A Project Report

On

An Analysis of Forest Fire Dataset

Submitted in Partial Fulfillment of the Requirements for the Award of

the Degree of

Bachelor of Technology in

Computer Science and Engineering

By

SANJEEV KUMAR (1406027)

RAHUL KUMAR (1406010)

ABHISHEK KUMAR (1406004)

PRSHANT KUMAR SINGH (1406011)

Under the Supervision of

Dr. Ditipriya Sinha

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

NATIONAL INSTITUTE OF TECHNOLOGY PATNA

JUL – DEC 2017

ABSTRACT

An Analysis of Forset Fire Dataset

Machine learning is a core sub-area of artificial intelligence; it enables computers to get into a mode of self-learning without being explicitly programmed. When exposed to new data, these computer programs are enabled to learn, grow, change and develop by themselves. Similarly, Predicting natural disasters like hailstorm, fire, rainfall etc. are infrequent and stochastic. Detection of these disasters should be fast and accurate as they may cause damage and destruction at large scale.

In this paper, comparison of various machine learning techniques such as SVM, logistic regression, decision trees, etc. has been done for prediction of forest fires. The proposed approach in this paper presents how different machine learning models works best for detection of forest fire with high accuracy. Fast detection of forest fires is done in this paper by taking less number of feature which are more importance reducing the energy usage for sensors, as compared to other techniques.

ACKNOWLEDGEMENT

We would like to take this opportunity to thank all our sources of aspirations during this project. First and foremost, I would like to thank our project Internal Guide, **Dr. Ditipriya Sinha**, for his continuous support during this project and related research, and for his patience, motivation, and immense knowledge. And during the most difficult times when writing this report, he gave us the moral support and the freedom we needed to move on.

In addition, I would like to thank my colleagues and project partners for their valuable contributions. Some of the open source contributions helped us to understand the platform better and finish this challenge. Last but not the least, I would like to thank the God, the Almighty.

Sanjeev Kumar	Rahul Kumar
B. Tech - C.S.E, 7th sem	B. Tech - C.S.E, 7th sem
1406027	1406010
Abhishek Kumar	Prashant Kumar Singh
B. Tech – C.S.E, 7th sem	B. Tech- C.S.E, 7th sem
1406004	1406011

राष्ट्रीय प्रौद्योगिकी संस्थान पटना
NATIONAL INSTITUTE OF TECHNOLOGY PATNA

DECLARATION

We hereby declare that this project work entitled **"An Analysis of Forest Fire DataSet"** has been carried out by us in the **Department of Computer Science and Engineering** of **National Institute of Technology, Patna** under the guidance of **Dr. Ditipriya Sinha, Assistant Professor.** No part of this work has been submitted for the minor project of degree or diploma to any other institute.

Sanjeev Kumar
B. Tech - C.S.E, 7th sem
1406027

Rahul Kumar
B. Tech - C.S.E, 7th sem
1406010

Abhishek Kumar
B. Tech – C.S.E, 7th sem
1406004

Prashant Kumar Singh
B. Tech- C.S.E, 7th sem
1406011

Place: Patna
Date: __/__/____

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

NATIONAL INSTITUTE OF TECHNOLOGY PATNA

JUL-DEC 2017

CERTIFICATE

This is to certify that **Sanjeev Kumar(1406027), Rahul Kumar(1406010), Abhishek Kumar (1406004), Prashant Kumar Singh(1406011)** have completed "**An Analysis of Forest Fire Dataset"** as Minor Project (7CS191) under my supervision.

We hereby recommend that this project is accepted as per the requirements of evaluation and for the award of B.Tech Degree in Computer Science & Engineering.

Dr. Ditipriya Sinha

(Project Guide)

Assistant Professor

Department of

Computer Science & Engineer

Dr. Prabhat Kumar

Assistant Professor &

Head of Department

Computer Science &

Engineering

Table of Contents

I. INTRODUCTION

Forest fire also known as bush fire or hill fire is an uncontrolled fire occurring wild or forest areas. It is very important to detect these kinds of fires as early as possible so as to prevent the damage from it to ecological system. Every year millions of acres of forest are burnt down. The land were forest is burnt it becomes impossible to grow vegetation over there. This is because soil becomes water repellent and accepts no more water, leading to reduction in ground water level. The Global Warming Report 2008 mentions forest fire as one of the major cause behind increase in global warming.

In recent year 2016 more than 4000 hectares of forest were burnt in the hills of Uttarakhand. Common causes of forest fire are lightning, extreme hot and arid weather and human carelessness. The use of wireless sensor in this paper presents one of the techniques for early forest fire detection.

Many solutions for detection of forest fire are presented and implemented in past few years. Video Surveillance System is most widely used for detection of forest fire[1]. It is divided into four categories : Video Cameras sensitive in visible spectrum based on recognition of smoke during day light and fire flames at night, Infrared(IR) Thermal Imaging cameras based on detection of heat flux from the fire, IR Spectrometer which identify spectral characteristics of smoke gases and Light Detection and Ranging (LIDAR) system which measures the laser light backscattered by smoke particles. The limitation of these systems was high false alarm rate due to atmospheric conditions such as presence of fog, shadows, dust particles etc.

II. Attribute Information

1.**X** - x-axis spatial coordinate within the Montesinho
parkmap: 1 to 9

2. **Y -** y-axis spatial coordinate within the Montesinho
park map: 2 to 9

3. **month** - month of the year: 'jan' to 'dec'

4. **day -** day of the week: 'mon' to 'sun'

5. **FFMC** - FFMC index from the FWI system:
18.7 to 96.20
6. **DMC** - DMC index from the FWI system: 1.1 to 291.3

