

FIAP GRADUAÇÃO

DATA SCIENCE

DATA GOVERNANCE & DATA SECURITY MANAGEMENT

Prof. Dr. Renê de Ávila Mendes

Objetivos da disciplina

DISCIPLINA: Data Governance & Data Security Management

OBJETIVOS: Descubra como funciona um **projeto de banco de dados** dentro de um ambiente corporativo, aplicando **técnicas de levantamento e documentação de requisitos**, aderente aos projetos de bancos de dados e aprenda a representar esses requisitos em arquiteturas de solução tecnológica para Data distribution e Data integration, modelos de estruturas de dados e dicionários de dados buscando **Data quality**. Garanta a qualidade dos dados de uma empresa para prover os melhores subsídios à tomada de decisão de negócio, praticando **Data cleaning** para limpar, harmonizar, complementar e corrigir dados inconsistentes, incompletos ou incorretos. Compreenda como funciona o **ciclo de vida da informação** e as responsabilidades administrativas sobre os dados de negócio, buscando qualidade, segurança e compatibilidade com políticas de administração de informação corporativas auditáveis, aplicando práticas atuais de **Data profiling** e conhecendo os princípios de **Data auditing**, de forma a atender a **Lei Geral de Proteção de Dados (LGPD)**.

Assuntos – 2º Semestre

- Qualidade em metadados
- Arquiteturas de integração e distribuição física de banco de dados
- Master Data Management e Data Hub
- **Qualidade de dados**
- Enterprise Data Management
- LGPD



Checkpoint 5

- **Data:** 04/09 (entrega até 09/09)
- **Local:** remoto



Checkpoint 6

- **Data: 18/09** (entrega no mesmo dia)
- **Local:** remoto

KDD - DESCOBERTA DE CONHECIMENTO EM BASES DE DADOS



O QUE KDD É

O KDD é o **processo não trivial** de identificar padrões válidos, novos, potencialmente úteis e compreensíveis nos dados.

(Fayyad, Piatetsky-Shapiro, and Smyth 1996)

Por não trivial, queremos dizer que alguma **pesquisa ou inferência** está envolvida; isto é, não é um cálculo direto de quantidades predefinidas, como calcular o valor médio de um conjunto de números.

FAYYAD, Usama; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic. From data mining to knowledge discovery in databases. **AI magazine**, v. 17, n. 3, p. 37, 1996.

O QUE KDD É

O KDD é o processo não trivial de identificar **padrões** válidos, novos, potencialmente úteis e compreensíveis nos dados.

(Fayyad, Piatetsky-Shapiro, and Smyth 1996)

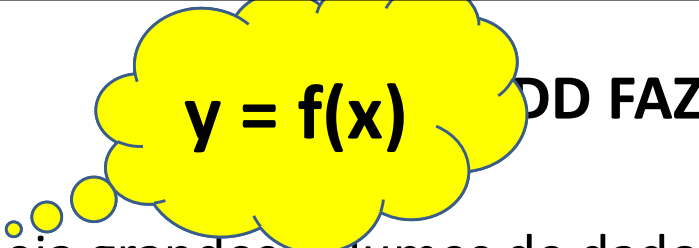
Aqui, dados são um conjunto de fatos (por exemplo, casos em um banco de dados) e padrão é uma **expressão em alguma linguagem** que descreve um subconjunto dos dados ou um modelo aplicável ao subconjunto.

FAYYAD, Usama; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic. From data mining to knowledge discovery in databases. **AI magazine**, v. 17, n. 3, p. 37, 1996.

O QUE KDD FAZ

Mapeia grandes volumes de dados em formas:

- Mais compactas (análise descritiva)
- Mais abstratas (padrão/processo de geração dos dados)
- Mais úteis (modelos preditivos)


$$y = f(x)$$

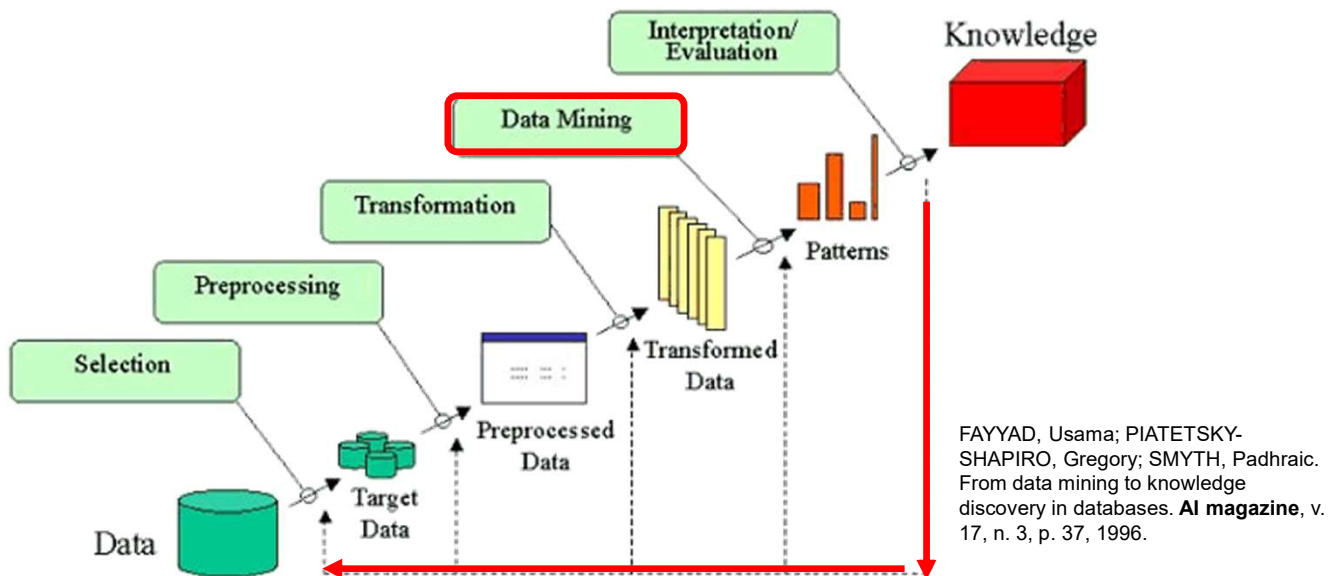
DD FAZ

Mapeia grandes volumes de dados em formas:

- Mais compactas (análise descritiva)
- Mais abstratas (padrão/processo de geração dos dados)
- Mais úteis (modelos preditivos)

Quando se pensa na palavra “mapear” deve-se pensar em uma função de mapeamento.

Descoberta de conhecimento em bases de dados



12

- Data Mining é **parte de um processo** de descoberta de conhecimento a partir dos dados
- Entre as tarefas de mineração está a classificação de dados
- Os passos de **preparação, seleção e limpeza dos dados** e a incorporação de conhecimento prévio são essenciais para garantir a utilidade dos resultados

MAS O DESEMPENHO DA CLASSIFICAÇÃO É CONHECIDO APENAS DEPOIS DA EXECUÇÃO

Processo custoso

Se a classificação for ruim, será necessário reiniciar o processo

MINERAÇÃO DE DADOS

Mineração de Dados (MD) pode ser definida como a aplicação de técnicas, implementadas por meio de algoritmos computacionais, capazes de receber como entrada **fatos** do mundo real e devolver como saída um **padrão** de comportamento.

(DA SILVA, Leandro Augusto; PERES, Sarajane Marques; BOSCARIOLI, Clodis. **Introdução à mineração de dados: com aplicações em R**. Elsevier Brasil, 2017, p11.)

Taxonomia das tarefas da Mineração (1º nível)

- **Tarefas descritivas**

Encontram padrões que **descrevem** os dados de maneira que possam ser interpretados mais facilmente.

- **Preditivas**

Usam valores de atributos descritivos para **prever** valores futuros ou desconhecidos de outros atributos.

(DA SILVA, Leandro Augusto; PERES, Sarajane Marques; BOSCAROLI, Clodis. **Introdução à mineração de dados: com aplicações em R**. Elsevier Brasil, 2017, p11.)





