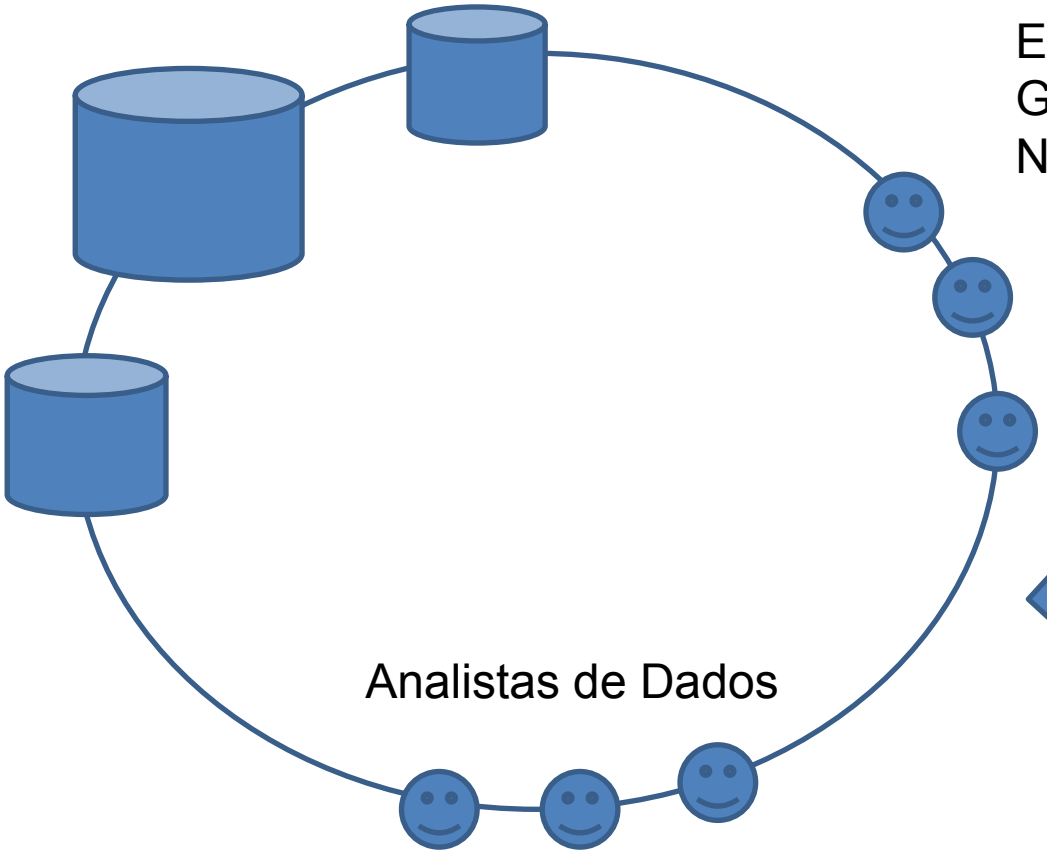
Pré-processamento de Dados

Ricardo Prudêncio

Relembrando KDD…

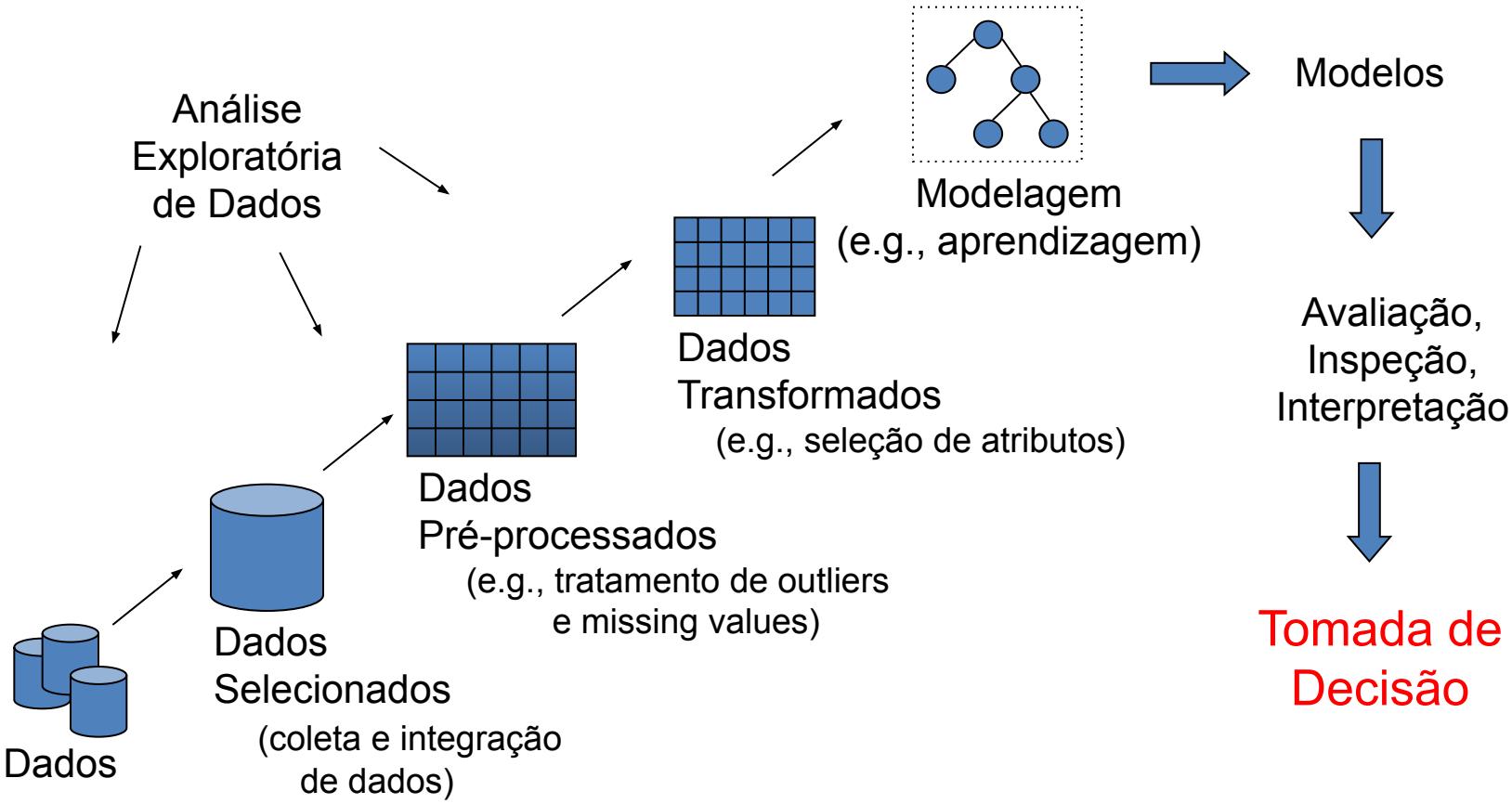
Bases de Dados,....



