Q1 - (iv)

The data we are working with contain measurements for normal patients and those with hyperthydroidism. The given data set is imbalanced data, which means that we have a minority and majority class. In cases like this it is understood to use SMOTE to up-sample the minority class to create better predictions. Accuracy is not a good measurement for imbalanced data, it is clear to only use the ROC as a measurement here.

As we can see from Table 1, from the first random split (highlighted in yellow) the AUC's of both kNN and LDA are equal to one and consequently the ROC curve shows a 90-degrees-angle (Figure 1), which mainly indicates no predictions of false positives and false negatives. Overall, Table 1 is showing a slightly higher result for AUC for the kNN method than LDA.

Considering the boxplot in Figure 2, it is evident that there is greater variability when applying LDA. LDA shows higher difference between upper quartile and lower quartile and a lower medium. Indication for the better model is a higher located medium and a smaller spread between the upper and lower quantile of the boxplot, which is in our case the kNN method, which emphasizes our findings above.

However, both models can be seen as highly effective as they show each a median above 99% and therefore, indicate that both are good classification methods for our data set.

One advantage of LDA is the calculated coefficients of linear discriminants. The sum of linear discriminants multiplied by the corresponding elements of predictors, gives a score for each respondent. This score is used to compute the posterior probability, which serves as base for this classification method.

KNN is more time consuming, as the method itself takes a lot of classification time. Furthermore, it is difficult to find the optimal k. Under the condition of our small data set, we can conclude that kNN is the better classification method, showing that the underlying structure of the real data is not entirely linear and therefore the more flexible approach of kNN allows us to classify patients better. Nevertheless, if the size of data set should increase massively, kNN might increase the computing time and therefore LDA might be the better option.

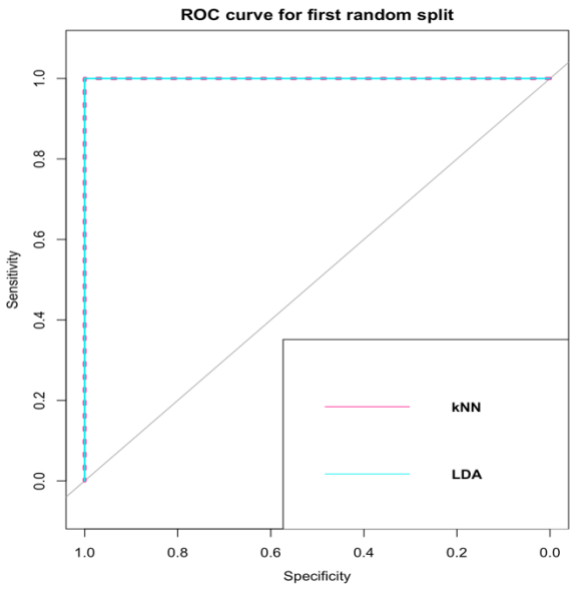

Figure 1: ROC curve for the first random split

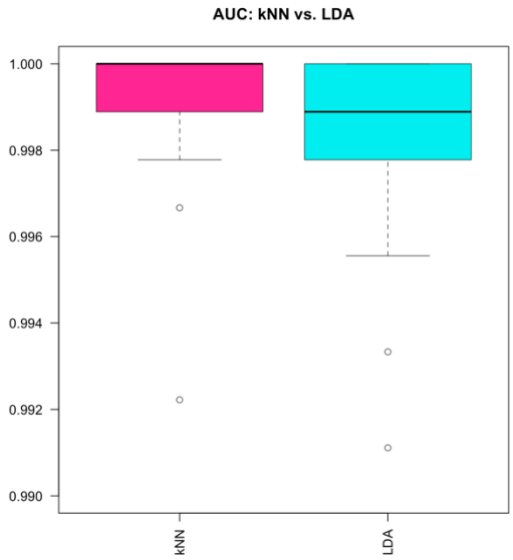

Figure 2: Box plot based on 20 values of AUC

# random split	AUC - kNN	AUC - LDA
1	1.0000000	0.9977778
2	1.0000000	1.0000000
3	1.0000000	1.0000000
4	1.0000000	0.9977778
5	1.0000000	1.0000000
6	1.0000000	0.9977778
7	1.0000000	1.0000000
8	1.0000000	0.9933333
9	0.9977778	0.9977778
10	0.9966667	0.9911111
11	0.9922222	0.9777778
12	1.0000000	1.0000000
13	1.0000000	1.0000000
14	0.9988889	1.0000000
15	0.9988889	0.9977778
16	1.0000000	1.0000000
17	1.0000000	1.0000000
18	1.0000000	0.9977778
19	0.9977778	0.9955556
20	1.0000000	1.0000000
Mean	**0. 9991111**	**0. 9972222**

Table 1: 20 AUC Values of kNN and LDA