

Curso Aprendizado de Máquina

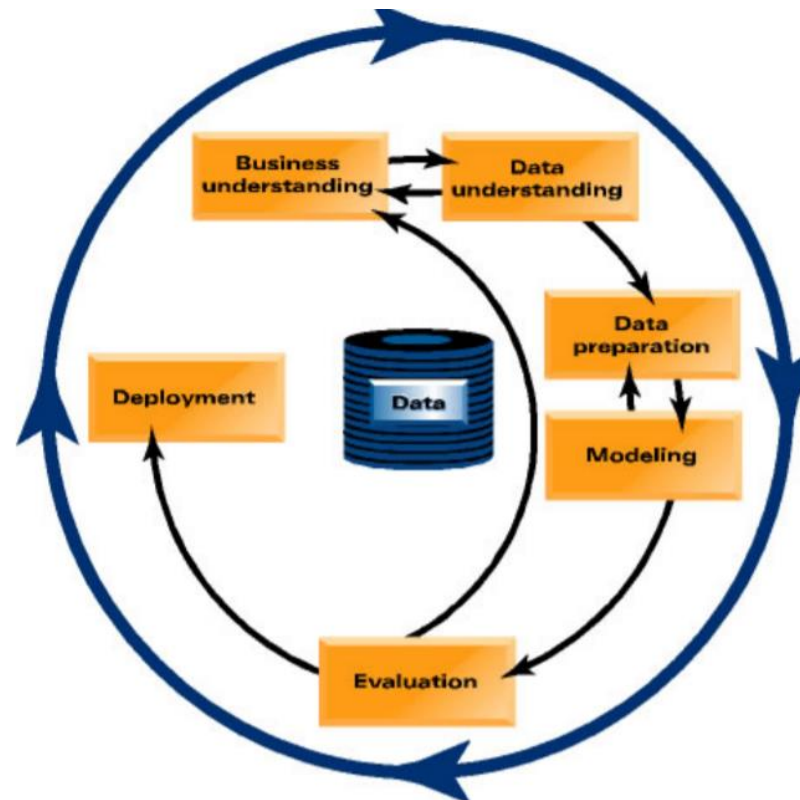
Alguns pontos da ultima aula

CONCEITOS FUNDAMENTAIS

- Extrair conhecimento útil dos dados para resolver os negócios problemas podem ser tratados sistematicamente seguindo um processo com etapas razoavelmente bem definidas.
- Formular soluções de mineração de dados e avaliar o resultados envolve pensar cuidadosamente sobre o contexto em quais eles serão usados.”

PROCESSO DE MINERAÇÃO DE DADOS

- Cross Industry Standard Process for Data Mining – CRISP-DM
- Processo padrão entre indústrias para mineração de dados



Introdução à modelagem preditiva

da correlação à segmentação supervisionada

Introdução à modelagem preditiva: da correlação à segmentação supervisionada

- Na aula anterior falamos de modelos e modelagem em alto nível.
- Agora vamos tratar um dos principais tópicos do aprendizado de máquina e da mineração de dados: modelagem preditiva.

Introdução à modelagem preditiva: da correlação à segmentação supervisionada

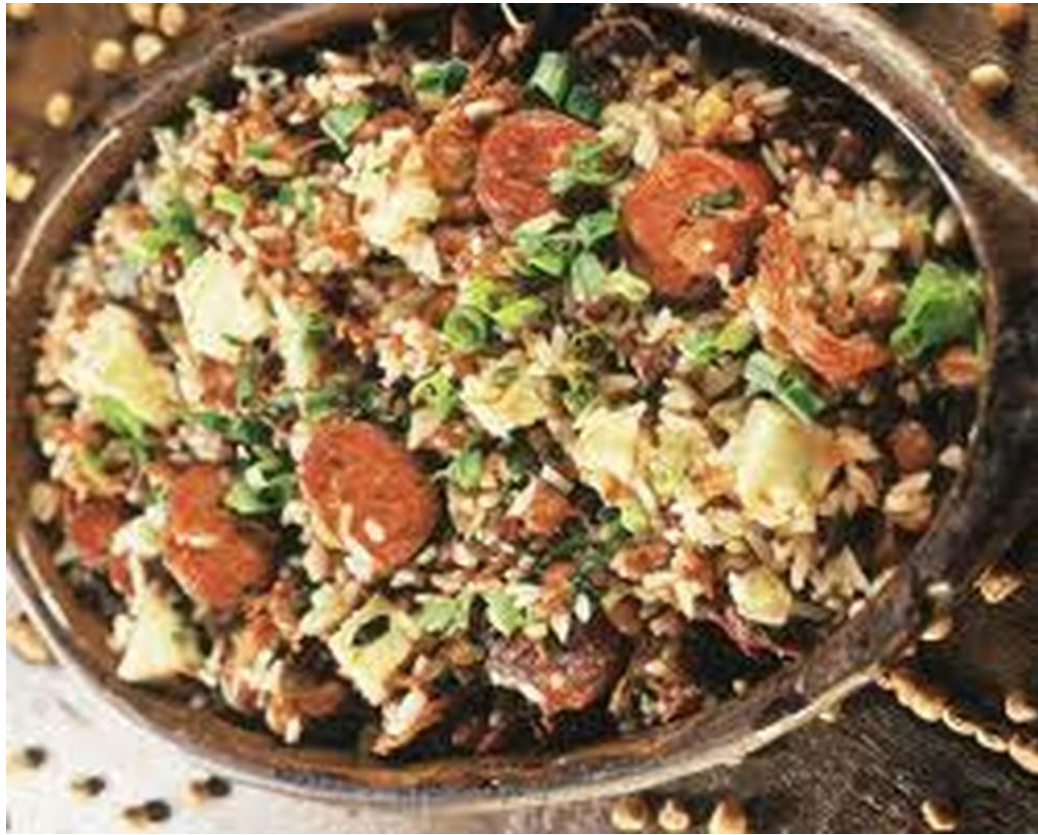
- O objetivo dessa previsão pode ser algo que gostaríamos de evitar, como:
- Quais clientes provavelmente deixarão a empresa quando seus contratos expirarem?
- Quais contas foram fraudadas?
- Quais clientes em potencial provavelmente não pagarão os saldos das suas contas (write-offs, como a inadimplência na conta telefônica ou no saldo do cartão de crédito)?
- Quais páginas da Web contêm conteúdo censurável?
- Quais consumidores têm maior probabilidade de responder a um anúncio ou oferta especial ?
- Quais páginas da Web são mais apropriadas para uma consulta de pesquisa?

Entendendo nos problema: Observe esta refeição com estilo japonês



- Classifique os componentes (ingredientes) dessa refeição até onde você conseguir

Entendendo nos problema: Observe esta refeição um Baião de Dois



- Classifique os componentes (ingredientes) dessa refeição até onde você conseguir

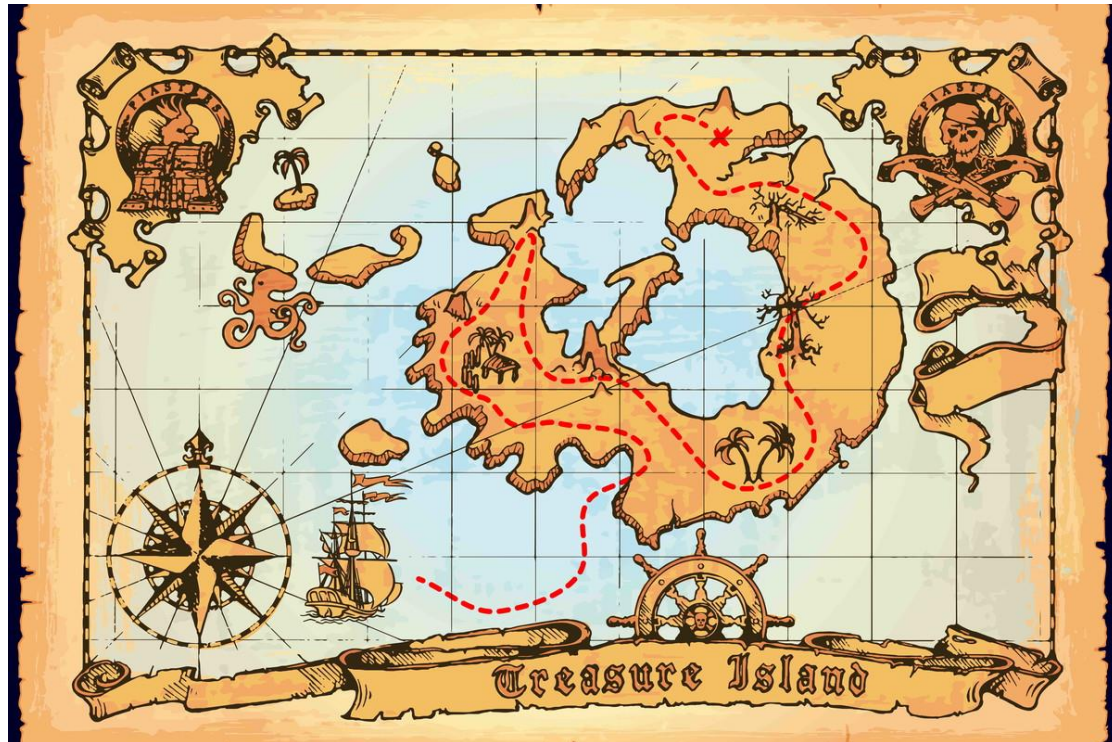
O conceito de informação

- No processo de discutir a segmentação supervisionada, apresentamos uma das idéias fundamentais é:
- Encontrar ou selecionar **variáveis importantes** ou **informativas** ou **"atributos"** das entidades descritas pelos **dados**.
- O que exatamente significa ser “informativo”?

O que exatamente significa ser “informativo”?

- Varia entre aplicativos, mas geralmente, a informação é uma quantidade que **reduz a incerteza** sobre algo.
- Portanto, se um velho pirata me fornece informações sobre onde está o tesouro dele, o que não significa que eu saiba com certeza onde está, isso significa apenas que minha incerteza sobre o local está escondida é reduzida.
- Quanto melhor a informação, mais minha incerteza é reduzida.

Informativo ?



A idéia de alvo

- A noção de mineração de dados "supervisionada" está relacionada a aprender através de exemplos.
- Uma chave para o aprendizado supervisionado de dados é que temos uma **quantidade alvo** que gostaríamos de prever ou entender melhor.
- Geralmente, essa quantidade é desconhecida ou no momento em que gostaríamos de tomar uma decisão:
- **Um cliente abandona logo após o vencimento do contrato ?**
- **Quais contas foram fraudadas?**

A idéia de correlação

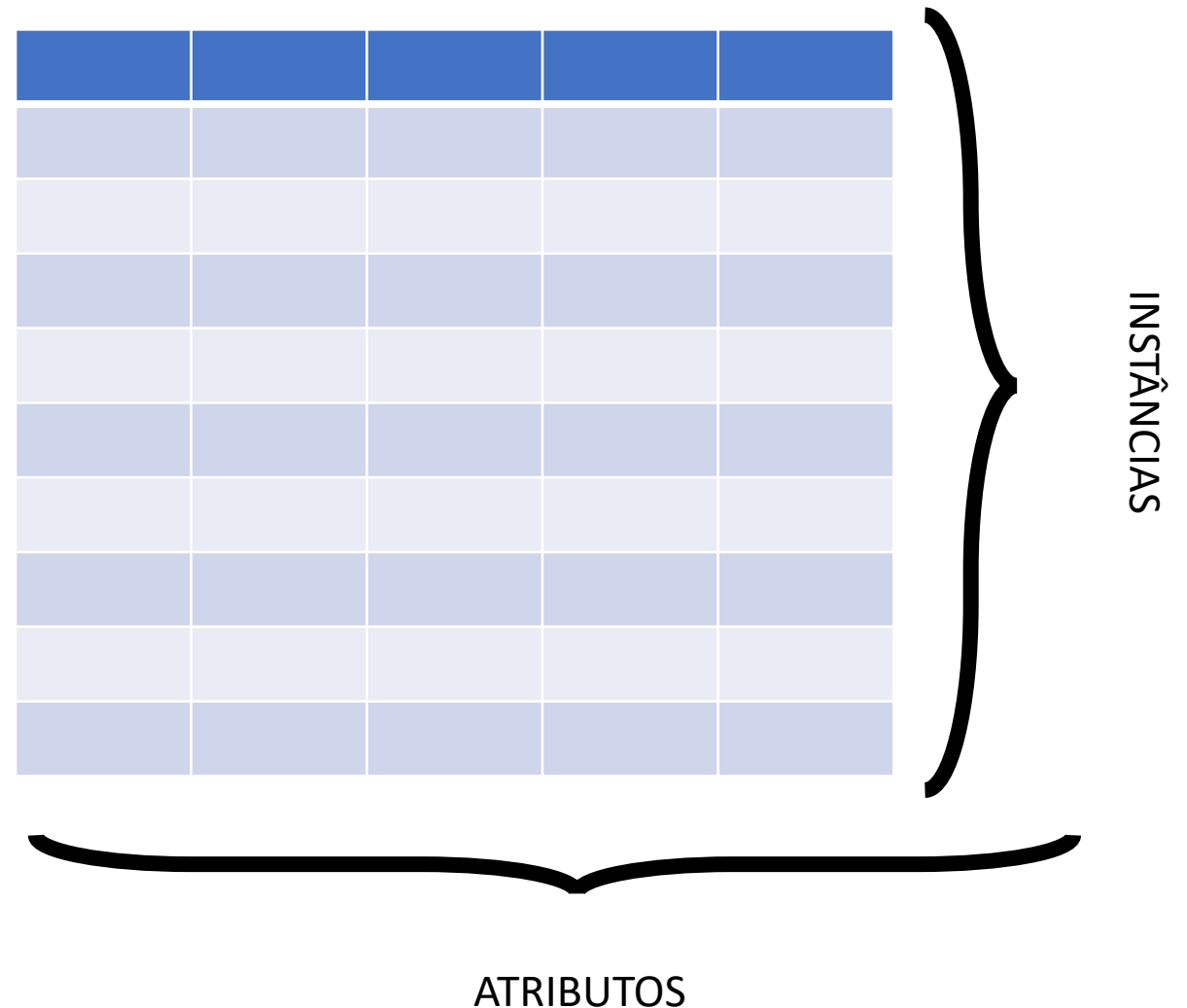
- Ter uma variável alvo que clarifica nossa noção de encontrar atributos informativos: há uma ou mais outras variáveis que reduzem nossa incerteza sobre o valor do alvo?
- Isso também fornece uma aplicação analítica comum da noção geral de correlação discutida acima: gostaríamos de encontrar atributos conhecíveis que se correlacionem com o alvo de interesse - **que reduzam nossa incerteza**.
- Apenas encontrar essas variáveis correlacionadas pode fornecer informações importantes sobre o problema de negócios.
- Encontrar atributos informativos também é útil para nos ajudar a lidar com bancos de dados e fluxos de dados cada vez maiores.
- Conjuntos de dados muito grandes apresentam problemas computacionais para técnicas analíticas, especialmente quando o analista não tem acesso a computadores de alto desempenho.

