oncase
power your business

Ferramentas de Mineração de Dados com o PDI

Usando o Pentaho Data Integration no Processo de Mineração de Dados

Mailson Filho
@maialson

Oncase

Treinamentos

Oficiais
In-company
Certificações

Suporte

Produto
Desenvolvedores
Mentoring

Consultoria

Architecture
Data Visualization
Cloud Computing

Fábrica de Inovações

Plugins
Apps de apoio
Produtos Analíticos

CERTIFICAÇÕES

CMMI-DEV	CMMI-SVC	MPS.BR
Maturity Level 2	Maturity Level 2	Maturity Level G

2013 PARTNER OF THE YEAR

pentaho Premium Partner

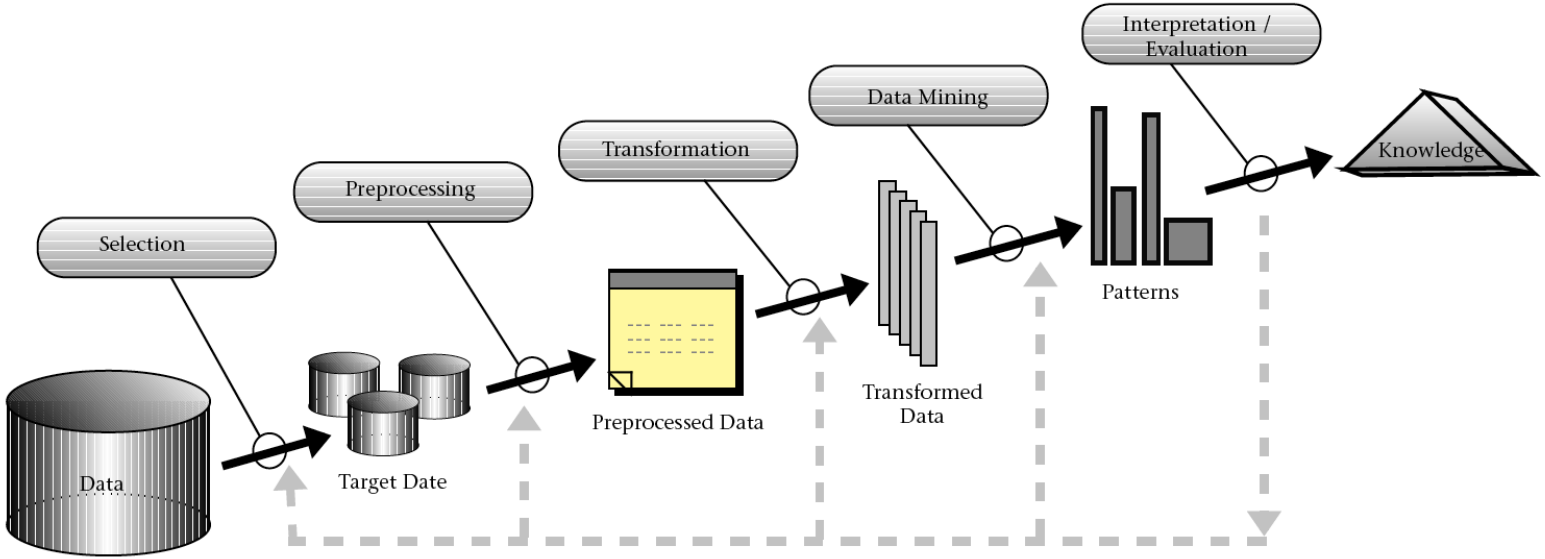
2014 PARTNER OF THE YEAR

oncase

O que é Data Mining?

Knowledge Discovery in Databases

A clássica


Interpretation / Evaluation
Data Mining
Transformation
Preprocessing
Selection
Knowledge
Patterns
Transformed Data
Preprocessed Data
Target Date
Data

Abordagens

Tarefas

- Descrição
- Classificação
- Estimação
- Predição
- Agrupamento
- Associação

Tarefa - Descrição

É a tarefa utilizada para descrever os padrões e tendências revelados pelos dados.

Exemplo: Traçar perfis comportamentais. Pessoas envolvidas em fraudes de cartão de crédito em gerão são hones, entre 25 e 40 anos, com um bom nível de instrução.

