TI TOTAL

ÁREA FISCAL E CONTROLE

Professor
Ramon Souza

Tecnologia da Informação

TEORIA

DATA MINING

SUMÁRIO

A nossa aula é bem esquematizada, então para facilitar o seu acesso aos **esquemas**, você pode usar o seguinte índice:

GLOSSÁRIO DE TERMOS

Algoritmo: sequência de ações que visam obter uma solução para um determinado tipo de problema.

Aprendizado de máquina: método que automatiza o desenvolvimento de modelos analíticos.

Classificação: descreve os dados e os categoriza em classes pré-definidas.

Cluster: grupo de elementos que apresenta similaridade.

Confiança ou força: probabilidade de existir relação entre itens.

Data mining ou mineração de dados: descoberta de padrões ou regras em dados.

Depuração: processo de encontrar ou reduzir defeitos.

Estratificar: separar em níveis.

Outlier ou anomalia: ponto fora da curva.

Particionamento: divisão de elementos em grupos.

Predição: previsão.

Processo não trivial: processo que não é facilmente realizado. Contrário de simples.

Suporte ou prevalência: frequência que um conjunto de itens ocorre.

1. DATA MINING

1.1 Noções de mineração de dados

A grande quantidade de dados gerada pelas organizações requer mecanismos mais voltad para auxiliar a tomada de decisões. Os gestores precisam analisar essa grande "massa" de dados e identificar padrões, regras, tendências e comportamentos excepcionais para que possam tomar decisões e agir para otimizar os negócios. Dada esta quantidade de dados crescente, o que torna inviável a análise humana e manual, a **mineração de dados (data mining)** é utilizada para auxiliar nessas análises.

A **mineração de dados (data mining)** refere-se à **mineração ou descoberta de novas informações em termos de padrões ou regras** com base em grandes quantidades dados. Dito de outro modo, o termo **mineração de dados** foi originalmente usado para descrever o processo pelo qual os **padrões anteriormente desconhecidos em dados são descobertos**.

Tecnicamente falando, a **mineração de dados** é um **processo que utiliza técnicas estatística, matemática e inteligência artificial para extrair e identificar informações úteis e subsequentes conhecimentos** (ou padrões) em grandes conjuntos de dados.

Em outro conceito, a **mineração de dados** é entendida como o **processo não trivial de identificar padrões válidos, novos, potencialmente úteis e**, em última instância, **compreensíveis** em dados armazenados em bancos de dados estruturados.

Os termos **extração de conhecimento, análise de padrões, arqueologia de dados, busc de padrões ou dragagem de dados** podem ser usados como sinônimos para mineração d dados.

É importante destacar que a **mineração de dados** pode ser utilizada junto com um data warehouse para ajudar com certos tipos de decisões. Porém, **não está restrita a um DW**, podendo ser aplicada também a bancos de dados operacionais com transações individuais **Alguns recursos de mineração são fornecidos por SGBDs relacionais, mas de forma limitada**.

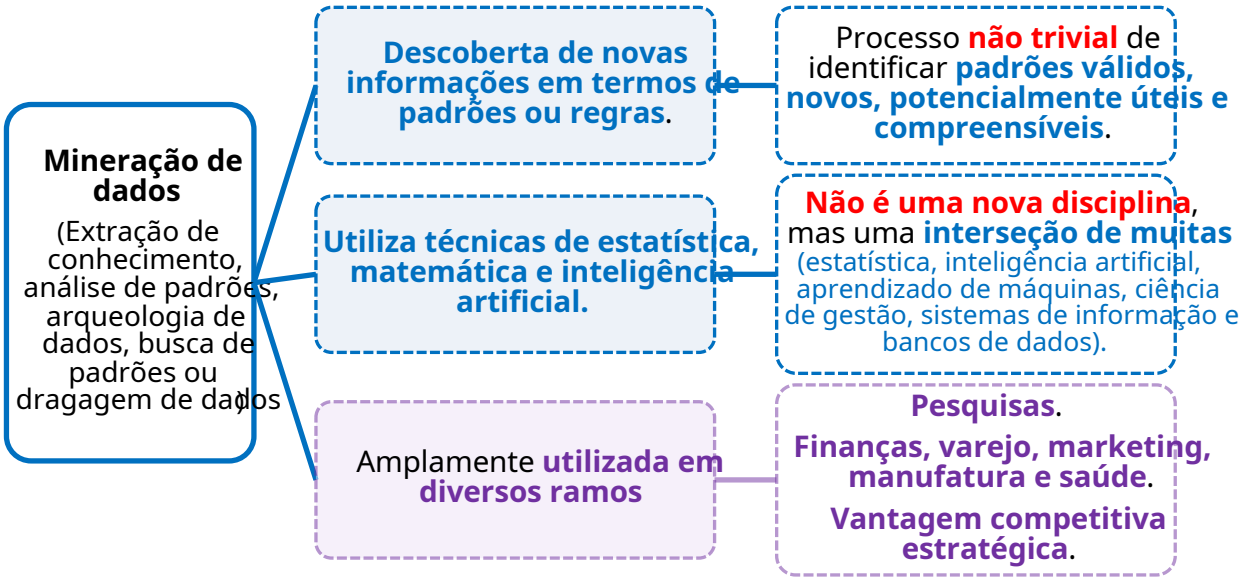
A **mineração de dados não é uma nova disciplina**, mas sim uma nova definição para o uso de muitas disciplinas. A mineração de dados está bem posicionada na **interseção de muitas disciplinas**,incluindo **estatística,inteligênciaartificial, aprendizadode máquinas, ciência de gestão, sistemas de informação e bancos de dados**. Usando avanços em todas essas disciplinas, a mineração de dados se esforça para avançar n extração de informações e conhecimentos úteis de grandes bancos de dados. É um campo emergente que atraiu muita atenção em pouco tempo.

A **mineração de dados** é **amplamente utilizada em diversos ramos**. Os dados gerados pela **Internet** estão aumentando rapidamente em volume e complexidade. Grandes quantidades de **dados genômicos** estão sendo gerados e acumulados em todo o mundo. Disciplinas como a **astronomia e a física nuclear** criam enormes quantidades de dados regularmente. **Pesquisadores médicos e farmacêuticos** constantemente geram e armazenam dados que podem ser usados em aplicativos de mineração de dados pa identificar melhores maneiras de diagnosticar e tratar com precisão doenças e descobrir novos e melhores medicamentos.

Do lado comercial, talvez o uso mais comum da mineração de dados tenha sido nos setore de **finanças, varejo, marketing, manufatura e saúde**. A mineração de dados é usada par detectar e reduzir atividades fraudulentas; para identificar os padrões de compra dos clientes; para identificar clientes rentáveis; para segmentar clientes; identificar regras de negociação a partir de dados históricos; e para auxiliar no aumento da rentabilidade usand a análise da cesta de mercado.

Uma organização que efetivamente aproveita as ferramentas e tecnologias de **mineração de dados** pode **adquirir e manter uma vantagem competitiva estratégica**. A mineração de dados oferece às organizações um ambiente indispensável para melhorar a decisão de forma a **explorar novas oportunidades pela transformação dos dados em uma arma estratégica**.

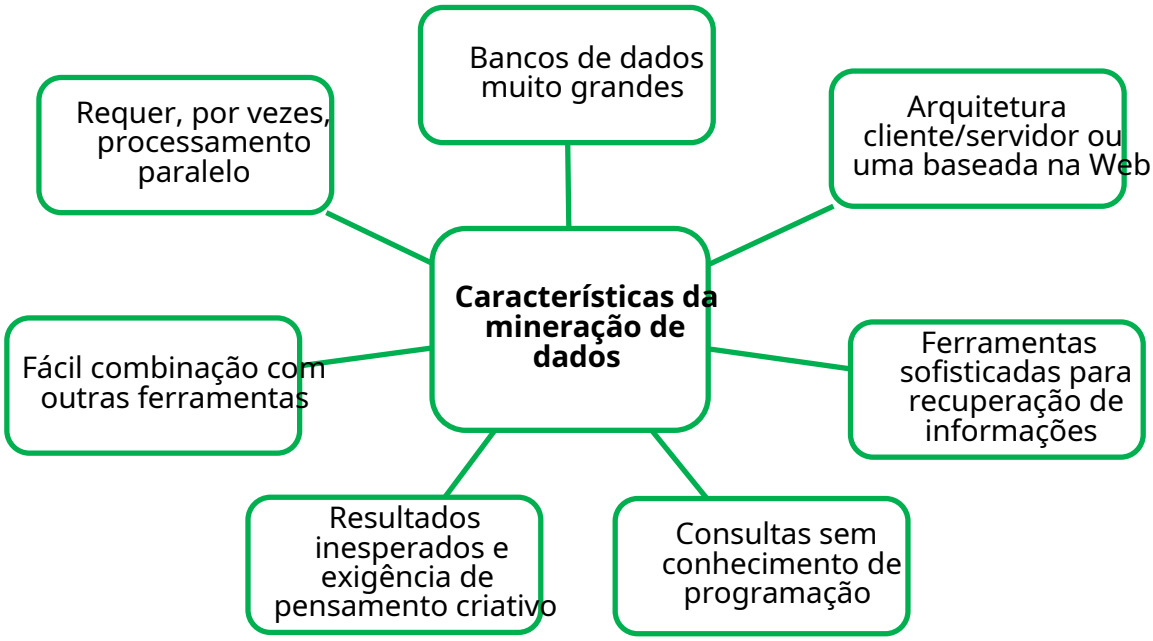
Vamos fixar o apreendido até aqui com um **esqueminha**!



Esquema 1 – Mineração de dados.

As **principais características** da mineração de dados são:

- Os dados são frequentemente dispostos em **bancos de dados muito grandes**, que às vezes contêm dados de vários anos. Em muitos casos, os dados são limpos e consolidados em um data warehouse.
- O ambiente de mineração de dados geralmente é uma **arquitetura cliente/servidor ou uma baseada na Web**.
- Novas ferramentas sofisticadas ajudam a **recuperar informações de arquivos corporativos e registros públicos** e podem **extrair dados de bancos não estruturados** (ex. bancos do Lotus Notes, textos e intranets empresariais).
- Os usuários podem realizar **consultas** com o uso de ferramentas especializadas **sem conhecimentos de programação**.
- Muitas vezes **encontram-se resultados inesperados** e **exige-se que os usuários finais pensem criativamente ao longo do processo**, incluindo a interpretação das descobertas.
- Ferramentas de mineração são **facilmente combinadas com planilhas e outras ferramentas de desenvolvimento de software**.
- Às vezes é necessário usar **processamento paralelo para suportar a carga** de grandes quantidades de dados e de consultas.
- **Alguns métodos de mineração são específicos para os tipos de dados** que manipulam. Fornecer-lhes tipos de dados incompatíveis pode levar a modelos incorretos ou a uma parada do processo de desenvolvimento do modelo.



Esquema 2 – Características da mineração de dados.

A mineração de dados costuma ser executada com alguns objetivos finais ou aplicações. Segundo Navathe, de um modo geral, estes objetivos se encontram nas seguintes classes:

- **Previsão**: a mineração de dados pode **mostrar como certos atributos dos dados se comportarão no futuro**. Para realizar a previsão (ou prognóstico), a **lógica de negócios é utilizada** em conjunto com a mineração de dados. Ex.: previsão d compras sob certos descontos.

- **Identificação**: os padrões de dados podem ser usados para **identificar a existência de um item, um evento ou uma atividade**. Ex.: intrusos tentando quebrar um sistema.

- **Classificação**: a mineração de dados pode **particionar os dados de modo qu diferentesclasses ou categorias**possamser identificadascom base em combinações de parâmetros. Ex.: segmentação de clientes.

- **Otimização**: um objeto relevante da mineração de dados pode ser **otimizar o uso de recursos limitados**, como tempo, espaço, dinheiro ou materiais e maximizar variáveis de saída como vendas ou lucros sob determinadas restrições.

Esquema 3 – Objetivos da mineração de dados.

1- (CESPE / CEBRASPE - 2020 - Ministério da Economia - Tecnologia da Informação - Ciência de Dados) No que se refere à mineração de dados, julgue o item a seguir.

Na etapa de mineração do data mining, ocorre a seleção dos conjuntos de dados que serão utilizados no processo de mining.

Resolução:

Que confusão hein? rsrsrs.Mineração de dados é a tradução de data mining, logo não é um etapa do processo. O examinador deve ter tido o objetivo de tratar da etapa de entendime dos dados do modelo CRISP-DM.

Gabarito: **Errado**.

2- (CESPE - 2018 - Polícia Federal - Perito Criminal Federal - Conhecimentos Básicos - Todas as Áreas) Acerca de banco de dados, julgue o seguinte item.

Descobrir conexões escondidas e prever tendências futuras é um dos objetivos da mineraç de dados, que utiliza a estatística, a inteligência artificial e os algoritmos de aprendizagem de máquina.

Resolução:

A **mineração de dados (data mining)** refere-se à **mineração ou descoberta de novas informações em termos de padrões ou regras** com base em grandes quantidades dados. Dito de outro modo, o termo **mineração de dados** foi originalmente usado para descrever o processo pelo qual os **padrões anteriormente desconhecidos em dados são descobertos**.

Tecnicamente falando, a **mineração de dados** é um **processo que utiliza técnicas de estatística, matemáticae inteligência artificial para extrair e identificar informações úteis e subsequentes conhecimentos** (ou padrões) em grandes conjuntos d dados.

Em outro conceito, a **mineração de dados** é entendida como o **processo não trivial de identificar padrões válidos, novos, potencialmenteúteis e**, em última instância, **compreensíveis** em dados armazenados em bancos de dados estruturados.

Gabarito: **Certo**.

3- (CESPE - 2018 - Polícia Federal - Agente de Polícia Federal) Julgue o item que segue, relativo a noções de mineração de dados, big data e aprendizado de máquina.

Pode-se definir mineração de dados como o processo de identificar, em dados, padr válidos, novos, potencialmente úteis e, ao final, compreensíveis.

Resolução:

A **mineração de dados (data mining)** refere-se à **mineração ou descoberta de novas informações em termos de padrões ou regras** com base em grandes quantidades dados. Dito de outro modo, o termo **mineração de dados** foi originalmente usado para descrever o processo pelo qual os **padrões anteriormente desconhecidos em dados são descobertos**.

Tecnicamente falando, a **mineração de dados** é um **processo que utiliza técnicas de estatística, matemáticae inteligência artificial para extrair e identificar informações úteis e subsequentes conhecimentos** (ou padrões) em grandes conjuntos d dados.

Em outro conceito, a **mineração de dados** é entendida como o **processo não trivial de identificar padrões válidos, novos, potencialmenteúteis e**, em última instância, **compreensíveis** em dados armazenados em bancos de dados estruturados.

Gabarito: **Certo**.

4- (FCC - 2019 - TRF - 4ª REGIÃO - Analista Judiciário - Infraestrutura em Tecnologia da Informação) Um Tribunal pretende analisar fatos (fatores ambientais e perfis profissionais, entre outros) que esclareçam por que alguns colaboradores se destaca profissionalmente enquanto outros não se desenvolvem e acabam por se desligar do órgã Para facilitar essa análise, o Tribunal solicitou um auxílio tecnológico que indique quais características nos fatos apresentam razões positivas que justifiquem investimentos mais robustos no treinamento de colaboradores que tendem a se destacar a médio e longo praz

Para tanto, o Analista implantará um processo de análise científica preditiva com base em dados estruturados, que consiste na obtenção de padrões que expliquem e descreva tendências futuras, denominado

a) snowflake.

b) drill over.

c) star schema.

d) slice accross.

e) data mining

Resolução:

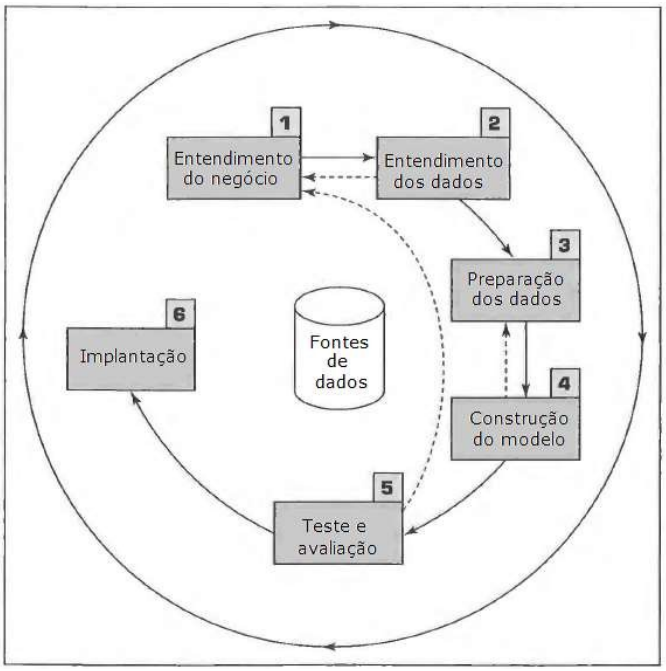
A **mineração de dados (data mining)** refere-se à **mineração ou descoberta de novas informações em termos de padrões ou regras** com base em grandes quantidades dados. Dito de outro modo, o termo **mineração de dados** foi originalmente usado para descrever o processo pelo qual os **padrões anteriormente desconhecidos em dados são descobertos**.