Da aula passada

- Vamos rever o exemplo da aula passada

A Estrutura de Dados de um Modelo

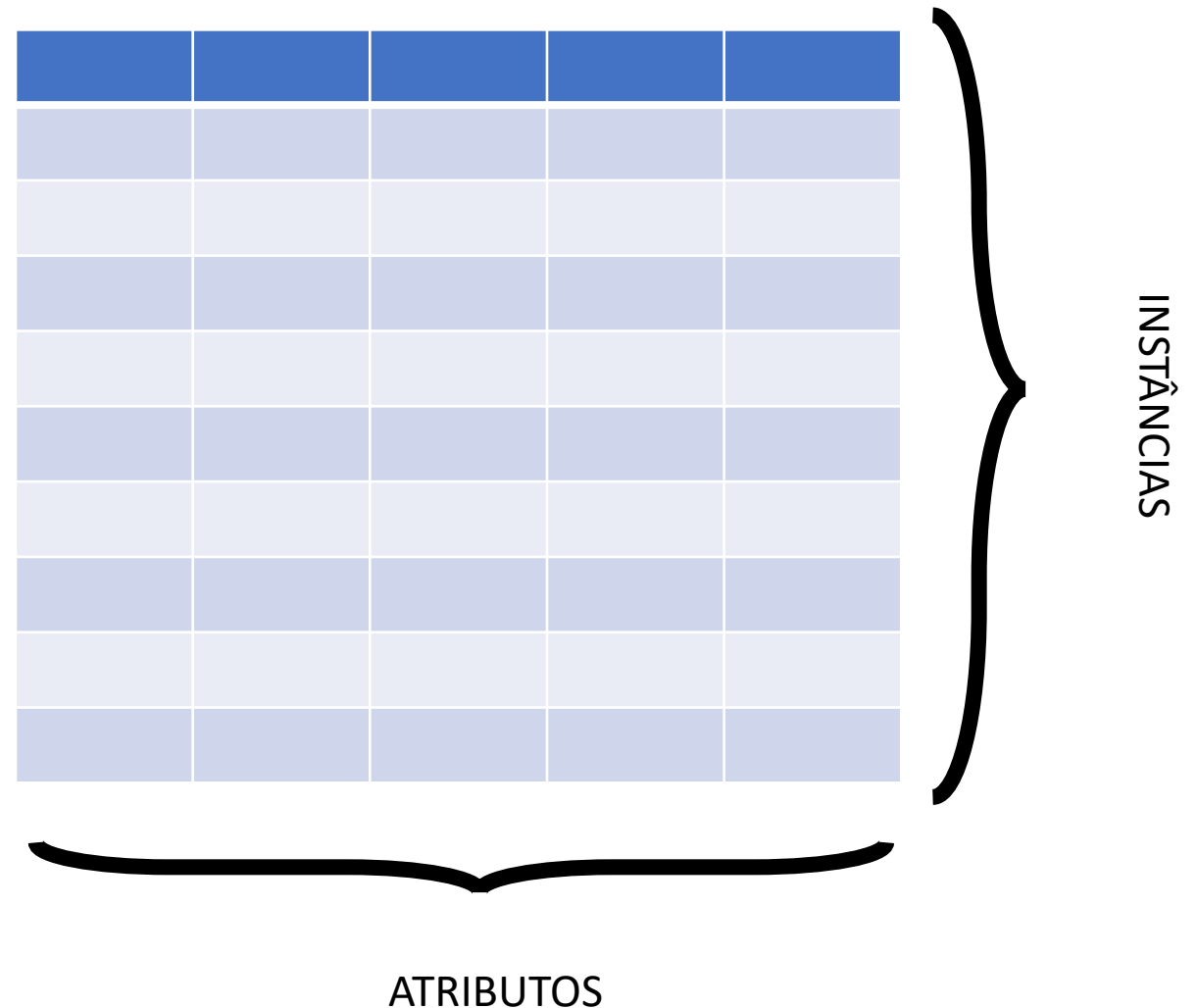
- Em geral a estrutura de dados de um modelo é entendida na forma de uma matriz (uma tabela)
- As linhas são chamadas instâncias
- As colunas são chamadas atributos



A Estrutura de Dados de um Modelo

ATRIBUTOS

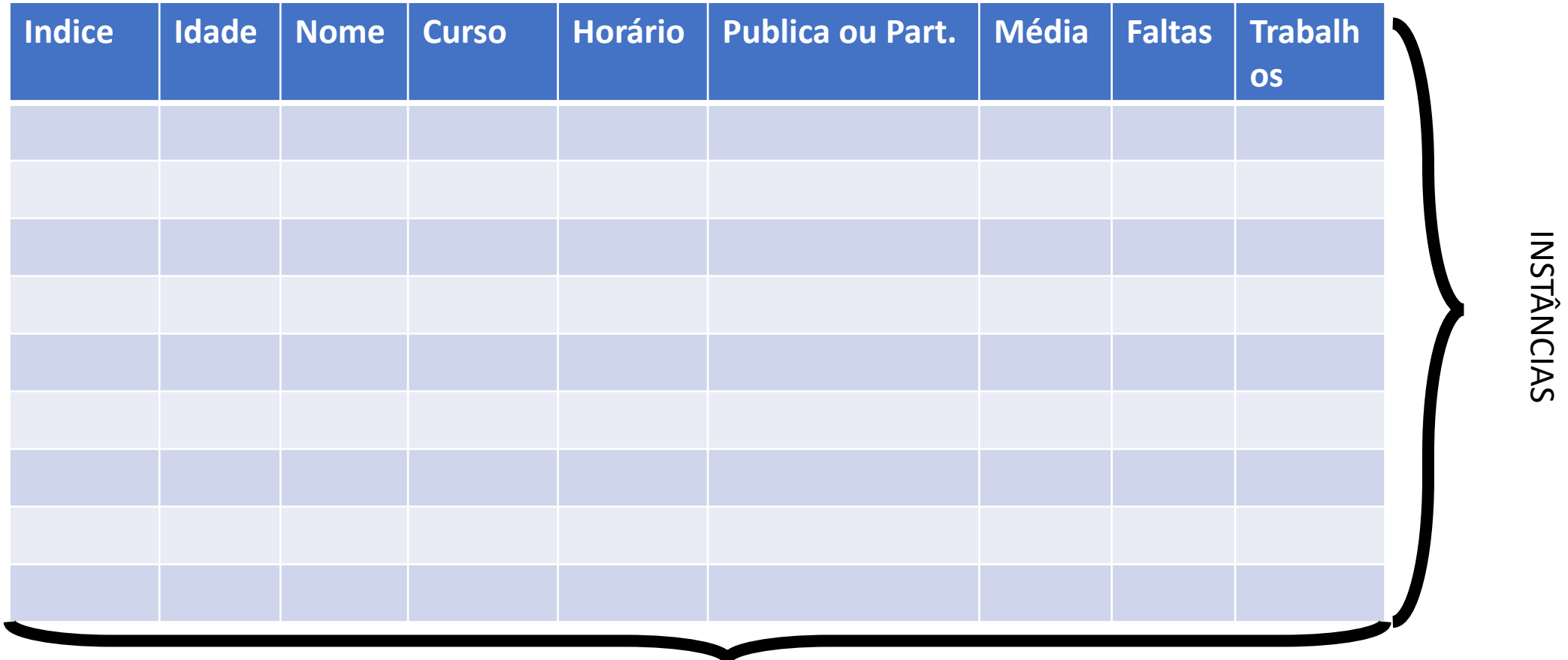
- Os atributos podem ser entendidos como as características que temos disponíveis para entender cada um dos dados de nosso conjunto de dados ou dataset
- Por exemplo num conjunto de dados ALUNOS suas características seriam:
 - índice
 - Idade
 - Nome
 - Local de nascimento
 - Curso
 - Horário
 - Se veio de uma escola publica ou particular
 - Média
 - Faltas
 - Número de trabalhos entregues
 - Etc..



A Estrutura de Dados de um Modelo

ATRIBUTOS

Indice	Idade	Nome	Curso	Horário	Publica ou Part.	Média	Faltas	Trabalhos



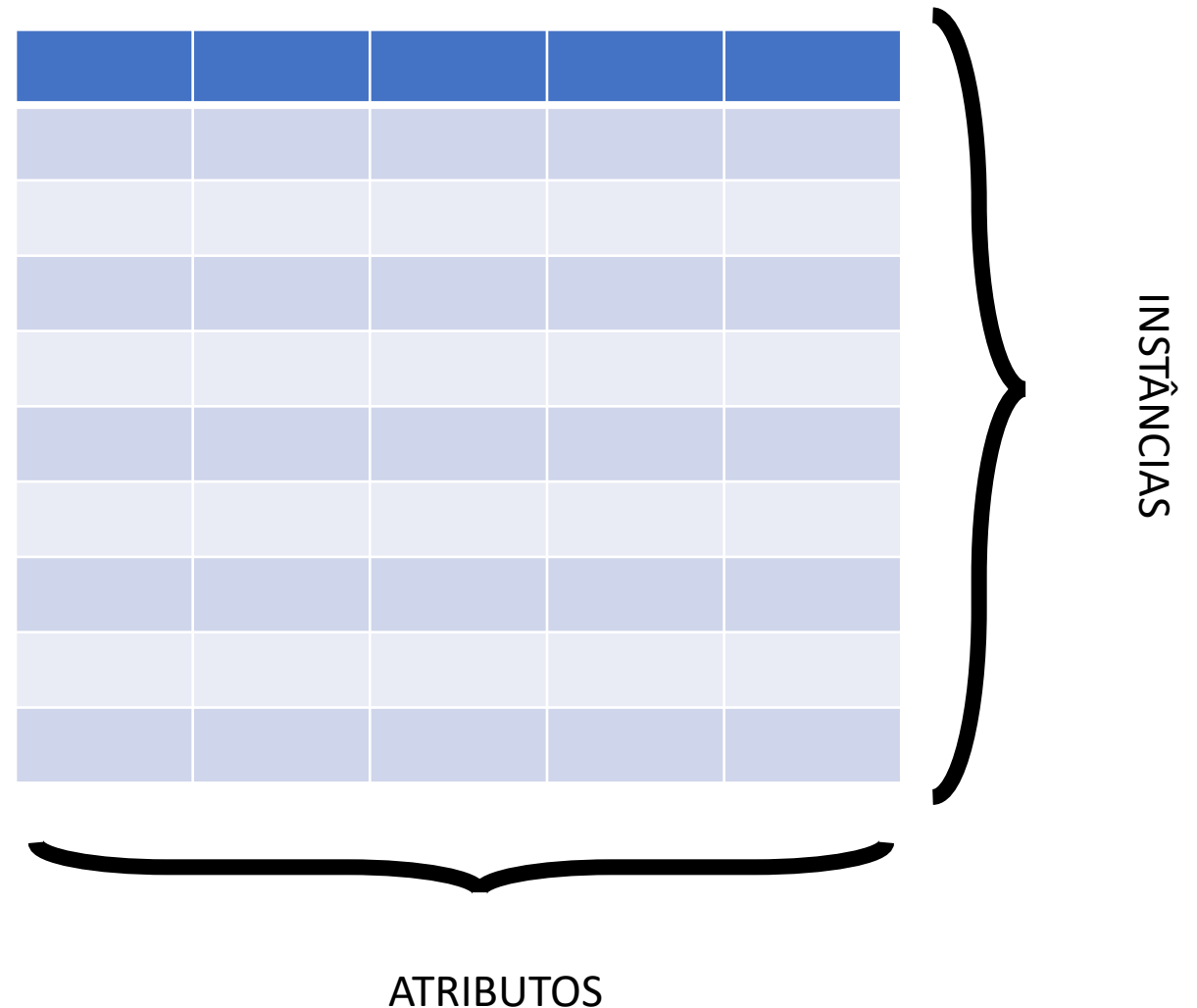
ATRIBUTOS

INSTÂNCIAS

A Estrutura de Dados de um Modelo

INSTÂNCIAS

- Os atributos podem ser entendidos como as características que temos disponíveis para entender cada um dos dados de nosso conjunto de dados ou dataset
- Por exemplo num conjunto de dados PARA CADA ALUNO teríamos as instâncias:
 - Índice [0,1,2,3,4,5,6,7,8,...]
 - Idade [21,22,31,27, 40, 49, 31, 22, ETC..]
 - Nome [JOÃO, JOSÉ, MARIA, CLAUDIO, TOMÁS, JACÓ, IRINEU,..]
 - Local de nascimento [São Paulo, Presidente Prudente, Osasco, Guarulhos, São Paulo, São Bernado, etc..]
 - Curso [CC, TADS, SI, CC, TADS, SI...]
 - Horário [D, N, D, N, N, N , N, etc...
 - Se veio de uma escola publica ou particular [Pb, Pa, Pb, Pb, PB, Pa, Pb, etc...
 - Média [10, 9, 9.5, 7, 7.5, 6.6, 8.1, etc..]
 - Faltas [2, 3 ,4 ,2 ,2 3 ,3 . 10, 3, 5, etc..]
 - Número de trabalhos entregues [10,8,10,9, 6, 5, 6, 8,etc..]
 - Etc..



Este exemplo completo

Indice	Idade	Nome	Curso	Horário	Publica ou Part.	Média	Faltas	Trabalho s
0	27	Artur	CC	D	PB	6,5	3	10
1	20	Maria	TADS	D	PT	8,9	2	7
2	35	Antonio	TADS	N	PB	7,9	4	9
3	47	Cleber	CC	N	PT	6,2	4	10
4	46	Silas	SI	N	PB	6,3	2	6
5	30	Mateus	TADS	N	PB	6,4	4	7
6	40	Rafael	TADS	N	PB	7,0	5	8
7	20	Edson	SI	N	PT	5,9	2	5
8	41	Ana	SI	N	PB	5,9	3	6
9	28	Paula	SI	N	PB	9,9	2	8

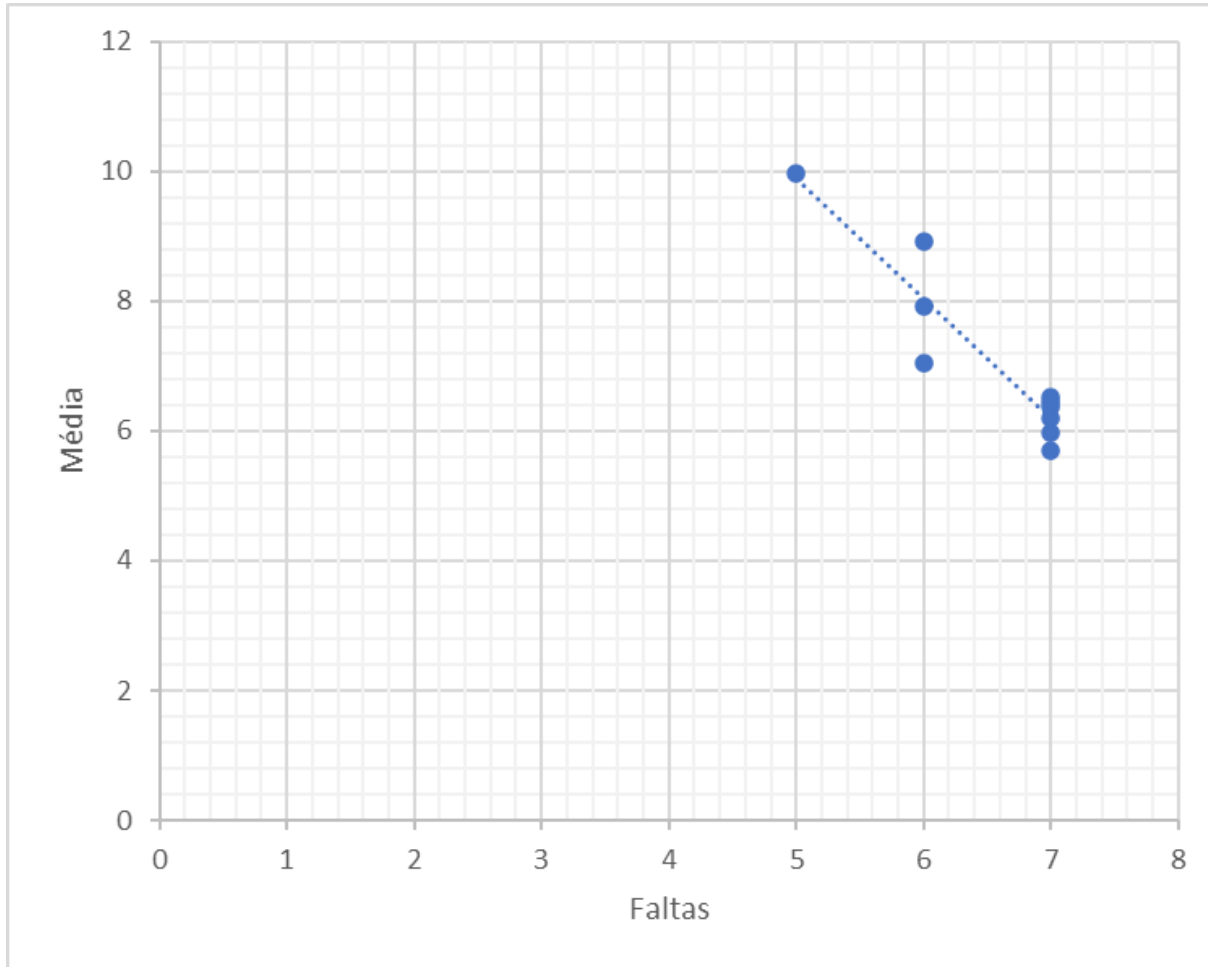
Colunas como variáveis para os modelos

- Por exemplo posso estar interessado se existe uma relação entre Média e Faltas
- Então eu seleciono estas duas colunas e desenvolvo um modelo que vai avaliara esta relação

