Wildfire Cause Prediction

Ganesh Thanu gthanu@ncsu.edu
Arun Gaonkar agaonka@ncsu.edu
Nikhil Prashant Patil npatil@ncsu.edu

Department of Computer Science, North Carolina State University
Raleigh, NC, USA

1 Background

1.1 Problem

Forests are an integral part of the terrestrial ecosystem and it has a major impact on the total biomass of the earth. Each year, in the US alone, around 60,000 wildfires can be seen and it causes impeccable damage to the earth. Although Wildfires are a natural occurrence, it is becoming extreme and widespread. 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, and uncontrolled fireworks are also major contributors to wildfire.

The objective of the project is to develop forest fire analysis and cause prediction techniques to aid in preventing future wildfire incidents. Technologies for finding the cause represent an essential tool to predict future forest fire risks,recommend forest fire monitoring procedures, activate prevention and security measures. Machine Learning models are implemented for the prediction of the cause of forest fires. The models can be used with Data Analysis techniques to forecast the cause of a wildfire using the dataset.[5]

To build the predictive model, the parameters being considered are forest fire size, geography, number of fires, etc. We are classifying the cause of wildfire into categories such as 'Natural', 'Accidental', 'Malicious' and 'Others'. The accuracies of ML techniques such as Decision Tree Classifier, Random Forest Classifier, AdaBoost classifier K-Nearest Neighbor classifier and Deep Learning models such as Convolutional Neural Networks, Bi-Directional LSTM are compared to identify the most accurate and efficient models that can be used for this specific use case and dataset.

1.2 Literature Survey

[1] **Piyush Jain, Sean C.P. Coogan, Sriram Ganapathi Subramanian, Mark Crowley, Steve Taylor, and Mike D. Flannigan. 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

Machine Learning and Deep Learning model are implemented to predict the cause of the wildfire given the values of existing parameters. This publication gives an idea about the different models that can be applied for each of our use cases. It also mentions the various research work done previously in this specific domain of wildfire and lists many citations and the common methods, techniques implemented for predicting the cause of wildfire.

[2] **Zhang, G., Wang, M. & Liu, K. Forest Fire Susceptibility Modeling Using a Convolutional Neural Network for Yunnan Province of China. Int J Disaster Risk Sci 10, 386–403 (2019).** https://doi.org/10.1007/s13753-019-00233-1

Convolutional Neural Network with deep architectures is used in this citation for the spatial prediction of forest fire susceptibility . Past forest fire locations from 2002 to 2010 were extracted and a set of 14 forest fire influencing factors are considered and optimized. The CNN architecture suitable

for the prediction of forest fire susceptibility in the study area is designed, and hyperparameters are optimized to improve the prediction accuracy. The obtained CNN model is compared with other ML techniques such as Random Forest with 160 decision trees, SVM with RBF as the Kernel function and 100 as the Penalty factor. The performance of the proposed model is compared with traditional ML methods using several statistical measures and the results of the paper concludes that CNN is the most accurate model for this use case.

[3] Marcos Rodrigues & Juan de la Riva. An insight into machine-learning algorithms to model human-caused wildfire occurrence, Environmental Modelling & Software, `https://doi.org/10.1016/j.envsoft.2014.03.003`

It proposes the use of ML within the context of fire risk prediction, and more specifically, in the evaluation of human-induced wildfires. It compares three ML algorithms—Random Forest (RF), Boosting Regression Trees (BRT), and Support Vector Machines (SVM) with traditional methods like Logistic Regression (LR). The results from the research work suggest that the use of any of these ML algorithms leads to an improvement in the accuracy,in terms of the AUC (area under the curve) when compared to LR outputs. The output from the research gave results of RF and BRT accuracy values as 0.746 and 0.730 respectively.

[4] Malik, Ashima, Megha R. Rao, Nandini Puppala, Prathusha Koouri, Venkata A.K. Thota, Qiao Liu, Sen Chiao, and Jerry Gao. 2021. "Data-Driven Wildfire Risk Prediction in Northern California" Atmosphere 12, no. 1: 109. `https://doi.org/10.3390/atmos12010109`

It compares the accuracy of Machine Learning algorithms in the context of fire risk prediction in Northern California. Unlike other existing models, these models are integrated models powered by machine learning algorithm such as Adaboost, Decision trees, Gradient descent, Multi-layered perceptron, Random Forest Tree (RF), and Long Short-Term Memory (LSTM) to address convoluted location-specific wildfire risk prediction. The final output by the research work concludes that in terms of accuracy of individual models, the Random forest algorithm outperforms all other algorithms when experimented with varied target labeling and subsets of the datasets.

