Assignment-1: Classification Project

Team:

Gayatri Purigilla - 20171309

Keshavan Seshadri - 20171051

Table of Contents

Problem

In this project, we use the KNN algorithm and Decision trees to predict the occurrence of an earthquake and its magnitude (Mw). An earthquake is said to occur if the magnitude of the earthquake in Mw is greater than T. We have chosen the value of T to be 4.5

Features Used: Latitude(N), Longitude(E), Depth(km), Intensity

We used a Training and Testing split of 80:20 on the cleaned data. Using the aforementioned features we try to predict the occurrence of the earthquake. Then we try to improve the model by choosing the parameters that give the highest ROC score. We also evaluate each model based on Accuracy, Precision, Recall, and F1-Scores. Then we try to process the features in the hope of improving the model.

Data Cleaning

Before we got to building the model, we had cleaned the data in the following ways.

We removed the unnecessary columns like serial number, time in IST, given that this can be derived from UTC, and the types of ways to represent the magnitude of an earthquake (except Mw) as again, they can all be derived from Mw.

We had to convert all the latitude and longitude values to 'float' data type and convert all the values such that they are uniform, I.e. convert all latitude values to N and all longitude values to E. We do this by finding all the rows in the latitude column with a value in S and convert it into the required value by first, extracting the exact numerical value after removing any trailing spaces and text (S) and then applying the formula,

N = 180 – S.

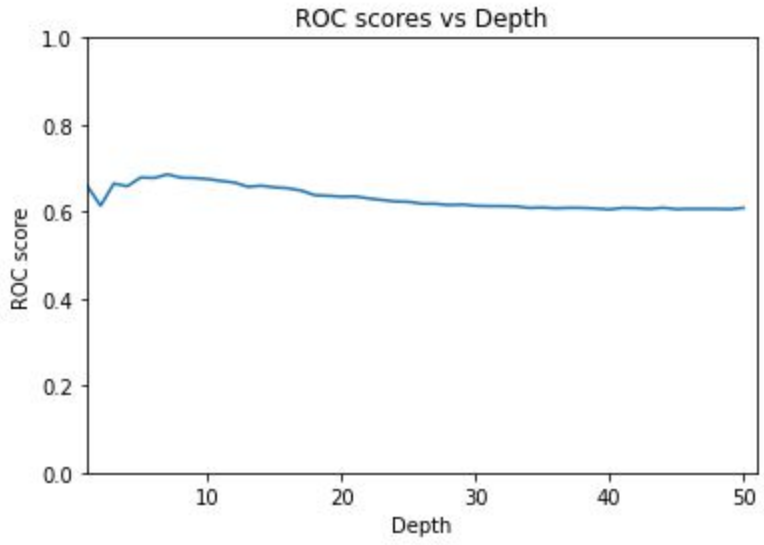
For converting longitudes to all in East direction, we used the formula E = 360 - W

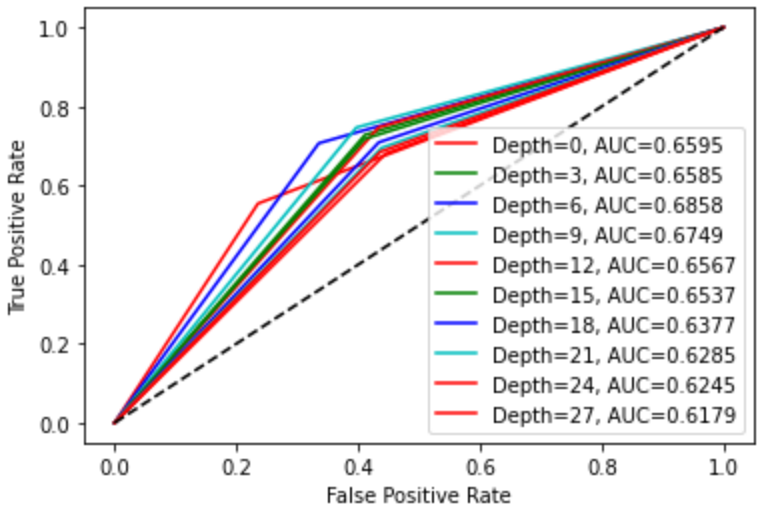
The data also consisted of trailing spaces, symbols like /, ?, 0, etc. and strings like N, E, W, S to indicate the direction for latitudes and longitudes. We parsed through the data and wrote a python script to remove all unnecessary characters.

The cells with a 'NULL' value are filled with the MEAN of all the values in their respective columns.