Indice	Média
0	6,5
1	8,9
2	7,9
3	6,2
4	6,3
5	6,4
6	7,0
7	5,9
8	5,9
9	9,9

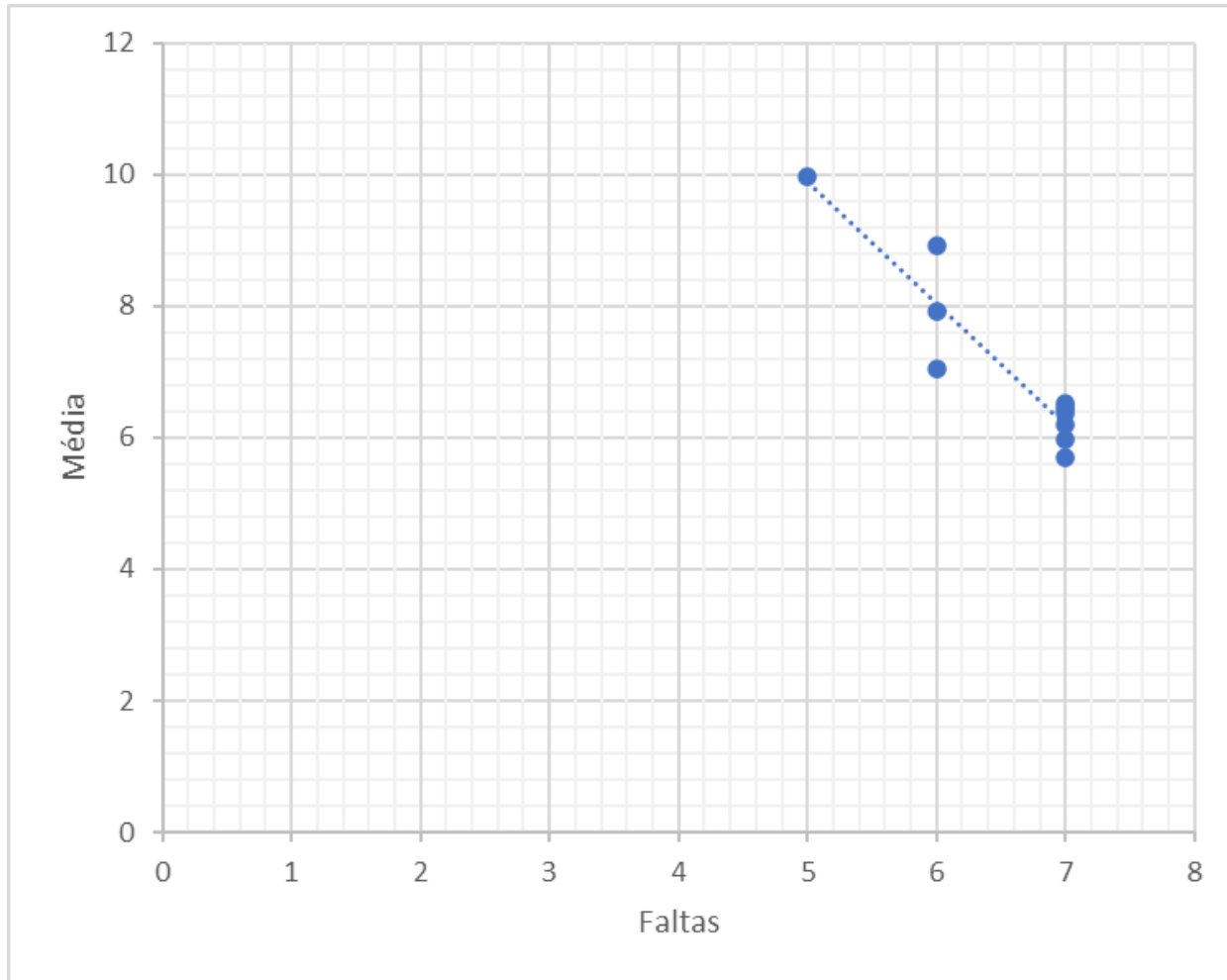
Indice	Faltas
0	3
1	2
2	4
3	4
4	2
5	4
6	5
7	2
8	3
9	2

Um primeiro modelo



- Temos um Gráfico onde X é o número de faltas dos alunos e y é a média
- O que podemos dizer ?

Um primeiro modelo



- Temos um Gráfico onde X é o número de faltas dos alunos e y é a média
- O que podemos dizer ?
- Que quando as faltas aumentam a média diminui
- Portanto nosso primeiro modelo $y = f(X)$ é:
- $Média = f(Faltas)$

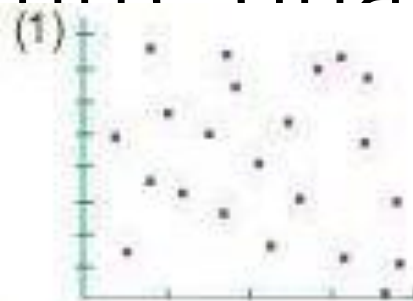
Visualizando correlação entre duas variáveis

Correlação

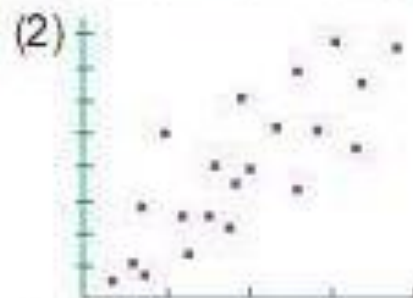
o Coeficiente de Correlação

- o Mede o grau de correlação linear entre duas variáveis numéricas

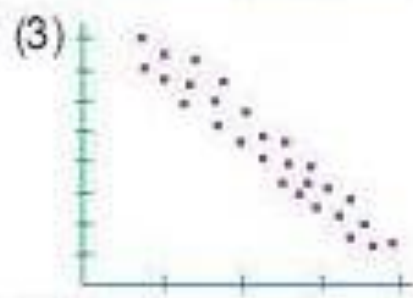
$$r = \frac{\sum XY - \frac{(\sum X)(\sum Y)}{n}}{\sqrt{\left[\sum X^2 - \frac{(\sum X)^2}{n}\right] \left[\sum Y^2 - \frac{(\sum Y)^2}{n}\right]}}$$



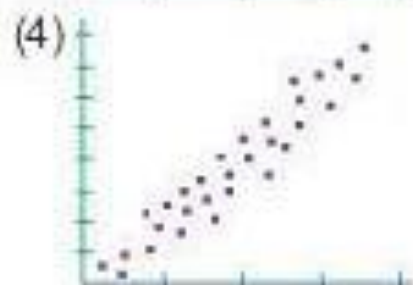
(3) Forte correlação negativa



(4) Forte correlação positiva

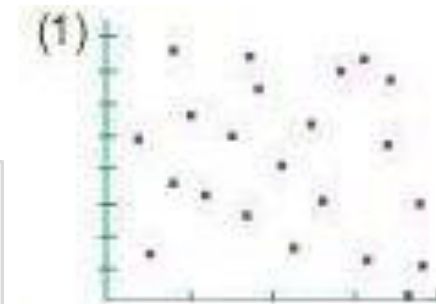
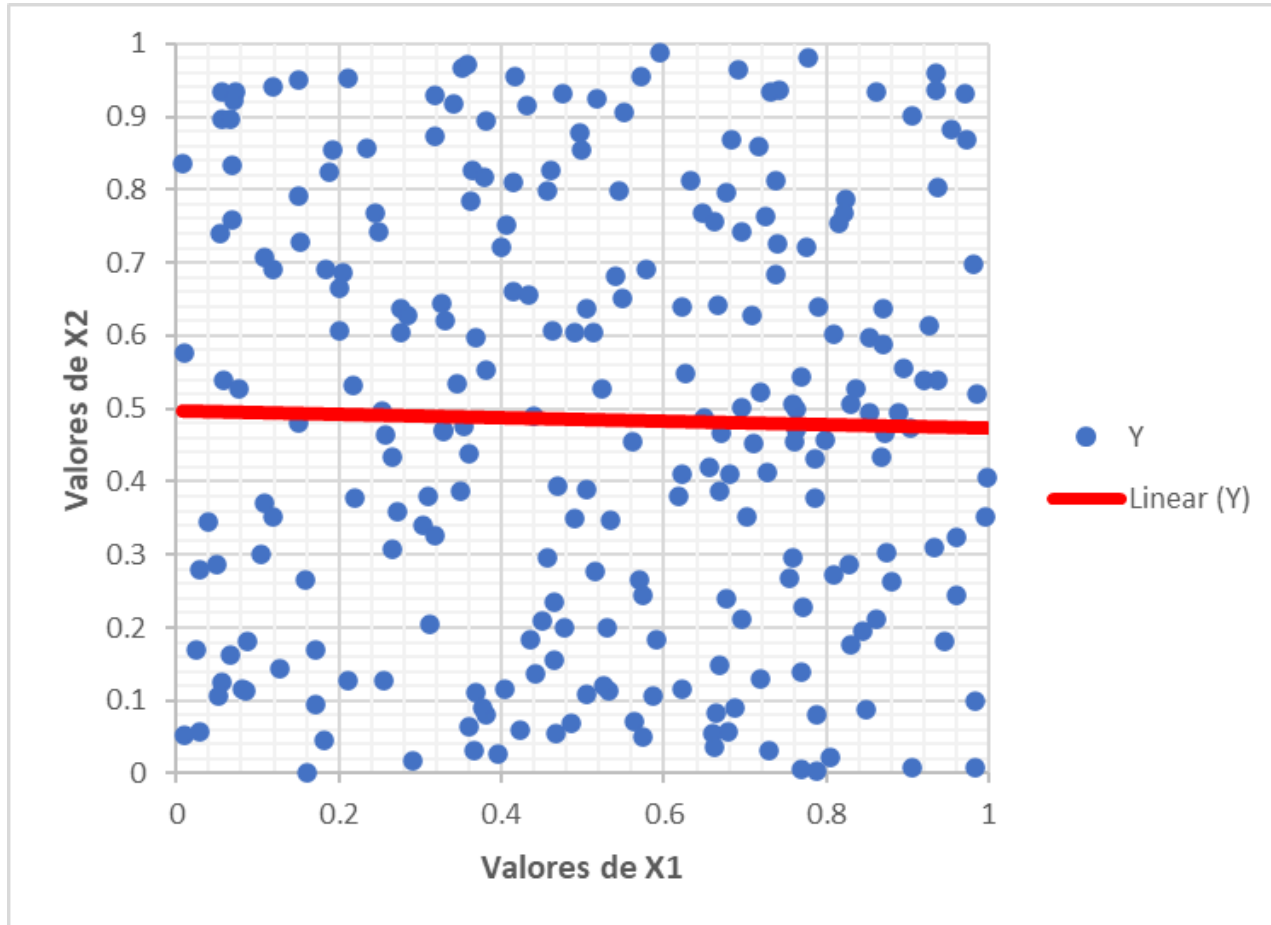


(2) Fraca correlação positiva

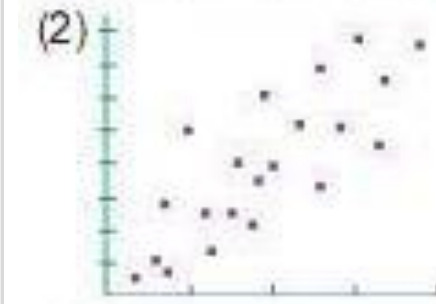


(1) Sem correlação

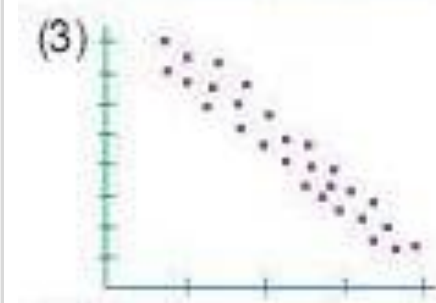
Que tipo de correlação temos para estes dados



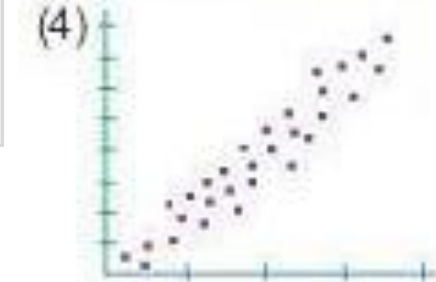
(F) Forte correlação negativa



(F) Forte correlação positiva

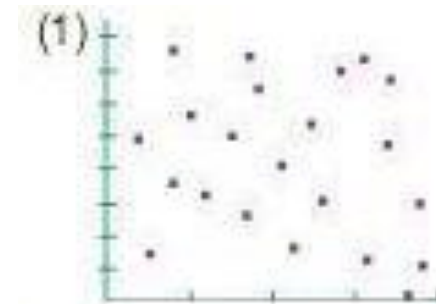
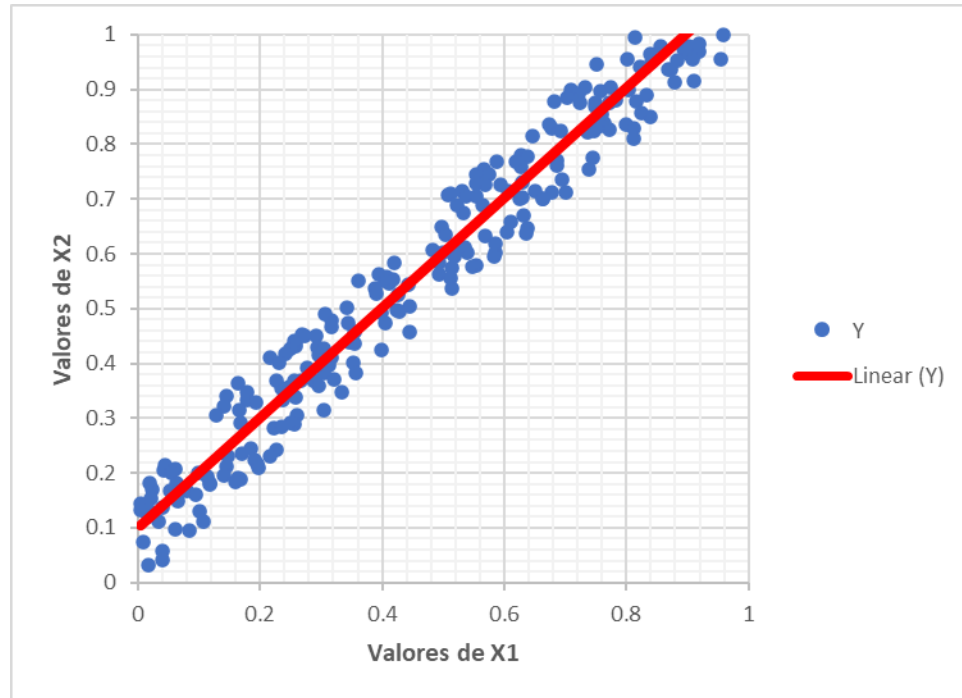


