



Licap

Formação Cientista de Dados



PUC Minas

Formação do Cientista de Dados

O Processo para Descoberta de Conhecimento – Módulo Básico

Luis Enrique Zárate

→ Conteúdo do Curso



Processo de descoberta de conhecimento em bases de dados – KDD

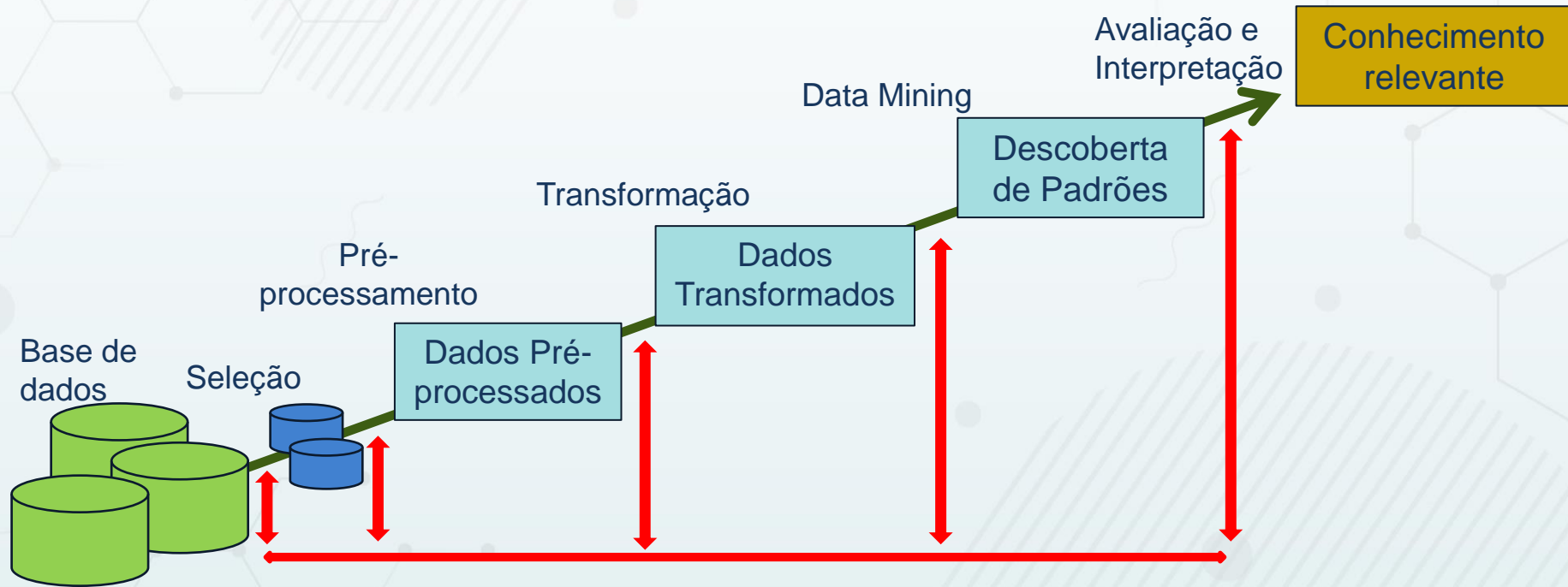
1. Detalhamento do processo KDD
2. Metodologias KDD
3. Tempo gasto para um projeto KDD
4. Importância de cada etapa de um projeto KDD

O que é Descoberta de Conhecimento em Base de Dados (KDD)?

“Processo não-trivial de identificação de padrões válidos, novos, potencialmente úteis e finalmente compreensíveis a partir de dados”

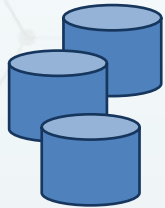
(Usama Fayyad)

