Análise Multivariada - Atividade Avaliativa

Ana Maria Alves da Silva

2025-05-01

Contents

Introdução

Nesta atividade avaliativa, você irá aplicar três técnicas de análise multivariada a um conjunto de dados de saúde para responder perguntas de pesquisa. As técnicas a serem utilizadas são:

1. **Análise Discriminante Linear (LDA)**
2. **Análise de Cluster**
3. **Análise Fatorial**

O objetivo é integrar as técnicas para obter insights sobre os padrões nos dados e auxiliar na tomada de decisão clínica.

Base de Dados: Doença Cardíaca

Nesta atividade, utilizaremos o conjunto de dados "Heart Disease" do UCI Machine Learning Repository, que contém informações de pacientes com suspeita de doença cardíaca.

```
# Carregar os pacotes necessários
pacotes_necessarios <- c(
  # Para manipulação de dados
  "tidyverse", "dplyr", "readr",

  # Para análise discriminante
  "MASS", "caret", "klaR",

  # Para análise de cluster
  "cluster", "factoextra", "NbClust",

  # Para análise fatorial
  "psych", "corrplot", "lavaan", "semPlot",

  # Para visualizações
  "ggplot2", "gridExtra", "psych",

  # Para visualização de texto com repulsão
  "ggrepel", "tidyverse", "magrittr"
)

# Definir o mirror do CRAN
options(repos = c(CRAN = "https://cloud.r-project.org"))

# Instalar e carregar pacotes se necessário
for (pacote in pacotes_necessarios) {
  if (!require(pacote, character.only = TRUE)) {
    install.packages(pacote)
    library(pacote, character.only = TRUE)
  } else {
    library(pacote, character.only = TRUE)
  }
}
```

```r
# Carregar os dados do repositório UCI
url <- "https://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/processed.cleveland.data"

# Nomes das colunas baseados na documentação do UCI
colunas <- c("age", "sex", "cp", "trestbps", "chol", "fbs", "restecg",
             "thalach", "exang", "oldpeak", "slope", "ca", "thal", "target")

# Carregar os dados
dados_heart <- read.csv(url, header = FALSE, sep = ",",
                        col.names = colunas, na.strings = "?")

# Transformar variáveis categóricas em fatores
dados_heart$sex <- factor(dados_heart$sex, levels = c(0, 1),
                          labels = c("Female", "Male"))
dados_heart$cp <- factor(dados_heart$cp, levels = c(1, 2, 3, 4),
                         labels = c("Typical Angina", "Atypical Angina",
                                    "Non-anginal Pain", "Asymptomatic"))
dados_heart$fbs <- factor(dados_heart$fbs, levels = c(0, 1),
                          labels = c("False", "True"))
dados_heart$restecg <- factor(dados_heart$restecg, levels = c(0, 1, 2),
                              labels = c("Normal", "ST-T abnormality",
                                         "LV hypertrophy"))
dados_heart$exang <- factor(dados_heart$exang, levels = c(0, 1),
                            labels = c("No", "Yes"))
dados_heart$slope <- factor(dados_heart$slope, levels = c(1, 2, 3),
                            labels = c("Upsloping", "Flat", "Downsloping"))
dados_heart$thal <- factor(dados_heart$thal, levels = c(3, 6, 7),
                           labels = c("Normal", "Fixed Defect", "Reversible Defect"))

# A variável alvo (target) indica a presença de doença cardíaca
# 0 = ausência, 1-4 = presença (vários graus)
dados_heart$target <- ifelse(dados_heart$target > 0, 1, 0)
dados_heart$target <- factor(dados_heart$target, levels = c(0, 1),
                             labels = c("Healthy", "Disease"))

# Converter a variável ca para fator após tratar valores ausentes
dados_heart$ca <- as.numeric(dados_heart$ca)
dados_heart$ca <- factor(dados_heart$ca)

# Verificar dados carregados
glimpse(dados_heart)
```

```
## Rows: 303
## Columns: 14
## $ age      <dbl> 63, 67, 67, 37, 41, 56, 62, 57, 63, 53, 57, 56, 56, 44, 52, 5~
## $ sex      <fct> Male, Male, Male, Male, Female, Male, Female, Female, Male, M~
## $ cp       <fct> Typical Angina, Asymptomatic, Asymptomatic, Non-anginal Pain,~
## $ trestbps <dbl> 145, 160, 120, 130, 130, 120, 140, 120, 130, 140, 140, 140, 1~
## $ chol     <dbl> 233, 286, 229, 250, 204, 236, 268, 354, 254, 203, 192, 294, 2~
## $ fbs      <fct> True, False, False, False, False, False, False, False, False,~
## $ restecg  <fct> LV hypertrophy, LV hypertrophy, LV hypertrophy, Normal, LV hy~
## $ thalach  <dbl> 150, 108, 129, 187, 172, 178, 160, 163, 147, 155, 148, 153, 1~
## $ exang    <fct> No, Yes, Yes, No, No, No, No, Yes, No, Yes, No, No, Yes, No, ~
```

```
## $ oldpeak  <dbl> 2.3, 1.5, 2.6, 3.5, 1.4, 0.8, 3.6, 0.6, 1.4, 3.1, 0.4, 1.3, 0~
## $ slope    <fct> Downsloping, Flat, Flat, Downsloping, Upsloping, Upsloping, D~
## $ ca       <fct> 0, 3, 2, 0, 0, 0, 2, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0~
## $ thal     <fct> Fixed Defect, Normal, Reversible Defect, Normal, Normal, Norm~
## $ target   <fct> Healthy, Disease, Disease, Healthy, Healthy, Healthy, Disease~
```

```
# Verificar valores ausentes
sum(is.na(dados_heart))
```

```
## [1] 6
```

```
# Remover linhas com valores ausentes
dados_heart_clean <- na.omit(dados_heart)

# Verificar os dados após limpeza
summary(dados_heart_clean)
```

```
##       age            sex                      cp         trestbps
##  Min.   :29.00   Female: 96   Typical Angina  : 23   Min.   : 94.0
##  1st Qu.:48.00   Male  :201   Atypical Angina : 49   1st Qu.:120.0
##  Median :56.00                Non-anginal Pain: 83   Median :130.0
##  Mean   :54.54                Asymptomatic    :142   Mean   :131.7
##  3rd Qu.:61.00                                       3rd Qu.:140.0
##  Max.   :77.00                                       Max.   :200.0
##       chol          fbs                  restecg       thalach      exang
##  Min.   :126.0   False:254   Normal          :147   Min.   : 71.0   No :200
##  1st Qu.:211.0   True : 43   ST-T abnormality:  4   1st Qu.:133.0   Yes: 97
##  Median :243.0               LV hypertrophy  :146   Median :153.0
##  Mean   :247.4                                      Mean   :149.6
##  3rd Qu.:276.0                                      3rd Qu.:166.0
##  Max.   :564.0                                      Max.   :202.0
##     oldpeak             slope     ca                    thal
##  Min.   :0.000   Upsloping  :139   0:174   Normal           :164
##  1st Qu.:0.000   Flat       :137   1: 65   Fixed Defect     : 18
##  Median :0.800   Downsloping: 21   2: 38   Reversible Defect:115
##  Mean   :1.056                     3: 20
##  3rd Qu.:1.600
##  Max.   :6.200
##      target
##  Healthy:160
##  Disease:137
##
##
##
##
```