Gabarito: Letra E.

1.2 Modelo de Referência CRISP-DM

A realização da mineração de dados de forma sistemática geralmente segue um processo geral. Com base nas melhores práticas, pesquisadores e profissionais de mineração de dad propuseram vários processos (fluxos de trabalho ou abordagens simples passo a passo) pa maximizar as chances de sucesso na realização de projetos de mineração de dados.

O **modelo de referência CRISP-DM** é provavelmente o mais popular e foi proposto por um consórcio de empresas europeu para servir como **metodologia padrão não proprietária** para a mineração de dados. O **CRISP-DM** é o **processo para condução da mineração de dados de forma sistemática** composto por **seis etapas que vão desde um boa compreensão do negócio e da necessidade do projeto de mineração até a implantação da solução para atender a esta necessidade**. A figura a seguir apresenta as seis etapas do modelo CRISP-DM.



ATENÇÃO!!!

Vamos detalhar cada uma das seis etapas do modelo CRISP-DM, mas antes é importante fazer uma ressalva: **embora estas etapas possuam uma natureza sequencial, geralmen há uma grande quantidade de retornos às fases anteriores**. Como podemos notar na figura, por exemplo, pode haver um retorno da etapa de construção do modelo par preparação dos dados caso seja necessário.

Como a mineração de dados é conduzida com base na experiência e experimentaçã dependendo da situação do problema e do conhecimento ou experiência do analista processo pode ser bastante iterativo e demorado. Outro importante destaque é que como os últimos passos são construídos sobre o resultado dos anteriores, deve-se prestar atenç extra às etapas anteriores, a fim de não colocar todo o estudo em um caminho incorreto desde o início.

Vale ressaltar que o CRISP DM é um processo para conduzir a mineração de dados. Logo, **todas as fases podem ser consideradas como fases da mineração** segundo esse proces

Vejamos então cada uma das fases do modelo CRISP-DM:

- **Entendimento do negócio**: o elemento-chave de qualquer iniciativa de mineração de dados é saber para o que ela serve. Esta fase inicial se concentra, portanto, na **compreensão das necessidades gerenciais e dos objetivos e requisitos de negócio** que devem ser atendidos pela mineração de dados. Em seguida, um **plano de projeto para busca de conhecimentos é desenvolvido**, especificando as pessoas responsáveis pela coleta dos dados, análise os dados e reporte dos resultados. Nesta fase inicial, um orçamento para apoiar o estudo também deve ser estabelecido, pelo menos em alto nível com valores aproximados.

- **Entendimento dos dados**: etapa que objetiva **identificar os dados relevantes das diferentes fontes de dados**. A fase de entendimento dos dados começa com uma **coleta inicial** de dados e prossegue com atividades para se **familiarizar com os dados**, **identificar problemas de qualidade** de dados, **descobrir insights** sobre os dados ou **detectar subconjuntos interessantes para formar hipóteses** para informações ocultas. O analista deve ser claro e conciso sobre a descrição da tarefa de mineração de dados para que os dados mais relevantes possam ser identificados e deve construir uma compreensão adequada das fontes de dados e da variáveis existentes.

- **Preparação dos dados (ou pré-processamento de dados)**: tem como propósito **carregar os dados identificados no passo anterior e prepará-los para análise** por métodos de mineração de dados. A fase de preparação de dados abrange todas as **atividades para construir o conjunto de dados final** (dados que serão alimentados na ferramenta de modelagem) **a partir dos dados brutos iniciais**. As tarefas incluem seleção de tabelas, registros e atributos, bem como transformação e limpeza de dados para inclusão nas ferramentas de modelagem. As tarefas de preparação de dados provavelmente serão realizadas várias vezes independentemente de ordem específica. Comparado com os outros passos no CRISP-DM, **a preparação de dados consome mais tempo e esforço** (cerca de 80% do tempo total), pois os dados do mundo real são geralmente incompletos (falta de valores de atributos, falta de certos atributos de interesse ou contendo apenas dados agregados), ruidosos (contendo erros ou valores atípicos) e inconsistente (contendo discrepâncias em códigos ou nomes).

- **Construção do modelo (ou modelagem)**: nesta etapa, **várias técnicas de modelagem são selecionadas e aplicadas em um conjunto de dados já preparado** para atender às necessidades específicas do negócio. Dependendo da necessidade do negócio, a tarefa de mineração de dados pode ser de uma predição (classificação ou regressão), uma associação ou uma clusterização, cada uma dessas tarefas podendo usar uma variedade de métodos ou algoritmos. A etapa de construção de modelo também **abrange a avaliação e análise comparativa dos vários modelos construídos**, pois como não existe um melhor método ou algoritmo universalmente conhecido para uma tarefa de mineração de dados, deve-se usar uma variedade de tipos de modelos viáveis, juntamente com uma experimentação bem definida e estratégia de avaliação para identificar o "melhor" método para um determinado propósito. Mesmo para um único método ou algoritmo, é necessário calibrar uma série de parâmetros para obter melhores resultados. Alguns métodos podem ter requisitos específicos na forma como os dados devem ser formatados; assim, voltar para o passo de preparação de dados é muitas vezes necessário.

- **Teste e avaliação**: **os modelos desenvolvidos são testados e avaliados quanto à sua precisão e generalidade**. Esta etapa **avalia o grau em que o modelo selecionado (ou modelos) atende aos objetivos comerciais**, podendo inclusive testar o(s) modelo(s) desenvolvido(s) em um cenário do mundo real se o tempo e as restrições orçamentárias permitirem. A etapa de teste e avaliação é uma **tarefa crítica e desafiadora**, pois nenhum valor é adicionado pela tarefa de mineração de dados até que o valor comercial obtido a partir de padrões de conhecimento descobertos seja identificado e reconhecido.

- **Implantação**: etapa em que o **conhecimento adquirido com a exploração dos dados é organizado e apresentado de forma que o usuário possa entendê-lo e tirar benefício dele**. Dependendo dos requisitos, a fase de implantação pode ser tão simples como gerar um relatório ou tão complexo quanto implementar um processo de mineração de dados repetitivo em toda a empresa. Em muitos casos, é o cliente, e não o analista de dados, que executa as etapas de implantação. No entanto, mesmo que o analista não realize o esforço de implantação, é importante que o cliente compreenda quais ações devem ser realizadas para realmente fazer uso dos modelos criados. A etapa de implantação **também pode incluir atividades de manutenção para os modelos implantados**, pois o negócio está em constante mudança e os dados que refletem as atividades comerciais também estão mudando.

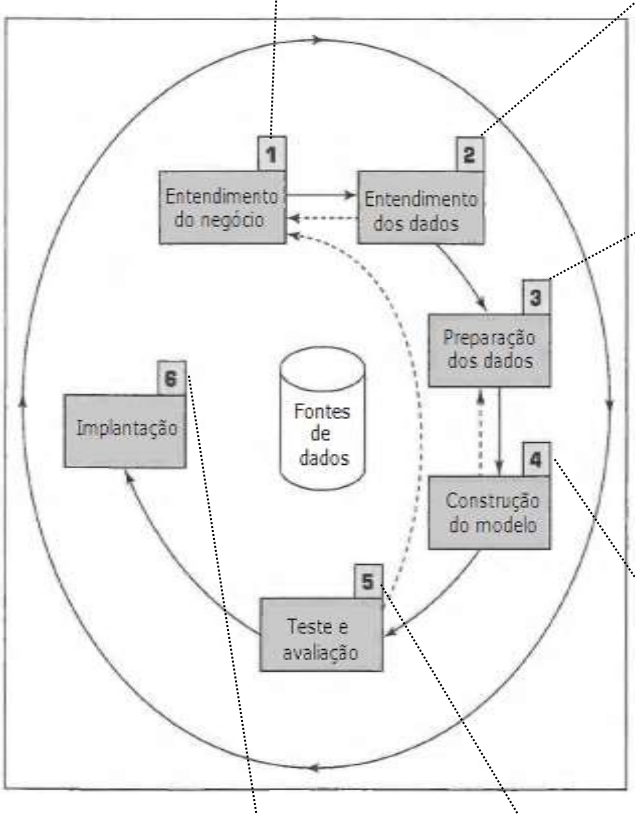
Para fixar as etapas do CRISP-DM, vamos utilizar um **esquema**!!!

❖ **Entendimento das necessidades gerenciais e das especificações de objetivos e requisitos de negócio**.

❖ Desenvolvimento de um **plano de projeto para busca de conhecimento** para a identificação das pessoas responsáveis por coletar, analisar e reportar dados. Um orçamento de alto nível pode ser estabelecido.

❖ **Identificação dos dados relevantes das diferentes fontes de dados**.

❖ Começa com a **coleta inicial** de dados e prossegue com atividades para se **familiarizar com os dados, identificar problemas de qualidade** de dados, **descobrir novos insights** sobre os dados ou **detectar subconjuntos interessantes para formar hipóteses**.



❖ **Carga dos dados e preparação para análise** por métodos de mineração de dados.

❖ Abrange as atividades para **construir o conjunto de dados final a partir dos dados brutos iniciais**.

❖ **Consome mais tempo e esforço** (~ 80% do tempo total).

❖ **Seleção e aplicação de técnicas de modelagem** para atender às necessidades específicas do negócio.

❖ Também **compreende a avaliação e comparação das análises** realizadas com os diversos modelos construídos.

❖ **Conhecimento adquirido** com a exploração dos é organizado e **apresentado de forma que o usuário possa entendê-lo e tirar benefício dele**.

❖ Também pode incluir **atividades de manutenção** para os modelos implantados.

❖ Os **modelos desenvolvidos são testados e avaliados quanto à sua precisão e generalidade**.

❖ Esta etapa **avalia o grau em que o modelo selecionado (ou modelos) atende aos objetivos comerciais**.

❖ Tarefa **crítica e desafiadora**.

Esquema 4 – CRISP-DM.

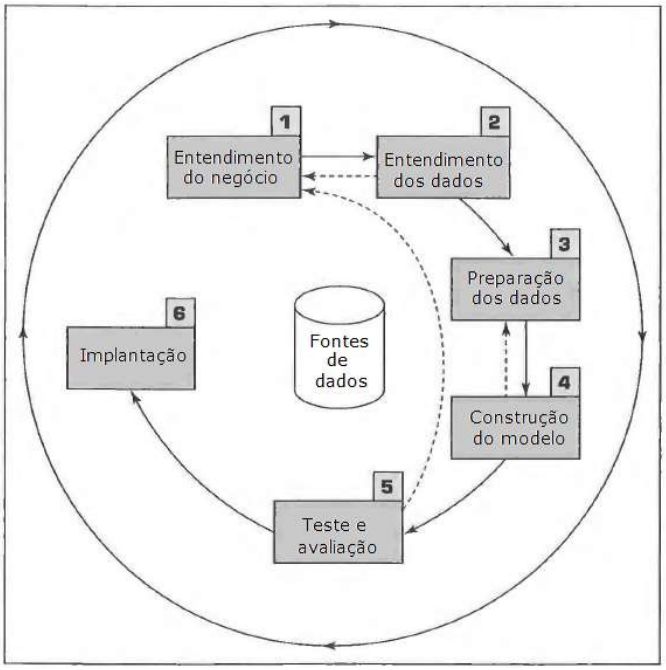
5- (CESPE / CEBRASPE - 2021 - TCE-RJ - Analista de Controle Externo - Especialidade: Controle Externo) Com relação a noções de mineração de dados e Big Data, julgue o item que se segue.

A fase de implantação do CRISP-DM (cross industry standard process for data mining) só deve ocorrer após a avaliação do modelo construído para atingir os objetivos do negócio.

Resolução:

Perfeitamente. Não se deve implementar um modelo sem antes avaliá-lo.

A figura a seguir apresenta as seis etapas do modelo CRISP-DM.



Gabarito: **Certo**.

6- (CESPE / CEBRASPE - 2021 - TCE-RJ - Analista de Controle Externo - Especialidade: Controle Externo) Com relação a noções de mineração de dados e Big Data, julgue o item que se segue.

Na primeira fase do CRISP-DM (cross industry standard process for data mining), há o entendimento dos dados para que se analise a qualidade destes.

Resolução:

O entendimento dos dados é realizado na segunda fase e não na primeira. A primeira fase de entendimento do negócio.

Gabarito: **Errado**.

7- (CESPE / CEBRASPE - 2020 - Ministério da Economia - Tecnologia da Informação - Ciência de Dados) Julgue o seguinte item, a respeito de big data.

A etapa de modelagem do modelo CRISP-DM permite a aplicação de diversas técnicas de mineração sobre os dados selecionados, conforme os formatos dos próprios dados.

Resolução:

Na etapa de **construção do modelo (ou modelagem)**, várias técnicas de modelagem são selecionadas e aplicadas em um conjunto de dados já preparado para atender às necessida específicas do negócio. Dependendo da necessidade do negócio, a tarefa de mineração de dados pode ser de uma predição (classificação ou regressão), uma associação ou um clusterização, cada uma podendo usar uma variedade de métodos ou algoritmos. Esta etap também abrange a avaliação e análise comparativa dos vários modelos construídos, pois como não existe um melhor método ou algoritmo universalmente conhecido para ur tarefa de mineração de dados, deve-se usar uma variedade de tipos de modelos viá juntamente com uma experimentação bem definida e estratégia de avaliação para identific o "melhor" método para um determinado propósito. Mesmo para um único método algoritmo, é necessário calibrar uma série de parâmetros para obter melhores resultados. Alguns métodos podem ter requisitos específicos na forma como os dados devem se formatados.

Gabarito: Certo.

8- (CESPE / CEBRASPE - 2020 - Ministério da Economia - Tecnologia da Informação - Ciência de Dados) No que se refere à mineração de dados, julgue o item a seguir.

No modelo CRISP-DM, a fase na qual se planejam todas as atividades para carga dos dados é denominada entendimento dos dados.

Resolução:

O planejamento das atividades é realizado na fase de entendimento de negócio.

A fase de entendimento dos dados visa identificar os dados relevantes das diferentes fonte de dados. A fase de entendimento dos dados começa com uma coleta inicial de dado prossegue com atividades para se familiarizar com os dados, identificar problemas d qualidade de dados, descobrir novos insights sobre os dados ou detectar subconjunt interessantes para formar hipóteses para informações ocultas. O analista deve ser claro e conciso sobre a descrição da tarefa de mineração de dados para que os dados mais relevan possam ser identificados e deve construir uma compreensão adequada das fontes de dado e das variáveis existentes.

Gabarito: **Errado**.

9- (FCC - 2018 - TCE-RS - Auditor Público Externo - Administração Pública ou de Empresas) O modelo de referência CRISP-DM tem seu ciclo de vida estruturado nas seguintes 6 fases:

a) Estruturação do Negócio, Limpeza dos Dados, Indicação das Métricas, Modelagem, Estimativa e Exportação dos Dados.

b) Otimização do Negócio, Redução dos Dados, Replicação dos Dados, Modelagem, Importação dos Dados e Backup.