Tarefa - Classificação

Uma das tarefas mais comuns, visa identificar a qual classe um determinado registro pertence.

- Determinar quando uma transação pode ser uma fraude
- Identificar em uma escola, qual a turma mais indicada para um determinado aluno
- Diagnosticar onde uma determinada doença pode estsar presente
- Identificar quando uma pessoa pode er uma emeaça para a segurança

Tarefa - Estimação

É similar à classificação, porém é usada quando o registro é identificado por um valor numérico e não um categórico

- Estimar a quantia a ser gasta por uma família de quatro pessoas durante a volta às aulas
- Estimar a pressão ideal de um paciente baseando-se na idade, sexo e massa corporal

Tarefa - Predição

A tarefa é similar às tarefas de classificação e estimação, porém ela visa descobrir o valor futuro de um determinado atributo.

- Predizer o valor de uma ação três meses adiante
- Predizer o percentual que será aumentado de tráfego na rede se a velocidade aumentar
- Predizer o vencedor do campeonato baseando-se na comparação das estatisticas dos times

Tarefa - Agrupamento

A tarefa visa identificar e aproximar os registros similares.

- Segmentação de mercado para um nicho de produtos
- Auditoria, separando comportamentos suspeitos

Tarefa - Associação

A tarefa consiste em identificar quais atributos estão relacionados. SE atributo X ENTÃO atributo Y

- Identificar os usuários de planos que respondem bem a oferta de novos serviços
- Determinar os casos onde um novo medicamento pode apresentar efeitos colaterais

Ferramentas

Tabela Periódica de Ferramentas

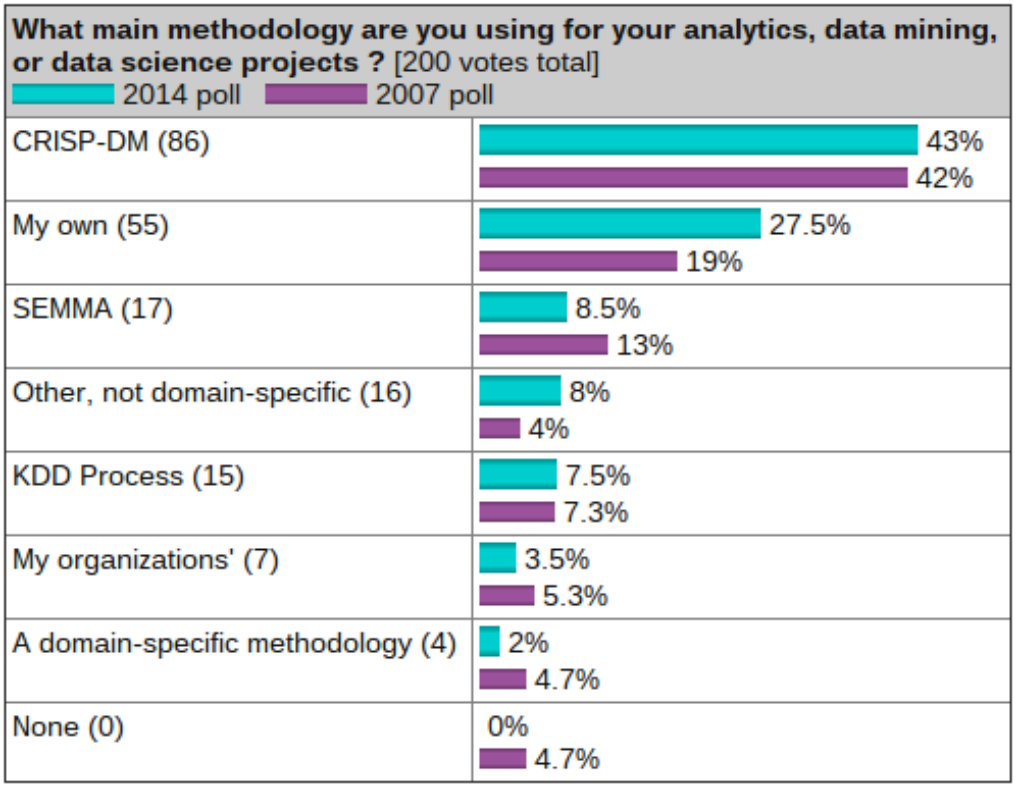
+ Ferramentas

A grande questão...

