Predicting the burned area in forest using Machine learning techniques

MSc Research Project
Data Analytics

SRINIVAS RAMASUBRAMANIAN
X16138805

School of Computing
National College of Ireland

Supervisor: Thibaut Lust

National College of Ireland
Project Submission Sheet – 2017/2018
School of Computing

Student Name:	SRINIVAS RAMASUBRAMANIAN
Student ID:	x16138805
Programme:	Data Analytics
Year:	2016
Module:	MSc Research Project
Lecturer:	Thibaut Lust
Submission Due Date:	11/12/2017
Project Title:	Predicting the burned area in forest using Machine learning techniques
Word Count:	5460

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

Signature:	
Date:	11th December 2017

PLEASE READ THE FOLLOWING INSTRUCTIONS:
1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. Please do not bind projects or place in covers unless specifically requested.
3. Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Contents

Predicting the burned area in forest using Machine learning techniques

SRINIVAS RAMASUBRAMANIAN
X16138805
MSc Research Project in Data Analytics

11th December 2017

Abstract

Predicting the amount of land burnt during forest fires is one of the challenging tasks.Forest fire causes serious damage to the flora and fauna of a country.This is one of major environmental issues which can also affect the economy of a country.Early prediction of fires saves large number of flora and fauna and prevents the ecosystem.By predicting the area burnt we can also classify whether the fire into small or big.The key motivation for this prediction is to help fire management team in proper resource allocation and to help the firefighters in a best possible way.The meteorological conditions of the forest are the key factors of the forest fire. These climatic data can be obtained using the local sensors which are incorporated in the nearest meteorological stations. This research proposes various Machine learning approaches such as Linear regression, logistic regression,SVR,Random forest,Gradient boosting and Bagging for predicting the amount of land burnt in the forest.Here the predictive model is build using the outbreaks of fire caused in the northeast region of Portugal.

1 Introduction

The wildfire is one of the serious events that needs to be controlled which can cause enormous damage.During summer forest fire is known to be a prominent event (Mateus and Fernandes; 2014).Around 4 million hectares of land were known to be burned every year and in the Mediterranean belt an average of 550 ha of land were burned (Ammann et al.; 2001).Smoke that emitted from the burned forest can serious heath issues such as nausea,mental illness,nausea,heart attack and even death (Ammann et al.; 2001).The fire fighters are the people who are exposed to these gases easily so they should be provided with all the necessary safety measures.They are the one who fights to save the people life and ecosystem of the forest.From 1990 to 2006 21.9 % of fire fighters death occurred due to heart attack (Mangan; 2007).The main idea of this paper is to help the (FMS) fire management system and the earlier prediction of fire in the forest help the forest managers in proper resource allocation such as providing enough air tankers and ground crews.There are many factors that can influence forest fire such as human intervention, Climatic changes, Forestry operations, Power line failure Improper forest policies , Lack of knowledge among the people to prevent the forest (Boubeta et al.; 2016).This approach uses the FWI system in Canada, it requires only small calculations and vlookup tables from the variables such as temperature,Wind speed,Relative humidity and rainfall.These variables details can be collected easily from nearby weather stations.Countries like Argentina and New zealand started using this FWI system (Taylor and Alexander; 2006).With the use of this historical data incorporated with DM techniques help in the analysis of forecasting the future fires and also for estimating the amount of area that will burn.The application of this data mining predictive model helps in analyzing the logical statements if the fire occurred in region Y then its most likely to spread towards region x.In this approach the data is modelled as a regression task where various algorithms are used and its efficiency is calculated by splitting the data into train and test

2 Related Work

Many Data mining and Machine learning techniques has been developed for the early detection of forest fire and estimation of the burned area in the forest.Forest fire is one major environmental issue that can cause serious damage to the ecosystem in various countries.The CFFDRS (Stocks et al.; 1989) is one of the fire danger rating system and it has been under development since 1968.The first introduced subsystem of CFFDRS is (FWI) which provides the numerical readings purely based on the weather.In 1990 the second major subsystem was evolved by CFFDRS called Fire behaviour system (FBP).The (FOP) The fire occurrence prediction is the recently introduced sub-system by CFFDRS (Stocks et al.; 1989).These are the systems used in the area of forest fire

2.1 A review of papers from 2005-2010

The primary cause of fire is smoke in 2005 with the use of satellite images to collect the statistics about smoke a model was developed.this model uses the SVM pixel classifier to identify the smoke mazzoni2005using.This classifier model provides around 75 % accuracy in finding the smoke over the region in north america.(Yu et al.; 2005) proposed a solution using wireless sensor for detecting the forest fire. This method mainly uses the wireless sensor instead of satellite based detection approach.In this approach a a large number of sensors are placed in the forest which is used to measure the environment temperature, relative humidity and smoke. Sensor node here captures all the information related to the environment conditions and location details.These details are shared to the cluster head.The neural network method is used to calculate the weather index and sends it to the manager node.The manager node is connected to sink using a wired network.This wired network is used to detect the wind speed.The data collected at the each node as packets.Once the cluster gets the packets from the node the data is processed.The evaluation of the performance is done by using in network processing.The main drawback in this model is failure of the sensor and cost involved in deploying these sensors in the forest. (Lafarge et al.; 2005) presented a textural kernel for SVM classification in forest fire detection.The important feature of this kernel is ability to weight the importance of texture by use of a tuning parameter.This ability helps to find various kinds of objects.The SVM here is used is to find the best classifier which is mainly produces the biggest margin of security.In this method the textural kernal is constructed using a markovian modeling.The radiometric information is balanced with the textural information using this kernel.the training set here used is a piece of cloud and a piece of smoke.The piece of cloud is used in the training set because the classifier should able to classify the smoke perfectly because both the smoke and cloud has a same texture.The main advantage is there is no over detection when there is a smokeless image.This model can find large range of patterns.However the drawback of the model here is it should be trained with wide range of training set images.In 2006 (Stojanova et al.; 2006) proposed a data mining method for find the forest fire detection.Here 3 different regions dataset was used because all the regions were under different climatic conditions. (Stojanova et al.; 2006) uses (J48) decision trees , logistic regression and bagging decision trees.The data is divided into 3 groups ALADIN data which contains the transpiration,speed and direction of the wind,humidity,temperature and sum energy.The ALADIN data is generated for every 3 hours which contains all ten weather attributes.The second group is the MODIS data which mainly dependence from day of the year.The third group contains the GIS data.This GIS data has all the time attributes related to the specific quadrant.Data mining tool WEKA is used in this methodology by (Stojanova et al.; 2006) for calculating the precision,Accuracy,Kappa and recall values of the algorithms. The validation of the predictive model is done by running the 10 fold cross.The two main uses of logistic regression is prediction of group membership and it also shows the importance of specific variable.In logistic regression the dependent variable is dichotomous.It provides the presence or absence and success or failure of the response variable.The classification algorithm used here is the decision tress called J48 which uses the greedy search techniques.Aggregation is the simplest way when we do classification because it takes the weighted vote.Average (ie weighted average) is used if we need to predict a numeric value.The boosting algorithm used here is used to distribute the weight over the training dataset.Bagging algorithm eliminates the drawback of over fitting.The bagging uses the concept of bootstraping techniques.The agreement between the predicted and observed values are evaluated by using a Kappa statistics.The bagging algorithm here gives out better value of precision,accuracy and kappa which is around 80 %.The main setback of this predictive model is the scanners and satellite usage which may cause some delays and mostly not adequate in all cases.

