Question 1.

1) To prune the tree, complexity parameter (cp) was used and the highest training accuracy of 0.737 was yielded by a cp of 0.33 (Figure1). Pruning tree any further will lower the accuracy and not pruning it will lead to overfitting.

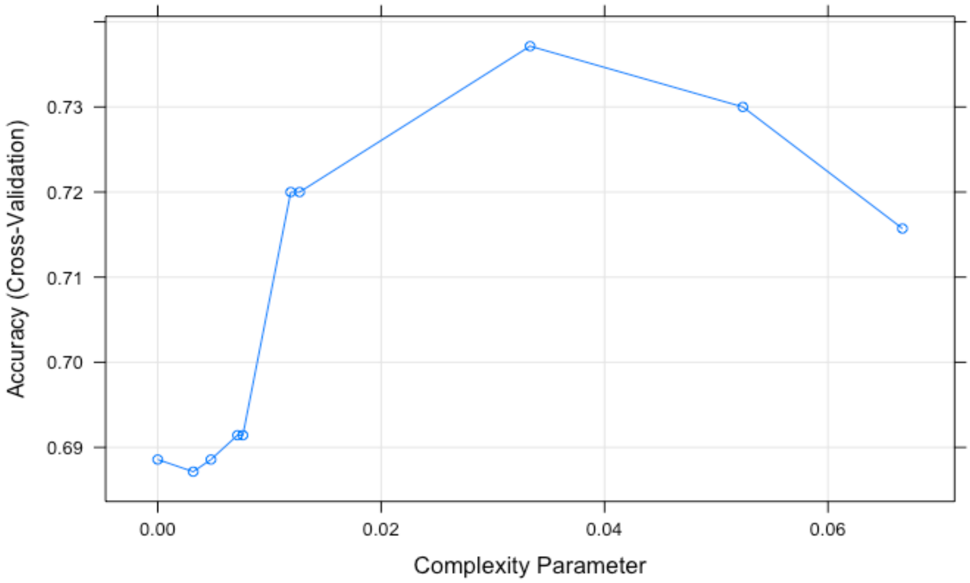

Figure 1: Decision Tree Cross-Validation

The final tree has used 3 features to create the splits and has 4 terminal nodes (Figure2). CheckingAccountStatus.none is the most important in splitting the feature space hence it appears first. If the observation is true according to this feature, then it flows through other internal nodes to be classified as Good or Bad, otherwise it is classified as Good. Therefore, Duration and CreditHistory.ThisBank.AllPaid are next best variables and thus are needed to continue splitting the feature space until the spaces comprise of pure/accurate enough classes. The other variables are not as important in classification according to this decision tree. The tree has erroneously classified 31% of the test set observations.

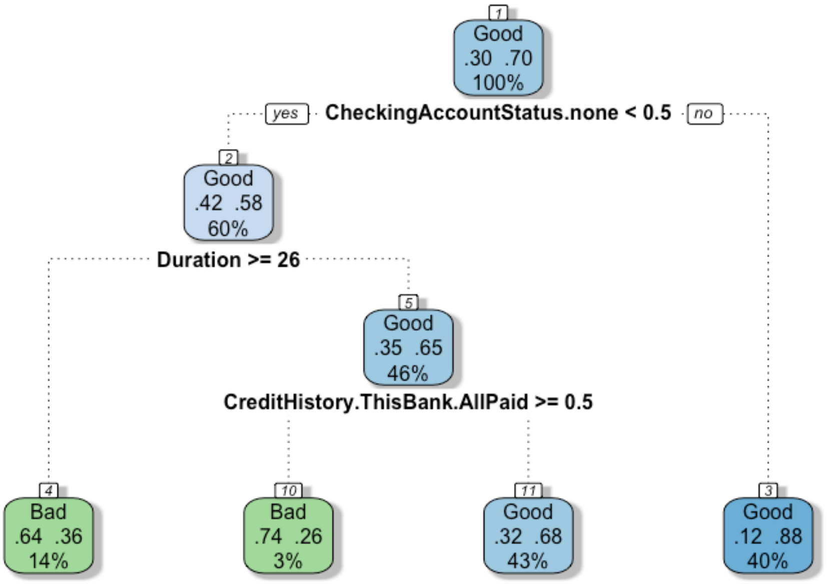

Figure 2: GermanCredit Decision Tree

2) There were 18 randomly selected predictors (Figure 3) per tree in the random forest and the resulting accuracy is 0.7628 (an improvement from the decision tree) – using more features will reduce accuracy. Whereas, in the test set, the random forest has an error rate of 0.2567 and accuracy of 0.7433 – the OOB error rate of 24.57% is close to this figure, hence showing consistent performance. This difference in accuracies may imply less chances of overfitting and that the random forest is overall a better classifier for GermanCredit data than decision tree.

