Primeira Prova de Introdução à Ciência de Dados

09 de dezembro de 2025

Entrega: deve-se entregar um arquivo .txt (Q1), um arquivo .R (Q2) e uma imagem com o print da matriz de confusão construída na Questão 2; deixe os cálculos da Q1 no arquivo .txt e os cálculos da Q2 como comentários no arquivo .R. Os materiais da prova estão no github.

Questão 1

As duas partes dessa questão devem ser entregues em um arquivo `txt`. O modelo desse arquivo está no github. O nome do arquivo com as duas soluções da Questão deve estar no formato `Q1_seu-nome.txt`. Você pode abrir, por exemplo, o arquivo `txt` no Rstudio.

Parte a)

Em sala de aula, discutimos a importância de padronizar ou normalizar variáveis numéricas calculando as estatísticas (média e desvio-padrão) apenas no conjunto de treinamento e, em seguida, aplicando essas mesmas estatísticas ao conjunto de teste, a fim de evitar o vazamento de informações.

Abaixo é fornecido um código em R que executa as etapas iniciais de pré-processamento e divisão dos dados para um problema de modelagem preditiva, onde se pretende usar um modelo de Vizinhos Mais Próximos (KNN).

```r
library(tidyverse)
library(class)

dados <- read_csv("dados_clientes.csv")

# A variável resposta está na primeira coluna
dados_padronizados <- scale(dados[,-1])

set.seed(42)
indices_treino <- sample(1:nrow(dados), 0.7 * nrow(dados))
treino <- dados_padronizados[indices_treino, ]
teste <- dados_padronizados[-indices_treino, ]

modelo_knn <- knn(train = treino, test = teste,
                  cl = dados$resposta[indices_treino], k = 1)
```

Identifique o erro de modelagem presente na etapa de pré-processamento e explique por que este erro compromete a avaliação do modelo.

Parte b)

A seguir, a diretriz de uma equipe de segurança de um provedor de e-mail: *"O objetivo primário do nosso filtro de spam é garantir que e-mails legítimos cheguem à caixa de entrada do usuário. Nesta aplicação, consideramos um erro crítico e de alto risco quando um e-mail legítimo (não-spam) é incorretamente classificado como spam (Falso Positivo), pois isso pode fazer com que o usuário perca comunicações importantes. Por outro lado, classificar um spam como legítimo (Falso Negativo) é considerado um erro de menor impacto, resultando apenas em inconveniência de ter que deletá-lo."*

A Tabela 1 apresenta duas matrizes de confusão para um problema de classificação de e-mails (Spam vs. Legítimo), usando o mesmo conjunto de teste em dois modelos diferentes. Para a análise, a classe 'Spam' é considerada a classe Positiva.

Tabela 1: Matrizes de confusão para classificação de e-mails (Linhas são as previsões)

(a) Modelo Árvore de Decisão

	Spam	Legítimo
Spam	35	15
Legítimo	5	145

(b) Modelo Naive Bayes

	Spam	Legítimo
Spam	40	10
Legítimo	10	140

Questão 2

Você recebeu um conjunto de dados contendo informações de primatas da família dos grandes símios. Cada linha representa um indivíduo classificado como **bonobo** ou **chimpanzé**, com as seguintes variáveis registradas:

- `especie`: espécie do primata (bonobo ou chimpanzé) — variável a ser predita.
- `altura_cm`: altura do primata, em centímetros.
- `peso_kg`: peso do primata, em quilogramas.
- `tamanho_cranio_cm`: tamanho aproximado do crânio, em centímetros.
- `genero`: sexo do primata (`masculino` ou `feminino`).
- `dia_semana_coleta`: dia da semana em que os dados foram coletados (1 = domingo, ..., 7 = sábado).

Seu objetivo é analisar o conjunto de dados e construir modelos capazes de prever a **espécie** do primata com base nas demais variáveis. É muito importante que as partes discursivas sejam bem fundamentadas.

Importação e exploração inicial

- Carregue o conjunto `dados_primatas.csv`.

Análise gráfica

- Produza visualizações que permitam **comparar bonobos e chimpanzés**.
- Discuta quais variáveis parecem ser mais úteis para distinguir as espécies.
- Apresente um pequeno texto resumindo as principais diferenças entre as duas espécies, com base nas variáveis do estudo.

Modelagem

Árvore de Decisão

- Utilize a semente 0912 e em seguida embaralhe o conjunto.
- Divida os dados em treino e teste (75% treino, 25% teste).
- Com os dados de treino, construa um modelo de **árvore de decisão** para prever a espécie dos primatas com base nas demais variáveis.
- Apresente a árvore resultante e comente as variáveis mais relevantes.
- Utilize esse modelo final para prever as espécies do conjunto de teste.
- Apresente a matriz de confusão a partir das classificações realizadas pelo modelo.
- Dado que um primata foi classificado como bonobo, qual a probabilidade de que ele seja realmente um bonobo? Utilize a matriz de confusão para responder.