Tarefas de Mineração de Dados - Associação



		x_{ij}						y_i	
ID		$i1$	$i2$	$i3$	$i4$	$i5$	$i6$		
\vec{x}_i		Matrícula	Nome	Sexo	Estado civil	Idade	Especialidade	Pretensão Salarial	Experiência
	1	01	João	M	Viúvo	65	Garçom	R\$ 5000	SIM
	2	02	Maria	F	Solteiro	19	Cozinheiro	R\$ 2000	SIM

	n	n	Pedro	M	casado	30	Entregador	R\$ 700	NÃO

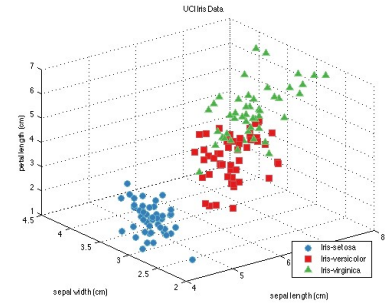
A tarefa de associação é definida como a busca por ocorrências frequentes e simultâneas entre elementos de um contexto.

(DA SILVA, Leandro Augusto; PERES, Sarajane Marques; BOSCARIOLI, Clodis. **Introdução à mineração de dados: com aplicações em R**. Elsevier Brasil, 2017, p14)

Taxonomia das tarefas da Mineração (2º nível)

Classificação e Estimação

Construção de modelos que avaliam a classe de um objeto não rotulado ou estimam o valor de um atributo.



Taxonomia das tarefas da Mineração (2º nível)

Classificação e Estimação

Construção de modelos que avaliam a classe de um objeto não rotulado ou estimam o valor de um atributo.




“O crédito será oferecido ou não ?”

“Qual o valor do crédito ?”

Tarefas de Mineração de Dados - Predição

		x_{ij}						y_i	
ID		$i1$	$i2$	$i3$	$i4$	$i5$	$i6$		
\vec{x}_i		Matrícula	Nome	Sexo	Estado civil	Idade	Especialidade	Pretensão Salarial	Experiência
	1	01	João	M	Viúvo	65	Garçom	R\$ 5000	SIM
	2	02	Maria	F	Solteiro	19	Cozinheiro	R\$ 2000	SIM

	n	n	Pedro	M	casado	30	Entregador	R\$ 700	NÃO



Tarefas de predição consistem na análise de um conjunto de dados nos quais estão presentes os dados, descritos por atributos, e seus rótulos associados. O objetivo dessa tarefa é descobrir um modelo capaz de mapear corretamente cada um dos dados aos seus rótulos.

(DA SILVA, Leandro Augusto; PERES, Sarajane Marques; BOSCARIOLI, Clodis. **Introdução à mineração de dados: com aplicações em R**. Elsevier Brasil, 2017, p13)

Taxonomia das tarefas da Mineração (2º nível)

Agrupamento

Agrupamento de conjuntos em classes de objetos similares, quando não se conhece o rótulo.



Tarefas de Mineração de Dados - Agrupamento

		x_{ij}							
ID		$i1$	$i2$	$i3$	$i4$	$i5$	$i6$	y_i	
\vec{x}_i		Matrícula	Nome	Sexo	Estado civil	Idade	Especialidade	Pretensão Salarial	Experiência
	1	01	João	M	Viúvo	65	Garçom	R\$ 5000	SIM
	2	02	Maria	F	Solteiro	19	Cozinheiro	R\$ 2000	SIM

	n	n	Pedro	M	casado	30	Entregador	R\$ 700	NÃO

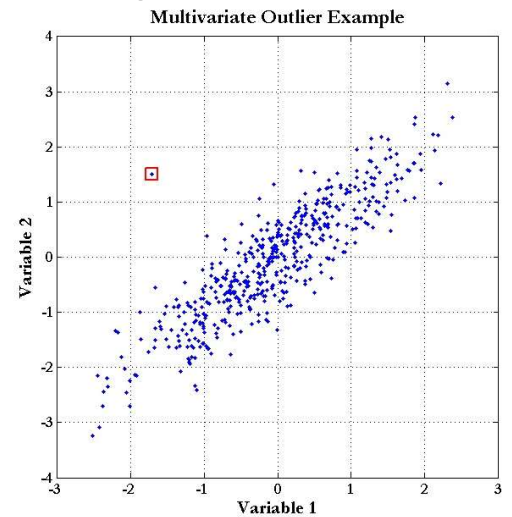
A tarefa de agrupamento de dados consiste na análise de conjuntos de dados em que estão presentes apenas as descrições dos dados. ... O objetivo na resolução dessa tarefa é descobrir relações entre os dados por meio de suas similaridades....

(DA SILVA, Leandro Augusto; PERES, Sarajane Marques; BOSCARIOLI, Clodis. **Introdução à mineração de dados: com aplicações em R**. Elsevier Brasil, 2017, p14)

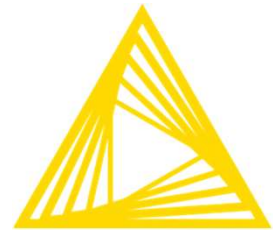
Taxonomia das tarefas da Mineração (2º nível)

Detecção de Anomalias

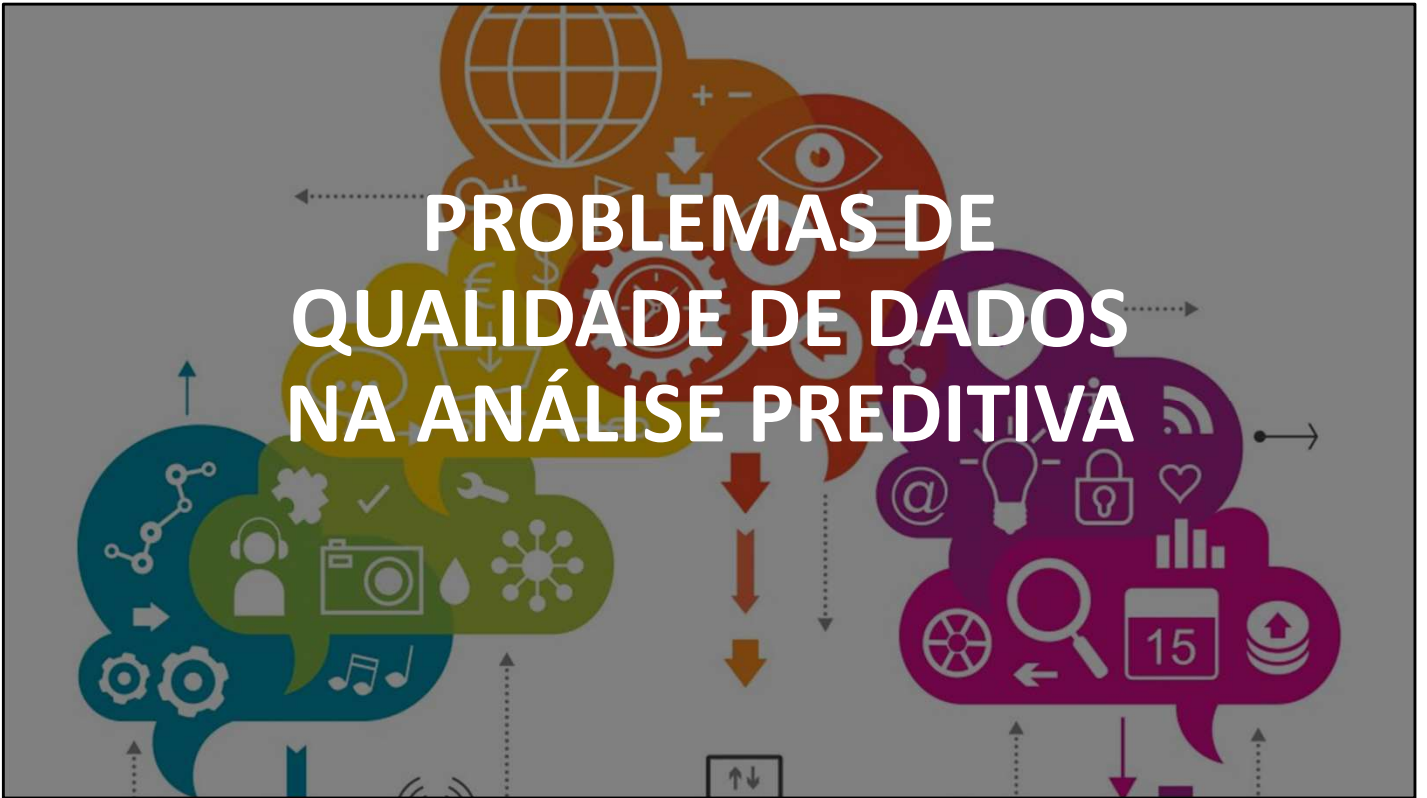
Identificação de objetos que não seguem um padrão característico comum aos dados. Outliers.



