Exercise 1: Classifying `spam`

a) Take a look at the `spam` dataset (`?mlr3::mlr_tasks_spam`). Shortly describe what kind of classification problem this is and access the corresponding task predefined in `mlr3`.

b) Use a decision tree to predict `spam`. Re-fit the tree using two random subsets of the data (each comprising 60% of observations). How stable are the trees?

(Hint: Use `rpart.plot()` from the package `rpart.plot` to visualize the trees.)

c) Forests come with a built-in estimate of their generalization ability via the out-of-bag (OOB) error.

   i) Show that the probability for each observation to be OOB in an arbitrary bootstrap sample converges to $\frac{1}{e}$.

   ii) Verify this result empirically by a small simulation. For this, draw 1000 bootstrap samples from a set of 1000 IDs and compute the average relative frequency of being OOB over all IDs.

   iii) Use the random forest learner `classif.ranger` to fit the model and state the out-of-bag (OOB) error.

d) You are interested in which variables have the greatest influence on the prediction quality. Explain how to determine this in a permutation-based approach and compute the importance scores for the `spam` data.

(Hint: use an adequate variable importance filter as described in
`https://mlr3filters.mlr-org.com/#variable-importance-filters`.)

Exercise 2: Decision boundaries

Simulate 500 samples from the `mlbench.spirals` data with a standard deviation of 0.1, and 4 cycles.

Visualize the decision boundaries of a random forest (`classif.ranger` learner from `mlr3learners`), using `mlr3viz::plot_learner_prediction`, for forest sizes $M \in (1, 2, 10, 100, 1000)$ trees. Explain what you see.