Prediction of Forest Fires using Machine Learning

18.11.2019

Abstract

In this project, we have got one dataset from UCI ML Repository [1].

Firstly, we perform some data preprocessing on the train and test data sets such as:

Checking if there are any missing data columns.

We also performed some visualization using:

1. Seaborn heatmap: To visualize the correlation between numerical columns in data.

2. Scatterplot: To visualize how closely two variables are correlated.

3. Pie chart: To see the recurring occurrence of the days.

4. Seaborn Barplot: To observe the comparison between two groups of data.

Now, before moving onto our model building stage, we perform label encoding on the categorical columns. Next, we split the given labelled data into training and testing data. This is done so that we can plot the Machine Learning models. Now, since this is a regression problem with clear outliers and cannot be predicted using anu reasonable model we compare different regression methods, we have done a comparison of the following models :-

(a) Linear Regression
(b) Lasso
(c) Ridge
(d) Elastic Net
(e) Support Vector Regression
(f) Random Forest Regression

Finally, we build our model based on the x_train and y_train data and perform model validation on the x_test, y_test data. Based on results of the cross-validation, we choose

(a) Support Vector Regression
(b) Random Forest Regression

to predict for the test set data.

Problem Questions:

In this assignment, we have addressed the following problem statements:

(1) Which days and months have the highest recorded forest fires
(2) How does the temperature vary in accordance to the days and months
(3) Which model is best suited for our data analysis
(4) How each variables affect the number of forest firing occuring
(5) Are there any outliers in dataset

Dataset Description:

The aim is to predict the burned area of forest fires, this data was collected from the Montesinho National Park in the northeast region of Portugal, by using meteorological and other data from January 2000 to December 2003 and data was recorded whenever there was a forest fire spotted (ref - UCI ML) [1]

- 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

The map of Montesinho was split into a 9x9 grid as shown below and the X and Y axis together specify the location of the forest fire.[2]

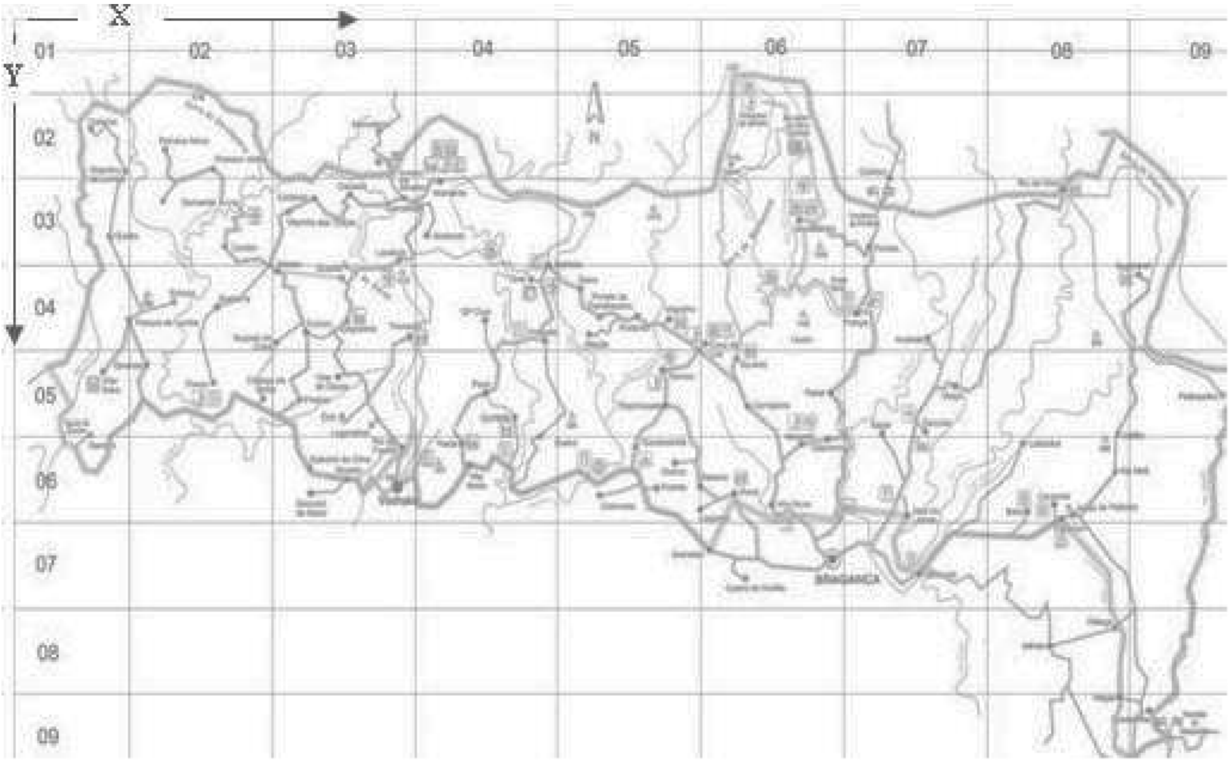

Fig 1: Map of Montesinho [2]

Month and Day were chosen as temporal variables because the average monthly conditions vary quite distinctly and the day of the week can also have an effect on forest fires as most of the fires have a human cause.