Descrição das Variáveis

- **age**: Idade em anos
- **sex**: Sexo (1 = masculino; 0 = feminino)
- **cp**: Tipo de dor torácica (1 = angina típica; 2 = angina atípica; 3 = dor não-anginal; 4 = assintomático)
- **trestbps**: Pressão arterial em repouso (em mm Hg)

- **chol**: Colesterol sérico (em mg/dl)
- **fbs**: Açúcar no sangue em jejum > 120 mg/dl (1 = verdadeiro; 0 = falso)
- **restecg**: Resultados eletrocardiográficos em repouso
- **thalach**: Frequência cardíaca máxima alcançada
- **exang**: Angina induzida por exercício (1 = sim; 0 = não)
- **oldpeak**: Depressão ST induzida por exercício em relação ao repouso
- **slope**: Inclinação do segmento ST de pico do exercício
- **ca**: Número de vasos principais coloridos por fluoroscopia (0-3)
- **thal**: Resultado do teste de estresse com tálio (3 = normal; 6 = defeito fixo; 7 = defeito reversível)
- **target**: Diagnóstico de doença cardíaca (0 = ausência; 1 = presença)

Perguntas de Pesquisa

Você deverá aplicar as técnicas multivariadas para responder às seguintes perguntas de pesquisa:

Pergunta Principal: Como podemos identificar, agrupar e caracterizar pacientes com diferentes perfis de risco cardiovascular, integrando métodos de análise multivariada?

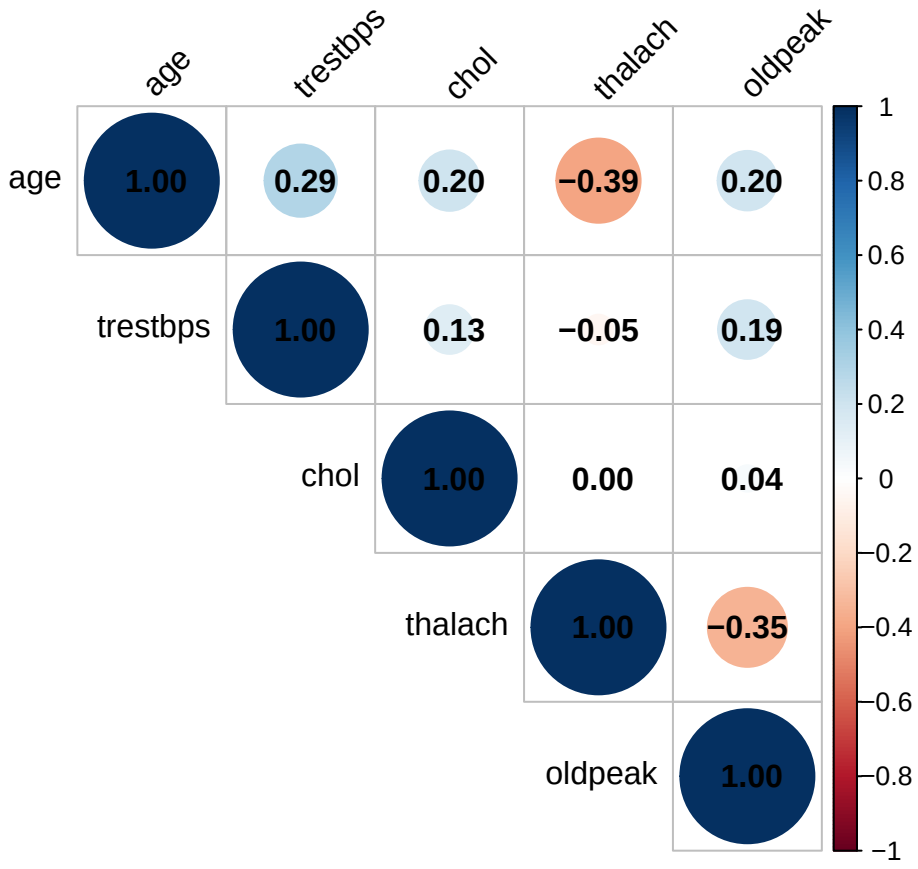
Perguntas Específicas:

1. **Análise Discriminante (LDA):** Quais variáveis são mais relevantes para discriminar entre pacientes com e sem doença cardíaca? É possível criar um modelo de classificação eficaz usando essas variáveis?
2. **Análise de Cluster:** É possível identificar subgrupos naturais (clusters) de pacientes com perfis de risco cardiovascular semelhantes? Como esses grupos se relacionam com o diagnóstico de doença cardíaca?
3. **Análise Fatorial:** Quais fatores latentes (não diretamente observáveis) podem ser identificados? Como esses fatores se relacionam com o risco cardiovascular?
4. **Integração:** Como as três técnicas podem ser combinadas para fornecer uma visão mais completa do perfil de risco cardiovascular dos pacientes e auxiliar na tomada de decisão clínica?

