Wildfire Cause Prediction

1_Nov15_7_ P24

CSC522 Project

- Ganesh Thanu (gthanu)
- Arun Gaonkar (agaonka)
- Nikhil Prashant Patil (npatil)

Contents :

1. Problem / Motivation
2. Proposed Solution
3. Dataset
4. Exploratory Data Analysis
5. Classification using Machine Learning methods
6. Classification using Deep Learning methods
7. Conclusion

Problem/ Motivation :

- ❑ Each year, around 60,000 wildfires can be seen in US which cause impeccable damage.
- ❑ Hotter, drier weather caused by climate change or natural events like lightning can cause a fire in the forest.
- ❑ Human activities like smoking, unattended campfires, uncontrolled fireworks are also a major contributors to wildfire.
- ❑ Due to this, Forest fire analysis and cause prediction measures have become increasingly important.

Proposed Solution :

- ❑ Building Machine Learning model that leverage historical wildfire data to predict the cause.
- ❑ Proposed solution will help to:
  - ❑ Predict future forest fire risks.
  - ❑ Recommend forest fire monitoring procedures.
  - ❑ Activate prevention and security measures.

Dataset :

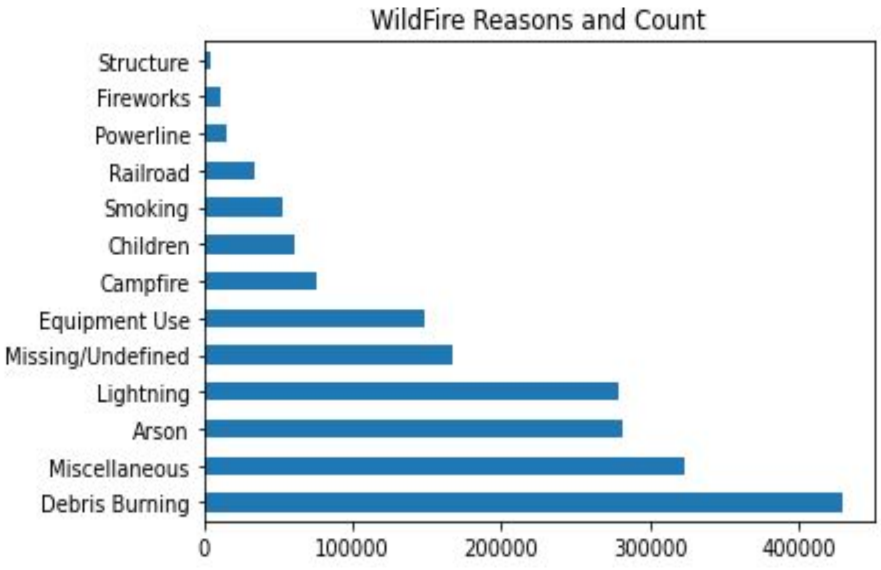
- ❏ Dataset is an SQLite database that contains the information as shown in the table.
- ❏ The Fire Program Analysis-Occurrence Database (FPA-FOD) contains geospatial records of wildfires that occurred in the USA from 1992 to 2015.
- ❏ Data consists of 1.88 million records and 14 columns.

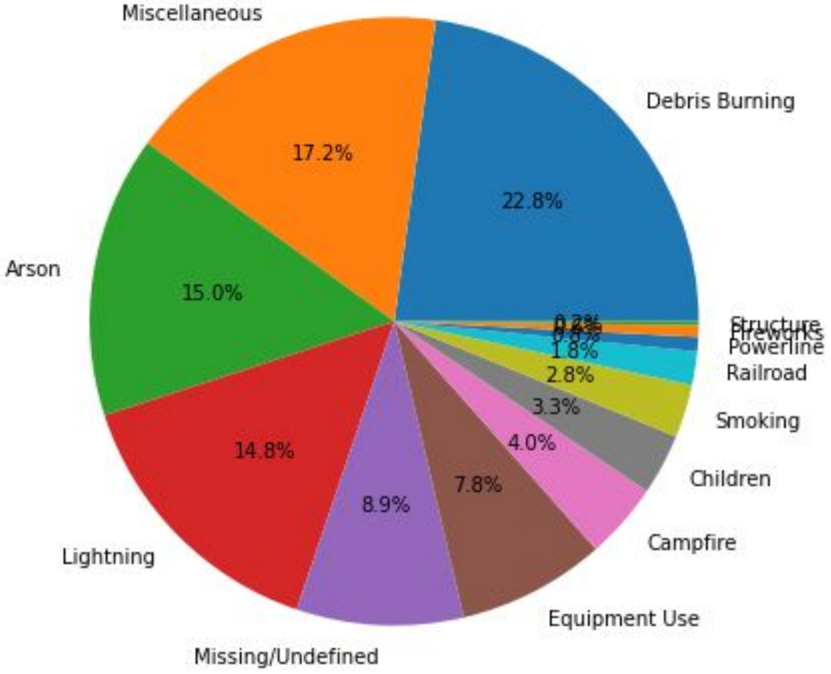
Source: https://www.kaggle.com/rtatman/188-million-us-wildfires

