DATA ANALYTICS -1

CLASSIFICATION PROJECT

EARTHQUAKE CLASSIFICATION

SUBMITTED BY:-

Anjali Bhatnagar 2019201012

TASKS UNDER CONSIDERATION

The aim of this project is understanding and implementing the classification models.

The models that we are going to consider are:

- Decision Trees
- KNN Classifier

The data given to us is about the occurrence of earthquakes in a geographical region. To begin with we will first and foremost clean the data, so that it becomes ready for training and we can make the best use out of it. We began this classification problem with cleaning the data and then choosing the features which we will use for prediction. We are given various features to be considered : the Depth of the earthquake, the location when it occurred, longitudinal and latitudinal coordinates and many other features from which we have to predict Mw which is the magnitude of the earthquake but only as a label that is we chose a threshold say T and if the value of the Mw >= T we classify it as 1 that is Earthquake else as 0.

Flow of the Project :

- Data Cleaning
- Selection of threshold and train test split parameter
- Training and testing the decision tree classifier for data
- Training and testing the KNN classifier for data
- Plotting ROC curves and drawing conclusions
- Comparison of the two classifiers
- Drawing conclusions from the analysis
- Feature Selections
- Analysis results

DATA DESCRIPTION

Drawing conclusions from the data which we will use for feature selection and data cleaning

The total no of data rows is 52989. The following are the various features of the data that are available to us . However we wont be using all of the data which is given rather we will clean the data and extract and use the relevant features from it.

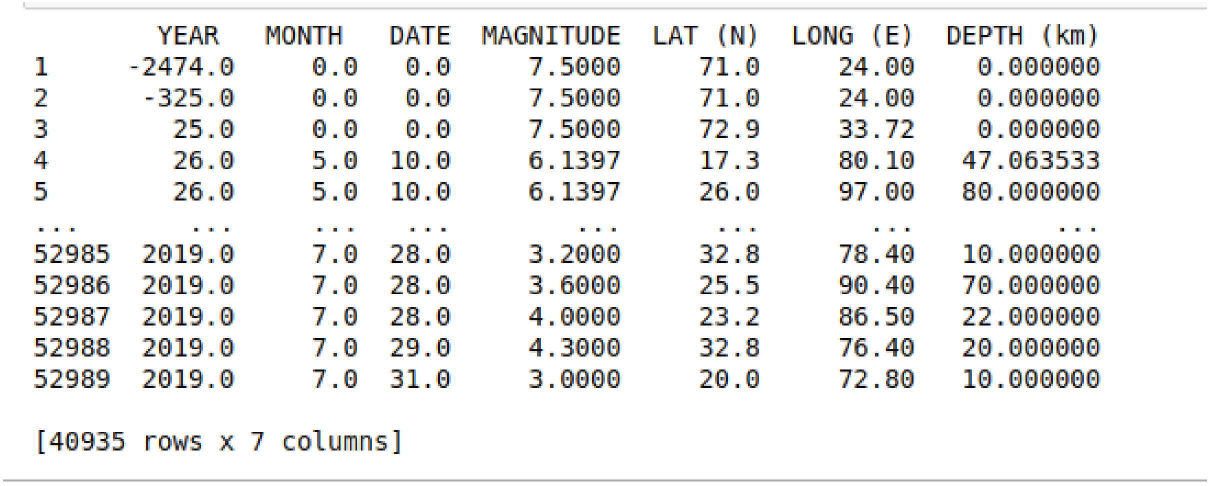
- Sl. No.: Serial Number
- Year, Month, Day: Date of a particular earthquake as per UTC (Coordinated Universal Time)
- Origin Time of earthquake in UTC and IST (Indian Standard Time) in [Hour: Minute: seconds] format
- Magnitude of Earthquake: There are a different way to represent the magnitude of an earthquake. For your study, you can consider Mw, since we are deriving other types from Mw only.
- GPS Location in terms of Latitude(Lat) and Longitude(Long) of earthquake
- Depth: Depth of occurrence of an earthquake in kilometre
- Location: Name of a region where an earthquake took place
- Source: The agency from which we have gathered the data, for e.g. IMD= Indian Meteorological Department, Min. of Earth Science, Government of India

DATA CLEANING

- Removing Irrelevant and redundant data columns which are not relevant for our analysis i.e columns such as serial number , time , reference(source) , location which do not play a major part in predicting the magnitude of an earthquake .
- Restricting our analysis to only relevant and important data columns : Year , Month , Date , Latitude , Longitude and Depth.
- For columns latitude and longitude if the value is empty string or not a number there is no point considering that data so we have removed these rows.
- The fields corresponding to latitude and longitude contained some special characters which were removed and the numerical value of the coordinates conserved.
- Rows for which date , month and year values were missing we filled them with 0 .
- Rows for which the column depth didn't have an appropriate value we substituted it by the mean of values of that column.
- For further processing the data all the values were converted to float .
- Rows which don't have any value for the attribute Magnitude were removed .

Data after the Cleaning Operation

This is how the data looks after the cleaning operation has been performed on it.



```
         YEAR  MONTH  DATE  MAGNITUDE  LAT (N)  LONG (E)  DEPTH (km)
1     -2474.0    0.0   0.0     7.5000     71.0     24.00    0.000000
2      -325.0    0.0   0.0     7.5000     71.0     24.00    0.000000
3        25.0    0.0   0.0     7.5000     72.9     33.72    0.000000
4        26.0    5.0  10.0     6.1397     17.3     80.10   47.063533
5        26.0    5.0  10.0     6.1397     26.0     97.00   80.000000
...       ...    ...   ...        ...      ...       ...         ...
52985  2019.0    7.0  28.0     3.2000     32.8     78.40   10.000000
52986  2019.0    7.0  28.0     3.6000     25.5     90.40   70.000000
52987  2019.0    7.0  28.0     4.0000     23.2     86.50   22.000000
52988  2019.0    7.0  29.0     4.3000     32.8     76.40   20.000000
52989  2019.0    7.0  31.0     3.0000     20.0     72.80   10.000000

[40935 rows x 7 columns]
```

Threshold Selection

The threshold is that value of magnitude of the earthquake above which we classify it as an earthquake i.e assign the label 1 and below it we assign the value 0 .