ROC Plots and Analysis

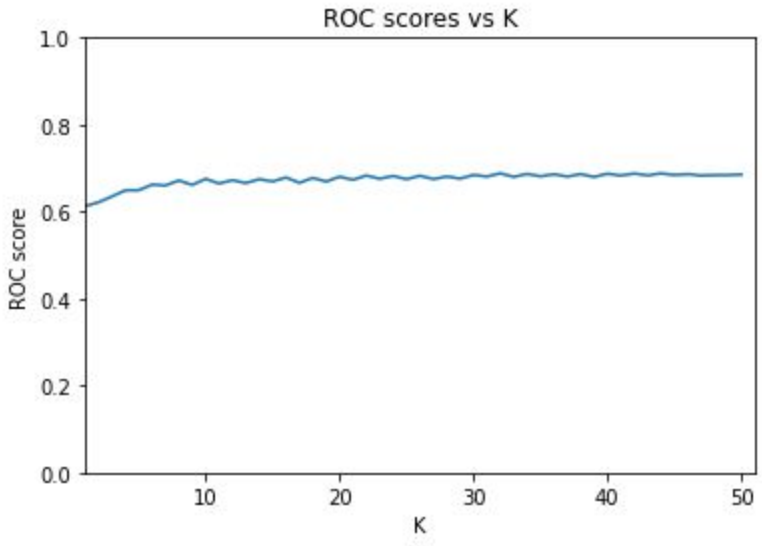
Decision Trees

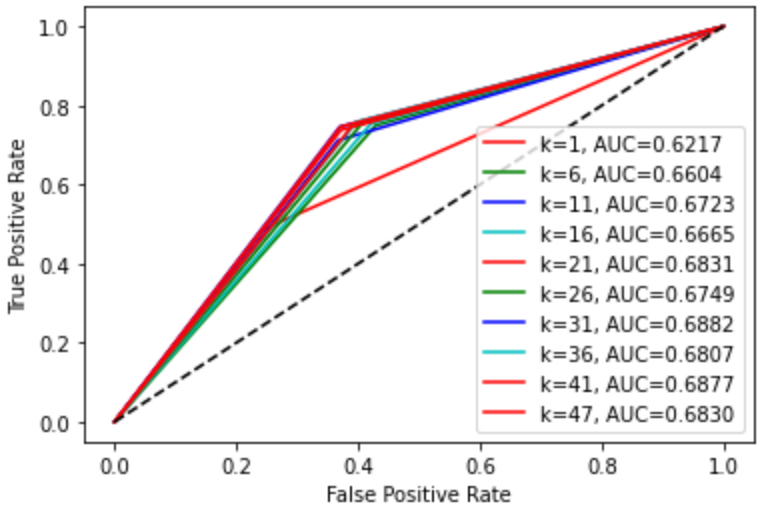




The best Decision Tree model is obtained with a ROC score of 0.6858397919538641 for a max pre-prune depth of 6.

KNN Classifier





The highest ROC score of 0.6881889047573484 for a K value of 31 when the KNN classifier is used.

Analysis

Classifier	Overall Accuracy	Precision	Recall	F1-Score
Decision Trees (Depth = 6)	0.68	0.69	0.68	0.68
KNN (K=31)	0.70	0.69	0.70	0.69

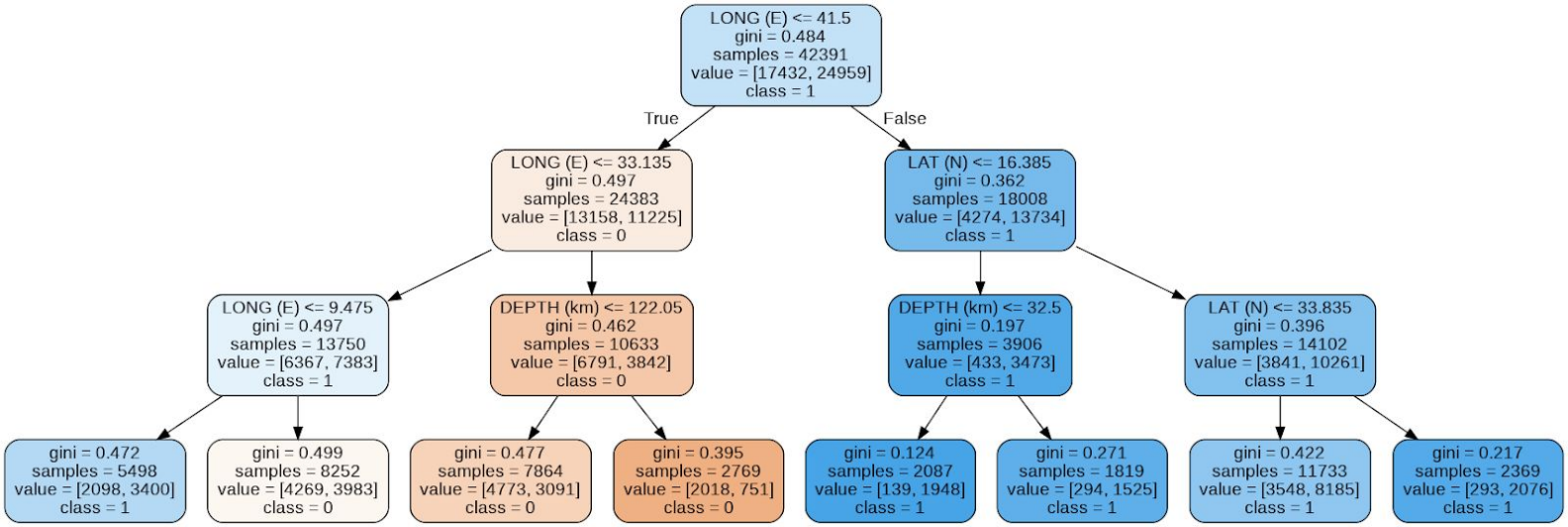
- Though decision trees can be visualized more easily and are faster in training and forecasting, in this case, KNN gives us slightly better results.
- This is because decision trees work better than KNN only in the presence of categorical variables that are not present in the features that we have considered.
- Moreover, it is known that KNN outperforms Decision Trees when it comes to rare occurrences.

