Análise do comportamento de alunos

Arthur T. L. de Santana

2025-04-03

Contents

Análise do comportamento de alunos e aplicaçao de árvore de decisao.

Introduçao

Este relatório apresenta a análise dos dados de alunos das classes de matemática e portugues de duas escolas distinas, combinando-as, explorando a estrutura dos dados e aplicando um modelo de arvore de decisao para identificar fatores que influenciam a intenção de cursar ensino superior. As variáveis contidas nos dataframes sao;

- **school** - Escola, Binária.
- **sex** - Sexo, Binária.
- **age** - Idade, Numérica.
- **address** - Endereço, Binária.
- **famsize** - Tamanho da familia, Binária.
- **Pstatus** - Se os pais vivem junsot ou separados, Binária.
- **Medu** - Educaçao da mae, Numérica.
- **Fedu** - Educaçao do pai, Numerico.
- **Mjob** - Trabalho da mae, Nominal.
- **Fjob** - Trabalho do pai, Nominal.
- **reason** - Razao para escolher essa escola, Nominal.
- **guardian** - Quem é o guardiao da criança, Nominal.
- **traveltime** - Tempo de viagem de casa até a escola, Numerica.
- **studytime** - Quantas horas por semana estuda. Numerica.
- **failures** - Numero de vezes que reprovou, Numerica.
- **schoolsup** - Se tem suporte educacional fora da escola, Binária.
- **famsup** - Se a familia é um apoio educacional, Binária.
- **paid** - Se faz banca (Math or Portuguese), Binária.
- **activities** - Se realiza atividades extracurriculares, Binária.
- **nursery** - Se ja foi na enfermaria da escola, Binaria.
- **higher** - Se quer cursar ensino superior, Binaria.
- **internet** - Tem acesso a internet, Binaria.
- **romantic** - Se esta namorando, Binaria.
- **famrel** - Qualidade do relacionamento com a familia, Numerica.
- **freetime** - Tempo livre depois da escola, Numerica.
- **goout** - Sai com amigos, Numerica.
- **Dalc** - Se consome alcool na semana, Numerica.
- **Walc** - Se consome alcool no final de semana, Numerica.
- **health** - Estado atual da saude, Numerica.
- **absences** - Numero de faltas, Numerica.
- **G1** - Nota do primeiro periodo, Numerica.
- **G2** - Nota do segundo periodo, Numerica.
- **G3** - Nota do ultimo periodo, Numerica

Análise Exploratória

- Carregando os pacotes que vamos usar, upload dos datasets, juntando os dois e excluindo alunos que estao cadastrados em ambos os bancos e verificaçao de valores nulos.

O dataset tinha 0 valores nulos e depois do tratamento passou a ter 669 linhas.