In our project the constraint on threshold was that it should lie between the range 4 to 5.

Threshold Value Selected : 4.1

Train – Test Split

The train test split in data defines how much of the data available to us is going to be used for training and how much data is going for testing our model. We have done random splitting to avoid bias.

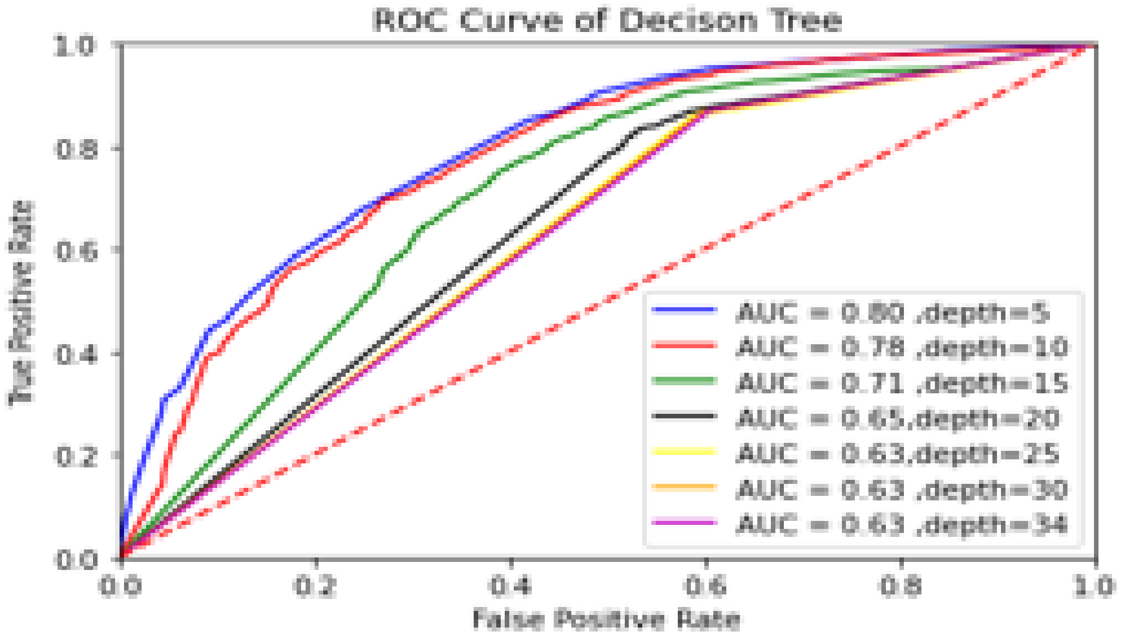
We have used 80 : 20 split in this classification project .

80 % of the total data after cleaning is our train data

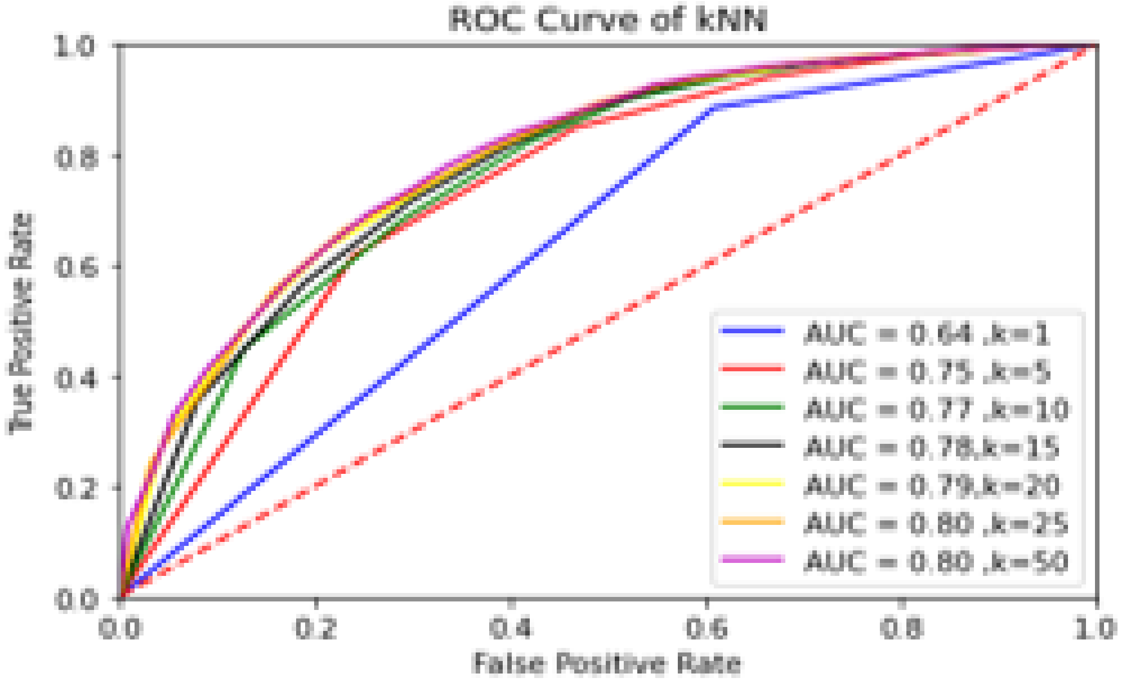
20 % of the total data after cleaning is our test data