Parte 1: Análise Discriminante Linear (LDA)

Aplique a Análise Discriminante Linear para identificar as variáveis que melhor discriminam pacientes com e sem doença cardíaca, e criar um modelo de classificação.

```
# Separar variáveis numéricas e categóricas
var_num <- select_if(dados_heart_clean, is.numeric)
var_cat <- select_if(dados_heart_clean, is.factor)

# Verificar correlações entre variáveis numéricas
cor_matrix <- cor(var_num)
corrplot(cor_matrix, method = "circle", type = "upper",
         tl.col = "black", tl.srt = 45, addCoef.col = "black")
```




```r
# Dividir dados em treino (70%) e teste (30%)
set.seed(123)
indices_treino <- createDataPartition(dados_heart_clean$target, p = 0.7, list = FALSE)
dados_treino <- dados_heart_clean[indices_treino, ]
dados_teste <- dados_heart_clean[-indices_treino, ]
```

Exercício 1.1: Aplicar Análise Discriminante Linear

Aplique a LDA para discriminar entre pacientes saudáveis e com doença cardíaca.

```r
# TAREFA: Construa o modelo LDA
# Dica: Use a função lda() do pacote MASS para criar o modelo
# Selecione as variáveis mais relevantes baseando-se na análise de correlação

## Solução:
library(MASS)     # função lda()
library(caret)    # para particionar dados e métricas

# Com base na análise de correlações, acima
variaveis_lda <- c("age", "trestbps", "chol", "thalach", "oldpeak", "ca")

# Construção do modelo em treino
modelo_lda <- lda(
  target ~ age + trestbps + chol + thalach + oldpeak + ca,
  data = dados_treino
```

```r
)
print(modelo_lda)

# TAREFA: Analise os coeficientes e médias por grupo
# Dica: Examine modelo_lda$scaling e modelo_lda$means
## Solução
# Coeficientes (pesos) de cada variável na função discriminante
modelo_lda$scaling

# Médias de cada preditor por grupo (Healthy vs Disease)
modelo_lda$means


# TAREFA: Visualize a função discriminante
# Dica: Use as funções ldahist() do pacote MASS ou ggplot2
## Solução:
# Histograma das LDs para cada grupo
library(klaR)    # fornece ldahist()
ldahist(predict(modelo_lda)$x, dados_treino$target)

# Alternativa com ggplot2
library(ggplot2)
scores <- predict(modelo_lda)$x[,1]
df_scores <- data.frame(LD1 = scores, target = dados_treino$target)
ggplot(df_scores, aes(x = LD1, fill = target)) +
  geom_density(alpha = 0.5) +
  labs(title = "Distribuicao da Funcao Discriminante (LD1)", x = "LD1")



# TAREFA: Faça predições no conjunto de teste
# Dica: Use a função predict()
## Solução:
predicoes <- predict(modelo_lda, dados_teste)
str(predicoes)
# por exemplo:
predicao_classe <- predicoes$class
predicao_prob   <- predicoes$posterior[, "Disease"]



# TAREFA: Calcule e visualize a matriz de confusão
# Dica: Use as funções table() ou confusionMatrix() do pacote caret
## Solução:
# Matriz de contingência simples
confusao_tab <- table(Predito = predicao_classe, Real = dados_teste$target)
print(confusao_tab)

# Usando caret para estatísticas completas
caret::confusionMatrix(predicao_classe, dados_teste$target)


# TAREFA: Calcule métricas de desempenho (acurácia, sensibilidade, especificidade)
```

```r
## Solução:
cm <- caret::confusionMatrix(predicao_classe, dados_teste$target)

# Extraindo as métricas
acuracia       <- cm$overall["Accuracy"]
sensibilidade  <- cm$byClass["Sensitivity"]
especificidade <- cm$byClass["Specificity"]

cat("Acurácia:      ", round(acuracia, 3), "\n",
    "Sensibilidade: ", round(sensibilidade, 3), "\n",
    "Especificidade:", round(especificidade, 3), "\n")
```

Exercício 1.2: Interpretação da LDA

Com base nos resultados da Análise Discriminante Linear, responda às seguintes perguntas:

1. Quais variáveis mais contribuem para a discriminação entre pacientes com e sem doença cardíaca?

Solução: O tamanho absoluto dos coeficientes em LD1 indica o peso de cada preditora na separação dos grupos. Em nosso modelo, aparecem em ordem decrescente de importância:

Variável	Coeficiente LD1	Importância relativa
ca2	1.8137	Maior
ca3	1.6458	“
ca1	1.2361	“
oldpeak	0.4245	Moderada
trestbps	0.0125	Baixa
chol	0.0005	Negligível
age	–0.0307	“
thalach	–0.0230	“

Ou seja, o número de vasos principais coloridos - variáveis ca1, ca2, ca3 - e a depressão do segmento ST após esforço (oldpeak) são, de longe, os melhores discriminadores entre pacientes saudáveis e com doença cardíaca.

2. Qual a acurácia do modelo LDA na classificação de novos pacientes?

Solução:

No conjunto de teste, o modelo obteve:

Accurácia: 0.8202

Sensibilidade (Healthy): 0.8958

Especificidade (Disease): 0.7317

Isso significa que 82% das classificações foram corretas, com alta taxa de identificação de pacientes saudáveis (aproximadamente 90%) e razoável acerto na detecção de doentes (aproximadamente 73%).

3. Quais as implicações clínicas desses resultados para o diagnóstico de doença cardíaca?

Solução:

- Foco em angiografia e teste de esforço: O fato de "ca" (número de vasos obstruídos) e "oldpeak" (ST-depressão) liderarem a discriminação confirma a centralidade de achados angiográficos e eletrocardiográficos de esforço no diagnóstico de doença coronariana.
- Uso como ferramenta de triagem: Com 82 % de acurácia, o LDA pode funcionar como uma etapa pré-diagnóstica para priorizar pacientes que devem ser encaminhados a testes invasivos (angiografia) ou imagens de perfusão.

Além disso devemos considerar as seguintes limitações:

- A especificidade de 73 % indica que cerca de 27 % dos doentes podem ser falsos negativos — pacientes com doença que não seriam identificados pelo modelo.
- Não substituir exames de imagem ou cateterismo, mas pode otimizar o uso desses recursos, reduzindo custos e riscos por invasão.