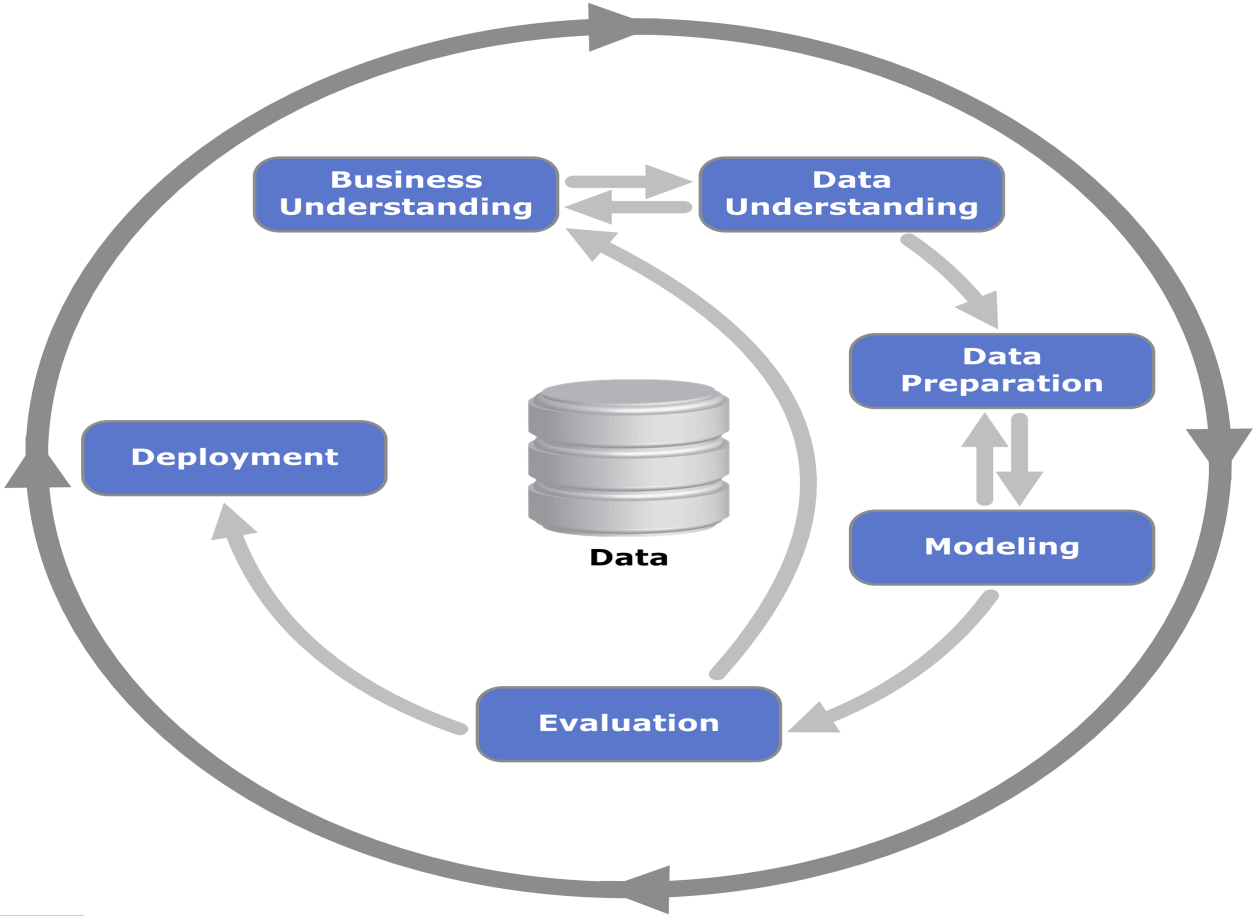
Qual é a melhor ferramenta?!