Processo de Descoberta de conhecimento em Base de Dados - KDD



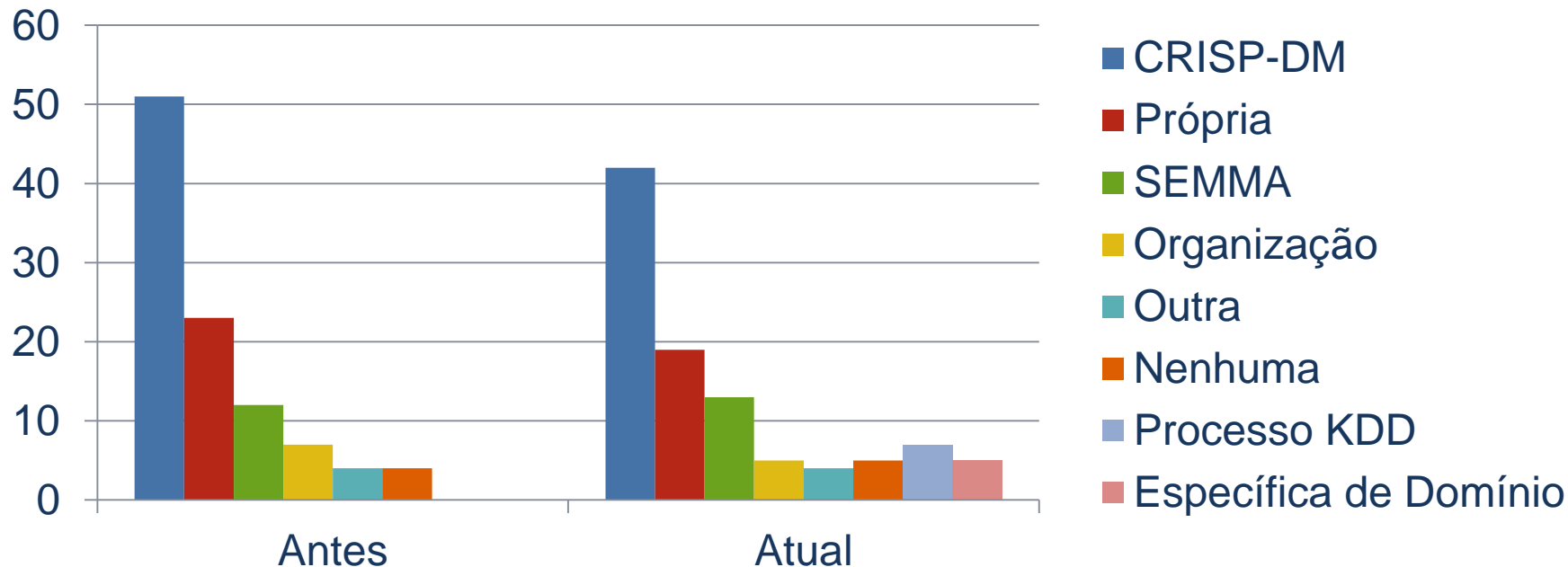
Etapas de um processo KDD - Adaptado de Fayyad et al (1996)

Detalhamento do Processo KDD



- *Modelamento do domínio de problema*
- *Montagem da base de dados*
- *Enriquecimento e Melhoramento da base de dados*
- *Limpeza de dados*
- *Análise de Outliers e de dados ausentes*
- *Integração e combinação de dados*
- *Discretização, Codificação ou Transformação*
- *Data Mining*
- *Validação de Padrões*
- *Visualização e Apresentação do Conhecimento*

Metodologias KDD



Fonte: Adaptado de (KDNUGGETS)

➡ Cross-industry standard process for data mining

CRISPDM: Padrão de projetos KDD (independente do problema de industria)

- ❑ Entendimento do Negócio: Entender o objetivo do projeto e as expectativas para o negócio.
- ❑ Entendimento dos Dados: Montagem da base de dados e exploração dos dados.
- ❑ Preparação dos Dados: Tratamento dos dados
- ❑ Modelagem do problema: Aplicação de técnicas de Aprendizado de máquina
- ❑ Avaliação: validação e teste do modelo
- ❑ Implantação: o conhecimento adquirido pelo modelo é organizado e apresentado de uma maneira que o cliente possa utilizar.



→ Metodologia – CRISP-DM



- ❑ De acordo com pesquisas, a metodologia mais utilizada é a CRISP-DM. Esta está vinculada à ferramenta de Mineração de Dados mais vendida do mercado, a SPSS-Clementine, hoje de propriedade da IBM.
- ❑ As ferramentas utilizadas em processos de KDD possuem uma excelente interface com o usuário, mas não possuem um método que sustente a aplicação das tarefas de Data Mining.

→ Metodologia – CRISP-DM



- ❑ CRISP-DM aparentemente controla a execução das tarefas, não entanto, a decisão de quais tarefas executar ainda é fortemente dependente do responsável pelo projeto.
- ❑ O CRISP-DM torna difícil a condução por profissionais menos experientes, desde que não há um detalhamento do que precisa ser feito ou de um fluxo a ser seguido de forma pedagógica.