Algumas ferramentas e linguagens



PROBLEMAS DE QUALIDADE DE DADOS NA ANÁLISE PREDITIVA



MD – Algumas convenções

- **Dataset** – conjunto de dados usado no processo de mineração;
- **Base de conhecimento** – outro nome para os dados utilizados na mineração, quando destes se obtém um modelo que os explique;
- **Instância, objeto, exemplar, unidade observacional** – cada linha de um conjunto de dados, formalmente representada pelo vetor \vec{x}_i

(DA SILVA, Leandro Augusto; PERES, Sarajane Marques; BOSCAROLI, Clodis. **Introdução à mineração de dados: com aplicações em R**. Elsevier Brasil, 2017, p9)

Dataset e objeto

		x_{ij}							
ID		$i1$	$i2$	$i3$	$i4$	$i5$	$i6$	y_i	
\vec{x}_i		Matrícula	Nome	Sexo	Estado civil	Idade	Especialidade	Pretensão Salarial	Experiência
	1	01	João	M	Viúvo	65	Garçom	R\$ 5000	SIM
	2	02	Maria	F	Solteiro	19	Cozinheiro	R\$ 2000	SIM

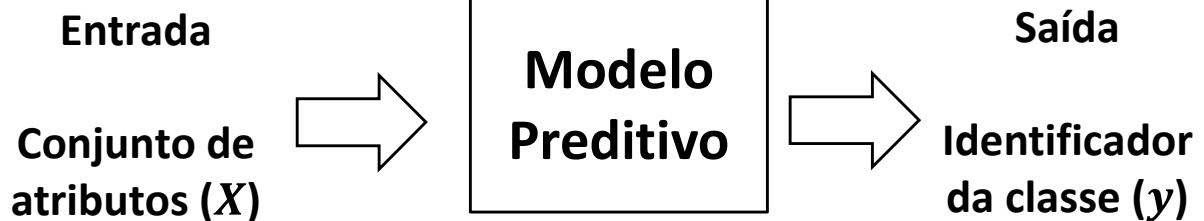
	n	n	Pedro	M	casado	30	Entregador	R\$ 700	NÃO

Dataset: $X = \{\vec{x}_1, \vec{x}_2, \dots, \vec{x}_i, \dots, \vec{x}_n\}$

Objeto: $\vec{x}_i = \{x_{i1}, x_{i2}, \dots, x_{ij}, \dots, x_{id}, \mathbf{y}\}$

(DA SILVA, Leandro Augusto; PERES, Sarajane Marques; BOSCARIOLI, Clodis. **Introdução à mineração de dados: com aplicações em R**. Elsevier Brasil, 2017, p10)

Análise Preditiva



Processo que permite descobrir o relacionamento entre exemplares de um dataset (objetos) e os rótulos a eles associados.


Processo que permite descobrir o relacionamento existente entre os exemplares de um conjunto de dados, descritos por uma série de características (atributos descritivos), e os rótulos a eles associados (atributos de classe).

(DA SILVA, Leandro Augusto; PERES, Sarajane Marques; BOSCARIOLI, Clodis. **Introdução à mineração de dados: com aplicações em R**. Elsevier Brasil, 2017, p77)

Análise Preditiva

		x_{ij}						y_i	
ID		$i1$	$i2$	$i3$	$i4$	$i5$	$i6$		
\vec{x}_i		Matrícula	Nome	Sexo	Estado civil	Idade	Especialidade	Pretensão Salarial	Experiência
	1	01	João	M	Viúvo	65	Garçom	R\$ 5000	SIM
	2	02	Maria	F	Solteiro	19	Cozinheiro	R\$ 2000	SIM

	n	n	Pedro	M	casado	30	Entregador	R\$ 700	NÃO



Processo que permite descobrir o relacionamento entre exemplares de um dataset (objetos) e os rótulos a eles associados.

Processo que permite descobrir o relacionamento existente entre os exemplares de um conjunto de dados, descritos por uma série de características (atributos descritivos), e os rótulos a eles associados (atributos de classe).

(DA SILVA, Leandro Augusto; PERES, Sarajane Marques; BOSCARIOLI, Clodis. **Introdução à mineração de dados: com aplicações em R**. Elsevier Brasil, 2017, p77)

E o que são os objetos e os rótulos ?

Objeto	Rótulo
Macarrão	Vinho tinto
Peixe	Vinho branco
Propaganda do Toyota Yaris	1.000.000 visualizações
Propaganda da Cultura Inglesa	50.000.000 visualizações

Objeto: evento no domínio da análise

Rótulo: (1) identificação da classe à qual o evento está associado, ou (2) um valor contínuo ao qual o evento está associado

Qual o prato adequado para consumir com vinho tinto ? E com vinho branco ? <- classificação (predição categórica)

Quantas visualizações um anúncio terá ? <- regressão (predição numérica)

(DA SILVA, Leandro Augusto; PERES, Sarajane Marques; BOSCARIOLI, Clodis. **Introdução à mineração de dados: com aplicações em R**. Elsevier Brasil, 2017, p77)

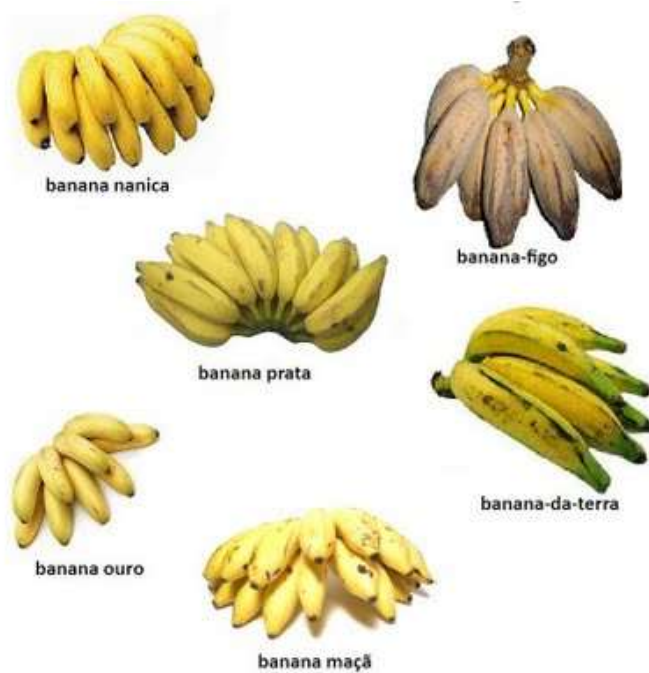
Predição categórica (classificação)

Objeto	Rótulo
Macarrão	Vinho tinto
Peixe	Vinho branco
Propaganda do Toyota Yaris	1.000.000 visualizações
Propaganda da Cultura Inglesa	50.000.000 visualizações

