# Estatística Aplicada
# EAD Ao Vivo

Tema da aula
**Árvore de Decisão**

LABDATA | FIA FUNDAÇÃO INSTITUTO DE ADMINISTRAÇÃO BUSINESS SCHOOL

21/10/2020

# ESTATÍSTICA APLICADA EAD AO VIVO

**Professora:**
Drª Karin Ayumi Tamura

**Coordenadores:**
Profª Drª Alessandra de Ávila Montini
Profª Dr.  Adolpho Walter Pimazoni Canton

# Currículo - Prof.ª Drª Karin Ayumi Tamura

FORMAÇÃO ACADÊMICA | EXPERIÊNCIA PROFISSIONAL



Profª Dra.
**Karin Ayumi Tamura**

**Contato:** karin.tamura@fia.com.br

- **FORMAÇÃO ACADÊMICA:** Pós-doutora (2015), Doutora (2012), mestre (2007) e bacharel (2003) em Estatística pelo Instituto de Matemática e Estatística da USP, tendo como área de pesquisa modelos de regressão, análise multivariada de dados e algoritmos de *machine learning*.
- **ATUAÇÃO PROFISSINAL:** Foi *Head* de *Analytics* por 14 anos, e atualmente é Conselheira Executiva e *Head* de Inovação na *Marketdata Solutions*, uma empresa do grupo WPP, e Professora Doutora no LABDATA FIA.
- **HISTÓRICO:** Atuação no mercado por 17 anos, com experiência profissional no segmento bancário (Bradesco) e consultoria *(Marketdata Solutions)*. Atuou como docente em cursos de pós-graduação (2010-16) no LABDATA FIA e ABEMD. Especialista em Estatística e *Advanced Analytics* trabalhando em projetos de diversos segmentos do mercado. Participante de congressos nacionais e internacionais voltados a área de Estatística, Dados e Algoritmos de *Machine Learning*.

*"Tenho duas paixões no meu trabalho: dados e pessoas. Voltar a lecionar no LABDATA FIA está sendo a realização de um sonho planejado desde a minha época de aluna de pós-graduação. Meu objetivo como professora é integrar a visão do mercado com as técnicas e tecnologias de análise de dados, por meio de uma atuação humanista no ensino aos alunos"*

## Projetos atendidos

# 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 Programático do Curso

21 AULAS AO VIVO COM PROFA. KARIN | 27 PLANTÕES AO VIVO COM PROF. STEPHAN, 7 LISTAS DE EXERCÍCIOS  E EAD VIDEO AULA EM PYTHON



| Dia | Mês | Aula | EAD Ao Vivo | Plantão Prof. Stephan |
|---|---|---|---|---|
| 5 | Agosto | Introdução ao Curso e Análise Exploratória de Dados | Aula Prof. Karin | 06/ago |
| 12 | Agosto | Análise Exploratória de Dados | Aula Prof. Karin | 13/ago |
| 19 | Agosto | Análise Exploratória de Dados - Introdução ao R | Aula Prof. Karin | 20/ago |
| 26 | Agosto | Lista de Exercícios em Sala de Aula (19hs-23hs - com presença obrigatória) | - | 27/ago |
| 2 | Setembro | Regressão Linear Simples | Aula Prof. Karin | 03/set |
| 9 | Setembro | Regressão Linear Simples e Múltipla | Aula Prof. Karin | 10/set |
| 16 | Setembro | Regressão Linear Simples e Múltipla | Aula Prof. Karin | 17/set |
| 23 | Setembro | Lista de Exercícios em Sala de Aula (19hs-23hs - com presença obrigatória) | - | 24/set |
| 30 | Setembro | Análise de Cluster | Aula Prof. Karin | 01/out |
| 7 | Outubro | Análise de Cluster | Aula Prof. Karin | 08/out |
| 14 | Outubro | Lista de Exercícios em Sala de Aula (19hs-23hs - com presença obrigatória) | - | 15/out |
| 21 | Outubro | Arvore de Decisão | Aula Prof. Karin | 22/out |
| 28 | Outubro | Lista de Exercícios em Sala de Aula (19hs-23hs - com presença obrigatória) | - | 29/out |
| 4 | Novembro | Regressão Logística | Aula Prof. Karin | 05/nov |
| 11 | Novembro | Regressão Logística | Aula Prof. Karin | 11/nov |
| 18 | Novembro | Lista de Exercícios em Sala de Aula (19hs-23hs - com presença obrigatória) | - | 19/nov |
| 25 | Novembro | estudo de caso | Aula Prof. Karin | 26/nov |
| 2 | Dezembro | estudo de caso | Aula Prof. Karin | 30/dez |
| 9 | Dezembro | estudo de caso | Aula Prof. Karin | 10/dez |
| 16 | Dezembro | Análise de Série Temporal  - modelo auto regressivo | Aula Prof. Karin | 17/dez |
| 23 | Dezembro | Lista de Exercícios em Sala de Aula (Frequência Liberada - véspera Natal) | - | - |
| Recesso Escolar | | EAD - INTRODUÇÃO AO PYTHON<br>EAD - INTRODUÇÃO AO PYTHON | EAD Video Aula<br>(8 horas) | -<br>- |
| 6 | Janeiro | Modelos estatísticos em Python | Aula Prof. Karin | 07/jan |
| 13 | Janeiro | Modelos estatísticos em Python | Aula Prof. Karin | 14/jan |
| 20 | Janeiro | Modelos estatísticos em Python | Aula Prof. Karin | 20/jan |
| 27 | Janeiro | Introdução a Big Data - Aplicações de Machine Learning e Deep Learning | Aula Prof. Karin | 28/jan |
| 3 | Fevereiro | Aplicações de Machine Learning | Aula Prof. Karin | 04/fev |
| 10 | Fevereiro | Aplicações de Machine Learning | Aula Prof. Karin | 11/fev |
| 17 | Fevereiro | Lista de Exercícios (Frequência Liberada - quarta de cinzas) | - | 18/fev |
| 24 | Fevereiro | EXERCICIOS DE REVISÃO - EAD (19hs e 23hs - com presença obrigatória) | - | 24/fev |
| 3 | Março | Prova (Plataforma On Line: 19hs e 23hs ) | - | |

# Conteúdo da Aula

1. Introdução
2. Teste Qui-Quadrado
3. Árvore de Decisão
4. Exercícios para casa
   CASE: *Churn* em Telefonia

# 1. Introdução

# Case *Churn* em Telefonia



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

**Aplicação**
Segmento Telecom.

**Exemplo**
Identificar o perfil dos funcionários que receberam promoção no ano anterior, e avaliar quais características que se destacam em relação a aqueles 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 & Nutrição.

# Case Perfil de Pacientes Crônicos

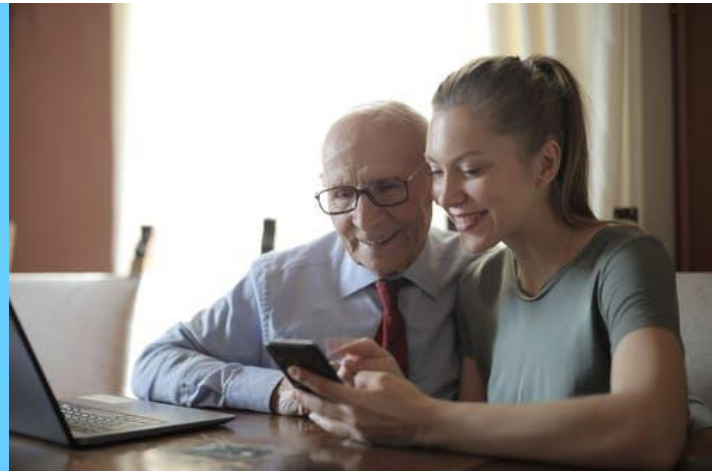


**Exemplo**
Identificar o perfil dos pacientes diabéticos, segundo seu hábito de vida: dieta, exercícios, se é fumante, idade, sexo,  peso, altura, etc.

