Analytics & Inteligência Artificial

Tema da aula

Árvore de Decisão e Teste Qui-quadrado

FIA Business School

NOSSOS DIFERENCIAIS | QUEM SOMOS

BUSINESS SCHOOL
Graduação, pós-graduação, MBA, Pós- MBA, Mestrado Profissional, Curso In *Company* e EAD

CONSULTING
Consultoria personalizada que oferece soluções baseada em seu problema de negócio

RESEARCH
Atualização dos conhecimentos e do material didático oferecidos nas atividades de ensino

FUNDAÇÃO INSTITUTO DE ADMINISTRAÇÃO

Líder em Educação Executiva, referência de ensino nos cursos de graduação, pós-graduação e MBA, tendo excelência nos programas de educação. Uma das principais escolas de negócio do mundo, possuindo convênios internacionais com Universidades nos EUA, Europa e Ásia. +8.000 projetos de consultorias em organizações públicas e privadas.

- Único curso de graduação em administração a receber as notas máximas
- A primeira escola brasileira a ser finalista da maior competição de MBA do mundo
- Única *Business School* brasileira a figurar no *ranking* LATAM
- Signatária do Pacto Global da ONU
- Membro fundador da ANAMBA - Associação Nacional MBAs
- Credenciada pela AMBA - *Association of MBAs*
- Credenciada ao *Executive MBA Council*
- Filiada a AACSB - *Association to Advance Collegiate Schools of Business*
- Filiada a EFMD - *European Foundation for Management Development*
- Referência em cursos de MBA nas principais mídias de circulação

O **Laboratório de Análise de Dados** – LABDATA é um Centro de Excelência que atua nas áreas de ensino, pesquisa e consultoria em análise de informação utilizando técnicas de ***Big Data***, ***Analytics*** e **Inteligência Artificial**.


Profª Drª Alessandra Montini

O LABDATA é um dos pioneiros no lançamento dos cursos de *Big Data* e *Analytics* no Brasil

Os diretores foram professores de grandes especialistas do mercado

+10 anos de atuação

+1000 alunos formados

Docentes

- Sólida formação acadêmica: doutores e mestres em sua maioria
- Larga experiência de mercado na resolução de *cases*
- Participação em Congressos Nacionais e Internacionais
- Professor assistente que acompanha o aluno durante todo o curso

Estrutura

- 100% das aulas realizadas em laboratórios
- Computadores para uso individual durante as aulas
- 5 laboratórios de alta qualidade (investimento +R$2MM)
- 2 Unidades próximas a estação de metrô (com estacionamento)

Conteúdo da Aula

1. Introdução
2. Teste Qui-Quadrado
3. Árvore de Decisão
4. Case

1. Introdução

Case: *Churn* em Telefonia

Exemplo
Identificar o perfil dos clientes que cancelam suas planos de telefonia, para desenvolver ações proativas de engajamento e relacionamento.

Aplicação
Segmento Telecom

Case: *People Analytics* - RH

Exemplo
Identificar o perfil dos funcionários que receberam promoção no ano anterior, e avaliar quais características se destacam em relação àqueles que não receberam.

Aplicação
Gestão de Pessoas

Case: Finanças e Economia

Exemplo
Identificar o perfil das empresas que faliram com base em variáveis: tempo de empresa, porte, quantidade de funcionários, ramo de atuação, região, etc.

Aplicação
Finanças e Economia

Case: Hábitos Alimentares

Exemplo
Identificar os hábitos de dietas alimentares dos países que possuem taxa de mortalidade abaixo da média mundial.

Aplicação
Áreas de Saúde e Nutrição

Case: Perfil de Pacientes Crônicos

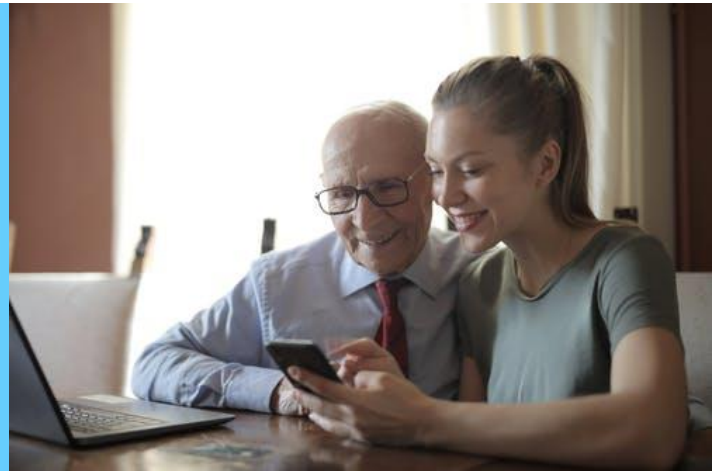
Exemplo
Identificar o perfil dos pacientes diabéticos, segundo seu hábito de vida: dieta, exercícios, se é fumante, idade, sexo, peso, altura etc.

Aplicação
Área Médica e de Nutrição

Case: CRM

Exemplo
Identificar o perfil dos clientes que realizam transações digitais, a fim de delinear ações de incentivo ao uso do canal para os que ainda não utilizam.

Aplicação
Área de Marketing e Comunicação



Case: Prospecção

Exemplo
Identificar o perfil de clientes que contratam um determinado serviço, dado um *mailing* de nomes comprados no mercado para a oferta do produto por Telemarketing Ativo.

Aplicação
Área de Marketing

Evento Binário

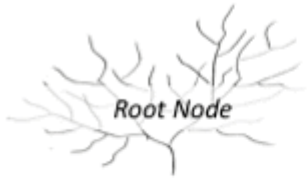
Uma característica comum dos cases apresentados anteriormente é que todos possuem um **evento de interesse**, e este evento é **binário**:

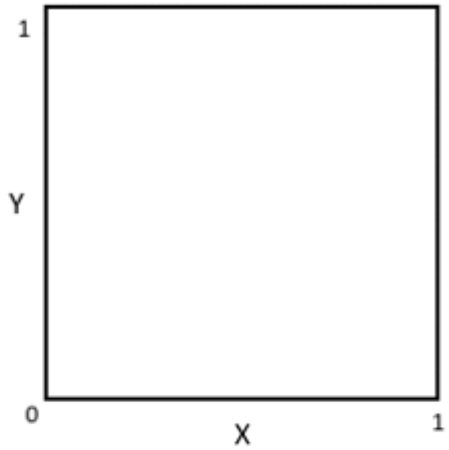
1 – apresentou o evento de interesse;
0 – não apresentou o evento de interesse.

Objetivo

A **Árvore de Decisão** é uma ferramenta muito utilizada para o apoio à tomada de decisão.

De forma visual, é possível encontrar a melhor **partição** da base, definida a partir de um conjunto de variáveis explicativas. Estes subgrupos são discriminados em função de um evento (variável resposta), de forma iterativa, até que um critério de parada seja satisfeito.





For more tutorials: annalyzin.wordpress.com

https://algobeans.com/2016/07/27/decision-trees-tutorial/

Uma *startup* especializada no ramo de Serviços possui dois perfis de profissionais: Tecnologia e Administrativo (RH, Financeiro, Comercial, etc). A empresa gostaria de conhecer o perfil dos funcionários de Tecnologia com base nas informações cadastradas no sistema de RH da empresa. Quais características explicam as diferenças entre os profissionais de tecnologia e os profissionais de áreas administrativas?

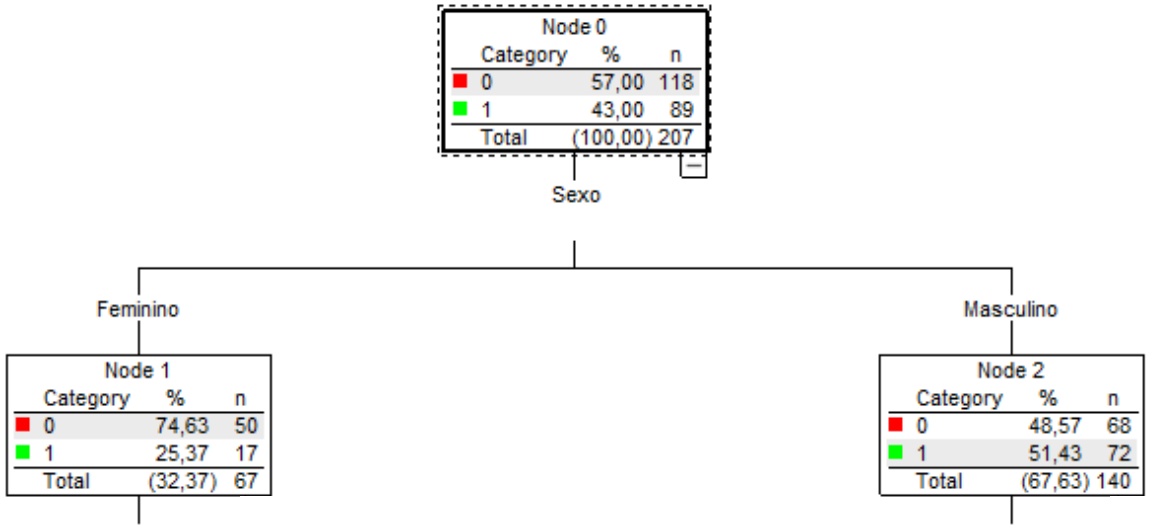
Evento (variável resposta):

- 1, caso o funcionário seja da área de Tecnologia;
- 0, caso o funcionário não seja da área de Tecnologia.

Variáveis explicativas:

- Idade
- Sexo
- Escolaridade
- Estado civil
- Cidade de nascimento

A Árvore de Decisão inicia-se com a variável resposta, e mostra a proporção do evento de interesse (0 ou 1). O algoritmo seleciona, entre todas as variáveis explicativas, aquela que melhor discrimina a variável resposta, segundo um critério de partição.



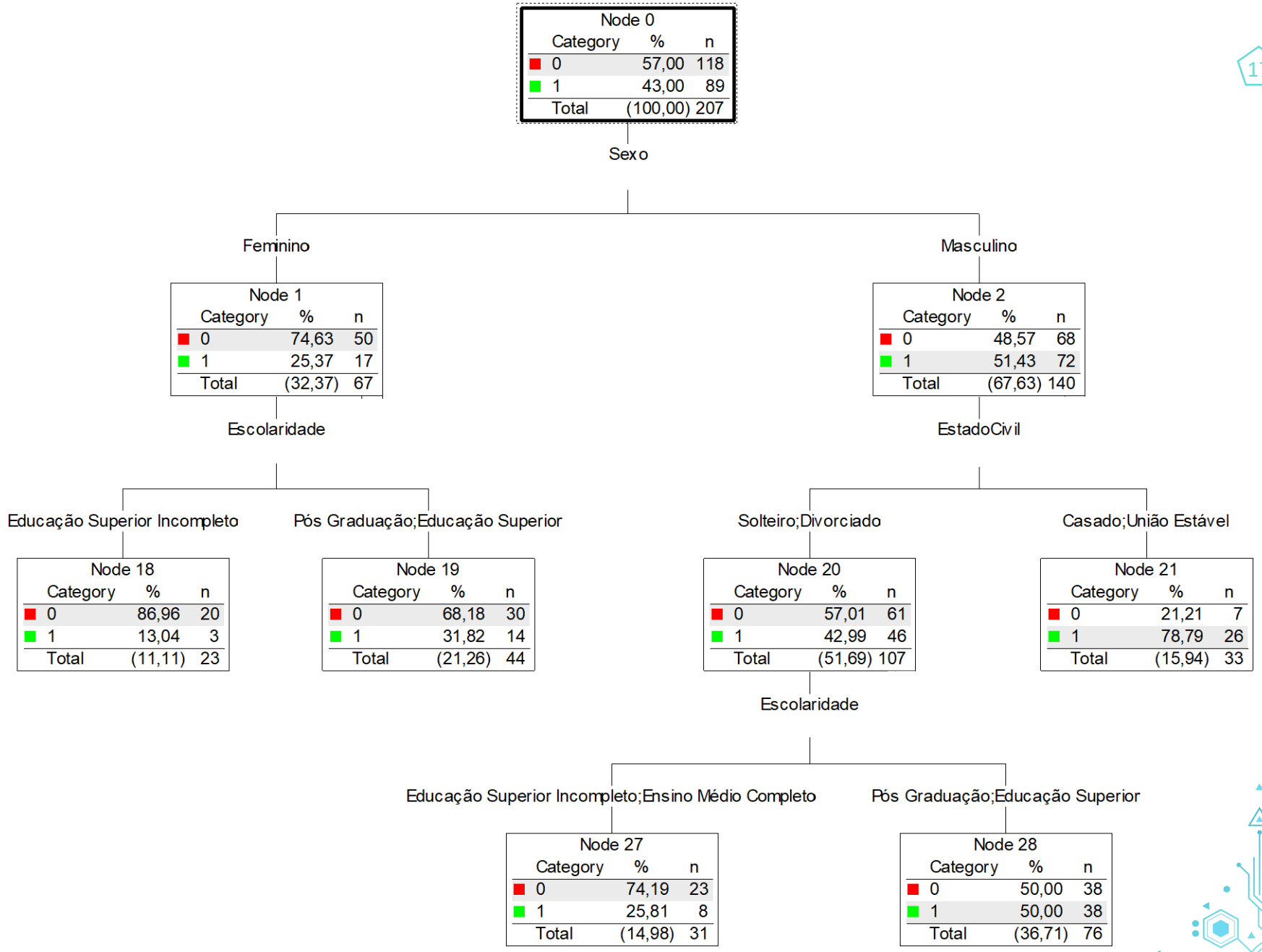
Interpretação: A variável resposta apresenta 43% de funcionários da área de Tecnologia, entre o total de 207 funcionários.

