**App Link -** [**https://forest-fire-predd.herokuapp.com**](https://forest-fire-predd.herokuapp.com)

**Data Collection:**

1. In data collection I observed that headers are coming from the second row and indexes have some labels so I need to take care of that thing with read\_csv options.
2. I also observed that in data set column names have some extra spaces leading and trailing spaces so we have to take care of that thing also, just remove the extra space to make unique columns

**EDA:**

**Missing values**:

1. I observe that there are only two missing values, so there are only two rows that have missing values, so we can remove them easily.

**Date Features:**

1.In the data set I observed that there are only three features that are related to date and year features having the same value ,that is 2012 so it will not impact our output so we can remove this year feature and for other features we can calculate the number of days.

1. I observed that number days is depend on the output feature so I plotted, And it is easily visible but number of days depend on the out feature

**Numerical features:**

1. I separate out numeric features and I try to figure it out, is there any discrete feature in that. So my discrete feature criteria is I am just checking the unique categories in numerical features if numeric category is less than 15 then I can consider as discrete feature, but in our case all the features of numeric feature so we don't have any discrete features
2. So here I observed that we don't have any discrete features so all the numeric features are containers in nature so I just try to analyze the distribution of those continuous features. Generally for machine learning algorithms we always prefer Gausian normal distribution so we just check the distribution. And all are almost follow Normal distribution

**Categorical features:** so in our data set there is only one feature that is of category in nature for that is classes so basically that category is of nominal type so we can use nominal encoding for that. Because this feature doesn't have any weight so we can use simply label encoding

**Variance:** I observed that many of a features has a very high variance so we can opt features scaling for at least those machine algorithms that have euclidean distance and manhattan distance concept

**Balanced/imbalanced dataset :** I verified that our data set is a balanced data set so we can use accuracy as a performance Matrix for classification problem

**Outlier:** I tried to analyze that , Is there any outlier in continuous features or not I observe that there are some outlier but our data set size is like only 240 rows so we can go with outliers