Assignment Report

Question 2.(c)

In the methods used i.e, Scikit-learn kNN and our implementation of kNN the accuracies are almost same but the time taken for our implementation is around 300 times more than the Scikit-learn kNN

We can improve the implementation as follows:

Here we are using an arbitrary value of k, it can be improved by taking an optimized value of k which gives maximum accurate predictions.

Here. we calculated the distance between each test instance and every single data point in our training set which is inefficient, and there exist alterations to kNN which subdivide the search space in order to minimize the number of pairwise calculations.

One more thing can be done that is by weighting the importance of specific neighbours based on their distance from the test case which allows closer neighbours to have more impact on the class voting process.

Handling of Missing Data

In the KNN dataset :

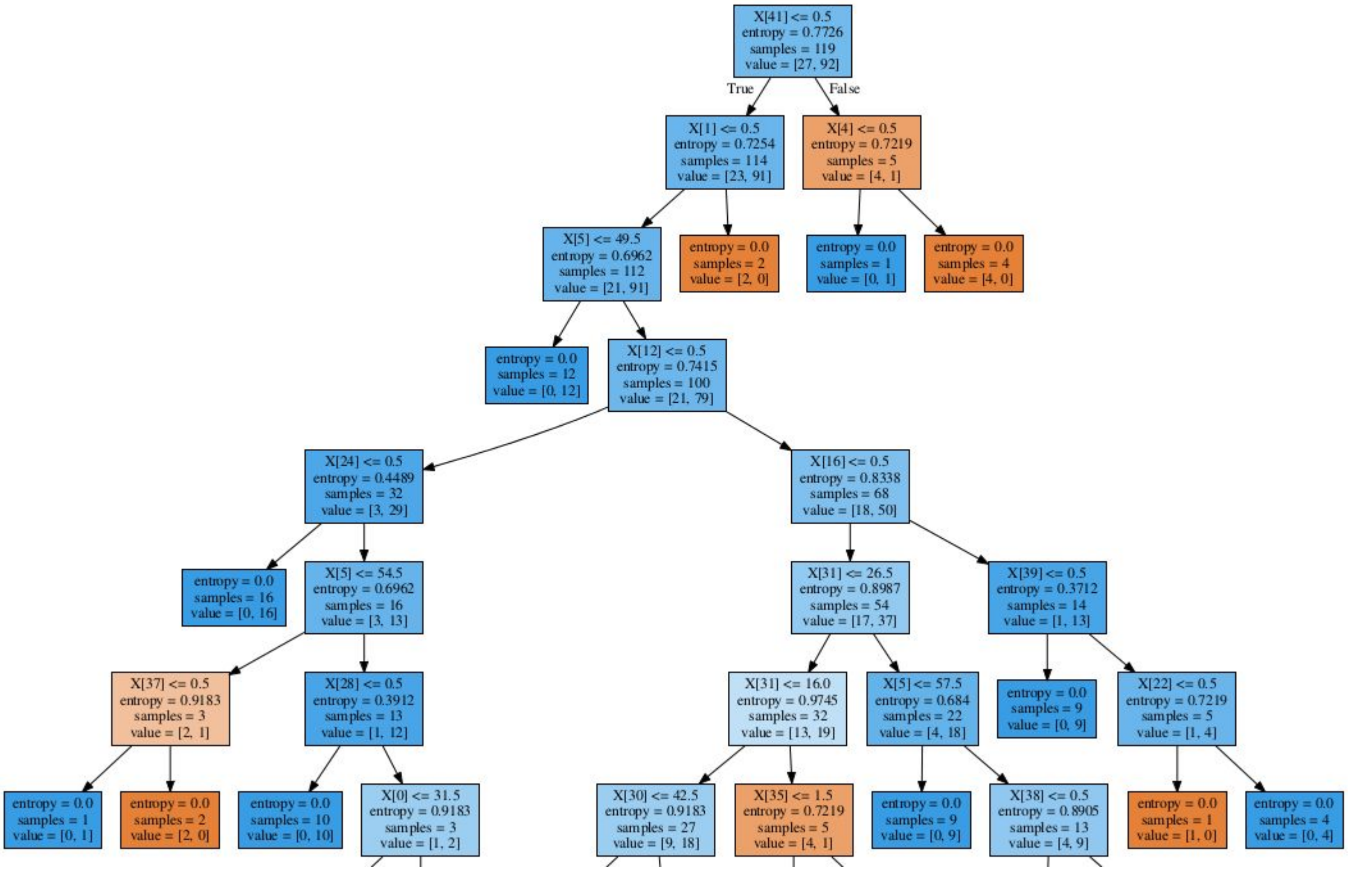
As the dataset is large the missing values in the form of question marks were removed. Other ways would have been to replace them with avg values.
It was found that accuracies didn’t vary much.