**Aplicação**
Áreas de Médica e Nutrição.

**Exemplo**
Identificar o perfil dos clientes que realizam transações digitais com ações de incentivo ao uso do canal para aqueles clientes 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**
Marketing.

# Evento Binário



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

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

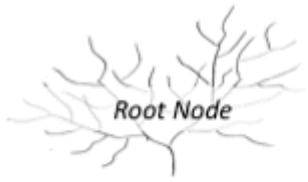
# Objetivo de Árvore de Decisão

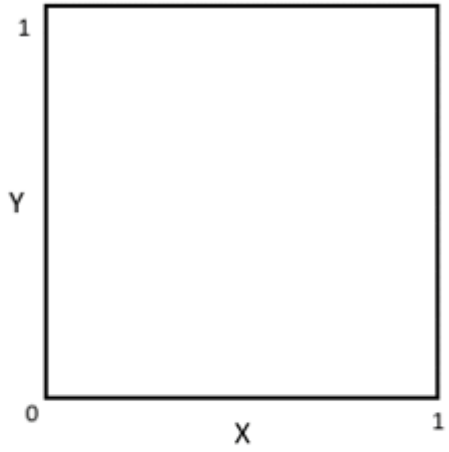


A Árvore de Decisão é uma ferramenta muito utilizada para o apoio à tomada de decisão.

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.






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

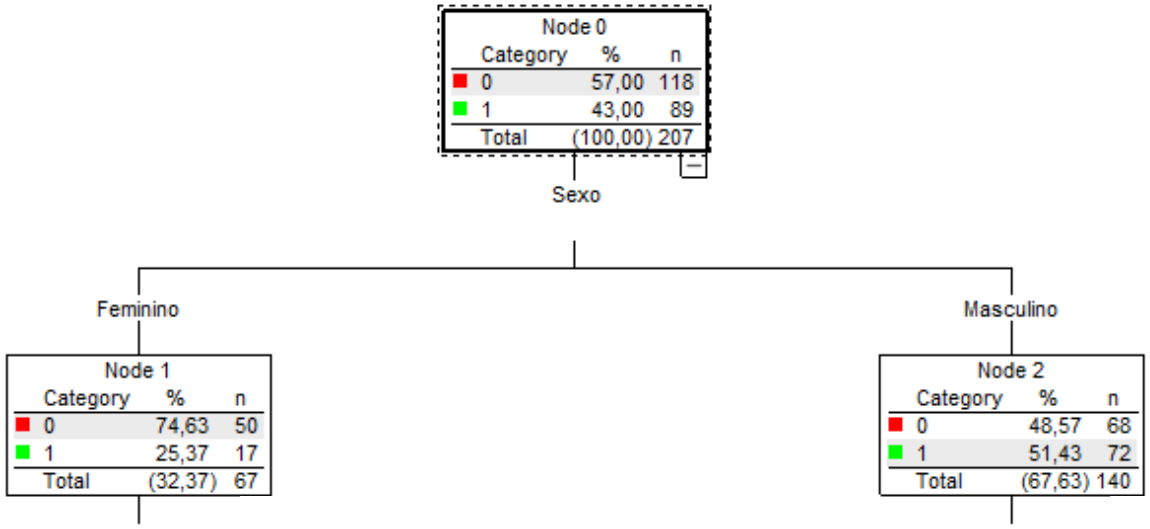
Uma *startup* especializada no ramo de Serviços via Aplicativo tem dois perfis de profissionais: Tecnologia e Administrativo (RH, Financeiro, Comercial, etc). A empresa gostaria de conhecer o perfil dos funcionários de Tecnologia com base nas informações cadastradas no sistema de RH da empresa. Quais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

**Evento (Variável resposta):**
1, caso funcionário seja da área de Tecnologia
0, caso funcionário não seja da área de Tecnologia

**Variáveis Explicativas:**
- Idade
- Sexo
- Escolaridade
- Estado Civil
- Cidade Nascimento

A Árvore de Decisão inicia com a variável resposta, e mostra a proporção dos eventos zeros e uns (evento de interesse).

O algoritmo seleciona dentre todas as variáveis explicativas, aquela que melhor discrimina a variável resposta, segundo um critério de partição.

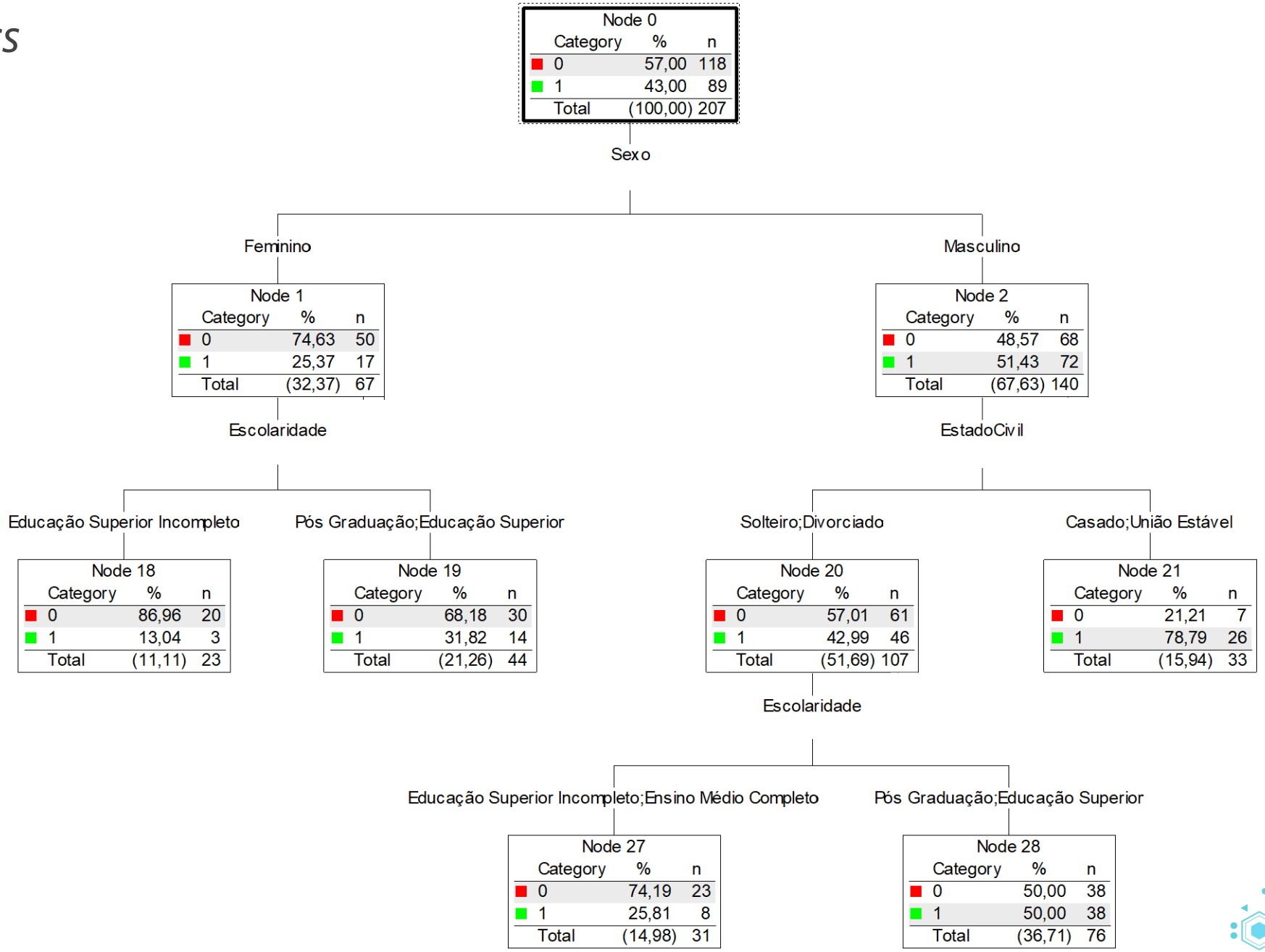



**Interpretação:** A variável resposta apresenta 43% dos funcionários da área de Tecnologia do total de 207 funcionários.

