Capstone Project

MACHINE LEARNING NANODEGREE
INSIGHT INTO FOREST FIRES

Project Overview:

One major environment concern is the occurrence of forest fires (also called wildfires), which affect forest preservation, create economic and ecological damage and cause human suffering is due to multiple causes (e.g. human negligence and lightnings) and despite an increasing of state expenses to control this disaster. Fast detection is a key element for a successful firefighting. Since traditional human surveillance is expensive and affected by subjective factors, there has been an emphasis to develop automatic solutions. 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 (e.g. warning the public of a fire danger) and to support fire management decisions.[1]. My main motivation to take this project is Forests cover 31 percent of the world's land surface, just over 4 billion hectares. As forest are main resource of environment and Forest fires are big problems in countries like Australia and America. I wanted to solve it by using some of the machine learning techniques.

One recent research paper based on the burned area under forest fires reference link: http://www3.dsi.uminho.pt/pcortez/fires.pdfThe main aim of this project is to predict the best suited regression for Insight in to forest fire dataset.

[1][1] www3.dsi.uminho.pt/pcortez/fires.pdf

Problem statement:

The main aim of my project is to find out which of the regression best suited for the dataset. For this I selected the data set compiled from a wide range of sources So, my goal is to predict the burned area based on features like temperature, humidity, rain, wind by below mentioned classifiers. Here I am using the regression models to find the mean squared error and variance of each model and select the best model which will have least mean squared error. The model created with the training dataset has been evaluated with the standard metrics such as accuracy, mean squared error. The experiment will be carried out using some classifier models, namely: Linear Regression, Decision tree, Support Vector Machine (SVM), Random forest, Lasso model are tested on the dataset. This is in view to finding out which of the regression best suits the dataset in terms of classifying the pre-processed data, trained data, testing.

METRICS:

I want to use accuracy score and mean squared error and variance as my evaluation metric for predicting the best regression for my dataset.

Mean Squared Error:

Mean Squared Error (MSE) it takes the average of the **square** of the difference between the original values and the predicted values. The advantage of MSE being that it is easier to compute the gradient. As, we take square of the error, the effect of larger errors become more pronounced then smaller error, hence the model can now focus more on the larger errors.[2]

$$MeanSquaredError = \frac{1}{N}\sum_{j=1}^{N}(y_j - \hat{y}_j)^2$$

Variance:

Variance is the amount that the estimate of the target function will change if different training data was used.The target function is estimated from the training data by a machine learning algorithm, so we should expect the algorithm to have some variance. Ideally, it should not change too much from one training dataset to the next, meaning that the algorithm is good at picking out the hidden underlying mapping between the inputs and the output variables.Machine learning algorithms that have a high variance are strongly influenced by the specifics of the training data.

- **Low Variance**: Suggests small changes to the estimate of the target function with changes to the training dataset.

- **High Variance**: Suggests large changes to the estimate of the target function with changes to the training dataset.

Examples of low-variance machine learning algorithms include: Linear Regression, Linear Discriminant Analysis and Logistic Regression.

Examples of high-variance machine learning algorithms include: Decision Trees, k-Nearest Neighbors and Support Vector Machines.[3]

Since this is a regression problem, the metric used will be “Coefficient of Determination”, in other words denoted as R2 (R squared) which gives a measure of the variance of target variable that can be explained using the given features.   For this project, I will use ‘r2_score()’ function of the metrics module of scikit-learn library.

$$R - Square = 1 - \frac{\Sigma(Y_{actual} - Y_{predicted})\hat{}2}{\Sigma(Y_{actual} - Y_{mean})\hat{}2}$$

These metrics are helpful for this problem because of the following reasons: i. It is a Regression based problem.

ii. R2 score will show the statistical robustness of the model.

iii. RMSE will give an idea about how accurate the predictions are to actual values.

Reference link:[2]https://towardsdatascience.com/metrics-to-evaluate-your-machine-learning-algorithm-f10ba6e38234

[3]https://machinelearningmastery.com/gentle-introduction-to-the-bias-variance-trade-off-in-machine-learning/

2.Analysis :

The objective of data analysis step is to increase the understanding of the problem from the data. There are two approaches to describe a given dataset. Summarizing and Visualizing data.

Data Exploration:

This dataset is public available for research. The details are described in [Cortez and Morais, 2007].The data can be used to test regression (difficult task), feature selection or outlier detection methods.

Dataset: https://archive.ics.uci.edu/ml/datasets/forest+fires

There are 517 instances and 13 attributes in my dataset.

Dataset information :

The attributes in the dataset include:

X - x-axis spatial coordinate within the Montesinho park map: 1 to 9

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

month - month of the year: "jan" to "dec"

day - day of the week: "mon" to "sun"

FFMC - FFMC index from the FWI system: 18.7 to 96.20

DMC - DMC index from the FWI system: 1.1 to 291.3

DC - DC index from the FWI system: 7.9 to 860.6

ISI - ISI index from the FWI system: 0.0 to 56.10

temp - temperature in Celsius degrees: 2.2 to 33.30

RH - relative humidity in %: 15.0 to 100

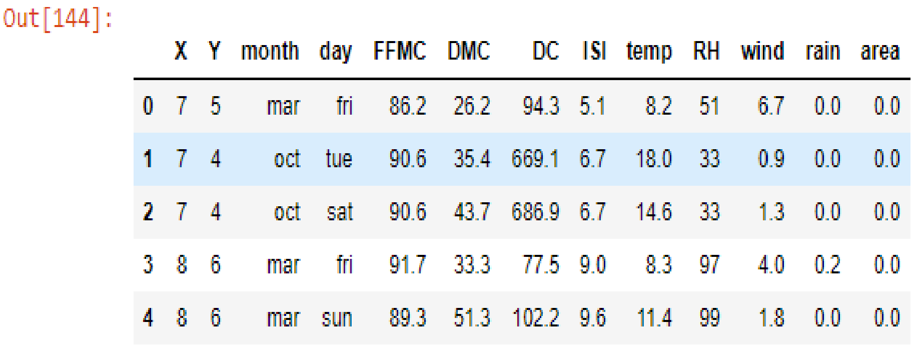
wind - wind speed in km/h: 0.40 to 9.40

rain - outside rain in mm/m2 : 0.0 to 6.4

area - the burned area of the forest (in ha): 0.00 to 1090.84

Descriptive statistics:

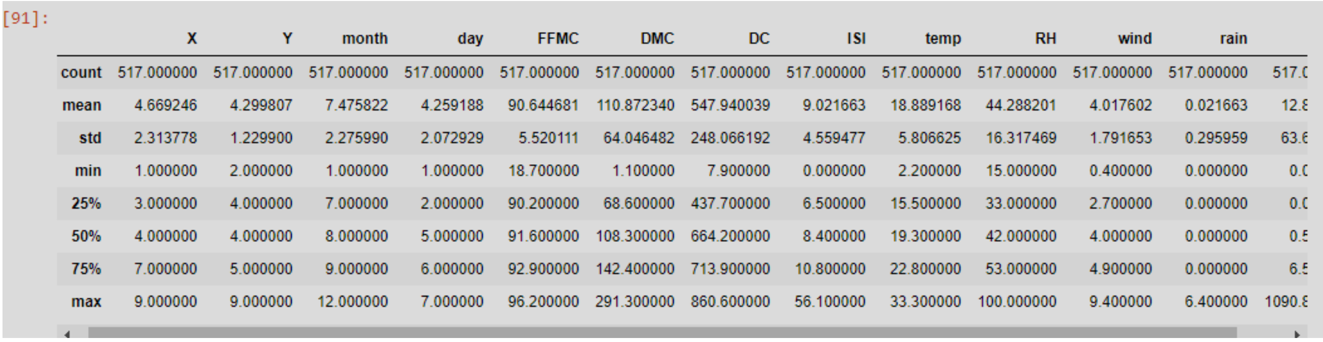
- Reading the data:



Out[144]:

	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
3	8	6	mar	fri	91.7	33.3	77.5	9.0	8.3	97	4.0	0.2	0.0
4	8	6	mar	sun	89.3	51.3	102.2	9.6	11.4	99	1.8	0.0	0.0

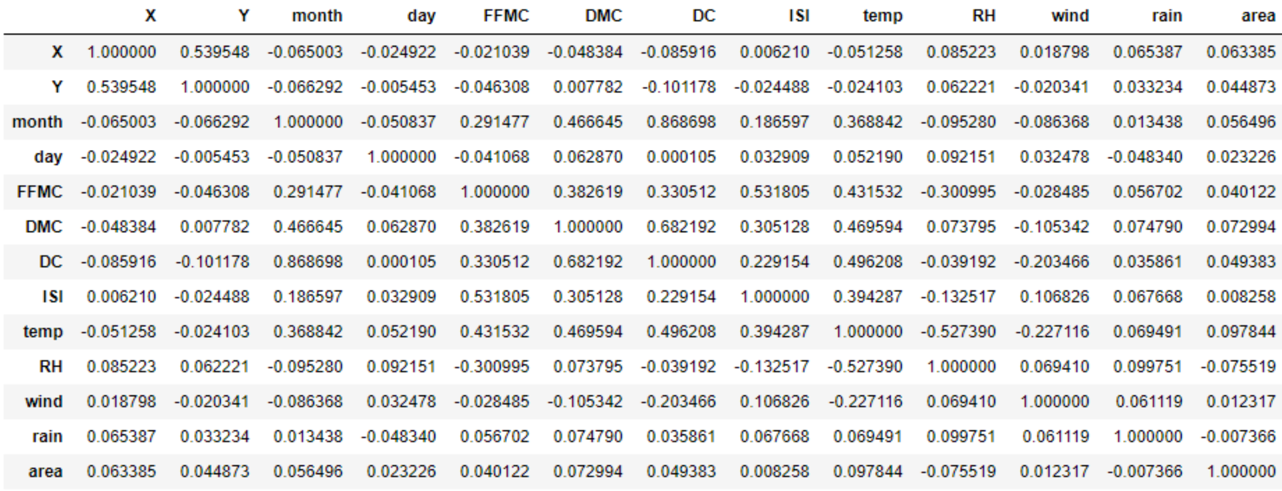
- Statistical analysis of dataset:



[91]:

	X	Y	month	day	FFMC	DMC	DC	ISI	temp	RH	wind	rain	
count	517.000000	517.000000	517.000000	517.000000	517.000000	517.000000	517.000000	517.000000	517.000000	517.000000	517.000000	517.000000	517.0
mean	4.669246	4.299807	7.475822	4.259188	90.644681	110.872340	547.940039	9.021663	18.889168	44.288201	4.017602	0.021663	12.8
std	2.313778	1.229900	2.275990	2.072929	5.520111	64.046482	248.066192	4.559477	5.806625	16.317469	1.791653	0.295959	63.6
min	1.000000	2.000000	1.000000	1.000000	18.700000	1.100000	7.900000	0.000000	2.200000	15.000000	0.400000	0.000000	0.0
25%	3.000000	4.000000	7.000000	2.000000	90.200000	68.600000	437.700000	6.500000	15.500000	33.000000	2.700000	0.000000	0.0
50%	4.000000	4.000000	8.000000	5.000000	91.600000	108.300000	664.200000	8.400000	19.300000	42.000000	4.000000	0.000000	0.5
75%	7.000000	5.000000	9.000000	6.000000	92.900000	142.400000	713.900000	10.800000	22.800000	53.000000	4.900000	0.000000	6.5
max	9.000000	9.000000	12.000000	7.000000	96.200000	291.300000	860.600000	56.100000	33.300000	100.000000	9.400000	6.400000	1090.8