7. **DC** - DC index from the FWI system: 7.9 to 860.6

8. **ISI** - ISI index from the FWI system: 0.0 to 56.10

9. **temp** - temperature in Celsius degrees: 2.2 to 33.30

10. **RH** - relative humidity in %: 15.0 to 100

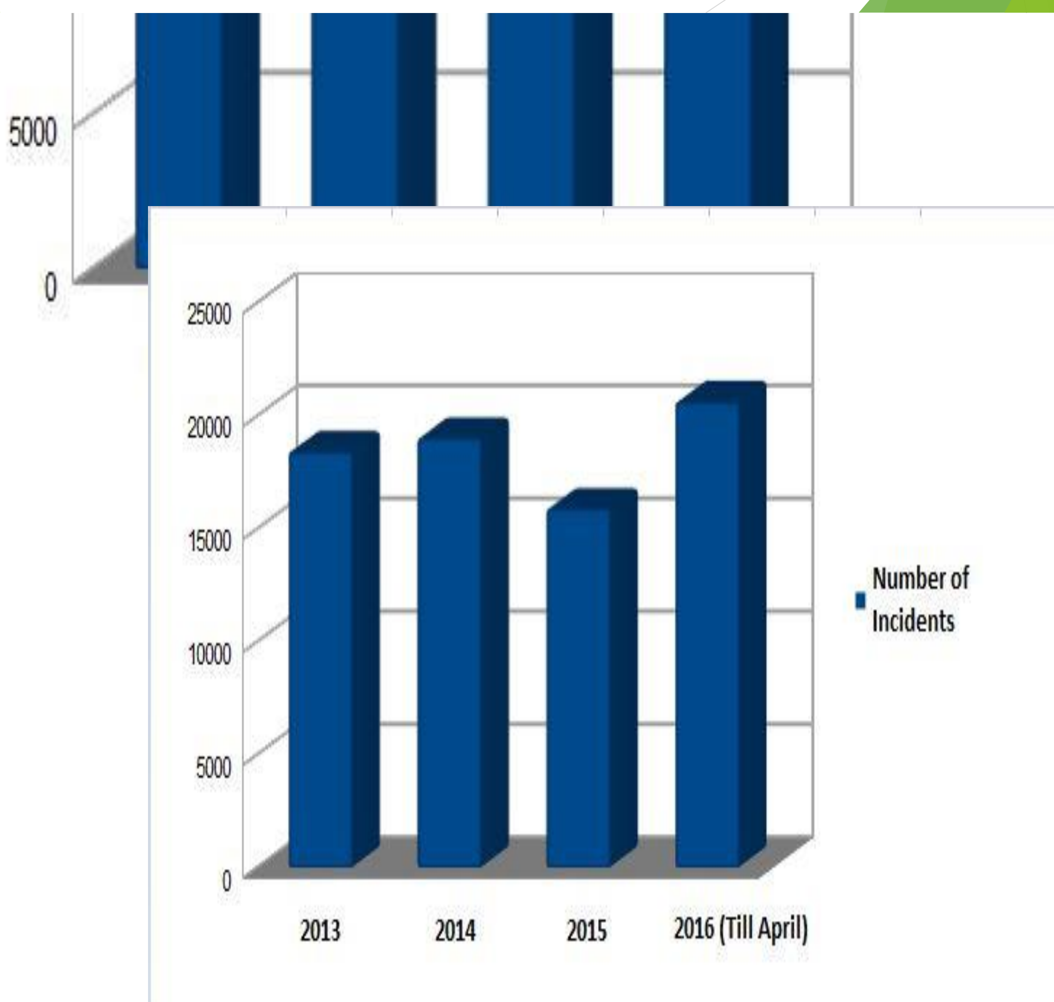
11. **wind -** wind speed in km/h: 0.40 to 9.40

12. **rain** - outside rain in mm/m2 : 0.0 to 6.4

13. **area** - the burned area of the forest (in ha): 0.00 to
1090.84
(this output variable is very skewed towards 0.0, thus
it may make sense to model with the logarithm transform)

Dataset Description

Title	Forest Fire Data Set
Data Set Characteristics:	Multivariate
Number of Instances:	517
Area:	Physical
Number of Attributes:	13
Associated Tasks:	Regression



III. Classes Description

AIM : To classify given dataset into three classes.

Active High

- It includes the most affected region by fire.
- Highest number of sensor to be deployed here.

Active Medium

- It includes the moderate affected region by fire.
- Lesser number of sensor to be deployed here.

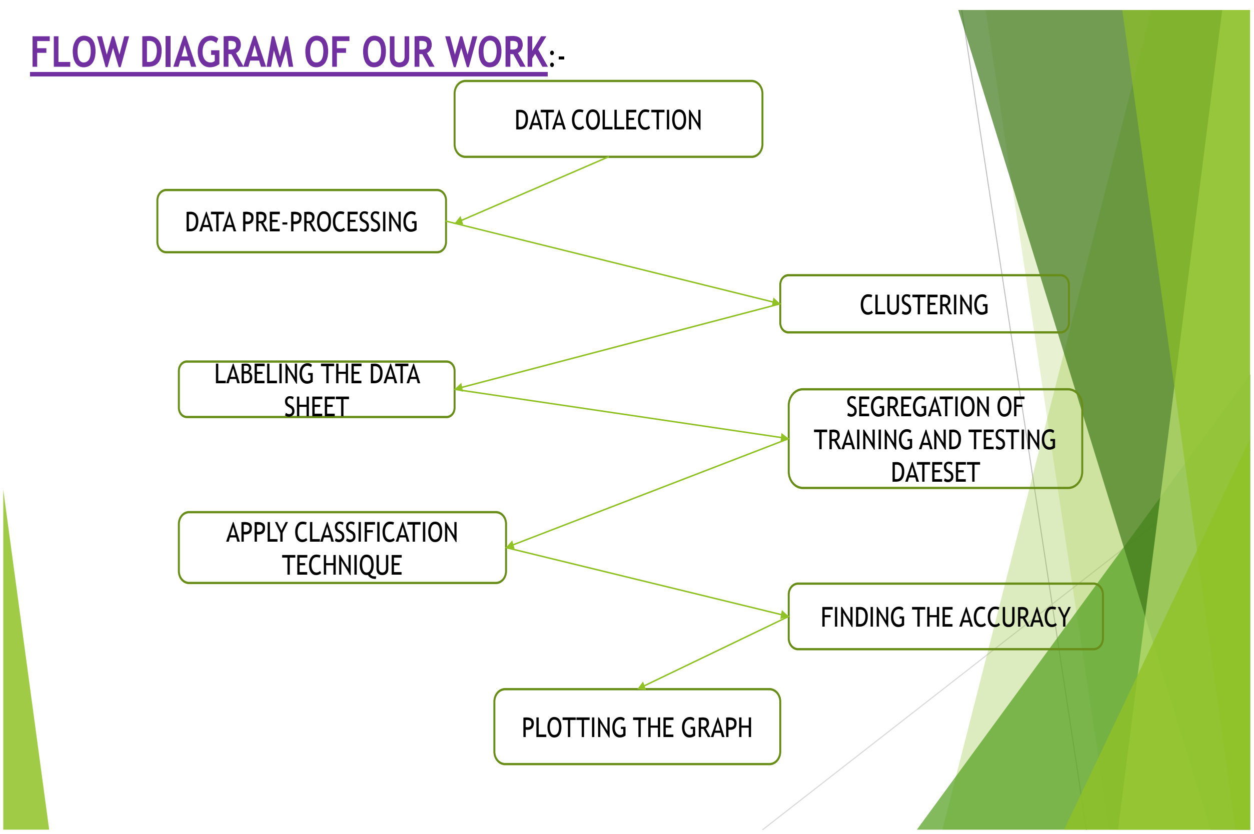
Active Low

- It includes the least affected region by fire.
- Very few number of sensor to be deployed here.