In 2007 cortez uses different data mining approach for estimating the burned area in the forest cortez2007data.This approach is purely based on the weather conditions that can be obtained by the local meteorological stations.He uses data available in the recent real world which is obtained from northeastern region of Portugal.With low cost employed in collecting the data from 162 meteorological stations in Portugal.The author uses FWI system which includes 3 fuel related codes and moisture codes.The wind speed was also recorded for every 30 minute period by the weather station located in the park.This approach is considered to be a regression task.The Multiple regression model was carried out by the author and its overall performance is calculated using the RMSE and MAD value.However RMSE values are largely sensitive to errors.It can only produce linear mappings.The decision tree model is used with the set of rules in hierarchical form.The decision tree can be transformed into if else statement which can be easily understood by the humans.The random forest model was also used by the author which is the ensemble algorithm of decision trees.SVM regression model was also build which transforms the input into high dimensional space using a non linear mapping.The popular radial kernel is used which avoids the difficulties of other kernel such as poly and linear.All these model were processed using Rminer which is mainly used for data mining tasks.The performance of these algorithms were measured by using the REC curve.The SVM algorithm outperforms all other in prediction accuracy when we use only meteorological variables.It performs better when its predicting small fires.The model predicts 46 % accurately when we consider a error till 1ha of land and the the value is increased to 61 % if we consider an error till 2ha of land. (zhang2008forest) proposed a zigbee method using wireless sensor for the forest fire detection.The connection network is consist of numerous sensor nodes and and data packets which have their own ability to calculate and communicate with other.On comparing with other wireless protocol zigbee has many advantage such as low cost ,low power consumption and reliable data transmission.The data such as relative humidity and temperature can be collected easily from any forest and analyzed using zigbee sensor.It uses cluster network topology so that energy loss between the data packets is reduced.This model is totally based on wireless sensor prediction.

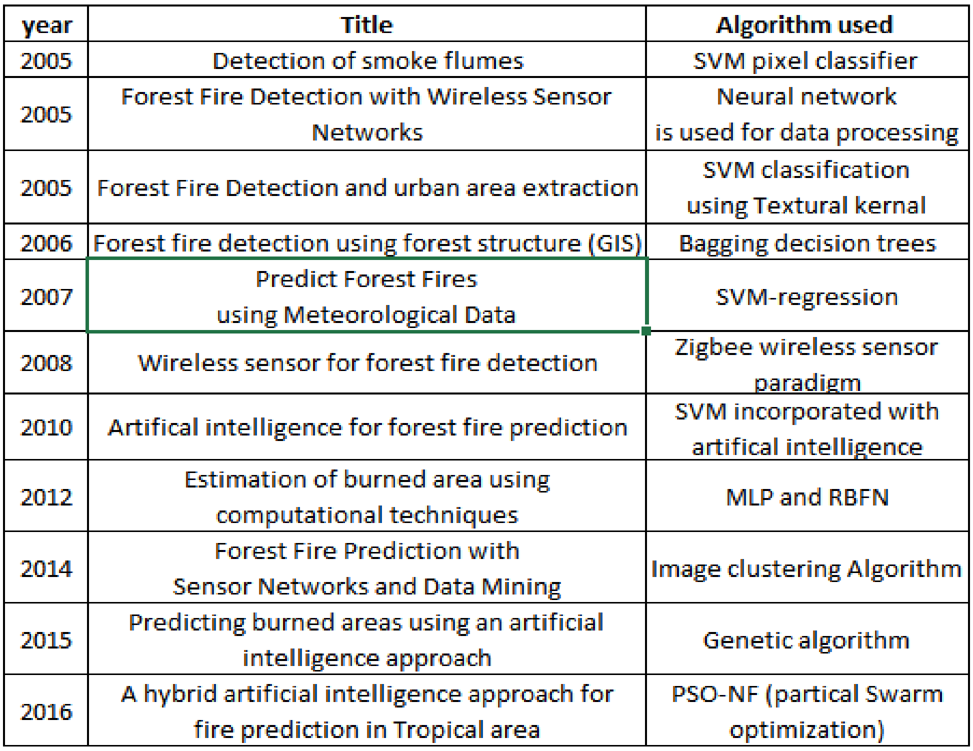
2.2 A review of papers from 2010-2017

The artificial intelligent system was used by (sakr2010artificial) for the prediction of forest fires.This approach uses SVM algorithm which uses the previous weather observations of a day in order to estimate the fire hazard level.The data here used is from Lebanese agricultural institute.The proposed implementation proposes a fire index between 1 to 4.The features used in this algorithm contains the Minimum and maximum temperature of the day,Average humidity, solar radiation,The precipitation level of a day, wind speed in a particular day.Since SVM produces a good generalization ability it used in this paper.The above method used has more than 2 class prediction architecture, it is a four class prediction architecture used across the months June to October.Each month has its own risk scale definition for example September has a risk scale of 3 and June or July has a risk scale of 4.By computing the average error of number of fire predicted the architecture performance is tested.This is a monthly prediction algorithm where the prediction accuracy for each month is calculated which shows the 96 % accuracy in the month of august with low error on number of fires.The computational intelligent techniques was used by (Özbayoğlu and Bozer; 2012) to predict the burned area in forest.Many data mining techniques such as SVM, RBFN, MLP and fuzzy logic were used for this predictions.The data is obtained from a period of 2000 to 2008 which cover over 7920 forest fires in Turkey.This method is also useful to predict the fire size whether it is small, medium or large fire.Here first a comparison was made in which the area burned is plotted along x-axis and number of fires is plotted along y-axis.This shows the 78 % land was less than 1 ha.The area burned were clustered using the K-means algorithm and various number of output clusters were tested.the centroids of each cluster was also calculated.The clusters were differentiated into very small fire, small fire, medium fire, big fire, large fire.The eighty percent of data represents fires that are less than 1ha.The data was split into 60 % for training and 20 % for cross validation and the remaining for testing purposes.The cluster 2, cluster 3 and cluster 5 were used for input.The training set uses 350 clusters as input.The performance was measured by comparing the output with the mean output value of the cluster.The RBFN performs very poorly in this method.The MLP was feed with 5 clusters and the MLP produces a overall performance of 53.02 %.The performance of MLP is increased to 62.89 % when we use 3 clusters.These different models were analyzed using error metrics such as MAE,RMSE and MAPE.A logarithmic function was introduced for transforming the burned area.SVM used 2 clusters (small and big) has input gives out best result.With more than 65 % success rate MLP turns out to be best model which produces a RMSE value of 7.33 nad MAE value of 3.66

The data mining techniques along with wireless sensor usage for predicting the forest fire is done by (Divya et al.; 2014).The image clustering algorithm was used in this paper which can detect the fires efficiently than the satellite based approach.Around 2000 hectares of land is destroyed by fires in India.The process of gathering the data is done using by Wireless sensor networks. The WSN is a cost effective device which is also used effective tracking.The images recorded by the WSN using Dv-cluster model was used for data processing.The clustering of images is used for predicting the fires.The data transmission is done by 2 ways one is from the nodes to the cluster head and from the cluster head to the base station.The securing clustering protocol is used for the security because the data should be transmitted securely from one cluster to analysis station.The Algorithm uses image pixel analysis where the extracted pixels are transmitted into values and used for segmentation.A incident matrix was created using these pixel values which contains the values as 0 and 1.The changes in the colour of image pixels was compared when there is orange yellow colour it shows the fire affected area and brown colour shows the area of smoke.By using this way the prediction is done.The main setback in the model is efficiency in the data transmission and memory used for the storage of the pixel values.