- Correlation analysis for the dataset:

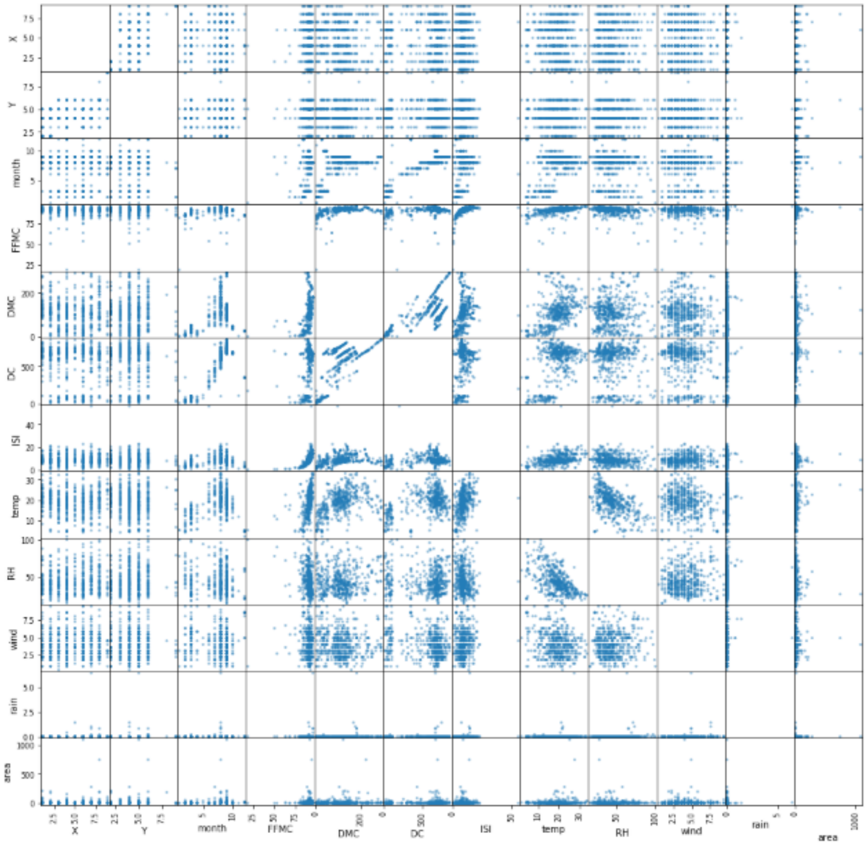


	X	Y	month	day	FFMC	DMC	DC	ISI	temp	RH	wind	rain	area
X	1.000000	0.539548	-0.065003	-0.024922	-0.021039	-0.048384	-0.085916	0.006210	-0.051258	0.085223	0.018798	0.065387	0.063385
Y	0.539548	1.000000	-0.066292	-0.005453	-0.046308	0.007782	-0.101178	-0.024488	-0.024103	0.062221	-0.020341	0.033234	0.044873
month	-0.065003	-0.066292	1.000000	-0.050837	0.291477	0.466645	0.868698	0.186597	0.368842	-0.095280	-0.086368	0.013438	0.056496
day	-0.024922	-0.005453	-0.050837	1.000000	-0.041068	0.062870	0.000105	0.032909	0.052190	0.092151	0.032478	-0.048340	0.023226
FFMC	-0.021039	-0.046308	0.291477	-0.041068	1.000000	0.382619	0.330512	0.531805	0.431532	-0.300995	-0.028485	0.056702	0.040122
DMC	-0.048384	0.007782	0.466645	0.062870	0.382619	1.000000	0.682192	0.305128	0.469594	0.073795	-0.105342	0.074790	0.072994
DC	-0.085916	-0.101178	0.868698	0.000105	0.330512	0.682192	1.000000	0.229154	0.496208	-0.039192	-0.203466	0.035861	0.049383
ISI	0.006210	-0.024488	0.186597	0.032909	0.531805	0.305128	0.229154	1.000000	0.394287	-0.132517	0.106826	0.067668	0.008258
temp	-0.051258	-0.024103	0.368842	0.052190	0.431532	0.469594	0.496208	0.394287	1.000000	-0.527390	-0.227116	0.069491	0.097844
RH	0.085223	0.062221	-0.095280	0.092151	-0.300995	0.073795	-0.039192	-0.132517	-0.527390	1.000000	0.069410	0.099751	-0.075519
wind	0.018798	-0.020341	-0.086368	0.032478	-0.028485	-0.105342	-0.203466	0.106826	-0.227116	0.069410	1.000000	0.061119	0.012317
rain	0.065387	0.033234	0.013438	-0.048340	0.056702	0.074790	0.035861	0.067668	0.069491	0.099751	0.061119	1.000000	-0.007366
area	0.063385	0.044873	0.056496	0.023226	0.040122	0.072994	0.049383	0.008258	0.097844	-0.075519	0.012317	-0.007366	1.000000