In the Decision tree Dataset :

Here we had a limited data, so the missing values were replaced with avg values of the respective columns for the integer types. There were specific columns of object type with few NaN values. For such types were deleted from the dataset.

`

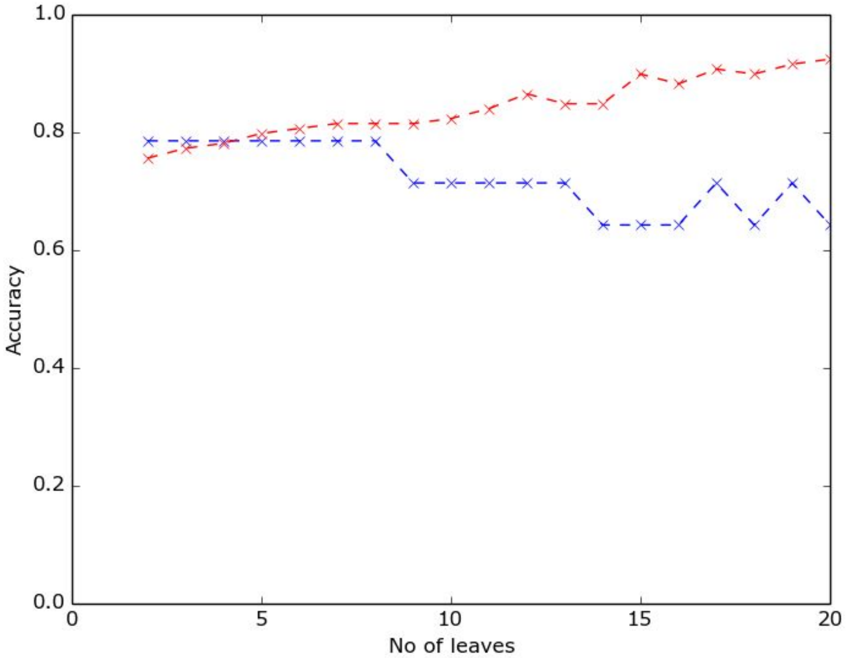
Question 3a)



Here we have handled the data using vectorization.
The datatree went 12 levels down.
The attributes selected are given blue colour.
The attributes with high entropy are reduced further and the lesser ones are resolved as leaves.
The attributes selected at each can be seen from the above tree.
First the attribute first selected is X[1] over X[5] as it has more Information gain than the other.
IG = Entropy total - Entropy local
Similarly this can be done for all nodes.

Question 3.(b)