The next four variables (FFMC,DMC,DC and ISI) come from the FWI (forest Fire Weather Index) System, It is a Canadian system to determine the intensity of the fire.

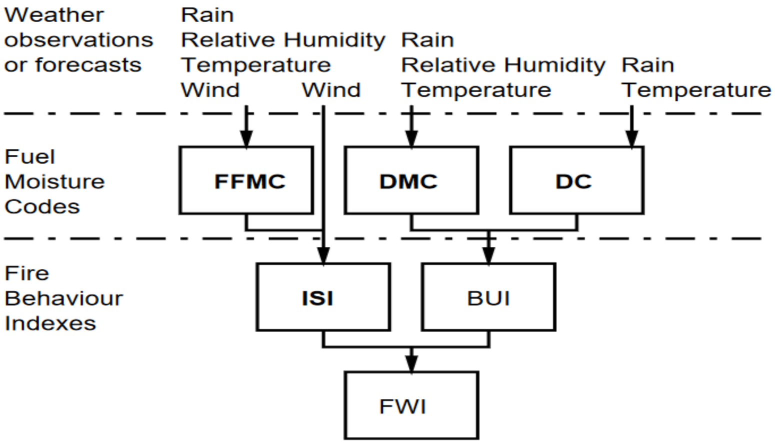

Fig 2: Forest fire weather index system [2]

The next four variables (temp, RH, wind, rain) are the four weather attributes used by the FWI system and were recorded by the meteorological station. These were recorded instantly with the exception of rain which denotes the accumulated precipitation within the previous 30 mins.

The area burned has 247 samples with a zero value that means the area burned was lower than 1*ha*/100 = 100 squares metres.

Data Storage Plan:

Stored in CSV files

Exploratory Data Analysis:

I. Data Preprocessing:

Before going through our dataset, we check for missing data that needs to be taken care of. This is because in case of the existence of missing data, it can be very difficult to recognise the real problem. Also, data analysis gets performed only on parts of the data which might not give accurate results. We also calculated the total percentage of missing data in our dataset, which turned out to be 0%. Therefore, there was no missing data present in our dataset.

We also use StandardScalar from Scikit-learn as part of our Data Preprocessing. Many Machine Learning estimators implemented in Scikit-Learn need dataset Standardization. It is to be made sure that each feature looks somewhat like standard normally distributed data; otherwise there exists a chance that they might behave poorly [3]. Also, since there is a chance of the test set leaking into the model, we have used fit transform on the training set after splitting the data.

Our dataset has 247 zero valued samples that denotes a positively skewed dataset. We also notice that there are many outliers in this dataset. Thus in order to reduce the skewness and make the symmetry of the data better, we use the logarithm function on the attribute area: -

$y = \ln(x + 1)$

II. Data Visualization

We have performed an extensive Exploratory Data Analysis on our dataset by plotting different charts and graphs. We shall go through them one by one:-

We first import the seaborn library and visualise the correlation matrix of the dataset. We check for multicollinearity as any Pearson Correlation Coefficient greater than 0.7 would give rise to the fact that the data is multicollinear [4]. However, such cases are not seen in our dataset.

We also plot a barplot of Temperature v/s Months.