Interpretação: Se isolarmos todos os funcionários do sexo masculino (67,6%) e selecionarmos aleatoriamente 1 funcionário, a probabilidade de ele ser da área de Tecnologia passa para a ser 51,4%.

Case: *People Analytics*

Após as quebras definidas no nível anterior da árvore, o algoritmo escolhe, entre as demais variáveis restantes, aquela qual é mais discriminante.

O processo continua de forma **iterativa**, até que um determinado critério de parada seja satisfeito.



Os nós finais podem ser classificados em: **propensos** ao evento de interesse (verde) e **não propensos** (vermelho).

O corte entre propensos e não propensos é dado pela proporção do evento resposta da base inicial (nó 0). Neste exemplo, 43%.

Case: *People Analytics*

Neste exemplo, foram encontrados **5 perfis**, sendo 2 propensos a serem de Tecnologia e 3 propensos às demais áreas administrativas.

Note que as variáveis *Idade* e *Cidade de Nascimento* não foram discriminantes no modelo final.

Exercício: *People Analytics*

Com base no modelo de árvore de decisão apresentado ao lado:

a) Qual a probabilidade do perfil mais propenso à ser de Tecnologia? Qual a representatividade deste perfil na empresa?

b) Qual o perfil com menor probabilidade de ser da área de Tecnologia?

Como a Árvore de Decisão é construída?

Uma vez entendendo o objetivo e como a árvore de decisão é interpretada, é importante entender como o algoritmo funciona.

Na literatura, são apresentadas algumas variações de critérios de partição e escolha de variáveis explicativas. Na aula de hoje, aprenderemos sobre o **método CHAID** (*Chi-Square Automatic Interaction Detection*), que é baseado no Teste Estatístico de Qui-Quadrado.

O tipo de variável resposta a ser estudada será uma resposta binária (0 ou 1), apesar de o método permitir o uso de variável resposta qualitativa com 3 ou mais categorias.

2. Teste Qui-Quadrado

Testar a hipótese de associação entre duas variáveis qualitativas

Ainda no case de *People Analytics*, pode-se avaliar se o *sexo* (feminino ou masculino) tem associação com o *tipo de departamento* (tecnologia ou administrativo). Para isso, poderíamos utilizar como critério de avaliação o **teste qui-quadrado**.

H_0 : Não existe associação entre 'sexo' e 'tipo de departamento'.
H_1 : Existe associação entre 'sexo' e 'tipo de departamento'.

Case: Estudo de doenças cardiovasculares

Um grupo de pesquisa de estudos cardiovasculares gostaria de investigar se existe relação entre o consumo de bebida alcoólica e pressão arterial. Para isso, realizou um estudo com 183 pacientes. Avaliou-se a frequência de consumo de bebida alcoólica, em 3 categorias (não consome; 1x/semana; 2x/semana ou mais), bem como o nível pressão, também em 3 categorias (baixa, normal e alta). Os dados são apresentados a seguir.

Tabela de frequências absolutas (**observada**):

Consumo de bebida alcoólica	**Pressão Arterial**			**Total**
	Baixa	Normal	Alta	
Não consome	20	21	11	**52**
1x/semana	21	19	13	**53**
2x/semana ou mais	21	22	35	**78**
Total	**62**	**62**	**59**	**183**

O primeiro passo é obter a tabela com os valores esperados.

Case: Estudo de doenças cardiovasculares

Um grupo de pesquisa de estudos cardiovasculares gostaria de investigar se existe relação entre o consumo de bebida alcoólica e pressão arterial. Para isso, realizou um estudo com 183 pacientes. Avaliou-se a frequência de consumo de bebida alcoólica, em 3 categorias (não consome; 1x/semana; 2x/semana ou mais), bem como o nível pressão, também em 3 categorias (baixa, normal e alta). Os dados são apresentados a seguir.

Tabela de frequências absolutas (**observada**):

Consumo de bebida alcoólica	**Pressão Arterial**			**Total**
	Baixa	Normal	Alta	
Não consome	20	21	11	**52**
1x/semana	21	19	13	**53**
2x/semana ou mais	21	22	35	**78**
Total	**62**	**62**	**59**	**183**

Tabela de frequências absolutas (**esperada**):

Consumo de bebida alcoólica	**Pressão Arterial**			**Total**
	Baixa	Normal	Alta	
Não consome	17,62	17,62	16,77	**52**
1x/semana	17,96	17,96	17,09	**53**
2x/semana ou mais	26,43	26,43	25,15	**78**
Total	**62**	**62**	**59**	**183**

O primeiro passo é obter a tabela com os valores esperados.

Sob a hipótese de independência, o **valor esperado é dado por** $\frac{62 * 78}{183} = 26{,}43$

Case: Estudo de doenças cardiovasculares

Tabela de frequências absolutas (**observada**):

Consumo de bebida alcoólica	**Pressão Arterial**			**Total**
	Baixa	Normal	Alta	
Não consome	20	21	11	**52**
1x/semana	21	19	13	**53**
2x/semana ou mais	21	22	35	**78**
Total	**62**	**62**	**59**	**183**

Tabela de frequências absolutas (**esperada**):

Consumo de bebida alcoólica	**Pressão Arterial**			**Total**
	Baixa	Normal	Alta	
Não consome	17,62	17,62	16,77	**52**
1x/semana	17,96	17,96	17,09	**53**
2x/semana ou mais	26,43	26,43	25,15	**78**
Total	**62**	**62**	**59**	**183**

$$\chi^2 = \sum_i \sum_j \frac{(O_{ij} - E_{ij})^2}{E_{ij}} =$$

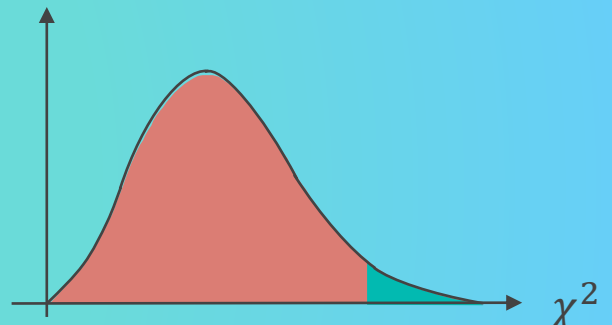
$$\frac{(20 - 17{,}62)^2}{17{,}62} + \frac{(21 - 17{,}62)^2}{17{,}62} + \ldots + \frac{(35 - 25{,}15)^2}{25{,}15} = 10{,}22$$

Este valor deve ser utilizado no **teste de hipótese qui-quadrado**

Distribuição Qui-Quadrado

O objetivo do teste qui-quadrado é verificar se **existe associação** entre duas variáveis qualitativas, com as seguintes hipóteses:

H_0: Não existe associação entre as variáveis;
H_1: Existe associação entre as variáveis.



Distribuição Qui-Quadrado

Estatística do teste

$$\chi^2 = \sum_i \sum_j \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Em que:

O_{ij} : frequência observada da categoria *ij*

E_{ij} : frequência esperada da categoria *ij*, dada por:

$$E_{ij} = \frac{(Total\ da\ linha\ i)(Total\ da\ coluna\ j)}{n}$$

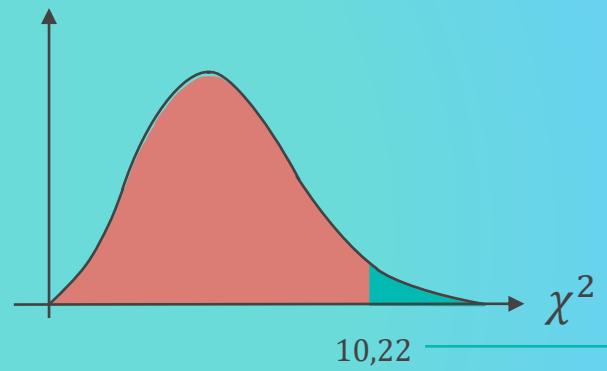
onde *n* é o tamanho amostral.

Distribuição Qui-Quadrado

Estatística do teste

$$\chi^2 = \sum_i \sum_j \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Distribuição Qui-Quadrado

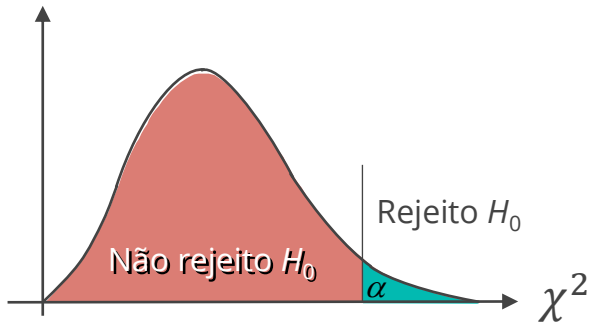


O valor 10,22, para um qui-quadrado com 4 graus de liberdade, fornece um nível descritivo (p-valor) de **0,0368**, que é a área abaixo da curva a partir do valor 10,22.

A estatística do teste segue a distribuição **qui-quadrado** com **(n-1)*(m-1) graus de liberdade**, sendo **n** o número de categorias da variável nas linhas e **m** é o número de categorias da variável nas colunas.

Regra de Rejeição

Distribuição Qui-Quadrado



Regra de rejeição

Pelo valor crítico: Rejeito H_0 se $\chi^2 \geq \chi^2{}_\alpha$

Pelo *p*-valor: Rejeito H_0 se *p*-valor $\leq \alpha$

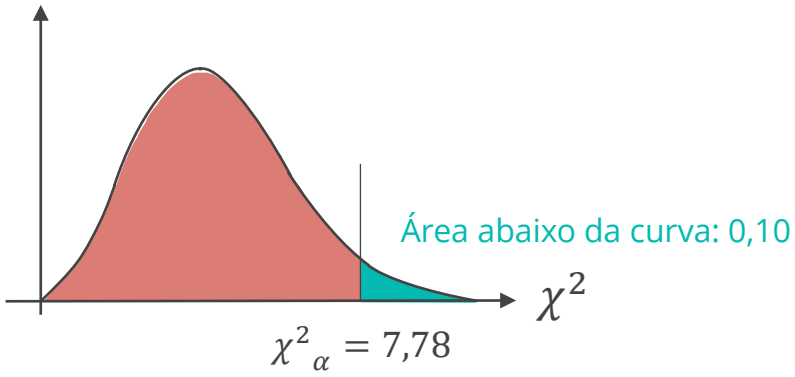
Em que α é o nível de significância e *k* são os graus de liberdade da estatística de Qui-quadrado.