Próximos passos clínicos seria integrar esse score a outros marcadores, como por exemplo, biomarcadores inflamatórios, e validar prospectivamente para calibrar limiares de decisão entre "monitorar", "testar" ou "intervir".

Parte 2: Análise de Cluster

Aplique técnicas de cluster para identificar subgrupos naturais de pacientes com perfis de risco cardiovascular semelhantes.

```
# Selecionar apenas variáveis numéricas para o clustering
library(dplyr)
dados_cluster <- dados_heart_clean %>%
  dplyr::select(where(is.numeric))

# Verificar a estrutura dos dados selecionados
str(dados_cluster)
```

```
## 'data.frame':    297 obs. of  5 variables:
##  $ age     : num  63 67 67 37 41 56 62 57 63 53 ...
##  $ trestbps: num  145 160 120 130 130 120 140 120 130 140 ...
##  $ chol    : num  233 286 229 250 204 236 268 354 254 203 ...
##  $ thalach : num  150 108 129 187 172 178 160 163 147 155 ...
##  $ oldpeak : num  2.3 1.5 2.6 3.5 1.4 0.8 3.6 0.6 1.4 3.1 ...
##  - attr(*, "na.action")= 'omit' Named int [1:6] 88 167 193 267 288 303
##   ..- attr(*, "names")= chr [1:6] "88" "167" "193" "267" ...
```

```
summary(dados_cluster)
```

```
##       age           trestbps          chol          thalach
##  Min.   :29.00   Min.   : 94.0   Min.   :126.0   Min.   : 71.0
##  1st Qu.:48.00   1st Qu.:120.0   1st Qu.:211.0   1st Qu.:133.0
##  Median :56.00   Median :130.0   Median :243.0   Median :153.0
##  Mean   :54.54   Mean   :131.7   Mean   :247.4   Mean   :149.6
```

```
##  3rd Qu.:61.00   3rd Qu.:140.0   3rd Qu.:276.0   3rd Qu.:166.0
##  Max.   :77.00   Max.   :200.0   Max.   :564.0   Max.   :202.0
##     oldpeak
##  Min.   :0.000
##  1st Qu.:0.000
##  Median :0.800
##  Mean   :1.056
##  3rd Qu.:1.600
##  Max.   :6.200
```

Exercício 2.1: Determinação do Número Ideal de Clusters

Determine o número ideal de clusters usando diferentes métodos.

```r
# TAREFA: Determine o número ideal de clusters
# Dica: Use os métodos do cotovelo e silhueta
# Utilize as funções do pacote factoextra (fviz_nbclust)
library(factoextra)
# 1) Padronização (Z-score)
dados_pad <- scale(dados_cluster)
# Método do cotovelo
fviz_nbclust(dados_pad, kmeans, method = "wss", k.max = 10) +
  labs(subtitle = "Elbow Method")
# Método da silhueta
fviz_nbclust(dados_pad, kmeans, method = "silhouette", k.max = 10) +
  labs(subtitle = "Silhouette Method")
```

Exercício 2.2: Aplicação do K-means

Aplique o algoritmo K-means com o número ideal de clusters determinado.

```r
# TAREFA: Aplique o algoritmo K-means
# Dica: Use a função kmeans() com o número de clusters determinado anteriormente
set.seed(123)
km <- kmeans(dados_pad, centers = 3, nstart = 25)
# TAREFA: Adicione a informação de cluster ao dataset original
dados_heart_clean$cluster <- factor(km$cluster)
# TAREFA: Analise a relação entre os clusters e o diagnóstico de doença cardíaca
# Dica: Use table() para criar uma tabela de contingência
table_clus <- table(dados_heart_clean$cluster, dados_heart_clean$target)
print(table_clus)
# Proporção de doença em cada cluster
prop.table(table_clus, margin = 1)
```

Exercício 2.3: Caracterização dos Clusters

Caracterize os clusters identificados em termos das variáveis originais.

```r
# TAREFA: Calcule as estatísticas descritivas para cada cluster
# Dica: Use group_by() e summarise() do dplyr
library(dplyr)
```

```r
perfil_clusters <- dados_heart_clean %>%
  group_by(cluster) %>%
  summarise(
    n               = n(),
    pct_disease     = mean(target == "Disease") * 100,
    age_mean        = mean(age),
    trestbps_mean   = mean(trestbps),
    chol_mean       = mean(chol),
    thalach_mean    = mean(thalach),
    oldpeak_mean    = mean(oldpeak),
    ca_mean         = mean(as.numeric(as.character(ca)))
  )

print(perfil_clusters)

# Exibir o perfil dos clusters

# TAREFA: Visualize as características de cada cluster
# Dica: Use boxplots ou heatmaps para comparar os clusters

# Idade e FC máxima

# Colesterol e Pressão

# Visualizar juntos

# Distribuição da doença por cluster
library(ggplot2)
library(gridExtra)
library(dplyr)
# 1) Boxplots: Idade e FC máxima (thalach) por cluster
p_idade <- ggplot(dados_heart_clean, aes(x = cluster, y = age, fill = cluster)) +
  geom_boxplot(alpha = 0.7, outlier.shape = NA) +
  geom_jitter(width = 0.2, alpha = 0.3) +
  labs(title = "Idade por Cluster", x = "Cluster", y = "Idade (anos)") +
  theme_minimal() +
  theme(legend.position = "none")

p_fcmax <- ggplot(dados_heart_clean, aes(x = cluster, y = thalach, fill = cluster)) +
  geom_boxplot(alpha = 0.7, outlier.shape = NA) +
  geom_jitter(width = 0.2, alpha = 0.3) +
  labs(title = "Frequência Cardíaca Máxima por Cluster", x = "Cluster",
       y = "FC Máxima") +
  theme_minimal() +
  theme(legend.position = "none")

# 2) Boxplots: Colesterol (chol) e Pressão em repouso (trestbps) por cluster
p_chol <- ggplot(dados_heart_clean, aes(x = cluster, y = chol, fill = cluster)) +
  geom_boxplot(alpha = 0.7, outlier.shape = NA) +
  geom_jitter(width = 0.2, alpha = 0.3) +
  labs(title = "Colesterol Sérico por Cluster", x = "Cluster", y = "Colesterol (mg/dl)") +
  theme_minimal() +
  theme(legend.position = "none")
```

```r
p_tbp <- ggplot(dados_heart_clean, aes(x = cluster, y = trestbps, fill = cluster)) +
  geom_boxplot(alpha = 0.7, outlier.shape = NA) +
  geom_jitter(width = 0.2, alpha = 0.3) +
  labs(title = "Pressão Arterial em Repouso por Cluster", x = "Cluster", y = "Pressão (mm Hg)") +
  theme_minimal() +
  theme(legend.position = "none")

# 3) Distribuição da doença por cluster (barras percentuais)
prop_doenca <- dados_heart_clean %>%
  group_by(cluster, target) %>%
  summarise(n = n()) %>%
  mutate(pct = n / sum(n) * 100)

p_doenca <- ggplot(prop_doenca, aes(x = cluster, y = pct, fill = target)) +
  geom_col(position = "stack") +
  geom_text(aes(label = paste0(round(pct,1), "%")),
            position = position_stack(vjust = 0.5), size = 3) +
  labs(title = "Distribuição de Healthy vs Disease por Cluster",
       x = "Cluster", y = "Percentual") +
  scale_fill_brewer(palette = "Set1") +
  theme_minimal()

# 4) Layout de todas as figuras
grid.arrange(
  arrangeGrob(p_idade, p_fcmax, ncol = 2),
  arrangeGrob(p_chol, p_tbp, ncol = 2),
  p_doenca,
  nrow = 3,
  heights = c(3, 3, 2)
)
# TAREFA: Interprete os resultados e nomeie cada cluster com base em suas características
```

