**TABLE OF CONTENTS**

**CHAPTER NO TITLE PAGE NO**

1. **FOREST FIRES PREDICTIONS 4**
   1. Introduction 4
   2. Objectives of Research 5
   3. Problem statement 5
   4. Industry Profile 5
2. **REVIEW OF LITERATURE 7**
   1. Introduction 7
   2. Human Surveillance 7
   3. Satellite surveillance 7
   4. Infrared/Smoke Scanners 7
   5. Local Scanners 7
   6. Data Mining Techniques 8
   7. Summary 8
3. **DATA COLLECTION 9**

3.1 Dataset 9

3.2 Explanation 9

3.3 Data Describing 11

1. **METHODOLOGY- 13**

EXPLORATORY DATA ANALYSIS 13

1. **FINDINGS AND SUGGESTIONS 26**
2. **CONCLUSION 27**
3. **BIBLIOGRAPHY AND REFERENCE 28**

**LIST OF FIGURES**

**FIGURE NO TITLE PAGE NO**

3.1 First five rows in dataset9

3.2 Fire Weather Index Structure 10

3.3 The pre-processed data attributes 11

3.4 Describing Data 11

4.1 Boxplots 13

4.1.2 Co-relation & Co-variance 14

4.2 Info and Null Values 15

4.3 Scatter Plot 16

4.4 Pair plot 22

4.5 Distribution Plot 23

4.6 Heat Map 25

5.1 Findings 26

**LIST OF ABBREVIATIONS**

FFMC Fuel Fine Moisture Code

DMC Duff Moisture Code

DC Drought Code

ISI Initial Spread Index

RH Relative Humidity

**CHAPTER 1**

**FOREST FIRES PREDICTION**

* 1. **INTRODUCTION**

Forest fires are the major environmental issue, creating environmental and ecological damage while endangering human lives. Fast detection is the key element for such phenomena. To achieve this, one alternative is to use automatic tools based on local sensors, such as provided by the meteorological stations. In effect, meteorological conditions (temperature, wind) are known to influence forest fires and several fire indexes, such as forest fire weather index (FWI). In this work, we explore a machine learning approach to predict the burned area of forests. Two different machine learning techniques are used (Simple linear and multilinear regressions).

One major environmental concern is the occurrence of forest fires which affect forest preservation, create economical and ecological damage and cause human suffering. Such phenomena is due to multiple cases (human negligence, lightnings). Despite of increasing state expenses to control this disaster, each year millions of hectares(ha) are destroyed all over the world. Fast detection is the key element for successful firefighting. Since traditional human surveillance is expensive and affected by subjective factors, there has been an emphasis to develop automatic solutions.

Some of the solutions are as follows:

Satellite-based, infrared/smoke scanners and local sensors are used. Satellites have acquisition costs, localization delays and the resolution is not adequate for all cases. Moreover, scanners have an high equipment and maintenance cost. Weather conditions, such as temperature and air humidity, are known to affect fire occurrence.

In the past, meteorological data has been incorporated into numerical indices, which are used for prevention and to support fire management decisions. In particular, the Canadian forest Fire Weather Index (FWI) system was designed in the 1970s when computers were scarce, thus it required only simple calculations using look-up tables with readings from four meteorological observations (i.e. temperature, relative humidity, rain and wind) that could be manually collected in weather stations. Nevertheless, nowadays this index highly used not only in Canada but also in several countries around the world (e.g. Argentina or New Zealand). Even though Mediterranean climate differs from those in Canada, the FWI system was correlated with ﬁre activity in southern Europe countries, including Portugal.

In contrast with these previous works, we present a novel machine learning approach where we can predict the amount area that can be burnt when forest fires takes place according to some input factors.

* 1. **OBJECTIVES OF RESEARCH**

The main purpose of this project is to solve one of the major problems of the present day natural disaster that is, forest fire. The problem can be solved with the help of a machine learning model that can predict how much area will be burnt with respect to the input factors we have taken. So here, we consider some of the input features like temperature, relative humidity, spatial coordinates of the area we are considering, wind, rain and other FWI components such as FFMC, DMC, DC, ISI.

**1.3 PROBLEM STATMENT**

In the present day, our major environmental concern is the occurrence of forest fires. So to reduce the damage of forests, we beforehand try to predict the that may be burnt if fire occurs using machine learning techniques.

* 1. **INDUSTRY PROFILE**

The goals and objectives of this project are as follows:

* To create a machine learning model which would predict how much area may be burned according to the input factors

The scope of this project could be extended to many different fields like forest departments, environmental sectors where meteorological data is stored or the data is given as input and the machine learning model can help in predicting how much area will be burned.

**CHAPTER -2**

**REVIEW OF LITERATURE**

**2.1 INTRODUCTION**

There are many types of models for predicting forest fires which are currently in market. In this chapter, each of the previous work related to this project will be discussed which are summarized below.

**2.2 HUMAN SURVEILLANCE**

Fast detection is the key element for a successful firefighting. Traditional human surveillance is expensive and affected by subjective factors. Therefore there has been an emphasis to develop automatic solutions.

**2.3 SATELLITE SURVEILLANCE**

Satellites have acquisition costs, localisation delays, and the resolution is not adequate for all cases. Sometimes the area it is predicting may not be accurate. Therefore, we try to opt other models in the coming future.

**2.4 INFRARED/SMOKE SCANNERS**

Scanners have high equipment and maintenance costs. Weather conditions such as temperature and wind are known to affect the fire occurrence. Since automatic meteorological stations are often available, such data can be collected in real-time at low costs.