Chi-square Distribution Table

d.f.	.995	.99	.975	.95	.9	.1	.05	.025	.01
1	0.00	0.00	0.00	0.00	0.02	2.71	3.84	5.02	6.63
2	0.01	0.02	0.05	0.10	0.21	4.61	5.99	7.38	9.21
3	0.07	0.11	0.22	0.35	0.58	6.25	7.81	9.35	11.34
4	0.21	0.30	0.48	0.71	1.06	7.78	9.49	11.14	13.28
5	0.41	0.55	0.83	1.15	1.61	9.24	11.07	12.83	15.09
6	0.68	0.87	1.24	1.64	2.20	10.64	12.59	14.45	16.81
7	0.99	1.24	1.69	2.17	2.83	12.02	14.07	16.01	18.48
8	1.34	1.65	2.18	2.73	3.49	13.36	15.51	17.53	20.09
9	1.73	2.09	2.70	3.33	4.17	14.68	16.92	19.02	21.67
10	2.16	2.56	3.25	3.94	4.87	15.99	18.31	20.48	23.21
11	2.60	3.05	3.82	4.57	5.58	17.28	19.68	21.92	24.72
12	3.07	3.57	4.40	5.23	6.30	18.55	21.03	23.34	26.22
13	3.57	4.11	5.01	5.89	7.04	19.81	22.36	24.74	27.69
14	4.07	4.66	5.63	6.57	7.79	21.06	23.68	26.12	29.14
15	4.60	5.23	6.26	7.26	8.55	22.31	25.00	27.49	30.58
16	5.14	5.81	6.91	7.96	9.31	23.54	26.30	28.85	32.00
17	5.70	6.41	7.56	8.67	10.09	24.77	27.59	30.19	33.41
18	6.26	7.01	8.23	9.39	10.86	25.99	28.87	31.53	34.81
19	6.84	7.63	8.91	10.12	11.65	27.20	30.14	32.85	36.19
20	7.43	8.26	9.59	10.85	12.44	28.41	31.41	34.17	37.57
22	8.64	9.54	10.98	12.34	14.04	30.81	33.92	36.78	40.29
24	9.89	10.86	12.40	13.85	15.66	33.20	36.42	39.36	42.98
26	11.16	12.20	13.84	15.38	17.29	35.56	38.89	41.92	45.64
28	12.46	13.56	15.31	16.93	18.94	37.92	41.34	44.46	48.28
30	13.79	14.95	16.79	18.49	20.60	40.26	43.77	46.98	50.89
32	15.13	16.36	18.29	20.07	22.27	42.58	46.19	49.48	53.49
34	16.50	17.79	19.81	21.66	23.95	44.90	48.60	51.97	56.06
38	19.29	20.69	22.88	24.88	27.34	49.51	53.38	56.90	61.16
42	22.14	23.65	26.00	28.14	30.77	54.09	58.12	61.78	66.21
46	25.04	26.66	29.16	31.44	34.22	58.64	62.83	66.62	71.20
50	27.99	29.71	32.36	34.76	37.69	63.17	67.50	71.42	76.15
55	31.73	33.57	36.40	38.96	42.06	68.80	73.31	77.38	82.29
60	35.53	37.48	40.48	43.19	46.46	74.40	79.08	83.30	88.38
65	39.38	41.44	44.60	47.45	50.88	79.97	84.82	89.18	94.42
70	43.28	45.44	48.76	51.74	55.33	85.53	90.53	95.02	100.43
75	47.21	49.48	52.94	56.05	59.79	91.06	96.22	100.84	106.39
80	51.17	53.54	57.15	60.39	64.28	96.58	101.88	106.63	112.33
85	55.17	57.63	61.39	64.75	68.78	102.08	107.52	112.39	118.24
90	59.20	61.75	65.65	69.13	73.29	107.57	113.15	118.14	124.12
95	63.25	65.90	69.92	73.52	77.82	113.04	118.75	123.86	129.97
100	67.33	70.06	74.22	77.93	82.36	118.50	124.34	129.56	135.81

Regra de Rejeição

Distribuição Qui-Quadrado (4 g.l.)



Decisão do teste (considerando 90 % de confiança)

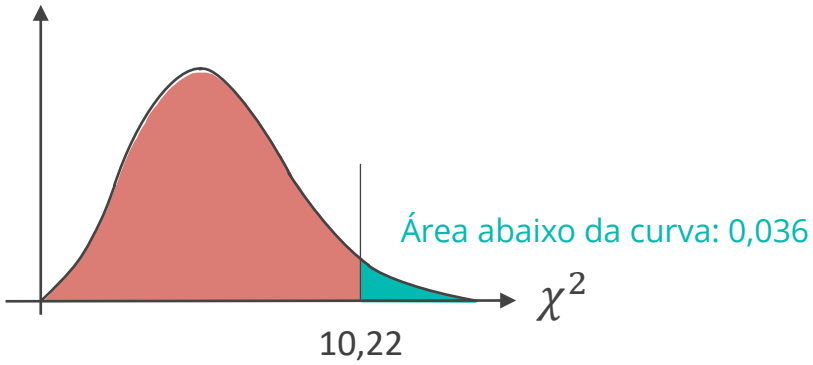
Como $\chi^2 = 10{,}22 > \chi^2{}_\alpha = 7{,}78$, rejeitamos H_0.
Ou seja, existe associação entre consumo de bebida alcoólica e pressão arterial.

Chi-square Distribution Table

d.f.	.995	.99	.975	.95	.9	.1	.05	.025	.01
1	0.00	0.00	0.00	0.00	0.02	2.71	3.84	5.02	6.63
2	0.01	0.02	0.05	0.10	0.21	4.61	5.99	7.38	9.21
3	0.07	0.11	0.22	0.35	0.58	6.25	7.81	9.35	11.34
4	0.21	0.30	0.48	0.71	1.06	7.78	9.49	11.14	13.28
5	0.41	0.55	0.83	1.15	1.61	9.24	11.07	12.83	15.09
6	0.68	0.87	1.24	1.64	2.20	10.64	12.59	14.45	16.81
7	0.99	1.24	1.69	2.17	2.83	12.02	14.07	16.01	18.48
8	1.34	1.65	2.18	2.73	3.49	13.36	15.51	17.53	20.09
9	1.73	2.09	2.70	3.33	4.17	14.68	16.92	19.02	21.67
10	2.16	2.56	3.25	3.94	4.87	15.99	18.31	20.48	23.21
11	2.60	3.05	3.82	4.57	5.58	17.28	19.68	21.92	24.72
12	3.07	3.57	4.40	5.23	6.30	18.55	21.03	23.34	26.22
13	3.57	4.11	5.01	5.89	7.04	19.81	22.36	24.74	27.69
14	4.07	4.66	5.63	6.57	7.79	21.06	23.68	26.12	29.14
15	4.60	5.23	6.26	7.26	8.55	22.31	25.00	27.49	30.58
16	5.14	5.81	6.91	7.96	9.31	23.54	26.30	28.85	32.00
17	5.70	6.41	7.56	8.67	10.09	24.77	27.59	30.19	33.41
18	6.26	7.01	8.23	9.39	10.86	25.99	28.87	31.53	34.81
19	6.84	7.63	8.91	10.12	11.65	27.20	30.14	32.85	36.19
20	7.43	8.26	9.59	10.85	12.44	28.41	31.41	34.17	37.57
22	8.64	9.54	10.98	12.34	14.04	30.81	33.92	36.78	40.29
24	9.89	10.86	12.40	13.85	15.66	33.20	36.42	39.36	42.98
26	11.16	12.20	13.84	15.38	17.29	35.56	38.89	41.92	45.64
28	12.46	13.56	15.31	16.93	18.94	37.92	41.34	44.46	48.28
30	13.79	14.95	16.79	18.49	20.60	40.26	43.77	46.98	50.89
32	15.13	16.36	18.29	20.07	22.27	42.58	46.19	49.48	53.49
34	16.50	17.79	19.81	21.66	23.95	44.90	48.60	51.97	56.06
38	19.29	20.69	22.88	24.88	27.34	49.51	53.38	56.90	61.16
42	22.14	23.65	26.00	28.14	30.77	54.09	58.12	61.78	66.21
46	25.04	26.66	29.16	31.44	34.22	58.64	62.83	66.62	71.20
50	27.99	29.71	32.36	34.76	37.69	63.17	67.50	71.42	76.15
55	31.73	33.57	36.40	38.96	42.06	68.80	73.31	77.38	82.29
60	35.53	37.48	40.48	43.19	46.46	74.40	79.08	83.30	88.38
65	39.38	41.44	44.60	47.45	50.88	79.97	84.82	89.18	94.42
70	43.28	45.44	48.76	51.74	55.33	85.53	90.53	95.02	100.43
75	47.21	49.48	52.94	56.05	59.79	91.06	96.22	100.84	106.39
80	51.17	53.54	57.15	60.39	64.28	96.58	101.88	106.63	112.33
85	55.17	57.63	61.39	64.75	68.78	102.08	107.52	112.39	118.24
90	59.20	61.75	65.65	69.13	73.29	107.57	113.15	118.14	124.12
95	63.25	65.90	69.92	73.52	77.82	113.04	118.75	123.86	129.97
100	67.33	70.06	74.22	77.93	82.36	118.50	124.34	129.56	135.81

Regra de Rejeição

Distribuição Qui-Quadrado (4 g.l.)



Decisão do teste (considerando 90 % de confiança)

Como o nível descritivo é 0,036 < 0,10, rejeitamos H_0.
Ou seja, existe associação entre consumo de bebida alcoólica e pressão arterial.

Chi-square Distribution Table

d.f.	.995	.99	.975	.95	.9	.1	.05	.025	.01
1	0.00	0.00	0.00	0.00	0.02	2.71	3.84	5.02	6.63
2	0.01	0.02	0.05	0.10	0.21	4.61	5.99	7.38	9.21
3	0.07	0.11	0.22	0.35	0.58	6.25	7.81	9.35	11.34
4	0.21	0.30	0.48	0.71	1.06	7.78	9.49	11.14	13.28
5	0.41	0.55	0.83	1.15	1.61	9.24	11.07	12.83	15.09
6	0.68	0.87	1.24	1.64	2.20	10.64	12.59	14.45	16.81
7	0.99	1.24	1.69	2.17	2.83	12.02	14.07	16.01	18.48
8	1.34	1.65	2.18	2.73	3.49	13.36	15.51	17.53	20.09
9	1.73	2.09	2.70	3.33	4.17	14.68	16.92	19.02	21.67
10	2.16	2.56	3.25	3.94	4.87	15.99	18.31	20.48	23.21
11	2.60	3.05	3.82	4.57	5.58	17.28	19.68	21.92	24.72
12	3.07	3.57	4.40	5.23	6.30	18.55	21.03	23.34	26.22
13	3.57	4.11	5.01	5.89	7.04	19.81	22.36	24.74	27.69
14	4.07	4.66	5.63	6.57	7.79	21.06	23.68	26.12	29.14
15	4.60	5.23	6.26	7.26	8.55	22.31	25.00	27.49	30.58
16	5.14	5.81	6.91	7.96	9.31	23.54	26.30	28.85	32.00
17	5.70	6.41	7.56	8.67	10.09	24.77	27.59	30.19	33.41
18	6.26	7.01	8.23	9.39	10.86	25.99	28.87	31.53	34.81
19	6.84	7.63	8.91	10.12	11.65	27.20	30.14	32.85	36.19
20	7.43	8.26	9.59	10.85	12.44	28.41	31.41	34.17	37.57
22	8.64	9.54	10.98	12.34	14.04	30.81	33.92	36.78	40.29
24	9.89	10.86	12.40	13.85	15.66	33.20	36.42	39.36	42.98
26	11.16	12.20	13.84	15.38	17.29	35.56	38.89	41.92	45.64
28	12.46	13.56	15.31	16.93	18.94	37.92	41.34	44.46	48.28
30	13.79	14.95	16.79	18.49	20.60	40.26	43.77	46.98	50.89
32	15.13	16.36	18.29	20.07	22.27	42.58	46.19	49.48	53.49
34	16.50	17.79	19.81	21.66	23.95	44.90	48.60	51.97	56.06
38	19.29	20.69	22.88	24.88	27.34	49.51	53.38	56.90	61.16
42	22.14	23.65	26.00	28.14	30.77	54.09	58.12	61.78	66.21
46	25.04	26.66	29.16	31.44	34.22	58.64	62.83	66.62	71.20
50	27.99	29.71	32.36	34.76	37.69	63.17	67.50	71.42	76.15
55	31.73	33.57	36.40	38.96	42.06	68.80	73.31	77.38	82.29
60	35.53	37.48	40.48	43.19	46.46	74.40	79.08	83.30	88.38
65	39.38	41.44	44.60	47.45	50.88	79.97	84.82	89.18	94.42
70	43.28	45.44	48.76	51.74	55.33	85.53	90.53	95.02	100.43
75	47.21	49.48	52.94	56.05	59.79	91.06	96.22	100.84	106.39
80	51.17	53.54	57.15	60.39	64.28	96.58	101.88	106.63	112.33
85	55.17	57.63	61.39	64.75	68.78	102.08	107.52	112.39	118.24
90	59.20	61.75	65.65	69.13	73.29	107.57	113.15	118.14	124.12
95	63.25	65.90	69.92	73.52	77.82	113.04	118.75	123.86	129.97
100	67.33	70.06	74.22	77.93	82.36	118.50	124.34	129.56	135.81

Exercício: Campanha de preferência de bebida

Uma empresa de cerveja produz e distribui três tipos de cerveja: *light*, comum e escura. Em uma análise dos segmentos de mercado das três cervejas, a equipe de pesquisa de mercado da empresa levantou a seguinte questão: As preferências pelos 3 tipos de cervejas são diferentes entre os consumidores do sexo feminino e masculino?