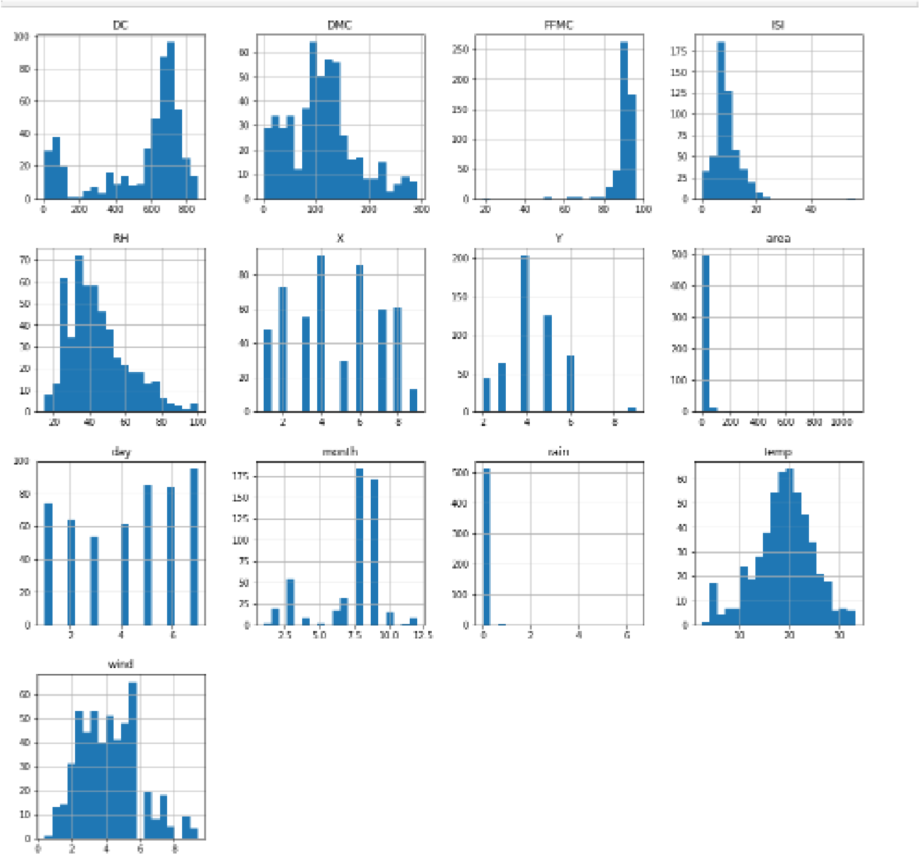
Exploratory Visualization:

Scatter Plot: A scatter plot matrix can be formed for a collection of variables where each of the variables will be plotted against each other.[5]

Here the graph is plotted between various attributes likehumid,temp,wind,X,Y etc.



- Histogram for each column:



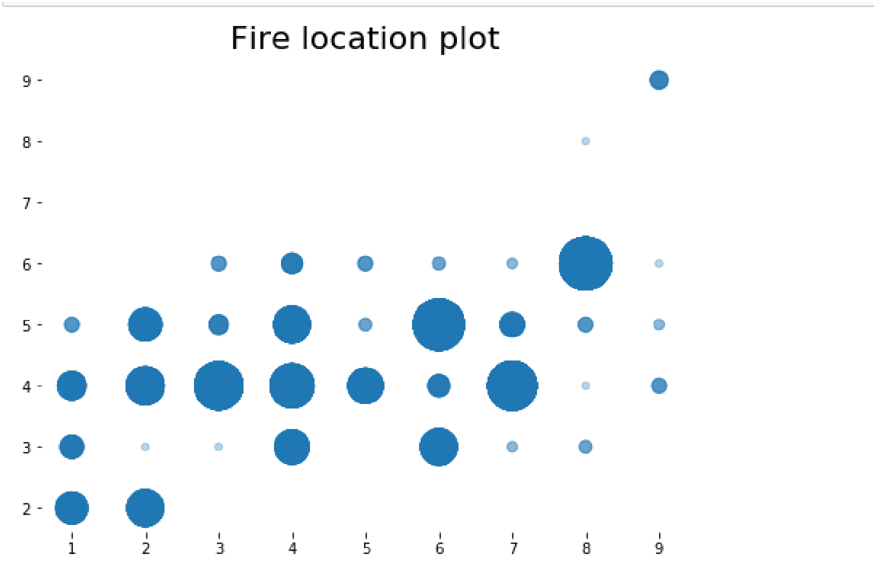
It can be observed from Histograms that: -

