Exercise 1:

Working hard to improve their machine learning skills, our medical research group has recently discovered **bagging**. An expert in the field, researcher Laetitia challenges her colleagues to explain the concept of bagging in one sentence, without using the following words: bootstrap, sample, aggregate, ensemble, base learner.

1) Define bagging without using the words mentioned above.

Random forests apply the principle of bagging to classification and regression trees (CART). This means that for each of the M bootstrap samples of the training data $\mathcal{D}$, a CART base learner $b^{[m]}$ is trained. These models $\hat{f}(x) = \hat{b}^{[m]},\ m = 1,\ \ldots,\ M$ are then aggregated to create the ensemble model $\hat{f}^{[M]}(\mathbf{x})$. However, researcher Laetitia notes that there is more to random forests than just merely applying bagging to trees.

2) Explain how random forests further modify the bagging approach to optimize performance.

Researcher Lisa is very skeptical towards using random forests for her research: *"Most of the time, we are more interested in explaining clinical outcomes rather than predicting them. I like linear models because they are highly interpretable. I can interpret a single tree as well - but hundreds of trees within a random forest?"*
Indeed, random forests are faced with challenges regarding their interpretability. However, the lecture introduced some first extra tools which have been developed to tackle this problem. To identify the importance of a feature variable in generating predictions within the random forest, two measures were introduced:

- Variable importance measure based on improvement in split criterion
- Variable importance measure based on permutations of OOB observations

3) Explain each measure in your own words, using as few sentences as possible. You can find the pseudocode for the algorithms used to calculate the measures in the lecture slides.