**2.5 LOCAL SCANNERS**

Inall cases. Moreover, scanners have a high equipment and maintenance costs. Weather conditions, such as temperature and air humidity, are known to affect ﬁre occurrence. Since automatic meteorological stations are often available (e.g. Portugal has 162 ofﬁcial stations), such data can be collected in real-time, with low costs.

In the past, meteorological data has been incorporated into numerical indices, which are used for prevention (e.g. warning the public of a ﬁre danger) and to support ﬁre management decisions (e.g. level of readiness, prioritizing targets or evaluating guidelines for safe ﬁreﬁghting). In particular, the Canadian forest Fire Weather Index (FWI) system was designed in the 1970s when computers were scarce, thus it required only simple calculations using look-up tables with readings from four meteorological observations (i.e. temperature, relative humidity, rain and wind) that could be manually collected in weather stations. Nevertheless, nowadays this index highly used not only in Canada but also in several countries around the world (e.g. Argentina or New Zealand). Even though Mediterranean climate differs from those in Canada, the FWI system was correlated with ﬁre activity in southern Europe countries, including Portugal.

**2.6 DATA MINING TECHNIQUES**

The interest in Data Mining (DM), also known as Knowledge Discovery in Databases (KDD), arose due to the advances of Information Technology, leading to an exponential growth of business, scientiﬁc and engineering databases. All this data holds valuable information, such as trends and patterns, which can be used to improve decision making. Yet, human experts are limited and may overlook important details. Moreover, classical statistical analysis breaks down when such vast and/or complex data is present. Hence, the alternative is to use automated DM tools to analyse the raw data and extract high-level information for the decision-maker .

Indeed, several DM techniques have been applied to the ﬁre detection domain. For example, Vega-Garcia et al, adopted Neural Networks (NN) to predict human caused wildﬁre occurrence. Infrared scanners and NN were combined in to reduce forest ﬁre false alarms with a 90% success.

**2.7 SUMMARY**

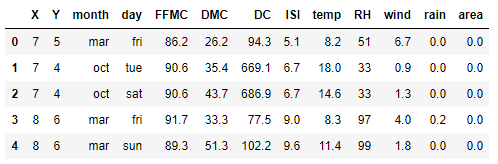
In contrast with these previous works, we present a novel ML forest fire approach, where the emphasis is the use of real time and non-costly meteorological data.

**CHAPTER-3**

**DATA COLLECTION**

**3.1 DATASET**

The data set we are using in this model to determine our project whereas follows.



**Fig 3.1- First five rows in dataset**

This dataset contains the columns which may or may not affect the target variable. Some of the columns are explained below-

**3.2 EXPLANATION**

The forest fire weather index is a Canadian system for rating fire danger and it includes six components.

FFMC - (fine fuel moisture code) denotes the moisture content surface litter a

And influences the ignition and fire spread

DMC&DC- (duff moisture code, drought code) represent the moisture content

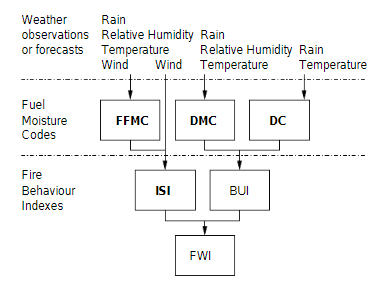
Shallow and deep organic layers which affect fire intensity.

ISI - (Initial spread index) the ISI is a score that correlates with fire spread

Velocity

RH – Relative Humidity

The FWI index is an indicator of ﬁre intensity and it combines the two previous components. Although different scales are used for each of the FWI elements, high values suggest more severe burning conditions. Also, the fuel moisture codes require a memory (time lag) of past weather conditions: 16 hours for FFMC, 12 days for DMC and 52 days for DC.



**Fig 3.2- Fire Weather Index Structure**

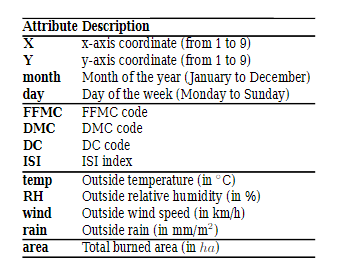
**(here we do not consider BUI and FWI)**

This study will consider forest fire data from a random forest department. The dataset contains X and Y columns which are the spatial coordinates of the forest area given. The X values range from 1 to 9 and Y values range from 2 to 9. The month and day column are the only categorical data in the dataset. It just show on which day and month how much area got burned depending on other columns. These 2 columns range with the values January to December and Monday to Friday respectively.

The FFMC column consists of a maximum value of 91.6 and 92.1. This column denotes the moisture content surface litter and influences ignition and fire spread. The DMC has a maximum values of 99.0 and a minimum values of 3.2. DMC denotes the moisture content of the deep and organic layers. The DC column contains a maximum values of 745.3 and a minimum values of many instances.

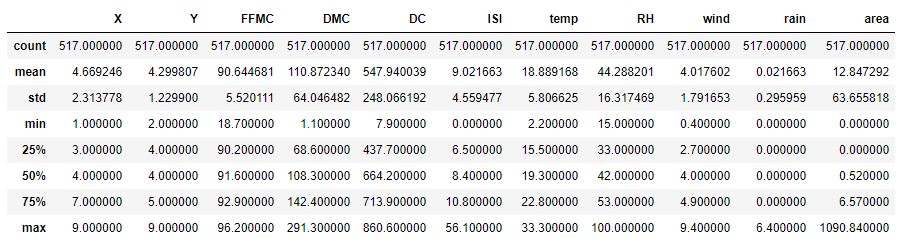
The ISI column contains the initial spread index value which tells the velocity of the fire spread. Other columns like rain, wind, temperature depicts, how much rainfall has occurred on that particular day in mm₂. Wind column contains the speed of the wind and the temperature can be of any value in degrees. Here we have to note that the rain column may contain the value 0, that means that there is no rainfall in that particular period. Here main point to be noted is 0 is not the missing data.