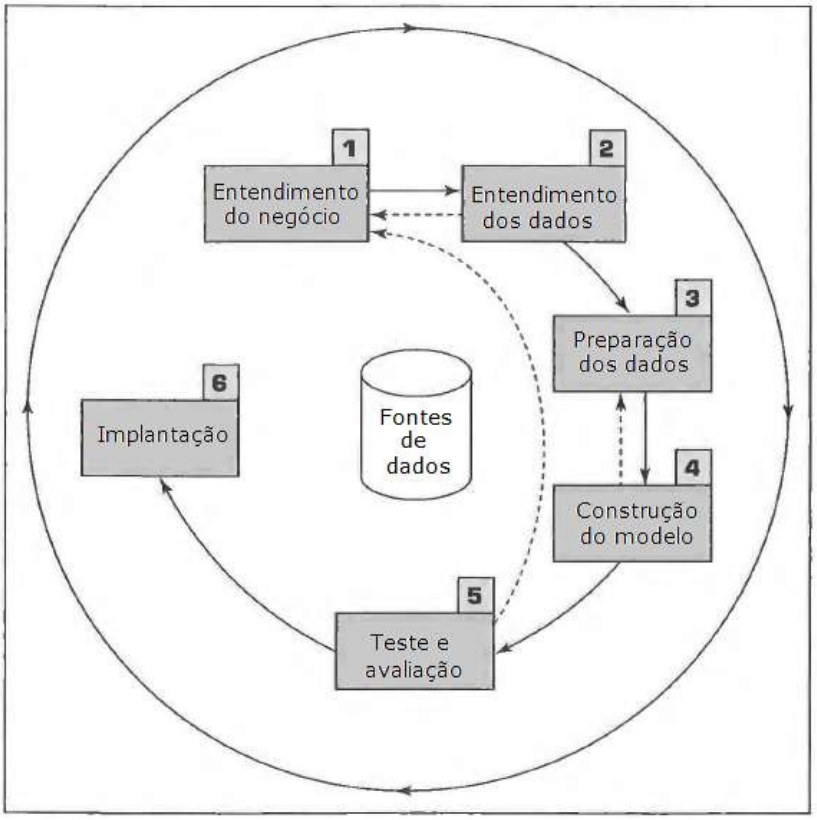
c) Entendimentodo Negócio, Entendimentodos Dados, Preparaçãodos Dados, Modelagem, Avaliação e Implantação.

d) Preparação do Negócio, Replicação dos Dados, Indexação dos Dados, Diagramação do Negócio, Estimativa e Organização.

e) Otimização do Negócio, Entendimento dos Dados, Indexação dos Dados, Exportação dos Dados, Organização e Importação dos Dados.

Resolução:

O **CRISP-DM** é o **processopara conduçãoda mineraçãode dados de forma sistemática** composto por **seis etapas que vão desde uma boa compreensão do negóci e da necessidade do projeto de mineração até a implantação da solução para atender a esta necessidade**. A figura a seguir apresenta as seis etapas do modelo CRISP-DM.



Gabarito: Letra C.

1.3 Técnicas para pré-processamento de dados

Meus caros, neste tópico detalhamentos as atividades que são realizadas antes de se aplica de fato as técnicas de mineração de dados.

Os dados disponíveis nas bases de dados existentes são altamente suscetíveis a ruíd perdas e inconsistências devido ao grande tamanho dessas bases e suas origens de múltip fontes heterogêneas. Se forem utilizados dados de baixa qualidade, os resultados da mineração serão de baixa qualidade e, portanto, estes dados precisam ser preparados ou p processados. **As técnicas de pré-processamento** buscam **melhorar a qualidade dos dados** e, consequentemente, da **eficiência e resultados da mineração**.

Diversas técnicas de pré-processamento podem ser aplicadas. Veremos nesta aula, as técnicas discutidas por Navathe e no modelo CRISP-DM. Embora os autores e modelos apresentem uma lista de técnicas diferentes, tenha em mente que estas **técnicas para pré processamento** estão voltadas para a **preparação dos dados para que estes sejam submetidos à mineração de dados**.

Técnicas de pré-processamento segundo Navathe

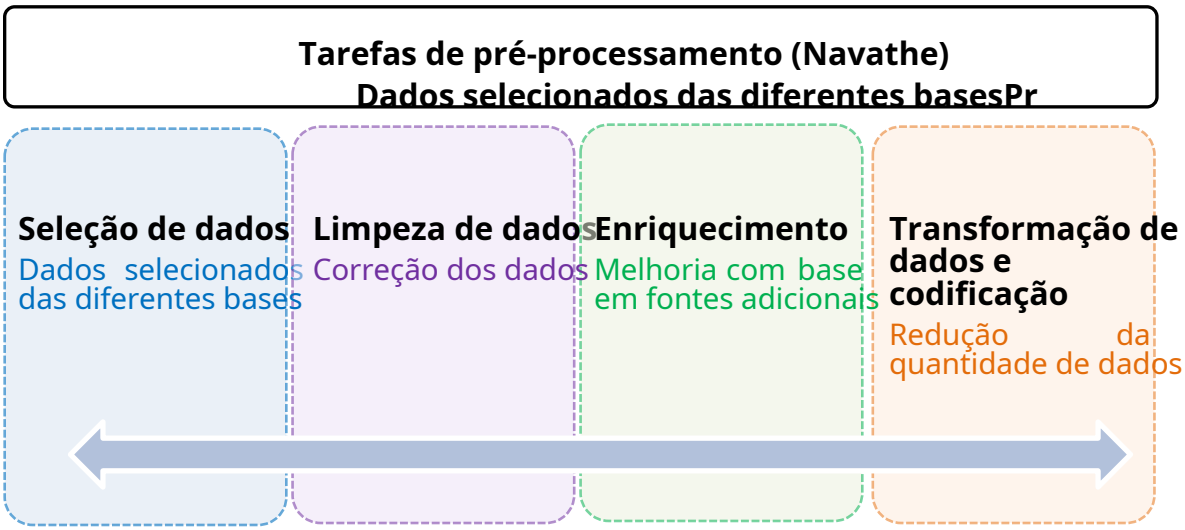
Navathe considera que a mineração de dados faz parte de um processo maior de descober de conhecimento nos bancos de dados, abreviado como **KDD (Knowledge Discovery in Databases – Descoberta de Conhecimento em Bancos de Dados)**.

O **processo de descoberta de conhecimento** compreende seis fases: seleção de dados, limpeza de dados, enriquecimento, transformação ou codificação de dados, mineração de dados e o relatório e exibição da informação descoberta. As **quatro primeiras fases são ditas de pré-processamento**, pois ocorrem anteriormente a mineração de dados propriamente dita.

Vejamos o que ocorre em cada uma das quatro fases de pré-processamento:

- **Seleção de dados**: os dados são **selecionados das diferentes bases de dados** de acordo com a necessidade do projeto de mineração.
- **Limpeza de dados**: **correção dos dados**, por exemplo, por meio da eliminação de redundâncias ou correção de códigos inválidos.
    - Se a mineração de dados for baseada em um data warehouse existente, **é possível que a limpeza já tenha sido aplicada** por meio de ETL.
- **Enriquecimento**: **melhoria dos dados com base em fontes de informações adicionais**.
- **Transformação de dados e codificação**: podem ser feitas para **reduzir a quantidade de dados**, por exemplo, por meio de agregações.

Vamos fixar as tarefas de pré-processamento com um **esquema**.



Esquema 5 – Técnicas de pré-processamento (Navathe).

Técnicas de pré-processamento segundo o CRISP-DM

O **modelo de referência CRISP-DM** dispõe da fase de preparação de dados ou também chamada de pré-processamento. Vamos ver as técnicas utilizadas nesta etapa em maiores detalhes.

A fase de **preparação de dados ou pré-processamento** tem como propósito **carregar os dados e prepará-los para análise** por métodos de mineração de dados. Esta fase abrange todas as **atividades para construir o conjunto de dados final a partir dos dados brutos iniciais**. Esta fase é composta pelas seguintes técnicas:

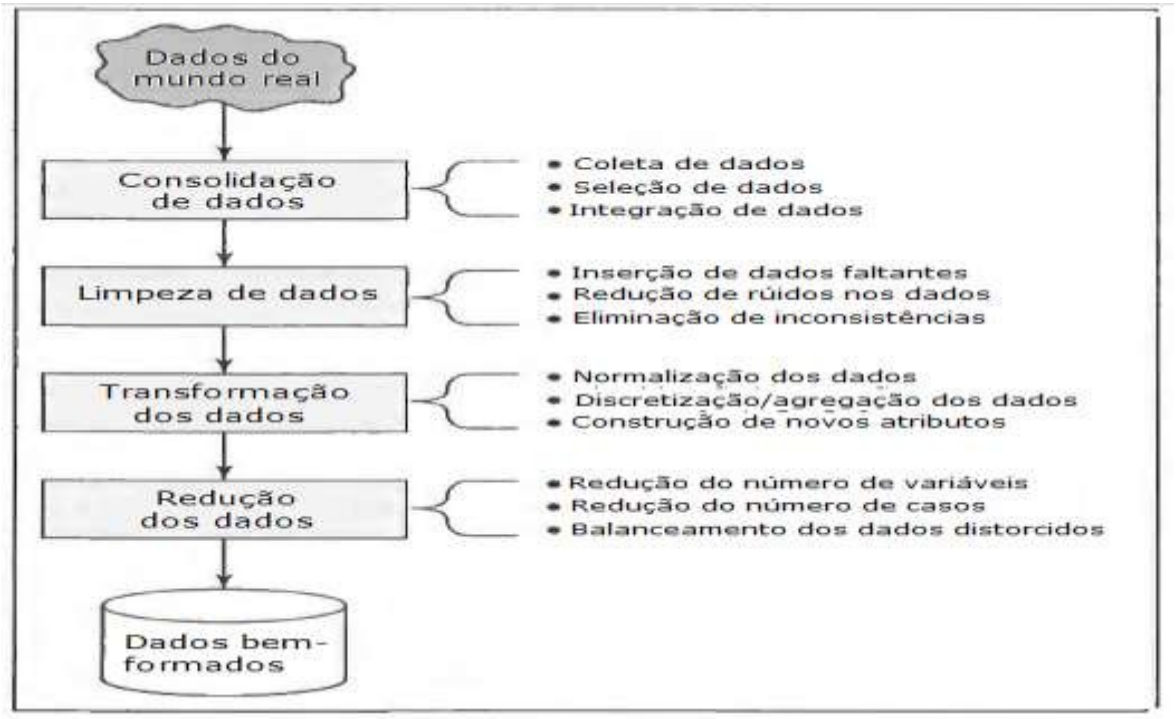
- **Consolidaçãode dados**: os **dados relevantessão coletados** das fontes identificadas, **os registros e variáveis necessários são selecionados** e os **registros provenientes de múltiplas fontes de dados são integrados**.

- **Limpeza de dados (ou depuração de dados)**: os **dados são limpos**. Em alguns casos, **os valores em falta** são uma anomalia no conjunto de dados, caso em que eles **precisam ser inseridos com o valor mais provável ou ignorados**; em outros casos, os valores em falta são uma parte natural do conjunto de dados. Nesta etapa, o analista também deve **identificar valores ruidosos nos dados (ou seja, os outliers) e suavizá-los**. Além disso, **as inconsistências** (valores incomuns dentro de uma variável) **nos dados devem ser tratadas** usando o conhecimento do domíni e/ou a opinião de especialistas.

- **Transformação de dados**: os **dados são transformados para um melhor processamento**. Em muitos casos, os **dados são normalizados entre um determinado mínimo e máximo para todas as variáveis**, a fim de mitigar o viés potencial de uma variável dominando outras variáveis com valores menores. Outra transformação que ocorre é **discretização e/ou agregação**, em que as variáveis numéricas são convertidas em valores categóricos e o intervalo de valores exclusivo de uma variável nominal é reduzido a um conjunto menor usando hierarquias conceituais para ter um conjunto de dados que seja mais acessível ao processament de computadores. Ainda assim, em outros casos, pode-se optar por **criar nova variáveis baseadas nas existentes para ampliar as informações encontradas** em uma coleção de variáveis no conjunto de dados.

- **Redução dos dados**: embora seja importante possuir todos os dados relevantes, muitos dados também são um problema. Em alguns casos, o número de variáveis pode ser bastante grande, e o analista deve **reduzir o número de variáveis para um tamanho gerenciável** (chamada **redução dimensional**, pois as variáveis são tratadas por dimensões). Em alguns casos, é necessário **reduzir o número de casos selecionando um subconjunto dos dados** para análise, desde que a amostra selecionada contenha todos os padrões relevantes do conjunto de dados completo. Além disso, é uma boa prática **equilibrar os dados altamente distorcidos** utilizando técnicas de amostragens capazes de realizar este equilíbrio.

O **esquema** a seguir sintetiza as técnicas de preparação:



Esquema 6 – Técnicas de pré-processamento (CRISP-DM).

ESCLARECENDO!!!

Qual o escopo da mineração de dados?

Engloba todo o CRISP-DM ou KDD?

Meus caros, aproveito essa seção para destacar que dependendo do modelo adotado ou d definição utilizada, a mineração de dados pode ter um escopo maior ou menor. Assim, ten se em mente que:

- **Conceito de mineração de dados:** 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. O conceito está restrito a extração de padrões com base em técnicas, mas sem destacar nenhuma metodologia.

- **CRISP-DM**: processo para condução da mineração de dados de forma sistemática composto por seis etapas que vão desde uma boa compreensão do negócio e da necessidade do projeto de mineração até a implantação da solução para atender a esta necessidade.
    - Segunda essa metodologia, **todas as seis etapas** são consideradas etapas de um processo de mineração de dados**.**

- **KDD**: O processo de descoberta de conhecimento compreende seis fases: seleção dados, limpeza de dados, enriquecimento, transformação ou codificação de dados, mineração de dados e o relatório e exibição da informação descoberta.
    - Nessa metodologia, a mineração de dados é **apenas uma das seis fases** possíve

De todo modo, sugiro que saibam os conceitos de forma independente e não tente relacioná-los, pois dependendo do autor ou do processo, serão considerados aspectos ou escopos diferentes. As questões indicam a qual linha estão se referindo e, sendo assim, voc irá conseguir resolvê-las. Geralmente as questões adotam as seguintes linhas:

- **Conceito de mineração de dados**: cobra que você conheça o conceito.
- **CRISP-DM**: cobra o conhecimento das fases da metodologia.
- **KDD**: cobra apenas que você saiba que a mineração de dados é uma de suas fases.

10- (CESPE / CEBRASPE - 2021 - TCE-RJ - Analista de Controle Externo - Especialidade: Controle Externo) Com relação a noções de mineração de dados e Big Data, julgue o item que se segue.

A descoberta de conhecimento em bases de dados, ou KDD (knowledge-discovery), é a etap principal do processo de mineração de dados.

Resolução:

A mineração de dados é que é uma das etapas do KDD. Segundo Navathe, o **processo de descoberta de conhecimento (KDD)** compreende seis fases: seleção de dados, limpeza de dados, enriquecimento, transformação ou codificação de dados, **mineração de dados** e o relatório e exibição da informação descoberta.

Gabarito: **Errado**.

11- (CESPE / CEBRASPE - 2020 - Ministério da Economia - Tecnologia da Informação - Ciência de Dados) Acerca de conceitos, premissas e aplicações de big data, julgue o item subsequente.

O objetivo das técnicas de pré-processamento de dados é preparar os dados brutos para serem analisados sem erros de incompletudes, inconsistências e ruídos.

Resolução:

Os dados disponíveis nas bases de dados existentes são altamente suscetíveis a ruíd perdas e inconsistências devido ao grande tamanho dessas bases e suas origens de múltip fontes heterogêneas. Se forem utilizados dados de baixa qualidade, os resultados da mineração serão de baixa qualidade e, portanto, estes dados precisam ser preparados ou p processados. As **técnicas de pré-processamento** buscam **melhorar a qualidade dos dados** e, consequentemente, da **eficiência e resultados da mineração**.

Gabarito: Certo.

12- (CESPE / CEBRASPE - 2020 - Ministério da Economia - Tecnologia da Informação - Ciência de Dados) No que se refere à mineração de dados, julgue o item a seguir.

O objetivo da etapa de pré-processamento é diminuir a quantidade de dados que se analisados, por meio da aplicação de filtros e de eliminadores de palavras.

Resolução:

Os dados disponíveis nas bases de dados existentes são altamente suscetíveis a ruíd perdas e inconsistências devido ao grande tamanho dessas bases e suas origens de múltip fontes heterogêneas. Se forem utilizados dados de baixa qualidade, os resultados da mineração serão de baixa qualidade e, portanto, estes dados precisam ser preparados ou p processados. As técnicas de pré-processamento buscam melhorar a qualidade dos dados e consequentemente, da eficiência e resultados da mineração.

