



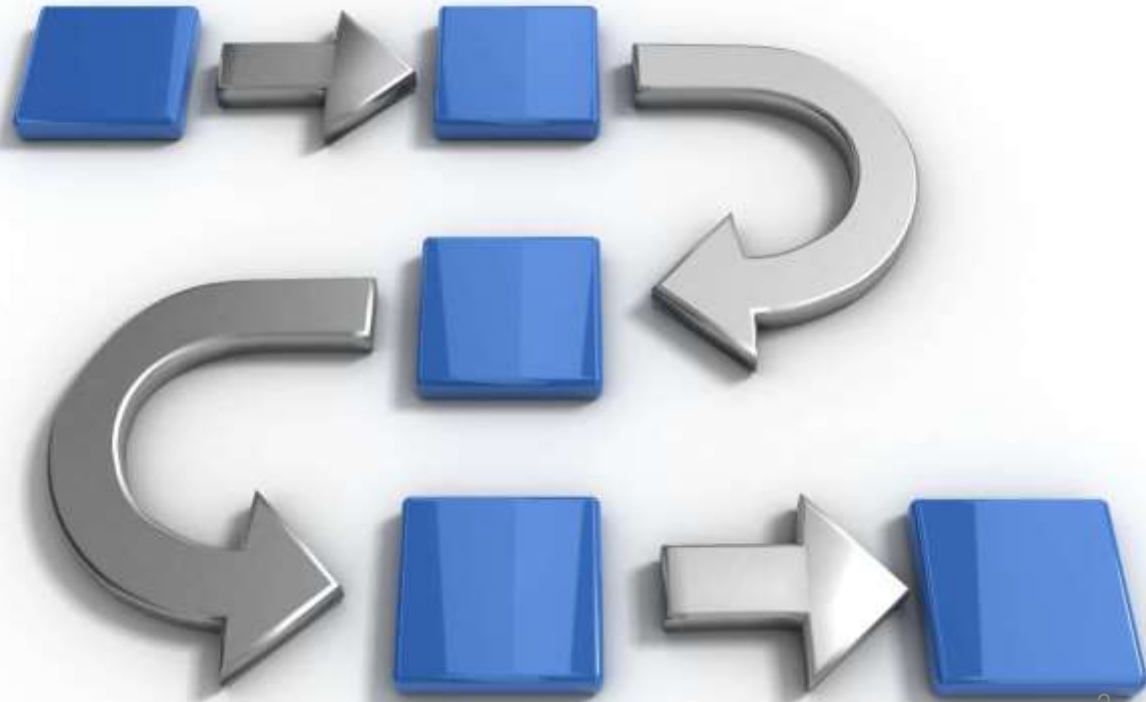
Universidade do Minho
Departamento de Informática

Metodologias de Análise de Dados

LEI/MiEI @ 2022/2023, 2º sem
[ADI³]

O que são Metodologias para Análise de Dados? (Extração de Conhecimento)

- Uma **Metodologia para Análise de Dados** descreve e cria **um conjunto de passos** pelos quais deverá passar o desenvolvimento de um **Projeto de Aprendizagem Automática (*Machine Learning*)** para a resolução de problemas.



■ Enquadrar um processo de Análise de Dados ao abrigo de uma metodologia:

- Garante maior robustez;
- Facilita a sua compreensão, implementação e desenvolvimento;
- Permite a replicação de processos;
- Auxilia no planeamento e na gestão do projeto;
- Confere “maturidade” ao processo;
- Encoraja a adoção de melhores práticas.



Que metodologias?