The target variable in our data set is area, that is our model predicts the amount of area burned when we give the input. Here the area column also consists of 0, it means that there is no area burnt in that specific conditions.

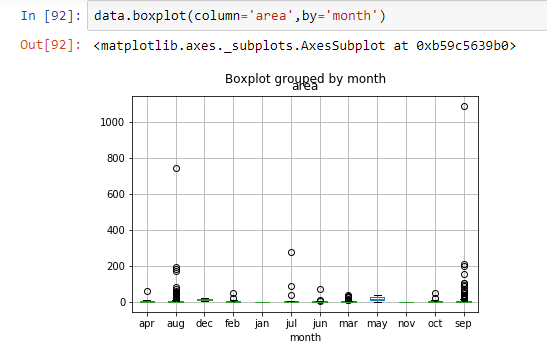


**3.3 Data Describing**

Here, now let us consider the description of the data, that is, get the max, min, count, 25%, 50%, 75%, mean.



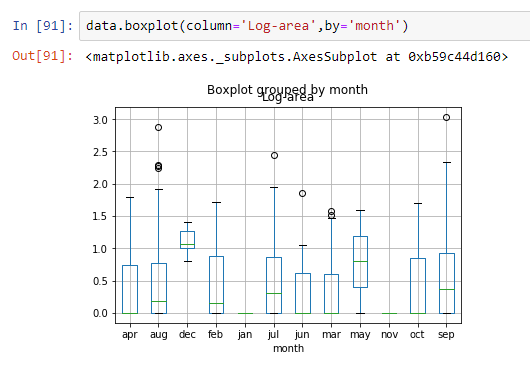
**Fig 3.3- The preprocessed Data Attributes**

****

**Fig 3.4.1- describing data**

Here we have taken every attribute from figure 3.3 , as each attribute is impacting the target variable (area- i.e., Burned area) . From fig 3.4.1, When we have drawn the box plot between the month and the target variable, we will get conjusted and compressed values as the target variable is having multiple inputs i.e from (0 to 1090.84). So for better view and understanding of boxplots between taget variabes and input variable, here we have reduced the target variable values using log formula ( Fig -3.4.2 )

**data['Log-area']=np.log10(data['area']+1)**

****

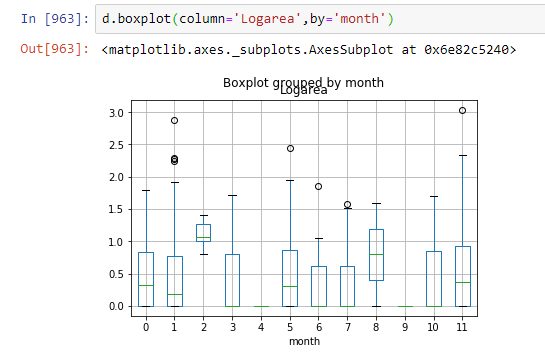
**Fig 3.4.2- describing data**

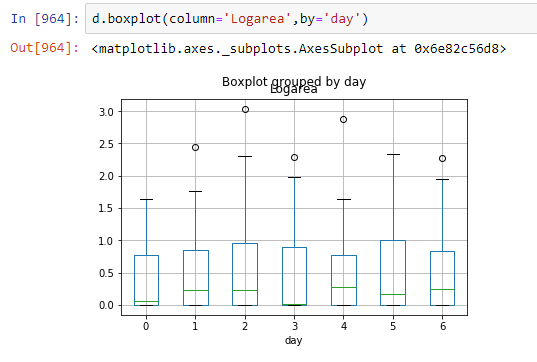
Here we have removed the outliers for every predictor variable

