Tecnologia da Informação

RESUMO

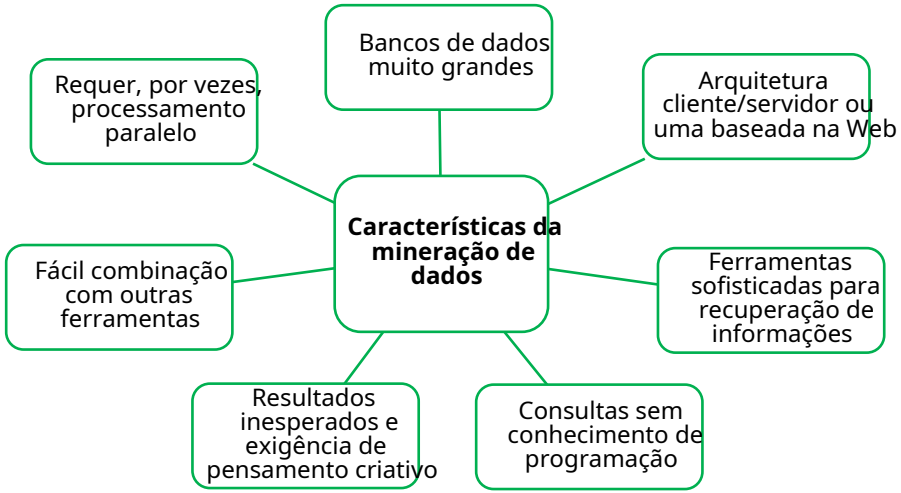
Data Mining

CONCEITOS DE MINERAÇÃO DE DADOS

A **mineração de dados (data mining)** refere-se:

- **mineração ou descoberta de novas informações em termos de padrões ou regras** com base em grandes quantidades de dados.
- processo pelo qual os **padrões anteriormente desconhecidos em dados são descobertos**.
- **processo que utiliza técnicas de estatística, matemática e inteligência artificial para extrair e identificar informações úteis e subsequentes conhecimentos** (ou padrões) em grandes conjuntos de dados.
- **processo não trivial de identificar padrões válidos, novos, potencialmente úteis e**, em última instância, **compreensíveis** em dados armazenados em bancos de dados estruturados.