(Castelli et al.; 2015) used a powerful Genetic programming based on genetic operators.He created a biological process using computational techniques.This approach mainly uses a tree structure to create new solutions called genetic operators.The offspring and parents of the genetic operator are identified.The GP is an iterative process the first step is to randomly generate an initial population and calculate the fitness of each population and then the iteration process continues till the maximum number of generations is reached.The best solutions are provided once the maximum iteration is reached.The Geometric operators uses unimodal fitness landscape consisting of input data into targets.The GP mutation operator creates a new solution by randomly replacing the subtree.The data used for this approach is from the park situated in Portugal.The author has studied the park in a very brief manner about all the vegetation and people living in and around the park.There are around 8000 people living in the villages situated in the park.The average rainfall is around 800mm to 1500 mm.The duration of drought period in this park is 4 months.By reviewing these details helps us in our research to implement a predictive model.(Castelli et al.; 2015) uses the month and day of the week as temporal variables.This research uses the FWI system data along with the weather attributes.The author performed a total of 50 runs.A different partition between train and test data is considered the fitness of the model is calculated by MAE (Mean absolute error).The GS-GP returned a 12.0 of MAE on the training set.A experimental comparison is done with various machine learning techniques for the prediction of burned area.The results obtained are the MAE of GS-GP for the train is 12.0 and for the test data is 12.9.Other algorithms like SVM produces MAE around 12.3 to 13.6.NN performed poorly.The find the statistical significance he used a bonferroni correction the final value is 0.014.The difference between the test and train fitness shows the GS-GP outperformed all the other algorithms.But the key drawback in this algorithm is it uses an unguided mutation.The mutation operator keeps on adding the number of parameters in the population.There is also over fitting of sample data and more optimization time. bui2017hybrid proposes a novel model using artificial intelligent approach.The model consist of particle swarm optimized Neural fuzzy logic for modeling forest fire.The data collected from the GIS system contains features like Tempearture, wind speed,vegetation index,rainfall, land use.The performance of the model was validated by ROC curve and AUC curve.The remporal analysis of the forest fire were done for each month which shows in the month of march and april the fire percentage was around 25-30 %. The neural fuzzy structure has 5 layers where the artificial neural networks were used to transfer the information from the input and output.Each layer is used for different purposes.The layer 1 is the input node where the ten forest fire ignition data is feed using a fuzzification method.The layer 2 is made of set of rules and layer 3 is for Normalization of data.The layer 5 has a fixed node which is used for aggregation purpose.The Positive predicted value is around 84.6 % which shows the proability of classificying the fire pixels.The negative predicted value is around 95.4 % which shows the probability of classifying non-fire region.The The proposed model produces a result of about AUC= 0.93 for the training dataset and 0.916 for the validation.

2.3 Summary of my Literature Review



year	Title	Algorithm used
2005	Detection of smoke flumes	SVM pixel classifier
2005	Forest Fire Detection with Wireless Sensor Networks	Neural network is used for data processing
2005	Forest Fire Detection and urban area extraction	SVM classification using Textural kernal
2006	Forest fire detection using forest structure (GIS)	Bagging decision trees
2007	Predict Forest Fires using Meteorological Data	SVM-regression
2008	Wireless sensor for forest fire detection	Zigbee wireless sensor paradigm
2010	Artifical intelligence for forest fire prediction	SVM incorporated with artifical intelligence
2012	Estimation of burned area using computational techniques	MLP and RBFN
2014	Forest Fire Prediction with Sensor Networks and Data Mining	Image clustering Algorithm
2015	Predicting burned areas using an artificial intelligence approach	Genetic algorithm
2016	A hybrid artificial intelligence approach for fire prediction in Tropical area	PSO-NF (partical Swarm optimization)

Figure 1: Literature review summary

3 Methodology

3.1 Method 1: Linear Regression

The problem here is modelled into a Regression task since over motive is to predict the area of the land burnt.The variable that we need to predict is in numerical value.The regression allows us to model mathematically the relationship between two or more variable.This linear regression model is used to find whether there is a positive or negative relationship between the variables.Normally a regression equation is Y(Dependent variable) = a (intercept) + b (slope of the line) * X (Independent or explanatory variable)

3.2 Method 2: Gradient boosting

Gradient boosting is a technique for producing regression models consisting of collections of regressor.It is an ensemble algorithm where the regressor predictions are combined usually by some sort of weighted average or vote in order to provide an over all prediction.Boosting is a method in which learners are learned sequentially with early learners fitting simple models to the data and analyzing the data for errors (Friedman; 2002).

3.3 Method 3: Bagging

Bagging is a classic ensemble method known as bootstrap aggregation.Bagging algorithm consist of many classifiers each uses only some portions of data in each iterations and then combining them through a model averaging techniques.The idea behind this is reduce the overfitting in the class of models.The bootstrap method in bagging creates a random subset of data from a given dataset by sampling (Breiman; 1996)

3.4 Method 4: Random forest

Random forest is a powerful algorithm which can be used for both regression and classification.The algorithm first creates bootstrap samples from the original data.A regression tree is developed from the each bootstrap samples.Then it randomly sample the number of predictors and the best split is chosen from the variables.Now the aggregation method is used for predicting the new data (Liaw et al.; 2002)

3.5 Method 5: SVM regression

SVM looks at the extremes of the dataset and draws a decision boundary known as hyperplane near the extreme points in the dataset.It is a method which uses epsilon loss function and performs linear regression in high dimensional space.The SVM always follows a kernel trick where we can use different kernels like RBF,linear,polynomial,Sigmoid

3.6 Method 6: Logistic regression

This method uses sigmod function where we get the probability of the given event.It just assign probability to every single event of area burned.This produces logistic coefficients from where we can find the proability of present or absent.The data should be distributed and there should be some relationship between the attributes in the data