(F) Fraca correlação positiva

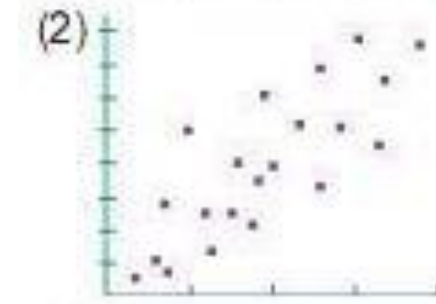


(V) Sem correlação

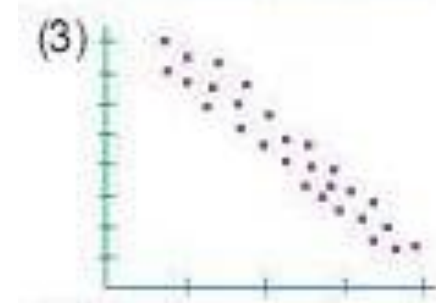
Que tipo de correlação temos para estes dados



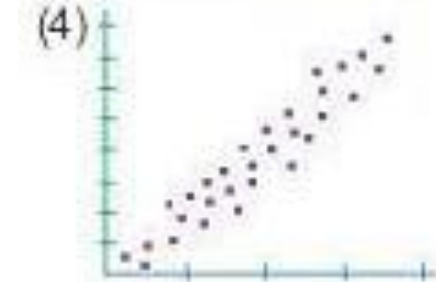
(**F**) Forte correlação negativa



(**V**) Forte correlação positiva

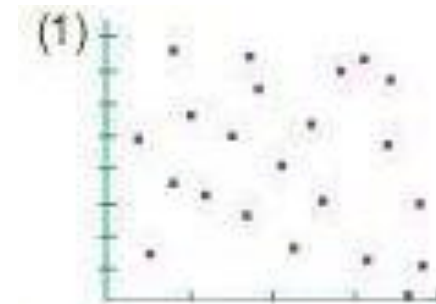
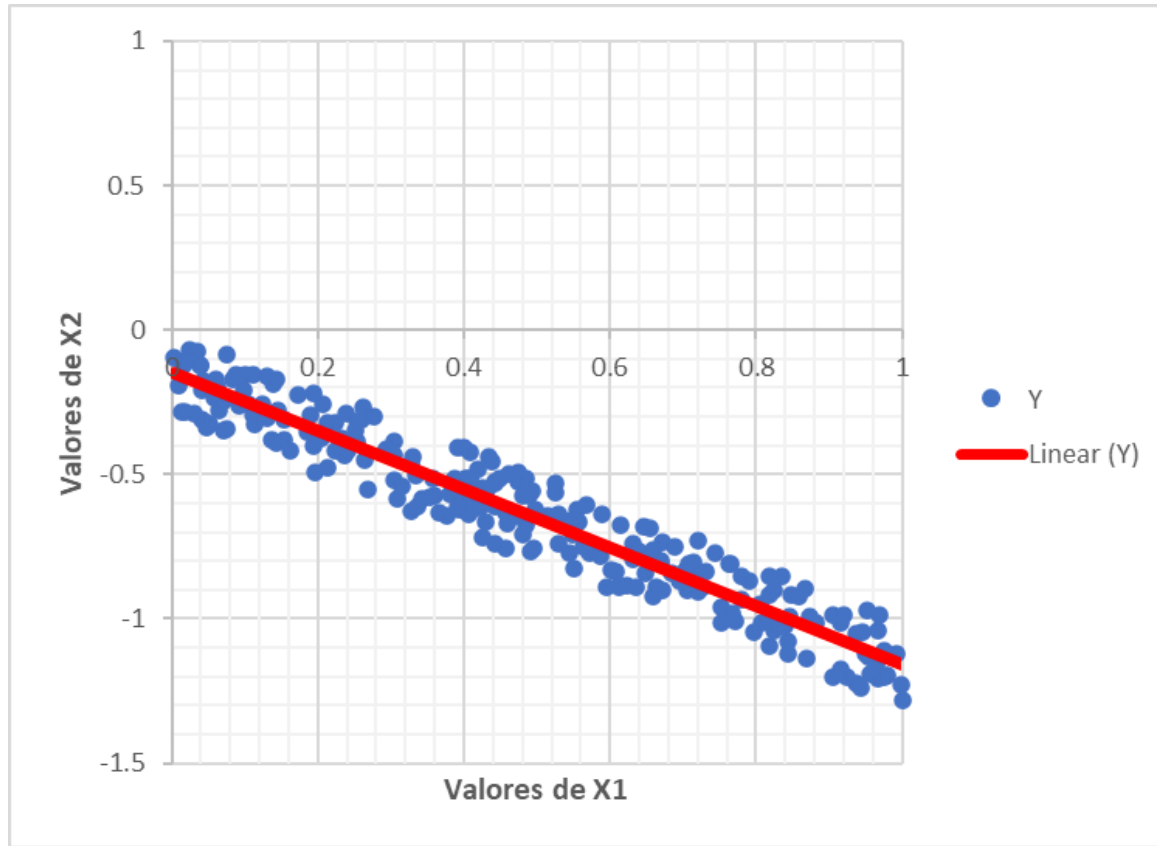


(**F**) Fraca correlação positiva

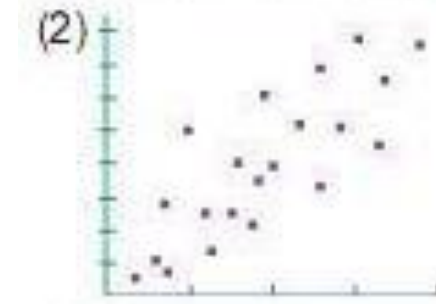


(**F**) Sem correlação

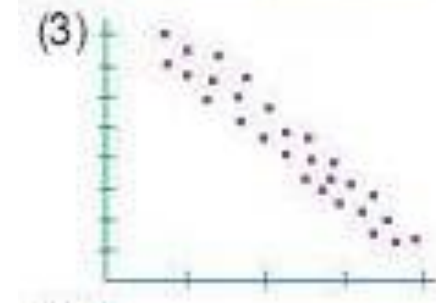
Que tipo de correlação temos para estes dados



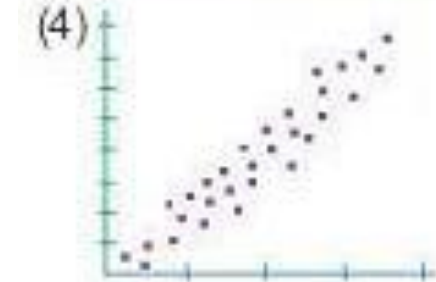
(**V**) Forte correlação negativa



(**F**) Forte correlação positiva

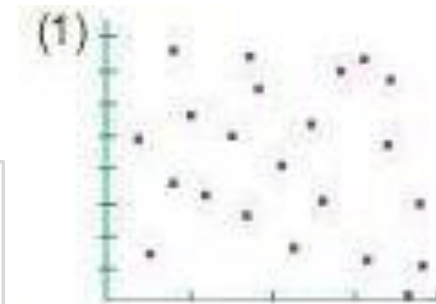
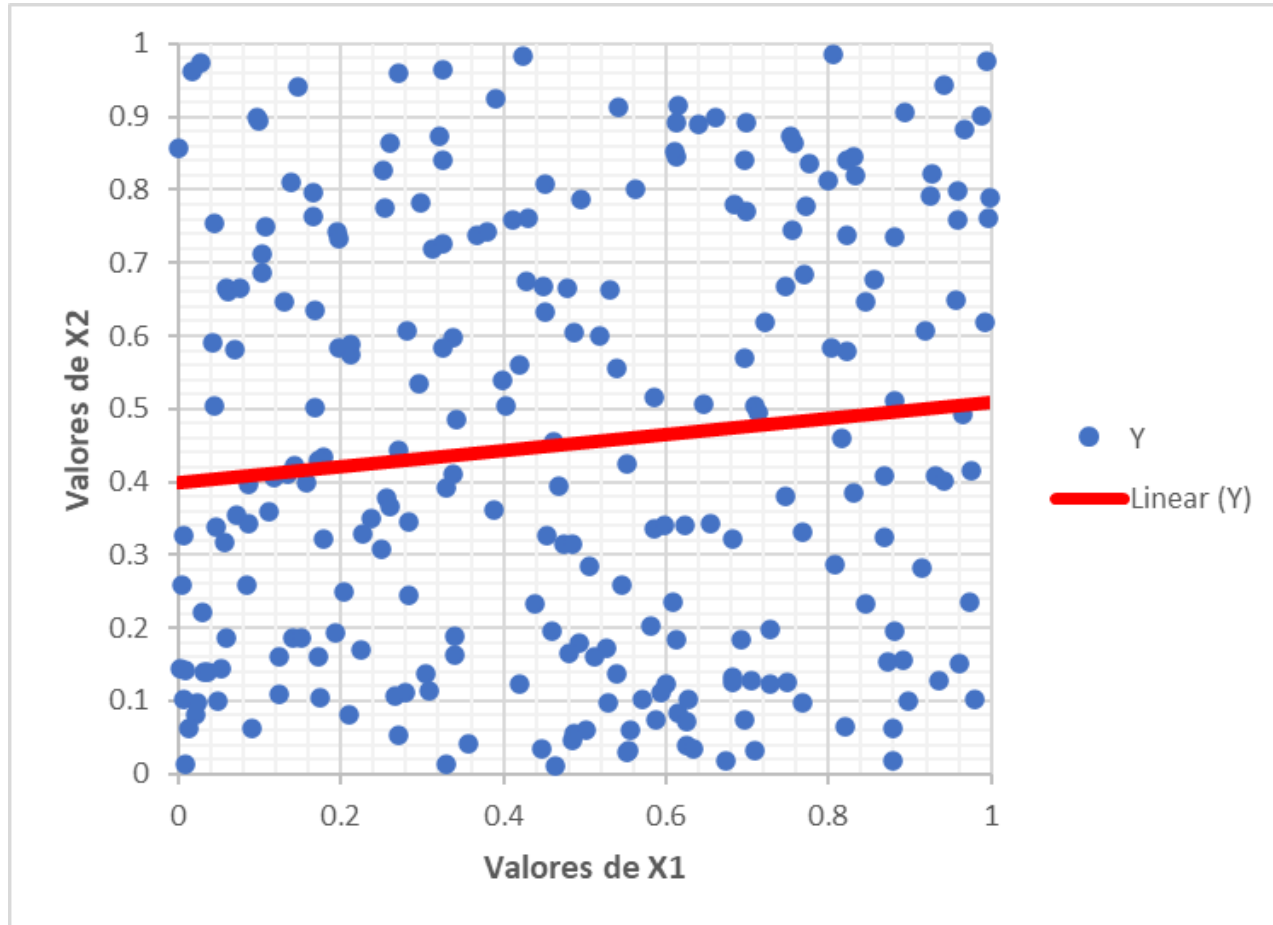


(**F**) Fraca correlação positiva

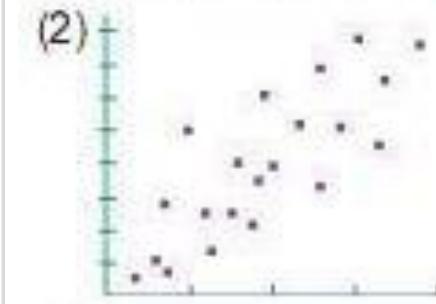


(**F**) Sem correlação

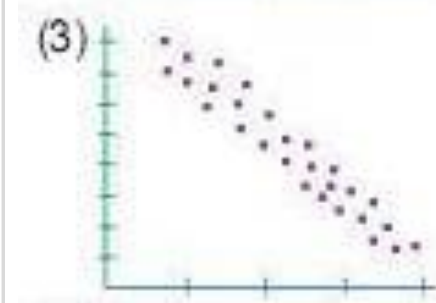
Que tipo de correlação temos para estes dados



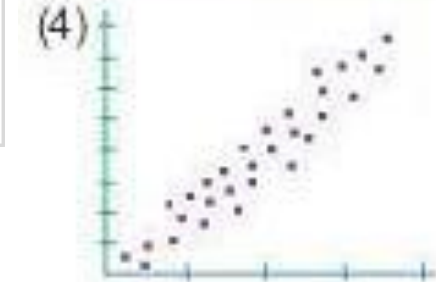
(F) Forte correlação negativa



(F) Forte correlação positiva



(V) Fraca correlação positiva



(F) Sem correlação

Conceitos de análise

- Um método testado e comprovado para analisar conjuntos de dados muito grandes é o primeiro a selecionar um subconjunto dos dados a serem analisados.
- A seleção de atributos informativos fornece um método "inteligente" para selecionar um subconjunto informativo dos dados.
- Além disso, a seleção de atributos antes da modelagem orientada a dados pode aumentar a precisão da modelagem, por razões que discutiremos mais a frente.

Conceitos de análise

- Encontrar atributos informativos também é a base para uma técnica de modelagem preditiva amplamente usada chamada indução em árvore, que apresentaremos mais a frente como uma aplicação desse conceito fundamental.
- A indução em árvore incorpora a ideia de segmentação supervisionada de maneira elegante, selecionando repetidamente atributos informativos.

Modelos, indução e previsão

- De um modo geral, um modelo é uma representação simplificada da realidade criada para servir a um propósito.
- É simplificado com base em algumas suposições sobre o que é e o que não é importante para uma finalidade específica, ou às vezes com base em restrições de informações ou capacidade de tratamento.
- Por exemplo, um mapa é um modelo do mundo físico. Ele abstrai uma quantidade tremenda de informações que o cartógrafo considerou irrelevantes para sua finalidade. Ele preserva e às vezes simplifica ainda mais as informações relevantes.
- Várias áreas têm tipos de modelos conhecidos: um projeto arquitetônico, um protótipo de engenharia, o modelo Black-Scholes de precificação de opções e assim por diante.
- Cada uma dessas simplificações elimina detalhes que não são relevantes para seu objetivo principal e mantém os que são.

Modelos em aprendizado de máquina

- Na ciência de dados, um modelo preditivo é uma fórmula para estimar o valor desconhecido do interesse: o alvo. A fórmula pode ser matemática, ou pode ser uma afirmação lógica, como uma regra. Muitas vezes, é um híbrido dos dois.
- Nessa aula, consideraremos modelos de classificação
- Isso contrasta com a modelagem descritiva, em que o objetivo principal do modelo não é estimar um valor, mas obter uma visão do fenômeno ou processo subjacente.
- Um modelo descritivo de comportamento de rotatividade nos diria como os clientes costumam se parecer.

Entendendo a estrutura dos dados do modelo

- O aprendizado supervisionado é a criação de modelos, em que o modelo descreve um relacionamento entre um conjunto de variáveis selecionadas (atributos) e uma variável predefinida chamada variável alvo (target).
- O modelo estima o valor da variável alvo como uma função (possivelmente uma função probabilística) dos atributos.
- Portanto, para o nosso problema de previsão de rotatividade, gostaríamos de criar um modelo de propensão à rotatividade em função dos **atributos** da conta do cliente, como idade, renda, duração do contrato com a empresa, número de chamadas para o atendimento ao cliente, cobranças adicionais, cobrança do cliente, dados demográficos, uso de dados e outros.

Terminologia

- A Figura ao lado ilustra algumas das terminologias que introduzimos aqui, em um exemplo simplificado de problema de previsão de crédito.
- Uma instância ou exemplo representa um fato ou um ponto de dados - nesse caso, um cliente histórico que recebeu crédito.
- Isso também é chamado de linha na terminologia do banco de dados ou planilha.
- Uma instância é descrita por um conjunto de atributos (campos, colunas, variáveis ou atributos).
- Às vezes, uma instância também é chamada de vetor de recurso, porque pode ser representada como uma coleção ordenada de comprimento fixo (vetor) de valores de recurso. Salvo indicação em contrário, assumiremos que os valores de todos os atributos (mas não o alvo) estão presentes nos dados.

Name	Balance	Age	Employed	Write-off
Mike	\$200,000	42	no	yes
Mary	\$35,000	33	yes	no
Claudio	\$115,000	40	no	no
Robert	\$29,000	23	yes	yes
Dora	\$72,000	31	no	no

This is one row (example).