Experts, Stakeholders, Gerentes, Analistas de Negócios,....

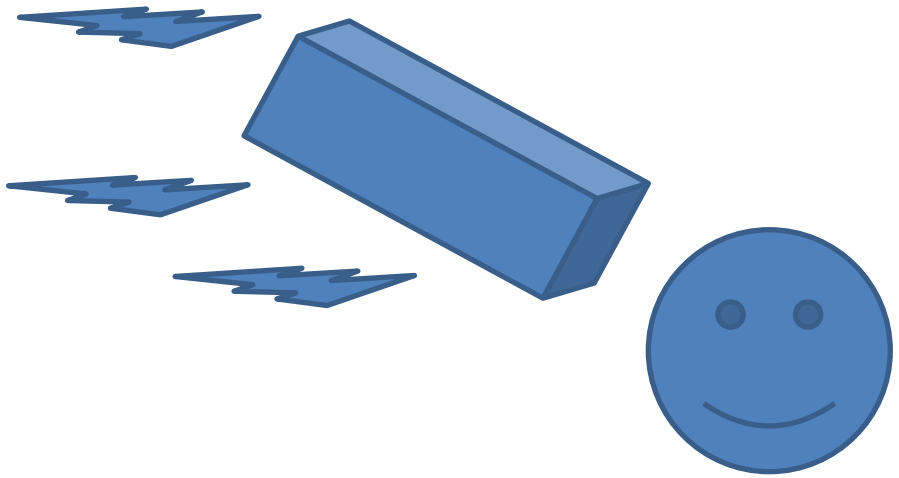
VALOR

KDD (Knowledge Discovery in Databases)