Dentre as tarefas de pré-processamento figura a limpeza dos dados, em que há a correção dos dados, por exemplo, por meio da **eliminação de redundâncias** ou correção de código inválidos.

Gabarito: Certo.

13- (CESPE - 2014 - TJ-SE - Analista Judiciário - Banco de Dados) Julgue os próximos itens, com relação a DataMining e ETL.

O processo de transformação de dados pode exigir que dados logicamente relacionados, mas fisicamente separados, sejam recompostos, ainda que envolvam registros distintos ou até mesmo estejam em bancos de dados operacionais distintos.

Resolução:

A Transformação do Dados é a fase do KDD que antecede a fase de Data Mining. Após serem selecionados, limpos e pré-processados, os dados necessitam ser armazenados formatados adequadamente para que os algoritmos possam ser aplicados.

Em grandes corporações é comum encontrar computadores rodando diferentes sistemas operacionais e diferentes Sistemas Gerenciadores de Bancos de Dados (SGDB). Estes dados que estão dispersos devem ser agrupados em um repositório único.

Gabarito: **Certo**.

14- (FCC - 2012 - BANESE - Técnico Bancário - Informática - Desenvolvimento) Data Mining é parte de um processo maior denominado

a) Data Mart.

b) Database Marketing.

c) Knowledge Discovery in Database.

d) Business Intelligence.

e) Data Warehouse.

Resolução:

Navathe considera que a mineração de dados faz parte de um processo maior de descober de conhecimento nos bancos de dados, abreviado como KDD (Knowledge Discovery Databases).

O **processo de descoberta de conhecimento** compreende seis fases: seleção de dados, limpeza de dados, enriquecimento, transformação ou codificação de dados, **mineração de dados** e o relatório e exibição da informação descoberta. As quatro primeiras fases são dita de pré-processamento, pois ocorrem anteriormente a mineração de dados propriame dita.

Gabarito: **Letra C**.

1.4 Técnicas e tarefas de mineração de dados

A mineração de dados constrói modelos para identificar padrões entre os atributos apresentados no conjunto de dados, usando dados existentes e relevantes. Os modelos são as representações matemáticas (relações entre as variáveis) que identificam os padrões em os atributos dos objetos descritos no conjunto de dados. Alguns desses padrões são explicativos (explicando as inter-relações e afinidades entre os atributos), e outros são preditivos (prevendo os valores futuros de certos atributos).

Os padrões ou o conhecimento descoberto durante a mineração de dados podem ser descritos com base em **regras de associação, hierarquias de classificação, padrões sequenciais, padrões dentro de série temporal e agrupamento (clusterização)**.

Estas formas de descrição dos padrões são chamadas de **tarefas (ou técnicas) da mineração de dados**. Turban as classifica em três categorias principais: **predição** (inclui a classificação e regressão), **associação** (inclui a análise de relacionamentos e a análise de sequências) e **agrupamento ou clusterização** (inclui a análise de *outliers*).

Vejamos estas tarefas:

- **Predição (ou previsão)**: busca **descrever a natureza de ocorrências futuras de certos eventos com base nos acontecimentos passados**. **Difere da adivinhação**, pois leva em consideração as experiências, opiniões e outras informações relevantes na condução da previsão. Dependendo da natureza da predição, podemos falar em classificação ou regressão.
    - **Classificação (ou indução supervisionada)**: tem como objetivo **criar uma hierarquia de classes com base em um conjunto existente de eventos ou transações**. É a **tarefa mais comum de mineração de dados**. Gera-se automaticamente um modelo que pode prever o comportamento futuro a partir da análise dos dados históricos armazenados em um banco de dados. Este modelo consiste em generalizações sobre os registros, distinguindo-os com base nas **classes pré-definidas**. Ex.: uma população pode ser dividida em cinco faixas de possibilidade de crédito com base em um histórico de transações anteriores.
    - **Regressão**: é uma **aplicação especial da regra de classificação**, que ocorre quando esta **regra de classificação é uma função sobre as variáveis** mapeando essas variáveis em uma variável de classe de destino. Ex.: identificar a probabilidade de um paciente sobreviver com base em variáveis como grau de infecção ou idade.

- **Associação (ou aprendizagem de regras)**: visa **descobrir relacionamentos entre variáveis** em grandes bancos de dados. Dito de outro modo, as regras de associação correlacionam a presença de um item com uma faixa de valores para um conjunto de variáveis diverso.
    - **Análise de ligações**: a **ligação entre os diversos objetos** de interesse é descoberta automaticamente. Ex.: quando um cliente do sexo masculino compra fraldas em supermercado, geralmente ele compra cerveja.
    - **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**. Ex.: se um paciente fuma excessivamente, provavelmente sofrerá com problemas pulmonares.
    - **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**. Ex.: os casacos de frio são mais baratos no verão e mais caros no inverno.

- **Agrupamento (agregação ou clusterização)**: **partição** de uma coleção de coisas, eventos ou itens **em segmentos cujos membros são características semelhantes**. Ao contrário da classificação, no agrupamento as **classes são previamente desconhecidas**. Ex.: uma população inteira de dados de transação sobre uma doença pode ser dividida em grupos com base na similaridade dos efeitos colaterais produzidos.
    - **Análise de outliers**: identificação dos **dados que não apresentam o mesmo comportamento** padrão da maioria. Ex.: identificação de pessoa com renda muito superior aos perfis de renda em determinada organização.

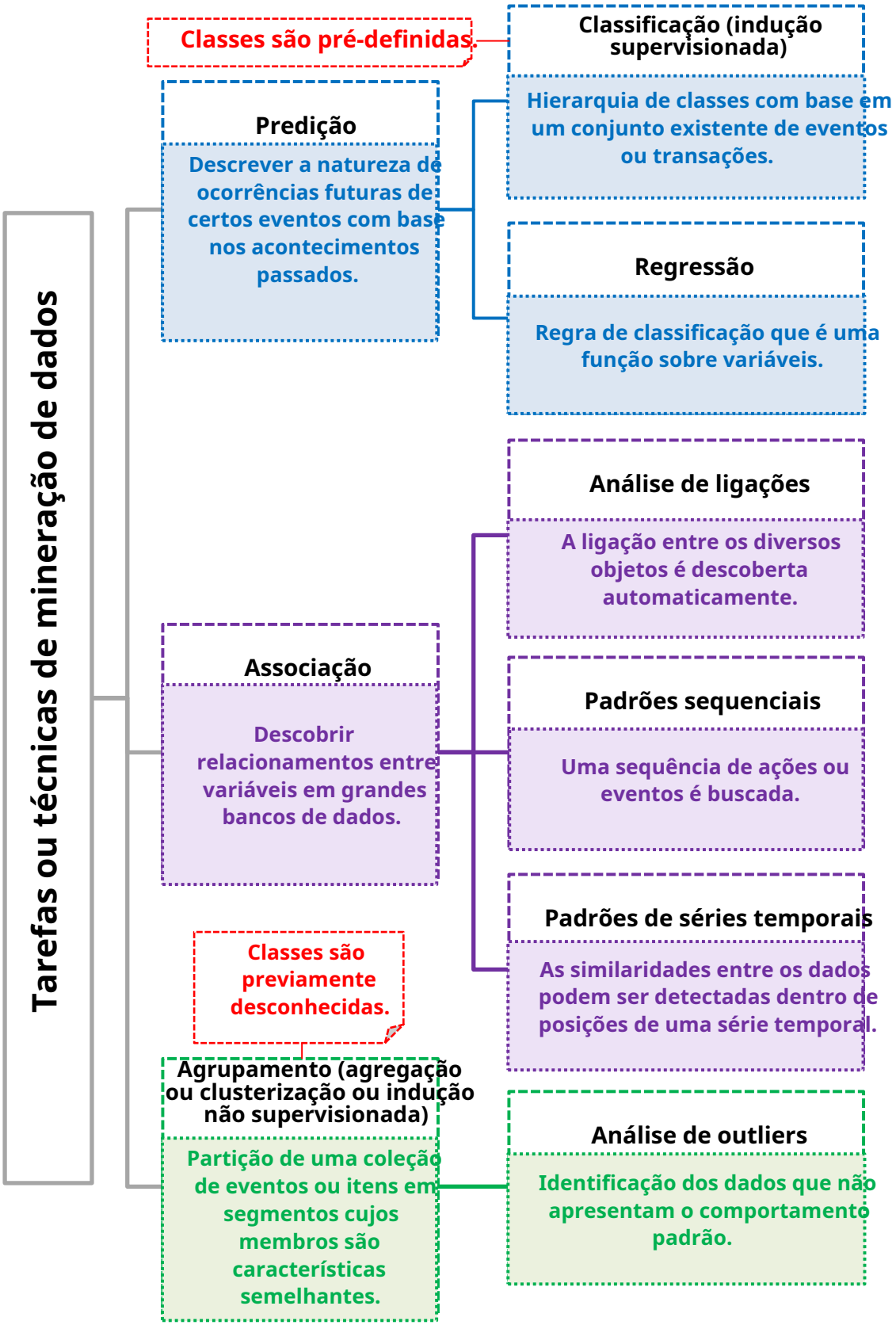
ATENÇÃO!!!

É importante ressaltar que embora a literatura indique a classificação como uma espécie de predição, muitas vezes é comum encontrarmos nas questões a classificação sendo tratada como um dos três gêneros, assim, teremos como tarefas de mineração: classificação, associação e clusterização.

DICA DO PROFESSOR!!!

Apresentamos as definições de cada uma das tarefas ou técnicas. Boa parte das questões cobra somente o entendimento da definição de cada uma destas técnicas, então fixe bem o conceito de cada uma destas tarefas.

Vamos fixar estas tarefas ou técnicas por meio de um **esquema**!!!



Esquema 7 – Técnicas ou tarefas de mineração.

15- (CESPE / CEBRASPE - 2021 - TCE-RJ - Analista de Controle Externo - Especialidade: Tecnologia da Informação) A respeito de mineração de dados, julgue o item que se segue.

No método de classificação para mineração de dados, a filiação dos objetos é obtida por m de um processo não supervisionado de aprendizado, em que somente as variáveis de entra são apresentadas para o algoritmo.

Resolução:

A classificação é um método de aprendizagem supervisionada. A aprendizagem não supervisionada é utilizada na clusterização.

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

Já na **clusterização**, o **objetivo é classificar casos** (por exemplo, pessoas, coisas, eventos) **em grupos ou clusters, de modo que o grau de associação seja forte entre os membro do mesmo cluster e fraco entre os membros de diferentes clusters**. Contudo, as **classe não são previamente definidas**, logo falamos em **aprendizado não supervisionado.**

Gabarito: **Errado**.

16- (CESPE / CEBRASPE - 2021 - CODEVASF - Analista em Desenvolvimento Regional - Tecnologia da Informação) Acerca de inteligência de negócios (business intelligence), julgue o item a seguir.

Em data mining, enquanto a classificação identifica possíveis agrupamentos dentro de uma massa de dados sem grupos predefinidos, a aglomeração reconhece modelos que identific o grupo a que um item pertence e os relaciona por meio do exame de itens já categorizado

Resolução:

Ocorreu a inversão das tarefas de mineração:

Em data mining, enquant~~o a classific~~ação **aglomeração** identifica possíveis agrupamentos dentro de uma massa de dados sem grupos prede~~finidos, a~~ aglomeração **classificaç** reconhece modelos que identificam o grupo a que um item pertence e os relaciona por me do exame de itens já categorizados.

Gabarito: **Errado**.

17- (CESPE / CEBRASPE - 2020 - Ministério da Economia - Tecnologia da Informação - Ciência de Dados) Julgue o seguinte item, a respeito de big data.

A técnica de agregação na mineração de dados atua em conjunto de registros que tenham sido previamente classificados.

Resolução:

Atenção para não confundir agregação com classificação. A técnica que utiliza classes previamente definidas é a classificação e não a agregação (agrupamentos ou clusterização).

Gabarito: **Errado**.

18- (FCC - 2020 - AL-AP - Analista Legislativo - Desenvolvedor de Banco de Dados) No contexto de data mining, considere o caso hipotético a seguir:

Uma financeira possui o histórico de seus clientes e o comportamento destes em relação ao pagamento de empréstimos contraídos previamente. Existem dois tipos de clientes: adimplentes e inadimplentes. Estas são as categorias do problema (valores do atributo alvo). Uma aplicação de mining, neste caso, consiste em descobrir uma função que mapeie corretamente os clientes, a partir de seus dados (valores dos atributos previsores), em uma destas categorias. Tal função pode ser utilizada para prever o comportamento de novos clientes que desejem contrair empréstimos junto à financeira. Esta função pode ser incorporada a um sistema de apoio à decisão que auxilie na filtragem e na concessão de empréstimos somente a clientes classificados como bons pagadores.

Trata-se de uma atividade denominada

a) sumarização.

b) descoberta de associações.

c) classificação.

d) descoberta de sequências.

e) previsão de séries temporais.

Resolução:

Como as categorias já são pré-definidas, então temos o uso da técnica da classificação.

A c**lassificação (ou indução supervisionada)** tem como objetivo **criar uma hierarquia de classes com base em um conjunto existente de eventos ou transações**. É a **tarefa mais comum de mineração de dados**. Gera-se automaticamente um modelo que pode prever o comportamento futuro partir da análise dos dados históricos armazenados em um banco de dados. Este modelo consiste em generalizações sobre os registros, distinguindo-os com base nas **classes pré-definidas**. Ex.: uma população pode ser dividida em cinco faixas de possibilidade de crédito com base em um histórico de transações anteriores.

Gabarito: Letra C.

1.4.1 Classificação

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

Quando o modelo é criado, ele pode ser usado para classificar novos dados. O primeiro pas – aprendizado do modelo – é realizado com um conjunto de treinamento de dados que já foram classificados. Cada registro nos dados de treinamento contém um atributo, chamad rótulo de classe, que indica a que classe o registro pertence.

A **classificação** é **talvez a mais comum de todas as tarefas de mineração de dados**. O objetivo da classificação é **analisar os dados históricos armazenados em um banco de dados e gerar automaticamente um modelo que pode prever o comportamento futuro**. Esse modelo induzido consiste em generalizações sobre os registros de um conjun de dados de treinamento, que ajudam a distinguir as classes predefinidas. A expectativa é que o modelo possa então ser usado para prever as classes de outros registros nã classificados e, mais importante, prever com precisão os eventos futuros reais.

EXEMPLIFICANDO!!!

Para entender melhor a classificação, imagine-se como um proprietário de um grande ban com uma infinidade de clientes correntistas. Você quer distribuir alguns cartões de crédito especiais entre estes clientes, mas quer correr o menor risco possível de crédito. Assim, nã seria útil se estes clientes estivessem separados com base no risco de "calote".

Dessa forma, os clientes do seu banco podem estar classificados em algumas categorias pr definidas:

Os clientes podem, então, ser dispostos nessas classes e, assim, você pode identifica facilmente para quem você irá "distribuir" os cartões especiais.

Perceba que como se trata de classificação, as categorias são definidas previamente para depois organizar os dados nelas.

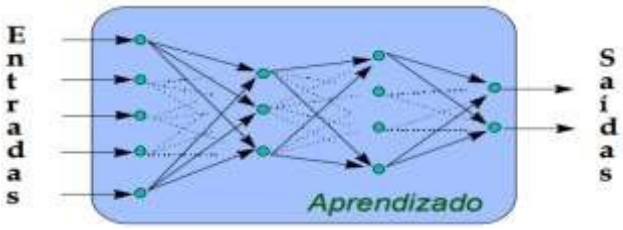
As ferramentas (por vezes chamados de algoritmos) comuns de classificação incluem redes neurais e árvores de decisão (da aprendizagem de máquina), regressão logística, métodos bayesianose análise discriminatória(das estatísticastradicionais)e ferramentas emergentes, como conjuntos aproximados, máquinas de vetores de suporte e algoritmos genéticos.

Vamos falar as duas principais: redes neurais e árvores de decisão.

As **redes neurais** envolvem o desenvolvimento de **estruturas matemáticas** (um tant parecidas com as redes neurais biológicas no 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 efetivas quando o número de variáveis envolvidas é bastante grande e as relações entre elas são complexas e imprecisas.

As **redes neurais** apresentam como principal desvantagem a **dificuldade de se interpreta as previsões feitas**. Além disso, as redes neurais tendem a **necessitar de treinamento considerável**, que demanda maior tempo à medida que aumenta a quantidade de dados.

EXEMPLIFICANDO!!!