Feature vector is: **<Claudio,115000,40,no>**

Class label (value of Target attribute) is **no**

Segmentação supervisionada

- Lembre-se de que um modelo preditivo se concentra na estimativa do valor de alguma variável alvo específica de interesse.
- Uma maneira intuitiva de pensar em extrair padrões de dados de maneira supervisionada é tentar segmentar a população em subgrupos que possuem valores diferentes para a variável alvo (e dentro do subgrupo as instâncias têm valores semelhantes para a variável alvo).
- Se a segmentação for feita usando valores de variáveis que serão conhecidas quando o alvo não for, esses segmentos poderão ser usados para prever o valor da variável alvo.

Segmentação supervisionada

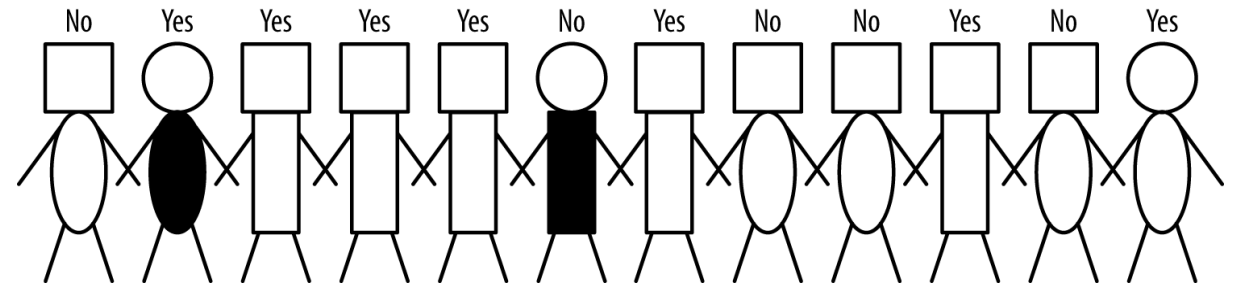
- Além disso, a segmentação pode ao mesmo tempo fornecer um conjunto de padrões de segmentação compreensível pelo homem. Por exemplo:
- "Profissionais de meia-idade que residem na cidade de Nova York, em média, têm uma taxa de rotatividade de 5%".
- Especificamente, o termo "profissionais de meia idade que residem na cidade de Nova York" é a definição do segmento (que faz referência a alguns atributos específicos) e "uma taxa de rotatividade de 5%" descreve o valor previsto da variável de alvo para o segmento.

Segmentação supervisionada

- Isso nos leva ao nosso conceito fundamental:
- **Como podemos julgar se uma variável contém informações importantes sobre a variável alvo? Quanto? Gostaríamos de obter automaticamente uma seleção das variáveis mais informativas em relação à tarefa específica em questão (a saber, prever o valor da variável de alvo).**
- Melhor ainda, gostaríamos de classificar as variáveis por quão boas elas são em prever o valor do alvo.

Selecionando atributos informativos

- Dado um grande conjunto de exemplos, como selecionamos um atributo para fazer uma de maneira informativa?
- Vamos considerar um problema de classificação binária (duas classes) e pensar no que gostaríamos de obter. Para ser concreto, a Figura ao lado mostra um simples problema de segmentação: doze pessoas representadas como figuras de palitos.
- Existem dois tipos de cabeças: quadrada e circular; e dois tipos de corpos: retangular e oval; e duas pessoas têm corpos cinzentos enquanto o resto é branco.
- Esses são os atributos que usaremos para descrever as pessoas.
- Acima de cada pessoa, há o rótulo de alvo binário, Sim ou Não, indicando (por exemplo) se a pessoa pode ou não receber um empréstimo.
- Poderíamos descrever os dados dessas pessoas como:



- Atributos:
 - cabeça: quadrada, circular
 - formato do corpo: retangular, oval
 - cor do corpo: cinza, branco
- Variável alvo:
 - Recebeu o empréstimo: Sim, Não ou ainda Positivo e Negativo

Selecionando atributos informativos

- Então, vamos nos perguntar: qual dos atributos seria melhor segmentar essas pessoas em grupos, de maneira a distinguir positivas das negativas?
- Tecnicamente, gostaríamos que os grupos resultantes fossem o mais puros possível.
- Por puro, queremos dizer homogêneo em relação à variável alvo.
- Se todos os membros de um grupo tiverem o mesmo valor para o alvo, o grupo será puro.
- Se houver pelo menos um membro do grupo que possua um valor diferente para a variável de alvo que o restante do grupo, o grupo estará impuro.

Selecionando atributos informativos

- Infelizmente, em dados reais, raramente esperamos encontrar uma variável que torne os segmentos puros.
- No entanto, se podemos reduzir substancialmente a impureza, podemos aprender algo sobre os dados (e a população correspondente) e, o que é mais importante para este capítulo, podemos usar o atributo em um modelo preditivo - em nosso exemplo, prevendo que os membros de um segmento terá taxas mais baixas ou mais altas do que as de outro segmento.
- Se pudermos fazer isso, poderemos, por exemplo, oferecer crédito àquelas com taxas de baixa previstas mais baixas ou oferecer diferentes termos de crédito com base nas diferentes taxas de baixa previstas.

Entropia

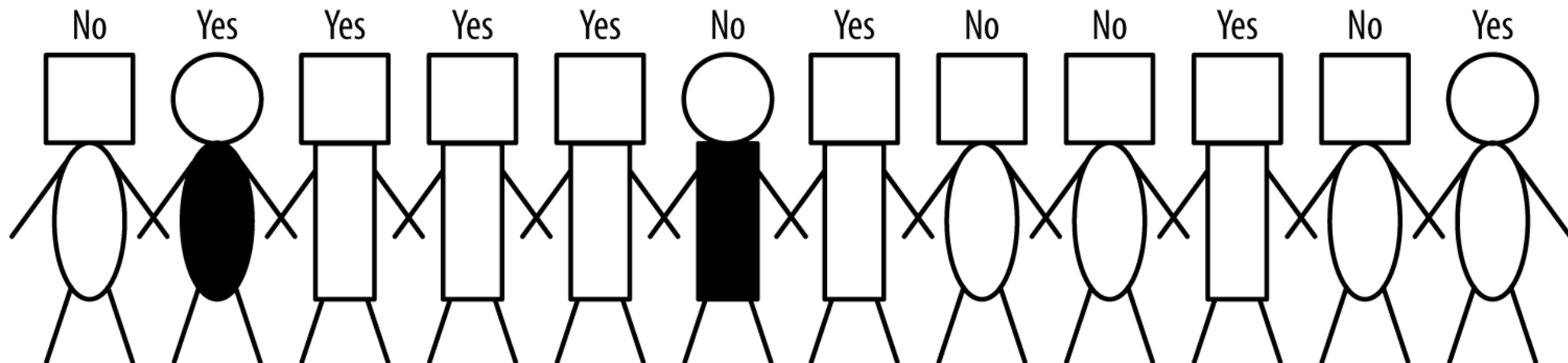
- Felizmente, para problemas de classificação, podemos resolver todos os problemas criando uma fórmula que avalie quão bem cada atributo divide um conjunto de exemplos em segmentos, com relação a uma variável de alvo escolhida. Essa fórmula é baseada em uma medida de pureza.
- O critério de divisão mais comum é chamado ganho de informação e é baseado em uma medida de pureza chamada entropia. Ambos os conceitos foram inventados por um dos pioneiros da teoria da informação, Claude Shannon, em seu trabalho seminal no campo (Shannon, 1948).
- Entropia é uma medida de **desordem** que pode ser aplicada a um conjunto, como um de nossos segmentos individuais.
- Considere que temos um conjunto de propriedades dos membros do conjunto e cada membro possui uma e apenas uma das propriedades.
- Na segmentação supervisionada, as propriedades do membro corresponderão aos valores da variável de alvo. A desordem corresponde à mistura (impura) do segmento com relação a essas propriedades de interesse. Assim, por exemplo, um segmento misturado com muitas baixas e baixas não-baixas teria alta entropia.

Entropia – Qual prato tem maior entropia

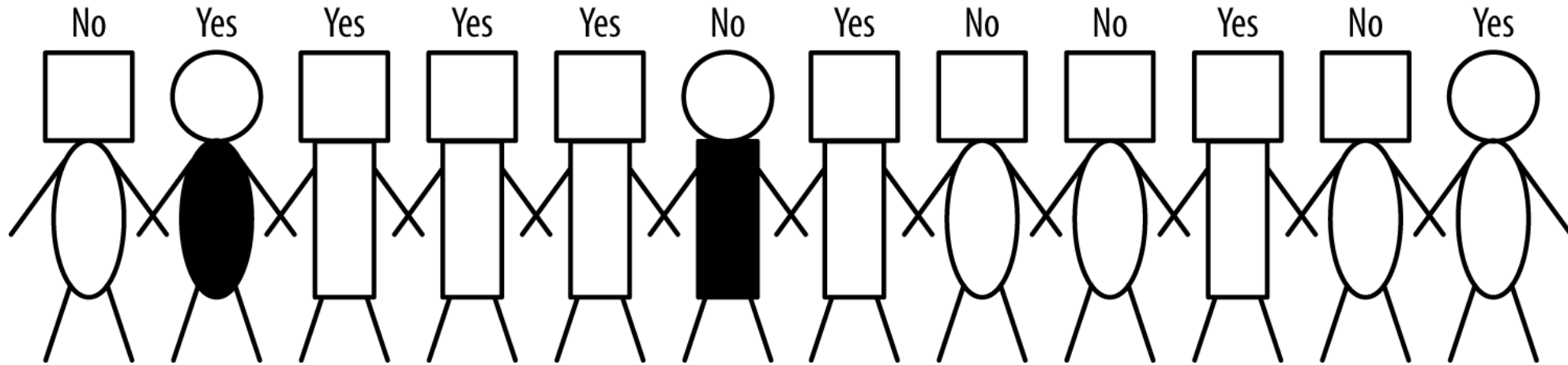


A indução de árvore trabalha com dois conceitos fundamentais

- Entropia
- A entropia tem a ver com o grau de pureza da informação de um subconjunto de dados
- Vamos ver como isso funciona neste conjunto de dados



Vamos ver o Atributo Cabeças



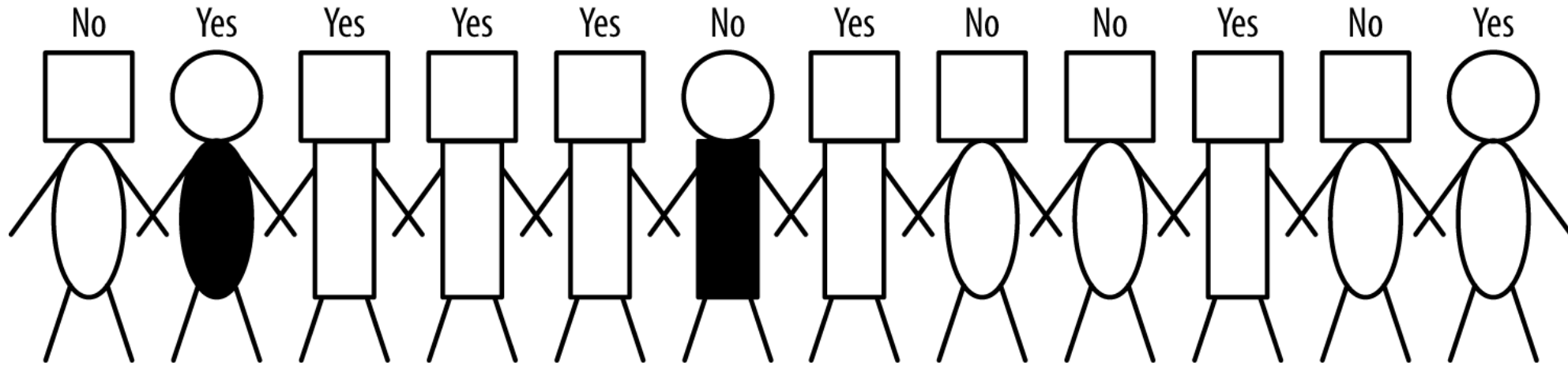
- Cabeça Quadrada
 - Total 9 instâncias
- Cabeça Quadrada e Sim
 - 5 Instâncias (55%)
- Cabeça Quadrada e Não
 - 4 Instâncias (45%)

- Cabeça Redonda
 - Total 3 instâncias
- Cabeça Redonda e Sim
 - 2 Instâncias (67%)
- Cabeça Redonda e Não
 - 1 Instância (33%)

O que vimos com relação às cabeças

- Parece que ambos os conjuntos não são muito bem organizados
- As porcentagens são relativamente altas neste conjunto de dados e fica difícil dizer que eles são puros e a entropia é alta.
- Ou seja os dados estão bagunçados, mas dá para dizer que a entropia em relação ao sim ou não (atributo alvo) é menor nas cabeças redondas do que nas cabeças quadradas

Vamos ver o Atributo Cabeças



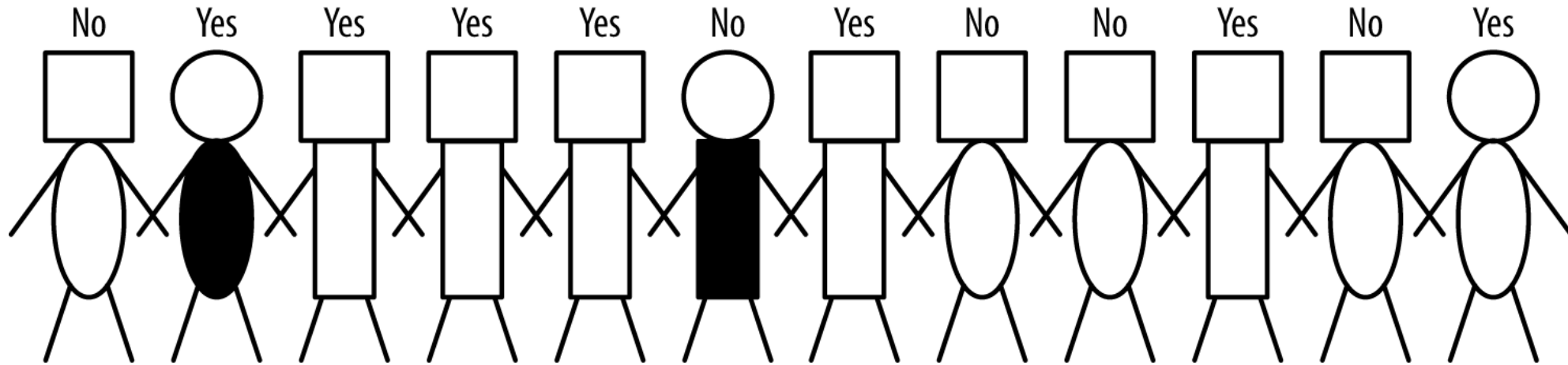
- Corpo Retangular
 - Total 6 instâncias
- Corpo Retangular e Sim
 - 5 Instâncias (83%)
- Cabeça Retangular e Não
 - 1 Instância (17 %)

- Corpo Oval
 - Total 6 instâncias
- Corpo Oval e Sim
 - 2 Instâncias (33%)
- Cabeça Oval e Não
 - 4 Instâncias (66 %)

O que vimos com relação aos corpos

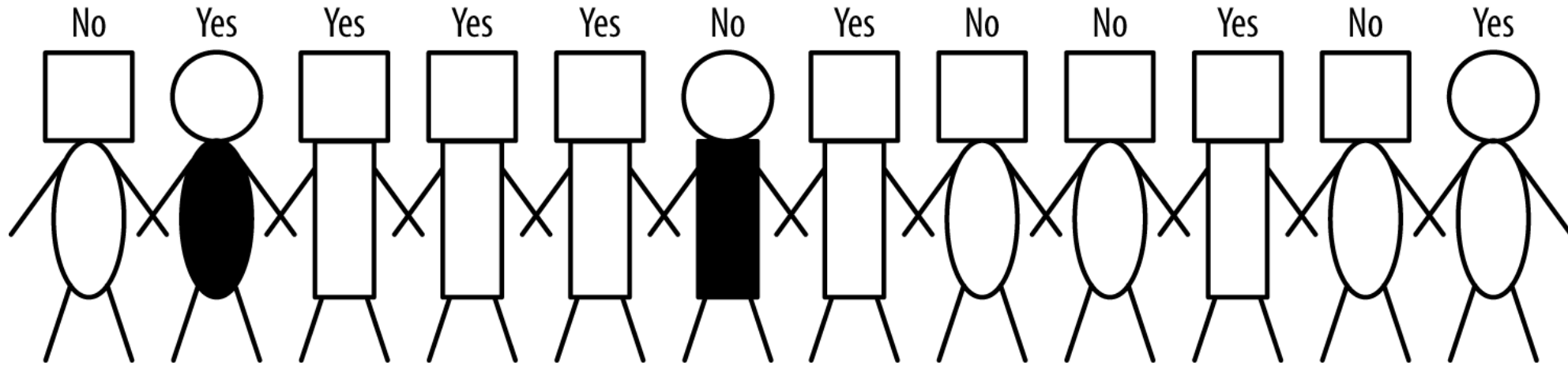
- Parece que ambos os conjuntos não são muito bem organizados
- Mas quando olhamos para o subconjunto dos corpos retangulares ele parece ter a menor entropia até agora parece ser um bom começo
- Que tal avaliarmos um subconjunto de dados dos corpos retangulares e as cabeças quadradas