Attribute	**Classification**	
FOD_ID	Nominal	Discrete
FIRE_YEAR	Interval	Discrete
MONTH	Ordinal	Discrete
START_DATE	Interval	Discrete
DISCOVERY_TIME	Interval	Continuous
END_DATE	Interval	Discrete
CONT_TIME	Interval	Continuous
FIRE_SIZE	Ratio	Continuous
FIRE_SIZE_CLASS	Ordinal	Discrete
LATITUDE	Interval	Continuous
LONGITUDE	Interval	Continuous
STATE	Nominal	Discrete
CAUSE	Nominal	Discrete

Exploratory Data Analysis

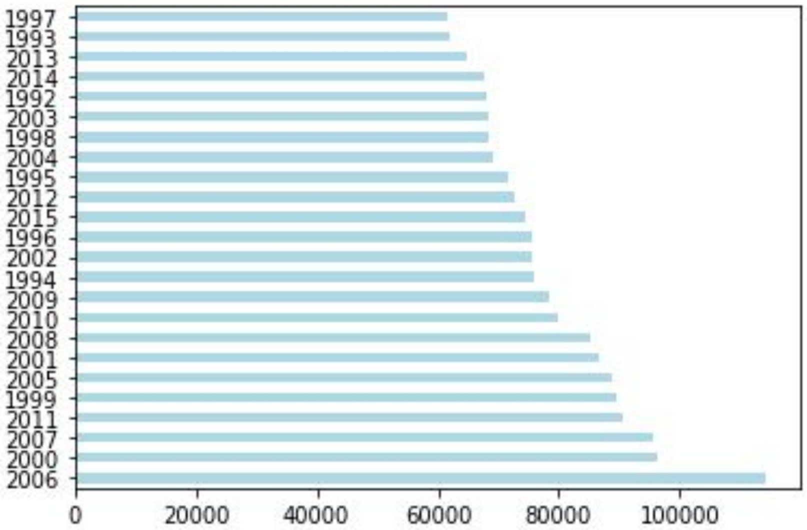
Exploratory Data Analysis:





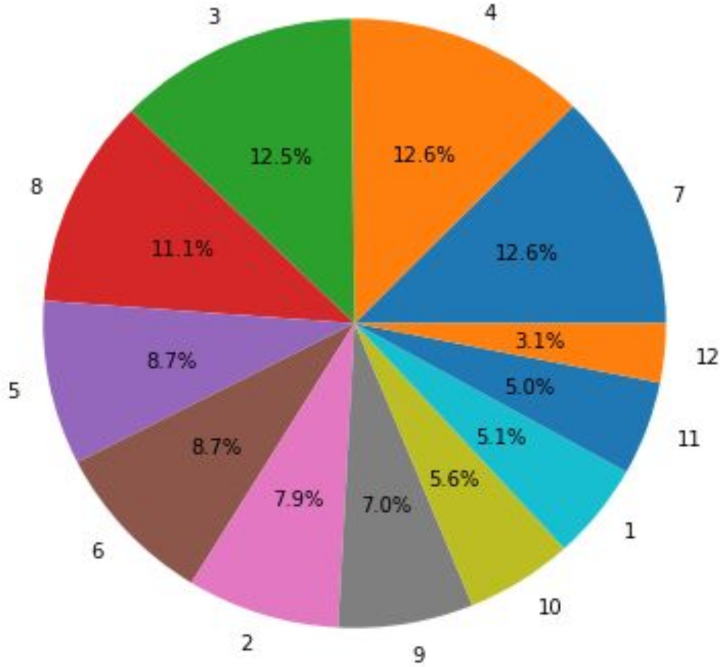
Debris burning, Arson, Lightning are the major causes of wildfire accounting for more than 50% of total wildfires.

Exploratory Data Analysis:



Year wise wildfire count

There is no significant difference in year on year wildfire count.

