CLASSIFICATION PROJECT: REQUIRED ANALYSIS AND ERRORS

Expectations :

Suitable data cleaning performed (though not considered for grading) .
A suitable threshold between 4 and 5 where there is a likely even distribution of positive and negative data points . A threshold of 4.3 - 4.7 preferred .
Q1 . Discrete ROCs . Plots connecting the single point to origin or (1.0,1.0) also accepted .
Q2 . Explanation backed by area under curve reasoning , other metric analysis (precision / recall).
Q3. Parameters for decision tree (depth) and knn (k) expected . Explanation backed by area under curve reasoning , other metric analysis (precision / recall).
Q4. A good number of analysis tests possible . (Decision tree visualisation , Random Forests, chi square tests) .
Q5. Certain modification on the columns, using modified functions, square root or logarithm of a linear combination of values in columns, with different weights to each column as required . Min max scaling . Introduction of a new column depicting the period of the earthquake .
Note : Apart from the analysis and depiction required in each and not just the proposal, comparison required along with the previous scheme of analysis on the basis of different metrics (accuracy, precision/recall/confusion matrix) .

Errors :

Picking thresholds on boundaries like 4 and 5 seem to be biased since the number of positives >> number of negatives on boundary 4 and the number of negatives >> number of positives on boundary 5. Picking so many have found almost perfect accuracy everytime and hence have preferred to avoid feature processing since no improvement is possible .
Q1. Continuous ROCs are considered for penalty(except line graph), some have bound few discrete curves and some continuous, considered for partial grading. [ROC should not be continuous because the parameters(K and depth) that we are looking at are discrete in nature.]
Q2. i) Simply verbose reasoning have met with penalisation
 ii) Answers based on continuous ROCs and no other metric analysis have been penalised .
Q3. 1st 2 points same as Q2 above
 iii) Some have provided the no. of leaves as a parameter and not the depth, some penalty attached .
Q4. i) Verbose reasoning have been penalised
 ii) Answers based on just experiments with some subset of features have been awarded a slight penalty .
Q5. i) Pure proposals for feature processing and no analysis have faced a penalty .
 ii) Some have just listed the features (the ones that were already present in the database) they took in their analysis as part of feature processing. No marks for this.
 iii) Some have just shown various combinations on subsets of features (already present) as feature processing. A deduction of marks for this too .
 iv) Some having not tried different metrics have met a small cut .