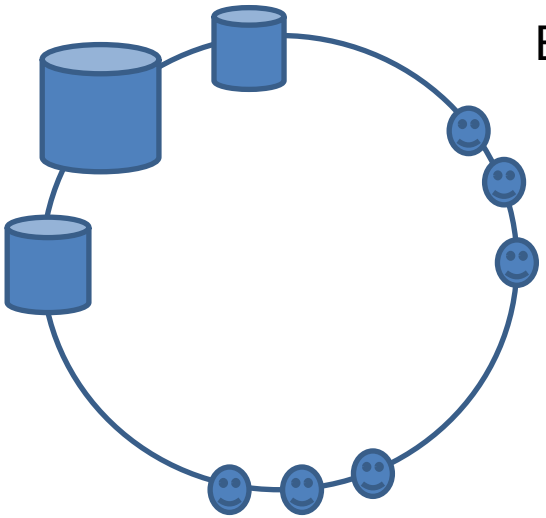


IMPORTANTE

- Boa parte do processo de aquisição de conhecimento (KDD) é feito na realidade ANTES da modelagem!!!



Bases de Dados

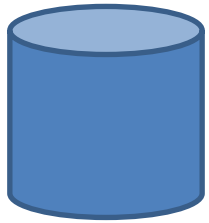


Experts

Analistas de Dados

- **Coleta de Dados**
- **Pré-Seleção dos Dados**
- **Padronização**
- **Transformações**
- **Enriquecimento de dados**
- **Seleção de atributos**
- **Seleção de instâncias**

Boa massa de dados!!!



MODELAGEM

Transformações nos Dados

- Uma vez coletados, os dados podem ser transformados para facilitar a análise
- Existem vários tipos de transformação para vários tipos de dados
  - E.g., atributos numéricos vs categóricos

Transformações nos Dados

- Tipos de Atributos
  - Numérico X Categórico
    - E.g., Peso (Kg) X Classe social (A, B, C, …)
  - Discreto X Contínuo
    - E.g., Idade X Temperatura
  - Ordinal X Nominal
    - E.g., Estatura (Alta, Media, Baixa) vs Cor (Azul, Verde)

Transformação de Dados Categóricos

- Boa parte dos algoritmos de aprendizagem não trata dados categóricos de forma direta
  - E.g., SVMs, Regressão Logística,…
- Necessidade de transformação para formato numérico tratável pelos algoritmos

Transformação de Dados Categóricos

- Abordagem ingênua:
  - Associar números inteiros às categorias
    - Ex.: Azul -> 1, Verde -> 2, Vermelho -> 3,…
  - Problema:
    - É inserida uma ordem nos valores que originalmente não existia (reflitam sobre o kNN!)
      - I.e. Azul < Verde < Vermelho

Transformação de Dados Categóricos

- Representação 1-of-N
  - Dado N categorias de uma variável, gerar N variáveis binária
  - Atribuir valor 1 para a variável binária associada a uma categoria e 0 para as demais
    - E.g. Azul -> 100, Verde ->010, Vermelho -> 001
  - Feito comumente em Regressão Logística

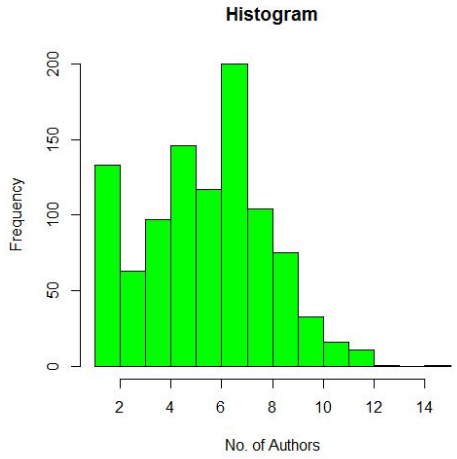
OBS.: Ver filtro NominalToBinary no WEKA