**Interpretação:** Se isolarmos todos os funcionários do sexo masculino (67,6%) e pegarmos aleatoriamente um funcionário, a probabilidade dele ser da Área de Tecnologia passa para 51,4%.

Dadas as quebras definidas pela primeira variável, o algoritmo escolhe dentre as demais variáveis restantes qual é a variável mais discriminante, dada a partição no nível anterior da árvore.

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

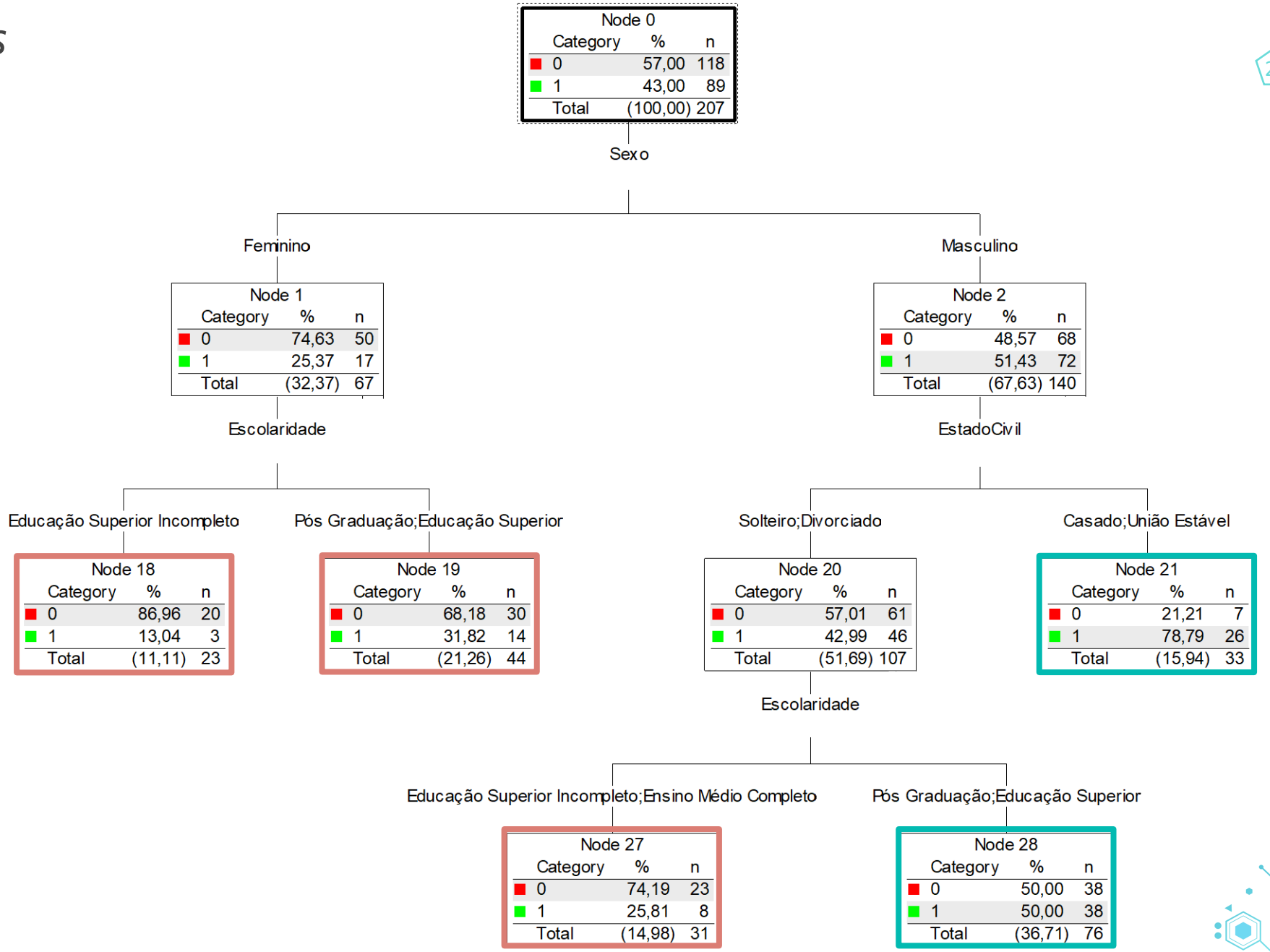


# Case: *People Analytics*

1. INTRODUÇÃO | ÁRVORE DE DECISÃO

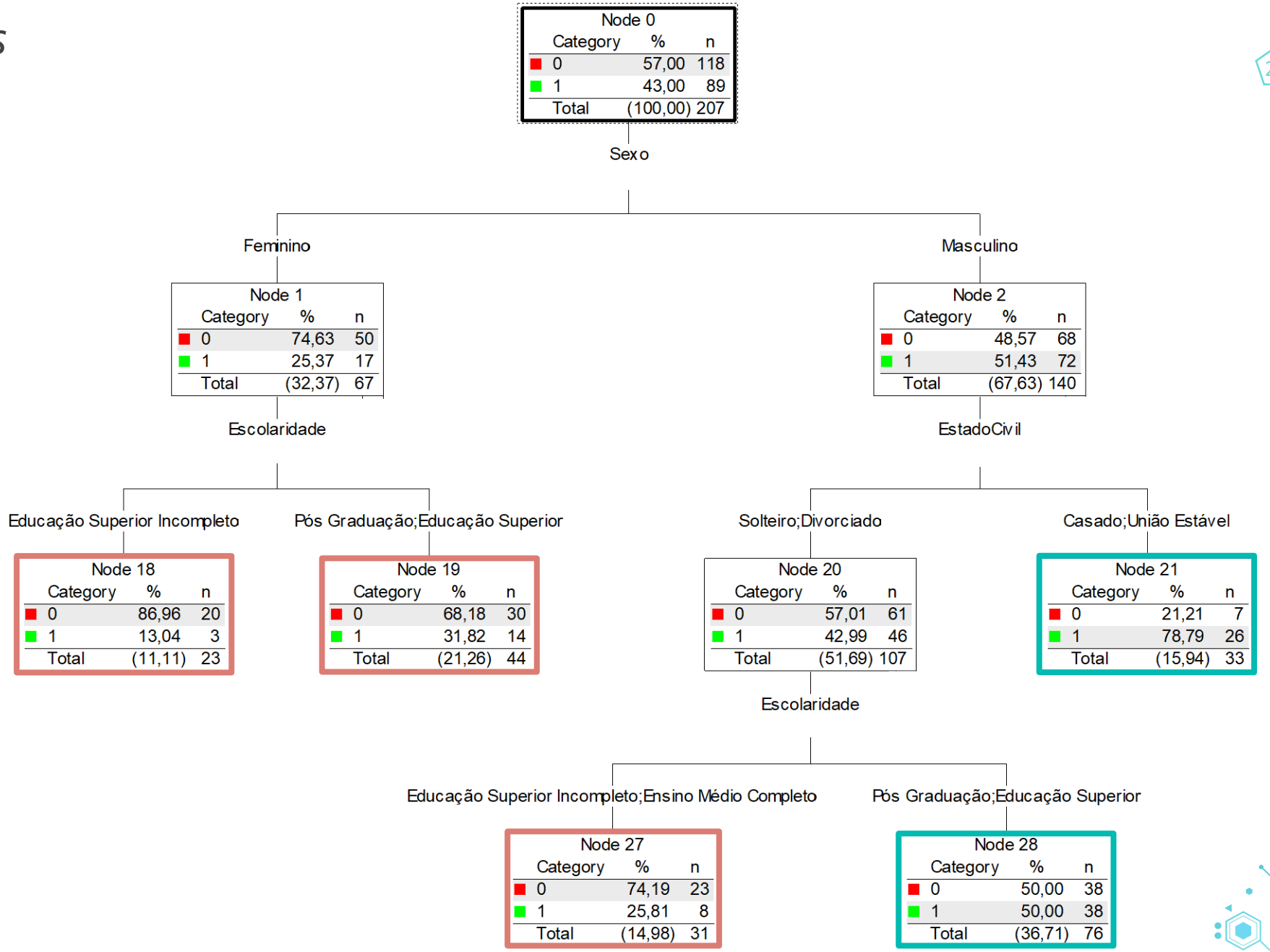
Os nós finais podem ser classificados em: propensos ao evento de interesse (verde) e não propensos (vermelho).