Vamos ver em Conjunto Atributo Corpo e Cabeças



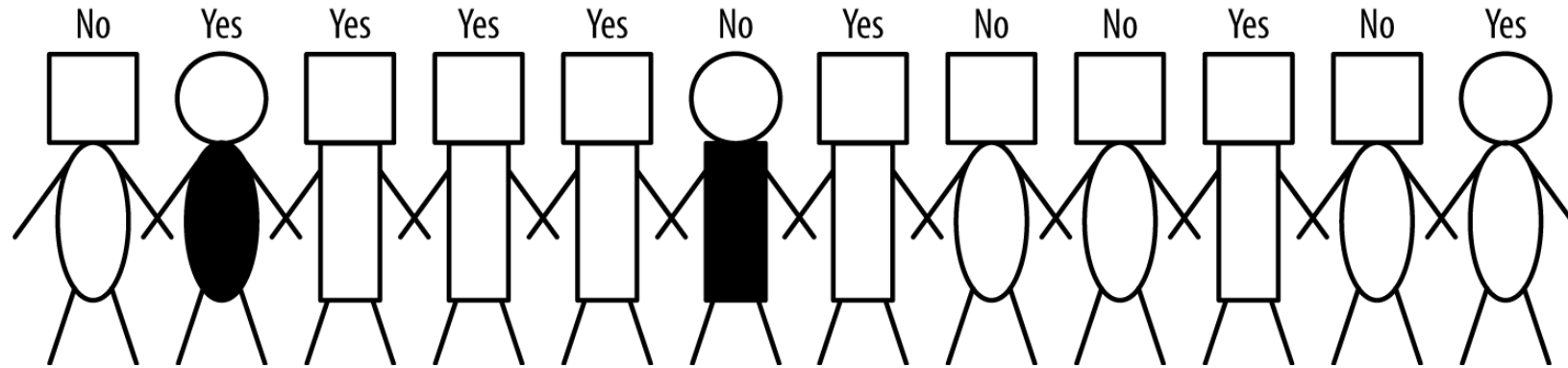
- Corpo Retangular
 - Total 6 instâncias
- Corpo Retangular e Sim
 - 5 Instâncias (83%)
- Cabeça Retangular e Não
 - 1 Instância (17 %)
- Corpo Retangular e Cabeça Quadrada
 - Total 5 instâncias
- Corpo Retangular e Cab. Quadrada e Sim
 - 5 Instâncias (100%)
- Corpo Retangular e Cab. Quadrada e Não
 - 0 Instância (0 %)

Vamos ver em Conjunto Atributo Corpo e Cabeças



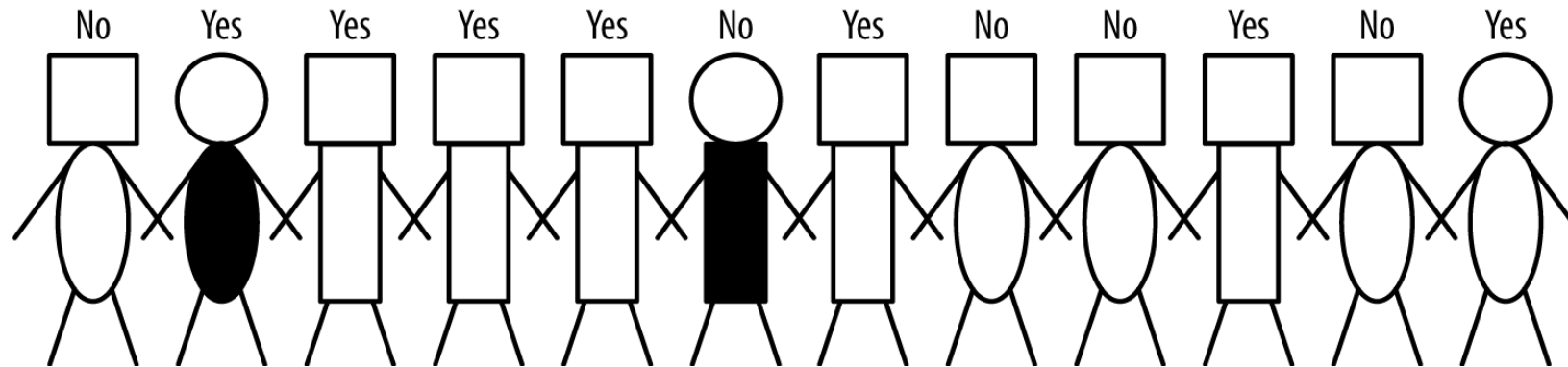
- Corpo Retangular
 - Total 6 instâncias
- Corpo Retangular e Sim
 - 5 Instâncias (83%)
- Cabeça Retangular e Não
 - 1 Instância (17 %)
- Corpo Retangular e Cabeça Redonda
 - Total 1 instâncias
- Corpo Retangular e Cab. Redonda e Sim
 - 0 Instâncias (0%)
- Corpo Retangular e Cab. Redonda e Não
 - 1 Instância (100 %)