Objeto: evento no domínio da análise

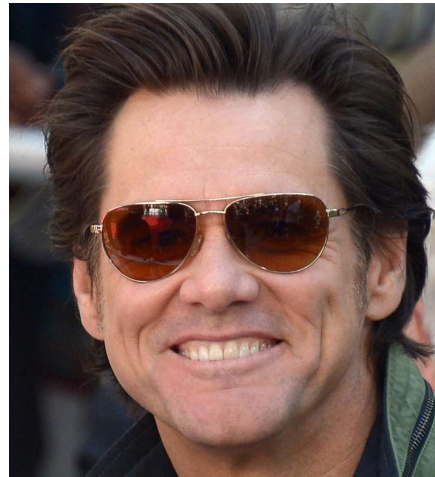
Rótulo: um valor discreto (numérico ou categórico) que identifica a classe à qual o evento está associado

(DA SILVA, Leandro Augusto; PERES, Sarajane Marques; BOSCARIOLI, Clodis. **Introdução à mineração de dados: com aplicações em R**. Elsevier Brasil, 2017, p77)



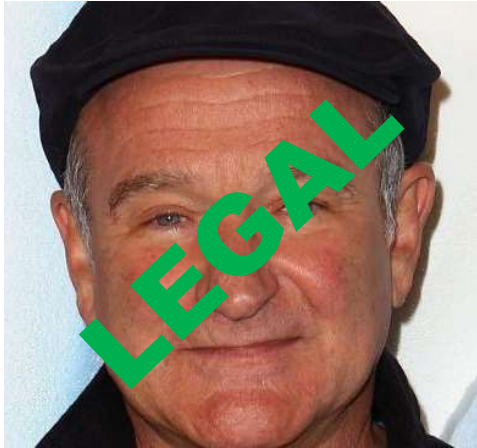
Exemplo: a classificação de uma banana qualquer entre uma dessas categorias dependerá das características da banana. As categorias são previamente conhecidas e nomeadas.

Predição categórica (classificação)



(DA SILVA, Leandro Augusto; PERES, Sarajane Marques; BOSCARIOLI, Clodis. **Introdução à mineração de dados: com aplicações em R**. Elsevier Brasil, 2017, p77)

Predição categórica (classificação)



(DA SILVA, Leandro Augusto; PERES, Sarajane Marques; BOSCARIOLI, Clodis. **Introdução à mineração de dados: com aplicações em R**. Elsevier Brasil, 2017, p77)

Predição numérica (regressão)

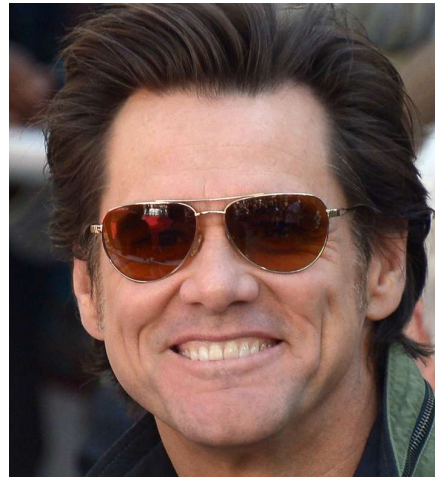
Objeto	Rótulo
Macarrão	Vinho tinto
Peixe	Vinho branco
Propaganda do Toyota Yaris	1.000.000 visualizações
Propaganda da Cultura Inglesa	50.000.000 visualizações

Objeto: evento no domínio da análise

Rótulo: um valor numérico contínuo associado ao objeto

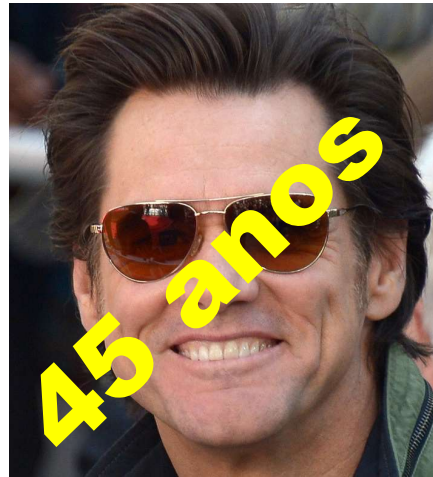
(DA SILVA, Leandro Augusto; PERES, Sarajane Marques; BOSCARIOLI, Clodis. **Introdução à mineração de dados: com aplicações em R**. Elsevier Brasil, 2017, p77)

Predição numérica (regressão)



(DA SILVA, Leandro Augusto; PERES, Sarajane Marques; BOSCARIOLI, Clodis. **Introdução à mineração de dados: com aplicações em R**. Elsevier Brasil, 2017, p77)

Predição numérica (regressão)



(DA SILVA, Leandro Augusto; PERES, Sarajane Marques; BOSCARIOLI, Clodis. **Introdução à mineração de dados: com aplicações em R**. Elsevier Brasil, 2017, p77)

Modelo preditivo

Relacionamento descoberto entre exemplares e rótulos, podendo ser descrito na forma de funções ou organizado em estruturas de dados.

O algoritmo adotado para a construção do modelo preditivo faz o **ajuste** dos parâmetros do modelo.

Uma vez determinado, o modelo preditivo pode ser usado para **predizer** o rótulo de exemplares desconhecidos.

(DA SILVA, Leandro Augusto; PERES, Sarajane Marques; BOSCARIOLI, Clodis. **Introdução à mineração de dados: com aplicações em R**. Elsevier Brasil, 2017, p77)

Modelo preditivo **treinamento**

Relacionamento descoberto entre exemplares e rótulos, podendo ser descrito na forma de funções ou organizado em estruturas de dados.

O algoritmo adotado para a construção do modelo preditivo faz o **ajuste** dos parâmetros do modelo.

Uma vez determinado, o modelo preditivo pode ser usado para **predizer** o rótulo de exemplares desconhecidos.

(DA SILVA, Leandro Augusto; PERES, Sarajane Marques; BOSCARIOLI, Clodis. **Introdução à mineração de dados: com aplicações em R**. Elsevier Brasil, 2017, p77)

Modelo preditivo

Relacionamento descoberto entre exemplares e rótulos, podendo ser descrito na forma de funções ou organizado em estruturas de dados.

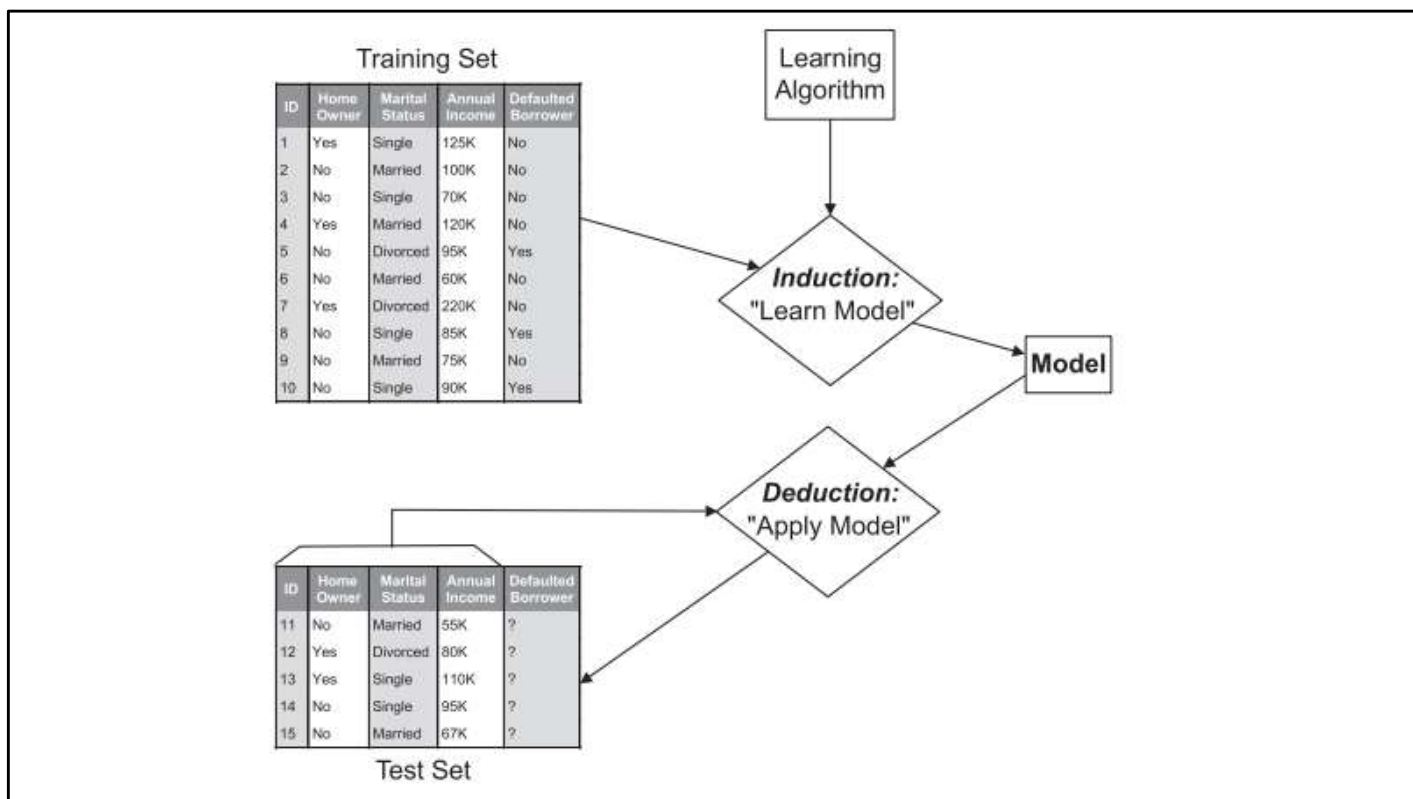
O algoritmo adotado para a construção do modelo preditivo faz o **ajuste** dos parâmetros do modelo.

