

Metodologias de Análise de Dados

Uma **Metodologia de Análise de Dados** descreve e cria um **conjunto de passos** pelos quais deverá passar o desenvolvimento de um projeto de *machine learning* para resolução de problemas.

Ao enquadrar um processo de análise de dados ao abrigo de uma metodologia: - garante maior robustez; - facilita a sua compreensão; - permite a replicação de processos; - auxilia no planeamento e na gestão do projeto; - confere maturidade ao processo; - encoraja a adoção de melhores práticas;

CRISP-DM

Cross-Industry Standard Process for Data Mining (CRISP-DM) é uma metodologia de análise de dados que descreve as fases de um projeto de *data mining*.

Os seus objetivos são: - 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 processos de análise de dados;

Este método é um modelo composto por seis fases: 1. Estudo do Negócio; 2. Estudo dos Dados; 3. Preparação dos Dados; 4. Modelação; 5. Avaliação; 6. Implementação;

SEMMA

Sample, Explore, Modify, Model, Assess (SEMMA) é uma metodologia de análise de dados que descreve as fases de um projeto de *data mining*.

Esta metodologia é composta por cinco fases: 1. **Sample**: - Exploração visual e/ou numérica dos dados; - Refinamento do processo de descoberta (data mining); - Técnicas estatísticas para a seleção de amostras; - Procura de padrões e relações nos dados; 2. **Explore**: - Análise exploratória dos dados; - Refinamento do processo de descoberta (data mining); - Técnicas estatísticas para a seleção de amostras; - Procura de padrões e relações nos dados; 3. **Modify**: - 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; 4. **Model**: - Definição de 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 momento; 5. **Assess:** - Avaliação dos modelos criados para *Data Mining*; - Aplicação do modelo e uma amostra de dados de teste; - Procedimento de ajuste do modelo;