TASK 1 : PLOTTING THE ROC CURVES FOR BOTH THE CLASSIFICATION MODELS

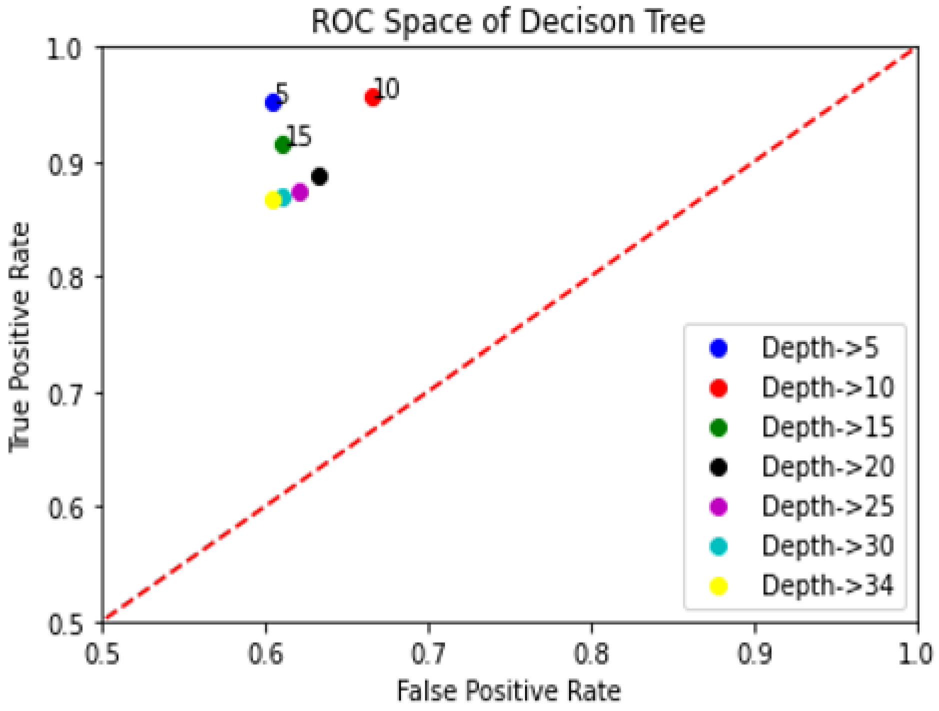
- Classification Model 1 : Decision Tree

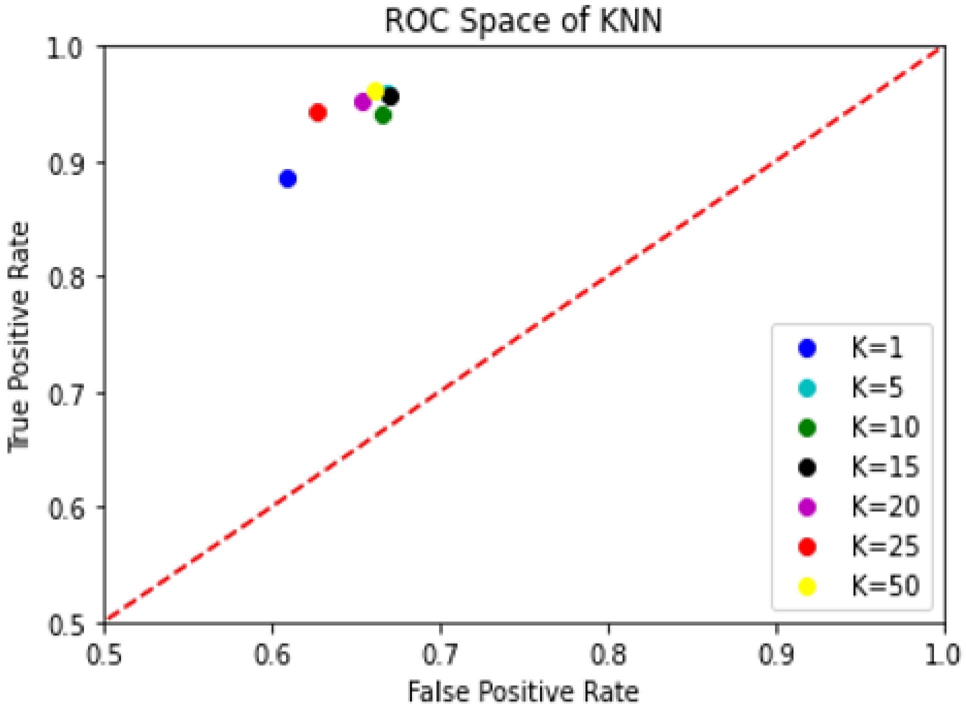


- Classification Model 2 : KNN Classifier



Besides plotting the ROC curve we also tried to have a deeper dive on the parameters and their performances by plotting the ROC space for these classification models. We plotted the TPR VS FPR for both these classification models in the form of a scatter plot .





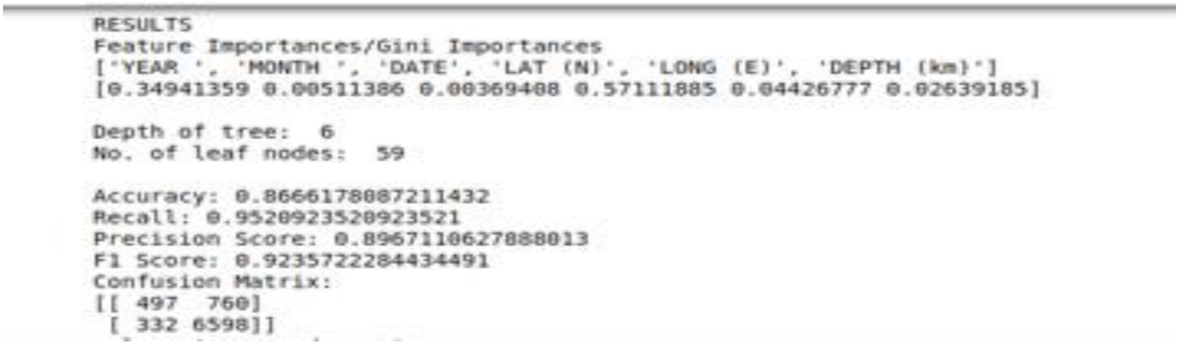
TASK 2 – COMPARISON BETWEEN THE TWO CLASSIFIERS FOR THIS DATA

For classification purpose we did analysis on our data using both the models. So before coming on to a conclusion as in which model to use lets have a brief look at the value of the evaluation parameters for both of these classifiers (the train-test split , threshold value, features used are same as the earlier stated in the beginning of the report)

MODEL 1 : DECISION TREE

In decision tree classifier we used the gini criterion(criterion is the parameter that decides upon how the impurity of a split will be measured), for splitter we used the best parameter which is also the default value and the parameter max depth i.e the depth of the decision tree was kept variable . Now we will perform analysis on the basis of varied values of the pre prune depth parameter .