Transformação de Dados Contínuos - Discretização

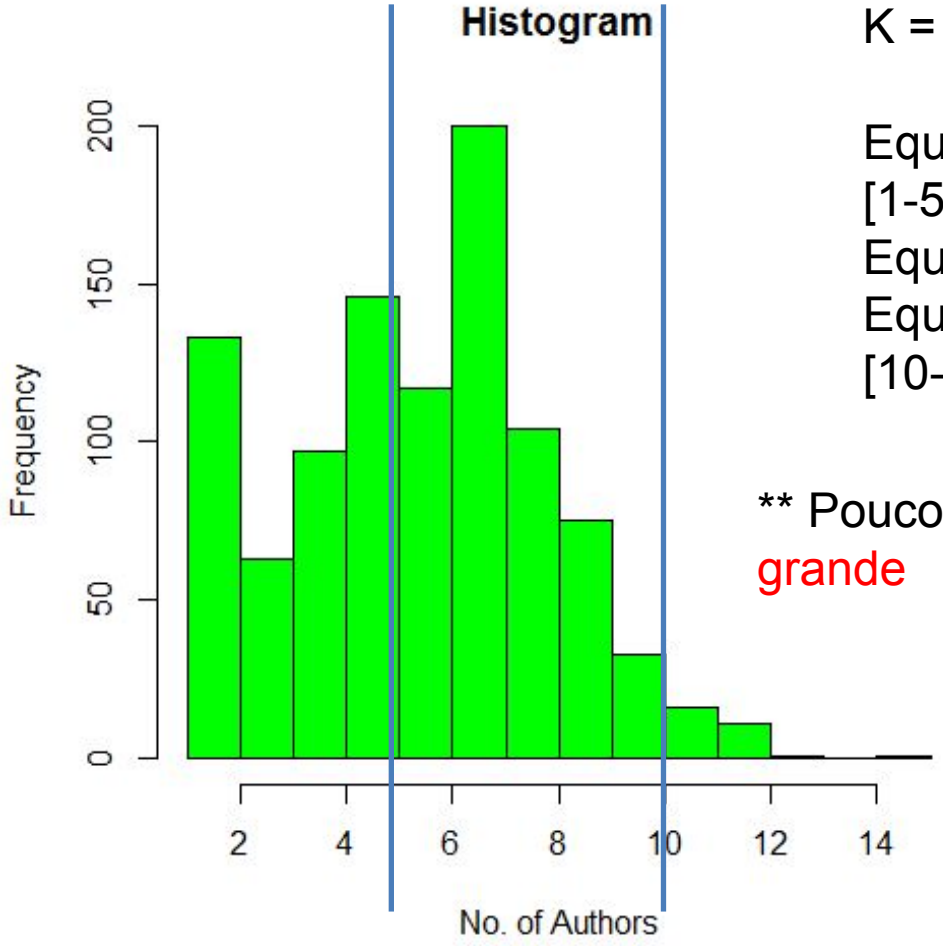
- Objetivo: criar informação mais fácil de ser tratada ou prevista
  - E.g.: número de bugs (presença, ausência) de bugs
  - E.g., índice da bolsa de valores tendência de queda (positiva ou negativa)
  - E.g., número de desenvolvedores tamanho da equipe (pequena, média ou grande)

Discretização – Equal-Width

- Divide o domínio do atributo em k intervalos de tamanho fixo
  - Nem sempre é trivial definir valor de k
  - Novos valores dos atributos podem ser distribuídos de forma desbalanceada



Discretização – Equal-Width



K = 3

Equipe pequena = [1-5]
Equipe média = [6-10]
Equipe grande = [10-15]