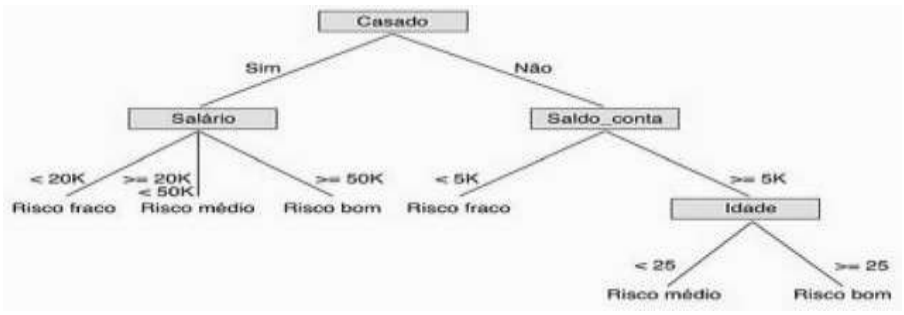
Uma rede neural utiliza uma série de entradas bem estruturadas para definir um comportamento ou realizar uma previsão.

As **árvores de decisão** **classificam os dados em um número finito de classes com base nos valores das variáveis de entrada**. As árvores de decisão são essencialmente u hierarquia de declarações se-então e, portanto, são significativamente mais rápidas do que as redes neurais. Elas são mais apropriadas para dados categorizados e intervalos de dado Portanto, incorporar variáveis contínuas em uma estrutura de árvore de decisão requer discretização; ou seja, converter variáveis numéricas de valor contínuo em intervalos categorias. A árvore de decisão auxilia no processo de **estratificação dos dados**, separand as classes com base nos valores de entrada.

A **árvore de decisão** pode ser entendida também como uma **representação gráfica da descrição de cada classe** ou, em outras palavras, uma **representação das regras d classificação**.

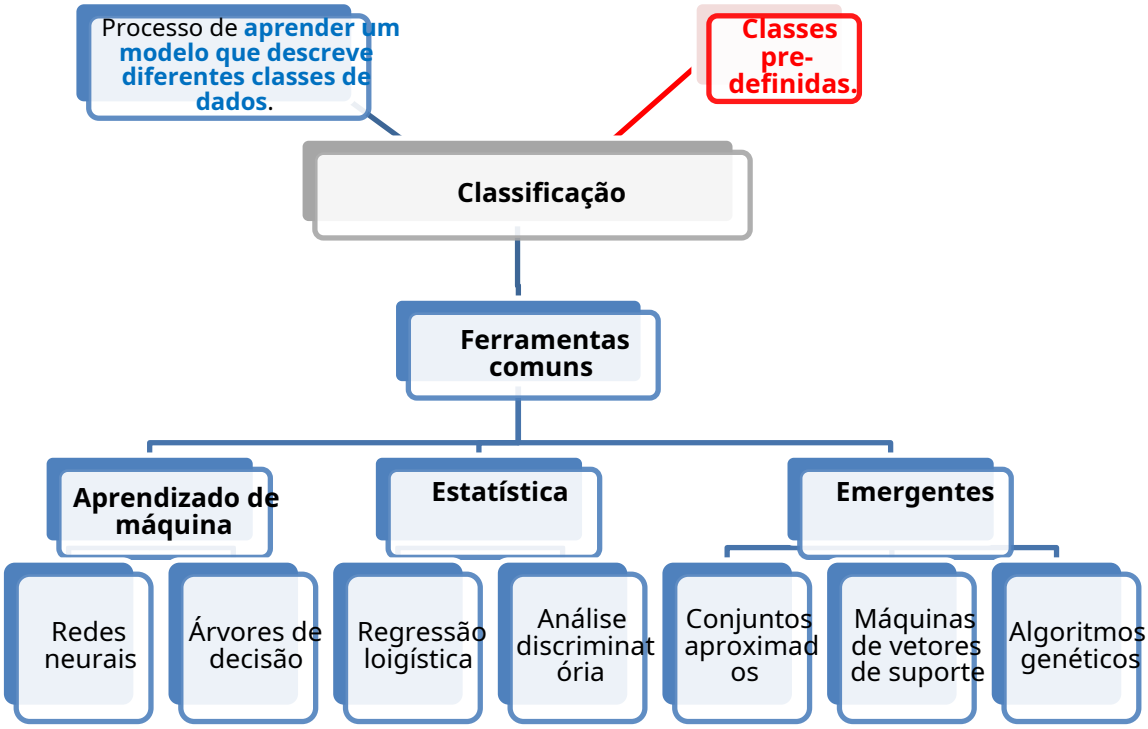
EXEMPLIFICANDO!!!

Visando entender a estrutura de uma árvore de decisão, vejamos o seguinte exemplo que visa classificar os clientes de um banco quanto ao risco para obter um cartão de crédito especial.



Neste exemplo, podemos perceber que um conjunto de regras é definido para categorizar os clientes nas classes “risco fraco”, “risco médio” e “risco bom”.

Note que ao percorrer a estrutura da árvore, saindo de sua raiz até os nós, forma as regras possíveis para uma classe. Por exemplo, se um cliente for casado e se o salário for >= 50K, então ele tem um risco bom para um cartão de crédito especial. Por outro lado, se o cliente for solteiro e o saldo de sua conta for menor que 5K, o gerente provavelmente não disponibilizar um cartão especial, pois ele possui risco fraco.



Esquema 8 – Classificação.

19- (CESPE / CEBRASPE - 2020 - Ministério da Economia - Tecnologia da Informação - Ciência de Dados) No que se refere à mineração de dados, julgue o item a seguir.

Estratificação é a abordagem da técnica de árvore de decisão que determina as regras para direcionar cada caso a uma categoria já existente.

Resolução:

As árvores de decisão classificam os dados em um número finito de classes com base nos valores das variáveis de entrada. As árvores de decisão são essencialmente uma hierarquia de declarações se-então e, portanto, são significativamente mais rápidas do que as redes neurais. Elas são mais apropriadas para dados categorizados e intervalos de dados. Portanto, incorporar variáveis contínuas em uma estrutura de árvore de decisão requer discretização; ou seja, converter variáveis numéricas de valor contínuo em intervalos categorias.

Estratificar significa separar em níveis. Logo, a árvore de decisão auxilia nesse processo, separando as classes com base nos valores de entrada.

Gabarito: Certo.

20- (CESPE - 2018 - Polícia Federal - Agente de Polícia Federal) Julgue o item que segue, relativo a noções de mineração de dados, big data e aprendizado de máquina.

Situação hipotética: Na ação de obtenção de informações por meio de aprendizado máquina, verificou-se que o processo que estava sendo realizado consistia em examinar as características de determinado objeto e atribuir-lhe uma ou mais classes; verificou-se também que os algoritmos utilizados eram embasados em algoritmos de aprendizag supervisionados.

Assertiva: Nessa situação, a ação em realização está relacionada ao processo de classificação.

Resolução:

A **aprendizagem supervisionada** ou **indução supervisionada** está relacionada à **classificação**, pois neste processo, as classes são definidas de forma prévia, antes de ser realizada a análise dos dados.

Gabarito: **Certo**.

21- (FCC - 2019 - SANASA Campinas - Analista de Tecnologia da Informação - Suporte de DBA-Banco de Dados) Considere que a SANASA busca realizar a gestão de recursos hídricos subterrâneos com base em parâmetros conhecidos que determinam poluição das águas subterrâneas. Um desses parâmetros, para exemplificar, seria o nitrato um indicador de poluição difusa de água subterrânea. Criando-se regras para realizar o aprendizado supervisionado do sistema de Data Mining utilizando-se uma certa técnica, chegar-se-á a um resultado que considera os diversos parâmetros para se descobrir se um certo aquífero tem água potável ou não, comparando-se com uma definição conhecida.

Nesse cenário, a técnica aplicada é denominada

a) Associação.

b) Classificação.

c) Clustering.

d) Regressão.

e) Prediction.

Resolução:

A questão está descrevendo uma situação de aprendizado supervisionado, logo classificaçã Outra forma de identificar é que há comparação com uma definição conhecida, ou seja, as classes são previamente definidas.

Gabarito: Letra B.

1.4.2 Regras de associação

As **regras de associação** são uma técnica popular para **descobrir relacionamentos interessantes entre variáveis** em grandes bancos de dados. Graças a tecnologias automatizadas de coleta de dados, o uso de regras de associação para descobrir os relacionamentos entre os produtos em transações de larga escala registradas nos sistemas de ponto de venda nos supermercados tornou-se uma tarefa comum de descoberta conhecimento no ramo varejista, em que é chamada de **análise de cesta de mercado**.

EXEMPLIFICANDO!!!

Qual a relação entre fralda e cerveja? Não é pegadinha e nem charada.

Uma das maiores redes de varejo dos Estados Unidos descobriu, em seu gigantesco armazém de dados, que a venda de fraldas descartáveis estava associada à de cerveja. Em geral, os compradores eram homens, que saíam à noite para comprar fraldas e aproveitava para levar algumas latinhas para casa. Os produtos foram postos lado a lado. Resultado: a venda de fraldas e cervejas disparou.

Neste caso, vemos claramente a ideia das regras de associação através do relacionamento entre duas variáveis de produtos. A descoberta de um padrão de relacionamento entre dois itens aparentemente não relacionados pode auxiliar bastante na tomada de decisões.

As derivações comuns das regras de associação são:

- **Análise de ligações**: a **ligação entre os diversos objetos** de interesse é descoberta automaticamente. Ex.: quando um cliente do sexo masculino compra fraldas em supermercado, geralmente ele compra cerveja.

- **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**. Ex.: se um paciente fuma excessivamente, provavelmente sofrerá com problemas pulmonares.

- **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**. Ex.: os casacos de frio são mais baratos no verão e mais caros no inverno.

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. Ex.: 30% das compras realizadas em um supermercado contém fraldas e cervejas.
- **Confiança ou força**: **probabilidade de que exista relação** entre itens. Ex.: 70% dos clientes que compram fraldas também compram cerveja.

Os algoritmos utilizados na mineração de regras de associação incluem o popular Apriori (onde itens de itens frequentes são identificados), PP-Growth, OneR, ZeroR e Eclat.

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 varej 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). O Apriori usa uma abordagem de baixo para cima, onde os subconjuntos frequentes são estendidos um item por vez (um método conhecido como geração de candidatos, pelo qual o tamanho dos subconjuntos frequentes aumenta subconjuntos de um item para subconjuntos de dois itens, subconjuntos de três itens, etc.) e grupos de candidatos em cada nível são testados em relação aos dados para sup mínimo. O algoritmo termina quando nenhuma outra extensão bem-sucedida é encontrad

EXEMPLIFICANDO!!!

Como um exemplo ilustrativo, considere o seguinte: uma pequena loja rastreia as transaçõ de vendas pela unidade de manutenção de estoque e, portanto, sabe quais itens normalm são comprados juntos. Cada unidade de manutenção de estoque no banco de dados transações corresponde a um produto, como "1 = manteiga", "2 = pão", "3 = água" e "4 = café". O banco de dados contendo as transações é mostrado em T1.

T1. Dados das transações

Transação	Unidade de estoque
1	1, 2, 3, 4
1	2, 3, 4
1	2, 3
1	1, 2, 4
1	1, 2, 3, 4
1	2, 4

T2. Conjuntos de itens individuais

Unidade de estoque	Suporte
1	3
2	6
3	4
4	5

T3. Conjunto de itens duplos

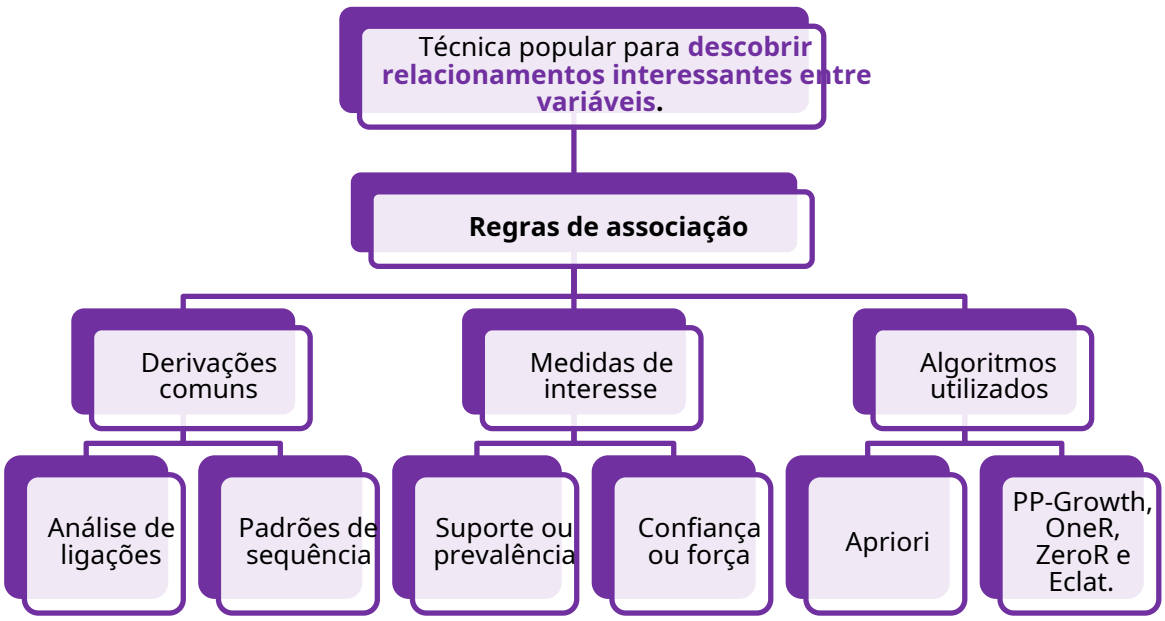
Unidade de estoque	Suporte
1, 2	3
1, 3	2
1, 4	3
2, 3	4
2, 4	5
3, 4	3

T4. Conjuntos de itens triplos

Unidade de estoque	Suporte
1, 2, 4	3
2, 3, 4	3

O primeiro passo é contar as frequências (suportes) de cada item individualmente. Neste exemplo simplificado, vamos definir o suporte mínimo para 3 (ou 50%). Como todos os conjuntos de itens individuais possuem pelo menos 3 na coluna de suporte, todos eles são considerados conjuntos de itens frequentes (T2). Se houvesse um conjunto de itens que nã fosse frequente, ele seria descartado da análise e não passaria para a análise de conjuntos duplos. Usando conjuntos de itens de um item, todos os conjuntos de dois itens são gerad e o banco de dados de transações é usado para calcular seus valores de suporte (T3). Com o conjunto de itens de dois itens {1, 3} tem um suporte menor que 3, ele não deve ser incluído nos conjuntos de itens que serão usados para gerar os conjuntos de itens próximo nível (conjuntos de itens de três itens) (T4). O algoritmo parece simples, ma apenas para pequenos conjuntos de dados. Em conjuntos de dados muito maiores, especialmente aqueles com grandes quantidades de itens presentes em pequenas quantidades e pequenas quantidades de itens presentes em grandes quantidades, a busca o cálculo se tornam um processo computacionalmente intensivo.

Vamos esquematizar as regras de associação.



Esquema 9 – Associação.

22- (CESPE / CEBRASPE - 2021 - TCE-RJ - Analista de Controle Externo - Especialidade: Tecnologia da Informação) A respeito de mineração de dados, julgue o item que se segue.

O fator de suporte e o fator de confiança são dois índices utilizados para definir o grau de certeza de uma regra de associação.

Resolução:

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. Ex.: 30% das compras realizadas em um supermercado contém fraldas e cervejas.
- **Confiança ou força**: **probabilidade de que exista relação** entre itens. Ex.: 70% dos clientes que compram fraldas também compram cerveja.

Gabarito: **Certo**.

23- (CESPE / CEBRASPE - 2021 - TCE-RJ - Analista de Controle Externo - Especialidade: Controle Externo) Com relação a noções de mineração de dados e Big Data, julgue o item que se segue.

As regras de associação adotadas em mineração de dados buscam padrões frequentes ent conjuntos de dados e podem ser úteis para caracterizar, por exemplo, hábitos de consumo de clientes: suas preferências são identificadas e em seguida associadas a outros potenciai produtos de seu interesse.