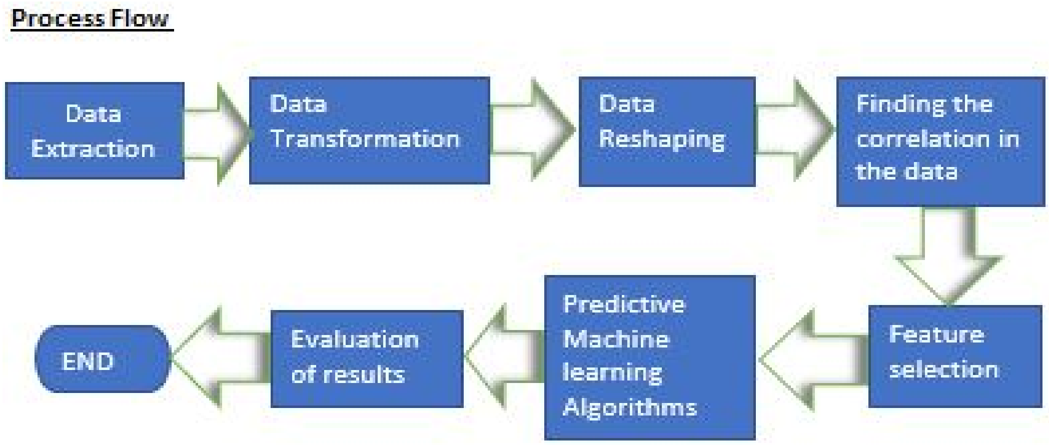


Figure 2: Process flow

4 Implementation

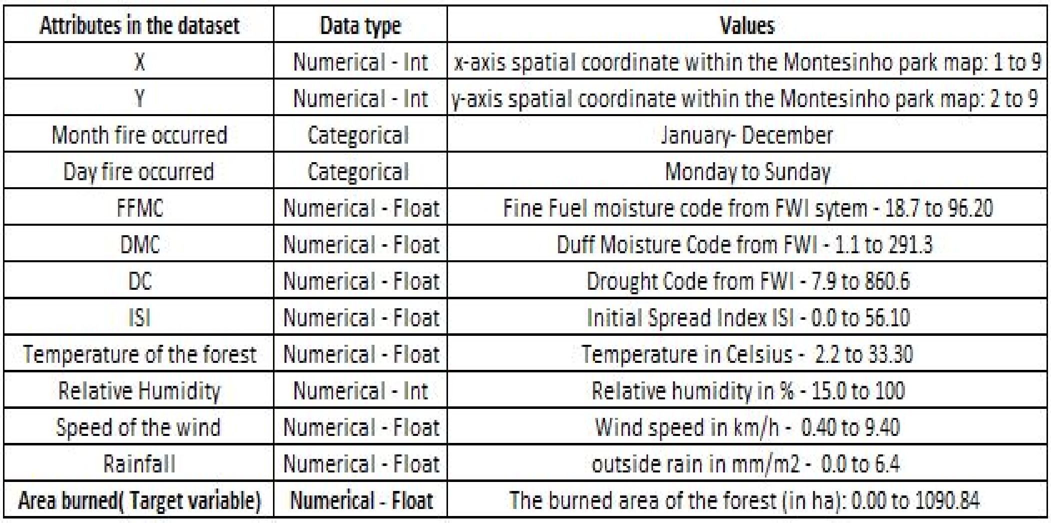
4.1 Process Flow Diagram

The process flow is step by step procedure that needs to be followed for implementing these models is briefly explained below sections

4.2 Data Extraction

The data is extracted from the UCI machine learning repository.The data consist of meteorological.FWI system data and amount of area burned during fires over a period of 2000-2003 in Portugal.The factors that mainly affect the forest fire are the climatic conditions of the forest.The data set has clear description of the climatic conditions such as Relative humidity,temperature of the forest,Wind speed and rainfall in the forest.These data is collected from the local sensors with are available in the Portugal.The Portugal has around 162 weather stations so getting this data is not a big deal.The FWI system is the which is widely used as a fire danger rating system.The data also contains the day , month and X and Y axis values where the fire occurred.The getting the day and month we can separate the fires into week day and weekend.The next FWI data is like moisture code,Fire index,Drought code and spread index which are mainly depend on the weather conditions.These values calculated by the FWI system is a direct indicator of the fire intensity.The Relative humidity value is a changing one because it will be high in the morning and keep reducing to the minimum value as hours past.The wind speed is a major factor since it can make the fire to spread rapidly.From looking at the data we can say when the wind speed is around 15/hr the chances of fire is high.One of the most important feature in the dataset is temperature of the forest which can cause fire.Below tables provides us a clear understanding of the dataset and FWI system.

The below table provides us a clear understanding of the FWI system (Stocks et al.; 1989)



Attributes in the dataset	Data type	Values
X	Numerical - Int	x-axis spatial coordinate within the Montesinho park map: 1 to 9
Y	Numerical - Int	y-axis spatial coordinate within the Montesinho park map: 2 to 9
Month fire occurred	Categorical	January- December
Day fire occurred	Categorical	Monday to Sunday
FFMC	Numerical - Float	Fine Fuel moisture code from FWI sytem - 18.7 to 96.20
DMC	Numerical - Float	Duff Moisture Code from FWI - 1.1 to 291.3
DC	Numerical - Float	Drought Code from FWI - 7.9 to 860.6
ISI	Numerical - Float	Initial Spread Index ISI - 0.0 to 56.10
Temperature of the forest	Numerical - Float	Temperature in Celsius - 2.2 to 33.30
Relative Humidity	Numerical - Int	Relative humidity in % - 15.0 to 100
Speed of the wind	Numerical - Float	Wind speed in km/h - 0.40 to 9.40
Rainfall	Numerical - Float	outside rain in mm/m2 - 0.0 to 6.4
Area burned(Target variable)	**Numerical - Float**	The burned area of the forest (in ha): 0.00 to 1090.84

Figure 3: Dataset attribute description

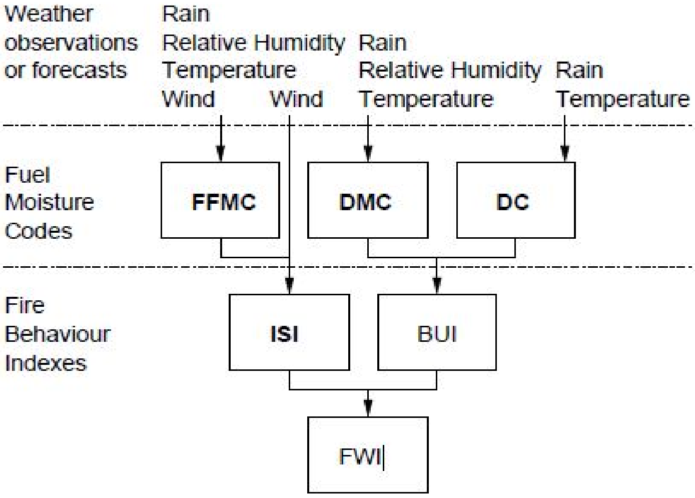


Figure 4: Fire Weather Index structure

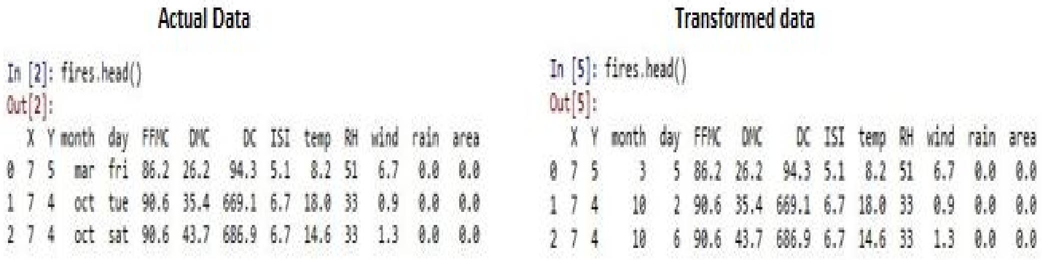
4.3 Tools used

Various components needed for this implementation are listed below

- Technical Environment: Spyder Python 3.6
- Excel 2016
- Python library such as numpy,pandas,calender,seaborn and Matlab plot
- Python Programming to create machine learning scripts

4.4 Data transformation

In this dataset we have categorical variables which needs to be converted into a numerical variable.The task here is modelled as a regression task.The data transformation is done by Python 3.6.First process is to load the data in the python and checking if it is loaded properly.Now we need to transform the month and day attribute to a numerical variable so it will be easy for us to implement the algorithms