Estatística descritiva

Table 1: Resumo Estatístico das Variáveis Numéricas

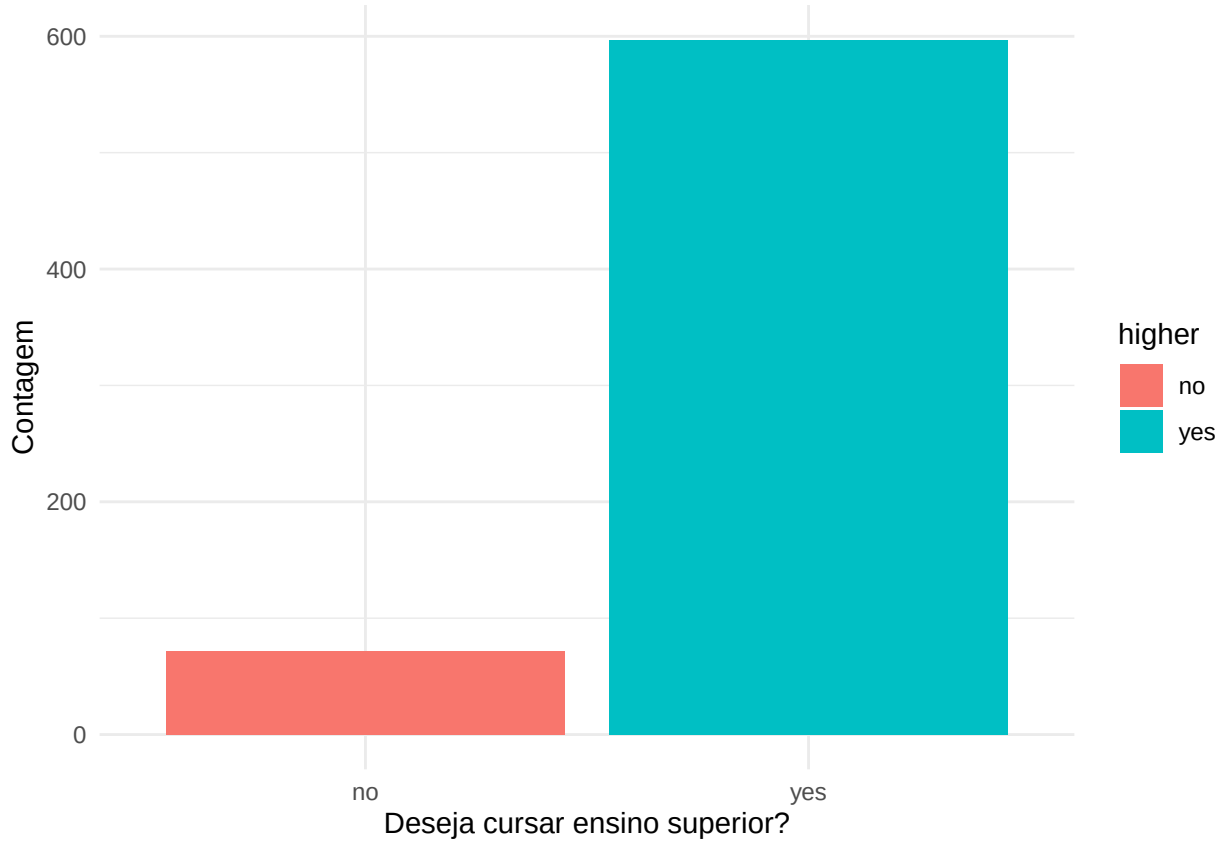
age	Min. :15.00	1st Qu.:16.00	Median :17.00	Mean :16.81	3rd Qu.:18.00	Max. :22.00
Medu	Min. :0.000	1st Qu.:2.000	Median :2.000	Mean :2.495	3rd Qu.:4.000	Max. :4.000
Fedu	Min. :0.000	1st Qu.:1.000	Median :2.000	Mean :2.294	3rd Qu.:3.000	Max. :4.000
traveltime	Min. :1.000	1st Qu.:1.000	Median :1.000	Mean :1.564	3rd Qu.:2.000	Max. :4.000
studytime	Min. :1.000	1st Qu.:1.000	Median :2.000	Mean :1.928	3rd Qu.:2.000	Max. :4.000
failures	Min. :0.0000	1st Qu.:0.0000	Median :0.0000	Mean :0.3318	3rd Qu.:0.0000	Max. :3.0000
famrel	Min. :1.000	1st Qu.:4.000	Median :4.000	Mean :3.934	3rd Qu.:5.000	Max. :5.000
freetime	Min. :1.000	1st Qu.:3.000	Median :3.000	Mean :3.182	3rd Qu.:4.000	Max. :5.000
goout	Min. :1.000	1st Qu.:2.000	Median :3.000	Mean :3.179	3rd Qu.:4.000	Max. :5.000
Dalc	Min. :1.000	1st Qu.:1.000	Median :1.000	Mean :1.502	3rd Qu.:2.000	Max. :5.000
Walc	Min. :1.000	1st Qu.:1.000	Median :2.000	Mean :2.278	3rd Qu.:3.000	Max. :5.000
health	Min. :1.000	1st Qu.:2.000	Median :4.000	Mean :3.531	3rd Qu.:5.000	Max. :5.000
absences	Min. : 0.000	1st Qu.: 0.000	Median : 3.000	Mean : 4.889	3rd Qu.: 7.000	Max. :75.000
G1	Min. : 3.00	1st Qu.: 8.00	Median :10.00	Mean :10.71	3rd Qu.:13.00	Max. :19.00
G2	Min. : 0.00	1st Qu.: 9.00	Median :11.00	Mean :10.68	3rd Qu.:13.00	Max. :19.00
G3	Min. : 0.00	1st Qu.: 9.00	Median :11.00	Mean :10.68	3rd Qu.:13.00	Max. :20.00