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



Neste exemplo, foram encontrados 5 perfis da empresa, sendo que 2 da Área de Tecnologia e 3 das demais Áreas Administrativas.

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



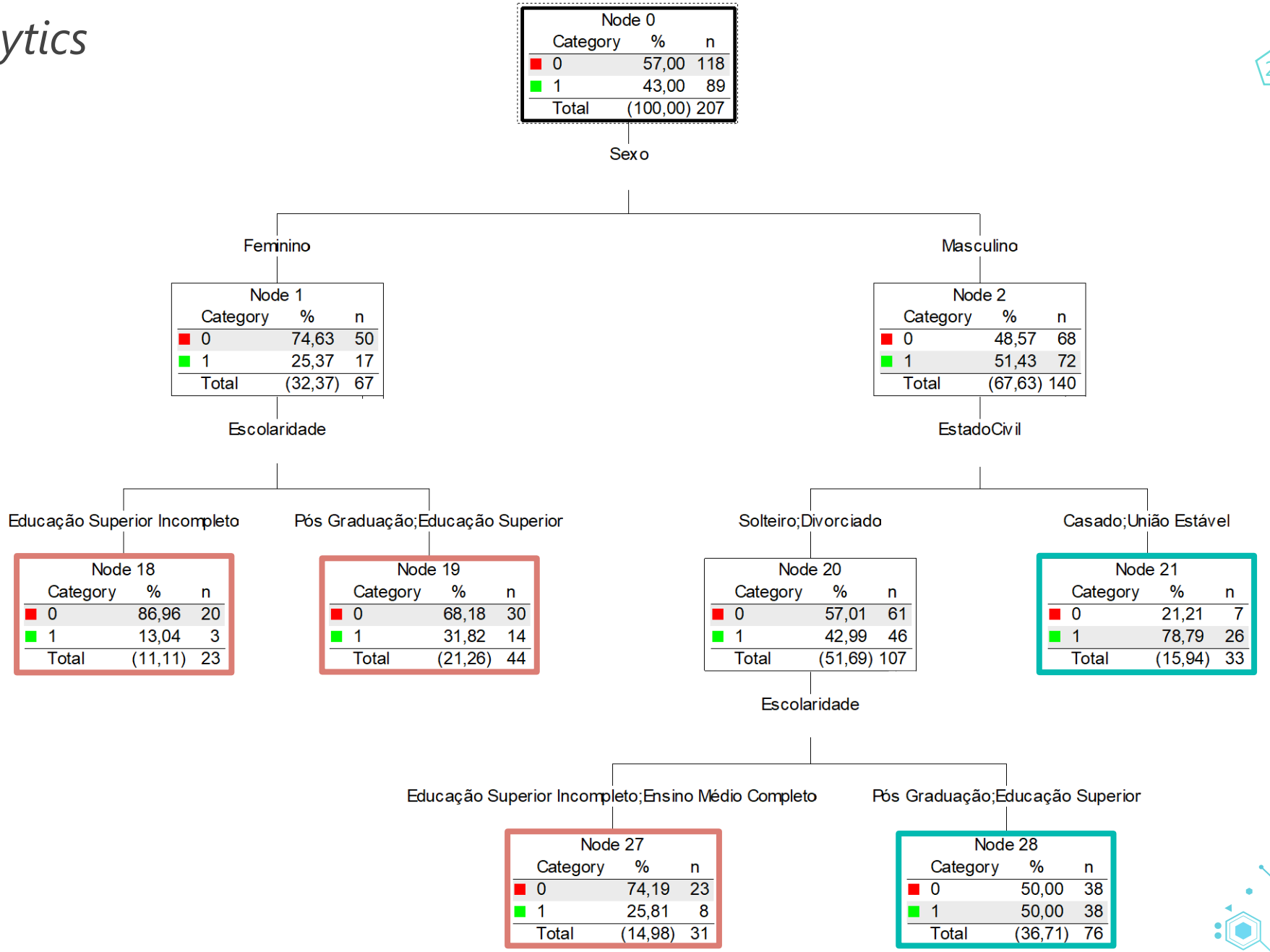
# Exercício: *People Analytics*



Com base no modelo de Árvore de Decisão apresentado ao lado:

**(a) Qual a probabilidade do perfil mais propenso a ser da Área de Tecnologia e qual a representatividade deste perfil na empresa?**

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



# Como a Árvore de Decisão é construída?



Dado que entendemos 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 Interation Detection*), que é baseado no Teste Estatístico de Qui-Quadrado.

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

# 2. Teste Qui-Quadrado

# Testar a hipótese de associação entre duas variáveis qualitativas



No case anterior de *People Analytics*, gostaríamos de avaliar se o Sexo (Feminino ou Masculino) tem associação com 'Tipo de Departamento' (Tecnologia ou Demais áreas), para isso podemos utilizar como critério de avaliação dessa associação 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, em 3 categorias avaliou-se a frequência de consumo de bebida alcoólica (não consome, 1x/semana e 2x/semana ou mais) e em 3 categorias o nível pressão (baixa, normal e alta). Os dados são apresentados a seguir.

Tabela de Frequências ABSOLUTA (observada):

| **Consumo de bebida alcoólica** | **Pressão Arterial** | | | **Total** |
|---|---|---|---|---|
| | Baixa | Normal | Alta | |
| Não consome | 20 | 21 | 11 | **52** |
| Uma vez por semana | 21 | 19 | 13 | **53** |
| Duas ou mais | 21 | 22 | 35 | **78** |
| **Total** | **62** | **62** | **59** | **183** |

O primeiro passo é obter uma tabela com os valores observados.

# 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, em 3 categorias avaliou-se a frequência de consumo de bebida alcoólica (não consome, 1x/semana e 2x/semana ou mais) e em 3 categorias o nível pressão (baixa, normal e alta). Os dados são apresentados a seguir.

