Exercise Sheet 8 — Recursive Partitioning

Problem 1. Titanic Passenger Survival
Consider the data set `titanic.csv` that you analysed in exercise sheet 6. The aim of this exercise is to compare the prediction error you obtain on the test data `titanic_test.csv` using the logistic model you fitted on exercise sheet 6 to trees and random forest.

(a) Impute missing ages and calculate the fare per person for `titanic_test.csv` as you did in exercise sheet 6.

(b) Fit a regression tree to the original data `titanic.csv` and prune it. Interpret the results.

(c) Fit a random forest on `titanic.csv`.

(d) Compute the prediction error rates based on

- your best logistic model of sheet 6
- the classification tree
- the random forest

on the test data set `titanic_test.csv`. You may use the missclassication rate

$$\frac{1}{n}\sum_{i=1}^{n}(y_i \neq \hat{y}_i),$$

where $\hat{y}_i$ is the predicted survival status for individual i and y_i is the true survival status for individual i, $i = 1, \dots, n$.