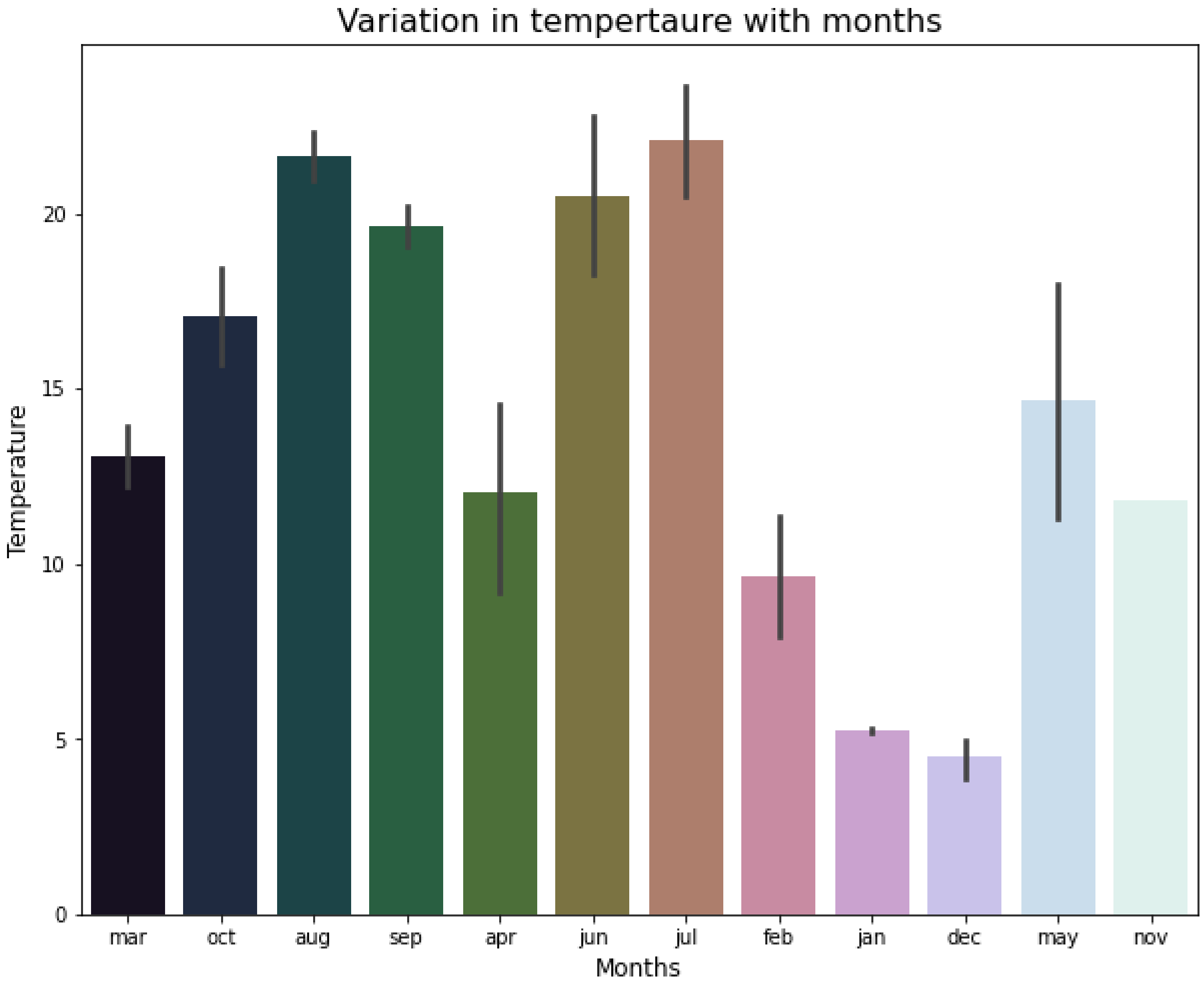

Fig 3 : Variation in temperature with months

OBSERVATION: Among the 12 months, it is seen that in the month of July, the temperatures were the highest. This can also be taken as a factor behind the most number of forest fires being caused in the month of August.

Next, we plot a pie chart that shows the temperature most occurring throughout the time of twelve months.

Here, we have taken the first 20 data points due to the large amount of data present in our dataset.

First, we shall plot a bar chart to show on which days the most number of forest fires had occurred.

