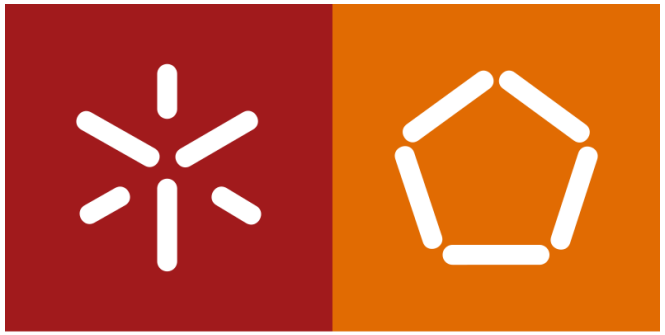


Universidade do Minho
Escola de Engenharia

S istemas de **A** prendizagem e **E** xtração de **C** onhecimento

José Machado

Diana Ferreira

PROGRAMA PRÁTICO

1 METODOLOGIA CRISP-DM

2 WEKA

3 PROCESSO DE DATA MINING

4 TRABALHOS PRÁTICOS

DATA MINING

DATA MINING (DESCOBERTA DE CONHECIMENTO A PARTIR DE DADOS):

Extracção de padrões ou conhecimentos de interesse (não triviais, implícitos, anteriormente desconhecidos e potencialmente úteis) de uma enorme quantidade de dados.

CARACTERÍSTICAS CHAVE:

- Combinação de Teoria e Aplicação;
- Processo de Engenharia:
  CRISP DM;
- Colecção de Funcionalidades:
  Diferentes Tarefas e Algoritmos;
- Área Interdisciplinar.

DATA MINING

APLICAÇÕES DO DATA MINING

- *Gestão do relacionamento com o cliente:* desenvolver a lealdade, implementar estratégias focadas no cliente;

- *Análise de dados financeiros:* encontrar padrões, causalidades e correlações em informações comerciais e preços de mercado;

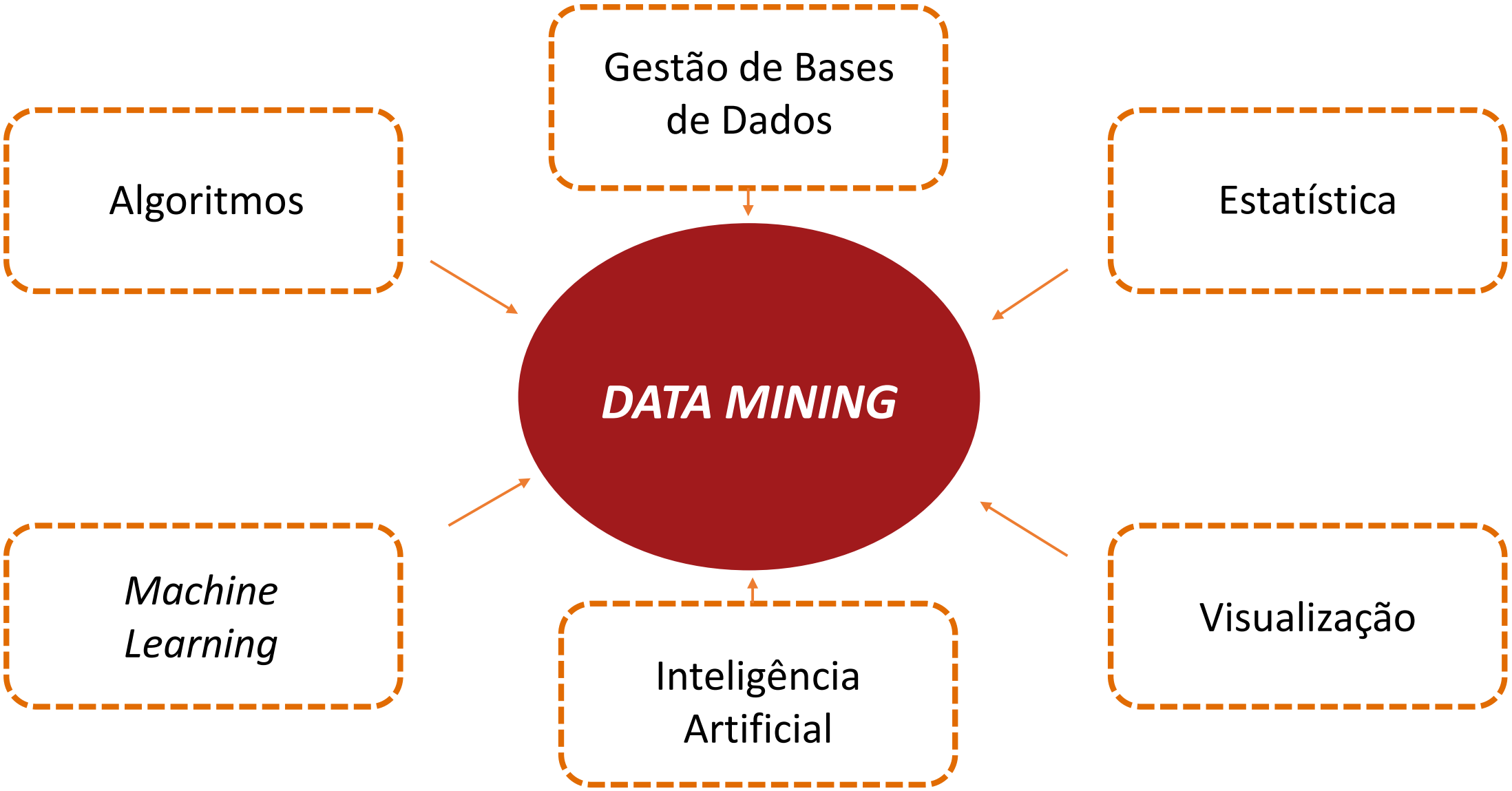
- *Análise de cestos de supermercado:* compreender as necessidades do comprador e alterar o layout da loja em conformidade;