Tabela de Frequências ABSOLUTA (observada):

| **Consumo de bebida alcoólica** | **Pressão Arterial** | | | **Total** |
|---|---|---|---|---|
| | Baixa | Normal | Alta | |
| Não consome | 20 | 21 | 11 | **52** |
| Uma vez por semana | 21 | 19 | 13 | **53** |
| Duas ou mais | 21 | 22 | 35 | **78** |
| **Total** | **62** | **62** | **59** | **183** |

Tabela de Frequências ABSOLUTA (**esperada**):

| **Consumo de bebida alcoólica** | **Pressão Arterial** | | | **Total** |
|---|---|---|---|---|
| | Baixa | Normal | Alta | |
| Não consome | 17,62 | 17,62 | 16,77 | **52** |
| Uma vez por semana | 17,96 | 17,96 | 17,09 | **53** |
| Duas ou mais | 26,43 | 26,43 | 25,15 | **78** |
| **Total** | **62** | **62** | **59** | **183** |

O segundo passo é obter uma 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 ABSOLUTA (observada):

| Consumo de bebida alcoólica | Pressão Arterial | | | Total |
|---|---|---|---|---|
| | Baixa | Normal | Alta | |
| Não consome | 20 | 21 | 11 | **52** |
| Uma vez por semana | 21 | 19 | 13 | **53** |
| Duas ou mais | 21 | 22 | 35 | **78** |
| **Total** | **62** | **62** | **59** | **183** |

Tabela de Frequências ABSOLUTA (**esperada**):

| Consumo de bebida alcoólica | Pressão Arterial | | | Total |
|---|---|---|---|---|
| | Baixa | Normal | Alta | |
| Não consome | 17,62 | 17,62 | 16,77 | **52** |
| Uma vez por semana | 17,96 | 17,96 | 17,09 | **53** |
| Duas ou mais | 26,43 | 26,43 | 25,15 | **78** |
| **Total** | **62** | **62** | **59** | **183** |

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

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

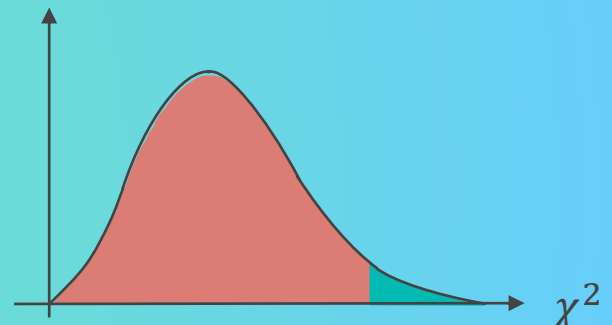
Este valor deve ser utilizado no **Teste de hipótese Qui-Quadrado**

# Distribuição Qui-Quadrado



O objetivo do teste do 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

# 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 $ij$

$n$ : tamanho amostral

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

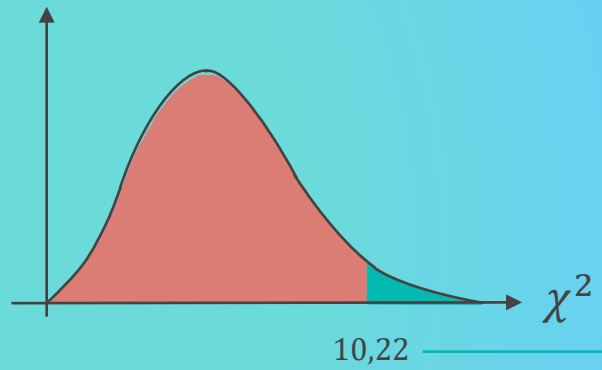
# Case: Estudo de doenças cardiovasculares



**Estatística do Teste**

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

**Distribuição Qui-Quadrado**



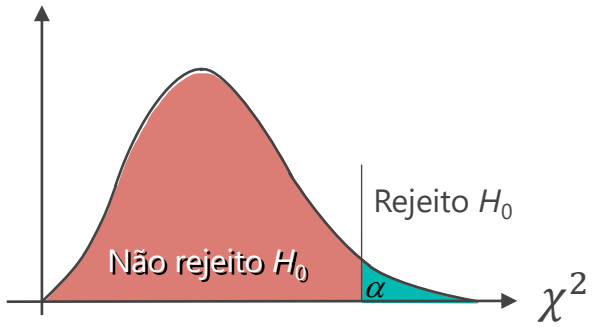

A estatística do teste segue a distribuição **Qui-quadrado** com **(n-1)(m-1) graus de liberdade**, sendo **n** o número de categorias na variável da linha e **m** a quantidade de categorias na variável da coluna.

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.

# 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 $\alpha$ é o nível de significância e *k* são os graus de liberdade da estatística de Qui-quadrado.

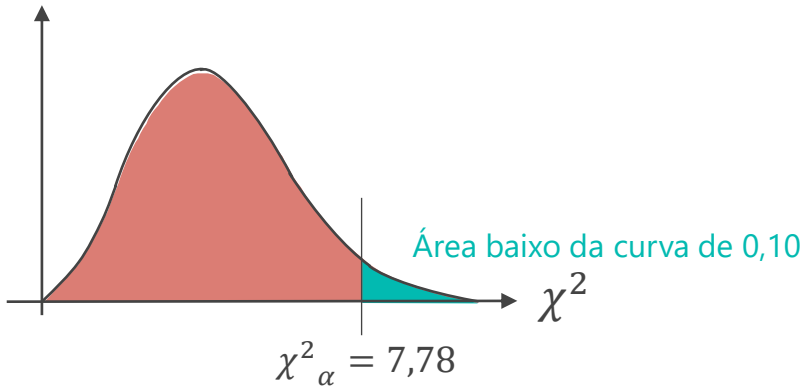
**Chi-square Distribution Table**

| 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 |

# Case: Estudo de doenças cardiovasculares



**Distribuição Qui-Quadrado (4 g.l.)**




**Regra de Decisão (considerando 90 % de confiança):** Como o $\chi^2 = 10{,}22 > \chi^2{}_\alpha = 7{,}78$, rejeitamos $H_0$, ou seja, existe associação entre consumo de bebida alcoólica e pressão arterial.

**Chi-square Distribution Table**

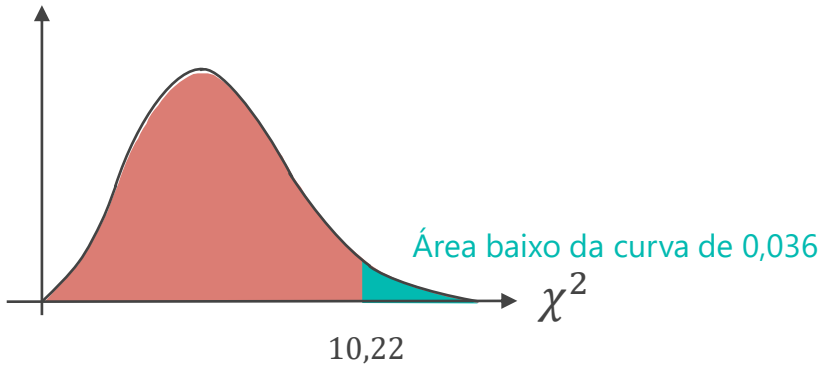
| d.f. | .995 | .99 | .975 | .95 | .9 | .1 | .05 | .025 | .01 |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.00 | 0.00 | 0.00 | 0.00 | 0.02 | 2.71 | 3.84 | 5.02 | 6.63 |
| 2 | 0.01 | 0.02 | 0.05 | 0.10 | 0.21 | 4.61 | 5.99 | 7.38 | 9.21 |
| 3 | 0.07 | 0.11 | 0.22 | 0.35 | 0.58 | 6.25 | 7.81 | 9.35 | 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 |

# Case: Estudo de doenças cardiovasculares





**Distribuição Qui-Quadrado (4 g.l.)**




**Regra de Decisão (considerando 90 % de confiança):** Quando 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: Estudo de Caso – 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 diferenciam entre os consumidores do sexo feminino e masculino? Se a preferência pela cerveja independe 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 mercados alvos. 150 consumidores de cerveja foram selecionados aleatoriamente, degustaram cada cerveja, e manifestaram sua predileção, ou primeira escolha. 80 eram do sexo masculino e 70 do sexo feminino. 50 optaram pela cerveja *light*, 70 pela comum e 30 pela escura. Dentre os *lights*, 20 eram homens e 30 mulheres. Dentre aqueles que optaram pela cerveja comum, 40 eram homens e 30 eram mulheres.