*There is only normal distribution for temp

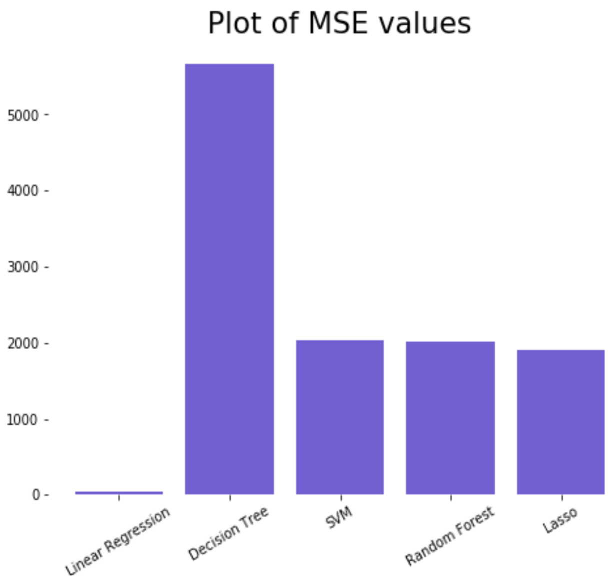
*There is no skewed data

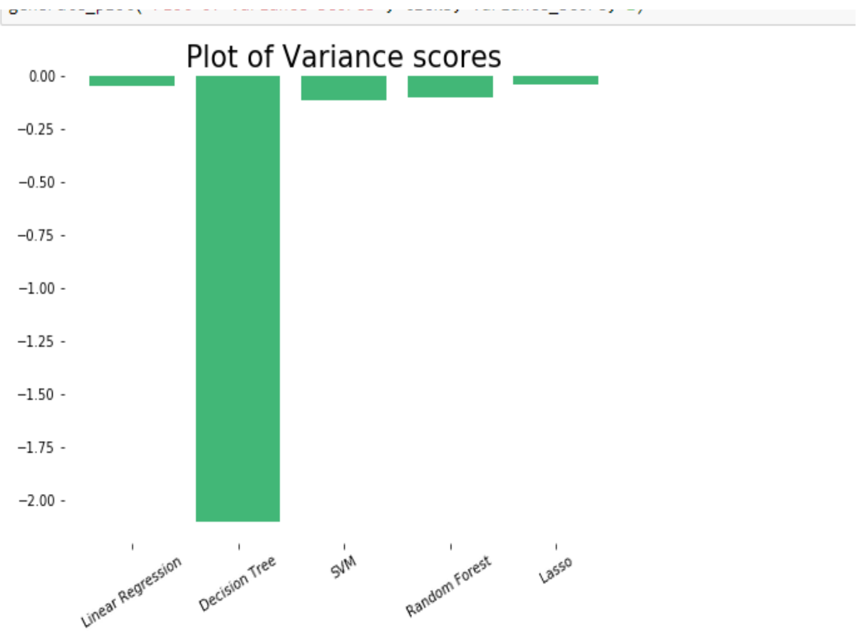
From this plot, it can be concluded that no columns have a distribution like the area column. Therefore, we can deny a linear relationship of any single feature independently with the area.

Scatter plot for the locations:The scatter() function makes a scatter plot with (optional) size and color arguments[4]



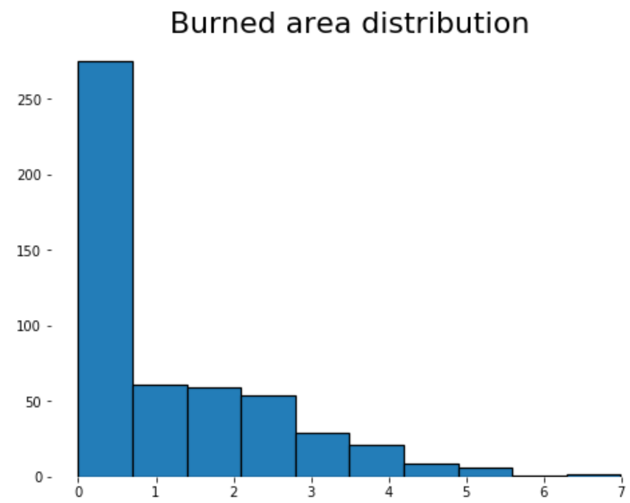
- Mean Squared Error and Variance of data before Processed:



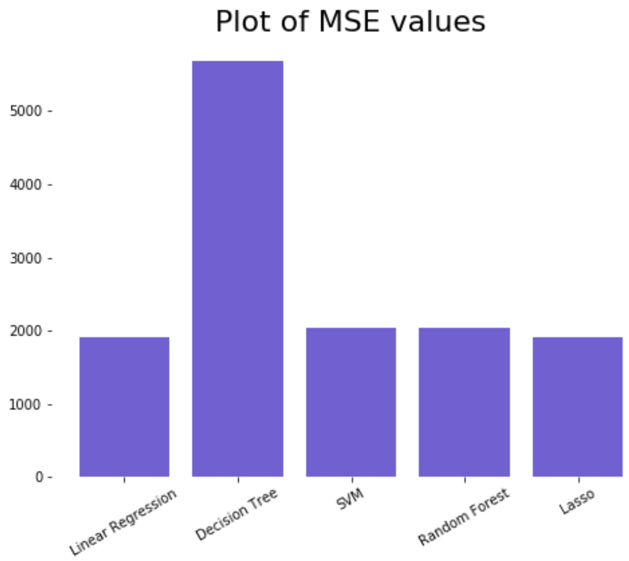


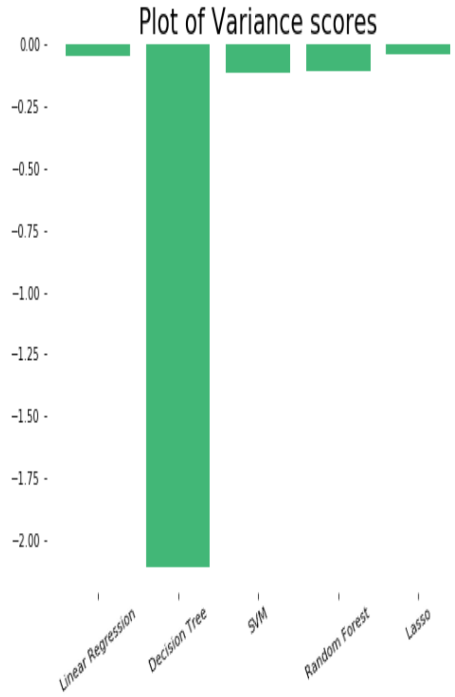
Applying Log-Transformation to the 'burned area' variable:

We can see that the errors in the prediction of burned areas from the given dataset is very high in the above mentioned model. A reason for this could be the high skewness of the 'Burned Area' variable is towards zero.