2 Methods

2.1 Approach

We started with importing dataset into pandas dataframe as our dataset was in the form of SQLite file. After that we proceeded with exploratory data analysis, plotting various graphs and correlation matrix in order to get the sense of data. Further, We prepared data by dropping certain columns, converting categorical variables to numerical and then applied machine learning and deep learning techniques.[9]

In machine learning classification techniques, First classification technique applied was the **Adaboost classifier**. Adaboost is a type of ensemble approach which builds a strong classifier by combining multiple poorly performing classifiers. Furthermore, we applied **K-Nearest Neighbors** (KNN) where classification is done based on k closest training examples. Next technique used was **Decision Tree.** Decision Tree is an intuitive approach in which tree's branches contain logic for a decision rule according to which data can be split. After Decision tree implementation, we went ahead with **Random Forest Classifier** as we know it consists of a large number of individual decision trees that operate as an ensemble and therefore, it is more robust than decision tree.

In Deep learning techniques, We first applied **Bidirectional LSTM** which enables additional training by traversing the input data twice (left-to-right and right-to-left). Finally, we used **Convolutional Neural Networks (CNN)**.

2.2 Rationale

During **Exploratory data analysis**[7], we plotted various graphs of each attribute against the target variable, analyzed correlation between attributes in order to determine significant attributes to include in the model, attributes which are strongly correlated and can be dropped.

Considering the large size of the dataset (1.88 million records)[5], We first started with approaches which were easy to implement and faster than other algorithms. So, we started with the Adaboost

classifier which was simple to implement with very few parameters and KNN as it is faster than other algorithms and only needs two parameters. Once we had the significant results, we moved forward with more robust algorithms like Decision Tree and Random Forest Classifier which was an intuitive choice after implementing the decision tree. Because, random forest reduces bias by aggregating multiple decision trees and can produce more accurate results.

SVM is also one of the robust approaches for classification. But, considering its computational complexity and large size of our dataset, we chose not to implement SVM.

Assuming Deep Learning gives better prediction results than traditional ML, we implemented Bi-LSTM, because of it's fast learning and additional training capabilities. Further we built a CNN model to get higher prediction results as it is efficient in extracting patterns from the dataset.

3 Plan and Experimental Setup

3.1 Dataset

The forest fire dataset *1.88 Million US Wildfires* [5] obtained from Kaggle[6] is used. This database contains geospatial records of wildfires that occurred in the US from 1992 to 2015. The wildfire records are taken from the reporting systems of federal, state, and local fire organizations. This database is referred to as *Fire Program Analysis Fire-Occurrence Database* (FPA-FOD).

This dataset is an SQLite database contains information about the latitude and longitude of the fire, fire start date, identification date (in Julian Date format), discovery time, fire ceased date, size of the fire, cause for the fire, geographic area, local Government and fire agency and etc, adding to total of 130 columns. The dates are in the Julian format are converted to Gregorian month-date format. The dataset is also checked for null values, and they are not found in the dataset.

Based on the initial analysis, multiple irrelevant columns for the prediction task are observed and 14 columns are extracted from the dataset. After doing Exploratory Data Analysis, few more redundant columns are removed. Finally the dataset, consists of 1.88 Million rows and 8 columns, viz.Start Day, Month, Fire Year, Fire Size, Latitude, Longitude, State, Cause. Each wildfire record is labelled into 13 different categories, viz.,Structure, Fireworks, Power line, Railroad, Smoking, Children, Campfire, Equipment use, Lightning, Arson, Debris Burning, Miscellaneous and Undefined. These labels are encoded for classification purposes.

3.2 Hypothesis

Hypothesis 1: Natural events such as hot weather, lightning are frequent, so the number of fires due to natural events is believed to be the highest.

Hypothesis 2: Drier months like July are likely to have more fires because of weather conditions, so during these months, fire counts are expected to be higher.