Responda as seguintes questões:

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 36.
g) Qual a conclusão do teste?
h) A empresa deve fazer campanhas diferentes de acordo com genêro do consumidor?

# 3. Árvore de Decisão

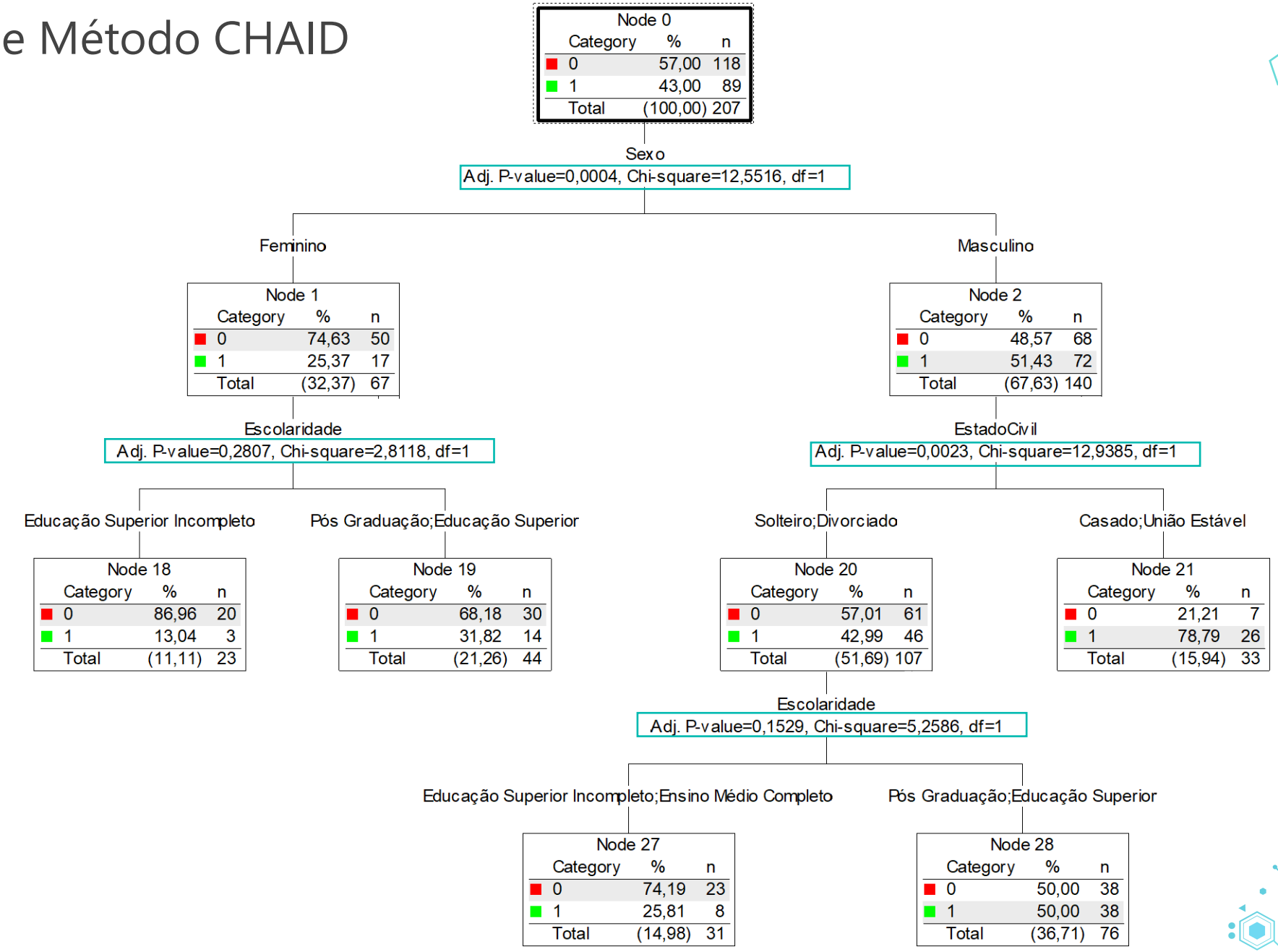
# Teste Qui-Quadrado e Método CHAID





Pelo método CHAID (*Chi-square Automatic Interaction Detection*), é calculada a Estatística Qui-quadrado, sendo a variável com o menor p-valor (e o maior valor do Qui-quadrado) é a primeira variável a ser sugerida para criar a primeira partição da base.

Para cada categoria da primeira variável quebra, é escolhida a próxima variável com o maior valor da Estatística Qui-Quadrado, até que um critério de parada seja satisfeito.



# Exercício: Estudo de Caso - *Churn* em Telefonia



Uma empresa de Telecom com 318.463 mil clientes está preocupada com o aumento do cancelamento voluntário das assinaturas de telefonia fixa. No período de 1 mês, existiram 2.766 cancelamentos voluntários (0,87% dos clientes cancelaram voluntariamente). Apesar deste número parecer pequeno, em um ano se nada for feito, 10% da base de clientes deixarão a companhia. Por meio do estudo da base de dados, o objetivo da empresa é identificar os clientes com maior probabilidade de cancelar voluntariamente suas linhas telefônicas e fazer ações segmentadas de marketing de acordo com seu perfil. Serão estudadas as características transacionais dos clientes para explicar o evento de cancelamento voluntário. O gestor da área de CRM gostaria de saber qual o perfil dos cientes que mais cancelam, quais suas características e se essa regra é eficaz para predizer o comportamento daqueles clientes que cancelam no próximo mês.

Fonte: base simulada, inspirada em problemas reais de consultoria da Profª Karin Ayumi Tamura.

**Evento (Variável resposta):**
1, cliente cancelou voluntariamente
0, cliente não cancelou voluntariamente

Avaliada em T+1 (por exemplo, em fev/20)

**Variáveis Explicativas:**

- Quantidade de minutos realizados em T0
- Tempo em que o cliente ingressou na empresa (meses)
- Quantidade de retenções nos últimos 6 meses
- Quantidade de produtos adicionais

Avaliadas em T0 (por exemplo, em jan/20)

Vamos fazer juntos?

RStudio

# Exercício: Estudo de Caso - *Churn* em Telefonia



## Análise das covariáveis

```
Minutos_realizados_T0 Tempo_casa      Qtd_retencao_6meses Qtd_prod
Min.   :   0.022      Min.   :  3.0   Min.   : 0.0000     Min.   : 0.0000
1st Qu.:  21.681      1st Qu.: 72.0   1st Qu.: 0.0000     1st Qu.: 0.0000
Median :  56.408      Median :101.0   Median : 0.0000     Median : 1.0000
Mean   : 130.239      Mean   :109.9   Mean   : 0.1872     Mean   : 0.7299
3rd Qu.: 136.458      3rd Qu.:149.0   3rd Qu.: 0.0000     3rd Qu.: 1.0000
Max.   :1474.066      Max.   :230.0   Max.   :45.0000     Max.   :30.0000
NA's   :22123
```

Como tratamos os *missing values*?

