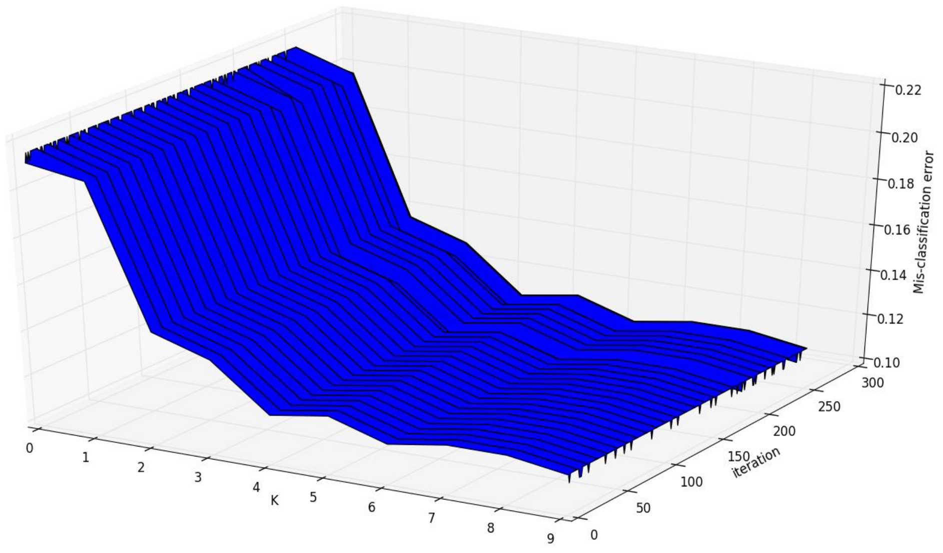
a. Write a k-NN algorithm, from scratch. Use this to pick a value of K. Based on your judgment and choice of k, delete or reclassify any points from the training set that you believe are noise points or are misclassified.
Please find attached HW09_Bharde_Manasi_ kNN_DecTree.py

b. You have some design decisions to make here:
What distance metric will you use?
I used Euclidean distance as distance metric.

Do you run k-NN on the data and delete or change points as you go?
Or, do you mark the points for deletion, and then delete them after marking all points for deletion?
Initially data is cleaned to remove points which do not have at least 5 points within 1 astronomical distance of it. Then reclassified the training data points as per kNN classes. There were some records in the data like (2.72, 4.05, 2.23, 0) for which decision tree was observed to be sensitive. After reclassification the point correctly as per kNN, the issue was resolved.

What value of k did you use?
5

How do you select a value for k?
What iterative loop will you use for this: design/code/test.
I ran hold one out algorithm on data for k as 1 to 10 and plotted the K vs. N vs Misclassification rate graph. The graph revealed that for k=5, misclassification rate is least. It can also be programmed to pass value of k that gives least cumulative misclassification rate to data cleaning algorithm.



Anywhere from 5% to 15% of these data points are classified incorrectly.
53 records were reclassified which is 12% of data points.

d. Run your decision tree code, as from a previous homework, to build a classifier on the training dataset to produce a decision tree. The mentoring program should write the decision tree program. Hint: The final decision tree should have only about 5 to 15 “if statements”.
Please find attached HW_09_Bharde_Manasi_Classifier.py. Decision tree created has 9 if statements. Weighted Gini index was chosen as measure of error. The decision tree program does not look ahead as asper studies decision trees that look ahead give less accuracy. Decision tree created uses all attributes however attribute 2 is of lesser importance. Decision tree created for cleaned data in Weka looks similar.

e. Run your decision tree classifier on all the data points in the training set, and have it generate an output file that gives the classifications. You will be graded based on these classifications.
Please find attached HW09_Bharde_Manasi_Classes.csv