DEPTH = 6

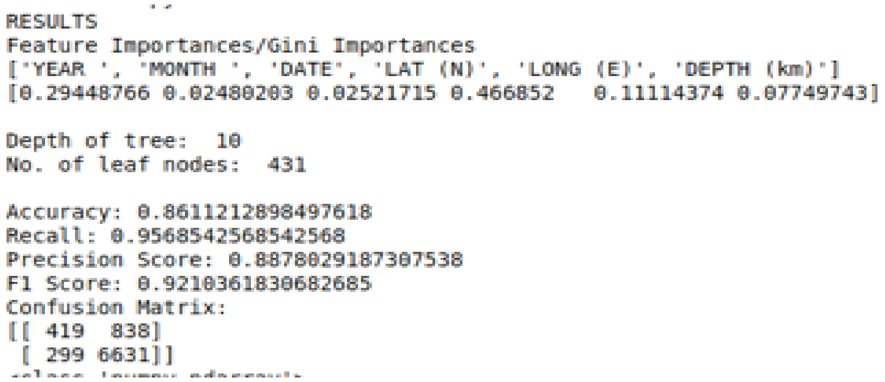


```
RESULTS
Feature Importances/Gini Importances
['YEAR ', 'MONTH ', 'DATE', 'LAT (N)', 'LONG (E)', 'DEPTH (km)']
[0.34941359 0.00511386 0.00369408 0.57111885 0.04426777 0.02639185]

Depth of tree:  6
No. of leaf nodes:  59

Accuracy: 0.8666178087211432
Recall: 0.9520923520923521
Precision Score: 0.8967110627888013
F1 Score: 0.9235722284434491
Confusion Matrix:
[[ 497  760]
 [ 332 6598]]
```

DEPTH = 10

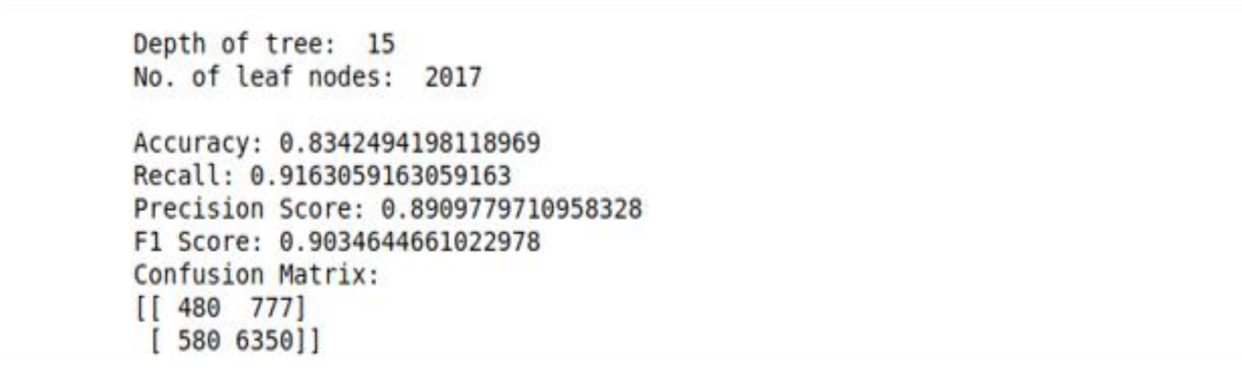


```
RESULTS
Feature Importances/Gini Importances
['YEAR ', 'MONTH ', 'DATE', 'LAT (N)', 'LONG (E)', 'DEPTH (km)']
[0.29448766 0.02480203 0.02521715 0.466852   0.11114374 0.07749743]

Depth of tree:  10
No. of leaf nodes:  431

Accuracy: 0.8611212898497618
Recall: 0.9568542568542568
Precision Score: 0.8878029187307538
F1 Score: 0.9210361830682685
Confusion Matrix:
[[ 419  838]
 [ 299 6631]]
```

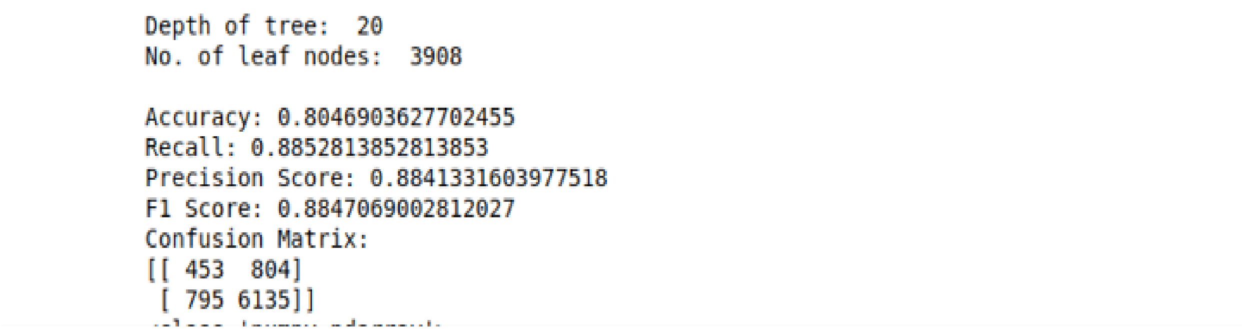
DEPTH = 15



```
Depth of tree:  15
No. of leaf nodes:  2017

Accuracy: 0.8342494198118969
Recall: 0.9163059163059163
Precision Score: 0.8909779710958328
F1 Score: 0.9034644661022978
Confusion Matrix:
[[ 480  777]
 [ 580 6350]]
```

DEPTH = 20



```
Depth of tree:  20
No. of leaf nodes:  3908

Accuracy: 0.8046903627702455
Recall: 0.8852813852813853
Precision Score: 0.8841331603977518
F1 Score: 0.8847069002812027
Confusion Matrix:
[[ 453  804]
 [ 795 6135]]
```

- So what we can see that our decision tree classifier is working quite well for our data and as previously seen in the ROC curve it is fairly evident that on increasing the value of depth of the tree as expected the accuracy on test data is decreasing comparatively . However keeping a reasonable depth we are able to see that we are getting pretty good accuracy , precision and recall scores .
- The major conclusion that we draw from the decision tree classifier from here is that it is giving us an accuracy of about 0.86 (depth :6 to 10) Now let us have a look at how does the kNN classifier performs .