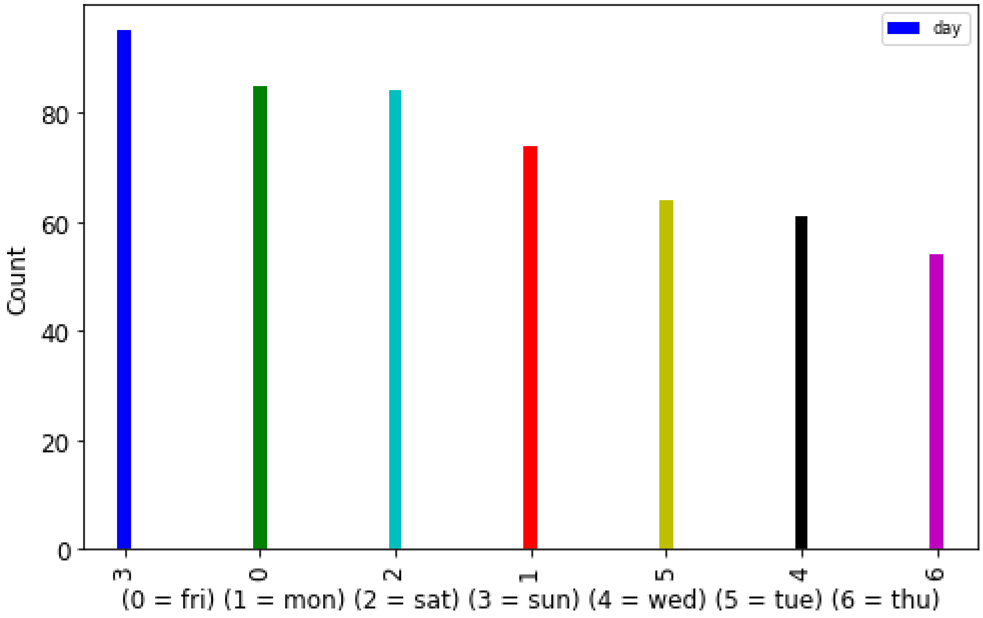

Fig 4: Bar plot of number of Forest fires with respect to days

OBSERVATION: We have label encoded the data that shall favour us when we will build our Machine Learning models. Thus, we can see that Sunday is the day when most of the forest fires have occurred.

This can also be depicted with the help of a pie chart that tells us the exact percentage of forest fires taking place on each day (of the first 20 data points).

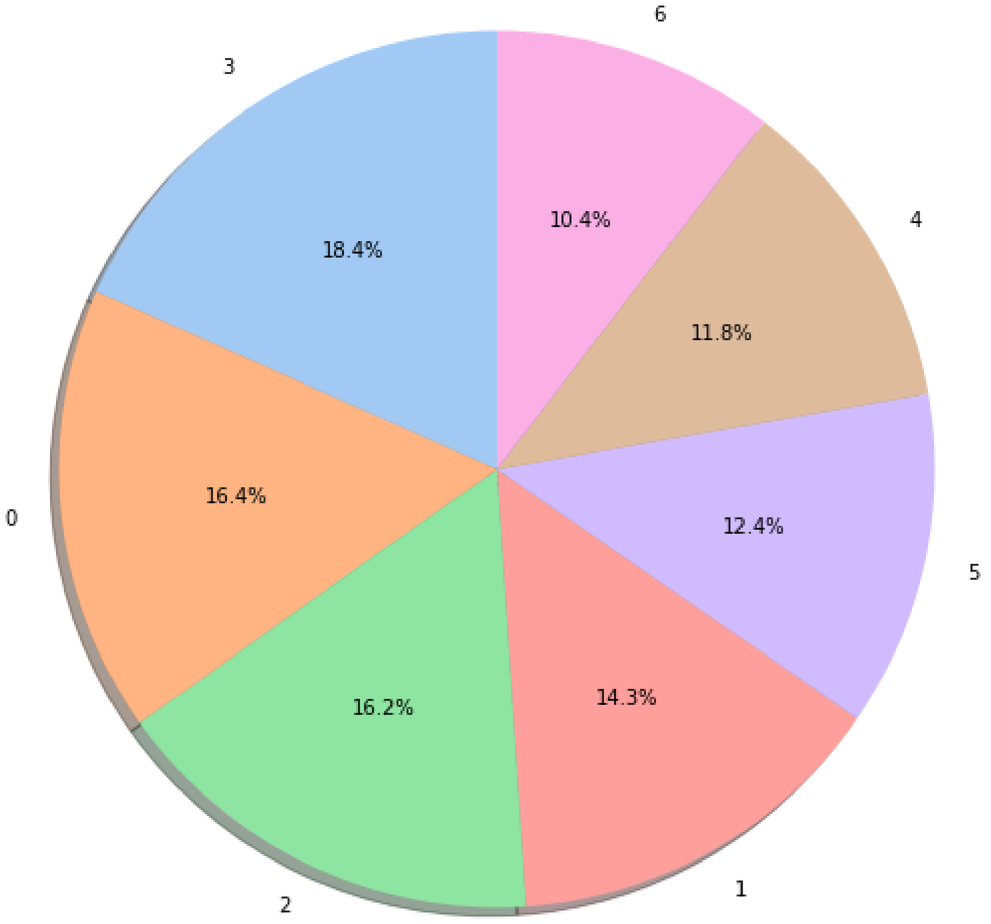

Fig 5: Pie chart depicting the days where the most number of forest fires occurred

OBSERVATION: It is noticed here that 18.4% of the forest fires took place on Sunday thereby supporting our previous bar plot. The least number of forest fires took place on Thursdays.