IV. Flow of Work

There are many steps to find the most burnt area by using given attributes which are as follows:

- First of all, the data is collected from the references.
- Then, the collected dataset is pre-processed before applying any further operations.
- Again, the dataset is clustered to different groups such that all objects in same group are similar.
- Further, the derived dataset is sampled to one or more labels.
- After that, the training and testing is done via separating into two dataset by applying the classification technique.
- Finally, the accuracy is calculated and their corresponding graph is plotted.



V. Clustering

Cluster analysis or **clustering** is the task of grouping a set of objects in such a way that objects in the same group (called a **cluster**) are more similar (in some sense or another) to each other than to those in other groups (clusters). It is a main task of exploratory data mining, and a common technique for statistical data analysis, used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, bioinformatics, data compression, and computer graphics.

Clustering is an unsupervised machine learning task that automatically divides the data into **clusters,** or groups of similar items. It does this without having been told how the groups should look ahead of time. As we may not even know what we're looking for, clustering is used for knowledge discovery rather than prediction. It provides an insight into the natural groupings found within data.

The resulting clusters can then be used for action. For instance, you might find clustering methods employed in the following applications:

- Segmenting customers into groups with similar demographics or buying patterns for targeted marketing campaigns.

- Detecting anomalous behavior, such as unauthorized network intrusions, by identifying patterns of use falling outside the known clusters

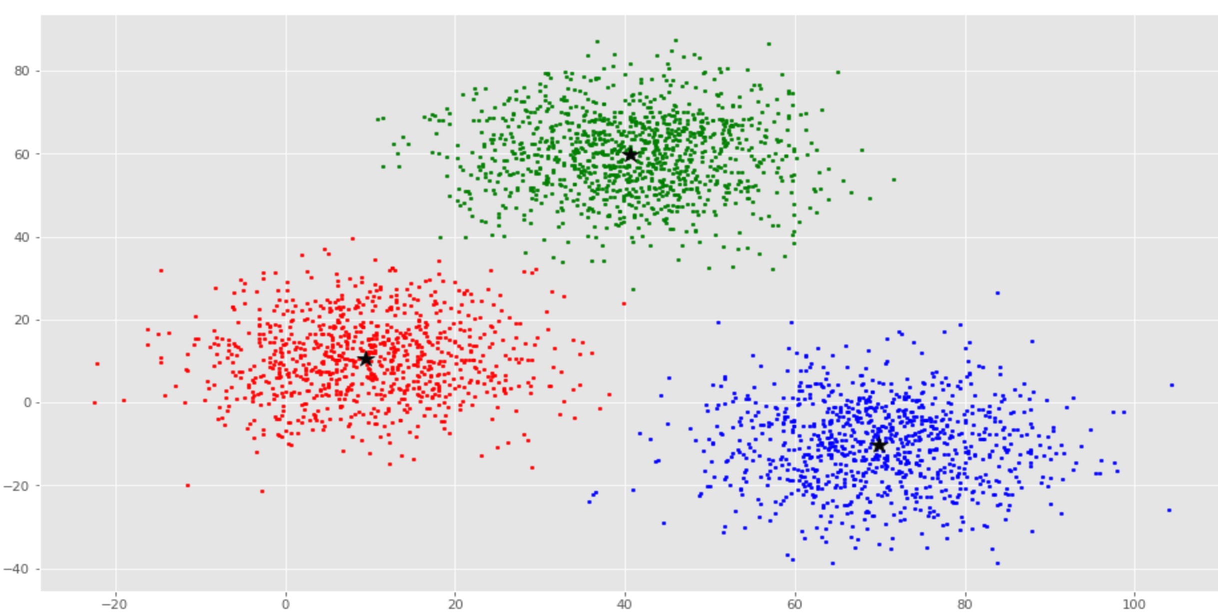
- Simplifying extremely large datasets by grouping features with similar values into a smaller number of homogeneous categories.

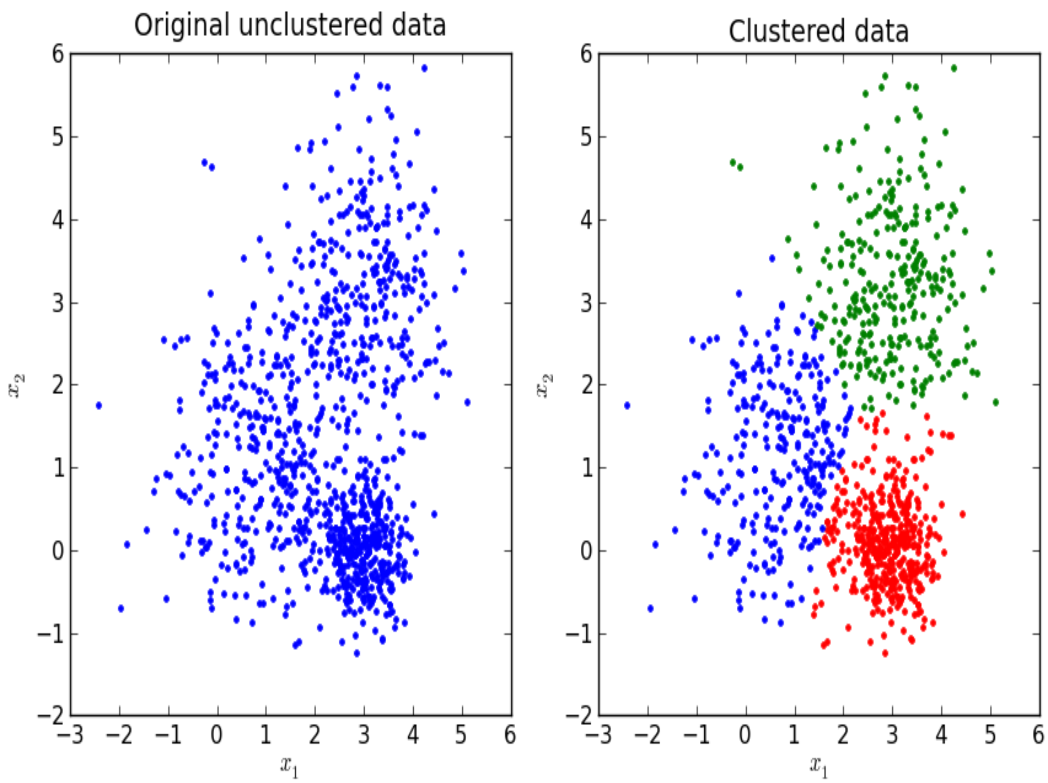
Overall, clustering is useful whenever diverse and varied data can be exemplified by a much smaller number of groups. It results in meaningful and actionable data structures that reduce complexity and provide insight into patterns of relationships.