Actual Data

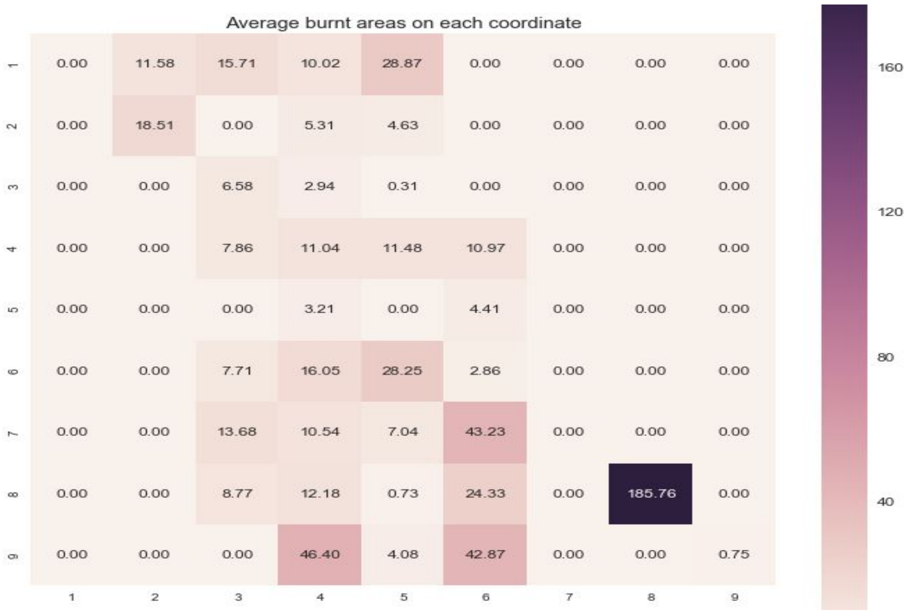
```
In [2]: fires.head()
Out[2]:
   X  Y month  day  FFMC   DMC     DC  ISI  temp  RH  wind  rain  area
0  7  5   mar  fri  86.2  26.2   94.3  5.1   8.2  51   6.7   0.0   0.0
1  7  4   oct  tue  90.6  35.4  669.1  6.7  18.0  33   0.9   0.0   0.0
2  7  4   oct  sat  90.6  43.7  686.9  6.7  14.6  33   1.3   0.0   0.0
```

Transformed data

```
In [5]: fires.head()
Out[5]:
   X  Y  month  day  FFMC   DMC     DC  ISI  temp  RH  wind  rain  area
0  7  5      3    5  86.2  26.2   94.3  5.1   8.2  51   6.7   0.0   0.0
1  7  4     10    2  90.6  35.4  669.1  6.7  18.0  33   0.9   0.0   0.0
2  7  4     10    6  90.6  43.7  686.9  6.7  14.6  33   1.3   0.0   0.0
```

Figure 5: Transformed Data

Now next step is get the insights from the data we have.We are producing a heat map for calculating the average area burnt on each coordinate.one of the valuable aspects of heat map is the ability to view the data tables in their entirety, because colour takes lesser space than numbers.The colour scheme here shows the density of the average area burnt here.Here from the heat map we can find the average of area burnt is well less than 50 along the x and Y axis.Since the majority of area burnt is less than 50 we can say that small fires are most common.The highest average area burnt is 185.76 along the x and y coordinate.From here we could find the area burnt is not normally distributed.Below is our Heatmap for the average area burnt



	1	2	3	4	5	6	7	8	9
1	0.00	11.58	15.71	10.02	28.87	0.00	0.00	0.00	0.00
2	0.00	18.51	0.00	5.31	4.63	0.00	0.00	0.00	0.00
3	0.00	0.00	6.58	2.94	0.31	0.00	0.00	0.00	0.00
4	0.00	0.00	7.86	11.04	11.48	10.97	0.00	0.00	0.00
5	0.00	0.00	0.00	3.21	0.00	4.41	0.00	0.00	0.00
6	0.00	0.00	7.71	16.05	28.25	2.86	0.00	0.00	0.00
7	0.00	0.00	13.68	10.54	7.04	43.23	0.00	0.00	0.00
8	0.00	0.00	8.77	12.18	0.73	24.33	0.00	185.76	0.00
9	0.00	0.00	0.00	46.40	4.08	42.87	0.00	0.00	0.75

Figure 6: Heat Map for the Average area burnt

4.5 Data Reshaping

The output variable is the area, here we could find the area is in a positive skew.Most of the values in the area attribute is equal to zero.The positive skew represents the majority of fires are small fires.The skew trait system is also available in countries like Canada (Malarz et al.; 2002).The values zero in the area column denotes that the amount of land burnt is less than 1ha.In mathematical terms 1ha/100 = 100 meter square was burned. For improving the accuracy and to reduce the skewness a log function was introduced in the area attribute.The transformed output area will be given in the figure below. From the below Figure 7 we can find area is normally distributed and the skewness is reduced after applying a log function.This transformed area is used as the output variable in the Machine learning models

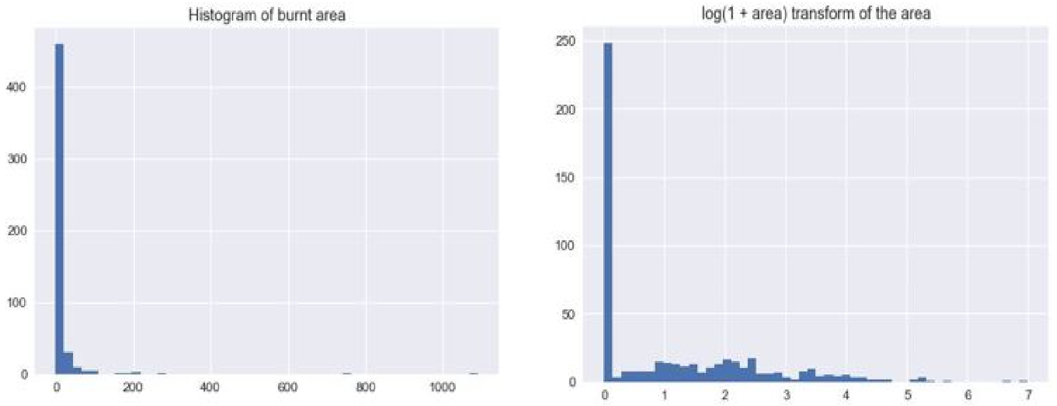
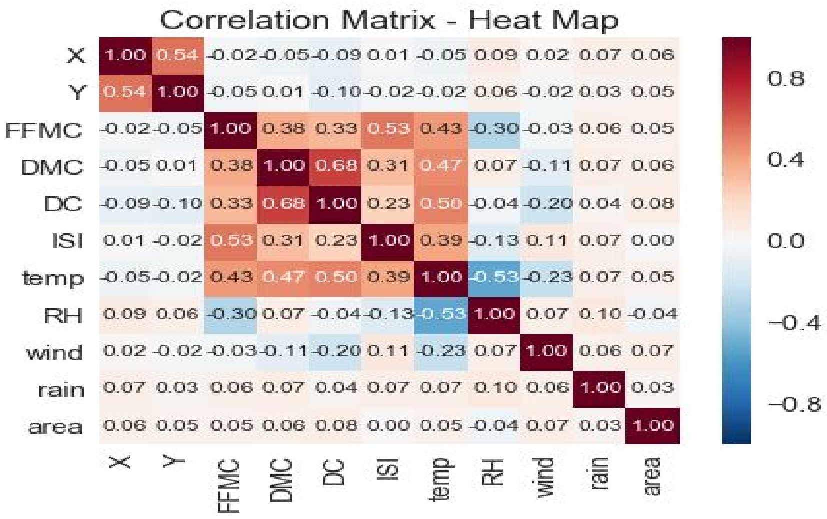


Figure 7: The histogram for the burned area (left) and respective logarithm transform (right)

4.6 Finding the correlation

The correlation matrix is used to find which attribute has a significant correlation on the output target variable.The correlation ranges from Negative and positive.If the correlation is zero then there is no relationship between two attributes.We can see that the temperature has more positive correlation with the area burnt.The wind has more negative correlation with the output variable.

