1. Baseado na partição do espaço de covariáveis apresentado na Figura 1:

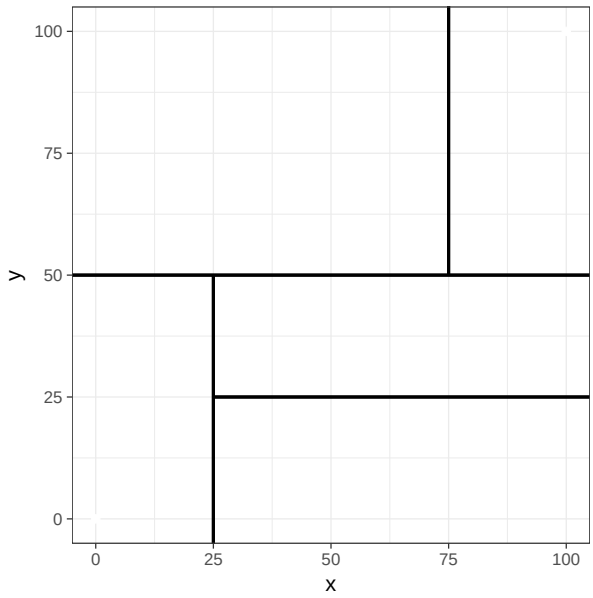

Figura 1: Partição do espaço das covariáveis

(a) Esboce a correspondente árvore de regressão. Deixe claro quais as variáveis usadas em cada partição e os respectivos pontos de corte;

(b) Na Tabela 1 é apresentada a amostra usada para ajuste da árvore:

Tabela 1: Observações usadas para ajuste da árvore

x	10	30	60	25	80	20	90	50	15	90
y	20	40	15	80	80	10	40	10	60	55
z	6	16	10	20	30	8	20	12	28	38

sendo z a variável resposta. As cinco observações apresentadas na Tabela 2, por sua vez, não foram utilizadas no ajuste, sendo reservadas para validação.

Tabela 2: Observações usadas para validação da árvore

x	50	20	75	15	90
y	75	60	30	5	10
z	25	30	18	10	13

Calcule a soma de quadrados de resíduos para a amostra de validação.

2. A Tabela 3 apresenta um resumo dos resultados da análise de validação cruzada aplicada à uma árvore de classificação. São apresentados os tamanhos das árvores obtidas variando o parâmetro de complexidade (**Tamanho**); o erro de classificação de cada árvore, avaliado por validação cruzada (**Erro-VC**) e os erros padrões associados a cada uma das estimativas dos erros de classificação (**Erro Padrão**).

Tabela 3: Resultados da análise por validação cruzada

Tamanho	1	2	3	4	5	6	8	10	12	14
Erro-VC	1	0.8	0.65	0.45	0.4	0.38	0.37	0.35	0.38	0.42
Erro Padrão	0.2	0.18	0.16	0.15	0.14	0.13	0.10	0.08	0.08	0.10

   (a) Esboce a curva de custo complexidade, representando todos os resultados dipostos na Tabela 3;

   (b) Aplique a regra empírica do *"1 erro padrão"*. Qual a árvore selecionada segundo essa regra? Justifique.

3. Um dos atrativos das técnicas de árvores de classificação e regressão é a forma com que são tratados os dados ausentes (dados *missing*). Qual dos itens abaixo melhor descreve a forma padrão de tratar dados ausentes usando CART?

   (a) Os dados ausentes são imputados usando a média das observações disponíveis para a variável;

   (b) Os dados ausentes são imputados usando uma predição baseada na regressão da variável para a qual não se dispõe dos dados e tomando as demais variáveis como preditoras;

   (c) Os dados não são imputados, mas sim utiliza-se um mecanismo de ponderação aos dados disponíveis, atribuindo maior peso às observações semelhantes àquelas para as quais não se dispõe dos dados;

   (d) Seleciona-se, dentre as variáveis para as quais se tem dados disponíveis, a partição que produz maior nível de concordância em relação àquela baseada na variável para a qual não se dispõe dos dados. Executa-se essa partição como alternativa.

   (e) As observações com dados ausentes são automaticamente eliminadas da base.