**CHAPTER-4**

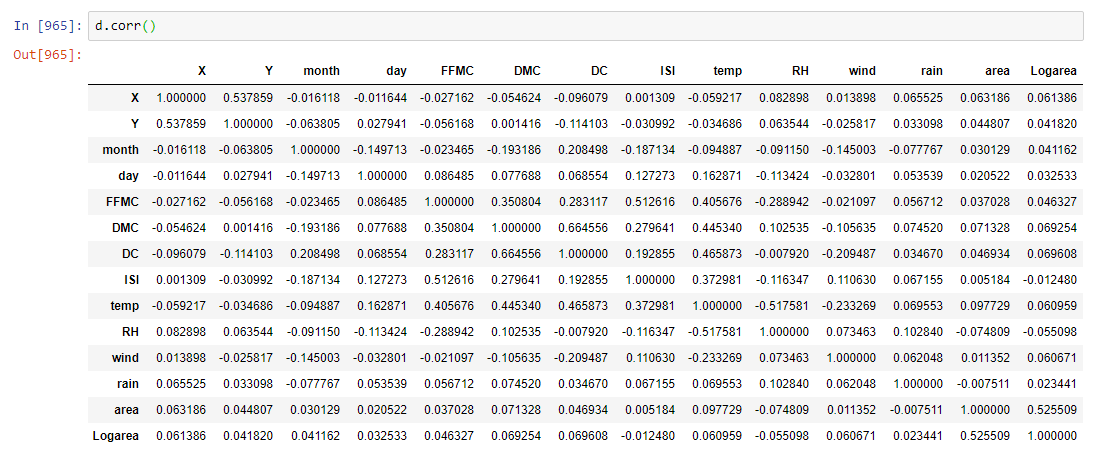
**EXPLORATORY DATA ANALYSIS**

**4.1 Boxplots**

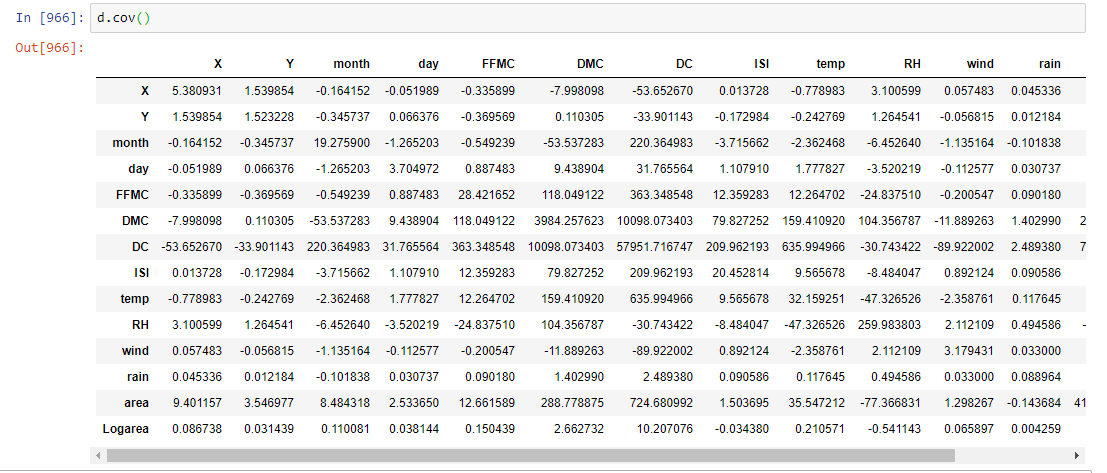
****

****

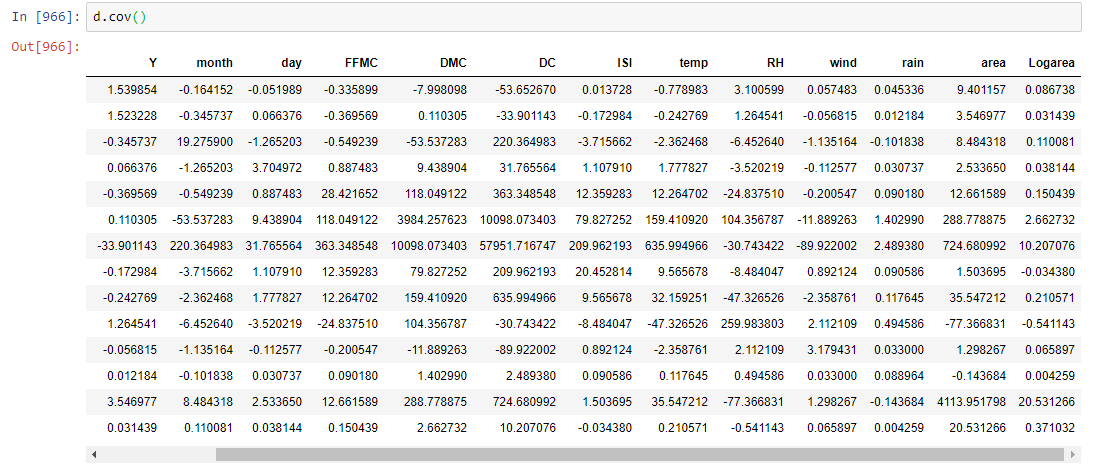
**4.2. Co-relation and Co-variance**

****

**Fig 4.2.1 Co-relation**

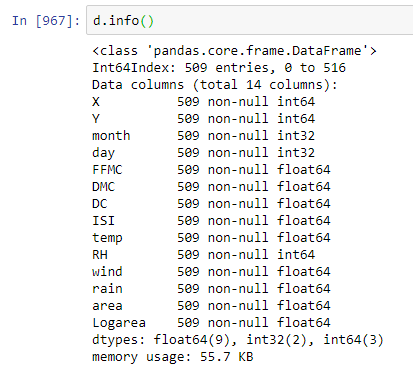


**Fig 4.1.2 Co-variance**

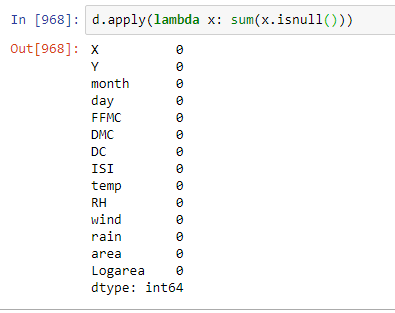
****

**Fig 4.1.3 Co variance**

**4.2 Info and Null values**

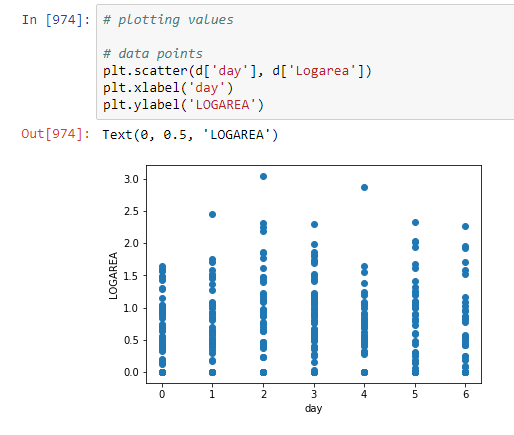
****

**Fig 4.2.1 Info**

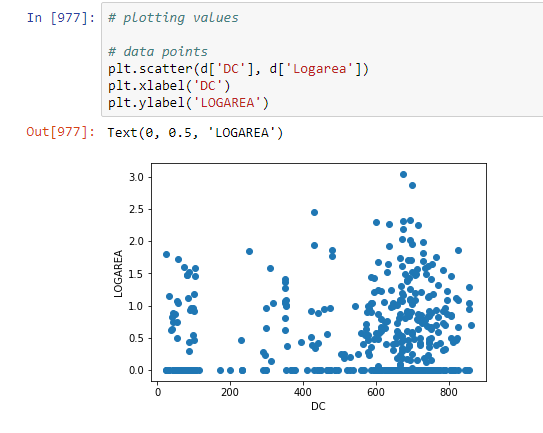
