Exercise 1:

Our medical research group has just learned about a new learner: Classification and Regression Trees (CART). Enthusiastic about its ability to be applied to both classification and regression tasks, the team is eager to use CART for their practical research.
However, there is still some discussion about some properties of the algorithm. Researcher Laetitia describes how the tree is grown by greedy optimization. This means each time a node $\mathcal{N}$ is split, all possible points t for all features x_j are compared in terms of their empirical risk $\mathcal{R}(\mathcal{N}, j, t)$. The split is then made with regards to the combination of feature and split point (j, t) that results in the lowest risk on the training data. Starting with the root node (containing all observations), this is recursively applied until a stopping criterion is hit.
Researcher Holger concludes this means that CART picks the optimal element of the hypothesis space $\mathcal{H}$, i.e. there can be no other element of $\mathcal{H}$ with lower empirical risk on the training data under consideration of the stopping criterion.

1) **Is his conclusion correct? Explain your choice.**

Furthermore, the team has learned that CART can perform feature selection. This describes the learners ability to discern between relevant and irrelevant features in model construction, and potentially select only a subset of the feature space for the final model.

2) **Explain how CART performs automatic feature selection. Is there a way to gauge how relevant or irrelevant some feature x_j could be after a CART model has been trained?**

Remember the logistic regression model our research group had implemented some weeks ago? The task was to predict whether a patient admitted to the hospital will require intensive care, which is a binary classification task with target space $\mathcal{Y} = \{0, 1\}$. The feature space is: $\mathcal{X} = (\mathbb{R}_0^+)^3$, with $\mathbf{x}^{(i)} = (x_{age},\ x_{blood\,pressure},\ x_{weight})^{(i)} \in \mathcal{X}$ for $i = 1, 2, \ldots, n$ observations. It turns out that although the logistic regression model itself generates useful predictions, its utility for the hospital is rather limited. This is because there is not always time for the nurses to ask patients about their weight, leaving x_{weight} as NA. As a consequence, the logistic regression model cannot be used as it has no possibility to reasonably deal with missing values. Researcher Lisa wonders whether using a CART model instead would allow for a more efficient use in an everyday hospital setting.

3) **Can a CART model handle missing feature values during prediction? Explain your choice.**

The group decides to train the CART model. However, they are no entirely sure how exactly the CART algorithm works in detail. For example, they do not know how many possible split points there are for splitting a node, e.g. the root node.

4) **Calculate an upper bound for the number of possible split points for splitting the root node.**

Exercise 2:

In this exercise you will build a small CART regression tree by hand on a tiny dataset.

Obs	X_1	Y
1	1.0	0
2	2.0	4
3	3.0	6
4	4.0	9

Parent stats: $\bar{y} = 4.75$, $\mathrm{SSE}_{\mathrm{parent}} = \sum_i (y_i - \bar{y})^2 = 42.75$.

Candidate splits (precomputed means, just plug into SSE):

Split on X_1	Left obs	Right obs	$\bar{y}_L$	$\bar{y}_R$
1.5	$\{1\}$	$\{2,3,4\}$	0	19/3 (≈ 6.33)
2.5	$\{1,2\}$	$\{3,4\}$	2	7.5
3.5	$\{1,2,3\}$	$\{4\}$	10/3 (≈ 3.33)	9

Compute $\mathrm{SSE}_{\mathrm{split}} = \sum_{i \in L} (y_i - \bar{y}_L)^2 + \sum_{i \in R} (y_i - \bar{y}_R)^2$ for each row.

(a) Sketch a scatter plot of (x_1, y) and the globally optimal 2-split tree. Show that the total SSE equals 2.

(b) Run CART's greedy search:

(i) List all candidate split points for X_1.

(ii) For each candidate, compute SSE and identify the best first split.

(iii) Grow the second split greedily and report the final 2-split tree.

(c) Compare the greedy 2-split tree with the optimal one. Are they the same?

Exercise 3:

Obs	X_1	X_2	Y
1	1.0	2.4	0
2	2.0	2.6	4
3	3.0	3.0	6
4	4.0	4.0	9

(a) Assume you fitted a CART of depth 1 which picks the root split $X_1 < 2.5$. Build a **surrogate split** at the root that mimics this primary split using only X_2. State (i) the surrogate rule and (ii) its *agreement* with the primary partition, where

$$\text{agreement} = \frac{\#\{\text{rows sent to the same child by both splits}\}}{\text{total rows in the node}}.$$

(b) A new observation has $X_1 = \mathrm{NA}$ and $X_2 = 2.30$. Using the surrogate idea, where is this observation routed and what would you predict?