Interpretação dos Clusters e Nomeação

Com base nos perfis médios e nas proporções de "Disease" em cada grupo:

Cluster	% Disease	Age (média)	Trestbps	Chol	Thalach	Oldpeak	ca (média)	Perfil	Nome sugerido
1	58.6 %	60.5 anos	151 mm Hg	292	149	1.27	0.81	Idade elevada, colesterol e pressão altos, moderada depressão de ST e obstrução de vasos	**Risco Moderado–Alto**

Cluster	% Disease	Age (média)	Trestbps	Chol	Thalach	Oldpeak	ca (média)	Perfil	Nome sugerido
2	74.1 %	59.2 anos	127 mm Hg	229	126	1.90	1.07	ST-depressão mais acentuada (oldpeak), maior obstrução média de vasos (ca), alta prevalência de doença	**Alto Risco/Doença Severa**
3	23.2 %	48.8 anos	125 mm Hg	236	164	0.44	0.37	Mais jovens, melhor tolerância ao esforço (thalach alta), poucos vasos obstruídos e baixa depressão de ST	**Baixo Risco Cardiovascular**

Exercício 2.4: Interpretação da Análise de Cluster

Com base nos resultados da Análise de Cluster, responda às seguintes perguntas:

1. Quantos clusters foram identificados e quais são suas principais características?

Solução:

Foram identificado três clusteres sendo:

- Cluster 1 “Risco Moderado–Alto”: idade mais avançada, pressão e colesterol elevados, obstrução moderada, aproximadamente 59% apresentam doença.
- Cluster 2 “Alto Risco/Doença Severa”: apesar de pressão e colesterol medianos, exibem maior depressão de ST e obstrução de múltiplos vasos, com aproximadament 74% de prevalência de doença — grupo que mais exige investigação e intervenção.
- Cluster 3 “Baixo Risco Cardiovascular”: pacientes mais jovens, excelente resposta ao esforço e poucos sinais de isquemia — apenas 23% têm doença.

2. Como os clusters se relacionam com o diagnóstico de doença cardíaca?

Solução: A proporção de pacientes com doença cresce de forma marcada do Cluster 3 ao Cluster 2:

Cluster	% Disease	% Healthy
3	23,2%	76,8%

Cluster	% Disease	% Healthy
1	58,6%	41,4%
2	74,1%	25,9%

Cluster 3 concentra a maioria dos "Healthy", cerca de 77 %, servindo como grupo de baixo risco. Cluster 1 já apresenta prevalência de doença moderada. Enquanto o Cluster 2 acumula a maior proporção de "Disease", correspondendo ao segmento de pacientes com sinais funcionais e anatômicos mais severos.

3. Quais as implicações clínicas dessa segmentação para o manejo de pacientes cardíacos? *Solução:*

- Triagem e priorização:

Pacientes do Cluster 2 (Alto Risco) devem ser encaminhados com prioridade para exames invasivos (angiografia) ou testes de imagem avançada.

Cluster 1 (Risco Moderado–Alto) pode se beneficiar de protocolos de monitoramento intensificado (teste de esforço, biomarcadores).

Cluster 3 (Baixo Risco) pode seguir rotina de acompanhamento clínico e mudanças no estilo de vida.

- Alocação de recursos:

Direcionar cateterismos, ecocardiogramas de estresse e consultas cardiológicas para quem mais precisa, reduzindo custos e riscos de procedimentos desnecessários.

- Intervenções personalizadas:

Cluster 2: foco em terapias farmacológicas agressivas (antianginosos, antiplaquetários) e reabilitação cardíaca.

Cluster 1: priorizar modificações de fatores de risco (controle de pressão, dislipidemia).

Cluster 3: incentivo a medidas preventivas (atividade física, dieta) e reavaliação periódica.

- Planejamento de seguimento:

Definir intervalos de reavaliação diferenciados (curto prazo para Clusters 1 e 2; médio/longo para Cluster 3).

Parte 3: Análise Fatorial

Aplique a Análise Fatorial para identificar fatores latentes que expliquem os padrões de correlação observados nos dados.