Table 2: Resumo Estatístico das Variáveis Binárias

schoolsup	famsup	paid	activities	nursery	higher	internet	romantic
no :599	no :263	no :475	no :348	no :136	no : 72	no :158	no :419
yes: 70	yes:406	yes:194	yes:321	yes:533	yes:597	yes:511	yes:250

Visualizações

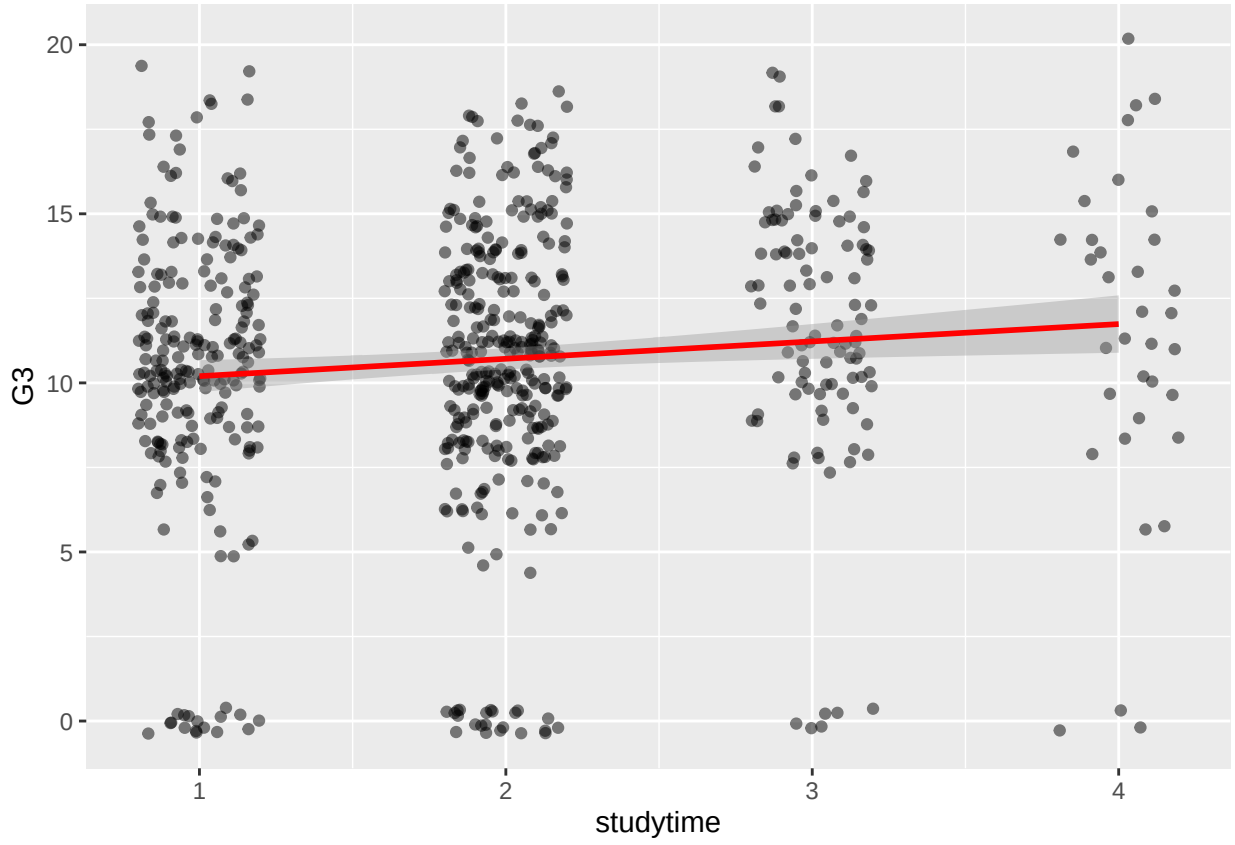
Gráfico 1



Intençao de Cursar Ensino superior

Como podemos observar os dados na nossa variável resposta estao bastante desbalanceados, vamos observar como o algoritmo se comporta antes de balancear