EXAMINING HOW EACH VARIABLE AFFECTS THE NUMBER OF FOREST FIRE OCCURRING (SEGREGATED WITH RESPECT TO DAYS):

Before plotting a histogram, it's necessary to identify the ranges and create bins of the. We then distribute it into a series of intervals [5]. Thus we have plotted detailed histogram charts to show how the number of forest fires occurring are affected by each variable in the dataset (such as RH, DMC etc.). To make it clearer to visualise, we have segregated this with every day of the week.

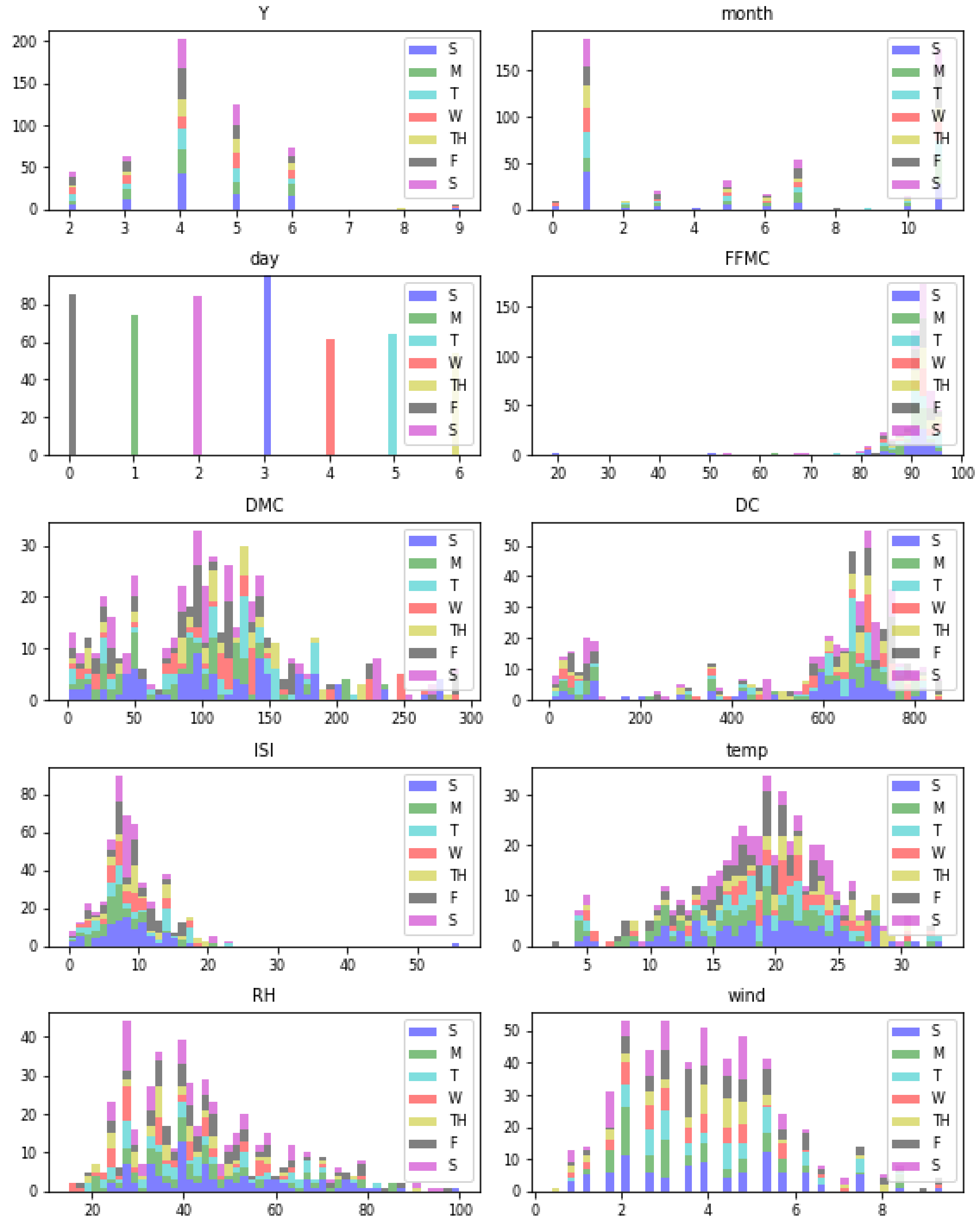

Fig 6: Histogram Distribution based on how the attributes of the forest fire dataset effect the number of forest fires caused

The X-axis signifies the range of the variables that the histogram is focusing on. The y-axis represents the number of forest fires it has had a major effect on. We have segregated each bar based on each day of the week starting from Sunday and ending with Saturday.

