Problem statement: Identify risky bank loans using decision trees and Random Forrest

The credit dataset includes 1,000 examples on loans, with a set of numeric and nominal features indicating the characteristics of the loan and the applicant.

A class variable (column 21 - attribute "default") indicates whether the loan went into default [1 - No default, 2 - Default].

Data set name: - credit.csv

Task

1. Split data into training and test datasets [70:30]

2. Build a decision tree using C50,rpart and Random Forest Model and identify the top three predictive features

3. Predict outcomes on the test data

4. Create confusion matrix and print the matrix (use CrossTable function to generate confusion matrix)

5. Try improving the performance of the decision tree (set the trials parameter in C50 and nTrees in Random Forest or you can do some data pre-processing)

6. C50 and Random Forest by default handles missing data. Randomly remove few features for some of the training samples and build the decision tree.

Create the confusion matrix and compare the performance of both the algorithms with and without missing data. (See na.action to understand missing features)