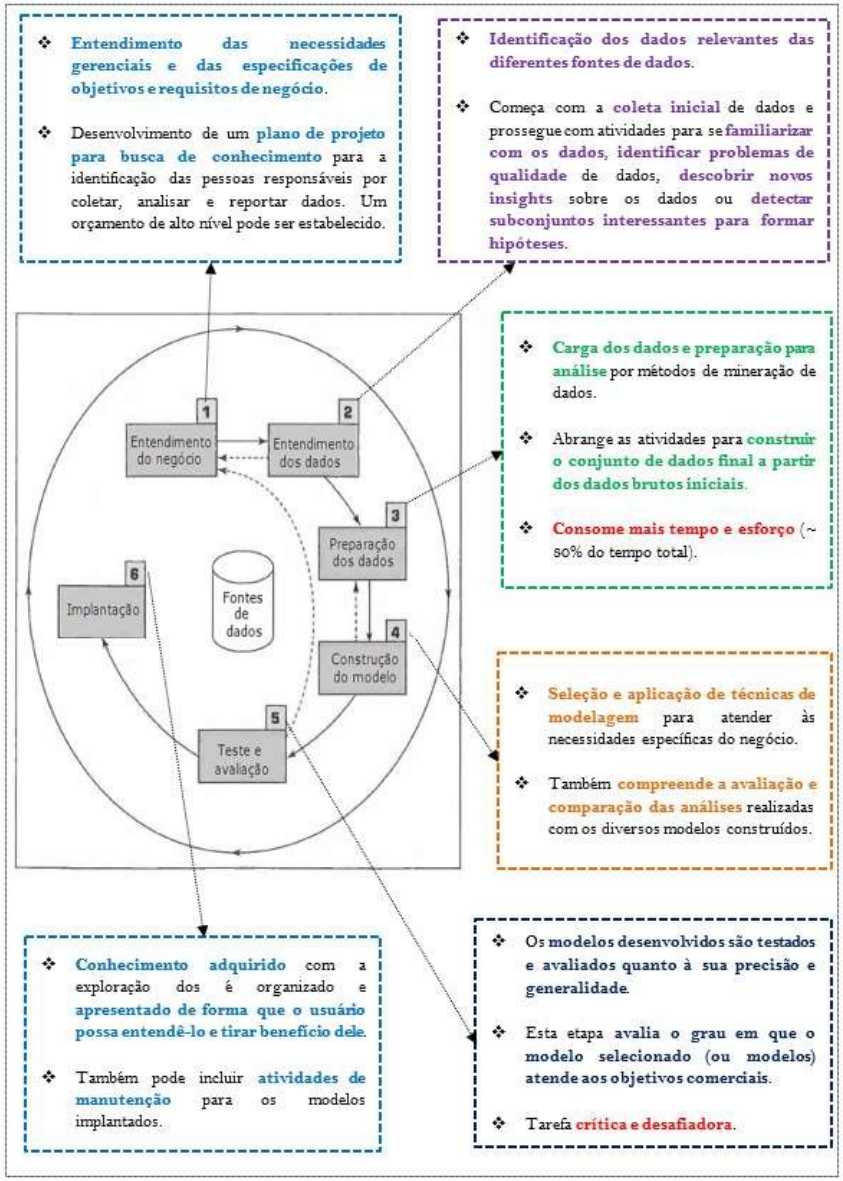
CARACTERÍSTICAS DA MINERAÇÃO



OBJETIVOS DA MINERAÇÃO

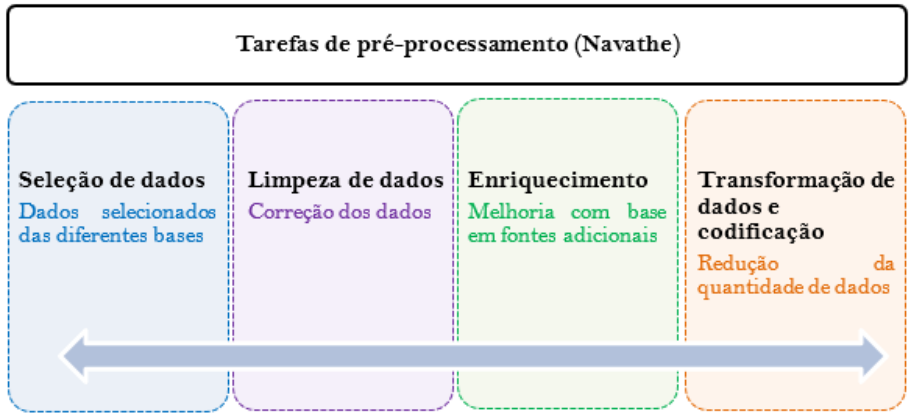
- **Previsão**: a mineração de dados pode **mostrar como certos atributos dos dados se comportarão no futuro**.
- **Identificação**: os padrões de dados podem ser usados para **identif a existência de um item, um evento ou uma atividade**.
- **Classificação**: a mineração de dados pode **particionar os dados d modo que diferentes classes ou categorias** possam ser identifica com base em combinações de parâmetros.
- **Otimização**: um objeto relevante da mineração de dados pode **otimizar o uso de recursos limitados**, como tempo, espaço, dinhe ou materiais e maximizar variáveis de saída como vendas ou lucros sob determinadas restrições.