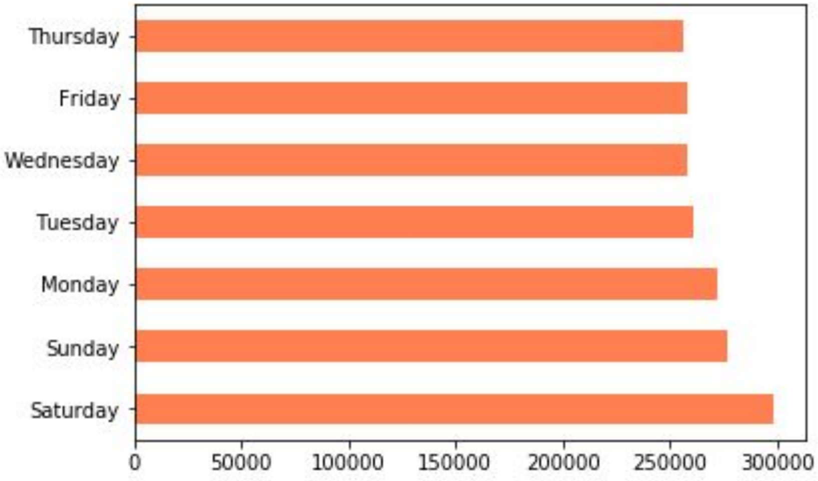


Month wise wildfire distribution

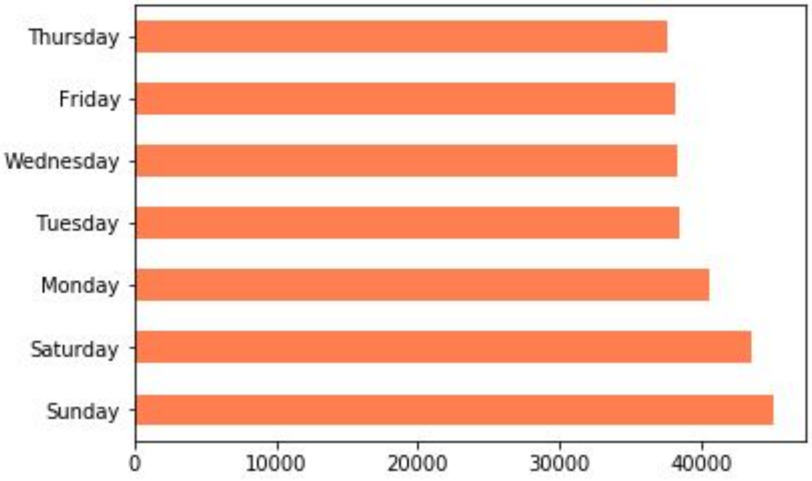
There is significant difference between wildfire count from October - January and February - September. It might be because of weather, since weather is cold from October - January which is not conducive for wildfire.

Exploratory Data Analysis:

For Arson category



Day wise wildfire count

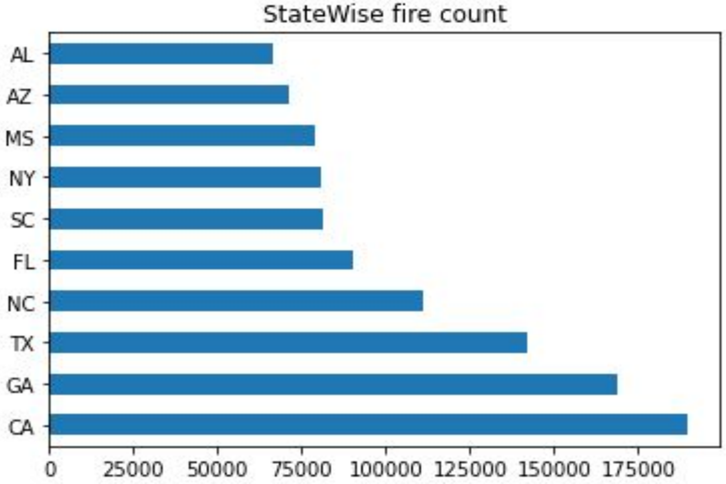


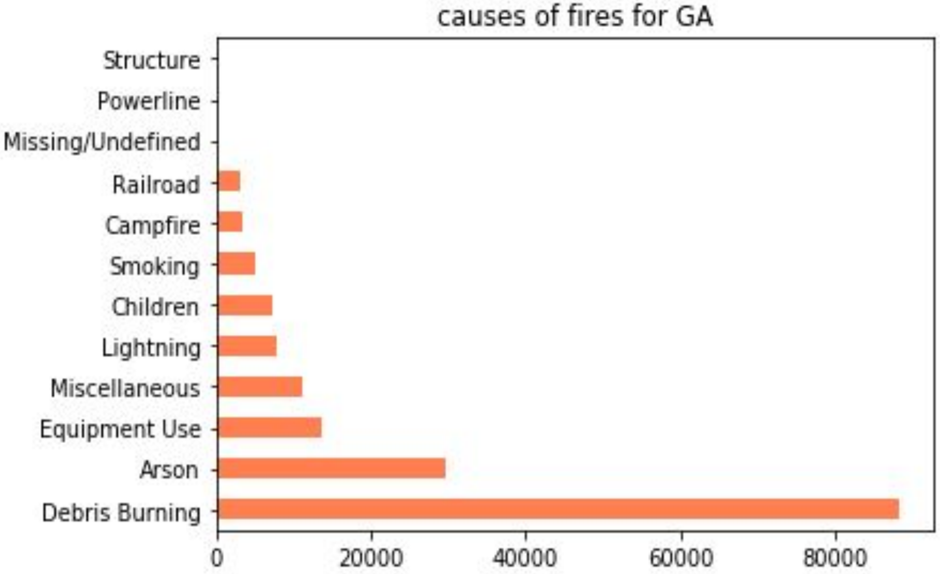
Day wise wildfire count

Fires are slightly more likely to start on the weekend. These fires might be due to people being careless with campfires or smoking or possibly malicious fires.

Arson is more likely at the weekend than during the week, an increase of around 30% of the average for weekdays.

Exploratory Data Analysis:

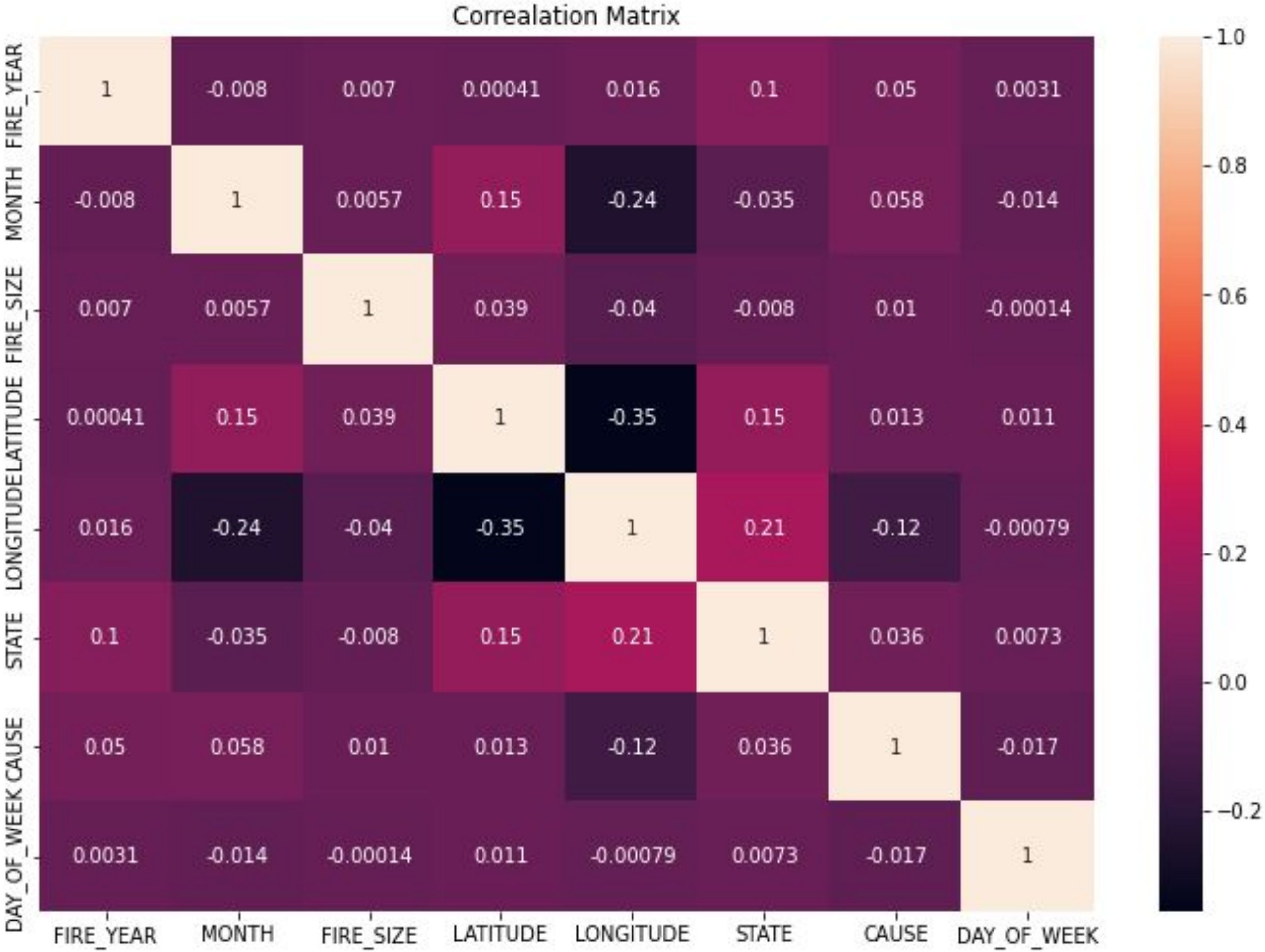




CA and TX are in the top 3, they are big states with dry climates. We are surprised to find GA in the top 3, as GA is not associated with wildfires.

Debris burning is the major cause of wildfires in GA which suggests location will be critical attribute in prediction.

Correlation Matrix



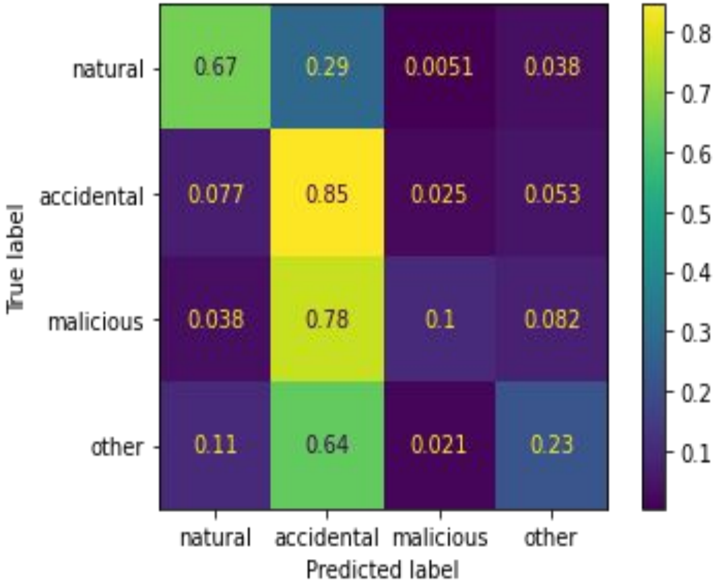
Classification using Machine Learning Methods

Ada Boost Classifier :

❑ n_estimators = 100

	precision	recall	f1-score	support
1	0.59	0.67	0.63	83316
2	0.53	0.85	0.65	249328
3	0.47	0.10	0.17	84170
4	0.59	0.23	0.33	147326
accuracy			0.55	564140
macro avg	0.55	0.46	0.45	564140
weighted avg	0.55	0.55	0.49	564140

Classification Report



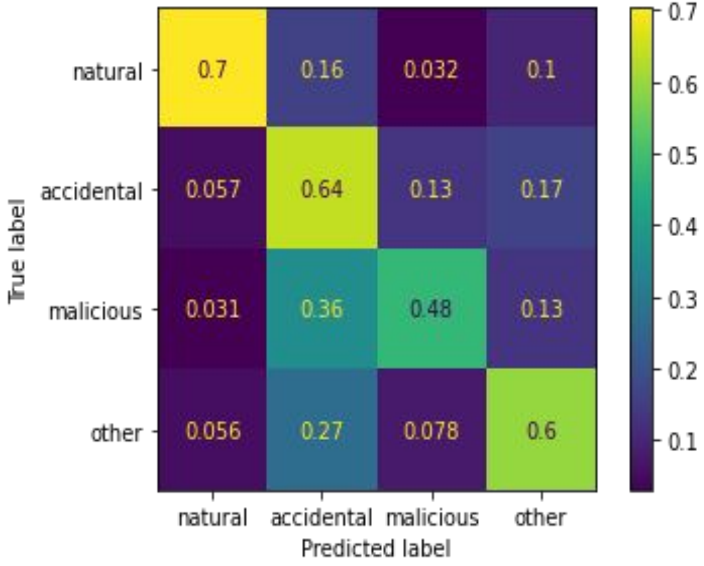
Confusion Matrix

Decision Tree Classifier :