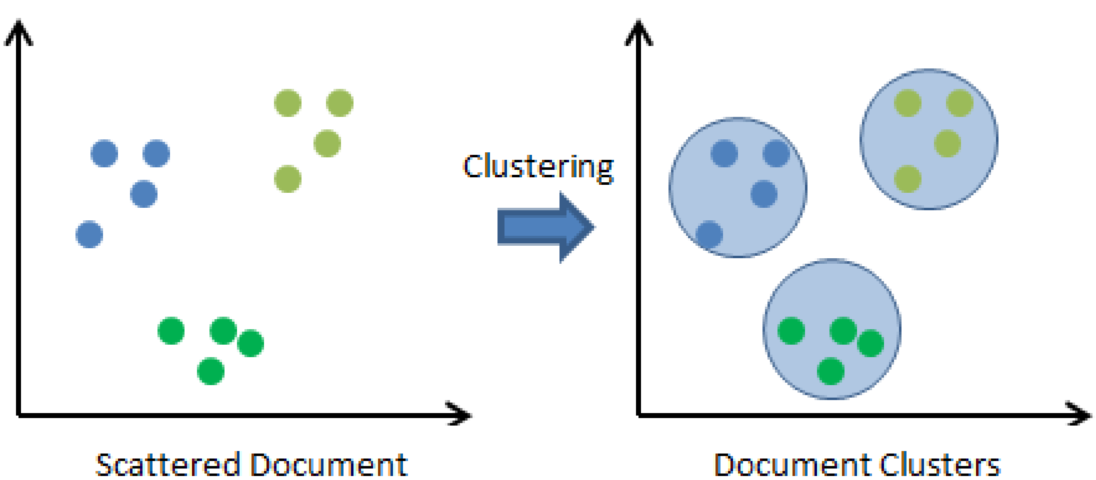
K-Means Clustering

After the necessary introduction, Data Mining courses always continue with K-Means; an effective, widely used, all-around clustering algorithm. Before actually running it, we have to define a distance function between data points (for example, Euclidean distance if we want to cluster points in space), and we have to set the number of clusters we want (k).

- Clustering is a type of **Unsupervised learning.**
- This is very often used when you don’t have labelled data.
- **K-Means Clustering** is one of the popular clustering algorithm.
- The goal of this algorithm is to find groups(clusters) in the given data.
- We have used python libraries (scikit learn) to implement K-means clustering.




Original unclustered data
Clustered data
x_1
x_2
−3 −2 −1 0 1 2 3 4 5 6
−2 −1 0 1 2 3 4 5 6


Clustering
Scattered Document
Document Clusters

VI. Labelling of Data

Labeled Data is a group of samples that have been tagged with one or more labels. Labeling typically takes a set of unlabeled data and augments each piece of that unlabeled data with meaningful tags that are informative. For example, labels might be indicate whether a photo contains a horse or a cow, which words were uttered in an audio recording, what type of action is being performed in a video, what the topic of a news article is, what the overall sentiment of a tweet is, whether the dot in an x-ray is a tumor, etc.

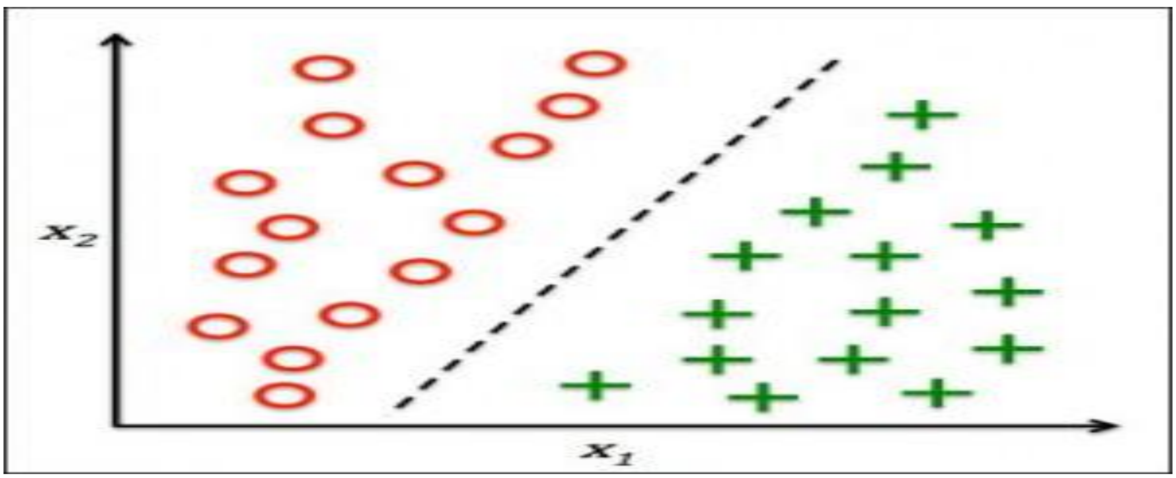
Labels can be obtained by asking humans to make judgments about a given piece of unlabeled data (e.g., "Does this photo contain a horse or a cow?"), and are significantly more expensive to obtain than the raw unlabeled data.

After obtaining a labeled dataset, machine learning models can be applied to the data so that new unlabeled data can be presented to the model and a likely label can be guessed or predicted for that piece of unlabeled data.

- We have clustered our dataset into three clusters. Now within each cluster we have to have label each record either as active low or active medium or active high.
- For doing this we have analysed our dataset and have done RECURSIVE FEATURE SELECTION.
- By doing RECURSIVE FEATURE ANALYSIS we have identified which features plays major role for our dataset analysis.
- We have found four features DC, RH, WIND and TEMP have more impact in our analysis.
- Using this features we have classify each records within each group as Active low, Active medium and Active high.
- We have given threshold values for each group.

VII. Classification

In machine learning and statistics, classification is the problem of identifying to which of a set of categories (sub-populations) a new observation belongs, on the basis of a training set of data containing observations (or instances) whose category membership is known. An example would be assigning a given email into "spam" or "non-spam" classes or assigning a diagnosis to a given patient as described by observed characteristics of the patient (gender, blood pressure, presence or absence of certain symptoms, etc.). Classification is an example of pattern recognition. An algorithm that implements classification, especially in a concrete implementation, is known as a classifier.



Classification is one of the most widely used techniques in machine learning, with a broad array of applications, including sentiment analysis, ad targeting, spam detection, risk assessment, medical diagnosis and image classification. The core goal of classification is to predict a category or class y from some inputs x. Through this course, you will become familiar with the fundamental models and algorithms used in classification, as well as a number of core machine learning concepts.