****

**Fig 4.2.2 Null values**

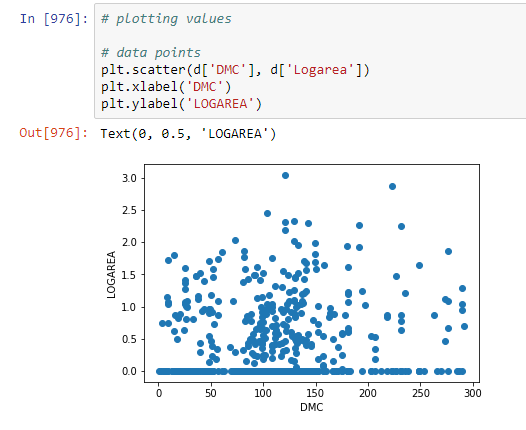
**4.3 Scatter plot between predictors and target variable**

****

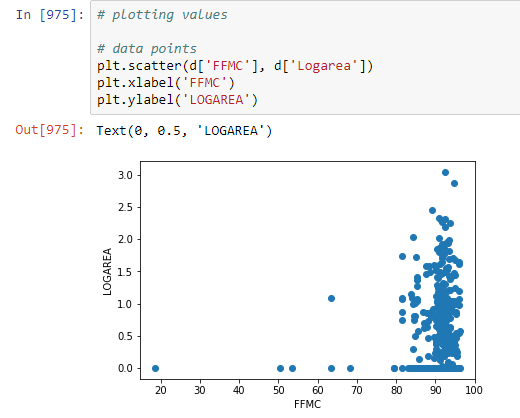
**Fig 4.3.1 Plot between Log area and day**

****

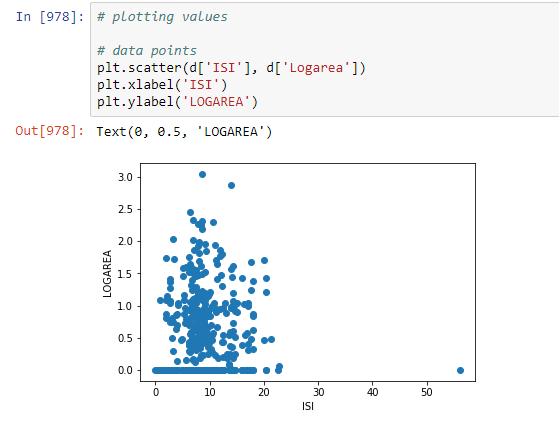
**Fig 4.3.2 Plot between Log area and DC**

****

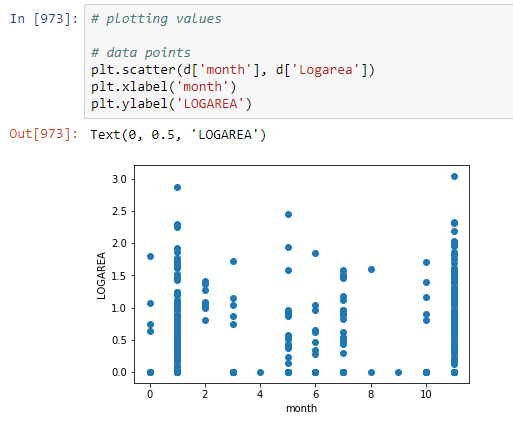
**Fig 4.3.3 Plot between Log area and DMC**

****

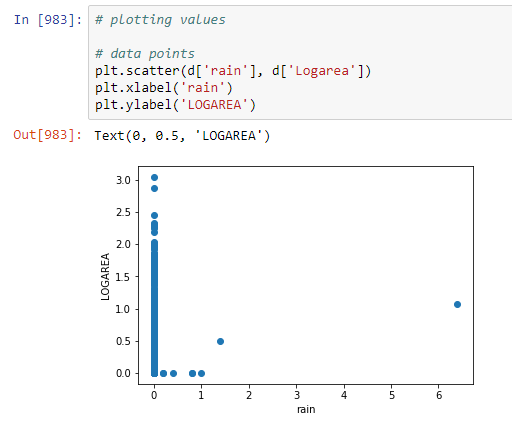
**Fig 4.3.4 Plot between Log area and FFMC**

****

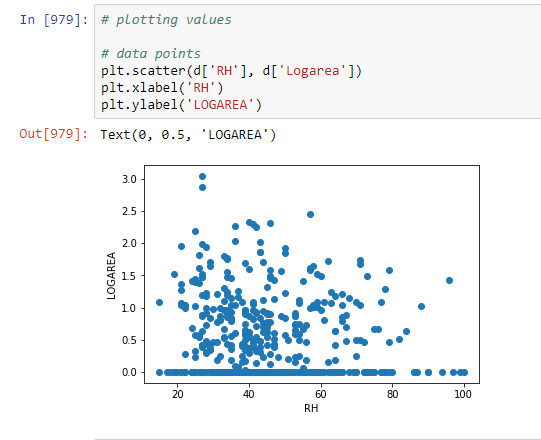
**Fig 4.3.5 Plot between Log area and ISI**

