Descrição Conjuntos de Dados

Os conjuntos de dados usados nos exemplos e nos exercícios propostos no texto são descritos a seguir. As variáveis são descritas na ordem em que aparecem em cada arquivo.

Capítulo 1

canc3.txt: tipo de tumor (0:benigno, 1:maligno), idade (em anos), sexo (1:masculino, 2:feminino), HL e FF (1:ausente, 2:discreta, 3:moderada, 4:intensa).

canc4.txt: grupo de passagem (0 a 28), presença de massa tumoral (1:sim, 0:não), caquexia (1:sim, 0:não) e tempo de sobrevivência (em dias).

capm.txt: taxa de retorno Tbill, retorno Microsoft, retorno SP500, retorno GE e retorno Ford.

censo.txt: unidade da federação, escolaridade média (anos de estudo) e renda média (em reais).

imoveis.txt: imposto do domicílio (em 100 USD), área do terreno (em 1000 pés quadrados), área construída (em 1000 pés quadrados), idade da residência (em anos) e preço de venda do imóvel (em 1000 USD).

reg1.txt: área (em mil pés quadrados) e preço (em mil USD).

reg2.txt: sigla do estado, taxa do combustível (em USD), porcentagem de motoristas licenciados, renda per capita (em USD), ajuda federal às

estradas do estado (em mil USD) e consumo per capita de combustível (em galões por ano).

reg3.txt: nome do estado, população estimada em julho de 75, renda per capita em 74 (em USD), proporção de analfabetos em 70, expectativa de vida 69-70, taxa de criminalidade em 76 (por 100000 habitantes), proporção de estudantes que concluíram o segundo grau em 70, número de dias do ano com temperatura abaixo de zero graus Celsus e área do estado (em milhas quadradas).

reg4.txt: x1, x2, x3, x4, e octanas. A resposta é o número de octanas.

salary.txt: salário anual (em mil USD), sexo, posição na empresa (escore de 1 a 9) e experiência (em anos).

trees.txt: diâmetro (em polegadas), altura (em pés) e volume da árvore (em pés cúbicos).

vendas.txt: total de telhados vendidos (em mil metros quadrados), gastos pela loja com publicidade (em mil USD), número de clientes cadastrados na loja (em milhares), número de marcas concorrentes do produto e potencial da loja.

Capítulo 2

claims.txt: valor do veículo (em 10000 dolares australianos), exposição do veículo, número de sinistros no período, custo total dos sinistros (em dolares australianos), tipo do veículo (em 11 categorias), idade do veículo (em 4 categorias), sexo do condutor principal, área de residência do condutor principal (em 6 categorias) e idade do condutor principal (em 6 categorias).

dfilme.txt: tempo de duração do filme (em horas) e densidade máxima do filme.

energy.txt: total de energia consumida num mês (em kilowatts-hora) e demanda de energia na hora de pico.

insurance.txt: valor pago do seguro (dolares australianos), representação legal (0:não, 1:sim), mês em que ocorreu o acidente e tempo operacional.

milho.txt: quantidade de nitrogênio, quantidade de fosfato e produtividade de milho (libras/acre).

pesca.txt: frota (Santos e Ubatuba), ano (95 a 99), trimestre (1 a 4), latitude (de $23{,}25^{o}$ a $28{,}25^{o}$), longitude (de $41{,}25^{o}$ a $50{,}75^{o}$), dias de pesca, captura (quantidade em kg de peixes capturados) e cpue (captura por unidade de esforço).

restaurante.txt: faturamento anual (em mil USD) e gastos com publicidade (em mil USD).

snack.txt: força necessária para o cisalhamento, tipo de snack (1:A, 2:B, 3:C, 4:D, 5:E), número de semanas.

sobrev.txt: número de células brancas, tempo de sobrevivência (em semanas) e característica morfológica (AG=1 positivo, AG=0 negativo).

turbina.txt: tipo de turbina (1 a 5) e tempo de duração do motor (em milhões de ciclos).

vidros.txt: tempo de resistência (em horas), voltagem (1:200, 2:250, 3:300, 4:350) e temperatura (1:170 graus Celsus, 2:180 graus Celsus).