MODEL 2 : KNN Classifier

So now we will evaluate the data for the KNN Classifier . We will change the values of K and see how the value of the evaluation parameters are being affected.
We will test for a few values of K and see the trend of the evaluation parameters .

```
-------
RESULTS  : Value of K =1
-------
Accuracy: 0.8095761573225846
Recall: 0.8854256854256854
Precision Score: 0.8891465005071729
F1 Score: 0.8872821921769937
Confusion Matrix:
[[ 492  765]
[ 794 6136]]
------
RESULTS  : Value of K = 5
-------
Accuracy: 0.8281739342860632
Recall: 0.9414141414141414
Precision Score: 0.8862926232848798
F1 Score: 0.9130221817927368
Confusion Matrix:
[[ 420  837]
[ 406 6524]]

------
RESULTS  : Value of K =10
-------
Accuracy: 0.8352583363869549
Recall: 0.9425685425685426
Precision Score: 0.8824716491323952
F1 Score: 0.9168362692118746
Confusion Matrix:
```

```
[[ 470  787]
[ 398 6532]]

-------
RESULTS : Value of K=15
-------
Accuracy: 0.8490448271650177
Recall: 0.9520923520923521
Precision Score: 0.8892183288409703
F1 Score: 0.9195818815331009
Confusion Matrix:
[[ 435  822]
[ 332 6598]]

-------
RESULTS : Value of K = 25
-------
Accuracy: 0.8535641871259313
Recall: 0.96002886002886
Precision Score: 0.887895369011077
F1 Score: 0.922554253622686
Confusion Matrix:
[[ 417  840]
[ 277 6653]]
```

- So from the KNN classifier we could see that a considerable increase in the value of k gave us an accuracy near to that of the decision tree.
- Other evaluation parameters also recite the same story that decision tree classifier is better

CONCLUSION :

The decision tree classifier is a better classifier than Knn for the given data .
The reasons for this conclusion are :

1. Decision Tree uses gini features to draw conclusions. From the gini features we are able to draw valuable insights about the features and hence they show us how the various features are contributing to the classification problem. Decision Tree gives more priority to the features which are useful towards classification. On the other hand KNN depends on the distance metrics of features which are present and does not have the

property to distinguish on the basis of the features which have higher importance in the classification task . This shortcoming of KNN is overcome by Decision Trees.

2. Now if we try drawing conclusions from the results that are visible to us that is our evaluation metrics : For less depth of the decision tree we are getting a better value of the evaluation parameters as compared to knn classifier where we need to have a significantly high value of K to have the evaluation parameters value near to that of the decision tree classifier.

3. Considering the most important aspect the features on the basis of which we are classifying the data . Now if we see which model is best suited for these features , Decision tree supports automatic feature interaction I.e. different orders of **feature** combinations of input **features** can be modeled. , whereas KNN cant as it is based on neighborhood distance and thus is a good model for our underlying classification task.

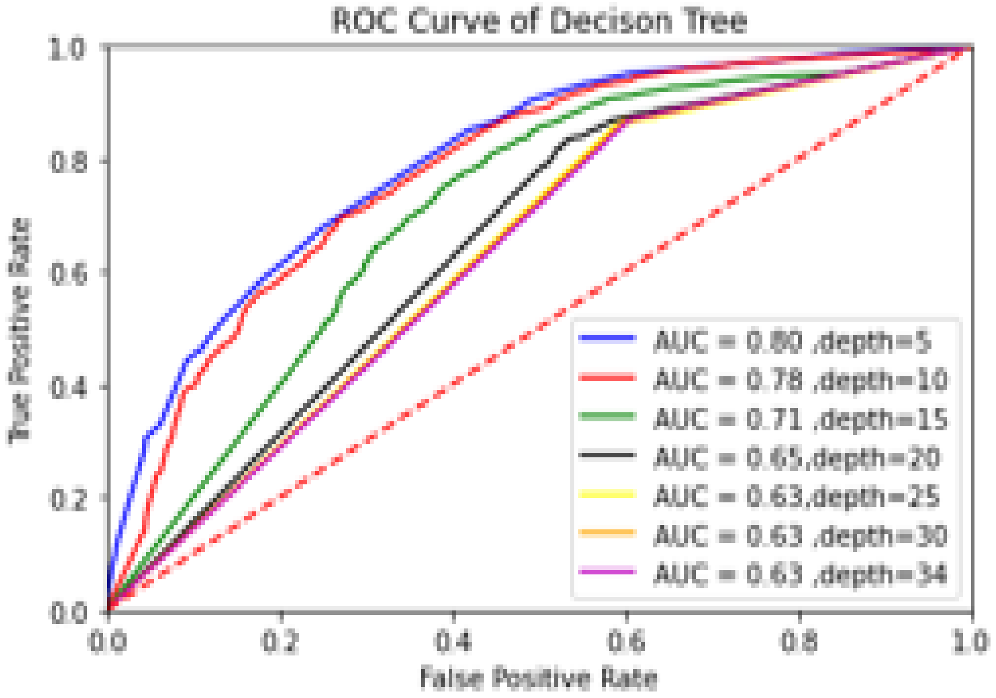
4. Considering the time factor to KNN has a higher cost involved as compared to Decision Trees.

Hence to conclude for the given data "Decision Tree " is the better classifier .