In Classification, We use different models which are as follows:

1. **<u>Logistic Regression</u>**:- It is a statistical method for analyzing a dataset in which there are one or more independent variables that determine an outcome. The outcome is measured with a dichotomous variable (in which there are only two possible outcomes).

2. **<u>Decision Tree</u>**:- A decision tree is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm that only contains conditional control statements.

3. **<u>Support Vector Machine(SVM):-</u>** An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible.

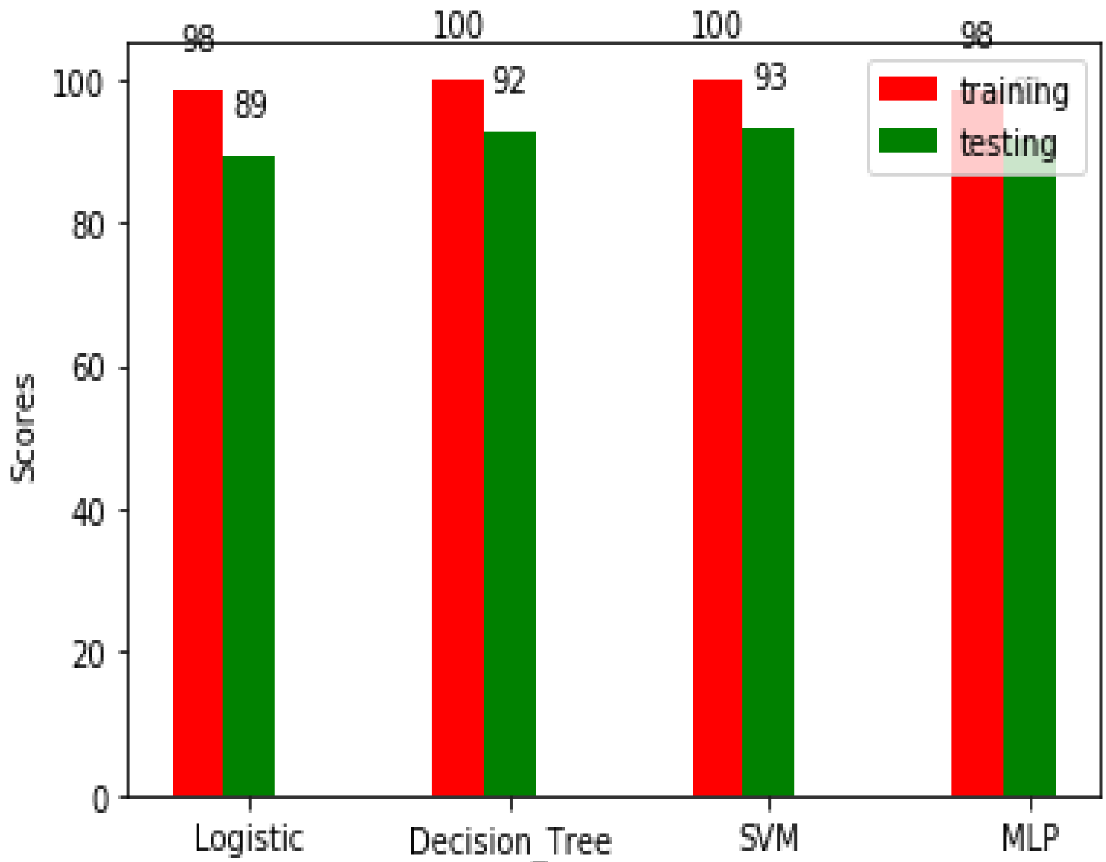
4. **<u>Multi-Layer Perceptron</u>:-**It is a class of feedforward artificial neural network. An MLP consists of at least three layers of nodes. Except for the input nodes, each node is a neuron that uses a nonlinear activation function. MLP utilizes a supervised learning technique called backpropagation for training.

VIII. Accuracy and Comparative analysis

Accuracy (in different models)

Models	Training Accuracy	Testing Accuracy
Logistic Regression	98.4	85.92
Decision Tree	100	92.71
SVM	100	93.2
MLP	98.6	82.52

Testing Dataset:- 361(70%)
Training Dataset:- 155(30%)


98
100
100
98
89
92
93
training
testing
100
80
60
40
20
0
Scores
Logistic
Decision_Tree
SVM
MLP

IX. Rule Based Classifier

Rule-based classifiers

- use a set of IF-THEN rules for classification
- if {condition} then {conclusion}
- if part - *condition* stated over the data
- then part - a class label, *consequent*

Each rule can be evaluated using these measures

- coverage: # of data points that satisfy conditions
- accuracy = # of correct predictions / coverage

Rule-based machine learning (RBML) is a term in computer science intended to encompass any machine learning method that identifies, learns, or evolves 'rules' to store, manipulate or apply. The defining characteristic of a rule-based machine learner is the identification and utilization of a set of relational rules that collectively represent the knowledge captured by the system. This is in contrast to other machine learners that commonly identify a singular model that can be universally applied to any instance in order to make a prediction.

- The main reason behind using this classifier is that efficiency is approximately same by using few attributes in comparison of using all attributes.

- Thus, it makes energy efficient and provides simplest selection to analyze the problem.

- It also reduces the Network Traffic by evaluating the selected attributes.

Models use in Rule Based Classifier

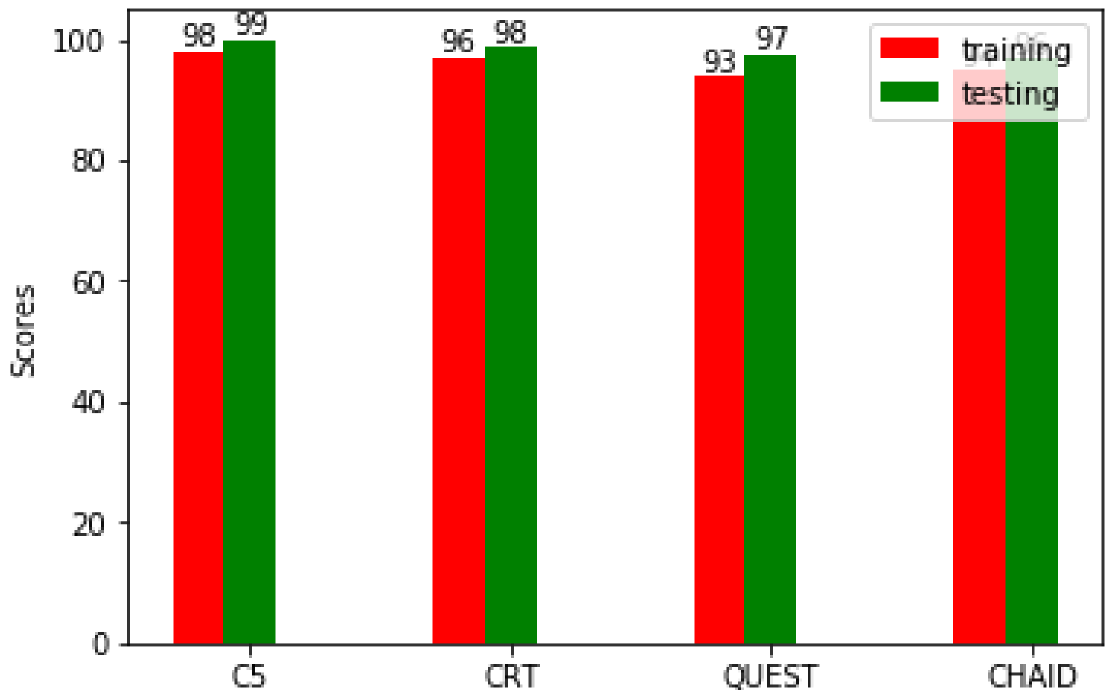
1. C5 Model
2. CRT Model
3. Chaid model
4. Quest Model