Caso a preferência pela cerveja seja independente do sexo do consumidor, será iniciada uma campanha publicitária para todos os consumidores de cerveja. Entretanto, se a preferência pelo tipo de cerveja depender do sexo do consumidor, a empresa modelará suas campanhas de acordo com os diferentes públicos alvo.

150 consumidores de cerveja foram selecionados aleatoriamente, degustaram cada cerveja, e manifestaram sua predileção (primeira escolha). Destes, 80 eram do sexo masculino e 70 do sexo feminino. 50 optaram pela cerveja *light*, 70 pela comum e 30 pela escura. Entre os *lights*, 20 eram homens e 30 mulheres. Entre os que optaram pela cerveja comum, 40 eram homens e 30 eram mulheres.

a) Qual hipótese está sendo testada? Explicite as hipóteses nula (H_0) e alternativa (H_1).
b) Construa a tabela de frequências observada.
c) Construa a tabela de frequências esperada.
d) Calcule a estatística do teste de independência.
e) Quantos graus de liberdade considera-se para a estatística qui-quadrado?
f) Adotando o nível de significância de 10%, aceita-se ou rejeita-se a hipótese nula? Utilize a tabela de qui-quadrado do slide anterior.
g) Qual a conclusão do teste?
h) A empresa deve fazer campanhas diferentes de acordo com o gênero do consumidor?

Exercício: Ouvidoria

Uma empresa do setor financeiro deseja identificar se há associação entre o gênero do consumidor e o fato de ele já ter apresentado reclamação ou não. O objetivo é entender o perfil do consumidor mais propenso à reclamação.

A tabela de valores observados está apresentada abaixo.

OBSERVADO

Sexo	**Reclamação**		**Total**
	Não	Sim	
Feminino	10	93	**103**
Masculino	35	80	**115**
Total	**45**	**173**	**218**

Exercício: Ouvidoria

Uma empresa do setor financeiro deseja identificar se há associação entre o gênero do consumidor e o fato de ele já ter apresentado reclamação ou não. O objetivo é entender o perfil do consumidor mais propenso à reclamação.

Responda as questões abaixo.

a) Qual hipótese está sendo testada? Explicite as hipóteses nula (H_0) e alternativa (H_1).
b) Construa a tabela de frequências esperada.
c) Calcule a estatística do teste de independência.
d) Quantos graus de liberdade considera-se para a estatística qui-quadrado?
e) Adotando o nível de significância de 10%, aceita-se ou rejeita-se a hipótese nula? Utilize a tabela de qui-quadrado do slide anterior.
f) Qual a conclusão do teste?

Exercício: Sinistro de automóveis

Uma seguradora deseja identificar se há associação entre a classe social do segurado e o fato de ele já apresentar ou não sinistro. O objetivo é entender o perfil do consumidor mais propenso à sinistrar para futuramente apoiar a precificação.

A tabela de valores observados está apresentada abaixo.

OBSERVADO

Classe Social	Não	Sim	Total
A	8	10	**18**
B	12	8	**20**
C	8	11	**19**
D + E	12	11	**23**
Total	**40**	**40**	**80**

Exercício: Sinistro de automóveis

Uma seguradora deseja identificar se há associação entre a classe social do segurado e o fato de ele já apresentar ou não sinistro. O objetivo é entender o perfil do consumidor mais propenso à sinistrar para futuramente apoiar a precificação.

Responda as questões abaixo.

a) Qual hipótese está sendo testada? Explicite as hipóteses nula (H_0) e alternativa (H_1).
b) Construa a tabela de frequências esperada.
c) Calcule a estatística do teste de independência.
d) Quantos graus de liberdade considera-se para a estatística qui-quadrado?
e) Adotando o nível de significância de 5%, aceita-se ou rejeita-se a hipótese nula? Utilize a tabela de qui-quadrado do slide anterior.
f) Qual a conclusão do teste?

Exercício: Varejo

Uma varejista deseja identificar se há associação entre o estado em que o cliente reside e o produto comprado. O objetivo é entender o perfil do consumidor.

A tabela de valores observados está apresentada abaixo.

OBSERVADO

Estado	Produto A	Produto B	Total
SP	20	21	**41**
RS	21	19	**40**
SC	21	22	**43**
Total	62	62	124

Exercício: Varejo

Uma varejista deseja identificar se há associação entre o estado em que o cliente reside e o produto comprado. O objetivo é entender o perfil do consumidor.

Responda as questões abaixo.

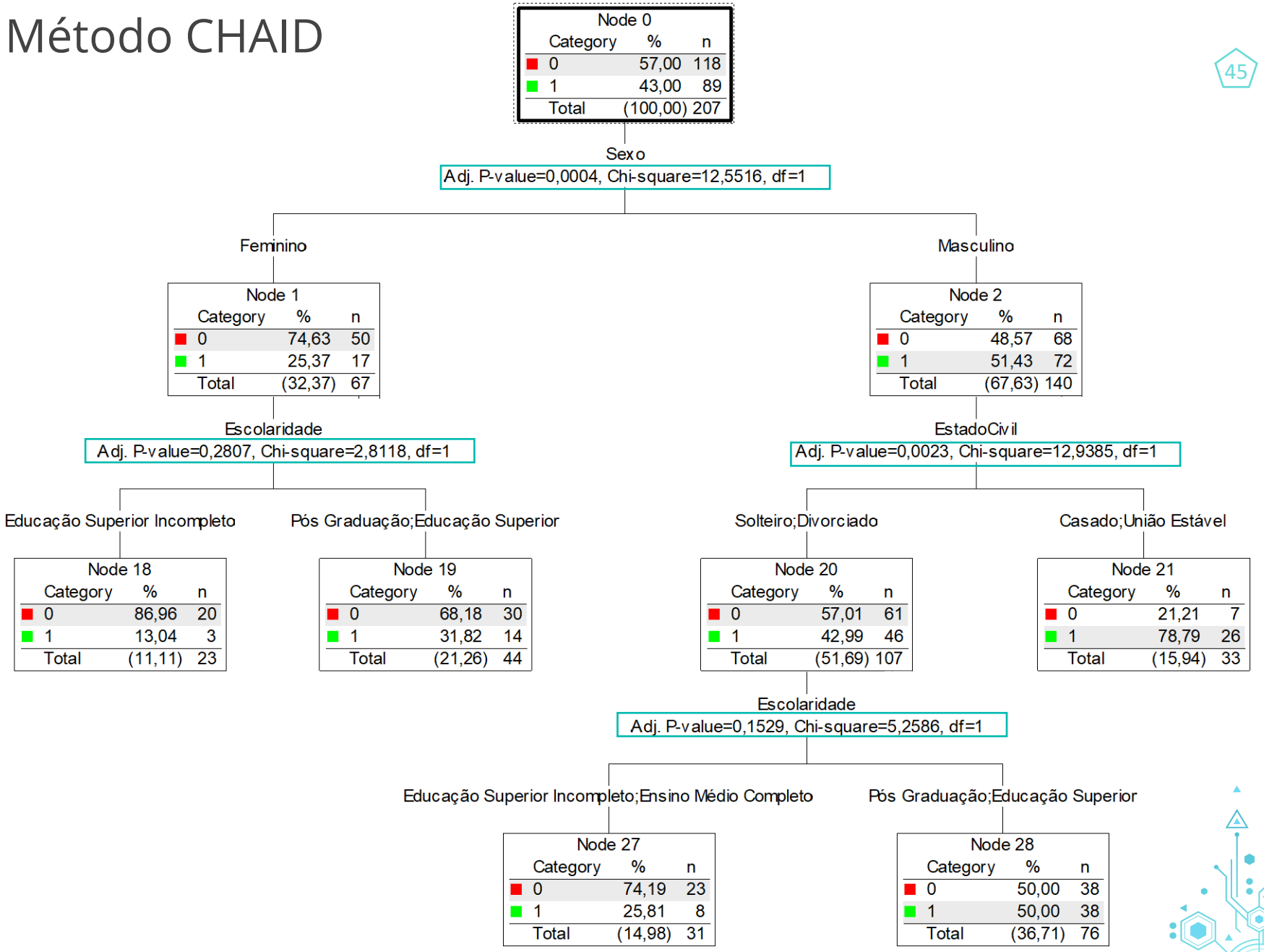
a) Qual hipótese está sendo testada? Explicite as hipóteses nula (H_0) e alternativa (H_1).
b) Construa a tabela de frequências esperada.
c) Calcule a estatística do teste de independência.
d) Quantos graus de liberdade considera-se para a estatística qui-quadrado?
e) Adotando o nível de significância de 5%, aceita-se ou rejeita-se a hipótese nula? Utilize a tabela de qui-quadrado do slide anterior.
f) Qual a conclusão do teste?

3. Árvore de Decisão

Teste Qui-Quadrado e Método CHAID

Pelo método CHAID, é calculada a estatística qui-quadrado. A variável com o menor p-valor (e o maior valor do qui-quadrado) é a variável sugerida para criar a **primeira partição** da base.

Para cada categoria da primeira variável ("quebra"), é escolhida a próxima variável com maior valor de qui-quadrado, até que um critério de parada seja satisfeito.



Como avaliar o desempenho da Árvore de Decisão?

Uma vez que o modelo interpretado faça sentido na visão do negócio, o próximo passo é avaliar o **acerto das regras do modelo**. Isso significa avaliar o resultado predito pelo modelo, em comparação com a resposta observada no passado.

Assim, pode-se ter uma ideia de se o modelo, na presença das variáveis explicativas, é capaz de explicar o evento resposta de forma satisfatória. Dessa forma, as regras do modelo poderão ser aplicadas em novas bases de dados.

Análise de Desempenho

A **tabela de classificação** apresenta o cruzamento da variável resposta observada, em comparação com a variável resposta predita pelo modelo. Ela também é conhecida como **matriz de confusão**.

Um modelo com bom ajuste apresenta grande concentração de casos na diagonal principal.

Tabela de classificação avaliada no ponto de corte:

		Variável resposta predita		
		0	**1**	**Total**
Variável resposta observada	**0**	VN	FP	VN + FP
	1	FN	VP	FN + VP
	Total	VN + FN	FP + VP	VN + FN + FP + VP

VP = verdadeiro positivo; VN = verdadeiro negativo; FP = falso positivo; FN = falso negativo

Análise de Desempenho

Tabela de classificação avaliada no ponto de corte:

		Variável resposta predita		
		0	**1**	**Total**
Variável resposta observada	**0**	VN	FP	VN + FP
	1	FN	VP	FN + VP
	Total	VN + FN	FP + VP	VN + FN + FP + VP

VP = verdadeiro positivo; VN = verdadeiro negativo; FP = falso positivo; FN = falso negativo

Os índices de **acurácia**, **sensibilidade** e **especificidade** variam de 0 a 1 (ou de 0% a 100%).

Valores acima de 50% indicam acerto superior ao aleatório (ausência de modelo). Valores acima de 60% são considerados índices satisfatórios. Já valores acima de 70%-75% indicam ótimo desempenho.

Acurácia

$$Acur = \frac{VP + VN}{VP + VN + FP + FN}$$

Sensibilidade

$$Sensib = \frac{VP}{VP + FN}$$

Especificidade

$$Espec = \frac{VN}{FP + VN}$$

Análise de Desempenho

Tabela de classificação – Case ***People Analytics***:

		Variável resposta predita		
		0	**1**	**Total**
Variável resposta observada	**0**	73	45	118
	1	25	64	89
	Total	98	109	207

Os índices de **acurácia**, **sensibilidade** e **especificidade** apresentaram desempenho satisfatório. De forma geral, é possível predizer o tipo de departamento de quase 70% dos funcionários.

Acurácia

$$Acur = \frac{73 + 64}{207} = 66{,}2\%$$

Sensibilidade

$$Sensib = \frac{64}{89} = 71{,}9\%$$

Especificidade

$$Espec = \frac{73}{118} = 61{,}9\%$$

Case: Pedidos de refeições

Uma empresa de tecnologia, atuante no ramo de entrega de refeições solicitadas *online*, gostaria de entender quais os perfis de consumidores que mais se deixam influenciar pela nota de avaliação dos restaurantes parceiros, no momento de realizarem seu pedido. Para entender esse aspecto, realizaram uma pesquisa com 2.744 clientes que fizeram algum pedido nos últimos 30 dias.