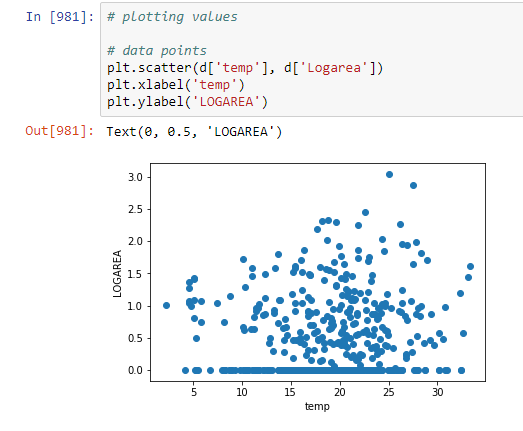
****

**Fig 4.3.6 Plot between Log area and MONTH**

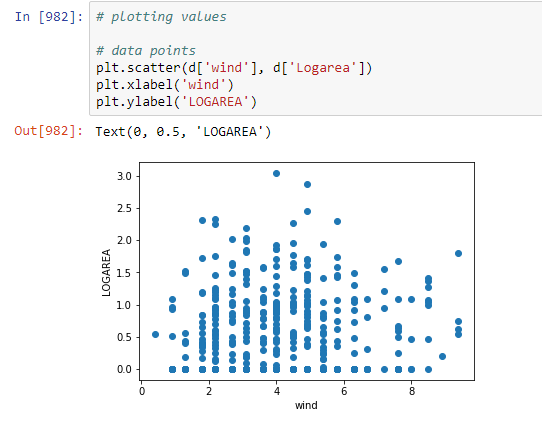
****

**Fig 4.3.7 Plot between Log area and rain**

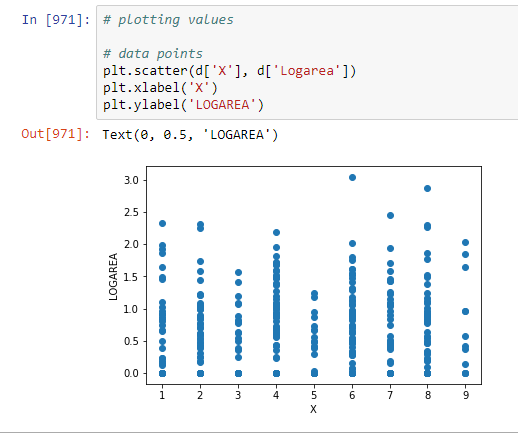
****

**Fig 4.3.8 Plot between Log area and RH**

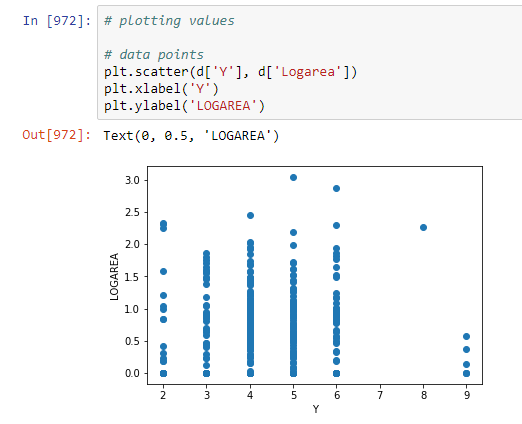
**Fig 4.3.9 Plot between Log area and temp**

