Activity 06

The cmc.data file corresponds to a sample survey on the prevalence of contraception methods in Indonesia in 1987. The purpose of classification is to predict the type of contraceptive method used (variable "V10" located in the last column of the table) based on a set of 9 attributes described in the file Data description.pdf. Load the data set (cmc.data) in RStudio. Data has no header, the separator is a comma (,). Submit your answers and the script used to generate the answers in a text file (word, pdf, etc). Upload the text file to webcursos. Very important: the file name must match your name.

Deadline: September 5, 9:30pm.

To begin, type and run the following script:

set.seed(1111)

Then, generate a training set (70% of the examples) and test (the remaining 30%). Answer the following questions

1. Train a decision tree using the training set and give the percentage of correct classifications (Accuracy) by evaluating the model with the test set. Give the resulting tree.
2. Based on the confusion matrix of the previous question, what kind of contraceptive method is more difficult for the tree to predict? Justify your answer.
3. Write in words the classification rules that are deduced from the tree for the type of contraceptive method = 3.
4. Train a naive Bayes classifier using the training set and give the percentage of correct classifications by evaluating the model with the test set. Based on the confusion matrix, what kind of contraceptive method is more difficult for the naive Bayes classifier to predict? Justify your answer.
5. Train a naive Bayes classifier using the training set but only using the attributes that appear in the tree obtained in Question 1. Give the percentage of correct classification when evaluating the model with the test set. Does the performance improve when compared with what was obtained in question 4?