Fonte: Kaggle (adaptado)
Link: https://www.kaggle.com/benroshan/online-food-delivery-preferencesbangalore-region

Variável resposta:

1, se o cliente respondeu que é influenciado pela nota do restaurante
0, se o cliente respondeu que não é influenciado pela nota do restaurante

10 variáveis explicativas:

- Idade
- Gênero
- Estado civil
- Ocupação
- Renda mensal declarada
- Grau de escolaridade
- Refeição mais frequente dos pedidos realizados no último ano
- Indicação de se já fez compra de comida saudável no último ano
- Indicação de se já fez alguma reclamação por atraso na entrega no último ano
- Nota média de avaliação dos restaurantes escolhidos no último ano

Arquivo: Pedidos_Refeicoes.xlsx

Case: Pedidos de refeições

Uma empresa de tecnologia, atuante no ramo de entrega de refeições solicitadas *online*, gostaria de entender quais os perfis de consumidores que mais se deixam influenciar pela nota de avaliação dos restaurantes parceiros, no momento de realizarem seu pedido. Para entender esse aspecto, realizaram uma pesquisa com 2.744 clientes que fizeram algum pedido nos últimos 30 dias.

Fonte: Kaggle (adaptado)
Link: https://www.kaggle.com/benroshan/online-food-delivery-preferencesbangalore-region

a) Faça a análise exploratória de cada variável, individualmente.
b) Categorize as variáveis quantitativas, a fim de utilizá-las na árvore de decisão.
c) Faça a análise bidimensional de cada variável explicativa *versus* a variável resposta. Quais variáveis parecem exercer maior influência sobre a resposta?

Construa uma Árvore de Decisão com **2 níveis** e avalie a sua saída gráfica.

d) Há quantos nós intermediários e nós finais?
e) Qual a frequência absoluta e relativa de cada nó final?
f) Quantos nós finais são mais propensos que a base geral a se deixarem influenciar pela nota?
g) Interprete todos os perfis obtidos, do mais propenso ao menos propenso.
h) Avalie o desempenho do modelo, por meio dos índices de sensibilidade, especificidade e acurácia.

Case: Pedidos de refeições

a) Faça a análise exploratória de cada variável, individualmente.

```
> summary(refeicoes$Idade)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  18.00   23.00   27.00   28.25   32.00   76.00
> round(prop.table(table(refeicoes$Genero)),3)

 Feminino Masculino
    0.432     0.568
> round(prop.table(table(refeicoes$Estado_civil)),3)

       Casado Não informado      Solteiro
        0.259         0.032         0.708
> round(prop.table(table(refeicoes$Ocupacao)),3)

       Autônomo Dono(a) de casa   Empregado CLT       Estudante
          0.157           0.020           0.545           0.277
> round(prop.table(table(refeicoes$Renda_mensal)),3)

    Até 5.000 Mais de 5.000
        0.283         0.717
> round(prop.table(table(refeicoes$Grau_educacao)),3)

    Doutorado  Ensino médio     Graduação Pós-Graduação Sem instrução
        0.058         0.032         0.443         0.461         0.006
> round(prop.table(table(refeicoes$Refeicao_mais_frequente)),3)

       Almoço Café da manhã        Jantar       Lanches
        0.312         0.294         0.137         0.257
> round(prop.table(table(refeicoes$Compras_saudaveis)),3)

  Não   Sim
0.813 0.187
> round(prop.table(table(refeicoes$Reclamacao_atraso)),3)

  Não   Sim
0.913 0.087
> round(prop.table(table(refeicoes$Avaliacao_media_anterior)),3)

 < 4,5 >= 4,5
 0.187  0.813
> round(prop.table(table(refeicoes$Influenciado_por_nota)),3)

  Não   Sim
0.292 0.708
```

O comportamento das variáveis é razoável? Existem indícios de inconsistências nos dados?

Case: Pedidos de refeições

b) Categorize as variáveis quantitativas, a fim de utilizá-las na árvore de decisão.

```
refeicoes$Idade_cat = quantcut(refeicoes$Idade, 4)
```

O comando *quantcut* do R **categoriza** uma variável quantitativa contínua em grupos de tamanho aproximadamente igual. Neste caso, foram criados 4 grupos de idade.

Case: Pedidos de refeições

c) Faça a análise bidimensional de cada variável explicativa *versus* a variável resposta. Quais variáveis parecem exercer maior influência sobre a resposta?

```
> round(prop.table(table(refeicoes$Idade_cat, refeicoes$Influenciado_por_nota), 1), 3)

           Não   Sim
  [18,23] 0.291 0.709
  (23,27] 0.307 0.693
  (27,32] 0.297 0.703
  (32,76] 0.271 0.729
> round(prop.table(table(refeicoes$Genero, refeicoes$Influenciado_por_nota), 1), 3)

             Não   Sim
  Feminino  0.291 0.709
  Masculino 0.292 0.708
> round(prop.table(table(refeicoes$Estado_civil, refeicoes$Influenciado_por_nota), 1), 3)

                 Não   Sim
  Casado       0.309 0.691
  Não informado 0.273 0.727
  Solteiro     0.287 0.713
> round(prop.table(table(refeicoes$Ocupacao, refeicoes$Influenciado_por_nota), 1), 3)

                   Não   Sim
  Autônomo       0.359 0.641
  Dono(a) de casa 0.375 0.625
  Empregado CLT  0.270 0.730
  Estudante      0.291 0.709
> round(prop.table(table(refeicoes$Renda_mensal, refeicoes$Influenciado_por_nota), 1), 3)

                 Não   Sim
  Até 5.000    0.420 0.580
  Mais de 5.000 0.241 0.759
```

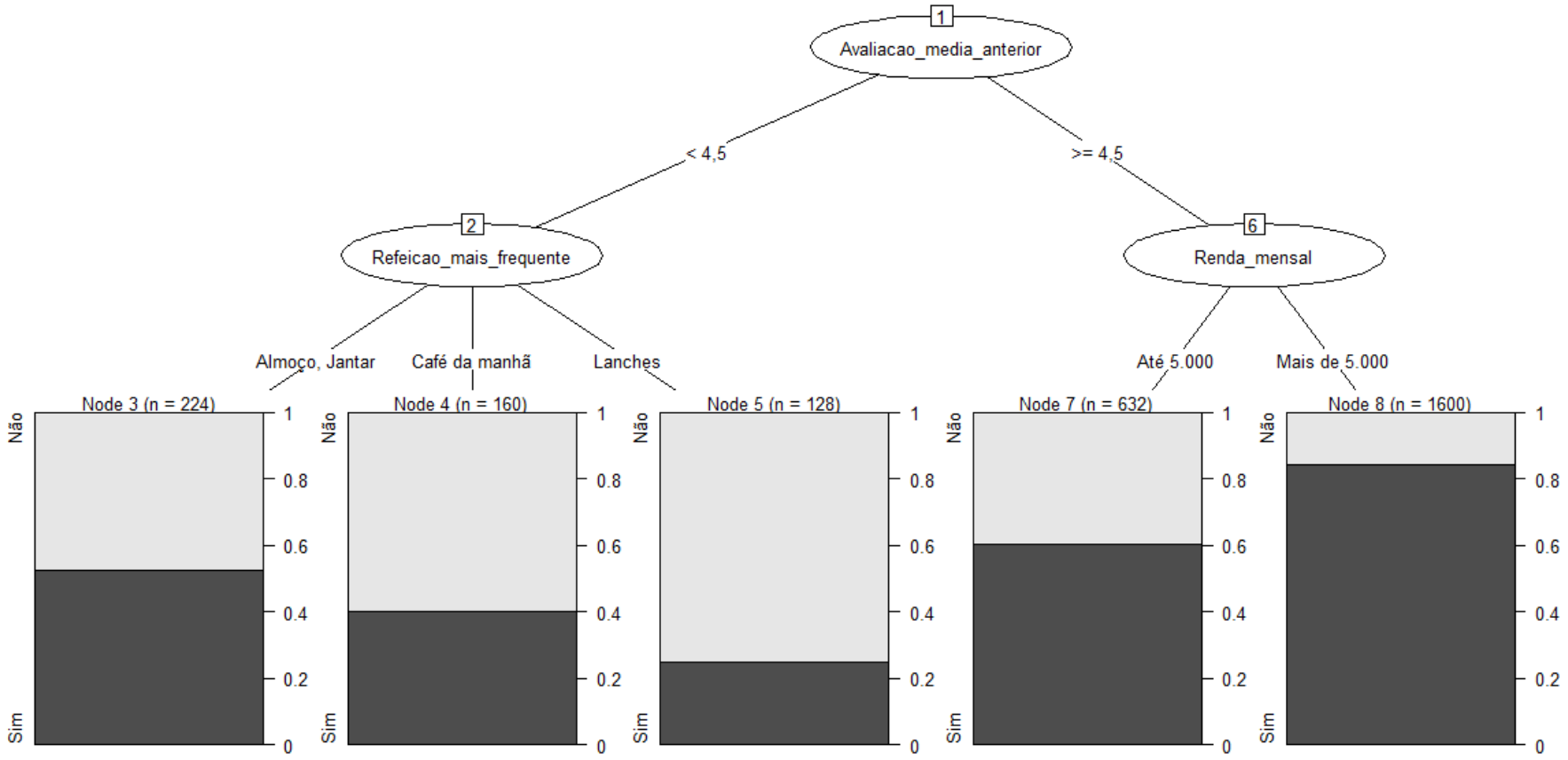
Case: Pedidos de refeições

c) Faça a análise bidimensional de cada variável explicativa *versus* a variável resposta. Quais variáveis parecem exercer maior influência sobre a resposta?

```
> round(prop.table(table(refeicoes$Grau_educacao, refeicoes$Influenciado_por_nota), 1), 3)

                 Não   Sim
  Doutorado     0.256 0.744
  Ensino médio  0.534 0.466
  Graduação     0.259 0.741
  Pós-Graduação 0.315 0.685
  Sem instrução 0.000 1.000
> round(prop.table(table(refeicoes$Refeicao_mais_frequente, refeicoes$Influenciado_por_nota), 1), 3)

                 Não   Sim
  Almoço        0.196 0.804
  Café da manhã 0.297 0.703
  Jantar        0.255 0.745
  Lanches       0.422 0.578
> round(prop.table(table(refeicoes$Compras_saudaveis, refeicoes$Influenciado_por_nota), 1), 3)

       Não   Sim
  Não 0.321 0.679
  Sim 0.166 0.834
> round(prop.table(table(refeicoes$Reclamacao_atraso, refeicoes$Influenciado_por_nota), 1), 3)

       Não   Sim
  Não 0.315 0.685
  Sim 0.046 0.954
> round(prop.table(table(refeicoes$Avaliacao_media_anterior, refeicoes$Influenciado_por_nota), 1), 3)

           Não   Sim
  < 4,5  0.584 0.416
  >= 4,5 0.225 0.775
```

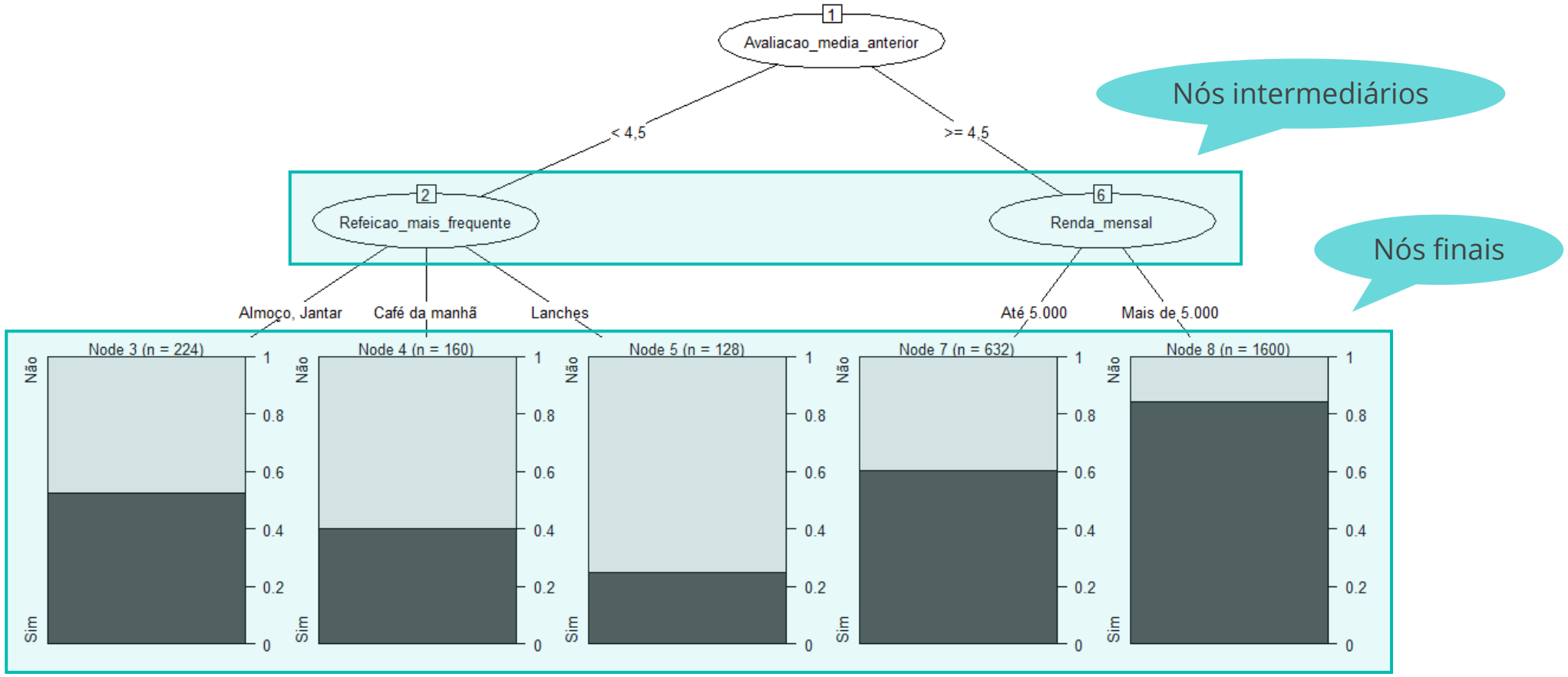
Case: Pedidos de refeições

Construa uma Árvore de Decisão com **2 níveis** e avalie a sua saída gráfica.