****

**Fig 4.3.10 Plot between Log area and wind**

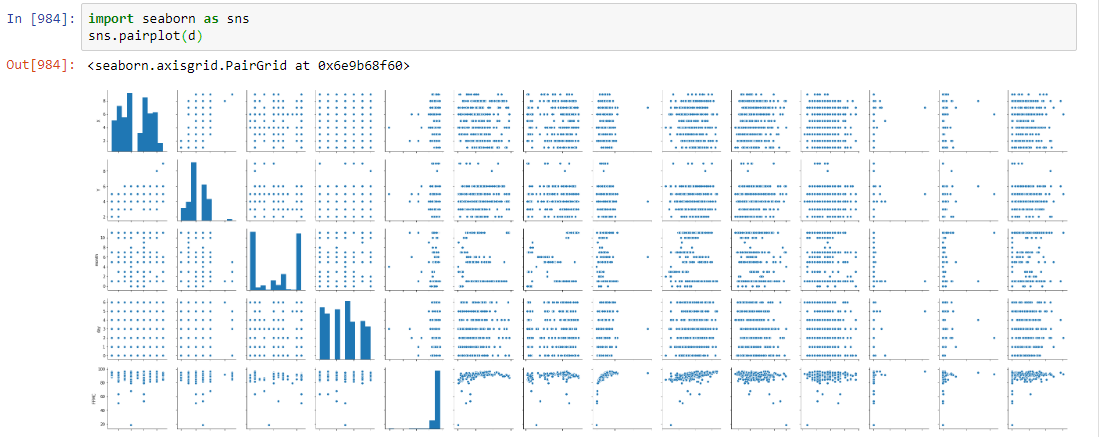
****

**Fig 4.3.11 Plot between Log area and X**

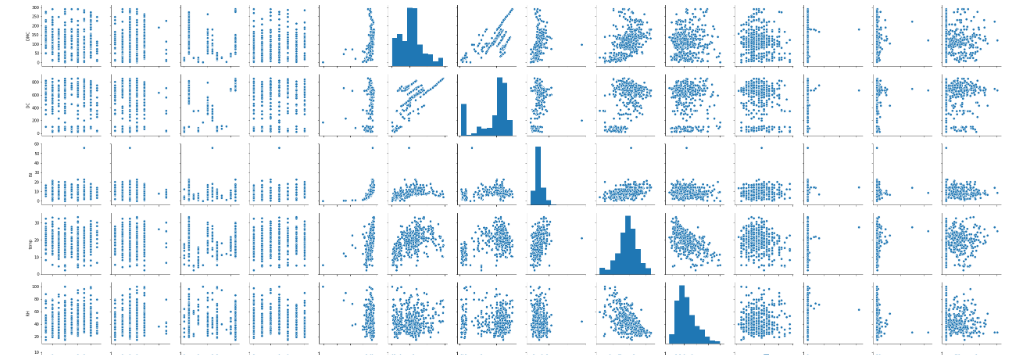
****

**Fig 4.3.12 Plot between Log area and Y**

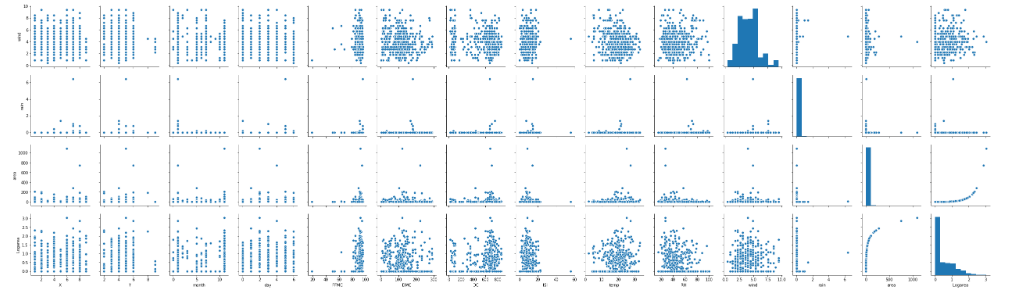
**4.4 Pairplot**

****

**Fig 4.4.1 pairplot for predictors and target variables**

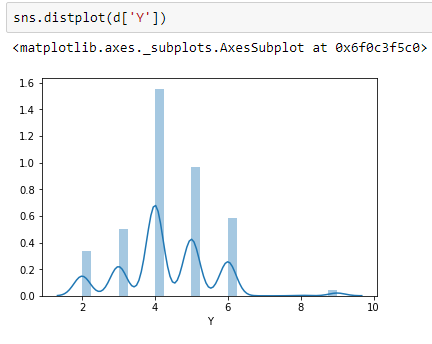
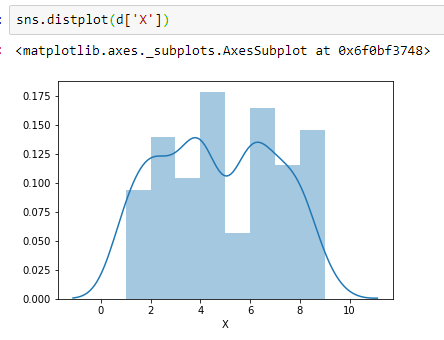
****

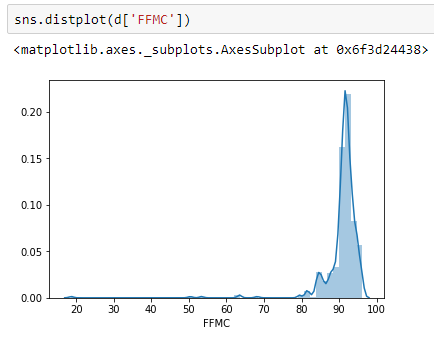
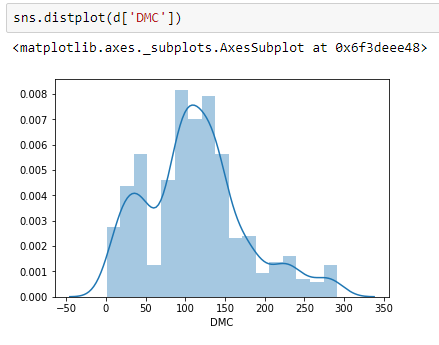
**Fig 4.4.2 pairplot for predictors and target variables**

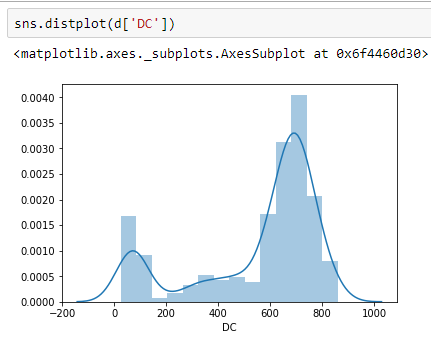
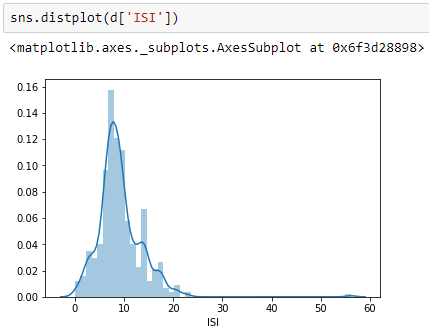
****

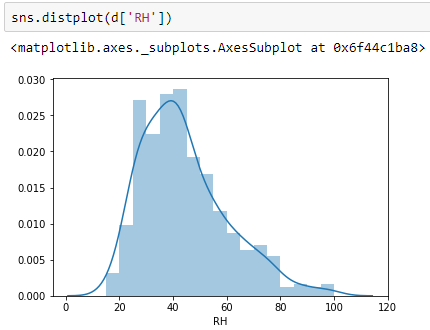
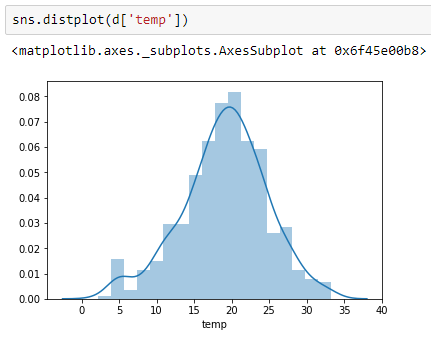
**Fig 4.4.3 pairplot for predictors and target variables**