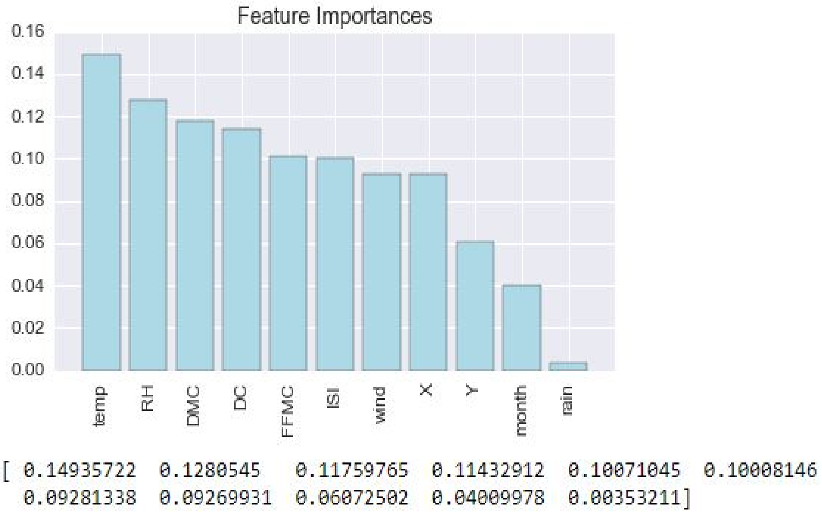


	X	Y	FFMC	DMC	DC	ISI	temp	RH	wind	rain	area
X	1.00	0.54	-0.02	-0.05	-0.09	0.01	-0.05	0.09	0.02	0.07	0.06
Y	0.54	1.00	-0.05	0.01	-0.10	-0.02	-0.02	0.06	-0.02	0.03	0.05
FFMC	-0.02	-0.05	1.00	0.38	0.33	0.53	0.43	-0.30	-0.03	0.06	0.05
DMC	-0.05	0.01	0.38	1.00	0.68	0.31	0.47	0.07	-0.11	0.07	0.06
DC	-0.09	-0.10	0.33	0.68	1.00	0.23	0.50	-0.04	-0.20	0.04	0.08
ISI	0.01	-0.02	0.53	0.31	0.23	1.00	0.39	-0.13	0.11	0.07	0.00
temp	-0.05	-0.02	0.43	0.47	0.50	0.39	1.00	-0.53	-0.23	0.07	0.05
RH	0.09	0.06	-0.30	0.07	-0.04	-0.13	-0.53	1.00	0.07	0.10	-0.04
wind	0.02	-0.02	-0.03	-0.11	-0.20	0.11	-0.23	0.07	1.00	0.06	0.07
rain	0.07	0.03	0.06	0.07	0.04	0.07	0.07	0.10	0.06	1.00	0.03
area	0.06	0.05	0.05	0.06	0.08	0.00	0.05	-0.04	0.07	0.03	1.00

Figure 8: Correlation Matrix

4.7 Feature selection

The feature selection here is used to find which attributes we need to use in our prediction model.From the Figure 10 we find the attributes Temperature and RH are the are the best features for doing a prediction model.By including the attributes like temperature,RH,DMC and DC in our prediction model and by splitting into training and testing set help us to get better prediction accuracy By doing this the overfitting of the model is reduced.The training time of the model is reduced because we are eliminating the attributes that are less contributing to the output variable(Area burnt).Here we are using random forest classifiers for the feature ranking of the attributes.The train data is now used to fit the model with random forest classifiers.The Important features in the dataset is plotted were explained in the below Figure 10



```
[ 0.14935722  0.1280545   0.11759765  0.11432912  0.10071045  0.10008146
  0.09281338  0.09269931  0.06072502  0.04009978  0.00353211]
```

Figure 9: Feature selection

4.8 Predictive machine Learning Algorithms

Before building a machine learning algorithm we need to split the data into Train and Test.This split is used to validate the model.The training part is used to create a model and testing part is used to verify the model created.Here the data split is done by separating 70 % to the test data and 30 % to the train data.Then a standard scaler function is introduced on the training set.Standardization of the dataset is the common usage for many machine learning algorithms.The standard transform is applied to both test and training set.Now the these data is loaded into a dataframe.Now the prediction models are implemented using this dataFrame.A confusion matrix is build from which we can calculate True positive, True negative,Type 1 error and Type 2 error

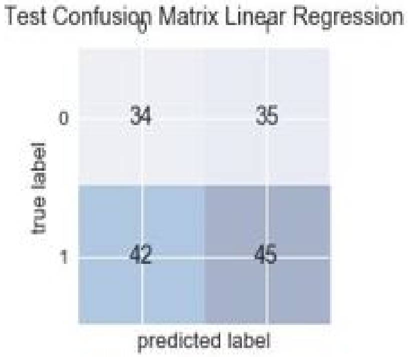
5 Evaluation

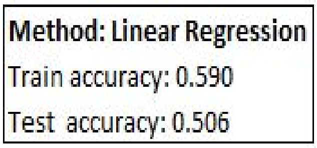
In the evaluation section we will be selecting the best model in terms of Accuracy Various results of the predictive models are given below

5.1 Regression models:

Linear Regression : We are creating a model fit for linear regression so that both train and test data.Once the prediction is done we are converting the result into 0 and 1 where 0 denotes Prediction output (Area) less than 0.5 and 1 denotes area greater than 0.5

Accuracy is calculated for each and every model.The accuracy for the both test and train data is calculated using the prediction model.A confusion matrix is build and plotted are given below




Method: Linear Regression

Train accuracy: 0.590

Test accuracy: 0.506

Figure 10: Confusion Matrix Linear Regression

Logistic Regression : In this method Logistic regression is done with PCA analysis with all the components.The PCA analysis is used to find directions of maximum variance in the data.PCA is used for compressing a data into something and capturing the essence of original data.The learning curve is also show in which the training and testing accuracy curves are plotted with the training samples.The below Figure 11 shows the learning curve for the logistic Regression in which the Training accuracy is gradually decreasing from 0.8 and the testing accuracy is around 0.5.