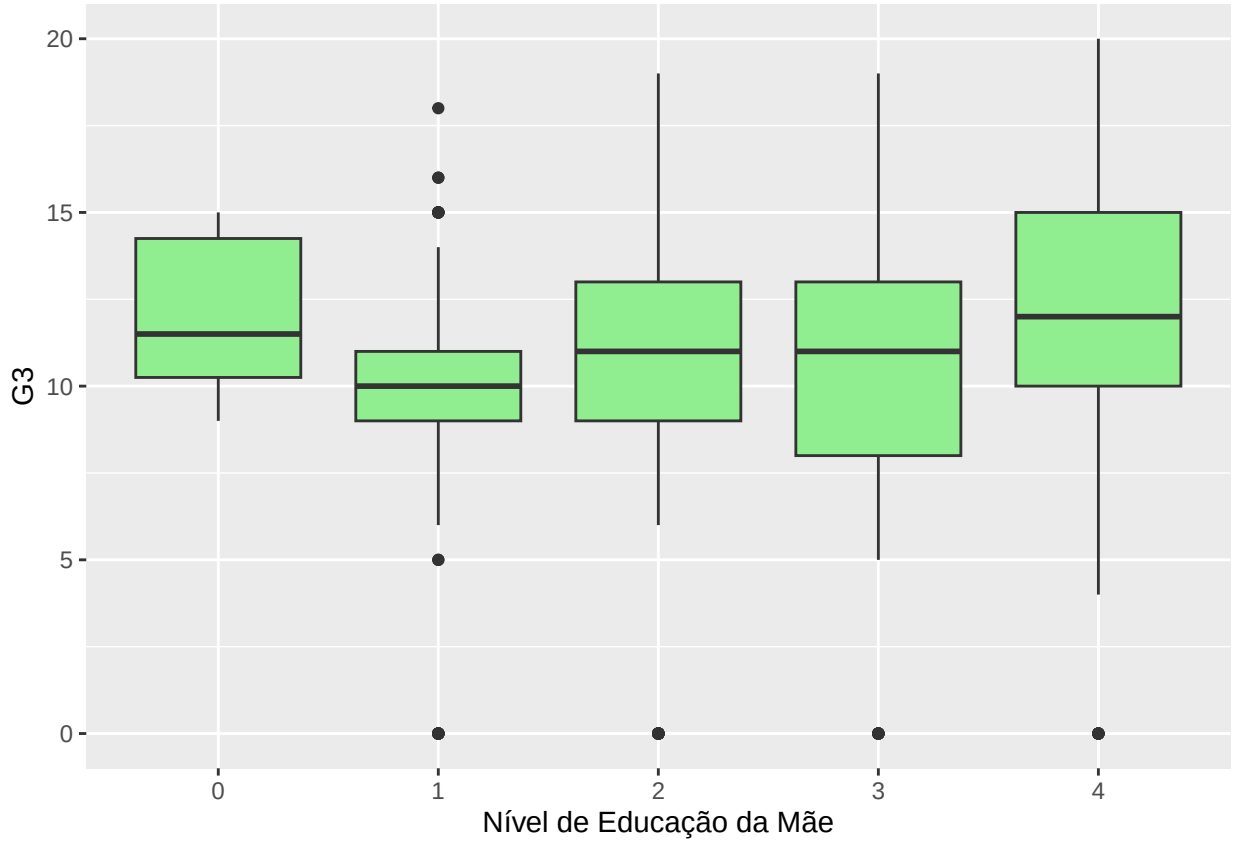
Gráfico 2



Relação entre Tempo de Estudo e Notas

Pouquissimos alunos tiraram notas entre 0 e 5, existe uma leve correlaçao positiva entre tempo de estudo e notas, porém é interessante notas notas que alguns alunos afirmaram estudar 4 ou mais horas por semana e ainda assim tiraram 0, oque nos leva a questionar a veracidade dos das respostas.

Gráfico 3



Notas por Educação da Mãe

Conseguimos observar que a média de notas é maior se a escolaridade da mae for básica ou inferior, e essa média é ultrapassada apenas em filhos de maes com pós graduaçao.

Modelo de Classificação

Call: C5.0.formula(formula = higher ~ ., data = train)

C5.0 [Release 2.07 GPL Edition] Thu Apr 3 00:50:45 2025

Class specified by attribute 'outcome'

Read 468 cases (33 attributes) from undefined.data

Decision tree:

paid = yes: yes (136/1) paid = no: :...schoolsup = yes: yes (33) schoolsup = no: :...age <= 17: yes (199/16) age > 17: :...G1 <= 6: no (12/2) G1 > 6: :...freetime <= 3: yes (58/7) freetime > 3: :...Mjob in {at_home,health,teacher}: no (8/1) Mjob in {other,services}: yes (22/6)

Evaluation on training data (468 cases):

```
    Decision Tree
  ----------------
  Size      Errors

     7   33( 7.1%)   <<


   (a)   (b)    <-classified as
  ----  ----
    17    30    (a): class no
     3   418    (b): class yes


Attribute usage:

100.00% paid
 70.94% schoolsup
 63.89% age