Resolução:

As **regras de associação** são uma técnica popular para **descobrir relacionamentos interessantes entre variáveis** em grandes bancos de dados. Graças a tecnologias automatizadas de coleta de dados, o uso de regras de associação para descobrir os relacionamentos entre os produtos em transações de larga escala registradas nos sistemas de ponto de venda nos supermercados tornou-se uma tarefa comum de descoberta conhecimento no ramo varejista, em que é chamada de **análise de cesta de mercado**.

Gabarito: **Certo**.

24- (CESPE / CEBRASPE - 2020 - Ministério da Economia - Tecnologia da Informação - Ciência de Dados) Julgue o seguinte item, a respeito de big data.

O objetivo da técnica de sequência de tempo é identificar a ocorrência de dois even diferentes no mesmo momento.

Resolução:

Com o uso de **padrões sequenciais (sequência de tempo)**, 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. Ex.: se um paciente fum excessivamente, provavelmente sofrerá com problemas pulmonares.

Os eventos ocorrem em momentos diferentes (e sequenciais) e não no mesmo momento.

Gabarito: Errado.

1.4.3 Análise de agrupamentos (clusterização)

A **análise de clusters (análise de agrupamentos ou análise de aglomerações ou análise de partições)** é um método de mineração de dados essencial para **classificar itens, eventos ou conceitos em agrupamentos comuns chamados de clusters**. O método é comumente usado em biologia, medicina, genética, análise de redes sociais, antropologia, arqueologia, astronomia, reconhecimento de caráter e até mesmo no desenvolvimento de sistemas de gerenciamento de informações. À medida que a mineração de dados aumentou em popularidade, as técnicas foram aplicadas aos negócios, especialmente ao marketing. A análise de cluster tem sido amplamente utilizada para detecção de fraude (fraude de cartão de crédito e de comércio eletrônico) e segmentação de mercado de clientes em sistemas de CRM contemporâneos.

A **análise de cluster** é uma ferramenta de análise exploratória de dados para resolver problemas de classificação. O **objetivo é classificar casos** (por exemplo, pessoas, coisas, eventos) **em grupos ou clusters, de modo que o grau de associação seja forte entre os membros do mesmo cluster e fraco entre os membros de diferentes clusters**. Cada cluster descreve a classe a que seus membros pertencem. No que diz respeito à mineração de dados, a importância da análise de cluster é que ela pode revelar associações e estrutura em dados que não eram anteriormente evidentes, mas são sensíveis e úteis uma vez encontradas.

As **classes não são previamente definidas**, mas muitas vezes, os algoritmos de cluster geralmente **requerem uma especificação do número de clusters** a serem encontrados. **Se este número não é conhecido previamente, ele deve ser escolhido de alguma forma**. Infelizmente, não há uma maneira ótima de calcular o número de cluster. Portanto, vários métodos heurísticos diferentes foram propostos como os critérios de informação bayesianos e akaikos.

EXEMPLIFICANDO!!!

Considere um grupo de pacientes de um hospital que tiveram um determinado medicamento receitado. Uma análise de clusters pode estabelecer determinados grupos de pacientes com reações semelhantes a estes medicamentos.

Perceba, para a análise de clusters não são definidos previamente os grupos possíveis. A partir da análise é que serão definidos os grupos com base nas semelhanças e diferenças entre as características dos pacientes.

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 individuais e os clusters são unidos** baseando-se em suas semelhanças.

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

- Os **métodos hierárquicos** tem como principal característica um algoritmo capaz de fornecer mais de um tipo de partição dos dados. Ele gera vários agrupamento possíveis, onde um cluster pode ser mesclado a outro em determinado passo algoritmo. Esses métodos não exigem que já se tenha um número inicial de clusters e são considerados **inflexíveis** uma vez que **não 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 **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:

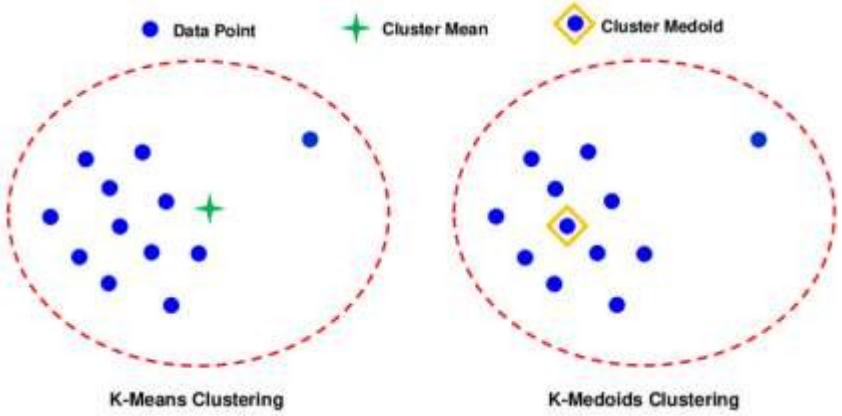
- **Métodos estatísticos**: k-means, k-modes, k-medoids, etc.
  - **K-means (k média)**: o algoritmo **atribui cada ponto de dados** (cliente, evento, objeto, etc.) ao **cluster cujo centro** (também chamado centróide) **é o mais próximo**. O **centro é calculado como a média de todos os pontos no cluster**; ou seja, suas coordenadas são a média aritmética para cada dimensão separadamente em todos os pontos do 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), usando novas medidas de similaridade para tratar com objetos categóricos, e usando um método baseado em frequência para atualizar as modas dos clusters.

  - **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). Assim, o elemento que melhor representa o cluster, é definido de acordo com seus atributos sem que haja muita influência dos valores próximos aos limites do cluster.

- **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.

  Veja de modo ilustrativo a diferença entre o k-means e o k-medoids:

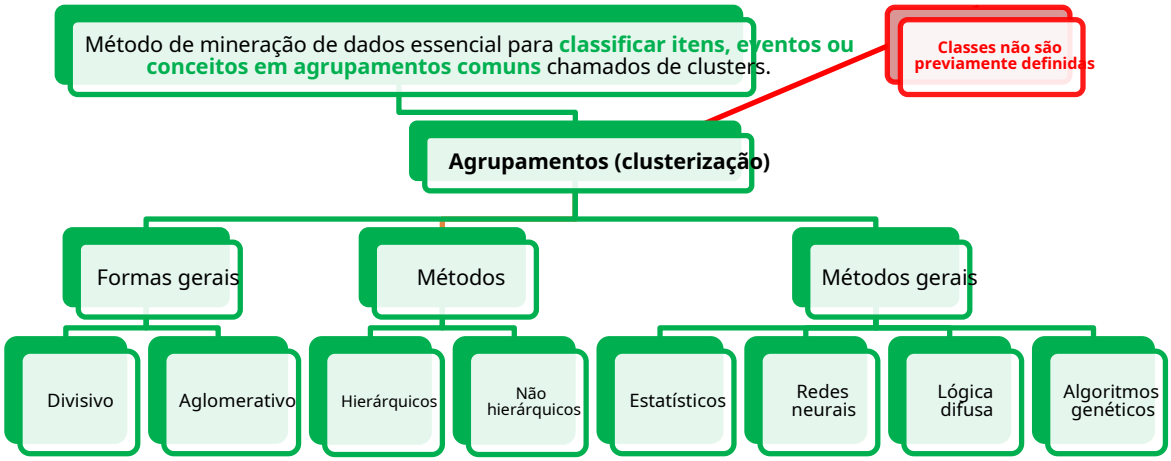


❖ **Redes neurais**: **estruturas matemáticas que têm a capacidade de aprender com experiências passadas apresentadas sob uma forma bem estruturada** dos conjuntos de dados.

❖ **Lógica difusa**: forma de **lógica multivalorada na qual os valores lógicos das variáveis podem ser qualquer número real entre 0 (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 falso.

❖ **Algoritmos genéticos**: são implementados como uma **simulação de computador em que uma população de representações abstratas de solução é selecionada em busca de soluções melhores**. A evolução geralmente se inicia a partir de um conjunto de soluções criado aleatoriamente e é realizada por meio de gerações. A cada geração, a adaptação de cada solução na população é avaliada, alguns indivídu são selecionados para a próxima geração, e recombinados ou mutados para formar uma nova população. A nova população então é utilizada como entrada para a próxima iteração do algoritmo.

E lá vem um **esqueminha** para memorizar a clusterização!!



Esquema 10 – Agrupamentos (clusterização).

25- (CESPE / CEBRASPE - 2021 - TCE-RJ - Analista de Controle Externo - Especialidade: Tecnologia da Informação) A respeito de mineração de dados, julgue o item que se segue.

No método de mineração de dados por agrupamento (clustering), são utilizados algoritmos com heurísticas para fins de descoberta de agregações naturais entre objetos.

Resolução:

Perfeitamente.

A **análise de clusters (análise de agrupamentos ou análise de aglomerações ou análise de partições)** é um método de mineração de dados essencial para **classificar itens, eventos ou conceitos em agrupamentos comuns chamados de clusters**.

O **objetivo é classificar casos** (por exemplo, pessoas, coisas, eventos) **em grupos ou clusters, de modo que o grau de associação seja forte entre os membros do mesmo cluster e fraco entre os membros de diferentes clusters**.

Gabarito: **Certo**.

26- (CESPE / CEBRASPE - 2021 - TCE-RJ - Analista de Controle Externo - Especialidade: Tecnologia da Informação) A respeito de mineração de dados, julgue o item que se segue.

Os principais métodos de análise de agrupamentos em mineração de dados incluem redes neurais, lógica difusa, métodos estatísticos e algoritmos genéticos.

Resolução:

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

- **Métodos estatísticos**: k-means, k-modes e k-medoids.

- **Redes neurais**: estruturas matemáticas que têm a capacidade de aprender com experiências passadas apresentadas sob uma forma bem estruturada dos conjuntos de dados.
- **Lógica difusa**: forma de lógica multivalorada na qual os valores lógicos das variáveis podem ser qualquer número real entre 0 (FALSO) e 1 (VERDADEIRO).
- **Algoritmos genéticos**: implementados como uma simulação de computador em qu uma população de representações abstratas de solução é selecionada em busca de soluções melhores.

Gabarito: **Certo**.

27- (CESPE / CEBRASPE - 2020 - Ministério da Economia - Tecnologia da Informação - Ciência de Dados) No que se refere à mineração de dados, julgue o item a seguir.

Na análise hierárquica de agrupamentos, é possível realocar um elemento que tenha sido alocado incorretamente no início do processo.

Resolução:

A assertiva está errada, pois na análise hierárquica de agrupamentos, não é possível realocação de elementos.

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

Os **métodos hierárquicos** tem como principal característica um algoritmo capaz de fornecer mais de um tipo de partição dos dados. Ele gera vários agrupamentos possíveis, onde um cluster pode ser mesclado a outro em determinado passo do algoritmo. Es métodos não exigem que já se tenha um número inicial de clusters e são considera **inflexíveis** uma vez que **não se pode trocar um elemento de grupo**.

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

Gabarito: Errado.

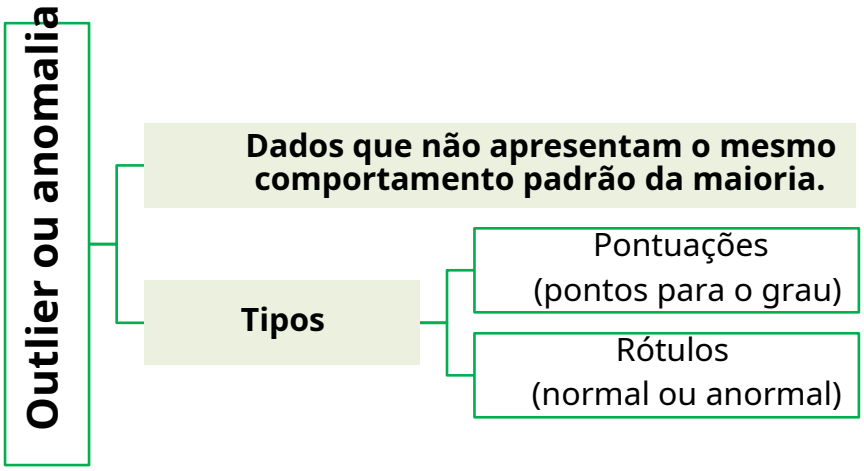
1.5 Detecção de anomalias

A **detecção de anomalias** consiste **na identificação de padrões em dados com um comportamento diferente do esperado**. Estes padrões são muitas vezes referidos como anomalias, *outliers*, exceções, aberrações, observações discordantes, entre outros, variando de acordo com o contexto.

No contexto da mineração de dados, a **análise de outliers** é uma técnica ou tarefa realizada na análise de clusters que consiste na identificação dos **dados que não apresentam mesmo comportamento** padrão da maioria. Ex.: identificação de pessoa com renda muito superior aos perfis de renda em determinada organização.

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. O analista pode optar por analisar as anomalias mais "pontuadas" ou usar um ponto de corte para as selecionar.
- **Rótulos**: os métodos usados atribuem um rótulo (normal ou anormal) para cada instância de teste.



Esquema 11 – Anomalias ou outliers.

28- (CESPE / CEBRASPE - 2020 - Ministério da Economia - Tecnologia da Informação - Ciência de Dados) Acerca de visualização e análise exploratória de dados, julgue o item seguinte.

Outlier ou anomalias são padrões nos dados que não estão de acordo com uma noção bem definida de comportamento normal.

Resolução:

Os outliers são dados que não apresentam o mesmo comportamento padrão da maioria. E identificação de pessoa com renda muito superior aos perfis de renda em determina organização.

Gabarito: Certo.

1.6 Modelagem preditiva

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

A **modelagem preditiva** utiliza de **estatísticas e modelos matemáticos para prever resultados futuros**. Basicamente, escolhe-se o melhor modelo fundamentado na probabilidade de um resultado ocorrer conforme um conjunto de dados de entrada. Esses modelos utilizam um ou mais classificadores que avaliam a probabilidade de um conjunto de dados pertencerem a outro conjunto. Assim, no nosso contexto de mineração de dados as **tarefas preditivas de classificação e regressão** são utilizadas com esta finalidade.

29- (CESPE / CEBRASPE - 2020 - Ministério da Economia - Tecnologia da Informação - Ciência de Dados) No que se refere à mineração de dados, julgue o item a seguir.

Modelagem preditiva é utilizada para antecipar comportamentos futuros, por meio do estudo da relação entre duas ou mais variáveis.

Resolução:

A predição busca descrever a natureza de ocorrências futuras de certos eventos com base nos acontecimentos passados. Difere da adivinhação pois leva em consideração as experiências, opiniões e outras informações relevantes na condução da previsão. Dependendo da natureza da predição, podemos falar em classificação ou regressão.

Gabarito: Certo.

1.7 Aprendizado de máquina

Aprendizado de Máquina (ou machine learning) é um **método de análise de dados que automatiza o desenvolvimento de modelos analíticos**. Usando algoritmos que aprendem interativamente a partir de dados, o aprendizado de máquinas permite que os computadores encontrem insights ocultos sem serem explicitamente programados pa procurar algo específico.

As tarefas e técnicas de mineração de dados estão bem relacionadas com o aprendizado d máquina, pois a **mineração de dados descobre padrões e conhecimento previamente desconhecidos** e o **aprendizado de máquina usa esses padrões e conhecimentos adquiridos**, aplicando isso a outros dados, e, em seguida, aplicando automaticamente esse resultados à tomada de decisões e ações.

O aprendizado de máquina é bastante utilizado para:

- Detecção de fraudes.
- Resultados de pesquisa na Web.
- Anúncios em tempo real em páginas da web e dispositivos móveis.
- Análise de sentimento baseada em texto.
- Pontuação de crédito e próximas melhores ofertas.
- Previsão de falhas em equipamento.
- Novos modelos de precificação.
- Detecção de invasão na rede.
- Reconhecimento de padrões e imagem.
- Filtragem de spams no e-mail.

30- (CESPE / CEBRASPE - 2020 - Ministério da Economia - Tecnologia da Informação - Ciência de Dados) Julgue o seguinte item, a respeito de big data.
Aprendizagem de máquina pode ajudar a clusterização na identificação de outliers, que são objetos completamente diferentes do padrão da amostra.

Resolução:

Os outliers são dados que não apresentam o mesmo comportamento padrão da maioria. E identificação de pessoa com renda muito superior aos perfis de renda em determina organização.

O aprendizado de Máquina (ou machine learning) é um método de análise de dados que automatiza o desenvolvimento de modelos analíticos. Usando algoritmos que aprendem interativamente a partir de dados, o aprendizado de máquinas permite que os computador encontrem insights ocultos sem serem explicitamente programados para procurar alg específico.