INSTITUIÇÕES DE SAÚDE

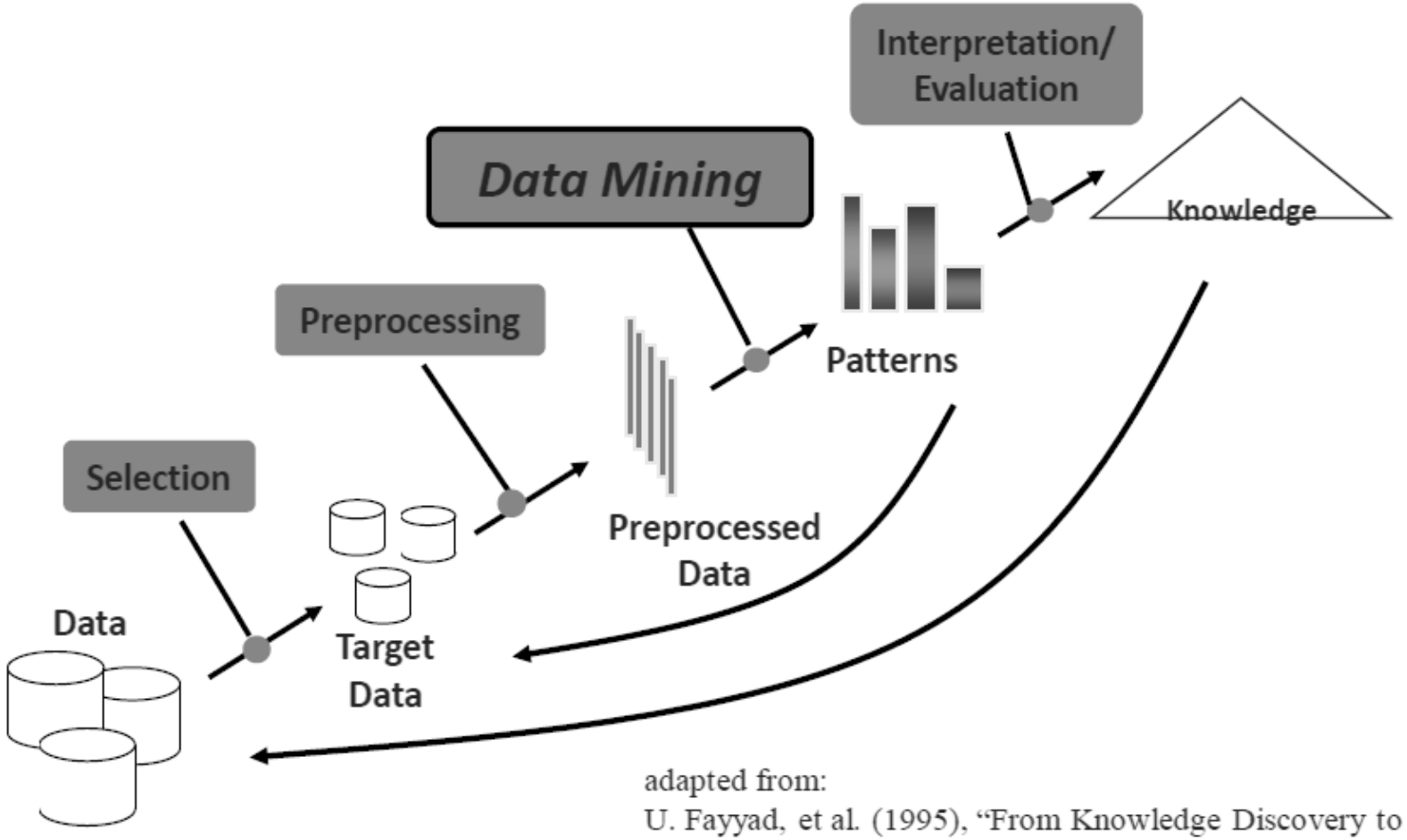
- Melhorar os cuidados de saúde;
- Reduzir custos;
- Prever váriáveis de interesse, por exemplo:

Prever o número de pacientes nas Urgências e o seu tempo de permanência e/ou tempo de espera.

DATA MINING


Gestão de Bases de Dados
Algoritmos
Estatística
DATA MINING
Machine Learning
Inteligência Artificial
Visualização

DATA MINING



adapted from:
U. Fayyad, et al. (1995), "From Knowledge Discovery to Data Mining: An Overview," Advances in Knowledge Discovery and Data Mining, U. Fayyad et al. (Eds.), AAAI/MIT Press

DATA MINING

- Cluster

- Classify

Categorical, Regression

- Summarize

Summary statistics, Summary rules

- Link Analysis / Model Dependencies

Association rules

- Sequence analysis

Time-series analysis, Sequential associations

- Detect Deviations

DATA MINING

- Concept description: Characterization and discrimination

– Generalizar, resumir e contrastar as características dos dados, por exemplo, regiões secas vs. regiões húmidas

- Association (correlation and causality)

– Fralda → Cerveja [0.5%, 75%]

- Classification and Prediction

– Construir modelos (funções) que descrevem e distinguem classes ou conceitos para previsão futura

Exemplo: classificar os países com base no clima

– Apresentação: decision-tree, classification rule, neural network

– Prever alguns valores numéricos desconhecidos ou em falta

DATA MINING

- Cluster analysis

– A etiqueta da classe é desconhecida: Agrupar dados para formar novas classes, por exemplo, cluster casas para encontrar padrões de distribuição
– Maximização da semelhança intra-classe e minimização da semelhança interclasse

- Outlier analysis

- Outlier: um objeto de dados que não está de acordo com o comportamento geral dos dados
- Ruído ou excepção? Não! útil na detecção de fraudes, análise de eventos raros

- Trend and evolution analysis

- Tendência e desvio: análise de regressão
- Mineração de padrões sequenciais, análise de periodicidade
- Análise baseada na similaridade

- Other pattern directed or statistical analyses

CRISP-DM

CRoss**I**ndustry **S**tandard **P**rocess for **D**ata **M**ining

Esforço financiado pela Comunidade Europeia para desenvolver um *framework* para o processo de *Data Mining*

OBJECTIVO:
- Encorajar ferramentas interoperáveis ao longo de todo o processo de *Data Mining;*
- Retirar conhecimento valioso de tarefas simples de *Data Mining.*

O processo de *Data Mining* deve ser confiável e repetível por pessoas com pouco conhecimento em DM!!

