IT836 Assignment 2: Advanced Analytics in R

In this assignment you will train a Naïve Bayes classifier on categorical data and predict individuals' incomes. Import the *nbtrain.csv* file. Use the first 9010 records as training data and the remaining 1000 records as testing data.

1. Read the nbtrain.csv file into the R environment.
2. Construct the Naïve Bayes classifier from the training data, according to the formula "income ~ age + sex + educ". To do this, use the "naiveBayes" function from the "e1071" package. Provide the model's a priori and conditional probabilities.
3. Score the model with the testing data and create the model's confusion matrix. Also, calculate the overall, 10-50K, 50-80K, and GT 80K misclassification rates. Explain the variation in the model's predictive power across income classes.
4. Use the first 9010 records as training data and the remaining 1000 records as testing data.
5. What is propose of separating the data into a training set and testing set?
6. Construct the classifier according to the formula "sex ~ age + educ + income", and calculate the overall, female, and male misclassification rates. Explain the misclassification rates?
7. Divide the training data into two partitions, according to sex, and randomly select 3500 records from each partition. Reconstruct the model from part (a) from these 7000 records. Provide the model's a priori and conditional probabilities.
8. How well does the model classify the testing data? Explain why.
9. Repeat step (b) 4 several times. What effect does the random selection of records have on the model's performance?
10. What conclusions can one draw from this exercise?