** Poucos dados receberão valor grande

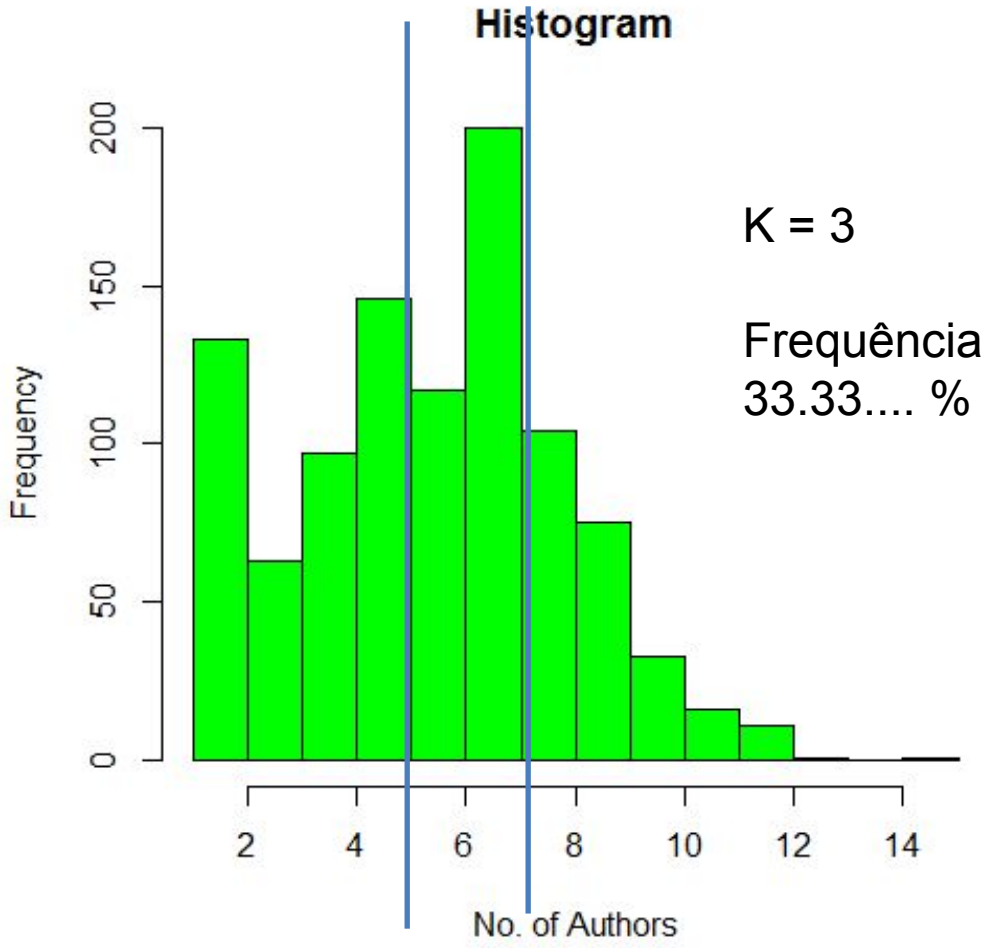
OBS.: Ver Discretize no WEKA

Discretização

- Equal-frequency discretization (EFD)
  - Divide domínio do atributo em k intervalos com mesmo número de exemplos

- Valores categóricos serão igualmente distribuídos

Discretização – Equal-Frequency



K = 3

Frequência de cada valor = 33.33.... %

Transformação de Dados - Normalização

- Objetivo: transformar dados contínuos em escalas diferentes
  - E.g.., idade, peso, altura,....

- Padronização
- Reescalonamento
- Normalização por Quantis

Transformação de Dados - Normalização

- Padronização

$$z = \frac{x - \mu}{\sigma}$$

- Reescalonamento

$$x' = \frac{x - \min}{\max - \min}$$

- Normalização por Quantil

$$r = rank(x)$$

	Idade	Altura
A	18	1,60
B	30	1,80
C	55	1,70
D	40	1,85
E	21	1,64

Media
32.8

Media
1.71

Padronização

	Idade	Altura
A	- 0.98	-1.11
B	- 0.18	+ 0.77
C	+ 1.47	-0,17
D	+ 0.47	+ 1,25
E	- 0.78	- 0,73