C5 Rule Set		**Rule (R)** **(No. of instances; Confidence factor)**
	Rules for Cluster 1 **R_{C5}1**	Rule 1 (5; 1.0) **R_{C5}11**
		Rule 2 (9; 1.0) **R_{C5}12**
		Rule 3 (4; 1.0) **R_{C5}13**
		Rule 4 (407; 1.0) **R_{C5}14**
		Rule 5 (5; 1.0) **R_{C5}15**
		Rule 6 (5; 1.0) **R_{C5}16**
	Rules for Cluster 2 **R_{C5}2**	Rule 1 (764; 1.0) **R_{C5}21**
		Rule 2 (8; 1.0) **R_{C5}22**
		Rule 3 (2; 1.0) **R_{C5}23**
	Rules for Cluster 3 **R_{C5}3**	Rule 1 (3; 1.0) **R_{C5}31**
		Rule 2 (208; 1.0) **R_{C5}32**
	Rules for Cluster	Rule 1 (3; 1.0) **R_{C5}41**
		Rule 2 (2; 1.0) **R_{C5}42**

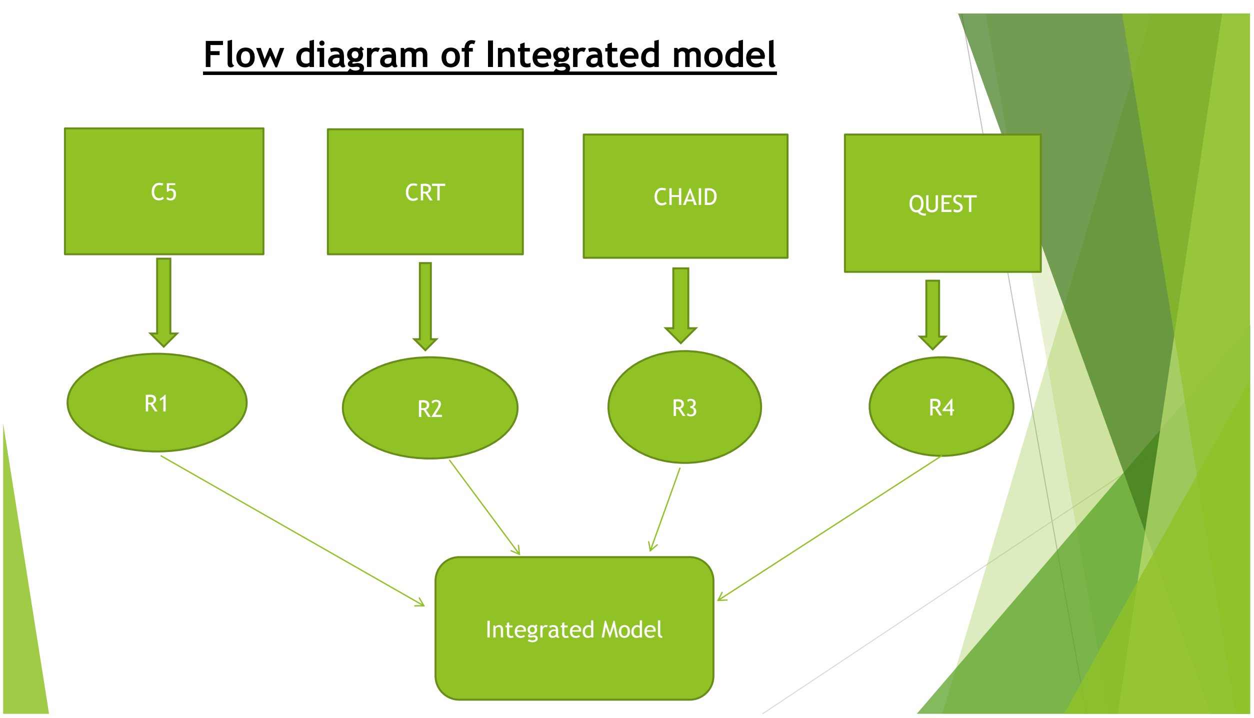
CRT Rule Set		Rule (No. of instances; Confidence factor)
R_{CRT}	Rules for Cluster 1 $R_{CRT}1$	Rule 1 (2; 1.0) $R_{CRT}11$
		Rule 2 (15; 0.6) $R_{CRT}12$
		Rule 3 (418; 0.995) $R_{CRT}13$
		Rule 4 (5; 1.0) $R_{CRT}14$
	Rules for Cluster 2 $R_{CRT}2$	Rule 1 (771; 1.0) $R_{CRT}21$
		Rule 2 (2; 1.0) $R_{CRT}22$
	Rules for Cluster 3 $R_{CRT}3$	Rule 1 (4; 0.75) $R_{CRT}31$
		Rule 2 (208; 1.0) $R_{CRT}32$
	Rules for Cluster 4 $R_{CRT}4$	Rule 1 (295; 0.976) $R_{CRT}41$
		Rule 2 (6; 1.0) $R_{CRT}42$
		Rule 3 (4; 1.0) $R_{CRT}43$
	Rules for Cluster 5 $R_{CRT}5$	Rule 1 (6; 0.833) $R_{CRT}51$
		Rule 2 (217; 0.991) $R_{CRT}52$
	Rules for Cluster 6 $R_{CRT}6$	Rule 1 (9; 0.667) $R_{CRT}61$
		Rule 2 (414; 1.0) $R_{CRT}62$
		Rule 3 (8; 0.625) $R_{CRT}63$
	Default: Cluster	