TASK 3 : FINDING THE BEST POSSIBLE VALUE OF THE MENTIONED PARAMETERS FROM THE ROC CLASSIFIER

The AUC score of the ROC curve determines how much better is a particular classifier for the given parameters. The higher the value of the AUC score the better the classifier.So we will use the AUC score to derive the conclusions.

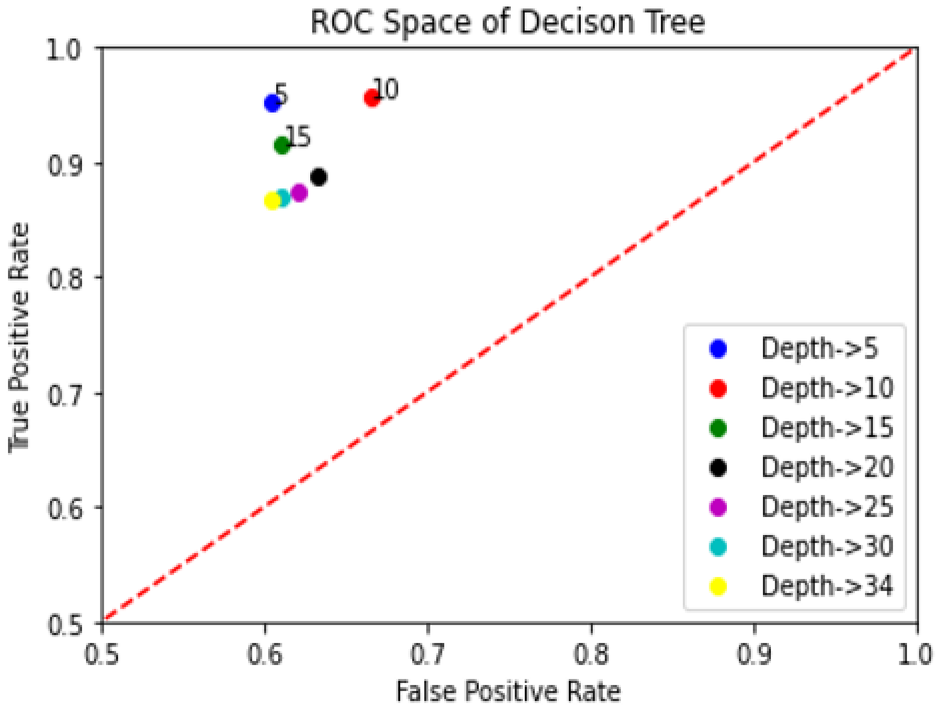
- Let us start with the decision tree classifier :



So looking at the ROC curve it is clearly evident that The decision tree performs considerably well for **depth =5** with an AUC score of 0.80 .

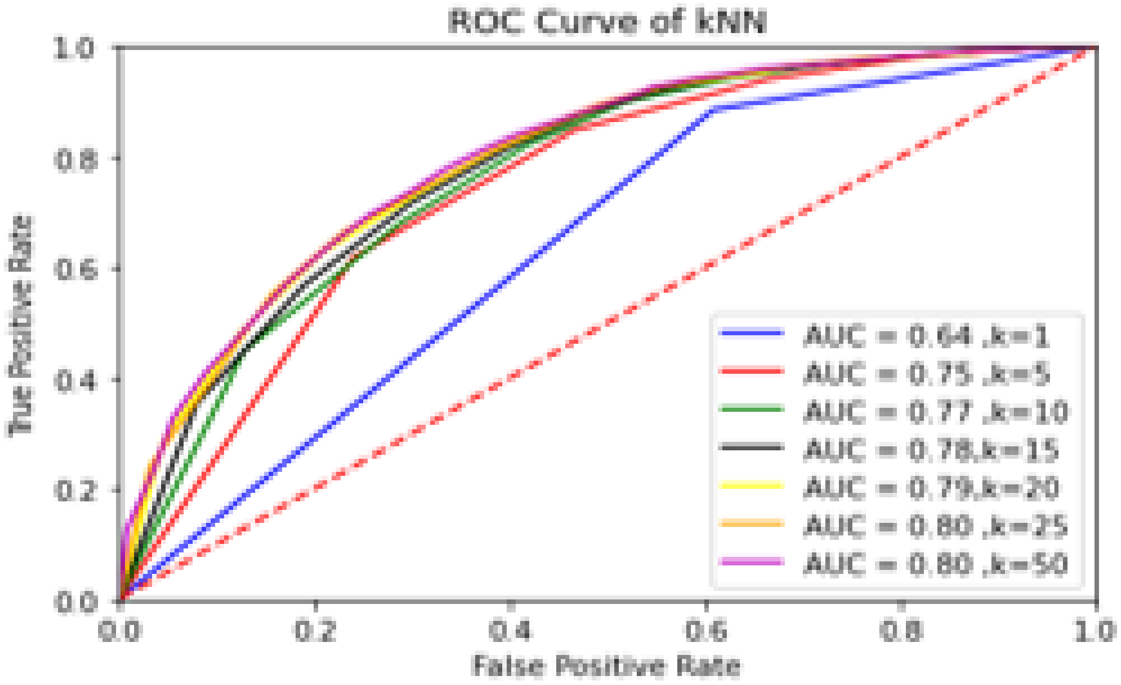
Further conclusion that we draw is that as the depth of the tree increases the auc score decreases due to overfitting of the model on the train data.

Let us also have a look at the ROC space plot to cross verify our conclusion.



Coming to the ROC space of the Decision Tree where we have plotted the values in terms of scatter plot , it is clearly evident and visible to us that the best value for the pre prune depth is : 5 . Thus both our ROC curve and ROC Space suggest that the best pre prune depth is 5 for this classification problem.