CRISP-DM

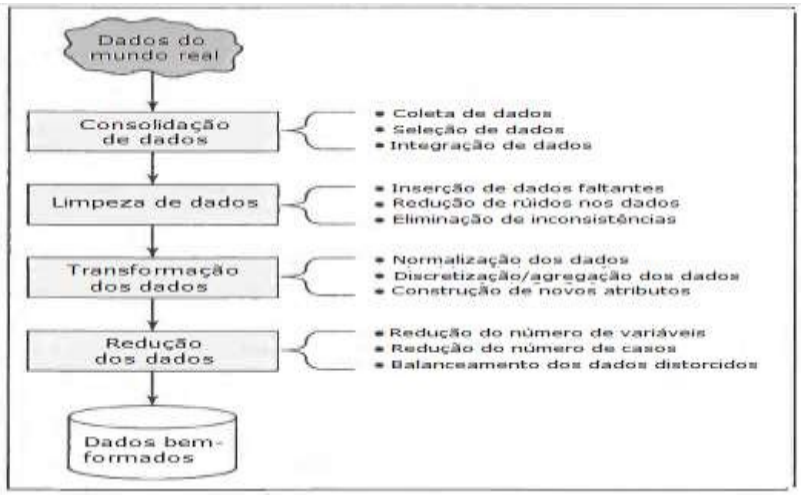


TÉCNICAS DE PRÉ-PROCESSAMENTO

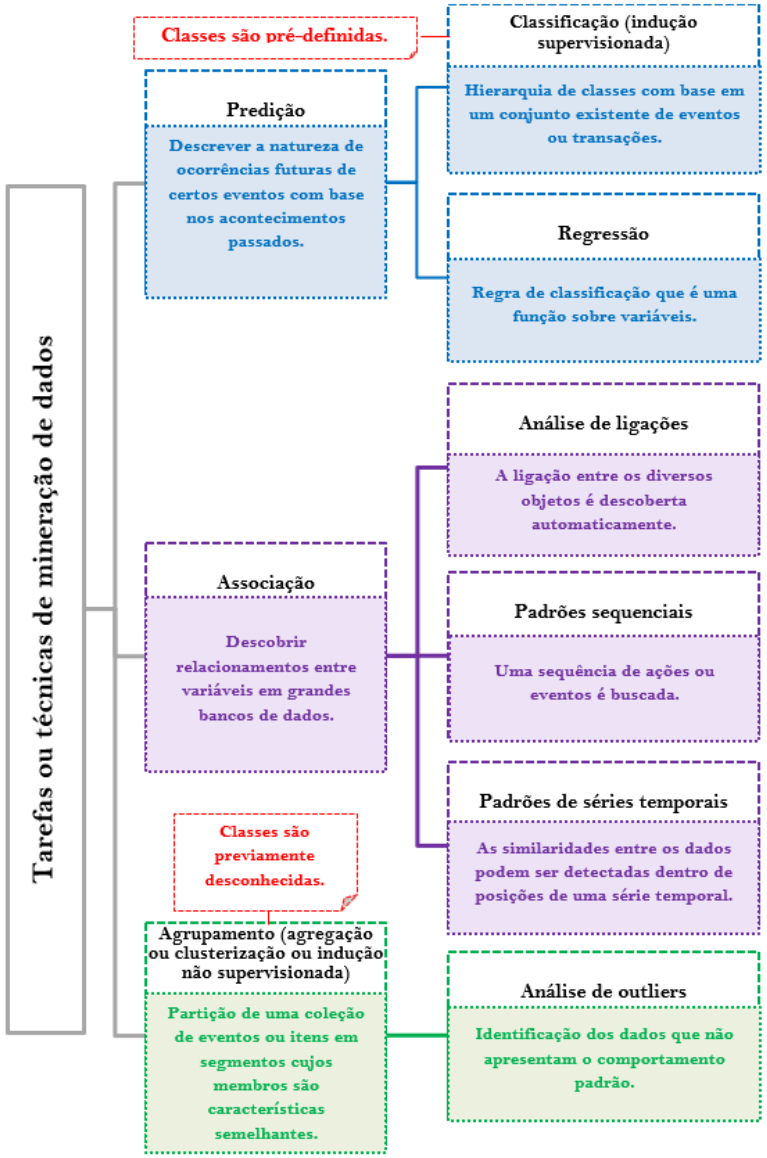
As **técnicas para pré-processamento** estão voltadas para a **preparação dados para que estes sejam submetidos à mineração de dados**.



