



UNIVERSIDADE FEDERAL DO PIAUÍ – UFPI

CENTRO DE TECNOLOGIA - CT

Programa de Pós Graduação em Engenharia Elétrica

DESCOBERTA DE CONHECIMENTO EM BASE DE DADOS

Flávio Henrique Duarte de Araújo
(flavio86@ufpi.edu.br)

MOTIVAÇÃO

- A maioria das empresas possui **armazenado** em seus Banco de Dados um **volume gigantesco** de dados contendo o **histórico de sua atividade**:
 - Transações eletrônicas;
 - Novos equipamentos científicos e industriais para observação e controle;
 - Dispositivos de armazenamento em massa;
- Os recursos de análise de dados tradicionais são inviáveis para acompanhar esta evolução:
 - *“Morrendo de sede por conhecimento em um oceano de dados”.*

MOTIVAÇÃO

- Solução:
 - Ferramentas de automatização das tarefas repetitivas e sistemática de análise de dados;
 - Ferramentas de auxílio para as tarefas cognitivas da análise;
 - Integração das ferramentas em sistemas apoiando o processo completo de descoberta de conhecimento para tomada de decisão.
- **Data Science:** habilidade para manipular, analisar e extrair valor dos dados.

EXEMPLOS DE SUCESSO

- Lojas brasileiras conseguiram reduzir os produtos oferecidos em suas prateleiras de 51000 para 14000 por meio da análise dos dados referentes as suas transações e produtos:
- Exemplos de anomalias identificadas:
 - Roupas de inverno encalhadas no nordeste;
 - Guardas chuvas encalhadas no período seco;
 - Batedeiras 110v a venda em SC onde a corrente é 220v.

EXEMPLOS DE SUCESSO

- O *Bank of America* utilizou os dados referentes a 36 milhões de clientes e descobriu que os casais que tinham filhos entre 18 e 21 anos possuíam o menor risco de dar calotes:
 - Resultado: em três anos o banco estima que lucrou cerca de 30 milhões de dólares com carteira de empréstimos.

EXEMPLOS DE SUCESSO

- Fraldas e cervejas:
 - O que as cervejas tem a ver com as fraldas ?
 - Homens casados, entre 25 e 30 anos;
 - compravam fraldas e/ou cervejas às sextas-feiras à tarde no caminho do trabalho para casa;
 - Wal-Mart otimizou às gôndolas nos pontos de vendas, colocando as fraldas ao lado das cervejas;
 - Resultado: o consumo cresceu 30% .

EXEMPLOS DE SUCESSO

- Logística e Varejo:
 - Data Science para predição de vendas e análise de outliers;
 - Aplicação na otimização e gestão de estoques;
 - Redução de custo de operação através de dados de geolocalização;
 - Segmentação de clientes através de dados de compras e informações de redes sociais.
-

Motivação

EXEMPLOS DE SUCESSO

- Mercado Financeiro:
 - Robôs que utilizam Machine Learning estão sendo construídos para operar no mercado de capitais;
 - Grandes instituições financeiras já estão utilizando Data Science para melhorar desempenho na bolsa de valores e outros segmentos.
 - Algoritmos estão sendo treinados recomendar carteiras de investimentos e identificar tendências no mercado.

EXEMPLOS DE SUCESSO

- Tecnologia e Entretenimento:
 - A Netflix faz recomendação de filmes e séries baseado nas preferências dos usuários;
 - O Spotify recomenda músicas baseado no histórico do usuário;
 - A Amazon com informações de milhares de clientes, faz recomendação de produtos ou serviços baseado em compras passadas.

Motivação

EXEMPLOS DE SUCESSO

- Saúde:
 - Data Science está sendo aplicado para predizer diagnósticos de pacientes;
 - Hospitais e empresas de planos de saúde estão usando Data Science para classificar o estado de saúde pacientes em tempo real;
 - Algoritmos estão sendo treinados para interagir com idosos, monitorar sinais vitais e alertar médicos e cuidadores.
 - Operadoras de planos de saúde estão usando Data Science para o bloqueio de solicitações de procedimentos suspeitas.