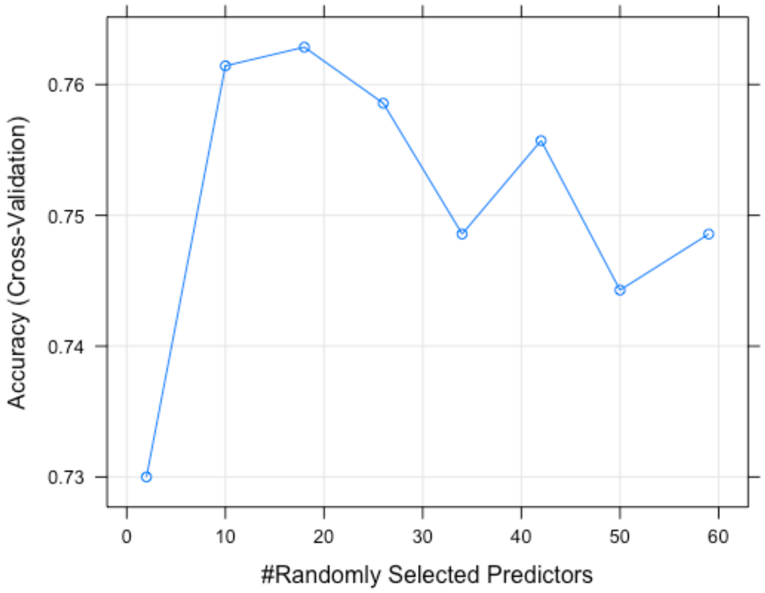

Figure 3: Random Forest Cross-Validation

According to the plot (Figure 4), Amount, Age, Duration, and CheckingAccountStatus.none are the 4 most important features in classifying the observations in the random forest. The importance is based on the overall choice of variables used by the trees within the forest; moreover, two of these variables are also ranked as important by the decision tree model for increasing the purity/accuracy of feature spaces.

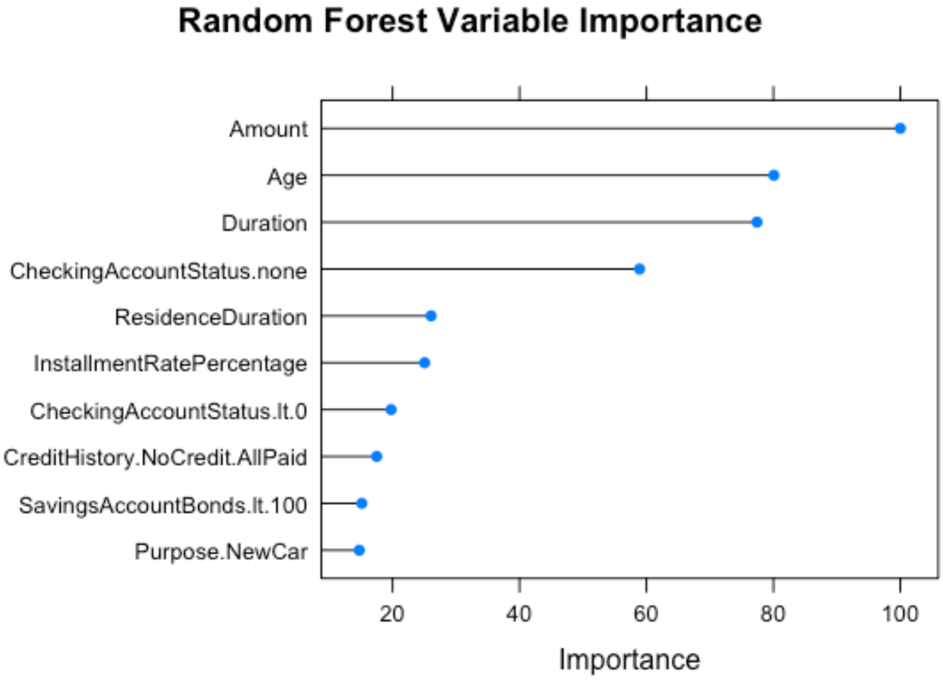

Figure 4: Variable Importance Random Forest

3) The discrete classification predictions were converted into probabilities to create the ROC curve and calculate the AUC. The random forest (AUC=0.78) does a much better job overall at correctly classifying observations of the GermanCredit dataset at different threshold values. The desired threshold will vary on the goals of the classification, but the decision tree (AUC=0.69) also does a better job than randomly classifying observations to either class. As seen in Figure 5, the random forest has different results for many thresholds, but for the decision tree there are very few thresholds where the FPR and TPR change much – perhaps because the probabilities are calculated differently in both the models (votes in random forest against the proportion of observations at each leaf in the decision tree). The random forest seems more robust here.