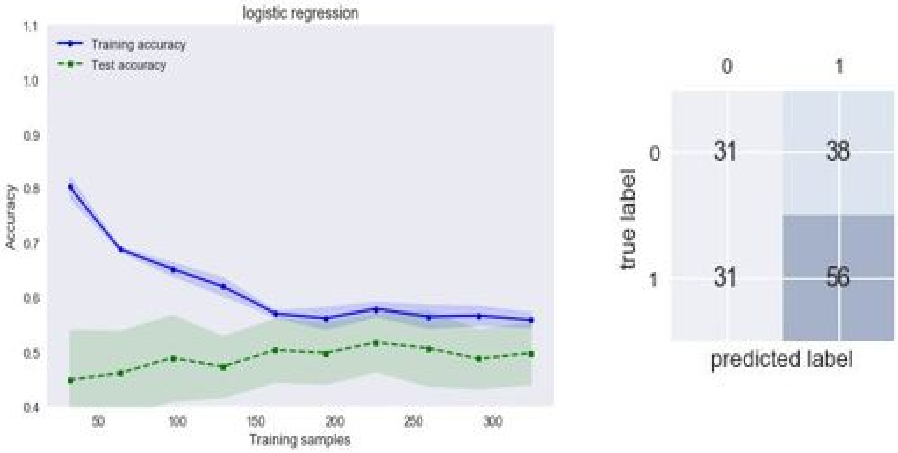


Figure 11: Learning curve and CM for Logistic Regression

SVM with RBF and Poly Kernels : SVM uses the kernel trick where we use RBF and Poly kernels for building a predictive model.SVM uses non-linear kernels (RBF and poly) for separating the groups efficiently.Here we use non-linear kernels because we have all the combinations of small, medium and large fires.The kernels were used with the gamma value of 0.2.The gamma defines how far the influence of a single training example reaches.We choose gamma has 0.2 since every point in the data has far reach.The learning curve of SVM was given in the Figure 12 which shows both poly and RBF Kernel has better Training accuracy (0.76)

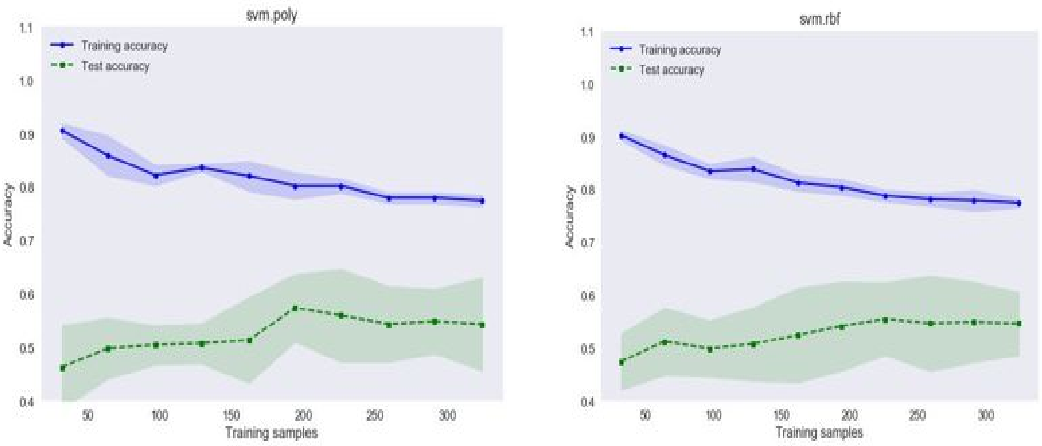
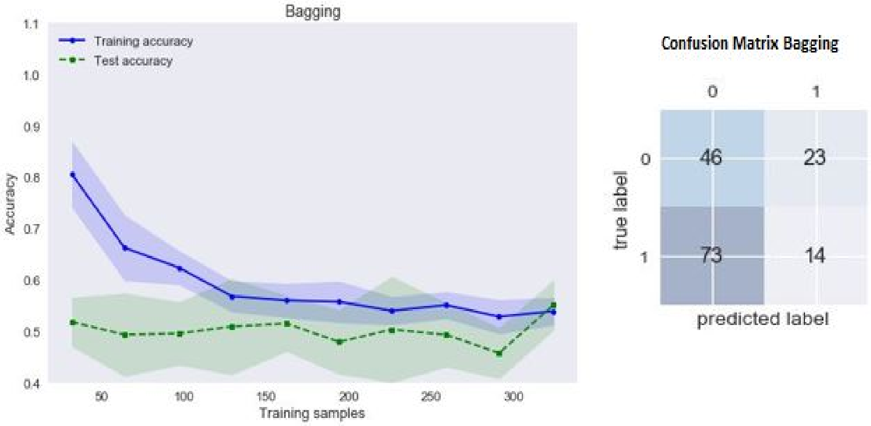


Figure 12: Learning curve for SVM poly and SVM RBF

5.2 Classification Models:

Bagging : It is a powerful ensemble algorithm which uses only portions of data in the train set and produce a set of classifiers.These classifiers are then combined into model average technique.Here we split the data into 70 % train and remaining to test ,Then introduce a scalar transformation for both train and test.Then the algorithm produces 10 base classifiers using bootstrapping technique this reduces the model overfitting.This bootstraping produces 10 subset of data from the original dataset.The Learning curve shows poor test accuracy 0.4.The confusion matrix shows more percentage of Type 2 error (False negative)



	0	1
0	46	23
1	73	14

Figure 13: Learning curve and CM for bagging

Random Forest : The prediction accuracy of the Random forest is very good both in terms of train and test data.This is one of the best algorithm which can provide better accuracy when the data is uneven.This reduces the noise in the data.The confusion matrix here shows RF produces a 44 % of True positive and 43 % of True negative which in-turn increases the accuracy of the model.Testing accuracy is around 0.947 which shows better prediction results

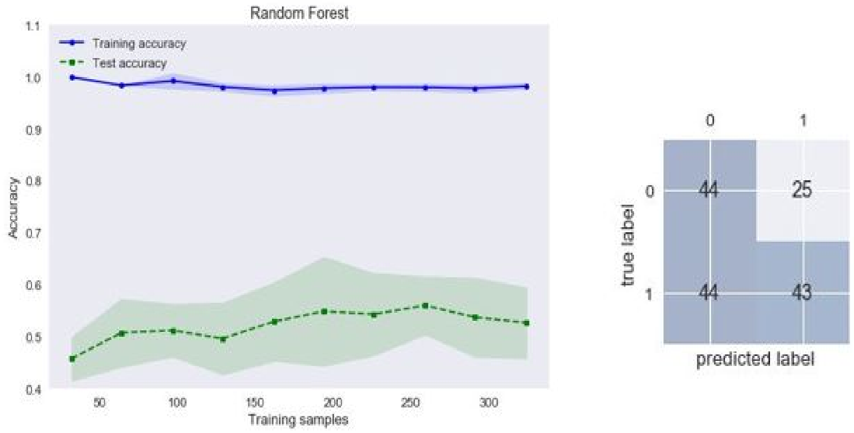


Figure 14: Learning curve and CM for Random Forest

Gradient Boosting : This algorithm uses collection of regressors and produces simple learners which avoid the data from errors.The GBM was modelled as max depth = 5 which shows the maximum depth of the tree.Since the value is higher the model is grow based on the relations which are very close to the samples. This is also avoid the overfitting.The accuracy of the GBM model was very good in terms of both train and test