- Criterion = GINI Index

	precision	recall	f1-score	support
1	0.70	0.70	0.70	83316
2	0.66	0.64	0.65	249328
3	0.46	0.48	0.47	84170
4	0.59	0.60	0.59	147326
accuracy			0.61	564140
macro avg	0.60	0.60	0.60	564140
weighted avg	0.62	0.61	0.61	564140

Classification Report



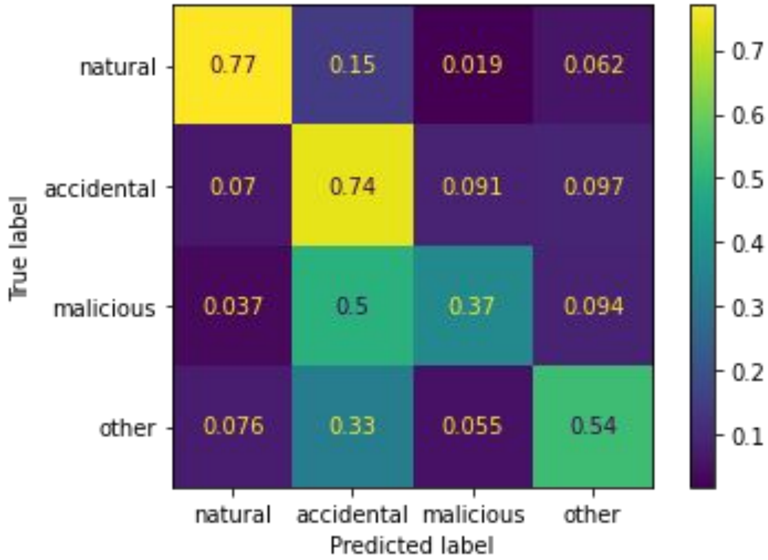
Confusion Matrix

K Nearest Neighbor Classifier :

- ❑ Number of Nearest Neighbors K = 5
- ❑ Metric = Minkowski

	precision	recall	f1-score	support
1	0.67	0.77	0.72	83316
2	0.64	0.74	0.69	249328
3	0.49	0.37	0.42	84170
4	0.68	0.54	0.60	147326
accuracy			0.64	564140
macro avg	0.62	0.61	0.61	564140
weighted avg	0.63	0.64	0.63	564140

Classification Report



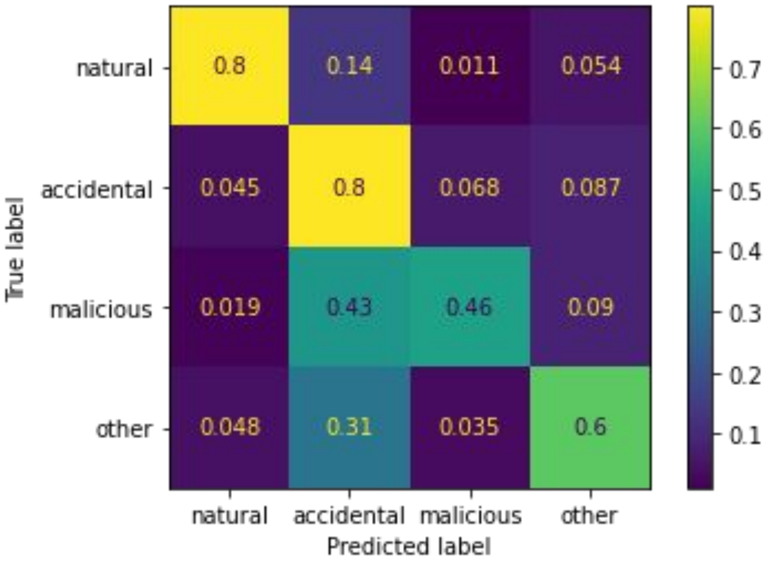
Confusion matrix

Random Forest Classifier :

- Number of Decision Trees = 50
- Criterion = GINI Index

	precision	recall	f1-score	support
1	0.77	0.80	0.78	83316
2	0.68	0.80	0.74	249328
3	0.63	0.46	0.53	84170
4	0.72	0.60	0.66	147326
accuracy			0.70	564140
macro avg	0.70	0.67	0.68	564140
weighted avg	0.70	0.70	0.69	564140