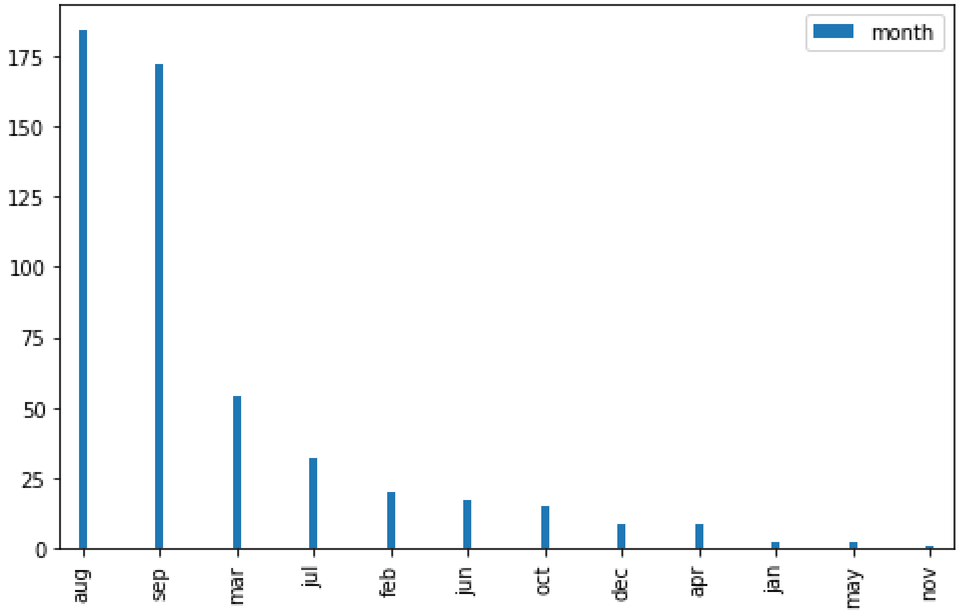

Fig 7: Barplot depicting the highest number of forest fires caused in each month

Observation: - We have also plotted a bar graph based on the months and the number of forest fires that occurred each month. We can see that here(above 175 of them), the month of August witnessed the most number of forest fires throughout the year , followed by the month of September. November however witnessed the least number of forest fires.

BOXPLOT:

```
dataset.boxplot(column='log_area',by = 'day')
```

(Log area is basically where we've taken the log of the area attribute)

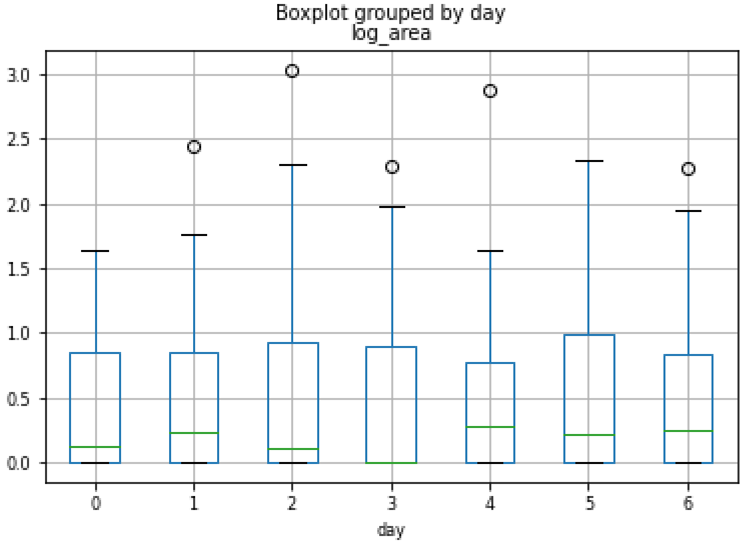

Fig 8: Boxplot grounded by day

OBSERVATION: The boxplot infers that there are many outliers present in the data. The boxplot for days Friday (0), Monday (1), Thursday (4), Wednesday (6) have a similar range which implies that these days have an equal chance of having a forest fire. Days Saturday (2) and Tuesday (5) have an equal chance of having forest fires.