Metodologias

Metodologias



CRIPS-DM

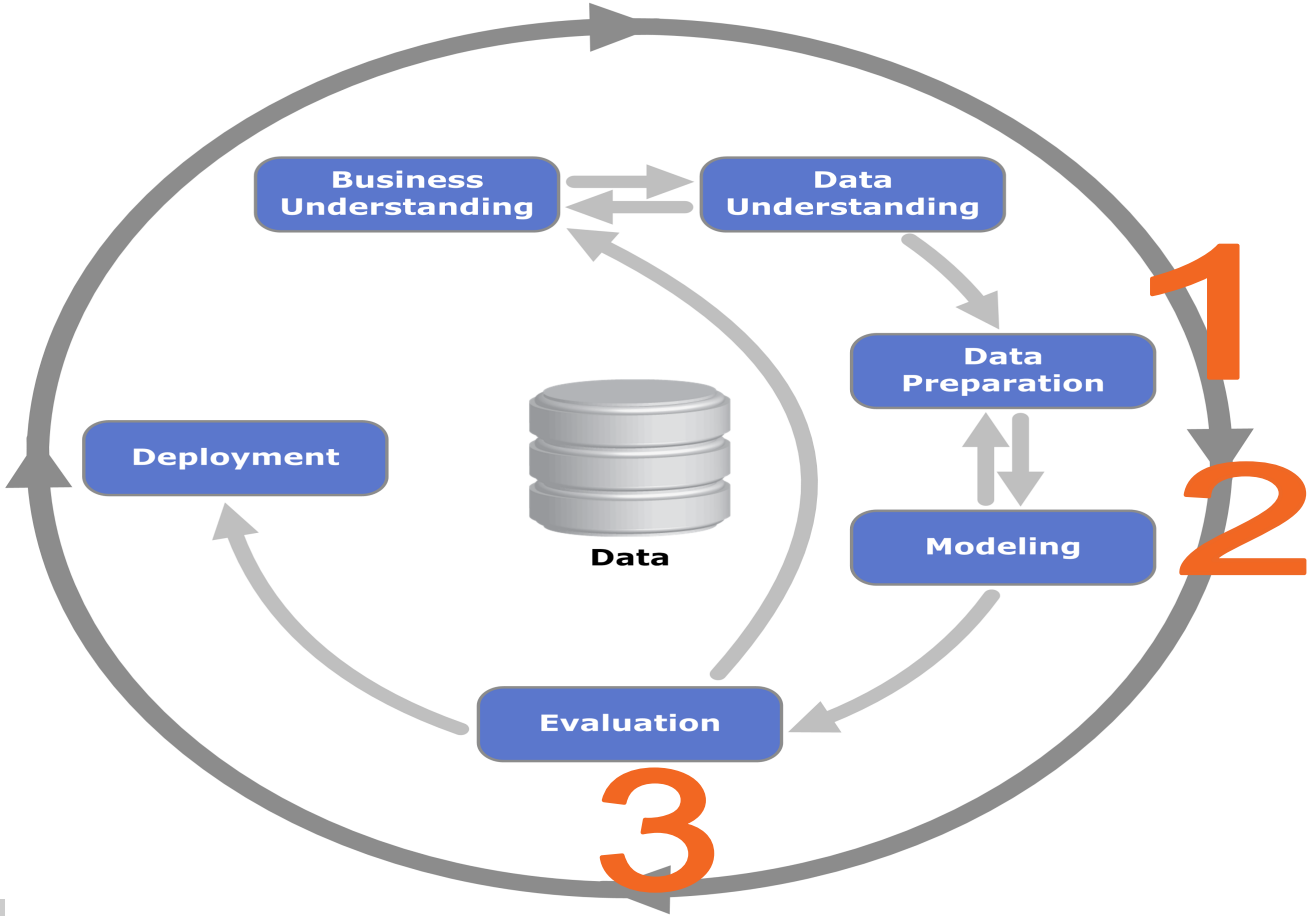

Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Data

A Solução

Uma Ferramenta para gestão do Fluxo de Dados

Pentaho Data Integration

A Integração



Data Preparation

Visa a preparação dos dados, que geralmente não estão dispostos em formato adequado, para a aplicação dos algoritmos de descoberta, análise e extração do conhecimento.