Classification Report

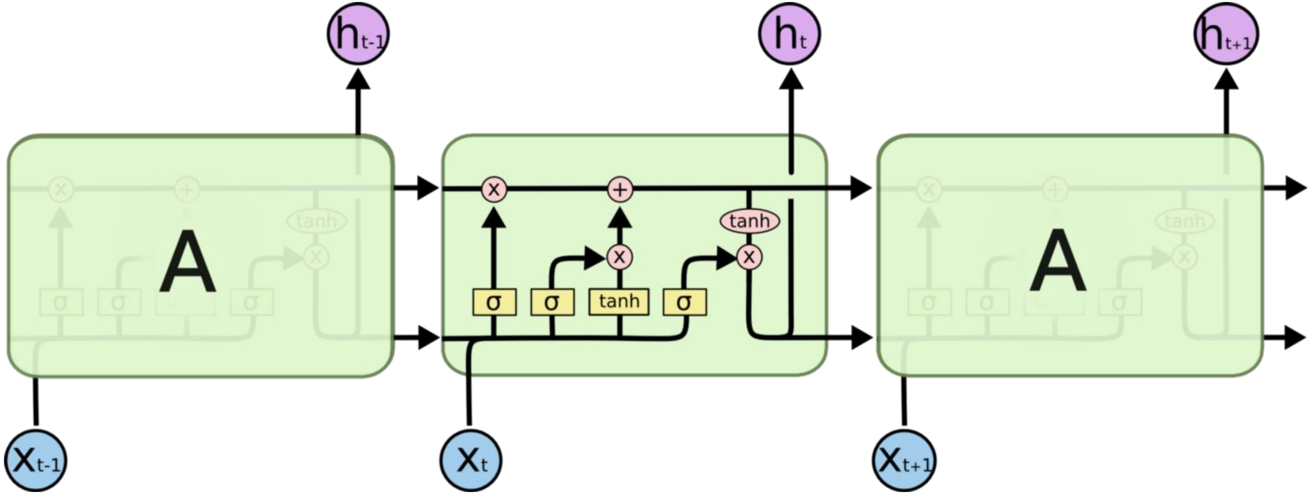


Confusion Matrix

Classification using Deep Learning Methods

Bi-directional LSTM :

- Consists of 2 LSTM layers:
  - One for taking inputs in forward direction.
  - One for progressing in backward direction.



Source : http://colah.github.io/posts/2015-08-Understanding-LSTMs/

Bi-directional LSTM :

```
Model: "sequential_1"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 bidirectional_1 (Bidirectio  (None, 256)              139264
 nal)

 dropout_1 (Dropout)         (None, 256)               0

 dense_3 (Dense)             (None, 128)               32896

 dense_4 (Dense)             (None, 64)                8256

 dense_5 (Dense)             (None, 5)                 325

=================================================================
Total params: 180,741
Trainable params: 180,741
Non-trainable params: 0
```

Model Parameters

	precision	recall	f1-score	support
1	0.61	0.73	0.66	55638
2	0.60	0.79	0.68	165779
3	0.54	0.22	0.32	56216
4	0.67	0.46	0.55	98459
accuracy			0.61	376092
macro avg	0.60	0.55	0.55	376092
weighted avg	0.61	0.61	0.59	376092

Classification Report

Convolutional Neural Network :

- ❑ CNN, is a deep learning neural network .
- ❑ Designed for processing structured arrays of data.
- ❑ Three main types of layers:
  - ❑ Convolutional layers.
  - ❑ Pooling layers.
  - ❑ Fully-connected layers.

Convolutional Neural Network :