Vamos olhar para o Geral



Conjunto de Dados
12 Instâncias

Vamos olhar para o Geral

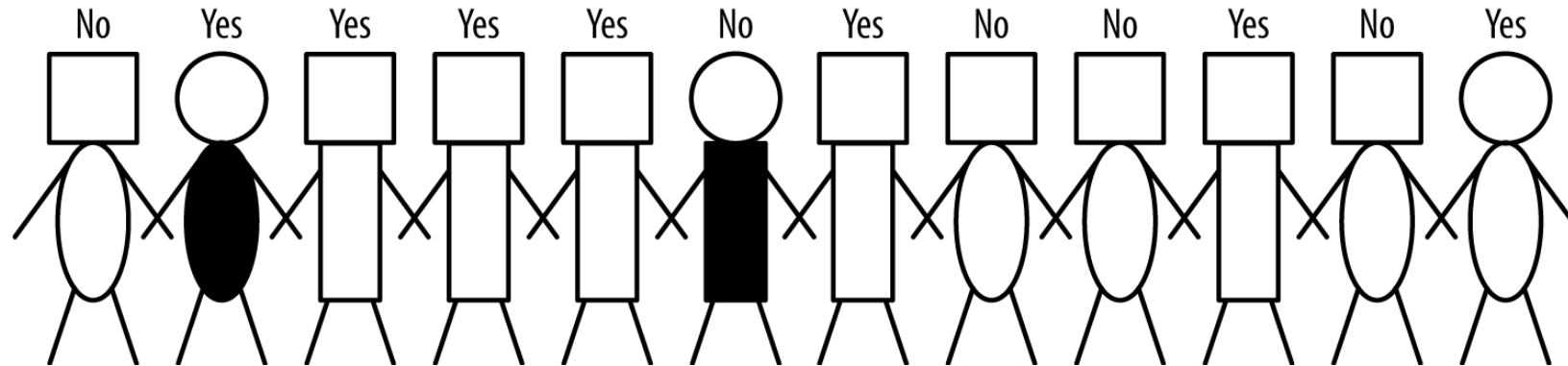


Conjunto de Dados
12 Instâncias

Corpo Retangular
6 Instâncias

Corpo Oval
6 Instâncias

Vamos olhar para o Geral



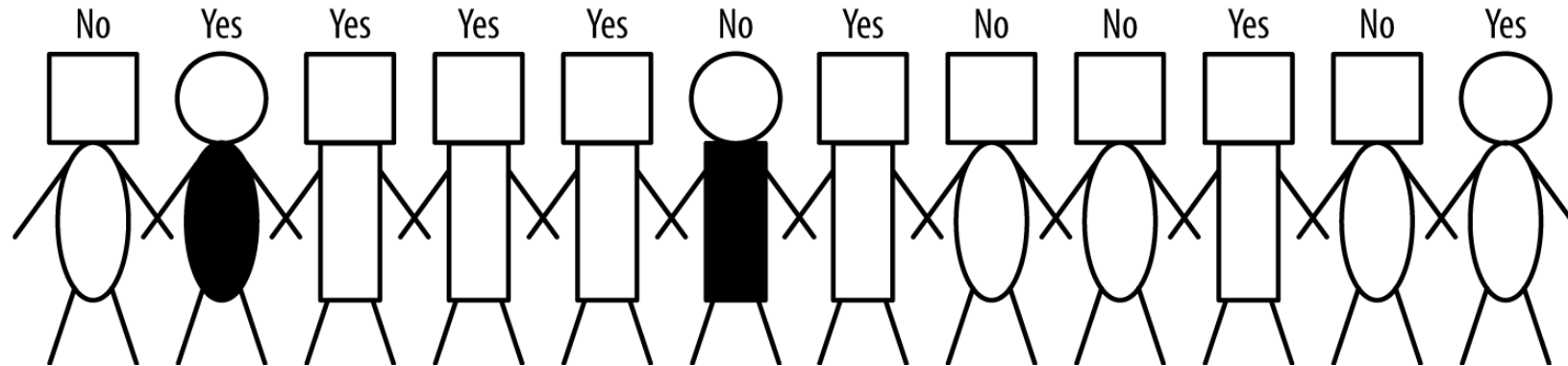
Conjunto de Dados
12 Instâncias

Corpo Retangular
6 Instâncias

Cabeça Quadrada
5 Instâncias

Cabeça Redonda
1 Instâncias

Vamos olhar para o Geral



Conjunto de Dados
12 Instâncias

Corpo Retangular
6 Instâncias

Cabeça Quadrada
5 Instâncias

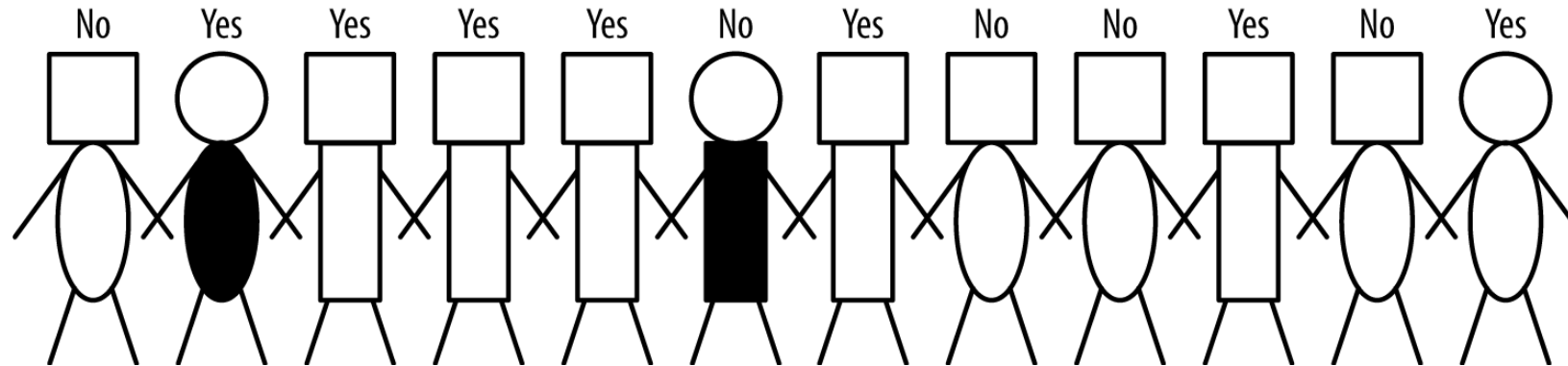
Cabeça Redonda
1 Instâncias

Sim
5 Instâncias

Não
0 Instâncias

Se o Corpo Retangular e Cabeça Quadrada então SIM

Vamos olhar para o Geral



Conjunto de Dados
12 Instâncias

Corpo Retangular
6 Instâncias

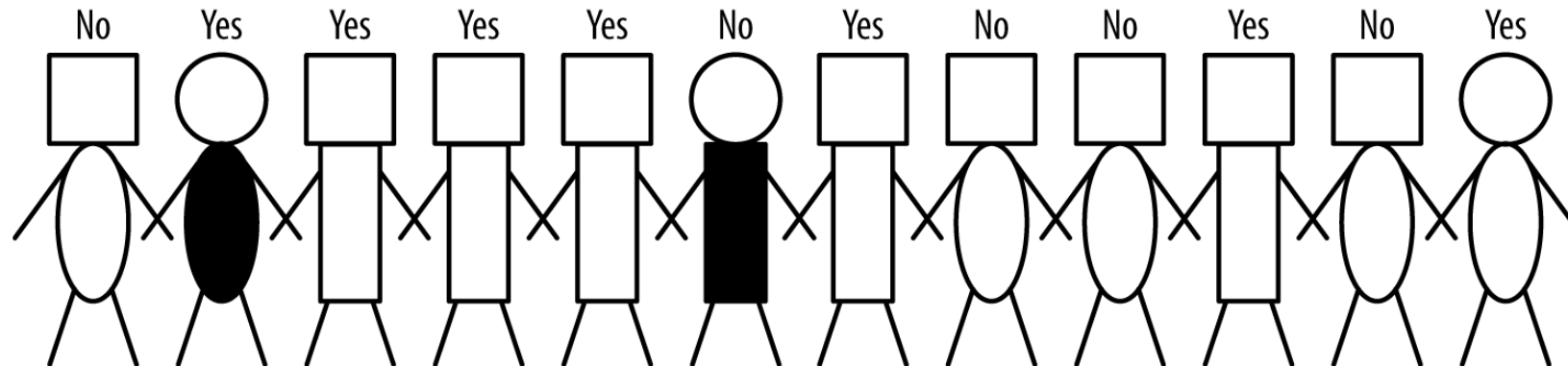
Cabeça Redonda
1 Instâncias

Sim
0 Instâncias

Não
1 Instâncias

Se o Corpo Retangular e Cabeça Redonda então NÃO

Vamos olhar para o Geral



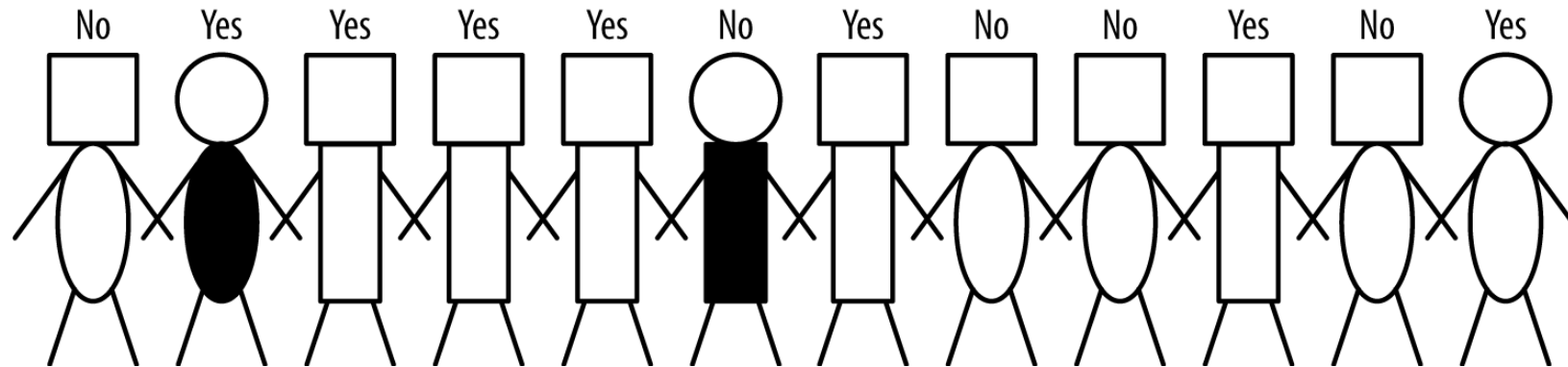
Conjunto de Dados
12 Instâncias

Corpo Oval
6 Instâncias

Cabeça Quadrada
4 Instâncias

Cabeça Redonda
2 Instâncias

Vamos olhar para o Geral



Conjunto de Dados
12 Instâncias

Corpo Oval
6 Instâncias

Cabeça Quadrada
4 Instâncias

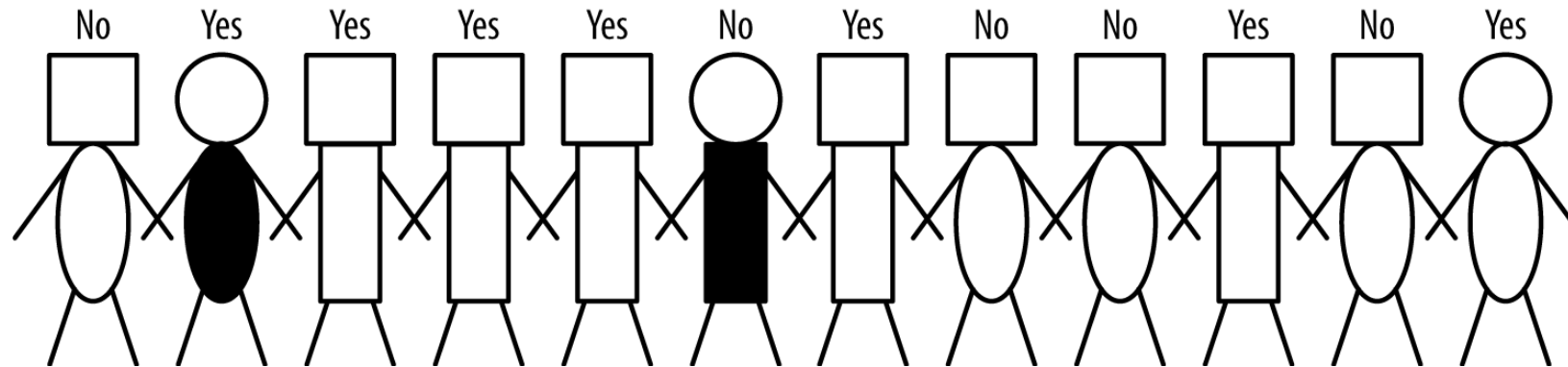
Cabeça Redonda
2 Instâncias

Sim
0 Instâncias

Não
4 Instâncias

Se o Corpo Oval e Cabeça Quadrada então NÃO

Vamos olhar para o Geral



Conjunto de Dados
12 Instâncias

Corpo Oval
6 Instâncias

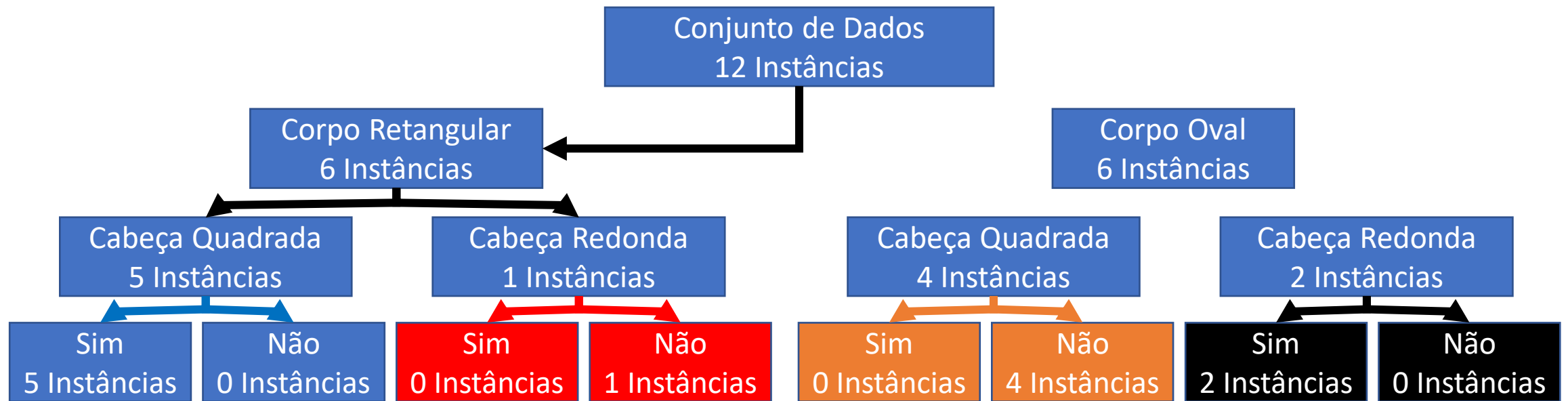
Cabeça Redonda
2 Instâncias

Sim
2 Instâncias

Não
0 Instâncias

Se o Corpo Oval e Cabeça Redonda então SIM

Vamos olhar para o Geral e ver as regras que encontramos



Se o Corpo Retangular e Cabeça Quadrada então SIM

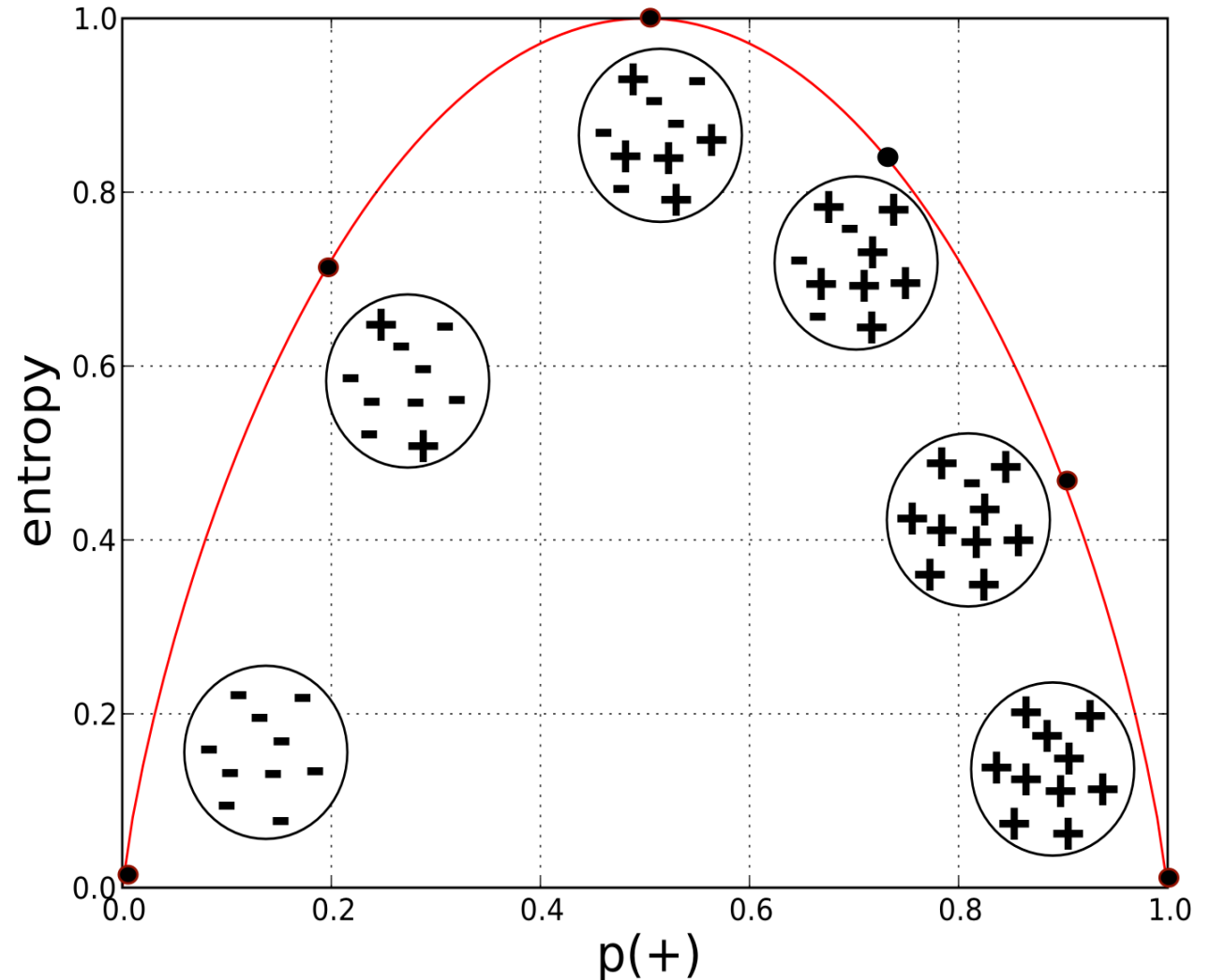
Se o Corpo Retangular e Cabeça Redonda então NÃO

Se o Corpo Oval e Cabeça Quadrada então NÃO

Se o Corpo Oval e Cabeça Redonda então SIM

Entendendo a fórmula da entropia

- Começando com todas as instâncias negativas no canto inferior esquerdo, $p_+ = 0$, o conjunto apresenta um distúrbio mínimo (é puro) e a entropia é zero.
- Se começarmos a alternar os rótulos de classe dos elementos do conjunto de - para +, a entropia aumentará.
- A entropia é maximizada em 1 quando as classes de instância são balanceadas (cinco de cada) e $p_+ = p_- = 0,5$.
- À medida que mais rótulos de classe são trocados, a classe + começa a predominar e a entropia diminui novamente.
- Quando todas as instâncias são positivas, $p_+ = 1$ e a entropia é mínima novamente em zero.

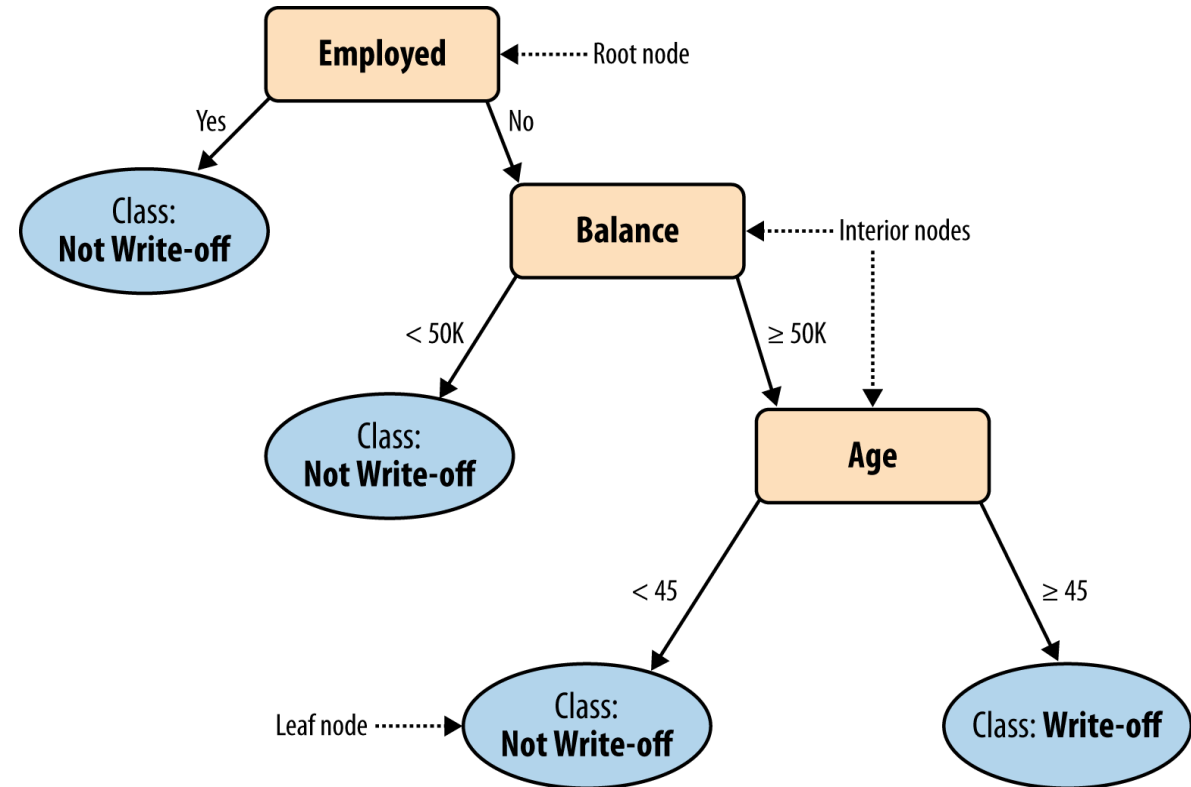


Ganho de informação (IG)

- Gostaríamos de medir o quão informativo um atributo é em relação ao nosso objetivo: quanto ganho de informação ele fornece sobre o valor da variável alvo. Um atributo segmenta um conjunto de instâncias em vários subconjuntos.
- A entropia nos diz apenas quão impuro é um subconjunto individual.
- Felizmente, com a entropia para medir a desordem de qualquer conjunto, podemos definir o ganho de informação (IG) para medir o quanto um atributo melhora (diminui) a entropia sobre toda a segmentação que ele cria.
- A rigor, o ganho de informação mede a alteração na entropia devido a qualquer quantidade de novas informações sendo adicionadas; aqui, no contexto da segmentação supervisionada, consideramos as informações obtidas dividindo o conjunto em todos os valores de um único atributo.
- Digamos que o atributo em que dividimos tenha k valores diferentes. Vamos chamar o conjunto original de exemplos de conjunto pai e o resultado da divisão nos valores de atributo que k filhos define.
- Assim, o ganho de informações é uma função de um conjunto pai e dos filhos resultantes de alguma partição do conjunto pai - quanta informação esse atributo forneceu? Isso depende de quão mais puros os filhos são do que os pais.
- Declarado no contexto da modelagem preditiva, se soubéssemos o valor desse atributo, quanto aumentaria nosso conhecimento do valor da variável de alvo?

Segmentação supervisionada com modelos estruturados em árvore

- Considere uma segmentação dos dados para assumir a forma de uma "árvore", como a mostrada na Figura ao lado.
- A árvore está de cabeça para baixo com a raiz no topo.
- A árvore é composta de nós, nós internos e terminais e ramificações que emanam dos nós internos.
- Cada nó interno na árvore contém um teste de um atributo, com cada ramificação do nó representando um valor distinto do atributo.
- Após as ramificações do nó raiz para baixo (na direção das setas), cada caminho termina eventualmente em um nó terminal ou folha.
- A árvore cria uma segmentação dos dados: cada ponto de dados corresponderá a um e apenas um caminho na árvore e, assim, a uma e apenas uma folha.
- Em outras palavras, cada folha corresponde a um segmento, e os atributos e valores ao longo do caminho fornecem as características do segmento.