Metodologia - SEMMA

- ❑ A metodologia SEMMA, desenvolvida pela SAS, foi a terceira metodologia mais utilizada nos anos citados.
- ❑ Embora SEMMA não apresente retornos em seu fluxo, espera-se que isto aconteça.
- ❑ Retornos podem levar ao aumento de custos (financeiro e de tempo), o que pode inviabilizar a continuidade do projeto.



→ Discussão acerca das Metodologias

- Metodologias “**Próprias**” é a segunda mais aplicada.
- Atualmente não existe **padrão universal** para processos KDD.
- A falta de **padronização** leva à falta de produtividade.
- Surgimento do ***Domain-Driven Data Mining (D3M)***
- É necessário o acompanhamento e validação das etapas por um especialista de negócio (**D3M**).

Discussão acerca das Metodologias

- Há necessidade de **documentação** para projetos KDD
- Há necessidade de metodologias para **Model Explainer**
- Atualmente, os projetos são desenvolvidos de **forma empírica** mais do que ciência.
- Há necessidade de projetos mais **aderentes** à realidade e necessidade das organizações.

→ Metodologia PICTOREA



É um novo método, pedagógico, para desenvolvimento, acompanhamento e documentação das etapas e atividades de projetos KDD.

PICTOREA modela o fluxo das etapas necessárias ao processo, bem como os atores responsáveis por cada uma delas.

ETAPA	Cientista de dados	Especialista de domínio
Exploração do espaço problema	x	x
Definição do espaço solução	x	x
Entendimento do domínio	x	x
Caracterização do problema por meio de atributos	x	
Montagem da base de dados	x	
Exploração dos dados e análise da representatividade da base de dados	x	
Seleção de atributos e redução de dimensionalidade	x	
Preparação dos dados	x	x
Mineração de dados	x	x
Validação estatística	x	x
Interpretação e Auditoria modelos	x	x
Visualização	x	x

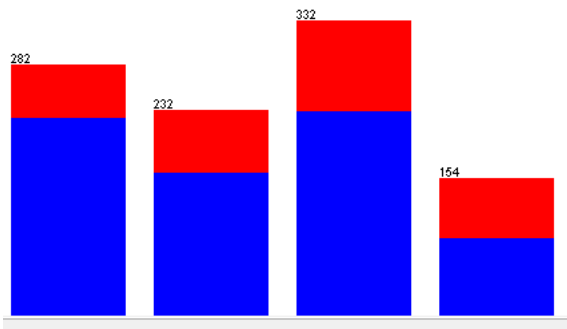
Exemplo Ilustrativo

Exploração do espaço problema:

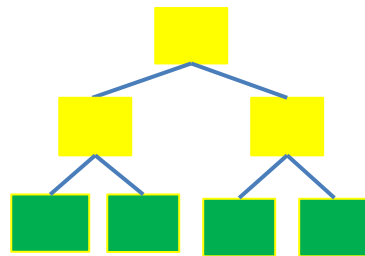
Problema	Importância	Dificuldade	Retorno	Resultado
PROB 1	5	3	2	3,75
PROB 2	2	1	4	2,25
PROB 3	1	2	6	2,25
PROB 4	6	6	3	5,25
PROB 5	3	4	1	2,75
PROB 6	4	5	5	4,5

PROB4 = <Perfil de crédito>

Exploração dos dados:



Exploração do espaço solução:

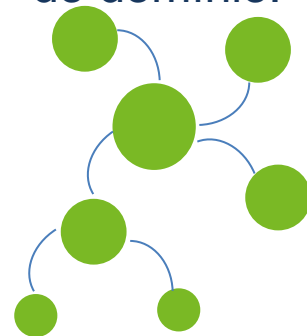


SOLUÇÃO= <Classificação>

Montagem da
base de dados



Entendimento
do domínio:



Caracterização
por atributos

E.C.= {S,C,V,D}
Idade=[25-55]
.....