CRISP-DM

CARACTERÍSTICAS:

- *Framework* para o registo de experiências;

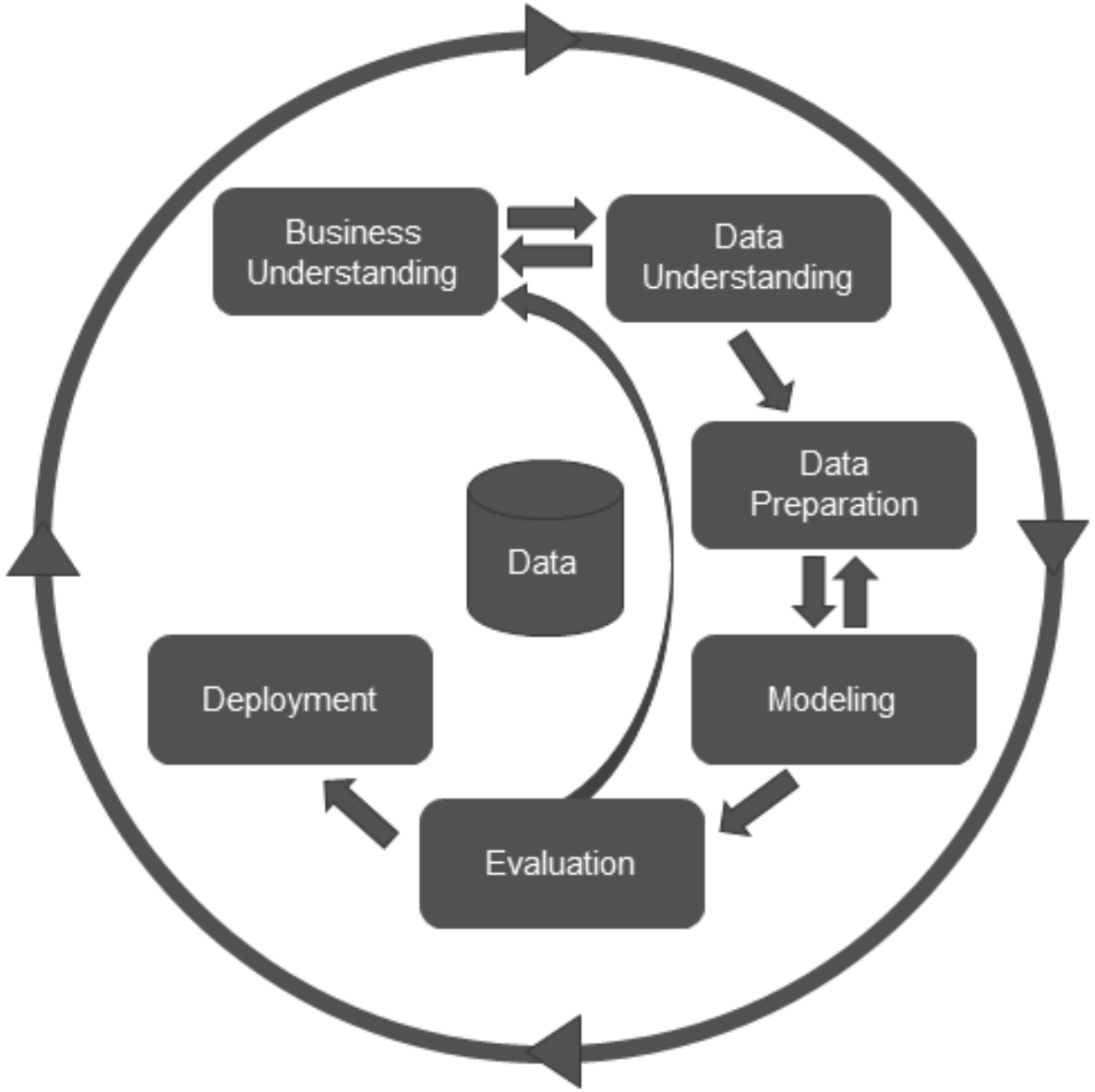
- Permite que os projetos sejam replicados;

- Ajuda no planeamento e na gestão de projectos;

- "Factor de conforto" para novos adoptantes;

- Demonstra a maturidade da Data Mining.