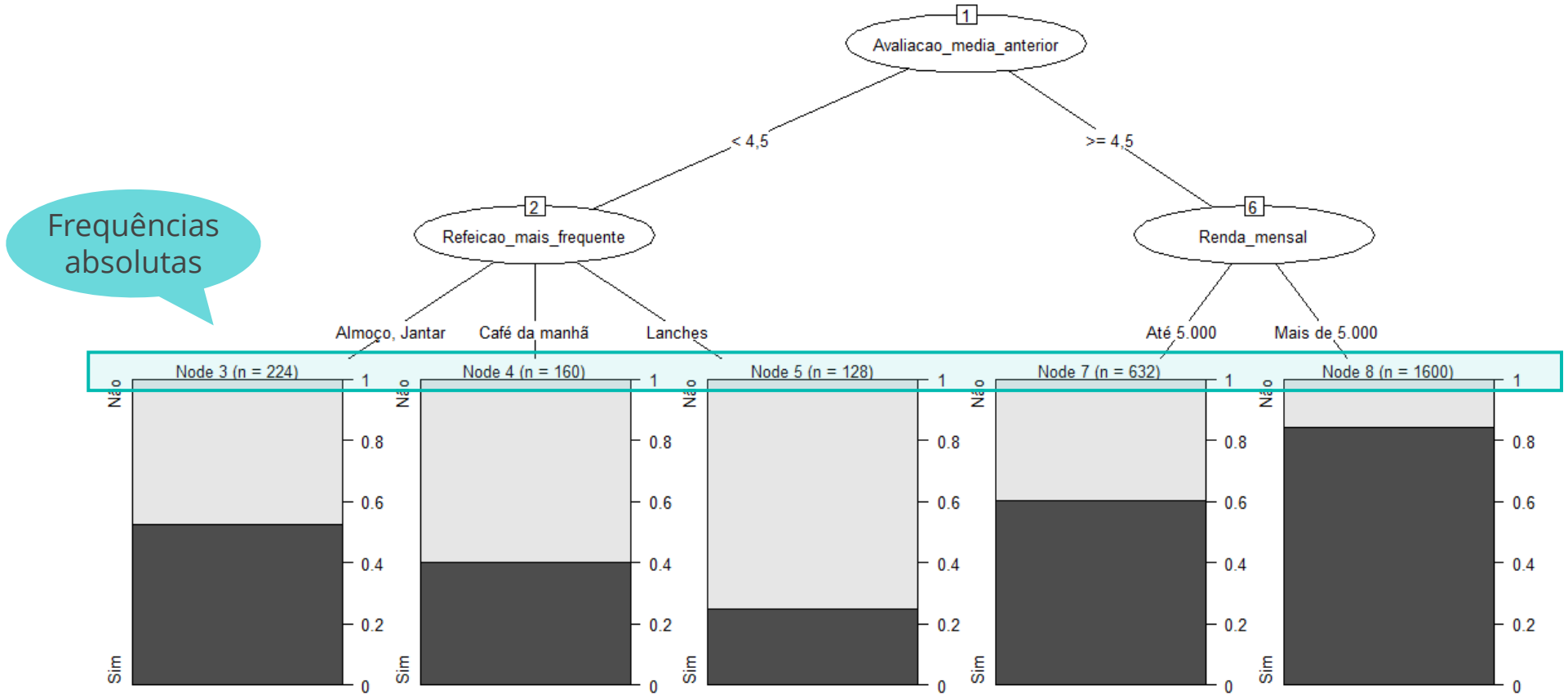
Case: Pedidos de refeições

d) Há quantos nós intermediários e nós finais?



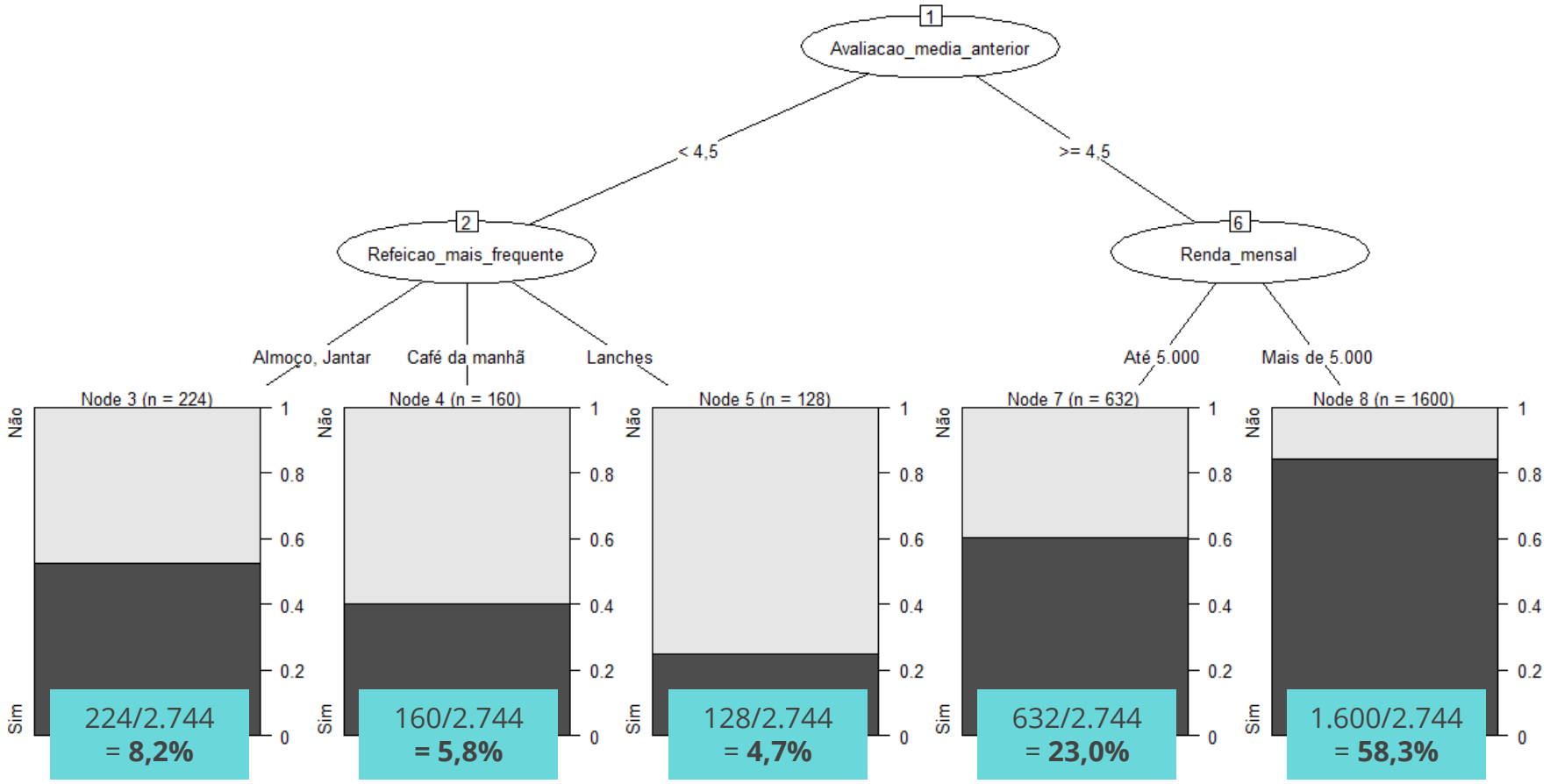
Case: Pedidos de refeições

e) Qual a frequência absoluta e relativa de cada nó final?



Case: Pedidos de refeições

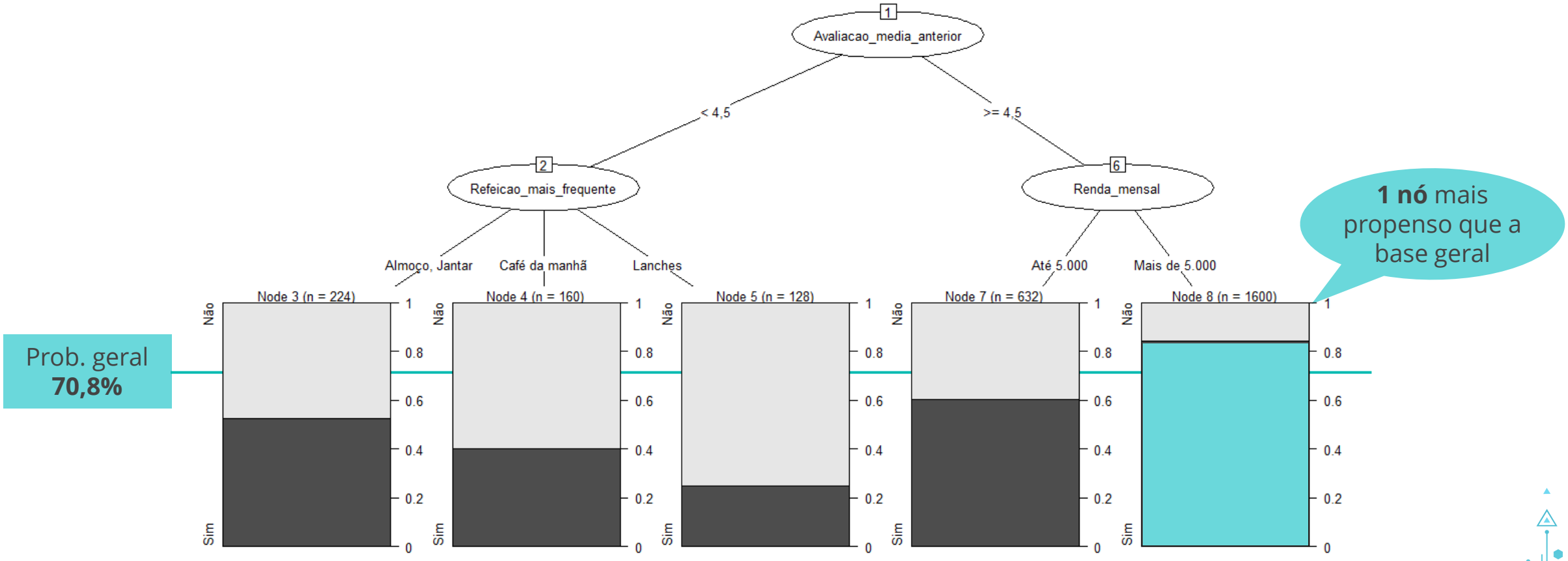
e) Qual a frequência absoluta e relativa de cada nó final?