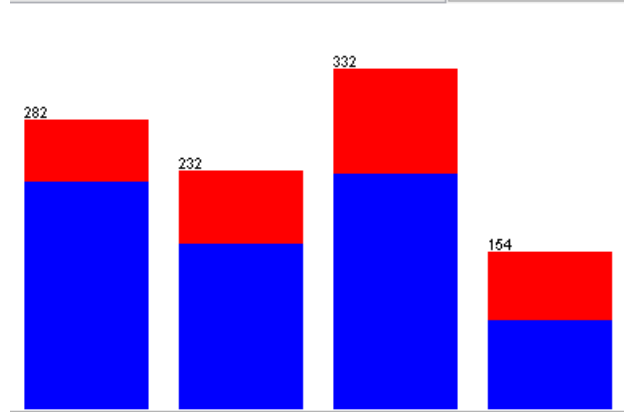
Exploração dos dados:

Selected attribute

Name: property_magnitude Type: Nominal
Missing: 0 (0%) Distinct: 4 Unique: 0 (0%)

No.	Label	Count
1	real estate	282
2	life insurance	232
3	car	332
4	no known property	154

Class: class (Nom) Visualize All

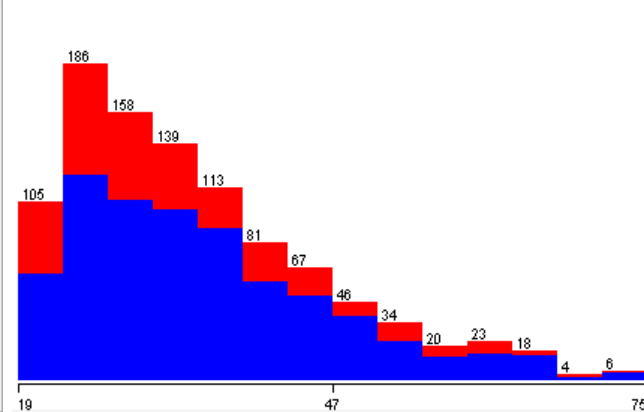


Selected attribute

Name: age Type: Numeric
Missing: 0 (0%) Distinct: 53 Unique: 1 (0%)

Statistic	Value
Minimum	19
Maximum	75
Mean	35.546
StdDev	11.375

Class: class (Nom) Visualize All

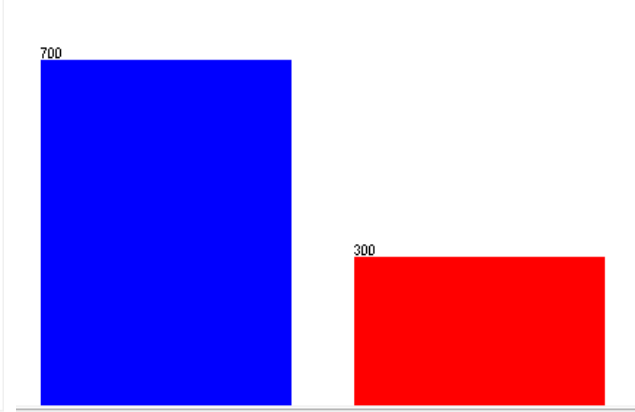


Selected attribute


Name: class Type: Nominal
Missing: 0 (0%) Distinct: 2 Unique: 0 (0%)

No.	Label	Count
1	good	700
2	bad	300

Class: class (Nom) Visualize All




Seleção de atributos:



No.	Name
1	checking_status
2	duration
3	credit_history
4	purpose
5	credit_amount
6	savings_status
7	employment
8	installment_commitment
9	personal_status
10	other_parties
11	residence_since
12	property_magnitude
13	age
14	other_payment_plans
15	housing
16	existing_credits
17	job
18	num_dependents
19	own_telephone
20	foreign_worker
21	class



No.	Name
1	duration
2	credit_history
3	purpose
4	credit_amount
5	employment
6	personal_status
7	residence_since
8	property_magnitude
9	age
10	housing
11	existing_credits
12	job
13	num_dependents
14	foreign_worker
15	class



Preparação dos dados:

No.	Name
1	duration
2	credit_history
3	purpose
4	credit_amount
5	employment
6	personal_status
7	residence_since
8	property_magnitude
9	age
10	housing
11	existing_credits
12	job
13	num_dependents
14	foreign_worker
15	class

Selected attribute

Name: employment

Type: Nominal

Missing: 0 (0%)