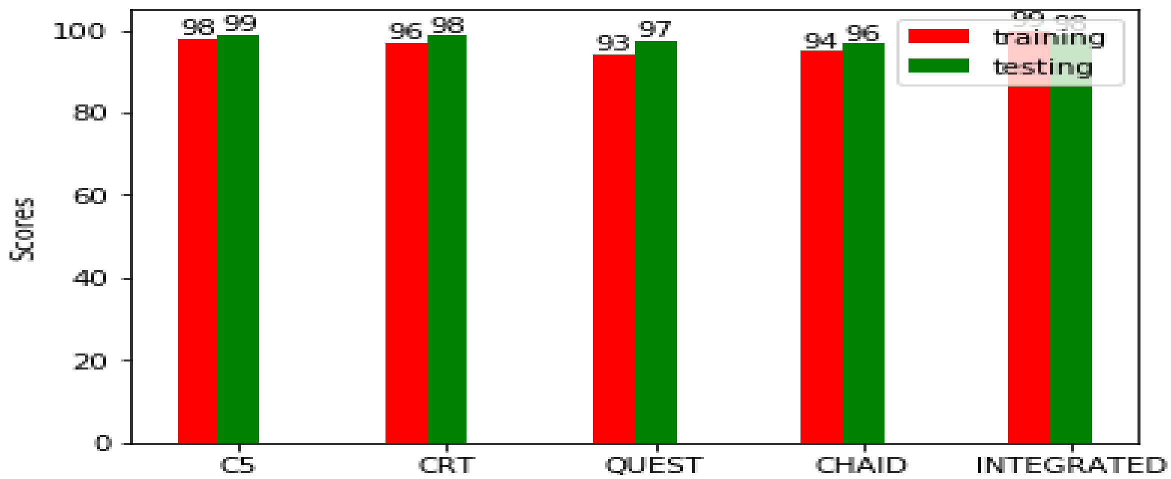
CHAID Rule Set		Rule (No. of instances; Confidence factor)
R_{CHAID}	Rules for Cluster 1 **R_{CHAID}1**	Rule 1 (189; 0.746) **R_{CHAID}11**
		Rule 2 (26; 1.0) **R_{CHAID}12**
		Rule 3 (238; 1.0) **R_{CHAID}13**
	Rules for Cluster	Rule 1 (715; 1.0) **R_{CHAID}21**
	Rules for Cluster	Rule 1 (238; 0.887) **R_{CHAID}31**
	Rules for Cluster 4 **R_{CHAID}4**	Rule 1 (239; 0.812) **R_{CHAID}41**
		Rule 2 (215; 0.507) **R_{CHAID}42**
	Rules for Cluster	Rule 1 (238; 0.752) **R_{CHAID}51**
	Rules for Cluster 6 **R_{CHAID}6**	Rule 1 (24; 1.0) **R_{CHAID}61**
		Rule 2 (238; 1.0) **R_{CHAID}62**
		Rule 3 (24; 0.833) **R_{CHAID}63**
	Default: Cluster	

QUEST Rule Set		Rule (No. of instances; Confidence factor)
	Rules for Cluster	Rule 1 (434; 0.961) R_{QUEST}**11**
	Rules for Cluster	Rule 1 (781; 0.986) R_{QUEST}**21**
	Rules for Cluster	Rule 1 (214; 0.963) R_{QUEST}**31**
	Rules for Cluster 4 R_{QUEST}**4**	Rule 1 (231; 0.944) R_{QUEST}**41**
		Rule 2 (81; 0.889) R_{QUEST}**42**
	Rules for Cluster	Rule 1 (206; 0.971) R_{QUEST}**51**
R_{QU}	Rules for Cluster	Rule 1 (437; 0.95) R_{QUEST}61
	Default: Cluster	

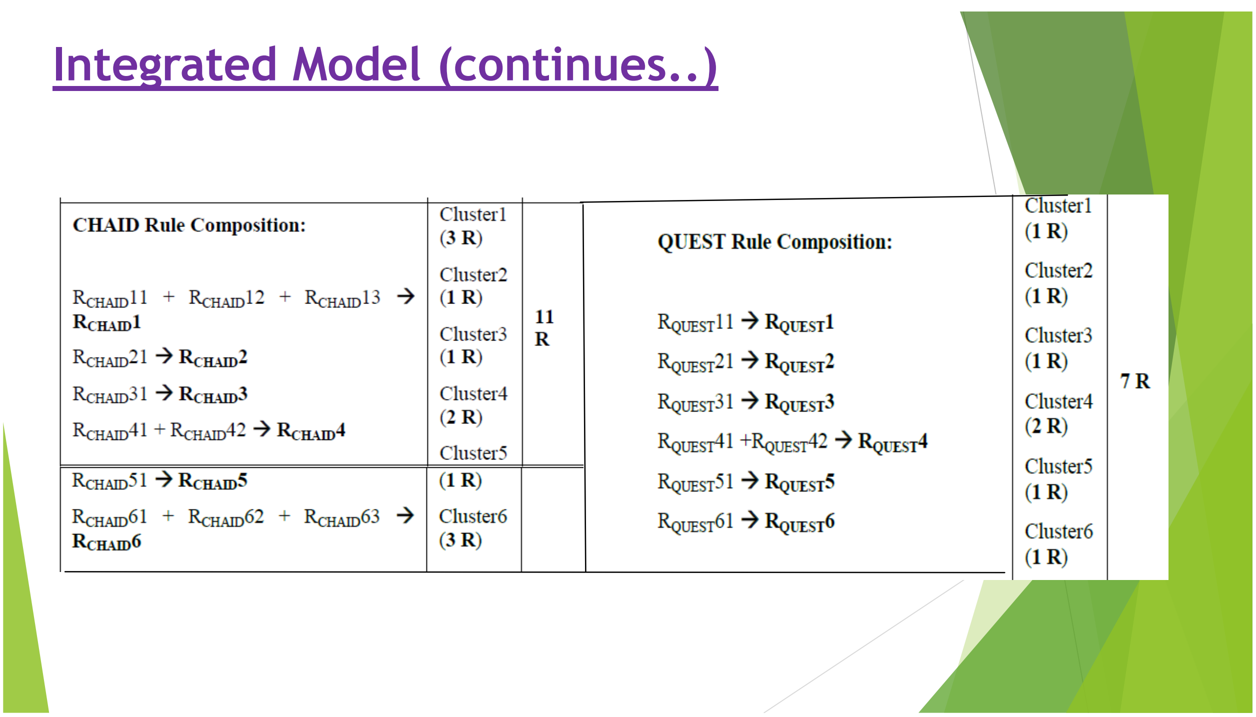


1.Using all features





2.Using attributes(DC,RH,TEMP,WIND)



Integrated Model (continues..)