Reescalonamento

	Idade	Altura
A	0.0	0
B	0.32	0.8
C	1.0	0.4
D	0.59	1.0
E	0.08	0.16

Normalização por Quantil

	Idade	Altura
A	0%	0%
B	50%	75%
C	100%	50%
D	75%	100%
E	25%	25%

Obs.: Ver Normalize e Stardardize no WEKA

Tratamento de Anomalias (Pontos Espúrios)

- Detecção de outliers:
  - Identificação de observações que diferem muito das outras
- Motivação:
  - Presença de outliers podem levar a modelos mal formados e resultados incorretos
  - Outliers podem representar situações de interesse do domínio de aplicação

Detecção de Anomalias

- Métodos estatísticos
  - Definem se um dado atributo de um exemplo é um outlier
  - Assumem uma distribuição dos dados e definem região de outliers
  - Exemplo: x é um outlier se

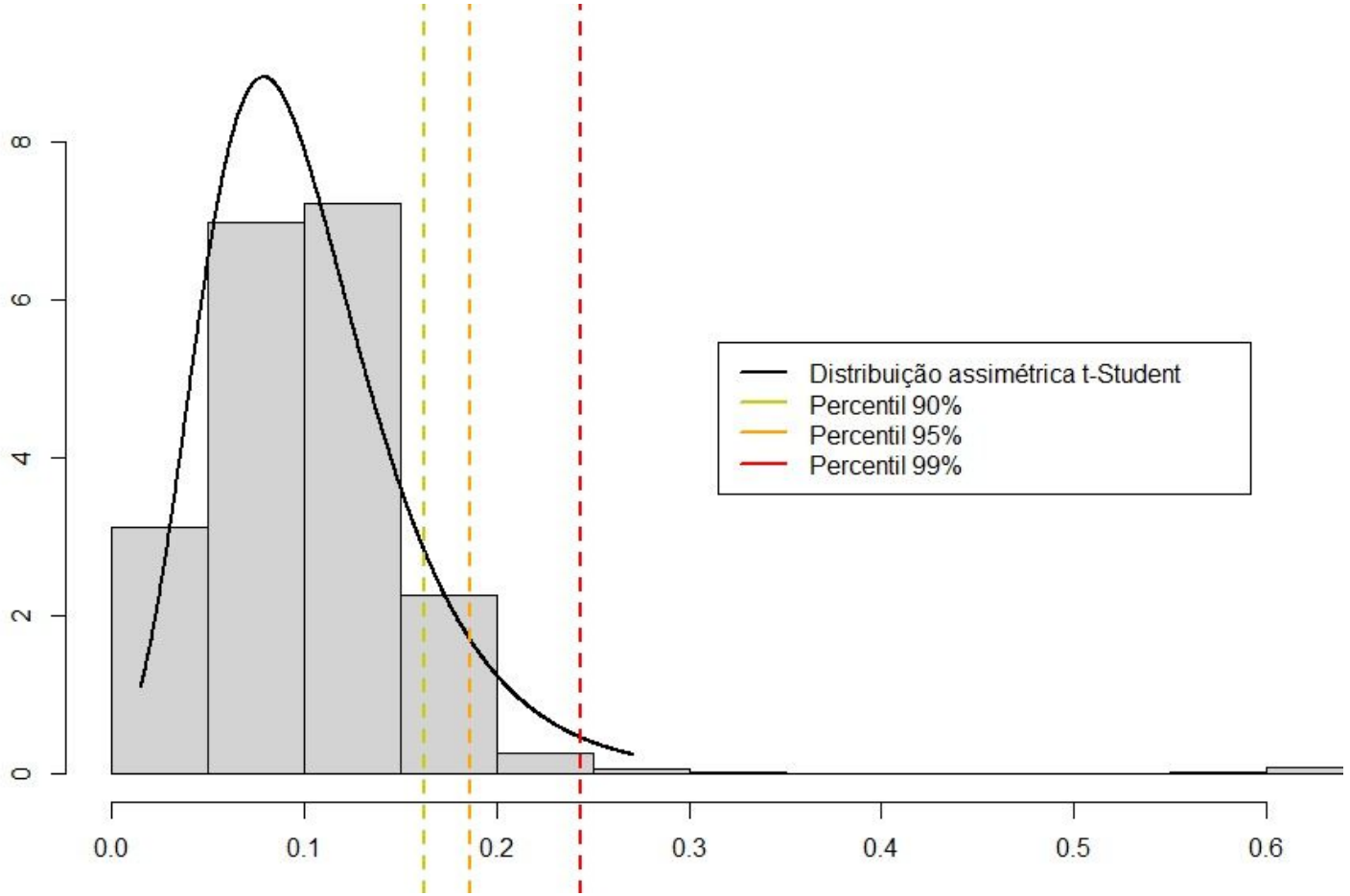
$$| x - \mu | > \rho \cdot \sigma$$

Média

Desvio Padrão

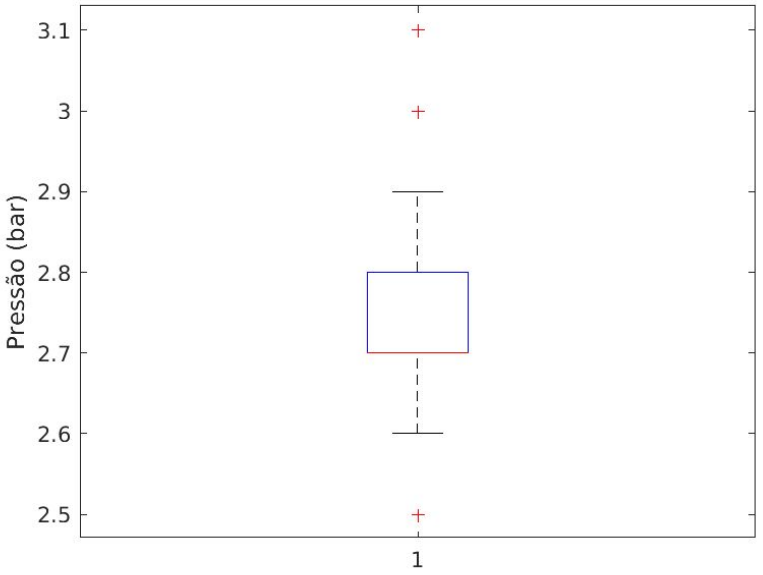
Parâmetro que define região de outlier (comumente igual a 3)

Detecção de Anomalias



Detecção de Anomalias

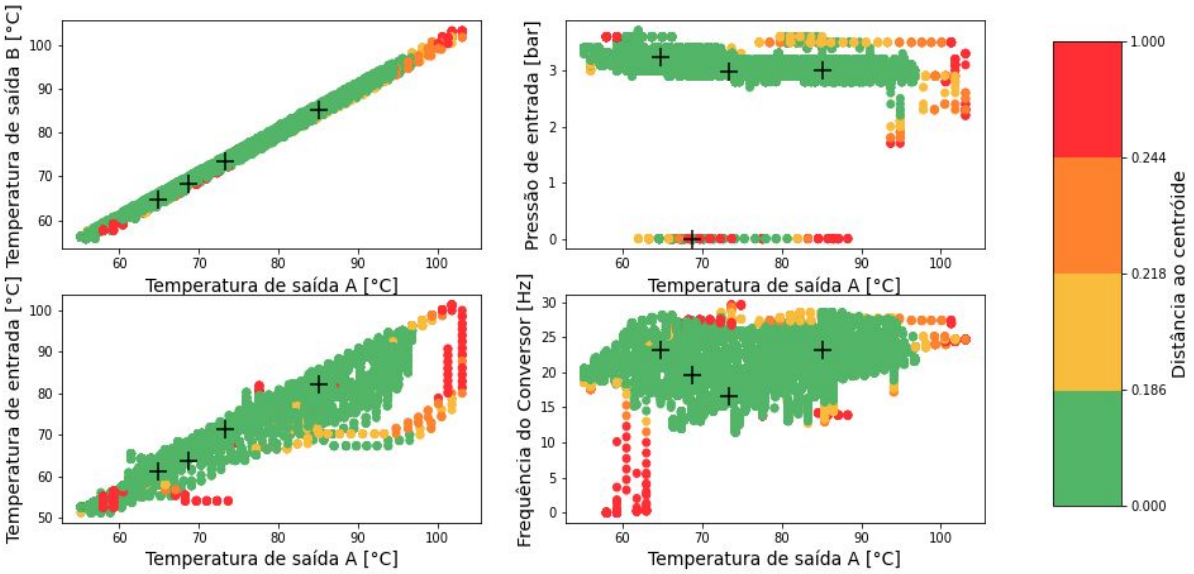
- Uso de gráficos Box Plot
  - Outlier é uma observação que está a 1.5*IRQ do primeiro ou do terceiro quartil dos dados
    - IRQ = Intervalo interquartil = Primeiro quartil menos terceito quartil