Outliers podem sim ser identificados com o auxílio de aprendizado de máquina.

Gabarito: Certo.

1.8 Mineração de texto

A **mineração de texto** (também conhecida como **mineração de dados de texto** ou **descobertade conhecimentoem bancos de dados textuais**) é o **processo semiautomático de extração de padrões (informações úteis e conhecimento) de grandes quantidades de fontes de dados não estruturadas**. Lembre-se de que a mineração de dados é o processo de identificação de padrões válidos, novos, potencialmer úteis e, finalmente,compreensíveisem dados armazenadosem bancos de dados estruturados, onde os dados são organizados em registros estruturados por variáveis categóricas, ordinais ou contínuas. A mineração de texto é semelhante a mineração dados, na medida em que tem o mesmo propósito e usa os mesmos processos; mas com a mineração de texto, **a entrada para o processo é uma coleção de arquivos de dados não estruturados ou semiestruturados**, como documentos do Word, arquivos PDF, trechos de texto, arquivos XML e assim por diante.

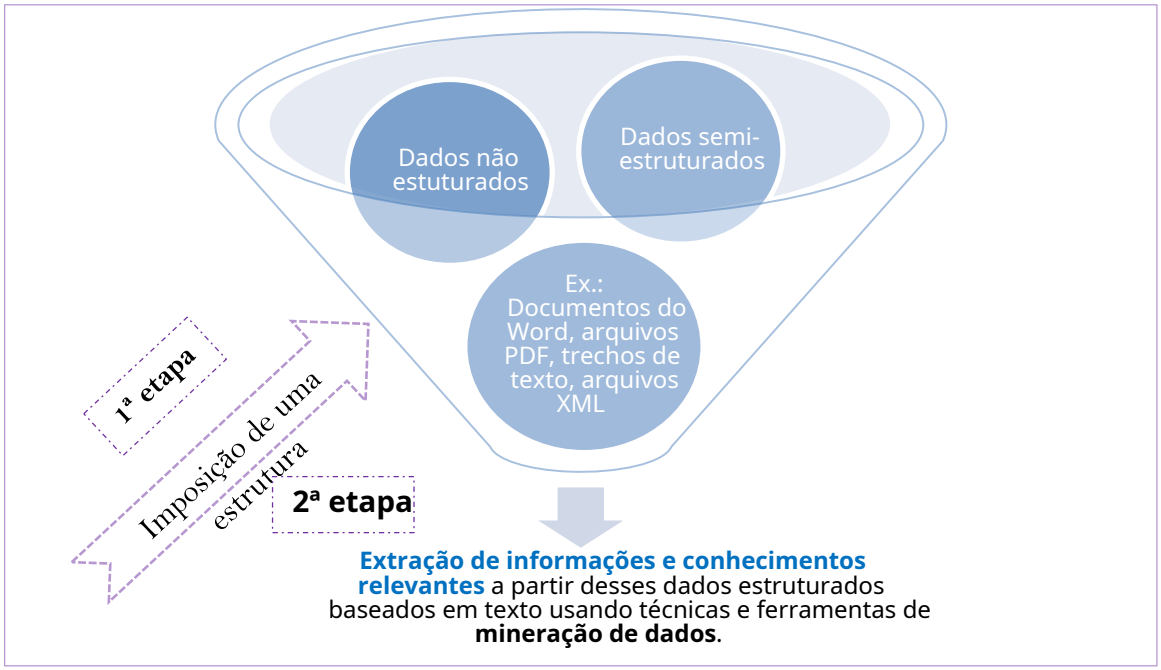
Em essência, a **mineração de texto** pode ser **pensada como um processo (com duas etapas principais)** que **começa com imposição de uma estrutura para as fontes dados baseadas em texto**, **seguindo da extração de informações e conhecimento relevantes** a partir desses dados estruturados baseados em texto usando técnicas e ferramentas de mineração de dados.

Os benefícios da mineração de texto são evidentes nas áreas em que grandes quantidades de dados textuais estão sendo gerados, como lei (ordens judiciais), pesquisa acadêm (artigos de pesquisa), finanças (relatórios trimestrais), medicamentos (sumários de alta), biologia (interaçõesmoleculares),tecnologia(arquivos de patentes)e marketing (comentários de clientes).

EXEMPLIFICANDO!!!

Por exemplo, as interações baseadas em texto de forma livre com clientes sob a forma de queixas (ou elogios) e reivindicações de garantia podem ser usadas para identificar objetivamente características de produtos e serviços que são consideradas imperfeita podem ser usadas como entrada para melhor desenvolvimento de produtos e alocações d serviços. Da mesma forma, programas de divulgação de mercado e grupos focais geram grandes quantidades de dados e, ao não restringir o feedback do produto ou do serviço à forma codificada, os clientes podem apresentar, em suas próprias palavras, o que el pensam dos produtos e serviços de uma empresa. Outra área em que o processam automatizado de textos não estruturados teve muito impacto é em comunicações eletrônic e e-mail. A mineração de texto não só pode ser usada para classificar e filtrar o e-m indesejável, mas também pode ser usada para priorizar automaticamente o email com bas no nível de importância, além de gerar respostas automáticas.

A mineração de texto pode ser sintetizada com base no seguinte **esquema**.



Esquema 12 – Mineração de texto.

31- (CESPE / CEBRASPE - 2020 - Ministério da Economia - Tecnologia da Informação - Ciência de Dados) Julgue o seguinte item, a respeito de big data.
A mineração de textos utiliza técnicas diferentes da mineração de dados, tendo em vista qu os textos representam um tipo específico de dado.

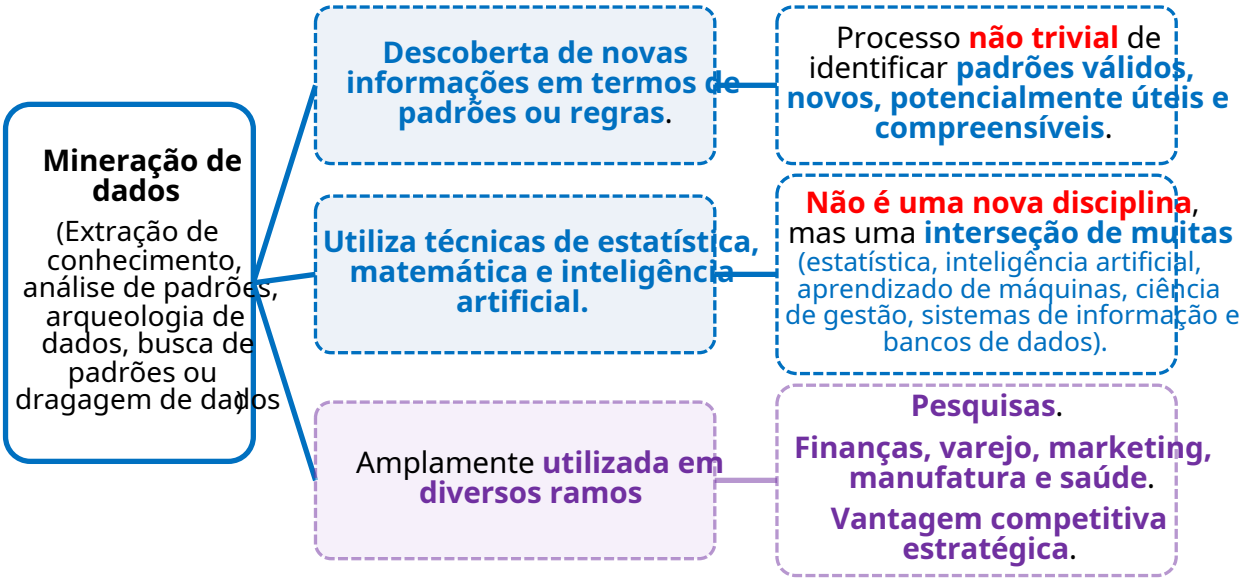
Resolução:

A mineração de texto é semelhante a mineração de dados, na medida em que tem o mesm propósito e usa os mesmos processos; mas com a mineração de texto, a entrada pa processo é uma coleção de arquivos de dados não estruturados ou semiestruturados, com documentos do Word, arquivos PDF, trechos de texto, arquivos XML e assim por diante.

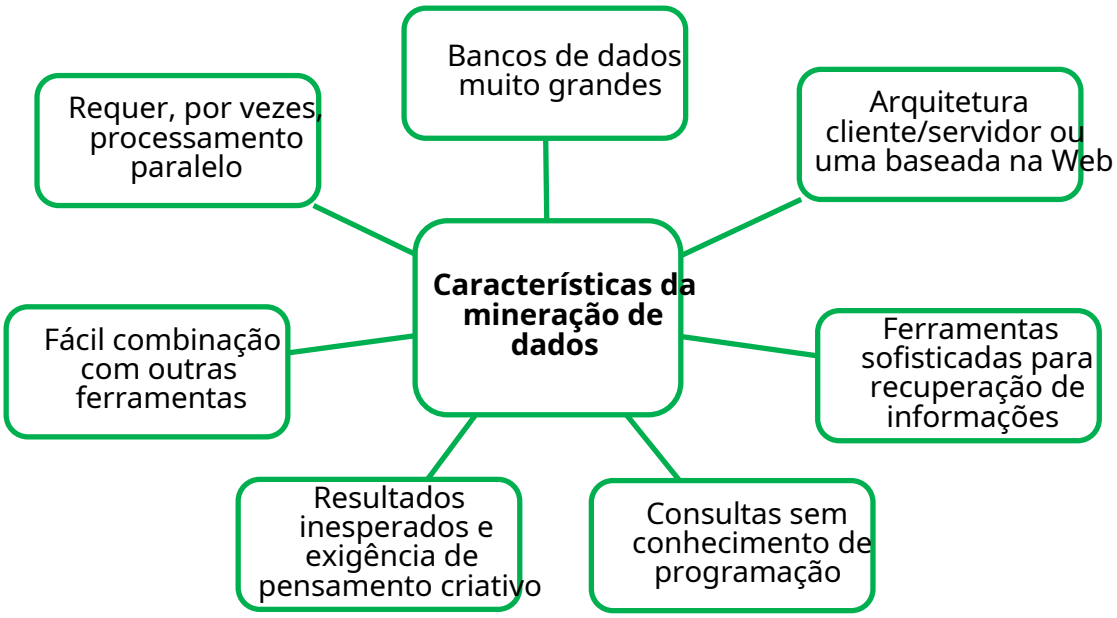
Gabarito: Errado.

2. ESQUEMAS DE AULA

Mineração de dados



Características da mineração de dados



Objetivos da mineração de dados

Objetivos finais ou aplicações da mineração de dados

Previsão	Identificação	Classificação	Otimização

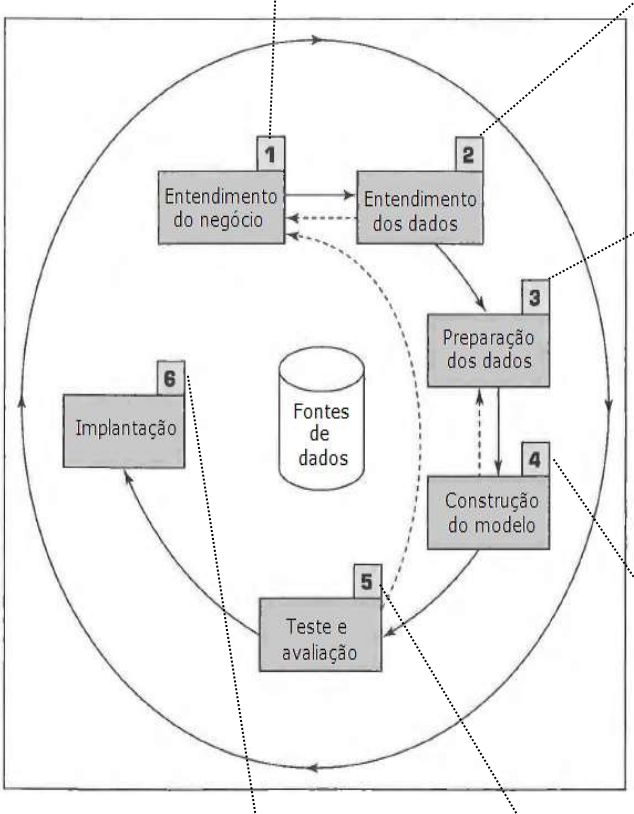
CRISP-DM

❖ **Entendimento das necessidades gerenciais e das especificações de objetivos e requisitos de negócio**.

❖ Desenvolvimento de um **plano de projeto para busca de conhecimento** para a identificação das pessoas responsáveis por coletar, analisar e reportar dados. Um orçamento de alto nível pode ser estabelecido.

❖ **Identificação dos dados relevantes das diferentes fontes de dados**.

❖ Começa com a **coleta inicial** de dados e prossegue com atividades para se **familiarizar com os dados, identificar problemas de qualidade** de dados, **descobrir novos insights** sobre os dados ou **detectar subconjuntos interessantes para formar hipóteses**.



❖ **Carga dos dados e preparação para análise** por métodos de mineração de dados.

❖ Abrange as atividades para **construir o conjunto de dados final a partir dos dados brutos iniciais**.

❖ **Consome mais tempo e esforço** (~ 80% do tempo total).

❖ **Seleção e aplicação de técnicas de modelagem** para atender às necessidades específicas do negócio.

❖ Também **compreende a avaliação e comparação das análises** realizadas com os diversos modelos construídos.

❖ **Conhecimento adquirido** com a exploração dos é organizado e **apresentado de forma que o usuário possa entendê-lo e tirar benefício dele**.

❖ Também pode incluir **atividades de manutenção** para os modelos implantados.

❖ Os **modelos desenvolvidos são testados e avaliados quanto à sua precisão e generalidade**.

❖ Esta etapa **avalia o grau em que o modelo selecionado (ou modelos) atende aos objetivos comerciais**.

❖ Tarefa **crítica e desafiadora**.