A subset of features

From the decision tree below, it can be seen that latitude, longitude, and depth are the three most important features. since these three features have the largest entropy.

Upon finding the correlation between different features in the dataset using the Pearson Correlation, we observe the largest negative correlation between LAT(N) and LONG(E). Therefore we consider only one of the features. The selection between latitude and longitude needs to be done as they are strongly dependent on each other, and using any one of the two is sufficient. We choose latitude over longitude since it has a lesser correlation with other values and is the most important feature according to the decision tree.

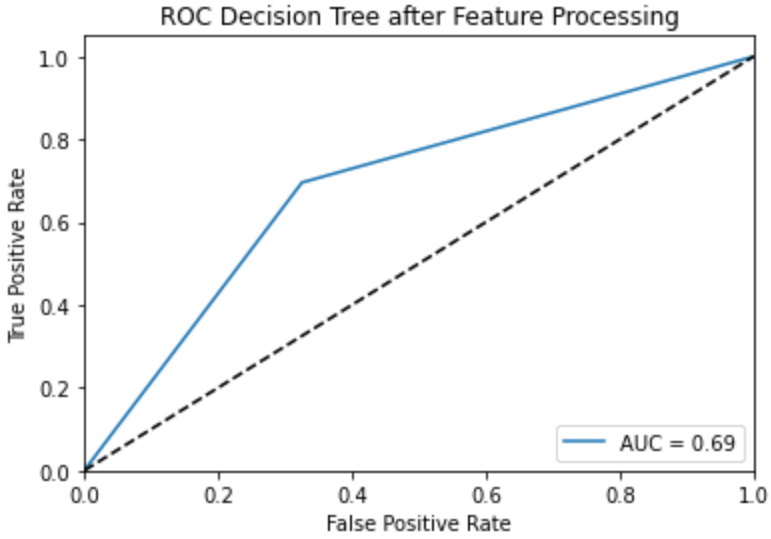
Therefore to consider a subset of two features, we use *Latitude* and *Depth.*

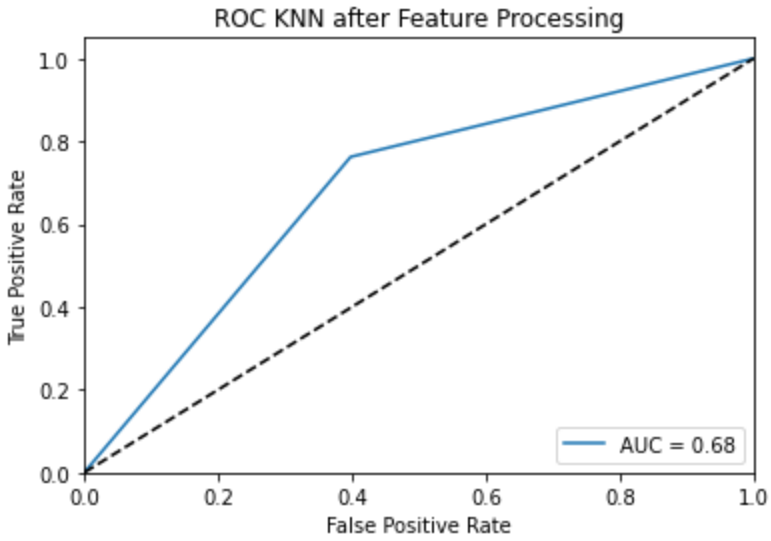


Feature Processing

- Originally, we had used latitude, longitude, depth, and intensity as features to build our model. To improvise the test results, we had included location as well as a reference as additional features.
- The location feature acts almost like a combination of the latitude and longitude features.
- LATITUDE + LONGITUDE = LOCATION
- The reference feature is useful as it indicates the credibility of the source.
- These are both categorical variables in string format and hence, needed preprocessing before usage.
- We use *One-Hot Encoding* to preprocess the strings.
- The accuracy given by the decision tree classifier increased from 0.68 to 0.69 after feature processing due to the presence of categorical variables
- However, there is no significant change in the accuracy after feature processing in the KNN Classifier. The similarity in the results before and after feature processing can be attributed to the fact that the latitude and longitude feature already capture most of the location information.
- In terms of overall improvement, we can see that metrics such as Precision and F1-Score have improved in both the models after feature processing.

We obtained a ROC score of 0.6851521423010999 for Decision Tree on max_depth=6 and a ROC score of 0.6821557277657485 for the KNN Classifier (k=31)





Classifier (after feature processing)	Overall Accuracy	Precision	Recall	F1-Score
Decision Trees	0.69	0.69	0.69	0.69
KNN	0.70	0.70	0.70	0.70