EXEMPLOS DE SUCESSO

- Ofertas de marketing de acordo com perfil do cliente;
- Identificação de spam;
- Quanto o cliente irá usar da franquia do plano;
- Estimativa do horário de chegada ao destino;
- Identificação de clientes semelhantes;
- Segmentar indivíduos baseados em itens em comuns;
- Bloqueio de transações financeiras.

Motivação

DESCOBERTA DE CONHECIMENTO EM BANCOS DE DADOS (DCBD)

- O processo de Descoberta de Conhecimento em Base de Dados (**DCBD**) (*Knowledge Discovery Database* - KDD) estuda formas de transformar grande volumes de dados em **informação útil e conhecimento**.
- Uma definição de DCBD: “Processo não trivial de extração de informações implícitas, anteriormente desconhecidas, e potencialmente úteis de uma fonte de dados”;
- “Torture os dados até eles confessarem”;
- O que é um padrão interessante ?

OBJETIVOS DO DCBD

- Entender o perfil dos clientes
- Desenvolvimento de novos produtos;
- Controle de estoque em postos de distribuição;
- Propaganda mal direcionada gera maiores gastos e desestimula o possível interessado a procurar as ofertas adequadas;
- Identificar terapias de sucessos para diferentes tratamentos
- Fraudes em planos de saúdes

OBJETIVOS

ATORES ENVOLVIDOS NO PROCESSO

- Frequentemente três atores estão envolvidos no DCBD:
 - **Analista de dados:** é aquele que entende das técnicas envolvidas no DCBD. Esse ator tem conhecimento sobre o funcionamento dos algoritmos e ferramentas utilizadas no processo, mas não necessariamente conhece o domínio ao qual os dados pertencem.
 - **Especialista no domínio:** conhece o domínio do problema que o processo será aplicado.
 - **Usuário:** é aquele que irá utilizar o resultado do processo. Normalmente o usuário não é somente uma pessoa, mas uma instituição, empresa ou departamento.
- Normalmente o papel de usuário e de especialista no domínio são exercidos pela mesma pessoa.



ATORES

ETAPAS DO PROCESSO

- Várias metodologias de DCBD existentes consideram que esse processo é estruturado em 5 fases:
 - Coleta;
 - Pré-processamento;
 - Transformação;
 - Mineração (*Data mining*);
 - Avaliação e Interpretação dos resultados.

ETAPAS

ETAPA DE COLETA

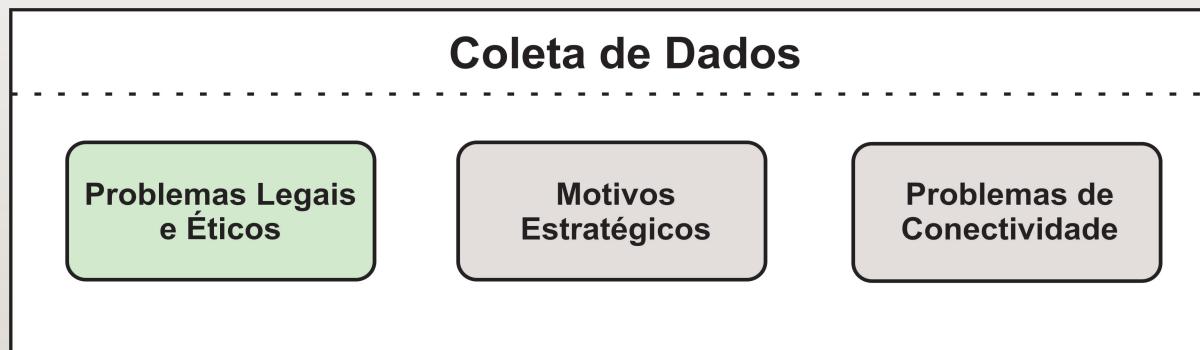
- Essa fase envolve inicialmente entrevistas com o especialista de domínio a fim de identificar os objetivos do projeto e os recursos disponíveis:
 - Pessoal e Dados;
 - Hardware e Softwares;
 - Existência de qualquer conhecimento prévio.
- O **especialista no domínio** pode fornecer ao analista de dados quais as **informações**, na sua opinião, são **mais relevantes** para a criação do modelo proposto;
- Entretanto, para **não limitar a originalidade** do conhecimento descoberto, o **analista de dados deve adicionar novas informações** e verificar a importância destas no conhecimento gerado.