Hypothesis 3: There is no correlation between the start day of the fire and the fire count. It is expected to be true because wildfires have no relation to days of the week.

Hypothesis 4: States that are dry are prone to wildfires, and have more fires than other normal or cold states, most probably because of the feasible climate conditions.

Hypothesis 5: The more number of categories to classify the data will make the model less accurate. When there are fewer categories, there is a higher chance of accurately predicting the category.

Hypothesis 6: For this task, the intuition is that the K Nearest Neighbor approach will outperform the Adaboost approach.

Hypothesis 7: The general assumption is that Random Forest Classifier with an optimal number of trees is better than Decision Tree for prediction tasks.

Hypothesis 8: The abstract belief is that Deep Learning methods will give better prediction results than traditional Machine Learning algorithms, because of its architecture to train and to predict.

3.3 Experimental Design

In the original dataset, we had 39 columns, all of which were not useful for prediction tasks. They were related to local reporting agency and owner information. We reduced the number of features to 14 columns, which are, FOD_ID, FIRE_YEAR, MONTH, START_DATE, DISCOVERY_TIME, END_DATE, CONT_TIME, FIRE_SIZE, FIRE_SIZE_CLASS, LATITUDE, LONGITUDE, STATE, COUNTY, STAT_CAUSE_DESCR.

We then performed EDA to select only those columns which can impact the cause prediction task. At the end of EDA, we have 8 feature columns left which are, Start Day, Month, Fire Year, Fire Size, Latitude, Longitude, State and Cause. We then decided to use first 7 features for X and the cause feature as Y. Both X and Y are then split into train-test set, with 70% being training and remaining 30% for testing set. After this splitting our train set is of size [(1316325, 7) and (1316325, 1)] and test set is of size [(564140, 7) and (564140, 1)].

We decided to reduce the number of categories in the Cause feature. So based on the cause categories provided in the dataset description , updated causes are 'natural', 'accidental', 'malicious' and 'other'. 'Lightning' is categorized to '**natural**'. We combined 'Structure', 'Fireworks', 'Powerline', 'Railroad', 'Smoking', 'Children', 'Campfire', 'Equipment Use' and 'Debris Burning' to '**accidental**' category. In '**malicious**' category, we added 'Arson'. 'Missing/Undefined' and 'Miscellaneous' are categorized to '**other**'. These 4 categories in the cause feature is then encoded to numerical values.

For training the dataset on **Adaboost classifier**, maximum number of estimators used for terminating the boosting is 100 and the learning rate of 1.0 is used. This model tested on testset. For the comparison purpose Classification report and confusion matrix are generated. Using **Decision Tree Classifier** with GINI as the splitting criterion, set is trained and then predicted on test set. For **KNN**, 5 nearest neighbors are considered and model is built. After training and testing, Classification report and Confusion matrix are generated.

The initial training dataset with cause having 13 categories, is fed into **Random Forest Classifier**[8] with 50 trees and the model is trained. The model is then used for prediction on the test set. The prediction accuracy is observed to be low. Further Random Forest Classifier model is trained with the updated cause feature having only 4 categories. Model is then predicted on the test set, the accuracy was found to be significantly increased and hence **hypothesis 5 is validated**. Classification report and confusion matrix are generated.

To implement Deep Learning methods (Bi-LSTM and CNN), we needed to have the three dimensional input. So we converted our 2D dataset to 3D data with an option to vary the timesteps. We have implemented **Bi-directional LSTM** using keras. The Bi-directional LSTM model is built with 256 LSTM units in the first layer. And then a dropout layer with probability of 0.6. Next 3 layers are dense layers with 128, 64 and 5 neurons at each layer respectively. Implemented model structure is shown in Figure 1. After training this model, it is used for prediction on the test dataset. Accuracy, F1 score, recall and precision are calculated.

We also trained **Convolutional Neural Network (CNN)**, on the 3 dimensional dataset, with timesteps as 40. The model is shown in the Figure 2. We started the model with one convolutional layer with 128 neurons and 'relu' as the activation is followed by a maxpooling layer. One more convolutional layer with 64 neurons is added and then the droupout layer with 0.5 probability. One more maxpool and then 2 dense layers with 256 and 5 units are added. Softmax is the activation function used for the final dense layer. The model is trained for 3 epochs, until when we observed, no increase in the accuracy for further training. Accuracy, Precision, F1-score and recall are calculated after predicting on the test set.