Capítulo 3

besouros.txt: besouros mortos, besouros expostos e dose.

caduquice.txt: escore no exame psicológico, ocorrência de caduquice (1:sim, 0:não).

camundongos: sexo (1:macho, 0:fêmea), tratamento (1:sim, 0:controle), casos e expostos.

dengue.txt: idade (em anos) do entrevistado, nível sócio-econômico (1:alto, 2:médio, 3:baixo), setor da cidade onde mora o entrevistado (1:setor 1, 2:setor 2) e diagnóstico da doença (1:sim, 0:não).

diabetes.txt: massa corporal, histórico familiar (1:presença, 0:ausência) e atividades físicas (1:presença, 0:ausência) para os casos e para os controles, respectivamente.

dose1.txt: dose, caramujos expostos e caramujos mortos.

dose2.txt: dose, caramujos expostos e caramujos mortos.

dose3.txt: dose, caramujos expostos e caramujos mortos.

equipamentos.txt: tempo, número de equipamentos expostos, número de equipamentos que falaharam.

gestantes.txt: idade (0:$<$ 30, 1:30 ou +), número de cigarros consumidos por dia (0:$<$ 5, 1:5 ou +), tempo de gestação (0:$<$=260 dias, 1:$>$ 260 dias), crianças não sobreviventes e crianças sobreviventes.

grahani.txt: número de lagartos da espécie grahani, total de lagartos, período do dia (1:manhã, 2:meio-dia, 3:tarde), comprimento da madeira (1:curta, 2:cumprida), largura da madeira (1:estreita, 2:larga) e local de ocupação (1:claro, 2:escuro).

insetic.txt: número de insetos mortos, número de insetos expostos, dose do inseticida, inseticida DDT, inseticida γ-DDT e inseticida DDT + γ-DDT (1:presença, 0:ausência).

leuce.txt: idade do paciente (em anos), mancha diferencial da doença, infiltração na medula, células com leucemia, malignidade da doença, temperatura máxima antes do tratamento, tratamento (1:satisfatório, 0:não), tempo de sobrevivência (em meses) e situação (1:sobrevivente, 0:não sobrevivente).

matched.txt: estrato, observação (1:caso, 2:controle), idade da paciente no momento da entrevista (em anos), diagnóstico (1:caso, 0:controle), tempo de escolaridade (em anos), grau de escolaridade (0:nenhum, 1:segundo grau, 2:técnico, 3:universitário, 4:mestrado, 5:doutorado), checkup regular (1:sim, 2:não), idade da primeira gravidez, idade do início da menstruação, número de abortos, número de filhos, peso (em libras), idade do último período menstrual e estado civil (1:casada, 2: divorciada, 3:separada, 4:viúva, 5:solteira). Observações perdidas são denotadas por NA.

meninas.txt: garotas menstruando, garotas entrevistadas e idade média.

morgan.txt: concentração (R, D, M), dose, insetos expostos, insetos mortos.

olhos.txt: cor dos olhos dos pais, cor dos olhos dos avós, número total de filhos e número de filhos com olhos claros.

prefauto.txt: preferência comprador tipo de automóvel (1:americano, 0:japonês), idade do comprador (em anos), sexo do comprador (0:masculino, 1:feminino) e estado civil do comprador (0:casado, 1:solteiro).

pregibon.txt: resposta (1:ocorrência, 0:ausência), volume e razão.

pulso.txt: pulsação em repouso (1:normal, 0:alta), hábito de fumar (1:sim, 2:não) e peso (em kg).

rotifers.txt: densidade, rotifers suspensos, rotifers expostos e espécie (1: Polyarthra, 0:Keratella).

sementes.txt: temperatura da germinação, nível da umidade, nível da temperatura, número de sementes que germinaram.

