Part 2)

1)

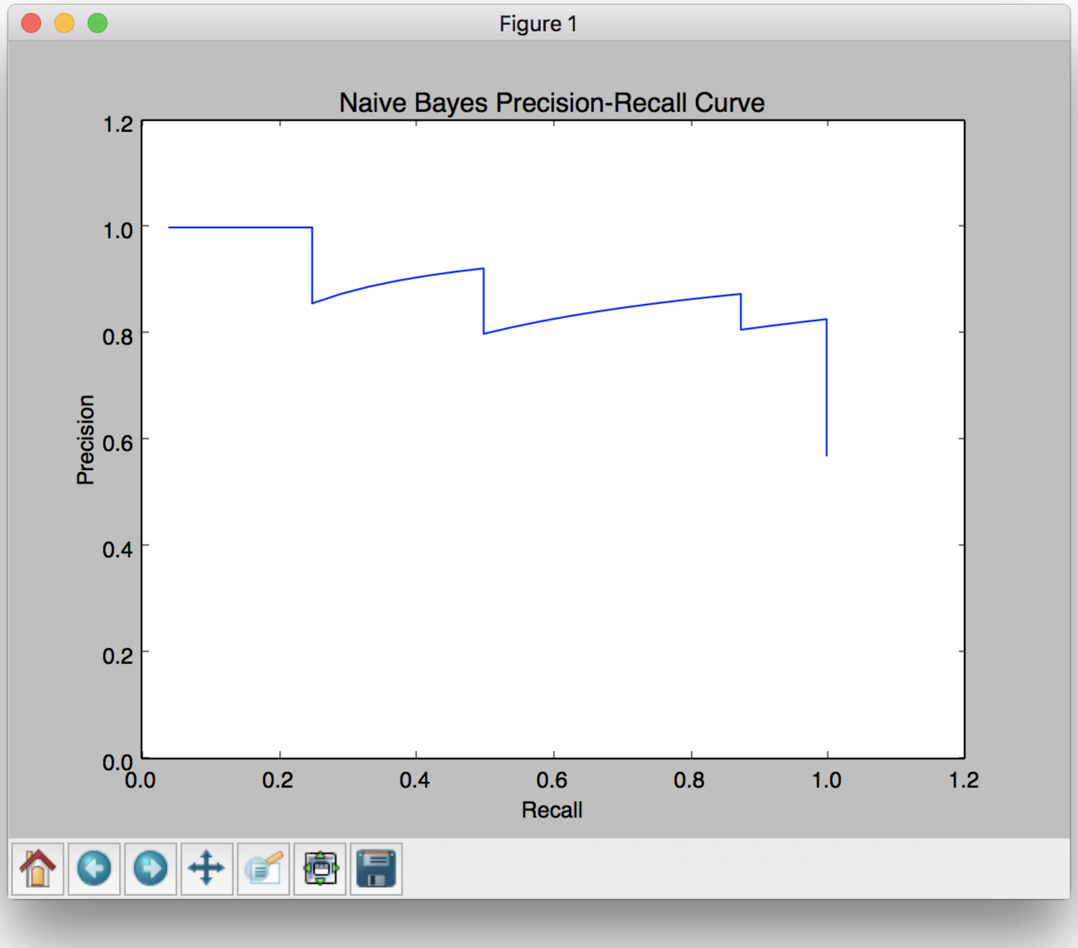


Figure 1
Naive Bayes Precision-Recall Curve
Precision
Recall
1.2
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
1.2

