Question - 3

A) Brief Implementation Details:

Here for implementation of decision tree the important part is to select the attribute on which we need to split the data. So every node of the tree does the following things:

-> Calculate the best value on which to split an attribute to obtain maximum gain of a attribute

-> Once the gains of all the attributes are calculated selecting the attribute with maximum gain and split the data set on that attribute

-> Now the splitted data set is sent to its left and right children and the same process is repeated

B) Accuracy via 10 fold cross validation(Function = “Gain”): 81.84% :

Here in this part the data set was divided into ten parts and for each epoch one part was left behind for test and 9 parts for train and the accuracy is the final average of this 10 epochs

C) Optimization:

So for optimization on the original implementation the gain function was replaced with gini index and the result are:

Accuracy via 10 fold cross validation(Function = “Gini”): 82.36%

So the original implementation of binary univariate split has an accuracy of 81.84% but when improved the function to gini the accuracy increased to 82.36%. The accuracy increased with gini because gini index shows the probability of wrong classification which is better then entropy.