Frequências relativas

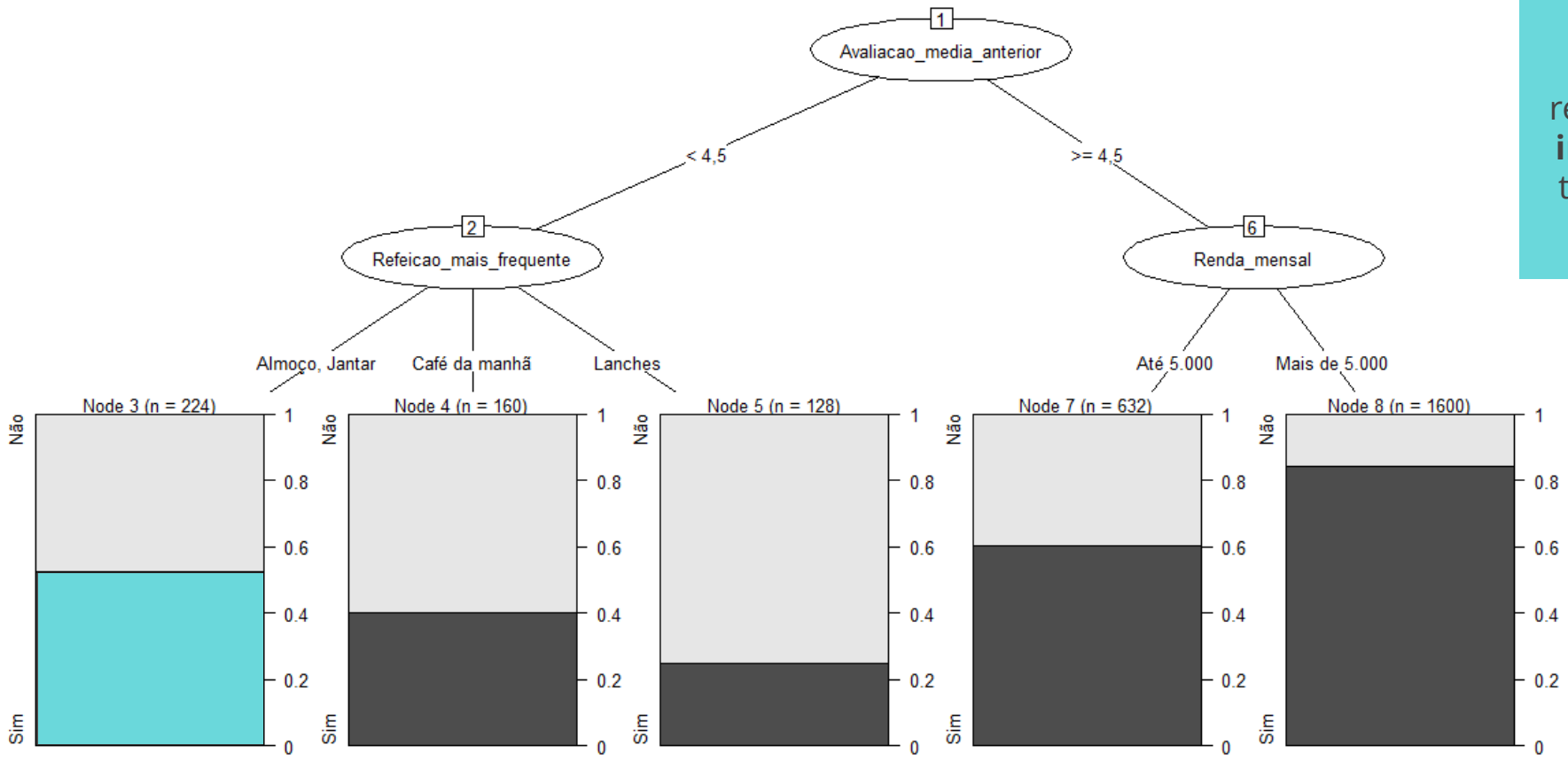
Case: Pedidos de refeições

f) Quantos nós finais são mais propensos que a base geral a se deixarem influenciar pela nota?



Case: Pedidos de refeições

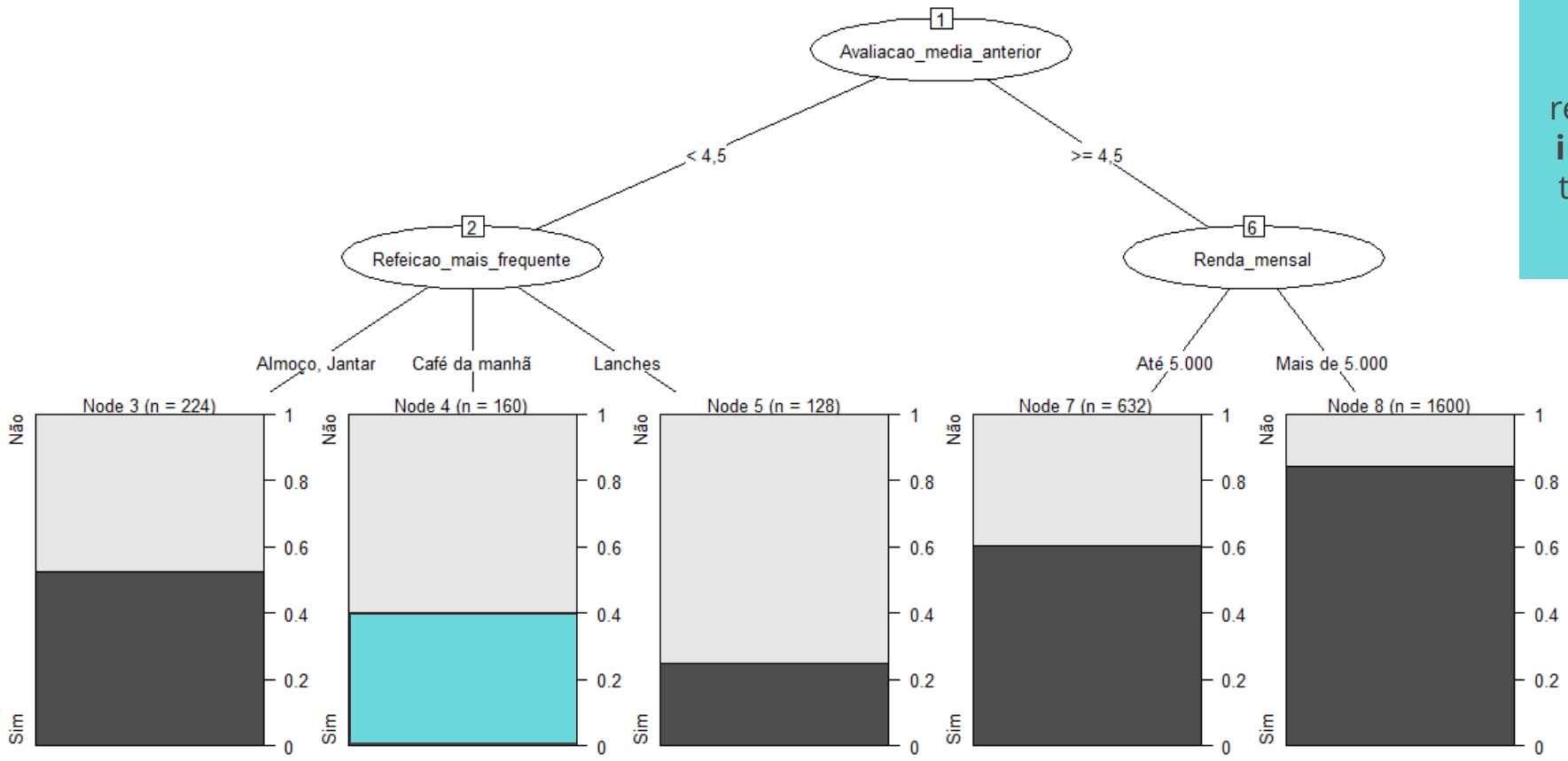
g) Interprete todos os perfis obtidos, do mais propenso ao menos propenso.



Clientes que, no último ano, realizaram pedidos em restaurantes avaliados com nota **inferior a 4,5**, em média; e cujo tipo de refeição mais frequente foi o **almoço** ou **jantar**.

Case: Pedidos de refeições

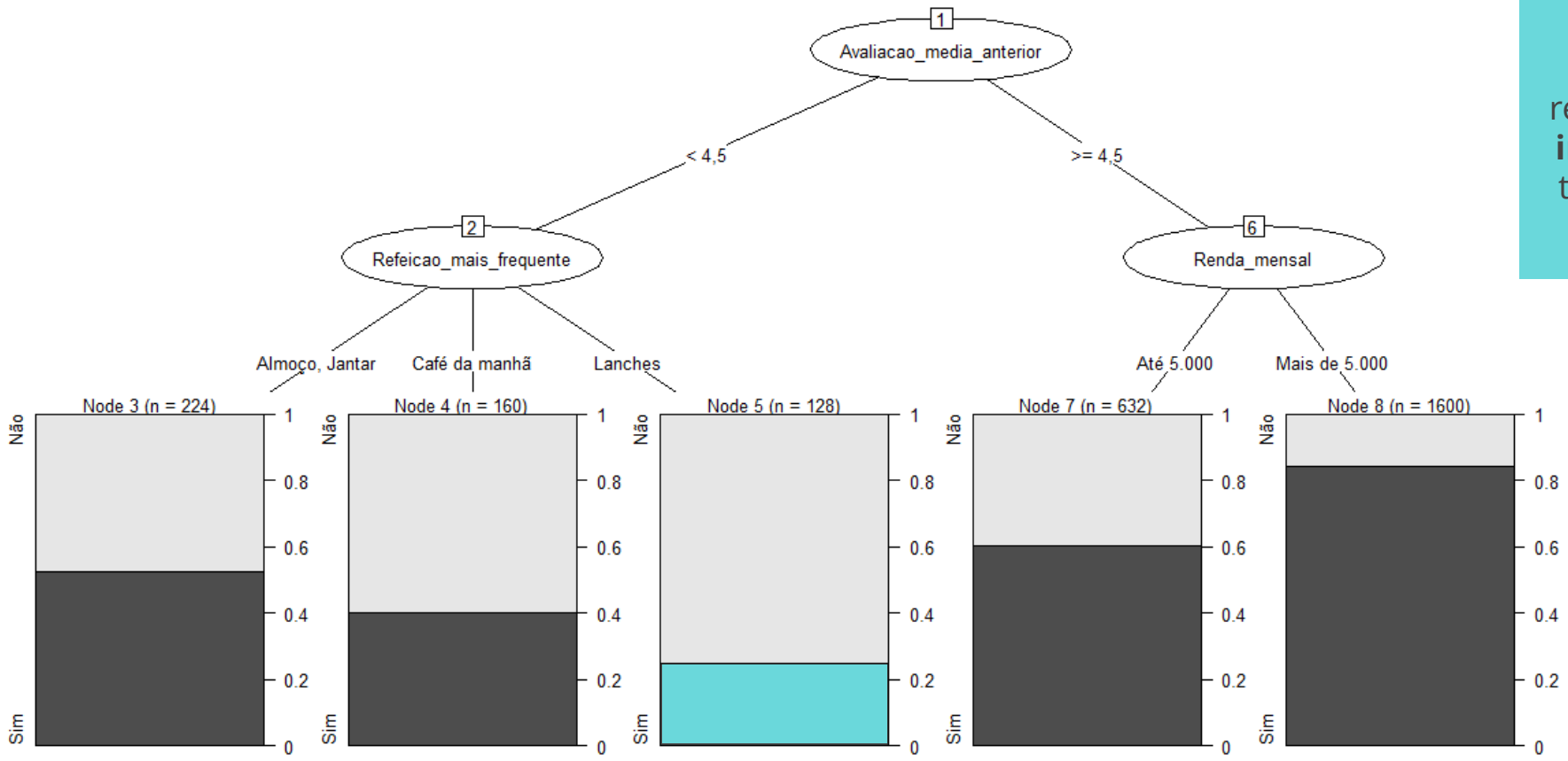
g) Interprete todos os perfis obtidos, do mais propenso ao menos propenso.



Clientes que, no último ano, realizaram pedidos em restaurantes avaliados com nota **inferior a 4,5**, em média; e cujo tipo de refeição mais frequente foi o **café da manhã**.