Distinct: 5

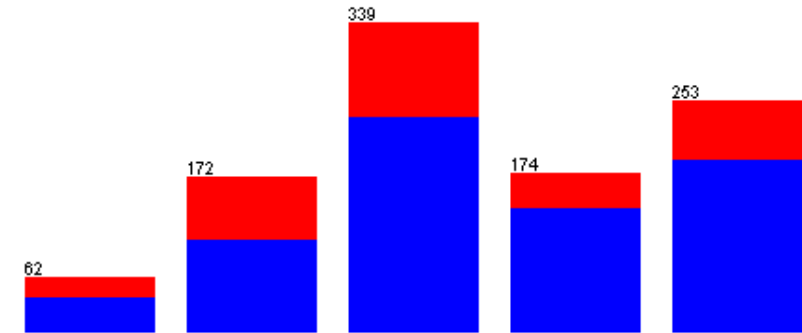
Unique: 0 (0%)

No.	Label	Count
1	unemployed	62
2	<1	172
3	1<=X<4	339
4	4<=X<7	174
5	>=7	253

Class: class (Nom)

Visualize All

Será que os intervalos são os mais adequados?



Mineração de dados: Árvore de Decisão J48

```

credit_history = existing paid
| credit_amount <= 8648
| | duration <= 40: good (476)
| | duration > 40
| | | num_dependents <= 1: bad (22)
| | | num_dependents > 1: good (5)
| credit_amount > 8648: bad (27)

```

Validação Estatística:

Clasificadas:

Coretamente 143 (71,5%)

Incoretamente 57 (28,5%)

```

credit_history = no credits/all paid
| housing = rent: bad (10)
| housing = own
| | duration <= 27: bad (14)
| | duration > 27: good (11)
| housing = for free: bad (5)

```

Precision	Recall	F-Measure	Class
0.763	0.856	0.807	good
0.545	0.393	0.457	bad

Desafios!!

Interpretação e Auditoria de Modelos:

credit_history = no credits/all paid
| housing = rent: bad (10)
| housing = own
| | duration <= 27: bad (14)
| | duration > 27: good (11)
| housing = for free: bad (5)

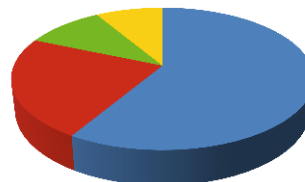
SE Nunca obteve crédito &
sempre pagou & mora em casa
Alugada = MAL PAGADOR

Parece discriminatório = ?

Visualização do
conhecimento descoberto:

O modelo é:
Representativo?
Possui restrições?
Está polarizado?
É interpretável?
Qual é a vida útil?
É moral?

Data 1	Data 2	Data 3
...
...
...



```
void main()  
{  
...  
}
```

→ Tempos gastos por projeto em Ciência dos Dados



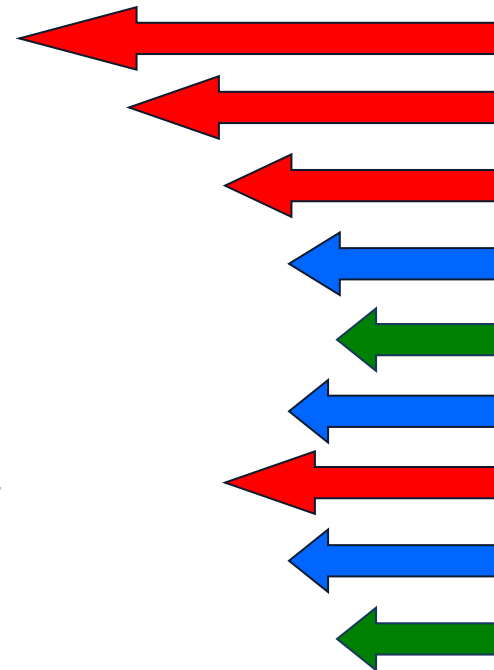
Importância de cada etapa num projeto de Ciência dos Dados



Etapas do KDD:

Definir e modelar corretamente o problema
Montagem e Representatividade da base de dados
Seleção de atributos
Limpeza de dados, outliers e dados ausentes
Integração e combinação de dados
Discretização, codificação e transformação
Data Mining
Validação de Padrões
Visualização e Apresentação do Conhecimento

Importância



- Não existem fatores desconhecidos que levem ao insucesso.
- Não existe ferramenta computacional ou metodologia única que garanta o sucesso.
- A padronização é resultado da aplicação do processo KDD para um específico problema. Embora seja possível generalizar alguns procedimentos.
- É necessária a formação de profissionais com domínio Data Science, KDD e DM.



Licap

Formação Cientista de Dados

Obrigado!



PUC Minas