	Class 1	Class 2	True Class	<0.2	<0.4	<0.6	<0.8	<0.9
1	0.2	0.6	2	1	2	2	2	2
2	0.8	0.3	1	1	1	1	1	2
3	0.55	0.44	2	1	1	2	2	2
4	0.52	0.49	2	1	1	2	2	2
5	0.88	0.46	1	1	1	1	1	2
6	0.55	0.3	1	1	1	2	2	2
7	0.1	0.7	2	2	2	2	2	2
8	0.4	0.7	2	1	1	2	2	2
9	0.55	0.45	1	1	1	2	2	2
10	0.7	0.33	2	1	1	1	2	2

Table with 5 different thresholds and the prediction for each instance:
Calculating the True positive rate and false positive rate for every threshold:

TPR 0.2: TP=4 TN=1 FN=0 FP=5 TPR= 4/4=1 FPR= 5/6=0.83

TPR 0.4: TP=4 TN=2 FN=0 FP=4 TPR= 4/4=1 FPR=4/6=0.67

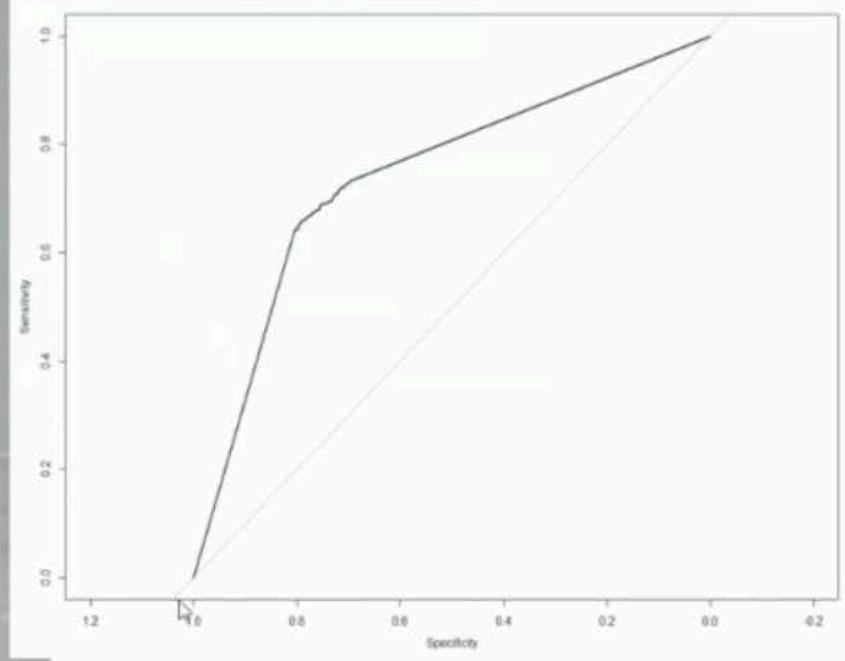
TPR 0.6: TP=2 TN=5 FN=2 FP=1 TPR= 2/4=0.5 FPR=1/6=0.17

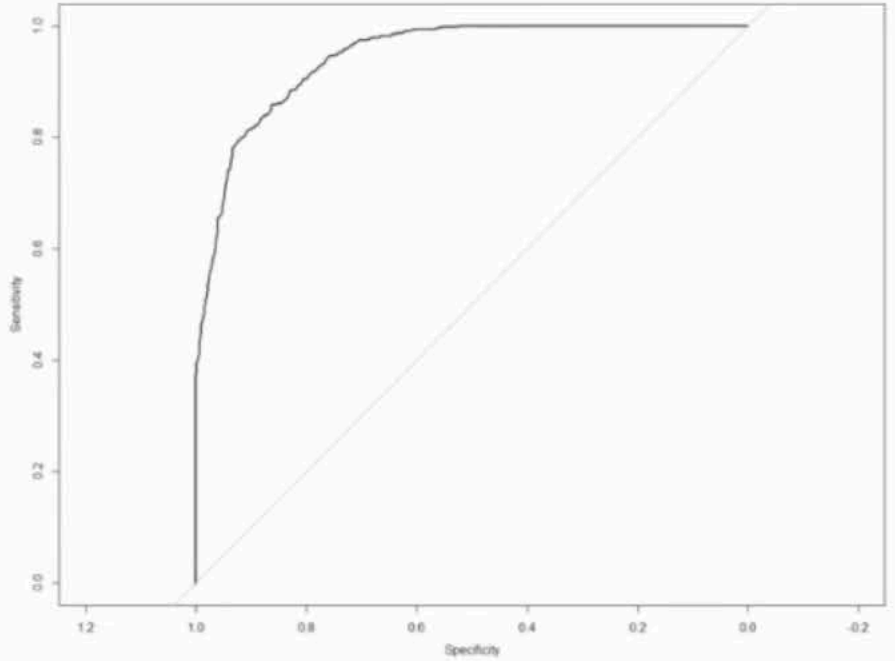
TPR 0.8: TP=2 TN=6 FN=2 FP=0 TPR= 2/4=0.5 FPR=0/6=0

TPR 1.0: TP=0 TN=6 FN=4 FP=0 TPR= 0/4=0 FPR=0/6=0

Question 1

- Which ROC curve is preferable?





The one on the right has a higher area under curve, meaning that there is a point at which a better TPR/FPR is found, the best ratio being in the upper left corner at 1 specificity (1-FPR), 1 sensitivity (TPR).

Question 1

- Explain the concept of AUC, and why the AUC is regarded to be a better metric than the accuracy.

The area under receiver operating curve shows the true positive/false positive rate for **various thresholds** to **classify** a binary dependent variable. It shows the accuracy the classifier will obtain in case that also different ratios of true positive and negatives are present, i.e., in case of a skewed distribution this can lead to majority prediction. It gives a better insight into how well the algorithm thus performs in extremer cases such as fraud detection, etc.

Question 2

- Make a comprehensive comparison of classification by using logistic regression and support vector machine. Describe how they achieve their outcomes, and what their strengths and weaknesses are.
- Word limit: 400

Question 3

- Consider the following HR data. Discuss what models you can build to predict whether someone is going to leave (attrition). What would they look like? What variables will be included? Build a decision tree using information gain with at least two levels. Check at least two variables per split. How well does your model perform?