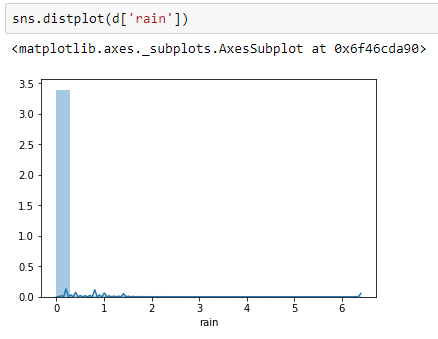
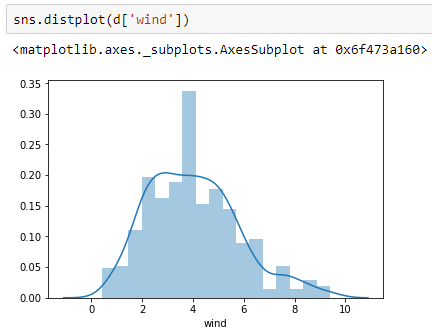
**4.5 Distribution Plots**

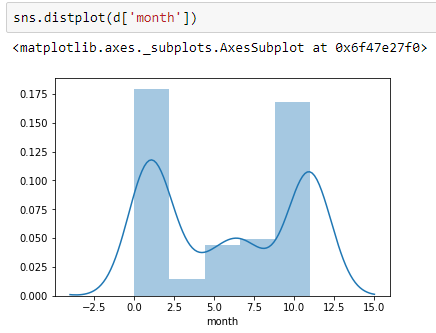
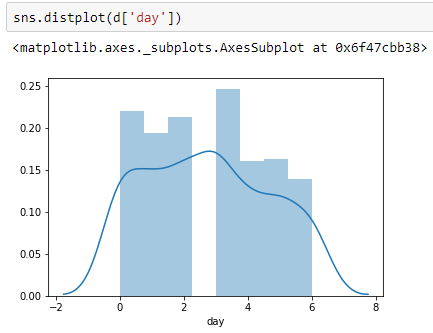
** **

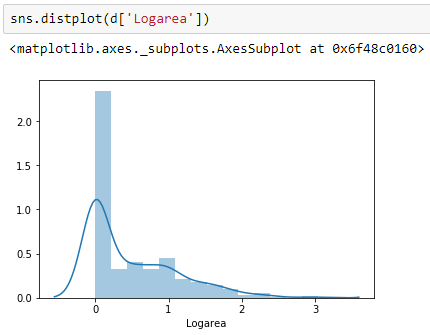
** **

** **

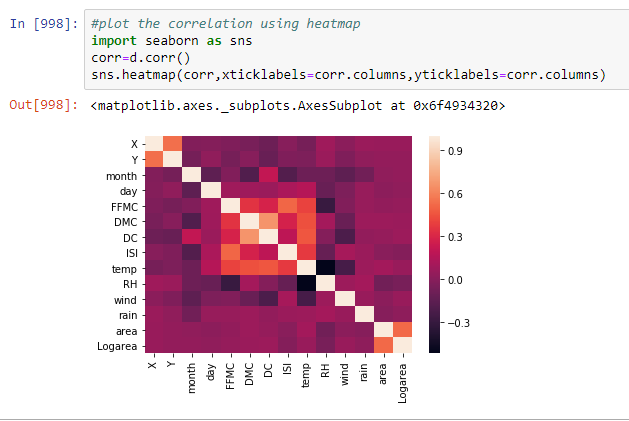
** **

** **

****  ****

****

**4.6 Heat Map**

****

**CHAPTER 5**

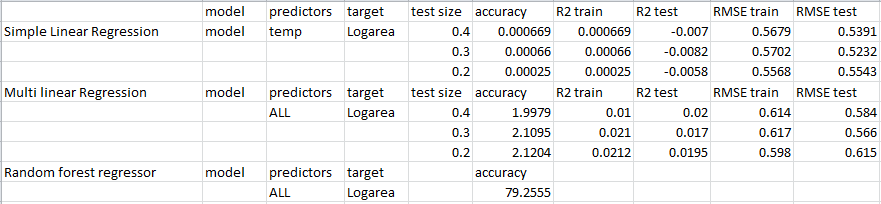
**FINDINGS AND SUGGESTIONS**

Here the data set has Continuous Data. So we have used 3 types of models

1. Simple Linear Regression Model

2. Multiple Linear Regression Model

3. Random Forest Regressor Model



**Fig 5.1 Findings**

From Fig 5.1 , We can say that Random Forest Regression model is the best suggested model for the Given Forest Fires Data Set as it has highest accuracy of 79%.

**CHAPTER 6**

**CONCLUSION**

From the Exploratory Data Analysis , Linear Regression Model and Random Forest Resgressor Model we can say that the Best way to Predict the area of forest burned is done by Random Forest Regressor Model, where we can get 79 % of correct predictions.

**CHAPTER 7**

**BIBLIOGRAPHY AND REFERENCES**

**Web References :**

1. <https://github.com/tirthajyoti/Machine-Learning-with-Python/blob/master/Real%20Dataset%20Problems/Forest_Fire_Prediction.ipynb>
2. <https://www.researchgate.net/publication/238767143_A_Data_Mining_Approach_to_Predict_Forest_Fires_using_Meteorological_Data>