- **CRISP-DM**

- **C**Ross **I**ndustry **S**tandard **P**rocess for **D**ata **M**ining
(Daimler Chrysler, SPSS, NCR)

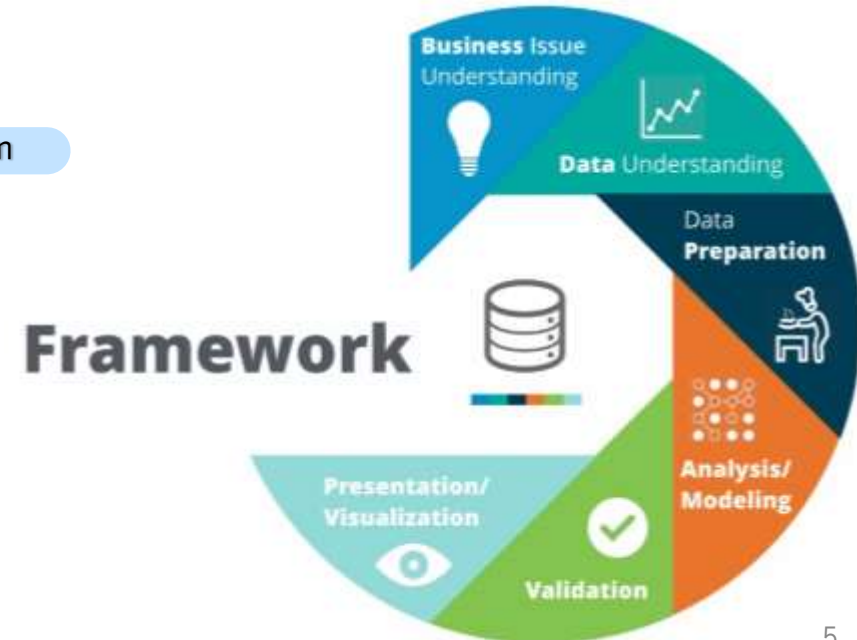
- **SEMMA**

- **S**ample, **E**xplore, **M**odify, **M**odel and **A**ssess
(SAS Institute Inc.)

- **PMML**

- **P**redictive **M**odel **M**arkup **L**anguage
(Angoss Software, Magnify, Univ. Illinois, NCR, SPSS)

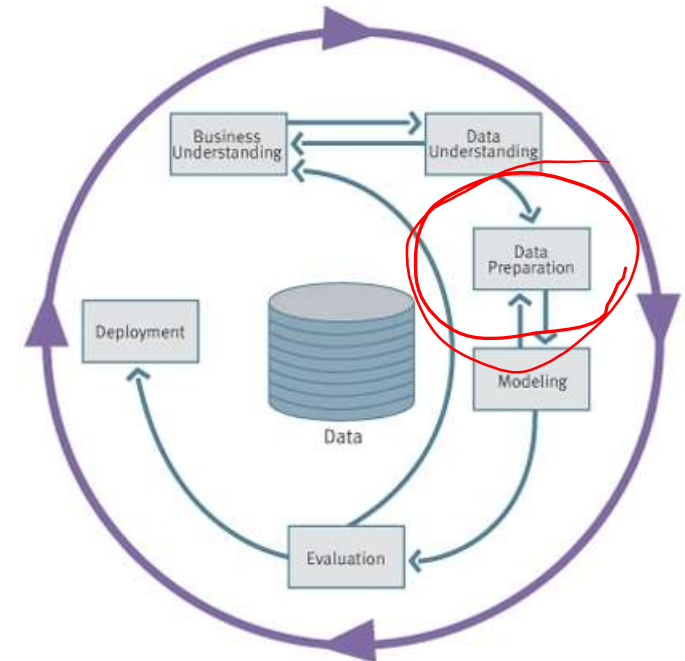
- **CR**oss **I**ndustry **S**tandard **P**rocess for **D**ata **M**ining
(Daimler Chrysler, SPSS, NCR)
- **Objetivos:**
 - Definir um processo de Análise de Dados para a indústria;
 - Construir e disponibilizar ferramentas de apoio;
 - Assegurar a qualidade dos projetos de Análise de Dados;
 - Reduzir os conhecimentos específicos necessários para conduzir um processo de Análise de Dados.