CRISP-DM


Business Understanding
Data Understanding
Data Preparation
Data
Modeling
Deployment
Evaluation

CRISP-DM

•Business Understanding
–Perceber os objectivos e requisitos do projeto
–Determinar o objectivo de *Data Mining*

•Data Understanding
– Recolha, exploração e familiarização com os dados
– Identificar problemas de qualidade nos dados

•Data Preparation
– Seleção de dados (critérios de inclusão/exclusão)
– Seleção e Criação de atributos
– Limpeza de dados

•Modeling
– Escolher os modelos de Data Mining
– Construção e avaliação dos modelos

•Evaluation
– Avaliar os resultados, i.e, determinar se os resultados cumprem os objetivos iniciais
– Rever o processo

•Deployment
– Colocar os modelos finais em prática
– Monitorização e manutenção dos modelos

CRISP-DM Stage 1 – Business Understanding

- Statement of Business Objective
  - States goal in business terminology
- Statement of Data Mining objective
  - States objectives in technical terms
- Statement of Success Criteria

Focuses on understanding the project objectives and requirements from a business perspective, then converting this knowledge into a data mining problem definition and a preliminary plan designed to achieve the objectives

What the client really wants to accomplish?

Uncover important factors (constraints, competing objectives)

CRISP-DM Stage 1 – Business Understanding

Determine business objectives

- Key persons and their roles? Is there a steering committee. Internal sponsor (financial, domain expert).
- Business units impacted by the project (sales, finance,...)? Business success criteria and who assesses it?
- Users' needs and expectations.
- Describe problem in general terms. Business questions, Expected benefits.

Assess situation

- Are they already using data mining.
- Identify hardware and software available. Identify data sources and their types (online, experts, written documentation).
- Identify knowledge sources and types (online, experts, written documentation)
- Describe the relevant background.

CRISP-DM Stage 1 – Business Understanding

Determine data mining goals

- Translate the business questions to data mining goals

  (e.g., a marketing campaign requires segmentation of customers in order to decide whom to approach in this campaign; the level/size of the segments should be specified).

- Specify data mining problem type

  (e.g., classification, description, prediction and clustering).

- Specify criteria for model assessment.

Produce project plan

- Define initial process plan; discuss its feasibility with involved personnel.
- Put identified goals and selected techniques into a coherent procedure.
- Estimate effort and resources needed; Identify critical steps.

CRISP-DM Stage 2 – Data Understanding

- Acquire the data
- Explore the data (query & visualization)
- Verify the quality

Starts with an initial data collection and proceeds with activities in order to get familiar with the data, to identify data quality problems, to discover first insights into the data or to detect interesting subsets to form hypotheses for hidden information.

CRISP-DM Stage 2 – Data Understanding

Collect data

- List the datasets acquired (locations, methods used to acquire, problems encountered and solutions achieved).

Describe data

- Check data volume and examine its gross properties.•
- Accessibility and availability of attributes. Attribute types, range, correlations, the identities.
- Understand the meaning of each attribute and attribute value in business terms.
- For each attribute, compute basic statistics (e.g., distribution, average, max, min, standard deviation, variance, mode, skewness).

CRISP-DM Stage 2 – Data Understanding

Explore data

Analyze properties of interesting attributes in detail

Distribution, relations between pairs or small numbers of attributes, properties of significant sub-populations, simple statistical analyses

Verify data quality

Identify special values and catalogue their meaning.

Does it cover all the cases required? Does it contain errors and how common are they?

Identify missing attributes and blank fields. Meaning of missing data.

Do the meanings of attributes and contained values fit together?

Check spelling of values (e.g., same value but sometime beginning with a lower case letter, sometimes with an upper case letter).

Check for plausibility of values, e.g. all fields have the same or nearly the

same values.

CRISP-DM Stage 3 – Data Preparation

Construct data

Derived attributes.

Background knowledge .

How can missing attributes be constructed or imputed?

Integrate data

Integrate sources and store result (new tables and records).

Format Data

Rearranging attributes (Some tools have requirements on the order of the attributes, e.g. first field being a unique identifier for each record or last field being the outcome field the model is to predict).

Reordering records (Perhaps the modeling tool requires that the records be sorted according to the value of the outcome attribute).

Reformatted within-value (These are purely syntactic changes made to satisfy the requirements of the specific modeling tool, remove illegal characters, uppercase lowercase).