teste

Uma vez determinado, o modelo preditivo pode ser usado para **predizer** o rótulo de exemplares desconhecidos.

O processo de aplicação do modelo é popularmente conhecido como teste e consiste em apresentar um novo exemplar para o modelo, que lhe fornecerá um rótulo de acordo como mapeamento previamente descoberto.

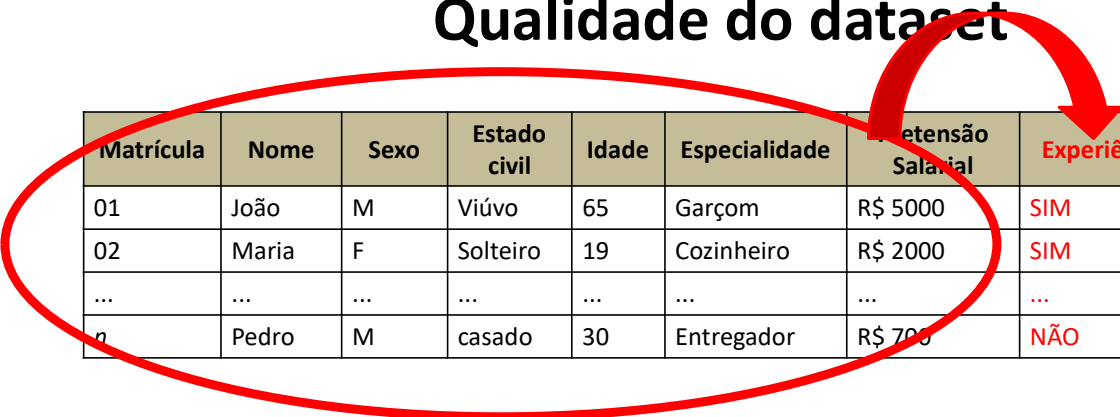
(DA SILVA, Leandro Augusto; PERES, Sarajane Marques; BOSCARIOLI, Clodis. **Introdução à mineração de dados: com aplicações em R**. Elsevier Brasil, 2017, p77)



TAN, Pang-Ning et al. **Introduction to data mining**. Pearson Education India, 2007.

Borrower: mutuário, pessoa que toma empréstimo

Qualidade do dataset

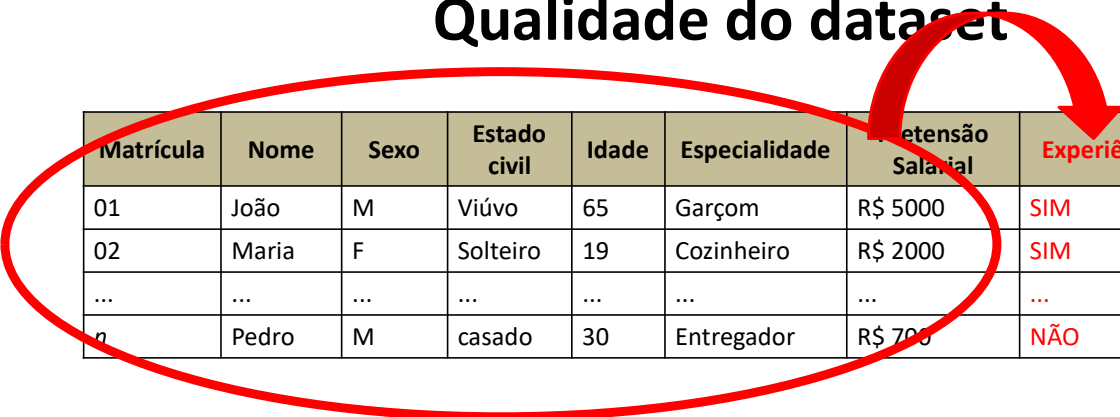


Matrícula	Nome	Sexo	Estado civil	Idade	Especialidade	Retensão Salarial	Experiência
01	João	M	Viúvo	65	Garçom	R\$ 5000	SIM
02	Maria	F	Solteiro	19	Cozinheiro	R\$ 2000	SIM
...
n	Pedro	M	casado	30	Entregador	R\$ 700	NÃO

Uma amostra (dataset) de **baixa qualidade** (atributos pouco representativos, poucos atributos, poucos exemplares, inconsistências) não será tão informativa a ponto de produzir um modelo de boa qualidade.

(DA SILVA, Leandro Augusto; PERES, Sarajane Marques; BOSCARIOLI, Clodis. **Introdução à mineração de dados: com aplicações em R**. Elsevier Brasil, 2017, p78)

Qualidade do dataset



Matrícula	Nome	Sexo	Estado civil	Idade	Especialidade	Retensão Salarial	Experiência
01	João	M	Viúvo	65	Garçom	R\$ 5000	SIM
02	Maria	F	Solteiro	19	Cozinheiro	R\$ 2000	SIM
...
n	Pedro	M	casado	30	Entregador	R\$ 700	NÃO



Qualidade



Erros de predição

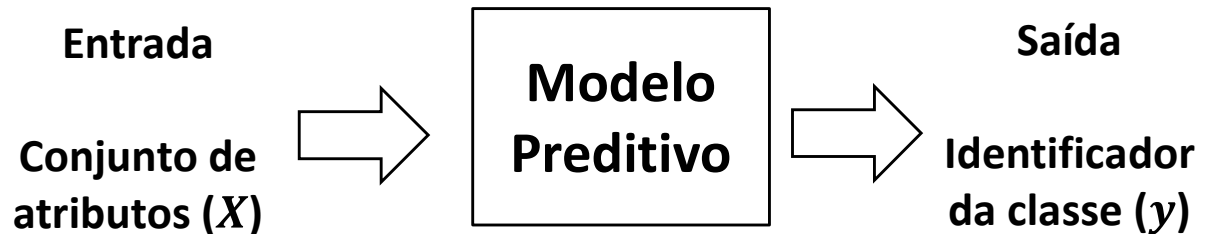
(DA SILVA, Leandro Augusto; PERES, Sarajane Marques; BOSCARIOLI, Clodis. **Introdução à mineração de dados: com aplicações em R**. Elsevier Brasil, 2017, p78)

O PROCESSO DE ANÁLISE DE DADOS



O processo de análise de dados

Data Mining: Análise Preditiva



DA SILVA, Leandro Augusto; PERES, Sarajane Marques; BOSCARIOLI, Clodis. **Introdução à mineração de dados: com aplicações em R**. Elsevier Brasil, 2017, p77

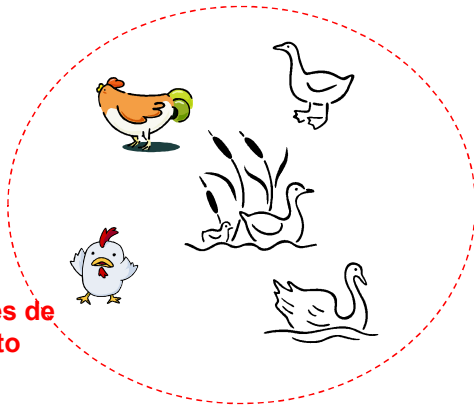
45

Processo que permite descobrir o relacionamento existente entre os exemplares de um conjunto de dados, descritos por uma série de características (atributos descritivos), e os rótulos a eles associados (atributos de classe).