Técnicas de pré-processamento (Navathe)

Tarefas de pré-processamento (Navathe)
Dados selecionados das diferentes basesPr

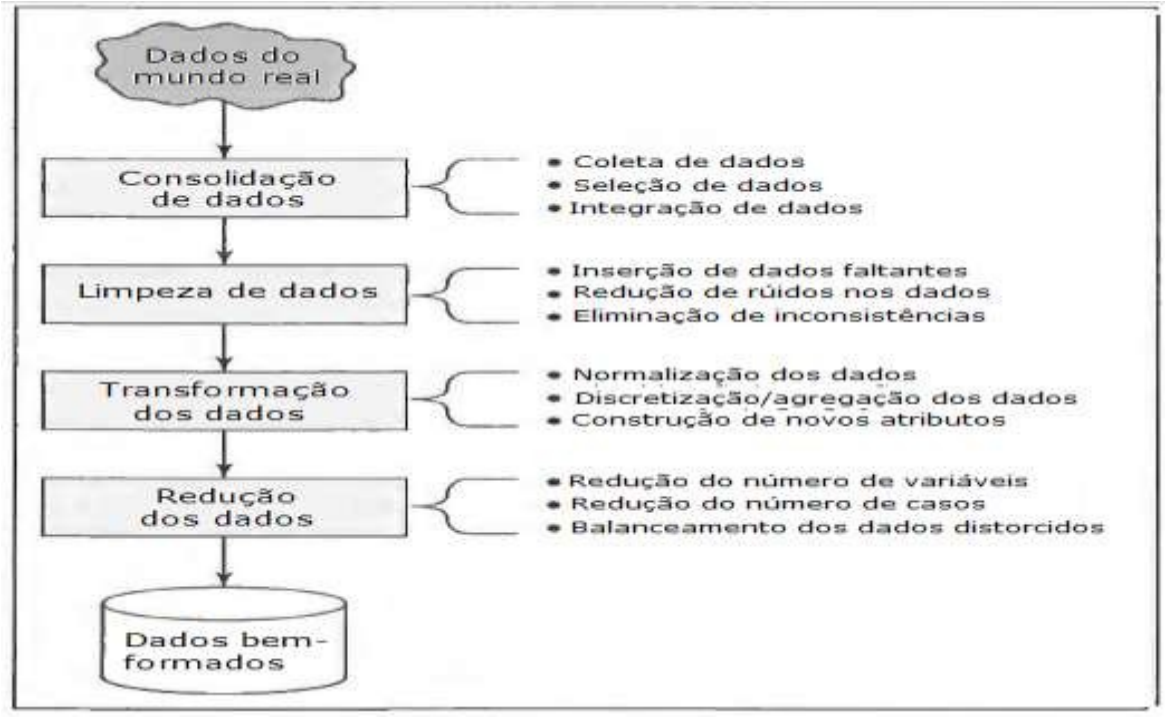
Seleção de dados
Dados selecionados das diferentes bases

Limpeza de dados
Correção dos dados

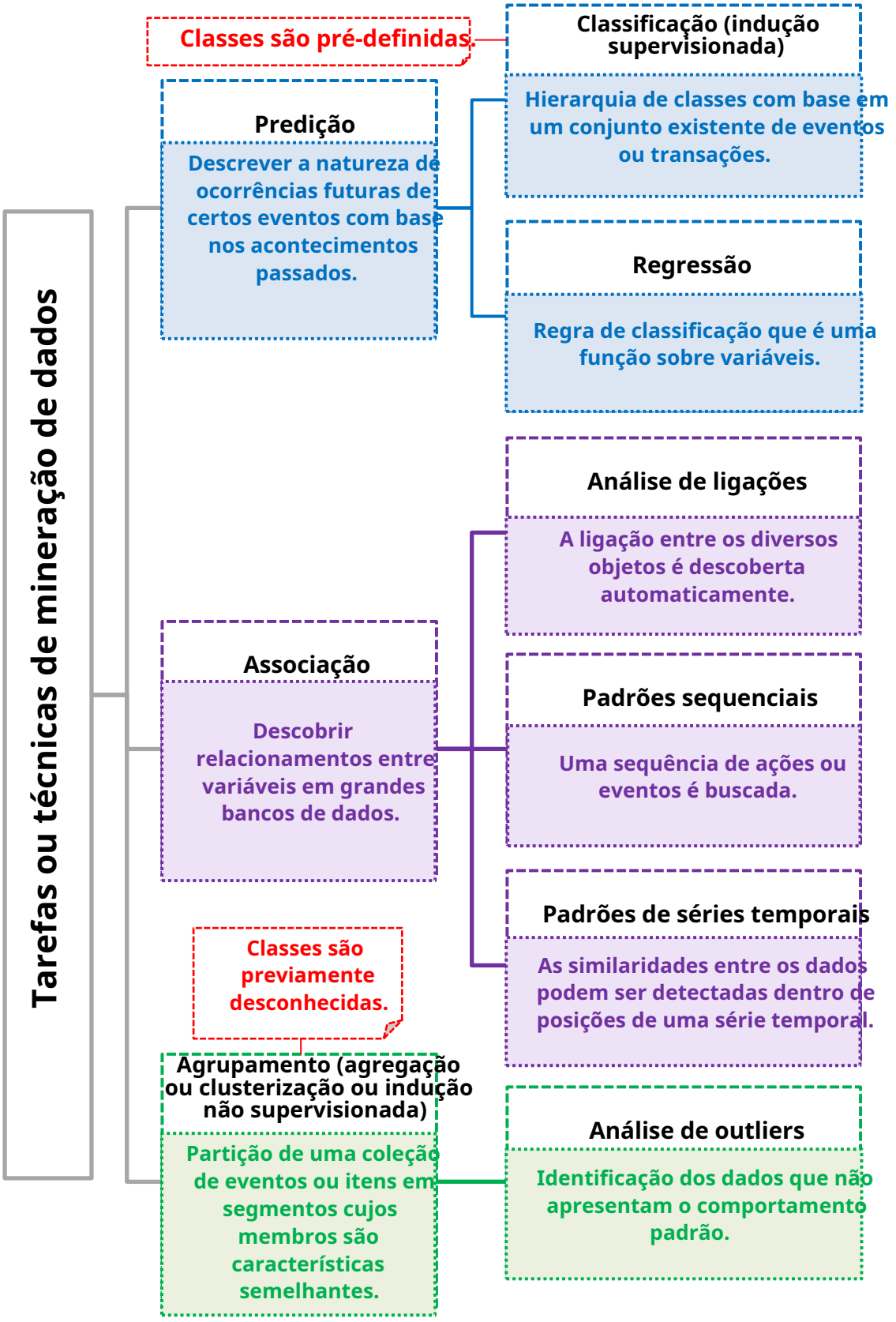
Enriquecimento
Melhoria com base em fontes adicionais

Transformação de dados e codificação
Redução da quantidade de dados

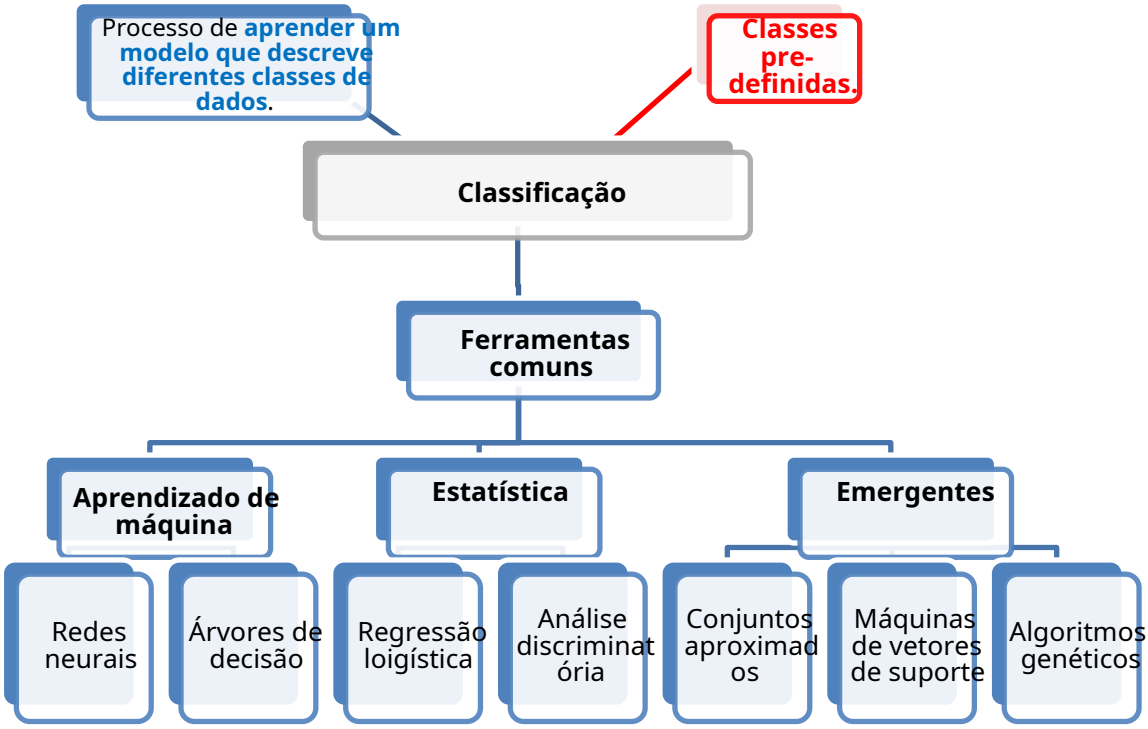
Técnicas de pré-processamento (CRISP-DM)



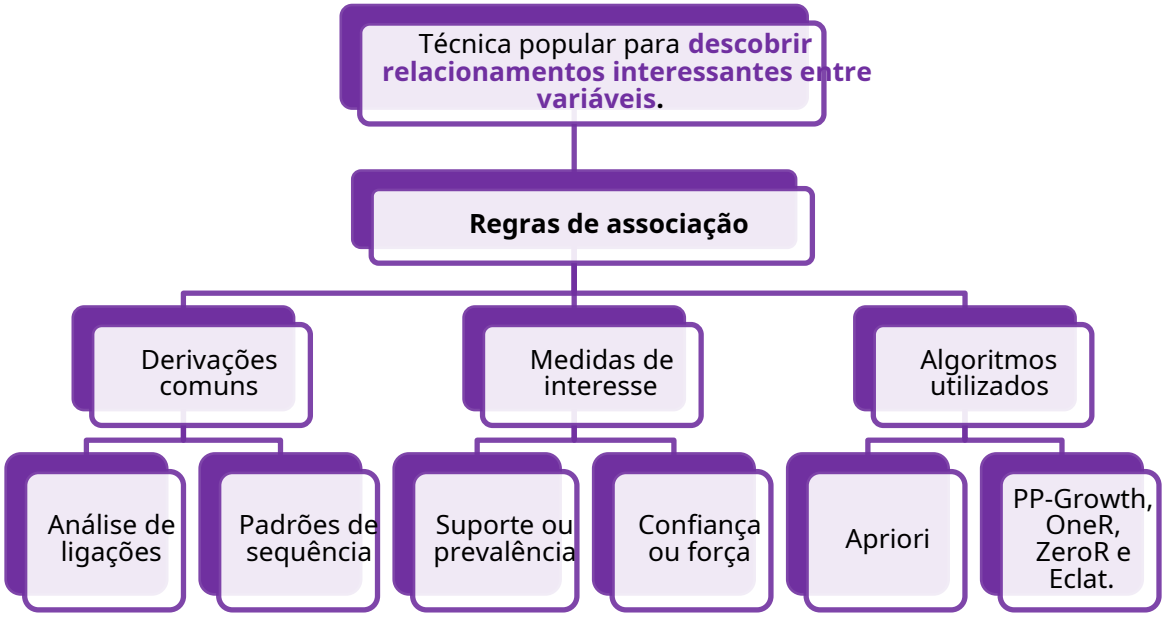
Técnicas ou tarefas de mineração



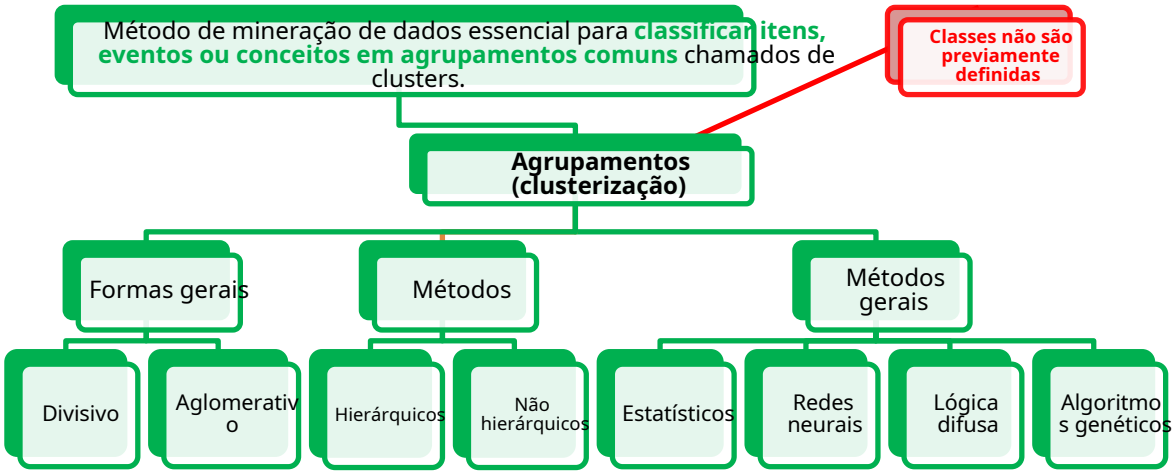
Classificação



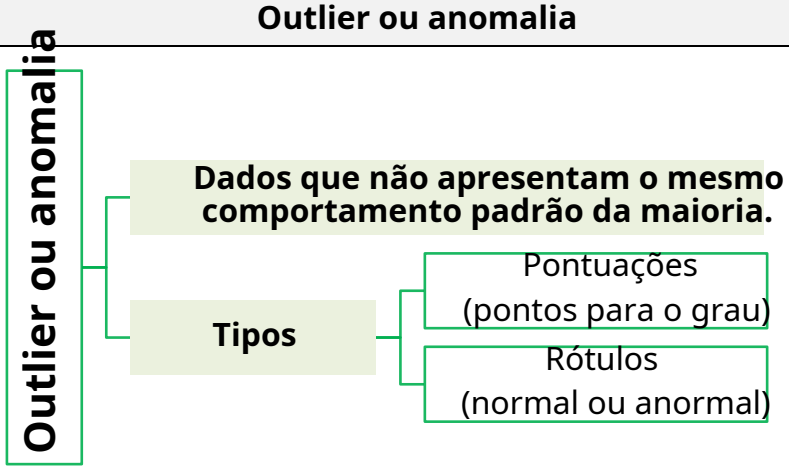
Associação



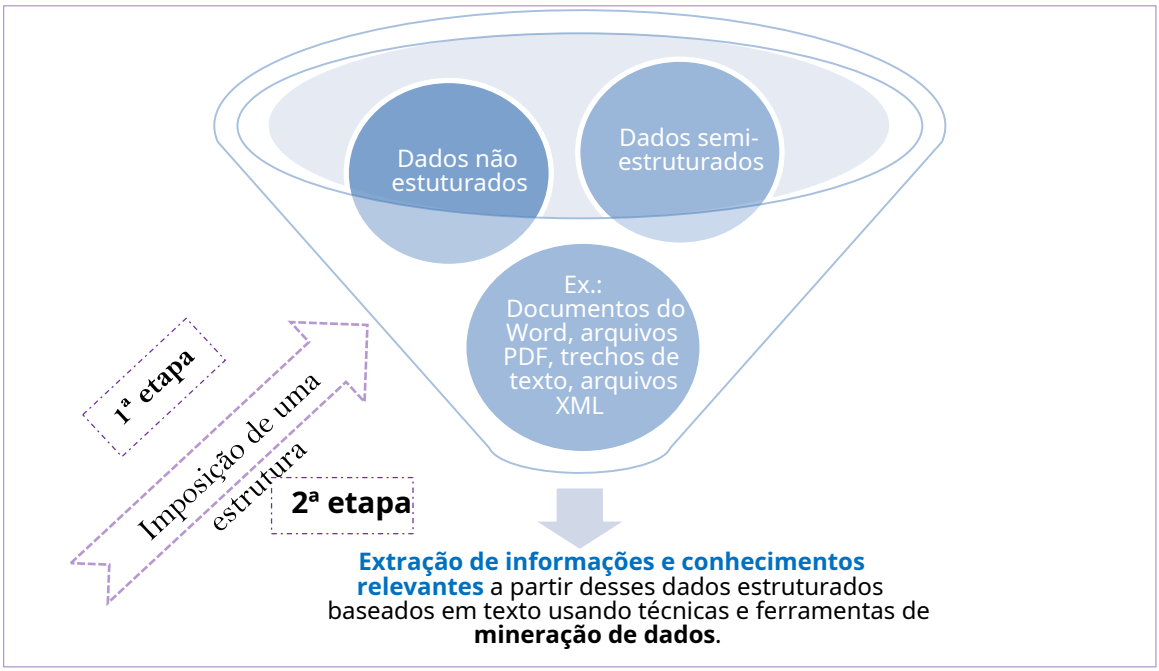
Agrupamentos (clusterização)



Outlier ou anomalia



Mineração de texto



3. REFERÊNCIAS

APRENDIS. **Detecção de anomalias**. Disponível em <http://aprendis.gim.med.up.pt/index.php/Detec%C3%A7%C3%A3o_de_anomalias>. Acesso em: 11 dez. 2017.

CHAPMAN, Pete et al. **CRISP-DM 1.0**: Step-by-step data mining guide. 2000.

ELMASRI, Ramez; NAVATHE, Shamkant B. **Sistema de Banco de Dados**. 6ed. São Paulo: Pearson Addison Wesley, 2011.

HAN, Jiawei; PEI, Jian; KAMBER, Micheline. **Data mining**: concepts and techniques. Elsevier, 2011.

IMASTERS. **Modelagem preditiva e produtos relacionados aos dados**. Disponível em <https://imasters.com.br/infra/modelagem-preditiva-e-produtos-relacionados-aos-dados/?trace=1519021197&source=single>. Acesso em: 11 dez. 2017.

SAS. **Machine Learning**: O que é e por que é importante?Disponível em <https://www.sas.com/pt_br/insights/analytics/machine-learning.html>. Acesso em: 11 dez. 2017.

TAN, Pang-Ning et al. **Introduction to data mining**. Pearson Education India, 2006.

TURBAN, Efraim et al. **Business intelligence**: A managerial approach. Upper Saddle River, NJ: Pearson Prentice Hall, 2008.