INFO411: Data Mining and Knowledge Discovery
Assignment 2

This assignment has a weighting of 12%. Submission of the answers must be done online by using the submission link that is associated with this subject for Assignment 2 on Moodle. The submission should have the following form:
"

- **One PDF document.** The PDF must contain typed text of your answers (do not submit a scan of a handwritten document). The document can include computer generated graphics and illustrations (hand drawn graphics and illustrations will be ignored). The size limit for this PDF document is 20MB. All questions are to be answered. A clear explanation needs to be provided with each answer.
- **One .R file** with code to reproduce any analyses done.

Submissions made after the due time will be assessed as late submissions and are counted in full day increments (i.e., 1 minute late counts as a 1 day late submission). There is a 25% penalty for each day after the due date. The submission site closes four days after the due date. No submission will be accepted after the submission site has closed. **This is an individual assignment. Plagiarism of any part of the assignment will result in having 0 marks for all students involved.**

Questions

1. (**No marks for question 1** - needs to be done in order to attempt the other questions.)

In this assignment we make use of the data **creditworthiness.csv** which was used in Task 2 of Assignment 1. As before, we wish to predict the credit rating that would be assigned to each individual. Recall that data on 2500 customers have been collected, and credit rating for 1962 of them has been assessed as either A, B, or C, coded as 1, 2, or 3, respectively, with the remaining 538 needing to be classified. Use the following code to split the dataset into a training and a test set:

```
# Here, cw.k = the data frame containing the
# dataset with known ratings.
cw.train <- cw.k[1:(nrow(cw.k)/2), ]
cw.test <- cw.k[-(1:(nrow(cw.k)/2)), ]
```

2. **(4.5 marks)** Using default settings, fit a decision tree to the training set to predict the credit ratings of customers using all of the other variables in the dataset.

(a) (0.5 mark) Report the resulting tree.

(b) (1 mark) Based on this output, predict the credit rating of a hypothetical "median" customer, i.e., one with the attributes listed in Table 1 (in the last page), showing the steps involved. **Text**

(c) (0.5 mark) Produce the confusion matrix for predicting the credit rating from this tree on the test set, and also report the overall accuracy rate.

(d) (1.5 marks) What is the numerical value of the gain in entropy corresponding to the first split at the top of the tree? (Use logarithms to base 2, and show the details of the calculation rather than just providing a final answer.)

(e) (0.5 mark) Fit a random forest model to the training set to try to improve prediction, and report the R output. **Compare c) and e)**

(f) (0.5 mark) Produce the confusion matrix for predicting the credit rating from this forest on the test set, and also report the overall accuracy rate.

3. **(2 marks)** Using default settings for svm() from the e1071 package, fit a support vector machine to predict the credit ratings of customers using all of the other variables in the dataset.

(a) (1 mark) Predict the credit rating of a hypothetical "median" customer, i.e., one with the attributes listed in Table 1 (in the last page). Report decision values as well.

(b) (0.5 mark) Produce the confusion matrix for predicting the credit rating from this SVM on the test set, and also report the overall accuracy rate.

(c) (0.5 mark) Automatically or manually tune the SVM to improve prediction over that found in question 3(b). Report the resulting SVM settings and the resulting confusion matrix for predicting the test set. (Any amount of improvement is acceptable.)

4. **(2 marks)** Fit the Naive Bayes model to predict the credit ratings of customers using all of the other variables in the dataset.

(a) (1 mark) Predict the credit rating of a hypothetical "median" customer, i.e., one with the attributes listed in Table 1 (in the last page). Report predicted probabilities as well.

(b) (1 mark) Reproduce the first 20 or so lines of the R output for the Naive Bayes fit, and use them to explain the steps involved in making this prediction.

5. **(1 mark)** Based on the confusion matrices reported in the preceding parts,

(a) (0.5 mark) Which of the classifiers look to be the best? (Be specific, and specify the figures you used to answer this question.)

(b) (0.5 mark) Are there any categories that all classifiers seem to have trouble with?

6. **(2.5 marks)** Consider a simpler problem of predicting whether a customer gets a credit rating of A or not.

(a) (0 mark) Fit a logistic regression model to predict whether a customer gets a credit rating of A using all of the other variables in the dataset, with no interactions.

(b) (0.5 mark) Report the summary table of the logistic regression model fit.

(c) (0.5 mark) Which predictors of credit rating appear to be significant? Which of them are likely to be spuriously so?

(d) (0 mark) Fit an SVM model of your choice to the training set.

(e) (1.5 marks) Produce an ROC chart comparing the logistic regression and the SVM results of predicting the test set. Comment on any differences in their performance.

(Table 1 is in the next page)

Table 1: Attributes of the median person in the credit-worthiness dataset.

functionary	0
re-balanced (paid back) a recently overdrawn current acount	1
FI3O credit score	1
gender	0
0. accounts at other banks	3
credit refused in past?	0
years employed	3
savings on other accounts	3
self employed?	0
max. account balance 12 months ago	3
min. account balance 12 months ago	3
avrg. account balance 12 months ago	3
max. account balance 11 months ago	3
min. account balance 11 months ago	3
avrg. account balance 11 months ago	3
max. account balance 10 months ago	3
min. account balance 10 months ago	3
avrg. account balance 10 months ago	3
max. account balance 9 months ago	3
min. account balance 9 months ago	3
avrg. account balance 9 months ago	3
max. account balance 8 months ago	3
min. account balance 8 months ago	3
avrg. account balance 8 months ago	3
max. account balance 7 months ago	3
min. account balance 7 months ago	3
avrg. account balance 7 months ago	3
max. account balance 6 months ago	3
min. account balance 6 months ago	3
avrg. account balance 6 months ago	3
max. account balance 5 months ago	3
min. account balance 5 months ago	3
avrg. account balance 5 months ago	3
max. account balance 4 months ago	3
min. account balance 4 months ago	3
avrg. account balance 4 months ago	3
max. account balance 3 months ago	3
min. account balance 3 months ago	3
avrg. account balance 3 months ago	3
max. account balance 2 months ago	3
min. account balance 2 months ago	3
avrg. account balance 2 months ago	3
max. account balance 1 months ago	3
min. account balance 1 months ago	3
avrg. account balance 1 months ago	3