CRISP-DM Ciclo de vida

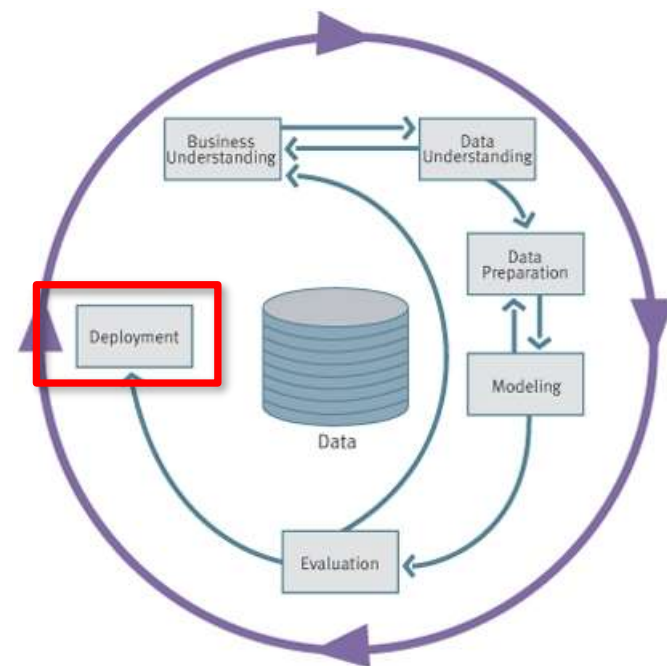
- O **CRISP-DM** é um modelo de processos com vista a definir um “guião” para o desenvolvimento de projetos de AD, que se desenrola em 6 etapas:

- Estudo do negócio;
- Estudo dos dados;
- Preparação dos dados;
- Modelação;
- Avaliação;
- Desenvolvimento.



CRISP-DM Ciclo de vida

- └ *Business Understanding/ Estudo do Negócio:*
 - Compreensão dos objetivos do projeto e definição do problema de AD;
- └ *Data Understanding/ Estudos dos Dados:*
 - Obter os dados e identificar a qualidade dos dados;
- └ *Data Preparation/ Preparação dos Dados:*
 - Seleção de atributos e limpeza dos dados;
- └ *Modeling/ Modelação:*
 - Experimentação com as ferramentas de AD;
- └ *Evaluation/ Avaliação:*
 - Comparação dos resultados com os objetivos do negócio;
- *Deployment/ Desenvolvimento:*
 - Colocação do modelo em produção.



CRISP-DM

Fases e Tarefas



- **Sample, Explore, Modify, Model and Assess;**
- Produto de *Data Mining* desenvolvido pelo SAS Institute Inc.;
- Definição SAS:
 - “*Data Mining* é o processo de **extrair conhecimento e relações complexas** de grandes volumes de dados.”
- **Motivação:**
 - necessidade de definir, padronizar e integrar sistemas ou processos de *Data Mining* nos ciclos de produção.
- Desenvolvimento focado na ferramenta [SAS Enterprise Miner](#).



■ Divide o processo de *Data Mining* em 5 etapas:

○ *Sample*/Amostragem:

- Extração de dados do universo do problema;
- Baseia o processo de *Data Mining* no conceito de “amostra” do problema;
- Amostra pequena e significativa;
- Proporciona flexibilidade e rapidez no tratamento dos dados.

○ *Explore*/Exploração;

○ *Modify*/Modificação;

○ *Model*/Modelação;

○ *Assess*/Avaliação.



O processo SEMMA

■ Divide o processo de *Data Mining* em 5 etapas:

- *Sample*/Amostragem;
- *Explore*/Exploração:
 - Exploração visual e/ou numérica das tendências;
 - Refinamento do processo de descoberta (*mining*);
 - Técnicas estatísticas: regressão linear, mínimos quadrados, distribuição de Poisson, etc.;
 - Procura de tendências imprevistas nos dados;
- *Modify*/Modificação;
- *Model*/Modelação;
- *Assess*/Avaliação.



O processo SEMMA

- Divide o processo de *Data Mining* em 5 etapas:
 - *Sample*/Amostragem;
 - *Explore*/Exploração;
 - *Modify*/Modificação:
 - Concentração de todas as modificações necessárias;
 - Inclusão de informação;
 - Seleção ou introdução de novas variáveis;
 - Objetivo: criar, selecionar e adaptar variáveis para a próxima etapa;
 - *Model*/Modelação;
 - *Assess*/Avaliação.