CRISP-DM Stage 4 – Modeling

- Select the modeling technique
  - Based upon the data mining objective
- Generate test design
  - Procedure to test model quality and validity
- Build model
  - Parameter settings
- Assess model (rank the models)

Various modeling techniques are selected and applied and their parameters are calibrated to optimal values. Typically, there are several techniques for the same data mining problem type. Some techniques have specific requirements on the form of data. Therefore, stepping back to the data preparation phase is often necessary

CRISP-DM Stage 4 – Modeling

Select modeling technique

- Select technique
- Identify any built-in assumptions made by the technique about the data (e.g. quality, format, distribution).
- Compare these assumptions with those in the Data Description Report and make sure that these assumptions hold.
- Preparation Phase if necessary.

Generate test design

- Describe the intended plan for train, test and evaluate the models.
- How to divide the dataset into training, test and validation sets.
- Decide on necessary steps (number of iterations, number of folds etc.).
- Prepare data required for test

CRISP-DM Stage 4 – Modeling

Build model

- Set initial parameters and document reasons for choosing those values.
- Run the selected technique on the input dataset. Post-process data mining results (eg. editing rules, display trees).
- Record parameter settings used to produce the model.
- Describe the model, its special features, behavior and interpretation.

Assess model

- Evaluate result with respect to evaluation criteria. Rank results with respect to success and evaluation criteria and select best models.
- Interpret results in business terms. Get comments by domain experts.
- Check plausibility of model.
- Check model against given knowledge base (discovered info. novel and useful?)
- Check result reliability. Analyze potentials for deployment of each result.

CRISP-DM Stage 5 – Evaluation

- More thoroughly evaluate model
- Decide how to use results

- Methods and criteria depend on model type:

e.g., coincidence matrix with classification models, mean error rate with regression models

Interpretation of model: important or not, easy or hard depends on algorithm

Determine if there is some important business issue that has not been sufficiently considered.

A decision on the use of the data mining results should be reached

CRISP-DM Stage 5 – Evaluation

- More thoroughly evaluate model
- Decide how to use results

- Methods and criteria depend on model type:

e.g., coincidence matrix with classification models, mean error rate with regression models

Interpretation of model: important or not, easy or hard depends on algorithm

Determine if there is some important business issue that has not been sufficiently considered.

A decision on the use of the data mining results should be reached

CRISP-DM Stage 5 – Evaluation

Evaluate results

- Understand data mining result. Check impact for data mining goal.
- Check result against knowledge base to see if it is novel and useful.
- Evaluate and assess result with respect to business success criteria
- Rank results according to business success criteria. Check result impact on initial application goal.
- Are there new business objectives? (address later in project or new project?)
- State conclusions for future data mining projects.

Review of process

- Summarize the process review (activities that missed or should be repeated).
- Overview data mining process. Is there any overlooked factor or task?
- (did we correctly build the model? Did we only use attributes that we are allowed to use and that are available for future analyses?)
- Identify failures, misleading steps, possible alternative actions, unexpected paths
- Review data mining results with respect to business success

CRISP-DM Stage 5 – Evaluation

Determine next steps

- Analyze potential for deployment of each result. Estimate potential for improvement of current process.
- Check remaining resources to determine if they allow additional process iterations (or whether additional resources can be made available).
- Recommend alternative continuations. Refine process plan.

Decision

- According to the results and process review, it is decided how to proceed to the next stage (remaining resources and budget)
- Rank the possible actions. Select one of the possible actions.
- Document reasons for the choice.

CRISP-DM

PORQUÊ?

- O processo de *Data Mining* deve ser confiável e repetível por pessoas com pouco conhecimento em DM

- CRISP-DM fornece um *framework* uniforme para

- diretrizes
- documentação de experiência

- CRISP- DM é flexível o suficiente para ter em conta:

- Problemas de negócio diferentes
- Dados diferentes

EXERCÍCIO – FE01

[1] Identifique um problema que possa ser enquadrado dentro do processo de *Data Mining*.
Para esse problema descreva sucintamente as seguintes fases do processo CRISP-DM:

[a] Business Understanding;
[b] Data Understanding;

[2] Que tipo de benefícios espera retirar da aplicação de *Data Mining*.