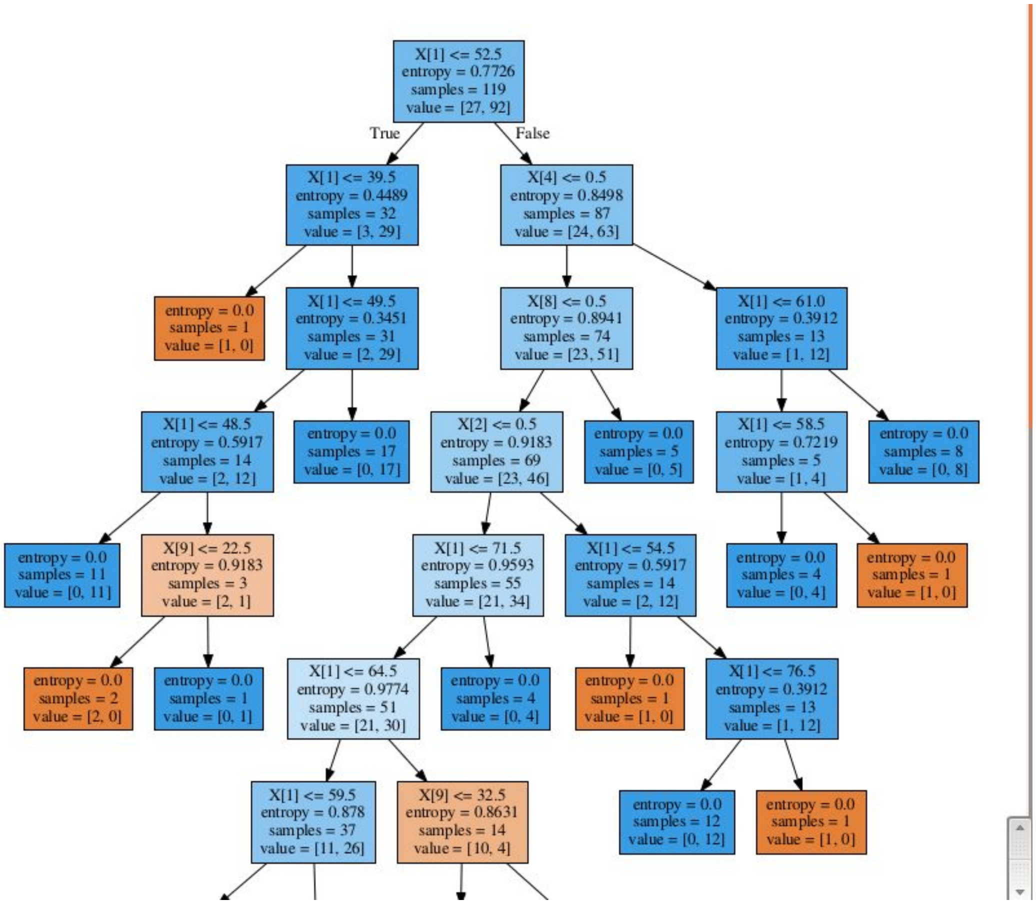
As the number of leaves in the decision tree grows, the accuracy of the training set increases and the performance of the test set decreases. This shows that overfitting is evident.



Question 3c)

The accuracy increases after pruning.

At each node in a tree it is possible to see the number of instances that are misclassified on a testing set by having the errors upwards from leaf nodes. This can be compared to the error-rate if the node was replaced by the most common class resulting from that node. If the difference is a reduction in error, then the subtree at the node can be considered for pruning. This calculation is performed for all nodes in the tree and whichever one has the highest reduced-error rate is pruned. The procedure is then recursed over the pruned tree until there is no possible reduction in error rate at any node.


X[1] <= 52.5
entropy = 0.7726
samples = 119
value = [27, 92]
True
False
X[1] <= 39.5
entropy = 0.4489
samples = 32
value = [3, 29]
X[4] <= 0.5
entropy = 0.8498
samples = 87
value = [24, 63]
entropy = 0.0
samples = 1
value = [1, 0]
X[1] <= 49.5
entropy = 0.3451
samples = 31
value = [2, 29]
X[8] <= 0.5
entropy = 0.8941
samples = 74
value = [23, 51]
X[1] <= 61.0
entropy = 0.3912
samples = 13
value = [1, 12]
X[1] <= 48.5
entropy = 0.5917
samples = 14
value = [2, 12]
entropy = 0.0
samples = 17
value = [0, 17]
X[2] <= 0.5
entropy = 0.9183
samples = 69
value = [23, 46]
entropy = 0.0
samples = 5
value = [0, 5]
X[1] <= 58.5
entropy = 0.7219
samples = 5
value = [1, 4]
entropy = 0.0
samples = 8
value = [0, 8]
entropy = 0.0
samples = 11
value = [0, 11]
X[9] <= 22.5
entropy = 0.9183
samples = 3
value = [2, 1]
X[1] <= 71.5
entropy = 0.9593
samples = 55
value = [21, 34]
X[1] <= 54.5
entropy = 0.5917
samples = 14
value = [2, 12]
entropy = 0.0
samples = 4
value = [0, 4]
entropy = 0.0
samples = 1
value = [1, 0]
entropy = 0.0
samples = 2
value = [2, 0]
entropy = 0.0
samples = 1
value = [0, 1]
X[1] <= 64.5
entropy = 0.9774
samples = 51
value = [21, 30]
entropy = 0.0
samples = 4
value = [0, 4]
entropy = 0.0
samples = 1
value = [1, 0]
X[1] <= 76.5
entropy = 0.3912
samples = 13
value = [1, 12]
X[1] <= 59.5
entropy = 0.878
samples = 37
value = [11, 26]
X[9] <= 32.5
entropy = 0.8631
samples = 14
value = [10, 4]
entropy = 0.0
samples = 12
value = [0, 12]
entropy = 0.0
samples = 1
value = [1, 0]