O processo SEMMA

▪ Divide o processo de *Data Mining* em 5 etapas:

- *Sample*/Amostragem;
- *Explore*/Exploração;
- *Modify*/Modificação;
- *Model*/Modelação:
 - Definição das técnicas de construção de modelos de *Data Mining*: redes neurais artificiais, árvores de decisão, regressão linear, etc.;
 - Dependente do tipo de dados presentes em cada modelo (p.ex., RNA são mais adequadas quando os dados do problema apresentam relacionamentos complexos);
- *Assess*/Avaliação.

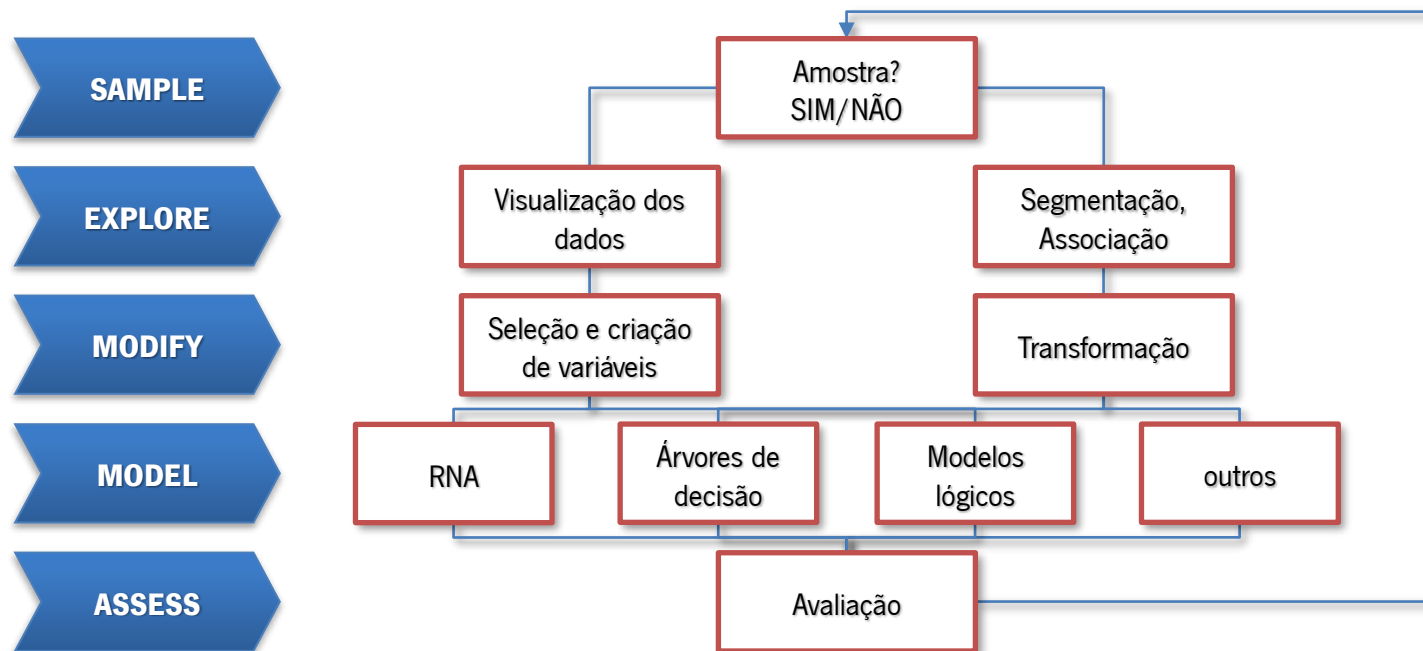


O processo SEMMA

- Divide o processo de *Data Mining* em 5 etapas:
 - *Sample*/Amostragem;
 - *Explore*/Exploração;
 - *Modify*/Modificação;
 - *Model*/Modelação;
 - *Assess*/Avaliação:
 - Aferição do desempenho do modelo construído para *Data Mining*;
 - Aplicação do modelo a uma amostra de dados de teste;
 - Procedimento de ajuste do modelo.