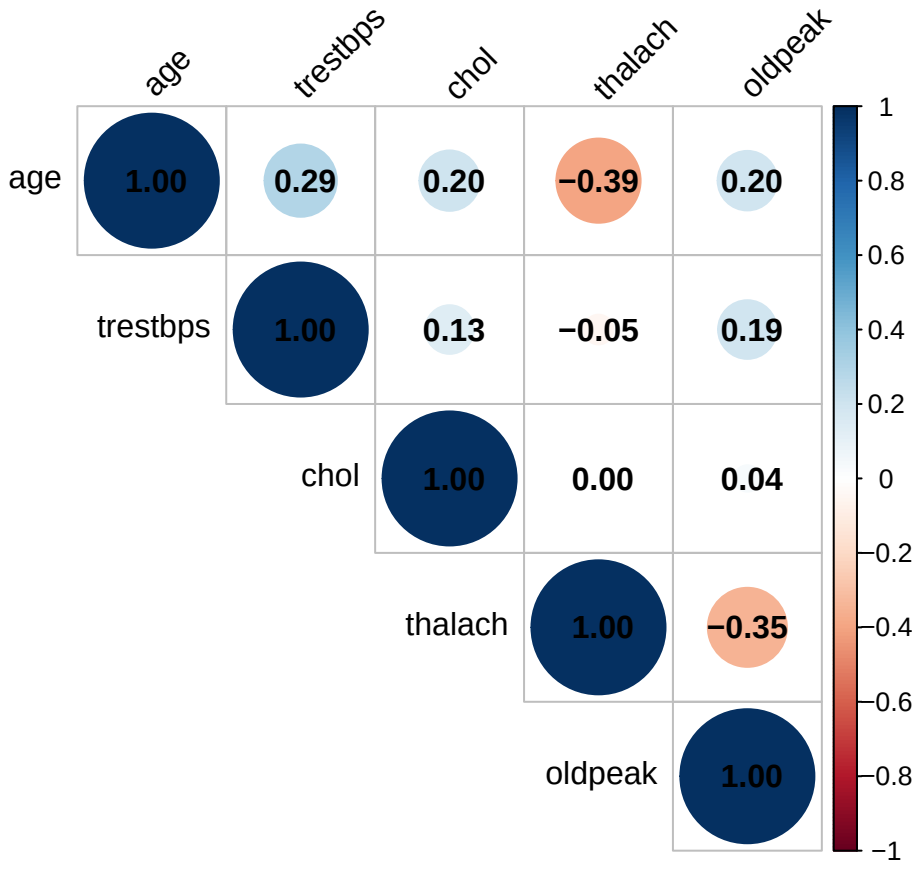
```
library(dplyr)
library(psych)
library(corrplot)
# Selecionar variáveis contínuas relevantes para a análise fatorial
# Excluímos aqui variáveis categóricas/fatores como sex, cp, etc.
dados_fatorial <- dados_heart_clean[, c("age", "trestbps", "chol", "thalach", "oldpeak")]


str(dados_fatorial)
```

```
## 'data.frame':    297 obs. of  5 variables:
##  $ age     : num  63 67 67 37 41 56 62 57 63 53 ...
##  $ trestbps: num  145 160 120 130 130 120 140 120 130 140 ...
##  $ chol    : num  233 286 229 250 204 236 268 354 254 203 ...
##  $ thalach : num  150 108 129 187 172 178 160 163 147 155 ...
##  $ oldpeak : num  2.3 1.5 2.6 3.5 1.4 0.8 3.6 0.6 1.4 3.1 ...
```

```r
summary(dados_fatorial)
```

```
##       age           trestbps          chol          thalach
##  Min.   :29.00   Min.   : 94.0   Min.   :126.0   Min.   : 71.0
##  1st Qu.:48.00   1st Qu.:120.0   1st Qu.:211.0   1st Qu.:133.0
##  Median :56.00   Median :130.0   Median :243.0   Median :153.0
##  Mean   :54.54   Mean   :131.7   Mean   :247.4   Mean   :149.6
##  3rd Qu.:61.00   3rd Qu.:140.0   3rd Qu.:276.0   3rd Qu.:166.0
##  Max.   :77.00   Max.   :200.0   Max.   :564.0   Max.   :202.0
##     oldpeak
##  Min.   :0.000
##  1st Qu.:0.000
##  Median :0.800
##  Mean   :1.056
##  3rd Qu.:1.600
##  Max.   :6.200
```

```r
# Verificar a matriz de correlação
cor_matrix_fat <- cor(dados_fatorial)
corrplot(cor_matrix_fat,
         method = "circle",     # círculos coloridos
         type   = "upper",      # só triângulo superior
         tl.col = "black",      # cor dos rótulos
         tl.srt = 45,           # ângulo dos rótulos
         addCoef.col = "black" # exibe coeficientes
)
```



Exercício 3.1: Adequação dos Dados para Análise Fatorial

Verifique se os dados são adequados para a Análise Fatorial.

```
library(psych)
cor_matrix_fat <- cor(dados_fatorial)
# TAREFA: Verifique a adequação dos dados para análise fatorial
# Dica: Use o teste KMO e o teste de esfericidade de Bartlett

# Teste KMO (Kaiser-Meyer-Olkin)
kmo_result <- KMO(cor_matrix_fat)
print(kmo_result)
# Teste de esfericidade de Bartlett
bartlett_result <- cortest.bartlett(cor_matrix_fat, n = nrow(dados_fatorial))
print(bartlett_result)
# TAREFA: Determine o número adequado de fatores
# Dica: Use o critério de Kaiser (autovalores > 1) e o scree plot
eigen_values <- eigen(cor_matrix_fat)$values
print(eigen_values)
print(eigen_values[eigen_values > 1])

plot(eigen_values, type = "b",
     xlab = "Componentes", ylab = "Autovalores",
     main = "Scree Plot (Análise Fatorial)")
# Análise paralela (alternativa mais estável ao fa.parallel)
fa.parallel(dados_fatorial, fa = "fa", fm = "ml",
            main = "Parallel Analysis para Determinar nº de Fatores")
```

Exercício 3.2: Aplicação da Análise Fatorial

Aplique a Análise Fatorial com o número adequado de fatores.