 21.37% G1
 18.80% freetime
  6.41% Mjob
```

Time: 0.0 secs

Avaliação do Modelo

Confusion Matrix and Statistics

```
      Reference
```

Prediction no yes no 2 6 yes 23 170

```
           Accuracy : 0.8557
             95% CI : (0.7994, 0.9012)
No Information Rate : 0.8756
P-Value [Acc > NIR] : 0.832649

              Kappa : 0.0648
```

Mcnemar's Test P-Value : 0.002967

```
        Sensitivity : 0.08000
        Specificity : 0.96591
     Pos Pred Value : 0.25000
     Neg Pred Value : 0.88083
         Prevalence : 0.12438
     Detection Rate : 0.00995
```

Detection Prevalence : 0.03980
Balanced Accuracy : 0.52295

```
   'Positive' Class : no
```

Balanceamento dos Dados

- Como vimos no gráfico 1, os dados estao um tanto desbalanceados, isso fez com que o nosso modelo atingisse uma acurácia de 85% e tendo um desempenho muito bom para identificar pessoas que querem cursar o ensino superior mas péssimo para prever alunos que nao irao, vamos balancea-los e aplicar novamente o algoritmo usando os dados de teste originais para avalia-lo e ver se conseguimos um resultado melhor

yes no 597 601

Modelo de Classificação

- Com os dados balanceados, vamos tentar aplicar o modelo novamente

Call: C5.0.formula(formula = higher ~ ., data = train2)

C5.0 [Release 2.07 GPL Edition] Thu Apr 3 00:50:45 2025

Class specified by attribute 'outcome'

Read 838 cases (33 attributes) from undefined.data

Decision tree:

G2 > 13: yes (97) G2 <= 13: :... paid = yes: :... address = U: yes (87) : address = R: : :... guardian = mother: yes (22) : guardian in {father,other}: : :... Fedu <= 2: no (18) : Fedu > 2: yes (4) paid = no: :... Medu > 3: :... G2 <= 6: no (13/1) : G2 > 6: yes (34) Medu <= 3: :... schoolsup = yes: :... Dalc > 3: no (10) : Dalc <= 3: : :... freetime <= 4: yes (24) : freetime > 4: : :... Mjob in {at_home,health,other,teacher}: no (4) : Mjob = services: yes (1) schoolsup = no: :... G3 > 12: :... reason in {course,home}: yes (17) : reason in {other,reputation}: : :... famsize = LE3: yes (4) : famsize = GT3: : :... health <= 1: yes (2) : health > 1: : :... age <= 15: yes (2) : age > 15: no (18/1) G3 <= 12: :... age <= 17: :... G2 <= 5: yes (14) : G2 > 5: : :... studytime > 1: : :... goout <= 4: : : :... studytime <= 3: yes (38) : : : studytime > 3: no (4) : : goout > 4: : : :... reason in {home,other, : : : reputation}: yes (3) : : reason = course: : : :... famsup = no: no (20) : : famsup = yes: : : :... guardian in {father, : : : other}: no (8) : : guardian = mother: yes (4) : studytime <= 1: : :... reason = reputation: yes (2) : reason in {course,home,other}: : :... guardian = other: yes (1) : guardian = father: : :... internet = no: yes (3) : : internet = yes: : : :... Mjob in {at_home,health,other, : : : teacher}: no (39/1) : : Mjob = services: yes (2) : guardian = mother: : :... internet = no: : :... Medu > 1: no (54/2) : : Medu <= 1: : : :... age <= 15: no (6) : : age > 15: yes (3) : internet = yes: : :... school = MS: yes (10) : school = GP: : :... traveltime > 2: yes (3) : traveltime <= 2: : :... Walc <= 3: no (24/2) : Walc > 3: yes (2) age > 17: :... freetime > 3: no (131/4) freetime <= 3: :... Dalc > 1: :... failures <= 0: yes (4) : failures > 0: no (53/1) Dalc <= 1: :... sex = M: yes (7) sex = F: :... Mjob in {health,teacher}: no (0) Mjob = services: yes (6) Mjob in {at_home,other}: :... traveltime > 2: yes (4) traveltime <= 2: :... traveltime <= 1: :... age <= 18: yes (5) : age > 18: [S1] traveltime > 1: :... Medu > 1: no (18) Medu <= 1: [S2]

SubTree [S1]

Mjob = at_home: no (5) Mjob = other: yes (1)

SubTree [S2]

Fjob in {at_home,health,services,teacher}: no (5) Fjob = other: yes (2)

Evaluation on training data (838 cases):

```
    Decision Tree
  ----------------
  Size      Errors

    47   12( 1.4%)   <<


   (a)   (b)    <-classified as
  ----  ----
   408    12    (a): class yes
         418    (b): class no


Attribute usage:

100.00% G2
 88.42% paid
 72.79% Medu
 67.18% schoolsup
 62.53% G3
 59.79% age
 32.22% freetime
 27.09% reason
 26.97% studytime
 24.22% guardian
 17.78% Dalc
 17.42% internet
 15.63% address
 10.98% Mjob
  9.19% goout
  8.23% traveltime
  6.80% failures
  6.32% sex
  4.65% school
  3.82% famsup
  3.10% famsize
  3.10% Walc
  2.63% Fedu
  2.63% health
  0.84% Fjob
```

Time: 0.0 secs

Avaliação do Modelo

Confusion Matrix and Statistics

```
      Reference
```

Prediction no yes no 25 12 yes 0 164

```
          Accuracy : 0.9403
            95% CI : (0.898, 0.9688)
No Information Rate : 0.8756
P-Value [Acc > NIR] : 0.001919

             Kappa : 0.7727
```