OBS.: Ver InterQuartileRange no WEKA

Detecção de Anomalias

- Métodos multivariados:
  - Permitem eliminar exemplos inteiros
  - E.g., usando medidas de distância



Tratamento de Dados Desbalanceados

- É comum nos depararmos com problemas de classificação com dados desbalanceados
  - I.e., Presença de classe majoritária com freqüência muito maior que as outras classes

- Desbalanceamento de classes pode ser prejudicial dependendo do problema e algoritmo

Desbalanceamento de Dados

- Conseqüência:
  - Maior tendência para a responder bem para as classes majoritárias em detrimento das minoritárias
  - Entretanto, em muitos casos, o que importa é ter um bom desempenho para as classes minoritárias!!!
    - Ver exemplos no próximo slide

Desbalanceamento de Dados

- Exemplo: Detecção de Fraude
  - Menos de 1% das transações de cartão de crédito são fraudes
  - Em um conjunto de exemplos relacionados a transações, teremos:
    - 99% dos exemplos para a classe negativa (não-fraude)
    - 1% dos exemplos para a classe positiva (fraude)

Desbalanceamento de Dados

- Exemplo: Detecção de Fraude
  - Classificadores terão uma tendência a dar respostas negativas para transações com fraude
    - I.e. Alto número de falsos negativos
  - Problema: o custo de um falso negativo é muito maior que o custo de um falso positivo
    - Falso negativo: fraude que não foi detectada em tempo
      - I.e., prejuízo a operadora de cartão
    - Falso positivo: transação normal bloqueada
      - I.e., aborrecimento para o usuário do cartão

Desbalanceamento de Dados

- Exemplo: Diagnóstico Médico
  - Pacientes doentes são em geral menos comuns que pacientes saudáveis
  - No diagnóstico médico, novamente a classe positiva (dos pacientes doentes) tem uma freqüência muito menor que a classe negativa (pacientes saudáveis)

Desbalanceamento de Dados

- Exemplo: Diagnóstico Médico
  - Classificadores terão uma tendência a classificar doentes reais como supostamente saudáveis
    - Novamente, alto número de falsos positivos
  - Conseqüência :
    - Diagnóstico tardio e dano para o paciente

Tratamento de Dados Desbalanceados

- Reamostragem aleatório:
  - Reamostragem dos exemplos de treinamento de forma a gerar conjuntos balanceados
  - Undersampling
    - Reduzir número de exemplos da classe majoritária
    - Pode acarretar em perda de informação
  - Oversampling
    - Replicar exemplos da classe minoritária
    - Se feito aleatoriamente, pode gerar overfitting

Abordagens para Tratamento de Dados Desbalanceados

- SMOTE
  - Oversampling usando exemplos sintéticos da classe minoritária
  - Exemplos sintéticos extraídos ao longo dos segmentos que unem vizinhos mais próximos da classe minoritária

Abordagens para Tratamento de Dados Desbalanceados

- Comparação entre métodos
  - Melhor método de sampling depende do algoritmo sendo utilizado e da métrica de avaliação
  - Comumente métodos de oversampling mais inteligentes (e.g., SMOTE) se dão bem