The accuracy, F1-score for different models are then compared.

4 Results

4.1 Results

Performing exploratory data analysis yielded interesting results. We analyzed wildfire causes and their respective fire counts. We also plotted month wise, day wise fire count, state-wise fire reasons for analyzing the dataset. Correlation between the attributes are visualized using correlation matrix[7].

```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 bidirectional (Bidirectional)  (None, 256)           139264

 dropout (Dropout)           (None, 256)               0

 dense (Dense)               (None, 128)               32896

 dense_1 (Dense)             (None, 64)                8256

 dense_2 (Dense)             (None, 5)                 325

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

Figure 1: Bi-directional LSTM Model

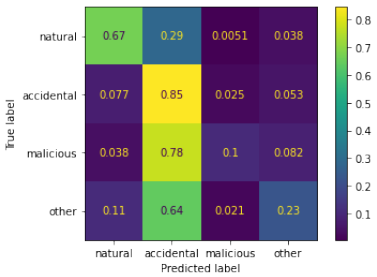
```
 Layer (type)                Output Shape              Param #
=================================================================
 conv1d (Conv1D)             (None, 38, 128)           2816

 max_pooling1d (MaxPooling1D)  (None, 19, 128)         0

 conv1d_1 (Conv1D)           (None, 17, 64)            24640

 dropout (Dropout)           (None, 17, 64)            0

 max_pooling1d_1 (MaxPooling1D)  (None, 8, 64)         0

 flatten (Flatten)           (None, 512)               0

 dense (Dense)               (None, 256)               131328

 dense_1 (Dense)             (None, 5)                 1285

=================================================================
Total params: 160,069
Trainable params: 160,069
Non-trainable params: 0
```

Figure 2: CNN Model

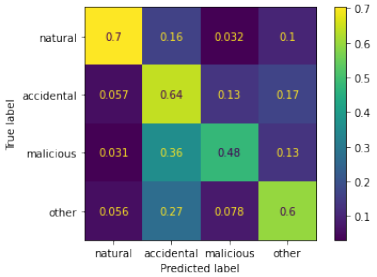
We ran the dataset through multiple models. With the Adaboost Classifier and fore-mentioned parameters, we achieved the accuracy of 55%. When the dataset is fed into Decision Tree and KNN (K =5), the prediction accuracy of 61% and 64% was observed respectively. The Random Forest Classifier with above mentioned model parameters, is giving the accuracy of 70%. Confusion matrices for each model is shown in Figure 3. Comparison for each model is shown in Figure 4.

With Bi-directional LSTM model, with the parameters as shown in Figure 1, accuracy was found to be 61%. When Convolutional Neural Network was used, with the layers and parameters as in Figure 2, the accuracy was found to be 92.37%. Comparison between Deep Learning models are shown in Figure 5.

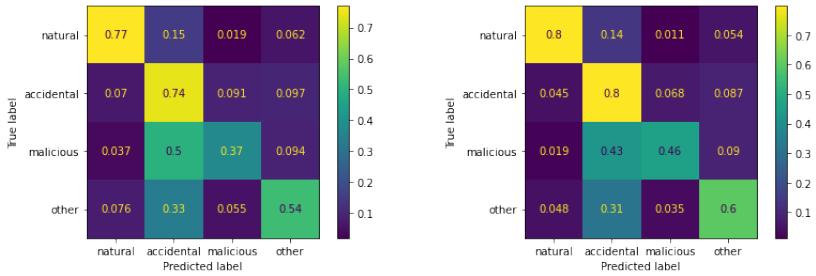
Comparing between the prediction accuracies for different models, we found that, CNN is the best model for wildfire cause prediction task.



(a) Adaboost Classifier



(b) Decision Tree



(c) K - Nearest Neighbor (d) Random Forest Classifier

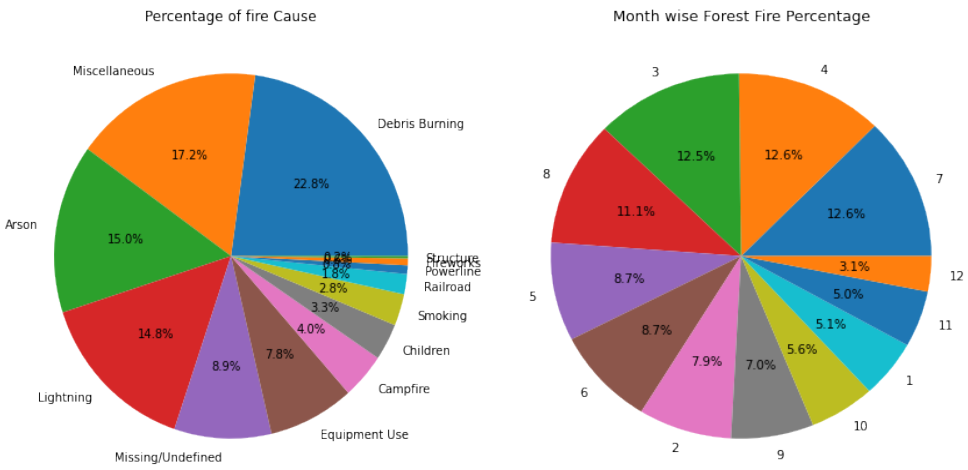