Mcnemar's Test P-Value : 0.001496

```
        Sensitivity : 1.0000
        Specificity : 0.9318
     Pos Pred Value : 0.6757
     Neg Pred Value : 1.0000
         Prevalence : 0.1244
     Detection Rate : 0.1244
```

Detection Prevalence : 0.1841
Balanced Accuracy : 0.9659

```
   'Positive' Class : no
```

Conclusão

- Além da nota da segunda e terceira unidade, a escolaridade da mae e se o aluno faz aulas extras, pagas ou nao, foram os fatores mais determinantes da árvore, acima de 60%. 0 modelo conseguiu atingir uma acurácia de 94% para prever a intenção de cursar ensino superior. Os resultados obtidos podem ser utilizados para melhorar a compreensão dos fatores que influenciam o desempenho acadêmico e as decisões educacionais dos alunos.

Código utilizado

```
#Carregando pacotes
library(tidyverse)
library(C50)
library(gmodels)
library(ROSE)
library(caret)
library(knitr)
library(kableExtra)

#carregando Dataframes
df1 <- read.table("C:/Users/Usuario/Documents/UFS/Matérias/2 periodo/Introdução ao software R/Trabalho I
df2 <- read.table("C:/Users/Usuario/Documents/UFS/Matérias/2 periodo/Introdução ao software R/Trabalho I

#Juntando os dois dataframes em um só eliminando alunos que apareciam nos
chaves <- c("school", "sex", "age", "address", "famsize", "Pstatus",
            "Medu", "Fedu", "Mjob", "Fjob", "reason", "guardian")
df_final <- df2 %>%
  anti_join(df1, by = chaves) %>%
  bind_rows(df1)

cat(sprintf("O dataset tinha %d valores nulos e depois do tratamento passou a ter %d linhas.",
            sum(is.na(df_final)), nrow(df_final)))

#Estatística descritiva
numericas <- df_final %>% select(where(is.numeric))
binarias <- df_final %>% select(where(~ all(. %in% c(0, 1, "yes", "no"))))
summary_transposed <- t(summary(numericas))

kable(summary_transposed, caption = "Resumo Estatístico das Variáveis Numéricas", format = "latex", bool
  kable_styling(latex_options = c("scale_down", "repeat_header"), full_width = FALSE, bootstrap_options

kable(summary(binarias), caption = "Resumo Estatístico das Variáveis Binárias")

#Gráfico 1
ggplot(df_final, aes(x = higher, fill = higher)) +
  geom_bar() +
  labs(x = "Deseja cursar ensino superior?",
       y = "Contagem") +
  theme_minimal()

#Gráfico 2
ggplot(df_final, aes(x = studytime, y = G3)) +
  geom_jitter(alpha = 0.5, width = 0.2) +
  geom_smooth(method = "lm", color = "red")

#Gráfico 3
ggplot(df_final, aes(x = factor(Medu), y = G3)) +
  geom_boxplot(fill = "lightgreen") +
  labs(x = "Nível de Educação da Mãe")

#Criando modelo
set.seed(123)
```

```r
index <- sample(1:nrow(df_final), 0.7 * nrow(df_final))
train <- df_final[index, ]
test <- df_final[-index, ]

modelo <- C5.0(higher ~ ., data = train)
summary(modelo)

#Avaliando modelo
pred <- predict(modelo, test)
conf_matrix <- confusionMatrix(pred, test$higher)
print(conf_matrix)

#Balanceando o modelo utilizando o pacote caret
df_balanced <- ovun.sample(higher ~ ., data = df_final, method = "over")$data
table(df_balanced$higher)

#Criando modelo 2
set.seed(123)
index <- sample(1:nrow(df_balanced), 0.7 * nrow(df_balanced))
train2 <- df_balanced[index, ]
test2 <- df_balanced[-index, ]

modelo2 <- C5.0(higher ~ ., data = train2)
summary(modelo2)

#Avaliando modelo 2

pred <- predict(modelo2, test)
conf_matrix2 <- confusionMatrix(pred, test$higher)
print(conf_matrix2)
```