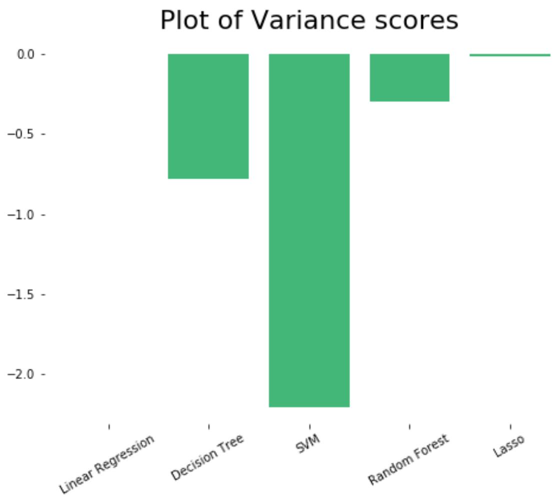


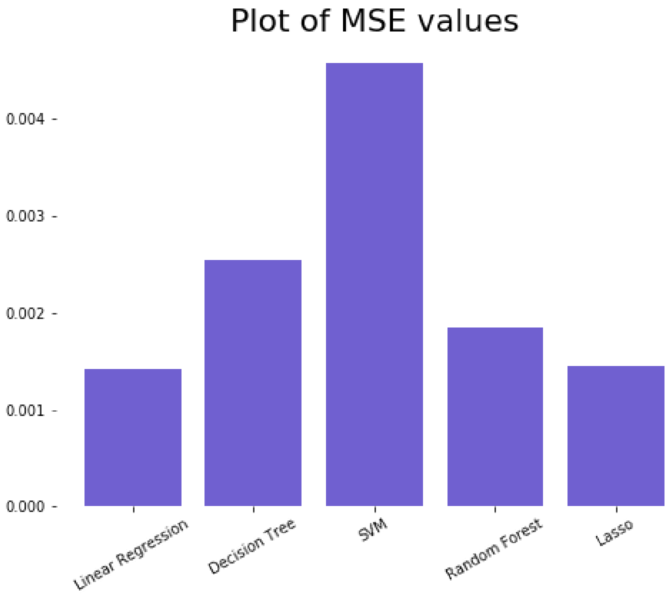
- Mean Squared Error and Variance of data Processed:


Plot of MSE values

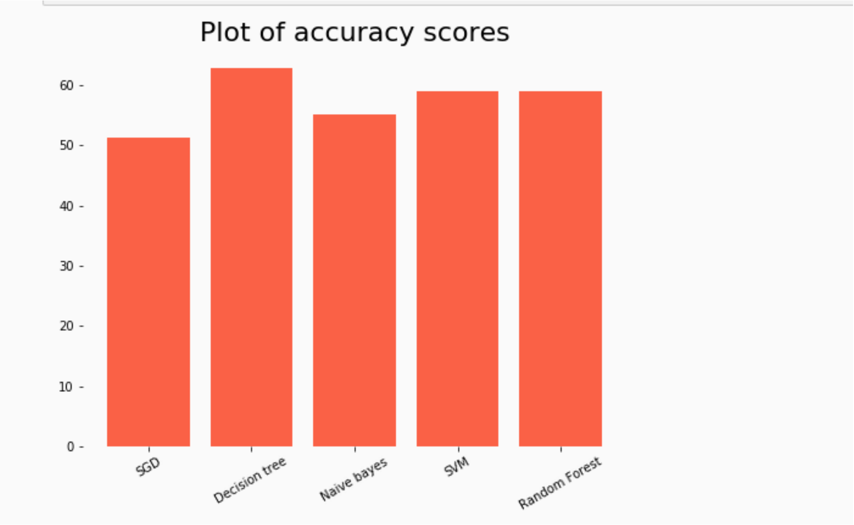

Plot of Variance scores

- Mean Squared Error and Variance for Normalized data:





After finding the best suitable regression then converting the target values to binary classes and finding accuracy among them



Reference link:https://matplotlib.org/tutorials/introductory/sample_plots.html[4]

https://subscription.packtpub.com/book/big_data_and_business_intelligence/9781784390150/4/ch04lvl1sec37/a-scatter-plot-matrix[5]

Algorithms and Techniques:

The algorithms which I am going to use are mentioned in my proposal namely Regression algorithm:

Linear Regression

Decision tree

SVM (Support Vector Machine)

Random forest

Lasso.

Regression: In regression tasks, the machine learning program must estimate – and understand – the relationships among variables. Regression analysis focuses on one dependent variable and a series of other changing variables – making it particularly useful for prediction and forecasting.

1.Linear Regression (Supervised Learning/Regression):
Linear regression is the most basic type of regression. Simple linear

regression allows us to understand the relationships between two continuous variables.

- Real time example: Linear Regression can be used to predict the sale of products in the future based on past buying behavior.
- Strength and Weakness: The main advantage is , the best fit line is the line with minimum error from all the points ,it has high efficiency but sometimes this high efficiency created disadvantage which is prone to overfitting of the data (i.e some noisy data also considered as useful data), and also it can't be used when the relation between dependent and independent variable is not linear.

http://www.stat.yale.edu/Courses/1997-98/101/linreg.htm[6]

- Parameters:https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html

2.Decision Trees (Supervised Learning – Classification/Regression)
A decision tree is a flow-chart-like tree structure that uses a branching method to illustrate every possible outcome of a decision. Each node within the tree represents a test on a specific variable – and each branch is the outcome of that test.