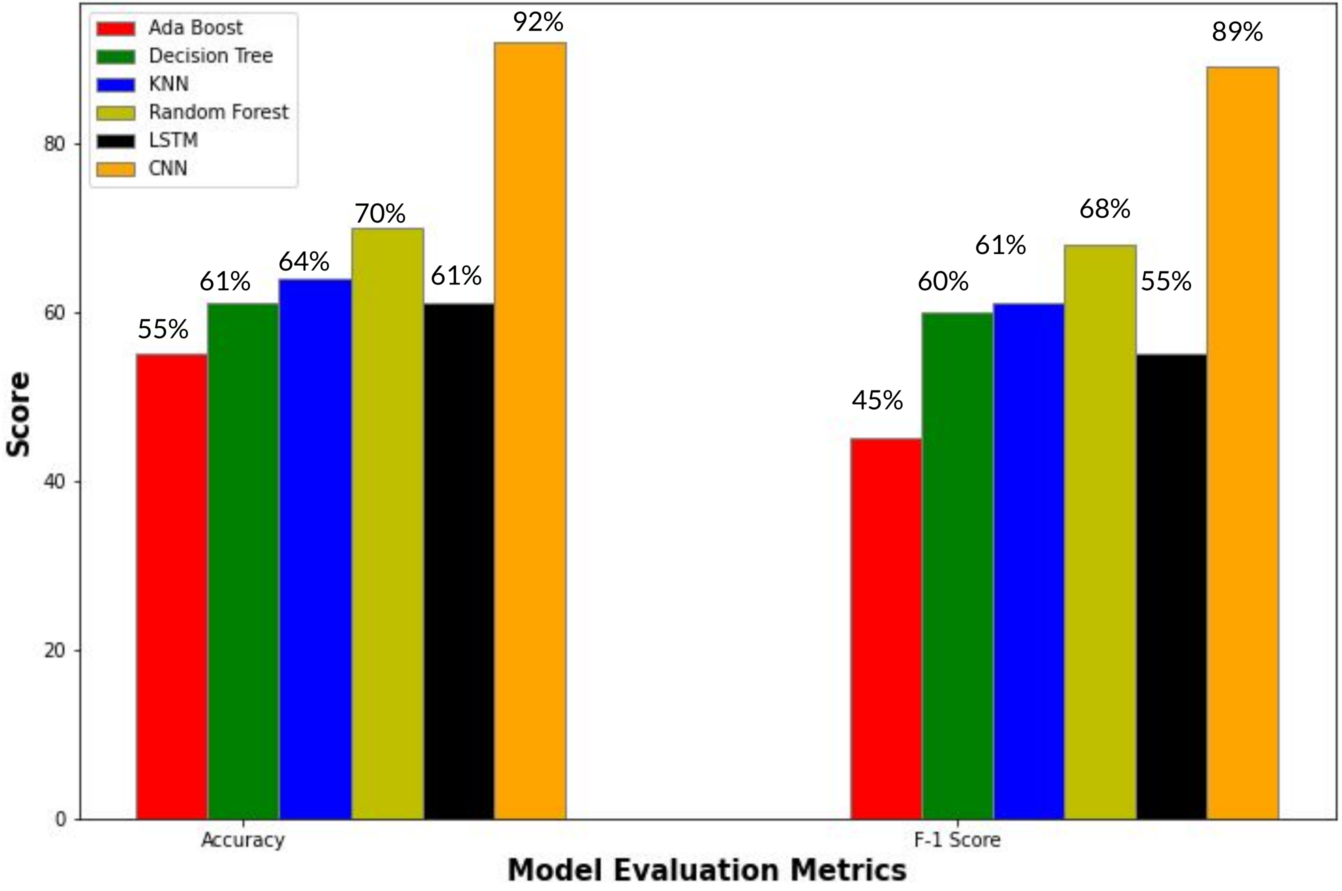
```
Model: "sequential_1"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 conv1d (Conv1D)             (None, 18, 128)           2816

 max_pooling1d (MaxPooling1D)  (None, 9, 128)          0

 conv1d_1 (Conv1D)           (None, 7, 64)             24640

 dropout_1 (Dropout)         (None, 7, 64)             0

 max_pooling1d_1 (MaxPooling1D)  (None, 3, 64)         0

 flatten (Flatten)           (None, 192)               0

 dense_3 (Dense)             (None, 256)               49408

 dense_4 (Dense)             (None, 5)                 1285

=================================================================
Total params: 78,149
Trainable params: 78,149
Non-trainable params: 0
_________________________________________________________________
```

Model Parameters

```
Precision score : 0.8533717876164396
F1_score        : 0.8871817313676837
Recall score    : 0.923781244460202
Accuracy        : 0.923781244460202
```

Classification Metrics

Conclusion:



Can we predict the cause of these wildfires using the data provided?

- ➔ Yes,
- ➔ With an accuracy of 70% using Random Forest.
- ➔ With an accuracy of 92% using CNN.
- ➔ Reducing the number of labels significantly improved the prediction score.
- ➔ The algorithms performed well while classifying data into ‘Natural’, ‘Accidental’ and ‘Other’ categories.
- ➔ However, trying to distinguish between ‘Accidental’ and ‘Malicious’ causes is not very accurate

Thank You