## Análise da variável resposta

Da base de **318.462** clientes, **0,86%** (2.766 clientes) **cancelaram** no próximo mês.

# Exercício: Estudo de Caso - *Churn* em Telefonia



| Quantidade de minutos realizados em T0 | Tempo em que o cliente ingressou na empresa (meses) | Quantidade de retenções nos últimos 6 meses | Quantidade de produtos adicionais |
|---|---|---|---|

| | 0 | 1 |
|---|---|---|
| [0,16.1] | 98.57 | 1.43 |
| (16.1,49.7] | 99.30 | 0.70 |
| (49.7,127] | 99.35 | 0.65 |
| (127,1.47e+03] | 99.31 | 0.69 |

Est. Qui-Quadrado: 394
Graus de liberdade: 3
P-valor: 0,00

| | 0 | 1 |
|---|---|---|
| [3,72] | 98.20 | 1.80 |
| (72,101] | 99.26 | 0.74 |
| (101,149] | 99.47 | 0.53 |
| (149,230] | 99.67 | 0.33 |

Est. Qui-Quadrado: 1221
Graus de liberdade: 3
P-valor: 0,00

| | 0 | 1 |
|---|---|---|
| 0 | 99.27 | 0.73 |
| (0,45] | 98.25 | 1.75 |

Est. Qui-Quadrado: 441
Graus de liberdade: 1
P-valor: 0,00

| | 0 | 1 |
|---|---|---|
| 0 | 98.97 | 1.03 |
| 1 | 99.06 | 0.94 |
| (1,30] | 99.86 | 0.14 |

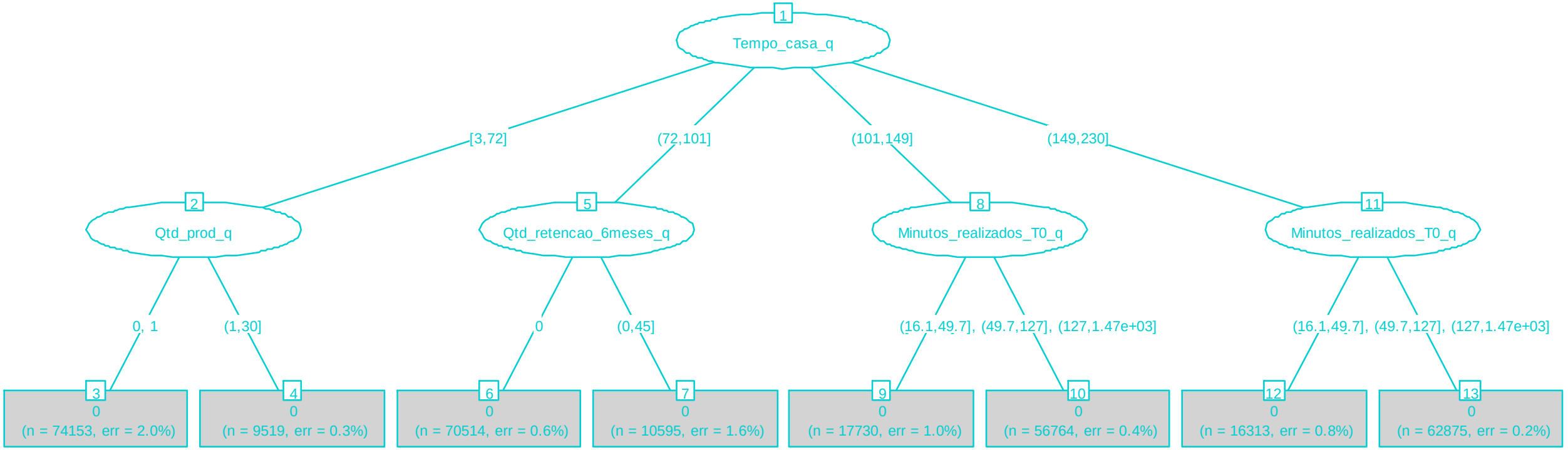
Est. Qui-Quadrado: 342
Graus de liberdade: 2
P-valor: 0,00

Para bases 'grandes', note que pelas características da Estatística Qui-Quadrado, o p-valor tende sempre a ser 'pequeno', próximo de zero.

- Pela **análise bivariada** das covariáveis *versus* a resposta, descritivamente, todas as variáveis parecem ter relação com o evento de cancelamento voluntário.
- O teste Qui-Quadrado identificou que todas as variáveis são significantes, ao nível de 5%, sendo a variável tempo em que o cliente está na empresa, a variável com o maior valor da estatística de Qui-quadrado.

# Teste Qui-Quadrado e Método CHAID






- Pelo **método CHAID**, a árvore apresentou 8 nós finais (perfis) de cliente, tendo 5 nós intermediários.
- Dado a proporção de cancelamento da base (0,86%), identificamos 3 perfis com propensão ao cancelamento e 5 sem propensão ao cancelamento.

# 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. Significa avaliar o resultado preditivo em comparação com a resposta observada no passado.

Assim, pode-se ter uma 'ideia' se o modelo, dadas as variáveis explicativas presentes, é capaz de explicar o evento resposta de forma satisfatória para que as regras do modelo possam ser aplicadas em bases mais recentes.

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

# 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 | | Total |
|---|---|---|---|---|
| | | 0 | 1 | |
| Variável Resposta Observada | 0 | VN | FP | VN+FP |
| | 1 | FN | VP | FN+VP |
| | Total | VN+FN | FP+VP | VN+FN+ FP+VP[1] |

[1] *VP*: verdadeiro-positivo; *VN*: verdadeiro-negativo; *FP*: falso-positivo e *FN*: falso-negativo

# Análise de Desempenho



**Tabela de Classificação** avaliada no ponto de corte:

| | | Variável Resposta Predita | | Total |
|---|---|---|---|---|
| | | 0 | 1 | |
| Variável Resposta Observada | 0 | VN | FP | VN+FP |
| | 1 | FN | VP | FN+VP |
| | Total | VN+FN | FP+VP | VN+FN+ FP+VP[1] |

[1] *VP*: verdadeiro-positivo; *VN*: verdadeiro-negativo; *FP*: falso-positivo e *FN*: falso-negativo

- Acurácia $$Acur = \frac{VP+VN}{VP+VN+FP+FN}$$
- Sensibilidade $$Sens = \frac{VP}{VP + FN}$$
- Especificidade $$Espec = \frac{VN}{FP + VN}$$

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

Esperamos que os índices sejam superiores a 50% (acima do acerto aleatório), sendo os valores mais próximos de 100% com maior poder preditivo.

Na prática, valores acima de 60% são considerados índices satisfatórios, sendo valores acima de 70%-75% já considerados com ótimo desempenho.

# Análise de Desempenho



**Tabela de Classificação** avaliada no ponto de corte (0,86%):

| Tabela de Classificação | | Predito 0 | Predito 1 | Total |
|---|---|---|---|---|
| Observado | 0 | 215.038 | 100.659 | **315.697** |
| | 1 | 947 | 1.819 | **2.766** |
| | **Total** | **215.985** | **102.478** | **318.463** |

- Acurácia: $\frac{215.038+1.819}{318.463} = 68\%$
- Sensibilidade $\frac{1.819}{2.766} = 66\%$
- Especificidade $\frac{215.038}{315.697} = 68\%$

Os índices de **acurácia**, **sensibilidade** e **especificidade,** apresentaram desempenho satisfatório, sendo possível predizer, de maneira geral, que quase 70% dos eventos de cancelamento são preditos pelo modelo corretamente.