COLETA

ETAPA DE COLETA

- Muitos dos **sistemas de gerenciamento de dados** que estão funcionando hoje possuem **documentação insatisfatória**, o que faz com que o processo de coleta de dados se torne difícil e **dependente do especialista de domínio**;
- Alguns dos principais desafios encontrados nessa fase:



COLETA

ETAPA DE COLETA

- **Problemas legais e éticos:** barreiras legais ou éticas podem impedir que dados sejam disponibilizados para análise:
 - Informações que identifiquem pacientes em bases de saúde, por exemplo;
- **Motivos estratégicos:** podem impedir o acesso a parte dos dados ou até mesmo a algumas estatísticas sobre eles:
 - A proporção de operações fraudulentas é uma informação estratégica mantida em absoluto segredo pelas companhias de cartão de crédito;
- **Problemas de conectividade:** surgem quando os sistemas armazenam os dados de forma descentralizada:
 - Uma empresa que possui filiais com bases de dados desconectadas.



COLETA

PRÉ-PROCESSAMENTO

- Uma **análise inicial** dos dados é feita para se ter sólidas definições sobre **estrutura de tabelas, valores de atributos, formatos e tipos de dados**, e qualquer operação necessária a escolha dos dados relevantes;
- Essa etapa representa um processo **semiautomático**, ou seja, o analista de dados deve identificar os principais problemas presentes nos dados e utilizar os métodos mais apropriados para trata-los.

Pré Processamento de Dados

Tratamento de
Valores Desconhecidos

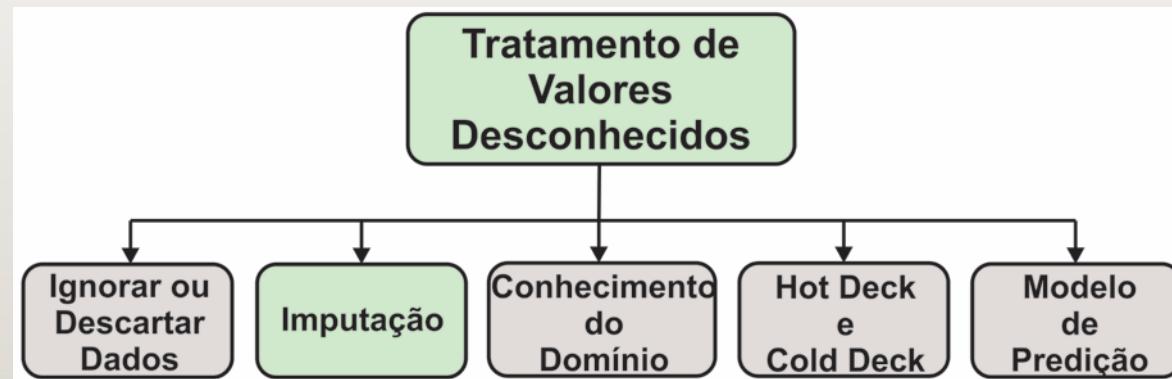
Tratamento de
Classes Desbalanceadas

Seleção de
Atributos

PRÉ-PROCESSAMENTO

TRATAMENTO DE VALORES DESCONHECIDOS

- Valores desconhecidos ou ausentes consistem no **não preenchimento** dos valores **de um atributo** para determinados casos;
 - Recusa por parte dos entrevistados em responder certas perguntas;
 - Não disponibilidade da informação no momento da entrada dos dados;
- O tratamento dos valores desconhecidos deve ser cuidadosamente pensado. As principais técnicas para isso são:



TRATAMENTO DE VALORES DESCONHECIDOS

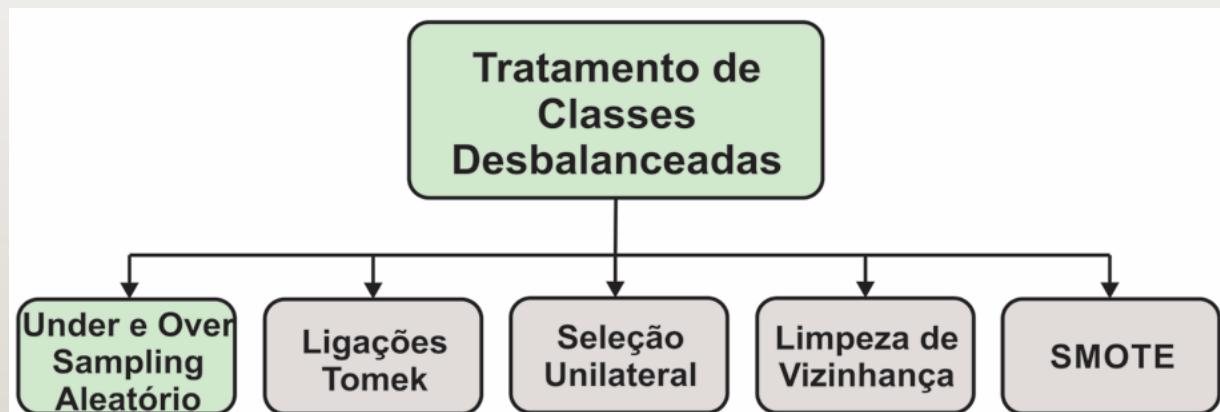
- **Ignorar ou descartar dados:** os exemplos com um ou mais valores desconhecidos são removidos ou os atributos com grande quantidade de valores desconhecidos são removidos;
- **Imputação:** consiste na substituição dos valores desconhecidos por valores estimados por meio de alguma informação extraída do conjunto de dados, como média ou moda dos atributos;
- **Conhecimento do domínio:** consiste em utilizar a experiência do especialista do domínio para a estimação de valores. Normalmente essa substituição é manual, e portanto restrita a bases com poucos valores desconhecidos.

TRATAMENTO DE VALORES DESCONHECIDOS

- ***Hot deck e cold deck***: os exemplos são particionados em grupos utilizando um método de aprendizado supervisionado, e cada exemplo com valores desconhecidos é associado a um desses grupos. Em seguida, os exemplos do grupo são utilizados para estimar os valores desconhecidos por meio da imputação. A diferença entre esses métodos é que no *hot deck* são utilizados todos os exemplos para gerar o agrupamento, já no *cold deck* são utilizados somente os exemplos que não possuem valores desconhecidos;
- **Modelos preditivos**: consiste na criação de um modelo preditivo para estimar valores que irão substituir os valores desconhecidos.

TRATAMENTO DE CLASSESS DESBALANCEADAS

- A maioria dos **algoritmos** de aprendizagem de máquina **tem dificuldades em criar** um modelo que **classifique** com precisão os exemplos da **classe minoritária**, e que normalmente possuem maior importância;
- As principais técnicas para tratar esse problema são:



PRÉ-PROCESSAMENTO

TRATAMENTO DE CLASSE DESBALANCEADAS

- ***Under e Over sampling aleatório:*** no primeiro método elementos da classe majoritária são eliminados de forma aleatória, já no segundo os elementos da classe minoritária são replicados aleatoriamente;
 - No ***Under sampling aleatório*** dados importantes podem ser eliminados;
 - Já o ***Over sampling aleatório***, pode ocasionar *overfitting*.

TRATAMENTO DE CLASSESS DESBALANCEADAS

- **Ligações Tomek:** nessa técnica ruídos e elementos próximos as bordas são removidos da classe majoritária;
 - Sejam e_i e e_j dois exemplos de classes diferentes;
 - Seja $d(\cdot)$ uma função de distância;
 - Um par de exemplos (e_i, e_j) constitui uma ligação *Tomek* se não existir um exemplo e_k tal que a distância $d(e_k, e_i) < d(e_i, e_j)$ ou $d(e_k, e_j) < d(e_i, e_j)$;
 - O exemplo da classe majoritária que formam ligações *Tomek* são removidos.

TRATAMENTO DE CLASSESS DESBALANCEADAS

- **Seleção unilateral:** consiste na aplicação do algoritmo para identificação de ligações *Tomek* seguido da aplicação do K-vizinhos mais próximos (KNN);
 - Ligações *Tomek* são utilizadas para verificar exemplos da classe majoritária que se sobrepõem a classe minoritária;
 - Em seguida, o KNN é utilizado para remover exemplos muitos distantes da fronteira de decisão.