CHAID Rule Composition:			QUEST Rule Composition:		
$R_{CHAID}11 + R_{CHAID}12 + R_{CHAID}13 \rightarrow$ **$R_{CHAID}1$**	Cluster1 (**3 R**)	**11 R**	$R_{QUEST}11 \rightarrow$ **$R_{QUEST}1$**	Cluster1 (**1 R**)	**7 R**
$R_{CHAID}21 \rightarrow$ **$R_{CHAID}2$**	Cluster2 (**1 R**)		$R_{QUEST}21 \rightarrow$ **$R_{QUEST}2$**	Cluster2 (**1 R**)	
$R_{CHAID}31 \rightarrow$ **$R_{CHAID}3$**	Cluster3 (**1 R**)		$R_{QUEST}31 \rightarrow$ **$R_{QUEST}3$**	Cluster3 (**1 R**)	
$R_{CHAID}41 + R_{CHAID}42 \rightarrow$ **$R_{CHAID}4$**	Cluster4 (**2 R**)		$R_{QUEST}41 + R_{QUEST}42 \rightarrow$ **$R_{QUEST}4$**	Cluster4 (**2 R**)	
$R_{CHAID}51 \rightarrow$ **$R_{CHAID}5$**	Cluster5 (**1 R**)		$R_{QUEST}51 \rightarrow$ **$R_{QUEST}5$**	Cluster5 (**1 R**)	
$R_{CHAID}61 + R_{CHAID}62 + R_{CHAID}63 \rightarrow$ **$R_{CHAID}6$**	Cluster6 (**3 R**)		$R_{QUEST}61 \rightarrow$ **$R_{QUEST}6$**	Cluster6 (**1 R**)	

X. Code Section:

In [ ]:
```
import pandas as pd
import numpy as np

c=0
```

In [ ]:
```
out=pd.read_csv('C:\\Users\\Rahul\\Downloads\\train.csv')
#out.head()
#out.columns.values
train=np.asarray(out)
cnt= 0
```

In [ ]:
```
for i in range(0,train.shape[0]):
    DC = train[i][2]
    RH = train[i][5]
    temp =train[i][4]
    DMC = train[i][1]
    FFMC = train[i][0]
    ISI = train[i][3]
    wind = train[i][6]
    if DC > 458.625 and DC > 545.618:
        c=c+1
    elif DC > 458.625 and RH <= 55.735:
        c = c+1
    elif DC > 458.625 and RH > 55.735 and temp <= 18.71 and temp > 16.864:
        c = c+1
    elif DC > 458.625 and RH > 55.735 and temp > 18.710:
        c = c+1
    elif DC > 458.625 and RH <= 55.735 and wind <= 6.477 and wind <= 4.469:
        c = c+1
    elif DC > 458.625 and RH <= 55.735 and wind <= 6.477 and wind > 4.469 and RH <= 48.083:
        c = c+1
    elif DC > 458.625 and RH > 55.735 and wind <= 2.399:
        c = 1
    elif DC > 458.625 and temp > 15.140  and temp <= 20.220 and temp <= 16.079 and temp > 15.588:
        c = 1
    elif DC > 458.625 and temp > 15.140 and temp <= 20.220 and temp > 16.079:
        c = c+1
    elif DC > 458.625 and temp > 15.140 and temp > 20.220:
        c = 1
    elif DC > 458.62 and temp > 15.140  and temp > 20.220:
        c = 1
    elif DC > 458.625 and RH <= 55.735 and wind <= 6.477 and wind > 4.469 and RH <= 48.083:
        c = 1
    elif DC > 458.625 and RH > 55.735 and wind <= 2.399:
        c = 1
```

```
        c = 1
    elif FFMC > 87.496 and FFMC > 90.440 and wind <= 5.745 and wind > 3.677 and FFMC > 91.385:
        c = 1
    elif FFMC > 87.496 and FFMC > 90.440 and wind > 5.745 and FFMC > 92.083:
        c = 1
    elif RH <= 68.333 and RH <= 59.821 and RH <= 44.983:
        c = 1
    elif RH <= 68.333 and RH <= 59.821 and RH > 44.983 and RH <= 57.250:
        c = 1
    elif temp <= 16.985 and temp > 14.287 and RH <= 55.827 and temp <= 15.990 and RH > 48.089:
        c = c  +1
    elif temp <= 16.985 and temp > 14.287 and RH <= 55.827 and temp > 15.990:
        c = 1
    elif temp > 16.985 and RH <= 58.962 and temp <= 20.831 and RH > 27.219:
        c = 1
    elif temp > 16.985 and RH <= 58.962 and temp > 20.831:
        c = 1
    elif temp <= 16.985 and temp > 14.287 and temp > 16.152:
        c = 1
    elif temp > 16.985 and temp <= 20.888 and temp > 17.609:
        c = 1
    elif DC > 458.625 and RH > 55.735 and RH <= 60.451 and FFMC > 91.691:
        c = 1
    elif DC > 458.625 and FFMC <= 87.741:
        c = 1
    elif DC > 458.625 and FFMC > 87.741 and FFMC > 89.071 and DC > 543.615:
        c = 1
    elif DMC > 65.344 and RH <= 57.907 and RH <= 43.647:
        c = 1
    elif DMC > 65.344 and RH <= 57.907 and RH > 43.647 and RH <= 52.575 and RH <= 50.571:
        c  = 1
    elif DMC > 65.344 and RH <= 57.907 and RH > 43.647 and RH > 52.575:
        c = 1
    elif DMC <= 65.344 and temp > 19.836:
        c = 1
    elif DMC > 65.344 and temp > 15.540:
        c = 1
    elif RH <= 68.333 and RH <= 59.821 and RH > 44.983 and ISI <= 6.878 and ISI > 5.506:
        c = 1
    elif RH <= 68.333 and RH <= 59.821 and RH > 44.983 and ISI > 6.878:
        c =1
    elif DMC > 53.700 and temp > 15.550 and FFMC > 90.600 and temp > 17.850:
        c = c +1
    elif DMC > 53.700 and temp > 15.550 and FFMC > 90.600 and temp > 17.850:
        c = 1
    elif temp > 15.850 and temp <= 18.250 and FFMC <= 91.800 and FFMC > 90.600 and FFMC <= 91.300:
        c = 1
```

All of the codes are present on Github:

https://github.com/puperfused/Forest-Fire-Dataset

XI. Use Cases

- Using the model so designed for use in routing protocol.

- Helps in early alerting system to the organization.

- Reducing the area being affected by fire.

XII. Future Work

1. Using this model in Routing protocol.

2. Design of model that requires minimum number of sensors for maximum coverage.

XIII. References

- https://archive.ics.uci.edu/ml
- http://www3.dsi.uminho.pt/pcortez/forestfires/
- https://machinelearningmastery.com/
- https://mubaris.com/2017/10/01/kmeans-clustering-in-python
- https://www.kaggle.com/datasets
- http://scikit-learn.org/stable/