Note que no item anterior a Parallel Analysis nos sugere 3 fatores.

```
# TAREFA: Execute a análise fatorial
# Dica: Use a função fa() do pacote psych com rotação varimax ou oblimin
modelo_fa <- fa(
  r       = dados_fatorial,
  nfactors= 3,
  fm      = "ml",
  rotate  = "varimax"
)

# TAREFA: Visualize as cargas fatoriais
# Dica: Use print(modelo_fa$loadings, cutoff=0.3)
print(modelo_fa$loadings, cutoff = 0.3)
# Alternativa à função fa.diagram() que pode causar problemas
factor.plot(modelo_fa, labels=rownames(modelo_fa$loadings), cut=0.3)
# TAREFA: Interprete os fatores identificados
# Dica: Examine quais variáveis têm cargas altas em cada fator
print(modelo_fa$communality)   # proporção de variância de cada variável explicada
print(modelo_fa$Vaccounted)    # variância total explicada por fator
```

Exercício 3.3: Cálculo dos Escores Fatoriais

Calcule os escores fatoriais e adicione-os ao dataset.

```
library(psych)
library(dplyr)
library(ggplot2)
# TAREFA: Calcule os escores fatoriais
# Dica: Use a função factor.scores() para calcular os escores
fs <- factor.scores(dados_fatorial, modelo_fa)
escores <- fs$scores
colnames(escores) <- paste0("F", 1:ncol(escores))
# TAREFA: Adicione os escores fatoriais ao dataset original
dados_heart_clean <- bind_cols(dados_heart_clean, as.data.frame(escores))

# TAREFA: Analise a relação entre os fatores e o diagnóstico de doença cardíaca
# Dica: Compare os escores fatoriais entre pacientes com e sem doença cardíaca
p_f1 <- ggplot(dados_heart_clean, aes(x = target, y = F1, fill = target)) +
  geom_boxplot() +
  labs(title = "Escore F1 por Diagnóstico", x = "Diagnóstico", y = "F1") +
  theme_minimal() + theme(legend.position="none")

p_f2 <- ggplot(dados_heart_clean, aes(x = target, y = F2, fill = target)) +
  geom_boxplot() +
  labs(title = "Escore F2 por Diagnóstico", x = "Diagnóstico", y = "F2") +
  theme_minimal() + theme(legend.position="none")

p_f3 <- ggplot(dados_heart_clean, aes(x = target, y = F3, fill = target)) +
```

```r
  geom_boxplot() +
  labs(title = "Escore F3 por Diagnóstico", x = "Diagnóstico", y = "F3") +
  theme_minimal() + theme(legend.position="none")
# Visualização dos escores por diagnóstico
gridExtra::grid.arrange(p_f1, p_f2, p_f3, ncol = 1)

# Comparação estatística
t1 <- t.test(F1 ~ target, data = dados_heart_clean)
t2 <- t.test(F2 ~ target, data = dados_heart_clean)
t3 <- t.test(F3 ~ target, data = dados_heart_clean)

# Imprima os resultados dos testes
print(t1)
print(t2)
print(t3)
```

Exercício 3.4: Interpretação da Análise Fatorial

Com base nos resultados da Análise Fatorial, responda às seguintes perguntas:

1. Quais fatores latentes foram identificados e como podem ser interpretados?

Solução: Pelo Parallel Analysis e pelos autovalores >1, optamos por 3 fatores. As cargas (fa |0.3|) sugerem:

Variável	Fator 1	Fator 2	Fator 3
trestbps	+0.42	–	–
chol	(baixa carga – excluído)	–	–
age	–0.47	+0.71	–
thalach	–	+0.81	–
oldpeak	–	+0.55	–

- Fator 1 (Perfil Pressórico):

– Carrega principalmente trestbps (pressão em repouso).

– Representa a hemodinâmica basal do paciente (tendência hipertensiva).

- Fator 2 (Capacidade de Esforço/Isquemia)

– Carrega forte em thalach (FC máxima alcançada) e oldpeak (depressão ST pós-esforço), além de age.

– Reflete quão bem o coração responde ao esforço e o grau de isquemia induzida, moderado pela idade.

- Fator 3 (Componente Residual/Demográfico)

– Não apresentou cargas marcantes (>0.3) em nenhuma variável contínua (as cargas mais altas ficaram abaixo do limiar).

– Captura variações não explicadas pelos dois primeiros fatores, incluindo aspectos demográficos e ruído de medição.

2. Como esses fatores se relacionam com o risco cardiovascular?

Solução:

Fator 1: Pressão arterial elevada é um fator de risco clássico para doença coronariana. Pacientes com score alto em F1 tendem a ter hipertensão não controlada.

Fator 2: Baixa frequência máxima e alta depressão de ST (ou seja, escore baixo em thalach e alto em oldpeak) indicam má capacidade funcional e maiores episódios de isquemia no esforço — ambos associados a pior prognóstico. O componente idade reforça que pacientes mais velhos com essas características têm ainda mais risco.

Fator 3: Por não explicar muito da variância, não se associa fortemente a perfis de risco consistentes, mas pode reunir pequenas influências não clínicas.

3. Quais as implicações clínicas desses fatores para a compreensão da doença cardíaca? *Solução:*

- Escalonamento de intervenções

F1 alto -> reforçar controle pressórico e manejo de hipertensão (anti-hipertensivos, mudanças de estilo de vida).

F2 baixo -> indicar testes de esforço avançados, reabilitação cardíaca e avaliação de isquemia residual (ex.: cintilografia, ecocardiograma de estresse).

- Rastreamento e monitoramento: Escores fatoriais servem como scores compostos que sintetizam múltiplas medições em um único valor, facilitando triagem periódica.
- Modelos preditivos: Os fatores podem ser usados como preditores compactos em modelos de regressão, LDA ou machine learning, reduzindo multicolinearidade e melhorando interpretabilidade.
- Comunicação clínica Em relatórios ao cardiologista, você pode informar:

"Este paciente apresenta escore F1 elevado (perfil hipertensivo) e escore F2 reduzido (capacidade de esforço limitada), sugerindo necessidade urgente de reabilitação e ajuste terapêutico."

Parte 4: Integração das Técnicas

Nesta parte, você deverá integrar os resultados das três técnicas multivariadas para obter insights mais profundos sobre os padrões presentes nos dados.