Capítulo 4

breslow.txt: número de casos de câncer, total de pessoas-anos, número de cigarros por dia (1:não fumante, 2:1-9 cigarros, 3:10-30 cigarros, 4:+ 30 cigarros) e faixa-etária (1:40-49 anos, 2:50-59 anos, 3:60-69 anos, 4:70-80 anos).

canc1.txt: idade no primeiro emprego com 4 níveis (1:<20, 2:20-27, 3:27.5-34.9, 4:35+ anos), ano do primeiro emprego com 4 níveis (1:<1910, 2:1910-1914, 3:1915-1919, 4:1920-1924), tempo decorrido desde o primeiro emprego com 5 níveis (1:0-19, 2:20-29, 3:30-39, 4:40-49, 5:50+ anos), número de casos de câncer e o total de pessoas-anos de observação.

detergente.txt: temperatura da água, uso de M, preferência (X,M), maciez da água, número de pessoas.

emprego.txt: nível de renda (1: < USD 6000, 2: USD 6000-15000, 3: USD 15000-25000, 4: > USD 25000), grau de satisfação (1:alto, 2: bom, 3: médio, 4: baixo) e número de indivíduos.

geriatra.txt: número de quedas no período, intervenção (0:educação somente, 1:educação e exercícios físicos), sexo (0:feminino, 1:masculino), balanço e força.

heart.txt: doença das coronárias (1:sim, 2:não), nível de colesterol (1:menor do que 200 mg/100 cc, 2:200-219, 3:220-259, 4:260 ou +),pressão arterial (1:menor do que 127 mm Hg, 2:127-146, 3:147-166, 4:167 ou +) e número de indivíduos.

navios.txt: tipo do navio (1:A, 2:B, 3:C, 4:D, 5:E), ano da fabricação (1:60-64, 2:65-69, 3:70-74, 4:75-79), período de operação (1:60-74, 2:75-79), tempo de operação (em meses) e número de avarias.

nitrofen: dosagem de nitrofen, total de ovos eclodidos.

quine.txt: etnia (A:aborígine, N:não aborígine), sexo (M:masculino, F: feminino), ano (F0:8ª série, F1:1º ano ensino médio, F2:2º ano ensino médio, F3:3º ano ensino médio), desempenho (SL:baixo, AL:normal) e dias ausentes no ano letivo.

recrutas.txt: hábito de nadar (ocasional, frequente), local onde costuma nadar (piscina, praia), faixa-etária (15-19, 20-25, 25-29), sexo (masculino, feminino) e número de infecções de ouvido.

rolos.txt: comprimento do tecido (em metros) e número de falhas.

store.txt: número de clientes, número de domicílios, renda média anual (em USD), idade média dos domicílios (em anos), distância entre a área e o competidor mais próximo (em milhas) e distância entre a área e a loja (em milhas).

tvcabo.txt: número de domicílios na área (em milhares), porcentagem de domicílios com TV a cabo, renda per capita (em USD) por domicílio com TV a cabo, taxa de instalação de TV a cabo (em USD), custo médio mensal de manutenção de TV a cabo (em USD), número de canais a cabo disponíveis na área e número de canais não pagos com sinal de boa qualidade disponíveis na área.

Capítulo 5

artrite.txt: paciente, ocasião (1:início, 2:1 mês, 3:2 meses, 4:3 meses), gênero (1:masculino, 0:feminino), idade (em anos), tratamento (0:placebo, 1:auronofin), resultado (1:ruim, 2:regular, 3:bom).

ataques.txt: indivíduo, período (1:antes do tratamento, 2:1º período após o tratamento, 3:2º período após o tratamento, 4:3º período após o tratamento), número de semanas em cada período, número de ataques em cada período e tratamento (0:placebo, 1:progabide).

cevada.txt: incidência da mancha (proporção), local (1 a 9) e variedade (1 a 10).

mosca.txt: número de ácaros coletados espécie2, espécie3, espécie6, espécie14, número de partes da placa, posição (1:lateral, 0:central), região (1:São Roque, 2:Pindamonhangaba, 3:Nova Odessa, 4:Ribeirão Preto) e temperatura (em graus Celsus).

mistura.txt: painel, dia, método, mistura, porcentagem de reflectância do pigmento.

ratosgee.txt: animal, período, quantidade de células brancas, quantidade de células vermelhas e número de colônias de células cancerosas.

respiratorio.txt: paciente, tratamento (0:droga ativa, 1:placebo), sexo (0:feminino, 1:masculino), idade (em anos), nível base (0:ausência, 1:presença) e condição do paciente nas visitas (0:boa, 1:ruim).

rinse.txt: voluntário, período (1:início, 2:após 3 meses, 3:após 6 meses), tratamento (1:placebo, 2:rinse A, 3:rinse B) e escore.