TRATAMENTO DE CLASSE DESBALANCEADAS

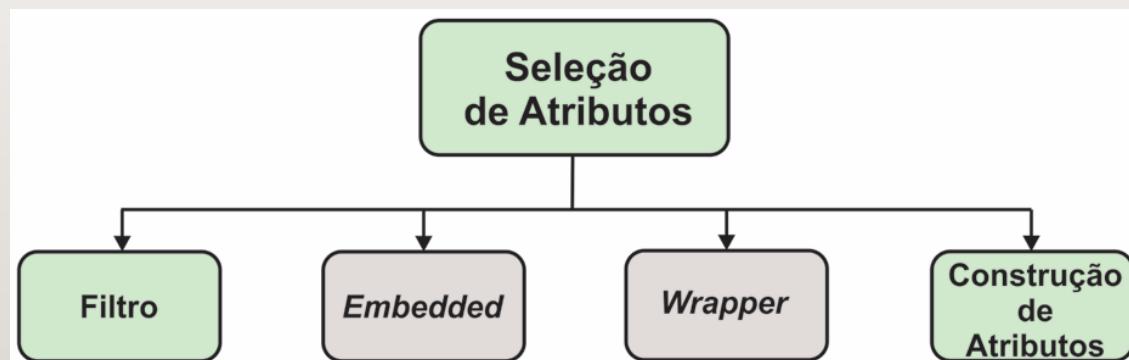
- **Limpeza da vizinhança:** esse método utiliza a regra do vizinho mais próximo para remover os exemplos da classe majoritária. São removidos os exemplos que possuem classe diferente de pelo menos 2 dos seus 3 vizinhos mais próximos;
- **SMOTE:** consiste em gerar casos sintéticos para a classe de interesse a partir de exemplos existentes;
 - Os exemplos da classe minoritária são interpolados aleatoriamente com os seus k-vizinhos mais próximos.

SELEÇÃO DE ATRIBUTOS

- O processo de **seleção de atributos** é necessário para identificar e **eliminar inconsistências** presentes nos dados, como:
 - **Poluição de dados:** presença de valores distorcidos, por utilizar o sistema além de suas funcionalidades originais, como um campo de sexo preenchido com PJ (Pessoa Jurídica);
 - **Valores preenchidos de forma default:** consiste no preenchimento com o mesmo valor para todos os casos que a informação não foi especificada;
 - **Informações duplicadas e redundantes:** consiste no armazenamento de informações idênticas em diferentes atributos. O maior dano causado pela redundância é o aumento no tempo de processamento dos algoritmos de aprendizado.

SELEÇÃO DE ATRIBUTOS

- Os **efeitos** imediatos da **seleção de atributos** para a aplicação são:
 - **Execução mais rápida** dos algoritmos utilizados nas fases seguintes;
 - **Melhoria da qualidade dos dados**, que consequentemente conduz a um melhor desempenho no processo de DCBD;
 - Aumento da **compreensibilidade dos resultado** obtidos.
- As abordagens para seleção de atributos podem ser divididas em:



SELEÇÃO DE ATRIBUTOS

- O método de **filtragem** utiliza **informações da própria base** de treinamento para escolher os atributos a serem utilizados;
- É possível utilizar uma métrica para **avaliar a qualidade** de cada **atributo individualmente** e ordena-los de acordo com esse valor, em seguida, são **removidos** os que possuem valor de importância **menor** que um **limiar**, ou mesmo, **fixar o número de atributos** utilizados;
- O algoritmo de **ganho de informação** é bastante utilizado para ranquear a importância dos atributos.

SELEÇÃO DE ATRIBUTOS

- Os métodos ***Embedded*** selecionam o conjunto de atributos no **próprio processo de construção do modelo de classificação** durante a fase de treinamento;
- As **árvores de decisão** são exemplos de algoritmos do tipo ***embedded***, pois elas selecionam internamente um **conjunto de atributos** para compor a árvore utilizando em cada etapa uma **função para avaliar qual é o atributo com maior capacidade de discriminação da classe**.