- Real time example:Direct Marketing,Fraud Detection
- Strengths :It is very easy to understand and interpret.The data for decision trees require minimal preparation.
- Weaknesses:Sometimes decision tree may become complex.The outcomes of decisions can be based mainly on your expectations.So this can lead to unrealistic decision trees. Since a decision tree can handle both numerical and categorical data, it’s a good choice of algorithm. The goal is to create a model that predicts the value of target variable by learning simple decision rules.
- https://www.hackerearth.com/practice/machine-learning/machine-learningalgorithms/ml-decision-tree/tutorial
- https://www.hackerearth.com/practice/machine-learning/machine-learning-algorithms/ml-decision-tree/tutorial

- Parameters: https://scikitlearn.org/stable/modules/generated/sklearn.tree.DecisionTree Classifier.html

3.Random Forests (Supervised Learning – Classification/Regression)
Random forests or 'random decision forests' is an ensemble learning method, combining multiple algorithms to generate better results for classification, regression and other tasks. Each individual classifier is weak, but when combined with others, can produce excellent results. The algorithm starts with a 'decision tree' (a tree-like graph or model of decisions) and an input is entered at the top. It then travels down the tree, with data being segmented into smaller and smaller sets, based on specific variables.

- Real Time Example: Random forest model can be applied in medical domain to identify a disease based on symptoms. Example: detection of Alzheimer's disease.
- Strengths and weaknesses.: Random forest runtimes are quite fast, and they are able to deal with unbalanced and missing data. Random Forest weaknesses are that when used for regression they cannot predict beyond the range in the training data, and that they may over-fit data sets that are particularly noisy.
- Parameters:https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

4.Support Vector Machine Algorithm (Supervised Learning - Classification)
Support Vector Machine algorithms are supervised learning models that analyse data used for classification and regression analysis. They essentially filter data into categories, which is achieved by providing a set of training examples, each set marked as belonging to one or the other of the two categories. The algorithm then works to build a model that assigns new values to one category or the other.

- Strengths and Weakness of the model: It has a regularisation parameter, which makes the user think about avoiding over-fitting. It uses the kernel trick, so you can build in expert knowledge about the problem via engineering the kernel.It is defined by a convex optimisation problem (no local minima) for which there are efficient methods.The parameters for a given value of the regularisation and kernel parameters and choice of kernel,kernelmodels can be quite sensitive to over-fitting the model selection criterion.

- https://medium.com/machine-learning-101/chapter-2-svm-support-vector-machine-theory-f0812effc72

https://stats.stackexchange.com/questions/24437/advantages-and-disadvantages-of-svm

- Parameters:https://scikit-learn.org/stable/modules/svm.html

5.Lasso Model:LASSO (Least Absolute Shrinkage Selector Operator), It uses L1 regularization technique.It is generally used when we have more number of features, because it automatically does feature selection.The black point denotes that the least square error is minimized at that point and as we can see that it increases quadratically as we move from it and the regularization term is minimized at the origin where all the parameters are zero .

- Strenth and Weakness:The biggest issue of L1 penalization is that, as any dimensionality reduction algorithm, it might lose some relevant independent variables along the way. This mainly depends on how much penalized the system.multicollinearity. L1 penalization tends to select one in a group of highly correlated variables. For many problems highly correlated variables should be selected or discarded as a group.[7]

https://www.analyticsvidhya.com/blog/2017/06/a-comprehensive-guide-for-linear-ridge-and-lasso-regression/

Parameters:https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.Lasso.html

Benchmark:

As benchmark is important to compare my results with some metric and decide which model suits the best . Here my bench mark is mean square error caluculated using the linear regression model and the bench mark is 1920

Select the best one.However I target to reach the less mean square error less than 1 in my attempt to solving this project by least mean squared error and high variance is the best model for predicting the Burned Area .

3.Methodology

Data preprocessing:

- Before data can be used as input for machine learning algorithms, it should often must be cleaned, formatted, and restructured — this is typically known as preprocessing.
- In my dataset there are no invalid or missing entries we must deal with. It refers to transformations applied to our data before feeding it to the algorithm.
- The technique that is used to convert raw data into a clean data set. The data as obtained from the UCI dataset repository have to be cleaned and to ensure that it is in the standard quality before the model creation is initiated.So clean the data i.e. removing unwanted data or replacing null values with some constant values or removing duplicates if any.
- Then finding the correlation for each features with the target variable.
- Data transformation is a very crucial process in data preprocessing.It involves normalization and aggregation.
- Normalization: Normalization makes training less sensitive to the scale of features, so we can better solve for coefficients.
- After preprocessed dataset is split into two halves of varying sizes at different times for use as training and testing datasets for model creation and selection of which of the models performs best. The data

set used for training is mainly a portion from the dataset from which the regression algorithm used learns the burned area of the model created from each model.

Implementation:

The implementation process can be split into two main stages.

1.On data before processing

Model	Before Processing(Mean Squared Error)	Before Processing(Variance)
Linear Regression	1913.263964775979	-0.04
Decision tree	5671.534384615384	-2.10
SVM	2039.790466031507	-0.11
Random forest	2011.557492115028	-0.10
Lasso	1904.404987844155	-0.04

Here Lasso has least mean squared error and high variance

2. On data after processing

Model	After Processing (Mean Squared Error)	After Processing (Variance)
Linear Regression	1913.26396477597	-0.04
Decision tree	5691.044186538461	-2.11
SVM	2039.790466031507	-0.11
Random forest	2031.306570461497	-0.11
Lasso	1904.4049878441	-0.04

	55	

Here Lasso has least mean squared error and high variance

3.ON NORMALIZATION OF DATA:

Model	Normalization(Mean Squared Error)	Normalization(Variance)
Linear Regression	0.001426223498840	0.00
Decision tree	0.002543007361276	-0.78
SVM	0.004579753060572	-2.20
Random forest	0.001853774628730	-0.30
Lasso	0.001443763815459	-0.01

Here Linear Regression has least mean squared error and high variance

Refinement:

- I started my project I have started using basic regression algorithm implemented by knowing about it from various sources.Then I explored for different algorithms for more better results and reached Lasso regression algorithm with least mean squared error and high variance.
- Later after normalization of data I reached finally a best suited algorithm i.e.,Linear Regression algorithm where I achieved least mean squared and high variance as I target to get.
- Later I have converted variables in to binary classes so that we can find accuracy of different model like Decision tree,Navie bayes,SVM,SGD,Random forest and I have find highest accuracy in decision tree with 62.820.