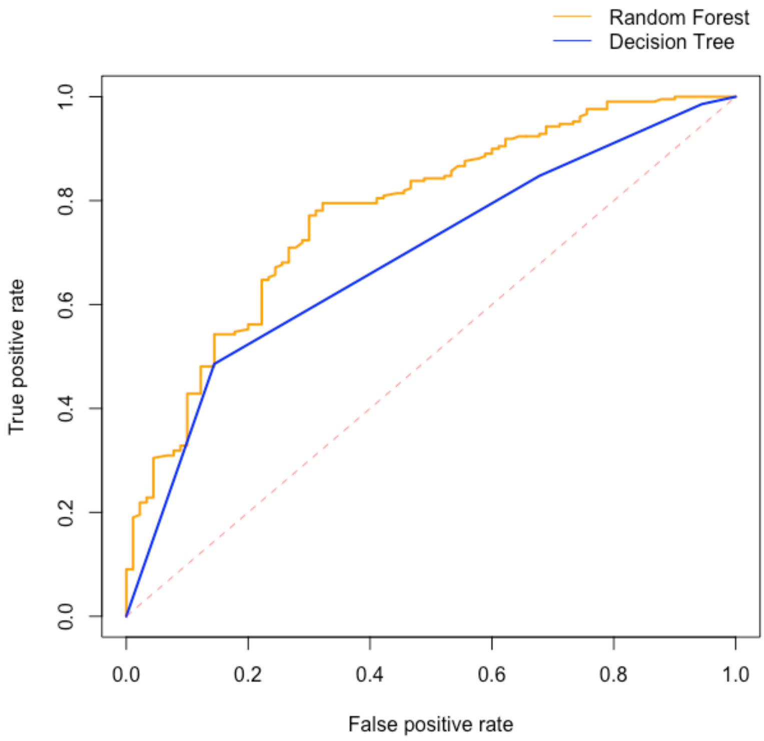

Figure 5: ROC curve for Decision Tree and Random Forest

Question 2.

1) Figure 6 shows the simulated non-linearly separable dataset with 3 classes and 2 features.

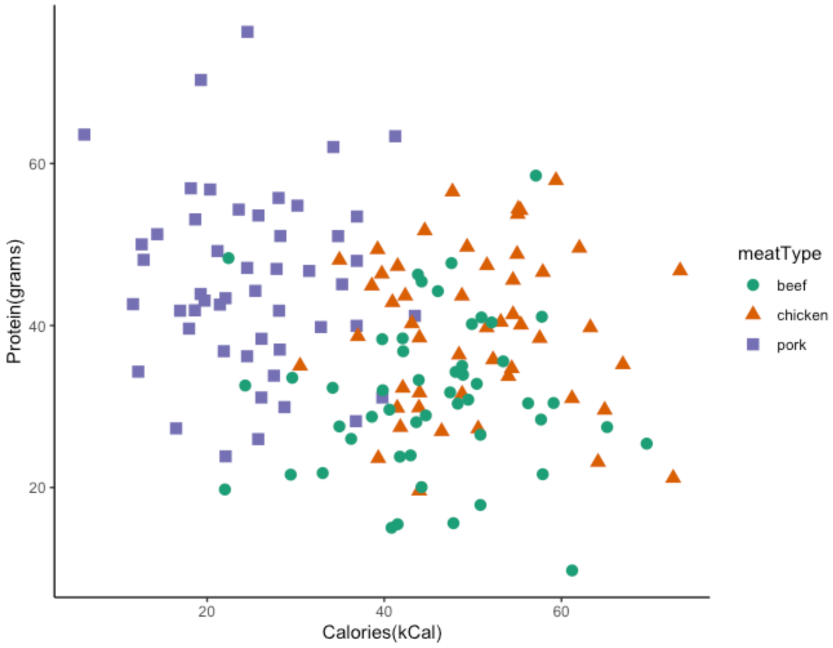

Figure 6: Scatterplot of simulated dataset

2) After splitting the dataset equally to test and training sets, training Support Vector Machine models on 3 different kernels with 5-fold Cross-Validation to tune parameters, the following results were found (Figure7):