SELEÇÃO DE ATRIBUTOS

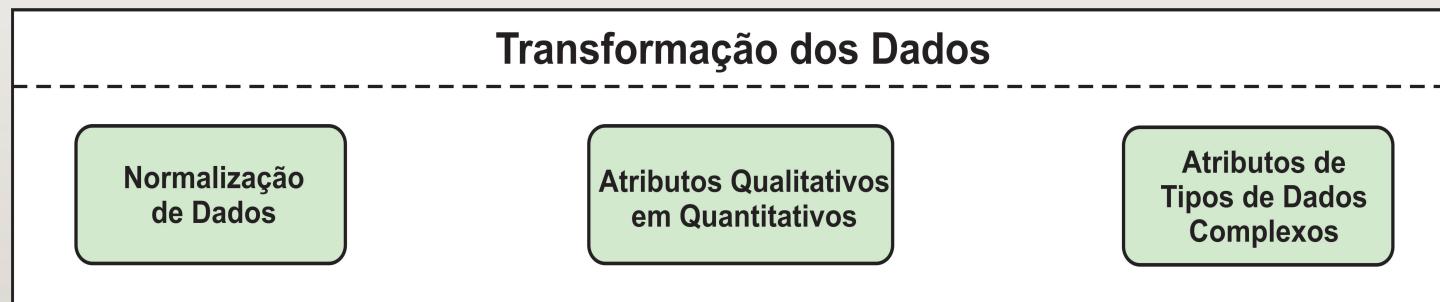
- Os métodos do tipo ***wrapper*** utilizam o próprio **algoritmo de classificação** (que será utilizado na fase de mineração) para avaliar os subconjuntos de atributos de acordo com a sua **capacidade preditiva**;
- Em resumo, essa abordagem consiste em um **método de força bruta** para testar todas as **combinações de atributos** possível e encontrar a que **produz melhor resultado**;
- ***Wrappers*** geralmente **produzem bons resultados**, no entanto, essa abordagem possui um **elevado custo computacional** e torna **inviável** seu uso para bases com um **grande número de atributos**.

SELEÇÃO DE ATRIBUTOS

- **Atributos fracamente**, indiretamente ou condicionalmente **relevantes** podem ser **individualmente inadequados**, entretanto, esses atributos **podem ser combinados** gerando novos atributos **altamente representativos** para a descrição do conceito;
- Dessa forma, a **construção de atributos** consiste em **produzir novos atributos relevantes** para a descrição do problema **a partir de atributos primitivos**;
- Normalmente é **utilizado o conhecimento** do usuário ou especialista de domínio para **guiar a composição dos atributos**.

TRANSFORMAÇÃO DOS DADOS

- O objetivo da **transformação de dados** é **modificar a forma** com que os dados estão representados, pois os **algoritmos utilizados** na fase de **Mineração não são capazes de analisar certos tipos de dados**;
 - Data e hora, por exemplo.
- Dessa forma atributos com esses tipos de dados são **transformados em outros atributos com a mesma informação**, sendo as mais comuns:



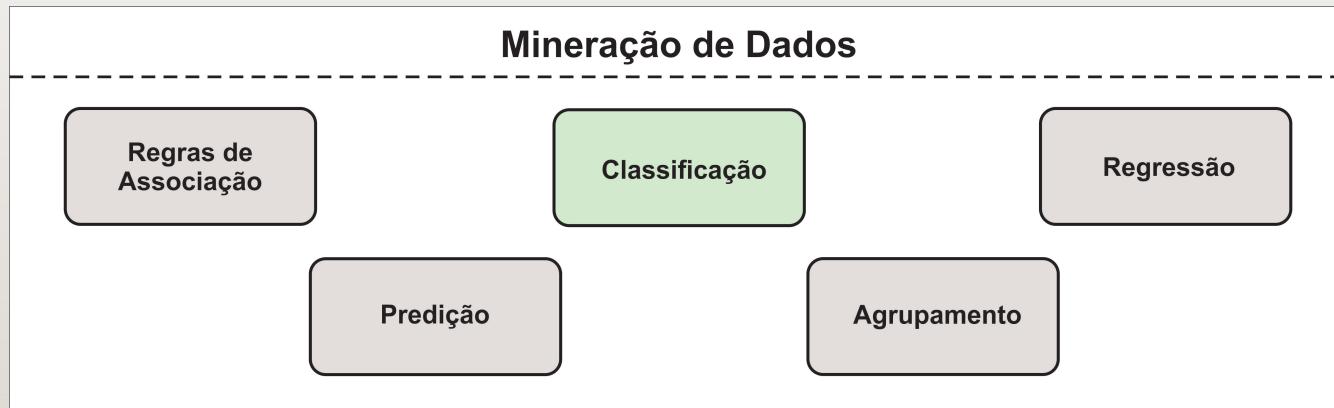
TRANSFORMAÇÃO

TRANSFORMAÇÃO DOS DADOS

- A **normalização** consiste em ajustar as faixas de valores dos atributos para o mesmo intervalo, tais como -1 a 1 ou 1 a 0. Essa técnica é importante quando os atributos possuem grandes intervalos de valores;
- A **transformação de atributos qualitativos em quantitativos** é necessária porque alguns classificadores só conseguem trabalhar com valores numéricos;
- Um atributo possui **tipo de dado complexo** quando ele não é considerado nem qualitativo nem quantitativo, como por exemplo data.

MINERAÇÃO DE DADOS

- O objetivo da fase de **mineração de dados** é a verificação de uma hipótese ou obtenção de novos padrões sobre os dados:
 - Tarefas de **predição** ou **descrição**.
- Dois fatores são fundamentais para o sucesso dessa fase:
 - **Atributos relevantes para análise e tarefa de mineração adequada.**



MINERAÇÃO

MINERAÇÃO DE DADOS

- **Regras de associação:** relaciona a ocorrência de um determinado conjunto de itens, ou seja, se uma transação X ocorre Y também tende a ocorrer;
 - “98% dos consumidores que adquiriram pneus e acessórios de automóveis também se interessaram por serviços automotivos”;
 - A análise de associação em um banco de dados pode gerar uma grande quantidade de regras, portanto devem ser definidos parâmetros de interesse para filtrar as regras que são importantes.
- **Classificação:** o objetivo dessa tarefa é descobrir relações existentes entre os atributos de predição e o atributo alvo, utilizando registros cuja classificação é conhecida.

MINERAÇÃO DE DADOS

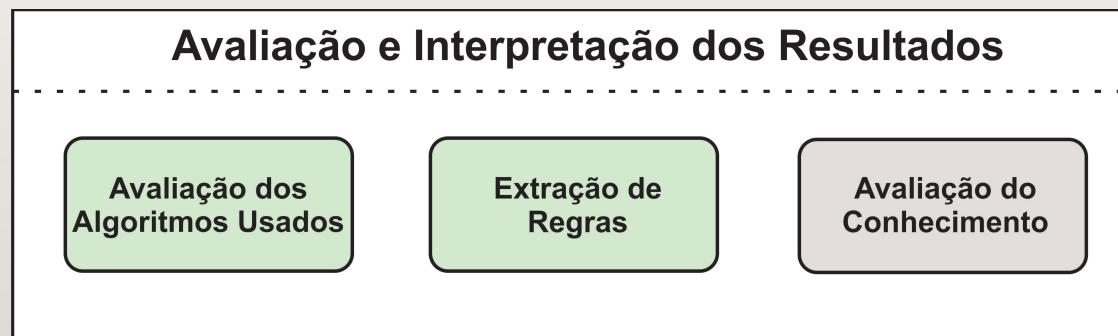
- **Régressão:** esse processo é semelhante ao de classificação, a principal diferença entre eles é que a classificação lida com valores discretos enquanto que a regressão com valores contínuos;
- **Predição:** também é semelhante aos processos anteriores, exceto pelo fato de que os registros possuem dados temporais e serão classificados de acordo com alguma predição de comportamento futuro ou predição de algum valor futuro;
- **Agrupamento:** consiste em dividir uma população heterogênea em um número de subgrupos mais homogêneos. Esses grupos não são pré-definidos e também não há exemplos rotulados.



MINERAÇÃO

AVALIAÇÃO E INTERPRETAÇÃO DOS RESULTADOS

- Essa etapa deve envolver todos os atores participantes do projeto:
 - **O analista de dados** tenta descobrir se os algoritmos utilizados atingiram as expectativas, avaliando os resultados de acordo com alguma métrica;
 - **O especialista no domínio** irá verificar a compatibilidade dos resultados usando o seu conhecimento sobre o problema em questão;
 - O usuário dará o julgamento final sobre a aplicabilidade do processo de DCBD.



AVALIAÇÃO