2nd BOXPLOT:

```
dataset.boxplot(column= 'log_area', by = 'month')
```

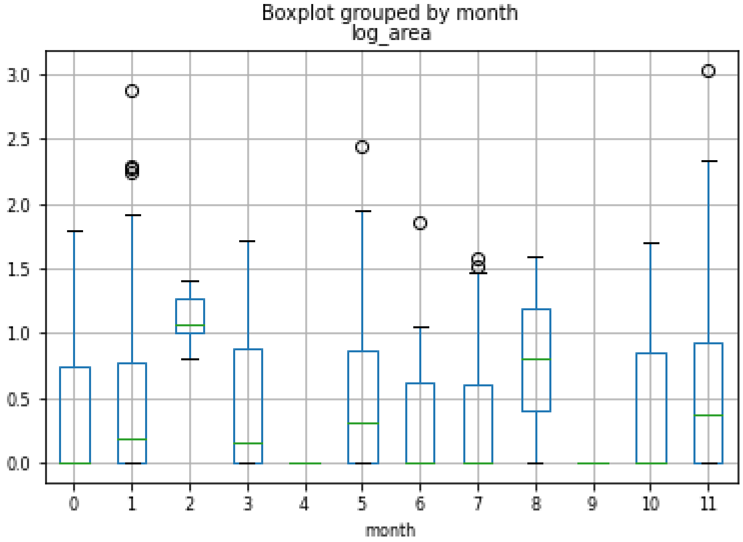

Fig 9: Boxplot grounded by months

OBSERVATION: The boxplot has many outliers in months August (1), June (6) and March (7). The months April (0), August (1), February (3), July (5), June (6), March (7), October (10) and September (11) have an equal chance of forest fires. The boxplot of August (1) is closely following a Normal distribution hence this month has a high threat of forest fire.

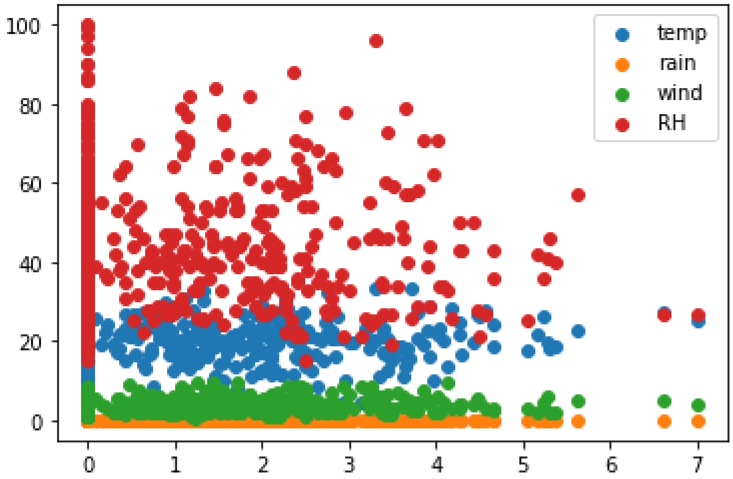

Fig 10: Scatterplot depicting the correlation

OBSERVATION: The correlation between Temperature, Rain, Wind, Relative Humidity with Forest fires is depicted in the scatter plot. The scatter plot shows that there is no relation between rainfall and forest fires, wind and forest fires, relative humidity and forest fires. There is a positive correlation between temperature and forest fires. Hence as the temperature increases, there are high chances of forest fires.

Model Design:

I. Choice of Model:

(1) Linear Regression
(2) Lasso
(3) Ridge
(4) Elastic Net
(5) Support Vector Regression
(6) Random Forest Regression

II. Validation:

Linear Regression:

For our model, we have deployed the Ordinary Least Squares method for Linear Regression. With the help of OLS, we can try to reduce the sum of squares residuals. It is used to understand the relationship between one dependent variable and one or more independent variables [6].

For our assignment, we have built a Linear Regression model so as to understand the relationship between the actual area burnt during the forest fires v/s the areas that were predicted to be burnt. We can see that most of the data-points that have been plotted would have been close enough to the line that Linear Regression attempts to draw to reduce the residual sum of squares of the points lying close to the regression line. We can also spot some outliers on our Regression Model.

Lasso Regression:

The Linear Regression model has a low bias while fitting the training data and thus introduces a high variance while fitting the testing data. Lasso regression introduces a little bias while fitting the training data which results in having a lot less variance when fitting the test data.

Formula:- sum of the residuals + λ(the slope)

When using Lasso regression on datasets with many parameters, by increasing the value of λ we can reduce the value of some less useful parameters to zero thus removing those parameters and making the equation a lot more simpler. Thus, Lasso Regression is useful in datasets where most of the parameters are not useful.