Case: Pedidos de refeições

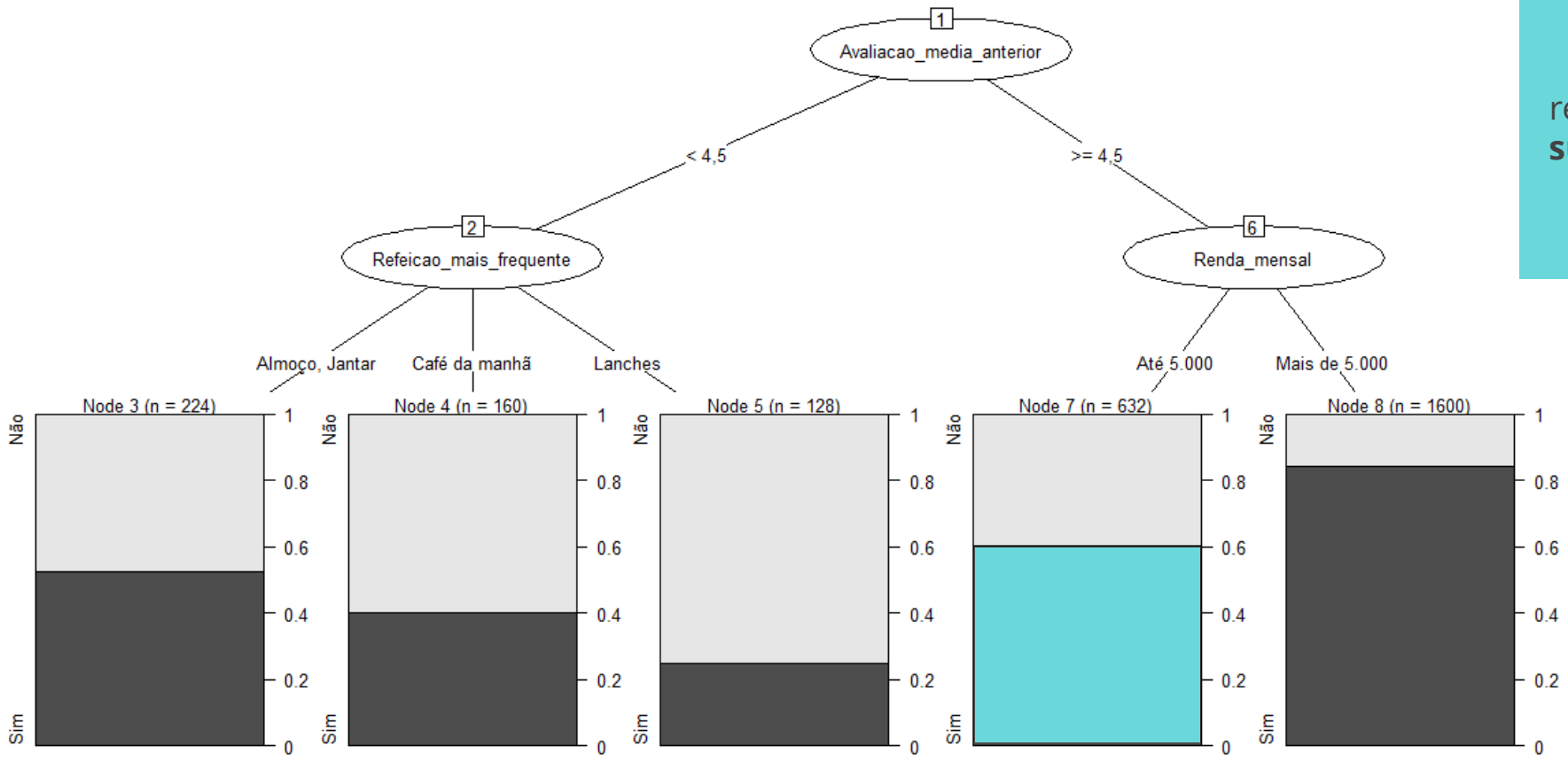
g) Interprete todos os perfis obtidos, do mais propenso ao menos propenso.



Clientes que, no último ano, realizaram pedidos em restaurantes avaliados com nota **inferior a 4,5**, em média; e cujo tipo de refeição mais frequente foi **lanches**.

Case: Pedidos de refeições

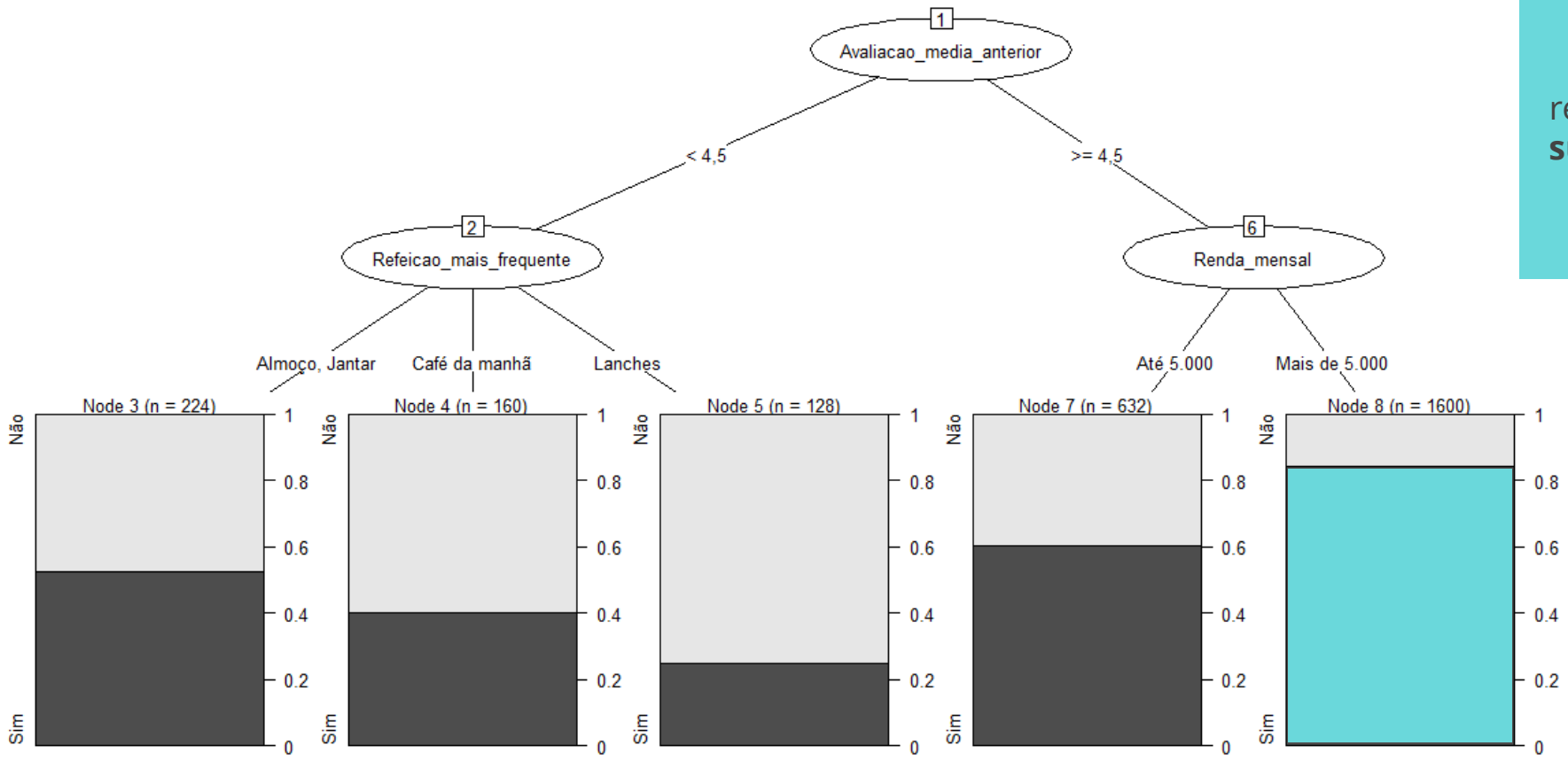
g) Interprete todos os perfis obtidos, do mais propenso ao menos propenso.



Clientes que, no último ano, realizaram pedidos em restaurantes avaliados com nota **superior a 4,5**, em média; e cuja renda mensal declarada é de **até R$ 5.000**.

Case: Pedidos de refeições

g) Interprete todos os perfis obtidos, do mais propenso ao menos propenso.



Clientes que, no último ano, realizaram pedidos em restaurantes avaliados com nota **superior a 4,5**, em média; e cuja renda mensal declarada é **superior a R$ 5.000**.

Case: Pedidos de refeições

h) Avalie o desempenho do modelo, por meio dos índices de sensibilidade, especificidade e acurácia.

Tabela de frequências absolutas

Real	**Predito**	
	Não influenciado	Influenciado
Não influenciado	550	251
Influenciado	594	1.349

Tabela de frequências relativas (soma 100% no total)

Real	**Predito**	
	Não influenciado	Influenciado
Não influenciado	20,0%	9,1%
Influenciado	21,6%	49,2%

Acurácia = 20,0% + 49,2% = **69,2%**

Tabela de frequências relativas (soma 100% em cada linha)

Real	**Predito**	
	Não influenciado	Influenciado
Não influenciado	68,7%	31,3%
Influenciado	30,6%	69,4%

Especificidade = **68,7%**

Sensibilidade = **69,4%**

Case: Pedidos de refeições

Uma empresa de tecnologia, atuante no ramo de entrega de refeições solicitadas *online*, gostaria de entender quais os perfis de consumidores que mais se deixam influenciar pela nota de avaliação dos restaurantes parceiros, no momento de realizarem seu pedido. Para entender esse aspecto, realizaram uma pesquisa com 2.744 clientes que fizeram algum pedido nos últimos 30 dias.

Fonte: Kaggle (adaptado)
Link: https://www.kaggle.com/benroshan/online-food-delivery-preferencesbangalore-region

Construa uma nova Árvore de Decisão, agora com **3 níveis*** e avalie a sua saída gráfica.

i) Refaça os itens (d) a (h).
j) Compare os desempenhos de ambos os modelos, com 2 níveis e com 3 níveis. Qual dos dois você escolheria para explicar o comportamento dos consumidores?

---

* Para esta base de dados, a fim de evitar nós muito pequenos, realize quebras apenas em nós que possuam ao menos 200 observações. Isso pode ser controlado, no R, por meio do comando:

controle <- chaid_control(maxheight = 3, minsplit = 200)

Case: Cross Sell

Uma financeira deseja oferecer um financiamento para seus clientes que possuem um crédito pessoal na instituição. Para isso, foi selecionada uma base de dados com 647 ofertas do financiamento.

Variável resposta - comprou:

1, se o cliente aceitou a oferta e adquiriu um financiamento
0, se o cliente não aceitou a oferta e adquiriu um financiamento

4 variáveis explicativas:

- Cargo
- Fatura do crédito pessoal: Mensal ou Semanal
- Cartao: Se possui ou não cartão de crédito
- Idade

Arquivo: Cross Sell.xlsx

Case: Cross Sell

Uma financeira deseja oferecer um financiamento para seus clientes que possuem um crédito pessoal na instituição. Para isso, foi selecionada uma base de dados com 647 ofertas do financiamento.

a) Faça a análise exploratória de cada variável, individualmente.
b) Faça a análise bidimensional de cada variável explicativa *versus* a variável resposta. Quais variáveis parecem exercer maior influência sobre a resposta?
c) Construa uma Árvore de Decisão e avalie a sua saída gráfica.
d) Há quantos nós intermediários e nós finais?
e) Qual a frequência absoluta e relativa de cada nó final?
f) Quantos nós finais são mais propensos que a base geral a se deixarem influenciar pela nota?
g) Interprete todos os perfis obtidos, do mais propenso ao menos propenso.
h) Avalie o desempenho do modelo, por meio dos índices de sensibilidade, especificidade e acurácia.

Case: Turnover

Uma área de gestão de pessoas deseja obter o perfil e prever os funcionários que irão sair da empresa.

Variável resposta - saiu:

1, se o colaborador saiu da empresa
0, se o colaborador não saiu da empresa

6 variáveis explicativas:

- Satisfacao: nota de satisfação do funcionário com a empresa
- horas_trabalhadas_media_mes : média de horas trabalhadas por mês
- acidente_trabalho: se teve acidente de trabalho (1) ou não (0)
- promocao_ultimos_5_anos: se foi promovido nos últimos 5 anos (1) ou não (0)
- departamento
- salario

Arquivo: Turnover.xlsx

Case: Turnover

Uma área de gestão de pessoas deseja obter o perfil e prever os funcionários que irão sair da empresa.

a) Faça a análise exploratória de cada variável, individualmente.
b) Faça a análise bidimensional de cada variável explicativa *versus* a variável resposta. Quais variáveis parecem exercer maior influência sobre a resposta?
c) Categorize as variáveis quantitativas pelo quartil.
d) Construa uma Árvore de Decisão e avalie a sua saída gráfica.
e) Há quantos nós intermediários e nós finais? Há a necessidade de diminuir o tamanho da árvore?
f) Crie a árvore com 2 níveis.
g) Qual a frequência absoluta e relativa de cada nó final?
h) Quantos nós finais são mais propensos que a base geral a saírem da empresa?
i) Interprete todos os perfis obtidos, do mais propenso ao menos propenso.
j) Avalie o desempenho do modelo, por meio dos índices de sensibilidade, especificidade e acurácia.

Referências

LIVROS-TEXTO

Anderson, R. A., Sweeney, J. D. e Williams, T. A. *Estatística Aplicada à Administração e Economia.* Editora Cengage. 4ª edição, 2019.

Agresti, A. (2002). *Categorical data analysis* (Vol. 359). Wiley-interscience.

Johnson, R. A. e Wichern, D. W. *Applied Multivariate Statistical Analysis*. Prentice-Hall Inc., 6th ed. 2007