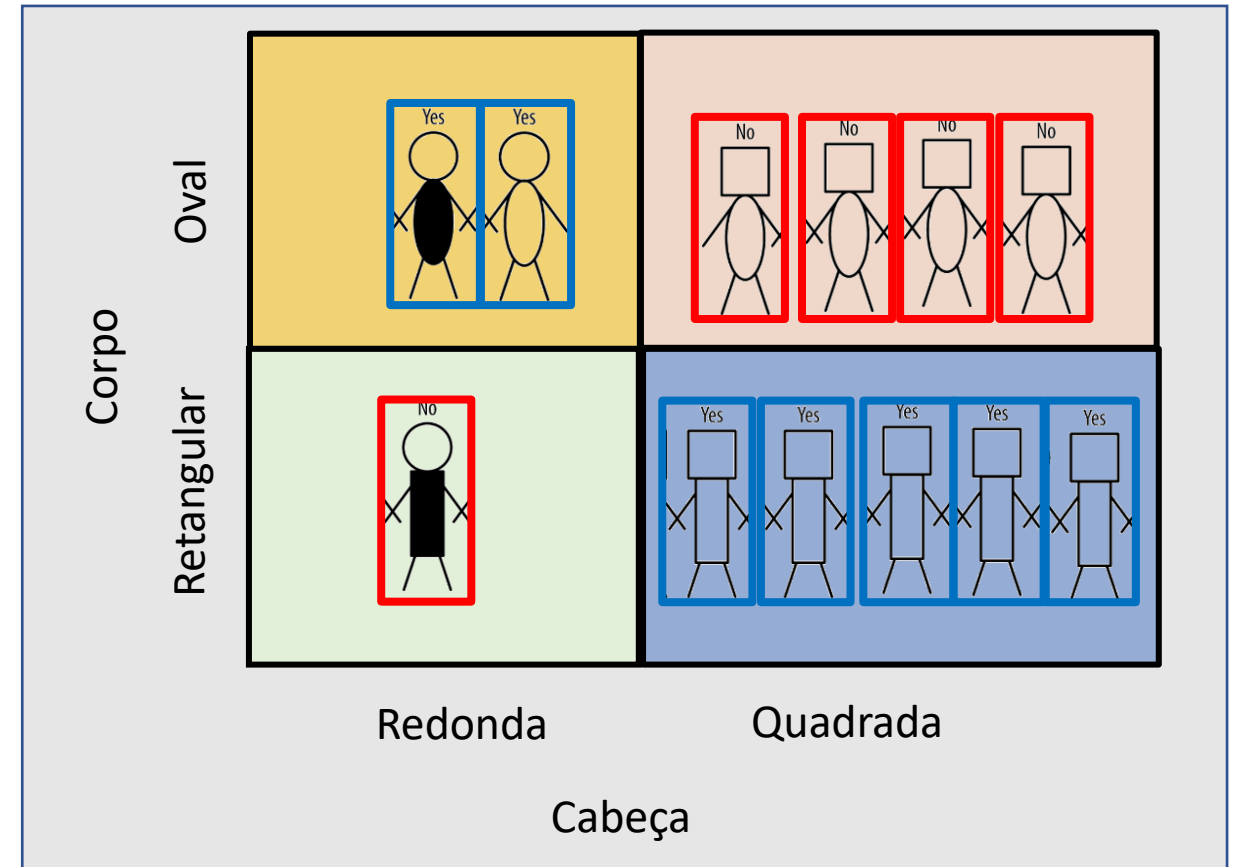
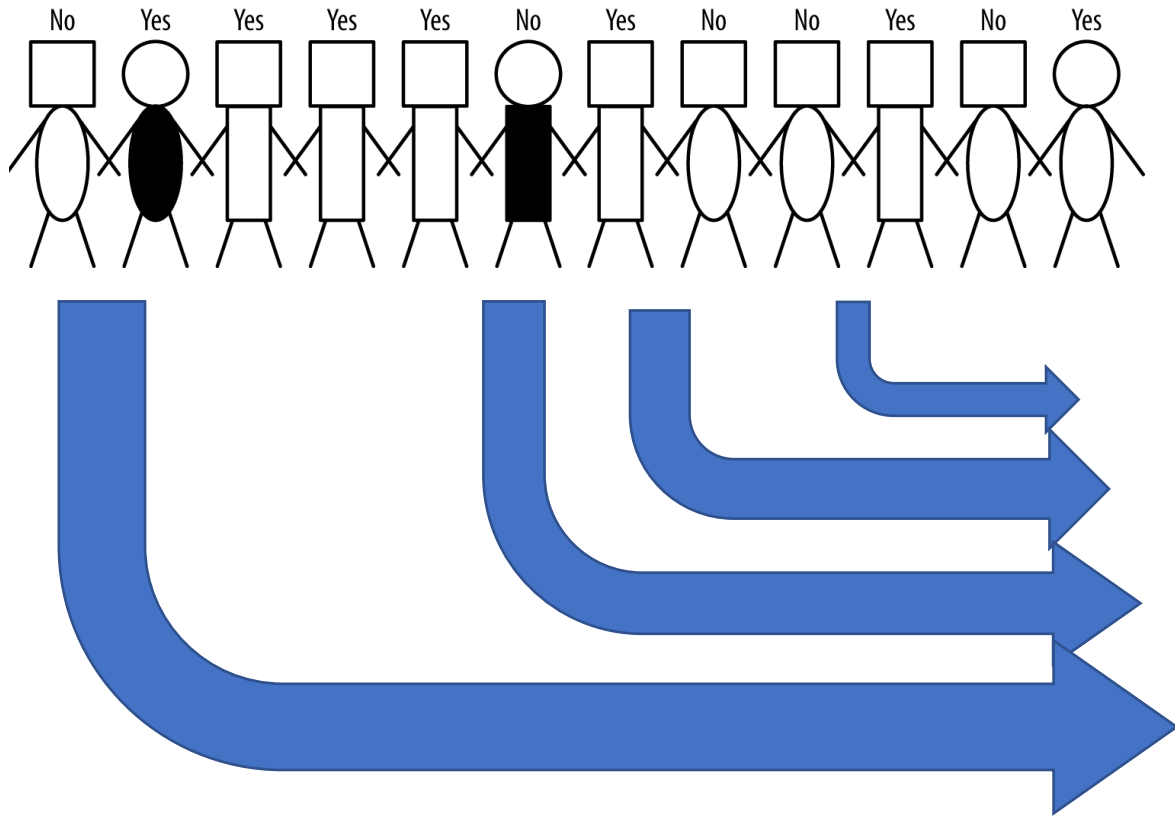


Portanto, o caminho mais à direita na árvore corresponde ao segmento **"Pessoas idosas, desempregadas e com saldos altos"**.

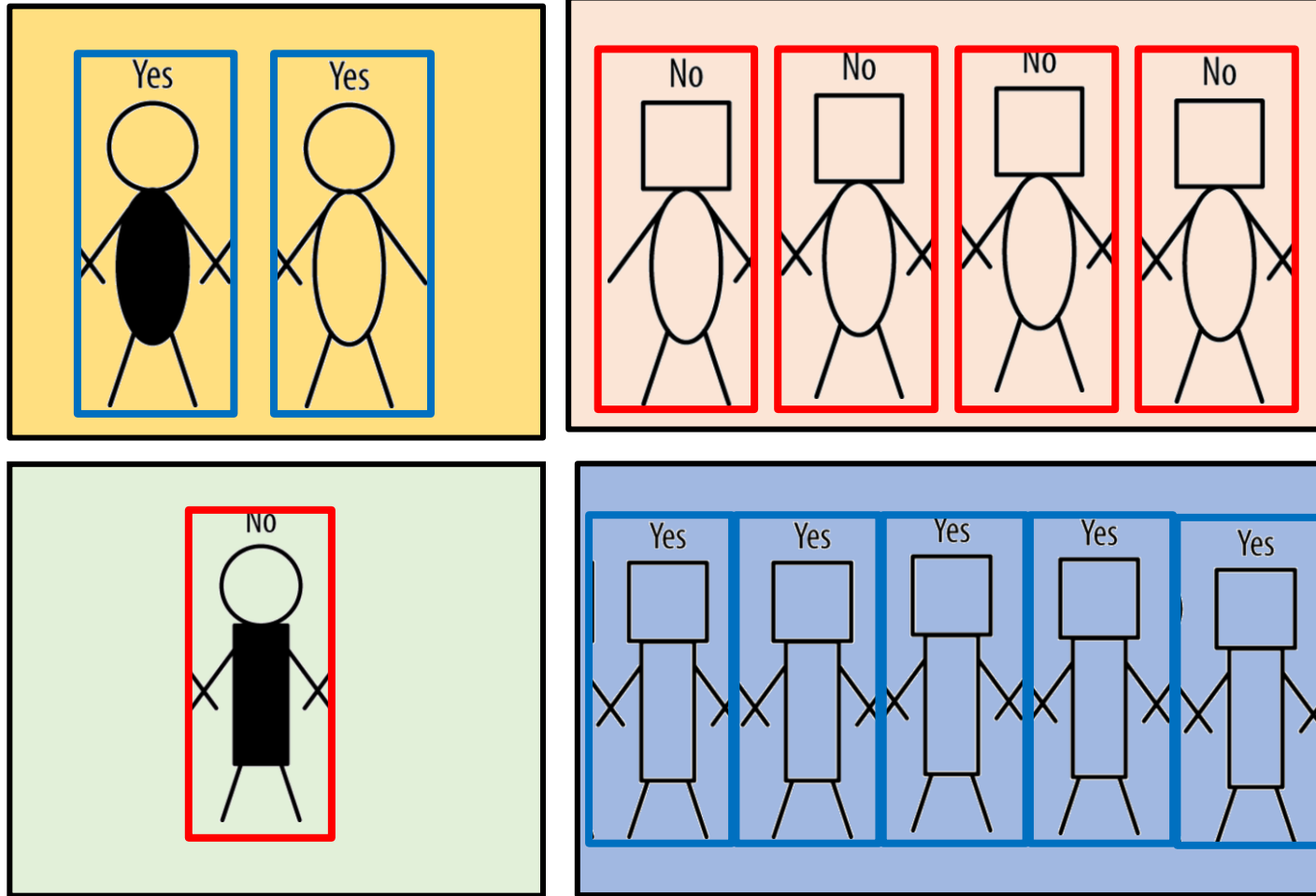
Visualização de segmentações

- Continuando com a metáfora da construção de modelo preditivo como segmentação supervisionada, é instrutivo visualizar exatamente como uma árvore de classificação particiona o espaço da instância.
- O espaço da instância é simplesmente o espaço descrito pelos recursos de dados. Uma forma comum de visualização do espaço da instância é um gráfico de dispersão em alguns pares de recursos, usado para comparar uma variável com outra para detectar correlações e relacionamentos.
- Embora os dados possam conter dezenas ou centenas de variáveis, só é realmente possível visualizar segmentações em duas ou três dimensões ao mesmo tempo.
- Ainda assim, visualizar modelos no espaço da instância em algumas dimensões é útil para entender os diferentes tipos de modelos, pois fornece insights que também se aplicam a espaços dimensionais mais altos.
- Pode ser difícil comparar famílias de modelos muito diferentes apenas examinando sua forma (por exemplo, uma fórmula matemática versus um conjunto de regras) ou os algoritmos que os geram.

Separando em Grupos

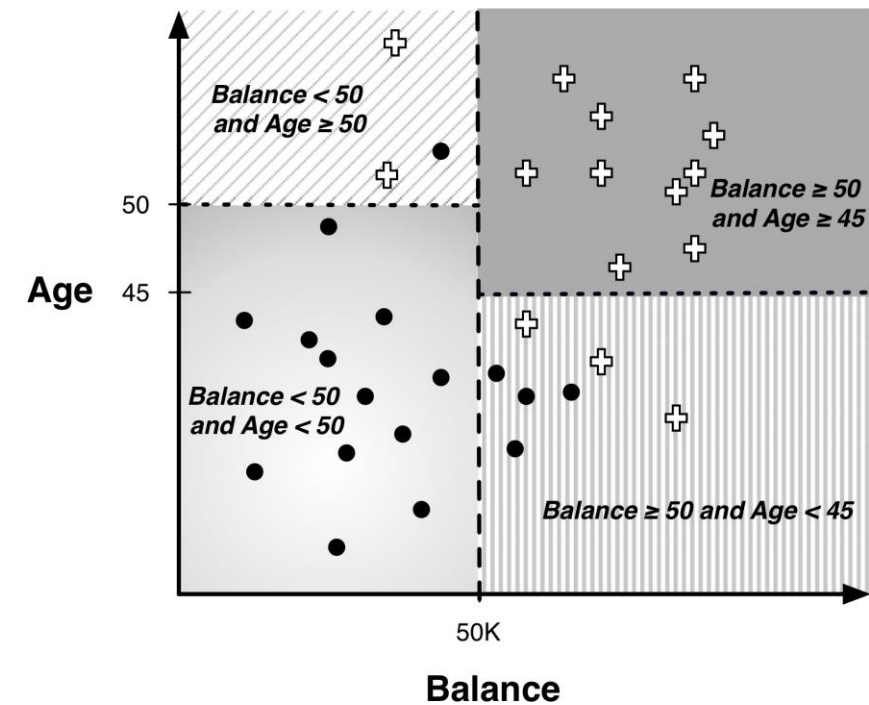
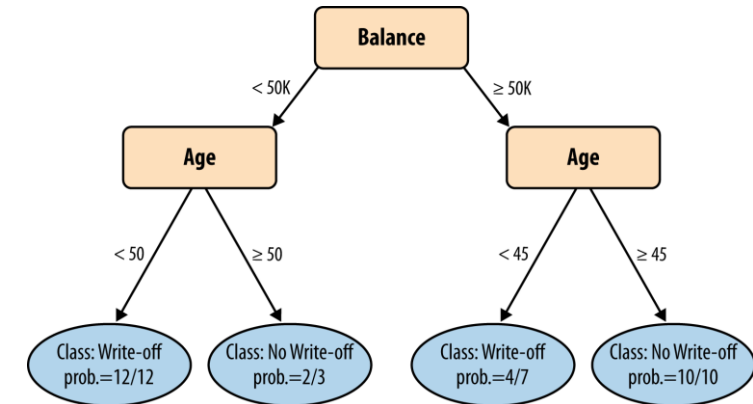


Separando em Grupos



Visualização de segmentações

- Geralmente, é mais fácil compará-los com base em como eles particionam o espaço da instância.
- Por exemplo, a Figura no próximo slide mostra uma árvore simples de classificação ao lado de um gráfico bidimensional do espaço da instância: Saldo no eixo x e Idade no eixo y.
- O nó raiz da árvore de classificação testa Saldo contra um limite de 50K.
- No gráfico, isso corresponde a uma linha vertical a 50K no eixo x que divide o plano em Saldo <50K e Saldo ≥ 50K.
- À esquerda dessa linha, estão as instâncias cujos valores de Saldo são menores que 50K; existem 13 exemplos de classe Baixa (ponto preto) e 2 exemplos de classe não-write-off (sinal de mais) nesta região.
- Na ramificação direita do nó raiz, há instâncias com Saldo ≥ 50K.
- O próximo nó na árvore de classificação testa o atributo Idade em relação ao limite 45.



Árvores como conjuntos de regras

- Antes de passarmos da interpretação das árvores de classificação, devemos mencionar a interpretação deles como afirmações lógicas.
- Considere novamente a árvore mostrada na parte superior da Figura no slide anterior.
- Você classifica uma nova instância invisível iniciando no nó raiz e seguindo os testes de atributo para baixo até chegar a um nó folha, que especifica a classe prevista da instância. Se rastrearmos um único caminho do nó raiz até uma folha, coletando as condições à medida que avançamos, geramos uma regra. Cada regra consiste nos testes de atributo ao longo do caminho conectado com AND.
- Começando no nó raiz e escolhendo os ramos esquerdos da árvore, obtemos a regra:
- SE (Saldo <50K) E (Idade <50) ENTÃO Classe = Write-off
- Podemos fazer isso para todos os caminhos possíveis para um nó folha. Nesta árvore, temos mais três regras:
- SE (Saldo <50K) E (Idade \geq 50) ENTÃO Classe = No Write-off
- SE (Saldo \geq 50K) E (Idade <45) ENTÃO Classe = Write-off
- SE (Saldo \geq 50K) E (Idade <45) ENTÃO Classe = No Write-off
- A árvore de classificação é equivalente a este conjunto de regras. Se essas regras parecerem repetitivas, é porque são: a árvore reúne prefixos de regras comuns em direção ao topo da árvore. Toda árvore de classificação pode ser expressa como um conjunto de regras dessa maneira. Se a árvore ou o conjunto de regras é mais inteligível é uma questão de opinião; Neste exemplo simples, ambos são bastante fáceis de entender. À medida que o modelo aumenta, algumas pessoas preferem a árvore ou o conjunto de regras.