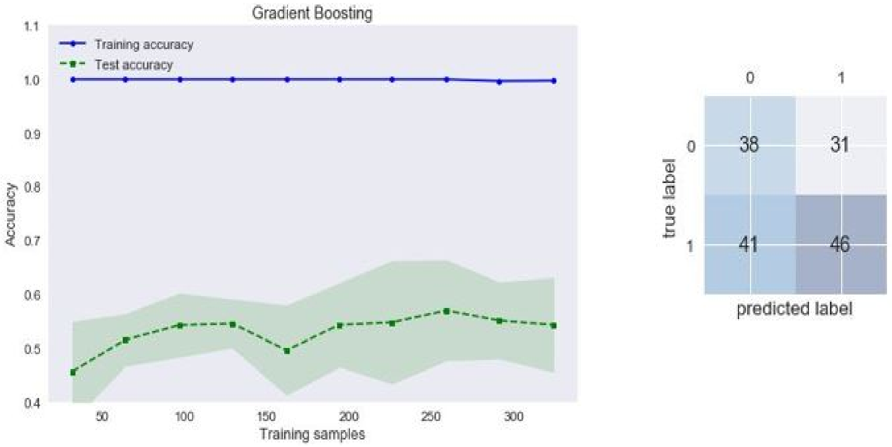
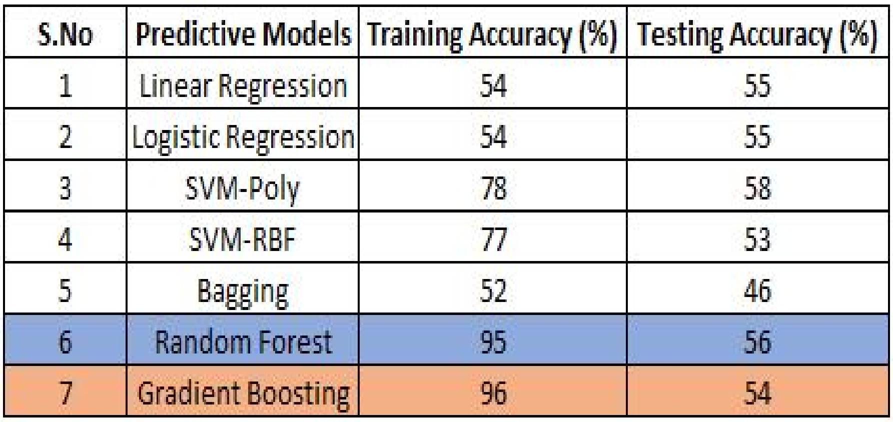


Figure 15: Learning curve and CM for Gradient Boosting

5.3 Comparative study of the Predictive Models

We implemented around 7 different machine algorithm for the forest fire data of Montesinho park in Portugal.We plotted the Learning curve with respective training and testing accuracy.From the analysis we could find the Random forest provides a Training Accuracy around 95 % and Testing Accuracy of 56 %.Even other model like Gradient boosting provides very good Results.The Accuracy of the ML models are listed in the below Table

The most accurate results were generated by Random forest.Even Gradient Boosting Accuracy is more or less equal to Random forest.The major difference between these 2



S.No	Predictive Models	Training Accuracy (%)	Testing Accuracy (%)
1	Linear Regression	54	55
2	Logistic Regression	54	55
3	SVM-Poly	78	58
4	SVM-RBF	77	53
5	Bagging	52	46
6	Random Forest	95	56
7	Gradient Boosting	96	54

Figure 16: Accuracy of the Models

algorithms are Random forest uses the concept of bootstraping and the trees developed are Independent.On the other hand Gradient boosting algorithm uses Boosting methods and the trees are dependent on each other. Both are ensemble tree algorithms which can be used for various classification and Regression problems

6 Acknowledgement

In completing this project,I am very much grateful to my Project supervisor **Mr.Thibaut Lust**,Lecture for his continuous guidance and valuable inputs throughout this project.I also wanted to thank National college of Ireland and my beloved friends on helping me in this research project

7 Conclusion and Future Work

The world is moving towards automation and this Big data era urges us to build more solutions for the complex problems.In this project with all the use of Big data and machine learning techniques we build a model for the prediction of area burned during the forest fires.This model should be incorporated in all the areas which have more probability of capturing fire.By further tuning the parameters and by adding some other attributes like vegetation of the forest,Forest cover,type of trees in the forest and Buildup Index we can improve the Accuracy of random forest and Boosting algorithm.This project mostly aims in developing a predictive model with the usage of climatic conditions.By detecting the area burned we can separate the fires into small and large.This classification of fires helps the FMS team to send adequate crews and air tankers to the following danger zone.

The future work in this project can be done by creating a probabilistic models that can identify the origin of fire by using some conditions.Those probabilistic models should be integrated with the model provided in this study to handle more risky conditions in the case of large or big fires.The use of GIS data and satellite view can also be included with this model which provides better accuracy

References

Ammann, H., Blaisdell, R., Lipsett, M., Stone, S. L., Therriault, S., Jenkins, J. and Lynch, K. (2001). Wildfire smoke: a guide for public health officials, *California Air Resources Board. http://www. arb. ca. gov/smp/progdev/pubeduc/wfgv8. pdf (accessed 06/02/08)* .

Boubeta, M., Lombardía, M. J., González-Manteiga, W. and Marey-Pérez, M. F. (2016). Burned area prediction with semiparametric models, *International Journal of Wildland Fire* **25**(6): 669–678.

Breiman, L. (1996). Bagging predictors, *Machine learning* **24**(2): 123–140.

Castelli, M., Vanneschi, L. and Popovič, A. (2015). Predicting burned areas of for-est fires: an artificial intelligence approach, *Fire ecology* **11**(1): 106–118.

Divya, T., Manjuprasad, B., Vijayalakshmi, M. and Dharani, A. (2014). An efficient and optimal clustering algorithm for real-time forest fire prediction with, *Communications and Signal Processing (ICCSP), 2014 International Conference on*, IEEE, pp. 312–316.

Friedman, J. H. (2002). Stochastic gradient boosting, *Computational Statistics & Data Analysis* **38**(4): 367–378.

Lafarge, F., Descombes, X. and Zerubia, J. (2005). Textural kernel for svm classification in remote sensing: Application to forest fire detection and urban area extraction, *Image Processing, 2005. ICIP 2005. IEEE International Conference on*, Vol. 3, IEEE, pp. III–1096.

Liaw, A., Wiener, M. et al. (2002). Classification and regression by randomforest, *R news* **2**(3): 18–22.

Malarz, K., Kaczanowska, S. and Kułakowski, K. (2002). Are forest fires predictable?, *International Journal of Modern Physics C* **13**(08): 1017–1031.

Mangan, R. J. (2007). *Wildland firefighter fatalities in the United States: 1990-2006*, The Center.

Mateus, P. and Fernandes, P. M. (2014). Forest fires in portugal: dynamics, causes and policies, *Forest Context and Policies in Portugal*, Springer, pp. 97–115.

Özbayoğlu, A. M. and Bozer, R. (2012). Estimation of the burned area in forest fires using computational intelligence techniques, *Procedia Computer Science* **12**: 282–287.

Stocks, B. J., Lynham, T., Lawson, B., Alexander, M., Wagner, C. V., McAlpine, R. and Dube, D. (1989). Canadian forest fire danger rating system: an overview, *The Forestry Chronicle* **65**(4): 258–265.

Stojanova, D., Panov, P., Kobler, A., Džeroski, S. and Taškova, K. (2006). Learning to predict forest fires with different data mining techniques, *Conference on Data Mining and Data Warehouses (SiKDD 2006), Ljubljana, Slovenia*, pp. 255–258.

Taylor, S. W. and Alexander, M. E. (2006). Science, technology, and human factors in fire danger rating: the canadian experience., *International Journal of Wildland Fire* **15**(1): 121–135.

Yu, L., Wang, N. and Meng, X. (2005). Real-time forest fire detection with wireless sensor networks, *Wireless Communications, Networking and Mobile Computing, 2005. Proceedings. 2005 International Conference on*, Vol. 2, IEEE, pp. 1214–1217.