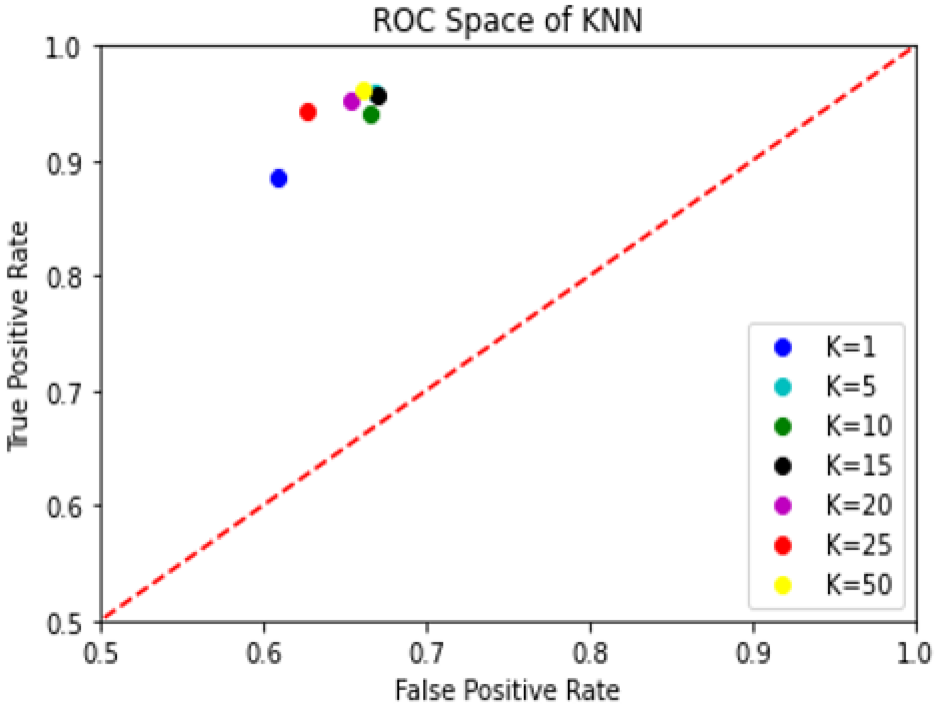
Moving on to the KNN classifier we have plotted the ROC curve for a wide range of K



As we keep on increasing the value of K the corresponding ROC Curve starts getting better.

So as we can see from the above ROC curve it can be concluded that the best parameter for KNN is **K= 25** (Though the value of auc score is same for both k=25 and k=50 we choose the lower value of k)

Now let us plot the ROC Space for KNN classifier and draw conclusions from it.



So the ROC space which is the scatter plot between TPR and FPR also adheres to the conclusion drawn from the ROC curve that best parameter i.e the most feasible value of K is 25.

CONCLUSION :

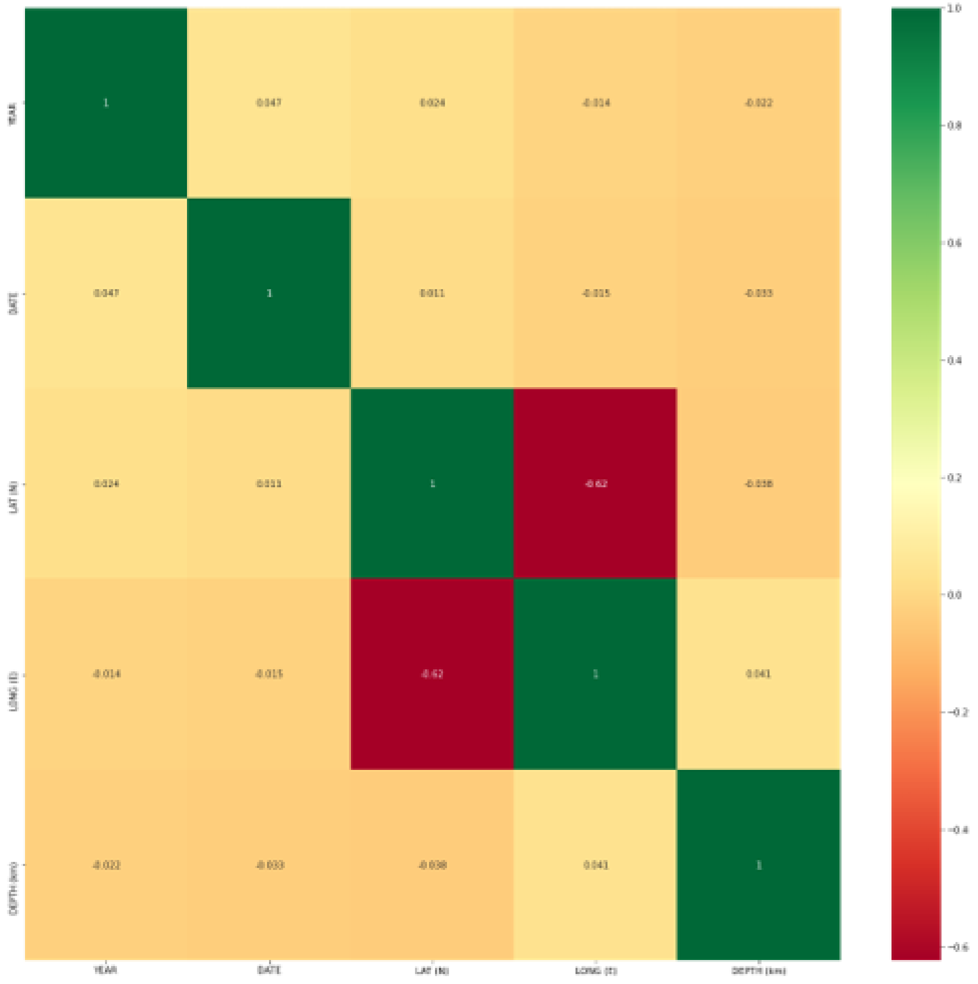
Value of Pre prune depth for decision tree : 5

Value of K for KNN Classifier : 25

TASK 4 : CHOOSING A SUBSET OF ONLY TWO FEATURES TO PREDICT THE EARTHQUAKE

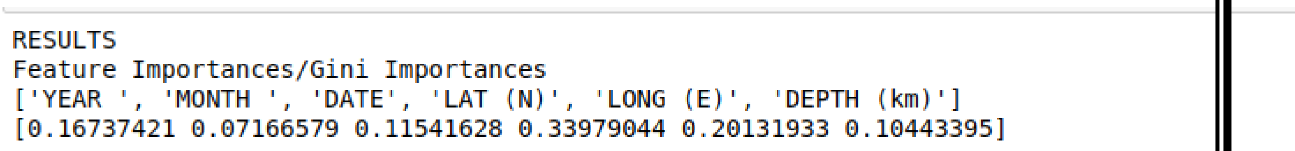
To find the best features out of the given features many methods are there which help us to decide on the best features. As we are using decision classifier we can easily use the gini features to decide on the importance of the various features which can be used in this case.