Ridge Regression:

Ridge Regression is similar to Lasso Regression in the sense that it also introduces some bias while calculating the least sum of residuals while fitting the training data which results in a lesser variance when fitting the test data. The difference is in the formula

Formula:- sum of residuals + λ(the slope)^2

The difference in formula makes a difference in the parameters used in fitting the training set. While increasing the value of λ, we can shrink the values of some parameters asymptomatically close to zero but not equal to zero. Thus, Ridge Regression is more useful in datasets where most of the parameters are useful.

We have observed multicollinearity in our dataset, thus we have used Ridge regression to analyse that.

Elastic Net Regression:

Elastic Net Regression is a combination of both Lasso and Ridge Regression models. It calculated on it's own the best fit for the training data based on the formula.

Formula:- sum of the residuals + λ1(the slope) + λ2(the slope)^2

The formula has two λ thus it uses cross validation to select the best values for both the λ to fit the training data. As we have previously used Elastic Net Regression in our assignment, we notice that there might be certain function loss while training the model. Thus we have used Elastic NEt Regression to regularize it.

SVR :

Considering the method of Ordinary Least Squares that is followed by Linear Regression, we know that the error is calculated based on the distance of the points from the regression line. However, in the case of SVR, instead of a line, a tube is pictured. That is known as the margin or error, or the amount of error that can be tolerated. This tube is known as the "Epsilon-Insensitive tube". We have used SVR as one of our Machine Learning models, to find the best fitting line.

Random Forest Regressor:

Random forest regression works by using "Ensemble Learning" which precisely means taking one algorithm multiple times to make something much more powerful than the original algorithm. As Random forest works by using multiple decision trees to finally give a result, we have used that instead of the decision tree model for our assignment as it can predict continuous outputs.

REC CURVE:

Regression Error Characteristic (REC) curves are a generalization of Receiver Operating Characteristic (ROC) curves for regression. ROC curves is a powerful visualization tool for comparing different classification models. REC curvers plot the absolute error (tolerance) on the x-axis versus the percentage of points predicted within the tolerance on the y-axis.

The Area Under the Curve (AUC) represents the estimate of the model's accuracy. The Area Over the Curver (AOC) represents the estimate of the error. The R^2 value can also be estimated by the ratio of the models AOC and the AOC of the nul-model.

INFERENCE:

We observe that the RMSE and R^2 scores for Linear regression were 1.16 and 1.13 respectively for train and test sets while the R2 score was near to 0.09.

the RMSE and R^2 scores for Lasso regression was 1.21 and 1.28 respectively for train and test sets and the R2 score was near to 0.09

Again, the RMSE and R^2 scores for Ridge regression were 1.16 and 1.31 respectively for train and test sets while the R2 score was near to 0.09

Similarly, for Elastic Net, the RMSE AND R^2 scores were 1.20 and 1.27 respectively for train and test sets while the R2 score was near to 0.14

For Support Vector Regressor the RMSE was 0.956 and 1.23 for the train and test set respectively while the R2 score was near to 0.20

Lastly, Random Forest Regressor scored an RMSE score of 0.044 and 0.036 for the train and test set respectively while the R2 score was near to 0.99

We know that the RMSE score lower than 0.5 shows that it is a good model, while the R^2 score is better if it's near to 1.
We observe that the Random Forest Regression model has fit both of these conditions perfectly and thus is the best model for this case.

References

[1] D. Aha, "UCI Machine Learning Repository," National Science Foundation, 2007. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Forest+Fires.

[2] Paulo Cortez and Anibal Morais, "A Data Mining Approach to Predict Forest Fires using Meteorological Data," *Research Gate,* p. 13, 2007.

[3] D. Cournapeau, "Scikit - learn," Microsoft, Zalando SE, The University of Sydney, Quansight Labs, 2007. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html.

[4] "Encyclopedia," Oxford University Press and Columbia Encyclopedia., 2019. [Online]. Available: https://www.encyclopedia.com/social-sciences/applied-and-social-sciences-magazines/ordinary-least-squares-regression.

[5] S. Jain, "Geeks for Geeks," [Online]. Available: https://www.geeksforgeeks.org/plotting-histogram-in-python-using-matplotlib/.

[6] "Encyclopedia," Oxford University Press and Columbia Encyclopedia., 2019. [Online]. Available: https://www.encyclopedia.com/social-sciences/applied-and-social-sciences-magazines/ordinary-least-squares-regression

[7] "GitHub," [Online]. Available: https://github.com/geeky-bit/SVR--Decision-Trees--Random-Forests--DeepNeuralNets--to-PREDICT-FOREST-FIRES/blob/master/Predict%20Forest%20Fires%20(SVR%2C%20Decision%20Trees%2C%20Random%20Forests%20%26%20DeepNets(Keras%20with%20tensorflow%20backend)%20).ipy.