O processo de análise de dados

Análise Preditiva: Exemplos

Exemplares de
treinamento



Exemplar
de teste

TAN, Pang-Ning et al. **Introduction to data mining**. Pearson Education India, 2007.

O processo de análise de dados

Análise Preditiva: Exemplos




Fonte da imagem: <https://www.mundoecologia.com.br>

O processo de análise de dados

Análise Preditiva: Como o *dataset* é estruturado

		x_{ij}						y_i	
ID		$i1$	$i2$	$i3$	$i4$	$i5$	$i6$		
\vec{x}_i		Matrícula	Nome	Sexo	Estado civil	Idade	Especialidade	Pretensão Salarial	Experiência
	1	01	João	M	Viúvo	65	Garçom	R\$ 5000	SIM
	2	02	Maria	F	Solteiro	19	Cozinheiro	R\$ 2000	SIM

	n	n	Pedro	M	casado	30	Entregador	R\$ 700	NÃO



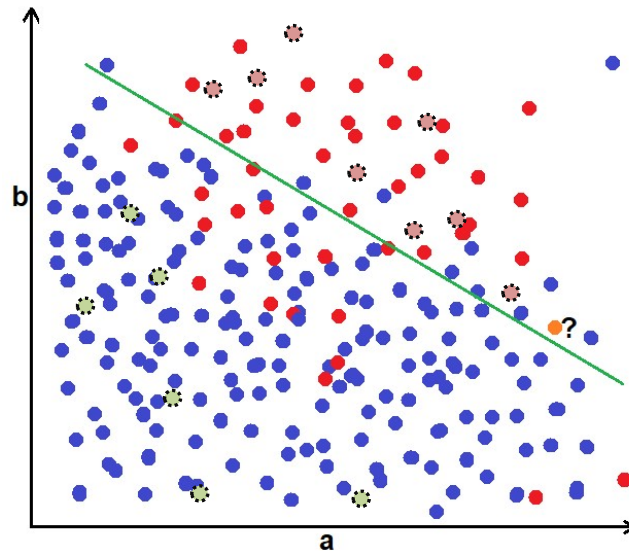
DA SILVA, Leandro Augusto; PERES, Sarajane Marques; BOSCARIOLI, Clodis. *Introdução à mineração de dados: com aplicações em R*. Elsevier Brasil, 2017, p77

48

Processo que permite descobrir o relacionamento existente entre os exemplares de um conjunto de dados, descritos por uma série de características (atributos descritivos), e os rótulos a eles associados (atributos de classe).

O processo de análise de dados

Análise Preditiva: Classificação de dados



49

A classificação de dados é uma tarefa de mineração de dados que consiste em ajustar parâmetros de um algoritmo por um conjunto de dados de treinamento usado com a finalidade de inferir a classe de um objeto (não classificado) em análise.

O processo de análise de dados

Definindo qualidade e complexidade de dados

QUALIDADE DE DADOS – Fidelidade com que os dados representam pessoas, objetos, eventos ou conceitos. Quanto maior a qualidade, maior a proximidade entre a representação e o objeto ou fato representado.

COMPLEXIDADE DE DADOS – Esforço necessário para descrever um conjunto de dados. Quanto maior o esforço, mais complexos são os dados.

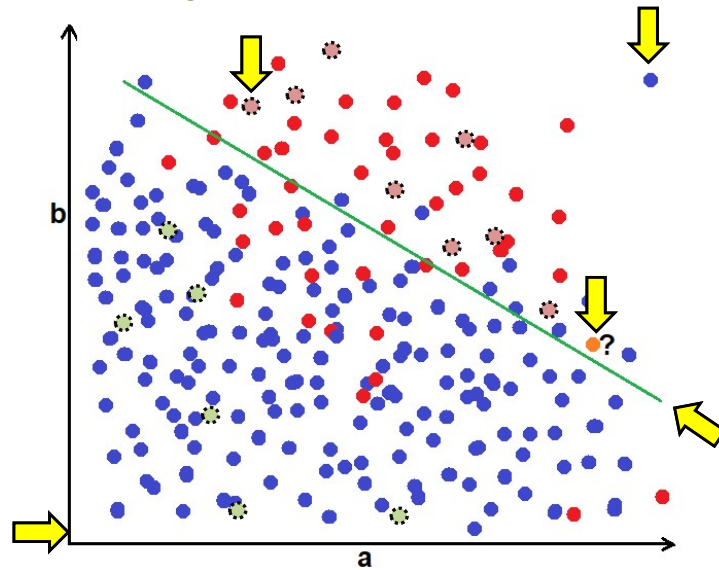
DE ÁVILA MENDES, Renê; DA SILVA, Leandro Augusto. Modeling the combined influence of complexity and quality in supervised learning. *Intelligent Data Analysis*, v. 26, n. 5, p. 1247-1274, 2022.

O processo de análise de dados

Análise Preditiva: Classificação de dados

Problemas intrínsecos nos dados:

- dados ausentes (missing values)
- dados discrepantes (outliers)
- sobreposição de atributos
- dimensionalidade do conjunto de dados



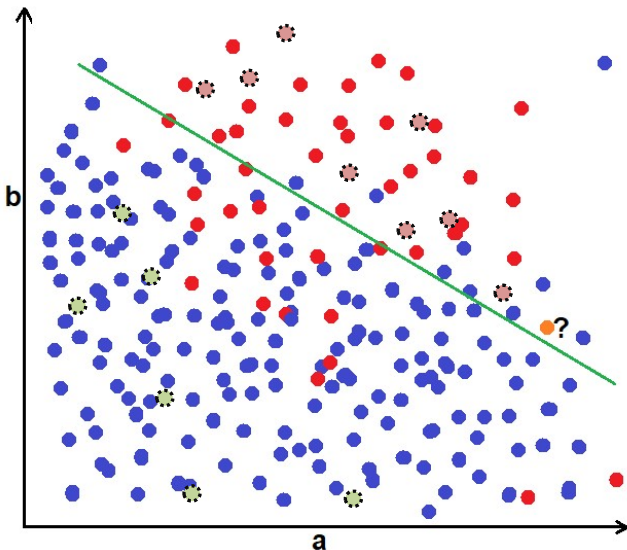
51

Problemas intrínsecos nos dados:

- dados ausentes (missing values)
- dados discrepantes (outliers)
- sobreposição de atributos
- dimensionalidade do conjunto de dados

O processo de análise de dados

O que afeta o desempenho da classificação



QUALIDADE DOS DADOS

Valores ausentes (*missing values*)

Valores discrepantes (*outliers*)