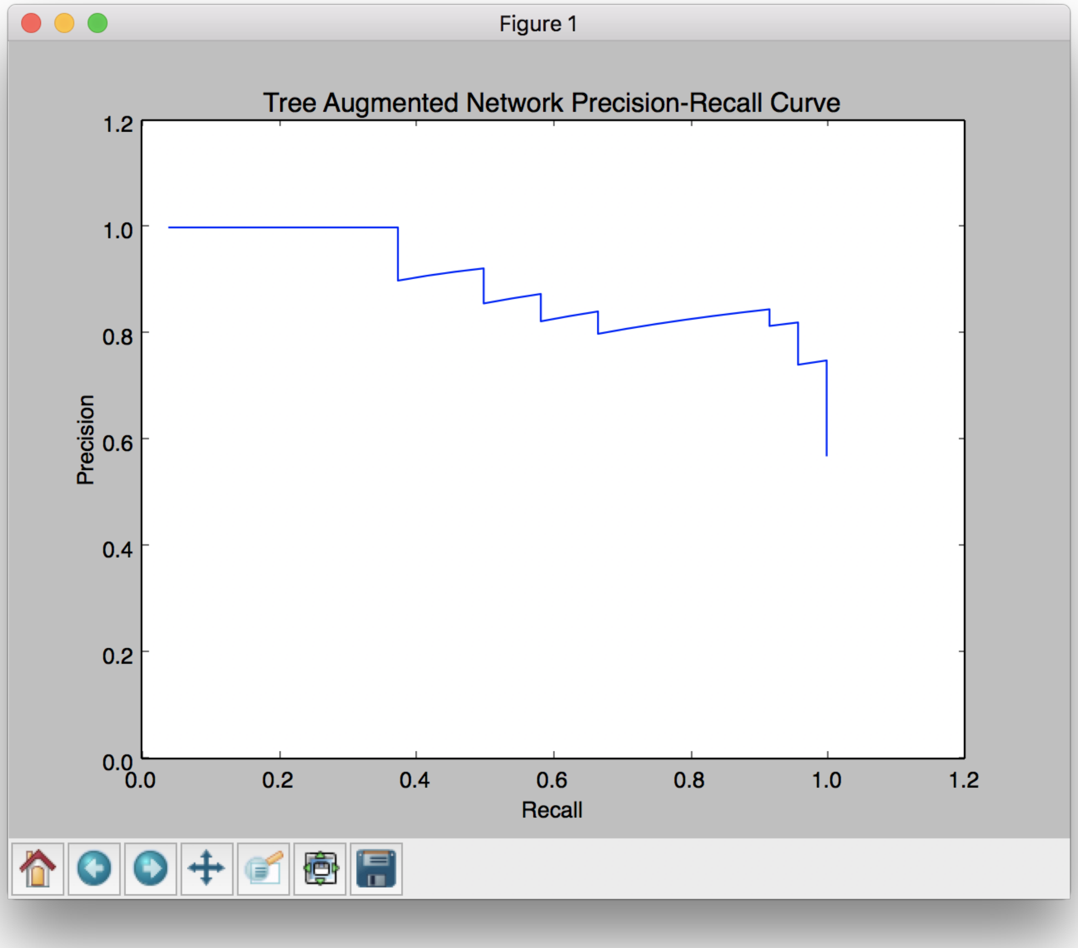
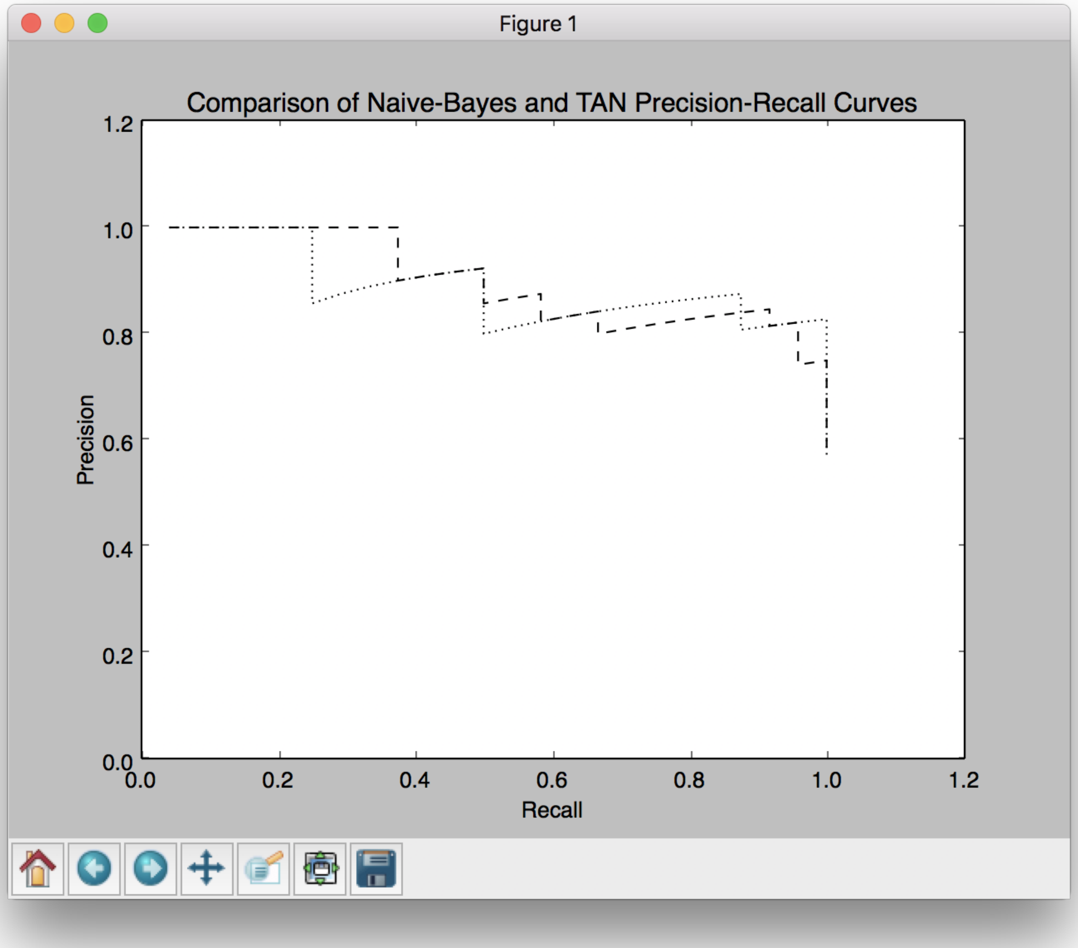


Figure 1
Tree Augmented Network Precision-Recall Curve
Precision
Recall
0.0
0.2
0.4
0.6
0.8
1.0
1.2
0.0
0.2
0.4
0.6
0.8
1.0
1.2



In the last image, dotted lines is the Precision-Recall curve for Naive Bayes classifier and dashed lines is the Precision-Recall curve for TAN classifier.