Segundo o CRISP-DM:



TÉCNICAS OU TAREFAS DE MINERAÇÃO



CLASSIFICAÇÃO

A **classificação** é o processo de **aprender um modelo que descrev diferentes classes de dados**. As **classes são predefinidas** e, portanto, es tipo de atividade é também chamado de **aprendizado supervisionado**.

A **classificação** é **talvez a mais comum de todas as tarefas de mineraç de dados**. O objetivo da classificação é **analisar os dados históricos armazenados em um banco de dados e gerar automaticamente u modelo que pode prever o comportamento futuro**.

Principais algoritmos de classificação:

- As **redes neurais** envolvem o desenvolvimento de **estruturas matemáticas** (um tanto parecidas com as redes neurais biológicas cérebro humano) **que têm a capacidade de aprender com experiências passadas apresentadas sob uma forma bem estruturada** dos conjuntos de dados. Elas tendem a ser mais efeti quando o número de variáveis envolvidas é bastante grande e relações entre elas são complexas e imprecisas.
- As **árvores de decisão classificam os dados em um número finito de classes com base nos valores das variáveis de entrada**. árvores de decisão são essencialmente uma hierarquia de declaraç se-então e, portanto, são significativamente mais rápidas do que as redes neurais. Elas são mais apropriadas para dados categorizados intervalos de dados. Portanto, incorporar variáveis contínuas em u estrutura de árvore de decisão requer discretização; ou seja, conve variáveis numéricas de valor contínuo em intervalos e categorias. A árvore de decisão auxilia no processo de **estratificação dos dados** separando as classes com base nos valores de entrada.

REGRAS DE ASSOCIAÇÃO

As **regras de associação** são uma técnica popular para **descobrir relacionamentos interessantes entre variáveis** em grandes bancos de dados. As derivações comuns das regras de associação são:

- **Análise de ligações**: a **ligação entre os diversos objetos** de interesse é descoberta automaticamente.
- **Padrões sequenciais**: uma **sequência de ações ou eventos é buscada**. A detecção de padrões sequenciais é **equivalente à detecção de associações entre eventos com certos relacionamentos temporais**.
- **Padrões dentro de série temporal**: as **similaridades** entre os dados podem ser **detectadas** dentro de posições de uma série temporal, que é uma **sequência de dados tomados em intervalos regulares**.

Uma regra de associação deve satisfazer alguma medida de interesse do analista de dados. Duas medidas comuns são o suporte e a confiança.

- **Suporte ou prevalência**: **frequência que um conjunto de itens específico ocorre no banco de dados**, ou seja, o percentual de transações que contém todos os itens em um dado conjunto.
- **Confiança ou força**: **probabilidade de que exista relação** entre itens.

O **algoritmo Apriori** é o algoritmo mais utilizado para descobrir regras de associação. **Dado um conjunto de conjuntos de itens** (por exemplo, conjuntos de transações de varejo com a listagem de itens individuais adquiridos), **o algoritmo tenta encontrar subconjuntos comuns a pelo menos um número mínimo de conjuntos de itens** (isto é, cumpre com um suporte mínimo).

CLUSTERIZAÇÃO

A **análise de clusters (análise de agrupamentos ou análise de aglomerações ou análise de partições)** é um método de mineração de d essencial para **classificar itens, eventos ou conceitos em agrupamento comuns chamados de clusters**. O **objetivo é classificar casos** (por exemplo, pessoas, coisas, eventos) **em grupos ou clusters, de modo que grau de associação seja forte entre os membros do mesmo cluster e f entre os membros de diferentes clusters**. As **classes não são previame definidas**, mas muitas vezes, os algoritmos de cluster geralmente **requer uma especificação do número de clusters** a serem encontrados.

A clusterização pode se proceder de duas formas gerais:

- **Divisivo**: todos os itens **começam em um cluster e são quebrados** em clusters menores.
- **Aglomerativo**: todos os itens **começam em clusters individua e os clusters são unidos** baseando-se em suas semelhanças.

