per the above improved machine learning models method, Random Forest and XGBoost algorithm performs well on default parameters. Therefore, the focus shifts to discovering the optimal combination of hyperparameter values for these algorithms, aiming for the highest possible results and a robust model.

The GridSearchCV method is used for hyperparameter tuning and finding out the best parameters. And now I am comparing all the default ML models and hyperparameter tuned models.

Comparing ML Models :

When comparing all the default parameters and hyperparameter tuned models, and evaluate each model's performance in unseen data using evaluation metrics like Accuracy, Precision, Recall and F1-Score. For identifying the cause of wildfire which is a multi classification problem, the most important metric which is considered are Accuracy and Recall. Accuracy measures overall performance of the model and Recall metric measures the correctly identified instances of specific class.

Random Forest with Best Parameters performing with highest accuracy which is 0.7992 and highest Recall 0.80 Let's check the confusion metric for the best algorithm.

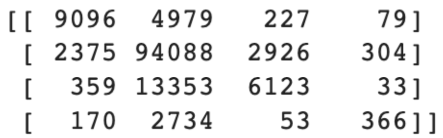


	MLA used	Accuracy	Precission	Recall	F1-Score
0	Decision Tree	0.7123	0.712	0.71	0.71
1	Random Forest	0.7984	0.798	0.80	0.80
2	Random Forest with Best Param	0.7992	0.799	0.80	0.80
3	XGBoost	0.7879	0.788	0.79	0.79
4	XGBoost with Best param	0.7957	0.796	0.80	0.80

Confusion Matrix :

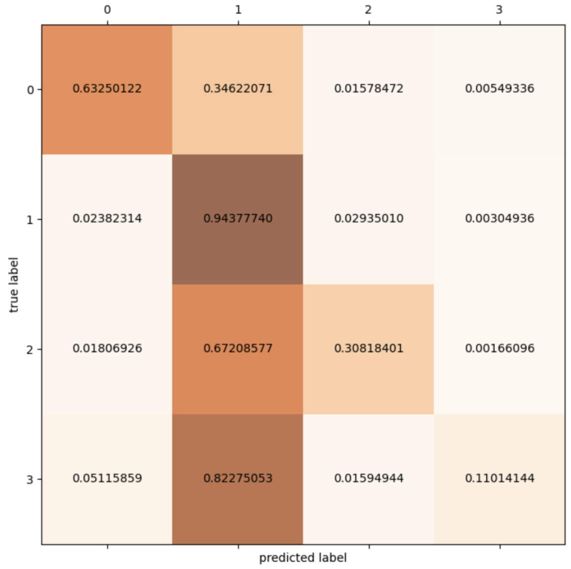
Confusion Matrix evaluates the performance of a classification algorithm on a set of data for which the true values are known. Below is the confusion matrix for predicting 4 different cause labels such as Natural, Accidentally, Arson and Other.

Confusion Matrix :



```
[[ 9096  4979   227    79]
 [ 2375 94088  2926   304]
 [  359 13353  6123    33]
 [  170  2734    53   366]]
```

Actual vs predicted label in Percentage :



Let's review the labels:

- Label 0 = Natural,   Label 1 = Accidentally, Label 2 = ARson, Label 3 = Other

- The model achieved a 63% accuracy when predicting natural causes.
- The model achieved a 94% accuracy when predicting Accidental causes.
- The model achieved a 31% accuracy when predicting Arson.
- The model achieved a 11% accuracy when predicting others.

Conclusion :

The aim of this project was to explore and predict the cause of wildfire in the United states from 1992 to 2020. A database of this project was originally generated by the national Fire Program Analysis (FPA) system. This project majorly focuses on visual explorations of the dataset.

The analysis shows the distribution of data and demonstrates the impact of wildfires across the country. Later on analyze the most wildfire impacted states California, Texas and georgia. Wildfires were analyzed on locations like most wildfire impacted states, least wildfire impacted states, month of the year, average fire size per cause, cause classification of wildfire and many more characteristics.

And predicted causes of wildfire with the overall accuracy is around 60%.
Thereafter, reducing the number of labels from twelve into four categories improved the prediction score to 80% for the random forest algorithm. As per the confusion matrix, unfortunately, the algorithm did not perform well when trying to distinguish between 'Arson' and 'Other'' causes. The algorithm did perform relatively well, however, in distinguishing natural causes and Accidental causes.

Recommendations :
To emphasize the practical value of this project, fire departments and respective agencies could use the resulting models to predict future wildfire causes in California, Texas and Georgia. The following recommendations would be used by fire reporting units and agencies to allocate resources appropriately to fight future fires.

- **Check and replace defective Electrical Utility** infrastructure like power lines which has contributed to the third-largest average size of fires, as illustrated in figure 5. Although these Electrical Utility related incidents are less frequent when analyzing total causes of fires, they have a significant impact, leading to some of the largest fires. Therefore, enhancing and upgrading electrical utility systems is crucial for reducing the severity of wildfires.
- **Increase awareness**, **particularly during the peak months** highlighted in figures 3 and 6. Human activities stand out as the primary cause of wildfires, with a significant surge in incidents from late spring to summer each year. As people spend more time outdoors during these seasons, the likelihood of accidental fires(campfires, Equipment/ Vehicle use etc) increases. So, fire departments and respective agencies should provide comprehensive guidelines outlining precautions for individuals engaging in activities such as campfires, smoking, and the use of specific equipment and vehicles.

Future Work :

In future consider expanding the dataset to include more comprehensive information on factors influencing wildfires. Also, refine and improve predictive models by exploring advanced machine learning techniques. Plus, I am planning to do from offline analysis to real-time deployment using Flask or Streamlit, allowing the model to provide timely predictions and insights.