If we try to find the covariances between the features we find maximum correlation between Latitude and Longitude for the given data .



We ran the model for different depths and drew conclusions from the particular gini features .

So lets have a view of the value various gini features available to us and their particular values(i.e. importance) :



```
RESULTS
Feature Importances/Gini Importances
['YEAR ', 'MONTH ', 'DATE', 'LAT (N)', 'LONG (E)', 'DEPTH (km)']
[0.16737421 0.07166579 0.11541628 0.33979044 0.20131933 0.10443395]
```

So from here we draw the conclusion that the 2 most important features for this data are : ‘LAT (N)’ and ‘LONG (E)’.

The third parameter that is also important is ‘Year’ So lets try to test our models using 2 of these features at a time and lets see what is the optimal result for us !

Let us now try evaluating our model for these features:

Features Used : Latitude and Longitude

Accuracy : 0.85

Precision: 0.95

Recall:0.89

F1 – Score:0.92

Features Used : Year and Longitude

Accuracy :0.835

Precision:0.924

Recall:0.88

F1 – Score:0.91

Features Used : Latitude and Year

Accuracy :0.848

Precision:0.95

Recall:0.88

F1 – Score:0.87

So the thing that is clearly evident is that the model performs best when all the features which we had obtained after cleaning are used i.e Year, Date, Month , Latitude , Longitude and depth and no two features can beat that performance.

But Latitude and Longitude when considered give us near about good values of evaluation parameters and better than any other 2 features taken as a subset . So we will be considering the two features as : **Latitude and Longitude**

TASK 5 : Considering test results of the best model after analysis and using feature processing to further improve on the result.

As we can conclude from the above analysis the best model is Decision Tree for this classification problem and it was also quite evident from the ROC curves .

So lets move to the Feature Processing Part .

Feature Processing means performing operations on the various features available to us to extract valuable insights and improvise the performance of our model.

- Feature Selection : Selected the most useful and meaningful features for our classification model
- Feature Processing :
  Processing the available features in a more meaningful manner :
  1. For the values of depth corresponding to which we had NAN or empty strings or missing values we filled it with the mean of the depth of all the values .
  2. As the source location is indirectly covered by the latitude and longitude we dropped it out.
  3. For those rows where the value of year , date and month were missing we processed it by filling 0 there as still we can make use of the other features for predictions instead of discarding the entire row .
  4. The entire data was converted to float to make it more suitable for processing .

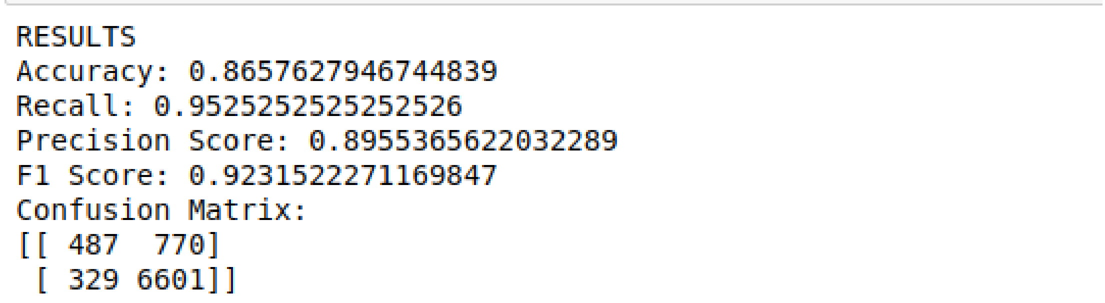
This is how we applied feature processing on our data and made it more meaningful and useful for prediction.

Features Used in Our Best Performing Model :

Model : Decision Tree

Features Used : [Latitude, Longitude, Depth, Year, Month, Date]

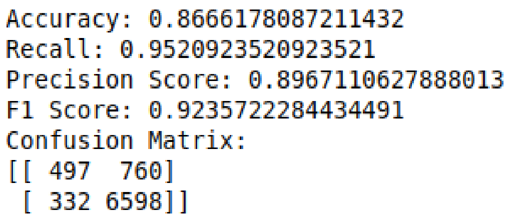
Results :



```
RESULTS
Accuracy: 0.8657627946744839
Recall: 0.9525252525252526
Precision Score: 0.8955365622032289
F1 Score: 0.9231522271169847
Confusion Matrix:
[[ 487  770]
 [ 329 6601]]
```

Model: Decision Tree

Features Used : [Longitude, Latitude, Depth]



```
Accuracy: 0.8666178087211432
Recall: 0.9520923520923521
Precision Score: 0.8967110627888013
F1 Score: 0.9235722284434491
Confusion Matrix:
[[ 497  760]
 [ 332 6598]]
```

Conclusion of Task 5: We could further improve upon our best model by considering only three features and applying the feature processing techniques discussed above . Though it is just a slight improvement that we saw in our data , if we want we can still go with all the features of our best model .

Conclusion of the Project

In this project we began with raw noisy data given to us which we cleaned further to extract the insights from it and use it for our classification problem in hand . We learned how to analyse data , how to clean it , how to process it and then perform analysis and prediction on it.

We used different metrics to draw conclusions from the data and analyse it efficiently . The results of the analysis were then applied in feature processing and model determination for the classification problem under discussion . We used various evaluation metrics and the ROC curve to conclude on the best model.

-----------------------------------Thank You--