COMPLEXIDADE DOS DADOS

Dimensionalidade

Sobreposição de atributos

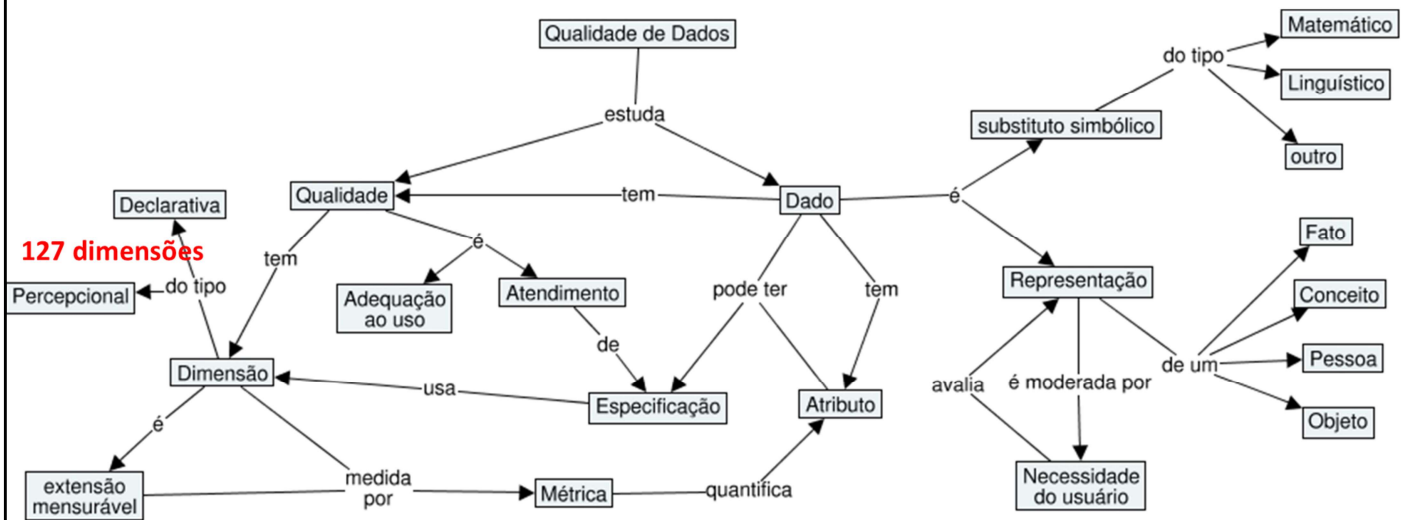
Separabilidade das classes

LORENA, Ana C. et al. How Complex is your classification problem? A survey on measuring classification complexity. *ACM Computing Surveys (CSUR)*, v. 52, n. 5, p. 1-34, 2019.

52

Uma parte significativa do desempenho do algoritmo de classificação depende da complexidade e da qualidade do conjunto de dados. A Complexidade dos Dados envolve a investigação dos efeitos da **dimensionalidade**, da **sobreposição de atributos descritivos** e da **separabilidade das classes**.

Qualidade de Dados



JAYAWARDENE, Vimuthki; SADIQ, Shazia; INDULSKA, Marta. **An analysis of data quality dimensions**. 2015.

53

As literaturas acadêmica e técnica apresentam dimensões da qualidade de dados, que podem ser agrupadas em duas categorias: declarativas e de uso;

DECLARATIVAS – intrínsecas aos dados; que em si mesmas explicam os dados
 DE USO – avaliações quanto à adequação ao uso

Mapa conceitual de qualidade de dados.



PRÉ-PROCESSAMENTO E ANÁLISE EXPLORATÓRIA

Pré-processamento

- Trata falhas e inconsistências
- Organiza os dados para a análise
- Pode ser a diferença entre o sucesso e o insucesso da análise
- Usa a análise exploratória para revelar problemas
 - Ausência de valores
 - Ruídos
 - Redundâncias

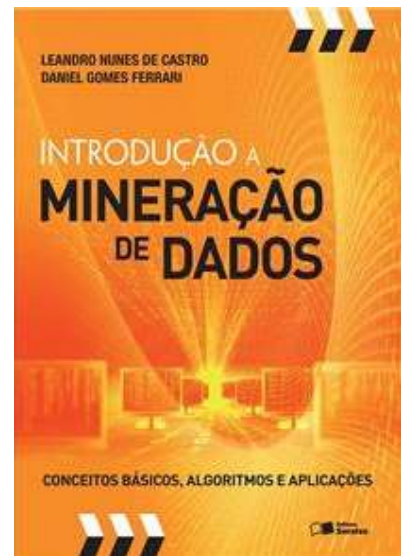
(DA SILVA, Leandro Augusto; PERES, Sarajane Marques; BOSCARIOLI, Clodis. **Introdução à mineração de dados: com aplicações em R**. Elsevier Brasil, 2017, p42)

Análise Exploratória dos Dados

- Ajuda a descrever o *dataset*
- Faz uso da estatística descritiva
- Pode identificar
 - ruídos
 - dados que precisam ser transformados
 - atributos que podem ser retirados
- Apoia a escolha da tarefa de mineração de dados
- Apoia a escolha do algoritmo

(DA SILVA, Leandro Augusto; PERES, Sarajane Marques; BOSCARIOLI, Clodis. **Introdução à mineração de dados: com aplicações em R**. Elsevier Brasil, 2017, p27)

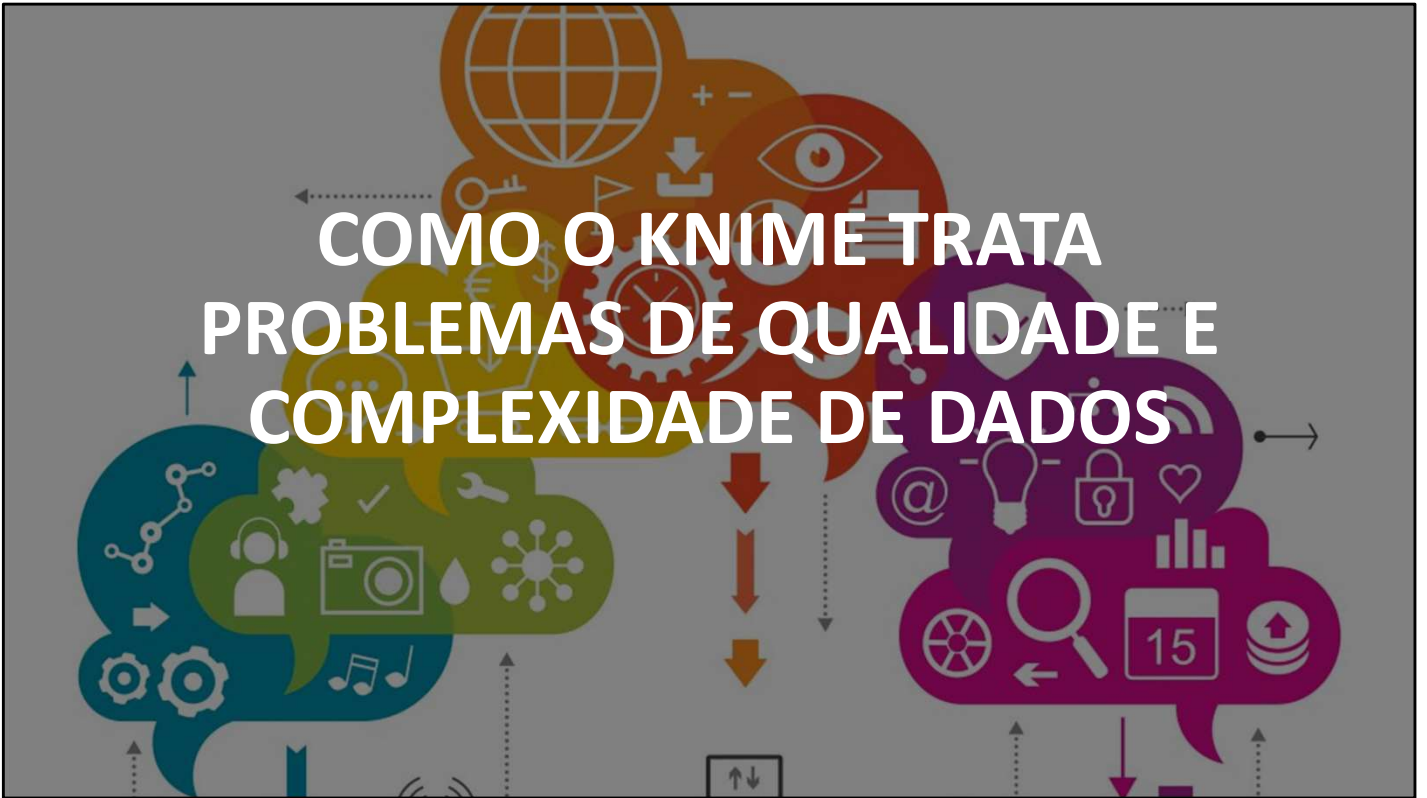
Pré-processamento (pdf, pp.
14 a 65)



CASTRO, Leandro Nunes de; FERRARI, Daniel Gomes. Introdução à Mineração de Dados: conceitos básicos, algoritmos e aplicações. **São Paulo: Saraiva**, 2016.