A clusterização pode ser realizada com métodos hierárquicos ou não-hierárquicos.

- Os **métodos hierárquicos** não exigem que já se tenha um núm inicial de clusters e são considerados **inflexíveis** uma vez que **nã se pode trocar um elemento de grupo**.

- Os **métodos não-hierárquicos** da análise de cluster são caracterizados pela necessidade de definir uma partição inicial e pela **flexibilidade**, uma vez que os **elementos podem ser trocados de grupo** durante a execução do algoritmo.

A análise de clusters pode ser baseada em um ou mais dos seguintes métodos gerais:

- **Métodos estatísticos**: k-means, k-modes, k-medoids, etc.
  - **K-means (k média)**: o algoritmo **atribui cada ponto de dados** (cliente, evento, objeto, etc.) ao **cluster cujo centro** (também chamado centróide) **é o mais próximo**. O **centro é calculado como a média de todos os pontos no cluster**.
  - **K-modes (k moda)**: estende o paradigma k-means para clusterizar dados categóricos (nominais) ao **trocar a média de clusters pela moda** (elementos que mais se repetem).
  - **K-medoids (k mediana)**: em relação a esse algoritmo, temos duas acepções possíveis.
    - **1ª acepção**: pode ser encontrado na literatura que o k-medoids **ao invés de usar a média para definir o centro dos clusters, utiliza a mediana** (valor mais ao centro do conjunto de dados). .
    - **2ª acepção**: é uma variação do k-means, mas não utiliza a média como centro do grupo, e sim, considera um problema onde um **objeto** é o **centro do próprio grupo**, chamado de objeto representativo ou medoide. O objeto central é **aquele com menor dissimilaridade média a todos os outros** objetos do grupo.
- **Redes neurais**: **estruturas matemáticas que têm a capacidade d aprender com experiências passadas apresentadas sob uma for bem estruturada** dos conjuntos de dados.
- **Lógica difusa**: forma de **lógica multivalorada na qual os valores lógicos das variáveis podem ser qualquer número real entr (FALSO) e 1 (VERDADEIRO)**. A lógica difusa foi estendida para lidar com o conceito de verdade parcial, onde o valor verdade pode compreender entre completamente verdadeiro e completamente fa
- **Algoritmos genéticos**: são implementados como uma **simulação computador em que uma população de representações abstrat de solução é selecionada em busca de soluções melhores**. A evolução geralmente se inicia a partir de um conjunto de solu criado aleatoriamente e é realizada por meio de gerações. A c geração, a adaptação de cada solução na população é avaliada, alg indivíduos são selecionados para a próxima geração, e recombinad ou mutados para formar uma nova população. A nova população e é utilizada como entrada para a próxima iteração do algoritmo.

DETECÇÃO DE ANOMALIAS

A **detecção de anomalias** consiste **na identificação de padrões em dados com um comportamento diferente do esperado**.

Os resultados produzidos pelos métodos de **detecção de anomalias** são de um dos dois tipos seguintes:

- **Pontuações**: os métodos de pontuação atribuem uma pontuação de anomalia para cada instância no teste de dados, dependendo do grau da anomalia.
- **Rótulos**: os métodos usados atribuem um rótulo (normal ou ar para cada instância de teste.

MODELAGEM PREDITIVA

A **modelagem preditiva** é uma técnica estatística para modelar e enc padrões,que **utiliza dados históricos para realizar previsões de tendências, padrões de comportamento ou eventos futuros**.

A **modelagem preditiva** utiliza de **estatísticas e modelos mater para prever resultados futuros**.

APRENDIZADO DE MÁQUINA

Aprendizado de Máquina (ou machine learning) é um **método de anál de dados que automatiza o desenvolvimento de modelos analíticos**.

A **mineração de dados descobre padrões e conhecimento previament desconhecidos** e o **aprendizado de máquina usa esses padrões e conhecimentos adquiridos**, aplicando isso a outros dados, e, em seguid aplicando automaticamente esses resultados à tomada de decisões e açõ

MINERAÇÃO DE TEXTO