Figure 3: Confusion Matrix comparison

Class	Adaboost			Decision Tree			K Nearest Neighbor			Random Forest Classifier		
	Precision	Recall	F1-score	Precision	Recall	F1-score	Precision	Recall	F1-score	Precision	Recall	F1-score
Natural	0.59	0.67	0.63	0.7	0.7	0.7	0.67	0.77	0.72	0.77	0.8	0.78
Malicious	0.53	0.85	0.65	0.66	0.64	0.65	0.64	0.74	0.69	0.68	0.8	0.74
Accidental	0.47	0.1	0.17	0.46	0.48	0.47	0.49	0.37	0.42	0.63	0.46	0.53
Other	0.59	0.23	0.33	0.59	0.6	0.59	0.68	0.54	0.6	0.72	0.6	0.66

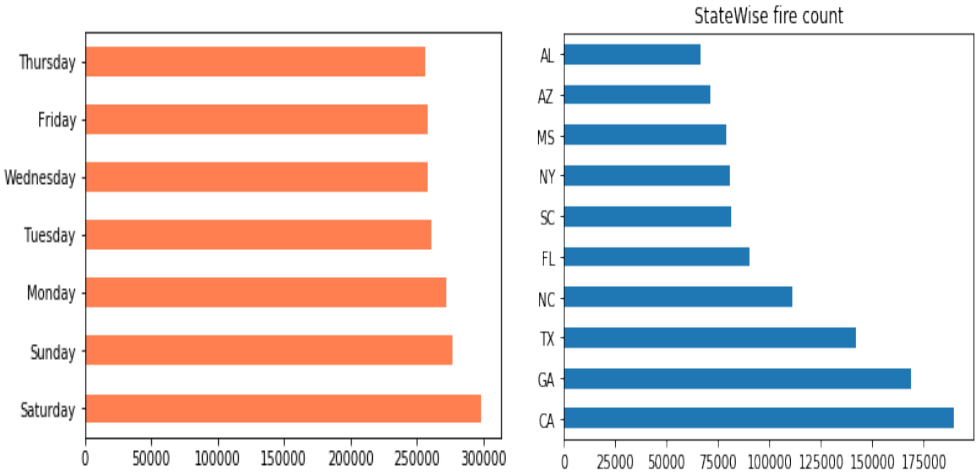
Figure 4: Classification Report Comparison for ML models

	Bi-LSTM	CNN
Precision	0.61	0.85
Recall	0.61	0.93
F1-Score	0.59	0.88
Accuracy	0.61	0.93

Figure 5: Classification Report Comparison for Bi-LSTM and CNN



(a) Fire cause and count parentage (b) Month-wise fire percentage



(c) Day-wise count parentage

(d) State-wise fire percentage

Figure 6: Exploratory Data Analysis plots

4.2 Critical Evaluation

From the exploratory data analysis and modelling part, we can evaluate our various hypothesis as follows:

- Debris burning, Arson and other accidental causes are the major causes of wildfire accounting for more than 50% of total wildfires which **negates hypothesis 1**, from Figure 6.(a).
- The period from February to September accounts for approximately 80% of the total number of wildfires which **validates hypothesis 2**,from Figure 6.(b).
- We got surprising results while visualizing the relationship between the start day of the fire and the fire count. The likelihood of fires on weekends is slightly higher. On further investigation, we found that it was due to fires from the arson category. On weekends, we observed an increase of around 30% of the average for weekdays. These results **contradict with our hypothesis 3**,from Figure 6.(c).
- From state wise wildfire count visualization, we found Drier states like California, Texas account for more wildfires which **satisfies hypothesis 4**,from Figure 6.(d).
- Prediction accuracy increased from 58% to 70% as we reduced the number of class labels from 13 to 4 which is **support of our hypothesis 5**. We think this is due to an increase in the number of samples per category as we clubbed some categories which results in higher accuracy.
- KNN gave better results than Adaboost as seen from classification report from Figure 4, which **validates hypothesis 6**.
- Accuracy and other confusion matrix results are better for Random Forest than Decision Tree which **validates hypothesis 7**, from Figure 4. This might be due to Random Forest relies on collecting various Decision Trees to arrive at any solution
- The model evaluation metrics for CNN are consistently higher than those of all the other classification algorithms implemented, from Figure 4,5. Therefore, **Hypothesis 8 is supported**.