Como o objetivo do negócio é predizer o evento 1, a sensibilidade passa a ser um índice importante a ser avaliado, com 66% de acerto do total de eventos 1 observados na base de dados.

# 4. Exercício para casa

# 4. Exercícios para casa

DATA DE ENTREGA 08/11/2020 | 1 EXERCÍCIO-CASE



CASE: *Churn* em Telefonia (10 pontos)

**Instruções importantes:**

- A lista vale nota (0-10) e deve ser entregue até 08/11/2020. Lista entregue até 15/11/2020 valerá 80% da nota. Posteriormente, não será mais aceita a lista para correção. Não serão aceitas listas parciais.
- O exercício será considerado como "realizado", quando tiver, além das análises, a interpretação do resultados.
- Disponibilização apenas do código, tabelas e gráficos mesmo se estiverem corretos, serão considerados na correção como "meio certo", pois o mais importante é a interpretação do resultado.
- Soluções técnicas "elegantes e mais completas" serão consideradas como ponto extra para o aluno (+0,5 na lista geral).
- A lista é individual. No caso de detecção de plágio, lista não será considerada para correção.

BOM ESTUDO ☺

# Case: *Churn* em Telefonia



Uma empresa de Telecom com 318.463 mil clientes está preocupada com o aumento do cancelamento voluntário suas assinaturas de telefonia fixa. No período de 1 mês, existiram 2.766 cancelamentos voluntários (0,87% dos clientes cancelaram voluntariamente). Apesar deste número parecer ser pequeno, em um ano se nada for feito, 10% da base de clientes deixarão a companhia. Por meio do estudo da base de dados, o objetivo da empresa é identificar os clientes com maior probabilidade de cancelar voluntariamente suas linhas telefônicas e fazer ações segmentadas de marketing de acordo com seu perfil. Serão estudadas as características transacionais dos clientes para explicar o evento de cancelamento voluntário. O gestor da área de CRM, gostaria de saber qual o perfil dos clientes que mais cancelam, quais suas características e se essa regra é eficaz para predizer o comportamento daqueles clientes que cancelam no próximo mês.

Fonte: base simulada, inspirada em problemas reais de consultoria da Profª Karin Ayumi Tamura.

Dando continuidade ao exercício realizado em sala de aula, considere agora **3 níveis da árvore**. Utilizar o mesmo código disponibilizado e mudar apenas: *controle <- chaid_control(maxheight = 3)*. Responda as seguintes perguntas.

(a) Identifique a quantidade de nós finais e nós intermediários.
(b) Calcule a frequência relativa de cada nó final e a proporção da variável resposta por nó final.
(c) Identifique quantos nós finais são propensos ao evento de interesse e qual a representatividade do grupo propenso.
(d) Interprete todos os perfis de clientes propensos ao cancelamento, por ordem de maior propensão para menor.
(e) Avalie o desempenho do modelo pelos índices sensibilidade, especificidade e acurácia.
(f) Comparando com o modelo rodado em sala de aula (slide 44), qual modelo você recomendaria para área de negócios? Apoie sua decisão de acordo com o desempenho do modelo e interpretação do negócio.
(g) Qual é o perfil de clientes que a empresa de Telefonia não precisa se preocupar, pois o cancelamento é muito baixo? Quanto ele representa da base e qual seu percentual de cancelamento?

Nos dois próximos slides são fornecidos os outputs para interpretação dos principais resultados.

```
Model formula:
resposta ~ Minutos_realizados_T0_q + Tempo_casa_q + Qtd_retencao_6meses_q + Qtd_prod_q

Fitted party:
[1] root
|   [2] Tempo_casa_q in [3,72]
|   |   [3] Qtd_prod_q in 0, 1
|   |   |   [4] Qtd_retencao_6meses_q in 0: 0 (n = 63287, err = 1.7%)
|   |   |   [5] Qtd_retencao_6meses_q in (0,45]: 0 (n = 10866, err = 3.6%)
|   |   [6] Qtd_prod_q in (1,30]
|   |   |   [7] Minutos_realizados_T0_q in [0,16.1]: 0 (n = 797, err = 1.0%)
|   |   |   [8] Minutos_realizados_T0_q in (16.1,49.7], (49.7,127], (127,1.47e+03]: 0 (n = 8722, err = 0.2%)
|   [9] Tempo_casa_q in (72,101]
|   |   [10] Qtd_retencao_6meses_q in 0
|   |   |   [11] Qtd_prod_q in 0: 0 (n = 41414, err = 0.8%)
|   |   |   [12] Qtd_prod_q in 1: 0 (n = 22253, err = 0.5%)
|   |   |   [13] Qtd_prod_q in (1,30]: 0 (n = 6847, err = 0.1%)
|   |   [14] Qtd_retencao_6meses_q in (0,45]
|   |   |   [15] Minutos_realizados_T0_q in [0,16.1]: 0 (n = 2836, err = 3.6%)
|   |   |   [16] Minutos_realizados_T0_q in (16.1,49.7], (49.7,127]: 0 (n = 4867, err = 1.0%)
|   |   |   [17] Minutos_realizados_T0_q in (127,1.47e+03]: 0 (n = 2892, err = 0.6%)
|   [18] Tempo_casa_q in (101,149]
|   |   [19] Minutos_realizados_T0_q in [0,16.1]
|   |   |   [20] Qtd_retencao_6meses_q in 0: 0 (n = 15720, err = 0.8%)
|   |   |   [21] Qtd_retencao_6meses_q in (0,45]: 0 (n = 2010, err = 2.3%)
|   |   [22] Minutos_realizados_T0_q in (16.1,49.7], (49.7,127], (127,1.47e+03]
|   |   |   [23] Qtd_prod_q in 0: 0 (n = 24225, err = 0.5%)
|   |   |   [24] Qtd_prod_q in 1: 0 (n = 20653, err = 0.4%)
|   |   |   [25] Qtd_prod_q in (1,30]: 0 (n = 11886, err = 0.1%)
|   [26] Tempo_casa_q in (149,230]
|   |   [27] Minutos_realizados_T0_q in [0,16.1]
|   |   |   [28] Qtd_retencao_6meses_q in 0: 0 (n = 14935, err = 0.6%)
|   |   |   [29] Qtd_retencao_6meses_q in (0,45]: 0 (n = 1378, err = 2.7%)
|   |   [30] Minutos_realizados_T0_q in (16.1,49.7], (49.7,127], (127,1.47e+03]
|   |   |   [31] Qtd_retencao_6meses_q in 0: 0 (n = 55766, err = 0.2%)
|   |   |   [32] Qtd_retencao_6meses_q in (0,45]: 0 (n = 7109, err = 0.6%)

Number of inner nodes:    13
Number of terminal nodes: 19
```

```
>#Mostra a tabela de desempenho: Predito x Resposta observada
> (tabela_desempenho<-table(telefonia$resposta,telefonia$predict))

        0      1
  0 231374  84323
  1   1048   1718
> #Calcula as medidas de desempenho: Sensibilidade, Especificidade e Acurácia
> (sensibilidade<-tabela_desempenho[2,2]/sum(tabela_desempenho[2,]))
[1] 0.6211135
> (especificidade<-tabela_desempenho[1,1]/sum(tabela_desempenho[1,]))
[1] 0.732899
> (accuracia<- sum(tabela_desempenho[1,1]+tabela_desempenho[2,2])/318463)
[1] 0.731928
>
```