The **TAN classifier seems to have a higher predictive power than Naive Bayes classifier** for the given data set for the following reasons.

Features of the PR-curves that led to the conclusion:

1) We can see that the area under the curve (AUC) for the TAN classifier is slightly higher than the AUC for the PR curve. The ideal PR curve has AUC of 1 square unit and a classifier with a higher AUC has a higher predictive power. Because of this feature, TAN classifier has a slightly higher predictive power than of Naive Bayes.
2) Since the two PR curves cross over each other quite frequently, it is not straightforward as to which has more predictive power. But since the TAN PR curve lies closer to the ideal curve, TAN's PR curve can be assumed to have more predictive power.

3) If diseased (i.e metastases which is a secondary tumour alongside a primary tumour) individual can be defined as a positive instance, then the critical component is false negative (i.e a test data item falsely classified as negative, since it is more important to identify all positive instances). Therefore, a model with a higher recall is the ideal choice for the given problem. Also for a given recall value, the model with the higher precision has more predictive power. It can be seen from the plots that (say for the range of recall from 0.2 to 0.4), TAN model shows a higher precision for the same recall than the Naive Bayes model and hence has a higher predictive power.

2) ROC Curves vs PR Curves

Advantages & Disadvantages:

1) ROC curves are advantageous when the misclassification costs vary by a larger margin. In the given problem, since classifying a negative instance as positive incurs a very less cost (penalty) than classifying a positive instance as negative, ROC curve provides an advantage.
2) PR curves are sensitive to datasets with different fractions of positive and negative instances (imbalanced-skewed datasets). Also PR curves are suited for tasks with a lot of negative instances. Since the given task is a medical domain task, the number of positive instances are very less compared to the number of negative instances and hence PR curves are ideal for this scenario.
3) PR curves do not consider false negatives in any of the computation (precision or recall). Since false negatives are not of importance in the given problem, PR curves are appropriate.

Since the PR curve is well suited for tasks with lot of negative instances and it gives more information with respect to imbalanced datasets (generally the case with medical domain), **PR curves are more informative**.

Part 3)

1) Sample mean of deltas = 0.00714285714286
2) t-statistic = 0.428571428571
3) p-value = 0.67836

The p-value that the mean of deltas would arise from null hypothesis is 0.67836. Since p<0.05 does not hold, null hypothesis holds and hence the **two systems have almost the same accuracies**.

Intermediate values in each of the 10-folds:

Iteration No:1
Accuracy - Naive Bayes: 0.785714285714 TAN: 0.714285714286
Delta in Naive Bayes accuracy and TAN Accuracy: 0.0714285714286

Iteration No:2

```
Accuracy - Naive Bayes: 0.785714285714 TAN: 0.785714285714
Delta in Naive Bayes accuracy and TAN Accuracy: 0.0

Iteration No:3
Accuracy - Naive Bayes: 0.785714285714 TAN: 0.785714285714
Delta in Naive Bayes accuracy and TAN Accuracy: 0.0
Iteration No:4
Accuracy - Naive Bayes: 0.857142857143 TAN: 0.928571428571
Delta in Naive Bayes accuracy and TAN Accuracy: -0.0714285714286

Iteration No:5
Accuracy - Naive Bayes: 0.928571428571 TAN: 0.928571428571
Delta in Naive Bayes accuracy and TAN Accuracy: 0.0

Iteration No:6
Accuracy - Naive Bayes: 0.857142857143 TAN: 0.857142857143
Delta in Naive Bayes accuracy and TAN Accuracy: 0.0

Iteration No:7
Accuracy - Naive Bayes: 0.928571428571 TAN: 0.857142857143
Delta in Naive Bayes accuracy and TAN Accuracy: 0.0714285714286

Iteration No:8
Accuracy - Naive Bayes: 0.714285714286 TAN: 0.785714285714
Delta in Naive Bayes accuracy and TAN Accuracy: -0.0714285714286

Iteration No:9
Accuracy - Naive Bayes: 0.928571428571 TAN: 0.857142857143
Delta in Naive Bayes accuracy and TAN Accuracy: 0.0714285714286

Iteration No:10
Accuracy - Naive Bayes: 0.9375 TAN: 0.9375
Delta in Naive Bayes accuracy and TAN Accuracy: 0.0
```