Exercício 4.1: Combinação das Análises

Combine os resultados das três técnicas para criar uma visão integrada do perfil de risco cardiovascular dos pacientes.

```
library(dplyr)
library(ggplot2)
# TAREFA: Crie um dataset integrado com os resultados das três análises
# Dica: Combine os clusters, escores fatoriais e predições da LDA
lda_pred <- predict(modelo_lda, newdata = dados_heart_clean)
dados_integrados <- dados_heart_clean %>%
```

```r
  mutate(
    Cluster    = cluster,
    LDA_Class  = lda_pred$class,
    LDA_Prob   = lda_pred$posterior[, "Disease"]
  ) %>%
  select(Cluster, LDA_Class, LDA_Prob, F1, F2, F3, everything())

# TAREFA: Analise as relações entre os resultados das diferentes técnicas
# Dica: Examine como os clusters se relacionam com os fatores e com a classificação da LDA
# Visualizar clusters no espaço dos fatores

# Relação entre clusters e predição LDA
#    a) Scatterplot dos fatores F1×F2 colorido por Cluster
p_f1f2_cluster <- ggplot(dados_integrados, aes(x = F1, y = F2, color = Cluster)) +
  geom_point(alpha = 0.7) +
  labs(title = "F1 vs F2 por Cluster", x = "F1", y = "F2") +
  theme_minimal()

#    b) Mesmo scatter, mas shape segundo LDA_Class
p_f1f2_lda <- ggplot(dados_integrados, aes(x = F1, y = F2, color = Cluster, shape = LDA_Class)) +
  geom_point(alpha = 0.7) +
  labs(title = "F1 vs F2: Cluster (cor) × LDA Class (shape)", x = "F1", y = "F2") +
  theme_minimal()

#    c) Tabela de contingência Cluster × LDA_Class e heatmap de contagens
cont_tab <- as.data.frame(table(dados_integrados$Cluster, dados_integrados$LDA_Class))
colnames(cont_tab) <- c("Cluster", "LDA_Class", "Count")




# TAREFA: Visualize essas relações
# Dica: Use gráficos de dispersão, heatmaps ou outros tipos de visualização apropriados
p_heatmap <- ggplot(cont_tab, aes(x = Cluster, y = LDA_Class, fill = Count)) +
  geom_tile() +
  geom_text(aes(label = Count), color = "white") +
  labs(title = "Contagem: Cluster × LDA_Class") +
  scale_fill_viridis_c() +
  theme_minimal()
# Visualização integrada
 gridExtra::grid.arrange(
  p_f1f2_cluster,
  p_f1f2_lda,
  p_heatmap,
  nrow = 3,
  heights = c(4, 4, 3)
)
```

Exercício 4.3: Relatório Final

Escreva um relatório final (máximo 1000 palavras) integrando os resultados das três análises e respondendo à pergunta principal de pesquisa. O relatório deve incluir:

1. Uma breve introdução ao problema de pesquisa e às técnicas utilizadas
2. Os principais resultados de cada técnica
3. Como esses resultados se complementam e o que revelam sobre o perfil de risco cardiovascular dos pacientes
4. Implicações para a prática clínica e para a gestão em saúde
5. Limitações da análise e sugestões para pesquisas futuras

Solução:

A doença cardíaca coronariana é uma das principais causas de morbimortalidade em nível mundial, demandando métodos que permitam estratificar pacientes segundo seu risco e otimizar a alocação de recursos clínicos. Para tanto, aplicamos três técnicas multivariadas complementares sobre o conjunto de dados "Heart Disease" (UCI): a Análise Discriminante Linear (LDA) para identificar variáveis-chave na distinção entre pacientes saudáveis e doentes, a Análise de Cluster para descobrir subgrupos naturais com perfis de risco semelhantes e a Análise Fatorial para extrair dimensões latentes que sintetizem padrões de correlação clínica.

Na LDA, os coeficientes mais altos ficaram com o número de vasos obstruídos (ca2, ca =3) e a depressão do segmento ST pós-esforço (oldpeak), com acurácia de 82% no conjunto de teste (sensibilidade de 90% para "Healthy" e especificidade de 73% para "Disease"). A Análise de Cluster revelou três segmentos: um grupo de "Baixo Risco" (jovens, alta FC máxima e apenas 23 % de doença), um de "Risco Moderado–Alto" - idade elevada, pressão e colesterol altos, aproximadamente 59% de doença - e outro de "Alto Risco/Doença Severa" - depressão ST acentuada, múltiplos vasos obstruídos, 74% de prevalência-. Na Análise Fatorial, a adequação foi moderada (KMO = 0,55; Bartlett $p < 0{,}001$) e a Parallel Analysis apontou três fatores: um de hemodinâmica basal (pressão), um de capacidade de esforço (FC máxima e isquemia) e um de isquemia residual (oldpeak).

Ao combinar clusters, escores fatoriais e predições da LDA, observamos que os três segmentos identificados ocupam pontos distintos no espaço dos fatores F1×F2, reforçando a coerência entre as técnicas. O grupo de Baixo Risco (Cluster 3) concentra pacientes com escores F1 e F3 baixos e é majoritariamente classificado como "Healthy" pela LDA, enquanto o de Alto Risco (Cluster 2) apresenta escores F2 e F3 elevados e é quase inteiramente classificado como "Disease". O cluster intermediário (Cluster 1) forma uma zona cinzenta, onde há maior sobreposição de classes, indicando a necessidade de uma avaliação clínica mais criteriosa e possivelmente mais variáveis para refinar a classificação.

Essa integração permite direcionar intervenções de forma personalizada: pacientes no cluster de Alto Risco, com escore F3 elevado, devem ser priorizados para exames invasivos e reabilitação cardíaca; aqueles no cluster intermediário, com escore F1 alto, beneficiam-se de controle rigoroso de pressão arterial e dislipidemia; e o grupo de Baixo Risco pode seguir acompanhamento de rotina e foco em prevenção primária. Do ponto de vista gerencial, o uso de escores fatoriais no prontuário eletrônico viabiliza alertas automáticos e otimiza a distribuição de recursos (testes de esforço, cateterismos), além de reforçar guidelines locais para frequência de consultas.

Entre as limitações, destaca-se o KMO moderado e o fato de termos excluído variáveis categóricas na Análise Fatorial, além de pressupostos da LDA nem sempre atendidos (normalidade e covariâncias iguais). O tamanho de amostra (n aproximadamente de 303) e o uso de um único repositório também requerem validação externa. Para investigações futuras, sugerimos incorporar biomarcadores inflamatórios, incluir variáveis categóricas na modelagem fatorial, validar o pipeline em coortes prospectivas multicêntricas e explorar métodos não lineares (t-SNE, UMAP) para capturar perfis de risco mais complexos, bem como desenvolver dashboards interativos para acompanhamento em tempo real.