Notes

Slide 1 - Outline - In this project I will give you the brief introduction of the dataset which I have imported from the Kaggle - 1.88 million us wildfires occurred in United States.

Slide2 - Wildfire is unwanted and unplanned fires in the forest

In this project, I will show you how to analyze the dataset, visualize the data and implement various Machine Learning Algorithms to classify the causes of wildfires. The whole model will take the fire part of this dataset and provide percentage ratings for all possible reasons that could cause wildfires.

Slide 3 - According to Kaggle, this data publication contains a spatial database of wildfires that occurred in the United States from 1992 to 2015

Slide 4- This dataset has a huge amount of data, I have focus on the table named “fire” since it contains the causes and other information of wildfires.

Slide 6 - I have implemented sqlite3 to import the dataset. This are the columns which we want for our model.

Slide 7 dataset content- Fire code stands for; Fire Name stands for

Slide 8- I am using 3 algorithms gaussian, decision tree and random forest to predict the causes of wildfires. I will show the implementation and explanation in the next slide.

Slide 9 - imported data from Kaggle and show head, I have selected the Fire Year, State, Fire\_Size, Longitude, Latitude, Day of the week, Month fields as attributes. And Remove null values - df.isna()

Slide 10- Number of wildfires per year •I have created a bar plot of fires per year.

•In 2006 maximum incidents of wildfires took place.

•Around 10,000 - 15,000 incidents of wildfire take place every year.

Slide 11- Number of fires for each class per year

We have column name fire size class, So I have plotted the wildfire occurrences based on the fire size class. To understand the spread of wildfire each year. Wildfire occurrence is more than 5000+ acres.

Slides -12 but what cause the high probability of fires in the United States is the question. after visualizing total fires per state, I found that CA, TX and GA were the top three states in United States, so for I decided to use this data to predict the causes of wildfires.

Slide 13 - Number of wildfires with causes in California

Slide 14 - Number of wildfires with causes in Texas

Slide 15- As we can see wildfire is initiated because of the debris and arson. There is less chance of initiating wildfire because of the natural causes.

Slide 16 - Correlation

•So next I have transferred the states and causes to numeric numbers.

•Tried to calculate the correlations but found out the dataset didn’t show a strong correlation.

•It is difficult to predict the fire causes, I used ML algorithms and implemented on the dataset to analyze and predict wildfires.

Slide 17- Naive Bayes is a statistical classification technique based on Bayes Theorem. It is one of the simplest supervised learning algorithms. Naive Bayes classifier is the fast, accurate and reliable algorithm. Naive Bayes classifiers have high accuracy and speed on large datasets.

Slide 18 - Imported the Gaussian bales model, created a Gaussian Classifier.

Trained the model and predicted the response for the test dataset. And finally came to an accuracy of around 25%, which is not good. So, moved to the next algorithm which is decision tree

Slide 19 - **Decision Trees (DTs)** are a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. A tree can be seen as a piecewise constant approximation.

Slide 20- Collected the data required and converted data to numeric numbers. The trained the data came out with a score of 33%. It still wasn’t what we want.

Slide 21

Random forests is a supervised learning algorithm. It can be used both for classification and regression. It is also the most flexible and easy to use algorithm. A forest is comprised of trees. It is said that the more trees it has, the more robust a forest is.

Slide 22- Imported the random forest classifier. Train and test accuracy for Random Forest

and got 58% accuracy which is better than the previous algorithms.

But in this project, I wanted to check the causes of wildfire so after getting the 58 % accuracy I have decide to divide the causes in 4 classes

natural- Lightning

accidental Structure','Fireworks','Railroad','Smoking','Children','Campfire','Equipment-Use','Debris Burning

malicious - Arson

Other - Missing/Undefined

Slide 23- After reducing the number of attributes model gave the accuracy 70% using random forest. But the accuracy is not good for malicious part which is Arson.

The previous model has the US wildfires and try to classify the causes into four classes. So, I have decided to perform random forest algorithm on single state with one fire cause, malicious.

Next step is to predict malicious fires in top three state. CA, GA and TX. In order to improve the accuracy, drop some columns to do that.9

Slide 24- The model did badly at the malicious part, so in the next step, I did a more precise work by focusing on one state to determine whether the cause of wildfires in this state is malicious. And finally got a 94% predicted score.

Slide 25 - Conclusion is that Random Forest is the best algorithm which gave accuracy of 94 % which is a great to predict the causes of wildfires in United States.