1. As grandes bases de dados são altamente susceptíveis a ruídos, valores faltantes e inconsistência
2. Dados limpos e consistente são requisitos básicos para sucesso da mineração de dados.
3. Esse processo tem como objetivo assegurar a qualidade dos dados.

Data Preparation

- Conhecimento sobre o domínio auxilia em todas as etapas do processo de KDD. **D3M** (Domain Driven Data Mining)

1. Seleção de Dados
2. Limpeza
3. Transformação
4. Integração de dados
5. Formatação sintatica

Data Preparation - Seleção de Dados

- Simplicidade do modelo gerado
- Relevância dos atributos
- Redundância entre atributos
- Aumento da Acurácia
- Principais formas:
  - Segmentação dos dados
  - Eliminação Direta
  - Amostragem aleatória
  - Agregação

Data Preparation - Limpeza

- Remoção de Ruidos
- Atributos incompletos ou sem informação
- Principais formas:
  - Exclusão de casos
  - Preenchimento de valores
  - Preenchimento com valores globais constantes
  - Preenchimento com medidas estatísticas
  - Preenchimento com métodos de mineração de dados

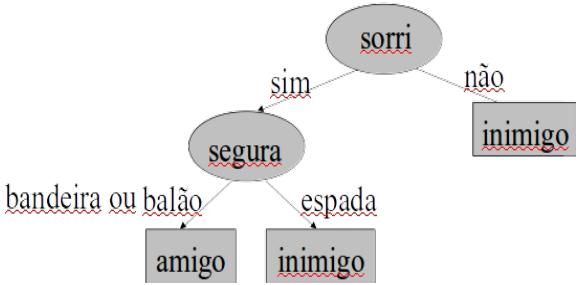
Data Preparation - Transformação de dados

- Generalização
- Normalização
- Transformação Númerico para categórico
- Transformação Categórico para Númerico

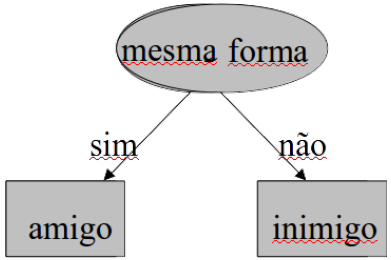
Data Preparation - Construção de dados

- Criação de novos atributos

Cabeça	Corpo	Sorri	Segura	Classe
Triangular	Triangular	Sim	Balão	Amigo
Quadrada	Quadrado	Sim	Balão	Amigo
Redonda	Redondo	Sim	Bandeira	Amigo
Quadrada	Triangular	Não	Espada	Inimigo
Triangular	Redondo	Sim	Espada	Inimigo
Redonda	Quadrado	Não	Bandeira	Inimigo



Cabeça	Corpo	Sorri	Segura	Mesma forma	Classe
Triangular	Triangular	Sim	Balão	Sim	Amigo
Quadrada	Quadrado	Sim	Balão	Sim	Amigo
Redonda	Redondo	Sim	Bandeira	Sim	Amigo
Quadrada	Triangular	Não	Espada	Não	Inimigo
Triangular	Redondo	Não	Espada	Não	Inimigo
Redonda	Quadrado	Não	Bandeira	Não	Inimigo



Modeling

- R Script Executor
- Weka Steps
- Ferramentas de Mineração de Dados!
  - Plugins
  - Chamadas Shell
  - Integração no nível de saída.

Evaluation

- Pós-processamento
- Criação de Data Mart com os Dados (Continuação do ETL pós modelagem)
- Desempenho avaliado diretamente no modelo OLAP, levando ao OLAM (OLAP + DataMining)
- Kolmogorov-Smirnov Curve (KS Test)
- ROC Curve

A Proposta


SQL
SAS
tools
Tableau
Unix
MATLAB
Python
paid
KNIME
including
IBM
Modeler
Data
Excel
R
or
Other
RapidMiner
Analytics
free
Hadoop
analytics/data
Enterprise
Miner
SAP
Revolution
Weka
Statistics
Microsoft
Server
SPSS
BusinessObjects/Sybase/Hana

oncase
power your business

Obrigado!

Mailson Filho
@maialson