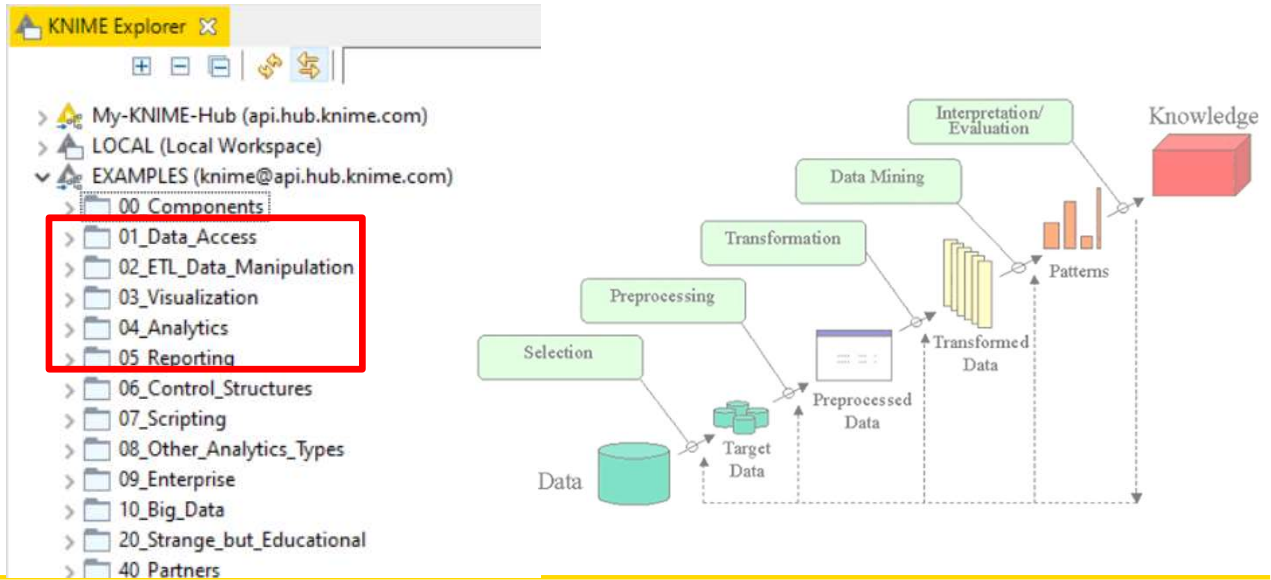
<https://pt.slideshare.net/Indecastro/2016-introduo-minerao-de-dados-conceitos-bsicos-algoritmos-e-aplicaes>

COMO O KNIME TRATA PROBLEMAS DE QUALIDADE E COMPLEXIDADE DE DADOS



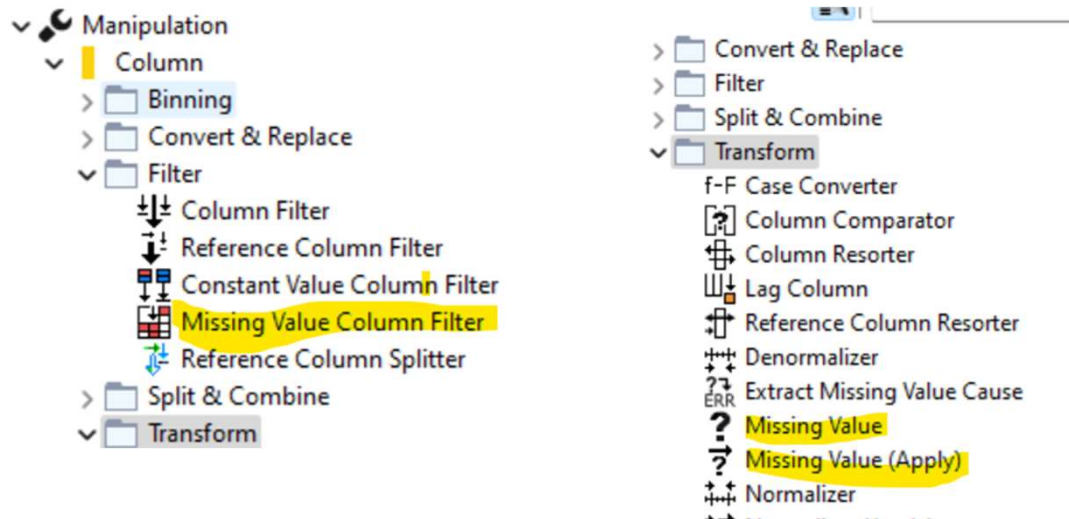
Apoiando o processo de análise

Os exemplos seguem a lógica do processo



Oferecendo nodes e workflows prontos

Para detectar e tratar *missing values*



Oferecendo nodes e workflows prontos

Para detectar e tratar *missing values*











Missing Value

- Define como manipular valores ausentes para todos os atributos de um determinado tipo
- Define como manipular valores ausentes para cada atributo do *dataset*

Oferecendo nodes e workflows prontos

Para detectar e tratar *outliers*

- ▼  EXAMPLES (knime@api.hub.knime.com)
 - >  00_Components
 - >  01_Data_Access
 - ▼  02_ETL_Data_Manipulation
 - >  00_Basic_Examples
 - ▼  01_Filtering
 - >  07_Four_Techniques_Outlier_Detection
 - ▲ 01_Validating_Datatables
 - ▲ 02_Column_Filter
 - ▲ 03_Row_Filtering
 - ▲ 04_Advanced_Row_Filters
 - ▲ 05_More_Row_Filter_Examples
 - ▲ 06_More_Column_Filter_Examples
 - ▲ 08_Filtering_Duplicates
 - >  02_Aggregations

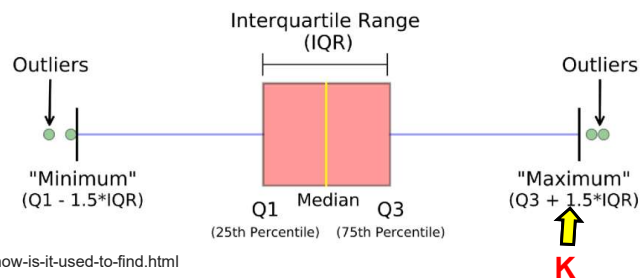
Oferecendo nodes e workflows prontos

Para detectar e tratar *missing values*

Numeric Outliers



- Detecta e trata outliers para cada atributo selecionado por meio de IQR (*interquartile range*)



<https://theprofessionalspoint.blogspot.com/2019/04/what-is-boxplot-how-is-it-used-to-find.html>

Permitindo a implementação em R

Para medir a complexidade do *dataset*

R Script

```
1 knime.out <- knime.in
2 library("ECoL")
3
4 dataset <- knime.in
5 knime.out <- as.data.frame(complexity(class ~ ., dataset))
6
```



<https://github.com/lpfgarcia/ECoL>

<https://cran.r-project.org/bin/windows/Rtools/>

KNIME Documentation > KNIME Integrations 4.6 > KNIME Interactive R Statistics Integration Installation Guide