Kernel	**Training Accuracy**	**Test Accuracy**
Linear *Cost=0.25, Support Vectors=54*	0.6933	0.706
Polynomial *Cost=0.5, Support Vectors=70, Degree=2, Scale=0.0325*	0.76	0.72
Radial *Cost=4, Support Vectors=63, Sigma=0.01*	0.76	0.706

Figure 7: Results of SVM on different kernels

The linear kernel used a cost of 0.25 from the ones cross-validated with, meaning that the classification hyperplane is the hardest as compared to polynomial and radial kernels. The radial kernel allows more violations of the margin with a cost of 4. Overall, the training accuracy is the same with radial and polynomial kernels, and lowest with linear; however, the test accuracy is best for polynomial. Counterintuitively, the test accuracy of linear is slightly higher than training accuracy and same as the test accuracy of radial - there seems to be some level of high variance. The relatively larger difference in the radial training and test accuracies also exhibit the variance. One of the possible reasons in both models could be the small number of observations to train with (75), especially for the radial kernel, which fits data well, but the high-dimensional projection makes it susceptible to overfitting the small dataset. Lastly, the second-degree polynomial kernel, which although was slow and hardest to tune, captured the non-linearity since it had the highest training and testing accuracy and the lowest variance. One drawback was that it used 70 support vectors out of 75 training observations, hence the cost can be further tuned to improve this.

Question 3.

1) Vectors of AUC values

kNN	0.9933	1	1	0.9933	1	1	1	0.9956	1	1
LDA	0.9911	0.9978	1	0.9956	1	1	1	0.9978	0.9978	0.9956

Figure 8: AUC comparison of kNN and LDA

The chosen k was 15 in most of the iterations. The LDA was not tuned because there were only 2 classes, and the only possible projection is one-dimensional (C-1).

2) Boxplots

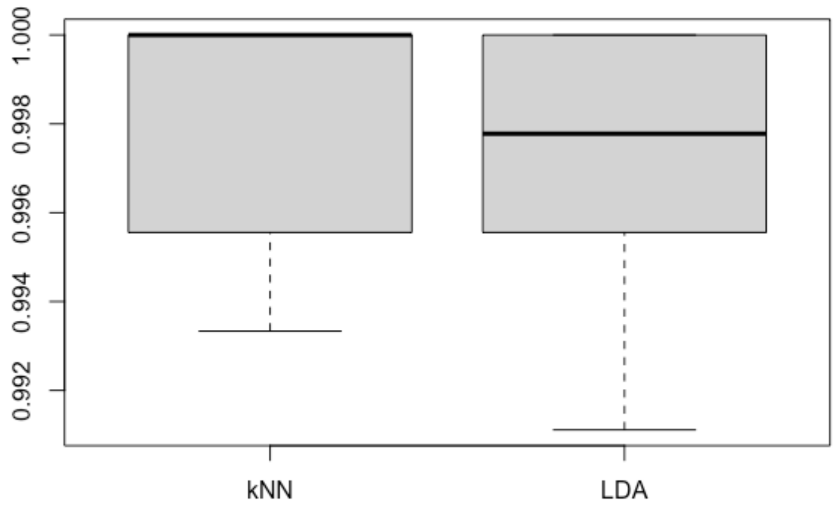

Figure 9: AUC boxplots of kNN and LDA

3) As per the boxplot (Figure9), both the models show high AUCs in all 10 training iterations. The methods of calculating the probabilities are very different in both the models – neighbour based and Bayesian. The AUC values vary less and are all above 0.99 (broadly similar). 1 is the mode for both the sets with kNN more consistent and centred around 1 while LDA centred around 0.997 and changing slightly across iterations. The AUC differences are minute and performance of LDA is high, so there may be a linear boundary between the classes after projection while meeting LDA assumptions of normality. kNN, prone to outliers and high variance, has classified well despite the five-feature space. The AUC may imply that observations within a class are positioned closely together to facilitate predictions based on distance to the neighbours, but without intuition into variable importance. Lastly, it is also essential to assess using other metrics such as accuracy to compare these discrete classification models more robustly.