5 Conclusions

We started with a question: Can we predict the cause of these wildfires using the data provided? The answer is **YES**. Through our experiments, we present various Machine Learning and Deep Learning models that can be used for wildfire cause prediction based on given data. From our results, we can

see that **Random Forest predicts the cause with an accuracy of 70%** and **CNN with an accuracy of 92%**. All 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. We think that is because training examples for malicious category are much less than accidental category and to improve further stratified sampling could have been used.

We learnt how to perform following tasks:

- Creating visualizations using matplotlib.
- Generating and analysing correlation matrix.
- Implementation of Adaboost, KNN, Decision tree, Random forest using sklearn.
- Implementation of Bi-LSTM, CNN using keras.
- Generating and analysing classification report representing various model evaluation metrics.

6 References

[1] Piyush Jain, Sean C.P. Coogan, Sriram Ganapathi Subramanian, Mark Crowley, Steve Taylor, and Mike D. Flannigan. 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

[2] Zhang, G., Wang, M. & Liu, K. Forest Fire Susceptibility Modeling Using a Convolutional Neural Network for Yunnan Province of China. Int J Disaster Risk Sci 10, 386–403 (2019). https://doi.org/10.1007/s13753-019-00233-1

[3] Marcos Rodrigues & Juan de la Riva. An insight into machine-learning algorithms to model human-caused wildfire occurrence, Environmental Modelling & Software, https://doi.org/10.1016/j.envsoft.2014.03.003

[4] Malik, Ashima, Megha R. Rao, Nandini Puppala, Prathusha Koouri, Venkata A.K. Thota, Qiao Liu, Sen Chiao, and Jerry Gao. 2021. "Data-Driven Wildfire Risk Prediction in Northern California" Atmosphere 12, no. 1: 109. https://doi.org/10.3390/atmos12010109

[5] Short, Karen C. 2017. Spatial wildfire occurrence data for the United States, 1992-2015 [FPAFOD20170508]. 4th Edition. Fort Collins, CO: Forest Service Research Data Archive. https://doi.org/10.2737/RDS-2013-0009.4

[6] Kaggle Dataset, https://www.kaggle.com/rtatman/188-million-us-wildfires

[7] Analysis on EDA, http://regclim.coas.oregonstate.edu/FireStarts/fpa-fod_RODBC_01.html

[8] J. Brownlee, "How to develop a random forest ensemble in Python," Machine Learning Mastery, 26-Apr-2021. [Online]. Available: https://machinelearningmastery.com/random-forest-ensemble-in-python/.

[9] P. Cortez and A. Morais, A data mining approach to predict forest fires using meteorological data, Proceedings of the 13th Portugese Conference on Artificial Intelligence, pp. 512-523, 2007.

7 Github link

Follow this link for the github repository.

https://github.ncsu.edu/agaonka/engr-ALDA-fall2021-P24