O processo SEMMA



in "Data Mining – Descoberta de Conhecimento em Bases de Dados"
Manuel Filipe Santos, Carla Azevedo

CRISP-DM versus SEMMA

■ Fases CRISP-DM:

- Estudo do negócio;
- Estudo dos dados;

- Preparação dos dados;
- Modelação;
- Avaliação;

- Desenvolvimento.

■ Processo SEMMA:

- Amostragem;
- Exploração;
- Modificação;
- Modelação;
- Avaliação.

- **Predictive Model Markup Language;**
- Desenvolvido por investigadores de *Data Mining* e várias empresas (NCR, SPSS, etc.);
- A especificação PMML encontra-se em fase de desenvolvimento e consolidação (versão 4.2.1);
- Utilizada por diversas aplicações (IBM DB2 Data Warehouse Edition v.10.5, SAS Enterprise Miner v.5.1, v.5.3, v.7.1, v.13.1, SPSS Statistics v.21);
(<http://www.dmg.org/products.html>)
- Expandir para transformá-la num padrão para o WWW;
- O PMML é uma linguagem para descrever modelos de DM;
- Utiliza XML para representar modelos de DM.



```

sepal_length,sepal_width,petal_length,petal_width,class
5.1,3.5,1.4,0.2,Iris-setosa
4.9,3.0,1.4,0.2,Iris-setosa
4.7,3.2,1.3,0.2,Iris-setosa
4.6,3.1,1.5,0.2,Iris-setosa
...
5.0,2.0,3.5,1.0,Iris-versicolor
5.9,3.0,4.2,1.5,Iris-versicolor
6.0,2.2,4.0,1.0,Iris-versicolor
6.1,2.9,4.7,1.4,Iris-versicolor
5.6,2.9,3.6,1.3,Iris-versicolor
6.7,3.1,4.4,1.4,Iris-versicolor
5.6,3.0,4.5,1.5,Iris-versicolor
5.8,2.7,4.1,1.0,Iris-versicolor
6.3,2.5,4.9,1.5,Iris-versicolor
6.1,2.8,4.7,1.2,Iris-versicolor
...
6.7,2.5,5.8,1.8,Iris-virginica
7.2,3.6,6.1,2.5,Iris-virginica
6.5,3.2,5.1,2.0,Iris-virginica
6.4,2.7,5.3,1.9,Iris-virginica
6.8,3.0,5.5,2.1,Iris-virginica
5.7,2.5,5.0,2.0,Iris-virginica
5.8,2.8,5.1,2.4,Iris-virginica
6.4,3.2,5.3,2.3,Iris-virginica
...

```

```

<PMML version="2.0">
-
<Header copyright="Copyright (c) 2001, Oracle Corporation. All rights reserved.">
<Application name="Oracle 9i Data Mining" version="9.2.0"/>
</Header>
-
<DataDictionary numberOfFields="1">
<DataField name="item" optype="categorical"/>
</DataDictionary>
-
<TransformationDictionary>
-
<DerivedField name="PETAL_LENGTH">
-
<Discretize field="PETAL_LENGTH">
-
<DiscretizeBin binValue="1-1.59">
<Interval closure="closedOpen" leftMargin="1.0" rightMargin="1.59"/>
</DiscretizeBin>
-
<DiscretizeBin binValue="1.59-2.18">
<Interval closure="closedOpen" leftMargin="1.59" rightMargin="2.18"/>
</DiscretizeBin>
-
<DiscretizeBin binValue="2.18-2.77">
<Interval closure="closedOpen" leftMargin="2.18" rightMargin="2.77"/>
</DiscretizeBin>

```

PMML: objetivos

- Permitir que aplicações utilizem diversas fontes de dados sem se preocuparem com as diferenças entre elas;
- Permitir a utilização combinada e/ou cooperativa de modelos de *Data Mining*;
- Permitir a administração de modelos de *Data Mining* baseados em áreas de negócio.