Model	Accuracy
SGD	51.28205128205128
Decision tree	62.82051282051282
Navie Bayes	55.12820512820513.
SVM	58.97435897435898.
Random forest	58.97435897435898.

Here SVM and Random forest has same accuracy and the SGD has least accuracy.

Result:

3.ON NORMALIZATION OF DATA:

Model	Normalization(Mean Squared Error)	Normalization(Variance)
Linear Regression	0.001426223498840	0.00
Decision tree	0.002543007361276	-0.78
SVM	0.004579753060572	-2.20
Random forest	0.001853774628730	-0.30
Lasso	0.001443763815459	-0.01

Here Linear Regression has least mean squared error and high variance

And to find the accuracy of target variable:

Model	Accuracy
SGD	51.28205128205128
Decision tree	62.82051282051282
Navie Bayes	55.12820512820513.
SVM	58.97435897435898.
Random forest	58.97435897435898.

Decision tree has highest accuracy.

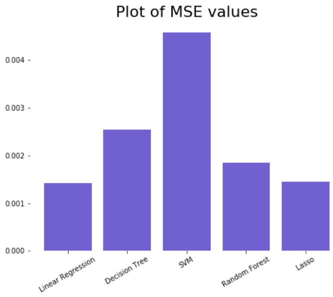
Justification:

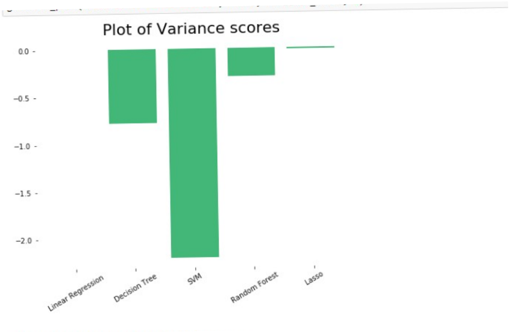
The models used in the burned area based on features were compared and researched, the results showed that all models have good predictive ability, and Linear Regression has least mean squared error and high variance and decision tree algorithm model showed the best accuracy on the target variable, sensitivity and specificity, with the best predictive ability, can be very good for burned area. The benchmark accuracy is around 1920.The optimised model has obtained mean squared error(0.00142622349884) better than my benchmark.So,it is performing well better.It is tell us that regression Report shows good results of mean squared error and variance.

5.Conclusion

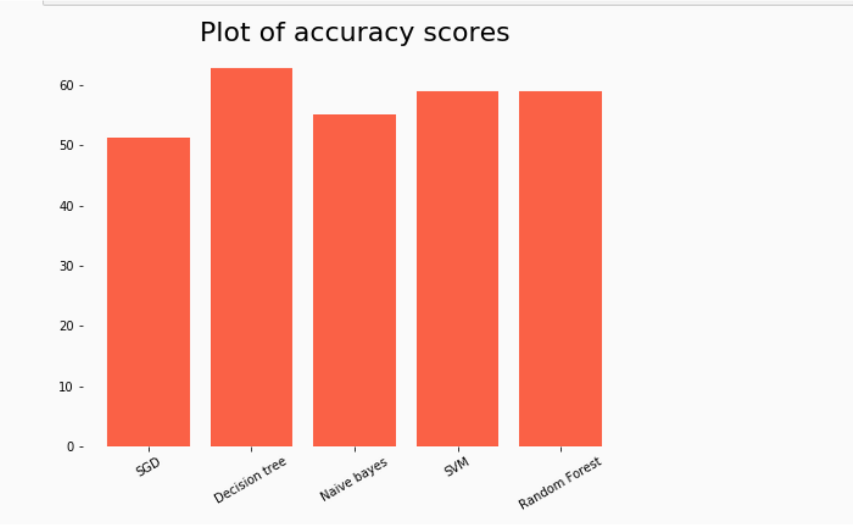
Free form Visualization:

Mean Squared Error and Variance for Normalized data:





After finding the best suitable regression then converting the target values to binary classes and finding accuracy among them



Reflection:

1.During my process of doing this project, I learnt how to visualize, and understand the data.

2.I have learnt that Data Cleaning plays crucial part in Exploratory Data Analysis(EDA).

3. Removing the data features that are not necessary in evaluating an model is most important.

4.I have come to know how to use best algorithm in different conditions for the data using appropriate techniques.”sklearn” helped me a lot in knowing a lot about the respective algorithms and their parameters.

5.I am aware of how to find accuracy to regression algorithm by converting them to binary classes

6 At last , I have learnt how to grab a data set from machine learning repository and applying techniques on it and to stick to best techniques to get good results.Finally I am glad that I can solve a problem and acquire a solution using machine learning concepts. Improvement: In this project, I have evaluated the different regression algorithm for Insight into forest firedataset. Based on the features like temperature,humidity,month,wind.The model is very appropiate to judge not only burned area under forest fires but also is used now a days to predict the

weather conditions.. In future, research can be use more refine technique to give more accuracy and deal with the some other issue like location and also used different regresion algorithm to get better results I hope to expect.